The School of Computing and Digital Technology
Faculty of Computing, Engineering and the Built Environment
Birmingham City University

Applications of Loudness Models in Audio Engineering

Dominic Ward

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy, March 2017

Abstract

This thesis investigates the application of perceptual models to areas of audio engineering, with a particular focus on music production. The goal was to establish efficient and practical tools for the measurement and control of the perceived loudness of musical sounds. Two types of loudness model were investigated: the single-band model and the multiband excitation pattern (EP) model. The heuristic single-band devices were designed to be simple but sufficiently effective for real-world application, whereas the multiband procedures were developed to give a reasonable account of a large body of psychoacoustic findings according to a functional model of the peripheral hearing system. The research addresses the extent to which current models of loudness generalise to musical instruments, and whether can they be successfully employed in music applications.

The domain-specific disparity between the two types of model was first tackled by reducing the computational load of state-of-the-art EP models to allow for fast but low-error auditory signal processing. Two elaborate hearing models were analysed and optimised using musical instruments and speech as test stimuli. It was shown that, after significantly reducing the complexity of both procedures, estimates of global loudness, such as peak loudness, as well as the intermediate auditory representations can be preserved with high accuracy. Based on the optimisations, two real-time applications were developed: a binaural loudness meter and an automatic multitrack mixer. This second system was designed to work independently of the loudness measurement procedure, and therefore supports both linear and nonlinear models. This allowed for a single mixing device to be assessed using different loudness metrics and this was demonstrated by evaluating three configurations through subjective assessment. Unexpectedly, when asked to rate both the overall quality of a mix and the degree to which instruments were equally loud, listeners preferred mixes generated using heuristic single-band models over those produced using a multiband procedure.

A series of more systematic listening tests were conducted to further investigate this finding. Subjective loudness matches of musical instruments commonly found in western popular music were collected to evaluate the performance of five published models. The results were in accord with the application-based assessment, namely that current EP procedures do not generalise well when estimating the relative loudness of musical sounds which have marked differences in spectral content. Model specific issues were identified relating to the calculation of spectral loudness summation (SLS) and the method used to determine the global-loudness percept of time-varying musical sounds; associated refinements were proposed. It was shown that a new multiband loudness model with a heuristic loudness transformation yields superior performance over existing methods. This supports the idea that a revised model of SLS is needed, and therefore that modification to this stage in existing psychoacoustic procedures is an essential step towards the goal of achieving real-world deployment.

Acknowledgements

Firstly, I must acknowledge my supervisor Cham Athwal. He has guided me throughout this long journey and always made himself available when I needed support, even at difficult times. Cham's continued encouragement, reassurance, motivation and belief in me was paramount to me completing this thesis.

I would like to express my gratitude to Joshua Reiss for his involvement in my work; his ideas and advice have been invaluable. Furthermore, a big thank you to Munevver Köküer for her attention to detail and teaching me the ways of LaTeX.

I'd like to acknowledge my colleagues at DMT Lab for fun times amidst interesting technical discussions: Ryan Stables, Matthew Cheshire, Yonghao Wang, Gregory Hough, Ian Williams, Sam Smith, Alan Dolhasz and Izzy MacLachlan. A special thanks to Sean Enderby for clarifying many technical issues and being a robust sounding board when battling concepts out with myself. I would also like to thank the Sound Engineering students at Birmingham City University, especially those who took the time to participate in my listening experiments.

Thanks to Esben Skovenborg for his patience and maintained interest, despite being attacked by my perpetual questions about regression models he fit 10 years ago. Similarly, thanks to Brecht De Man for many fruitful discussions about our research. Furthermore, I am grateful for the invaluable techniques and skills I have developed during my time at The University of Birmingham. Alan Wing, Mark Elliot, Winnie Chua and Caroline Palmer thanks for all your support. I would also like to acknowledge the following people for sharing code and clarifying implementation-level details: Harish Krishnamoorthi, Brian Moore and Brian Glasberg.

A massive thank you to my family for their continued love and support, without you I would not be writing this final page. Special thanks to Charl for reading through my work, despite telling me you didn't understand the majority of it. Finally, thanks to Frank, a nightmare of a dog but my best friend.

Contents

1 Introduction **1**

1.1 Loudness models 1

1.1.1 Applications in audio engineering 2

1.2 Problems 4

1.3 Thesis objectives 4

1.4 Thesis structure 5

1.5 Publications 6

2 The Hearing System **7**

2.1 Basics of sound 7

2.1.1 Level measurement 8

2.1.2 Spectrum and density 9

2.2 Hearing physiology 10

2.2.1 The outer ear 11

2.2.2 The middle ear 11

2.2.3 The inner ear 13

2.2.4 Modelling the auditory periphery 20

2.3 Hearing psychology 20

2.3.1 Absolute thresholds 21

2.3.2 Masking 22

2.3.3 Loudness perception 34

2.3.4 Spatial hearing 43

2.4 Summary 45

3 Loudness Models **48**

3.1 Historical overview 49

3.1.1 Multiband models for stationary sounds 49

3.1.2 Dynamic multiband models 52

3.1.3 Dynamic single-band models 53

3.2 Single-band models 55

3.2.1 Stevens' power law 55

3.2.2 Equivalent continuous sound level 56

3.2.3 ITU-R BS.1770 58

3.2.4 TC Electronic's LARM 61

3.3 Multiband models 62

3.3.1 Zwicker's model 63

3.3.2 Chalupper and Fastl's dynamic loudness model 65

3.3.3 The Cambridge loudness model . . . 66
3.3.4 Chen and Hu's dynamic loudness model . . . 76
3.3.5 Dynamic partial-loudness estimation . . . 80
3.3.6 Comparison of stationary model predictions . . . 84
3.3.7 Nonsimultaneous spectral loudness summation . . . 93
3.4 Summary . . . 102

4 Efficient Loudness Estimation 105
4.1 Model overview . . . 105
4.1.1 Previous work . . . 106
4.1.2 Profiling the model . . . 107
4.2 Optimising the excitation transformation . . . 109
4.2.1 Evaluation procedure . . . 110
4.2.2 Level per equivalent rectangular bandwidth . . . 113
4.2.3 Filter and component reduction . . . 114
4.3 Pre-cochlear filtering . . . 126
4.4 Parameterisation . . . 130
4.5 Real-time excitation-pattern binaural loudness meters . . . 133
4.5.1 Motivation . . . 133
4.5.2 Overview of the loudness meters . . . 135
4.5.3 Profiling the models . . . 138
4.5.4 Approximations and evaluation procedure . . . 139
4.5.5 Results . . . 141
4.5.6 Discussion . . . 150
4.6 Summary . . . 154

5 Loudness Driven Automatic Music Mixing 156
5.1 Related work . . . 157
5.2 This work . . . 159
5.3 Implementation . . . 159
5.3.1 The models . . . 160
5.3.2 Dynamic gate . . . 161
5.3.3 Peak normalisation . . . 162
5.3.4 Fader control . . . 162
5.3.5 Master section . . . 164
5.4 Subjective evaluation . . . 164
5.4.1 Assessment procedure . . . 165
5.4.2 Results . . . 166
5.5 Further analyses . . . 172
5.5.1 Comparison of gain settings . . . 173
5.5.2 Comparison of offline loudness model mix balances . . . 177
5.6 Discussion . . . 180
5.6.1 Empirical findings . . . 180
5.6.2 Partial loudness . . . 182
5.7 Summary . . . 185

6 Measuring the Relative Loudness of Musical Sounds 186
6.1 Measuring loudness . . . 188

6.1.1 Equal-loudness matching 189
6.1.2 Scaling procedures 191
6.2 Experiment 1: preliminary loudness assessment 192
6.2.1 Method 193
6.2.2 Results 194
6.2.3 Summary 201
6.3 Experiment 2: main loudness assessment 202
6.3.1 Method 202
6.3.2 Results 211
6.4 Discussion 219
6.5 Summary 223

7 Evaluation of Loudness Models Applied to Musical Sounds 224

7.1 The subjective reference datasets 224
7.2 The loudness models 228
7.2.1 Global-loudness descriptor 230
7.2.2 A linear multiband loudness model 230
7.2.3 Model configuration 232
7.3 Evaluation procedure 234
7.3.1 Testing the procedure 235
7.3.2 The error metrics 237
7.4 Results 239
7.4.1 Dataset $A_{16,\ 14}$ 239
7.4.2 Dataset $B_{4,\ 8}$ 242
7.4.3 Dataset $C_{12,\ 13}$ 243
7.4.4 Dataset $D_{10,\ 4}$ 245
7.4.5 Overall performance 246
7.4.6 Inter-model differences 249
7.5 Discussion 251
7.5.1 Overview 251
7.5.2 Global-loudness descriptor 252
7.5.3 General performance 253
7.5.4 Spectral effects 254
7.5.5 Related work 259
7.6 Global-loudness estimation of multitrack audio 269
7.7 Summary 272

8 Conclusions 274

8.1 Summary of research 274
8.1.1 Chapter 2: The Hearing System 274
8.1.2 Chapter 3: Loudness Models 275
8.1.3 Chapter 4: Efficient Loudness Estimation 276
8.1.4 Chapter 5: Loudness Driven Automatic Music Mixing 277
8.1.5 Chapter 6: Measuring the Relative Loudness of Musical Sounds 277
8.1.6 Chapter 7: Evaluation of Loudness Models Applied to Musical Sounds . . . 278
8.1.7 Key findings 279
8.2 Contributions 280

8.3 Critique and further work . . . 281
8.3.1 Spectral loudness summation . . . 281
8.3.2 Global-loudness estimation . . . 281
8.3.3 Loudness estimation of multitrack audio . . . 282
8.3.4 Partial loudness . . . 283
8.3.5 Intelligent Music Production . . . 283
8.3.6 Adaptive efficient loudness modelling . . . 283

A Validation of ANSI S3.4 2007 **285**

B Level change prediction using sone-to-phon mapping **289**

C Between- and within-subject confidence intervals **292**
C.0.1 Bootstrapping . . . 296

Abbreviations **298**

Bibliography **301**

Chapter 1

Introduction

Most normal-hearing people are well-versed in adjusting the level of their television or radio to achieve a comfortable listening experience. If the level of a subsequent commercial or song increases considerably, the listener will soon attenuate the reproduction to re-stabilise their sensory environment. The 'level-correction' process is subjective—how loud is comfortable for me? Though this may be the case, there is a consensus among listeners when judging the perceived strength - *the loudness* - of sound (Stevens 1955; Florentine 2011). Without this inter-listener consistency, our daily interactions would be very awkward indeed!

This thesis investigates objective methods to predict the subjective loudness of sound, such that practical tasks, including volume adjustment, can be performed automatically to a standard that falls within the general agreement of a group of listeners. In particular, this project concentrates on the modelling of loudness of musical sounds commonly found in popular music, and developing tools to aid audio engineers with the many complex processes involved in music production.

The next section introduces perceptual models for loudness estimation and their application in audio engineering and music production. Following this overview, problems are highlighted to arrive at the thesis aim and objectives, an outline of the overall thesis is given, and the publications arising from the research are listed.

1.1 Loudness models

Hearing models can be divided into two types: perceptual (or psychoacoustic) models and physiological models. The former describe and predict the relationship between the stimulus and the percept, whereas the latter draw correlations between changes in the physical properties of the stimulus and the physiological responses they invoke (Florentine 2011). The two types of model are not mutually exclusive; although perceptual models are designed to account for psychophysical data, they too draw from knowledge gained from physiological measurements of the hearing organ, but to a far lesser degree than models of hearing anatomy and physiology. From a scientific perspective, the end goal of auditory modelling is to arrive at a general theory of hearing that encapsulates pertinent physiological and psychophysical knowledge (Meddis et al. 2010). Yet, sophisticated models of the auditory system are not necessary for many practical applications.

At present, computational models of loudness fall into the first category of hearing model (Florentine 2011). This is because there is currently a limited understanding of the physiology of loudness. Thus, these procedures are founded on a large body of empirical data obtained from subjective listening tests, but, as mentioned, also consider established physiological modelling strategies for scientific support and deeper understanding. There is a great deal of variation in the level of detail of present perceptual loudness models. For example, the growth in loudness of the pure tone is adequately described by a simple power-law function of stimulus intensity. However, as more variables are added to the experimental recipe, such as the frequency of the tone, the psychological response becomes more complex, requiring a more refined model with multiple parameters. This bottom-up approach leads to procedures of considerable complexity as scientists continue to test hypotheses concerning the underlying mechanisms, uncovering a plethora of perceptual effects along the way. The most sophisticated of methods are those that characterise the three stages of auditory periphery: the outer, middle and inner ear. In short, the frequency dependency of loudness, as characterised by the equal-loudness contours, is primarily accounted for by the transmission of sound through the outer and middle ear. Then, a decomposition of the sound at the cochlea into different frequency bands—the critical bands of the hearing system—is necessary to predict the effect of spectral bandwidth on loudness. This latter phenomenon, known as spectral loudness summation (SLS), is theorised to be the result of the active filtering taking place inside the inner ear. Although primarily a predictor of loudness, the modelling of cochlear activity, or neural excitation, yields intermediate representations of sound, such as *excitation* and specific loudness (SL) patterns. Therefore, in addition to predicting various loudness phenomena, these excitation pattern (EP) loudness models can also be used to calculate auditory masking effects. This peripheral-ear modelling approach also facilitates the simulation of impaired loudness perception in individuals with reduced cochlear sensitivity.

Another type of perceptual loudness model is the frequency-weighted energy measurement. These devices are essentially physical measures of sound intensity, but include a frequency-dependent attenuation to approximate the ear's sensitivity at different frequencies. Once such device is the sound pressure level (SPL) meter, typically used for quantifying the loudness of environmental and industrial noise. Unlike the hearing system, these devices operate linearly with respect to sound level, and as such they offer different pre-emphasis filters, such as A and C weightings, which are to be manually activated by the user depending on the noise level. The A curve, for example, mimics the reduced sensitivity of the ear to low frequencies for softer sounds, whereas the C curve is comparatively broadband, designed for measuring the loudness at relatively high intensities. Because these devices do not include an elaborate filter bank, they are inherently simpler in design compared to the multiband approach, but consequently cannot predict complex psychoacoustic effects such as SLS.

1.1.1 Applications in audio engineering

Hearing models in general have served a myriad of audio applications, such as speech recognition (Cohen 1989; Ghitza 1994) and enhancement (Kleinschmidt et al. 2001; Hu and Loizou 2004), onset detection (Supper et al. 2006), audio coding (ISO/IEC 11172-3 1993) and audio quality assessment (Karjalainen 1985; Paillard et al. 1992; Thiede et al. 2000; Tan et al. 2004). Such applications incorporate computer modelled perceptual sound features to better mimic the internal representation of sound in human listeners over purely physical signal features. For example, audio coding algorithms exploit psychoacoustic principles to predict the masking of quantisation noise invoked by the input signal, such that the audio can be encoded using fewer bits. Because multiband loudness models also generate (modelled) cochlear representations of sound, these algorithms have also been used for a great deal of audio processing tasks, including onset detection (Klapuri 1999), fitting and encoding cochlear implants (McDermott et al. 2003; McDermott and Varsavsky 2009; Francart and Mcdermott 2012) and hearing aids (Moore and Glasberg 1998; Moore et al. 1999a,b; Moore 2000), speech enhancement (Shin and Kim 2007), and dynamics processing (Seefeldt 2007).

Perhaps the most elementary yet prevalent application of a loudness model is loudness normalisation or alignment, e.g. given sound A and sound B, make them equally loud. Such a scheme is often an important aspect of experiment design in order to counteract potential biasses introduced by differences in inter-stimulus loudness. It is rarely the case that physical signal metrics are accurate for loudness alignment (see Bech and Zacharov (2006)[Ch.8]). The practical importance of loudness normalisation is evident by the extensive work conducted by the Special Rapporteur Group (SRG) of the International Telecommunication Union Radiocommunication Sector (ITU-R), to address the problem of loudness fluctuations of programme material used in broadcasting, such as dialogue and music. This resulted in an internationally standardised objective loudness measure (ITU-R BS.1770 2006) which was found to work reliably well on a range of programme content (Soulodre et al. 2003; Soulodre and Norcross 2003; Skovenborg and Nielsen 2004b; Soulodre 2004; Soulodre and Lavoi 2005) and was later revised by the European Broadcast Union (EBU) for industry deployment (EBU R 128 2010; EBU Tech 3341 2011). The subsequent ITU-R BS.1770 (2015) single-band processor includes an adaptive gate to exclude quiet periods from the foreground, and can be applied to multichannel setups, e.g. surround sound, making it both robust and versatile. Despite its simplicity, the levelling paradigm has been embraced by the broadcasting and music production communities as a shield to defend against excessive dynamics compression and to realise consistent listening levels for the end listener (Camerer 2010).

One particular field of research where loudness models have received great interest is Intelligent Music Production (IMP), the goal of which is to develop intelligent systems to assist both amateur and professional audio engineers with music production. Here, the engineer must *listen*, manage and process multiple sound sources, called tracks, simultaneously, often in front of a live audience, such that the mixture is aesthetically pleasing to the listener. This technically demanding and creative task requires years of practical experience and domain-specific knowledge to execute to a high standard. Consequently, new forms of multichannel signal processors that analyse the relationships between the different tracks have been developed to assist humans with this complex time-consuming task (Reiss 2011). An example of this is the automatic fader controller, a device that sets the levels of a multitrack recording automatically by comparing inter-track sound features. Rather than use simplistic signal metrics, efficient single-band loudness models have been incorporated to better approximate the engineer's hearing sensitivity to the different instruments comprising the mix (Gonzalez and Reiss 2009; Mansbridge et al. 2012b). The majority of IMP systems, especially those capable of real-time online processing, are based on single-band loudness models, such as the BS.1770 algorithm and EBU variants (EBU Tech 3341 2016).

A few investigators have explored the use of more elaborate loudness models to better capture the perception of the mixing engineer (Aichinger 2009; Lopez 2010; Ward et al. 2012; Terrell et al. 2013). Such approaches allow for more perceptually orientated mixing objectives to be realised. For example, Terrell et al. (2013) used an established EP loudness model to realise the static level balance of a multitrack recording according to a user specified relative-loudness balance, e.g. 'set the vocal twice as loud as the rhythm section, with everything else half as loud'. Such perceptual objectives cannot be solved using single-band estimators as they do not measure loudness on an absolute scale, but instead output decibel readings generally based on measures of sound energy. Ward et al. (2012) used the same model to identify tracks affecting the audibility of a lead instrument, and used this information to re-balance the mix under masking constraints. Ford et al. (2015) developed a real-time interactive visualisation tool driven by similar measures of loudness and masking, helping mixing engineers to quickly identify perceptually conflicting instruments. Interestingly, these models have even been used to objectively evaluate the performance of automatic mixing tools that target masking minimisation using traditional time-frequency signal representations (Hafezi and Reiss 2015).

1.2 Problems

There is a clear need for models of perceived loudness, yet with different modelling strategies and domains of application, it is not yet apparent which methods are best suited to music production. At first sight, the solution appears to incorporate algorithms that have been developed and validated on broadcast material, especially considering their simplicity. But, broadcast content is *not* the primary type of sound worked upon by music producers. Indeed, Pestana and Alvaro (2012) discovered inadequacies of the BS.1770 algorithm when applied to single instruments. Similarly, multiband models show high predictive power when simulating psychophysical experiments involving stimuli synthesised in the laboratory, but music producers do not mix pure tones (fortunately). The current state-of-the-art in IMP assumes that the elaborate psychophysical models are better suited to complex musical sounds because they are grounded in auditory theory and can predict masking effects. Yet, these perceptually-motivated mixing devices have not been subjectively evaluated or compared against alternative features.

One big advantage of employing a sophisticated perceptual model is the ability to quantify masking when combining different sounds (Aichinger et al. 2011). Frequency masking is a known hurdle for novices of music production, because it requires knowledge of the spectro-temporal interactions between all elements in the mix. Furthermore, frequency masking is hard to combat without accurate use of frequency equalisation, dynamics processing and good level/image balancing. Senior (2011, p. 172) succinctly emphasises the complexities brought on by auditory masking when discussing the importance of equalisation: ‘The ramifications of frequency masking for mixing are enormous.’ In practice, the computational complexity of these multiband algorithms often restricts their application to real-world problems, especially in music production where multiple tracks play simultaneously.

To summarise, a promising yet underdeveloped area of research is the use of loudness models to quantify perceptual phenomena for use in audio engineering. In theory, procedures that include an inner-ear stage should better capture the experience of the listener because they have proved essential for describing a large body of psychophysical data. Unfortunately, the ability of such methods to generalise beyond the laboratory to the real-world, especially in music production, is relatively unknown. In addition, their high computational complexity makes them unattractive candidates for many practical applications (Krishnamoorthi 2011; Burdiel et al. 2012; Francombe et al. 2015; Wichern et al. 2015; De Man et al. 2017).

1.3 Thesis objectives

The aim of this thesis is to evaluate and extend current loudness modelling strategies for practical use in audio engineering, with a particular emphasis on music production. Not only does this research hope to contribute to improving the ear of the digital mixing assistant, but also advance scientific knowledge in loudness modelling. Accordingly, the primary objectives of this thesis are as follows:

1. To evaluate the performance of state-of-the-art loudness models when applied to musical sounds.

2. To develop flexible, computationally efficient auditory models for calculating loudness.

3. To establish new validated procedures for measuring the loudness of single instruments.
4. To develop a real-time mixing framework that accommodates both linear and nonlinear loudness estimators to perform perceptually-motivated music mixing.

1.4 Thesis structure

Chapter 2 presents an overview of sound acoustics, the peripheral hearing system, and core principles of psychoacoustics. This review provides the necessary information to understand the basic mechanisms of the ear, and therefore the architecture of modern hearing models. Empirical psychoacoustic findings are reported on to give insight into perceptual phenomena such as loudness and masking.

Chapter 3 provides a historical overview of loudness models before detailing computational methods, namely single-band and multiband procedures. Methods for analysing steady-state and time-varying sounds are discussed. In the latter case, a detailed analysis of their implementation and functionality is carried out—a necessary prerequisite for Chapter 4. The final sections compare the predictions of both standardised and recent multiband procedures to gain a deeper understanding of their practical differences when predicting the loudness of the type of sounds they were designed for. In addition, hypotheses relating the processing of the ear to the measured subjective psychological responses are investigated using different modelling techniques.

Chapter 4 presents an efficient realisation of a well-established multiband model for computing internal representations of dynamic sounds and measures of loudness. Prior work is presented, extended and evaluated using musical instruments and speech as test stimuli. Alternative low-complexity approaches to pre-cochlear filtering are given which are shown to outperform the original specification in terms of predicting absolute thresholds of pure tones. A parameterised implementation of the model is then presented and practical implications of different configurations are discussed. Finally, these techniques are used to realise fast adaptations of two different loudness models in the form of a real-time binaural loudness meter. The meter is assessed through an objective analysis of the discrepancies in the auditory features introduced by the proposed techniques.

A real-time automatic fader controller for mixing multitrack audio is presented and evaluated in Chapter 5. The system is assessed through subjective listening tests, giving a preliminary insight into the predictive nature of the multiband model first adapted in Chapter 4 when applied to musical instruments. Additional considerations such as inter-track masking and relative loudness are discussed within the context of music mixing. The results of this preliminary evaluation motivate subsequent work in this thesis.

Chapter 6 presents two loudness-matching experiments involving musical sounds for the purpose of creating a subjective dataset for evaluating loudness models. Different experimental paradigms are reviewed and the specifications of the balanced pair-matching design are examined. A statistical model is applied to the matching data in order to estimate the relative loudness levels of the stimuli. Within-subject consistency and between-subject variability are discussed and related to prior work.

Chapter 7 compares the performance of seven loudness models using the empirical data derived in Chapter 6. Three multiband models and two perceptual single-band models are compared, using the equivalent continuous sound level as a baseline predictor. A new multiband algorithm is presented and included in the assessment. Different approaches to global-loudness estimation are explored for offline and online applications. The chapter concludes by proposing enhancements to existing approaches to loudness calculation and drawing comparisons between the results obtained and those reported in prior work.

The body of work conducted in this thesis is summarised in Chapter 8 by highlighting key findings and discussing them in relation to the objectives introduced in Section 1.3. The thesis contributions are presented and directions for future work are suggested based on a critical analysis of the methods employed.

1.5 Publications

The following publications are associated with the work carried out in this thesis:

1. Ward, D., Reiss, J. D., and Athwal, C. (2012). "Multi-track mixing using a model of loudness and partial loudness". In: *Proceedings of the 133rd Audio Engineering Society Convention.*

2. Ward, D., Athwal, C., and Kokuer, M. (2013). "An efficient time-varying loudness model". In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*

3. Ward, D., Enderby, S., Athwal, C., and Reiss, J. D. (2015). "Real-time excitation based binaural loudness meters". In: *Proceedings of the 18th International International Conference on Digital Audio Effects*

4. Ward, D. and Reiss, J. D. (2016). "Loudness algorithms for automatic mixing". In: *Proceedings of the 2nd Audio Engineering Society Workshop on Intelligent Music Production*

In addition, the algorithms presented in Chapters 3, 4 and 7 have been packaged as an open-source software library for loudness and auditory modelling: https://github.com/deeuu/loudness.

Chapter 2

The Hearing System

When conversing with a friend it is necessary to adapt to our environment, perhaps by raising our voices in a populated area or when waiting for a train to pass at the station platform. In order to understand such behaviour, we need to turn to the science of hearing. An understanding of the senses establishes a theoretical framework from which hearing models can be developed for predicting and understanding sensory experience. At the core of this thesis is the application of perceptual models to enhance the listener experience by drawing on key auditory phenomena, namely loudness and masking. It is therefore the purpose of this chapter to establish a theoretical background on physiological and psychological aspects of hearing. The anatomy of the auditory system is composed of the outer, middle and inner ear, followed by the auditory nerve and central pathways, but this review concentrates on the first three peripheral components. Four major areas of psychoacoustics most relevant for this work are then discussed: absolute thresholds, the concept of the critical band and masking, loudness perception, and spatial perception.

The majority of information used to compile this chapter was extracted from the following textbooks:

- Hartmann, W. (1998). *Signals, sound and sensation.* Corrected fifth printing 2005. New York: Springer
- Moore, B. C. J. (2013). *An introduction to the psychology of hearing.* 6th ed. Leiden: Brill
- Fastl, H. and Zwicker, E. (2007). *Psychoacoustics: facts and models.* 3rd ed. Berlin Heidelberg: Springer
- Yost, W. A. (2007). *Fundamentals of hearing: an introduction.* 5th ed. Massachusetts: Academic Press, Elsevier
- Pickles, J. (2008). *An introduction to the physiology of hearing.* 3rd ed. Bradford: Emerald

For brevity, these sources are only cited in text to support statements regarding empirical data and when figures have been reused or adapted.

2.1 Basics of sound

Most objects in our environment have the ability to vibrate and therefore generate sound. The vibrating object produces a sound wave that can travel through any elastic medium that has inertia. Air is the medium for sound transmission in our natural environment. As the object vibrates, the air molecules are set in motion and a sound wave propagates towards the ear, causing the eardrum to vibrate, initiating the process of hearing. In this case, the sound is said to be audible because it is perceived by the listener. As the density of the air molecules changes with the displacement of the object, the pressure also changes about the static air pressure in a given area.

The instantaneous pressure $x(t)$ that a vibrating object exerts on an area is proportional to its velocity, meaning that sound pressure is proportional to the object's displacement.[1] The magnitude of $x(t)$ is measured in Pascals (Pa), where $1\,\mathrm{Pa}$ is equal to a $\mathrm{N\,m^{-2}}$. Power is defined as the rate at which work is done, or energy per unit time. *Sound intensity* is a measure of power for cases dealing with sound. The instantaneous intensity $I(t)$ of an acoustic signal is defined as acoustic power per unit area, with units Watts per square metre ($\mathrm{W\,m^{-2}}$). It is given by the product of the pressure signal and velocity $u(t)$ of the transmission fluid:

$$I(t) = x(t)u(t), \tag{2.1}$$

with fluid velocity related to pressure by

$$u(t) = x(t)/z. \tag{2.2}$$

In this equation, z is the specific acoustic impedance of the transmission medium, which determines the propagation of a sound wave through that medium. Acoustic impedance, given in Rayl units, is the product of the density of the medium, e.g. air, and the speed of sound in it. Instantaneous intensity is therefore equivalent to

$$I(t) = x^2(t)/z. \tag{2.3}$$

2.1.1 Level measurement

In the field of psychoacoustics, it is convenient to express sound intensity as a ratio in decibels (dB). For example, the peak intensity level quantifies the maximum instantaneous intensity of a source,

$$L_{peak} = 10\log_{10}\left(\frac{\max(I(t))}{I_0}\right) \tag{2.4}$$

where I_0 is the reference intensity of $0.964 \times 10^{-12}\mathrm{W\,m^{-2}}$, commonly selected as the absolute intensity threshold for normal-hearing humans. In practice, it is common to work with *average intensity* rather than instantaneous intensity. Average intensity is calculated by averaging the intensity over a duration of time T which must be long enough that variations in the signal do not markedly influence the measurement:

$$\bar{I} = \frac{1}{T}\int_0^T x^2(t)/z\,\mathrm{d}t. \tag{2.5}$$

For a periodic signal, such as the pure tone, an averaging time equal to the period of the waveform is sufficient. The root mean square (RMS) pressure of a signal is the value of a constant that would lead to the same average power of the signal:

$$x_{RMS}^2/z = \bar{I} = \frac{1}{T}\int_0^T x^2(t)/z\,\mathrm{d}t. \tag{2.6}$$

In the case of the pure tone, the RMS value is equal to $A/\sqrt{2}$, where A is the absolute maximum or peak of the sinusoidal waveform. In practice, the RMS pressure is commonly measured by an

[1] In digital systems, continuous time t is replaced by the sample index n of a continuous waveform sampled at a rate defined by the systems sampling frequency f_s. The sampling interval is thus $1/f_s$.

acoustic device that combines a pressure-sensitive microphone with a temporal integrator, and the measurement is then displayed as an intensity level or—equivalently—a sound pressure level (SPL) in dB:

$$\begin{aligned} L &= 10\log_{10}\left(\frac{\bar{I}}{I_0}\right) \\ &= 10\log_{10}\left(\frac{(x_{RMS}^2/z)}{(p_0^2/z)}\right) \\ &= 20\log_{10}\left(\frac{x_{RMS}}{p_0}\right), \end{aligned} \tag{2.7}$$

where the reference pressure p_0 is 20 µPa.

When measuring combinations of sounds, the basic rule is that the pressures of the constituent sounds sum linearly, e.g.

$$x_3(t) = x_1(t) + x_2(t). \tag{2.8}$$

The instantaneous intensity of the summed signal is the square of the sum,

$$\begin{aligned} I_3(t) &= x_3^2(t)/z \\ &= (x_1(t) + x_2(t))^2/z \\ &= (x_1^2(t) + x_2^2(t) + 2x_1(t)x_2(t))/z \\ &= I_1(t) + I_2(t) + (2x_1(t)x_2(t))/z. \end{aligned} \tag{2.9}$$

The average intensity of the sum $\bar{I}_3$ is therefore equal to sum of $\bar{I}_1$ and $\bar{I}_2$ plus the average intensity of the cross-product term $2x_1(t)x_2(t)$. This cross product makes $\bar{I}_3$ dependent on the relationship of the two input signals. For example, if they are identical, then $\bar{I}_3 = 4\bar{I}_1$, i.e. if the amplitude of an acoustic signal is doubled, its intensity increases by a factor of four. In contrast, when the two signals are completely uncorrelated, the integral of the cross product is zero, and consequently $\bar{I}_3 = \bar{I}_1 + \bar{I}_2$. In short, when two signals are added together, the average intensity of the sum could be anything between zero (correlation = -1) and $4\bar{I}_1$ (correlation = 1). For two separate sound sources, one usually assumes that the two intensities simply sum.

2.1.2 Spectrum and density

The spectral representation of a sound made up of discrete spectral components, i.e. a periodic signal, describes the amplitude and phase of those individual components. Thus, the intensity level of each component and indeed the overall signal level can be determined from the discrete amplitude spectrum. For continuous spectra such as noise, there are no discrete frequency components and so the concept of the spectral density is used. The spectral density describes the amount of a specified quantity, such as electrical power or acoustic intensity, in a band that is 1 Hz wide. The intensity spectrum I_s is a density describing the distribution of intensity over frequency. Therefore, the total intensity $\bar{I}_b$ in a band between a lower frequency f_l and an upper frequency f_u is calculated through integration:

$$\bar{I}_b = \int_{f_l}^{f_u} I_s(f)\,\mathrm{d}f. \tag{2.10}$$

For convenience, the intensity density is often characterised by its intensity spectrum level, defined as

$$L_s = 10\log_{10}(I_s/I_0), \tag{2.11}$$

where I_0 is the reference intensity of $\approx 10^{-12}\ \mathrm{W\,m^{-2}}$ in a 1 Hz wide band.

If the distribution of noise, over a given frequency interval of width B, is uniform (white), the intensity density is constant and so the total sound intensity in that band is given by (using Equation 2.10)

$$\begin{aligned} \bar{I}_b &= I_s(f_u - f_l) \qquad (2.12) \\ &= I_s B. \qquad (2.13) \end{aligned}$$

Rearranging for the intensity density gives

$$I_s = \frac{\bar{I}_b}{B}, \qquad (2.14)$$

which means that the spectrum level L_s of a white noise band can be calculated using

$$\begin{aligned} L_s &= 10\log_{10}\left(\frac{\bar{I}_b}{I_0 B}\right) \\ &= L_b - 10\log_{10}(B), \end{aligned} \qquad (2.15)$$

where L_b is the overall intensity level of the sound within the band.

2.2 Hearing physiology

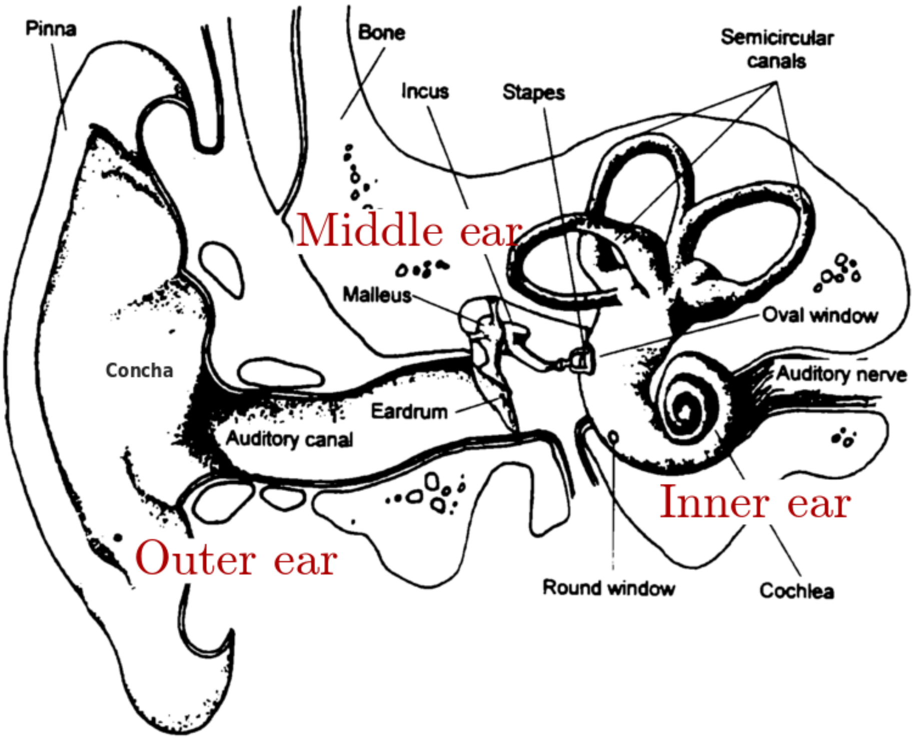


Figure 2.1: Structure of the outer, middle and inner ear. Adapted from Moore (2013).

Figure 2.1 depicts the peripheral auditory system comprised of the outer, middle and inner ear. The outer ear is primarily responsible for capturing acoustic energy as well as sound localisation. The middle ear establishes a bridge between the outer and inner ear by correcting for the mismatch in impedance between the air outside the cochlea and the liquid inside it. The inner ear encompasses the cochlea which translates incoming mechanical vibrations into neural responses in the 8th cranial nerve. Details of these three components are discussed in the following sections.

2.2.1 The outer ear

The pinna (auricle) and auditory canal (meatus) make up the outer ear. The pinna is the visible part of the ear outside of the head and is made up of different structural components. In addition to protecting the middle and inner ear, the pinna collects sound waves over a large area and directs them to the eardrum. This, along with the resonances created by the cavities of the external ear, creates a frequency-dependent gain in pressure at the eardrum, increasing the energy transfer to the middle ear. The ear canal, approximately 2.3 cm in length in humans, is often modelled as a quarter-wavelength resonator with one open end and one closed end. The concha extends the effective length of the tube, leading to a major resonance at 2.5 kHz with a peak gain between 15 and 20 dB. A second peak at 5.5 kHz is present due to the resonance of the concha. The complex structure of the external ear yields a relatively steady increase in pressure over 2–7 kHz. Figure 2.2 shows the average pressure gain of the outer ear as a function of frequency according to Shaw's summary of acoustic measurements performed in 12 different studies (Shaw 1974; Shaw and Vaillancourt 1985). The peak at 2.5 kHz is independent of the direction of the sound relative to the head. At higher frequencies, however, the transfer function changes with the angle of incidence because of the interaction between incoming sound and the complex structures of the pinna. As described in Section 2.3.4, these direction-dependent spectral contours provide important localisation cues, especially for identifying the elevation of a sound.

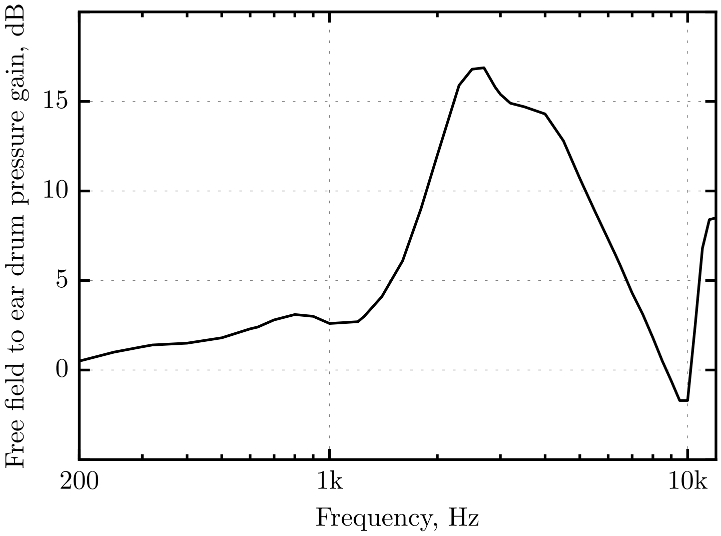


Figure 2.2: Transfer function of the outer ear (ratio of sound pressure in the free field to sound pressure at the eardrum) at an azimuth of 0° (directly in front of the head). Data from Shaw (Shaw 1974; Shaw and Vaillancourt 1985).

2.2.2 The middle ear

When a sound wave meets a high-impedance medium, such as sea water, most of its acoustic energy is reflected. The purpose of the middle ear is to reduce this effect when transferring vibrations at the eardrum (low-impedance air) to the cochlea of the inner ear (high-impedance fluids). Thus, the middle ear serves as an acoustic impedance matching device, coupling acoustic energy from the outer ear to the inner ear. This is primarily achieved by a high-pressure gain (about 25 dB) resulting from the large surface area of the eardrum relative to that of the oval window. The middle ear comprises three ossicular bones: the malleus, the incus, and the stapes. Collectively, the ossicles form a lever action that increases force and decreases velocity at the stapes which connects to the oval window, causing it to vibrate which in turn causes the fluids of the inner ear to move. This lever action also contributes, though to a lesser extent, to the transmission of acoustic energy to the cochlear fluids.

The transmission of sound through the middle ear varies with the frequency of the stimulus. At low to mid frequencies, pressure transfer is determined by the stiffness of the middle ear structures (elasticity in tympanic membrane and ligaments of ear bones) and by the compression and expansion of air inside the cavity. Transmission at high frequencies is limited by the mass of the ossicular structure and is also affected by the vibratory patterns of the eardrum breaking up into zones. This frequency dependency can be examined by calculating the transfer function of the middle ear, defined as the ratio of the sound pressure in the scala vestibuli of the cochlea to the sound pressure in the ear canal. For example, Figure 2.3 shows the average transfer function of the middle ear in 11 human ears (Aibara et al. 2001). The average transfer function over 0.05–10 kHz is given by the solid curve and shows a bandpass characteristic with a peak of 23.5 dB at 1.2 kHz. The dashed lines represent an average standard deviation (SD) derived from Aibara et al. (2001). The authors reported approximate slopes of 6 dB/octave from 0.1 to 1.2 kHz and -6 dB/octave above 1.2 kHz. Interestingly, more recent data from the same laboratory showed a shallower high-frequency roll-off of about -2 dB/octave (O'Connor and Puria 2006).

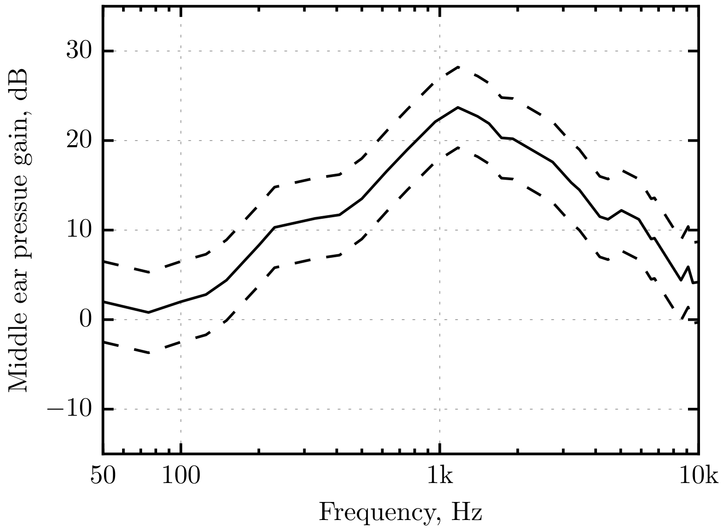


Figure 2.3: Mean (solid) and SD (dashed) transfer function of 11 human middle ears. Data based on Figure 6 (A) of Aibara et al. (2001).

Although the middle ear is often modelled as a linear time-invariant filter, muscles attached to the ossicles contract in response to loud sounds, stiffening the ossicular structure. The effect of this is an attenuation of the incoming sound, especially at frequencies below 1.5 kHz. The transmission response of the middle ear is therefore level dependent. It is believed that the acoustic reflex may help to reduce damage to the cochlea when exposed to loud sounds, although the response time is not sufficient to protect against very transient sounds. Other functions of the reflex include reducing the audibility of self-generated sounds such as speech and lowering masking at high SPLs.

2.2.3 The inner ear

The inner ear comprises three parts: the semicircular canals, the vestibule and the cochlea. The semicircular canals (superior, posterior and lateral) open into the vestibule, utricle and saccule. These structures are collectively known as the vestibular system which affects the sense of balance. The stapes footplate of the middle ear connects to the oval window of the vestibule—the central cavity of the inner ear. The primary organ of the auditory system is the cochlea, the role of which is to transform vibrations at the stapes into neural information and is the main topic of this section.

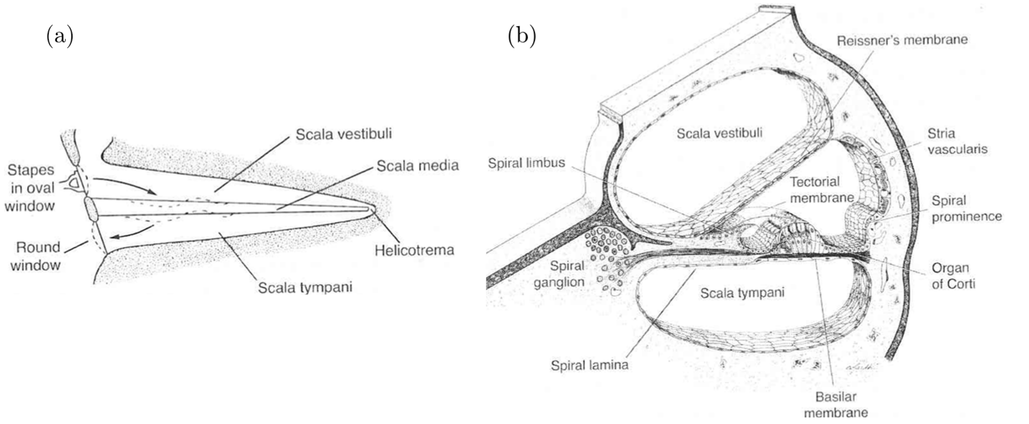


Figure 2.4: (a) Path of vibrations in the unrolled cochlear duct. (b) Cross section of the cochlear duct depicting the three scalae and related structures. Adapted from Pickles (2008).

The cochlea is a coiled structure about 1 cm wide and 3 cm when elongated (see Figure 2.4(a)). It is divided by the basilar membrane and Resissner's membrane into three scalae: the scala vestibuli, the scala tympani and the central scala media or cochlear duct (see Figure 2.4(b)). The two outer scalae (vestibuli and tympani) contain the fluid perilymph, which is near the potential of the surrounding bone, and are joined at the apex by an opening called the helicotrema. In contrast, the inner compartment, the scala media, contains the fluid endolymph which is at a high positive potential (+80 mV) and does not communicate directly with the other scalae. This electrochemical potential difference is responsible for the generation of neural activity in the hair cells and auditory nerve.

The basilar membrane

Eardrum vibrations are transmitted to the oval window by the middle ear bones, causing a displacement of the cochlear fluids towards the round window which opens onto the scala tympani (see Figure 2.4(a)). This displacement of fluid induces a travelling wave on the basilar membrane—a structural element that forms the floor of the cochlear duct. The basilar membrane is about 34 mm long and is wider and more flaccid at the apical end compared to at the base. It is the variation in elasticity and width that determines the position and pattern of the wave on the basilar membrane. Although every point along the membrane vibrates at the same frequency as the input stimulus, Békésky (1960) showed the location of *maximum* displacement to depend on input frequency: the lower the frequency of the tone, the farther the peak moves towards the apex. Travelling waves generated by high-frequency tones therefore peak at the base of the cochlea. Furthermore, membrane displacement is not confined to a single point, but rather increases in amplitude as the wave travels towards the apex and then decays rapidly beyond reaching a specific point. This is demonstrated in Figure 2.5 which shows the vibratory patterns and corresponding envelopes of cochlear travelling waves produced by a single tone at two different frequencies. The figure illustrates two

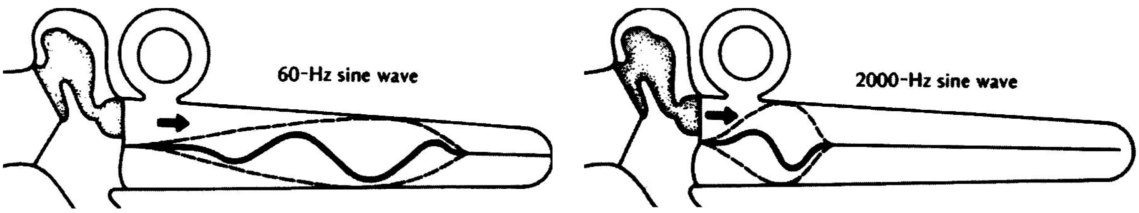


Figure 2.5: Schematic of the cochlea showing instantaneous patterns (solid) and envelopes (dashed) of a 60 Hz and 2 kHz travelling wave. Peak displacement is shown by the highest point of the envelope, which is near the apex for low frequencies and near the base for high frequencies. Adapted from Yost (2007).

important points:

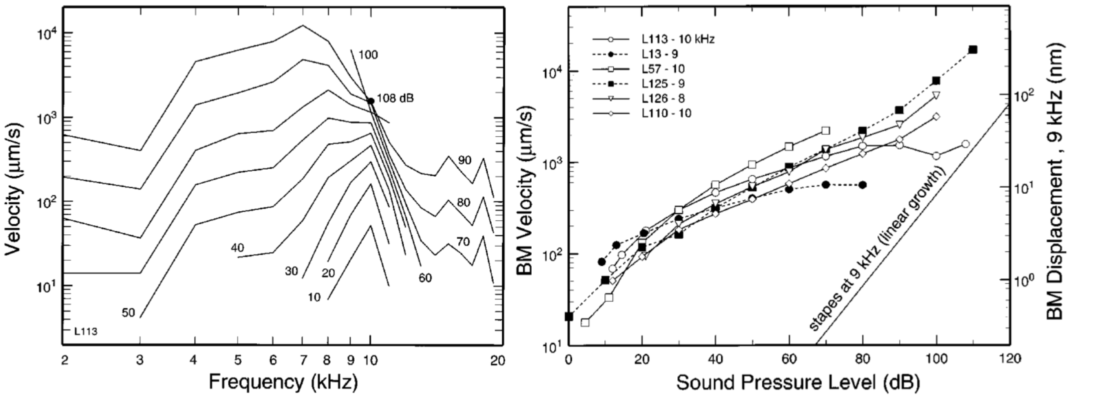
1. As the frequency of the tone increases, the position of maximum displacement shifts from the apical end towards the basal end of the cochlea.

2. Because lower-frequency tones travel farther along the membrane, they stimulate both the basal region as well areas near the apex (at peak excitation). In contrast, high-frequency stimulation is concentrated at the base of the cochlea.

Consider the second schematic of Figure 2.5 and imagine a probe measuring the response of the basilar membrane at the location corresponding to the peak excitation (envelope maximum in the schematic). For a pure tone stimulus presented at a constant input level to the cochlea, maximum displacement would occur at a frequency of 2 kHz. Because lower frequencies also generate waves that travel past the same location (see the response at 60 Hz), membrane displacement would also be measured at the same location for tone frequencies lower than 2 kHz, though to a lesser extent. However, frequencies above 2 kHz do not stimulate the measurement point to the same degree because the apical end of the envelope decays much rapidly relative to the basal end. Therefore, the frequency response of the basilar membrane at a particular location shows a bandpass characteristic.

Vibratory patterns can be studied using measurement techniques such as Mossbauer and laser interferometry. For example, Ruggero et al. (1997b) used the latter technique to measure basilar-membrane responses to pure tones at a site 3.5 mm from the base of the chinchilla cochlea. The frequency of greatest sensitivity, known as the characteristic frequency (CF), at this position was 10 kHz. Figure 2.6(a) shows the peak velocity (displacement per unit time) of the basilar membrane at CF in response to tone pips at different frequencies for low (10 dB SPL) to high (90 dB SPL) intensities. For input levels of 10–50 dB SPL the peak response occurs when the stimulus frequency matches CF and shows a distinctive bandpass shape, particularly at the lowest input level. As the stimulation level increases, the area of excitation broadens and the sharp tuning diminishes where the point of greatest sensitivity shifts towards lower frequencies. This happens because the response at CF site to nearby frequencies is compressive, but linear at lower frequencies. This illustrates that the frequency selectivity at given point on the basilar membrane is level dependent.

As the intensity of the stimulus is raised, the response grows nonlinearly around the peak of the travelling wave. This compressive nonlinearity is reflected by the reduced growth in velocity around CF as the level of stimulation increases by 10 dB increments, i.e. the slope of the basilar membrane input-output function decreases with increasing input level. Figure 2.6(b) shows velocity-intensity functions of the basilar membrane in response to tones with frequency equal to or less than CF (10 kHz). The functions at 9 and 10 kHz show a strong compressive behaviour (compare the slope before and after about 20 dB). This nonlinearity is significantly reduced as tone frequency



(a) (b)

Figure 2.6: (a) Velocity of chinchilla basilar-membrane response to tone pips as a function of stimulus frequency and intensity. The measurement was taken at 3.5 mm from the base of the cochlea, which corresponds to a CF of 10 kHz. (b) Velocity of basilar-membrane response to tones with frequency equal to and lower than CF. The straight dashed line has a slope of 1 (dB/dB). Adapted from Ruggero et al. (1997a).

decreases, becoming linear at very low frequencies, as reflected by the straight dotted line. Two primary functional consequences of the nonlinearity near CF is the reduction in input dynamic range inside the cochlea and the production of distortion products. A compressed dynamic range allows perception to operate over a wide range of stimulus intensities but consequently generates audible distortion products which influence perception.

As mentioned, the sharp tuning at CF occurs at low intensities and diminishes at high intensities where the response becomes much broader and takes on a lowpass function. A popular explanation for the mechanism driving this effect is that the healthy cochlea makes use of an active process determined by the outer hair cells. At low stimulus intensities, energy is fed back into the travelling wave, increasing the amplitude of the input vibration. This localised amplification generates a sharply tuned response. As the level of stimulation increases, amplification decreases, causing the response to grow at a rate that is slower than the input. At high levels of stimulation, the cochlear amplifier provides little further contribution and so the input-output behaviour returns to linearity. Because compression mainly occurs near the peak of the vibratory pattern, only frequencies near CF experience this nonlinearity.

In short, each point on the basilar membrane in a healthy cochlea is sharply tuned at low levels, responding to a select range of frequencies. At mid-range levels, the response of the membrane is compressive for stimulus frequencies near CF, effectively broadening the tuning. It should be noted that the pattern of responses are less consistent at the apical end of the basilar membrane. The amount of compression appears to be less than at the base and does not depend as much on the input frequency relative to CF; the frequency-dependency of compression is more broadly tuned.

The organ of Corti

The organ of Corti, contained within the scala media, lies on the surface of the basilar membrane. It is made up of one row of inner hair cells, three rows of outer hair cells, supporting cells and the pillar cells which form the tunnel of Corti separating the inner and outer hair cells. There are about 3500 inner hair cells, each with 40 hairs or stereocilia arranged in two rows in the shape of a ‘U’. There are about 12 000 outer hair cells, each with 150 cilia organised in three of more rows in the shape of a ‘V’. The tectorial membrane is positioned above the organ of Corti and contacts the stereocilia of the outer hair cells.

The hair cells and supporting cells contain the necessary proteins for establishing high sensitivity of the cochlear structures to biomechanical vibration. Vibration of the stapes causes a travelling wave on the basilar membrane which leads to a bending of the cilia of hair cells. The shearing of stereocilia initiates a neural response in the auditory nerve. It is the difference in pivot points between the tectorial membrane and the basilar membrane that cause this shearing. The cilia of the outer hair cells are bent as a consequence of direct contact with the tectorial membrane. In contrast, it is the fluids trapped between the stereocilia and tectorial membrane that affects the inner hair cells. The shearing of stereocilia is transduced into neural activity by a change in permeability of the hair cell. Tip links are strength-enhancing cross bridges that connect to stereocilia and help to open and close ion channels at the top of the hair cell. This allows for the transport of sodium and potassium ions, thus initiating the neural transduction process.

The majority (85–95%) of afferent nerve fibres are radial fibres which connect to the inner hair cells. These fibres are responsible for carrying sensory information from the organ of Corti to the brainstem and brain. The inner hair cells are therefore the major mechanism for conveying aural information about the incoming stimulus.

When stimulated, the outer hair cells expand and contract. Their change in length influences the vibratory pattern occurring within the cochlear structures. As discussed, this establishes an active mechanism whereby the outer hair cells respond to movement by changing the mechanical coupling between the basilar membrane and tectorial membrane. The sensitivity of the cochlea is amplified as a result of this feedback, thus transducing small vibrations into neural impulses via the inner hair cells. In other words, the outer hair cells serve to amplify the incoming sound with a sharp tuning. They are the active component responsible for the nonlinearities observed in basilar membrane responses, as discussed above.

Neural responses

The auditory nerve carries information from the cochlea to the central nervous system. There are about 30 000 nerve fibres in each ear, the majority of which make synaptic contacts with the inner hair cells. 20 afferent fibres innervate each inner hair cell. Electrical potentials produced in the inner hair cells initiate potentials in the auditory nerve fibres which are then propagated along the nerve fibre to the hebenula perforata (openings in the cochlea that allow neural fibres to pass). Here, action potentials or spikes are produced that travel along the auditory nerve to the cochlear nucleus. The rate of these neural spikes is proportional to the velocity of basilar membrane motion.

The action potential generated by an individual neuron is independent of stimulus amplitude. It is the timing and number of spikes from a combination of auditory nerves that conveys such information. Interestingly, electrophysiologists have observed spontaneous neural discharges in individual neurons in absence of external stimulation. The discharge rate under spontaneous activity can be measured and thus used as a baseline to measure the effects of acoustic signals on neural responses. The spontaneous discharge rate ranges between 0–100 spikes per second and the maximum number of spikes is restricted by a refractory period which determines the minimum time interval between consecutive firings. The refractory period is approximately 1 ms, yielding a maximum discharge rate of 1000 spikes per second, though most nerve fibres do not exceed a rate of about 500. The threshold of a neuron determines the minimum intensity of a stimulus required to increase the discharge rate beyond the spontaneous rate. Neurons with high spontaneous rates are generally associated with lower thresholds.

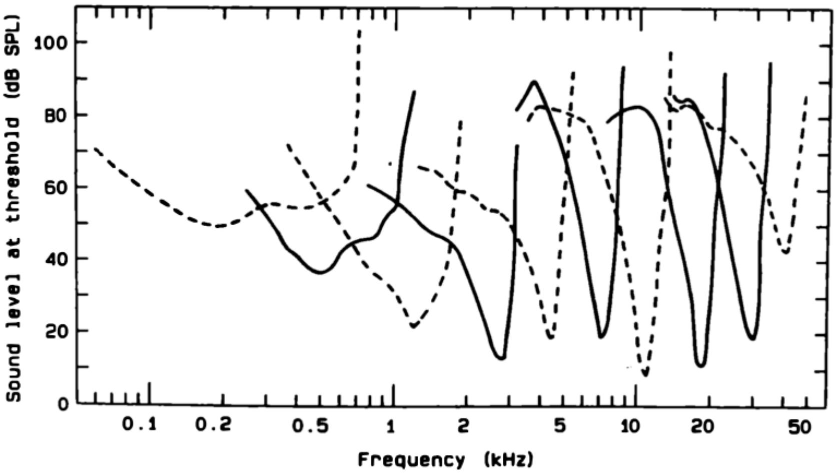


Figure 2.7: Tuning curves measured from single neurons in the auditory nerve of anaesthetised cats. Image from Moore (2013).

The frequency selectivity of a single neuron is best described by a tuning curve which plots threshold as a function of frequency. As with measures of basilar membrane selectivity, the frequency at which a neuron is most sensitive (lowest threshold level) is also called the characteristic frequency (CF). Figure 2.7 shows typical tuning curves of neurons with different CFs. The tuning curves show a sharply tuned tip where sensitivity is greatest. The fibres show a bandpass characteristic with a steeper high-frequency slopes compared to the low-frequency slopes (on a logarithmic frequency axis). Because each fibre responds to vibration at a single point on the basilar membrane, the tuning curves reflect similar frequency selectivity as measured on the basilar membrane. In short, the response of each auditory nerve shows a bandpass characteristic, the width of which increases with centre frequency (in hertz) and the level of the input stimulus.

Rate-level functions

The spike rate of a neuron can be plotted as a function of the input stimulus level to produce a rate-level function. The dynamic range of a neuron is typically about 20 to 50 dB above its threshold, after which saturation occurs. However, neuronal dynamic range decreases with increasing spontaneous activity. Thus, neurons with low thresholds tend to have reduced dynamic range compared to those with higher thresholds. Figure 2.8 shows the discharge rate of single neurons as a function of stimulus level. Curves (a)–(c) show responses from neurons with low to high thresholds, respectively. Notice that high-threshold fibres show a shallower rate-level function and thus a wider dynamic range at CF. The shallower slopes of these fibres reflect the strong compressive nonlinearity of the basilar membrane intensity functions to CF tones at higher stimulus levels.

Temporal response

Post-stimulus-time histograms (PSTHs) display the number of action potentials or spikes an auditory nerve fibre discharges at each moment in time to a repeated stimulus. The spike count is high at the stimulus onset, but decreases rapidly over the first 10–20 ms of stimulation and then drops more steadily. This adaptation effect is illustrated in Figure 2.9 which shows a PSTH to a pure tone burst repeated many times. After the tone offset, the number of spikes falls *below* the count produced by spontaneous activity, before recovering to the normal rate.

For nerve fibres below about 4–5 kHz, nerve spikes tend to be phase locked (synchronised)

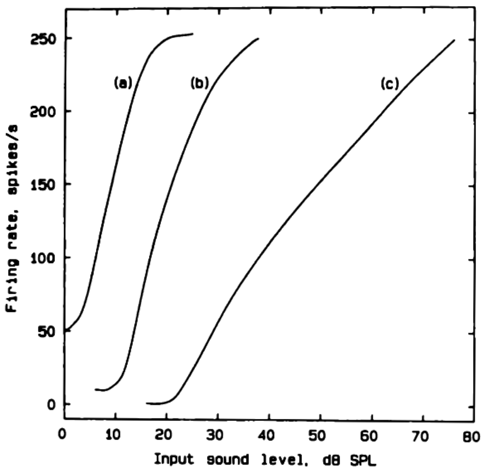


Figure 2.8: Discharge rate of single neurons as a function of level. Curves (a), (b) and (c) illustrate the response of neurons with high, medium and low spontaneous discharge rates, respectively. The stimulus corresponds to a sinusoid with frequency equal to the CF of each neuron. Image from Moore (2013).

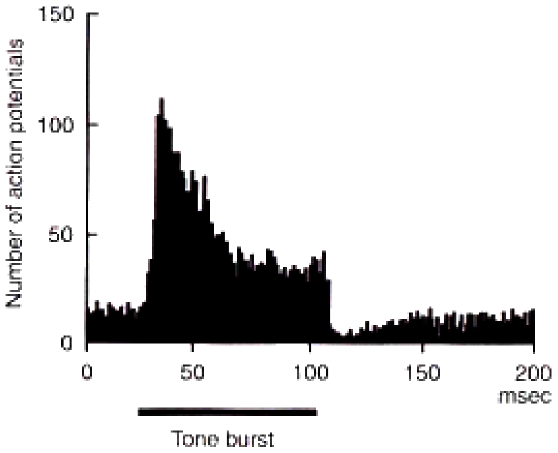


Figure 2.9: Single auditory nerve fibres show an initial sharp onset response to a tone burst, followed by a decrease and finally a transient suppression of spontaneous activity at the offset of the stimulus. This PSTH was formed by repeatedly presenting a tone burst and counting the number of action potentials at each point in time. Image from Pickles (2008).

to a pure tone stimulus. Due to the stochastic nature of the neuronal function, this does not mean that the fibre fires once per cycle, but that when they do fire, they occur in only one phase of the stimulating waveform. As mentioned, the firing rate is limited by the refractory period (roughly 1 ms) and thus fibres are unable to discharge to every cycle of a sinusoid with a frequency above 1 kHz. Furthermore, a neuron only fires when stereocilia are sheared in one direction of the stimulating signal. Phase locking can therefore be seen as a stochastic bandlimited detection process applied to the half-wave rectified input stimulus. Figure 2.10 demonstrates phase locking by means of a period histogram, also referred to as a phase-locked PSTH. The occurrence of each spike is plotted in time and the time axis is reset in every cycle of the stimulating waveform. Notice that the period histogram follows the half-wave rectified version of the tone and that no squaring takes place despite saturation at levels exceeding 70 dB SPL. The ability of a fibre to synchronise with the waveform can be derived from the period histogram. If the period histogram changes sinusoidally, then neural synchrony is said to be high. Conversely, if the neural patten is evenly

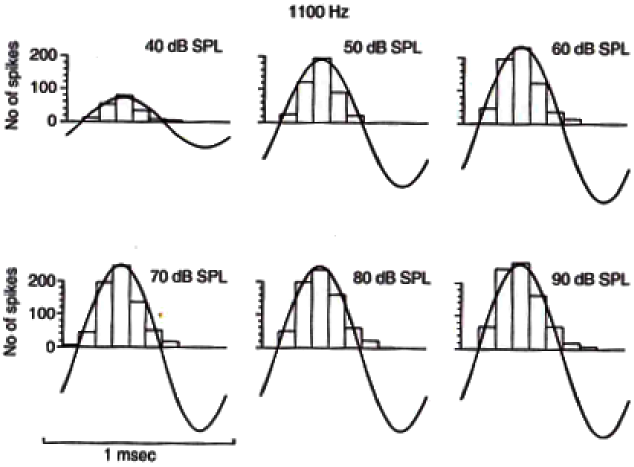


Figure 2.10: Period histogram of an auditory nerve fibre given a low-frequency pure tone. The sinusoids of best fit show that the neural pattern is synchronised with the stimulating waveform, even at high levels where the firing is saturated. Image from Pickles (2008).

distributed throughout the waveform period, no synchronisation takes place because the neuron is unable to follow the temporal trajectory of the stimulus.

Two-tone suppression

Discussion thus far has concentrated on neural activity in response to sinusoids presented at different frequencies and intensities. Presenting a single pure tone with a frequency near a neuron's CF will cause it to fire above its spontaneous activity rate. In absence of any other sounds, a fibre's response is always excitatory. However, if two tones are presented simultaneously, then, depending on frequency and level, the spike rate generated by the neuron in response to the first tone may actually decrease. The second tone therefore inhibits or *suppresses* the activity invoked by the first tone. This phenomenon is called two-tone suppression and is illustrated in Figure 2.11(a). The area bounded by white circles shows the excitatory tuning curve of a single neuron and the open triangle denotes the level of the probe tone which is typically 10 dB above threshold at the fibre's

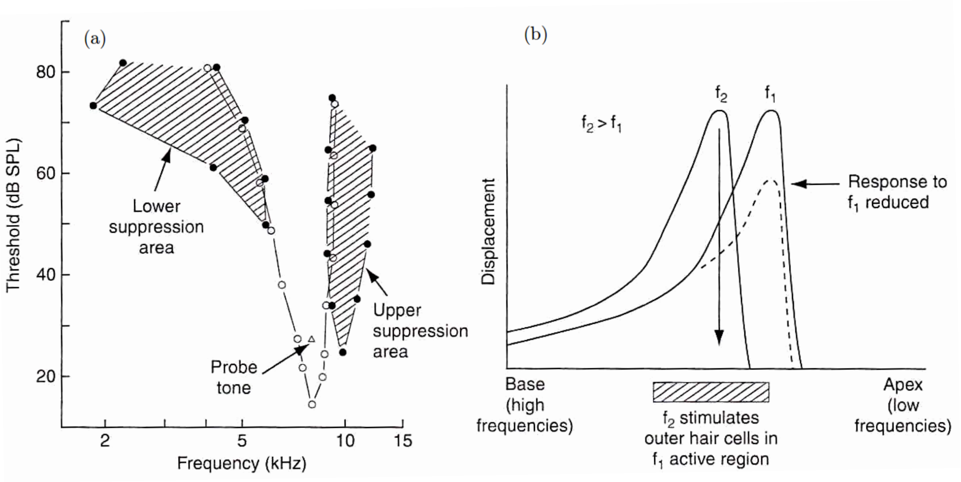


Figure 2.11: (a) Excitatory tuning curve of an auditory nerve fibre (white circles). The shaded areas bounded by the black circles correspond to the fibre's upper and lower suppression areas. (b) Demonstration of two-tone suppression where the suppressor tone is of a higher frequency f_2 than the suppressed tone with frequency f_1. See text for explanation. Adapted from Pickles (2008).

CF. The rate of discharge increases when a second tone with frequency and level falling within the excitatory area is presented. However, when frequency and level of the second tone are varied to within the suppression regions (shaded areas), the average firing rate recorded with the probe drops by at least 20%.

Two-tone suppression occurs at the mechanical stage and is most likely explained by the active behaviour of the outer hair cells. As the travelling waves generated by two sinusoids overlap, the active amplification near the CF corresponding to the excitatory tone is effectively reduced. The travelling wave produced by the suppressor is filtered out by cochlear mechanics, leaving only the wave due to the suppressed tone, but now at a reduced amplitude. This is illustrated in Figure 2.11(b). Note that this explanation of suppression is only appropriate when the excitatory areas of the suppressor and suppressed tones overlap in the upper suppression area. Low-frequency two-tone suppression is currently not well understood.

2.2.4 Modelling the auditory periphery

Computer models of the auditory periphery have been developed to simulate the response of the ear to arbitrary acoustic signals and gain a better understanding of the mechanisms driving the hearing process. A typical auditory model comprises three key stages. First, the transfer function of the outer and middle ear is modelled, typically using linear time-invariant filters. Some models accommodate different sound fields and allow the direction of the incoming source to be specified, e.g. free field with frontal incidence. Second, the frequency selectivity of the basilar membrane is modelled using a bank of bandpass filters with filter bandwidths that increase with centre frequency in accordance with shapes of isovelocity tuning curves at different points along the basilar membrane. The filters are spaced with a near-logarithmic mapping of stimulus frequency to place of peak excitation on the basilar membrane, called the tonotopic map (see Section 2.3.2). A common filter used to mimic both the bandpass filtering action of the basilar membrane and psychophysical measures of frequency selectivity is the Gammatone filter (Boer and Kuyper 1968; Johannesma 1972; Boer and Jongh 1978; Schofield 1985). The Gammatone was developed to simulate the impulse response of auditory nerve fibres and is advantageous in that it has a simple and efficient time-domain implementation (Slaney 1993). More sophisticated models of the hearing system employ nonlinear and asymmetric filters such as the Gammachirp filter (Irino and Patterson 1997; Irino and Patterson 2001) and the dual-resonance nonlinear filter (Lopez-Poveda and Meddis 2001; Meddis et al. 2001), to more accurately simulate the vibratory patterns on the basilar membrane. In particular, they simulate the change in CF and bandwidth at different stimulus levels at a given cochlear site. In contrast, the Gammatone filter is linear in level and has a symmetric frequency response. The final stage of the auditory periphery model mimics the transduction of mechanical vibration to neural signals at the output of each auditory channel. A simplified model involves half-wave rectification followed by lowpass filtering applied to the output of each channel of the filter bank. As with models of basilar membrane filtering, more complex computational devices have been developed to better represent the stochastic nature of the mechanical to neural transduction that takes place between the inner hair cells and individual auditory-nerve fibres, such as Meddis' hair cell model (Meddis 1986, 1988).

2.3 Hearing psychology

Psychoacousticians study the relationship between the physical attributes of sound and hearing sensation. This involves designing a psychophysical experiment in which psychological responses to an external stimulus are measured. This procedure can be viewed as measuring an unknown system (the human listener) with inputs and outputs. By carefully manipulating the properties of the input acoustic signal, systematic changes in the response of the listener can be captured and correlations can be made. By conducting different experiments and relating findings with physiological measurements, one can develop an understanding of the mechanisms driving perception and establish models for predicting subjective responses to new sounds. The aim of this section is to provide the reader with an overview of key findings from experiments concerned with cognitive responses to sound. In particular, the fundamentals of psychoacoustics are summarised: perceptual thresholds, critical bands and masking, loudness perception, and spatial hearing. An understanding of these topics is imperative for developing models and applications founded on auditory perception.

2.3.1 Absolute thresholds

The absolute threshold of a sound is the minimum detectable SPL of the sound in the absence of any other sounds. There are two common methods for measuring absolute threshold. The first uses a probe microphone placed close to the listener's eardrum with the stimulus typically presented through headphones. When measured this way, the threshold is called the minimal audible pressure (MAP). The second approach is to measure the threshold of the sound presented via loudspeaker in a controlled sound field, such as an anechoic chamber. In the absence of the listener, the sound pressure measurement is made at the point corresponding to the centre of the listener's head when present. This form of threshold is called the minimal audible field (MAF). Thresholds of pure tones at different frequencies produce audibility curves (either MAP or MAF) that characterise the frequency-sensitivity of the human ear in a given listening condition.

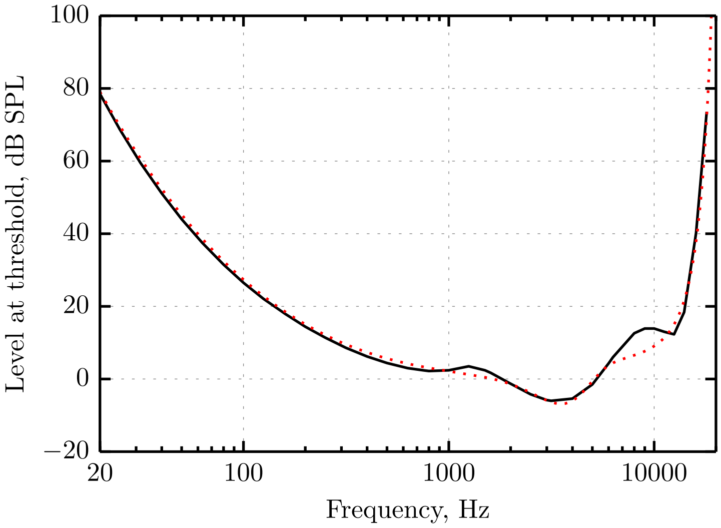


Figure 2.12: Threshold of hearing for pure tones presented binaurally in the free field with frontal incidence. The black solid curve shows the threshold data of ISO 389-7 (2005). The red dotted curve is the approximation given by Equation 2.16.

The solid curve in Figure 2.12 shows the reference MAF threshold of hearing for pure tones under free-field binaural listening (frontal incidence) as a function of frequency as per ISO 389-7 (2005). The dotted curve shows an approximation of the threshold data (root mean square error

(RMSE) = 2.1 dB) according to the formula

$$\begin{aligned} A_{ThQ}(f) = &-5.92\log_{10}(f)^3 + 66.34\log_{10}(f)^2 - 243.84\log_{10}(f) + 296.12 \\ &+ e^{\left(\frac{f}{6900}\right)^{1.5}} - 10e^{-5\left(\frac{f}{3600}-1\right)^2}, \end{aligned} \tag{2.16}$$

where f is frequency in Hz. The approximation is a modification of the function proposed by Terhardt (1979) to better account for the more recent data published by the International Organization for Standardization (ISO) in 2005.

The MAF curve illustrates that the human ear is not equally sensitive to pure tones at different frequencies; if it were, the threshold would be a constant level irrespective of frequency. Hearing sensitivity drops off at low and high frequencies and so the audibility curve shows a smooth bandstop characteristic. For example, as the tone frequency decreases below about 1 kHz, the stimulus must be presented at an increasingly higher intensity in order to be detected by the listener. The acoustic effects of the head, torso and middle ear are partly responsible for this nonuniform frequency sensitivity. Recall that the outer ear enhances sound level in the 2–7 kHz region, peaking at around 2.5 kHz (Figure 2.2). In combination with the optimum efficiency of the middle ear at mid frequencies (Figure 2.3), this explains why the audibility curve shows a global minimum near 3 kHz where hearing sensitivity is highest. The marked increase in threshold at very low and high frequencies is partly accounted for by the reduced efficiency of the middle ear. However, the active cochlear amplifier is less effective at frequencies below 500 Hz, which explains why the audibility curve is steeper than the transfer function of the middle ear at low frequencies (Moore et al. 1997).

The MAF (sound-field) and MAP (headphones) contours differ in shape because the former are dependent on the physical characteristics of the listener. In particular, the MAP is somewhat smoother than the MAF at frequencies above 1 kHz, whereas the MAF shows a marked dip at 3–4 kHz. Furthermore, MAP thresholds represent monaural-listening conditions and MAF represent binaural-listening conditions. Compared to the average monaural threshold, the average MAF threshold is about 2 dB lower as a result of both ears being used. It is also worth noting that there is large inter-subject variability in measurements of absolute threshold—some individuals differ by as much as ±20 dB from the reference contour shown in Figure 2.12. Note also that because of this strong between-subject variation, sound levels are often made relative to the audibility threshold of a specific individual and are expressed as a *sensation level* in dB. Thus, two sensation levels of 10 dB measured from two different subjects indicates that the stimulus was presented 10 dB above their respective absolute threshold, which may not be equivalent in SPL. This representation effectively normalises the data for inter-subject differences in detectability.

2.3.2 Masking

Frequency sensitivity refers to a listener's ability to detect pure tones at different frequencies. In contrast, frequency *selectivity* describes the ability of the ear to resolve the constituent sinusoidal components of a complex sound. In psychoacoustics, frequency selectivity is typically measured through masking experiments. Masking determines the amount by which the threshold of audibility of a sound (the maskee) is raised by the presence of another sound (the masker). For example, the average absolute threshold of a 1 kHz pure tone presented binaurally in free field with frontal incidence is 2.4 dB SPL (ISO 389-7 2005). However, if a masking sound such as a narrowband noise centred at 1 kHz is presented simultaneously, the audibility of the pure tone may be significantly reduced as a consequence of frequency masking. The level of the 1 kHz tone would then have to be increased to achieve detectability; thus, the threshold of audibility has been raised. If this new threshold—the masked threshold—is now, for example, 10 dB SPL, then the amount of masking is the difference between the masked and unmasked (in quiet) threshold. In this example, the amount of masking is 7.6 dB.

There are two types of auditory masking: simultaneous masking and nonsimultaneous (or temporal) masking. The first refers to the case where both signal and masker are presented at the same time, and it is primarily the spectral relationship between the two that determines the amount of masking. The second type of masking happens when the signal and masker do not co-exist in time; the masker can still affect the perception of the signal if the masking stimulus is presented before the onset or after the offset of the signal. In this section, the concept of the critical band is introduced, followed by a summary of perceptual phenomena relating to the two types of masking.

Critical bands

The frequency resolution of the hearing system is primarily determined by the tuning of the neural transduction process. As discussed in Section 2.2.3, different frequencies excite different positions on the basilar membrane, with low frequencies stimulating the apical end of the cochlea, and high frequencies stimulating the basal end. Because the hair cells lie along the length of the membrane, this frequency-to-place conversion establishes a neural encoding of sound that is tonotopically organised. Furthermore, the response of the basilar membrane at a given site is frequency selective or tuned, and therefore the sensitivity of a single auditory nerve fibre is also dependent on the frequency of the stimulating waveform. As discussed, neural tuning curves show a bandpass characteristic with maximum sensitivity at a specific frequency. The hearing system is therefore said to behave as though it contains a set of spatially organised bandpass filters known as the auditory filters or critical-band filters. This filtering is responsible for the frequency-resolving power of the ear.

Fletcher (1940) assumed that the auditory filter was rectangular and conducted experiments in which the threshold of a pure tone was measured in the presence of a narrowband masking noise. Threshold was measured as a function of the noise bandwidth, keeping the spectral density constant, i.e. the overall level of the noise increased with bandwidth. Fletcher found that signal threshold increases with the bandwidth of the noise, but flattens off as the bandwidth exceeds a critical point, known as the critical bandwidth. In other words, increasing the noise bandwidth beyond a specific point no longer affects the detectability of the tone, despite the overall intensity of noise continuing to grow. An example threshold measurement according to this procedure is shown in Figure 2.13. Although Fletcher's band-widening experiment leads to systematic errors in the estimation of the bandwidth of the auditory filter, it is nonetheless important because it has led to the widely accepted power-spectrum model of masking (Patterson and Moore 1986). In essence, this model assumes that the auditory system is made up of a series of overlapping bandpass filters and that, when detecting a signal in noise, the listener makes use of a single filter centred close to the signal frequency. Furthermore, it is assumed that only components falling within the auditory filter centred on the signal frequency contribute to masking. Signal threshold is determined by the signal-to-noise ratio at the output of the filter, based entirely on long-term-power spectra (component phases and short-term fluctuations are discounted).

If the transfer function of the auditory filter is denoted $H(f)$ and the spectral density of the

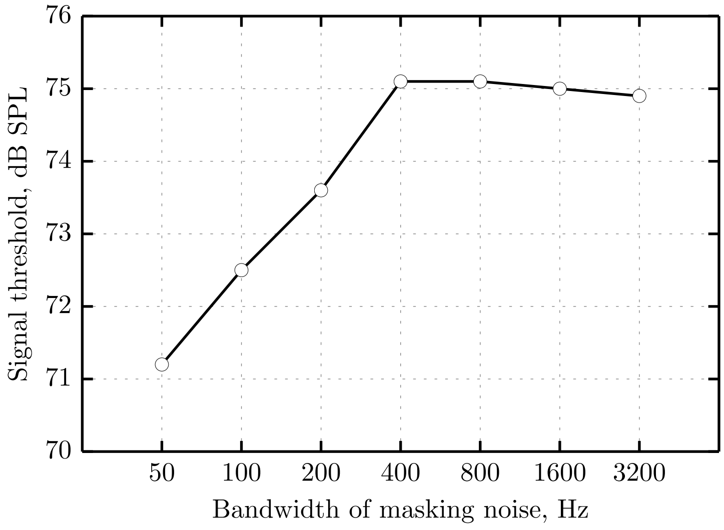


Figure 2.13: Threshold of a 2 kHz pure tone plotted as a function of the bandwidth of a noise masker centred at 2 kHz. Image redrawn from Moore (1995).

white noise masker is denoted N_0, then the total power in the filter is given by

$$P_N = N_0 \int_0^\infty |H(f)|^2 \, \mathrm{d}f. \tag{2.17}$$

Given that the filter is rectangular, the power can be expressed as

$$P_N = N_0 B, \tag{2.18}$$

where B is the filter bandwidth. Fletcher assumed that, at masked threshold, the signal-to-noise ratio k was constant,

$$k = \frac{P_S}{P_N}, \tag{2.19}$$

and therefore that

$$B = \frac{P_S}{k N_0}. \tag{2.20}$$

By assuming that $k = 1$, Fletcher could measure the critical bandwidth using P_S/N_0, known as the critical ratio. Unfortunately, the critical ratio is known to vary with measurement method and more recent experiments have shown the value of k to be around 0.4 and also to vary with frequency (Moore 1995).

In reality, the critical bands are not rectangular but this representation is convenient when expressing the bandwidths of real filters. The critical bandwidth is a measure of the effective bandwidth of the auditory filter if it were rectangular in shape with a maximum value of one. The equivalent rectangular filter should therefore pass the same amount of power as the real filter. It is for this reason that the bandwidth is also called the equivalent rectangular bandwidth (ERB) of the auditory filter.

Psychophysical tuning curves

One technique for better measuring the shape of the auditory filter is based on the process used to derive neural tuning curves of auditory nerve fibres. Recall that these curves characterise the sensitivity of a single neuron to different stimulating frequencies; a neuron is most sensitive to a particular frequency, and as the frequency of stimulus changes, higher levels of stimulation are required to reach threshold. Similarly, the selectivity of the ear at a given place can be measured psychophysically through simultaneous masking. A psychophysical tuning curve (PTC) is typically measured by using a low-intensity pure tone as the target stimulus and either a short-duration sinusoid or narrowband noise as the masking stimulus. The masked threshold is then measured as a function of level and frequency of the masker. It is assumed that the masker produces a constant output from the auditory filter centred at the frequency of the target stimulus. The PTC therefore defines the level of the masker required to produce a constant output from the auditory filter as a function of frequency. Example PTCs are shown in Figure 2.14.

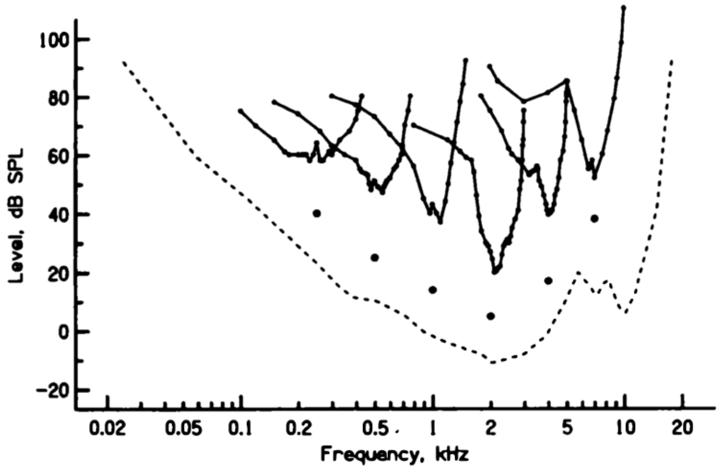


Figure 2.14: PTCs derived from a simultaneous masking experiment involving 50 ms pure tones fixed at a sensation level of 10 dB. The solid circle below each curve indicates the SPL and frequency of the target tone. The masker was a sinusoid and the masked threshold was measured with the masker set at different frequencies. The dashed line corresponds to the absolute threshold for the target tone. Image from Moore (2013).

It is important to stress the difference between tuning curves measured physiologically and psychophysically. Neural tuning curves represent the response of a single neuron to a single input stimulus. Psychophysical tuning curves sample the response of a group of neurons when both the target and the masker are presented simultaneously. The fact that multiple neurons each with slightly different CFs are excited means that the listener can make use of auditory filters centred below or above the signal frequency, whenever it is advantageous to do so. For example, when the frequency of the masking tone is above that of the target tone, the listener would do better at detecting the target by attending to a filter positioned below the target frequency. This deliberate shift in attention, called ‘off-frequency listening’, is done to make use of filters that maximise the signal-to-noise ratio. The result of off-frequency listening is a sharpening of the PTC; the level required for threshold is higher than if off-frequency listening did not take place. A further complication with using PTCs to measure the frequency selectivity of the peripheral auditory system is the level-dependent shape of the auditory filters. In short, this method leads to an underestimation of the slope of lower skirt of the filter and an overestimation of the slope of the upper skirt.

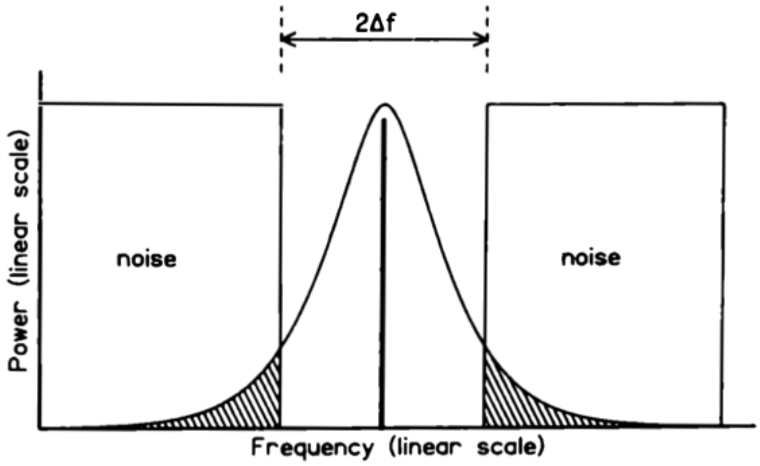


Figure 2.15: Illustration of the notched-noise method for measuring the shape of the auditory filter. Image from Moore (2013).

The notched-noise method

Patterson (1976) established a technique called the 'notched-noise method' to measure the shape of the auditory filter, whilst controlling for off-frequency listening. In this paradigm, illustrated in Figure 2.15, a notched-noise masker is centred on the frequency of the pure tone. Signal power at detection threshold is then recorded as a function of the notch width $2\Delta f$. Because noise is placed either side of the tone frequency, the listener is forced to use an auditory filter located in the notch, presumably close to the frequency of the signal. As the width of the notch increases, less noise passes through the filter and so threshold level decreases (less masking). Compared to Fletcher's method, the notched-noise experiment has the advantage of measuring the shape of the filter and providing an estimate of the signal-to-noise ratio k at threshold.

At moderate noise levels, the shape of the auditory filter is near enough symmetric on a linear frequency scale. The power of the signal at threshold P_S can then be obtained by integrating the power of the notched noise, centred at frequency f_c with a width of $2\Delta f$, passing through both sides of the filter

$$P_S = kN_0 \int_0^{f_c-\Delta f} \mathrm{d}f |H(f)|^2 + kN_0 \int_{f_c+\Delta f}^{\infty} \mathrm{d}f |H(f)|^2. \tag{2.21}$$

The integral represents the total area of the noise passing through the filter, depicted by the shaded areas in Figure 2.15. Because this function measures the integral of the filter at a given frequency deviation Δf, the shape of the filter $|H(f)|^2$ can be derived directly from the slope of the threshold curve. Patterson and Nimmo-Smith (1980) argued that the rounded exponential (roex) function provided the most successful representation of the filter shape. The simplest of the family of roex functions is

$$|H(f)|^2 = (1 + pg)e^{-pg}, \tag{2.22}$$

where g is the normalised deviation of the frequency from the centre frequency of the filter

$$g = \frac{|f - f_c|}{f_c}, \tag{2.23}$$

and p defines the critical bandwidth. The filter is called the roex(p) filter because it has only one parameter (p). The larger the value of p, the more sharply tuned the filter. The area under the function is $4/p$ and so the ERB is $4f_c/p$.

To model asymmetry in the shape of the filter, p can take on two different values, one for the lower side p_l and one for the upper side p_u. The ERB of the roex(p_l, p_u) filter is then $(2/p_l+2/p_u)f_c$. Results show that the upper skirt of the filter becomes slightly steeper with increasing level, though this effect varies with frequency, and that the lower skirt becomes shallower (Moore and Glasberg 1987; Glasberg and Moore 1990).

Critical bandwidth

Since the 1950s, many experiments have been conducted to measure the critical bandwidth or ERB of the auditory filters. The critical bandwidth describes the width of the critical-band filter at a given frequency. The most popular critical bandwidths are those defined by the Bark scale published by Zwicker (1961). The ERB according to Zwicker is

$$\mathrm{ERB}_m(f) = 25 + 75\left(1 + 1.4\left(\frac{f}{1000}\right)^2\right)^{0.69}, \tag{2.24}$$

where f is the frequency of the critical band in Hz. Following a similar convention as Hartmann (1998), the subscript m is used to emphasise that this is a measure of bandwidth for the 'Munich' critical bands as measured in Germany.

The Bark scale (in units of Bark) defines the number of critical bands below a given frequency and as such is a transformation of linear frequency,

$$z_m(f) = 13\arctan(0.00076f) + 3.5\arctan\left((f/7500)^2\right). \tag{2.25}$$

It is assumed that a constant increment in Bark corresponds to a constant distance along the basilar membrane. Equation 2.25 therefore establishes a relationship between stimulus frequency and place of peak excitation on the basilar membrane. The width of each critical band corresponds to a fixed distance along the basilar membrane of 1.3 mm. There are a total of 24 abutting critical bands according to the Bark scale.

Moore and Glasberg (1983) summarised results from masking experiments where the shape of the auditory filter was measured using the notched-noise and the rippled-noise method. Their findings indicated that unlike Zwicker's expression of the critical bandwidth, the width of the auditory filter continues to decrease at centre frequencies below 500 Hz. For normal-hearing listeners, their revised ERB function is (ANSI S3.4 2007)

$$\mathrm{ERB}_c(f) = 24.673(0.004368f + 1), \tag{2.26}$$

where the subscript c is used to differentiate from Equation 2.24, since the measurements of the auditory filter shapes were carried out in Cambridge, UK.

The two ERB formulae are compared in Figure 2.16. At frequencies above 500 Hz, the bandwidths defined by Zwicker (Equation 2.24) and Moore and Glasberg (Equation 2.26) are similar except that the latter are narrower. At lower frequencies, the function according to the Bark scale tapers towards a fixed value of 100 Hz, whereas ERB_c continues to decrease. Moore and Glasberg (1997) state that at low frequencies, the methods from which Zwicker's critical bandwidth function was derived are affected by factors other than frequency sensitivity and that later studies employing more controlled methods show a continuing decrease in bandwidth for centre frequencies below 500 Hz.

Similarly to the Bark scale, the Cam scale is an absolute frequency scale of auditory perception. However, the Cam scale is based on Moore and Glasberg's ERB, and defines the number of filters

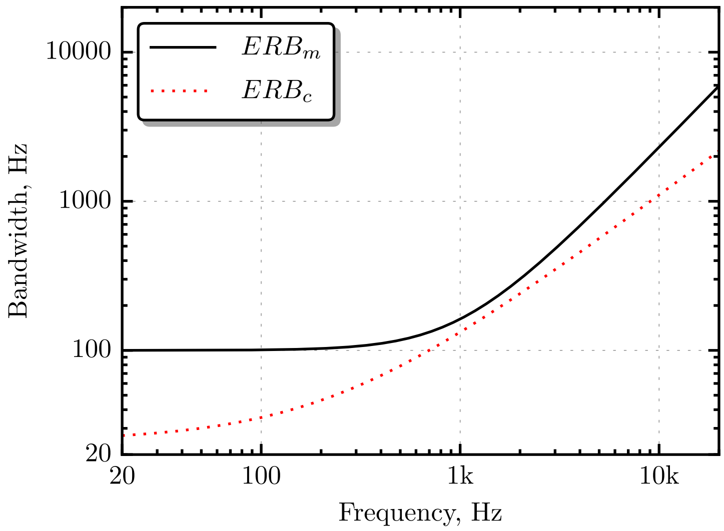


Figure 2.16: ERB as a function of centre frequency according to Zwicker (ERB$_m$), and Moore and Glasberg (ERB$_c$).

below a given centre frequency,

$$z_c(f) = 21.366 \log_{10}(0.004368f + 1). \tag{2.27}$$

As a result of the narrower bandwidths defined by Equation 2.26, more (non-overlapping) filters are required to cover the audible frequency range - approximately 41 on the Cam scale compared to 24 on the Bark scale.

Masking patterns

Simultaneous masking is also referred to as frequency or spectral masking, because the amount of masking is dependent on the spectral relationship between the masker and signal. For example, a narrowband noise centred at the same frequency of a pure tone is much more effective at masking the tone compared to a noise centred an octave higher. Figure 2.17 illustrates this by showing *masking patterns* derived from an experiment whereby a narrowband noise centred at 410 Hz served as the masking stimulus and the threshold of a pure tone was measured as a function its frequency. In this case, as was done in many early experiments investigating frequency masking, it is the signal frequency that is varied while the masker is held constant.

Masking is most effective when the frequency of the tone is close to the centre frequency of the masking noise. Additionally, the masking pattern becomes progressively asymmetrical as the level of the masker increases: the upper slope of the masking pattern becomes shallower. This means that an increment in the level of the masker raises the masked threshold by a greater amount at higher frequencies. Consequently, high-frequency signals are more susceptible to masking caused by lower frequency maskers than vice versa, especially at high-intensity levels. This nonlinear growth in masking is referred to as ‘the upward spread of masking’.

Tonal masking

The findings of experiments in which the masker and maskee are both tonal signals, called tonal masking, also reflect the filtering taking place inside the cochlea. An example of this is the PTC,

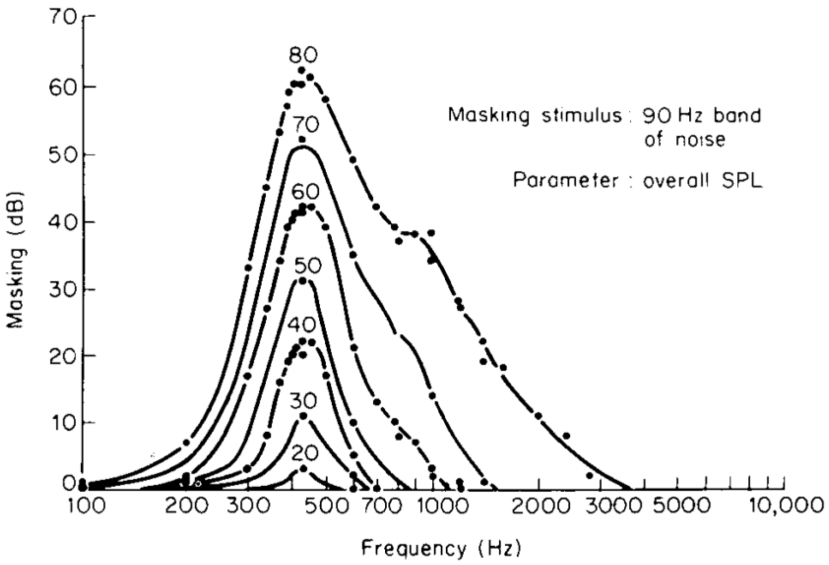


Figure 2.17: Masking patterns for a 90 Hz wide noise centred at 410 Hz. Curves shows the elevation in audibility threshold of a pure tone at different frequencies. Image from Moore (1996).

discussed earlier, where the masked threshold of a pure tone is measured as a function of the level and frequency of a masking tone. In this paradigm, it is preferable to use short-duration tones (30 ms) or replace the masking stimulus with a narrowband of noise in order to reduce temporal modulations that help the listener to detect the target stimulus. Signal detection under tonal masking is also influenced by the presence of aural harmonics and combination tones generated by the nonlinearities of the auditory system.

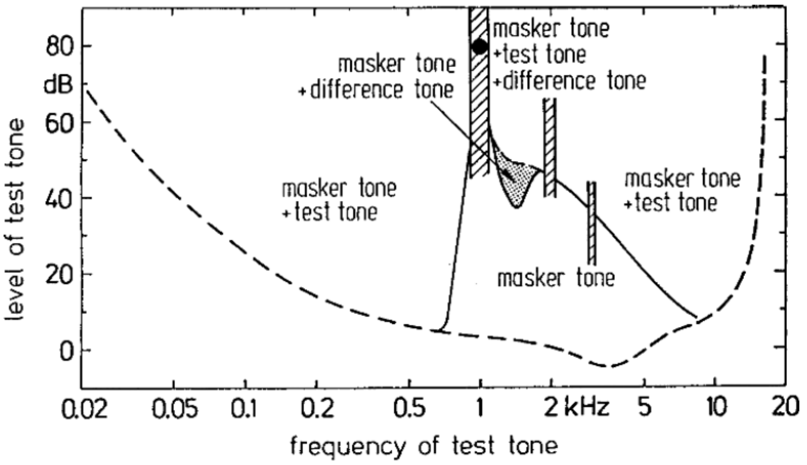


Figure 2.18: Masked threshold of a pure tone in the presence of a 1 kHz masking tone as a function of frequency. The different areas correspond to different hearing sensations. See text for explanation. Image from Fastl and Zwicker (2007).

Tonal masking effects are illustrated in Figure 2.18 which shows the threshold of a pure tone as a function of frequency when masked by a 1 kHz masking tone presented at 80 dB SPL. The dashed line represents threshold in quiet; when the level of the test tone (the target) is below this threshold, the subject only hears the masking tone. When the frequency and level of the test tone is outside of the bounded areas and exceeds threshold in quiet (denoted ‘masker tone + test tone’ in the figure), the listener hears both stimuli. When the frequency and level of the test tone falls within the white area bounded by the solid line (‘masker tone’), the test tone is completely masked and the listener hears only the 1 kHz masker tone.

The solid curve is somewhat similar to the masking pattern produced by the narrowband noise masker as shown in Figure 2.17. The situation is, however, more complex when the parameters defining the target tone fall within the marked areas. For example, the cross-hatched areas correspond to sensations in which the listener experiences a tone that fluctuates in loudness or beats. The beating frequency is equal to the difference between the frequency of the masking tone and that of the target tone. This effect is especially marked when the frequencies of the masker and target are close, but is also present at the harmonics of the masker (corresponding to the cross-hatched regions at 2 kHz and 3 kHz in the figure). This occurs when the masking tone is presented at an intensity level high enough to elicit aural harmonics as a result of the active cochlear mechanism. These harmonics, generated by the hearing system, interact with the target tone to produce audible beats. In addition, combination tones may also be generated by nonlinear distortions in the ear. In this example, at test-tone frequencies near 1.4 kHz, the listener hears both the masking tone and a difference tone near 600 Hz. As a result, the threshold is lower than expected because the listener detects the difference tone rather than the signal tone. This is shown by the shaded area labelled ‘masker tone + difference tone’ in Figure 2.18. Combination tones arise as a result of nonlinear distortions in the hearing system.

The excitation pattern

A masking pattern is crude indication of the neural EP invoked by the masker (Moore 1996). It is approximate because the shape of the masking pattern is affected by off-frequency listening and the detection of combination tones resulting from the interaction of the masker and signal.

Moore and Glasberg (1983) suggested that a more accurate estimate of the neural EP evoked by an arbitrary stimulus can be taken as the output of each auditory filter expressed as a function of centre frequency. The input spectrum is processed through a bank of overlapping roex filters equally spaced on the Cam scale. The EP is formed by plotting the output of each filter as a function of centre frequency; therefore, the resolution of the pattern is determined by the filter spacing. Figure 2.19 demonstrates this procedure using a bank of roex filters applied to a 1 kHz input tone at an effective cochlear level of 40 dB SPL (top panels) and 80 dB SPL (bottom panels). The roex filters were calculated according to ANSI S3.4 (2007) where the parameter defining the filter bandwidth and slope is dependent on both the centre frequency of the filter and the level of the input stimulus. This follows from the general physiological and psychophysical measures of frequency-selectivity at a given CF on the basilar membrane.

The top panel of the figure shows the shape of the roex filters and their output—the EP—in response to a 1 kHz tone presented at 40 dB SPL (red dashed vertical line). Note that the filter shapes are symmetrical but the bandwidth increases with centre frequency (linear frequency scale). As a result, filters located above the tone frequency are excited by the 1 kHz tone to a greater extent than filters positioned at the same distance in hertz below the tone. The excitation therefore spreads to auditory filters located below and above the frequency of tone, but the high-frequency side is slightly shallower than the low-frequency side. The coloured circles show the sampling points of each auditory filter, i.e. where each filter and tone intersect.

The bottom panel of the Figure 2.19 shows the same information but for a 1 kHz tone presented at 80 dB SPL. This time, the lower-frequency skirt of roex filters located above the frequency of the tone are shallower than their respective upper-frequency skirt. This occurs because the *low-frequency* slope p_l decreases with increasing input level, whereas the upper-frequency slope p_u is level-invariant. Accordingly, the EP shows a greater spread of excitation towards *higher frequencies*

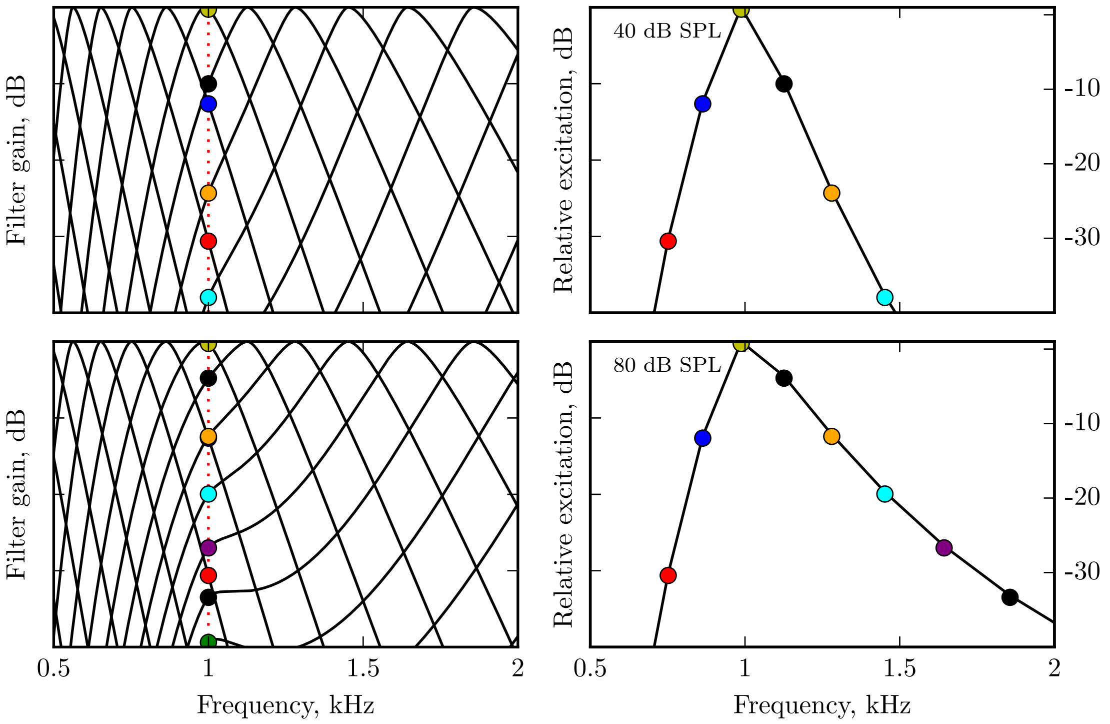


Figure 2.19: Auditory filter shapes (left) and EPs (right) corresponding to a 1 kHz pure tone (red dotted line) at 40 dB SPL (top panels) and 80 dB SPL (bottom panels). The coloured circles denote the locations where the tone intersects each filter shown on the left. EPs are expressed relative to peak excitation (40 and 80 dB SPL) and all functions are plotted on a linear frequency scale.

for the more intense tone. Therefore, by using the concept of the auditory filter, this excitation-pattern model explains the upward spread of masking as reflected by masking patterns.

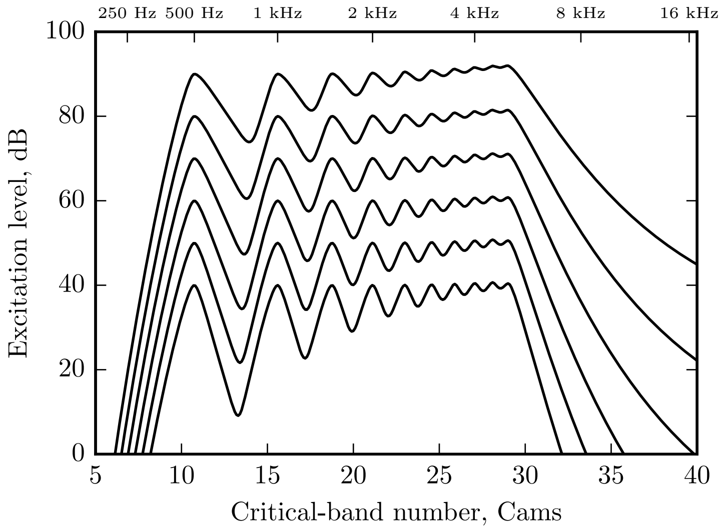


Figure 2.20: EP of a multitone complex made up of ten harmonics of a 500 Hz fundamental, with overall SPL as parameter. The component intensities were equal and EPs were computed with the overall SPL ranging from 50 to 100 dB SPL in steps of 10 dB.

Finally, Figure 2.20 shows EPs of a multitone complex presented at five different overall intensity levels ranging from 50 to 100 dB SPL. The stimulus is made up of equally intense harmonics with a fundamental frequency of 500 Hz and maximum frequency of 5 kHz. In comparison to Figure 2.19, the EPs are plotted on a Cam scale; closer to the auditory representation. It can be seen that the upper slope of the EP is strongly dependent on the input level, making differences between peaks and valleys less distinct as the overall level increases. Furthermore, the ripples between adjacent partials become progressively compressed and smoothed out as one moves from the first component to the last, especially at the highest level. Note that at the highest frequencies (near 5 kHz), the partials fuse together and are not easily resolved. This reflects both the tonotopic encoding of sound in the ear and the broadening of the critical band filters with increasing frequency and input level.

It is important to stress, however, that this model of auditory processing is not entirely accurate. Recall that a given site on the basilar membrane is compressive for frequencies presented near CF, but linear for those farther way (see Figure 2.6). Thus, changes in peak excitation with increasing sound level would not realistically grow linearly as shown in Figure 2.20.

Temporal masking

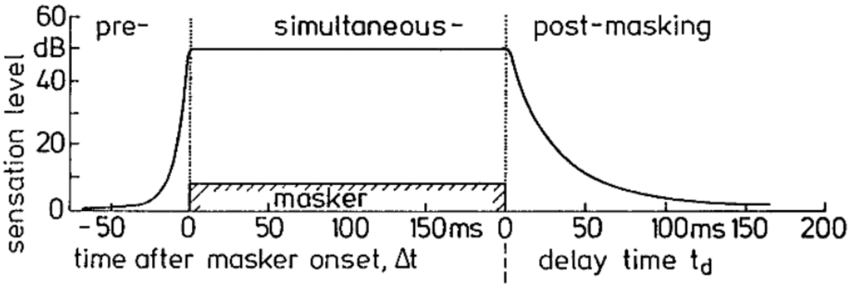


Figure 2.21: Illustration of pre-, simultaneous-, and post-masking. Image from Fastl and Zwicker (2007).

Temporal masking occurs when the audibility of a signal is reduced by a stimulus that does not coexist in time. Post-masking, or forward masking, refers to the masking effect when the masker temporally precedes the signal. Pre-masking, or backward masking, refers to the reverse arrangement: the signal is rendered inaudible by a masker presented *after* the signal has ended. These two phenomena are illustrated in Figure 2.21, which shows a 200 ms masker (typically a tone burst or short pulse) affecting the audibility of a short tone burst with a very small duration. The first time scale Δt shown on the x-axis specifies the temporal separation between the onset of the signal and the onset of the masker. The second time scale t_d is the delay time of the signal after the end of the masker. Clearly the amount of masking (as quantified by sensation level in the figure) is dependent on the temporal gap between the two signals. By comparing the short rise of the masking envelope against the persistent decay, it is clear that post-masking is more effective than pre-masking, with the former lasting up to about 100–200 ms.

The main findings of post-masking are summarised in Figure 2.22. Subplot (a) shows that the amount of masking (in dB) is a linear function of $\log(t_d)$, and this relationship holds for different levels of the masker. The slope is negative, meaning masking decreases as the delay time increases. In Figure 2.22(b), the amount of masking is plotted as a function of masker spectrum level for three delay times. As with simultaneous masking, the amount of masking increases with the level of the masker. In post-masking, however, the rate of increase is less than that commonly found in simultaneous masking, and decreases as the time between the start of the signal and end of

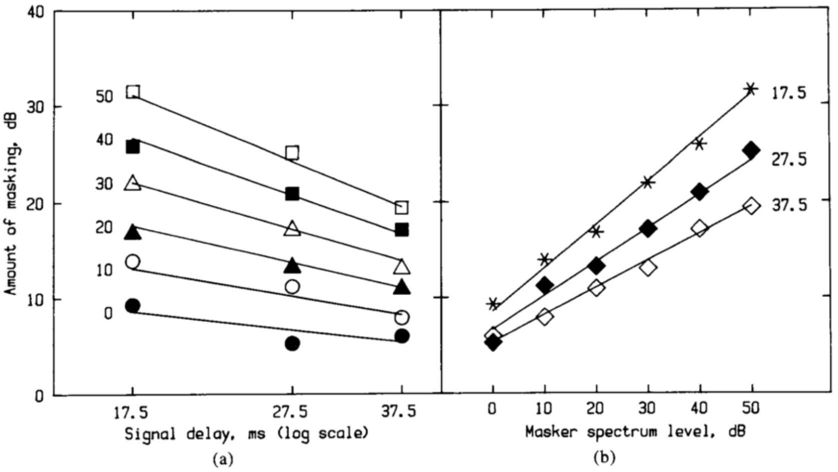


Figure 2.22: (a) Amount of post-masking of a brief 2 kHz signal as a function of the time delay of the signal after the end of the masker. The noise spectrum level is the parameter. (b) Same as (a) but data plotted as a function of masker spectrum level. These functions are called growth of masking functions. In this subplot, the parameter is the signal delay. Image adapted from Moore (1996).

the masker increases. This means that the level of the masker has a weaker effect on threshold in temporal masking compared to simultaneous masking. Additional properties of forward masking include an increase in masking with the duration of the masker (up to 200 ms) (Zwicker 1984), and the frequency relationship between the signal and the masker (Elliott 1967). The latter property is true of both forms of temporal masking.

One successful account of temporal processing in the auditory system is the temporal window model (Moore et al. 1988; Plack and Moore 1990; Plack and Oxenham 1998; Plack and Oxenham 2002). The general idea is to integrate the intensity at the output of a nonlinear auditory filter using a sliding asymmetrical window, centred at time zero. The compressive auditory filtering resembles the active filtering that takes place at a given cochlear site. This intensity signal is then smoothed using a temporal window such as the one displayed in Figure 2.23. This window moves forwards

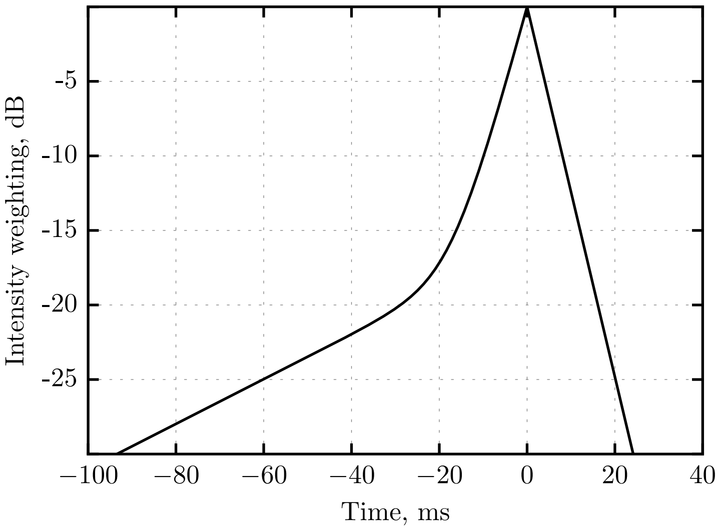


Figure 2.23: The shape of the sliding temporal window used to model temporal masking in the hearing system.

in time, performing a weighted running average at the output of the active auditory filter. Note that the window makes use of a shorter time-constant for times after its centre (peak) compared to times before (negative time). This asymmetry is necessary to account for the build up and decay of energy in the hearing system that explains backward and forward masking. For example, in the case of backward masking, the energy of a short-duration signal may still be growing when a masking stimulus, following the signal, starts to enter the moving integrator. If the masker is sufficiently more intense compared to the signal, the output of the window is dominated by the masker, rendering the signal inaudible. Note that, prior to integration, it is necessary to incorporate a compressive function in order to account for 'on-frequency' nonlinearities in temporal masking data (Plack and Oxenham 2002). However, by using a more complete model of the auditory filter, e.g. the algorithm proposed by Meddis et al. (2001), off-frequency phenomena such as upward spread of masking and suppression can also be accounted for.

2.3.3 Loudness perception

The loudness of a sound is a psychological description of the magnitude of an auditory sensation (Fletcher and Munson 1933). Loudness should not be confused with the physical intensity of an acoustic source; rather, it is a subjective quantity used to describe the perceived strength of a sound. The two quantities are indeed correlated, for example a listener can increase the loudness of a sound by increasing the physical intensity. However, there are many other characteristics of sound that influence loudness perception and the relationship between the physical and psychological domain is often complex and nonlinear. Because loudness is a subjective quantity, psychophysical methods such as scaling and matching techniques are used to measure the sensation (see Section 6.1 for a review). This section provides an overview of loudness perception, which is elaborated on in Chapter 3 where models of loudness perception are discussed in detail.

Loudness level

Loudness level is not an absolute measure of loudness, but a relative quantity that states the physical level (in dB SPL) that a 1 kHz tone must be presented at in order to be as loud as the sound in question. The listener is presented with a reference tone and a 1 kHz comparison tone. The task of the subject is to adjust the level of the comparison tone such that both tones are equally loud. To prevent frequency masking, the tones are presented alternately (in isolation) with a sufficiently long inter-stimulus gap. This is the essence of a loudness-matching paradigm. The level of the 1 kHz tone at the point of subjective equality (PSE) determines the loudness level of the reference tone, specified in units of phons. For example, a 100 Hz reference tone at 80 dB SPL might have a loudness level of 62 phons, which means that the subject set the level of the 1 kHz tone to 62 dB SPL in order to match the loudness of the two tones. Note that according to the definition of the phon scale, the loudness level of a 1 kHz tone is equal to its physical intensity level.

The level of the 1 kHz tone may also be fixed in level, making it the reference stimulus. The listener then adjusts the level of the comparison tone with a specified frequency. When repeated for different frequencies, this variation of the procedure generates an equal-loudness contour or phon contour. The solid curves in Figure 2.24 are the equal-loudness contours defined by the ISO 226 standard (ISO 226 2003). These latest contours were based on a review of results from 12 studies in which equal-loudness curves were measured (Suzuki and Takeshima 2004). The revised contours were established to address conflicting data measured in earlier work (Fletcher and Munson 1933; Robinson and Dadson 1956; Fastl and Zwicker 1987). The dashed curves in Figure 2.24 show

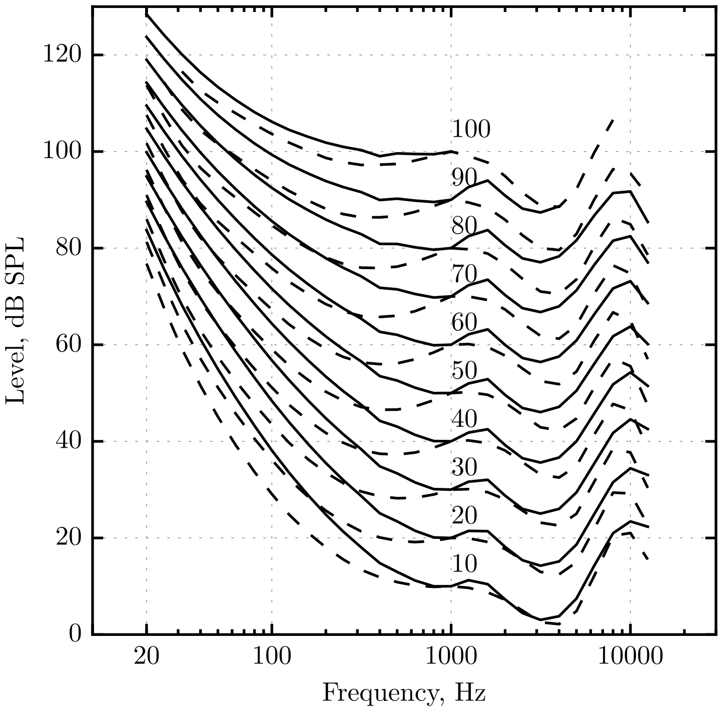


Figure 2.24: Equal-loudness contours as described by the revised ISO 226 standard (ISO 226 2003) (solid) and the 1987 standard (ISO 226 1987) (dashed).

contours according to the 1987 ISO standard (ISO 226 1987) which were based on the contours measured by Robinson and Dadson (1956). It can be seen that the largest discrepancies between the old and new ISO standard occur at low frequencies. Such deviations from more recent data may be attributed to poor calibration of headphone-to-free-field SPLs in the early studies and differences in measurement procedure (Lydolf and Møller 2000; Suzuki and Takeshima 2004).

Each equal-loudness contour describes the (relative) sensitivity of the ear to different frequencies at a given loudness level. If the SPL of the comparison tone is greater than its respective loudness level, the ear is less sensitive to that frequency compared to at 1 kHz. A lower SPL means that the ear is more sensitive than at 1 kHz, since the level of the comparison was set below the level of the reference. As discussed, differences in hearing sensitivity across frequency are primarily attributed to the transfer function of the outer and middle ear. The phon contours are similar in shape to the curve describing absolute threshold (see Figure 2.12) but become flatter at higher levels. The level-dependent shape of the contour can be explained by both internal noise and the active compression that takes place in the cochlea. At very low sound levels, internal noise effectively masks lower frequencies, meaning that greater reproduction levels are needed to establish audibility (or equal loudness) than at higher frequencies. This masking effect diminishes when cochlear excitation is dominated by the external stimulus rather than internal noise. The frequency-dependent compression taking place on the basilar membrane means that the rate of growth of loudness with sound level is also frequency-dependent and, consequently, the shape of the contour changes with each loudness level. For example, at 1 kHz, the change in loudness level from 70 phon to 90 phon corresponds to a change in sound level of 20 dB (by definition). In contrast, the same increase in loudness level requires a gain of only 13.7 dB SPL at 100 Hz. This difference occurs because the loudness of a 1 kHz tone grows at a slower rate, i.e. it is more compressive, than at 100 Hz. As a result, the phon contours become flatter at lower frequencies at higher levels. In short, if the spacing between two contours, at a given frequency, is smaller than the interval at 1 kHz, then loudness growth is faster at that frequency, i.e. less compression takes place.

Sone scale

The phon scale is a frequency-compensated decibel scale. This measure does not scale with the magnitude of sensation: a sound with a loudness level of 80 phons is not twice as loud as the another sound with a level of 40 phons. To construct a linear scale of loudness, scaling procedures such as magnitude estimation or magnitude production are used (Stevens 1957b). For example, in magnitude estimation, the subject is presented with a series of stimuli at different levels and asked to assign a number to each sound using any scale she or he chooses. The numerical estimates assigned by the subject are proportional to a power of signal intensity, meaning equal ratios of subjective magnitudes correspond to equal ratios of signal intensities. Stevens (Stevens 1955, 1957b, 1961b) summarised this phenomenon using the expression

$$\Psi = kI^p, \tag{2.28}$$

where Ψ is the perceived loudness, I is the signal intensity and k and p are constants.[2] When plotted in log-log co-ordinates, the slope of this function is equal to the exponent p. Stevens (1955) reviewed data describing the relationship between loudness and stimulus intensity and found an exponent of 0.3 to best describe the growth of loudness of a 1 kHz tone. This leads to the convenient rule that a *ten-fold* (10 dB) increase in signal intensity leads to a *two-fold* increase in loudness. Stevens proposed the *sone* as the unit of loudness, with 1 sone chosen as the absolute loudness of a 1 kHz tone presented at 40 dB SPL in the free field with frontal incidence. By definition, the loudness level of a 1 kHz tone is equal to its intensity level. As such, an increase in loudness level of 10 phon at any frequency will double the loudness (see Table 2.1). For levels below about 40 dB SPL, loudness changes more rapidly with sound level, and so the power law no longer holds. Models describing loudness near threshold are described in Chapter 3.

Phons	40	50	60	70	80	90
Sones	1	2	4	8	16	32

Table 2.1: Relationship between loudness level in phons and loudness in sones for signal intensities above 40 dB SPL.

Loudness grows differently for broadband and narrowband sounds. The loudness-growth function of a uniform exciting noise is most interesting because it produces equal power per critical band. For levels above about 60 dB SPL, the loudness of UEN can be approximated by a power law with an exponent of 0.23 (Fastl and Zwicker 2007). This means that, compared to narrowband sounds, the loudness of a broadband stimulus grows more slowly with sound level. The faster growth rate for narrowband sounds can be attributed to the upward spread of masking, whereas the loudness of broadband sounds is mainly caused by the main excitation per critical band. Figure 2.25 shows the loudness growth function of a UEN and of a 1 kHz pure tone (with N denoting loudness in the figure). Observe that both functions deviate from a power law (straight line): below 60 dB SPL for the UEN and below 40 dB SPL for the tone. Furthermore, the UEN is about 3.5 times louder than the tone when both sounds are at the same (moderate) sound level. This effect is known as spectral loudness summation (SLS).

[2] In this thesis, Ψ is used to represent the loudness of a stimulus in sone units. However, the symbol N is

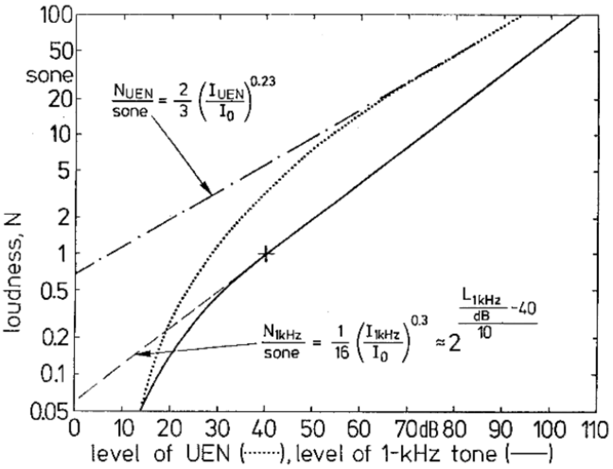


Figure 2.25: Loudness growth function of a 1 kHz pure tone (solid) and a UEN (dotted). Power-law approximations are shown by the dashed (pure tone) and dashed-dotted (noise) lines along with their respective equations. Image from Fastl and Zwicker (2007).

Spectral loudness summation

Natural sounds are complex in that they are not defined by a single frequency component, but instead are characterised by intricate spectral distributions. In order to measure the loudness of such sounds, it is important to investigate the effects of spectral bandwidth on loudness. For example, if the loudness of a multitone complex is equal to the sum of the loudnesses of the individual components, i.e. linear superposition holds, a simple and efficient model of loudness can be formed. Unfortunately, experiments have shown that summation of loudness across frequency depends on the distribution of the components.

Zwicker et al. (1957) conducted a loudness-matching experiment in which subjects were instructed to vary the level of a pure tone such that its loudness matched that of a four-tone complex stimulus (and vice versa). This was repeated with the average frequency (centre frequency) of the complex tone set to 0.5, 1 and 2 kHz, with the spacing of the four components varied symmetrically about each frequency. This experiment showed that the loudness of the complex tone is constant when its bandwidth is less than the critical bandwidth, and increases when its bandwidth exceeds that of the critical band. Figure 2.26 demonstrates this effect by plotting the matching data for the condition in which the centre frequency of the four-tone complex stimulus was 1 kHz. The figure also shows that SLS is dependent on sound level, with greater summation occurring at medium intensities than at low and high intensities. Zwicker et al. (1957) repeated this experiment but this time involving noise. A comparison sound (tone or noise) was matched in loudness to a bandlimited noise of constant overall level but variable bandwidth. The results supported the same conclusion: loudness increases when the bandwidth of the stimulus exceeds the critical bandwidth. Numerous later studies have confirmed these results (Scharf 1959; Scharf 1961; Florentine et al. 1978; Verhey and Kollmeier 2002; Leibold et al. 2007). A working model of SLS is one in which the intensities of all components within a critical band are summed and then compressed, before integrating the sub-band loudnesses across the perceptual frequency scale.

Temporal loudness integration

The effect of stimulus duration on loudness can be summarised as follows: constant energy leads to constant loudness for durations up to about 80 ms. This means that the perceived loudness of

commonly used in psychophysics following the pioneering study of Fletcher and Munson (1937).

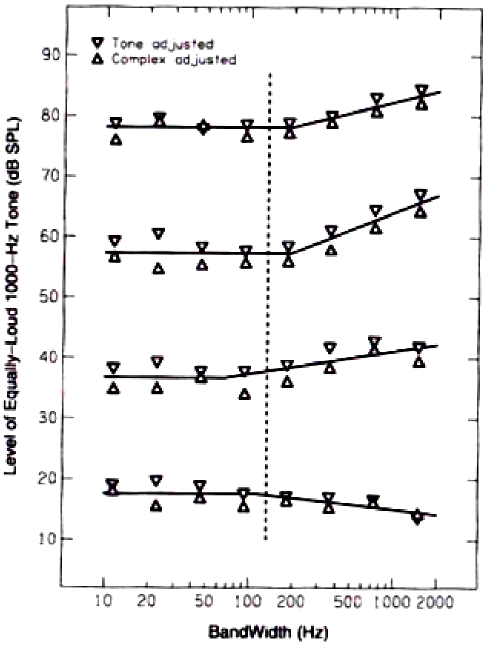


Figure 2.26: Loudness matching data from Zwicker et al. (1957) showing the level of a 1 kHz tone when as loud as a four-tone complex as a function of bandwidth. The parameter is the overall level of the complex tone. The lines were fitted to the data with the assumption that loudness was independent of bandwidth for bandwidths less than 100 Hz. The geometric mean of the breakpoints, shown by the dashed vertical line, is 129 Hz which is close to ERB at 1 kHz (132 Hz) (Glasberg and Moore 1990). Image from Jesteadt and Leibold (2011).

a short-duration sound, with a fixed sound level, increases with duration up to a critical point, after which the growth in loudness begins to diminish. This effect is greater for tones and noises at medium sound levels than at low and very high sound levels (Florentine et al. 1996; Buus et al. 1999).

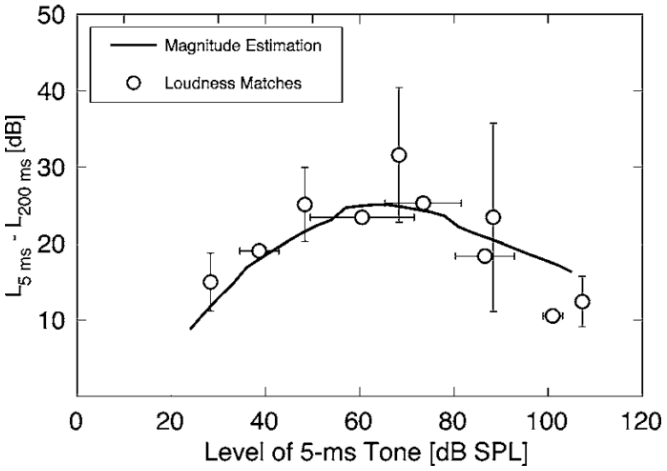


Figure 2.27: Level difference between a 5 ms and 200 ms pure tone at the point of equal loudness. The solid line corresponds to loudness matching data obtained by Epstein and Florentine (2005). The white circles are the averages from nine subjects using magnitude estimation (Epstein and Florentine 2006). Error bars indicate the SD of the means of nine listeners. Image from Epstein and Florentine (2006).

Figure 2.27 shows the amount of temporal integration, defined as the level difference between a 5 ms and a 200 ms pure tone at the point of equal loudness, as a function of level using loudness matching (Epstein and Florentine 2005) or magnitude estimation (Epstein and Florentine 2006). The positive level difference indicates that the level of the 5 ms tone must be higher than the level of the 200 ms tone in order for both tones to be equally loud. The bulge around 60 dB SPL suggests that temporal integration in the auditory system is level-dependent, being highest at medium intensities.

Florentine et al. (1996) suggested that the variation of temporal integration with level can be explained by the compression that takes place on the basilar membrane; the velocity of the basilar membrane response grows more slowly at medium sound levels than at low and high levels. If the loudness growth functions (log loudness versus stimulus level) for the short and long tones are assumed to be parallel, then the horizontal distance between the two functions would give the level difference at the point of equal loudness, i.e. the amount of temporal integration as a function of level. The vertical distance, describing the loudness ratio between the long and short tone at an equal SPL, would be constant at all levels. This property is known as the 'equal-loudness-ratio hypothesis' (Florentine et al. 1996). Figure 2.28 shows the loudness growth function of a 5 ms tone and of a 200 ms tone obtained by cross-modality matching using string length. In this paradigm, the subject is asked to cut a piece of string to a length proportional to the loudness of the tone. Both loudness functions appear to be parallel to one another and are shallower at moderate levels, consistent with velocity measurements of basilar membrane vibration. This explains the increase in level difference, i.e. increase in temporal integration, around 60 dB SPL observed in Figure 2.27. As shown by the heavier line, the loudness ratio between the 200 ms and 5 ms tone is approximately constant over a wide range of levels, and therefore supports the equal-loudness-ratio hypothesis. Importantly, these functions deviate from a simple power law, in a manner that is consistent with the input-output behaviour of the mammalian basilar membrane (Ruggero et al. 1997b; Robles and Ruggero 2001).

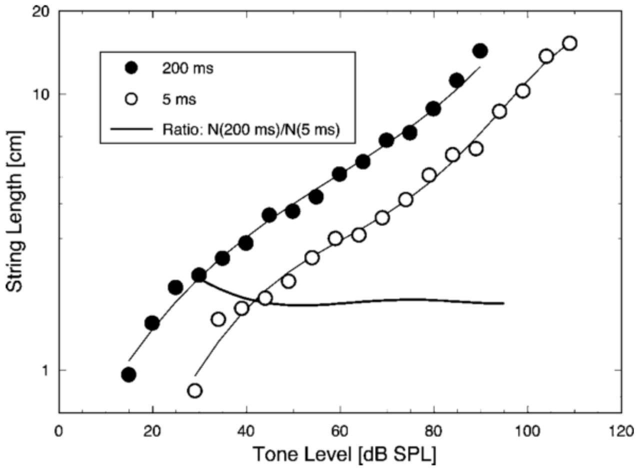


Figure 2.28: Loudness growth functions for short and long duration tones derived from cross-modality matching. The thick line is the vertical distance between the two functions and represents the loudness ratio between the two tones at equal SPL. Image from Epstein and Florentine (2005).

Effect of modulation

Everyday sounds, such as speech, music and roadside noise, fluctuate in complex ways. Unfortunately, the majority of studies conducted in the laboratory involve abstract stimuli of a steady nature. One approach to studying the effect of temporal fluctuations is to present the listener with a stationary sound, such as a pure tone, and an amplitude-modulated sound. The listener must adjust the level of one of the sounds such that both are equally loud. At PSE, the level difference ΔL between the RMS of the unmodulated stimulus and modulated stimulus is measured. A positive ΔL means that the modulation increased the loudness, whereas a negative difference indicates a decrease in loudness. An increase in loudness means that the loudness of the modulated signal is determined by a value higher than the RMS level, such as its peak. Glasberg and Moore (2002) reviewed a number of studies that investigated the loudness of modulated sounds (Bauch 1956; Fastl 1975; Hellman 1985; Zhang and Zeng 1997; Moore et al. 1998, 1999c) and although some findings were contradictory, perhaps influenced by measurement procedure, the following trends were drawn:

1. For signals modulated at rates below 10 Hz, the perceived loudness is determined by a level between the RMS and peak.

2. For modulation rates lower than about half of the critical bandwidth (at the carrier frequency), loudness corresponds to a level slightly below the RMS level.

3. For sinusoids modulated at high-modulation rates, the detection of spectral sidebands leads to SLS and therefore an increase in loudness. Because the critical bandwidth increases with centre frequency, this effect is dependent on the carrier frequency.

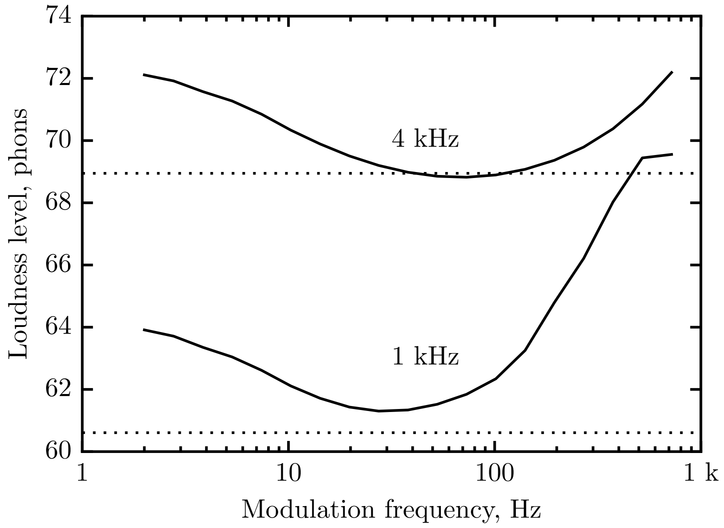


Figure 2.29: Loudness level of 100% (linearly) amplitude modulated 1 kHz and 4 kHz pure tones as a function of modulation frequency. The horizontal dotted lines show the loudness level of the unmodulated tones at 60 dB SPL (free field with frontal incidence). All tones had the same RMS level and 20 ms raised cosine ramps applied to their onset and offset. Loudness levels were predicted using the model of Glasberg and Moore (2002).

The above trends are depicted in Figure 2.29, generated using the loudness model of Glasberg and Moore (2002). The plot shows the loudness level (in phons) of a 1 kHz tone and a 4 kHz tone as a function of amplitude modulation rate. The horizontal dotted lines show the loudness level of the unmodulated tones, and all tones had an RMS level of 60 dB SPL. It can be seen that the loudness of the modulated tone increases at low- and high-modulation rates in accordance with empirical findings. Above intermediate rates, the loudness of the modulated 1 kHz tone rises before that of the 4 kHz tone because the ERB at 1 kHz is about 132 Hz compared to 456 Hz at 4 kHz (Equation 2.26). The sidebands corresponding to the 1 kHz carrier are therefore resolved by the auditory filters at lower modulation rates, leading to loudness summation across critical bands.

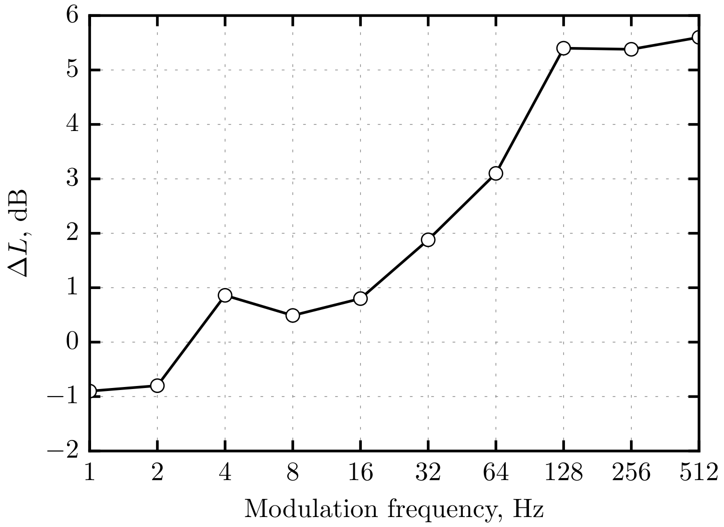


Figure 2.30: Level difference ΔL between unmodulated and equally loud modulated carrier at 1.5 kHz as a function of modulation frequency. Data determined from Figure 7 of Rennies et al. (2010).

In addition to amplitude modulation, frequency modulation also affects perceived loudness. For example, Zwicker (1974) measured the level difference required to match the loudness of a 1.5 kHz tone and a frequency-modulated tone with a carrier of the same frequency. The frequency deviation was 700 Hz, giving a minimum and maximum instantaneous frequency of 800 Hz and 2.2 kHz, respectively. The level of both sounds was 50 dB SPL. Figure 2.30 shows the loudness-matching data, where a positive ΔL means that the level of the stationary tone had to be increased in order to match the loudness of the frequency-modulated tone. For rates below about 16 Hz, the two tones are essentially equal in loudness. As the modulation rate increases, the loudness of the fluctuating sound increases, thus requiring a positive gain correction to maintain equal loudness between the two stimuli. At about 128 Hz, the level difference reaches an asymptote, with no further increase in loudness thereafter. Zwicker explained this effect in terms of the temporal integration occurring within a critical band. At low-modulation rates, the change in instantaneous frequency yields strong temporal fluctuation in a given critical band. As such, excitation is concentrated in one or two critical bands at any given time (Fastl and Zwicker 2007). At higher rates, post-masking plays a greater role in that the decay of filter activity is long relative to the modulation frequency. Thus, excitation is relatively constant in every auditory channel, leading to greater loudness summation.

Partial loudness

When two sounds with similar frequency content are presented at the same time, spectral masking occurs (see Section 2.3.2). The audibility of the target sound is dependent on the level of the masker; there is a transition from audible to inaudible as the level of the masker increases from low to high. The loudness of the target sound is therefore reduced (relative to loudness in quiet) as the level of the masker increases. This effect is known as 'partial masking of loudness', where partial loudness describes the loudness of a sound in the presence of other sounds.

Figure 2.31 shows loudness matches between a 1 kHz pure tone in masked and unmasked conditions obtained from a study by Steinberg (1937). The subjects were presented with a tone in

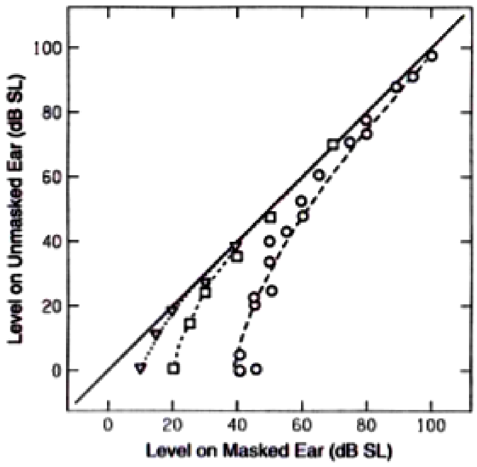


Figure 2.31: Loudness matching data for a 1 kHz tone in masked and unmasked conditions (Steinberg 1937). The straight line denotes the unmasked condition: equal loudness = equal level. The triangles, squares and circles correspond to threshold shifts of 10, 20 and 40 dB, respectively, as produced by a masking noise. Image from Jesteadt and Leibold (2011).

quiet in one ear and a tone in noise in the other. The level of the noise was set to produce either 10, 20 or 40 dB of masking. Listeners were instructed to adjust the level of one of the tones such that its loudness matched that of the other. As shown in the figure, the presence of the masking noise introduces a shift in the threshold of quiet (0 dB). As the level of the tone in noise increases above the masked threshold, loudness grows rapidly before converging to the loudness of the unmasked tone at higher levels. The partial masking effect might be explained by a complex partitioning of the auditory excitation evoked by the noise and target sound (Moore et al. 1997).

Binaural loudness summation

When a sound is presented to both ears (diotic presentation), the perceived loudness is greater than when the same sound is presented to one ear (monaural presentation). In a survey of data on binaural loudness summation, Moore and Glasberg (2007) concluded that a diotic sound is louder than the same sound presented monaurally, but by a factor of less than two. Following previous modelling proposals, e.g. Gigerenzer and Strube (1983), the authors employed the concept that the neural activity in one ear inhibits the activity in the other, and that this inhibition is broadly tuned across frequency. They developed a binaural loudness model tuned to give fairly accurate predictions of empirical data from a number of different studies.[3] The model predicts a diotic-to-monaural loudness ratio of 1.5, in agreement with studies in which binaural loudness summation was found to be less than perfect (Scharf 1969; Zwicker and Zwicker 1991; Edmonds and Culling 2009).

Intensity coding

Intensity discrimination refers to a listener's ability to detect intensity differences of sounds. A number of different measurement techniques have been used to measure the just noticeable difference or the difference limen (DL) for intensity, and the general trend is consistent across findings. The threshold of detecting changes in stimulus intensity is expressed in decibels, taking the form

$$\Delta L = 10 \log_{10} \left(\frac{I + \Delta I}{I} \right), \tag{2.29}$$

[3] This model is presented in Chapter 5.

where I is the intensity of the stimulus and ΔI is the increment in intensity required for threshold. For example, if the smallest change in intensity was a doubling of sound intensity ($\Delta I = I$), then the DL would be 3 dB. For broadband noise, the change in intensity is a constant proportion of stimulus intensity, i.e. $\frac{\Delta I}{I}$ is constant. This ratio is known as the Weber fraction and defines Weber's law: the intensity difference at threshold is proportional to intensity itself. Accordingly, the DL is also constant, regardless of the absolute level of the stimulus used to start with. For broadband sounds, the DL is about 0.5–1 dB and holds for levels greater than about 20 dB SPL (Houtsma 1980). For the pure tone and narrowband noise, however, the DL steadily decreases with stimulus intensity, ranging from about 1.7 dB near threshold, to 0.5 dB at 80 dB SPL (Riesz 1928; Houtsma 1980). This is known as the 'near miss to Weber's law' (McGill and Goldberg 1968). The near miss to Weber's law for pure tones is explained by the nonlinear changes in excitation with level (upward spread of masking) and the ability of subjects to aggregate information from different parts of the EP.

2.3.4 Spatial hearing

The ability to localise a sound object is primarily based on a comparison of signals arriving at the two ears. For example, a sound source located to the left of the head will impinge on the left ear before reaching the right ear. However, other cues must exist in order for individuals to distinguish between a sound arriving from directly in front or behind the head because the sound reaching both ears at the same time and level. Horizontal directions are expressed as angles of azimuth around the head. For example, a sound generated by an object positioned straight ahead would have an azimuth of 0°, and a sounds coming from directly behind the head would have an azimuth of 180°. An object located midway between these two positions on the right side of the head corresponds to an angle of 90°. Similarly, vertical directions are expressed as angles of elevation, with 0° straight ahead, 90° above the head, and 180° behind the head. Finally, a sound can be localised in the near-far dimension, allowing listeners to judge the distance of an object.

The primary auditory cues responsible for sound localisation are interaural time differences (ITDs) and interaural level differences (ILDs) (also called interaural intensity differences when dealing with intensity directly). Consider a sound source located at 45° in the horizontal plane at an azimuth of 0°. The sound reaching the left ear is delayed in time relative to the sound reaching the right ear due to the greater physical distance the sound must travel. This introduces an ITD in the arrival of the sound. In addition, because the sound arrives at the left ear after it reaches the right ear, it has travelled farther, meaning the sound intensity will be lower at the left ear as a result of the inverse square law. Furthermore, the head works as an acoustic barrier casting an acoustic shadow over the ear farthest away from the source, further reducing the level of the sound in the left ear. However, frequencies with wavelength exceeding the dimensions of the head are able to diffract around it to reach the opposite ear. Thus, ILDs are frequency dependent; the higher the frequency, the shorter the wavelength and thus the greater the cast of the head's acoustic shadow. In short, ILDs are useful cues for the localisation of high-frequency sounds. The idea that the auditory system makes use of ITDs at low frequencies and ILDs at high frequencies for localisation is called the duplex theory (Rayleigh 1907).

To understand why ITDs provide localisation information at low frequencies, the phase of the stimulus must be considered. The average ear-to-ear distance is roughly 22–23 cm, meaning that the maximum ITD is around 660 µs, occurring when a sound is 90° in the horizontal plane (presented at the opposite one ear). The ITD therefore varies with the azimuth of the source, peaking at 90° (Feddersen 1957). For frequencies below about 1500 Hz, the wavelength is greater than the diameter of the spherical head, and so the phase of the two waveforms at each ear can be compared to establish an interaural phase difference. However, the phase difference becomes ambiguous (after the first wavelength) as the frequency approaches 1500 Hz, where the wavelength approaches that of the distance between the two ears. Thus, ITDs become less effective for the localisation of tones above about 750 Hz where the wavelength is twice the diameter of the head.

Spectral cues generated by interactions between the sound source and head and torso play an important role in the perception of elevation, where ITDs and ILDs are constant and thus are ambiguous cues. These interactions modify the phase and magnitude spectrum of the incoming sound, especially at high frequencies where the wavelength is comparable or greater than the physical features of the head, e.g. the pinna. These spectral alterations depend on the position of the source relative to the body and can therefore be used as auditory templates for localising a source. They are best characterised by head related transfer functions (HRTFs) which are used to measure spectral contours at different spatial positions. Magnitude responses of transfer functions measured at four elevations for both ears of a subject are shown in Figure 2.32. The sound source was positioned at an azimuth of 0° or midway between the ears in the mid-sagittal plane and therefore the overall level in each ear was roughly the same, as was the distance between the source and each ear. The shape of the HRTF does change, however, with the elevation of the source, especially around 10 kHz where the size of the spectral valley decreases and shifts in frequency slightly as the source moves above the head. In summary, direction-dependent spectral contours provide important information about the location of a sound source, especially when ITDs and ILDs are ambiguous.

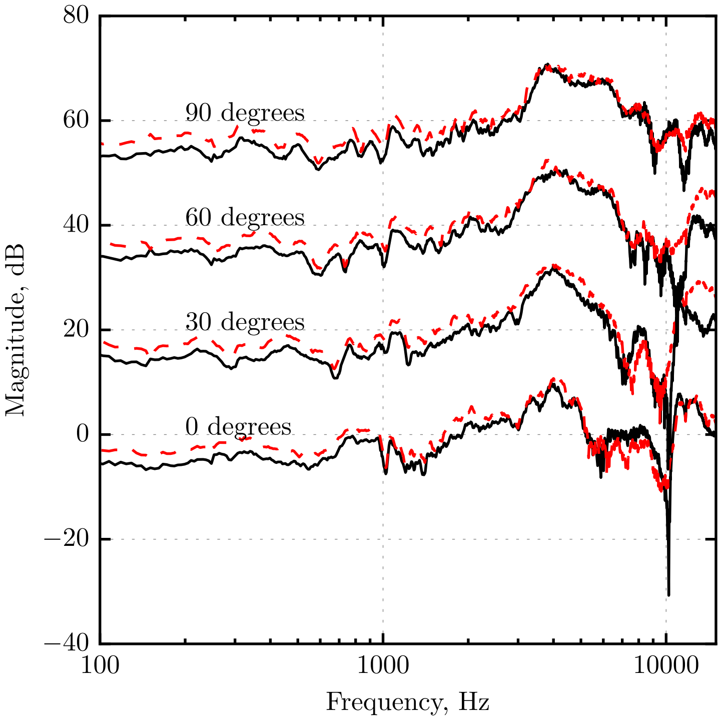


Figure 2.32: HRTF for the left ear (solid black lines) and right ear (dotted red lines) of a subject. There are four sets of binaural HRTFs corresponding to measurements made at different elevations with the source positioned at 0° in the mid-sagittal plane. Each binaural contour has been offset by +20 dB for visual clarity. The data are from IRCAM's Listen HRTF database (IRCAM 2003) and the image is based on Figure 12.9 of Yost (2007).

In everyday life, the perceived location of a sound source is influenced by the listener's environment. Complex interactions between the sound, room and listener provide additional cues for determining the spatial position of the source. For example, the ratio of direct sound to reflected sound can be used to estimate the distance of a sound source. Such reflections tend to be generated by multiple objects, thus taking an indirect route to the listener's ears. Despite hearing a direct sound followed by a series of successive related sounds (echoes) the individual perceives a single sound source coming from the direction of the sound arriving at the ears first. This is because the direct sound contains information about the location of the sound source, whereas reflections provide information about the listener's surroundings. This phenomenon is known as the precedence effect, Hass effect (Hass 1951), or the 'law of the first wavefront' (Blauert 1996). Experiments studying the perception of echoes, e.g. (Wallach et al. 1949), reveal that two sounds are perceived as a single sound if the time interval between them is, depending on the type of source, in the range of 5–40 ms and that the location of this fused sound is primarily determined by the location of the first sound. The precedence effect is only apparent for sounds of a discontinuous or transient nature and is most effective when both sounds have similar characteristics.

2.4 Summary

The peripheral auditory system is composed of the outer ear, middle ear and inner ear. The outer ear directs sound towards the tympanic membrane, modifying the spectrum of the sound through resonances introduced by its complex physical structures. The middle ear works as a mechanical impedance matching device, coupling the acoustic energy at the eardrum with the fluids and tissues of the inner ear. Sound transmission through the middle ear is frequency dependent, showing a bandpass characteristic centred near 1 kHz. The transmission response of the outer and middle ear is typically modelled using a linear filter with frequency-dependent gain, although the middle ear muscles do contract in response to loud sounds, thus introducing some level dependency in the response.

The cochlea is the main auditory component of the inner ear and is responsible for transforming vibrations at the stapes into neural activity processed by the brain. Along the length of the cochlea runs the basilar membrane, which moves in response to incoming sound. The place of maximum displacement of the travelling wave is frequency dependent, with low frequencies peaking near the apical end and high frequencies peaking close to the basal end. The inner hair cells are sheared by the interaction of motion between the basilar membrane and the tectorial membrane. This shearing causes neural discharges in the inner hair cells which contact afferent nerve fibres (sensory nerve fibres). The outer hair cells are motile and influence the tuning of the vibratory patterns on the basilar membrane such that sensitivity is increased and tuning is selective in frequency. The frequency selectivity of the inner ear can be examined by measuring the velocity of basilar membrane motion at a given site in response to pure tones at different frequencies, or by measuring the threshold of single auditory neurons as a function of frequency. In the latter case, the frequency that causes a single neuron to fire above its spontaneous rate at the lowest input level is called the characteristic frequency (CF) of that neuron. Neural tuning curves are bandpass in shape, showing a maximum sensitivity at CF of the nerve fibre and a highly tuned frequency selectivity at low input levels. Neurons with high CFs are excited by a wider range of frequencies compared to those with low CFs. Nerve fibres centred below 4–5 kHz are synchronised with the periodicity of tones of low frequency, as captured by phase-locked post-stimulus time histograms. The magnitude of a stimulating waveform is encoded from the combined response of many neurons, and phase and dynamic temporal information is captured by only those neurons that are able to synchronise with the stimulus. The presence of multiple tones can, however, reduce the average firing rate of a neuron as a result of the active response of the outer hair cells. This phenomenon is known as two-tone suppression, and the degree of suppression in single-unit discharge rate depends on the excitatory interactions on the basilar membrane.

The sensitivity of the hearing system is frequency dependent. This is revealed by audibility curves which plot detection threshold as a function of frequency. This function is similar in shape to the *inverse* transmission response of the outer and middle ear. The shape of this function is very much dependent on the method of presentation as well as the angle of incidence of the sound.

Masking describes the audibility of a sound in the presence of other sounds. The masked threshold defines the level of the target sound, typically a pure tone in psychoacoustics, required for detection in the presence of a masking sound such as bandlimited noise. Both masking patterns and psychophysical tuning curves are used to quantify the amount of masking at different frequencies. The general finding is that low-frequency tones are more effective at masking high-frequency tones than vice versa. This effect is revealed clearly in tonal masking patterns and is referred to the 'upward spread of masking'. The upward spread of masking can be modelled using the concept that the cochlea contains a bank of overlapping bandpass filters called the critical bands. Under a number of assumptions, collectively known as the power-spectrum model, the shape of the auditory filters can be systematically measured. The filter bandwidth increases with centre frequency, and becomes progressively asymmetrical with increasing sound level: the lower skirt becomes shallower than the upper skirt at high input levels. In essence, the auditory filter describes the frequency selectivity of the cochlea at a given place on the basilar membrane using psychophysical methods. When the power spectrum of a sound is fed through this filter bank, the result is a smeared version called an excitation pattern (EP) which approximates the internal representation of that sound. For tonal signals, the EP is asymmetric, especially at high levels, with the upper slope, about the peak excitation, being shallower than the lower slope. Many psychoacoustic phenomena can be explained using the concept of the critical band.

Loudness is a subjective quantity describing the perceived magnitude of a sound. Early experiments on loudness perception compared the loudness of tones at different frequencies. The results are typically summarised using equal-loudness contours which are similar in shape to the absolute threshold curve but become flatter with increasing sound level. This means that the perceived spectral balance of a complex sound strongly depends on the presentation level. In the 1950s, direct measures of loudness, obtained using techniques such as magnitude estimation, were used to establish an absolute scale of loudness called the sone scale. This scale forms a useful map relating the physical intensity of sound to the perceived magnitude sound, typically in the form of a power law. The spectral content of a stimulus also influences loudness: broadband sounds are louder than narrowband sounds when presented at the same overall intensity. This effect is known as spectral loudness summation and can be explained by the frequency-dependent compressive filtering that takes place in the cochlea. Finally, loudness increases with the duration of the stimulus up to a limit of about 80 ms. Although the findings of different studies vary, they do suggest that the loudness of short-duration sounds is based primarily on energy integration.

Finally, various localisation cues important for spatial perception have been discussed. Monaural listening refers to the presentation of sound in one ear alone. In binaural listening, both ears are used, and the sound at the two ears may be identical (diotic) or different (dichotic). In everyday listening, sound is perceived dichotically. The localisation of sounds in the horizontal plane is based on interaural time differences at low frequencies, and interaural level differences at high frequencies. For sounds in the vertical plane, differences in spectra at the two ears play an essential role in localisation. Such differences arise from the transmission of sound through the head and torso. Head movements may aid with the localisation of elevated sources based on a comparison of spectral templates. The perception of distance is most likely governed by source loudness and environmental cues such as reverberation.

Chapter 3

Loudness Models

Marozeau (2011) states that any model should have three objectives:

1. Describe a complex phenomenon with simple parameters.
2. Predict the output from known inputs.
3. Test hypotheses on mechanisms underlying a phenomenon.

Objective one refers to a model's ability to describe and simplify experimental data, without overfitting a particular dataset. If residual error is nonsystematic and cross-validation is satisfactory, the model can then be used to predict the response of the system to an input (objective two). In addition to prediction, models are often used to gain an understanding of the underlying process responsible for the output response. Thus, different mathematical models can be compared to test the explanatory power of different variables pertaining to specific scientific hypotheses.

The purpose of this chapter is to review models of loudness perception in normal hearing individuals. In the previous chapter a number of factors that affect perceived loudness were discussed, namely: sound intensity, frequency, spectral bandwidth, stimulus duration and amplitude/frequency modulation. Thus, a mature loudness model should account for the observed psychological effects invoked by the physical characteristics of sound. Objective models that are able to predict subjective phenomena serve a broad range of applications, many of which differ in the level of accuracy required. Thus, models of varying complexity have been proposed.

Loudness models can be divided into two types: single-band and multiband algorithms. The latter feature similar processing stages as the former, but with the added complexities of simulating mechanisms of the inner ear. More specifically, multiple frequency bands are used to analyse the incoming sound in a manner consistent with the frequency selectivity of the basilar membrane, as discussed in the preceding chapter. Models for stationary and nonstationary sounds exist within each category. Stationary models are those suitable for estimating the loudness of steady-state sounds, such as pure tones and noise, which can be defined by their long-term frequency spectrum. Dynamic models mimic the temporal characteristics of the ear in order to predict time-varying effects on loudness, such as the influence of stimulus duration and amplitude modulation. Stationary models have a long history and are well grounded in auditory theory whilst models for time-varying sounds are somewhat underdeveloped due to a comparative lack of empirical data. In this chapter we review both single-band and multiband loudness calculation procedures for stationary and nonstationary sounds, including detailed descriptions of sophisticated dynamic loudness models applicable to sounds with complex spectro-temporal behaviour. Before presenting and comparing the workings of these algorithms, a brief historical overview of the different approaches to loudness calculation is given.

3.1 Historical overview

Models describing the magnitude of an auditory sensation go as far back as the late 19th century where Fechner attempted to establish a mathematical relationship between the physical intensity of a sound and its perceived magnitude. Founded on Weber's work, Fechner used measures of the difference limen (DL) for intensity to create a scale of sensory magnitude, which subsequently led to the logarithmic law that states that subjective loudness is proportional to the logarithm of sound intensity (Fechner 1860). Many studies on the measurement of loudness were conducted in the early 20th century. Unfortunately, as shall be discussed, Fechner's law did not capture data from direct measurements of loudness, thereby failing to quantify perceived magnitude. One pioneering study was that of Fletcher and Munson (1933), who carried out a number of measurements using the method of adjustment (MOA), in which subjects were asked to vary the intensity of a 1 kHz tone until its loudness matched that of a reference stimulus. The *loudness level* (in phons) corresponding to the reference was set to the *intensity level* (in dB SPL) of the 1 kHz at the point of equal loudness.

This measurement was conducted for both pure and complex tones, presented to one or both ears. A series of loudness level growth functions (loudness level as a function of intensity level) and equal-loudness contours were derived from the empirical measurements. For example, it was shown that the level of a pure tone presented to one ear had to be 10 dB higher to match the loudness of the same tone presented binaurally. Thus, under the assumption of additivity in separate auditory channels, i.e. that loudness sums linearly across the ears, a 10 phon increase corresponds to a doubling of loudness. In contrast to Fechner's law, this suggests that loudness is a power function of stimulus intensity with an exponent of 0.3. Richardson and Ross (1930) used magnitude estimation, in which the listener is asked to assign numbers to sounds of different intensities, to obtain ratio measurements between tones that varied in frequency and level to a reference tone that was assigned a fixed numerical value. A series of similar experiments employing this technique followed, the findings of which were used by Stevens (1936) to establish a scale of loudness that he called the sone scale. This too was a power function relating sensory magnitude and signal intensity which was later revised in light of further empirical data (Stevens 1955, 1956b). The sone scale defines 1 sone as the loudness of a 1 kHz pure tone presented binaurally with frontal incidence in the free field at 40 dB SPL. The revised function became the International Organization for Standardization (ISO) recommendation 131 for calculating loudness (ISO 131 1959).

3.1.1 Multiband models for stationary sounds

Fletcher and Munson (1937) proposed the first multiband loudness model and use of the symbol N to denote total loudness, although Ψ is used here. The procedure is based on the formula

$$\Psi = \int F(M)\, dx, \tag{3.1}$$

where M is a measured threshold (in dB) in the presence of a masking sound relative to threshold at quiet at some position x on the basilar membrane, i.e. the amount of masking at a given frequency. When M is plotted as a function of frequency (or position x), a masking audiogram is produced. The function F describes the corresponding amount of excitation in the auditory nerves per unit length at that position. Fletcher and Munson assumed $F(M)$ to represent the loudness contributed per unit length of the basilar membrane, which today is called specific loudness (SL). Therefore, Equation 3.1 states that the total loudness of a sound is proportional to the area under the SL pattern invoked by that sound. The procedure requires a chart to convert the intensity spectrum of the incoming sound to a masking audiogram. A lookup table (LUT) is then used to convert the masking level in a given frequency band to a loudness value ($F(M)$), such that the loudness density can be integrated. The authors noted that their procedure was best suited to noises that approach a continuous spectrum rather than stimuli defined by discrete frequency components.

Following Fletcher and Munson (1933, 1937), several authors established methods for estimating the loudness of complex sounds (Beranek and Peterson 1948; Beranek et al. 1951; Mintz 1952). The equivalent-tone method involves decomposing the spectrum of a sound into different frequency bands and treating each intensity band level as that of a single pure tone, the loudness values of which are then estimated from the corresponding loudness levels. Similar to Fletcher and Munson's proposal, contributory loudnesses are summed to estimate the total loudness of the sound. The underlying assumption with these models is that the overall loudness is equal to the sum of the loudnesses corresponding to the frequency components comprising the sound. Stevens (1956a) questioned this assumption and presented empirical data showing the sum of band loudnesses to be greater than the loudness of the combined bands. He presented a formula that accounted for inter-band loudness inhibition. That is, the combination of loudnesses from multiple bands yields an increase in overall loudness, but each band reduces the contribution of neighbours nearby. This means that total loudness is not a simple sum of the per-band loudness values. The reduction in loudness caused by introducing additional energy bands is a consequence of frequency masking. A formula for loudness estimation using the concept of loudness inhibition along with various charts was provided for calculating the loudness of spectra defined in octave, half-octave and third-octave bands. In later work (Stevens 1957a, 1961a), various approximations were made to simplify the procedure, though the fundamental formula remained the same. These revisions led to the procedure used for part A of ISO 532 (1966) and the subsequent British standard BS 4198 (1967). In essence, the user was required to look up a set of loudness indices from a chart by inputting the octave-band levels of a sound reported by a spectral analyser. These indices were then used in the loudness summation formula to arrive at the total loudness in units of sones. A further chart was needed to convert from absolute loudness to a level in phons.

Part B of ISO 532 (1966) (and BS 4198 (1967)) is founded on the work of Zwicker (Zwicker 1960, 1961) who was, around the same time as Stevens, also studying the loudness of complex sounds (Zwicker and Feldtkeller 1955; Zwicker et al. 1957). Unlike Stevens' method which is entirely based on an octave-band specification, Zwicker divided the spectrum into the 24 critical bands of the auditory system (Zwicker 1961). Again, a graphical procedure was developed in which the user was required to specify third-octave band levels of a sound as the input, from which the critical-band loudnesses were estimated. Level-dependent spectral masking, which influences the total loudness estimate, were modelled by appending downward sloping curves to the critical-band levels. Zwicker termed the loudness per critical band the specific loudness given in sones. As with all multiband procedures, the area under the SL pattern was obtained through integration. Although Zwicker's approach required considerably more work from the user compared to that of Stevens', his method allowed for the estimation of arbitrary input spectra for sounds presented in both the free field and the diffuse field, and may be considered a more realistic simulation of peripheral auditory filtering. On the contrary, Stevens' method was limited to broadband spectra of sounds presented in the diffuse field.

Stevens and Zwicker's excitation pattern (EP) models formed the recipes for parts A and B of the first international standard ISO 532 (1975). As the years passed, part A of the standard lost practical interest whilst part B continued to develop. For example, Zwicker's method, which also became the German standard DIN 45631 (1967), was refined in 1991 (DIN 45631 1991) to improve predictions of the ISO 226 (1987) equal-loudness contours. A computer program written in BASIC was published along-side the revised German standard by Zwicker et al. (1991), making the procedure much more practical to implement compared to the preceding graphical methods.

More recent developments of stationary multiband loudness models can be found in the work of Moore and co-workers (Moore and Glasberg 1997; Moore et al. 1997). The 1996 publication outlines an updated version of Zwicker's model which employs a different outer- and middle-ear filter, an alternative technique for deriving neural EPs based on earlier modelling work (Moore and Glasberg 1983; Glasberg and Moore 1990), and a modified specific loudness transformation. The authors showed their new model accounted more accurately for the 1987 equal-loudness contours, predicted the loudness of partially masked sounds without requiring correction factors and avoided discontinuities in calculations. Amendments were later proposed by Moore et al. (1997) to improve predictions of binaural loudness, absolute thresholds, the shape of the equal-loudness contours at low frequencies and thresholds of complex sounds. This model can be considered more flexible than its predecessors in that the frequency response at the eardrum can be specified in numerous forms such as third-octave bands, discrete frequency components and bands of white or pink noise. The steady-state model became the basis for the ANSI S3.4 (2005). In 2006, Glasberg and Moore (2006) modified the middle-ear transfer function to improve predictions of absolute thresholds (ISO 389-7 2005) and better describe the updated equal-loudness contours (ISO 226 2003). This revision resulted in an update to ANSI S3.4 (ANSI S3.4 2007). In addition to examples of loudness predictions for a range of phenomena, the standard reports this model to predict absolute thresholds to within 0.2 dB of the ISO data and, at medium levels, the predicted loudness growth of a 1 kHz tone to be accurate to within 5% of the sone values reported in ISO 532 (ISO 532 1975). Taken together, this set of refinements to Zwicker's procedure is known as the Cambridge loudness model, because it was developed in Cambridge, UK.

At the time of writing, the ISO 532 is currently under development and two separate standards are being prepared (Scheuren 2014). The first (ISO 532-1) is based on Zwicker's method and is expected to be compliant with DIN 45631 (DIN 45631/A1 2010).[1] ISO 532-2 is based on the Cambridge loudness model but includes the binaural inhibition model proposed by (Moore and Glasberg 2007) for improving the loudness of dichotic sounds.

A more recent EP loudness model for steady-state sounds was published by Chen et al. (2011). Unlike previous approaches, the model does not require a separate transformation from excitation to SL. Instead, the filter bank used to perform the critical-band decomposition incorporates a compressive nonlinearity such that the output of each auditory filter, a compressed intensity, is proportional to SL. This means that the area under the EP can be used as a direct estimate of total loudness. As stated by Moore (2014) it is unrealistic to model the frequency selectivity and active compression of the auditory system as two separate stages. As discussed in Section 2.2.3, physiological measurements of basilar membrane responses to tones at a given cochlear site show compressive rates of growth at medium SPLs (Ruggero et al. 1997b; Robles and Ruggero 2001). The rate of growth is level dependent and hence correlated with the selectivity of the membrane at a given place. The proposed compressive filter employed by the model of Chen et al. (2011) is based on the concept that the response of the basilar membrane at a specific location is determined by a passive-wideband filter and an active-narrowband filter (Glasberg et al. 1999). The outer-ear transfer function follows that of ANSI S3.4 (2007), but the transmission response of the middle ear

[1] The most recent DIN standard is the same as the 1991 method but includes a procedure for estimating the loudness of time-varying sounds.

was adapted according to a fitting procedure which aimed to achieve a good fit to both absolute thresholds (ISO 389-7 2005) and notched-noise masking data (Baker et al. 1998; Glasberg and Moore 2000). Chen et al. (2011) demonstrated that the model could be used to predict absolute thresholds (ISO 389-7 2005), equal-loudness contours (ISO 226 2003), loudness as a function of level and as a function of spectral bandwidth. Predictions of absolute thresholds were not as accurate those given by the ANSI S3.4 (2007) model, although the authors stated that predictions of equal-loudness contours were comparable. The authors reported that their model provided a better account of the ISO 226 (2003) equal-loudness contours compared to Zwicker's model (DIN 45631 1991) which has been shown to give poor predictions (Charbonneau et al. 2009). This is to be expected considering Zwicker's method was originally developed around the time of the ISO 226 (1987).

3.1.2 Dynamic multiband models

The models discussed so far were designed to predict the loudness of stationary sounds, such as pure tones or noise with no long-term fluctuation in level. Such procedures are able to predict the way in which loudness varies with physical properties of sound and give a good understanding into the mechanisms underlying the peripheral auditory system. In comparison, dynamic models are designed to deal with time-varying sounds and are generally conceived by extending the steady-state procedure. This typically involves employing a time-frequency transform as the front-end to the stationary model and appending a temporal integration stage to the output. The prediction objectives include those of the stationary models, but also aim to account for the way in which loudness depends on stimulus duration and amplitude and frequency modulation.

In 1977, Zwicker published a model for temporally-variable sounds in the form of an analogue electronic loudness meter (Zwicker 1977). In order to conform to ISO 532 (1975), critical-band filtering was approximated by employing a bank of third-octave filters. The loudness calculation procedure is essentially the same as the standard except for the use of smoothing filters to simulate post-masking effects and overall integration of loudness over time. Spectral loudness summation (SLS) is modelled according to the stationary procedure, but a temporal integrator is used to smooth the total instantaneous loudness (IL). Zwicker demonstrated how the device could be used to quantify the loudness of tone bursts, narrowband noise, amplitude- and frequency-modulated tones, temporally-masked tone bursts and continuous speech. In most cases, Zwicker used the peak of the loudness-time function as the decision variable for estimating the overall loudness impression of each fluctuating sound.

Zwicker's dynamic model was the basis for the first amendment 'Calculation of the loudness of time-variant sound' of DIN 45631/A1 (2010), which is expected to form the procedure of the new ISO 531-1 (in preparation). Chalupper and Fastl (2002) refined Zwicker's 1977 model to predict the loudness of stationary and nonstationary sounds for normal and hearing-impaired listeners. The basic architecture is the same, except for a modification to the SL formulae in order to simulate hearing loss in individual listeners. In addition, 24 Gammatone filters equally spaced on the Bark scale are used rather than the third-octave filters employed by Zwicker.[2] With further refinements in signal processing, the authors showed that the model could explain a wide range of phenomena as measured empirically on subjects with and without hearing loss.

The initial revision to Zwicker's loudness model proposed by Moore and Glasberg (1997) was first adapted to tackle time-variant sounds in 1997 (Stone et al. 1997), but, following the refined

[2] Zwicker himself suggested critical-band filters, but simplified the design in order to conform with the original ISO 532 procedure.

stationary model (Moore et al. 1997), a more sophisticated algorithm was published five years later (Glasberg and Moore 2002). This dynamic model makes use of a fixed digital finite impulse response (FIR) filter to simulate the combined transfer function of the outer and middle ear, and six parallel fast Fourier transforms (FFTs) in order to approximate the time-frequency resolution of the hearing system. On a frame-by-frame basis, the multi-resolution power spectrum is then fed to the 1997 steady-state loudness procedure. After SLS, the total IL is smoothed by two temporal-integration stages to quantify the build up and decay of loudness of short-duration sounds and also the loudness of amplitude-modulated sounds. Unlike Zwicker's and Chalupper and Fastl's algorithms, Glasberg and Moore did not include a temporal-masking stage in each critical band. This means that, in some cases, for sounds with fluctuating spectra, the predicted amount of SLS could be less than that of the other models because there is no time decay of SL in each critical band. Indeed, in a study comparing Glasberg and Moore's dynamic model with that of Chalupper and Fastl's, Rennies et al. (2010) concluded that the latter is likely to provide more accurate predictions of loudness for signals with strong spectral variations over time. However, Rennies et al. also pointed out one of strengths of the Glasberg and Moore model is its ability to estimate the loudness of binaural sounds, since Chalupper and Fastl's model is limited to diotic sounds. Furthermore, the authors emphasised that the temporal integration stage employed by the former model is more sophisticated compared to other methods which is useful for estimating the loudness of different types of sounds. For example, Glasberg and Moore (2002) suggested using the peak of the fast changing short-term loudness (STL) for short-duration sounds and the mean of the more slowly fluctuating long-term loudness (LTL) when estimating the overall loudness of sounds such as speech.

The most recent auditory model for calculating the loudness of time-varying sounds is one proposed by Chen and Hu (2012) which is an elaboration of their steady-state procedure (Chen et al. 2011). The spectral decomposition and temporal integration stages are similar to those used by Glasberg and Moore's model, although the two were not compared. Predictions of temporal phenomena such as stimulus duration and amplitude modulation were demonstrated, as were predictions of the stationary effects as shown in their earlier publication.

3.1.3 Dynamic single-band models

Thus far, single-band and multiband approaches to loudness estimation have been discussed from the perspective of psychoacoustics. The formulae used in these algorithms are derived from fits to empirical data as measured in the laboratory. These experiments involve stimuli such as sinusoids and bandlimited noise which are seldom encountered in the real-world. Furthermore, the complex methods of the early models required charts and tables to arrive at the final measurement, making them impractical for everyday use. Sound level meters, dating back as far 1932 (Scott 1957), were developed as a convenient means to objectively quantify industrial and environmental noise. These early instruments made use of a microphone, amplifier and indicator to display peak and root mean square (RMS) pressure measurements of the incoming sound. Following knowledge of the equal-loudness contours, early developers started to experiment with different frequency-weighting filters to account for the sensitivity of the human ear at different frequencies. Such filters take the form of an inverse equal-loudness contour at a specified loudness level. The architecture of today's standardised meters has not changed: 'Generally, a sound level meter is a combination of a microphone, a signal processor, and a display device' (BS 61672 2003)[p.10]. The signal processor comprises an amplifier, a frequency-weighting function, a device to square the weighted signal and some form of temporal integration. Three weighting curves are typically used, two of which are the A and C weightings which approximate the inverse shape of the equal-loudness contours at low and high sound levels, respectively. The third (Z-weighting) is simply a broadband response. In BS 61672 (2013), the specifications of three kinds of meters are given:

- A device that measures exponentially-time-weighted frequency-weighted sound levels.
- A device that measures linearly-weighted time-averaged frequency-weighted sound levels.
- A device that measures frequency-weighted sound exposure levels.

Time-averaged measurements (linear temporal weighting) or the equivalent continuous sound level (L_{eq}), and are useful for recording the average power of a fluctuating sound over some defined time period. Traditional meters, employing exponential-time weighting, are problematic in many cases because of the continuously changing reading. Such devices are examples of single-band algorithms because no decomposition into frequency bands takes place. These objective measurements are displayed in decibels and thus do not represent an absolute measure of loudness. They are useful however, for approximating the relative loudness (as a level difference) between two sounds with similar long-term properties. Although originally developed for monitoring noise and establishing noise exposure regulations at work, measurements based on signal energy have proven successful when applied to loudness matching paradigms. For example, Soulodre and Norcross (2003) found high correlations between the predictions of different L_{eq} based metrics and a subjective dataset derived from an equal-loudness-matching experiment involving broadcast material. In a second study involving a more diverse dataset, Soulodre (2004) evaluated several loudness meters, including Zwicker's multiband model, and concluded that a frequency-weighted L_{eq} metric outperformed all other algorithms. In a separate but related study Skovenborg and Nielsen (2004b) evaluated twelve loudness models against a large corpus of loudness matches of music and speech as performed by listeners in two independent assessments (Skovenborg et al. 2004). Although a multiband loudness model developed by the authors ranked first in terms of minimum error across several performance metrics, they too found the L_{eq} approach to work very well and was found to outperform variants of Zwicker's model. Details of the authors' best-performing algorithm were not disclosed, but the procedure involved extracting different time-frequency features from the stimulus and combining them using a nonlinear function derived from a parameter tuning stage. Thus, although a filter bank was employed, the proposed model was not designed to simulate the peripheral auditory system.

In 2006, the International Telecommunication Union Radiocommunication Sector (ITU-R) published recommendation BS.1770 which specified an algorithm to measure audio programme loudness (ITU-R BS.1770 2006). This algorithm is based on a variant of the L_{eq} metric proposed by Soulodre. Over the past few years, this recommendation was superseded, with the most recent algorithm incorporating an adaptive gate to minimise quiet aspects of the signal biassing the measurement (ITU-R BS.1770 2015). Because of its simplicity and effectiveness as a loudness-matching device for broadcast sounds, the European Broadcast Union (EBU) published recommendation 128 in 2010 for loudness normalisation according to the BS.1770 procedure (EBU R 128 2010, 2014). The EBU essentially proposed a target loudness level for the purpose of average loudness normalisation and a supplementary descriptor of loudness dynamics called loudness range. As a supplement to R 128, EBU Tech 3341 (2016) presented the requirements for an 'EBU mode' compliant loudness meter such that loudness normalisation can be realised in live production. Owing to industry standardisation, the EBU meter has received wide-spread popularity amongst the audio engineering community and is manufactured by several leading audio companies.

3.2 Single-band models

This section takes a more detailed look into some of the most popular single-band methods for estimating the loudness of stationary and time-varying sounds.

3.2.1 Stevens' power law

The simplest and earliest models of loudness are those designed to estimate the perceived strength of pure tones as a function of sound intensity, typically expressed in decibels. Although logarithms are useful for expressing equal ratios as equal intervals, this decibel representation does not correspond to a linear scale of perceived loudness. In other words, loudness is not a linear function of sound level.

As discussed in Section 2.3.3, the DL is the smallest detectable change in a sensory stimulus. For example, measuring the change in stimulus intensity ΔI required to elicit the smallest perceptual difference is a measure of the intensity DL for that particular stimulus. Fechner (1860) assumed that perceived loudness could be calculated by counting the number of DLs, leading to the following single-band model:

$$\Psi = kL, \tag{3.2}$$

where k is a constant and L is intensity level. The problem is that one intensity DL does not have the same perceptual magnitude as another, and thus the logarithm model cannot be applied.

Stevens (Stevens 1955, 1957b, 1961b) proposed that subjective loudness grows as a power function of stimulus intensity (not level) raised to the power 0.3. Although this model was introduced in the previous chapter (Equation 2.28), the power law is repeated here for completeness:

$$\Psi = kI^{0.3}. \tag{3.3}$$

This power function has been found to hold in the case of pure tones at medium sound levels and differs from Fencher's law which states that loudness is proportional to log intensity. As shown by the solid line in Figure 3.1, on log-log co-ordinates, the result is a straight line, meaning that loudness ratios are proportional to intensity ratios or increments in level. A common example is that the loudness of 1 kHz pure tone doubles for every 10 dB increase in intensity. The slope of log-intensity vs log-loudness is therefore 0.3 (see Marozeau (2011) for a derivation of this). As shown by the ordinate label, loudness sensation is specified in units of sones. Each curve describes the relationship between sones and the level of a 1 kHz pure tone presented binaurally in the free field with frontal incidence. Note that the constant k in Equation 2.28 is configured such that 1 sone corresponds to the binaural loudness of a 1 kHz sinusoid at 40 dB SPL.

Florentine and Epstein (2006) point out two known deviations from Stevens' power law:

1. The loudness function is steeper near absolute threshold.

2. The loudness function is less steep at moderate sound levels than at low and high levels.

Consequently, alternative models such as the inflected exponential function (Buus and Florentine 2002) have been proposed to better describe how the loudness of a tone grows with intensity. This concept however, is not new. Stevens (1966), for example, suggested the absolute threshold corrected power law

$$\Psi = k\left(I - I_{\mathrm{ThQ}}\right)^{0.3}, \tag{3.4}$$

where I_{ThQ} is the intensity of the tone at threshold in quiet, owing to the presence of internal noise. This has the effect of outputting zero when the input intensity matches the intensity at threshold. Zwislocki and Hellman (1960) presented an alternative model that compresses both signal intensity and the intensity of the sound at threshold separately, prior to the subtraction,

$$\Psi = k\left(I^{0.3} - I_{\mathrm{ThQ}}^{0.3}\right). \tag{3.5}$$

At high intensities, all three functions are similar, but differ as the intensity of the sound approaches the hearing threshold. The final equation commonly cited alongside these functions is the one given by Zwislocki (1965), which subtracts the loudness contributed by internal noise from the total loudness,

$$\Psi = k\left(I + 2.5I_{\mathrm{ThQ}}\right)^{0.3} - 2.5I_{\mathrm{ThQ}}^{0.3}. \tag{3.6}$$

The loudness-growth functions described above are all plotted in Figure 3.1. The value of I_{ThQ} was set to give the reference threshold of 2.4 dB SPL (ISO 389-7 2005). Equation 3.6 is similar to Equation 3.5 except near the threshold where Zwislocki's formula predicts greater loudness. The loudness function used to derive the standardised model for equal-loudness contours (ISO 226 2003) is based on Equation 3.5 but employs an exponent of 0.27 as typically obtained from measurements using magnitude estimation.

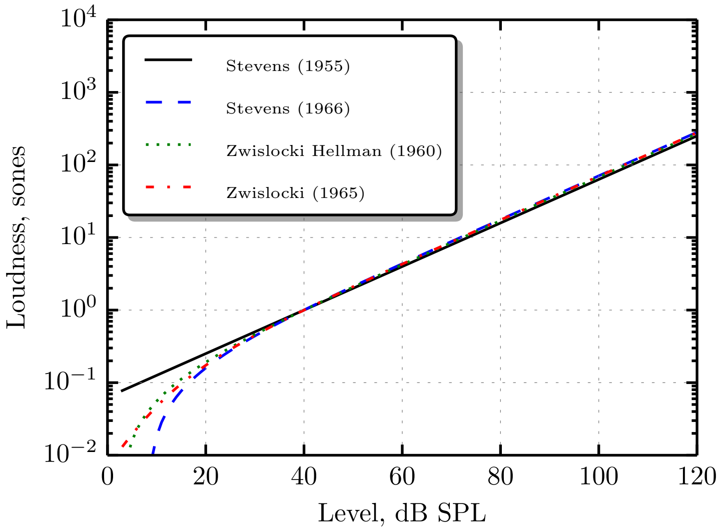


Figure 3.1: Loudness models for a 1 kHz pure tone defined as a compressive function of sound intensity. The functions cited in the legend, from top to bottom, are defined by Equations 3.3-3.6. The parameters were set to give an absolute threshold of 2.4 dB SPL (ISO 389-7 2005) and such that 40 dB SPL corresponds to exactly 1 sone.

3.2.2 Equivalent continuous sound level

The effect of frequency on loudness is illustrated by the equal-loudness contours (refer to Figure 2.24). The listener is first presented with a 1 kHz sinusoid at a known sound level and then the frequency is changed. The listener is then asked to adjust the intensity of the tone until the perceived loudness matches that of the 1 kHz tone. When this measurement is carried out with the variable tone set to a range of frequencies, an equal-loudness contour is generated. Because the sone scale is configured such that a 1 kHz tone at 40 dB SPL corresponds to 1 sone, there is a direct mapping between loudness level and loudness in sones. For pure tones at medium and high

SPLs, this relationship can be approximated as

$$L_{phons} = 40 + 10\log_2 \Psi. \tag{3.7}$$

The fact that loudness also varies with frequency necessitates a model that captures the sensitivity of the ear across the audible spectrum. As discussed, a simple approximation of such a filter is an inverse equal-loudness contour. Commercially available single-band models offer various frequency weightings to approximate the shape of an inverse loudness at a given loudness level. For example, the A, B and C frequency weightings employed by traditional SPL meters are based on the inverse shapes of the 40, 70 and 100 phon contours, respectively. In practice, these functions are realised as time-domain filters, followed by a time-averaged intensity measurement before being converted to decibels for display. For example, the $\mathrm{L_{eq}}$ is defined as

$$L_{eq_{T,w}}[n] = 10\log_{10}\left(\frac{1}{Tx_0^2}\sum_{m=0}^{T-1} x_w^2[n-m]\right), \tag{3.8}$$

where subscript w denotes the frequency weighting used and $x_w[n]$ is the frequency-weighted input signal. x_0 is the reference constant, and T defines the time interval over which linear averaging takes place. $\mathrm{L_{eq}}$ is a frequency-weighted RMS-level detector implemented as a simple moving average filter. Thus, if no frequency weighting is used and the measurement period T is equal to the duration of a signal, $\mathrm{L_{eq}}$ is equivalent to the RMS level of that signal. Alternatively, continuous sound level can be computed using the exponential moving average (EMA) filter applied to the square of the input with

$$l_{\tau,w}[n] = \alpha x_w^2[n] + (1-\alpha) l_{\tau,w}[n-1], \tag{3.9}$$

$$L_{\tau,w}[n] = 10\log_{10} l_{\tau,w}[n], \tag{3.10}$$

where τ denotes the time constant of the exponential in seconds. The coefficient α is calculated using the relation

$$\alpha = 1 - e^{\frac{-1}{f_s \tau}}, \tag{3.11}$$

where f_s is the sampling frequency in Hz. As the time constant increases (leading to greater smoothing) the value of the coefficient approaches zero, minimising the contribution of new input values. There are three time constants typically used in an exponentially time-weighted SPL detector (Matelijan 2014):

1. Slow: 1 000 ms.
2. Fast: 125 ms.
3. Impulse: 35 ms (attack phase), 1500 ms (decay phase).

Figure 3.2 shows the A, B and C frequency-weighting curves, along with the step responses of an EMA filter for the three smoothing configurations. For the Impulse setting, the attack and decay phase of the integrator are configured such that the detector reacts quickly to rising inputs but applies greater averaging as the signal level drops, yielding a sort of long-lasting memory effect. Both linear and exponential integrators are RMS detectors. In contrast to Stevens' model (Equation 3.3), such level meters display readings on a logarithmic scale and therefore do not describe loudness in absolute terms. However, according to the power law, one may loosely (noting the deviations from the power law) interpret equal increments in sound level as equal-loudness ratios, at least for narrowband sounds.

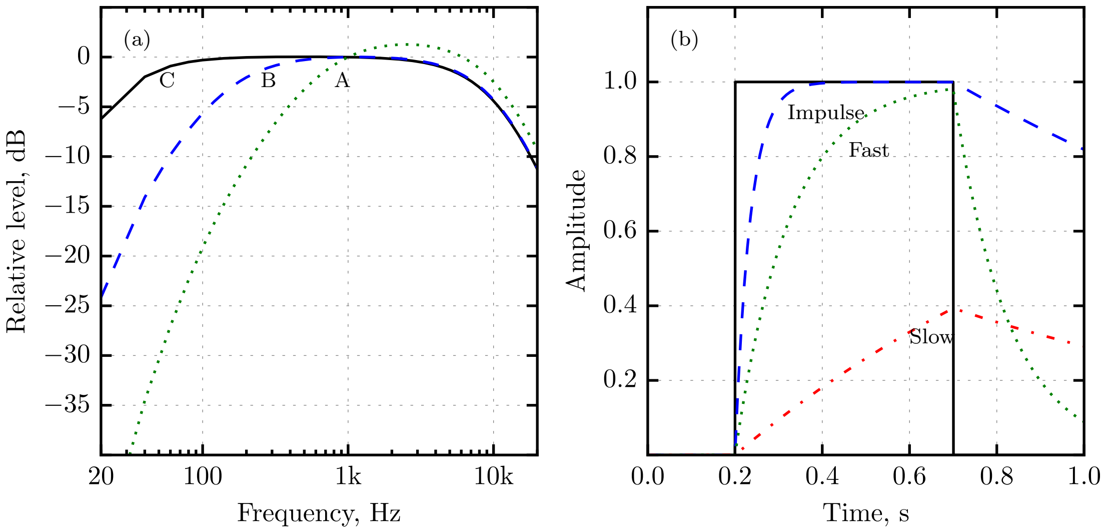


Figure 3.2: Subplot (a) shows, from left to right, the C (black solid), B (blue dashed) and A (green dotted) frequency-weighting functions as commonly used in a time-integrating SPL meter. The step responses of the EMA filter, as used to estimate sound level, with the time constant set to Impulse (blue dashed), Fast (green dotted), and Slow (red dashed-dotted) are shown in subplot (b). The solid line shows the step function input to the integrator.

3.2.3 ITU-R BS.1770

In 2003/2004 the Special Rapporteur Group (SRG) within the ITU-R investigated objective methods of measuring the perceived loudness of programme material typically used in broadcasting. Soulodre made significant contributions to this body of work, in which he developed a testing procedure for measuring the subjective loudness of monophonic sounds (Soulodre et al. 2003). These experiments were later conducted at numerous sites to establish a larger subjective database for evaluating several loudness models (Soulodre 2004). Soulodre's performance evaluation of the models revealed that a simple $\mathrm{L_{eq}}$ using a frequency-weighting function known as the revised low-frequency B curve (RLB) gave the best fit to the data. This work led to a standardised algorithm for predicting the subjective loudness of multichannel programme audio (ITU-R BS.1770 2015). It should be stressed, however, that these algorithms are not measuring loudness per se, but are essentially frequency-compensated energy measurements expressed in dB. This is useful for loudness alignment in broadcasting and related areas where it is crucial to control for variations in loudness across a range of commercial programmes, thus stabilising the sensory experience of the end listener. As a result, the ITU-R highlighted that the BS.1770 algorithm was designed for programme material and is not suitable for estimating the relative loudness of pure tones. Although the BS.1770 algorithm originally stemmed from evaluations involving only monophonic signals, later studies found the meter to perform well on stereo and multichannel signals (Soulodre and Lavoi 2005; Seefeldt and Lyman 2006). The algorithm is now presented in detail.

Each input channel x_i of a five channel surround sound system, is passed through a cascade of two biquad filters: a *pre-filter* which is a highshelf filter designed to mimic the acoustic influence of the head, followed by the RLB highpass filter.[3] The filter cascade is termed the K-filter. The magnitude response of the K-filter is shown in Figure 3.3.

The filter coefficients given in the technical report are only specified for a sampling frequency f_s

[3] $i \in \{L, R, C, L_s, R_s\}$ corresponds to the left, right, centre, left surround and right surround speakers, respectively. A five-channel signal is not compulsory: the algorithm simply supports up to five channels.

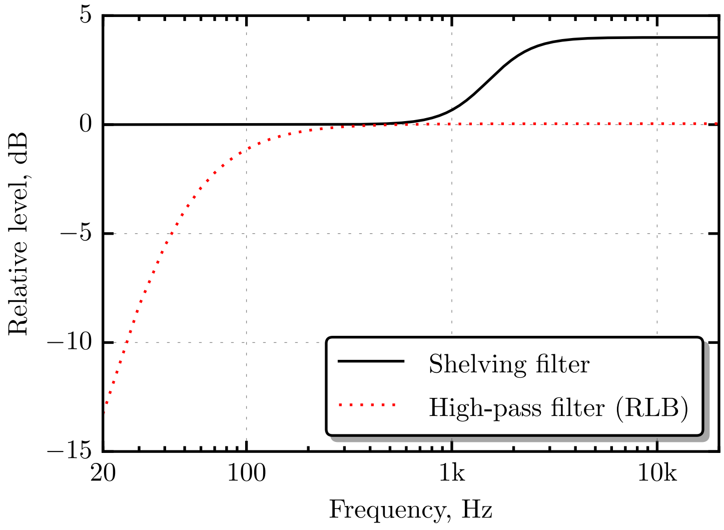


Figure 3.3: Magnitude response of the K-filter as used in ITU-R BS.1770 (2015), comprised of a shelving filter (black solid) and the RLB filter (red dotted).

of 48 kHz, but other sampling rates can be supported by acknowledging the continuous biquadratic transfer function of a second-order filter (Neunaber 2008)

$$H(s) = \frac{V_H s^2 + V_B \frac{\omega}{Q} s + V_L \omega^2}{s^2 + \frac{\omega}{Q} s + \omega^2}, \tag{3.12}$$

where V_H, V_B and V_L denote the highpass, bandpass and lowpass gain, respectively. Q is the quality factor and ω is the centre or cut-off frequency f_c in radians per second, depending on the filter type. The filter coefficients for the second-order filters given in (ITU-R BS.1770 2012) are specified using the form

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}}, \tag{3.13}$$

so it is necessary to map the transfer function of Equation 3.12 to the digital domain by substituting $s = \frac{1-z^{-1}}{1+z^{-1}}$ and $\omega \rightarrow \Omega = \tan(\pi \frac{f_c}{f_s})$. This yields the rational transfer function

$$H(z) = \frac{\left(V_L \Omega^2 + V_B \frac{\Omega}{Q} + V_H\right) + 2\left(V_L \Omega^2 - V_H\right) z^{-1} + \left(V_L \Omega^2 - V_B \frac{\Omega}{Q} + V_H\right) z^{-2}}{\left(Q^2 + \frac{\Omega}{Q} + 1\right) + 2\left(\Omega^2 - 1\right) z^{-1} + \left(\Omega^2 - \frac{\Omega}{Q} + 1\right) z^{-2}}. \tag{3.14}$$

As the filter coefficients are known for $f_s = 48$ kHz, a system of five equations with five unknowns for the above variables can be solved (Neunaber 2008):

$$f_c = \frac{f_s}{\pi} \arctan\left(\sqrt{\frac{1 + a_1 + a_2}{1 - a_1 + a_2}}\right), \tag{3.15}$$

$$Q = \frac{\sqrt{\left(a_2 + 1\right)^2 - a_1^2}}{2|1 - a_2|}, \tag{3.16}$$

$$\tag{3.17}$$

$$V_L = \frac{b_0 + b_1 + b_2}{1 + a_1 + a_2}, \tag{3.18}$$

$$V_B = \frac{b_0 - b_2}{1 - a_2} \quad \text{and} \tag{3.19}$$

$$V_H = \frac{b_0 - b_1 + b_2}{1 - a_1 + a_2}. \tag{3.20}$$

Given a new sampling frequency, the filter coefficients are then determined as follows

$$b_0 = V_L \Omega^2 + V_B \frac{\Omega}{Q} + V_H, \tag{3.21}$$

$$b_1 = 2\left(V_L \Omega^2 - V_H\right), \tag{3.22}$$

$$b_2 = V_L \Omega^2 - V_B \frac{\Omega}{Q} + V_H, \tag{3.23}$$

$$a_0 = Q^2 + \frac{\Omega}{Q} + 1, \tag{3.24}$$

$$a_1 = 2\left(\Omega^2 - 1\right) \quad \text{and} \tag{3.25}$$

$$a_2 = \Omega^2 - \frac{\Omega}{Q} + 1, \tag{3.26}$$

which can then be normalised by a_0 to realise the transfer function of Equation 3.13. The variables characterising the constituent second-order filters of the K-filter are given in Table 3.1.

	Highshelf (pre)	Highpass (RLB)
f_c	1681.9744510	38.1354709
Ω	0.1105318	0.0024960
Q	0.7071752	0.5003270
V_L	1	0
V_B	1.2587209	0
V_H	1.5848647	1.0049949

Table 3.1: Parameters of the highshelf and highpass filters defining the K-filter used in ITU-R BS.1770 (2015) for $f_s = 48$ kHz. Given a new sampling frequency, the only parameter that needs updating is $\Omega = \tan(\pi \frac{f_c}{f_s})$.

After filtering the i^{th} input channel, the average power of the filtered signal y_i is measured in blocks of length $T = 400$ ms, to the nearest sample. The j^{th} block overlaps with the previous block by 75%, giving a hop size of $H = 100$ ms. The average power of block j is computed as

$$P_{i,j} = \frac{1}{T} \sum_{n=jH}^{jH+T-1} y_i[n]^2. \tag{3.27}$$

The total average power for a given block is thus

$$P_j = G_i \sum_{i=0}^{I-1} P_i, \tag{3.28}$$

where G_i is a channel weight. The gain factors are unity for the left, right and centre channels and 1.41 for left and right surround speakers. The next step is to measure gated loudness by summing powers over all blocks that exceed an absolute threshold τ_{abs} of -70 dB (with units loudness K-weighted relative to Full Scale (LKFS)) and transforming the result to decibels:

$$L_{abs} = 10 \log_{10} \left(\frac{1}{K_{abs}} \sum_{j \in J_{abs}} P_j \right) - 0.691. \tag{3.29}$$

In this equation J_{abs} is the set of indices where the block power exceeds the absolute threshold of -70 LKFS and K_{abs} is the number of indices. The -0.691 dB offset is to compensate for the response of the K-filter at 1 kHz. A relative threshold τ_{rel} is then taken as the gated loudness (a scalar) minus 10 dB:

$$\tau_{rel} = L_{abs} - 10, \tag{3.30}$$

and the final gated loudness is recalculated but this time using the relative threshold determined from the input signal:

$$L_{rel} = 10 \log_{10} \left(\frac{1}{K_{rel}} \sum_{j \in J_{rel}} P_j \right) - 0.691. \tag{3.31}$$

Thus, only the powers of blocks above the relative threshold contribute to the final measurement. The purpose of this second threshold is to reduce the influence of very quiet periods influencing the average. This gating procedure was initially proposed by the EBU which was later adopted by the ITU-R for BS.1770. The final reading is expressed in units of LKFS, which is equivalent to a decibel scale: a 1 dB increment in sound intensity corresponds to an increase of 1 LKFS. The EBU stipulated that the unit be called Loudness Units relative to Full Scale (LUFS) rather than LKFS, based on inconsistencies between ITU-R BS.1770 (2006) and ITU-R BS.1771 (2006), and also to establish consistency with the conventions used by other standards. However, the ITU-R did not change their naming, so LKFS and LUFS are equivalent in practice. As a final note, when describing the relative loudness of two sounds, the level difference is given in Loudness Units (LU), e.g. sound A is 3 LU louder than sound B.

The EBU later proposed the ‘EBU-mode’ meter (EBU Tech 3341 2016) for real-time loudness monitoring when measurements are performed according to ITU-R BS.1770 (2015). In particular, three measurement methods were proposed:

1. momentary loudness (ML): using a sliding rectangular time window of 0.4 seconds. The measurement is not gated.

2. STL: using a sliding rectangular time window of 3 seconds. The measurement is not gated and update rate should be at least 10 Hz.

3. Integrated loudness: employs the gating procedure described above. The update rate should be at least 1Hz.

The integrated loudness performs the BS.1770 measurement in an online setting, requiring the relative threshold to be updated with every new block. When applied to an audio programme of fixed length, the final reading is therefore equivalent to the BS.1770 measure (a scalar), and is called the programme loudness since it is the average loudness of the entire programme. In addition to the three time-scale requirements, the EBU proposed a supplementary descriptor called the Loudness Range which quantifies the variation in loudness of time-varying signals (EBU Tech 3342 2016).

3.2.4 TC Electronic’s LARM

Following initial studies conducted by the SRG within the ITU-R, Skovenborg et al. (2004) carried out a series of loudness-matching experiments involving a broad range of speech and music. 12

loudness models, made up of both single-band and multiband models, were later evaluated using the subjective dataset obtained from the experiments (Skovenborg and Nielsen 2004b). One of top two best performing models was a new single-band algorithm proposed by TC Electronic called LARM. Interestingly, LARM outperformed the $\mathrm{L_{eq}}$(RLB) algorithm which ranked first in Soulodre's study (Soulodre 2004).

LARM was implemented by combining the properties of a peak programme meter (PPM) and the frequency-weighted $\mathrm{L_{eq}}$ metric. The PPM is essentially an envelope follower with a fast attack time and slow release time, such that peaks in the envelope are accentuated. An asymmetrical lowpass filter can be used to mimic this behaviour and thus extract the amplitude envelope x_{env} of a given input x:

$$x_{env}[n] = \alpha|x[n]| + (1-\alpha)|x[n-1]|, \tag{3.32}$$

where the value of the smoothing coefficient α depends on the state of the input signal. If the signal is rising ($|x[n]| > |x[n-1]|$) the attack coefficient is used, otherwise the release coefficient is used. The time constants characterising LARM were not explicitly reported by Skovenborg and Nielsen (2004b), but a reference was made to the German DIN 45406 standard (DIN 45406 1966). This standard specifies a peak level of -1 dB full scale (FS) for a full scale tone-burst of 10 ms, and -4 dB FS for a 3 ms tone-bust (Texas Instruments 2013). From experimenting with different time constants, these criteria are well approximated with an attack time of 14.5 ms and a release time of 650 ms. After extracting the amplitude envelope in this way, the overall loudness measurement according to LARM is calculated using

$$L = 20\log_{10}\sqrt[p]{\frac{1}{N}\sum_{n=0}^{N-1}|x_{env,w}[n]|^p}. \tag{3.33}$$

In this equation, $x_{env,w}$ is the envelope of the frequency-weighted input signal, with subscript w denoting the weighting curve, such as the RLB. Although the exponent p is a free parameter, the authors discovered a value of 1.5 after fitting the algorithm to the subjective data, and highlighted that this value yields a quantity between the average magnitude and the RMS value of the input signal.

3.3 Multiband models

As discussed in Section 3.1.1, multiband models were developed to account for more complex spectral effects on loudness perception. For example, when the frequency interval between two pure tones at a constant intensity exceeds the critical bandwidth, the perceived loudness begins to increase until it reaches the sum of their individual loudnesses. This phenomenon is called SLS and describes how loudness depends on the frequency distribution of complex sounds (Zwicker et al. 1957). However, spectral masking also plays an important part, because a strong tonal component may well mask other frequencies in neighbouring critical bands. It is therefore assumed that loudness is dependent on the amount of neural activity transmitted to the brain (Fletcher and Munson 1933). Thus, unlike the aforementioned single-band models, further to frequency weighting, the signal is decomposed into a series of frequency bands using a bank of auditory filters. The per-band intensities are compressed in order to approximate the effective neural excitation transmitted from the cochlea.

3.3.1 Zwicker's model

Fundamental approaches to the modelling of SLS are founded on the work of Zwicker (Zwicker 1958; Zwicker and Scharf 1965; Fastl and Zwicker 2007) which formed part B of ISO 532 (1975) standard and the German standard DIN 45631 (1991). As discussed, the model operates on the long-term spectrum of the input signal and is thus limited to stationary sounds. The general structure of Zwicker's loudness model is given in Figure 3.4.

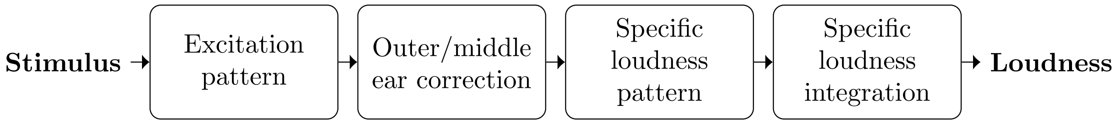

Stimulus
Excitation pattern
Outer/middle ear correction
Specific loudness pattern
Specific loudness integration
Loudness

Figure 3.4: Block diagram of Zwicker's loudness model (Zwicker 1958; Zwicker and Scharf 1965; Fastl and Zwicker 2007).

Cochlear EPs are representations of the distribution of energy along the basilar membrane. Zwicker and Scharf (1965) derived EPs corresponding to arbitrary input spectra using standard masking patterns for narrowbands of noise with bandwidths less than or equal to a critical band. In their model, sounds wider than a critical band are represented by combining these basis masking patterns. According to part B of ISO 532 (1975), the input spectrum is specified in third-octave bands, approximating the critical-band filters above 300 Hz (Fastl and Zwicker 2007). At lower frequencies, the critical-band levels have to be modified by combining the two or three third-octave bands. Refer to Section 2.3.2 for a review of critical bands.

Let $E(i)$ denote the total intensity within critical-band i. In this implementation, the spread of energy along the basilar membrane is modelled at the SL stage, so here, E is said to be the *main* excitation per critical band. The transfer function of the outer and middle ear is modelled using a fixed filter. The model provides a choice for either free- or diffuse-field presentation. For practical convenience, the linear filter defining the presentation mode is applied directly to the EP. The frequency-weighted EP is then transformed to an SL pattern by applying a psychophysical equation similar to Stevens' power law (Equation 3.3). In this model, SL represents the loudness per critical band with units sone/Bark, and for levels well above the absolute threshold is given by

$$\psi(i) = \left(\frac{E_c(i)}{E_0}\right)^{\alpha}, \tag{3.34}$$

where $E_c(i)$ and $\psi_(i)$ is the outer- and middle-ear corrected excitation and the SL in the i^{th} critical band, respectively, and E_0 is the excitation corresponding to a reference intensity of $10^{-12}\mathrm{W\,m^{-2}}$. α is a compressive exponent equal to 0.23, i.e. SL is approximately the fourth root of the normalised excitation. This value is based on the observed loudness-growth function of a broadband noise that equally excites every critical band called uniform exciting noise (Fastl and Zwicker 2007). Notice that the exponent is lower than the one used for characterising the rate of growth in loudness of a 1 kHz tone, which employs an exponent of 0.3.

As discussed in Section 3.2.1, the function describing loudness growth is steeper at very low levels and thus the slope of the SL function at critical-band levels near threshold in quiet E_{ThQ} should also increase:

$$\psi(i) = \beta\left(\frac{E_{\text{ThQ}}(i)}{s(i)E_0}\right)^{0.23}\left[\left(1 - s(i) + s(i)\frac{E_c(i)}{E_{\text{ThQ}}(i)}\right)^{0.23} - 1\right]. \tag{3.35}$$

In this equation $s(i)$ is the ratio between the intensity of the just-audible test tone and the intensity of the internal noise within the critical band centred at the test tone's frequency (Fastl and Zwicker 2007). For frequencies in the vicinity of 1 kHz, $s = 0.5$. The constant β is used to scale the function to give 1 sone for a 1 kHz sinusoid presented at 40 dB SPL in the free field with frontal incidence.

In Zwicker's formulation, upward spread of masking is modelled by appending slopes to the SL values. These slopes approximate the transformed upper excitation slopes, similar to the masking patterns measured empirically using pure tones (refer to Section 2.3.2). Finally, total loudness (in sones) is calculated by summing the SL pattern over the Bark scale,

$$\Psi = \sum_{i=0}^{N-1} \psi(i)\Delta z, \tag{3.36}$$

where N is the number of critical-band filters and Δz is the filter interval in Bark (typically 0.1 Bark). The transformation to loudness level in phons is given by the equation

$$L_{phons} = \begin{cases} 40 + 10\log_2 \Psi, & \Psi \geq 1 \\ 40\Psi^{0.35}, & \text{otherwise.} \end{cases} \tag{3.37}$$

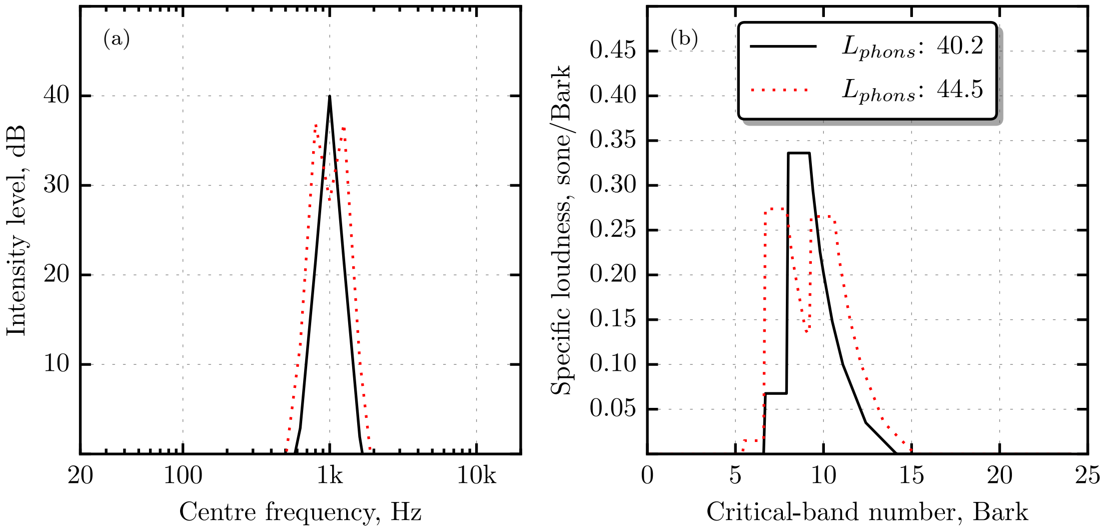


Figure 3.5: Subplot (a) shows the third-octave band levels of a 1 kHz tone (black solid line) and a two-tone complex with frequencies separated by 2 Barks (red dotted line). Both the pure tone and complex tone had an overall intensity level of 40 dB SPL. Subplot (b) shows the corresponding SL patterns estimated using Zwicker's procedure.

Figure 3.5(a) shows the third-octave band levels and (b) the SL patterns estimated by Zwicker's model given two input signals: a 1 kHz pure tone and a two-tone complex with component frequencies of 839 Hz and 1181 Hz (a 2 Bark separation). The overall intensity level was 40 dB SPL for each signal. Despite the two signals having the same presentation level, the loudness level of the complex tone (red dotted lines) is around 4 phons higher than that of the pure tone (black solid lines). This occurs because the frequency interval between the two pure tones comprising the complex tone is greater than a critical band, so the two-tone complex excites a wide range of critical bands which—after compression—yields greater area than the pattern produced by the 1 kHz tone with energy concentrated in a single critical band. The calculated loudness corresponding to the complex tone continues to increase with frequency separation until the overall loudness is equal to the sum of the individual loudnesses of each tone. This effect is, however, level dependent.

Excitation near threshold results in a steep SL slope, meaning that the energy in neighbouring critical bands becomes highly attenuated. Therefore, at very low levels, the peak excitation for the 1 kHz tone would lead to greater loudness than the two-tone complex.

Zwicker's dynamic loudness model

Multiband loudness models have traditionally been developed for estimating the loudness of stationary sounds only. In order to calculate the loudness of time-varying sounds, a time-frequency decomposition is added to the front of the steady-state model, and a temporal-integrator is appended to reflect post-masking and memory effects. For example, Zwicker's (1977) model for temporally-variable sounds was designed as a practical electronic loudness meter conforming to the ISO 532 procedure outlined above. However, a highpass filter with a cut-off frequency near 50 Hz is first applied to mimic the reduced sensitivity of the ear at low frequencies. The required critical-band excitation levels are then approximated by smoothing the square-law rectified output channels of a third-octave filter bank and converting them to logarithmic values. Temporal smoothing is achieved using lowpass filters with a time constant of 2 ms.

The SL transformation is computed as per the ISO standard, using nonlinear analogue converters to account for threshold in quiet. The upper slopes of the main excitations are modelled using a special network with level-dependent characteristics, whilst the steeper lower slopes are ignored. An IL is then calculated by summing the 24 SL values and smoothing this quantity over time using a nonlinear lowpass filter with a fast attack time and a decay time that depends on the duration of the input. A final third-order lowpass filter is used to simulate temporal masking. Although not included in the original paper, Fastl and Zwicker (2007) perform nonlinear smoothing of the SL-time functions *prior* to the summation across the perceptual scale. The idea is that long signals elicit an extended decay of SL compared to very short signals, i.e. the fall time is duration dependent. This modification was introduced by Zwicker (Zwicker et al. 1979; Zwicker 1984) as an analogue circuit that produced a time constant of about 3.5 ms for very short masker durations (5 ms) and 11 ms for long masker durations (200 ms).

3.3.2 Chalupper and Fastl's dynamic loudness model

Chalupper and Fastl (2002) proposed the first loudness model applicable to nonstationary sounds for both normal and hearing-impaired listeners. In principle, it is similar to Zwicker's dynamic model as summarised above, but with refinements at each stage, including modified SL equations to simulate hearing loss.

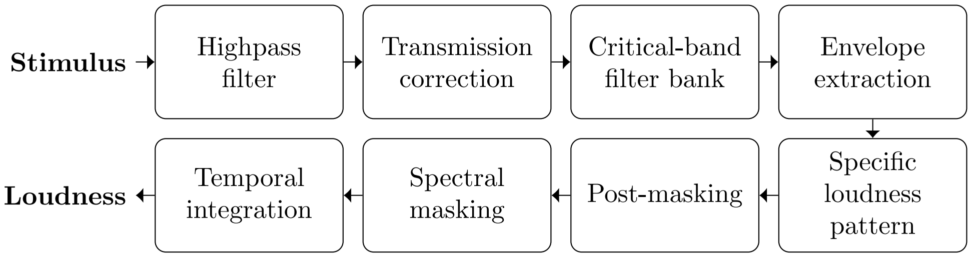


Figure 3.6: Block diagram of Chalupper and Fastl's (2002) dynamic loudness model.

A diagram of the model is shown in Figure 3.6. Following Zwicker's suggestion, a highpass Butterworth filter is first installed to account for the rise in absolute threshold at low frequencies falling in the first critical band. Chalupper and Fastl employ a bank of 24 Gammatone filters (Patterson et al. 1987; Slaney 1993) with an equivalent rectangular bandwidth (equivalent rectangular bandwidth (ERB)) of 1 Bark. The centre frequencies are spaced equally on the Bark scale, following the critical bands defined by Zwicker (1961). Note that Zwicker (1977) also suggested the use of a filter bank closer to that of the hearing system, but this was not done in his procedure as the focus was to establish a method that complied with part B of ISO 532 (1975).

Envelope extraction is performed on every frequency channel using the temporal window proposed by Plack and Moore (1990), configured to have an equivalent rectangular duration of 4 ms and a time step of 2 ms. The critical-band levels are thus obtained by computing a running RMS level in each channel. The critical-band levels are then corrected for free-field presentation as described in DIN 45631 (1991) and ISO 532 (1975). The transformation from excitation to SL follows Equation 3.35 but parameters E and E_{ThQ} are made dependent on the amount of hearing loss (in dB) and a recruitment factor k for a given individual. The hearing loss effectively shifts the loudness growth function measured in normal-hearing individuals to the right whilst the factor k adjusts the slope. Thus, the dynamic model accounts for the elevated absolute thresholds and loudness recruitment observed in the hearing impaired.

Post-masking is modelled using a discrete-time implementation (Widmann et al. 1998) of Zwicker's analogue electronic circuit (Zwicker 1984) as discussed in the preceding section. Thus, a temporal smoothing of SL takes place in each critical-band, prior to spectral summation. Similarly, spectral masking follows the implementation published in DIN 45631, in which SL patterns are generated with a resolution of 0.1 Bark. The loudness pattern is summed to estimate the IL, which is then smoothed by a first-order lowpass Butterworth filter with a cut-off frequency of 8 Hz.

3.3.3 The Cambridge loudness model

Moore and Glasberg (1997) proposed a revision of Zwicker's steady-state loudness model, the Cambridge loudness model, to account more accurately for the way that equal-loudness contours change with level and to provide a more satisfactory explanation of loudness summation in terms of hearing physiology. As discussed in Section 3.1.1, further improvements to the Cambridge model were made in later work (Moore et al. 1997; Glasberg and Moore 2006; ANSI S3.4 2007). The key revisions that differentiate this more recent algorithm from Zwicker's method are:

1. The auditory frequency scale is based on the Cam scale and not the Bark scale (see Section 2.3.2).
2. The use of level-dependent rounded exponential (roex) filters (Patterson 1976) to calculate EPs directly (Moore and Glasberg 1983; Glasberg and Moore 1990).
3. Modifications to the transformation from excitation to specific loudness with added partial loudness calculations.
4. The assumed transfer function of the outer and middle ear, and its placement prior to the EP transformation.

The first three points are the most important because they influence the shape of the SL pattern which consequently leads to different predictions of SLS. Because the critical-band filters are narrower when using the Cambridge formula (refer to Figure 2.16), more filters are required to cover the audible spectrum, e.g. approximately 41 on the Cam scale compared to 24 on the Bark scale. A larger number of filters results in a greater amount of SLS owing to the fact that loudness is determined by the area under the loudness density (refer to Section 2.3.3). Not only are the critical bandwidths defined differently, but so is the method in which EPs are calculated. More specifically, whereas Zwicker's method of deriving EPs is based on empirically measured masking patterns (Zwicker and Scharf 1965), the Cambridge method is grounded in more direct measurements of the auditory filter shape (Moore and Glasberg 1983) using a technique that controls for confounding factors. Additionally, the relationship between excitation and SL differs between the two models, introducing further differences in the way that loudness is summed across frequency.

A detailed description of the Cambridge multiband loudness model is given in the following sections. It is presented as an implementation-level guide applicable to time-varying sounds (the dynamic model), not only because this is useful for understanding its architecture, but also because the model is used extensively throughout this thesis and so a deeper understanding of its functionality is important for later chapters. Appendix A provides validation results of the implementation, where the example data given in ANSI S3.4 (2007) is used as reference data.

The dynamic model

Stone et al. (1997) first extended their steady-state procedure (Moore and Glasberg 1997) to deal with dynamic sounds for use as a real-time loudness meter. This involved computing a 1024-point discrete Fourier transform (DFT) via the FFT to obtain a short-term power spectrum, which was then fed to the loudness procedure for stationary sounds. Frames were updated every 30 ms and first-order temporal smoothing was applied to the frequency-integrated loudness measurements. Similarly, Glasberg and Moore (2002) adapted their 1997 stationary model (Moore et al. 1997) to deal with time-varying sounds using more sophisticated signal processing techniques compared to the real-time predecessor in order to more accurately model the loudness of complex fluctuating sounds. In what follows, the details of the 2002 procedure are given in combination with refinements proposed in subsequent work (Glasberg and Moore 2006; ANSI S3.4 2007).

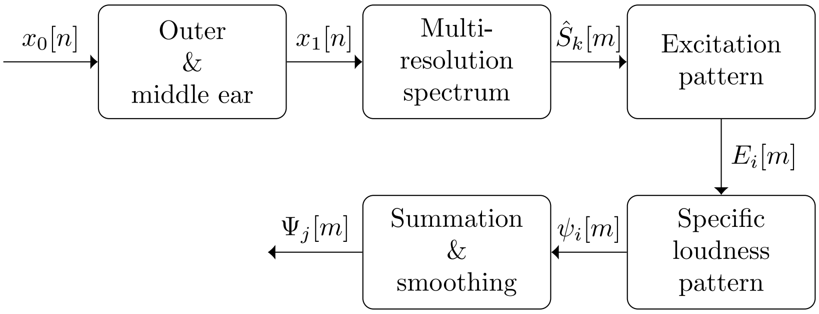


Figure 3.7: Block diagram of Glasberg and Moore's (2002) dynamic loudness model.

Index Notation	Reference	Resolution
b	Frequency band	N/A
m	Spectral frame	1 ms
n	Time sample	0.03125 ms
i	Auditory filter	0.25 Cams
j	Loudness function	1 ms
k	Frequency component	15.625 Hz

Table 3.2: Details of the notation used in Figure 3.7. Example time/frequency resolutions are given for a sampling frequency of 32 kHz.

The structure of the Glasberg and Moore's dynamic loudness model is depicted in Figure 3.7. The architecture is similar to that of Zwicker's model, but differs in the technicalities of implementation. The stimulus input x_0, a monophonic digital waveform, is expected to be in Pascals. A three-channel loudness signal Ψ_j is output by the model in sones. Details of the time and frequency notation used in the figure are specified in Table 3.2.

Pre-cochlear filtering

A 4096th order FIR filter is used to simulate the combined outer ear to eardrum and eardrum to oval window (middle ear) transfer function. The free-field outer ear transmission response follows the function measured by Shaw (1974). Based on the assumption that the cochlea is equally sensitive between 500 Hz and 15 kHz, the middle-ear transmission response was derived from the absolute-threshold curve (intensity level of a pure tone presented in quiet at threshold as a function of frequency) given in ISO 389-7 (2005). Below 500 Hz, Glasberg and Moore (Moore et al. 1997; Glasberg and Moore 2006) incorporated a steep roll-off in accordance with the data of Puria et al. (1997) and Aibara et al. (2001). The function was then modified such that the complete model gave good predictions (within 0.2 dB) of the audibility data. Figure 3.8 shows the transfer function of the outer and middle ear as specified in ANSI S3.4 (2007) for free- and diffuse-field environments. For stimuli presented via headphones, the outer-ear transfer function is to be replaced by that of the headphone as measured at the eardrum.

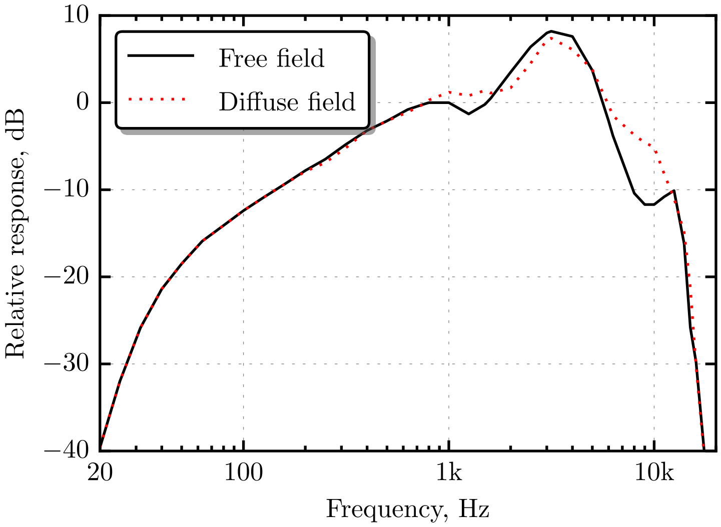


Figure 3.8: Transmission function of the outer (free/diffuse field) and middle ear according to the data tabulated in ANSI S3.4 (2007).

Multi-resolution power spectrum

In order to derive cochlear EPs from an arbitrary stimulus, it is necessary to decompose the input into critical bands. To approximate the time-frequency resolution of the ear, a multi-resolution DFT is computed using six parallel FFTs. Six Hann analysis windows of duration 64, 32, 16, 8, 4 and 2 ms are used, corresponding to frequency bands 20–80 Hz, 80–500 Hz, 500–1250 Hz, 1250–2540 Hz, 2540–4050 Hz and 4050–15 000 Hz, respectively. Table 3.3 summarises the multi-resolution DFT specification where $f_k = k\frac{f_s}{N_{FFT}}$ denotes the frequency (in Hz) of the k^{th} spectral component computed using an FFT of length N_{FFT}.

All segments are generated using a Hann window of sample length N_b,[4] where b denotes the

[4] The Hann window is specified in its periodic form as per Harris (1978, Eq. 27b).

Window size (ms)	Band edges (Hz)
64	$20 \leq f_k < 80$
32	$80 \leq f_k < 500$
16	$500 \leq f_k < 1250$
8	$1250 \leq f_k < 2540$
4	$2540 \leq f_k < 4050$
2	$4050 \leq f_k \leq 15\,000$

Table 3.3: Multi-resolution DFT specification proposed by Glasberg and Moore (2002). f_k refers to the frequency of the k^{th} spectral component.

frequency band index ($b \in \{0, 1, \ldots, 5\}$). The Hann window is defined as

$$w_b[n] = 0.5 - 0.5\cos(2\pi n/N_b), \quad 0 \leq n < N_b. \tag{3.38}$$

The windows are normalised to give an average energy of one,

$$\hat{w}_b = w_b\sqrt{N_b/\sum w_b^2}. \tag{3.39}$$

To align the temporal centres of the normalised windows, O_b zeros are prepended and appended to the start and end of each window,

$$O_b = \lceil\frac{N_0 - 1}{2}\rceil - \lceil\frac{N_b - 1}{2}\rceil, \tag{3.40}$$

which are then used to segment the input signal into time-aligned, amplitude-tapered blocks. The FFT buffer length used for *all* windowed segments is chosen to be the first power of two equal to or greater than the length of the largest window N_0,

$$N_{FFT} = 2^{\lceil\log_2(N_0)\rceil}. \tag{3.41}$$

Taken together, this means that zero-padding is employed such that all spectral bands are sampled at frequency intervals of $\Delta f = \frac{f_s}{N_{FFT}}$. For a given FFT output, components outside of its designated frequency range are discarded. The running complex spectrum $X_{b,k}[m]$ is computed using a time step (or hop size) of 1 ms. The multi-resolution power spectrum is computed using

$$\hat{S}_{b,k}[m] = c_b X_{b,k}^2[m], \tag{3.42}$$

where c_b is an average-power normalisation factor for the one-sided spectrum,[5]

$$c_b = \frac{2}{N_{FFT} N_b p_0^2}. \tag{3.43}$$

In short, summing the normalised power spectrum over a given frequency band gives the mean square energy in that band. The factor $p_0 = 2 \times 10^{-5}$ is incorporated in order to normalise the input signal by a common reference sound pressure. Because intensity is proportional to the square of the pressure of a travelling wave in the free field, the constant of proportionality vanishes when expressed in normalised units, which is convenient for subsequent calculations.

[5] Note that DC and Nyquist components are excluded from the compiled spectrum, so the numerator of two is valid here.

Excitation pattern

Following pre-cochlear filtering and time-frequency decomposition, an EP is calculated. Moore and Glasberg (1983) proposed a method of calculating EPs from the auditory filters directly, the shapes of which were estimated from experiments employing the notched-noise method (Patterson 1976). The EP is defined as the output of each auditory filter expressed as a function of filter centre frequency. In subsequent work (Moore and Glasberg 1987; Glasberg and Moore 1990), the authors showed how the auditory filter shapes can be calculated for varying input levels, and as a result, how EPs for arbitrary spectra can be obtained. This approach has the advantage over Zwicker's in that masking audiograms provide only a crude indicator of the EP, and are influenced by numerous factors, e.g. off-frequency listening and interactions between the signal and masker (Moore and Glasberg 1997).

In what follows, let f_{c_i} denote the centre frequency and ERB_i be the corresponding ERB (in Hz) of the i^{th} auditory filter as defined by Equation 2.26, respectively. Glasberg and Moore (2002) specify a filter interval (ΔCam) of 0.25 Cams, yielding 149 filters spanning 50 to 15 000 Hz. This contrasts to 372 filters used in ANSI S3.4 (2007) ($\Delta\text{Cam} = 0.1$) which can afford higher resolution because the procedure requires only a single power spectrum representing a stationary sound.

A bank of roex filters are used to filter the power spectrum and arrive at the EP. The procedure for computing auditory EPs was given by Glasberg and Moore (1990) and is summarised here.

The shape of the i^{th} auditory filter at the frequency of input component k is

$$W_{i,k} = (1 + p_{i,k} g_{i,k}) e^{-p_{i,k} g_{i,k}}, \tag{3.44}$$

where $g_{i,k}$ is the normalised absolute deviation from the filter centre frequency,

$$g_{i,k} = \frac{|f_k - f_{c_i}|}{f_{c_i}}, \tag{3.45}$$

and p determines the filter bandwidth and shape of the filter skirts. The normalised frequency g is limited to a maximum of two when computing Equation 3.44, which gives sufficient resolution for the upper skirt of the auditory filter in normal-hearing persons. At high levels the filter shape is asymmetric, and so some level dependency is introduced by letting parameter p vary,

$$p_{i,k} = \begin{cases} p_{l_{i,k}}, & f_k < f_{c_i} \\ p_{u_i}, & \text{otherwise,} \end{cases} \tag{3.46}$$

where p_{l_i} and p_{u_i} refer to the lower and upper slopes of the filter centred at f_{c_i}, respectively. Glasberg and Moore (1990) reanalysed the data summarised in their previous work (Moore and Glasberg 1987), amongst two additional studies, to investigate the variation of auditory filter shape with level. They concluded that the low-frequency filter skirt becomes less sharp with increasing level, though changes in the high-frequency slope with varying level were not consistent across studies and therefore can be ignored. However, Moore (1995) emphasised that at high levels, where the filter is markedly asymmetric, there may be systematic errors in the estimates of the sharpness of the high skirt. The upper and lower slopes of the i^{th} filter are determined using

$$p_{u_i} = \frac{4 f_{c_i}}{ERB_i}, \quad \text{and} \tag{3.47}$$

$$p_{l_{i,k}} = p_{u_i} - 0.35 \left(\frac{p_{u_i}}{p_{1kHz}} \right) (L_{\text{ERB}}[k] - 51), \tag{3.48}$$

where p_{1kHz} is the value of p_u at 1 kHz and $L_{\mathrm{ERB}}[k]$ is the intensity level of the intensity sum of all components falling within 1 ERB around the k^{th} input component, i.e. the level per ERB. Symmetrical and level independent roex filters centred on every input component are used to obtain the level per ERB (ANSI S3.4 2007):

$$L_{\mathrm{ERB}}[j] = 10\log_{10}\left(\sum_{k=0}^{K-1} W_{j,k}\hat{S}_m[k]\right), \tag{3.49}$$

where K is the total number of components comprising the multi-resolution spectrum (omitting the band index for brevity). $W_{j,k}$ denotes a symmetrical roex filter centred on frequency component j, using Equation 3.47 to define the (fixed) slope parameter.

The final step is to calculate the EP in linear power units using the filter shapes obtained using Equation 3.44. This transformation involves sampling the power spectrum on the Cam scale using four filters per ERB. Therefore, the overall excitation in critical band i at time step m is given by

$$E_i[m] = \sum_{k=0}^{K-1} W_{i,k}\hat{S}_m[k]. \tag{3.50}$$

It is convenient to express linear excitation E relative to the excitation produced by a 1 kHz sinusoid at 0 dB SPL (free field, frontal incidence) at the output of the auditory filter centred at 1 kHz. However, because the combined response of the modelled outer and middle ear is 0 dB at 1 kHz and the power spectrum is in normalised power units, no further normalisation is required.

SL transformation

With the EP of a spectral frame extracted, the next stage is to convert it into an SL pattern ψ_i which represents the loudness in a given critical band. This is modelled as a piecewise power function of the EP,

$$\psi = \begin{cases} \beta\left[(GE+A)^{\alpha} - A^{\alpha}\right], & E_{\mathrm{ThQ}} \leq E \leq 10^{10} \\ \beta\left(\frac{2E}{(E+E_{\mathrm{ThQ}})}\right)^{1.5}\left((GE+A)^{\alpha} - A^{\alpha}\right), & E < E_{\mathrm{ThQ}} \\ \beta\left(\frac{E}{1.04\times10^{6}}\right)^{0.5}, & E > 10^{10}. \end{cases} \tag{3.51}$$

In this equation all parameters are a function of the filter centre frequency which has been omitted for brevity. The equation for high intensities (where $E > 100$ dB) was revised in ANSI S3.4 (2007) such that compression is maintained rather than reduced:

$$\psi = \beta\left(\frac{E}{1.0707}\right)^{0.2}. \tag{3.52}$$

As with Zwicker's derivation of the transformation, Moore et al. (1997) started with the assumption that at medium to high levels, excitation and specific loudness are related by the compressive function βE^{α}, with $\alpha < 1$. Parameter A was introduced to increase the slope of the SL for excitations near threshold and thus takes a similar form to Equation 3.6 (in fact the value of A is equal to twice the peak excitation produced by a sinusoid at absolute threshold). G represents the low-level gain of the cochlear amplifier at a given frequency, relative to the constant unity gain at 500 Hz and above, and E_{ThQ} is the peak excitation of a sinusoid required for absolute threshold at a given centre frequency (both in linear units).

The parameters in Equation 3.51 are tabulated in ANSI S3.4 (2007), meaning that LUTs and

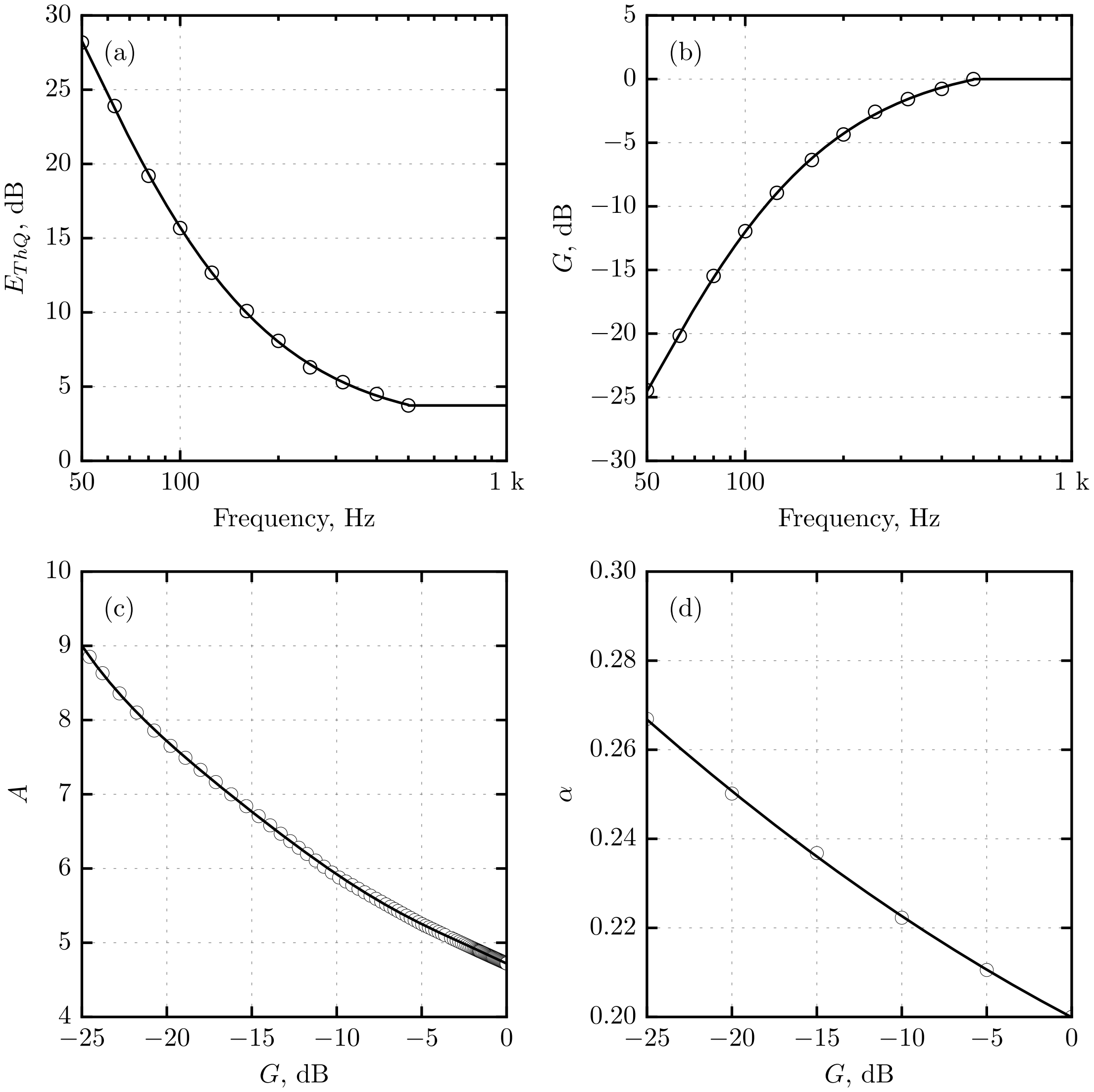


Figure 3.9: The curves in plots (a) to (c) show the SL parameters defined by Equations 3.53-3.56, respectively. Circles are the data points given in ANSI S3.4 (2007), on which the polynomials were modelled.

careful interpolation is needed. To simplify the implementation, polynomials of varying degree were developed in this thesis and fit to the data using penalised goodness-of-fit metrics such as the Akaike information criterion and adjusted-R^2 to prevent overfitting. For frequencies below 500 Hz, the parameters in Equation 3.51 can be approximated using the following equations:[6]

$$E_{ThQ_{\mathrm{dB}}} = 10^{0.430688\log_{10}(f_c)^3 - 2.79761\log_{10}(f_c)^2 + 5.07385\log_{10}(f_c) - 1.20606}, \tag{3.53}$$

$$G_{\mathrm{dB}} = 3.73 - E_{ThQ_{\mathrm{dB}}}, \tag{3.54}$$

$$\begin{aligned} A = & -1.0706497 \times 10^{-6} G_{\mathrm{dB}}^5 - 6.0648487 \times 10^{-5} G_{\mathrm{dB}}^4 \\ & - 0.0012047327 G_{\mathrm{dB}}^3 - 0.0068190418 G_{\mathrm{dB}}^2 \\ & - 0.11847826 G_{\mathrm{dB}} + 4.7138722 \quad \text{and} \end{aligned} \tag{3.55}$$

$$\alpha = -2.686429 \times 10^{-5} G_{\mathrm{dB}}^2 - 0.002002336 G_{\mathrm{dB}} + 0.1999311, \tag{3.56}$$

where $E_{ThQ_{\mathrm{dB}}}$ and G_{dB} denote the values of E_{ThQ} and G in decibels. For frequencies $\geq$ 500 Hz: $EThQ_{\mathrm{dB}} = 3.73$, $G_{\mathrm{dB}} = 0$, $A = 4.72096$ and $\alpha = 0.2$. The value of the constant β is 0.046871 (ANSI S3.4 2007). Figure 3.9 shows the approximations along with the target data points, which apply for frequencies below 500 Hz.

The transformation from excitation to SL for auditory filters centred at high to low frequencies (outer to inner curves) is depicted in Figure 3.10. For frequencies below 500 Hz, a higher excitation level is required to establish the same SL as at higher frequencies. This can be explained in terms of the parameters defining the SL transformation: filters centred at lower frequencies are subject to greater internal noise, less compression and reduced cochlear gain than at higher frequencies.

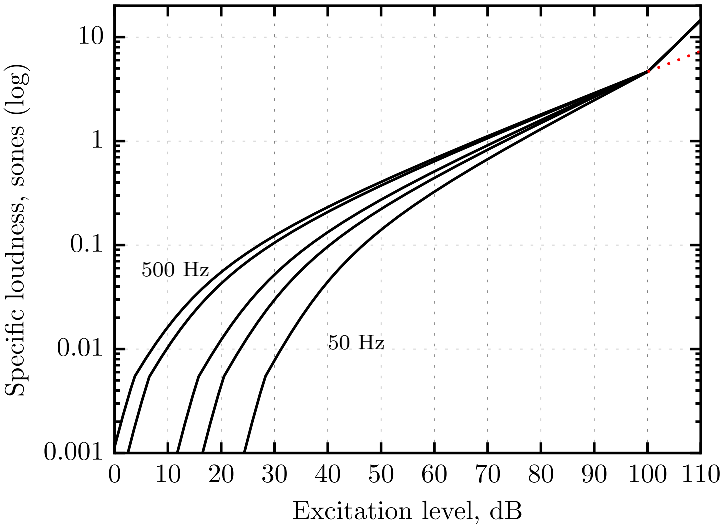


Figure 3.10: Functions relating excitation level to SL at auditory filters centred at 500, 250, 100, 75 and 50 Hz, corresponding to the outer to inner curves. The outer curve applies for all frequencies of 500 Hz and above. These curves were generated using the original formula as specified by Moore et al. (1997) with the frequency-dependent loudness parameters approximated using Equations 3.53-3.56. The red dotted curve shows the revised transformation at excitation levels above 100 dB (ANSI S3.4 2007).

[6] The coefficients are provided to two decimal places beyond the number of significant figures of the tabulated data.

Loudness-time functions

The total IL of the current analysis frame is computed by summing the SL over all critical bands:

$$\Psi_0[m] = 0.25 \sum_{i=0}^{I-1} \psi_i[m]. \tag{3.57}$$

The constant 0.25 refers to the interval between adjacent filters on the Cam scale. For a monaural input signal presented to both ears (diotic presentation), binaural loudness can be approximated by multiplying the IL value by two. For dichotic presentation, it is necessary to calculate the SL in each ear separately. Note that the Ψ_0 is calculated once every millisecond and is assumed unavailable to conscious perception (Glasberg and Moore 2002). The IL is therefore smoothed by two temporal integration stages resulting in two perceptual measures: the STL, describing the loudness perceived at any moment; and the LTL, reflecting overall loudness judgements and memory effects. The STL is calculated using a one-pole lowpass filter:

$$\Psi_1[m] = \alpha_S \Psi_0[m] + (1 - \alpha_S)\Psi_1[m-1], \tag{3.58}$$

where the filter coefficient is dependent on the attack or release state of the integrator:

$$\alpha_S = \begin{cases} 1 - e^{\frac{-T}{\tau_{S_a}}}, & \Psi_0[m] > \Psi_1[m-1] \\ 1 - e^{\frac{-T}{\tau_{S_r}}}, & \text{otherwise,} \end{cases} \tag{3.59}$$

where T is the time step in seconds (here $T = 0.001$).[7] Similarly, the LTL is given by

$$\Psi_2[m] = \alpha_L \Psi_1[m] + \alpha_L \Psi_2[m-1], \tag{3.60}$$

again with

$$\alpha_L = \begin{cases} 1 - e^{\frac{-T}{\tau_{L_a}}}, & \Psi_1[m] > \Psi_2[m-1] \\ 1 - e^{\frac{-T}{\tau_{L_r}}}, & \text{otherwise.} \end{cases} \tag{3.61}$$

The time constants in Equation 3.59 and 3.61 can be obtained by substituting the smoothing coefficients 0.045, 0.02, 0.01 and 0.0005, as given by Glasberg and Moore (2002), for α in the formula

$$\tau = \frac{-T}{\log(1-\alpha)}. \tag{3.62}$$

With $T = 1$ ms, this gives time constants of 21.7 ms, 49.5 ms, 99.5 ms and 1999.5 ms corresponding to τ_{S_a}, τ_{S_r}, τ_{L_a} and τ_{L_r}, respectively. The STL time constants were chosen so as to predict a doubling of loudness per doubling of duration for signals less than 100 ms. For such short duration sounds, Glasberg and Moore (2002) suggest using the maximum of the STL as a measure of overall loudness. For relatively steady sounds, such as pure tones and stationary noise, the mean LTL was recommended. The longer time constants were selected to give good correspondence to the trends observed in experiments on the loudness of amplitude modulated tones, although in a later paper Moore et al. (2003) reduced the value of τ_{L_r} was to 199.5 ms to give a better fit to loudness-matching data involving segments of speech.

Figure 3.11 shows (a) the excitogram and (b) the three loudness-time series output by the model in response to a 200 ms 8 kHz pure tone presented binaurally in the free field at 60 dB SPL. The excitogram is similar to the spectrogram but instead of plotting the log-magnitude of each

[7] More generally $T = \frac{H}{f_s}$, where H is the desired hop size in samples.

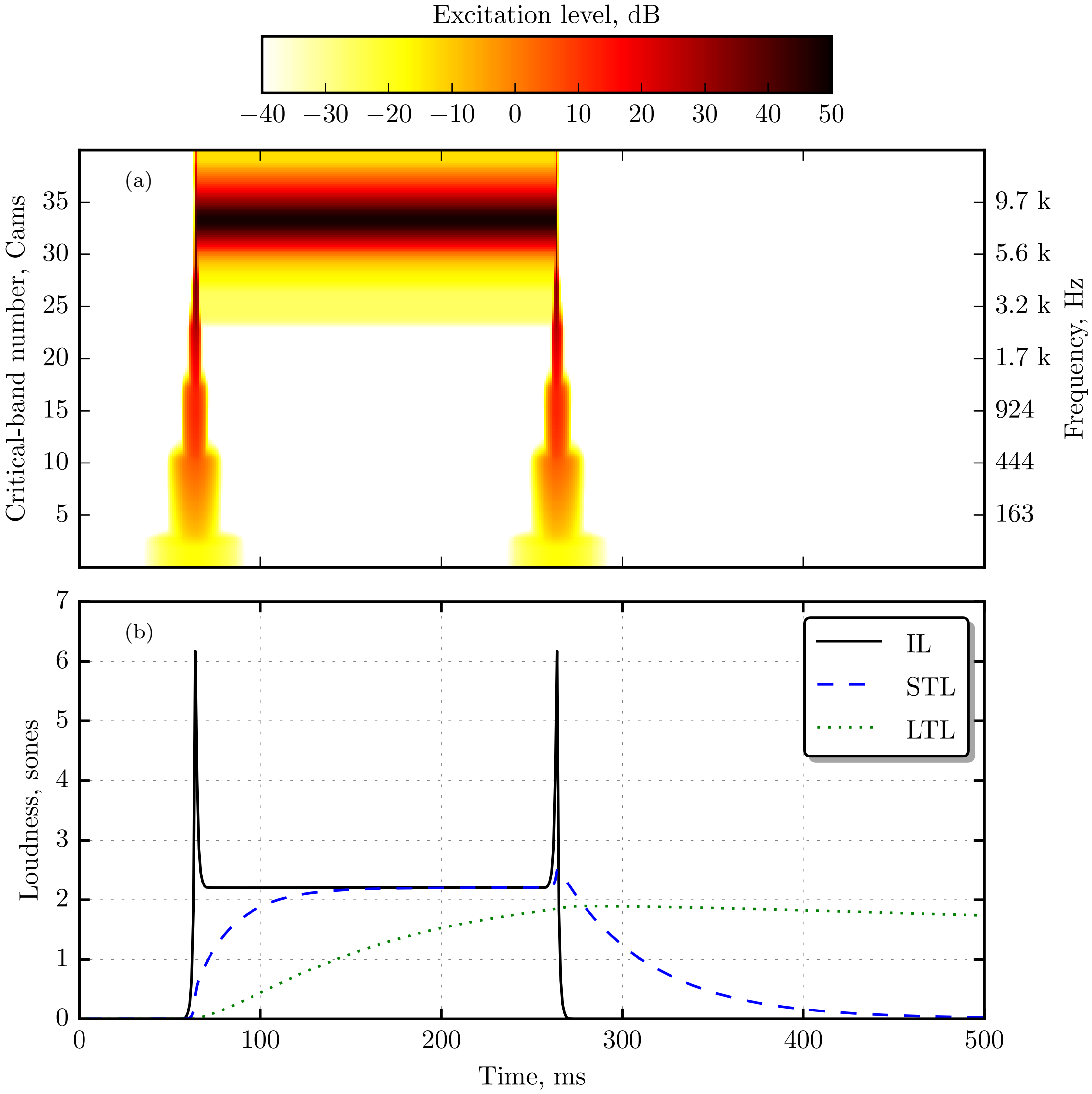


Figure 3.11: (a) Excitogram of a 200 ms 8 kHz pure tone presented at 60 dB SPL in the free field with frontal incidence. The corresponding IL, STL and LTL of the tone are shown in subplot (b).

DFT bin as a function of time, the critical-band excitation levels are plotted over time. In other words, each vertical time slice represents the EP of a segment at a given time step. Compared to the spectrogram, this representation of sound is therefore closer to that represented in the auditory system. An 8 kHz tone with no ramp applied to the onset or offset was deliberately selected to reveal two things. First, the abrupt onset and offset of the pure tone broadens the spectrum, which consequently excites a number of critical-band filters across the majority of the audible frequency scale.[8] This leads to greater SLS at the onset and offset of the stimulus, which translates to peaks in the IL. Notice that the IL peaks after allowing for the group delay of the FIR filter (64 ms with $f_s = 32$ kHz). The STL follows the same trajectory as the IL but smooths the transients in a perceptually-relevant manner. Due to duration of the stimulus being short, as well as the long attack and release times of the LTL integrator, the LTL fails to reach the asymptotic loudness and falls below absolute threshold after several seconds. This highlights the second major point: there is little time smoothing in the excitation or SL domain. There is some temporal smearing taking place below about 5 kHz, which increases rather distinctly with decreasing frequency. These regions correspond to the excitation derived from different temporal-analysis windows used for computing the multi-resolution spectrum. The lack of critical-band smoothing is another aspect that distinguishes this model from that of Zwicker (1977) and Chalupper and Fastl (2002), as both of these latter algorithms incorporate nonlinear filters to simulate post-masking effects. There is evidence that some form of temporal persistence of SL is needed to account for the loudness of frequency-modulated sounds (Grimm et al. 2002; Rennies et al. 2010). This is discussed further in Section 3.3.7.

3.3.4 Chen and Hu's dynamic loudness model

Chen et al. (2011) proposed a new procedure for estimating the loudness of stationary sounds. This approach is interesting because unlike those based on Zwicker's classic model, no transformation from excitation to specific loudness is required. Instead, the proposed filter bank mimics basilar-membrane compression, consistent with the idea that both frequency selectivity and cochlear compression are the result of a single active process (Robles and Ruggero 2001).

Following tradition, the stationary model was extended to predict the loudness of time-varying sounds a year later (Chen and Hu 2012). The signal processing of this dynamic model is similar to that of Glasberg and Moore's, except that the outer/middle-ear filter is applied after extracting power spectra, and the transformation to SL stage is omitted as a result of the new filter bank. For conciseness, the following sections describe the modifications made to Glasberg and Moore's procedure and the implementation of the active filter bank.

Multi-resolution power spectrum

Chen and Hu (2012) use six FFTs in parallel to execute the multi-resolution DFT, based on window sizes of 128, 64, 32, 16, 8 and 4 ms corresponding to the six frequency bands 20–80 Hz, 80–500 Hz, 500–1250 Hz, 1250–2540 Hz, 2540–4050 Hz and 4050–16 000 Hz, respectively. Although the frequency bands are identical to those of Glasberg and Moore (2002), the windows are double in length (refer to Table 3.3). The windows are all temporally aligned via zero padding, yielding approximately twice as many frequency points as used in Glasberg and Moore's model. Chen and Hu decided to double the frequency resolution because the DFT specification reported by Glasberg and Moore resulted in a spectral bandwidth that exceeded the critical bandwidth at 1 kHz. Consequently, the predicted total loudness of pure tones is larger than expected. Table 3.4

[8] Spectral broadening is also referred to as spectral splatter.

shows the loudness in sones predicted by both dynamic models in response to a 1 kHz pure tone presented binaurally at 40 dB SPL in the free field. It can be seen that the prediction of Glasberg and Moore's model (GM02) is about 0.7 dB higher than that of Chen and Hu's (CH12).

GM02	CH12	ΔL (dB)
1.11	1.06	0.67

Table 3.4: Predictions, in sones, of the dynamic loudness models of Glasberg and Moore (2002) (GM02) and Chen and Hu (2012) (CH12) in response to a 1 kHz pure tone presented binaurally in the free-field at 40 dB SPL. The level difference needed for equal loudness is given in the final column.

Pre-cochlear filtering

The middle-ear transfer function is based on the one given in ANSI S3.4 (2007) but modified to better fit *both* absolute thresholds (ISO 389-7 2005) and notched-noise data (Baker et al. 1998; Glasberg and Moore 2000). The modified transfer function is shown in Figure 3.12 along with the function defined in the ANSI standard.[9] Although the two responses are similar, the one used by Chen et al. applies a gentler roll-off at frequencies below 1 kHz, but slightly more attenuation in the 2–10 kHz region. The free field to eardrum transfer function is the same as the one given in the ANSI standard. Unlike Glasberg and Moore's procedure, the transfer function of the outer and middle ear is applied in the frequency domain by weighting the power spectrum. Because of the increased frequency resolution, Chen and Hu's model does not suffer from the discrepancies at low frequencies when the spectral weighting approach is taken and (presumably) this is why they opted against the increased computational complexity associated with the FIR filter.

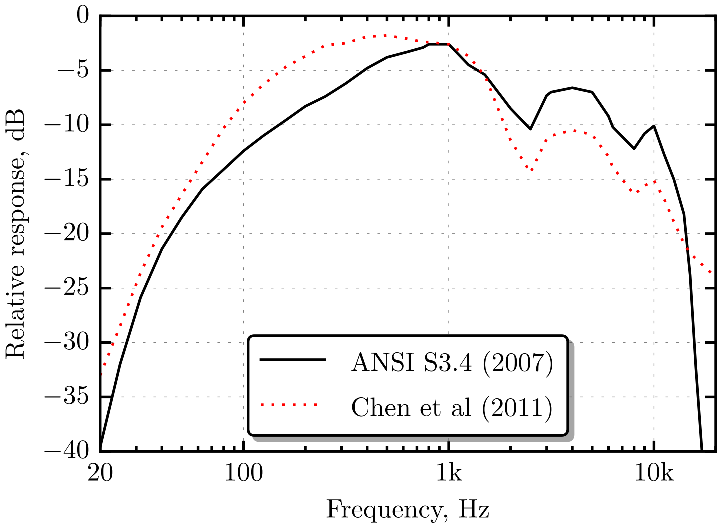


Figure 3.12: Transfer function of the middle ear according to Chen et al. (2011) and ANSI S3.4 (2007).

Excitation pattern

The transformation from power spectrum to EP is carried out using the double-roex filter, depicted in Figure 3.13. The filter consists of a passive-wideband roex filter and an active-narrowband roex

[9] The data points were determined from a visual inspection of the curve shown in Figure 2 of Chen et al. (2011).

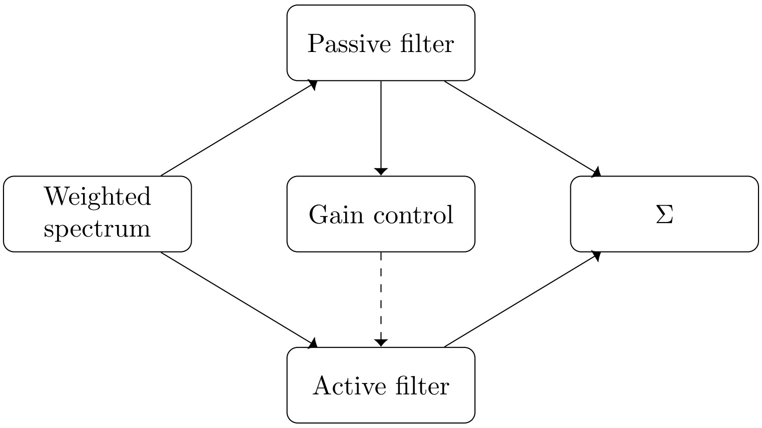


Figure 3.13: Block diagram of the double roex filter model.

filter, with the gain of the latter being controlled by the output of the former. Using the same notation as in Table 3.2, index i refers to the i^{th} auditory filter and k refers to the k^{th} frequency component of the multi-resolution spectrum. In this procedure, a filter spacing of 0.1 Cams is used with filters spanning 40–17 000 Hz, giving a total of 388 filters.

The passive filter is given as

$$W_{P_{i,k}} = (1 + g_{i,k}t_i)e^{-g_{i,k}t_i}, \tag{3.63}$$

where $g_{i,k}$ is the normalised absolute deviation from the centre frequency f_c of the i^{th} filter and t_i defines the filter slope:

$$t_i = \begin{cases} \frac{f_{c_i}}{0.108 f_{c_i} + 2.33}, & f_k < f_{c_i} \\ 15.6, & \text{otherwise.} \end{cases} \tag{3.64}$$

The passive filter therefore has a different but fixed slope for frequencies below and above the centre frequency. The total output power of the passive filter in response to the short-term power spectrum $\hat{S}$ at analysis frame m is given by

$$E_{P_i}[m] = \sum_{k=0}^{K-1} W_{P_{i,k}} \hat{S}_m[k]. \tag{3.65}$$

The excitation *level* L_P of the passive filter determines the gain G (in dB) of the active filter centred at f_c. At low excitation levels ($L_P \leq 30$), the gain is given by (omitting frame number here)

$$G_i = G_{\max_i}\left(1 - \frac{1}{1 + e^{-0.05\left(L_{P_i} - \left(100 - G_{\max_i}\right)\right)}} + \frac{1}{1 + e^{0.05\left(100 - G_{\max_i}\right)}}\right), \tag{3.66}$$

and at higher levels,

$$G_i = G_{\max_i}\left(1 - \frac{1}{1 + e^{-0.05\left(L_{P_i} - \left(100 - G_{\max_i}\right)\right)}} + \frac{1}{1 + e^{0.05\left(100 - G_{\max_i}\right)}}\right) - 0.003(L_{P_i} - 30)^2, \tag{3.67}$$

where $G_{\max}$ denotes the maximum gain of the filter which is frequency dependent,

$$G_{\max_i} = \frac{f_{c_i}}{0.0191 f_{c_i} + 1.1}. \tag{3.68}$$

The active filter is defined as

$$W_{A_{i,k}} = G_i'(1 + g_{i,k}p_i)e^{-g_{i,k}p_i}, \tag{3.69}$$

where the G' is the linear gain corresponding to G (in dB) and p defines the filter slope for frequencies below and above the centre frequency,

$$p_i = \begin{cases} \frac{f_{c_i}}{0.027 f_{c_i} + 5.44}, & f_k < f_{c_i} \\ 27.9, & \text{otherwise.} \end{cases} \tag{3.70}$$

Finally, the EP is calculated as

$$E_i[m] = \sum_{k=0}^{K-1} \hat{S}_m[k] \left(W_{P_{i,k}} + W_{A_{i,k}} \right). \tag{3.71}$$

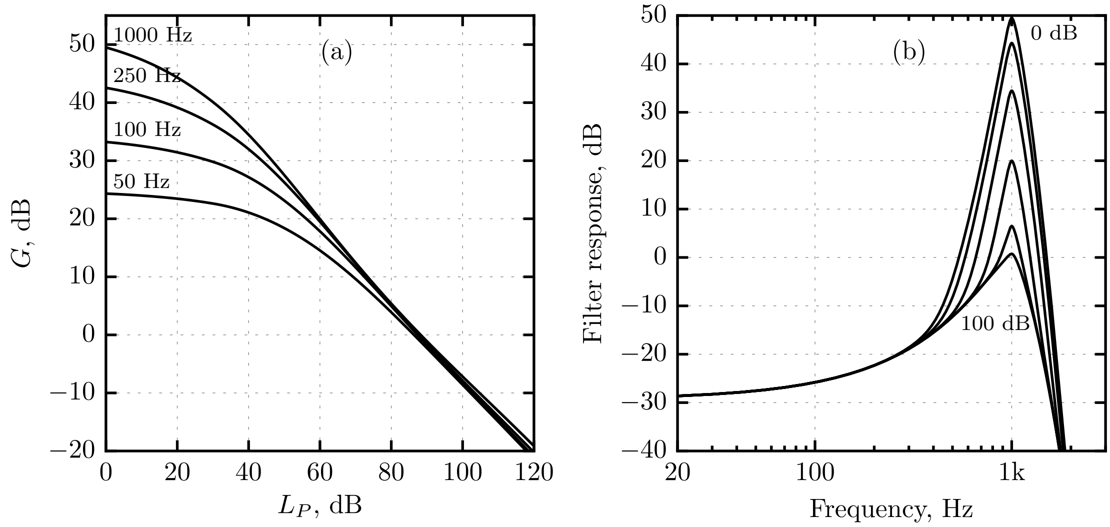


Figure 3.14: (a) The active filter gain expressed as a function of the excitation level at the output of the passive filter. The parameter is the centre frequency of the auditory filter. (b) The response of the double-roex filter centred at 1 kHz for passive filter output levels of 0, 20, 40, 60, 80 and 100 dB. Higher input levels give rise to broader filters owing to the decrease in gain of the active filter.

Figure 3.14(a) shows the gain of the active filter as a function of the output of the passive filter, for centre frequencies of 1 kHz, 250 Hz, 100 Hz and 50 Hz from top to bottom. It can be seen that the that maximum gain decreases with decreasing centre frequency, in agreement with the assumptions of (Moore and Glasberg 1997; Moore et al. 1997) (see Figure 3.9). At all frequencies, the filter gain stays near maximum at very low input levels but decreases with increasing input level, demonstrating level-dependent compression. Subplot (b) shows the response of the double-roex filter centred at 1 kHz for low to high excitation levels measured at the output of the passive filter. At low input levels, where the active filter gain is high (peak at 50 dB), the combined filter is much narrower than at higher input levels because the active-narrowband filter dominates the total output. As the gain decreases with increasing excitation measured in the passive filter, the active filter contributes less to the overall output (applies negative gain; see response at 0 dB) and so the passive filter dominates the output, and thus broadening the overall bandwidth.

Loudness-time functions

The IL is computed on a frame-by-frame basis using a time step of 1 ms. However, due to the level-dependent gain of the active-narrowband filter, the total loudness is measured directly from the area under the (already compressed) EP. In practice, the total IL for the current frame is computed by summing the excitation in each critical band,

$$\Psi_0[m] = k0.1 \sum_{i=0}^{I-1} E_i[m], \tag{3.72}$$

where $k = 1.53 \times 10^{-8}$ and the factor 0.1 refers to the filter spacing.

The STL and LTL are computed as explained in Section 3.3.3 except for the time constants used. The attack and release times of the STL integrator are 16 ms and 32 ms, respectively. For the LTL, the attack and release times are 100 ms and 2000 ms. Note that the STL time constants are slightly lower than those proposed by Glasberg and Moore, which is likely to compensate for the longer temporal analysis windows used in the multi-resolution DFT.

3.3.5 Dynamic partial-loudness estimation

When multiple sounds are presented simultaneously their individual loudnesses may decrease, depending on the spectral and temporal relations between them. When this happens, masking has occurred (refer to Section 2.3.2). Along with loudness, the intelligibility and timbre of a partially-masked sound are also affected. EP loudness models assume that the overlap of excitation within a critical band between the signal and masker is the primary source of energetic masking. Thus, if the spectral characteristics of the signal and masker are available, this allows the masked loudness, or partial loudness, of the signal to be estimated. Such a procedure has many practical applications such as predicting the audibility of real-world objects in the presence of noisy backgrounds. This is particularly pertinent to musical applications where multiple instruments are mixed together, and therefore these models are of interest to researchers in the field of Intelligent Music Production (IMP) (Gonzalez and Reiss 2008; Lopez and Janer 2010; Aichinger et al. 2011; Ward et al. 2012; Terrell et al. 2013; Ma et al. 2014; Terrell et al. 2014; Ford et al. 2015; Wichern et al. 2015).

Two approaches to measuring the partial loudness of multiple time-varying sounds are now presented. The first is an extension of Glasberg and Moore (2002), as published by Glasberg and Moore (2005), and the second is an extension of the procedure given by Chen et al. (2011) for dynamic sounds. Both implementations are presented within the context of multiple concurrent sound objects, rather than simply a single target stimulus and background stimulus.

Glasberg and Moore's method

The Cambridge hearing group developed a procedure for calculating loudness, partial loudness, and masked thresholds of stationary sounds (Moore and Glasberg 1997; Moore et al. 1997). The ability to take into account time-varying background noise when modelling loudness was developed later by Glasberg and Moore (2005), although the focus of that work was predicting masked thresholds, rather than validating predictions of partial loudness. The algorithm for estimating the partial loudness of multiple dynamic sounds is now presented, where the term 'mix' is used to refer to the mixture of all sounds at a given point in time (analysis frame).

1. Apply an appropriate outer- and middle-ear transfer function to each input.

2. Estimate the short-term power spectrum of each source.

3. Calculate the EP of the mix and store the shapes of the roex filters.

4. Using the stored roex filters, calculate the EP of each sound.

5. Using the stored roex filters, calculate the EP of each combination of mix minus source, i.e. the EP of every 'background'.

6. Use the EPs of the maskers to calculate the masked threshold of each sound.

7. Calculate partial specific loudness (PSL) patterns from the EPs and masking thresholds.

8. Sum the PSL over the Cam scale to arrive at the instantaneous partial loudness (IPL).

9. Smooth the IPL to obtain the short-term partial loudness (STPL).

10. Smooth the STPL to obtain the long-term partial loudness (LTPL).

The majority of these steps are detailed in Section 3.3.3, but the implementation of stages 3–7 are slightly different when calculating partial loudness. In particular, the excitation transformation is more involved compared to the single sound procedure because the perceptual effects of the mixture must now be accounted for. To derive the appropriate shapes of the auditory filters, the level per ERB used to compute the slope of the roex filter (Equation 3.48) is calculated using the power spectrum of the mix $\hat{S}_{\text{mix}}$, rather than the spectrum of each input individually, i.e.

$$L_{\text{ERB}}[j] = 10\log_{10}\left(\sum_{k=0}^{K-1} W_{j,k}\hat{S}_{\text{mix}}[k]\right). \tag{3.73}$$

Therefore, the same set of auditory filters are used for *all* sounds. In theory, this approach should be used for estimating the loudness of all sounds comprising a mixture, regardless of the transformation relating excitation to SL, since the auditory filters are excited by the entire mix at any point in time. This is also advantageous from a computational perspective, since the shapes of the filters need only be computed once for a given frame.

After generating all N EPs, the next stage is to compute masked thresholds. Since all sounds are analysed independently, each one has $N-1$ potential maskers. The EP corresponding to the masker of the n^{th} sound is therefore estimated as

$$E_{\text{mask},n} = \sum_{\substack{0\le m\le N-1\\ m\ne n}} E_m. \tag{3.74}$$

The masked threshold of sound n is then calculated using

$$E_{ThM,n} = kE_{\text{mask},n} + E_{\text{ThQ}}, \tag{3.75}$$

where k is the signal-to-noise ratio at the output of the auditory filter required for threshold, and E_{ThQ} is the threshold in quiet as given by Equation 3.53. Parameter k is shown pictorially by Moore et al. (1997), but are well approximated in dB using the following polynomial (see Figure 3.15)

$$k(fl) = 317.075 - 418.886f_l + 212.891f_l^2 - 49.680f_l^3 + 4.468f_l^4, \tag{3.76}$$

where fl is the base 10 logarithm of the centre frequency of the auditory filter. For frequencies above 1 kHz, k is limited to -3.01 dB.

With the EPs and masked thresholds computed for each sound, the transformation from excitation to PSL can now be carried out. Moore et al. (1997) give four equations for this transformation, the choice of which depends on the relative excitation between the sound and the mixture, as well as the total amount of excitation in a given critical band. These equations define the partitioning

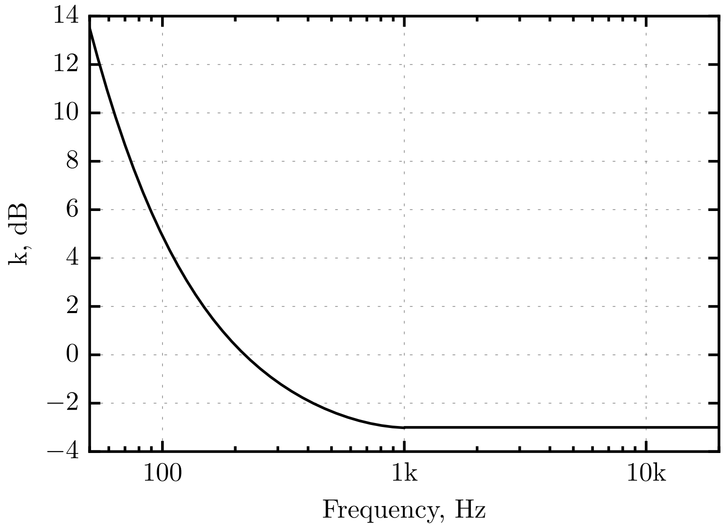


Figure 3.15: Signal-to-noise ratio at the output of the auditory filter at masked threshold as a function of centre frequency, according to Equation 3.76.

of excitation between the two signals, in accordance with loudness growth data of tones partially-masked by noise. Let E and E_{mix}, denote the EP of a given source and the mix, respectively. As above, E_{ThM} is the masked threshold and E_{ThQ} is the threshold in quiet. All variables are a function of critical-band number, which has been omitted for brevity. The PSL, denoted ψ', is then calculated using the following formulae:

- When $E \geq E_{\mathrm{ThM}}$ and $E_{\mathrm{mix}} \leq 100$ dB:

$$\begin{aligned}\psi' = {} & \beta((GE_{\mathrm{mix}} + A)^{\alpha} - A^{\alpha}) \\ & - \beta((G(E_{\mathrm{mask}} + E_{\mathrm{ThM}}) + A)^{\alpha} \\ & - (GE_{\mathrm{ThQ}} + A)^{\alpha}) \left(\frac{E_{\mathrm{ThM}}}{E}\right)^{0.3}\end{aligned} \tag{3.77}$$

- When $E < E_{\mathrm{ThM}}$ and $E_{\mathrm{mix}} \leq 100$ dB:

$$\begin{aligned}\psi' = {} & \beta\left(\frac{2E}{E + E_{\mathrm{ThM}}}\right)^{1.5} \\ & \times \frac{(E_{\mathrm{ThQ}}G + A)^{\alpha} - A^{\alpha}}{(G(E_{\mathrm{mask}} + E_{\mathrm{ThM}}) + A)^{\alpha} - (E_{\mathrm{mask}}G + A)^{\alpha}} \\ & \times ((GE_{\mathrm{mix}} + A)^{\alpha} - (GE_{\mathrm{mask}} + A)^{\alpha})\end{aligned} \tag{3.78}$$

- When $E \geq E_{\mathrm{ThM}}$ and $E_{\mathrm{mix}} > 100$ dB:

$$\begin{aligned}\psi' = {} & \beta_2(E_{\mathrm{mix}})^{0.5} \\ & - \beta_2((E_{\mathrm{mask}} + E_{\mathrm{ThM}})^{0.5} \\ & - (GE_{\mathrm{ThQ}} + A)^{\alpha} + A^{\alpha}) \left(\frac{E_{\mathrm{ThM}}}{E}\right)^{0.3}\end{aligned} \tag{3.79}$$

- When $E < E_{\text{ThM}}$ and $E_{\text{mix}} > 100$ dB:

$$\begin{aligned}\psi' = \beta &\left(\frac{2E}{E + E_{\text{ThM}}}\right)^{1.5} \\ &\times \frac{(GE_{\text{ThQ}} + A)^{\alpha} - A^{\alpha}}{(E_{\text{mask}} + E_{\text{ThM}})^{0.5} - (E_{\text{mask}})^{0.5}} \\ &\times ((E_{\text{mix}})^{0.5} - (E_{\text{mask}})^{0.5})\end{aligned} \tag{3.80}$$

The constants in these equations are $\beta = 0.046871$ and $\beta_2 = \beta/(1.04 \times 10^6)^{0.5}$. Note that the last two equations apply when the level of the mixture is very high, whereby Moore et al. (1997) chose to model the relationship between excitation and loudness as being more linear that at medium levels. For consistency with the revised procedure given in ANSI S3.4 (2007), where a compressive relationship is maintained (see Figure 3.10), the exponent of 0.5 in these two equations should be replaced by a value of 0.2, and $\beta_2 = \beta/(1.0707)^{0.2}$.

Chen and co-workers' method

As discussed in Section 3.3.4, Chen et al. (2011) presented a more recent loudness model for stationary sounds which was later developed to handle time-varying stimuli (Chen and Hu 2012). In the first paper, the authors presented a simplified version of the partial loudness method described by Moore et al. (1997) and reported similar predictive accuracy with the exception that the partial loudness of complex signals with components falling below threshold is not accounted for. The model makes the assumption that only signal excitation above the masked threshold contributes to partial loudness. The dynamic model is here extended to compute time-varying partial loudness based on the stationary-sound implementation. Using the same notation as above, the PSL is therefore calculated as

$$\psi' = \max(0, E - kE_{\text{mask}}), \tag{3.81}$$

where parameter k defines the detection efficiency of auditory filter at a given centre frequency. In this model, parameter k (in dB) is described by the function (Chen et al. 2011)

$$k(f_c) = 6.51(1.37 \log_{10}(0.00437 f_c + 1) - 1)^2 - 1.93. \tag{3.82}$$

The terms inside the brackets of this equation correspond to a normalised version of the Cam scale where zero represents a frequency of 1 kHz.

Note that because the excitation transformation used in this model is compressive, no further transformation to PSL is required. The authors do not comment on whether the gain of the active filter of the double-roex filter should be derived from the power spectrum of the combined signals, but in theory this would be the case.[10] Experiments by the author suggest that this is indeed necessary to replicate Figure 12 of Chen et al. (2011). Thus, the total power used to derive the gain of the active filter should be calculated from the output of the passive filter in response to all sounds (refer to Section 3.3.4), i.e.

$$E_{P_i} = \sum_{k=0}^{K-1} W_{P_{i,k}} \hat{S}_{\text{mix}}[k]. \tag{3.83}$$

Once the PSL has been calculated, the IPL, STPL and LTPL are computed using the standard

[10] This also simplifies the computational cost of the algorithm when calculating the partial loudness of multiple sources.

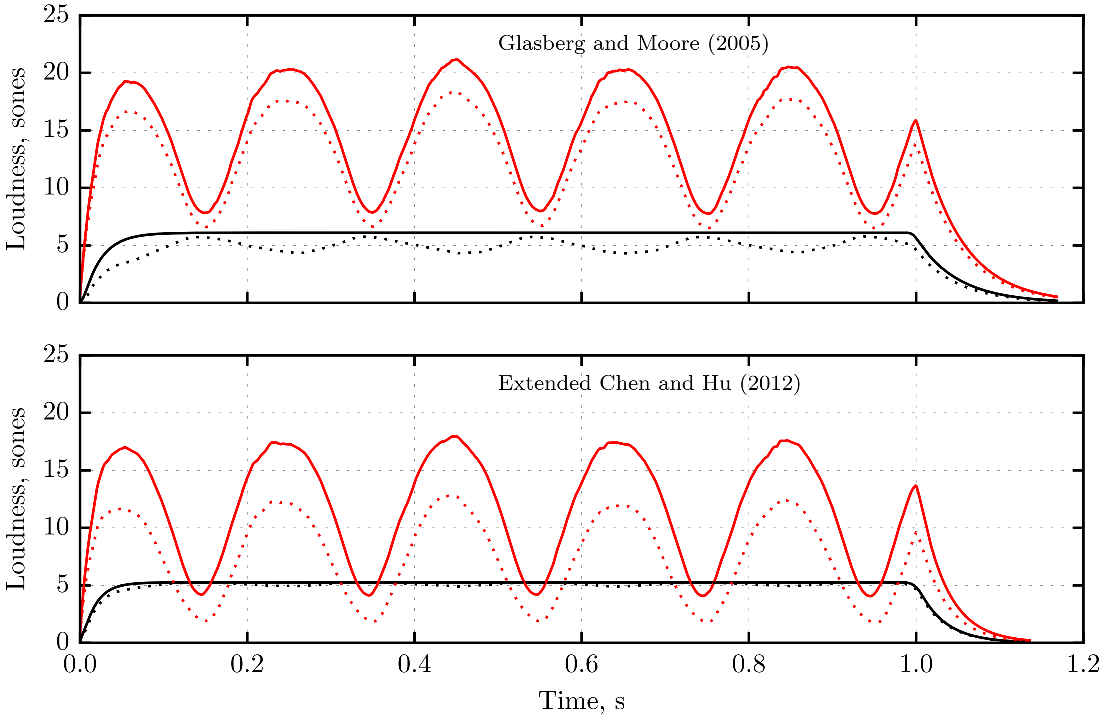


Figure 3.16: STL (solid lines) and STPL (dotted) of a 1 kHz pure tone (black lines) and amplitude-modulated white noise (red), as predicted by the model of Glasberg and Moore (2005) (top) and the proposed extension of Chen and Hu (2012) (bottom).

protocol.

Figure 3.16 demonstrates the two dynamic partial loudness models by showing the STL and STPL of a 1 kHz pure tone and an amplitude-modulated white noise, both of which were presented at 65 dB SPL in the free field with frontal incidence. It can be seen the model of Glasberg and Moore predicts that both stimuli, when presented in isolation, are louder than predicted by the model of Chen and Hu. This is partly explained by differences in the spectral broadening introduced by the two configurations of the multi-resolution DFT. Both models predict a change in loudness dynamics as introduced by energetic masking: it can be seen that the reduction in loudness of the tone is most pronounced when the noise is at its highest in level. The amount of masking, inferred from the reduction in loudness, also appears to differ between the two procedures: the model of Glasberg and Moore shows greater masking of the tone compared to the noise, whereas the model of Chen and Hu shows the opposite effect.

3.3.6 Comparison of stationary model predictions

In the preceding sections three important multiband loudness models for predicting the loudness of stationary sounds were discussed. The first was developed by Zwicker in the 1960s and became the basis for international (ISO 532 1975) and German standards (DIN 45631 1991). In the late 1990s, Moore and co-workers revised Zwicker's procedure in light of recent masking data measured using systematic psychophysical techniques, as well as further empirical evidence of perceptual phenomena as measured in more recent studies. The updated procedure was standardised in 2005 and further revised in 2007 (ANSI S3.4 2007). The two models are currently under review by the ISO standards committee and it is expected that Zwicker's model will form part A of the new ISO 532 standard, and Moore and Glasberg's model will form part B. Using an author/standard and publication year convention, for simplicity, the procedures of Zwicker and Moore and Glasberg are referred to as DIN1991 and ANSI2007, respectively. The third approach to loudness estimation is a recent model proposed by Chen et al. (2011) and differs from the two standardised techniques in that a separate stage for the transformation from excitation to SL is not needed. Instead, the functionality of the inner ear is simulated in a physiologically realistic manner using a filter bank incorporating a compressive nonlinearity. Therefore, both frequency selectivity and the active cochlear mechanism are modelled as a single process. This model is termed CHGM2011.

In order to give some insight into the predictive behaviour of these three procedures, model predictions for stationary sounds were compared. In particular, the following phenomena were investigated:

1. Loudness growth of a 1 kHz tone presented binaurally in the free field with frontal incidence.
2. Loudness growth of a broadband noise presented binaurally in the free field with frontal incidence.
3. Loudness as a function of spectral bandwidth.
4. Loudness as a function of frequency according to the ISO standardised equal-loudness contours.

Implementation and procedure

The implementation of DIN1991 follows the BASIC code published by Zwicker et al. (1991). This loudness model takes third-octave band levels as input, meaning that the stationary signal of interest must be processed by a fractional-octave band filter bank. To simplify the implementation and maintain consistency within a single modelling framework, a processing module was developed to transform a power spectrum into third-octave bands. In essence, the spectrum is defined by discrete components each with a specified frequency and intensity, which is then passed through the transfer function of a set of third-order bandpass Butterworth filters. Each bandpass filter has a one-third octave bandwidth with nominal centre frequencies spanning 25 Hz to 12.5 kHz, yielding 28 bands (Zwicker et al. 1991). The filters were implemented in accordance with the specifications of ANSI S1.11 (1986).

Except for the multi-resolution spectrogram and temporal integration stages, the implementation of ANSI2007 and CHGM2011 follows Sections 3.3.3 and 3.3.4, respectively. As with the implementation of Zwicker's model, the input to both of the models is a single intensity spectrum.

In psychophysical experiments, the point of subjective equality (PSE) is typically measured using a loudness-matching paradigm in which the subject varies the gain of a comparison sound until its loudness matches that of a reference or target sound. The dependent variable is therefore the difference in intensity level between the comparison and reference sounds at the point of equal loudness. To make predictions for such cases, an iterative procedure was used to search for the optimal level difference or matching gain. Let Ψ_t and Ψ_c denote the loudness in sones of the target sound and the comparison sound, respectively. With the listening conditions specified, the loudness of the target is estimating using the model and its value is stored. The comparison sound is then input to the model at some initial intensity level L_c, and the ratio of the target loudness to the comparison loudness is treated as an error to minimise in decibels,

$$\varepsilon = 10\log_{10}(\Psi_t/\Psi_c). \tag{3.84}$$

The level of the comparison sound is then updated on the n^{th} iteration according to

$$L_{c,n} = L_{c,n-1} + \lambda\varepsilon_{n-1}, \tag{3.85}$$

where λ defines the proportion of the error to increment the input level. This iterative process ends when either a maximum number of iterations has been reached or the absolute error is less than a predetermined threshold. An error tolerance of ± 0.01 dB was used in this assessment.

This operation can also be used to estimate the loudness level, in phons, of any input source without the need for a sone-to-phon converter based on a LUT which may compromise accuracy. The LUT approach, for example, is based on sampling the loudness growth function of a 1 kHz tone presented binaurally in the free field with frontal incidence, and then applying interpolation in order to estimate the loudness level corresponding to an arbitrary input loudness given in sones. In contrast, the above iterative procedure involves setting the target sound to the stimulus of interest, with a 1 kHz tone serving as the comparison sound. After running the procedure with a specified tolerance, the resulting intensity level of the 1 kHz tone denotes the loudness level of the target stimulus. Conversely, the reverse of this method can be used to find the intensity (or relative level adjustment) of an arbitrary sound at a specified target loudness level (Appendix B addresses the accuracy of using a sone-to-phon approximation for such a task). Equal-loudness contours can therefore be estimated using this method. Such iterative procedure have been used for loudness normalising arbitrary sounds (Moore et al. 2003; Simpson and Reiss 2011; Ward et al. 2012; Terrell 2013). The method described here is essentially the same as the one used by Terrell et al. (2013) for optimising the relative-loudness balance of a musical mixture.

Results

Loudness as a function of level

Figure 3.17(a) shows the output of the models in response to a 1 kHz pure tone presented binaurally in the free field at sound levels from 0 to 120 dB SPL. For reference, the black line corresponds to the loudness-growth function specified in ISO 226 (2003), as derived from the equal-loudness contours. It can be seen that both the DIN1991 and ANSI2007 agree with the ISO function, especially over the middle range of input levels. The ANSI2007 shows a slightly better match with the ISO data below 30 dB SPL. Predictions of the CHGM2011 show a similar trajectory as the ISO curve up to about 75 dB where the slope of the loudness function begins to increase. The exponent of a power law relating the intensity of the tone and the loudness can be estimated from the slope of the log(sone) vs log(intensity) function. Between 100 and 120 dB SPL the exponent was 0.99 for the CHGM2011 and 0.27 for the ISO function, with the latter being constant above about 40 dB SPL. Few studies have reported on the loudness at very high levels, so it is not clear how the disagreement between the models at high sound levels should be treated. Chen et al. (2011) cite the work of Viemeister and Bacon (1988) who reported significantly larger exponents (around 0.45) at levels above 100 dB SPL. Recall that in the 1997 version of Moore and Glasberg's procedure (Moore et al. 1997), the function relating excitation to SL was steeper above 100 dB SPL compared to lower levels, but this was later removed in ANSI S3.4 (2007) based on a lack of data.

At medium levels, the loudness of a 1 kHz pure tone doubles for every 10 dB increase in intensity (see Section 3.2.1). Between 40 and 90 dB SPL (in steps of 1 dB), the predictions of the two standardised procedures were found to be within 5.2% of the expected sone values. The model of CHGM2011, however, deviates from Stevens' power law for sensory magnitude to a greater extent, especially at levels above 75 dB, as shown in Figure 3.17(a).

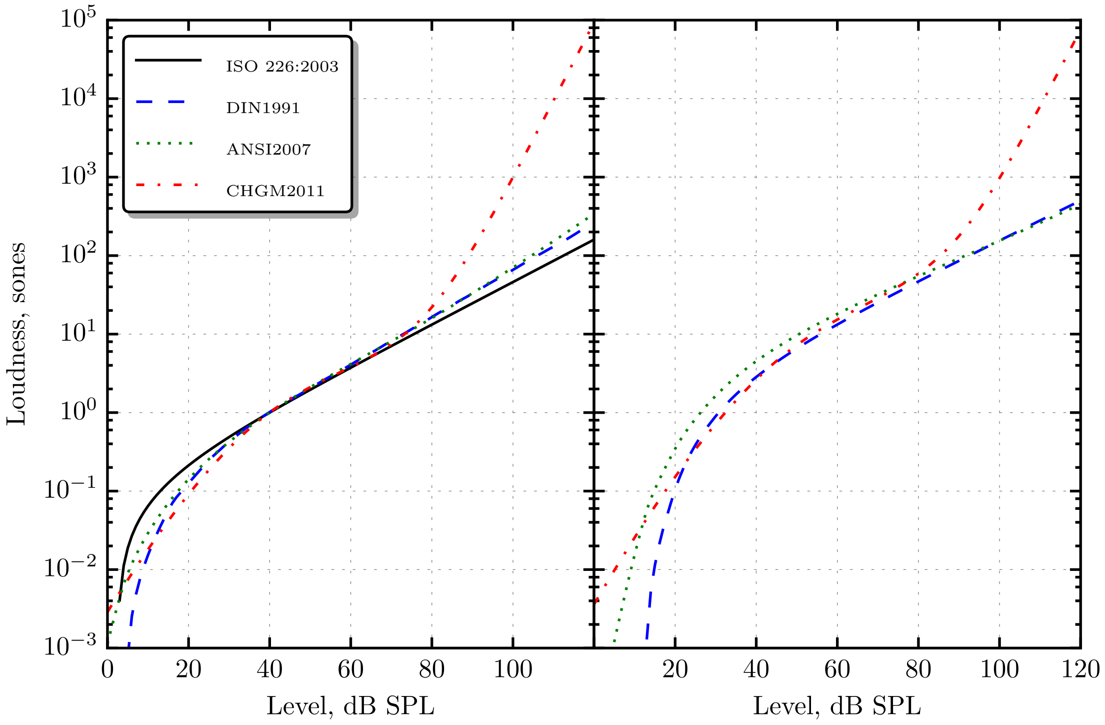


Figure 3.17: Predicted loudness in sones for (a) a 1 kHz pure tone and (b) a band of pink noise spanning 50 Hz to 15 kHz, as a function of overall sound level. Both stimuli were presented binaurally in the free field with frontal incidence.

The loudness-growth function according to ISO 389-7 (2005) predicts a value of zero when the input level falls below 2.4 dB SPL (the threshold of a 1 kHz sinusoid at threshold). The DIN1991 also predicts zero near threshold as very low levels yield negative estimates that get clipped in the model. In contrast, both ANSI2007 and CHGM2011 output finite loudness and thus are able to predict subthreshold values. This is important when estimating thresholds of multitone complexes when the level of the individual components are below their respective thresholds, yet the complex as a whole is audible (Moore et al. 1997).

Figure 3.17 (b) shows the model predictions of a band of pink noise with a frequency range of 50 Hz to 10 kHz. The overall intensity of the noise ranged from 0 to 120 dB SPL. It can be seen that, compared to the DIN1991, the ANSI2007 predicts greater loudness over all sound levels up to about 90 dB SPL. For levels between 40 and 90 dB SPL, the predictions of the ANSI2007 were approximately 4 dB (a level difference) higher than those of the DIN1991, but the differences are more marked at lower levels. This is consistent with the findings of Fastl et al. (2009) who found that the ANSI2007 pink noise predictions were around 5 dB higher than those predicted by DIN1991.[11] Note that the growth functions for the 1 kHz pure tone are, however, very similar for these two models, suggesting that the ANSI2007 calculates greater SLS compared to the DIN1997, consistent with both theory and other empirical work. Schlittenlacher et al. (2011), for example, conducted an experiment in which subjects had to match the loudness of a pink noise and a 1 kHz tone (and vice versa) over a range of levels. When comparing the loudness matches to the predictions of DIN1991 and ANSI2007, the latter performed the worst in that predictions were notably higher than the experimental data. As described by Schlittenlacher et al. (2012a), a possible reason for the overestimation of broadband sounds by ANSI2007 is that this model assumes

[11] Fastl et al. (2009) used a different specification for the pink noise compared to the one used here.

narrower auditory filters compared to those employed by the DIN1991, which consequently leads to greater SLS. Interestingly, for levels between 20 and 80 dB, the CHGM2011 shows slightly better agreement with the DIN1991 compared to the ANSI2007.

Loudness as a function of bandwidth

In this section the predictions of the three models are compared with the data of Leibold et al. (2007) who conducted a band-widening experiment to determine the effect of the bandwidth of a multitone complex on perceived loudness. The level of a 1 kHz tone when equal in loudness to a five-tone complex was estimated using a two-interval forced choice loudness-matching procedure. Loudness matches were obtained for six different bandwidths as defined by the logarithmic spacing between adjacent components. The geometric mean was always 1 kHz. Six frequency ratios were used: 1.012, 1.023, 1.059, 1.120, 1.260, and 1.586; corresponding to spectral bandwidths of 46, 92, 231, 456, 965 and 2119 Hz, respectively. The level of each component was set to 53 dB SPL, giving an overall intensity level of 60 dB SPL.

Model predictions were made using the aforementioned iterative loudness-matching procedure with the five-tone complex assigned as the target sound and the 1 kHz tone as the comparison sound. The level of the 1 kHz tone at the point of equal loudness was recorded. Each complex tone was geometrically centred at 1 kHz at one of 50 bandwidths spanning 30 Hz to 3 kHz. Levels of the components comprising the complex sound were set to 53 dB SPL. The models were configured with a diffuse-field transmission to approximate the effects of the Sennheiser HD 250 linear 2 headphones which have a diffuse-field response.

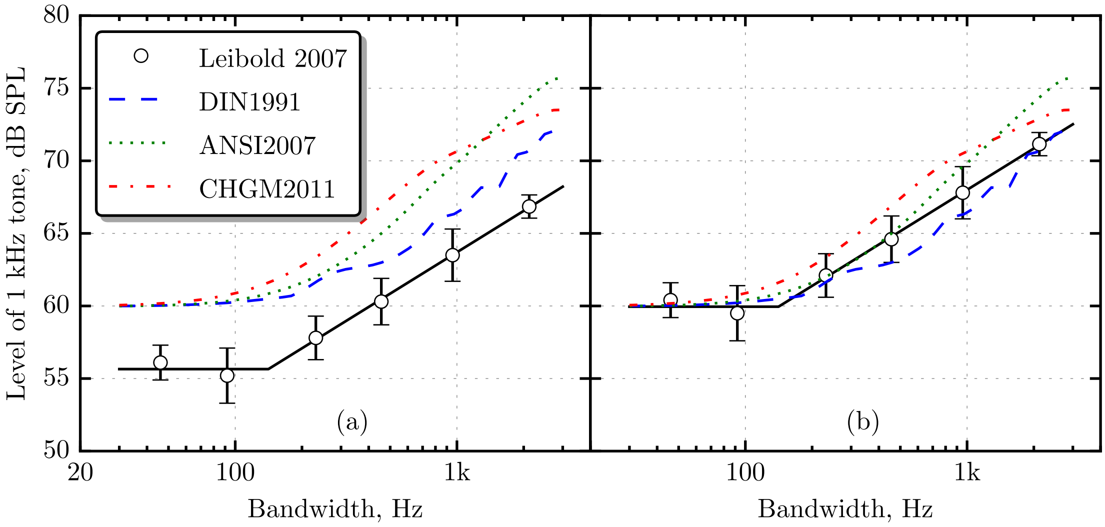


Figure 3.18: Loudness level of a five-tone complex as a function of bandwidth. (a) The circles are the mean level, averaged across seven subjects, of a 1 kHz tone judged equal in loudness to the complex tone at 60 dB SPL according to Leibold et al. (2007). Error bars show the standard error of the mean level across subjects. (b) Same as (a) but with experimental data shifted to simulate perfect intensity summation within the critical-band.

The circles in Figure 3.18 represent the mean level, averaged across seven subjects, of a 1 kHz pure tone judged equal in loudness to the complex tone presented at 60 dB SPL. The black solid lines are those that best fit the data with the constraint that one of the lines is horizontal (Leibold et al. 2007). The point of intersection was 141.5 Hz, which is close to the ERB of 132.4 Hz (Equation 2.26).

For the two narrowest bandwidths (46 and 92 Hz), the loudness of the multitone complex is constant. This occurs because the ERB at 1 kHz is 132.4 Hz which means that, for these two bandwidths, the constituent components of the complex sound fall within a single critical band. In contrast, when the bandwidth of the complex tone exceeds the critical bandwidth, as in the four remaining conditions, components falling within separate filters are summed and then compressed *within* their respective bands, and then summed across all critical bands. Due to the compression, the total loudness is greater when across critical-band loudness summation takes place.

Turning to the model predictions, all three give estimates that are initially about 4 dB higher than the empirical data. Given that the overall level of the multitone complex was 60 dB SPL, one would expect the same level for the 1 kHz tone in the first two conditions under the assumption of intensity addition within the critical band as discussed above (see also Fastl and Zwicker (2007)). Although previous studies have reported less than perfect within-band intensity summation, Leibold et al. (2007) reported that their data were biassed by the performance of two subjects. Subplot (b) represents the same data if loudness summation within a critical band were perfect. In this representation, the ANSI2007 gives an excellent fit to the first four data points but the slope of the function increases thereafter. At bandwidths greater than 200 Hz, the predictions of the DIN1991 were, on average, 2.5 dB and 3.1 dB lower than those given by the ANSI2007 and CHGM2011, respectively. This is partly explained by the broader filters employed by the DIN1991. SLS is therefore lowest for the DIN1991 but its predictions are in slightly better agreement with the subjective data. Additional differences may arise from the form of the transmission effects of the outer ear and the procedure used to calculate spectral masking. Also note that the DIN1991 contour shows irregularities beyond 200 Hz compared to those output by the more recent procedures. This can be explained by the higher number of filters used in the more recent models: the ANSI2007 uses 372, the CHGM2011 uses 388; compared to the set of 28 third-octave bands employed by the DIN1991. Furthermore, unlike the DIN1991, which makes use of LUTs, both ANSI2007 and CHGM2011 calculate EPs using analytical formulae which produces smooth changes when the parameters defining the input stimulus are varied. Finally, the slope of the CHGM2011 contour becomes shallower at bandwidths greater than 1 kHz, meaning that SLS decreases as the components begin to span several critical bands. The ANSI2007 does show a slight drop-off in loudness at the very highest bandwidth, but the trend is not as marked as with the CHGM2011. Such differences are attributable to the responses of the auditory filters to the input spectra, the amount of critical-band compression, and again differences due to the assumed outer- and middle-ear transfer functions. One must also acknowledge that as the level of the pure tone approaches higher intensities, the CHGM2011 becomes more linear, meaning the growth of the loudness function becomes steeper for the pure tone as shown in Figure 3.17.

The models are now compared using more recent data from an experiment investigating the subjective loudness of bandpass filtered pink noise. Figure 3.19 shows the subjective results and corresponding model predictions for the first experiment of Schlittenlacher and Ellermeier (2015).[12] In this experiment, 20 subjects matched the loudness of a bandpass filtered pink noise to that of a 1 kHz pure tone presented at different intensities. If the level of the tone is higher than that of the pink noise at the point of equal loudness, SLS is assumed. Model predictions were made by measuring the matching level required to equate the loudness of a band of pink noise to the loudness of a 1 kHz reference tone presented at levels of 40–90 dB SPL in steps of 5 dB. As shown by plots (a) to (c), the following three bandwidths were used: 0.125–1 kHz, 0.5–2 kHz and

[12] The empirical data were derived by visual inspection of Figure 1 in the paper. The model predictions were computed using the stationary loudness implementations of the models developed in this chapter. The predictions of the DIN1991 and ANSI2007 are consistent with those displayed by Schlittenlacher and Ellermeier (2015).

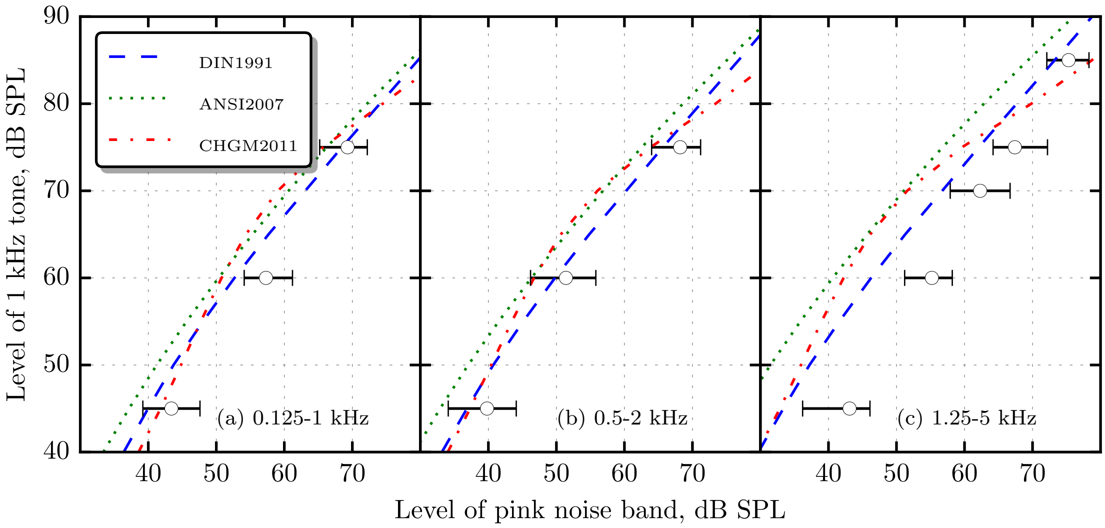


Figure 3.19: Level of a pink noise band required to match the loudness of a 1 kHz pure tone. Results for the low, mid and high spectral bands are shown from left to right. White circles show the mean of 20 subjects and error bars show the IQRs as given in Schlittenlacher and Ellermeier (2015).

1.25–5 kHz. In general, all three models over predict the loudness of the pink noise, particularly for the high-frequency noise band. In other words, the algorithms predict too much SLS. The DIN1991 gives the best fit overall in all three conditions, especially for the mid-frequency band where its predictions fall within the three IQRs of the subjective adjustments. The CHGM2011 sits somewhere between ANSI2007 and DIN1991, being closer to the former when the tone is presented in the region of 60–70 dB SPL. At levels above this range, the amount of SLS predicted by this algorithm decreases (smaller level differences) more substantially compared to the other models, especially for the high-frequency noise band. This is explained by the active auditory filter used by the CHGM2011 which is most compressive at moderate intensity levels (see Figure 3.14(a)). Thus, the loudness of the pure tone increases more rapidly when the level exceeds about 70 dB SPL, as shown in Figure 3.17(a), and as a result, the level of the pink noise band must be higher than when the tone is presented at moderate levels. In contrast, the curvature is less pronounced for the ANSI2007 and DIN1991, and the predictions of these two models are most comparable for the low-frequency noise band.

Equal-loudness contours

Equal-loudness contours were predicted using the automatic loudness-matching algorithm. A 1 kHz pure tone presented binaurally in free field with frontal incidence was used as the target sound. The comparison sound was a pure tone with frequency set to one of the 29 covering 20 to 12 500 Hz as given in Annex B of ISO 226 (2003). For frequencies below 50 Hz, the DIN1991 did not converge at loudness levels of 30, 40, 60, 70 and 100 phons.

The solid black curves in Figure 3.20 show the equal-loudness contours according to ISO 226 (2003). The red dotted curves show those predicted by each model. It can be seen that the predictions of the DIN1991 are too shallow below about 400 Hz, especially at loudness levels below 70 phons where discrepancies begin to exceed 10 phons. The grey dashed curves in the top subplot are the old equal-loudness contours specified in ISO 226 (1987). Note that the DIN1991 gives a better fit to these data which represented the ISO standard at the time that the computer

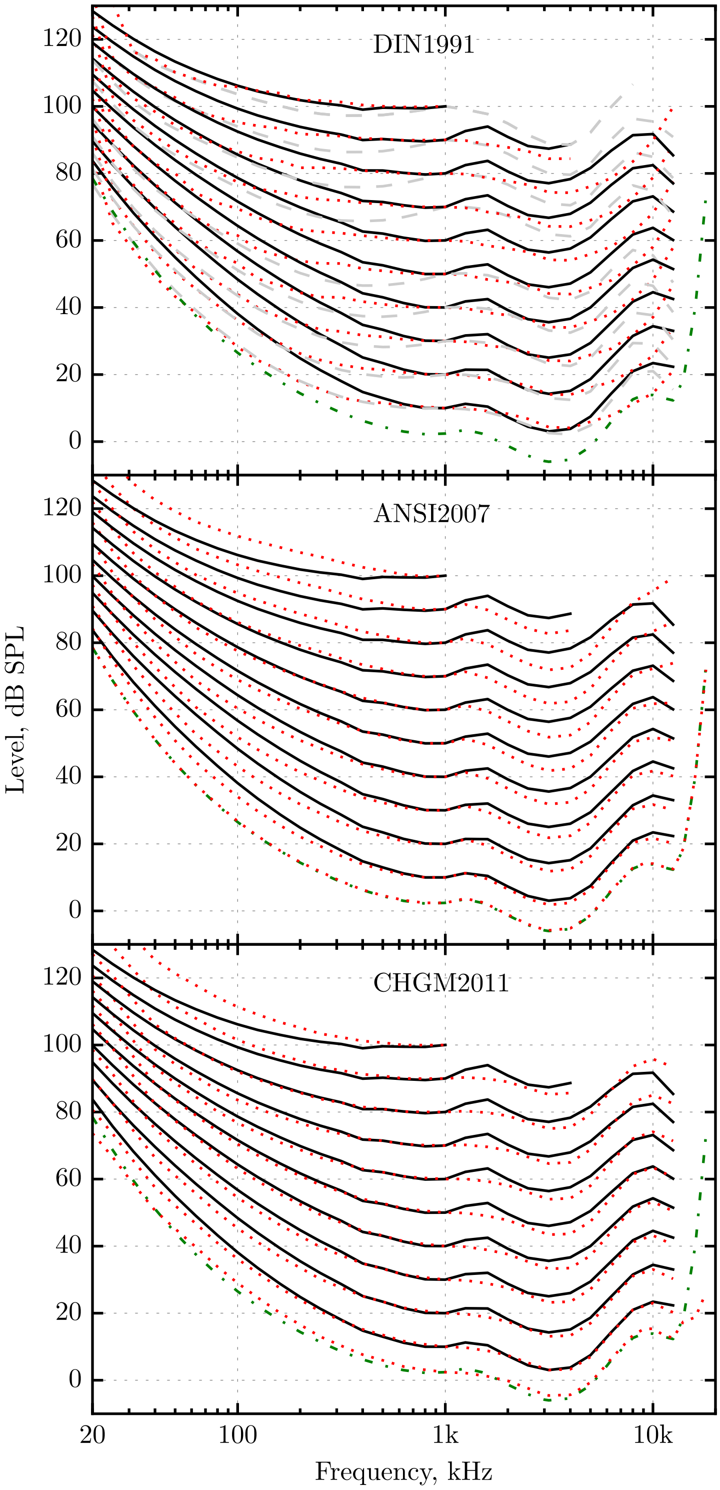


Figure 3.20: Comparison of equal-loudness contours between those specified in ISO 226 (2003) (black solid curves) and those predicted by three loudness models (red dotted curves). The grey dashed curves in the top subplot correspond to the equal-loudness contours specified in ISO 226 (1987). The green dashed-dotted curve is the absolute threshold data given in ISO 389-7 (2005).

program for the DIN1991 was published. The 1987 curves are generally flatter than those given by the revised contours, especially at frequencies between 100 and 1000 Hz. These observations are consistent with those of Charbonneau et al. (2009) who compared the predictions the two standardised procedures.

Both ANSI2007 and CHGM2011 models fit the ISO contours very well, which is not surprising considering that both models were designed to give a good fit to ISO 389-7 (2005) and also audibility thresholds for pure tones (Glasberg and Moore 2006; Chen et al. 2011). Predictions of the ANSI2007 between 2 and 5 kHz are notably lower than the reference values. Conversely, the CHGM2011 gives slightly better fit over all loudness levels in this frequency range. This can be explained by the greater compression used by the CHGM2011 at centre frequencies in this region (at medium levels). In other words, there is greater variation in loudness growth across frequency for the CHGM2011 compared to the ANSI2007. Recall from Section 3.3.3 that the exponent used by ANSI2007 is constant ($\alpha = 0.2$) at frequencies above 500 Hz. This was done such that the ANSI2007 model accurately predicted the growth in loudness for a 1 kHz tone over 40–80 dB SPL, i.e. a loudness function with a slope of 0.3 in accordance with the definition of the sone scale (Moore et al. 1997).

Both the ANSI2007 and CHGM2011 show large deviations from the ISO data at very low frequencies at medium to high loudness levels. For loudness levels above 80 phons, the models also tend to underestimate the loudness of frequencies above 8 kHz where the ANSI2007 no longer maintains the curvature of the reference contours. In terms of overall fit, the average root mean square error (RMSE) between the ISO data and the predicted values (converged only) for loudness levels from 10 to 100 phons was 6.0 (95% CI = 1.4), 3.4 (95% CI = 0.8) and 2.9 dB (95% CI = 1.2 dB) for the DIN1991, ANSI2007 and CHGM2011, respectively.[13]

The two most recent loudness models were developed to give a good fit to hearing thresholds for pure tones as specified in ISO 389-7 (2005). It is therefore interesting to compare the predictions for these data. The green dashed-dotted curve in Figure 3.20 shows the absolute threshold curve as per the standard. Note that no predictions are shown for the DIN1991 as this model was unable to produce stable loudness estimates at low sound levels using the iterative procedure. In any case, the predictions for the 10 phon contour are markedly worse compared to the two recent loudness models and so the goodness-of-fit are not of interest here. According to ANSI S3.4 (2007), predictions of absolute threshold of an arbitrary spectrum can be calculated by varying the input gain until a reading of 0.003 sones (2.2 phons) is measured (ANSI S3.4 2007). Similarly, Chen et al. (2011) state that the loudness at threshold (according to the model) corresponds to 0.004 sones or 2 phons. Thus, the iterative procedure was adapted by specifying a target loudness of either 0.003 sones or 0.004 sones, depending on the model. Predictions were made for frequencies from 50 Hz to 12500 Hz at values specified in Table 1 of ISO 389-7 (2005). For frequencies between 50 Hz and 12500 Hz, the RMSE was 0.11 dB for the ANSI2007 and 1.04 dB for the CHGM2011. The maximum absolute errors were 0.20 dB and 1.46 dB. These errors are comparable to those reported by the respective authors of the two algorithms (Glasberg and Moore 2006; Chen et al. 2011).

To reduce discrepancies between the data of ISO 389-7 (2005) and the predictions of ANSI2007 for frequencies around 3 kHz, Schlittenlacher et al. (2012a) proposed a revised function that takes into account variation in the loudness growth at higher frequencies. To show this, the frequency dependent parameter α, defining the amount of compression per critical band, used in the transformation from excitation to SL by ANSI2007 was allowed to vary across the entire frequency range as done by Schlittenlacher et al. (2012a).[14] After experimenting with various functions, the following

[13] Confidence intervals were thus based on the between-contour variation in RMSE. See appendix C for an overview of confidence intervals.

[14] Note that in addition to a modification to the exponent used to calculate the SL, the extension proposed by Schlittenlacher et al. (2012a) also included a transformation from the Cam scale to the Bark scale. The modification here is purely to show how the frequency-dependent compression may be adjusted to give better fit to equal-loudness

roex function was found to describe the desired variation in the exponent across frequency:

$$\alpha(g) = (1 + pg)e^{-pg} + c, \tag{3.86}$$

where g is the normalised deviation between the critical-band number of the auditory filters, in Cams, and the critical-band number $\hat{z}_c$ where the compression is lowest:

$$g_i = \frac{z_{c,i} - \hat{z}_c}{\hat{z}_c}. \tag{3.87}$$

Parameter p defines the slope of the function and c is an offset. Asymmetry is introduced by allowing the slope of the function to vary on both sides of the minimum. The four parameters of this function were found using a constrained optimisation routine which minimised the average RMSE across absolute thresholds and ten equal-loudness contours spanning 10–100 dB SPL. The function was constrained such that α was fixed at 0.2 at 1 kHz to preserve the doubling of loudness per 10 dB increase in level for a 1 kHz pure tone. The optimal function is shown in Figure 3.21(a). Notice that the amount of compression increases with frequency up to 11 kHz (the frequency corresponding to $\hat{z}_c$), consistent with the need for greater compression to account for the discrepancies around 3–4 kHz (Glasberg and Moore 2006; Schlittenlacher et al. 2012a). Above 11 kHz, the curve rapidly increases in order to account for the large discrepancies at very high frequencies presented at loudness levels above 60 phons. Predictions of the ISO 389-7 (2005) absolute thresholds and ISO 226 (2003) equal-loudness contours are shown in Figure 3.21(b), where the black lines show the reference data, and the green dotted and red dashed lines correspond to the ANSI2007 with and without the modified exponent shown in subplot (a), respectively. It can be seen that the 'dip' around 3.5 kHz is less pronounced for the modified model due to the incorporation of greater compression up to 11 kHz. The discrepancies are also less marked as a result of the sharp growth in α, although it is unlikely that this sudden increase is physiologically realistic. Noting that the accuracy of absolute threshold predictions are essentially preserved (RMSE = 0.7 dB), it may be concluded that the equal-loudness predictions of ANSI2007 can be improved if the α parameter, used to compress the EP, is allowed to vary above 500 Hz. See Schlittenlacher et al. (2012a) for a more complete refinement of this loudness model.

To conclude, the loudness model of Chen et al. (2011) gives the best overall fit to the current equal-loudness contours (ISO 226 2003), whilst the model of Moore and Glasberg (ANSI S3.4 2007) matches the absolute thresholds curve (ISO 389-7 2005) exceptionally well. The fit of the latter model to the equal-loudness contours can be improved if the amount of compression applied to the EP is allowed to vary across all critical bands. Zwicker's published loudness model (Zwicker et al. 1991) gives a closer fit to the now outdated equal-loudness contours (ISO 226 1987) but shows large deviations from the revised contours.

3.3.7 Nonsimultaneous spectral loudness summation

A systematic review of the dynamic loudness models of Glasberg and Moore (2002) (denoted GM02 from hereafter) and Chalupper and Fastl (2002) (denoted CF02) was conducted by (Rennies et al. 2010). The authors found that both models captured the main trends of empirical data collected from many psychophysical studies on loudness, but clear discrepancies were observed in some cases. For example, the predicted amount of temporal integration of repeated tone bursts was too low in both models, indicating the need for extended release time constants in the STL integrator. The authors also showed that the CF02 was able to better predict the loudness of

contours across a wide frequency range independently of other extensions.

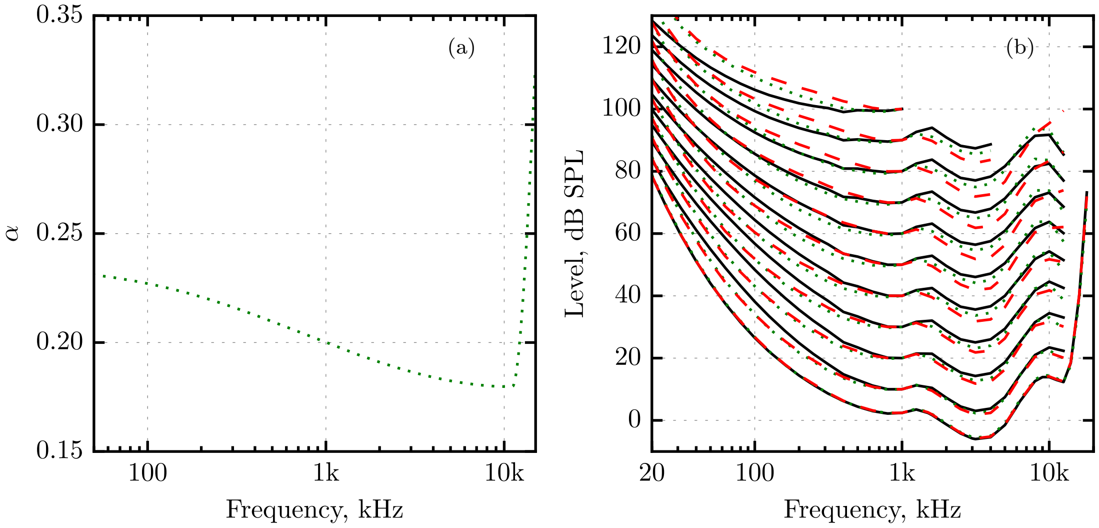


Figure 3.21: (a) Modified exponent defining the amount of critical-band intensity compression. (b) ISO 389-7 (2005) absolute thresholds and ISO 226 (2003) equal-loudness contours (back solid curves) and those predicted by the ANSI2007 with (green dotted lines) and without (red dashed lines) the modified exponent.

pulse trains with systematic spectro-temporal patterns, as well as frequency-modulated tones, than the GM02. The authors associated this with the simulation of SL decay within each critical band (Zwicker 1984). This additional stage captures the SLS of frequency components presented asynchronously in time. Although the GM02 and the model of Chen and Hu (2012) do not explicitly include such a stage, nonsimultaneous SLS still occurs due to the time-analysis windows used to compute the multi-resolution power spectrum. For mid- to high-frequency spectra, however, multi-resolution windowing does not sufficiently capture (nor was it designed to) the SLS arising from the nonsimultaenous presentation of tones with different frequency components (Rennies et al. 2010). In this section, different modifications to the GM02 are investigated to identify potential improvements when predicting the loudness of such stimuli.

Decay of specific loudness

Zwicker (1984) measured forward-masking curves using a 5 ms and 200 ms uniform masking noise, and test-tone impulses with frequencies of 2 kHz and 8 kHz. The results revealed strong differences in the post-masking decay between the threshold curves generated using the two maskers. Marked differences in thresholds were observed at masker-to-signal offset intervals between 5 and 40 ms. Zwicker transformed the masking data to SL-time functions, concluding that the decay of SL first follows an exponential decay with a time constant that decreases with masker duration, before slowly approaching the SL corresponding to threshold in quiet. Zwicker concluded that the value of the time constant can be obtained from an exponential function, characterised by a constant of 70 ms, of masker duration. The duration-dependent decay of SL was modelled using a nonlinear resistor-capacitor network, which was subsequently used for the post-masking stage of Chalupper and Fastl's (2002) dynamic model.

A more recent model of temporal processing in the auditory system, also based on the persistence of neural activity, is the temporal window model (for a review of temporal masking see Section 2.3.2). In short, a sliding window integrates the *compressed* intensity at the output of a given auditory filter. Forward masking occurs because the signal is masked by the persistence of neural activity evoked by the masker, as captured by the decay of the window. The shape of the temporal window is fixed, and it is the compressive device that follows the bandpass filter that is responsible for the nonlinear aspects of forward masking. The integration is, therefore, a linear process. Oxenham and Moore (1994) used this model to simulate the forward masking of Zwicker (1984) reasonably well. Oxenham and Plack (2000), conducted further forward-masking experiments, using the temporal-window model to simulate various conditions. They also found that the model was able to explain the dependence of masked threshold on masker duration, and attributed this effect to the level of the signal at threshold. This is because as the masker *duration* increases, the *level* at the output of the window also increases (up to a limit), so the rise in signal level needed for threshold is determined by the compression of the peripheral filter.

GM02 extension

If SL is seen to reflect the compressed intensity following the nonlinear compression performed in the cochlea, the temporal-window model might be used to model the persistence of SL needed to capture the SLS introduced by fluctuating sounds. Although at first sight this seems plausible, including the temporal window *after* the transformation from excitation to SL does not simulate the effects of nonsimultaneous SLS, because effectively, the window is convolved with every critical-band signal. This does not, at any point in time, lead to an increase in the area of the specific-loudness pattern for sequences of pure tones of different frequency, relative to the area produced by a tone of just one frequency. Hence, this approach does not introduce SLS. However, if the temporal window operates on the *excitation prior* to compression, then a sequence of nonsimultaneous frequency components effectively gives a broadband EP, which subsequently increases the area of the resulting SL pattern to a greater extent than for a single tone. A similar effect would arise if the input spectrum was smoothed over time prior to loudness analysis.

Based on the above, two approaches were used to test the hypothesis that temporal smoothing of SL (either directly or indirectly) improves the predictive ability of the GM02. The first approach was based on the temporal-window model of masking (Moore et al. 1988; Plack and Moore 1990; Plack and Oxenham 1998):

$$W(t) = \begin{cases} 0.975e^{t/4} + 0.025e^{t/29} & : t < 0 \\ e^{-t/3.5} & : t \geq 0, \end{cases} \tag{3.88}$$

where W is the intensity-weighting function and t is time (in ms) relative to the centre of the window. Thus, the temporal window is a weighted combination of a fast and slow exponential for times before the centre of the window, and a single fast exponential for times after. Although both skirts of the window describe the amount of temporal integration, the former (negative time) simulates forward masking and the latter (positive time) simulates backward masking. In practice, the time-integrated output is calculated by flipping this function about time zero and convolving it with the input signal (in units of intensity). A digital implementation of the temporal window yields an effective duration of roughly 16 500 samples ($f_s = 48$ kHz). Since the FIR filter must be applied to every auditory channel, this significantly increases the computational cost of the complete model. Härmä (1998) presented an efficient implementation by breaking W into a causal and noncausal part (see Lokki and Karjalainen (2002)). The extension to the GM02 makes use of the causal infinite impulse response (IIR) filter to simulate the decay of excitation in a given critical band, i.e. only forward masking is modelled. This decision was partly based on maintaining the efficiency of the original procedure, but primarily because the loudness model is based on the short-term power spectrum, which leads to a trade-off in time-frequency resolution that prohibits

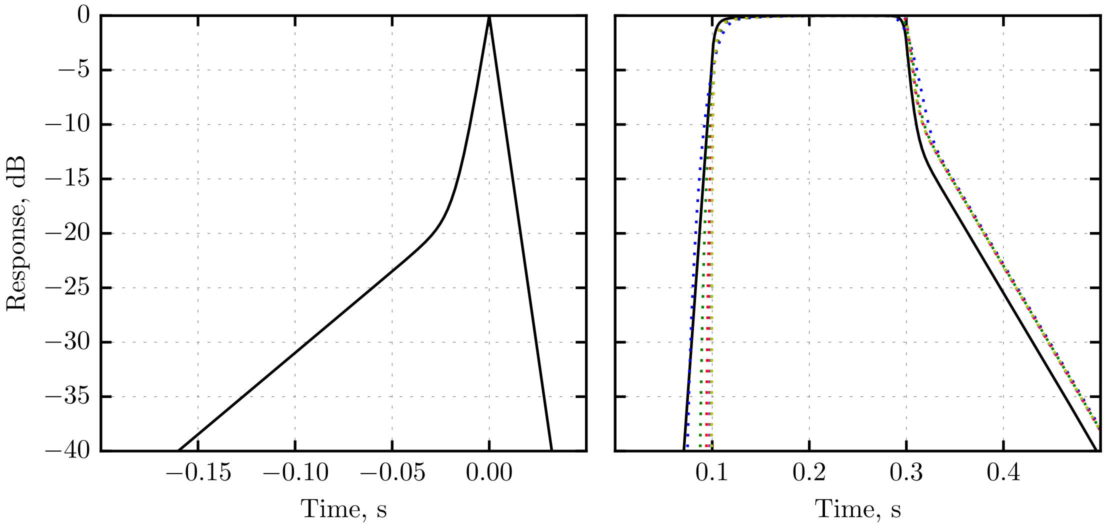


Figure 3.22: The black solid lines show (a) the temporal window and (b) its step response. The coloured dotted lines in (b) correspond to the step responses of the six Hann windows cascaded with the filter defined by Equation 3.89.

accurate modelling of backward masking (and thus the complete temporal window). It should be stressed that the temporal window is applied to the uncompressed excitation in each critical-band filter rather than the SL (as would be assumed based on recent theories of temporal masking).

After auditory filtering, the resulting excitogram was smoothed in time using the following recursive filter:

$$H(z) = \frac{1 + b_1 z^{-1}}{1 + a_1 z^{-1} + a_2 z^{-2}}, \tag{3.89}$$

where

$$\begin{aligned} b_1 &= -0.975e^{-1/(0.029f_s)} - 0.025e^{-1/(0.004f_s)}, \\ a_1 &= -e^{-1/(0.004f_s)} - e^{-1/(0.029f_s)}, \text{ and} \\ a_2 &= e^{-1/(0.004f_s)} e^{-1/(0.029f_s)}. \end{aligned}$$

The complete temporal window defined by Equation 3.88 is shown in Figure 3.22(a). The response of the window to a step function defined over $t = [0.1, 0.3]$ is shown by the solid black line in subplot (b). Notice that the rise time is faster than the decay time, which is necessary to explain backward and forward masking. The six coloured dotted lines in subplot (b) show the step response of the Hann windows, as used by the GM02, cascaded with $H(z)$ of Equation 3.89 in order to describe the step response of the loudness model in different frequency bands. The blue dotted line corresponds to a window of 64 ms and shows a build up at the onset of the input similar to the complete temporal window (black line). In contrast, the shorter windows show a faster initial response, but better match the decay trajectory of the temporal window. To preserve the rise time of the STL in response to tone bursts as originally proposed by Glasberg and Moore (2002), the attack time was reduced from 22 ms to 20 ms.

The second approach to modelling the decay of SL was to simply append a peak follower to the output of each auditory filter, following the SL transformation:

$$\psi_i'[m] = \begin{cases} \psi_i[m] & : \psi_i[m] > \psi_i'[m-1] \\ \psi_i'[m-1] + \alpha(\psi_i[m] - \psi_i'[m-1]) & : \text{otherwise.} \end{cases} \tag{3.90}$$

In this equation $\psi_i[m]$ is the SL in the i^{th} channel at frame m, and ψ' is the output of the peak follower. The release time determines the amount of decay in each critical band as the input falls. This approach is essentially a simplified version of Zwicker's (1984) nonlinear lowpass filter which makes use of duration-dependent decay times. Since the peak follower rises instantaneously, no modifications were made to the STL-time constants, although the total decay time in the STL is slightly extended. As a first approximation, the time constant used for α in Equation 3.90 was determined by fitting the extended GM02 to condition three of a subset of Zwicker's data (see below), taking into account the model's inherent discrepancy when predicting simultaneous SLS in condition one. The optimal time constant was 17 ms.

Results with the modifications

Predictions of data from two different studies (Zwicker 1969; Heeren et al. 2011) investigating nonsimultaneous SLS were made using the GM02, with and without the proposed modifications, and the CF02. The latter model was included because, as discussed, it includes a post-masking stage prior to SLS. In all simulations, the maximum of the STL-time function was used to measure the overall loudness of the input stimulus. Predictions were made using the iterative procedure described in Section 3.3.6.

Data of Zwicker (1969)

The same simulation of a subset of data from Zwicker's (1969) experiment as used by Rennies et al. (2010) to compare the two dynamic loudness models was used to test the proposed modifications. In Zwicker's experiment, subjects matched the loudness of a reference stimulus and test stimulus comprised of tone pulses with different but systematic complex spectro-temporal patterns. The shortest tone duration was 20 ms and, in this study, 4 ms raised-cosine ramps were applied to the stimulus onset and offset in order to reduce spectral broadening. Shorter ramp durations, such as 2.5 ms as used by Rennies et al. (2010) were found to have 1–2 dB influence on the model predictions, due to the multi-resolution DFT used by the GM02. Rennies et al. (2010) focused on four conditions involving three stimuli, which are demonstrated pictorially in Figure 3.23.

Condition one (C1) uses a 1.85 kHz 100 ms tone as the reference stimulus (top panel) and the simultaneous presentation of five 100 ms tones of different frequency presented simultaneously (see ordinate tick labels) for the test stimulus (bottom panel). In all conditions, the sound level of the individual tones was 70 dB SPL, and the level differences at the point of equal loudness were calculated as the difference between the levels of the individual tones comprising the test and reference stimuli. Therefore, a negative difference means that the test stimulus is louder than the reference stimulus. In theory, the loudness of the complex tone is expected to be louder than the single tone based on (a) the total intensity of the complex tone being greater than that of the pure tone, and (b) the constituent components of the test stimulus exciting different critical bands, leading to SLS.

In the first condition, Zwicker found that, for equal loudness, the level of each tone in the test stimulus had to be 23 dB lower than the level of the reference, of which about 16 dB was due to SLS (since the overall level of the test stimulus was roughly 7 dB higher). Condition two (C2)

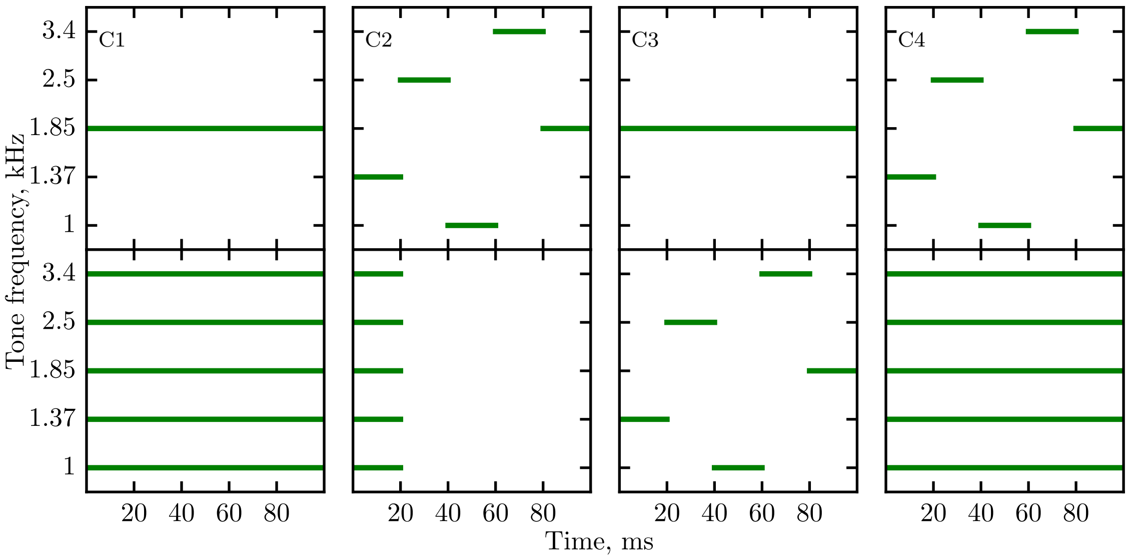


Figure 3.23: Diagram showing the stimuli used in four conditions (C1-C4) of Zwicker's (1969) experiment. Reference stimulus top, test stimulus bottom. Image based on Figure 8 of Rennies et al. (2010).

compares the loudness of a train of five 20 ms pulses each of different frequency, with that of the same five pulses presented simultaneously. If the ear fully integrates over both time and frequency, the pulse train, with a total length of 100 ms, should be louder than the 20 ms complex tone. The average level difference between the test and reference stimuli was -3 dB meaning that the integration of loudness over different critical bands was incomplete for the pulse train, and also that complex tone was softer than in the first condition due to its shorter duration.

Condition three (C3) compares the reference stimulus of C1, a single tone, with aforementioned pulse train. If full time-frequency integration is assumed, the pulse train should be louder than the single tone, since the two stimuli are equivalent in overall length. Zwicker reported a level difference of -11 dB, indicating nonsimultaneous SLS. The effect is not as large as in the simultaneous case (C1), where a -23 dB difference was measured. A proportion of this difference is, however, attributed to differences in the loudness of the tones at different frequencies (the tones were equal in level, not loudness). The final condition (C4), compares the pulse train with the complex tone used as the test stimulus in the first condition. The average level submitted by the six subjects was -11 dB, which again indicates nonsimultaneous SLS. In short, the experimental data support the idea that there is a temporal persistence in each auditory channel, prior to spectral integration.

Model	C1	C2	C3	C4
GM02	1.4	-4.8	7.8	-7
GM02 + Temporal window	1.4	1.1	1.1	0
GM02 + SL decay	1.4	-0.7	1.5	-0.4
CF02	4.5	-3.8	8	-4.4

Table 3.5: Error (dB) between the predictions of different configurations of the GM02, the CF02 and the level differences measured in four conditions of Zwicker's (1969) experiment.

The deviations between the predictions of the different dynamic models and the mean data in each of the four experimental conditions are given in Table 3.5. The three configurations of the GM02 give the same predictions for the first condition because the temporal smoothing of SL does not affect the predicted level difference between stimuli with stationary spectra, i.e. the concurrent tones of different frequency. The GM02 underestimates the loudness of the complex tone by about 1.4 dB compared to the experimental data. It should be noted that the discrepancies corresponding to the original GM02 are not in line with those reported by Rennies et al. (2010), in that those shown in Table 3.5 indicate a better correspondence with the data. Part of the reason is due to the smaller ramps (2.5 ms) used in that study, which introduces more SLS in the calculated loudness. For conditions 2–4, it can be seen that the two adaptations of the GM02 are much closer to the experimental data. The temporal window extension achieves slightly better performance than the simple SL decay approximation, the time constant of which was tuned to condition three. The two approaches are, however, based on different mechanisms.

The CF02 did not predict as much SLS in condition one as measured experimentally. The result of the third condition suggests either a lack of temporal integration of the test stimulus (the pulse train), insufficient SL decay, or a combination of both. The error for condition three is similar to that of the GM02: both predict lower SLS than measured empirically. The CF02 performed better than the GM02 in condition four, presumably due to the inclusion of a post-masking stage prior to summing the loudness over critical bands. Further tests show that CF02 performed slightly better with a ramp duration of 2.5 ms, as used by Rennies et al. (2010), since this broadens the spectrum.

The temporal window smooths the excitation in each auditory channel first, which is subsequently compressed according to Equations 3.51 and 3.52. The pulse train comprising tones of different frequency is effectively a broadband signal with similar overall energy to the single frequency tone, but becomes louder when summed *after* intensity compression. This effect would not arise had the temporal window been placed after the SL transformation.[15] Note, however, that the chosen order of placement contradicts that of literature (Oxenham and Plack 2000; Plack and Oxenham 2002) whereby the temporal window was developed to predict forward masking data. This discrepancy suggests a separate process at work.

The peak follower approach appends tails to the SL-time functions after stimulus offset such that the bandwidth of the SL pattern is greater at the onset of successive pulses. In contrast to the temporal window model used here, this extension therefore operates directly on the compressed excitation at the output of each auditory filter. Note that the 1.5 dB discrepancy in condition three is primarily associated with the miscalculation of spectral information only, i.e. the same error measured in condition one. This could be alleviated for condition three by increasing the decay time slightly, but this would not be valid because the error is associated with either a miscalculation of simultaneous SLS or differences caused by the assumed frequency response used for reproduction. The SL decay approximation is more in accord with Zwicker's (1984) nonlinear lowpass filter, but is greatly simplified in that here only one time constant is used.

Data of Heeren et al. (2011)

To verify the improvements offered by the extensions, a second simulation was run to predict the more recent data of Heeren et al. (2011). They conducted an extended version of Zwicker's (1969) experiment by using sequences of tone pulses of different durations and with varying interpulse intervals (Zwicker only formed pulse trains using 20 ms segments with no pause, as shown in Figure 3.23). In the first two experiments of Heeren et al. (2011), 12 subjects matched the loudness of a reference sequence of tone pulses with a frequency of 2.5 kHz to that of a test sequence comprising tone pulses with frequencies of 1.37, 1.85, 2.5, 3.4 and 4.8 kHz. Unlike Zwicker's paradigm, the individual tones of different frequencies were equally loud rather than equal in level. The five tone pulses of the test sequence were presented simultaneously in one condition (similar

[15] Further tests were conducted to verify this.

to Figure 3.23 C1), and nonsimultaneously in another (similar to Figure 3.23 C3). Pulse durations of 10, 20, 50 and 100 ms were used, and the inter-pulse interval varied between 0 and 200 ms. Raised-cosine ramps of 2.5 ms were applied to the pulses.

Heeren et al. (2011) measured SLS as the level difference between the 2.5 kHz pulse in the test and reference stimuli; a negative difference indicates that the test stimulus was louder than the reference stimulus. The main idea is that for large inter-pulse intervals, the SL following the offset of a tone pulse will have completely decayed, thereby reducing SLS entirely, i.e. a level difference of 0 dB. A third experiment investigated the effect of overall stimulus duration, which was increased from 1 s to 2 s. In the final experiment, ten listeners matched the loudness of the sequence of 10 ms concurrent tone pulses to the sequence involving a different frequency in each 10 ms pulse. In other words, the reference stimulus involving a single-frequency tone was replaced by the sequence of complex tone pulses. The results of this study are summarised as follows:

1. SLS is larger for short than for long pulses.

2. SLS of pulse trains therefore depends on pulse duration and inter-pulse interval.

3. Zwicker's hypothesis that a persistence of SL leads to SLS for nonsimultaneous frequency components was supported by the data. There is, however, a residual effect at long inter-pulse intervals that is not accounted for by current models incorporating this temporal persistence.

4. Current established multiband loudness models cannot fully account for these results unless the critical-band compression is duration-dependent.

The complexity of all of these issues is beyond the scope of this work, so instead the focus is on the benefit of including some form of temporal masking stage prior to summation across frequency bands (point three above). To investigate the model predictions of nonsimultaneous SLS in isolation of other effects, the results from the first two experiments of Heeren et al. (2011) were used to derive loudness matches between the simultaneous and nonsimultaneous tone pulse sequences. This essentially simulates the final experiment of Heeren et al. (2011), but using all four pulse durations, with the effect of repetition rate on simultaneous SLS removed. This was done because the loudness models discussed in this chapter do not model duration-dependent SLS.[16] The data for the 10 ms pulse duration were obtained from experiment four of Heeren et al. (2011), as the level differences were measured experimentally rather than indirectly using the data from the first two experiments. Simulations were performed using 12 log-spaced inter-pulse intervals as well as with no pause (0 ms interval). The frequencies of the pulses in the nonsimultaneous tone pulse sequence followed the temporal order: 3400, 1850, 2500, 1370 and 4800 Hz. Individual tones were equalised in loudness according to each of the model predictions given the long-term spectra.

Figure 3.24 shows the mean of the experimental data, along with the predictions of the three GM02 variants and of the CF02. The ordinate shows the level difference required for equal loudness between the single-tone pulse sequence and the complex-tone pulse sequence. It can be seen that the subjective matches are all positive, meaning that the level of the nonsimultaneous frequency components L_{nonsim} had to be higher than the level of the simultaneous frequency components L_{sim} in order be equally loud. This level difference increases with inter-pulse interval, with one exception for a pulse duration of 10 ms, because there is greater temporal separation between the tone pulses of different frequency, resulting in less spectral integration. At fast repetition rates (low inter-pulse intervals), the loudness of the nonsimultaneously presented tones are notably louder at

[16] Rennies. J. et al. (2009) presented extended versions of the CF02 to capture this phenomenon, but this extension is not reviewed here.

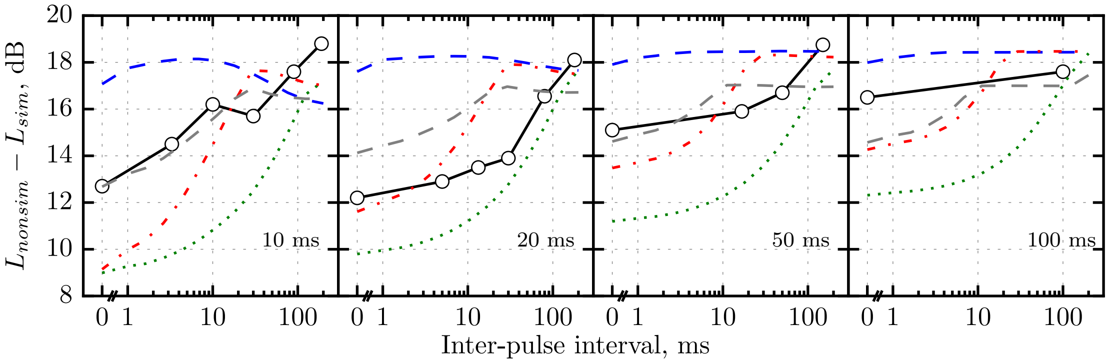


Figure 3.24: Level difference, as a function of inter-pulse interval, between a pulse train with *nonsim*ultaneous frequency components and another with *sim*ultaneous frequency components. The black lines with white markers show the mean data derived from Heeren et al. (2011). The blue dashed, green dotted and red dashed-dotted lines show the predictions of the GM02 in original form, with the temporal window extension, and with SL decay extension, respectively. The grey dashed lines show the predictions of the CF02.

pulse durations of 10 ms and 20 ms compared to durations of 50 ms and 100 ms. This is explained by the greater overlap in SL among successive tones of shorter length.

First, consider the predictions of the CF02 (grey dashed lines). This model predicts the observed monotonic increase in level difference with increasing inter-pulse interval particularly well for the 10 ms pulse duration. As the pulse duration increases, the model estimates less nonsimultaneous SLS (corresponding to a greater level difference), which is expected based on the reduced overlap in SL at the introduction of each new frequency component. Comparing the experimental data at pulse durations of 10 ms and 20 ms, this expectation is not observed. In addition, the predictions reach a plateau at inter-pulse intervals above 30 ms, whereas the data continue to rise. This suggests the need for a more prolonged persistence of SL following stimulus offset, as highlighted by Heeren et al. (2011). In general, the CF02 captures the reduction in dynamic range of the level difference contour with increasing pulse duration.

The blue dashed lines in Figure 3.24 show the predictions of the unmodified GM02. Unexpectedly, the predicted level differences are nonmonontic at pulse durations of 10 ms and 20 ms. The initial increase in level difference is partly attributed to the asynchronous tone pulses of different frequencies falling within the same DFT analysis window, leading to a broader excitation across frequency. There is also a spectral broadening introduced by the multi-resolution DFT when processing multiple 10 ms segments quasi-simultaneously. As the gap between successive pulses increases, this effect reduces and so the predicted loudness of the nonsimultaneous frequency components also decreases, leading to a greater level difference. Due to the six differently-sized analysis windows, the temporal order of the tone frequencies also influences the deviation from the expected flat line. At larger inter-pulse intervals, the decay of the STL integrator plays less of a role in determining the overall loudness, and the level difference is based on the loudest pulse. As a consequence of the multi-resolution DFT, the spectral bandwidth of individual tone pulses is frequency-dependent, meaning that the different frequency components are not equally loud at small pulse durations where spectral broadening becomes an issue.

The green dotted lines correspond to the predictions of the GM02 with the temporal window. This adaptation captures the trend of the experimental data, but generally leads to an overestimation of nonsimultaneous SLS, since the level difference is too low with respect to the subjective data. This effect is observed at all pulse durations, and is most notable at fast repetition rates.

The fact that contours do not taper off to a constant at large inter-pulse intervals, as done by the CF02, is explained by the use of two exponential functions describing the initial and later characteristic of forward-masking curves (refer to Equation 3.88).

The predictions of the GM02 with the SL decay extension are shown by the red dotted lines. They are in line with Zwicker's notion: loudness summation across critical bands decreases with inter-pulse interval. At 0 ms, this adaptation works best for a pulse duration of 20 ms, which is not surprising given that the time constant was tuned to condition three of Zwicker's (1969) experiment which made use of 20 ms pulses with no inter-pulse gap. The time constant is, however, too slow for the data obtained with 10 ms pulses, suggesting the need for duration-dependent decay factors as proposed by Zwicker (1984) and subsequently used in the CF02. An alternative modelling strategy is to incorporate an asymmetrical lowpass filter, such as the one used by the GM02 when calculating the STL and LTL, in each auditory channel prior to loudness summation. That way, short-duration tone pulses would not reach a maximum before decaying, resulting in a lower effective overlap in SL across frequency for successive tones, and thus less STL than in the method presented here.

To conclude, the CF02 and the GM02 with SL decay give the best absolute fit to the data, but are inconsistent across different pulse durations. Additionally, the persistence of SL after stimulus onset is not sufficiently long in either model to capture the continued trend in reduced SLS at inter-pulse intervals greater than 10–30 ms. In this respect, the more prolonged decay of the temporal window as used for simulating forward masking appears to be more beneficial. In any case, the GM02 in original form is not able to reproduce the effects of frequency modulation on SLS.

3.4 Summary

A historical and technical review of the various approaches to loudness calculation has been given. Single-band and multiband approaches to loudness modelling have been introduced and realisations of those applicable to both stationary and nonstationary sounds have been discussed. In psychophysics, the single-band model takes the form of a power law describing the relationship between the physical intensity of a sound and the sensory magnitude associated with it. In engineering and broadcasting, the single-band approach encapsulates a device to perform intensity-time averaging such that the loudness of real fluctuating sounds can be approximated. The frequency sensitivity of the human ear may also be taken into account by referring to the equal-loudness contours. The multiband approach involves decomposing the input stimulus into multiple frequency bands in a way that is consistent with empirical measures of the critical-band filters of the auditory system. This provides a satisfactory explanation of spectral loudness summation (SLS), which describes the effect of spectral bandwidth on loudness. In later years, the parameters of these sophisticated loudness estimators have been extended by incorporating models of temporal integration designed to predict loudness as a function of duration and amplitude/frequency modulation rate. When taken as a whole, modern multiband loudness models can be used to predict a broad range of empirical data, including energetic masking, and give insight into the underlying mechanisms responsible for the systematic effects observed in studies investigating loudness perception.

Glasberg and Moore's loudness model is a popular multiband procedure that has been shown to predict a wide range of perceptual phenomena from experiments involving both steady-state and time-varying sounds (Moore et al. 1997; Glasberg and Moore 2002). Although the model has been modified and extended over recent years, its architecture is fundamentally based on Zwicker's excitation pattern (EP) model proposed many years earlier (Zwicker 1958, 1960; Zwicker and

Scharf 1965; Zwicker 1977). The different procedures for calculating the loudness of stationary sounds were standardised, with Glasberg and Moore's forming the ANSI S3.4 (2007) standard, and Zwicker's the DIN 45631 (1967) and ISO 532 (1975) standards. In addition, a more recent and physiologically realistic model of the peripheral hearing system capable of calculating the loudness of stationary and dynamic sounds was also introduced (Chen et al. 2011; Chen and Hu 2012). A comparison of steady-state predictions given by these three loudness models (ANSI S3.4, DIN 45631, and the model of Chen et al.) has been carried out to gain further insight into the effects of different modelling strategies. The three models were implemented within the context of a single framework to establish a controlled testing procedure. A comparison of predictions for pure tones, multitone complexes and pink noise revealed that the three models show similar trends, but that systematic differences exist, the magnitudes of which are stimulus and level dependent. For applications outside of psychophysics, where accurate loudness estimation is required, such contrariness hinders model selection. However, comparisons of the models offers reasons for the differences in performance and thus highlights potential solutions. The results of this comparative study are summarised as follows:

1. The ANSI S3.4 predicts the loudness growth of a 1 kHz pure tone most accurately, followed by the DIN 45631.
2. The ANSI S3.4 gives an excellent fit to recent estimates of the absolute threshold of pure tones at different frequencies (ISO 389-7 2005).
3. Both the ANSI S3.4 and the model of Chen et al. provide a good account of the equal-loudness contours (ISO 226 2003), with the latter model giving a slightly better overall fit. The DIN 45631 shows large discrepancies when compared to the revised contours.
4. ANSI S3.4 predictions of the equal-loudness contours can be improved by incorporating greater critical-band compression above 1 kHz. However, further compression may lead to even worse predictions of broadband sounds, especially at higher frequencies.
5. All of the loudness models predict too much spectral loudness summation, meaning that the loudness of broadband sounds, especially in the 1–5 kHz frequency range, are generally overestimated. The DIN 45631 appears to show the best agreement with empirical data, which can be explained by the use of broader critical-band filters.
6. In general, the model of Chen et al. gives loudness predictions that vary the most with intensity level. This is due to this model employing a strongly level-dependent active filter bank. There is little data on loudness at high sound levels, although Viemeister and Bacon (1988) provide some evidence for the increase in the rate of growth of loudness, a characteristic also incorporated in the model of Moore et al. (1997).

An up to date implementation-level guide to Glasberg and Moore's time-varying model was presented to establish a deeper understanding of state-of-the-art loudness modelling techniques and also highlight the level of complexity involved in more elaborate psychoacoustic procedures. Unlike the latest incarnations of Zwicker's (1977) model for dynamic sounds, such as the one proposed by Chalupper and Fastl (2002), the model does not include a post-masking stage prior to SLS. This means that sounds with time-varying spectra are processed at a finer time-resolution than by the ear, leading to an underestimation of SLS for frequency components presented at different times. Different modelling strategies were compared using experimental data from studies involving simultaneous and nonsimultaneous tone pulses of different frequencies presented at different repetition rates. The results indicate that a simple decay of specific loudness (SL) captures the decrease in SLS as repetition rate is reduced, but discrepancies still exist at large inter-pulse intervals and across different pulse durations. An alternative extension was also tested which smooths the excitation in each critical band over time, prior to compression or SL transformation, using a temporal window derived from studies on forward masking. This too introduced SLS for frequency components presented asynchronously, but deviated from the experimental data by 2–4 dB. This approach did, however, confirm the need for longer time constants in order to capture the continued SLS at very slow repetition rates, which the model of Chalupper and Fastl (2002) was also unable to predict. It is not clear whether the duration-dependent model of SL decay is the best approach, but modifications are certainly needed to fully capture the observed trends at low repetition rates. Further tests are needed to understand the psychophysical mechanisms at work and relate them to current theories of temporal masking. This final study also identified problems introduced by the multi-resolution DFT in the model of Glasberg and Moore (2002) when predicting the loudness of short-tone pulses.

The next chapter investigates the computational cost of using the loudness models of Glasberg and Moore (2002) and Chen and Hu (2012) to extract auditory features from music and speech on a modern day personal computer. The first model was selected because the core algorithm is founded on the ANSI S3.4 (2007) standard which, as will be shown, can be extended to better predict the loudness of binaurally presented sounds (Moore and Glasberg 2007). The current state-of-the-art in loudness metering is therefore enhanced by developing a real-time binaural loudness meter capable of displaying both time-varying loudness in sones and auditory patterns as calculated using a protocol derived from psychophysical measures of energetic masking. As such, this particular loudness model is also suited to applications where predictions of auditory masking are needed or related measures of inter-source auditory interactions would be beneficial, such as in music production. The model of Chen and Hu (2012) was also selected because, though based on similar signal processing, it takes a physiologically realistic approach to loudness calculation. It is therefore interesting to identify if the techniques used to extend and optimise the dynamic model of Glasberg and Moore can also be used to devise an effective real-time auditory processor driven by a more recent model with different predictive behaviour.

Chapter 4

Efficient Loudness Estimation

In the previous chapter a number of complex physiologically motivated procedures for predicting the loudness of time-varying sounds were presented. In contrast to the simplistic single-band models, multiband algorithms require more intricate signal processing in order to simulate the functionality of the inner ear. This computationally intensive processing is unattractive for applications that require fast estimates of loudness, e.g. real-time loudness metering and automatic loudness control. It is therefore beneficial to establish computationally efficient multiband loudness models to allow for accurate and fast measures of loudness. This chapter focuses on optimising the loudness model developed by Glasberg and Moore (2002) by exploring its core processing parameters as well as integrating alternative low-complexity signal processing. A systematic approach was taken by profiling the model and assessing the performance of the proposed optimisations that target the bottleneck. A corpus of 30 sounds comprising speech and music were used to evaluate the error introduced in the auditory features estimated by the model.

The decision to concentrate on this particular model was based on the core algorithm being internationally standardised (ANSI S3.4 2007) as a procedure for predicting the loudness of a broad range of stationary stimuli (Moore et al. 1997), its potential to process dichotic sounds (Moore and Glasberg 2007), and its ability to predict audibility thresholds of time-varying sounds based on partial loudness (Glasberg and Moore 2005). In addition, the model outputs two loudness-time functions which facilitates more flexible modelling of global-loudness perception compared to other procedures (Rennies et al. 2010; Terrell 2013, Ch. 7). Although the proposed optimisations are primarily demonstrated using this algorithm, the second half of this chapter demonstrates how the techniques can be incorporated into other perceptual models based on the short-term power spectrum. In particular, the proposed techniques were aggregated to realise a real-time binaural loudness meter using the more recent excitation pattern based loudness model of Chen et al. (2011). As far as the author is aware, this is the first time this second algorithm has been optimised for real-time binaural processing.

4.1 Model overview

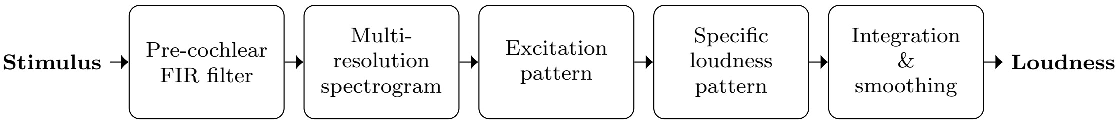


Figure 4.1: Block diagram of Glasberg and Moore's (2002) dynamic loudness model.

A block diagram of the loudness model is given in Figure 4.1. The loudness model transforms the time-domain input signal (in Pascals) into three loudness-time series: the instantaneous loudness (IL), the short-term loudness (STL), and the long-term loudness (LTL). These measures are in sone units and the interval between successive measurements, i.e. the hop size or time step, is 1 ms.

The first processing module is a 4096th order finite impulse response (FIR) filter, simulating the transfer of sound through the outer and middle ear. Subsequent processing is carried out by a multi-resolution spectrogram in order to obtain a frequency-domain representation of the stimulus at each time step. The short-term power spectrum is transformed into an excitation pattern (EP), which approximates the excitation along the length of the basilar membrane. The EP is defined as the output of a set of rounded exponential (roex) filters (Moore and Glasberg 1983). To arrive at the specific loudness (SL) pattern, which represents the loudness per critical band, the EP is passed through a compressive nonlinearity. Loudness is determined on a linear scale with units sones, and the perceptual frequency scale has units Cams. The area under the SL pattern gives the IL, which is assumed to be unavailable for conscious perception (Glasberg and Moore 2002). Finally, the IL is smoothed by two cascaded temporal integrators; the output of the first produces the STL, and the output of the second gives the LTL. The global loudness of the input stimulus can be estimated by applying a summary statistic, such as the maximum or mean, to the STL or LTL. The choice of time series and summary statistic is dependent on the type of signal (Glasberg and Moore 2002).

4.1.1 Previous work

A number of researchers have optimised specific stages of this particular loudness model in order to speed up performance. Most previous works have concentrated on reducing the number of spectral components and auditory filters involved in the calculation of EPs.

Frequency pruning is a technique applied to the input spectrum in order to simplify the excitation transformation. The general idea is to formulate a compact power spectrum by grouping the powers of multiple components in a perceptually motivated manner. Since the equivalent rectangular bandwidth (ERB) of the auditory filter increases with centre frequency (see Equation 2.26), the number of components falling within a given filter's passband also increases. Thus, the spectral fine structure becomes increasingly smoothed out with frequency, and therefore perceived differences between dense spectral ripples are harder to distinguish. Furthermore, the loudness of noise with a bandwidth less than that of the critical band is equivalent to the loudness of a single component with the same overall intensity (Fastl and Zwicker 2007). Thus, as long as the total power in each critical band is maintained, fewer components can be used to represent the spectrum whilst preserving total loudness.

The principle of frequency pruning based on internal representations was used by Stone et al. (1997) to reduce the output of a 1024-point fast Fourier transform (FFT) to 120 bins that cover the entire audible spectrum. This served to reduce computation time when computing auditory filter shapes and EPs, allowing for real-time loudness metering using a more primitive version of the 2002 model. A Matlab implementation of Glasberg and Moore's loudness model developed for the PsySound3 project (Cabrera et al. 2007) also employs a compressed spectrum prior to computing EPs. Components are grouped together such that the maximum distance between adjacent samples is less than 0.1 Cams in order to preserve the spectral fine structure along the auditory scale. Krishnamoorthi (Krishnamoorthi et al. 2008, 2009; Krishnamoorthi 2011) introduced an adaptive algorithm to prune the input spectrum for the same purpose. This procedure achieves a sparse set of components by summing the power of all bins in each critical band. In order to preserve the shape of the auditory patterns, the position of a 'representative' component within each band is placed at the frequency of the bin with maximum power. Auditory filters are then centred on these newly positioned components to best capture peaks in the EP. A second pruning procedure was presented which involves lowpass filtering the power spectrum in order to identify tonal and noise bands. The general concept is that strong tonal components will mask neighbouring components and therefore masked components can be bypassed at the auditory filtering stage. Noise bands are decomposed into smaller sub-bands whose total intensity is preserved using a single component, following a similar approach as Stone et al. (1997) and Cabrera et al. (2007). Rounded exponential filters are positioned adaptively according to local extrema in the smoothed power spectrum. The sampled EP is then interpolated to estimate the reference EP.

In Section 3.3.3 it was shown how Glasberg and Moore compute a multi-resolution spectrogram via six FFTs, approximating the time-frequency resolution of the human ear. Cassidy and Smith III (2007) proposed the hopping Goertzel discrete Fourier transform (HGDFT) to address redundancy associated with the parallel FFT approach. In this approach, Goertzel resonators are tuned to harmonics of the fundamental frequency of each DFT, meaning that the spectrum is sampled nonuniformly such that the sampling interval increases with frequency. By avoiding spectral interpolation, the total number of frequency points is significantly reduced. Although the HGDFT spectrogram proved to be significantly faster than the parallel FFT implementation, the authors did not address the performance of the model as a whole and thus the influence of nonuniform spectral sampling on loudness predictions remains unclear.

Finally, Burdiel et al. (2012), within the context of Glasberg and Moore's (2002) dynamic model, studied the effects of various hop sizes, filter spacings and sampling frequencies on loudness predictions, as well as a replacing the FIR filter used for pre-cochlear filtering by the spectral-weighting method. Loudness errors and comparisons of execution speed between the original model and its parameterised counterpart were discussed, concluding that such parameters should be source dependent, pointing towards an adaptive algorithm similar to those proposed by Krishnamoorthi (2011). For example, using a large filter spacing is more likely to introduce larger prediction errors for tonal signals than for broadband sounds. Additionally, a very small hop size may be too conservative for stationary sounds compared to those of a more dynamic nature. In this same vein, Krishnamoorthi et al. (2008) presented a time-varying algorithm in which comparisons between adjacent spectral frames were also used to predict auditory filter responses rather than computing them on every iteration, although key parameters were not specified.

In this study, a detailed assessment of the loudness model is carried out by implementing and systematically evaluating many of the above optimisations using broad range of time-varying sounds. New approaches to simulating pre-cochlear filtering are proposed and assessed against those used in previous work.

4.1.2 Profiling the model

In order to investigate the real-world performance of an algorithm, it is useful to profile its implementation under typical application. This potentially flags up computational bottlenecks which can be optimised using knowledge derived from a Big-O description of algorithmic complexity.

A modular block-based processing implementation of the loudness model was developed in C++ for the following reasons:

- The user can access the output of each processing module: useful for auditory feature extraction.

- The modular approach is suitable for real-time processing where audio is buffered into blocks of samples.
- Interchanging modules is simple which helps to minimise the trade-off between research and development time.

The implementation consists of five modules in accordance with the structure shown in Figure 4.1. A simple program was developed to extract the loudness from a 10 second music excerpt which was sampled at 32 kHz and had a peak level of 89.9 dB SPL.[1] An input buffer size of 32 samples was set in accordance with the 1 ms hop size specified by Glasberg and Moore (2002). The filter spacing was set to 0.1 Cams rather than 0.25 Cams (as used by Glasberg and Moore) for reasons described in Section 4.2.1. The simulation was based on causal block-based processing, with additional frames added to the end of the stimulus to account for the 2048 sample group delay introduced by the FIR filter. The average user CPU time of the program was 9.9 minutes, indicating extremely slow non-real-time performance.[2] In order to determine which stage of the model consumed the most processing time, the sample-based profiler OProfile (2015) was used to identify critical parts of the program. OProfile works by interrupting the CPU at regular intervals and assigning one sample to the process running at that time. Performance bottlenecks can be identified by inspecting the percentage of samples counted in each process relative to the total number of samples in the program or entire system. The program was run for 10 iterations to increase the sample size. The most important execution times were extracted from the system wide profile report and further examined in the program report as shown in Table 4.1.[3]

Application/symbol	Samples (%)
Program	99.1
`libm-2.19.so`: exp	91.7
Excitation transformation	7.2
FIR filter	0.3
`fftw3.so`	0.2

Table 4.1: Breakdown of the program profile report.

The first row shows the percentage of samples associated with the main application 'Program' with respect to the entire system: the program consumed 99% of the entire system run time. This is decomposed into the primary processes (including function calls from external libraries) consumed by the application identified by OProfile using symbol names. For example, it was found that the exponential function from the `libm` math library accounted for 92% of the time spent in the program. This can be attributed to the calculation of the roex filter shapes (see Equation 4.1). Additional computations within the excitation transformation stage consumed about 7% of the CPU time, with subsequent processing modules taking less than 1%. `fftw3.so` refers to the FFT library developed by Frigo and Johnson (2005). In conclusion, the procedure for calculating EPs is the cause for the slow execution time of this loudness model—auditory filtering is the performance bottleneck. These findings are in agreement with the hotspots identified by previous researchers (Stone et al. 1997; Cabrera et al. 2007; Krishnamoorthi et al. 2009).

[1] In this study, the sampling frequency was set to 32 kHz as used by Glasberg and Moore (2002), although the implementation developed here supports any reasonable audio rate.

[2] Timings were measured on an idle Intel Core 2 Quad CPU Q8200 operating at 2.33 GHz with 4 GB of RAM running Ubuntu 14.04. Profiling was conducted on a range of different sound sources but only small variations in the relative timings were observed.

[3] Relative times $\leq 0.1\%$ were excluded.

4.2 Optimising the excitation transformation

From profiling the loudness model, the transformation from power spectrum to EP was identified as the most computationally demanding process. The EP is defined as the output of each auditory filter in response to the input intensity spectrum. The roex filter is central to this process and so it is reintroduced:

$$W(g) = (1 + pg)e^{-pg}, \tag{4.1}$$

where p defines the filter slope and g is the normalised deviation from the centre frequency f_c of the filter, i.e.

$$g(f) = \frac{|f - f_c|}{f_c}, \tag{4.2}$$

where the input frequency f and f_c are in Hz.

In what follows, let K denote the number of input components comprising the spectrum and A denote the number of auditory filters used to calculate the EP of a single analysis frame. The complete excitation transformation stage involves three main steps:

1. Calculating the level per ERB using a symmetrical roex filter centred on each input component. Parameter p is a linear function of the filter centre frequency (there is no level dependency at this stage). With K input components, this procedure requires K filters. Therefore, both Equation 4.1 and the calculation of the filter outputs in response to each input component are associated with a time complexity of $O(K^2)$.

2. Using the power per ERB obtained from stage one, estimate the (now level dependent) slope of each auditory filter in response to each input component. With A filters used for the actual EP, the time complexity of this procedure is $O(AK)$.

3. Calculating the output power of each auditory filter in response to each input component. The roex is again re-evaluated due to the level-dependent slope and each filter is applied to the spectrum according to the centre frequency. The centre frequencies are mapped uniformly on the Cam scale at a resolution of 0.1 Cams. Both filter shape and EP computations require $O(AK)$ operations.

For a detailed mathematical description of the above steps, refer to Section 3.3.3. Stages 1–3 are evaluated on a frame-by-frame basis and so the running time of these stages are also linearly dependent on the number of frames and therefore the hop size. Although processing time can be reduced by increasing this parameter, Burdiel et al. (2012) found that reducing the number of auditory filters was preferable over increasing the hop size in terms of minimising error. Hop size is therefore considered separately in Section 4.4. Of the three procedures formulating the excitation transformation, the level per ERB computation is the most complex, since in most applications $K >> A$. The profiling results flagged the exponential function as the performance bottleneck, which is explained by the K^2 call time. Therefore, the most critical part of the algorithm is stage one.

In this section, a number of methods are introduced that aim to reduce the number of operations required to compute auditory patterns. The primary goal is to simplify the level per ERB computation and then to investigate various techniques for reducing the number of input components and auditory filters to allow for fast loudness predictions within tolerable error bounds.

4.2.1 Evaluation procedure

The proposed modifications were evaluated by extracting perceptual features from 30 monophonic tracks taken from the sound quality assessment material (SQAM) database (EBU Tech 3253 2008) and assessing the error introduced. The SQAM dataset was selected because it contains a variety of sound material recommended by the European Broadcast Union (EBU) for subjective test purposes and it is freely available as a research and development tool allowing comparable experiments to be carried out using identical stimuli. Related to this last point, both Krishnamoorthi (Krishnamoorthi et al. 2008, 2009) and Burdiel et al. (2012) used selected tracks from the SQAM corpus to assess the performance of their proposed efficient implementations of Glasberg and Moore's loudness model. Furthermore, the present study is concerned with addressing loudness estimation algorithms applied to musical sounds, rather than broadband programme material or abstract stimuli commonly used in psychophysics.

Each of the 30 selected recordings belonged to one of the following stimulus categories (following the classification used in EBU Tech 3253 (2008)): single instruments, vocal, speech, solo, orchestra and pop music. Because the original loudness model operates at a sampling rate of 32 kHz, the selected signals were re-sampled using the free software library `libsamplerate`.[4] Representative sound segments of the selected tracks were spliced manually with 0.1 s raised-cosine ramps applied to the onsets and offsets, and each segment was root mean square (RMS) calibrated to a random level between 70 and 80 dB SPL, giving a spread of common listening levels across the 30 sources. Specific details of each sound segment are listed in Table 4.2.

The reference model was configured as follows:

- Sampling frequency: 32 kHz.
- Outer- and middle-ear filter: 4096th order FIR filter (free field, frontal incidence).
- Spectrogram: Uniformly sampled multi-resolution DFT via six parallel FFTs with a hop size of 1 ms.[5]
- Auditory filters: 1.8 to 38.9 Cams in steps of 0.1 (ANSI S3.4 2007).
- SL transformation: as per ANSI S3.4 (2007).
- Time constants: Those proposed by Glasberg and Moore (2002).

This setup follows the specification used by Glasberg and Moore (2002) except that the filter density has been increased from four to ten filters per ERB. This way the loudness error is made relative to a high-resolution implementation following the procedure outlined in ANSI S3.4 (2007) and can be assessed as a function of filter spacing.

For each of the approximations, the primary interest was to investigate the error between the reference and estimated loudness-time series (either IL, STL or LTL; in sones) of a given stimulus, denoted Ψ and $\hat{\Psi}$ respectively. The error e at frame number m is

$$e_m = \Psi_m - \hat{\Psi}_m. \tag{4.3}$$

The mean absolute error (MAE) between the reference loudness and its approximation is given by

$$\mathrm{MAE}_\Psi = \frac{1}{M}\sum_{M=1}^{M} |e_m|, \tag{4.4}$$

[4] http://www.mega-nerd.com/SRC/

[5] All frequency bands were configured as shown in Table 3.3.

Track no.	Category	Contents	Duration (s)	L_{RMS} (dB)	L_{pk} (dB)
08	Single instruments	Violin	10.2	75.1	92.7
11	Single instruments	Double-bass	10.2	70.7	83.6
17	Single instruments	Bass-clarinet	8.3	77.8	89.7
24	Single instruments	Tuba	10.2	75.9	88.3
25	Single instruments	Harp	10.0	70.3	87.7
26	Single instruments	Claves	7.2	78.4	101.8
27	Single instruments	Castanets	6.7	78.1	106.5
30	Single instruments	Kettle-drums	7.8	78.4	99.9
32	Single instruments	Triangle	5.1	79.5	96.6
34	Single instruments	Tubular bells	8.7	71.1	89.0
35	Single instruments	Glockenspiel	7.6	73.4	92.3
36	Single instruments	Xylophone	8.0	75.8	93.1
37	Single instruments	Vibraphone	3.8	75.4	90.7
39	Single instruments	Grand piano	6.4	72.2	87.1
40	Single instruments	Harpsichord	5.4	78.0	93.9
42	Single instruments	Accordion	7.9	75.8	91.3
44	Vocal	Vocal (soprano)	11.1	74.7	88.4
47	Vocal	Vocal (bass)	11.6	75.0	92.2
49	Speech	Female speech	4.1	74.2	90.1
50	Speech	Male speech	4.7	72.6	88.7
53	Speech	Female speech	3.9	70.3	88.8
55	Solo	Trumpet	26.5	70.6	89.6
56	Solo	Organ	10.4	71.9	85.5
58	Solo	Guitar	11.9	77.7	97.7
59	Solo	Violin	18.7	75.1	93.1
60	Solo	Piano	17.1	74.9	95.9
66	Orchestra	Wind ensemble	13.1	77.8	94.8
68	Orchestra	Orchestra	23.4	74.3	94.2
69	Pop music	ABBA	11.9	74.0	93.9
70	Pop music	Eddie Rabbitt	14.9	78.2	96.2
Mean (and standard deviation)			10.2 (5.4)	74.9 (2.7)	92.4 (4.8)

Table 4.2: Description of the sound segments extracted from the SQAM database used to form the evaluation dataset.

where M is the number of (non-zero) frames and the subscript Ψ is used to emphasise loudness error. The root mean square error (RMSE), which assigns greater weight to extreme deviations, was also used to assess the performance of the approximations:

$$\mathrm{RMSE}_{\Psi} = \sqrt{\left(\frac{1}{M}\sum_{m=1}^{M} e_m^2\right)}. \tag{4.5}$$

The peak error (PE) in the loudness-time function was used as an indication of worst-case performance and is given by Equation 4.6.

$$\mathrm{PE}_{\Psi} = \max_{1 \leq m \leq M}(|e_m|). \tag{4.6}$$

In order to obtain a scale-free evaluation metric, all measures were normalised by the average of the reference loudness-time series $\overline{\Psi}$. For example, the normalised root mean square error (nRMSE) for loudness was computed as

$$\mathrm{nRMSE}_{\Psi} = \frac{\mathrm{RMSE}_{\Psi}}{\overline{\Psi}}, \tag{4.7}$$

which corresponds to the coefficient of variation of the RMSE (Reddy 2011). This metric is similar to the nRMSE used by Burdiel et al. (2012) to quantify the loudness-time series error, but here the RMSE is normalised by the average loudness rather than the peak loudness.

The error in the EPs was measured using the normalised mean absolute error (nMAE) between the reference EPs and estimated EPs,

$$\text{nMAE}_E = \frac{\sum\limits_{m=1}^{M}\sum\limits_{i=1}^{A}|E_{m,i} - \hat{E}_{m,i}|}{\overline{E}}, \tag{4.8}$$

where $E_{m,i}$ and $\hat{E}_{m,i}$ are the excitation and approximated excitation of frame m in critical band i, in linear power units. $\overline{E}$ is the total excitation over all frames calculated using the reference algorithm. This metric is similar to the linear quantity used to compute the average error energy metric proposed by Krishnamoorthi et al. (2008) for evaluating EP error, except that Equation 4.8 sums the absolute error over multiple frames and normalises this by the total excitation.

It is also insightful to explore the perceptual implications of model performance within the context of practical application for different parameter sets. For example, say the reference model was used to align the overall loudness impression of set of stimuli to a target loudness value—a common task in audio engineering and broadcasting.[6] An alternative implementation of the same model is likely to introduce some error when estimating the global loudness of the same stimuli. In many practical applications an alignment error 0.5 dB would be of no concern, but errors exceeding 10 dB would not be tolerated.

The difference in level required to equate the overall loudness as predicted by the reference model and the approximation was estimated for all stimuli according to two global-loudness descriptors: the average LTL, denoted Ψ_μ, and the peak STL, denoted Ψ_{pk}. The mean LTL and peak STL descriptors were selected based on the suggestions of Glasberg and Moore (2002) for quantifying the overall loudness of relatively long-duration sounds such as speech, and short-duration sounds such as noise bursts, respectively. Global-loudness descriptor error is represented as a relative gain in dB. For example, the gain required for equal loudness according to the peak STL is approximated using the sone ratio:

$$\Delta L(\text{dB}) = 10\log_2\left(\frac{\Psi_{pk}}{\hat{\Psi}_{pk}}\right), \tag{4.9}$$

where Ψ_{pk} and $\hat{\Psi}_{pk}$ is the peak STL of the reference and approximation, respectively. ΔL does not measure error in terms of the relative loudness levels of all segments in the dataset, but instead estimates the disparity in absolute loudness as a level difference. For example, according to Equation 4.9, a sone ratio of two means that the peak loudness measured using the approximation was 10 dB below that of the reference prediction—even if the loudness of *every* stimulus was estimated as being 10 dB below that of the corresponding reference loudness. Equation 4.9 is equivalent to transforming the global-loudness measures to loudness levels in phons (an approximation) and taking the difference between the two loudness levels. Thus, for convenience of measuring errors on a large dataset with many experimental variables, the assumption of a linear relationship between gain in decibels and log sones was made. The magnitude of the error associated with this approximation is discussed in Appendix B. All descriptor errors reported here refer to absolute values, i.e. the magnitude of the level change.

[6] Overall or global loudness refers to a single quantity denoting the perceptual magnitude of a sound.

4.2.2 Level per equivalent rectangular bandwidth

The level per ERB spectrum is a high resolution EP obtained by centring fixed symmetrical roex filters on every input component. This is a prerequisite for the final (primary) EP which requires auditory filters with level dependent shapes. The problem is that the first stage requires considerably more operations than the second; the number of operations is on the order of K^2 compared with AK, where $A < K$.

The power in each critical band can be approximated using a rectangular window rather than the roex function (Glasberg and Moore 1990). This approximation affects the filter slope used to construct the secondary filter shapes. The argument for this approximation is that the upper slopes of the EP are formed primarily by components falling within 1 ERB around the peaks in the EP, e.g. tonal spectra. However, given a white spectrum, the approximation is fair because the rectangular window has the same equivalent bandwidth as the fixed roex filter. A rectangular window should not be used in the second and primary stage of filtering, however, because then nonlinearities in the upper slopes of the EP would not be captured.

Given the short-term multi-resolution power spectrum $\hat{S}$, the intensity level in the ERB around component j is estimated using (Glasberg and Moore 1990)

$$L_{\mathrm{ERB},j} = 10\log_{10} \sum_{k\in D_j} \hat{S}_k, \quad D_j = \{k | z_{c_j} - 0.5 \leq z_{c_k} \leq z_{c_j} + 0.5, k = 1 \ldots K\}, \tag{4.10}$$

where z_{c_k} denotes the critical-band number, in Cams, of the k^{th} frequency component. D_j is the set of components falling within 1 ERB of component j. The summation can be performed using a sliding rectangular window which requires no more than $2K$ additions, with $K \leq \frac{N}{2} + 1$ and N the DFT length. In contrast to the initial cochlear filtering procedure of ANSI S3.4 (2007), this gives an efficient run-time complexity of $O(K)$ with no multiplications or costly exponentials needed.

Glasberg (personal communication, August 19, 2014) explained that, in the implementation of their dynamic loudness model of (Glasberg and Moore 2002), a 2001-sample lookup table (LUT) with nearest-neighbour interpolation was used to approximate the roex filter (Equation 4.1) in order to speed up the level per ERB calculation. As the level per ERB stage has been modified to make use of a rectangular window, it is interesting to measure the additional computational savings and error introduced by such an LUT when approximating the response of the secondary set of roex filters used to obtain the final EP. Therefore, a table of length 1024 samples was used to store pre-calculated values of the roex function, which was found to preserve the filter shapes over a dynamic range of ~75 dB with less than 1% error. The significance of using a rectangular window to estimate the level per ERB and employing an LUT for the final roex filtering is now addressed.

Evaluation

Method	Loudness error (%)			EP error (%)
	nMAE	nRMSE	nPE	nMAE
Rectangular window	0.04	0.07	0.37	0.40
Rectangular window + LUT	0.07	0.11	0.60	0.42

Table 4.3: Median of the normalised relative errors across all stimuli.

Table 4.3 shows the median of the relative errors in the IL across all stimuli for the sliding window alone and when combined with the LUT. The IL errors are all very low, with the PE being 0.37% for the rectangular window and 0.60% when combined with the LUT. The median nMAE in the EPs across all sound segments was less than 0.5% for both methods. A larger LUT and/or an improved interpolation algorithm is not likely to offer significant improvements as the rectangular window was responsible for the majority of the error introduced, especially in terms of EP error. In addition to the performance metrics displayed in Table 4.3, the median and maximum of the global-loudness descriptor errors over all stimuli were measured to estimate the level differences between the stimuli input to the reference and approximation (rectangular window + LUT) at equal (objective) loudness. The median of the global-loudness descriptor errors was 0.01 dB for both mean LTL and peak STL, with the highest discrepancy being 0.12 dB for the peak STL. Taken together, both the time series errors and descriptor errors indicate a minimal effect.

The running time and speedup factor of the reference and approximations as measured on the aforementioned desktop personal computer (PC) are shown in Table 4.4.[7] Linearising the complexity of the level per ERB procedure by means of a sliding rectangular window results in huge gains in speed and the bottleneck shifts to the calculation of EPs using roex filters where the LUT provides additional savings.

Method	CPU time (s)	Speedup factor
Reference	592.3	1
Rectangular window	113.6	5.2
Rectangular window + LUT	35.4	16.7

Table 4.4: Running time and speedup factors obtained when processing a 10 s music excerpt.

Image/Symbol	Samples (%)
Program	99.0
Excitation	86.8
FIR	4.9
`libm-2.19.so`	3.8
`fftw3.so`	3.5

Table 4.5: Breakdown of the second program profile report.

The new profile report after implementing the power per ERB as a running sum and employing an LUT to avoid computing exponentials is shown in Table 4.5. This second program profile report shows a greater contribution to overall running time from other processing modules as a consequence of optimising the areas highlighted in the initial report. Despite the large increase in speed, the loudness model did not execute in real time (see Table 4.4). The new profile report shows that the bottleneck still resides in the excitation transformation stage which accounts for approximately 87% of the running time.

4.2.3 Filter and component reduction

In the preceding section it was determined that the vast majority of processing time was spent calculating EPs. After alleviating the initial bottleneck, model performance improved considerably with little error introduced. After returning to the analysis of algorithmic complexity after administering these modifications, two parameters that determine the speed-accuracy trade-off become evident:

1. Compute the level per ERB using a sliding window: $O(K)$.

[7] Speedup factor is the ratio of the execution time for the reference configuration to that of the approximation.

2. Compute the filter slope: $O(AK)$.

3. Compute the filter shapes and their output in response to the input spectrum: $O(AK)$.

Here, K and A denote the number of spectral components and auditory filters, respectively. For the reference configuration, $K = 960$ and $A = 372$. Three procedures are proposed to reduce the execution time of the model based on these two variables:

1. Increase the spacing between auditory filters (reduce A).

2. Apply spectral compression to the multi-resolution spectra (reduce K).

3. Sample the spectrum nonuniformly using the HGDFT (reduce K).

For a single analysis frame, the total complexity reduction (CR) of the $O(AK)$ stages dominating the excitation transformation stage is defined as (Krishnamoorthi et al. 2009):

$$CR = 1 - \frac{\hat{A}\hat{K}}{AK}, \tag{4.11}$$

where $\hat{A}$ and $\hat{K}$ denote the number of auditory filters and frequency components used in the approximation, respectively.

1. Auditory filter spacing

Glasberg and Moore (1990) sampled the power spectrum using roex filters uniformly spaced on the Cam scale. The interval between adjacent filters, sometimes called the ERB step or Cam step, should be low enough to ensure that peaks in the EP are captured as well as preserving spectral loudness summation (SLS). For example, given a single input component, the maximum power must be represented as well as the slopes of the excitation as captured by filters located farther away from the peak excitation (see Figure 2.19 for EPs of a 1 kHz tone presented at low and high sound levels). The choice of filter spacing, denoted α, is a compromise between accuracy and complexity, and also application dependent (Burdiel et al. 2012). For example, the procedure for calculating the loudness of stationary sounds given in ANSI S3.4 (2007) employs a filter spacing of 0.1 Cams, whereas Stone et al. (1997) used a higher interval of 0.33 Cams in order to achieve real-time processing. Glasberg and Moore (2002) used a spacing of 0.25 Cams for the dynamic model, but no justification was given. Alternative approaches are based on adaptive filters which sample the spectrum nonuniformly in a stimulus-dependent manner (Krishnamoorthi et al. 2009).

A simple approach to ease the workload of the model but improve the EP estimation is to interpolate the sampled pattern at a higher resolution. From experimenting with various input signals and interpolation schemes, cubic spline interpolation applied to the log EP was found to improve the detection of filter centre frequencies responsible for peak excitations as well as capturing the general shape of the pattern. An example is shown in Figure 4.2 where three EPs derived from a single analysis frame of a violin recording have been calculated. The black solid and red dashed lines correspond to the patterns computed with roex filters spaced at 0.1 and 1 Cam intervals, respectively, with the latter interpolated using the aforementioned scheme. White circles show the outputs of filters sampled at increments of 1 Cam. From visual inspection alone, interpolation yields a better representation of peaks and captures the rounded shape of the high resolution pattern.

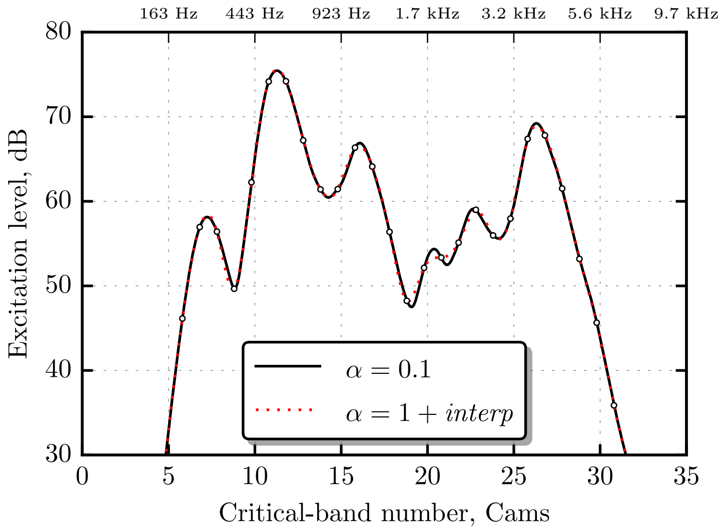


Figure 4.2: EPs of a segment of track 8 (violin) of the SQAM collection, calculated using roex filters spaced at 0.1 and 1 Cam intervals. The open circles correspond to the excitation levels produced by filters uniformly positioned at increments of 1 Cam.

2. Component reduction by compression

A flexible method, inspired by Stone et al. (1997) and Krishnamoorthi et al. (2008) and adapted from Cabrera et al. (2007), for reducing the number of spectral components was developed. A compressed power spectrum is obtained by summing bins into wider rectangular bands, the bandwidths of which increase with frequency. The idea is that fewer components can be used at higher frequencies because the fine spectral structure is smoothed out as a result of increasing auditory filter bandwidth (see Figure 2.20 and associated text). The compression criterion, denoted β, is chosen to preserve a predefined maximum spacing between adjacent components on the Cam scale. For example, $\beta = 1$ means the distance between any two components should be as large as possible but not exceed 1 Cam. Each composite bin is assigned a frequency equal to the average of those components included in the sum. The amount of data reduction is thus dependent on the compression criterion which affects the shape of the resulting auditory patterns. Cabrera et al. (2007) employed a spacing of 0.1 Cams, reducing the number of multi-resolution components from 960 to 346 (64% reduction). Stone et al. (1997) used a similar technique to obtain 120 bins from a 1024-point DFT in order to speed up the excitation transformation. From a perceptual standpoint, the criterion should not exceed the critical bandwidth in order to preserve the total energy per ERB and thus not sacrifice predictions of SLS.

Figure 4.3 shows the distribution of the real components of a 2048-point DFT on the Cam scale, with and without spectral compression. When the compression algorithm is bypassed, the components are equally spaced on the linear frequency scale which translates (approximately) to a logarithmic distribution on the Cam scale. For this configuration ($\Delta f = 15.625$ Hz), spectral compression takes place at frequencies above 1200 Hz for $\beta = 0.1$, and above 484 Hz for $\beta = 0.2$. Above these limits, components are averaged such that the frequencies of adjacent composite bins do not exceed their respective compression criterion. Larger criteria produce wider bands and hence greater data reduction. In this example, $\beta = 0.1$ covers the audible frequency range using 346 components whereas $\beta = 0.2$ requires just 199 components. Smoothing the spectrum in this manner also has the effect of reducing the depth of spectral ripples and shifting tonal components

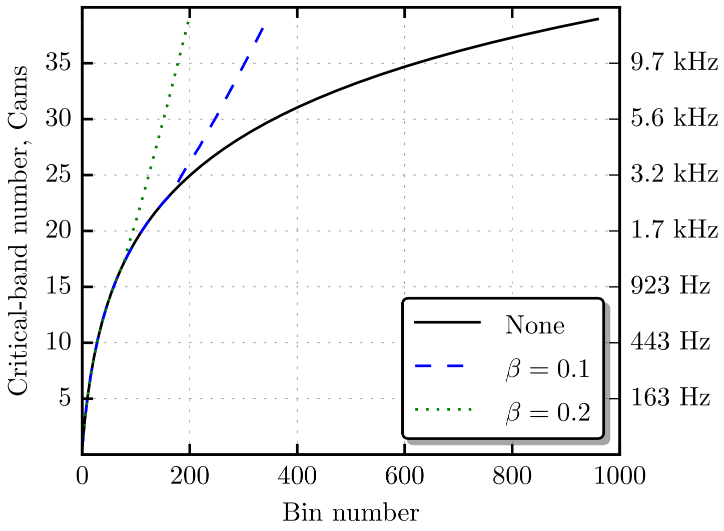


Figure 4.3: Cam frequency distribution of the real components of a 2048-point DFT and of the composite components obtained with spectral compression applied using criteria of 0.1 and 0.2 Cams.

which subsequently distorts the shape of the auditory patterns.

3. Component reduction by nonuniform spectral sampling

Glasberg and Moore (2002) used six parallel FFTs instead of level-dependent time-domain filters, e.g. the Gammachirp (Irino and Patterson 1997), allowing for more efficient processing, and conformance to their steady-state procedure (Moore et al. 1997). In order to approximate the time-frequency resolution of the human ear, the six FFTs each have different temporal resolutions meaning that the spectral resolution is inversely proportional to frequency across the bands. The multi-resolution spectrogram comprises six Hann-windowed segments of durations 64, 32, 16, 8, 4 and 2 ms used to compute the power of components in frequency bands 20 to 80 Hz, 80 to 500 Hz, 500 to 1250 Hz, 1250 to 2540 Hz, 2540 to 4050 Hz and 4050 to 15 000 Hz, respectively. The sampling interval is constant across all frequency bands: an N-point DFT is computed for *all* segments, requiring the windowed data to be padded with zeros. The original model operates at 32 kHz, so a transform of length $N = 2048$ is used for all bands ($\Delta f = 15.625$ Hz), which results in 960 non-redundant components.[8]

Cassidy and Smith III (2007) proposed the HGDFT as an efficient alternative to the parallel FFTs in order to deal with the redundancy associated with discarding spectral components from the six bands. The idea is to sample inversely proportional to the window length by tuning complex resonators to only those frequencies falling within each of the band limits. Although not emphasised by the authors, nonuniform spectral sampling yields significantly fewer frequency components compared to the original implementation, which is desirable for targeting the computational complexity of the excitation transformation. This means that the sampling interval doubles for each subsequent band, i.e. $\Delta f_b = \frac{f_s}{N_b}$, where N_b is the length of the window used for band number b. The k^{th} spectral component in a given frequency band of an N_b-point DFT can be computed using

[8] Here, redundancy refers to components from different bands of the same frequency. These components are discarded if they are outside their respective band limits.

the sliding discrete Fourier transform (SDFT) (Cassidy and Smith III 2007):

$$X_{b,k}[m] = \sum_{n=0}^{N-1} x[n + (m - (N_b - 1))]e^{-j2\pi kn/N_b}, \tag{4.12}$$

where m indicates the frame number. Following Jacobsen and Lyon (2003) and Jacobsen and Lyons (2004), the SDFT can be implemented as

$$X_{b,k}[n] = e^{j2\pi k/N_b}(X_{b,k}[n-1] + x[n] - x[n - N_b]). \tag{4.13}$$

This is a length N_b comb filter followed by a complex resonator. This recursive filter operates on a sample-by-sample basis, requiring one complex multiply and two real adds for each output sample. The transfer function of this filter is

$$H_{SDFT}(z) = \frac{e^{j2\pi k/N_b}(1 - z^{-N_b})}{1 - e^{j2\pi k/N_b}z^{-1}}. \tag{4.14}$$

To save on complex multiplications, the numerator and denominator are multiplied by the conjugate of the denominator, i.e. $1 - e^{-j2\pi k/N_b}z^{-1}$, which simplifies to

$$H_{SGFT}(z) = \frac{(e^{j2\pi k/N_b} - z^{-1})(1 - z^{-N_b})}{1 - 2cos(2\pi k/N_b)z^{-1} - z^{-2}}, \tag{4.15}$$

which is the transfer function of the sliding Goertzel discrete Fourier transform algorithm. The corresponding difference equation is

$$\begin{aligned} X_{b,k}[n] = e^{j2\pi k/N_b}&(x[n] - x[n - N_b]) - x[n-1] + x[n - (N_b + 1)] \\ &+ 2cos(2\pi k/N_b)y[n-1] - y[n-2], \end{aligned} \tag{4.16}$$

which can be realised using the state-variable equations (Cassidy and Smith III 2007)

$$v[n] = x[n] - x[n - N_b] + 2\cos(2\pi k/N_b)v[n-1] - v[n-2], \quad \text{and} \tag{4.17}$$

$$X_{b,k}[n] = e^{j2\pi k/N_b}v[n] - v[n-1]. \tag{4.18}$$

Notice that because Equation 4.18 is decoupled from Equation 4.17, the complex-by-real multiplication need only be computed every R samples (the hop size). The HGDFT is simply the sliding form when the feedforward part (Equation 4.18) is computed every R samples with $R > 1$. This allows one to efficiently compute the DFT for a single spectral coefficient at different time points:

$$X_{b,k}[m] = \sum_{n=0}^{N_b-1} x[n + (mR - (N_b - 1))]e^{-j2\pi kn/N_b}. \tag{4.19}$$

The complexity of the HGDFT is summarised as follows:

- Three real additions and one real multiplication per input sample (Equation 4.17).
- One real addition and one complex-by-real multiplication per hop size of R samples (Equation 4.18).
- Summing the above for K_b spectral components in band b and a hop size of R samples, gives $K_b(R+2)$ multiplications and $K_b(3R+1)$ additions for every R samples. The total number of operations per frame is then $K_b(4R+3)$.

The complexity of the HGDFT for a select frequency band and single analysis frame is then $O(K_b R)$ and $O(N_b \log_2 N_b)$ for the radix-2 FFT. When the hop size R is less than $\log_2 N_b$, the HGDFT may well be more efficient than the FFT, especially when one is interested in computing only a subset of the N_b output samples, i.e. when $K_b < \frac{N_b \log_2 N_b}{R}$, with $R \geq 1$. As the hop size increases, the HGDFT will most likely no longer be the optimal choice when a fixed number of components are desired.

To reduce the influence of spectral leakage on loudness estimates, windowing can be performed efficiently in the frequency domain by convolving adjacent spectral samples with the DFT of the Hann window (Jacobsen and Lyon 2003):[9]

$$X_{b,k}[m] = -0.25X_{b,k-1}[m] + 0.5X_{b,k}[m] - 0.25X_{b,k+1}[m]. \tag{4.20}$$

In order to satisfy Equation 4.20 within the context of the multi-resolution spectrogram, each spectral band requires an additional 2 components to complete the convolution at component frequencies corresponding to the extrema of the band: a total of $K_b + 2$ resonators are needed. Allowing for a gain of four, Equation 4.20 can be computed using one real-by-complex multiplication ($2X_{b,k}$). In this case, frequency-domain windowing using the Hann function requires K_b real-by-complex multiplications and $4K_b$ real additions per frame, and can be more efficient than windowing in the time domain when the number of Goertzel resonators is less than the window size.

In summary, the six spectral bands proposed by Glasberg and Moore can be realised using six comb filters each feeding a set of Goertzel resonators linearly spaced in frequency with respect to window length. Because the spacing doubles for each subsequent frequency band, the result is a nonuniform multi-resolution spectrogram. Notice that compared to the FFT approach, in which the majority of computed DFT bins are discarded, the only redundancy associated with using a bank of HGDFT resonators is the two additional bins (per band) needed for frequency domain windowing. Based on Table 4.1, opting to employ the nonuniform sampling scheme may indeed look like premature optimisation, since the excitation transformation is the primary stage impeding real-time execution. However, the nonuniform spectrogram gives 69 non-redundant components covering 20–15 000 Hz, compared to 960 as used in the original model. This is a significant reduction (93%) in the number of components involved in the computation of EPs. Note that when sampling nonuniformly the band edges can be modified in order to maximise proximity between bins in each subsequent band. For example, the upper limit of the fourth band could be changed from 2540 Hz to 2750 Hz and the upper limit of the fifth band could be increased from 4050 Hz to 4500 Hz. However, it was decided to use the original DFT specification because this lead to lower error in both the loudness-time series and EPs. The alternative specification could, however, show an improvement when estimating the loudness of pure tones near the band edges. It is also worth noting that because the HGDFT does not require N_b to be a power of 2, the number of components is independent of the sampling frequency. However, although the per band spectral resolution is identical in both sampling methods, zero-padded spectra has the advantage of increasing the sampling density and so spectral peaks are captured with greater precision. It is therefore important to assess the impact of the nonuniform spectrogram on both EP and integrated loudness estimates.

Evaluation

In this section the effects of varying the auditory filter spacing α, spectral compression criterion β, and incorporating the HGDFT into the loudness model are assessed. All modifications to the roex

[9] Cassidy and Smith III (2007) did not discuss the case of windowing but only compared the execution speed of

filter shape calculations described in section 4.2.2 apply to the approximations here. The stimuli, performance metrics and configuration of the reference model are the same as those detailed in Section 4.2.1. Configuration of the approximations and treatment of data specific to this evaluation are now detailed.

First, the nMAE in the IL and EPs are examined as a function of filter spacing, similarly to the approach of Burdiel et al. (2012), and then as a function of spectral compression criterion. Filter spacings ranged from 0.25 Cams (as used by Glasberg and Moore (2002)) to 4 Cams, in steps of 0.25 Cams. Spacings were tested with and without cubic interpolation estimating the response at 0.1 Cam intervals. Spectral compression values ranged from 0.1 Cams (as employed by Cabrera et al. (2007)) to 1 Cam, in steps of 0.1 Cams.

The base 10 logarithm of the stimulus errors was taken in order to reduce positive skew in the error distributions (and thus biassing the arithmetic mean), as well as making the trend in error visually coherent. Data points therefore correspond to the arithmetic means of the log normalised MAE across stimuli. Thus, discussion generally refers to proportional changes in error between different configurations, although back transformed values are also highlighted to emphasise the scale of the error with respect to the reference features. The shaded areas surrounding plotted lines are the 95% confidence intervals (CIs) of the means and were estimated using a bootstrapping procedure based on 5000 samples (Hesterberg et al. 2003). These intervals have been corrected to eliminate between-stimulus variability which biases the sampling error in repeated-measures designs (see Appendix C). For example, the intervals shown in Figure 4.4(a) were calculated according to a 2 (interpolants) x 16 (filter spacings) within-stimulus design.

After assessing the effects of filter spacing and spectral compression, error statistics for the HGDFT are reported and discussed. This is followed by a statistical assessment of the global-loudness descriptor errors introduced when combining the proposed techniques. Computational savings of these different parameter sets are also discussed. This was done to give insight into the practical implications of the parameterised loudness model. For this final experiment, filter spacings of 0.25, 0.5, 1 and 2 Cams were each tested using spectral compression values of 0.1, 0.2, 0.4 and 0.8 Cams. The HGDFT was also tested at the same four filter configurations, giving a total of 20 unique parameter sets.

Error as a function of filter spacing

Figure 4.4(a) shows the average log nMAE for the IL as a function of filter spacing with (blue dotted line) and without (black solid line) cubic interpolation applied to the EPs. The relative error increases with filter spacing which can be explained by the effects of sampling the EP sparsely:

- Spectral peaks are less well-defined in the resulting EP, meaning salient tonal components contribute less to the total loudness.
- The prediction of SLS is degraded as a consequence of fewer filters per ERB.
- The complex nature of the EP is approximated with fewer sampling points and so the accuracy of the area under the specific loudness pattern is compromised.

The blue dotted line shows the average (log) error when the EPs are interpolated at points defined by the reference filter spacing of 0.1 Cams. As shown by the non-overlapping CIs, the benefit of cubic interpolation appears to be limited to filter spacings below 1.25 Cams, after which the two functions coincide. Note that a log nMAE of -3.0 corresponds to an average nMAE of

the HGDFT against that of the parallel FFTs.

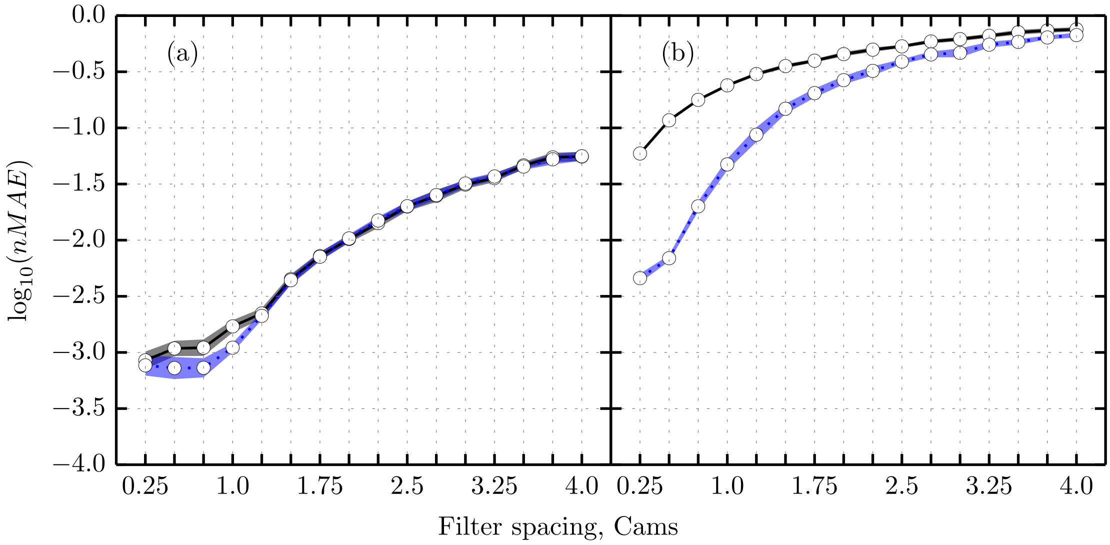


Figure 4.4: The average log nMAE of (a) the IL and (b) the EP as a function of filter spacing in Cam units. The circles show the arithmetic mean of the log stimulus errors and the shaded areas surrounding the lines represent the 95% within-stimulus CI of each estimate. The blue dotted line is the average error as a result of interpolating the approximate EPs at filter locations defined by the reference model.

only 0.1% of the mean IL time series, which is expressed in sones. Thus, such small improvements in loudness estimation may not be of practical importance. Reinforcing this point, the *maximum* global-loudness descriptor error at 1 Cam was only 0.15 dB, occurring for the peak STL. This means that all estimates of global loudness at 1 Cam were within 0.15 dB of those predicted by the reference loudness model.

Figure 4.4(b) shows the trajectory of the average log nMAE in the EPs as filter spacing increases. The black solid line corresponds to the error between the reference and approximation using nearest-neighbour interpolation to quantify error, i.e. approximating the target pattern using a series of rectangles. Both the size and rate of growth in the EP error function is greater than those in the IL error function, especially for intervals below 1 Cam. As expected, cubic interpolation shows a clear improvement over the nearest-neighbour approach. Based on the CIs, it may be concluded that the error functions increase monotonically with filter spacing.

Error as a function of compression criterion

Figure 4.5 shows IL and EP log nMAE as a function of compression criterion. Recall that this parameter determines the amount of smoothing applied to the multi-resolution spectra in order to form a compact representation for subsequent processing. For this test, the filter interval was set to 0.1 Cams in accordance with the target model. As revealed in subplot (b), the EP error increases monotonically with the amount of spectral compression. Peaks and valleys in the auditory patterns become less defined as a result of the averaging. The magnitude of these errors are comparable to those covered by filter spacings of 0.25–1.25 Cams.

The effect of spectral compression on loudness estimates is not so clear: the IL error function is relatively flat over 0.1–0.3 Cams, with a very low mean log error, but shows a sharp rise over 0.3–0.5 Cams before declining. This trend was also reflected in the loudness descriptor errors (not shown). For example, the median (and maximum) of the peak STL descriptor error at 0.3 Cams, 0.5 Cams and 0.6 Cams was 0.01 (0.06), 0.09 (0.33), and 0.03 dB (0.23 dB), respectively. This

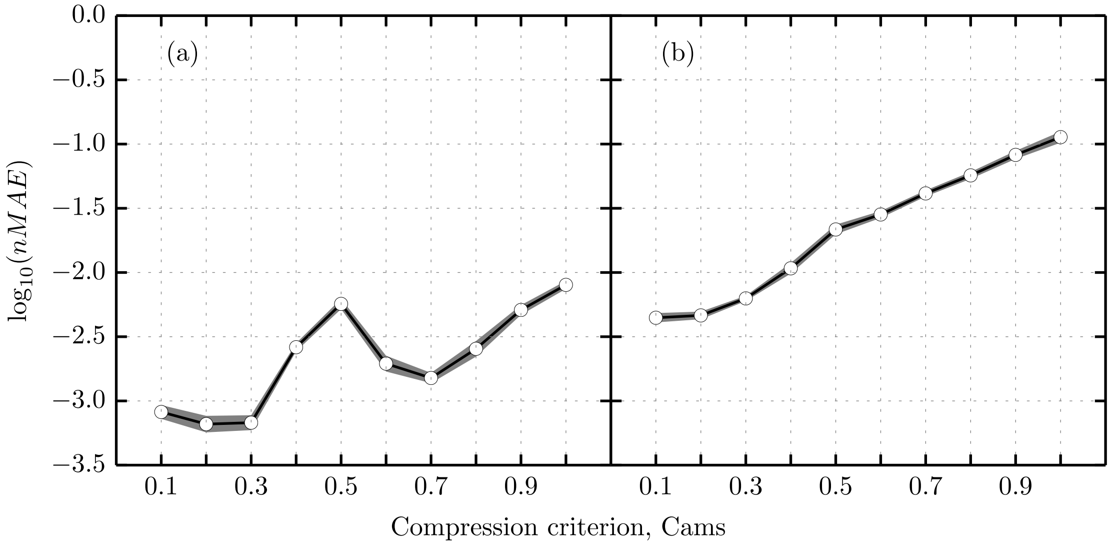


Figure 4.5: The average log nMAE of (a) the IL and (b) the EP as a function of compression criterion in Cam units. The circles show the arithmetic mean of the log stimulus errors and the shaded areas surrounding the lines represent the 95% within-stimulus CI of each estimate.

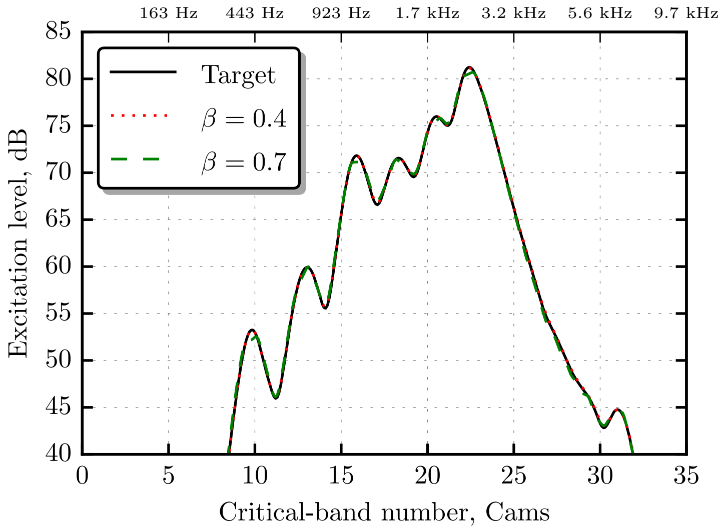


Figure 4.6: Calculated EPs of a segment of track 34 (tubular bells) of the SQAM collection, using the reference model (target) and an approximation with compression values of 0.4 and 0.7 Cams.

non-monotonic trend in error means that although there are greater discrepancies introduced in the EPs (as shown in subplot (b)), the SL transformation and subsequent integration over frequency can result in loudness estimates that are *closer* to the target loudness. Figure 4.6 shows the EPs extracted from a single frame of a tubular bells recording using three configurations: the reference model, $\beta = 0.4$ and $\beta = 0.7$. Despite the 0.7 Cam criterion yielding the worst approximation to the target EP (note the lack of definition near peak excitations), the frequency-integrated SL was actually closer to the reference: the IL as calculated by the three configurations was 40.39 (target), 40.57 ($\beta = 0.4$) and 40.31 ($\beta = 0.7$) sones.

Performance of the HGDFT

A one-way repeated-measures analysis of variance (ANOVA) was used to compare the log nMAE introduced by the HGDFT with that introduced by the spectral compression algorithm, keeping the filter spacing at 0.1 Cams.[10] The ANOVA detected differences in log nMAE for both the IL ($F(10, 290) = 234.1$, $p < 0.0001$) and EP ($F(10, 290) = 612.6$, $p < 0.0001$). Post-hoc pairwise comparisons showed that the HGDFT introduced systematically higher IL error compared to all compression criteria ($p < 0.0001$, Bonferroni corrected for ten tests) except at 0.5 Cams and 0.9 Cams ($p = 1$), and reliably lower error compared to a compression size of 1 Cam ($p < 0.0001$). The difference in mean rank error was smallest at 0.9 Cams. The HGDFT introduced reliably higher EP error than the error at all compression criteria less than 0.8 Cams ($p < 0.0001$, Bonferroni corrected), and reliably lower than at 0.9 and 1 Cam ($p < 0.0001$).

The HGDFT reduces the number of spectral components from 960 to 69, corresponding to a 92.8% CR. A spectral compression criterion of 0.7 Cams yields only 64 components giving a slightly higher reduction of 93.4%. Note, however, that the compression technique was found to result in reliably lower error in both IL and EP domains. Tables 4.6 and 4.7 provide a more detailed comparison between the two approaches with the filter spacing set to 0.1 Cams. Bootstrapped 95% between-stimulus CIs are shown for each mean estimate. To correct for normality, all errors were log transformed prior to calculating the mean error and CIs, which were then back transformed. The resulting averages therefore correspond to the geometric means of their respective distributions.

	Loudness error (%)			EP error (%)	Descriptor error (dB)	
	nMAE	nRMSE	nPE	nMAE	Ψ_{pk}	Ψ_{μ}
Max	1.764	2.201	14.688	13.235	0.337	0.284
Mean	0.510	0.686	3.280	6.370	0.026	0.029
95% CI	[0.450, 0.588]	[0.604, 0.800]	[2.844, 3.881]	[5.708, 7.094]	[0.016, 0.043]	[0.019, 0.044]

Table 4.6: Maximum error and geometric average error with 95% between-stimulus CI, for the HGDFT.

	Loudness error (%)			EP error (%)	Descriptor error (dB)	
	nMAE	nRMSE	nPE	nMAE	Ψ_{pk}	Ψ_{μ}
Max	0.474	0.721	6.854	5.703	0.119	0.102
Mean	0.151	0.205	0.968	4.129	0.019	0.010
95% CI	[0.13, 0.18]	[0.174, 0.246]	[0.791, 1.205]	[3.891, 4.366]	[0.013, 0.029]	[0.007, 0.016]

Table 4.7: Maximum error and geometric average error with 95% between-stimulus CI, for $\beta = 0.7$.

With exception of peak STL Ψ_{pk}, the CIs do not overlap, agreeing with the post-hoc tests which showed that a compression criterion of $\beta = 0.7$ was reliably superior to the HGDFT. The descriptor errors are, however, very low for both techniques, and the expected nRMSEs are less 1%. Therefore, the practical importance of this degraded performance may not be crucial for most audio applications based on loudness metering or normalisation procedures. However, signal processing devices that require accurate estimation of auditory patterns, such as modelling auditory masking or hearing impairment, can benefit from the flexibility of the spectral compression algorithm. Better performance can be expected via spectral compression because the reference multi-resolution spectra are summed into composite bins prior to auditory filtering, whereas the HGDFT directly samples a subset of the total 960 frequency bins. Thus, the EPs derived from compressed spectra are more positively correlated with those output by the reference model.

[10] Because the residuals were non-normally distributed—severely so in the case of the EP nMAE—it was decided to employ the Aligned Rank Transform (ART) for nonparametric factorial ANOVA (Wobbrock et al. 2011). This

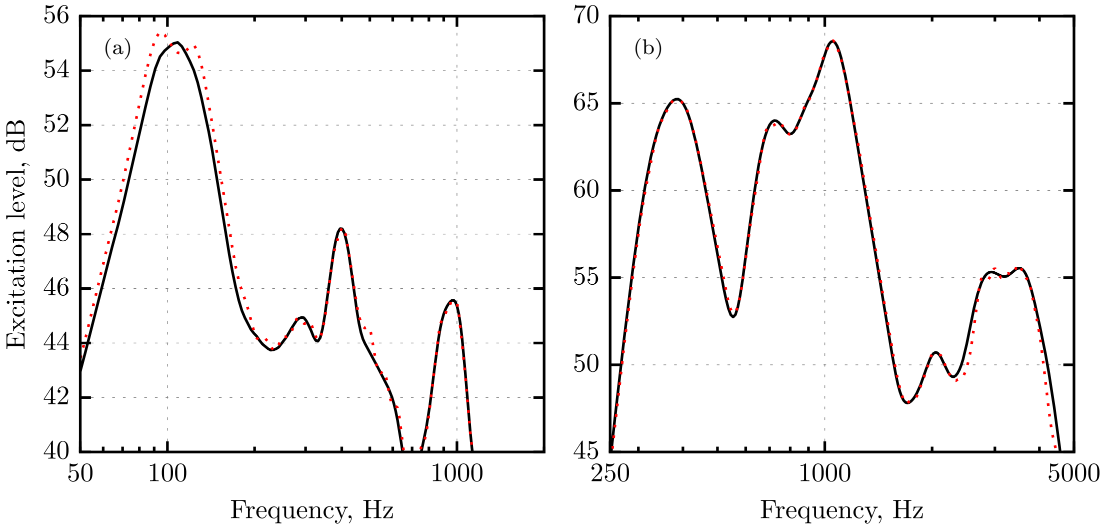


Figure 4.7: Long-term EP of (a) double bass and (b) vocal soprano, taken from track 11 and 44 of the SQAM collection, respectively. The double bass had the highest nMAE; the vocal had the lowest. The black solid lines show the reference EPs and red dotted lines show the approximations obtained using the HGDFT.

A good example of the EP distortions introduced through sampling nonuniformly is illustrated in Figure 4.7(a), which shows the long-term EP extracted from track 11 (double bass) of the SQAM database. This segment received the highest EP error (13.2% as reported in Table 4.6). In contrast, a recording of a vocal soprano (track 44) was associated with the minimum error; the long-term EP is shown in subplot (b). Although the worst-case error is not huge in terms of deviation in decibels, there are two discernible peaks centred about 100 Hz, whereas the reference pattern shows a single broad peak. This occurs because (a) the HGDFT samples every other frequency component, with respect to the reference sampling, in the second spectral band contributing to the multi-resolution spectrum; (b) the auditory filters are narrower at low frequencies. Thus, differences in the spectral distribution between the two approaches are amplified by narrow auditory filters. Such differences could be minimised by increasing the sampling density at lower frequencies, i.e. shifting the upper edge of the first spectral band to a higher frequency. It can be seen that the HGDFT has minimal impact on the EP of the vocal segment, primarily because the maximum input component happened to coincide with the bin centred at 1062.5 Hz (the nMAE is based on discrepancies in linear power so assigns greater weight to higher intensities). Additional issues arise for tonal components near the band edges of multi-resolution spectra, but this is less of a concern for real-world sounds.

Combining the parameters

Figure 4.8 shows the log nMAE measured in the final experiment whereby different combinations of filter spacing (with interpolation) and spectral reduction techniques were combined. The HGDFT introduces the largest error in both loudness and excitation domains. Compression criteria of 0.1 and 0.2 Cams perform similarly across the four filter intervals, with IL and EP nMAEs less than 0.1% and 1%, respectively, at spacings of 0.25 and 0.5 Cams. For these two spectral configurations, the EP error rises quite drastically when the filter spacing is set to 1 Cam, indicating a loss of peak excitation. The IL errors generated by spectral compression values of 0.4 and 0.8 Cams

procedure is based on centring and ranking the data before applying a traditional ANOVA.

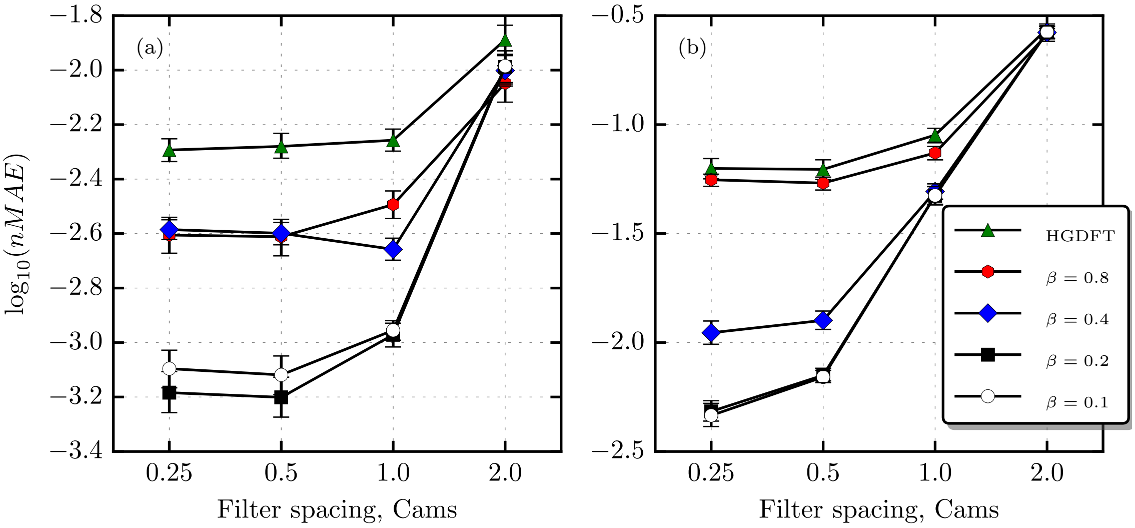


Figure 4.8: The average log nMAE of (a) the IL and (b) the EP as a function of filter spacing, with component reduction method as the parameter. The markers show the arithmetic mean error of the log stimulus errors and error bars show the 95% within-stimulus CI of each estimate.

are comparable, with large overlap between the CIs, although there appears to be a systematic difference at a 1 Cam filter spacing. In contrast, a compression criterion of 0.8 Cams introduces markedly higher initial EP error. Differences in log nMAE between the different spectral reduction techniques become negligible at a filter spacing of 2 Cam, especially in terms of preserving the shape of the auditory patterns.

α	β/HGDFT	CR (%)	Time (s)	nRMSE$_\Psi$ (%)	Descriptor error (dB)	
					Ψ_{pk}	Ψ_{μ}
0.5	0.2	95.8	6.528 6.299 [6.260, 6.356]	0.724 0.099 [0.074, 0.133]	0.118 0.010 [0.006, 0.016]	0.079 0.009 [0.006, 0.013]
1	0.2	97.9	5.183 5.183 [5.166, 5.207]	0.714 0.149 [0.128, 0.178]	0.115 0.009 [0.005, 0.015]	0.074 0.004 [0.002, 0.007]
1	0.4	98.8	4.862 4.751 [4.726, 4.782]	1.026 0.287 [0.243, 0.343]	0.160 0.038 [0.029, 0.050]	0.129 0.033 [0.027, 0.041]
1	0.8	99.4	4.536 4.442 [4.200, 4.469]	0.999 0.411 [0.365, 0.466]	0.200 0.019 [0.010, 0.031]	0.089 0.014 [0.010, 0.021]
1	HGDFT	99.3	3.430 3.333 [3.305, 3.365]	2.123 0.728 [0.640, 0.844]	0.314 0.036 [0.023, 0.055]	0.276 0.032 [0.022, 0.047]

Table 4.8: Performance of the model with different combinations of parameters. The final four columns with 3 values per cell correspond, from top to bottom, to the maximum, mean and 95% between-stimulus CI for the mean.

Table 4.8 assesses the performance of the model when configured with five different parameter sets. The parameter combinations were selected by sorting the different combinations by complexity, dividing this distribution into quartiles, and, within each quartile, selecting the combination with the lowest IL error. Note that although the HGDFT was not selected using this procedure, it has been included in the table as a potential alternative to $\beta = 0.8$ because it offers superior performance at the multi-resolution stage.

The expected RMSEs are less than 1% of the mean IL for all parameter sets. LTL descriptor errors are lower than those for peak STL, but both are very low across all configurations, with the maximum being only 0.31 dB. Thus, a huge reduction in the number of spectral components and auditory filters is possible—as shown by the final column—with negligible impact on estimates of global loudness. Column seven ('Time') shows the absolute processing time of the model when applied to 10 s music clip using the same PC used to first profile the algorithm. All parameter combinations achieve real-time feature extraction (time <10 s), and all configurations give reliably different execution times, inferred from the CIs. Note that despite the HGDFT using slightly more spectral components compared to $\beta = 0.8$ (compare CR percentages), it runs roughly 1 s faster. This is explained by the efficiency of the HGDFT compared to the six parallel FFTs, when $R = 1$ ms.

Parameters	Symbol	Samples (%)
$\alpha = 1$, $\beta = 0.8$	FIR filter	36.2
	`fftw3.so`	24.9
	Spline interpolation	13.9
	Excitation transformation	6.3
$\alpha = 1$, HGDFT	FIR filter	51.2
	Spline interpolation	31.7
	HGDFT	2.1
	Excitation transformation	10.1

Table 4.9: Breakdown of the program profile report for two parameter sets.

Trimmed profile reports for the two fastest parameter sets are given in Table 4.9. In both cases, the performance of the loudness model is no longer limited by the excitation transformation, which previously accounted for 86.8% of the execution time. The pre-cochlear filter now represents the most consuming component, followed by the six FFTs in the case of the first configuration, and then spline interpolation. This ordering is explained by the lower complexity of the HGDFT compared to the parallel FFTs. It might surprise the reader that spline interpolation was so costly, but recall that the interpolation is carried out on the *log* EP, which then have to be transformed back to linear power units.[11] However, as discussed in Section 4.2.3, interpolation seems to be of little practical importance when loudness measures are the primary interest.

4.3 Pre-cochlear filtering

Glasberg and Moore (2002) employed a single 4096$^{\text{th}}$ order FIR filter to approximate the transmission through the outer and middle ear, opting against the more efficient approach of weighting the power spectrum due to discrepancies at low frequencies. This high-order filter is required to meet the requirements of the middle-ear transfer function which is particularly steep below about 75 Hz. Unfortunately, filters of this order are computationally expensive, memory hungry and introduce group delay, all of which are unattractive for real-time application. Because the loudness model is based on the short-term power spectrum, linear phase is not a requirement, making infinite impulse response (IIR) filter realisations candidate solutions. However, the loudness model of Chen and Hu (2012), which also uses six FFTs, shows good performance with the weighted spectrum (WS), indicating a time-domain filter may not be necessary. Note, however, that Chen and Hu (2012) compute the spectrum with twice the frequency resolution used by Glasberg and Moore,

[11] The proportion of samples for spline is approximate because the number of calls to `pow`, `exp` and `log` from `libm-2.19.so` could not be entirely distinguished from other modules.

which has the effect of increasing the running time of the subsequent excitation transformation, thus counteracting the main objective of this chapter.

Implementation

The following four filters were designed to simulate the combined free field to eardrum and middle-ear transfer function as specified in ANSI S3.4 (2007):

1. 4096$^{\text{th}}$ order FIR.
2. 20$^{\text{th}}$ order IIR cascaded with a third order Butterworth highpass filter.
3. Weighted spectrum (WS).
4. Hybrid filter.

Following Glasberg and Moore (2002), the FIR filter was designed using the window method (Smith III 2011). For frequencies between 0.03–15 kHz, the maximum deviation between the filters transfer function and the interpolated target response as per ANSI S3.4 (2007) was only 0.15 dB.

The IIR filter was realised by converting the desired magnitude response to a minimum phase spectrum and then using an equation-error method for the filter design (Smith III 2008). The equation-error method minimises the frequency-weighted output error with respect to the feedforward and feedback filter coefficients in the least-squares sense. It was discovered that an order of 120 yields similar accuracy to the FIR filter (within 1 dB over 0.03–15 kHz). However, for frequencies above about 100 Hz, a good fit was achieved with filters of a much lower order. Although alternative filter design methods exist that assign greater weight to lower frequencies, e.g. Härmä et al. (2000), a simple and elegant solution is to approximate the low-frequency roll-off of the hearing system using a highpass filter, and approximate the remaining transmission response using a second filtering stage (Chalupper and Fastl 2002; Robinson 2002, Appendix K). Here, a third-order highpass Butterworth filter with a centre frequency of 50 Hz and a gain of -14.6 dB was used. The gain of this highpass filter was obtained by minimising the mean squared error between the filter response and middle-ear transfer function given in ANSI S3.4 (2007) over 20–75 Hz using an iterative brute force procedure. A 20$^{\text{th}}$ order IIR was then designed with the middle-ear target response configured to be flat below about 50 Hz, offset by 14.6 dB and cascaded with the highpass filter. At frequencies spanning 0.05–15 kHz, the maximum errors from the (interpolated) combined outer- and middle-ear response were 0.8, 0.89 and 1.06 dB at sampling frequencies 32, 44.1 and 48 kHz, respectively. The magnitude response of the 4096$^{\text{th}}$ order FIR and the IIR/highpass filter cascade along with the target response are shown in Figure 4.9.

The WS method approximates the target response by multiplying the power spectrum with the desired transfer function. Thus, the accuracy of this approach is dependent on both the DFT frequency resolution and the trajectory of the weighting curve. The hybrid filter is the same as method WS, but employs a third-order highpass Butterworth filter to improve the fit at low frequencies, prior to spectral modification. In this case, the specifications of the weighting curve and highpass filter are similar to those used to design the IIR filter.

Evaluation

In line with the fitting procedure of Glasberg and Moore (2006), the four filters, when incorporated into the original dynamic loudness model, were assessed in terms of their ability to predict the free-field absolute thresholds for 36 pure tones with frequency spanning 20 Hz to 14 kHz (ISO 389-7 2005). Pure tones were 2.3 seconds in length, each with a 0.1 second raised cosine ramp applied

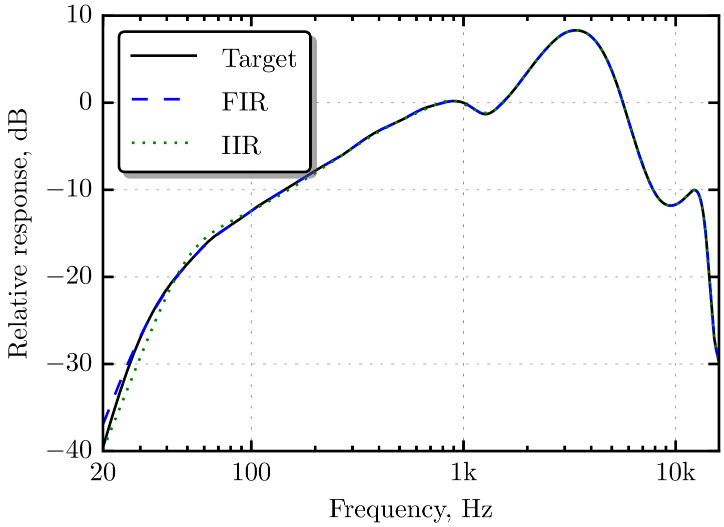


Figure 4.9: Magnitude response of the 4096th order FIR and 20th order IIR cascaded with a third-order highpass Butterworth filter. The target response refers to the interpolated outer- and middle-ear data given in (ANSI S3.4 2007).

to the onset and offset, thus minimising the effect of spectral splatter on loudness measures. The model was configured as per Glasberg and Moore (2002) but with roex filters equally spaced from 1.8 to 38.9 Cams in steps of 0.1 Cam for increased resolution. Loudness predictions were made by averaging the LTL between 0.6 and 2 seconds after the start of each sinusoid (Glasberg and Moore 2002). Threshold estimates were made using the iterative procedure described in Section 3.3.6. A target loudness of 0.003 sones, corresponding to loudness at threshold (ANSI S3.4 2007), and an error tolerance of ± 0.01 dB was used.[12]

Results

	FIR	IIR	WS	Hybrid
Max	4.74	2.12	19.30	3.21
RMSE	1.19	0.80	4.09	0.99
95% CI	[0.66, 1.71]	[0.60, 1.00]	[1.22, 6.52]	[0.64, 1.30]

Table 4.10: Absolute threshold prediction errors for the four filters. A bootstrapped 95% CI is provided for each RMSE. All measures are in dB.

Table 4.10 shows the maximum absolute errors and RMSE (with 95% CI) between the ISO 389-7 absolute thresholds (in dB) and those predicted by the model when configured with five different filters. For the FIR filter, the 4.74 dB maximum can be explained by the deviation from the target magnitude response below 30 Hz. At low frequencies, the IIR filter yields a better fit to the data even though its *magnitude* response is less accurate below 100 Hz compared to that of the FIR filter (refer to Figure 4.9). It is assumed that this is attributed to the influence of spectral sidelobes because, below 35 Hz, the slight negative deviation from the target middle-ear response reduces SLS caused by the DFT sidelobes resulting in a *better* fit to the absolute threshold data. For frequencies between 50 Hz and 14 kHz, the RMSE was 0.62 dB for the FIR filter and 0.77 dB for the IIR filter. The maximum absolute error for the FIR filter over this frequency range was 1.36 dB, compared

[12] Threshold predictions were also made using a sone-to-phon converter as described by Glasberg and Moore (2006) and the results were essentially the same as those obtained using the iterative procedure.

to 4.74 dB which occurred when making predictions at very low frequencies. Thus, ignoring single maximum discrepancies, the FIR and IIR filter realisations yield similar performance in terms of the average threshold prediction. As expected, spectral weighting is problematic at low frequencies as a consequence of broad spectral sidelobes relative to the bandwidth of the auditory filters, which are weighted by a gain function with a steep low-frequency roll-off. The hybrid approach corrects for this by incorporating a highpass filter and hence gives significantly lower maximum error.

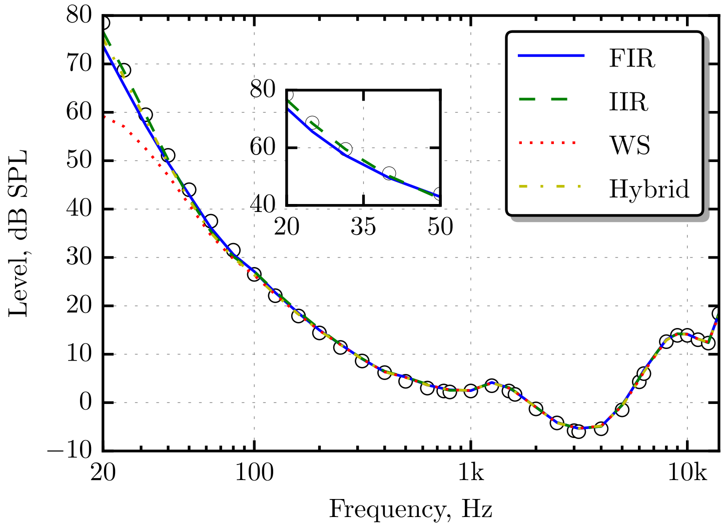


Figure 4.10: The circles show the free-field absolute threshold values given in ISO 389-7 (2005). The lines are the predicted thresholds output by the loudness model using different pre-cochlear filters. The inset shows the ISO data and predictions for the FIR and IIR filter at very low frequencies as discussed in the text.

Figure 4.10 the shows model predictions for each filter configuration. The lines are hard to distinguish except at frequencies below 50 Hz where the approximation WS fails to capture the transmission loss of the middle ear. Burdiel et al. (2012) suggested using the frequency-domain weighting function for real-time loudness modelling, but clearly this choice is application specific; weighting low frequencies with insufficient DFT resolution should be avoided if accurate predictions of low-frequency narrowband sounds are required. The performance of the FIR and IIR filters at low frequencies is exemplified by the inset of Figure 4.10 which reveals the absolute deviation from the measured data between 20 and 50 Hz. At frequencies above this range, the deviations from the ISO data are consistent across all filters.

Both execution time and memory resources can be decreased by using either the IIR or hybrid filter, whilst preserving the response of the pre-cochlear filter. However, one must be cautious whenever spectral-weighting functions are used, because the resulting accuracy is dependent on the DFT length and the trajectory of the desired filter curve. For example, the frequency response of headphones can differ significantly at low frequencies and so approximating their response by modifying the shape of the power spectrum may indeed result in erroneous loudness predictions. Ignoring the highpass filter, which is common to both methods, the IIR filter requires only 41 coefficients compared to the K coefficients used by the hybrid filter, where K is the number of components forming the multi-resolution spectra. In terms of computation, the IIR filter becomes advantageous whenever $\frac{K}{R} > 41$, where R is the hop size in samples. Stated another way, the workload of the spectral-weighting methods is dependent on both the DFT resolution and the hop size. However, differences in execution time between these two approaches were found to be of little concern for practical configurations of the model.

4.4 Parameterisation

In the preceding sections, a number of techniques were introduced that targeted faster computation of EPs and low-latency, efficient pre-cochlear filtering. A final assessment was carried out to investigate the performance of the model when these techniques are combined. In particular, the IIR/highpass filter cascade described in Section 4.3 was selected to simulate the transmission through the outer and middle ear as this gave the best fit to absolute threshold data over the full range of frequencies tested. The excitation transformation was then configured according to the parameter sets listed in Table 4.8, but this time with no interpolation.[13] Furthermore, the effect of hop size R was also explored by testing each configuration at hop sizes of 1, 2, 4, 8 and 16 ms. Thus, five different configurations of the excitation transformation were each tested at five hop sizes. For hop sizes above 1 ms, time series error was measured using a sample and hold procedure, i.e. using the previous loudness value to predict current (non-sampled) time points.

Subplots (a) and (b) of Figure 4.11 show the mean log nMAEs measured in the STL and LTL, respectively, for the 25 model variants. Note that the IL errors are not shown because they were found to be highly affected by hop sizes above 2 ms, thus making the errors indistinguishable. The plots reflect this in that the STL error functions are both lower and shallower than their respective LTL errors, which can be explained by the greater smoothing performed by the LTL integrator. From comparison with Figure 4.8(a), which shows the IL error as a function of filter spacing, a doubling of the hop size yields greater proportional growth in nMAE compared to a doubling of the filter spacing, at least for spacings less than 1 Cam. Note that using the *reference* model, the geometric average STL nMAE at 2 ms was 0.28% which corresponds to a log error of -2.56. Thus, as shown in Figure 4.11(a), superior performance can be achieved by combining filter spacings of 1 Cam with compression criteria of up to 0.4 Cams. In actuality, a compression criterion of 0.7 Cams would likely be optimal in this respect since the loudness error is non-monotonic with spectral compression (see Figure 4.6).

Subplots (c) and (d) of Figure 4.11 show the maximum of the descriptor errors across the 30 sound segments, which denotes the level change ΔL required to match the peak STL or mean LTL between the reference and approximation. For hop sizes of 4 ms and below, the level differences are all less than 0.3 dB. These worst-case errors rise quite drastically at 8 ms, with the maximum error being 2.5 dB in the case of peak STL for all configurations. This discrepancy occurred for castanets recordings (track 27 of the SQAM collection), an instrument characterised by a series of percussive impulses. Figure 4.12 shows the STL and LTL (in phons) for this segment as measured using the reference model (black solid lines) and various approximations (red dotted lines). The approximations were made using a 1 Cam filter spacing, the HGDFT, the newly proposed IIR filter and four different hop sizes (one per subplot). For this segment, small yet identifiable discrepancies are revealed around the peaks of the STL for the case $R = 4$ ms. Errors in the vicinity of numerous consecutive transients cause noticeable deviations in LTL as a consequence of the two-second release time constant used by the temporal integrator. This is especially noticeable at around 3.5 seconds for the case $R = 8$ ms. Although further research is needed to evaluate the necessary resolutions required by engineers when monitoring loudness, hop sizes up to 4 ms seem more than adequate for capturing the development of loudness of the corpus of sounds tested here, especially when errors in peak STL of 0.4 dB can be tolerated.

[13] This final experiment was repeated with cubic interpolation but no overall improvement was found.

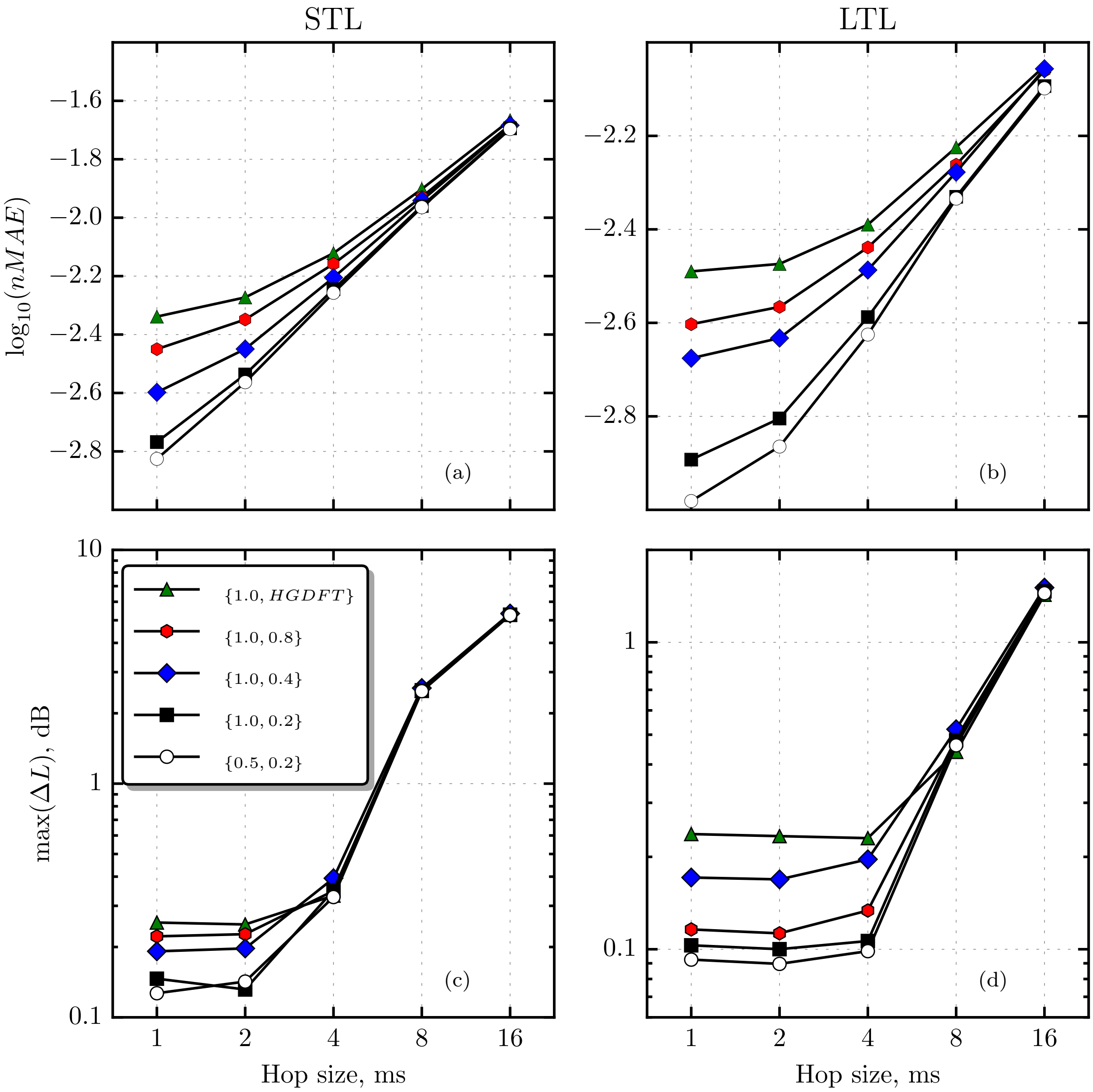


Figure 4.11: As a function of hop size: mean log nMAE for (a) the STL and (b) the LTL; maximum absolute level change required for equal loudness according to (c) peak STL and (d) mean LTL. Filter spacing (α) and spectral configuration (β or HGDFT) combination are specified in the legend as $\{\alpha, \beta\}$.

Figure 4.13 shows the mean time to compute a 10 s audio file using the different approximations on the same desktop PC used to profile the model throughout this study. Although computational times are implementation and hardware specific, the reader is reminded that the reference implementation took 9.9 minutes to extract the loudness of the same audio file. Note that the 95% CIs were roughly the same size as the markers shown in the plot so have been omitted for clarity. It can be seen that every model configuration achieves real-time processing.

As expected, the HGDFT combined with a 1 Cam filter spacing executes the fastest at all hop sizes, at the expense of introducing the most error as measured using all metrics. Comparing the compression based optimisations at a given hop size suggests that the rate of change in execution time is not constant with proportional changes in the number of spectral components and auditory filters. This is expected because the bottleneck is no longer the excitation transformation, but now the multi-resolution spectrogram. The running times for the top four lines are essentially parallel, averaging a two-fold speedup per doubling of hop size. This occurs because a number

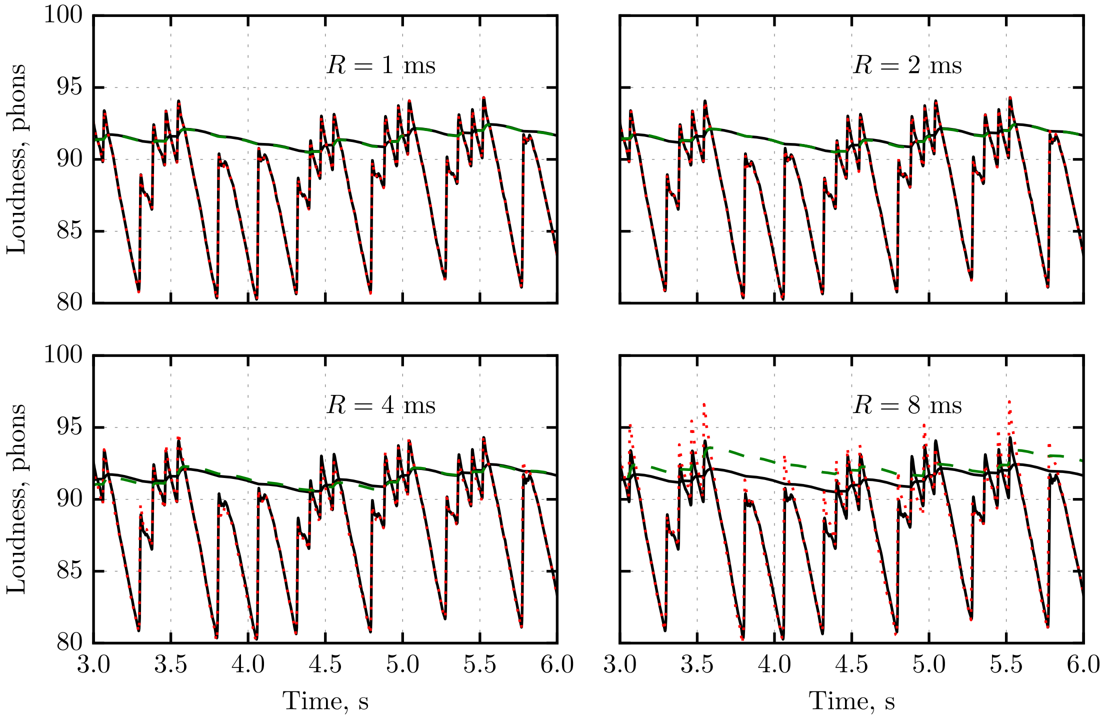


Figure 4.12: Approximations of the STL (red dotted lines) and LTL (green dashed lines) of a segment of castanets recording for hop sizes R of 1, 2, 4 and 8 ms. The filter spacing was 1 Cam and spectra were obtained via the HGDFT. In each plot, the black lines show STL and LTL output by the reference model with $R = 1$ ms.

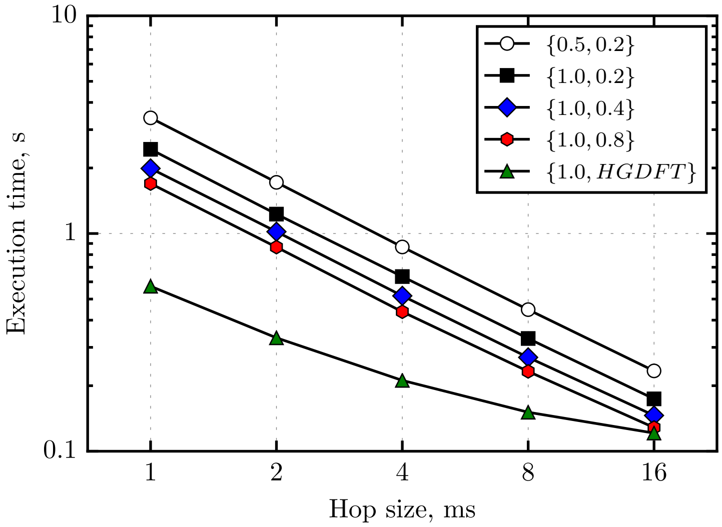


Figure 4.13: Mean time in seconds to extract the loudness of a 10 s audio file using the different approximations as a function of hop size. Filter spacing (α) and spectral configuration (β or HGDFT) combination are specified in the legend as $\{\alpha, \beta\}$.

of different stages simultaneously benefit from fewer analysis frames. In contrast, lower frame rates do not benefit the HGDFT to the same extent because this algorithm requires sample-based processing. For this implementation, initial increments in hop size lead to greater proportional speedup compared to at larger values, primarily because the excitation transformation stage still dominates the running time at $R = 1$ ms for this variant.

4.5 Real-time excitation-pattern binaural loudness meters

In the previous sections, the architecture and complexity of a complex auditory model designed to calculate auditory patterns and perceived loudness of time-varying signals was discussed in detail. Alternative signal processing methods were presented and incorporated into the algorithm to achieve efficient calculation of perceptual sound features. Existing proposals from prior work were extended and compared, using representative musical signals as the primary source of input for evaluation. In this section, these techniques are consolidated to realise a real-time *binaural* loudness meter to help both experienced and amateur audio engineers with music production and to facilitate psychoacousticians and acousticians when measuring, calibrating and aligning the perceived intensity of sound. In addition, the generalisability of these techniques are further demonstrated by using them to reduce the computational complexity of the more recent EP loudness model of Chen and Hu (2012).

4.5.1 Motivation

The need to measure perceived loudness is imperative within fields such as psychoacoustics and audio engineering. In particular, accurate loudness alignment of complex stimuli is crucial when conducting controlled listening experiments and configuring multichannel systems (Zacharov 1998). In broadcasting, the loudness of a wide range of programme material must be consistent yet natural in order to maintain a comfortable listening experience across multiple platforms.

Following the ITU-R broadcast standard (ITU-R BS.1770 2015) first published in 2006, subsequent EBU recommendation (EBU R 128 2014) and metering specifications (EBU Tech 3341 2016), and related standards (ATSC A/85 2013), a number of commercial loudness meters have appeared over the past few years to help broadcasters monitor and normalise the loudness of audio content. These meters make use of single-band loudness algorithms which generally comprise a fixed linear filter to simulate the frequency-sensitivity of the human ear at typical listening levels, followed by an RMS-level detector. Such efficient estimators have proven to work well on broadcast material (Skovenborg and Nielsen 2004b; Soulodre 2004) and have been extended to capture the perception of single instruments (Pestana et al. 2013) making them attractive candidates for real-time meters (Skovenborg and Nielsen 2007). Indeed, following the work of the ITU-R and EBU, a spawn of real-time loudness meters have appeared on the digital market: TB's EBULoudness[14], Klangfreund's LUFS meter[15], TC electronic's LM6[16], Gimm Audio's LevelView[17] and Waves' WLM Plus[18]. Some general features offered by these meters are:

1. Compliance with international loudness recommendations (ATSC A/85 2013; EBU R 128 2014; ITU-R BS.1770 2015; EBU Tech 3341 2016).

2. The measurement of short-term, momentary and integrated loudness as per EBU Tech 3341 (2016).

3. Loudness descriptors such as Loudness Range (EBU Tech 3342 2016), Programme Loudness (ITU-R BS.1770 2015), and Peak Loudness (EBU R 128 2014).

[14] http://www.toneboosters.com/tb-ebuloudness/
[15] https://www.klangfreund.com/lufsmeter/
[16] http://www.tcelectronic.com/lm6-plug-in/
[17] http://www.grimmaudio.com/pro-products/software/levelview/
[18] http://www.waves.com/plugins/wlm-loudness-meter

4. Mono, stereo and 5.1 surround sound metering.
5. True-peak metering following EBU R 128 (2014) and ITU-R BS.1770 (2015).

In contrast to single-band loudness models, multiband models founded on empirical measurements of auditory filters and EPs (Zwicker et al. 1957; Patterson 1976; Glasberg and Moore 1990) have received little interest outside of psychoacoustics where these algorithms are commonly used by researchers to calibrate the loudness of stimuli in preparation for controlled listening experiments or to support hypotheses following observed effects arising from empirical studies. Like single-band algorithms, multiband models also correct for the response of the outer and middle ear, but go a step further by accounting for the frequency selectivity of the cochlea.

The primary drawback of multiband models is their computational demand which consequently limits their application. Furthermore, the efficient single-band approach has proven to be robust when measuring the loudness of programme material (Skovenborg and Nielsen 2004b; Soulodre 2004). However, these tools may not be suitable for measuring the loudness of sounds with very different spectral profiles or those presented over a wide listening level and are incapable of providing frequency-domain statistics describing the distribution of loudness across frequency, nor can they be used to quantify masking effects between different sound sources. To model such phenomena, it is necessary to obtain auditory representations of sounds by simulating peripheral filtering. At the same time, it remains to be seen how state-of-the-art auditory models compare to simpler loudness estimators such as the ITU-R BS.1770 algorithm when applied to everyday sounds. From a music production perspective, audio engineers often find spectral representations of sound useful for comparing and manipulating spectral content. Compared to FFT-based spectrograms, the auditory patterns output by multiband loudness models give a more accurate indication of frequency masking and can be used to identify perceptually-salient frequency regions. In addition, because of their physiological basis, these models are versatile and can be tuned to account for cochlear hearing loss (Moore and Glasberg 2004; Chen et al. 2011) and therefore used to enhance hearing aid compression schemes and fitting procedures (Moore and Glasberg 1998).

Here, the dynamic EP loudness models of Glasberg and Moore (2002) and Chen and Hu (2012) are modified to obtain fast and efficient estimates of perceived loudness whilst maintaining agreement with empirical data. Following an author-year naming convention, these two loudness models are referred to as the GM02 and CH12, respectively.[19] Although real-time loudness meters based on the core loudness estimation algorithm used by the GM02 have been developed (Stone et al. 1997; Tuomi and Zacharov 2000), a full binaural dynamic loudness meter incorporating more recent loudness theory has not yet been established. Here, using the established signal processing techniques established earlier in this chapter, efficient parameterised adaptations of the GM02 and CH12 are developed and presented as real-time binaural loudness meters, and the errors introduced by the approximations are statistically analysed. Unlike previous work, the models have been realised as software plugins for use in digital audio workstations (DAWs) and are thus not restricted to specific host environments or hardware. The following sections, therefore, extend the exploratory work presented in the first half of this chapter by comparing the performance of two different yet related models of auditory periphery that have been tailored for real-time processing of binaurally presented sounds.

[19] For a detailed analysis of both models, see Sections 3.3.3 and 3.3.4.

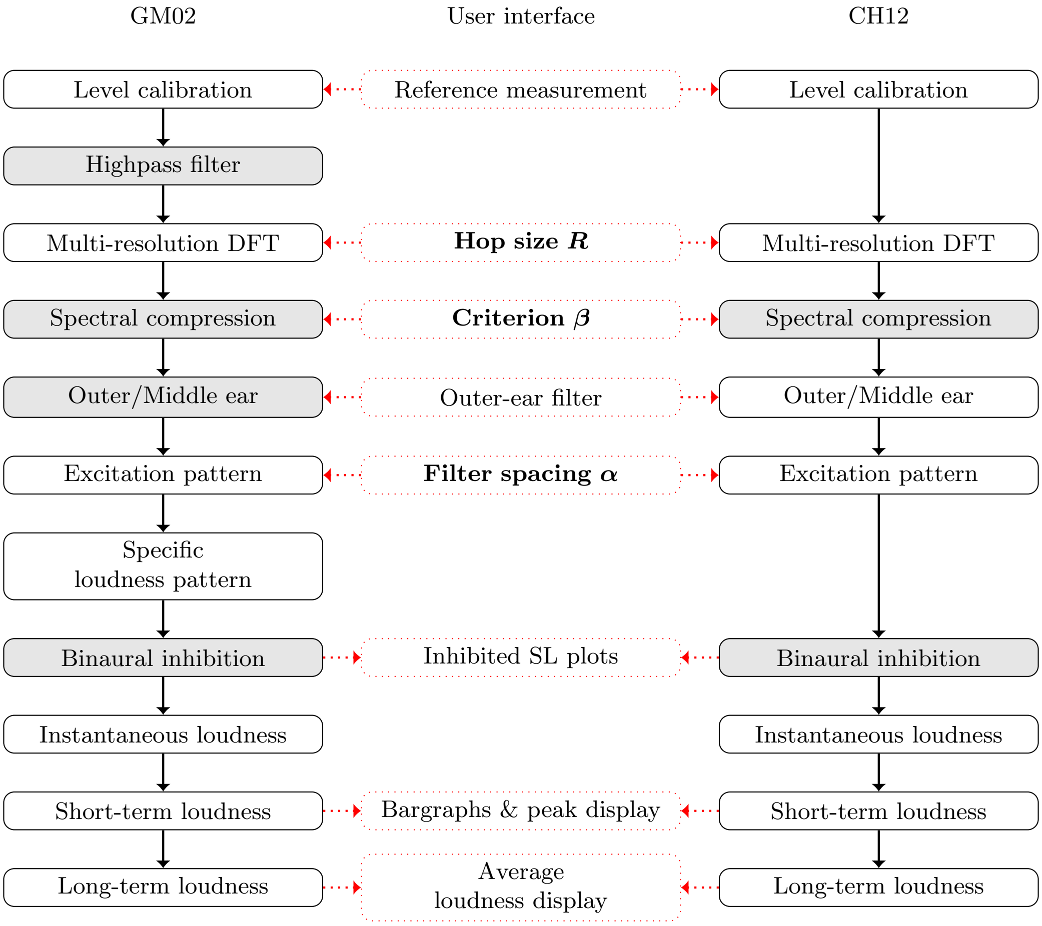


Figure 4.14: Block diagram of the loudness models within the context of the loudness meter. Bold text indicates parameters investigated in this study. Shaded blocks denote additional stages that have been modified/added for efficient processing and binaural functionality.

4.5.2 Overview of the loudness meters

An overview of both the GM02 and CH12 within the context of the binaural loudness meter is shown by Figure 4.14. The general structure is similar to the one described by Figure 4.1, except that the two algorithms operate on stereo information and incorporate an additional stage to better model binaural loudness perception. Note also that the pre-cochlear filter used by original GM02 was replaced with a hybrid filtering scheme presented in Section 4.3 to allow the meter to function with minimal overhead in applications where the sampling frequency is unknown. The GM02 meter was also implemented using the modifications to the excitation transformation stage presented in Section 4.2.2.

The CH12 differs from the GM02 in the following ways:

1. It uses twice the frequency resolution (largest window size of 128 ms compared to 64 ms for the GM02).

2. The pre-cochlear filter is performed entirely by weighting the power spectrum.

3. The middle-ear transfer function follows the one given by Chen et al. (2011).

4. It uses a set of *double-roex* filters (Chen et al. 2011) equally spaced from 1.5 to 40.2 in steps of 0.1 Cams (388 filters).

5. No transformation from excitation to SL is required because the double-roex filter incorporates a compressive nonlinearity.

Chen and Hu (2012) decided to double the frequency resolution because the DFT specification reported by Glasberg and Moore (2002) resulted in a spectral bandwidth that exceeded the critical bandwidth at 1 kHz. Consequently, the predicted total loudness of pure tones is larger than expected (refer to Table 3.4). Unlike the GM02, the filter bank used by the CH12 incorporates compression and thus the area under the EP is proportional to loudness. The output of the filter bank was therefore scaled by the constant of proportionality in order to arrive at the SL pattern, which is then fed to the binaural inhibition stage.

The binaural inhibition algorithm

Moore and Glasberg (2007) surveyed data on binaural loudness summation which indicated that a diotic sound is less than twice as loud as the same sound presented monaurally. Subsequently, they proposed a modification to their steady-state loudness model (Moore et al. 1997), which assumes a simple addition of loudness across ears, to account for this phenomenon. The general idea is that the auditory patterns of the signals at the two ears interact in such a way that the loudness at both ears is reduced in a frequency-dependent manner. The inhibition of loudness is broadly tuned such that loudness reduction takes place even when the spectra at each ear are separated by several critical bands.

Let $\psi_l(z_c)$ and $\psi_r(z_c)$ denote the SL in critical-band number z_c, on the Cam scale, in the left and right ear, respectively. The two patterns are smoothed by convolution with the following Gaussian kernel:

$$W(\Delta z_c) = e^{-(0.08\Delta z_c)^2}, \tag{4.21}$$

where Δz_c is the deviation from the critical-band number of the filter expressed in Cams, and the constant 0.08 determines the spread of inhibition along the Cam scale.[20] The smoothed patterns, denoted $\hat{\psi}_l$ and $\hat{\psi}_r$, are then used to define the two gain contours g_l and g_r for scaling the original left and right SL patterns, respectively:

$$g_l(z_c) = \frac{2}{1 + \operatorname{sech}\left(\frac{\hat{\psi}_r(z_c)}{\hat{\psi}_l(z_c)}\right)^p} \quad \text{and} \tag{4.22}$$

$$g_r(z_c) = \frac{2}{1 + \operatorname{sech}\left(\frac{\hat{\psi}_l(z_c)}{\hat{\psi}_r(z_c)}\right)^p}, \tag{4.23}$$

where $p = 1.5978$. The inhibited SL patterns are then given by

$$\psi_l'(z_c) = \frac{\psi_l(z_c)}{g_l(z_c)} \quad \text{and} \tag{4.24}$$

$$\psi_r'(z_c) = \frac{\psi_r(z_c)}{g_r(z_c)}. \tag{4.25}$$

An example of this procedure integrated with the ANSI S3.4 (2007) stationary loudness model is shown in Figure 4.15. The stimulus at the left ear was a 1 kHz tone presented at 60 dB SPL (free field with frontal incidence) and the stimulus at the right ear was a two-tone complex comprising a 300 Hz tone and a 3 kHz tone, both at 60 dB SPL. The SL patterns and inhibited SL patterns are

[20] Note that the Gaussian defined by Moore and Glasberg (2007) is a function of normalised deviation from the centre frequency of the filter which would suggest that the spread of inhibition varies with Cam frequency. The correct formulation, as presented here, is given in the draft of the new ISO 532 standard (ISO 532-2 2014).

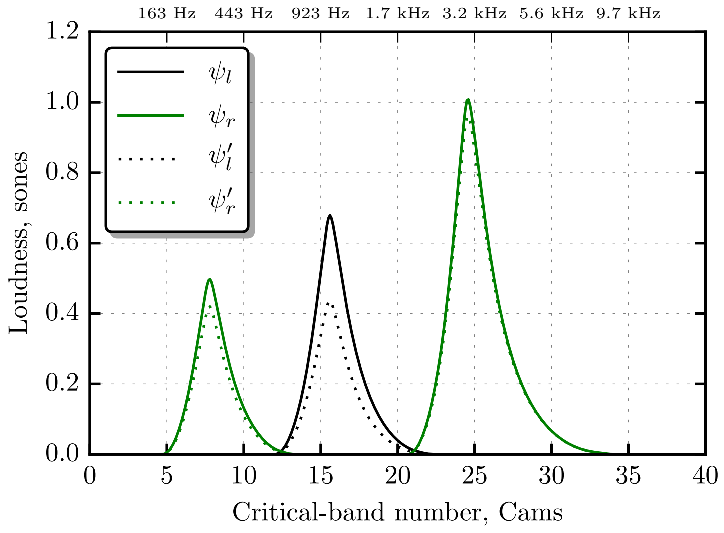


Figure 4.15: Example of inhibitory interactions between the SL patterns at the two ears.

shown by the solid and dashed lines, respectively. It can be seen that loudness of the 1 kHz tone in the left ear (black lines) is reduced by both components at the right ear, even though the auditory patterns barely overlap. The 1 kHz tone in the left ear affects the loudness of the 300 Hz tone in the right ear (green lines) to a greater extent than the 3 kHz tone because it is closer in frequency and has a greater effective excitation after passing through the transfer function of the outer and middle ear. The loudness of the 1 kHz tone in the left ear is largely affected by dominant tone at 3 kHz presented to the right ear. Binaural loudness was measured as 7.0 sones for the original loudness model, and 5.9 sones for the revised model incorporating binaural inhibition. Henceforth, SL refers to the loudness per critical band at the output of the binaural inhibition model.

Meter interface and configuration

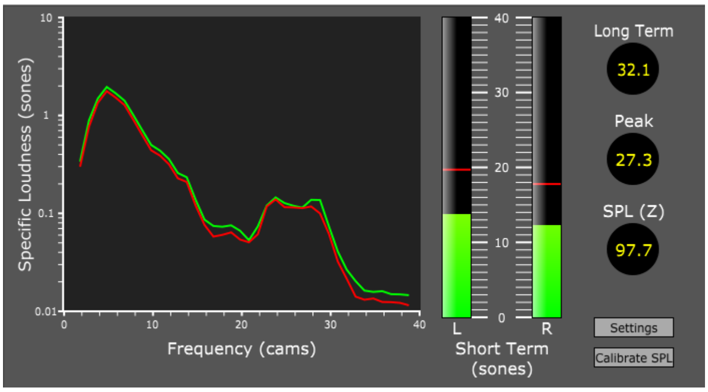


Figure 4.16: Loudness meter user interface.

The loudness meter interface, shown in Figure 4.16, displays SL on a logarithmic scale as a function of frequency on the Cam scale. Both left and right patterns are shown to facilitate visual comparison. In addition, bargraphs have been added to show the STL (in sones) for each ear, which are useful for comparing and aligning the loudness at the two ears. Finally, the binaural LTL, calculated according to the original specifications, and peak STL are displayed for purposes of measuring the overall loudness impression of a sound. Although both the GM02 and CH12 output STL and LTL, the peak STL was added because (a) maximum loudness is important for determining the global loudness of speech (Zwicker 1977), and (b) both LTL and peak STL can be used to infer the loudness range or perceived crest factor of a sound. Peak loudness is calculated by applying a peak follower with a decay time of 2 s to the binaural STL (see Equation 5.10).

The third number box displayed by the meter gives the average unweighted SPL which is necessary for system calibration. Two additional screens can be activated using the appropriate buttons. The first is a settings screen for configuring the efficiency of the models in accordance with the exposed parameters highlighted in bold in Figure 4.14. The second is a calibration window, enabling the user to enter SPL measurements and calculate calibration gains. By default, the meter assumes that 0 dB full scale (FS) corresponds to a peak level of approximately 94 dB SPL.

4.5.3 Profiling the models

The three parameters governing the speed-accuracy trade-off are: hop size R, filter spacing α and spectral compression criterion β. A program to extract the loudness from 30 stereo sounds was developed for both models, each configured with their respective default parameter set (see Section 4.5.4). The sample-based profiler OProfile (2015) was used to identify the most computationally intensive parts of each model. The programs were run on an idle Intel Core 2 Quad CPU Q8200 operating at 2.33 GHz with 4 GB RAM running Ubuntu 14.04 over 10 executions to increase the sample size. The profile reports are summarised in Table 4.11, which shows the number of samples accumulated by the profiler in the most time-consuming processing modules.

Process	GM02	CH12
Excitation transformation	2.7×10^8 (87.3)	1.3×10^8 (67.2)
FFT	1.4×10^7 (4.6)	3.4×10^7 (17.6)
Binaural Inhibition	8.1×10^6 (2.6)	8.8×10^6 (4.6)

Table 4.11: Program profile report showing the absolute number of samples attributed to each of the main processing modules (and percentage of total application samples collected).

As expected from the previous profiling of the original GM02, both algorithms spend the majority of time transforming short-term power spectra to an EP. Compared to the CH12, the GM02 collected a larger number of samples in this module, likely due to the added overhead of calculating variables pertaining to the roex filter shapes—the double-roex filters used by the CH12 can be pre-calculated and stored in memory. Furthermore, the number of frequency points used by the CH12 is twice that of its predecessor and thus the time to execute the multi-resolution DFT is greater. The time required to compute the SL values was also longer for the CH12, which can be attributed to the use of slightly more auditory filters.

Importantly, the time complexity of the most computationally expensive stages involved in the excitation transformation of both models for a given frame are linear functions of the number of frequency components K and auditory filters A. For both models, the dot product of the auditory filters and power spectrum is associated with the highest complexity being $O(AK)$. Note that for the GM02, the filter slopes and therefore their shapes (also of $O(AK)$ complexity) must be computed on a frame-by-frame basis. Thus, although hop size reduces complexity by discarding frames, the filter spacing and spectral compression technique target the bottleneck of these models directly.

4.5.4 Approximations and evaluation procedure

Using an evaluation methodology based on the work of Burdiel et al. (2012), the three free parameters were varied independently for each model to reduce the computational workload and the errors between the reference and approximated auditory features. The parameter values used for each reference model and the approximations are given in Table 4.12 using the notation [start: delta: end].

Parameter	Reference	Approximation
Hop size (R)	1 ms	[2:2:16] ms
Filter spacing (α)	0.1 Cams	[0.25:0.25:4] Cams
Spectral compression (β)	NA	[0.1:0.1:1] Cams

Table 4.12: Reference and approximation parameter values.

Note that the reference configurations do not make use of spectral compression; instead, the frequency points are linearly spaced over a broad frequency range at a resolution determined by the largest window. The models were configured to process stereo recordings using the outer-ear (free field) (ANSI S3.4 2007) and middle-ear (ANSI S3.4 2007; Chen et al. 2011) transfer functions. For both models, the effect of filter spacing was measured with and without cubic interpolation applied to the (log) EP, as described in Section 4.2.3.

Stimuli

Each model variant was evaluated by extracting loudness features from 30 stereo tracks taken from the SQAM database (EBU Tech 3253 2008) and assessing the error introduced. The source material was similar to those listed in Table 4.2, but different segments were taken and all were stereo and sampled at 44.1 kHz. Representative sound segments of the selected tracks were edited manually (average duration of 4.4 s, SD = 0.5 s) and then peak calibrated to a random level between 84 and 94 dB SPL giving a spread of typical listening levels across the 30 sources (average RMS of 72.2 dB SPL, SD = 4.4 dB; average peak of 89.1 dB SPL, SD = 2.7 dB).

Errors

For this assessment, the nRMSE metric, defined by Equation 4.7, was used to assess the error in the STL time series. Similarly, the error in the SL was measured by applying the same statistic to the frame-wise concatenated SL patterns, i.e.

$$\text{nRMSE}_{\psi'} = \frac{\sqrt{\frac{1}{MA}\sum_{m=1}^{M}\sum_{i=1}^{A}(\psi'_{m,i} - \hat{\psi}'_{m,i})^2}}{\overline{\psi'}}, \tag{4.26}$$

where $\psi'_{m,i}$ is the total SL in auditory filter i at frame m, obtained by summing the corresponding SL across both ears. $\hat{\psi}'$ is the same feature measured by the approximation and $\overline{\psi'}$ represents the average SL used to normalise the error. Here, the decibel was used to log transform the nRMSE measured on a given stimulus:

$$\text{nRMSE(dB)} = 10\log_{10}\text{nRMSE}. \tag{4.27}$$

In order to calculate the error at hop sizes greater than 1 ms, a sample and hold procedure was applied to the approximated STL and SL time series. For filter spacings above 0.1 Cams without interpolating EPs, the SL error was computed using nearest-neighbour interpolation.

For each stimulus, the level change ΔL required for equal overall loudness between the reference and approximated loudness predictions was estimated using the sone ratios obtained from three objective measures of global loudness: the average LTL Ψ_μ (Glasberg and Moore 2002), the peak STL Ψ_{pk} (Zwicker 1977) and the 95th percentile of the STL distribution Ψ_{95} (Fastl and Zwicker 2007). In contrast to the previous evaluation procedure (see Section 4.2.1), where ΔL was approximated using loudness level in phons, an iterative procedure was employed to find the optimal gain required for equal loudness according to each global-loudness descriptor.[21] Specifically, ΔL was obtained for each descriptor by minimising the sone ratio in decibels using a tolerance of 0.01 dB. For example, the level change required to equate the peak STL of the reference and approximation was found by minimising the magnitude of

$$\epsilon = 10\log_{10}\left(\frac{\Psi_{pk}}{\hat{\Psi}_{pk}}\right). \tag{4.28}$$

Statistical methods

In this work, the stimuli are treated as the experimental units, e.g. as the subjects used in psychology studies, which represent a random sample of the population of interest. As such, the errors can be viewed as the dependent variable in a repeated-measures design (because the stimuli are passed through the models multiple times, each with a different experimental configuration). Such a design was supported by fitting a linear regression model to the data and inspecting the residuals by stimulus, revealing a clear ‘stimulus effect’ caused by stimulus-specific errors.

Despite the logarithmic transform applied to the nRMSEs (Equation 4.27), some model configurations lead to non-normally distributed data with moderate skew, making formal parametric statistical inferences untrustworthy, given their assumptions, e.g. normally distributed residuals. Figure 4.17 shows a Q-Q plot of the standardised residuals extracted from a two-way ANOVA used to test the effect of hop size (eight settings) and model (GM02 or CH12) on STL error (in dB). Clearly the residuals are long-tailed and therefore the assumption of normality has been violated.

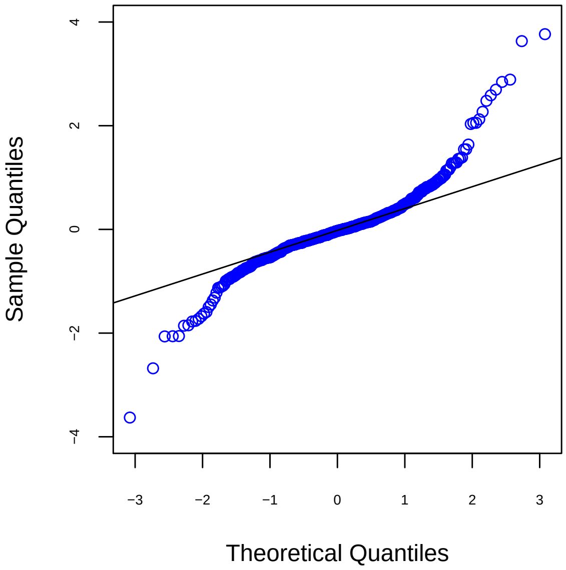


Figure 4.17: Q-Q plot of standardised residuals extracted from a repeated-measures ANOVA used to analyse the STL error data obtained from the hop size experiment.

[21] This was practically possible because the ΔL was not calculated for as many approximations.

It was decided to employ nonparametric ANOVAs for each parameter (hop size, spectral compression and filter spacing). The statistical analyses were performed using the ART for nonparametric factorial ANOVA (Wobbrock et al. 2011). Unlike other nonparametric methods, e.g. the Friedman test, this algorithm supports factorial experiment designs with repeated measures. For each time series error (STL or SL), two-way ANOVAs were used to analyse the effect of each parameter on the two loudness models. It was decided to analyse STL and SL separately because the main interest was to inspect the trend in error for each feature individually and draw comparisons between the two models.

The results of the factorial ANOVAs are supported by graphical representations in which the arithmetic mean of the nRMSEs (as used as input data to the ANOVAs), with 95% bootstrapped within-stimulus CIs are plotted as a function of each parameter.[22] Again, error bars were constructed to reflect the structure of the two-way ANOVAs, e.g. following an 8 (hop size) × 2 (loudness model) repeated-measures design.

ANOVA significance level and follow up tests

A significance level of 0.05 was used for all tests. Any two-way interactions were followed up using simple main effect analysis where the levels of one experimental factor are compared at each level of the other factor. Yuend's 20% trimmed means paired test (Wilcox 2012) was used to conduct the post-hoc analyses, using the Bonferroni correction for multiple comparisons. For example, given an $M(\text{factor A}) \times N(\text{factor B})$ design, α would be reduced by a factor of $(N \times M \times (M-1)/2)$ for the comparisons made in factor A, and $M \times N \times (N-1)/2)$ for the comparisons made in factor B.

The hypothesis that interpolating EPs improves the estimates of the auditory features was tested by comparing the errors produced by each model with and without EP interpolation at each filter spacing. Based on Figure 4.4, it was anticipated that, for *both* models, the interpolation of EPs would yield lower STL error for filter spacings up to a limit, before contributing no further improvement. In addition, it is interesting to know if larger filter spacings with interpolation outperform lower filter spacings without interpolation. For example, Glasberg and Moore (2002) use a 0.25 Cam spacing rather than 0.1 Cams as used for the stationary model (ANSI S3.4 2007), but the two configurations might be comparable after incorporating interpolation at the excitation transformation stage. These questions were addressed using a two-way ART ANOVA on each model-feature combination to identify for a main effect or interaction between hop size and interpolation (the two factors). When necessary, a series of dependent post-hoc tests were used to compare the errors at each filter spacing (with no interpolation) with those resulting after interpolation at all filter spacings equal to and above the reference. For example, the error at 0.5 Cams was compared with all errors at 0.5 Cams and above with interpolation applied. With a total of 16 filter spacings, this gives $16 \times 15/2 + 16 = 136$ comparisons per analysis (a subset of all possible comparisons). Thus, the significance level $\alpha = 0.05$ was reduced to 0.05/136 for each of the four ANOVAs.

4.5.5 Results

Hop size

Figure 4.18 shows the mean time series error as a function of hop size for the STL and the SL. The solid black lines show the average errors across stimuli for the GM02 and the red dotted lines show those for the CH12. The effect of hop size on both STL and SL is clear; increasing the hop

[22] The bootstrapped distributions of the errors (in dB) tended to be normal.

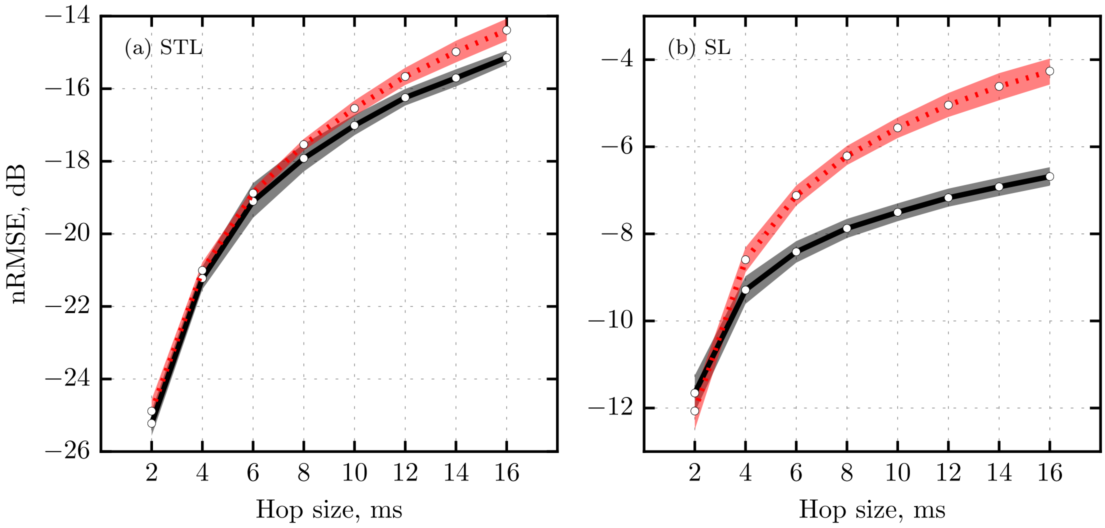


Figure 4.18: nRMSE plotted as a function of hop size for (a) the STL and (b) the SL. The data points show the arithmetic means of the stimulus errors (in dB) and shaded areas surrounding the lines represent the 95% within-stimulus CI of each estimate for the GM02 (black solid lines) and CH12 (red dotted lines).

size beyond the reference value of 1 ms widens the time interval between the calculation of power spectra, leading to greater error in the approximation of auditory features. For steady-state sounds such as pure tones and stationary noise, the expected errors would be minimal, especially for the STL, but evidently the processing rate has a clear effect on typical time-varying sounds such as music and speech. As expected, the error functions all increase monotonically with hop size. The STL errors are much lower than corresponding SL errors, which is attributed to the time-frequency smoothing of the SL patterns.

The effect of hop size on STL and SL features was tested using two separate two-way ART ANOVAs with factors hop size and loudness model. For the STL, the main effect of hop size was found to be statistically significant ($F(7, 203) = 318.4$, $p < 0.001$) but model was not ($F(1, 29) = 3.6$, $p = 0.067$). The hop size × model interaction was statistically significant ($F(7, 203) = 2.1$, $p = 0.044$). These findings are in line with Figure 4.18(a): the trend of increasing error with hop size is similar for both models although the difference between the two error functions is greater at larger hop sizes. The non-significant main effect of model indicates that no reliable difference detected in STL error between the two models was found when the data are collapsed over all hop sizes.

Paired Yuen's tests on trimmed means were performed to compare the average STL errors between the models at each hop size. No reliable difference in mean error was found between the two models at any of the hop sizes tested ($p > 0.006$; Bonferroni correction for eight tests). Paired tests were then conducted on the STL error functions within each model. For the GM02, all comparisons except for 8–10 ms and 12–14 ms were statistically significant ($p < 0.0009$; Bonferroni correction for 56 tests). For the CH12, all pairwise comparisons were found to be statistically significant ($p < 0.0009$).

The mean STL error at 2 ms was -25.2 dB (between-stimulus 95% CI = [-26.1, -24.3]) for the GM02 and -24.9 dB (between-stimulus 95% CI = [-25.9, -23.7]) for the CH12. Thus, the RMSE

[23] This is not the case when outer and middle-ear filtering is performed in the time domain according to Glasberg and Moore (2002), but spectral weighting is used here.

introduced in the STL was approximately 0.32% of the average STL value. Doubling the hop size halves the number of processing frames, lowering the execution time by a factor of approximately two.[23] Knowing that a doubling of hop size increases error systematically, the size of this effect is worthy of discussion. Table 4.13 shows the difference in trimmed means of the STL errors. In general, doubling the hop size increases the STL error for both models by about 2.3–4.5 dB, but the change in error appears to decrease at larger hop sizes.

Pair	GM02		CH12	
(ms)	Δ	95% CI	Δ	95% CI
4–2	3.97	3.37–4.48	3.80	3.63–3.99
8–4	3.24	2.72–3.76	3.52	3.18–3.86
16–8	2.63	2.26–3.01	3.02	2.69–3.35

Table 4.13: Trimmed mean difference in STL nRMSE (dB) between two-fold related hop sizes for both models. All differences were statistically significant ($p < 0.001$; Bonferroni corrected for 56 tests.).

The above procedure (two ANOVAs followed by simple main effect analysis) was repeated for the SL feature. The ANOVA revealed that both main effects and their interaction were statistically significant (hop size: $F(7, 203) = 197.6$, $p < 0.001$; model: $F(1, 29) = 52.8$, $p < 0.001$; hop size × model: $F(7, 203) = 67.3$, $p < 0.001$). Again, the detected interaction is in agreement with the Figure 4.18(b): the SL errors systematically increase with hop size but vary in a model-specific manner. Paired Yuend's tests on trimmed means showed that the GM02 and CH12 errors were not reliably different at 2 ms (trimmed mean difference = 0.37, 95% CI = $[-0.14, 0.89]$, $p = 0.148$) but were systematically different at all other hop sizes tested ($p < 0.006$; Bonferroni correction for eight tests). For the GM02, all pairwise comparisons between the different hop sizes were found to be reliably different ($p < 0.0009$; Bonferroni correction for 56 tests), except for pairs 12–14 ms (95% CI = $[-0.35, -0.06]$, $p = 0.01$) and 14–16 ms (95% CI = $[-0.38, -0.06]$, $p = 0.011$). In contrast, all pairwise comparisons were found to be reliably different for the CH12 ($p < 0.0009$).

Filter spacing

Figure 4.19 shows the STL and SL errors for both loudness models running without EP interpolation. The SL errors are markedly larger than the STL errors, which occurs because the STL sums loudness over the critical band filters before being further smoothed through temporal integration. The error in the intermediate auditory patterns is effectively averaged out. Both subplots suggest that the number of critical band filters is more important for the CH12 compared to the GM02. Lower filter spacings (0.25–0.75 Cams) appear to impact the STL estimates to a greater extent for the CH12.

The ANOVA for the STL showed that main effects of filter spacing ($F(15, 435) = 493.3$, $p < 0.001$) and model ($F(1, 29) = 696.6$, $p < 0.001$) were statistically significant, as was the interaction of the two factors ($F(15, 435) = 65.3$, $p < 0.001$). The main effect of model refers to the STL errors in the CH12 being greater than the STL errors in the GM02, on average. The exception to this generalisation occurs at lower filter spacings. Paired Yuend's tests on trimmed means showed that the STL errors for the CH12 were reliably higher than those for the GM02 ($p < 0.003$; Bonferroni corrected for 16 tests) at all filter spacings except at 0.25 Cams (trimmed mean difference = -1.87, 95% CI = $[-3.49, -0.26]$, $p = 0.025$) and 0.5 Cams (trimmed mean difference = 2.82, 95% CI = $[0.55, 5.10]$, $p = 0.018$). The CI for the trimmed mean error at 0.25 Cams is however more suggestive of the CH12 having lower error than the GM02, in agreement with the data shown

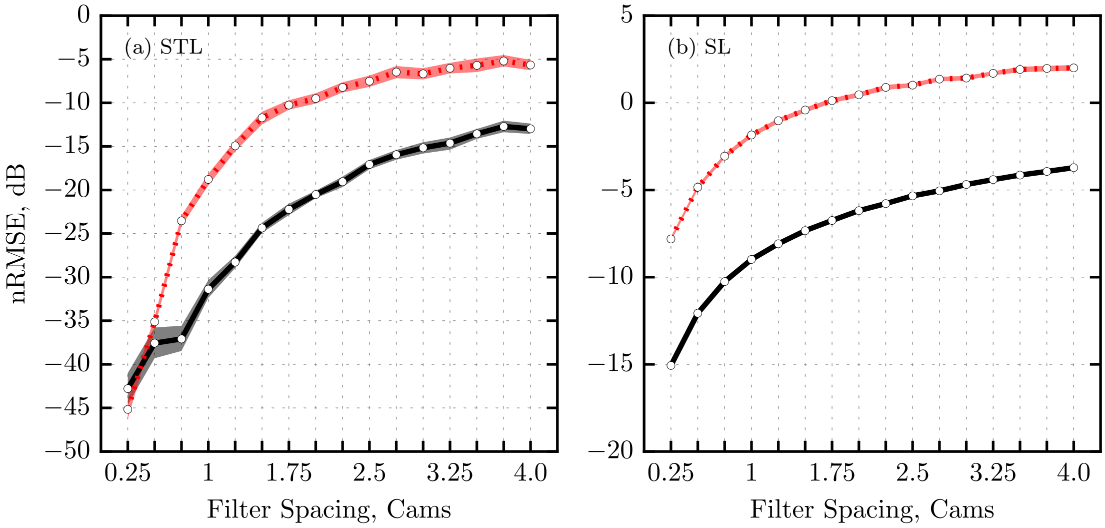


Figure 4.19: nRMSE plotted as a function of filter spacing for (a) the STL and (b) the SL. The data points show the arithmetic means of the stimulus errors (in dB) and shaded areas surrounding the lines represent the 95% within-stimulus CI of each estimate for the GM02 (black solid lines) and CH12 (red dotted lines).

in subplot (a). A series of paired tests were then conducted to identify differences between the errors within each model (Bonferroni corrected for 240 tests). For both models, all statistically significant trimmed mean differences between the upper and lower filter spacings were positive, indicating that the most likely effect of filter spacing is to increase the error, as expected. For the GM02, no systematic difference was detected between the trimmed mean error at 0.75 Cams and 0.5 Cams (trimmed mean difference = -0.11 95% CI = $[-2.03, 1.82]$, $p = 0.91$), but this increment in filter spacing did have a systematic and larger impact on the CH12 (trimmed mean difference = 11.6 95% CI = [10.72, 12.45], $p < 0.0002$).

The two-way ANOVA for the SL also revealed statistically significant main effects (filter spacing: $F(15, 435) = 455.9$, $p < 0.001$; model $F(1, 29) = 2329.9$, $p < 0.001$) and a significant interaction ($F(15, 435) = 62.3$, $p < 0.001$). As with spectral compression, the main effect of model can be interpreted here in that the CH12 errors were all higher than those measured for the GM02. This was confirmed with a series of Yuend's tests comparing the trimmed errors at each filter spacing ($p < 0.003$). The interaction says that the difference in average error between the two models is dependent on the number of auditory filters. Again, this is apparent from comparing the two lines shown in the Figure 4.19(b). As with the STL comparisons, detected statistically significant differences in the trimmed means within each model were based on positive differences in average error measured between the upper and lower filter intervals.

The hypothesis that interpolating EPs improves the approximation of SL and integrated loudness was tested for each model using a series of one-way ART ANOVAs followed by planned comparisons. An ANOVA was run on each model-feature combination, yielding four separate analyses with factors filter spacing (16 levels) and EP interpolation (two levels). The results of these four omnibus tests are listed in Table 4.14. It can be seen that the main effects and their interaction are statistically significant for both features within each model. This indicates that EP interpolation influences both SL and STL features, and that this effect varies with filter spacing.

Yuend's trimmed mean paired two-tailed t-test was used to investigate the planned comparisons given in Section 4.5.4. Figure 4.20 addresses the question 'does EP interpolation improve the

Model	Feature	Filter spacing	Interpolation	Interaction
GM02	STL	$F(15, 928) = 648.4$	$F(1, 928) = 11.9$	$F(15, 928) = 11.9$
	SL	$F(15, 928) = 375.6$	$F(1, 928) = 2596.12$	$F(15, 928) = 153.0$
CH12	STL	$F(15, 928) = 286.4$	$F(1, 928) = 41.1$	$F(15, 928) = 8.0$
	SL	$F(15, 928) = 165.7$	$F(1, 928) = 709.3$	$F(15, 928) = 68.2$

Table 4.14: Main effects of filter spacing and interpolation and their interaction for each combination of loudness model and feature. Each omnibus test was found to be statistically significant ($p < 0.001$).

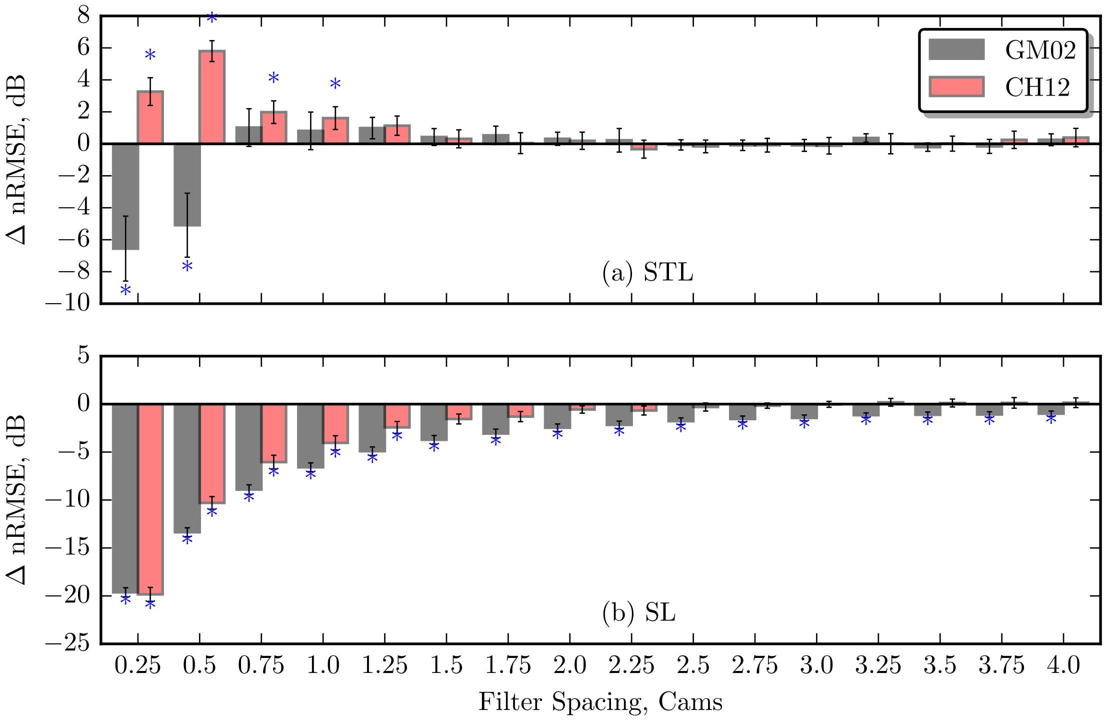


Figure 4.20: Change in trimmed mean nRMSE error for (a) the STL and (b) the SL after applying EP interpolation. A negative difference indicates reduced error. Error bars show the 95% between-stimulus CI of each estimate. An asterisk indicates that the trimmed mean difference is reliably different from zero ($p < 0.05$).

approximation of loudness features at each filter spacing?' Each plot shows the change in the trimmed mean feature errors at each filter spacing as a result of EP interpolation; a negative change in nRMSE denotes lower error. It can be seen that interpolation improves the estimation of the SL at all filter spacings for the GM02 (grey bars) and that this improvement is reliable on statistical grounds. The improvement carries over to the STL domain, but is only statistically significant at filter spacings of 0.25 Cams and 0.5 Cams. Note that the improvement appears to diminish with increasing filter spacing, which explains the interaction term picked up by the ANOVA. The most interesting and unexpected finding is that for the CH12, interpolation improves the SL estimates (reliably up to 1.75 Cams), but degrades the STL measures. Interpolation did not reliably reduce the STL error compared to lower filter spacings without interpolation (for either model). For the GM02, cubic interpolation improved the SL estimation for every two successive filter spacings up to 2.75 Cams, after which the improvement was found to be reliably lower for every consecutive filter spacing. For example, both 0.5 and 1 Cam spacings with interpolation lead to lower error compared to at 0.25 Cams with no interpolation, whereas only 3.25 Cams with interpolation outperformed 3.0 Cams with none. Interpolation improved the SL estimates of the CH12 for each successive spacing up to 1 Cam. These effects are further discussed in Section 4.5.6.

Spectral Compression

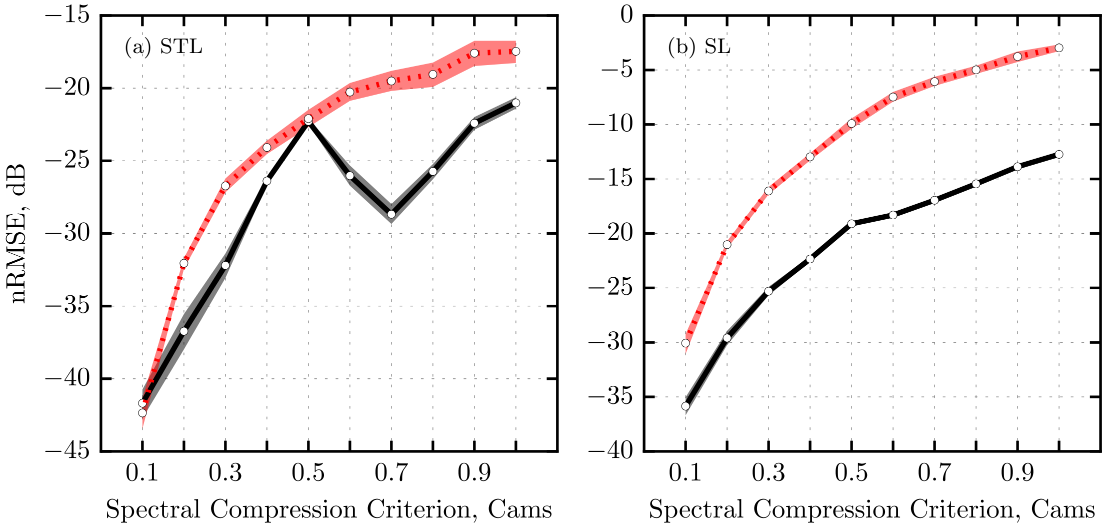


Figure 4.21: nRMSE plotted as a function of spectral compression criterion for (a) the STL and (b) the SL. The data points show the arithmetic means of the stimulus errors (in dB) and shaded areas surrounding the lines represent the 95% within-stimulus CI of each estimate for the GM02 (black solid lines) and CH12 (red dotted lines).

The STL and SL error functions for spectral compression criterion are plotted in Figure 4.21 for the GM02 (black lines) and CH12 (red dotted lines). For both models, spectral compression distorts the loudness patterns and subsequently introduces error in the frequency and time integrated measurements. This is to be expected, given that the input components are summed into sub-critical-bands prior to auditory filtering. This process alters the distribution of the power spectra fed to the EP transformation stage. The CH12 errors are generally larger than those of the GM02, especially in the SL domain. Additionally, initial increments of compression criterion appear to yield greater change in error for the CH12, indicating that this model is more sensitive to spectral modifications. Summing the erroneous SL pattern leads to a non-monotonic STL error function for the GM02; the error function shows a peak at 0.5 Cams and continues to increase above 0.7 Cams. This trend is not observed for the CH12 over the ten criteria tested.

Two ANOVAs were performed for the STL and SL errors with factors compression criterion and loudness model. The ANOVA for the STL revealed that both spectral compression ($F(9, 261) = 395.6$, $p < 0.001$) and model ($F(1, 29) = 156.7$, $p < 0.001$) main effects were statistically significant, as was their interaction ($F(9, 261) = 62.8$, $p < 0.001$). Paired Yuend's tests on trimmed means confirmed that the GM02 and CH12 errors were not reliably different at 0.1 Cams (trimmed mean difference = -0.44 dB 95% CI = $[-0.57, 1.46]$, $p = 0.372$) or at 0.5 Cams (trimmed mean difference = -0.61 dB 95% CI = $[-1.32, 0.11]$, $p = 0.091$) but were systematically different at all other criteria ($p < 0.005$; Bonferroni corrected for 10 tests). The CH12 errors were all higher than the GM02 errors in these cases. The simple main effect analysis of spectral compression within each model confirmed the observed non-monotonic form of the error function for the GM02. For example, increasing the amount of compression up to 0.5 Cams systematically increased the trimmed mean error ($p < 0.0005$, Bonferroni corrected for 90 tests), but further compression up to 0.8 Cams systematically decreased the mean error relative to 0.5 Cams. Furthermore, compression settings of 0.8, 0.9 and 1 Cam were all reliably higher than the error at 0.7 Cams. This effect was not observed for the CH12; all statistically significant differences were based on increasing in error with the amount of compression.

The ANOVA for SL also showed statistically significant main effects (compression criterion: $F(9, 261) = 2114.5$, $p < 0.001$; model: $F(1, 29) = 663.7$, $p < 0.001$) and a significant interaction of the two ($F(9, 261) = 44.6$, $p < 0.001$). Since both interaction terms were significant for the STL and SL ANOVAs, the effect of spectral compression on both features is evidently model dependent. These findings are in accordance with the different trends shown in Figure 4.21. The main effect of spectral compression *can* be interpreted here because the CH12-SL errors are greater than the GM02-SL errors at all parameter settings, although the magnitude of these differences vary with the amount of compression. As expected, paired Yuend's tests on trimmed means revealed that the GM02 and CH12 errors were reliably different at all compression criteria tested ($p < 0.005$; Bonferroni corrected for ten tests). Pairwise comparisons of the errors across compression criteria within each model were all reliably different ($p < 0.0005$, Bonferroni corrected for 90 tests). The significant two-way interaction therefore suggests that the rate of (increasing) change in SL error functions is different across the two models, with the CH12 being more sensitive to this parameter.

Targeting the bottleneck

The effect of hop size R, filter spacing α and spectral compression β on STL and SL features have been presented for two loudness models. When configured with the reference parameter sets (see Table 4.12), the models did not execute in real time. Although computational speed can be increased approximately two-fold by doubling the hop size, Table 4.11 shows that the bottleneck resides in the excitation transformation. Additionally, the findings thus far indicate that the initial increments in hop size markedly affect the feature approximations when compared to a range of α and β values. For example, the 95% between-stimulus CI for the trimmed mean difference in STL error between the hop size approximation $R = 2$ ms and the filter spacing approximation $\alpha = 0.25$ Cams was $[15.3, 19.5]$ dB for the GM02, and $[18.0, 20.9]$ dB for the CH12. Similarly, the 95% CIs for the trimmed mean difference in error between the $R = 2$ ms approximation and the spectral compression approximation $\beta = 0.1$ Cam was $[15.3, 17.5]$ dB (GM02) and $[15.3, 18.8]$ dB (CH12).

By combining spectral compression with filter spacings beyond 0.1 Cams, the bottleneck can be alleviated, allowing a significant reduction in execution time to be achieved without having to resort to large hop sizes. To demonstrate this, an experiment was conducted in which the STL errors were measured at all permutations of α (no interpolation) and β, with $R = 1$ ms. A threshold of -25.2 dB and -24.8 dB was set for the GM02 and CH12, respectively, as determined using a hop size of 2 ms. For each model, the error surface generated by different combinations of α and β (with $R = 1$ ms) was searched for all errors falling below this threshold. From this subset, the combination of parameters that maximised the CR at the excitation transformation stage, as defined by Equation 4.11, was selected.[24]

Table 4.15 summarises the performance of the models operating at a hop size of 1 ms with the optimal parameters according to the above procedure. When $f_s = 44100$, the number of spectral components K used by the reference configuration was 1393 and 2971 for the GM02 and CH12, respectively. The number of auditory filters A was 372 (GM02) and 388 (CH12). It can be seen that the selected parameter sets lead to a considerable reduction in these two quantities: the total

[24] This procedure was also conducted using trimmed means but this had no impact on the resulting parameters.

Model	α	$\hat{A}$	β	$\hat{K}$	$\max \Delta L$ (dB)
GM02	1.25	30 (92%)	0.7	63 (95 %)	0.08
CH12	0.5	78 (80%)	0.3	148 (95 %)	0.11

Table 4.15: Number of auditory filters $\hat{A}$ (and percentage reduction) and frequency components $\hat{K}$ obtained with the optimal combination of filter spacing α and spectral compression criterion β for a hop size $R = 1$ ms. The maximum global-loudness descriptor error is also given.

CR was 99% for both models. Indicative of worst-case performance, the final column shows the maximum absolute level change required for equal loudness between the predictions of the reference and approximation across all stimuli and all three global loudness descriptors (Ψ_{μ}, Ψ_{pk} and Ψ_{95}). This gives some insight into the size of the effects when the models are used for predicting the overall loudness of dynamic sounds.

For the GM02 with the selected parameters, the mean STL error was 0.78 dB (95% CI = [-0.49, 1.97]) lower than that measured using the reference model with $R = 2$ ms. For the CH12, the mean STL error was 2.15 dB (95% CI = [0.91, 1.97]) lower. Although both mean differences were positive, only that for the CH12 was found to be statistically significant (CH12: $t(29) = 3.37$, $p = 0.002$; GM02: $t(29) = 1.24$, $p = 0.23$).[25] Although the mean error for the GM02 with the new parameter set was not reliably different, the CI is asymmetrical, tending towards positive values.

Process	GM02	CH12
FFT (Frigo and Johnson 2005)	1.4×10^7 (73.7)	3.4×10^7 (78.9)
Excitation transformation	1.6×10^6 (8.1)	1.6×10^6 (3.8)
Binaural inhibition	1.1×10^5 (0.6)	4.7×10^5 (1.1)

Table 4.16: Program profile report with the optimal parameters.

The profile reports associated with the optimal parameters are shown in Table 4.16. It can be seen that the processing time consumed by the excitation transformation has been significantly reduced (by two orders of magnitude for the GM02) and consequently, the bottleneck has shifted to the computation of the multi-resolution DFT. The mean stimulus speedup factor (CPU time of the reference divided by the CPU time of the approximation averaged across ten executions per stimulus) was 16.09 (SD = 0.01) for the GM02 and 4.55 (SD = 0.02) for the CH12. This demonstrates that large speedups can be achieved before having to resort to incrementing the hop size independently. It should be noted that these parameters were selected using the STL error at $R = 2$ ms as target. Running the same procedure using the LTL yields more conservative α and β values because this feature is less influenced by hop size. When the LTL feature was used to establish a threshold, the parameter sets were $\{\alpha = 1, \beta = 0.3\}$ for the GM02, and $\{\alpha = 0.5, \beta = 0.2\}$ for the CH12.

Finally, the mean STL nRMSE and maximum global-loudness descriptor errors across all sounds are given in Table 4.17 for the two models configured with the parameter values listed in Table 4.15. The models were evaluated at four hop sizes and worst-case speedup factors across all program executions and all stimuli are listed. The row in bold indicates that the CH12 did not achieve real-time performance on the computer used in this study at the reference hop size of 1 ms.

Generalisation and effect of level

A few questions remain:

[25] The differences of the two distributions were found to be normally distributed so a parametric paired t-test was

R (ms)	max ΔL (dB)			nRMSE (dB)		Speedup
	Ψ_{μ}	Ψ_{pk}	Ψ_{95}	μ	95% CI	
1	0.05	0.08	0.08	-26.0	[-26.8, -25.2]	15.4
	0.06	**0.03**	**0.11**	**-27.0**	**[-27.8, -26.4]**	**4.4**
2	0.05	0.07	0.10	-23.5	[-24.3, -22.8]	30.5
	0.06	0.03	0.11	-23.7	[-24.5, -22.8]	8.7
4	0.12	0.81	0.22	-20.8	[-21.7, -19.8]	63.2
	0.08	0.21	0.20	-20.7	[-21.6, -19.7]	17.7
8	0.29	2.57	0.24	-17.9	[-18.9, -16.9]	124.8
	0.12	1.54	0.42	-17.6	[-18.5, -16.3]	34.4

Table 4.17: Performance of the optimised models at four hop sizes R. From left to right: maximum global-loudness descriptor error (mean LTL, peak STL, and 95$^{\text{th}}$ percentile of STL), mean STL nRMSE (with 95% between-stimulus CI), and minimum speedup factor across all stimuli. The first and second row within each hop size cell correspond to the GM02 and CH12, respectively. The row in bold indicates the only parameter set where real-time processing was not achieved.

1. Do the approximations perform consistently when applied to other musical sounds?

2. Do the approximations perform consistently at different reproduction levels?

To share some insight into these questions, the performance of the two loudness models, configured using the parameters given in Table 4.15, was evaluated using a 30 new six second sound segments. These segments were taken from a larger corpus specifically collated for loudness-matching experiments involving multitrack audio (Ward and Reiss 2016). The selected samples involved a roughly equal mix of percussion, electric/acoustic guitar, bass guitar, piano and bowed stringed instruments, and synthesized sounds. As before, the reference models were configured using a filter spacing of 0.1 Cams with no spectral compression. A 1 ms hop size was used for all models. To test for any systematic influence of sound level, the time series and global-loudness descriptor errors were measured with the RMS level of each sound segment set to 55, 65, 75, 85, and 95 dB SPL.

The results are shown in Figure 4.22. Subplot (a) shows the mean STL nRMSE (and 95% within-stimulus CIs) as a function of stimulus level, for the two algorithms. To assess the generalisation of the approximations, a comparison between the trimmed means of the stimulus errors at 75 dB SPL measured on the new dataset, and those measured on the previous SQAM dataset was conducted. A positive difference indicates that the approximations performed worse on the new data. The trimmed mean difference was 0.85 dB 95% CI = [2.0, -0.3] (Cohen's d = 0.41) for the GM02, and -0.71 dB 95% CI = [0.5, -1.9] (Cohen's d = -0.35). Thus, the GM02 performs worse on the new data with a medium effect size, whereas the CH12 performs slightly better. The top subplot shows, however, that the errors are level dependent. For the GM02, the error in the STL time series is consistent over medium levels, but increases as the sound level exceeds 75 dB SPL. The opposite effect is observed for the CH12: the error is stable over 55–75 dB SPL, but *decreases* as the stimulus intensity approaches the highest level of 95 dB SPL, where the average error reaches -32 dB. In all cases the GM02 errors are higher than those measured using the CH12. These observations were confirmed by a two-way ART ANOVA, which revealed a statistically significant effect of model ($F(1, 29) = 161.8$, $p = 0$), no reliable main effect of level ($F(4, 116) = 161.8$, $p = 0.124$), and a statistically significant interaction of these two factors ($F(4, 116) = 107$, $p = 0$). Post-hoc tests on trimmed means showed that the error in the GM02 approximation were reliably higher than of the CH12 at all stimulus levels ($p < 0.01$; Bonferroni correction for five tests). For the GM02, no reliable differences were found between the trimmed mean errors over 55–75 dB SPL, but

used (no trimming).

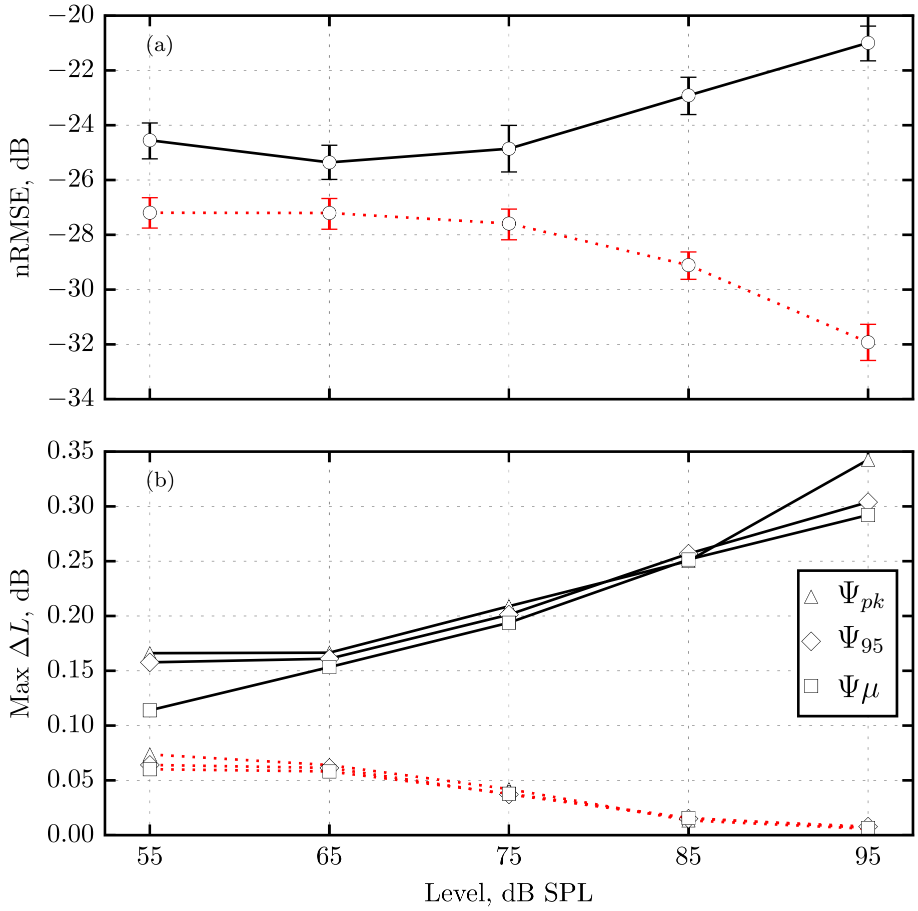


Figure 4.22: Performance of the GM02 (black solid lines) and CH12 (red dotted lines) with the parameters listed in Table 4.15 applied to the multitrack audio. (a) Mean STL nRMSE, with 95% within-stimulus CI, and (b) maximum global-loudness descriptor error, both plotted as a function of stimulus level.

all comparisons with those at 85 and 95 dB were statistically significant ($p < 0.0025$; Bonferroni corrected for 20 tests). All but two pair-wise comparisons (55 vs 65 dB SPL and 55 vs 75 dB SPL) performed on the CH12 errors were found to be statistically significant.

Subplot (b) shows, for each reproduction level, the maximum error in the three global-loudness descriptors. These worst-case discrepancies also reflect the trend in STL nRMSE. The level differences are less than 0.1 dB for the CH12 across all levels and stimuli. In contrast, those for the GM02 vary from about 0.15 dB at 55 dB SPL, to a maximum of 0.35 dB at 95 dB SPL. Half of the stimuli input to the GM02 had peak STL absolute errors less than 0.1 dB over 55–85 dB SPL, and less than 0.2 dB at 95 dB SPL. The median at each level were all below 0.02 dB for the CH12.

4.5.6 Discussion

Parameters

Two dynamic loudness models have been parameterised by hop size, filter spacing and spectral compression. Increasing the hop size beyond 1 ms reduces the computation load by discarding analysis frames and, as a result, introduces Error into the perceptual features—the STL and the SL. The effect of hop size on STL error was comparable in both models, although the CH12 appeared to be more sensitive compared to the GM02, especially in the SL domain where the errors were reliably higher for hop sizes of 4 ms and above. This can be attributed to differences in the auditory patterns generated by the two procedures, as those output by the GM02 show a broader profile compared to those output by the CH12 (refer to later discussion). The SL patterns of the GM02 are likely more positively correlated over time, yielding lower error when previous calculations are used to predict the patterns of successive frames. Differences in STL error were found to be less marked compared to those in the SL domain, which is explained by spectral and temporal loudness summation. As shown in Table 4.17, hop sizes exceeding 4 ms can be detrimental when estimating the loudness of short-duration sounds, which rely on accurate estimates of peak loudness (Glasberg and Moore 2002). Zwicker (1977) also highlighted the importance of maximum loudness for quantifying the perceived loudness of impulse and speech sounds. Indeed, of all the stimuli used in the primary tests, the castanets segment occurred most frequently in the top 5% of the STL error distributions across all hop sizes tested in both models. Compared to the peak STL descriptor, the mean LTL and percentile STL descriptors were within 0.5 dB of the reference loudness estimates for hop sizes up to 8 ms.

The choice of filter spacing determines the accuracy of the excitation and SL patterns. Too few filters per critical band can be severe for tonal signals where peak excitations that dominate the IL are less likely to be captured. Large intervals can also have a marked effect on relatively broadband sounds in that the area under the SL pattern is estimated from an under-sampled EP. The trends in error displayed in Figure 4.19 indicate that, compared to the GM02, the CH12 is more sensitive to changes in filter spacing and that the majority of deviations from the target data were more extreme. This can be attributed to differences in DFT frequency resolution and filter bank architecture. For the CH12, the DFT mainlobes are half the bandwidth of those used by the GM02, and so filter spacing plays a greater role in preserving frequency-localised energy. The SL patterns output by the two reference models were compared across stimuli and it was observed that those generated by the CH12 had markedly larger peak-to-valley ratios compared to those produced by the GM02. This is demonstrated by Figure 4.23 which shows the SL pattern taken from a single analysis frame of a bass clarinet recording (track 17 of the SQAM corpus). It can be seen that the SL patterns generated by the GM02 are more broadband and so larger filter spacings are less detrimental to SL and STL features. In contrast, the EPs produced by the CH12 must be sampled more often to accurately capture peak excitations. Considering that the RMSE metric emphasises large deviations, it is of no surprise that the SL errors measured on the CH12 were notably high.

When cubic interpolation was applied at the excitation transformation stage, a systematic decrease in SL error was observed for both models. It should be noted that these discrepancies were measured with respect to nearest-neighbour interpolation applied to the SL patterns. This reduction in error only carried over to the STL for the GM02 at filter spacings of 0.25 and 0.5 Cams, where a systematic reduction in error was measured. However, empirical investigation suggests that these improvements can be considered negligible in terms of STL. Figure 4.24 shows the STL and LTL output by the GM02 for a portion of track 11 of the SQAM database (double bass). This segment was chosen from the corpus because its STL error was largest at 0.5 Cams and showed a large reduction in nRMSE when the sampled EPs were interpolated. In each plot, the black lines show the STL and LTL time series generated by the GM02 with a filter spacing of 0.1 Cams (the reference configuration). The red dotted lines correspond to the loudness-time functions estimated using filter spacing of 0.5 Cams (left plot) and 0.5 Cams with interpolation (right plot). At this time scale, the discrepancies between the approximation and reference time series are visible in the left plot, but are around only 0.1 phons at most. Although these discrepancies are less apparent

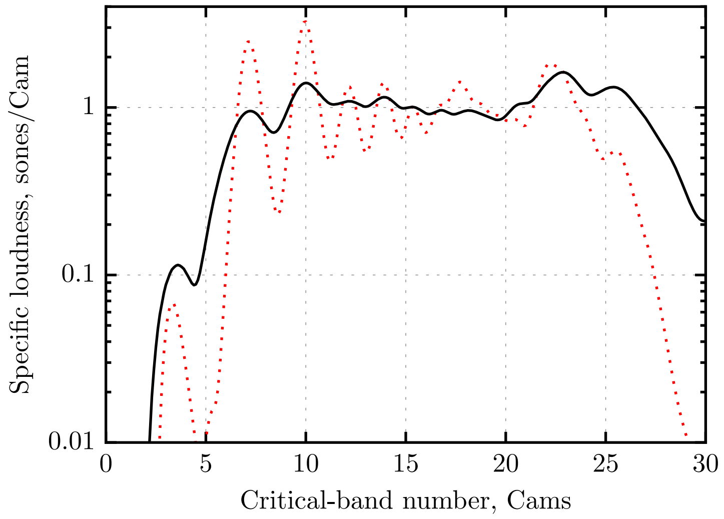


Figure 4.23: SL pattern of a single analysis frame of track 17 (bass clarinet) of the SQAM database calculated using the GM02 (solid black line) and CH12 (dotted red line) with the reference parameter sets.

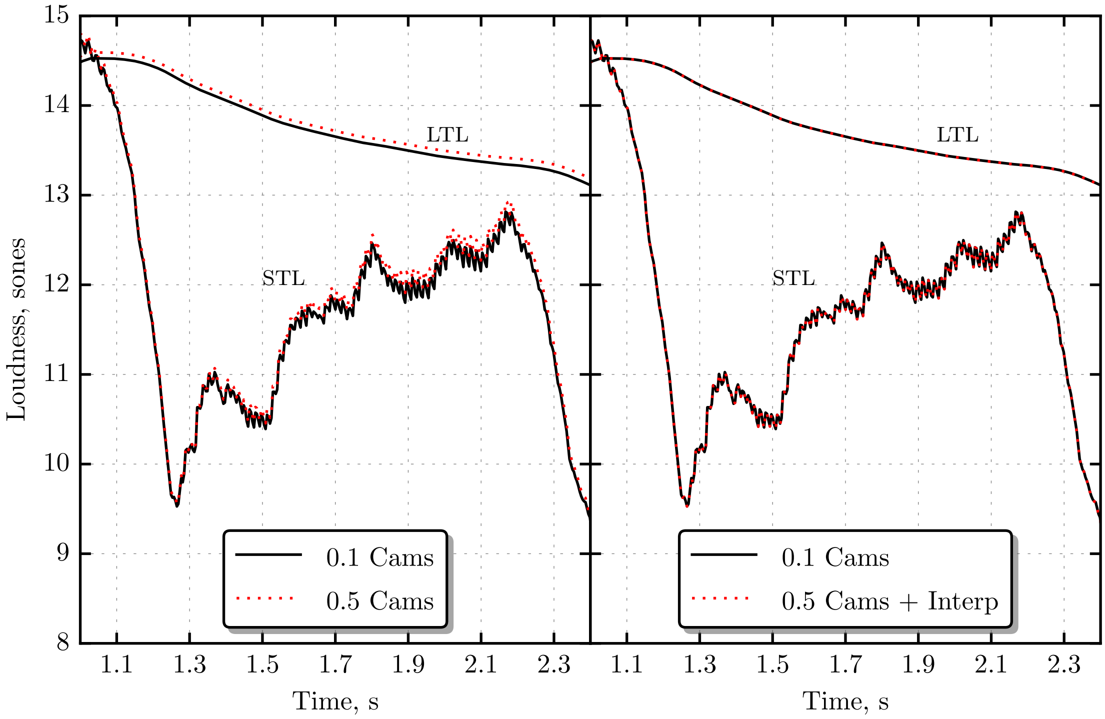


Figure 4.24: STL and LTL measured using the GM02 for different excitation transformation configurations.

when EP interpolation is used, the three global-loudness descriptors measured without it (left plot) were all within 0.1 dB of the reference values.

For the CH12, a systematic *increase* in STL error was observed when interpolating EPs. The cause for this discrepancy is that interpolation is applied in the log domain which biasses the SL. That is, the SL error $e = \psi' - \hat{\psi}'$ was not centred on zero. Thus, when the SL is summed across frequency, an offset or bias is introduced into the frequency-integrated IL. This bias is more extreme for the CH12 than for the GM02, due to differences in filter bank architecture and the second SL transformation stage of the GM02 which is not a feature of the CH12. To conclude, auditory patterns may be interpolated in isolation of total loudness calculation, for the purpose of enhancing their display, or when a given application requires frequency-domain summary statistics or masking calculations. If integrated loudness is of primary interest, as in most situations, then EP interpolation appears to be of little practical benefit to either model. More sophisticated integration algorithms for estimating the area under the SL pattern are likely to improve the total loudness estimate at a lower computation cost compared to the interpolation scheme proposed here.

Finally, a compressed spectrum was utilised to reduce the number of frequency components whilst attempting to maintain sufficient energy in the auditory filters to preserve the loudness density. The grouping of component intensities does introduce error in the auditory patterns, the extent of which is dependent on the bandwidth used to average the spectrum. The GM02 requires a wider bandwidth to achieve a similar percentage reduction in the number of components as the CH12. For example, when $\beta = 0.1$ Cams the compression criterion is satisfied above 264 Hz for the CH12 but not until 764 Hz for the GM02. This is because the GM02 uses half the frequency resolution as used by the CH12. In combination with knowledge that most musical sounds are predominantly in the low-mid frequency region, this explains why the error functions derived using CH12, as shown in Figure 4.21, are generally higher than those measured using the GM02. The non-monotonic STL error function measured on the GM02 was expected based on the initial tests conducted earlier in this chapter where a similar trend in error was observed for the IL time series but not in the EP domain. However, of the three parameters examined in this study, spectral compression is the only one that does not discard information, but instead redistributes it; the total power input to the excitation transformation is preserved but the distribution of energy across frequency is altered. The shape of the auditory patterns is highly influenced by this redistribution of spectra, but the error in the total IL depends solely on the area under those patterns.

Faster performance

In conjunction with Figures 4.21 and 4.19, the parameter sets listed in Table 4.15 provide a good guideline for speeding up the models based on the initial bottleneck. Both models see a significant reduction in execution time by combining spectral compression with filter spacings above 0.1 Cams. Further speedup factors of ~ 2 can be obtained by doubling the hop size at the cost of roughly two-fold increase in STL error (see Table 4.17). The results in Table 4.17 indicate that for a hop size of 2 ms, combined with the selected α and β values, both models can achieve real-time performance with estimates of global loudness preserved to within 0.2 dB of the reference predictions for a range of music and speech. Considering that discrimination thresholds for intensity can be as high as 0.2 dB for pure tones and 0.5 dB for most broadband noises (Fastl and Zwicker 2007), it is unlikely that there will be noticeable differences between *identical* stimuli aligned in loudness by the reference and proposed implementation. More importantly, Soulodre et al. (2003) found the average reproducibility of subjective relative loudness judgements involving *different* programme material to be 1.24 dB. This suggests that greater error is tolerable when the faster implementation is used for balancing the loudness of typical audio content.

Burdiel et al. (2012) showed that, based on STL errors of the GM02, a filter spacing of 1 Cam could be used to establish real-time performance on their test machine. In the current study, a finer range of parameter values have been explored and an algorithm for reducing the number of spectral components has been integrated to facilitate greater computational savings. To demonstrate the added benefit, the average STL nRMSE measured using a 1 Cam filter spacing was compared to the average error measured using all combinations of spectral compression criteria and filter spacings of 0.25, 0.5 and 0.75 Cams. The STL error was found to be reliably lower for the combination $\{\alpha = 0.75, \beta = 0.2\}$ ($t(29) = 4.4, p < 0.0017$; Bonferroni corrected for 30 tests[26]). This parameter set yields a CR of 98.0%. In contrast, a 1 Cam filter spacing gives a CR of 89.8%. Therefore, the spectral compression algorithm can be combined with filter spacing to provide superior performance than approximations based on filter spacing alone. Furthermore, the real-time proposal of Burdiel et al. (2012) included a frequency-domain weighting function to replace the FIR filter originally used by Glasberg and Moore (2002). However, it was shown in Section 4.3 that spectral weighting is problematic at low frequencies, mainly in terms of absolute threshold predictions, and hence why the implementation here utilises a highpass filter to improve the response. The implementation of the GM02 presented here is thus robust to different types of stimuli, without having to switch filtering techniques or calculate appropriate impulse responses for different sampling frequencies.

A final test was conducted to test the performance of the models with the empirically derived parameter sets on a new selection of audio samples taken from a different multitrack recordings. The experiment was run at a range of SPLs and revealed a clear systematic effect of sound level on both the STL time series and the global-loudness descriptors. At comfortable listening levels of 75 dB SPL and below,[27] the approximations were fairly consistent with those found in the previous evaluation, with effect sizes of -0.35 for the CH12 (lower error on the new dataset) and 0.4 for the GM02 (greater error). The unstandardised effect sizes (STL nRMSE) of roughly 1 dB are most likely not of practical concern for purposes of loudness normalisation, given that the global-loudness descriptors were still in line with the reference values. However, at higher levels the performance of the GM02 degraded, whereas the CH12 improved. The GM02 approximation was indeed more vulnerable to greater error on the test data, since it made use of higher spectral compression and fewer auditory filters than the CH12. Further investigations indicate that the cause of these level-dependency discrepancies are a result of the changes to the power-spectrum introduced by spectral compression, in combination with the model specific level-dependent excitation transformation. A lower compression criteria, such as 0.2 Cams, can be used to reduce this level-variant effect for the GM02.

As a final point, the measurements of both time series error and global-loudness descriptor error made throughout this chapter suggest a general rule that may be useful for future work: mean STL nRMSEs of up to -24 dB can be expected to preserve estimates of global loudness (mean, peak and 95th percentile) as generated by the reference algorithm, to within 0.3 dB. Average discrepancies in predicted global loudness can be expected to be 0.1 dB or less, depending on the algorithm.

4.6 Summary

In this chapter, Glasberg and Moore's (2002) dynamic loudness model has been optimised and parameterised to facilitate real-time loudness feature extraction on modern computing platforms. The primary interest was to preserve the loudness-time series which are fundamental to the estimation of loudness dynamics and overall loudness perception, but errors in the intermediate auditory patterns were also explored. Statistical profiling and complexity analysis were used to identify performance bottlenecks and establish ways to alleviate them. Techniques for reducing execution

[26] Again, the assumption of normality was valid for 28 of the 30 comparisons so a parametric paired t-test was used.

[27] This is subjective, but informal listening tests involving three experienced audio engineers agreed that a loudspeaker reproduction level of 73 dB SPL was comfortable listening level for the stimuli collected by Ward and Reiss (2016), despite inter-stimulus fluctuations in global loudness.

time were presented and evaluated by analysing the errors introduced in the auditory features of 30 sounds comprising speech and musical instruments.

It was determined that the majority of running time was spent transforming the short-term power spectrum to an auditory representation called the excitation pattern (EP). The complexity of the level per critical band calculation was linearised and an LUT was employed to yield significant speedup with minimal error introduced. Following previous work, an algorithm for reducing the number of spectral components involved at the excitation transformation stage was adapted and incorporated into the loudness model. Two parameters were then used to simplify the computation of EPs: the spacing between adjacent rounded exponential (roex) filters and the perceptual criterion defining the amount of smoothing applied to the power spectra. The hopping Goertzel discrete Fourier transform (HGDFT) was also presented, extended and evaluated for the first time in the context of a complete loudness model. Finally, a number of alternatives to the original high order FIR pre-cochlear filter were proposed and assessed by comparing the goodness-of-fit to empirical data.

The proposed efficient model makes use of a simplified EP procedure, a low-order IIR filter and a filter spacing and spectral compression criterion of 0.5 Cams and 0.2 Cams, respectively. For the types of sounds tested in this study, a hop size of 4 ms yields acceptable performance in terms of preserving measures of global loudness within tolerable error bounds. Furthermore, because the model has been fully parameterised, the speed-accuracy trade-off can easily be adjusted according to the requirements of a given application. Large speedup can be obtained by using the HGDFT to execute the multi-resolution spectrogram, which greatly simplifies the implementation of the algorithm at the cost of being less flexible.

These techniques were then used to develop a real-time binaural loudness meter driven by either Glasberg and Moore's method or the more recent auditory model of Chen and Hu (2012). Both algorithms were modified to process stereo signals in a way that agrees with empirical findings and recent auditory theory on binaural loudness perception. The meter displays both left and right channel loudness measurements, in sones, along with the plots of the calculated internal representations. The effects of hop size, filter spacing and spectral compression on STL and SL features were systematically investigated to understand how the approximations affect the resulting measurements and identify differences in functionality between the two loudness estimators. A set of parameters were empirically derived for both models that give good performance in terms of metrics commonly used for loudness alignment at typical listening levels. Additional tests using multitrack audio confirmed this, but did reveal a level-dependent effect that is likely explained by the amount of compression applied to the spectrum prior to auditory filtering.

Preliminary investigations indicate that a time step of 4 ms is more than adequate for the purpose of loudness visualisation and normalisation, although subjective experiments are warranted to measure the sensory resolutions necessary for real-world applications. Improvements offered by interpolating EPs have been quantified and discussed, but better numerical integration techniques are likely optimal in terms of improving integrated loudness measurements with low filter densities. It is imperative to note that the model parameters are, however, source dependent. Adaptive loudness modelling is thus a fruitful topic for future research and should be compared to the static approach presented in this work, especially with respect to the overhead they introduce.

Chapter 5

Loudness Driven Automatic Music Mixing

The previous chapter introduced computationally efficient approaches to multiband loudness estimation based on two established models of auditory periphery. This enables the embedding of sophisticated hearing models into audio applications that may benefit from perceptually-relevant descriptions of the incoming sound to be processed. This chapter, therefore, applies an accelerated adaptation of Glasberg and Moore's (2002) dynamic loudness model to the problem of real-time automatic music mixing. The purpose of this study is to identify if a more complete auditory model can improve the state-of-the-art in real-time automatic fader control in which single-band models, such as the EBU Tech 3341 (2016) loudness meter, are used to balance the instruments within a mixture (Gonzalez and Reiss 2009; Mansbridge et al. 2012b). The proposed system is primarily targeted towards (and presented here within the context of) assisting music producers by automating mundane yet fundamental procedures when mixing music, but may serve other fields such as live sound reinforcement, audio for games and broadcasting.

A fundamental procedure in music mixing is the fader balance. In conjunction with highpass filtering and panning, the fader settings are responsible for obtaining a good mix balance (Senior 2011). The primary objective is to ensure that every instrument is clearly audible throughout the duration of a song and that the overall mixture is at a comfortable listening level. In the most basic form, the mix is defined as a sum of amplitude-weighted time-varying musical signals x_k, also known as 'tracks':

$$\text{mix}[n] = \sum_{k=1}^{K} f_k x_k[n], \tag{5.1}$$

where n is the discrete-time index, K is the total number of tracks and f defines the weight or contribution of each track as controlled using the *fader*. A typical first step in mixing is to establish the track weights, i.e. the fader levels, such that:

- The listener can hear each track consistently.
- The desired mix balance is preserved across different reproductions systems.
- The overall level is sufficiently high to optimise the signal-to-noise ratio, thus minimising the need for large gains at the output device.
- No clipping takes place.

- There is sufficient headroom in the mix to allow for input level dynamics, further signal processing and easy alteration of mix balance.

The points above combine both technical, perceptual and subjective demands. For example, the first two statements are concerned with the audibility and relative loudness of each element in the mix, but also consider the specification of the end playback medium(s). The remaining statements address some important technical considerations involved when mixing music and can be treated as constraints on the gains f_k in Equation 5.1. For example, since digital signals are limited to the range $[-1, 1]$, a conservative constraint to prevent output clipping is to ensure that the total gain applied to the input tracks is less than 1, i.e. $\sum f_k < 1$.

5.1 Related work

Automatic mixing techniques have been developed to solve problems such as those listed above with minimal human interaction. When human interaction is not required for operation, these systems are said to be *autonomous*, whilst the term *automatic* is reserved for devices that may be configured or instructed by a human. For brevity, the term automatic is used here to generalise devices capable of performing a mixing task with little or no influence from the user. Many automatic systems have been developed to balance the levels of a multitrack recording (Dugan 1975; Julstrom and Tichy 1983; Dannenberg 2007; Gonzalez and Reiss 2009; Terrell and Reiss 2009; Mansbridge et al. 2012b; Ward et al. 2012; Terrell et al. 2013). Dugan (1975) developed one of the first automatic mixers for controlling the levels of a multi-microphone setup. Two systems were proposed to achieve maximum gain before feedback: one for speech and music, and a simplified yet elegant design applicable to speech only. The main goal was to provide a stable yet flexible system that could manage the gain structure of multiple microphones in live situations. For speech applications, where intelligibility is of prime importance, the author focused on dealing with technical difficulties such as feedback prevention and noise removal. Adaptive threshold gating was proposed to detect active microphones. Julstrom and Tichy (1983) also developed an automatic multi-microphone mixer but instead used specially designed microphones and circuitry to achieve more reliable gating over methods based on level measurement alone. These early systems were primarily developed for improving the intelligibility of speech in live sound reinforcement applications, such as conferences, church meetings and broadcasting.

Little further advances were made in the field of automatic mixing until the rise of the digital age where the digital audio workstation (DAW) and its software extensions, known as plugins, became more accessible. Moorer (2000, p. 498) proposed the arrival of intelligent assistants, allowing computer programs to 'take over the mundane aspects of music production, leaving the creative side to the professionals, where it belongs.' To clarify, audio assistants can perform the technical aspects of the mixing process automatically, much like the design goals of Dugan (1975). White (2008) made more explicit proposals such as automatic equalisation, level balancing and dynamic range compression based on genre-dependent templates. The author also emphasised the problem of musicians having to alternate between creative and technical roles.

Dannenberg (2007) proposed an offline auto-level balancer as part of a larger intelligent-audio editor in which the root mean square (RMS) levels of the different tracks are brought to a target nominal level. The author used low-level signal features for simplicity, but did suggest enhancements such as incorporating perceptually motivated measures. Terrell and Reiss (2009) presented an optimisation framework to find the static gains for achieving a set of target monitor mixes subject to various constraints such as maximum and minimum sound pressure levels (SPLs), and avoiding acoustic feedback. This approach is less concerned with real-time performance and perceptual phenomena, but more so the physical difficulties encountered in live sound.

Perceptually-based sound features have also been employed to harness the power of psychoacoustic principles. Gonzalez and Reiss (2009), for example, implemented a real-time level setter by measuring loudness from each input track and adapting the gains according to a perceptually motivated objective function. Rather than solve for the set of static gains that achieve a common RMS level (Dannenberg 2007), the authors developed a cross-adaptive effect to dynamically adjust the gains such that the track loudnesses target an average. This criterion was based on the assumption that '... a mixture in which loudness per channel tends to the overall average loudness is a well-balanced mixture with optimal inter-channel intelligibility' (Gonzalez and Reiss 2009, p. 1). In this audio effect, loudness is measured by filtering each track using approximations of the level-dependent (inverted) ISO 226 (2003) equal-loudness contours, followed by short-term amplitude averaging. Probability mass functions are employed to obtain a set of representative loudness values, from which a target average is derived. Gonzalez and Reiss (2009) also considered technical issues, e.g. guaranteeing system stability by continuously normalising the signal gains. Mansbridge et al. (2012b) refined the automatic fader controller, which was subsequently combined with an algorithm to automatically pan the sound sources comprising a mix (Mansbridge et al. 2012a). These two systems make use of the EBU Tech 3341 (2016) loudness meter which itself is founded on the ITU-R BS.1770 (2015) algorithm for measuring loudness. Subjective evaluation showed that the system outperformed its predecessor (Gonzalez and Reiss 2009) with mix quality rated consistently higher for nine songs of different genres, including metal, pop, rock, folk, indie and R&B. The device's success can be attributed to the use of an improved and validated loudness model applicable to programme material, as well as further advances to signal conditioning.

Ward et al. (2012) developed an offline multitrack mixer using Glasberg and Moore's loudness and partial loudness model (Glasberg and Moore 2002, 2005), allowing for the consideration of masking interactions between the sources. This model was first presented in the context of multitrack audio mixing by Aichinger et al. (2011), but a mixing tool was not realised.[1] The authors established an equal-loudness mix by dividing each input track into segments of musical activity and performing loudness equalisation on all overlapping segments. Measures of partial loudness were then used to restore the masked (i.e. in the mix) loudness of a user-prioritised track, which was demonstrated using a singing voice embedded in a rock song. Ward et al. (2012), informally reported that the mixing system underestimated the loudness of low-frequency instruments, suggesting that the psychoacoustic loudness model may not be suited to songs comprising a diverse range of instruments.

A similar system was presented by Terrell et al. (2013) which allowed a target *relative-loudness balance* to be specified by the user. For example, the user of the mixer can simply state that they want the lead vocal to be twice as loud as the rhythm guitar, and the system solves for the optimal fader levels that realise the balance. Thus, mixes other than equal track loudness (as per (Gonzalez and Reiss 2009; Mansbridge et al. 2012b; Ward et al. 2012)) are easily configured with this approach. Unlike the system of Ward et al. (2012), this mixing device is based entirely on partial loudness.[2] Equal-partial-loudness mixing was demonstrated by specifying a unity track-to-average loudness ratio across all tracks. By using a model which measures the perceived magnitude of the input in sones, the authors were able to optimisation a mix based on loudness ratios. Finally, Wichern et al. (2015) compared the offline auto-levellers of Ward et al. (2012) and Terrell et al.

[1] Aichinger et al. (2011) used the model as part of a subjective test to investigate the intelligibility of musical instruments under varying degrees of energetic masking.

[2] Problems with this approach are touched upon towards the end of this chapter.

(2013) with the ITU-R BS.1770 (2015) single-band loudness model.[3] A set of instrument-dependent target-loudness templates were identified for each of the three loudness features, which were then used to generate different automatic mixes of unseen multitrack content. The subjective evaluation revealed that the mixes realised using the single-band loudness feature were preferred over those generated using more complex psychoacoustic models.

5.2 This work

This chapter presents a new device to automatically adjust the levels of a multitrack recording in real-time. Unlike prior work, the system is presented as a framework in which different loudness models can be used to calculate the track gains. The advantage of this system is that the implementation is not bound to a particular algorithm and therefore interfaces with different features, including nonlinear loudness estimators. This is demonstrated by employing two different models to realise the mix balance. In particular, the system was used to mix five multitrack recordings using loudness features based on the specifications of EBU Tech 3341 (2016) and Glasberg and Moore's (2002) excitation pattern (EP)-based loudness model. The main goal of this study was to investigate whether the use of more sophisticated features grounded in auditory theory can be used to improve the quality of automatic computer-generated mixes relative to recent developments.

5.3 Implementation

In audio engineering, a channel or track is a discrete-time signal often representing a music or speech source. An audio mixer is a device that accepts multiple input channels and sums them to a reduced number of output channels for reproduction. For example, in live sound reinforcement, on stage microphones are connected to a mixing device which combines the incoming signals for delivery over two or more front-of-house speakers. The proposed mixing device places no limit on the number of input tracks and all inputs are assumed to be monophonic. Let x_k denote the k^{th} channel of a K channel mixer. A mix is then defined as

$$mix[n] = m[n] \sum_{k=1}^{K} p_k[n] f_k[n] x_k[n], \tag{5.2}$$

where p_k and f_k are the pre-amplifier and fader gains of channel k, respectively, and m is the master gain. For clarity, n denotes the current sample number of an audio-rate signal or the interpolated sample index of a control signal operating at the frame rate. This definition is similar to the one given by Equation 5.1, but the single gain has been expanded to express the design of the system. The pre-amp control signal is used to normalise the input channels such that dynamic range is optimised whilst avoiding signal distortion (Gonzalez and Reiss 2009). In live applications, the pre-amplifiers are often configured during the initial sound-check and left static during the true performance, enabling the engineer to focus on dynamically adjusting the fader levels to balance the loudness of each sound source. The master gain is used to prevent overall clipping (mix level exceeding unity) and facilitate change in overall mix level.

The software was developed in the audio programming language SuperCollider (McCartney 1996) which operates on a frame-by-frame basis, where a frame typically comprises 64 samples. The time-varying control signals in Equation 5.2 are derived from features extracted from the inputs. Sudden changes in these features can lead to undesirable audible artefacts and so it is

[3] The system developed in this chapter was evaluated prior to the publication of Wichern et al. (2015).

necessary to smooth out such fluctuations (Reiss 2011; Mansbridge et al. 2012b). Erratic scaling can be minimised through the use of lowpass filters such as the exponential moving average (EMA) filter:

$$y[n] = \alpha x[n] + (1 - \alpha)y[n - 1]. \tag{5.3}$$

For a given sampling frequency f_s, the relationship between the smoothing coefficient α and filter time constant τ (in seconds) is given by Equation 3.11. For control-rate signals, the smoothing coefficient is calculated by substituting the frame rate for f_s.

5.3.1 The models

The level balancer presented here expects a real-time online loudness algorithm whereby multiple instances can be run simultaneously in order to extract the loudness of every track. If the model is linear with respect to signal level, decibels readings are expected. For nonlinear loudness models, it is assumed that the readings are given in sones. Two loudness models were employed in this study:

- An adapted version of the short-term loudness (STL) meter specified by EBU Tech 3341 (2016). From here on in, this model is referred to as EBU, after the European Broadcast Union.
- An adapted version of Glasberg and Moore's (2002) dynamic loudness model. This model is referred to as GM02.

The EBU model was implemented similarly to the efficient realisation of Mansbridge et al. (2012b). In short, the signal is filtered by two cascaded biquad filters, collectively termed the K-filter (see Section 3.2.3). EBU Tech 3341 (2016) specifies that the STL be calculated using a sliding rectangular window of length 3 s. The frequency-weighted level measurement was approximated using an EMA filter applied to the square of the K-filtered audio. The smoothing coefficient α is selected according to Equation 5.4 which relates the averaging performed by a P (sample) simple moving average and an EMA (Brown 2004).

$$\alpha = \frac{2}{P + 1} \tag{5.4}$$

Thus, the EBU STL was implemented according to the equations

$$y_k[n] = \alpha x_{w,k}^2[n] + (1 - \alpha)y_k[n - 1], \quad \text{and} \tag{5.5}$$

$$L_k[n] = 10 \log_{10} (y_k[n]) - 0.691, \tag{5.6}$$

where $x_{w,k}$ is the K-weighted signal corresponding to track k, and the smoothing constant based on a 3 second average period. The loudness measurement L_k is expressed in decibels with Loudness Units relative to Full Scale (LUFS).

The implementation of the GM02 was based on the efficient realisation developed by Ward et al. (2013). It uses the hybrid filtering technique proposed in Section 4.3 to simulate the transmission of sound through the outer and middle ear, and performs the multi-resolution spectrogram using the hopping Goertzel discrete Fourier transform (HGDFT). This model was realised as a SuperCollider processing module, called a unit-generator, using a hop size of 128 samples (2.7 ms at $f_s = 48$ kHz). Given that the GM02 displays readings in sones, the mixer takes the long-term loudness (LTL) of each track and transforms it to a ratio in dB:

$$L_k[n] = 10 \log_{10} \Psi_k[n], \tag{5.7}$$

where Ψ is the LTL in sones and L_k is the corresponding level[4]. The LTL was used following the suggestions of Glasberg and Moore (2002) and Moore et al. (2003). The release time of the LTL was set to 2 s.

5.3.2 Dynamic gate

Modifications to signal gain should only occur when there is justification to do so; there is no reason to scale a signal that is inactive.[5] Therefore, a gain adjustment is only permitted if an active state of musical activity is registered by the system. This avoids incorrect level adjustments caused by background noise. For example, when a lead vocalist stops singing into the microphone, it would be unwise to increase the signal level based on features extracted at the present time.

A Schmitt trigger is a dual-comparator circuit used to remove noise from an input signal, providing a stable logical output signal. The Schmitt trigger employs a low threshold t_l and upper threshold t_u to reduce output oscillations caused by noise in the input which would otherwise arise when using a conventional comparator based on a single threshold. The input signal must exceed t_u in order to produce an output of one, and fall below t_l to output zero. For intermediate values, the output signal is fixed at its previous state and so variations in signal level have less impact on the gated output. This is particularly useful for dynamic music signals.

For an online system, both the dynamic range and overall signal level of an input track are unknown. The use of fixed thresholds is therefore difficult, especially for signals suffering from high amounts of instrument bleed or spillage (common in live performance). Here, a Schmitt trigger with dynamic thresholds is proposed to solve this problem.

An EMA filter is used to estimate the short-term signal energy $l_{1,k}$ of each input channel using a time constant of 100 ms. A second EMA filter with a 5 s time constant is used to produce the long-term average $l_{2,k}$. A binary control signal is then calculated to determine stable states of activity:

$$a_k[n] = \begin{cases} 1 & : l_{1,k} > t_{u,k}[n] \\ 0 & : l_{1,k} < t_{l,k}[n] \\ a_k[n-1] & : \text{otherwise.} \end{cases} \tag{5.8}$$

The lower threshold is set 5 dB below the upper threshold and the initial value of $t_{u,k}$ is set to -50 dB full scale (FS), which is subsequently updated according to

$$t_{u,k}[n] = \begin{cases} c \times l_{2,k}[n] & : c \times l_{2,k}[n] > t_{u,k}[n-1] \\ t_{u,k}[n-1] & : \text{otherwise,} \end{cases} \tag{5.9}$$

where c is a scaling factor. Typically, the signal-to-background noise ratio will be greater than one when the source is active. Thus, the upper threshold will approach the (scaled) average energy of the desired signal, ensuring that background noise will not interfere with the activity state. An empirically derived value of $c = -6$ dB was chosen to allow for natural variation in the average signal level, since otherwise a logical output of one only occurs at the highest of signal levels. The dynamic gate is demonstrated in Figure 5.1 which shows the short-term energy l_1 and gating

[4] For clarity, L_k is used to denote the loudness of a track in dB regardless of the algorithm used to calculate it.

[5] Signal gating is an exception to this rule.

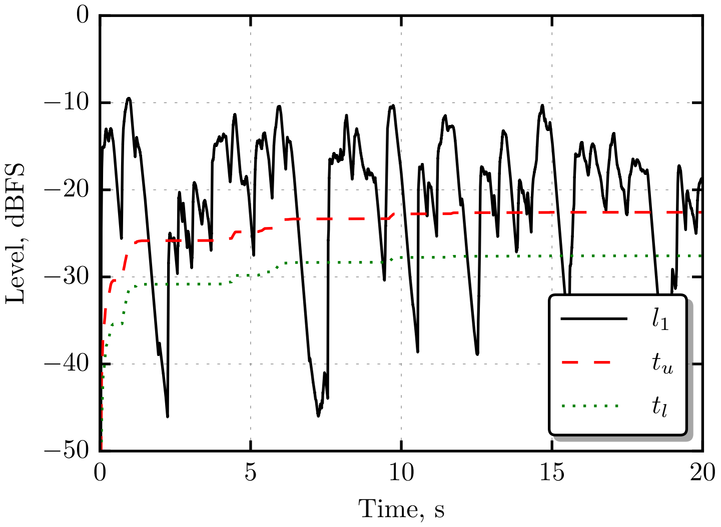


Figure 5.1: Demonstration of the dynamic gate applied to a recording of a male singer.

thresholds calculated from of a recording of a male singer. As time progresses, the thresholds converge to the average level of the desired input signal. Any noise falling below about -27 dB FS is treated as silence and therefore no longer affects subsequent measurements.

5.3.3 Peak normalisation

Peak normalisation scales a signal by a constant such that the maximum absolute value (peak) is equal to a target value, typically 0 dB FS. This ensures no signal will exceed the digital limit whilst maximising the source level. Aligning the signal levels also ensures that the average loudness measurements are not biased by very quiet tracks, which may otherwise necessitate human intervention to optimise the mix level. In an offline system, the multitrack audio are known in advance and so peak normalisation is generally computed using a gain factor of $1/\max(|x|)$. One approach to online peak normalisation is to monitor signal peaks using a peak follower:

$$x'_k[n] = \begin{cases} |x_k[n]| & : |x_k[n]| > x'_k[n-1] \\ x'_k[n-1] + \alpha(|x_k[n]| - x'_k[n-1]) & : \text{otherwise.} \end{cases} \tag{5.10}$$

A time constant of 10 s was selected to ensure the peak of pure tones amplitude modulated at a rate of 1 Hz could be tracked reliably with little variation. The pre-amplifier gain p_k is then equal to the reciprocal of the average peak and, if the track is active, subsequently smoothed using an EMA filter which holds its previous value if the corresponding input track becomes inactive (Mansbridge et al. 2012b).

5.3.4 Fader control

The peak normalised signals must now be balanced to produce a subjectively pleasing mix. As with previous approaches (Gonzalez and Reiss 2009; Mansbridge et al. 2012b; Ward et al. 2012) a first approximation to a balanced mix is to equalise the loudness of each track dynamically over time, such that no single instrument dominates. When using linear models such as the EBU, the fader gain can be calculated using

$$f_k[n] = 10^{e_k[n]/20}, \tag{5.11}$$

where e_k is the error between the target relative loudness level $L'_{tgt,k}$ and the actual relative loudness level L'_k of track k:

$$e_k[n] = L'_{tgt,k}[n] - L'_k[n], \quad \text{where} \tag{5.12}$$

$$L'_k[n] = L_k[n] - \bar{L}[n], \tag{5.13}$$

with L_k and $\bar{L}$ the measured absolute loudness level (in LUFS) of track k and the average absolute loudness level of all *active* tracks, respectively. The target relative-loudness balance L'_{tgt} can be specified by the user as a 'floating' vector of levels, which is then normalised to have zero mean. For example, for a mix involving just two tracks, the user might enter $\{-10, 0\}$ to make the first instrument 10 LU lower than the second. In this case, the mean is subtracted to give the target balance $L'_{tgt} = \{-5, 5\}$. Similarly, an equal-loudness balance is generated by simply setting $L'_{tgt} = 0$ for all k. The fader adjustments are smoothed in dB by an EMA filter prior to linear transformation and application.

For the GM02, the relationship between signal gain and loudness (here in dB) is nonlinear: a gain increment of X dB does not necessarily increase the loudness reading by αX dB, where α is a constant. The system therefore incorporates feedback to correct the error between the target balance and the actual balance *after* application of the gains. This is achieved by placing the loudness meters post fader. With the fader *level* (dB) of the previous frame denoted $F_k[n-1]$, the current fader factor is computed using

$$f_k[n] = 10^{(F_k[n-1]+e_k[n])/20}, \tag{5.14}$$

which requires smoothing over a sufficiently long time period to avoid gain oscillation and ensure convergence of the track loudnesses. Both feedforward and feedback topologies are open to any loudness feature. If feedforward mode is specified with meters that calculate loudness in sones, a sone-to-phon transformation can be used to approximate the required level change in Equation 5.13. Appendix B gives some insight into the relative error associated with this approach. When in feedback mode, the sone values are converted to dB (not phons), which means the target balance is specified as a vector of loudness ratios as proposed by Terrell et al. (2013). Note that Mansbridge et al. (2012b) used a feedforward scheme to perform equal-loudness mixing, but feedback mixing may be preferable, even for linear loudness estimators, for two reasons. First, the faders are not influenced by the channel count. For example, consider three active channels in a feedforward equal-loudness mixer. If a single track becomes inactive after the three gains have stabilised, the two remaining tracks will shift in level as the target average updates. In feedback mode, however, the levels would stay fixed because the tracks are already measured as equally loud. If the third track re-enters the mix with little change in loudness, the feedback mixer would still satisfy the target relative-loudness balance, whereas the feedforward mixer would have to recompute the previous three-track mix. Second, the feedback topology allows the per track loudness meters to display *post-fade* loudness, which is preferable from the perspective of the user.

In both mixer configurations, the initial pre-amp levels are set to 0 dB and the faders levels default to $20\log_{10}(1/K)$. This ensures that the master output does not exceed 0 dB FS in the case of peak normalised input signals and, in practice, provides sufficient mix headroom for the pre-amplifiers to operate safely.

5.3.5 Master section

It is important that the automatic faders do not clip the master level. Furthermore, it is desirable to have some mix headroom in order to allow for natural dynamics and unexpected variation in level. The system fulfils these requirements by overriding the master gain as needed:

$$m[n] = \begin{cases} (1-\alpha)m[n-1] & : |mix[n-1]| > h \\ m[n-1] & : \text{otherwise}, \end{cases} \tag{5.15}$$

where h denotes the headroom in linear amplitude, which defaults to 0.5 (6 dB headroom). Thus, the system attempts to keep the signal level below -6 dB FS. If desired, the master gain can be offset by the user to decrease or increase the overall listening level, but the gain will automatically adjust the master level as per Equation 5.15.

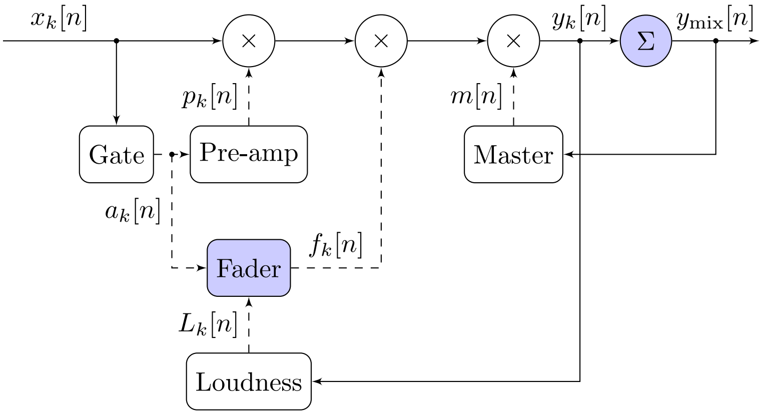


Figure 5.2: Block diagram showing the per channel processing stages of the automatic mixer in feedback mode. Solid lines represent audio signals and dashed lines represent control signals. Blue filled nodes denote multichannel input processors.

A flow diagram of the mixing system in feedback mode is depicted in Figure 5.2. Note that all gain modifications occur prior to the loudness estimator, ensuring that the level-dependent GM02 always operates in a calibrated environment and so the relative-loudness balance is preserved regardless of listening level. Both the 'Fader' and 'Σ' stages, highlighted in blue, derive their inputs from all tracks comprising the mix. For example, the fader computer requires the loudness of all tracks in order to process the signals cross-adaptively. The default target relative-loudness balance is set to unity to target equal-loudness mixing, but, in addition to the overall mix level, the balance can be modified by the user in real-time.

5.4 Subjective evaluation

A subjective evaluation was conducted in which automatic mixes of five multitrack recordings, generated by the proposed system, were compared to those performed by four human mixing engineers. The five multitrack projects were obtained from the Free Multitrack Download Library (Cambridge Music Technology 2013). Five songs of different genres were selected for the evaluation: barbershop vocal quartet (four tracks), acoustic rock (eight tracks), classical chamber (four tracks), blues (seven tracks) and reggae/gospel (eight tracks). In most cases each track was a recording of an individual instrument, with a few exceptions: both the rock and the blues multitracks included a drum overhead microphone and the classical multitrack included a room microphone capturing all instruments simultaneously. A representative 20 s segment was extracted from each multitrack and the monophonic stems were peak normalised.

Human mixes were collected from ten sound engineers who were either undergraduates or graduates of the Sound Engineering and Production degree course at Birmingham City University. The engineers were asked to produce a well-balanced mix of each song, using only the fader levels and keeping the master fader at unity. Mixes were performed using the DAW Logic Pro 8. Although level automation was permitted, the engineers felt this to be unnecessary considering the relatively consistent dynamics over the 20 s period. The engineers were asked to optimise the levels of each mix such that the peak mix level was maximised but did not exceed 0 dB FS. The engineers were therefore instructed to mix iteratively to ensure a good final balance (rather than increasing the fader levels at the end of each mix session which is likely to alter the perceived balance). This was done so that mixes were produced at a similar listening level. No limit was placed on the overall duration of the sessions and regular breaks as enforced. Playback was output through a pair of BeyerDynamic DT990 headphones and the exact same reproduction settings were used in every session. The decision to use headphones for reproduction was for practical convenience as (a) it was not possible to use the same listening room for both the mixing engineers and listening participants, and (b) the transfer function of the headphones had been measured and implemented as part of the GM02.

A subset of the mixes were used for the subjective assessment (see Section 5.4.1), as including all ten human mixes would require listeners to make an unrealistic number of comparisons. In particular, the mixes performed by the four engineers with the most mixing experience were used for evaluation against the automatic mixes. The author listened to the selected mixes and inspected the software projects to ensure that they were fairly representative of the mix-balance task, with no obvious problems, e.g. distortion or use of audio effects.

For each multitrack, three computer-generated mixes were recorded using the software described above: one using the EBU, another using the GM02, and the third using a simple unweighted-frequency level measurement. The latter model is identical to the EBU but with the initial K-filter bypassed. Since this third model is an RMS-level detector, it is simply referred to as RMS. The outer-ear transfer function required by the GM02 was set to the transfer function of the same headphones used by the engineers. For each auto-mix, the pre-amplifiers were activated and then deactivated once the gains had stabilised. The automatic faders were then activated and the mix headroom was set to 0 dB. After several continuous playback loops of a given mix, the automatic fader levels were relatively stationary throughout the 20 s of each song and the master gain had eliminated clipping (if any). At this point the mix was recorded. For the first listening task (see below) the system was configured with the default relative-loudness balance of 0 dB across the tracks, i.e. equal track-loudness. For the second task, the lead vocal in each song was increased by approximately 4 dB above the level at equal loudness. This was done to enhance the quality of the automatic mixes when compared against human mixes where the lead vocal tends to be more prominent compared to other instruments. For example, Pestana (2013) found that the typical listener prefers the vocals to be around 4–6 LU higher than any other instrument when the remaining tracks are equally loud.

5.4.1 Assessment procedure

A test similar to the multiple stimuli with hidden reference and anchor (MUSHRA) assessment procedure (ITU BS.1534 2015) was used to obtain subjective ratings of the different mixes for each multitrack according to two criteria:

- Task 1: *'Rate the following mixes according to the equal loudness of all sources.'*
- Task 2: *'Rate the following mixes according to the overall quality.'*

These two tasks were used by Mansbridge et al. (2012b) to assess the performance of their automatic fader controller and were also considered appropriate for evaluating the system developed here. The first task targets the ability of the device to mix for equal track-loudness, whilst the second addresses overall mix quality. In each experiment, subjects were asked to rate the mixes on a quality scale from 0–100 based on the above criteria. The grading scale recommended by ITU BS.1534 (2015) is divided into five equal intervals with the following adjectives: bad (0–20), poor (20–40), fair (40–60), good (60–80) and excellent (80–100). In this study, all subjects were encouraged to make use of the full scale by assigning at least one mix in the lowest category and another in the highest category. The presentation order of the trials was randomised as was the order of the mixes on screen within each trial. The user interface also featured a scrub bar allowing subjects to quickly navigate each stimulus and easily compare selected points across the mixes.

The MUSHRA protocol requires subjects to rate the quality of different stimuli against a reference source on a scale of 0–100. A hidden reference and hidden anchor is included in the test corpus to establish fixed points, stabilising the scale, and to allow the reliability of the test results to be assessed. A reference stimulus was not deemed appropriate in these experiments since no single mix could be objectively selected as having the best equal-loudness balance or highest overall quality. This variation of the MUSHRA design has been used in previous work for obtaining subjective quality ratings of music (Maddams et al. 2012; Mansbridge et al. 2012b). However, in this study, it was decided to include a hidden anchor in both experiments to ensure that at least one of the mixes would be consistently rated low in both experiments, allowing the reliability of each subject's ratings to be verified (Sporer et al. 2009). The hidden anchor consisted of a mix in which one instrument was set markedly louder than the other instruments and that this artefact was perceptually obvious in comparison to the mixes produced by the mixing engineers. It was felt that the use of this anchor would introduce contrast into the quality of the mixes, helping the listeners to make use of the entire rating scale, and also provide a baseline indication of worst-case performance in post-analyses.

14 individuals participated in the listening assessment. 13 of the subjects were male and one was female, with ages ranging between 21 and 26 (median = 22). 12 of the listeners had audio engineering experience (mixing, recording and critical listening) ranging from two to seven years (median = 3.5), whilst the other two participants reported no experience in this area. No subject reported having any known hearing impairment. The same equipment and configuration used to generate the mixes was used for the listening assessments. Thus, listeners heard the mixes under the same conditions as the mixing engineers. Informal listening by the author indicated that the overall loudness of the different mixes were fairly consistent with one another. Prior to the main experiment, subjects undertook a short training phase in which they were able to explore the full-range of mixes and ask any questions regarding the grading procedure.

5.4.2 Results

Post-screening

Post-screening of the subjective ratings was performed by analysing the scores assigned to the anchors, checking for full use of the rating scale (in accordance with the instructions), and performing correlations between the individual subject ratings and the median of all ratings. This procedure revealed that three subjects (S1, S10 and S13) rated the anchor for the classical song

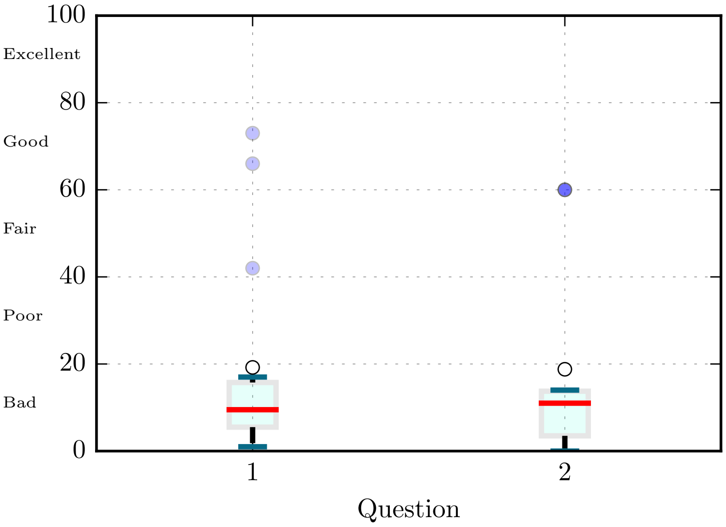


Figure 5.3: Boxplots showing the distribution of the subjective ratings assigned to the anchor mix of the classical song. The red horizontal line shows the median and the white circle shows the mean for each task. The blue shaded circles correspond to the outlying observations discussed in the text.

between 40 and 80 (fair-good) in task one, and in task two they assigned the same mix a rating of 60 (borderline fair-good). Boxplots of the scores for these two conditions are shown in Figure 5.3, which indicate that the ratings from these three subjects are potential outliers and clearly bias the mean (white circles). Further inspection revealed that, for this condition, S1 and S10 did rate the anchor the lowest relative to all other test items, but failed to make use of the entire scale. S13, however, rated two of the human mixes of the classical piece below the anchor in both tasks. In trial 'Task 1-Vocal Quartet', S1 and S13 failed to make use of the lowest category, and S13 rated the anchor second highest (70 = good) as well as failing to make use of the highest category. In this same trial, S1 scored their highest mix at 69. Thus, despite the on-screen instructions to make use of the entire scale, a few of the subjects still felt strongly against assigning scores in the extrema of the rating scale. Sporer (1996) also found the use of the quality rating scale to be highly subjective, which has been accentuated in more recent work (Sporer et al. 2009).

Pearson and Spearman correlations between the individual ratings and the medians of the subjective ratings across all conditions revealed that S13 had the lowest correlation in the group.[6] This subject happened to be one of the two participants with no audio engineering experience. This shows the importance of pre-screening subjects to ensure reliable and consistent results. Taken together, it was decided to remove S13 from the dataset, whilst acknowledging a few influential observations generated by the failure of some subjects to utilise the entire rating scale.

Task 1: Equal-loudness mix ratings

In this task, the subjects were asked to rate eight mixes of a given song based on the degree of equal loudness between the different instruments. If all elements were judged to be equally loud, a maximum rating of 100 would be expected. The subjective ratings of the eight mixes for each of the five songs are displayed using boxplots in Figure 5.4. The mixes produced by the four human

[6] Sporer et al. (2009) reported individual correlations below 0.85 warrant inspection but the correlations measured here are lower than this value for all subjects. Pestana (2013) also reported lower correlations when screening subjects after conducting multi-stimulus tests for subjective mix assessment.

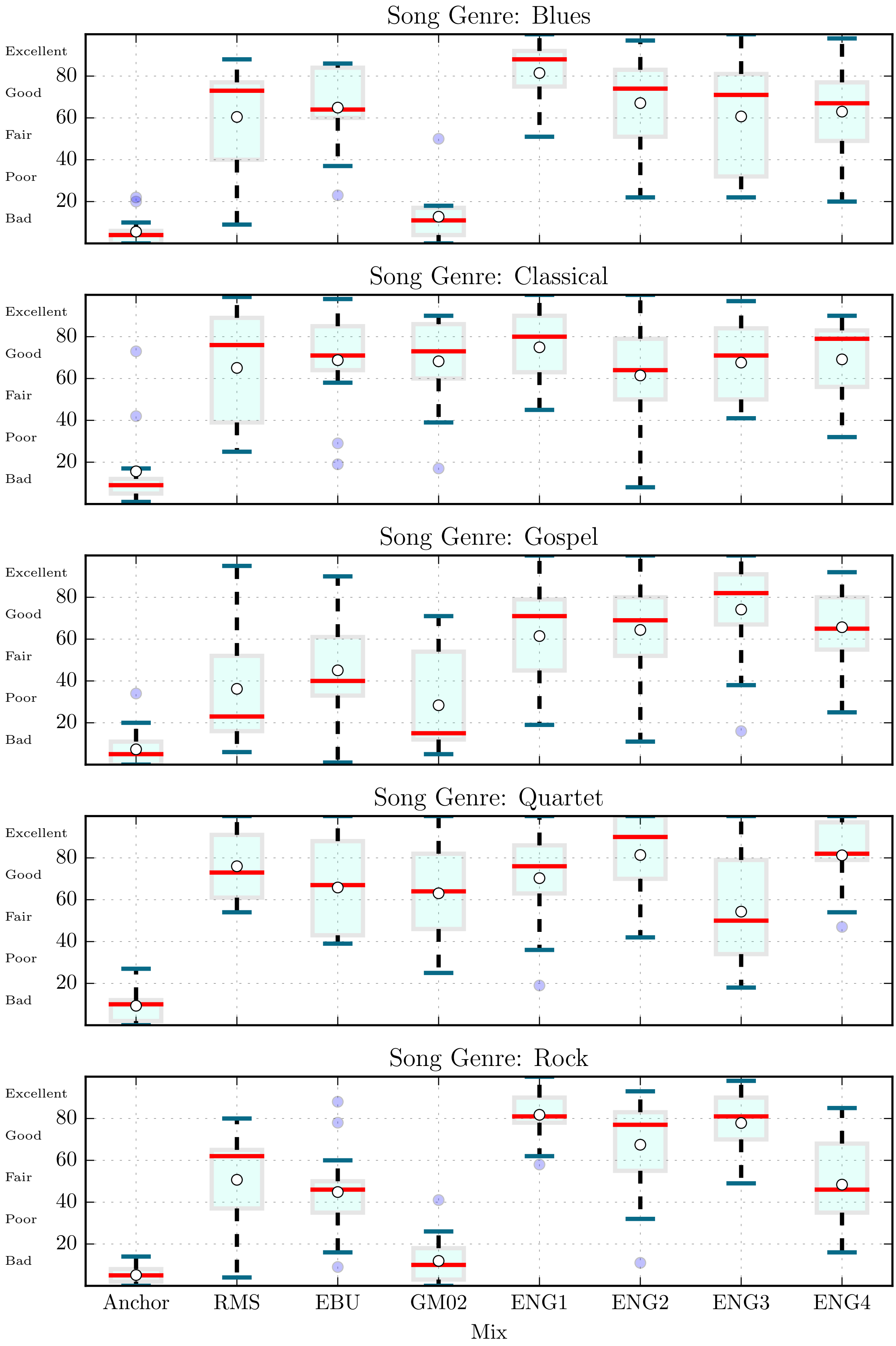


Figure 5.4: Boxplots showing the distribution of the subjective ratings for the different mixes of each song in response to task 1: 'Rate the following mixes according to the equal loudness of all sources.' For a given mix, the red horizontal line shows the median score and the white circle shows the mean score. The blue shaded circles correspond to the outlying observations.

engineers are denoted ENG1, ENG2, ENG3 and ENG4.

Focusing on medians of the mix ratings, key observations are summarised as follows:

- The anchor was consistently rated lowest and categorised as bad in every song.
- For the blues and rock genres, the GM02 auto-mixes fell in the lowest category with lower variance compared to the distributions of the non-anchor mixes. The success of this mixer appears to depend on song.
- Both RMS and EBU auto-mixes performed similarly, with 50% of the subjects scoring the mixes as at least fair/good in all but the gospel song.
- The human mixes were consistently rated in the upper half of the quality scale, performing good-excellent in most cases.
- With exception of a few outliers, the automatic mixes appear to be comparable in average rating to the human mixes for the classical and vocal quartet multitracks.

For each song a one-way analysis of variance (ANOVA) was conducted using a mixed-effects regression model with 'subject' treated as a random effect and 'mix type' as the fixed effect but inspection of residuals indicated that parametric statistics were not appropriate in all trials. In particular, the spread of residuals tended to increase in the middle range of the fitted values and were also non-normally distributed for certain songs. The homoscedasticity observed in the residuals was attributed to the unequal variance of the mix ratings as reflected by the boxplots in Figure 5.4. Instead, a Friedman test was used to assess for statistical differences between the mixes of each song. The Friedman test is based on average ranks and was therefore also deemed suitable for these data because the subjects, in addition to assigning a score to each condition, also performed an inherent ranking test (Sporer et al. 2009). By effectively treating the data as ordinal, i.e. ignoring the magnitude of the absolute differences between the scores and instead focussing on rank order, problems relating to the way in which subjects make use of the rating scale are mitigated. Pairwise comparisons were performed using the Nemenyi post-hoc test (Nemenyi 1963) which is appropriate following significance of the Friedman test.[7] The Nemenyi test compares the mean ranks of two conditions as output by the Friedman test.

The Friedman test showed a statistically significant effect of mix on the subjective *rankings* for all five songs: blues ($\chi^2(7) = 54.4$, $p < 0.0001$), classical ($\chi^2(7) = 32.2$, $p < 0.0001$), gospel ($\chi^2(7) = 54.4$, $p < 0.0001$), vocal quartet ($\chi^2(7) = 48.4$, $p < 0.0001$) and rock ($\chi^2(7) = 67.1$, $p < 0.0001$). This confirms the observations drawn from Figure 5.4 that the listeners preferred some mixes of a song over others. Table 5.1 shows the mean rank differences for all significant pairwise comparisons as determined using Nemenyi's test. Only mixes that were found to be statistically significantly different from other mixes (at the 0.05 significance level) have been included. A negative difference indicates the mix listed in the row (here called the comparison) was ranked lower than the mix listed in the column (called the reference). The number in the parentheses is the number of subjects (out of 13) that rated the comparison lower or higher than the reference, according to the sign of the mean rank difference. For example, the first data cell of Table 5.1 is -3.6 (13), meaning that all 13 listeners ranked the anchor 3.6 places lower, on average, than the RMS auto-mix.

In conjunction with Figure 5.4, the statistical analysis on ranks indicates that the three automatic mixes were ranked reliably higher than the anchor for the classical and the vocal quartet

[7] The R-package `PMCMR` (Pohlert 2016) was used to perform the Nemenyi post-hoc test. Because this test is conservative, no adjustment was made to the resulting p-values to account for multiple testing.

Song	Comparison	RMS	EBU	GM02	ENG1	ENG2	ENG3	ENG4
Blues	Anchor	-3.6*(13)	-4.1*(13)	-0.5 (8)	-5.3*(13)	-4.4*(13)	-3.6*(13)	-3.8*(13)
	GM02	-3.2*(12)	-3.7*(13)	- - - -	-4.8*(13)	-4.0*(12)	-3.1*(13)	-3.3*(13)
Classical	Anchor	-4.1*(13)	-3.9*(13)	-3.7*(13)	-4.7*(13)	-3.5*(12)	-3.7*(13)	-3.8*(13)
Gospel	Anchor	-2.5 (13)	-3.0*(13)	-1.8 (11)	-4.5*(13)	-4.5*(13)	-5.7*(13)	-4.7*(13)
	RMS	- - - -	-0.5 (7)	0.7 (5)	-2.0 (11)	-2.0 (10)	-3.2*(12)	-2.2 (11)
	GM02	-0.7 (7)	-1.2 (8)	- - - -	-2.8 (12)	-2.8 (11)	-4.0*(13)	-3.0*(12)
Quartet	Anchor	-4.7*(13)	-3.8*(13)	-3.4*(13)	-3.9*(13)	-5.3*(13)	-2.1 (13)	-4.8*(13)
	ENG3	-2.6 (10)	-1.7 (10)	-1.3 (10)	-1.8 (10)	-3.3*(12)	- - - -	-2.8 (10)
Rock	Anchor	-3.3*(12)	-2.9 (13)	-0.5 (8)	-5.8*(13)	-4.5*(13)	-5.4*(13)	-2.8 (13)
	GM02	-2.7 (12)	-2.3 (12)	- - - -	-5.3*(13)	-4.0*(13)	-4.8*(13)	-2.3 (13)
	ENG1	2.5 (12)	2.9*(12)	5.3*(13)	- - - -	1.3 (10)	0.4 (8)	3.0*(12)

Table 5.1: Mean rank difference between each comparison mix (rows) and the reference mixes (columns) for paired comparisons deemed statistically significant (denoted by *, at level 0.05) by Nemenyi's test for task 1. The number in parentheses is the number of subjects who rated the comparison mix less than or greater than the reference mix, according to whether sign of the mean rank difference is negative or positive, respectively.

songs. Recall that the anchor was deliberately mixed such that the loudness of one instrument overpowered the others, and so negative differences are to be expected. Excluding the anchor, the degree to which equal track-loudness was fulfilled by the automatic mixes was not reliably different from any other mix of these two songs. Since the automatic mixer was not prioritising any single instrument in this task, the results suggests that the engineers were more reserved when mixing these recordings. This is especially pertinent for the vocal quartet because the mixes are indeed fairly evenly balanced with respect to the perceived loudness of the four singing voices (as judged by the author). ENG3 did, however, rank reliably lower than ENG2 for this song, and the boxplot of ratings does suggest that the subjects tended to assign lower scores to this mix. In no song was the EBU mix ranked higher (on statistical grounds) than the RMS approach, which is not surprising as both loudness estimators driving the automatic faders employ a 3 s energy measurement. Furthermore, the K-filter, as used by the EBU algorithm, is generally conservative above about 100 Hz. It must be stressed that this task was not a controlled loudness-matching experiment, so only stimuli that highlight marked differences in the design and parameters of the models would likely influence the preferences of the subjects.

Most surprising is that the equal-loudness mixes of the blues, gospel and rock songs generated using the GM02 were not ranked reliably higher than their respective anchors, and tended to be ranked below the human mixes. It should be noted that 11 of the subjects did, however, rank the GM02 mix of the gospel song above than the corresponding anchor, but this difference was not found to be statistically significant. The GM02 mixer performed on par with the anchor for the blues multitrack (both mixes rated as bad and a non-significant mean rank difference of -0.5), whereas all other mixes - including those produced using the RMS and EBU - outperform it. Even for the rock song, the simple RMS detector was preferred over the GM02 and it may be surmised that this approach would probably reach statistical significance if this task was repeated with a larger sample size ($p = 0.08$).

Task 2: Mix quality ratings

In this task, the subjects were asked to rate the overall quality of eight different mixes for each of the five songs. Recall that for these automatic mixes, the level of the vocal was around +4 dB up from the equal-loudness balance set by each algorithm. The subjective ratings are shown in Figure 5.5. The trend in the medians across the different mixes within each song is generally consistent with those shown in Figure 5.4, with the exception that the interquartile ranges are

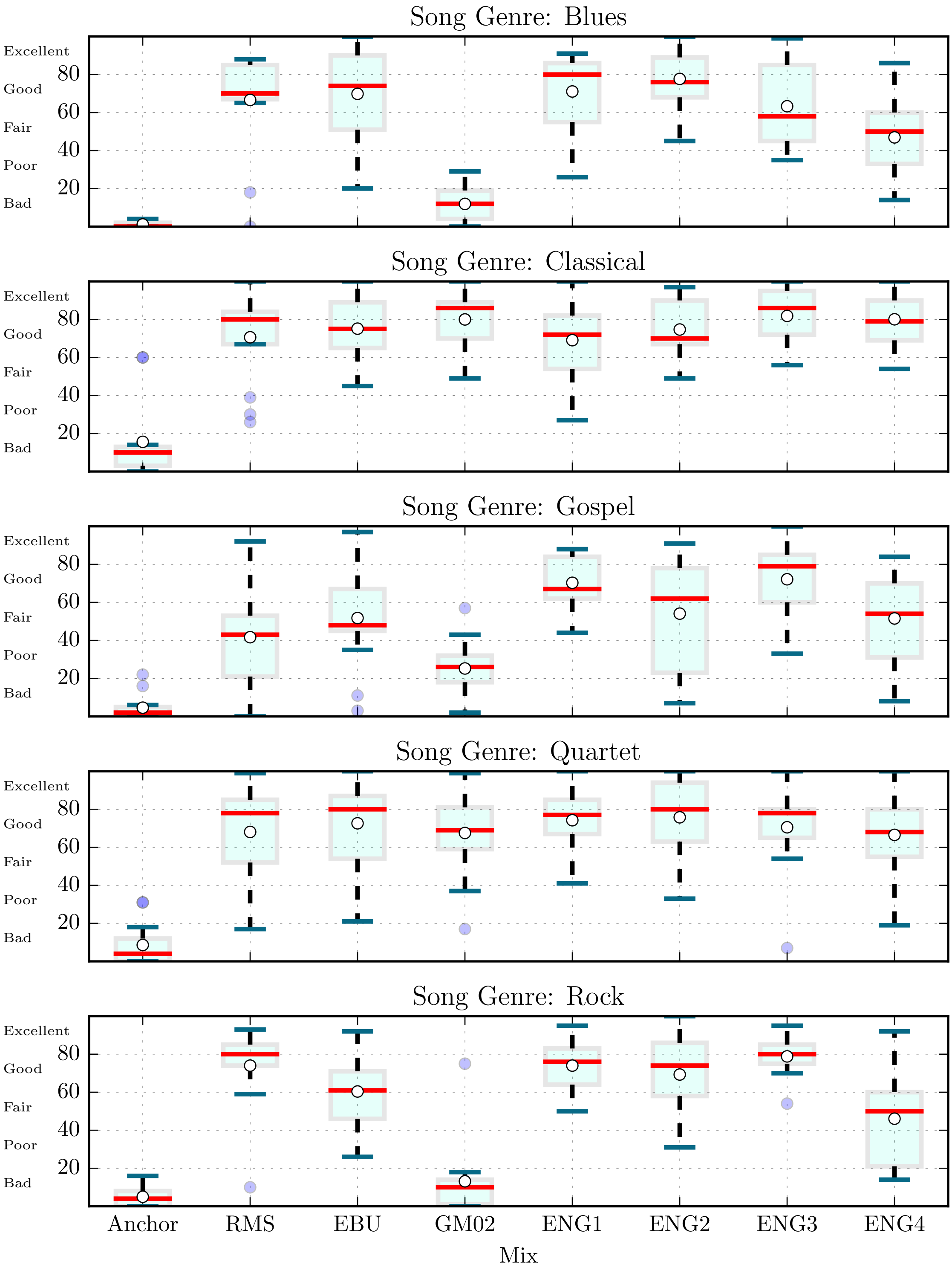


Figure 5.5: As in Figure 5.4 but for task 2: ‘Rate the following mixes according to overall quality.’

generally tighter here than those obtained in the first task. This suggests that the subjects were in greater agreement with one another when rating differences in overall quality than differences in equal loudness. Again, task 1 was not a controlled loudness-matching task and so a certain amount of inter-subject variability is to be expected.

The Friedman tests revealed a statistically significant effect of mix on the subjective rankings for all five songs: blues ($\chi^2(7) = 59.3$, $p < 0.0001$), classical ($\chi^2(7) = 37.5$, $p < 0.0001$), vocal gospel ($\chi^2(7) = 47.9$, $p < 0.0001$), quartet ($\chi^2(7) = 34.0$, $p < 0.0001$) and rock ($\chi^2(7) = 62.5$, $p < 0.0001$). Table 5.2 shows the mean rank differences for all significant pairwise comparisons as determined using Nemenyi's test. The table follows the same convention as in Table 5.1.

Song	Comparison	RMS	EBU	GM02	ENG1	ENG2	ENG3	ENG4
Blues	Anchor	-4.3*(12)	-4.8*(13)	-0.7 (10)	-4.8*(13)	-5.1*(13)	-3.9*(13)	-2.8 (13)
	GM02	-3.6*(12)	-4.2*(13)	- - - -	-4.1*(13)	-4.4*(13)	-3.2*(13)	-2.1 (13)
Classical	Anchor	-3.6*(13)	-4.0*(13)	-3.8*(13)	-3.1*(13)	-3.6*(12)	-5.1*(13)	-4.5*(13)
Gospel	Anchor	-2.6 (11)	-3.3*(12)	-1.3 (12)	-5.1*(13)	-3.7*(12)	-5.2*(13)	-3.4*(12)
	GM02	-1.3 (9)	-2.1 (11)	- - - -	-3.8*(13)	-2.4 (12)	-3.9*(12)	-2.2 (11)
Quartet	Anchor	-3.5*(13)	-4.7*(13)	-3.7*(13)	-4.3*(13)	-3.6*(13)	-4.1*(13)	-4.2*(13)
Rock	Anchor	-4.7*(13)	-3.6*(13)	-0.4 (8)	-4.4*(13)	-4.6*(13)	-5.3*(13)	-2.6 (13)
	GM02	-4.3*(13)	-3.2*(13)	- - - -	-4.0*(13)	-4.2*(13)	-4.9*(13)	-2.2 (13)

Table 5.2: As in Table 5.1 but for task 2.

As in the first task, the anchor was ranked reliably lower than the majority of other mixes for both the classical and the vocal quartet songs. No statistical difference between the remaining (non-anchor) mixes for these two multitracks was found, which is also apparent in Figure 5.5 where the median values of the ratings are generally within 20 points of one another for all but the anchor.

No reliable difference in mean rank was found between the RMS and EBU automatic mixes for any of the songs. This further supports the idea that the K-filter may not be necessary for practical applications such as automatic music mixing, although further tests involving a broader selection of music, preferably comprising instrumentation spanning a wide frequency range, should be conducted to fully address this. A more pronounced high-shelf filter, such as the modified K-filter proposed by Pestana et al. (2013), may yield superior performance over an unweighted loudness measurement. Importantly, the RMS and EBU mixes were not deemed to be statistically different compared to the four human mixes in any of the five songs, although there is considerable variability in the ratings. In contrast, the GM02 mixes of the blues and rock songs were both ranked reliably lower in overall quality compared to the automatic mixes generated using the RMS and EBU models, as well as those mixed by three of the engineers (mean rank differences are negative for all comparisons). Note that the mean rank of ENG4 was not found to be systematically different from that of the GM02 mix for either the blues or rock songs, although there is a fair amount of variability in the ratings for of the former compared to the latter. Table 5.2 does, however, show that all 13 subjects ranked the GM02 blues and rock mixes lower than the those of ENG4, but statistical significance was not observed using Nemenyi's paired test. Presumably, the inter-subject variability in the rankings yields too much uncertainty to identify a reliable difference despite every subject assigning a higher rank to ENG4. The GM02 mix of the gospel song was ranked reliably lower than ENG1 and ENG3, and judging by boxplots, the ratings are indicative of inferior quality compared to the RMS and EBU approaches (albeit not statistically significant). Again, the majority of subjects rated the GM02 lower for this song.

5.5 Further analyses

This section presents two investigations comparing the predictions of different loudness algorithms within the context of automatic level balancing. First, the relative gains set by the real-time automatic mixers for three songs used in the subjective evaluation are compared with those of a human mixing engineer to better understand the relationship between the fader balance and the subjective preference of a given mix. Second, more sophisticated offline approaches to loudness estimation are used to extract the relative-loudness balance of the same three songs but using the entire dataset collected from all ten human engineers. The purpose of this second analysis was to examine three offline approaches commonly used to describe the loudness relations of the instruments comprising a mix in order to give further insight into inter-model differences when describing a mix.

5.5.1 Comparison of gain settings

The GM02-based mixes of a blues and a rock song were rated poorly by 13 listeners when asked to judge different mixes in terms of two criteria: equal track-loudness and overall mix quality. For the blues song, the GM02 mixes were ranked lower than two other automatic mixes generated using the same mixing framework but simplified measures of loudness (RMS and EBU features). Furthermore, no evidence was found to suggest that the listeners preferred the GM02 blues and rock mixes over their respective anchors—a mix in which equal track-loudness was purposefully violated. Even though the statistical analyses are based on mean rank differences, the boxplots do show that the listeners tended to rate the GM02 and anchor mixes in the lowest category for these two songs, i.e. the GM02 and anchor mixes received comparable scores. Unlike the RMS and EBU algorithms, the GM02 is a hearing model grounded in established auditory theory and was developed to predict a range of perceptual phenomena (see Chapter 3 for a review). It is therefore both surprising and unclear as to why the subjective quality of the automatic mixes degraded when this model was utilised to inform the faders.

As a preliminary investigation a subset of the mixes were analysed: the RMS, the GM02 and the ENG1 gain settings were compared for the blues, rock and vocal quartet songs. The mixes produced by ENG1 were selected because they were rated highly for each song, and the RMS-based mixes were selected because, unlike the GM02, this configuration outperformed the anchor for the blues and rock songs in both tasks. Additionally, the automatic blues mix generated using the simple RMS detector was also found to be superior to the one produced using the GM02. Tables 5.3, 5.4 and 5.5 show the zero-centred gains set by the different mixers for the vocal quartet, the blues song and the rock song, respectively. For the automatic mixes, these levels were obtained by averaging the combined gain of the pre-amp and fader applied to each stem overtime, and then subtracting the mean of the track levels. Since the engineers did not automate the temporal profile of the faders, the static levels were zero-centred to allow for comparison. For reference, recall that the input tracks were peak normalised prior to any mixing.

As expected from the rating data, the four levels used to balance the vocal quartet are similar across the three mixes: the levels are within 1.1 dB of one another. Considering that the task 1 ratings of these mixes were much higher than those of the anchor, and no single track appears to take priority with respect to the original peak-normalised inputs, it is fair to assume equal track-loudness. Indeed, the average (and standard deviation (SD)) of the amplitude envelope derived using the RMS mixer for the bass, baritone, tenor and lead voice was -21.8 (1.2), -21.8 (0.8), -21.7 (0.8) and -21.6 dB FS (0.7), respectively. For the GM02, the averages (and SDs) of the loudness-level envelopes for the same tracks were 89.0 (1.2), 88.9 (1.3), 89.0 (1.4) and 89.1 phons (1.3). Taken together, when the input tracks share similar physical characteristics, such as those in a four piece male vocal harmony, equal-RMS mixing gives a good approximation to equal-loudness mixing. This is consistent with Bech and Zacharov (2006) who recommend a sound-pressure metric for aligning the loudness of sounds with similar spectra.

The gains of the RMS mixer are more in line with those set by the human engineer for both the blues and rock songs, although there are a few notable discrepancies. For example, the equal-RMS

	Bass	Baritone	Tenor	Lead
RMS	0.1	-0.8	0.4	0.3
GM02	1.2	-0.3	-0.7	-0.2
ENG1	0.35	-0.35	-0.35	0.35

Table 5.3: Gain values (dB) applied to the peak-normalised multitrack recording of a vocal quartet as set by the automatic mixing system targeting equal RMS level or equal GM02 LTL, and by a human mixing engineer.

	Kick	Drums	Bass	A.Guitar	E.Guitar	Keys	Voice
RMS	1.5	7.0	-2.0	3.5	-3.3	-1.0	-5.7
GM02	4.8	3.2	10.6	-2.0	-5.8	-3.0	-7.8
ENG1	2.5	2.8	-3.2	1.6	-3.9	1.8	-1.7

Table 5.4: As Table 5.3 but for the blues song.

	Kick	Snare	Drums	Bass	A.Guitar	E.Guitar	Keys	Voice
RMS	3.4	11.4	-0.8	-3.1	0.8	-5.1	-6.6	-0.1
GM02	12.1	6.2	-6.7	10.1	-2.4	-8.8	-9.3	-1.2
ENG1	1.3	4.1	-0.2	-2.1	0.7	-3.1	-4.1	3.3

Table 5.5: As Table 5.3 but for the rock song.

mixer set the drum overhead 4.2 dB higher than the human engineer in the blues song, and set the snare drum 7.3 dB higher in the rock song. Indeed, a few subjects did report that the snare drum was too loud but not overly problematic when listening to the automatic rock mix. Note that the relative level of the drum overhead in the rock song was only 0.6 dB below the engineers. The overhead recording in the blues song was, however, dominated by the snare drum, in contrast to the overhead recording in the rock mix which was captured using a distant room microphone.

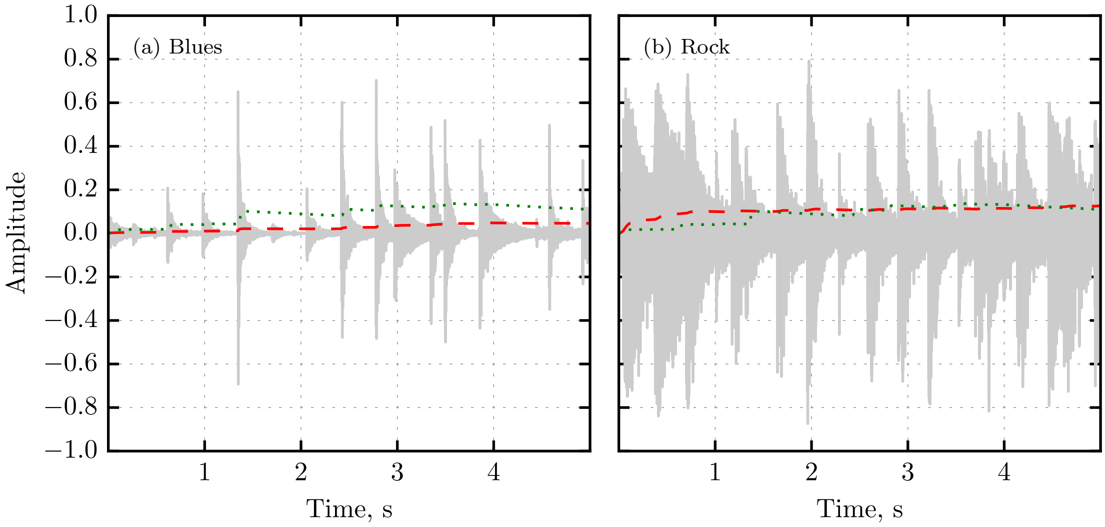


Figure 5.6: Amplitude waveform of the drum overhead recording used in (a) the blues song and (b) the rock song. Red dashed and green dotted lines show the two amplitude envelopes of each recording, approximating those used by RMS and GM02 mixers, respectively (see text for details).

This is demonstrated in Figure 5.6 which shows the amplitude waveform of the drum overhead in each song. The denser waveform shown in subplot (b; Rock) is attributed to a combination of faster tempo, more frequent use of cymbals and hi-hat, and greater room reverberation. The waveforms are overlaid with two amplitude envelopes: the red dashed line denotes the envelope obtained using a RMS detector with a 3 s time constant, and the green dotted line denotes the envelope calculated using two cascaded asymmetrical lowpass filters with time constants as per the GM02. Note that the two amplitude envelopes are most separated (≈ 10 dB) for the drums in the blues song, shown in subplot (a). This occurs because the asymmetrical smoothing filter assigns greater weight to new input values and thus, compared to the 3 s RMS detector, responds more quickly to the attack of the snare drum. This helps to understand why the RMS- (and EBU-) driven auto-mixer generally boosts the level of the drum overhead track in the blues song and the snare drum track in the rock song to a greater extent than the GM02-driven mixer.

In both the blues and the rock songs, the RMS and GM02 mixers assign less weight to the voice compared to the mixing engineer. Further inspection found this to be true across all human mixes collected for the assessment (see Section 5.5.2). This is not surprising, given that the engineers were simply asked to produce a well-balanced mix, rather than target equal track-loudness as done by the automatic mixers in task 1. The GM02 mixer assigns a great deal of gain to the bass in both songs, and considerable weight to the kick drum in the rock track. Indeed, some participants reported that the GM02 mixes of the blues, rock and gospel songs were notably 'bass heavy' compared to the other test items.

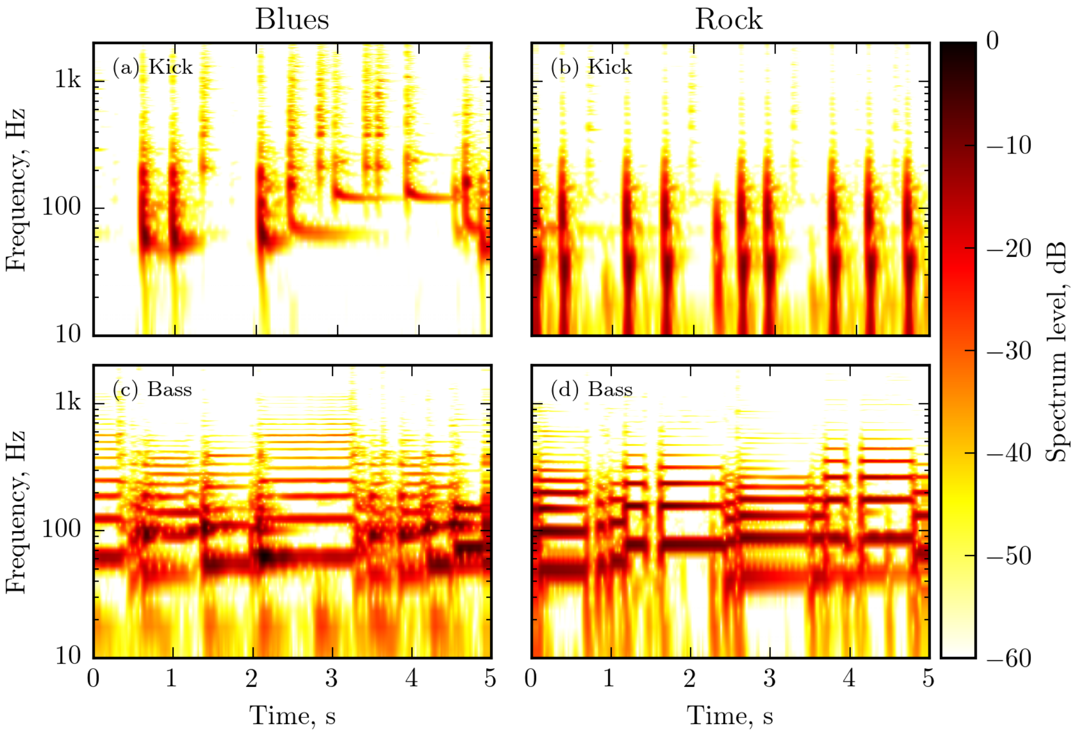


Figure 5.7: Power spectrograms of (a)–(b) kick drum and (c)–(d) bass guitar excerpts taken from the blues and rock songs. The spectrum levels are relative to the maximum power measured across all four spectrograms.

Power spectrograms of the kick drum and bass guitar from the blues and rock songs are shown in Figure 5.7. In both songs, the kick drum spectra shown in subplots (a) and (b) are broadest in frequency at the transient and, in the case of the blues song, there are noticeable (and audible) resonances around 55–65 Hz. The kick drum in the rock song is more transient with very little resonance, and the energy is concentrated in two central regions: 0–45 Hz and 70–150 Hz. The spectrograms of the bass guitars shown in subplots (c) and (d) show clear harmonics up to about 1 kHz, with fundamental frequencies of individual notes spanning 40–120 Hz. Considering that the GM02-driven system performed better when mixing the vocal quartet and classical multitracks, and the respective gain settings are comparable to those extracted from the other mixes, these findings suggest that the GM02 algorithm significantly underestimates the loudness of low-frequency instruments relative to instruments that occupy a broader range of frequencies in the mid to upper-frequency region.

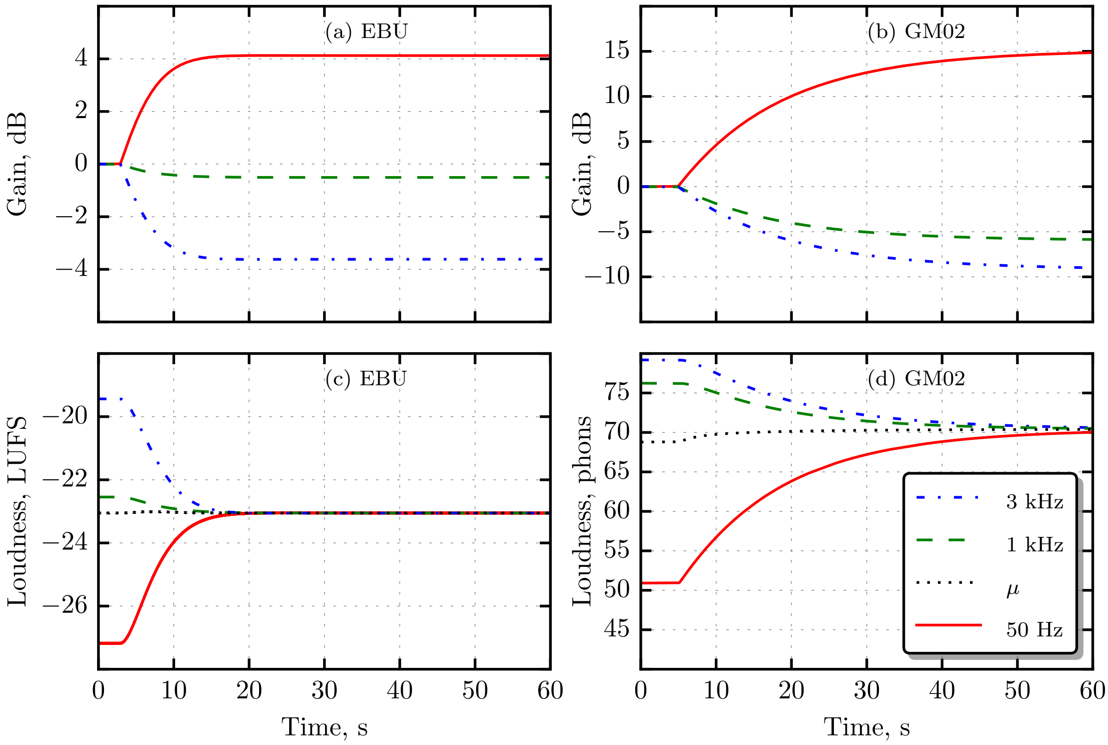


Figure 5.8: Gain (top) and loudness (bottom) contours output by the automatic mixer, using the EBU or GM02 loudness algorithm, in response to three pure tones each presented at 78 dB SPL. Note the difference of scale in vertical axes between subplots (c) and (d).

Such frequency-dependent effects are further demonstrated in Figure 5.8 which shows the gain and loudness contours when the feedback-mode auto-mixer was instructed to equate the loudness of three pure tones with frequencies of 50 Hz, 1 kHz and 3 kHz, each presented at an effective cochlear level of 78 dB SPL, using either the EBU or GM02 loudness models.[8] The outer-ear filter required by the GM02 was set to the transfer function of a pair of BeyerDynamic DT990 headphones. The EBU mixer was included here to highlight differences in frequency sensitivity between the models—the RMS-driven mixer would not modify the faders since no frequency weighting is used. Upper subplots (a) and (b) show that both systems assign progressively less weight to the loudness measurement as the frequency of the tone decreases, resulting in greater gain to compensate. This is in accordance with the frequency sensitivity of the human ear as expressed by the equal-loudness contours (ISO 226 2003). As a result, the mixer—in both configurations - attenuates the 1 kHz and 3 kHz tones, and boosts the 50 Hz tone in order to target the average loudness shown by the black dotted line in subplots (c) and (d). Note that the convergence times and temporal profiles of the gains differ between the two models which is primarily attributed to the differences in the way that signal intensity is averaged over time. Both algorithms predict a level change of 3–4 dB to match the loudness of the 1 kHz and 3 kHz tones. However, the levels differ considerably when

[8] Note that the GM02 does not output 78 phons for the 1 kHz because of the middle-ear transmission loss of 2.6 dB.

the 50 Hz tone is considered. The EBU mixer estimates an 8 dB level difference to equate the loudness of the 50 Hz tone with that of the 3 kHz tone, whereas the GM02 estimates a level change of about 25 dB. This illustrates that the GM02 is markedly less sensitive to very low frequencies compared to the EBU algorithm which employs the same frequency weighting as used in ITU-R BS.1770 (2015) (refer to Figure 3.3). Considering that the RMS-driven automatic mixer is frequency insensitive, this gives some indication as to why the GM02 mixer assigned very high gains to the bass instruments used in this study.

5.5.2 Comparison of offline loudness model mix balances

This section analyses the relative-loudness balances of the blues, rock and gospel songs as mixed by ten human engineers. Here, the definition of the relative-loudness balance follows the one proposed by De Man et al. (2014):

$$L'_k = L_k - L_{\text{mix}}, \tag{5.16}$$

where L_k is the predicted *global loudness* of track k, and L_{mix} is the global loudness of the mix. The loudness levels of each track are therefore expressed relative to the loudness of the mix, meaning that louder instruments are closer to zero. For models that measure loudness in sones, the global-loudness readings are transformed to a vector of loudness levels in phons, which are then normalised using Equation 5.16.

Recall that for task 1 of the subjective evaluation, the mixers were configured to target equal track-loudness, i.e. $L'_k = 0$ for all k. Did the human engineers mix according to this criterion? To give some insight into this question and broaden the analysis beyond the implementation of the real-time mixing system, relative-loudness balances were extracted using different offline variants of the EBU and GM02 models. These descriptions of mix perception can then be compared to identify fundamental differences between the loudness estimation procedures and therefore how they might influence the subjective quality of a computer-generated mix.

Three offline loudness algorithms were used to compute the relative-loudness balances: the EBU, the GM02 and the GM02 with consideration of inter-track masking (denoted GM02-ITM). The latter model refers to the partial-loudness variant of the GM02, details of which can be found in Section 3.3.5 (see also Glasberg and Moore (2005)). This model outputs two time series: the short-term partial loudness (STPL) and the long-term partial loudness (LTPL).

The global loudness of each track in a given song was measured using two different descriptors, as shown in Table 5.6. For the EBU model, the integrated momentary loudness (ML), also called programme loudness (EBU R 128 2014), is identical to that of the ITU-R BS.1770 (2015) measurement, whilst the integrated STL is similar but makes use of a 3 s rectangular window rather than a 400 ms length window. This was done to approximate the temporal averaging used by the EBU meter in the real-time mixer presented in Section 5.3. The peak STL/STPL and mean LTL/LTPL features were used for the two GM02 algorithms, as these two global-loudness descriptors are associated with this model (Glasberg and Moore 2002) and have been used in previous work (Ward et al. 2012; Terrell 2013, Chapter 7).

Figure 5.9 shows the relative-loudness balances of the three songs according to each model/descriptor combination. The six configurations reveal that the lead vocal (L.Vox) is generally the loudest instrument in each song and that, from comparing the 95% confidence intervals (CIs), the between-subject variance is the lowest. This reduced variance is to be expected considering that the lead vocal dominates the mix. For the integrated ML (BS.1770 measure), the vocal-to-mix level was -1.9, -3.7 and -1.6 LU for the blues, rock and gospel song, respectively. The peak STL level of these three lead vocals, as estimated by the GM02, followed a similar magnitude and trend: -3.3

Model	Descriptor	Details
EBU	Integrated ML Integrated STL	Integrated ML according to ITU-R BS.1770 ITU-R BS.1770 type gating applied to STL K-filter Time step of 2 ms
GM02	Peak STL Mean LTL	STL attack time: 22 ms STL release time: 50 ms LTL attack time: 100 ms LTL release time: 2 s Headphone transmission accounted for Filter spacing of 0.5 Cams Spectral compression criterion of 0.2 Cams Time step of 2 ms
GM02-ITM	Peak STPL Mean LTPL	As above but using partial loudness based on ITM

Table 5.6: Details of the offline loudness models used to compute the relative-loudness balances of the human mixes.

(blues), -4.6 (rock) and -3.3 phons (gospel). For the blues and rock songs, the drum elements tend to be associated with greater variance compared to the other instruments. Importantly, all three models, using either descriptor, indicate that the different tracks are not equally loud, although it should be noted that drum components are inter-related due to microphone bleed. This suggests that, for these three songs, equal track-loudness is not a realistic mixing objective.

Considering the variability in the data, the two EBU descriptors are consistent with one another, although the short-term measure gives slightly lower estimates, differing by 1–2 dB in some cases. This is attributed to the larger temporal window used to calculate signal energy. Greater differences can be expected for percussive instruments or rhythmic elements with relatively long silences between successive hits/notes where the loudness of individual beats would be underestimated by the 3 s average. The choice of loudness descriptor appears to be more important for the two GM02 models, with the peak short-term measures generally leading to higher relative levels compared to the mean long-term measures. This is particularly relevant for the drum and percussive elements, e.g. the djembe and high-frequency percussion (H.Perc) in the gospel song, where there is strong contrast between the peak STL and mean LTL. Such discrepancies are explained by the short attack time of the STL integrator and the use of the maximum (rather than the mean) to quantify overall loudness.

Differences in the magnitude of the balances are most pronounced when comparing those of the GM02-ITM with those of the other two models. For example, the GM02-ITM predicts that the acoustic guitar (A.Guitar) in the rock song is about 27 phon softer than the loudness of the mix (peak STPL). In comparison, the GM02 predicts a 6.5 phon difference. Such large differences reflect the amount of masking, and thus reduction in loudness, calculated by the GM02-ITM given the different instruments comprising the mix. Note that the partial-masking effect is not constant across the different instruments, and as a result, the objective measures of relative loudness between different pairs of instruments are inconsistent with those of the GM02. In other words, the balance predicted by the GM02-ITM is not simply the GM02 balance negatively offset by a fixed amount. This effect is attributed to the spectro-temporal relations between the EPs of the different instruments. The GM02-ITM predicts a huge masking (a 50–60 phon loudness reduction) of the kick drum in both the blues and rock songs, which is likely due to the interactions between the kick drum and drum overhead tracks. Indeed, relatively high correlations between the power envelopes (in dB) of the kick drum and drum overhead tracks were measured for the blues song

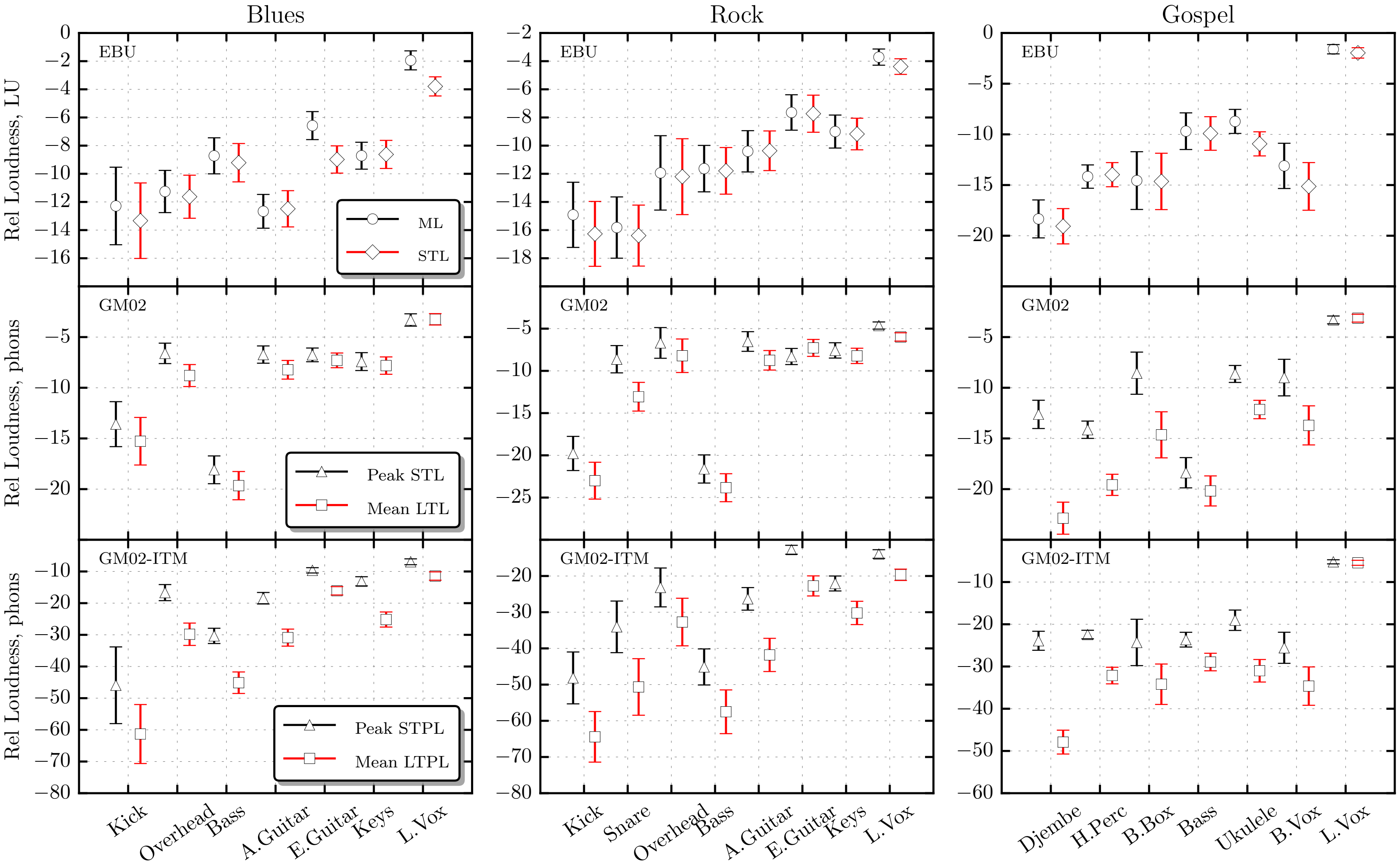


Figure 5.9: Relative-loudness balances of the blues (left), rock (middle) and gospel (right) songs according to three loudness models each configured with two different measures of global loudness. Error bars show the 95% between-subject CI of each mean.

($r = 0.69$) and the rock song ($r = 0.79$). The GM02-ITM predicts the least amount of masking for the lead vocal in all three songs, with the greatest reduction in vocal loudness calculated in the rock song. Such instrument-dependent masking explains the greater variance in the relative loudness levels according to the GM02-ITM (compare the scale of the ordinates across models for the blues and rock songs). The SD of the mean relative-loudness balance of the GM02 for the blues, rock and gospel song was 5.1, 6.5 and 4.8 phons, respectively. The corresponding SDs of the mean balances generated by the GM02-ITM were 13.6, 13.2 and 7.1 phons.

The GM02 predicts that the mix is roughly 20 phons louder than the bass guitar in all three songs (regardless of descriptor), which corresponds to a loudness ratio of four ($2^{20/10}$). For the blues and rock songs, the GM02 predicts a very low contribution from both the kick drum and the bass guitar to the mix. This is further emphasised by comparing, in these two songs, the level differences between the bass and acoustic guitars determined using the GM02 with the level differences measured using the EBU. In both cases the GM02 predicts that the bass guitar is markedly softer (10–15 phons, both descriptors) than the acoustic guitar. The EBU does not predict as large a difference, and in fact measures the bass to be around 4 dB *louder* than the acoustic guitar in the blues song. The inclusion of energetic-masking effects does not appear to reduce this effect: the bass guitar is still estimated to play a minor role in the mix. For the gospel song, however, the consideration of ITM does appear to align the levels of the bass guitar, ukulele and backing vocals (B.Vox). It should be noted that there is a considerable difference (about 30 phon for mean LTPL) between the loudness of the lead vocal and the other tracks.

5.6 Discussion

A real-time automatic mixer has been developed to control the levels of a multitrack recording according to the loudness of each track. The mixer accommodates different loudness models and arbitrary relative-loudness balances. For nonlinear level-dependent loudness estimators, the mixer employs a feedback topology to realise the desired relative-loudness balance regardless of listening level. The system could easily be extended to optimise the balance according to a target mix loudness as done by Terrell et al. (2013). The user could specify, for example, ‘set the overall mix loudness to 16 sones’. Instead, the real-time system here simply ignores total loudness and uses the time-varying average loudness to find the optimal fader levels, but does allow the user to increase or decrease the master gain during playback. Overall mix loudness was not used here because (a) the computational cost of running multiple auditory models simultaneously is expensive, and (b) the additional constraint may not be warranted in practice. Regarding the first point, the average CPU usage was around 5% in the case of an 8-track EBU-based mixer, which increased to about 45% when using the GM02.[9]

5.6.1 Empirical findings

The performance of the automatic mixer was assessed by conducting a listening assessment in which subjects compared computer-generated mixes of five songs with mixes produced by four human mixing engineers taken from a larger dataset. For the automatic mixes, the faders were controlled using either an RMS-level detector, the EBU STL or the GM02 LTL. The EBU STL measure was employed in the automatic fader controller developed by Mansbridge et al. (2012b), and the latter was proposed by Glasberg and Moore (2002) for measuring the loudness of relatively long fluctuating signals such as speech.

[9] MacBook Pro Intel Core Duo @ 2.8 GHz with 8 GB Ram running SuperCollider 3.7 on Ubuntu 14.04.

In the first task, subjects were required to assess the mixes in terms of the perceived loudness equality across instruments. Subjects were then asked to rate the overall quality of the mixes. The automatic mixes performed as well as the human mixes for the two multitracks with the least number of tracks, where the dynamics and spectral content were relatively homogeneous across the instruments. For the songs encompassing a wider range of instruments, the RMS and EBU level-setting approaches continued to perform fairly successfully in terms of overall quality, although further testing is required to truly gauge the magnitude of the perceived differences between the automatic and human mixes.

The GM02-based mixes scored poorly in both tasks for songs involving bass instruments. In combination with subject feedback, an inspection and comparison of the mixer settings indicated that this EP-based model underestimates the relative loudness of bass. This was also supported by the relative-loudness balances extracted from the complete dataset involving the mixes from all 10 engineers, using offline variants of the same algorithm (see the middle subplots in Figure 5.9). This phenomenon was also reported by Ward et al. (2012) who used the GM02 to automatically control the levels of multitrack recordings using an offline-noncausal approach. More recently, Wichern et al. (2015) compared the static-fader mixes generated using the ITU-R BS.1770 (2015) algorithm (similar to the EBU mater tested here), and two variants of the GM02 algorithm: one with consideration of ITM (Glasberg and Moore 2005) and the other without. The authors used instrument-specific target gains rather than the equal-loudness mixing objective, in accordance with their finding that human engineers generally do not mix for equal track-loudness. In an analysis of 48 multitrack recordings, they discovered that the relative loudness of the instruments within a mix were predicted differently by the three models, especially for the electric bass where both adaptations of the GM02 gave lower measures of relative loudness compared to the BS.1770 algorithm. In the final subjective evaluation, the authors found that listeners preferred mixes generated using BS.1770 over the EP-based models, parallel to the findings presented here.

Relative-loudness balances of the three most instrument-diverse songs as mixed by ten engineers were extracted using variants of the EBU and GM02 algorithms. The balances were in agreement with the conclusions drawn from other work supporting the idea that engineers do not mix for equal instrument loudness (Terrell 2013, Chapter 8; Pestana 2013; De Man et al. 2014; Pestana and Reiss 2014; Wichern et al. 2015; Wilson and Fazenda 2015). Nonetheless, in absence of target balances or mix templates, the equal-loudness mixing objective appears to be a robust first approximation when using the RMS or EBU features. The vocal-to-mix loudness level differences measured using the integrated ML approach, i.e. the BS.1770 measure, were also consistent with those reported by others. The balances produced by the GM02 with and without ITM also supported a prioritisation of the lead vocal, but the variance in the balances (across instruments) was notably larger than those generated using the EBU, especially when modelling masking. As mentioned, the GM02 balances, with and without ITM, predicted a very low contribution from the kick drum and bass guitar in the blues and rock songs. This effect may be responsible for the poor performance of the GM02 in the listening assessment.

The two main global-loudness descriptors proposed for the GM02 also play an important role when describing mix balance. For example, peak STL assigns greater weight to transients, making the relative levels of percussive instruments higher than those measured using the mean LTL. Further improvements to the summary statistic used to quantify the global loudness of multitrack audio is warranted. For the models used in the real-time mixing system, relatively long time averaging was used to obtain a smooth time-varying loudness estimate. However, this can lead to an underestimation in the loudness of transient signals where peak loudness is salient. For the automatic mixer, one must also consider the temporal smoothing applied to the fader control signal post loudness measurement. One approach is to combine the dynamic gate proposed here with a faster changing loudness, such as the STL generated by the GM02, and formulate a cumulative-probability distribution for each source. When the gate is open, signifying musical activity, the STL should populate a cumulative histogram from which percentiles can be estimated. A moving circular buffer could also be used to compute a running percentile based on active frames over a given time period.[10] Alternatively, a peak detector could be employed to assign greater weight to loudness maxima. In addition, the fully-fledged EBU integrated measurement (EBU Tech 3341 2016) which employs a dynamic adaptive gate, could also be incorporated into the real-time mixing device.

5.6.2 Partial loudness

The loudness features used to control the fader levels did not account for partial loudness: the perceived loudness of a sound in the presence of other sounds. Thus, when masking occurs, greater signal gain is needed to restore the audibility of the maskee and therefore restore its loudness. At present, only EP-based loudness models that take into account frequency information, such as the partial loudness version of the GM02 (Glasberg and Moore 2005), are able to simulate such effects. However, the relative-loudness balances shown in the bottom subplots of Figure 5.9 suggest huge reductions in loudness when real sources are combined. Wichern et al. (2015) observed similar reductions when analysing a much larger set of multitrack audio using the GM02 with masking calculations. An inflation of the variance in relative loudness levels (both across and within instrument) was observed when incorporating ITM, an effect also observed in the relative balances extracted in this chapter.

The EP models assume that the loudness of a signal in 'noise' has little to no contribution to the overall loudness as the signal falls below the masked threshold (Moore et al. 1997; Chen et al. 2011). This approach gives accurate predictions when applied to tones in stationary noise, but how about when loudness inhibition is less apparent? There are many real-world situations where two different signals are competing to be heard, but their individual loudnesses are not necessarily reduced due to uncertainty in the overall percept. Indeed, Ma et al. (2014) found that the GM02 with ITM predicted too much energetic masking (and thus loudness reduction) when compared against subjective loudness matches of musical sounds. The authors reduced the assumed signal-to-noise ratio of the auditory filters at the masked threshold by as much as 10 dB to obtain a better agreement with the experimental data. This effectively allows a greater amount of excitation to be transferred to the loudness transformation. Furthermore, when listening in real-world environments, additional masking—unrelated to energetic masking—called informational masking takes place which influences the ability to detect a signal (Durlach et al. 2003). Glasberg and Moore (2005) used their partial loudness procedure to predict the *audibility* (not loudness) of real-world sounds presented in different background sounds and discovered discrepancies that were attributable to informational masking. It seems therefore that a sophisticated set of stimulus and context-dependent features are necessary to model the perceived loudness of multiple sounds interacting over time. Bradter and Hobohm (2008) acknowledged the need to incorporate more advanced features such as gestalt principles (Bregman 1994) when computing the loudness of single objects in a mixture. Bradter and Hobohm (2008, p. 6) presented a simple strategy to deal with two real-world sounds:

> 'When masking conditions [between two EPs] are obvious, masked excitation of the object of interest is discarded. If masking is not obvious, either excitation of the

[10] The P^2 algorithm (Jain and Chlamtac 1985) would be a fitting choice for real-time percentile tracking.

mixture of both objects is considered, or special rules for excitation selection have to be applied.'

Demonstration

To demonstrate some characteristics of current partial loudness models when applied to musical sounds, loudness features were extracted from two 3 second segments of an acoustic and electric guitar taken from an acoustic rock song. The two segments were first normalised to an RMS level of 75 dB SPL and then loudness matched according to the ITU-R BS.1770 (2015), which calculated a slight level correction of 0.3 dB. These two segments were judged by the author to be perceptually matched in loudness. The segments were then input to two dynamic partial loudness models: the GM02 (Glasberg and Moore 2002, 2005) and Chen and Hu's 2012 dynamic loudness model (CH12) (Chen et al. 2011; Chen and Hu 2012). For a detailed account of these two algorithms refer to Section 3.3.5. Estimates of global loudness for each sound in quiet Ψ and when mixed together Ψ', as calculated using the mean LTL and mean LTPL, respectively, are given in the upper partition of Table 5.7.

Instrument	GM02		CH12		Unit
	Ψ	Ψ'	Ψ	Ψ'	
Acoustic	20.8	8.5	18.2	7.6	Sones
Electric	20.1	11.2	18.0	9.9	
Acoustic	83.7	70.7	78.6	70.4	Phons
Electric	83.2	74.8	78.5	73.2	
	$\Delta L(\rightarrow)$	$\Delta L(\leftarrow)$	$\Delta L(\rightarrow)$	$\Delta L(\leftarrow)$	dB
Acoustic	-14.2	7.3	-11.7	4.6	
Electric	-9.7	6.3	-8.3	4.5	

Table 5.7: Loudness Ψ and partial loudness Ψ' predictions of two 3 s guitar segments, with intensity level differences ΔL required to match the loudness of the unmasked sound to that of the masked sound ($\rightarrow$) or vice versa ($\leftarrow$).

In agreement with the BS.1770 measure, both models predict that the two guitars are roughly equally loud (within 0.5 phon) when using the average LTL. The two models also predict similar absolute loudness values (in sones) for the sounds in quiet, although the corresponding loudness levels (in phons) are less in line with one another. This is because the loudness growth function of a 1 kHz tone, as predicted by the CH12, grows more rapidly at higher levels compared to the GM02 (see Figure 3.17). This means that when both models predict the same loudness ratios between two sounds, the CH12 generates smaller loudness level differences. When the two guitars are mixed together, the GM02 predicts reductions in global loudness of 12.3 and 8.9 phons for the acoustic and electric guitar, respectively. The CH12 predicts reductions of 8.2 and 5.3 phons. Unlike the GM02, these reductions in loudness level are lower than expected based on the loudness ratios alone, e.g. $10\log_2(18.2/7.6) = 12.6$ phons, where 18.2 is the loudness of the acoustic guitar presented in isolation and 7.6 is the corresponding partial loudness (both in sones). Again, this is due to differences in the way that loudness grows with intensity within the two models. After masking, the CH12 predicts that the electric guitar is about 3 phon louder than the acoustic guitar, whereas the GM02 predicts a slightly higher difference of about 4 phon.

The bottom portion of Table 5.7 shows the intensity level differences ΔL (in dB) between each guitar when presented in isolation and when mixed together, at the point of equal loudness, and vice versa. The gain corrections were determined using an iterative procedure. The level differences are negative when going from unmasked to masked loudness ($\Delta L(\rightarrow)$), and positive when going from masked to unmasked loudness ($\Delta L(\leftarrow)$). The important point is that the level differences are asymmetrical, being greater in magnitude for the unmasked to masked adjustment. This is explained by the faster rate of growth in loudness of the signal when under inhibition: an increment in signal level under masking yields a greater change in loudness level than that introduced by the same level change in absence of the masker.

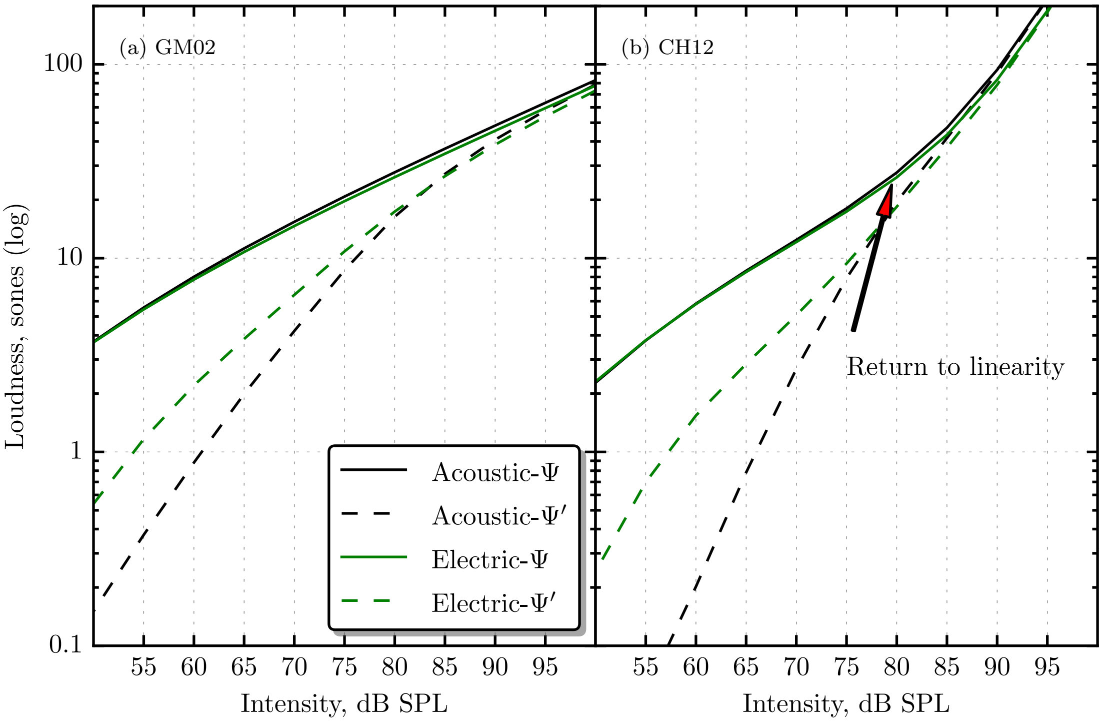


Figure 5.10: Loudness Ψ and partial loudness Ψ' growth functions of the acoustic and electric guitar when masked by one another according to (a) the GM02 and (b) the CH12. Each loudness function was constructed by setting the level of the masker to 75 dB SPL, and varying the level of the signal from 50 to 100 dB SPL in steps of 5 dB.

The loudness versus intensity level functions of both guitar segments, as measured when the other masking sound is fixed at 75 dB SPL, are plotted for the GM02 and CH12 in Figure 5.10. It can be seen that partial loudness rises more rapidly compared to loudness in quiet, and that the two functions, for a given sound, converge at around 100 dB SPL for the GM02 and 90 dB SPL for the CH12. The difference in slope between the loudness and partial loudness growth functions of each stimulus is responsible for the asymmetry observed in the ΔL values given in Table 5.7. The figure also highlights the increase in slope of the loudness growth function of the CH12 as the sound level exceeds 75–80 dB SPL. This is explained by the diminishing compression of the active component of the double-rounded exponential (roex) filter, formulating the filter bank of the CH12, at relatively high levels. In contrast, the GM02 maintains compression at high sound levels (ANSI S3.4 2007).

Informal listening comparisons performed by the author suggest that when matching the loudness of each guitar to its corresponding partial loudness, a level difference of about -4 dB and -2 dB seems appropriate for the acoustic and electric guitar, respectively. These gain adjustments are more in line with those predicted when the masked instrument is varied ($\Delta L(\leftarrow)$ in Table 5.7), owing to the faster rate of change in inhibited loudness growth. The loudness matching task is not easy due to the onsets of the two instruments coinciding, in addition to both instruments sharing characteristic resonances near 120 Hz. Despite this, the masking effect clearly has a more pronounced effect on the timbre and loudness of the acoustic guitar, as captured by the models.

Qualitatively, the predicted partial loudness of both signals are underestimated, indicating that the partial specific loudness (PSL) transformation derived from experiments involving laboratory stimuli may not generalise to more complex sounds where there is room for ambiguity and/or a plethora of other perceptual cues are available to the listener. The over-sensitivity to the effects of inter-source excitation interaction on loudness is supported by more quantitate studies (Ma et al. 2014), as well as the unrealistic relative-loudness balances of multitrack audio previously measured. It should be remembered that the partial loudness calculation procedures are founded on comparisons of instantaneous EPs, and do not take into account correlations between the signals involved. This is particularly important when considering multitrack projects comprising highly-correlated signals. In such cases, a grouping mechanism may be necessary to organise the audio content prior to analysis.

5.7 Summary

In this chapter, the state-of-the-art in automatic music mixing was further developed by establishing a real-time mixing system that functions independently of the features used to inform the cross-adaptive signal processor. The mixer adjusts the fader levels to achieve a user-specified relative-loudness balance and accommodates both heuristic single-band loudness algorithms and more elaborate level-dependent psychoacoustic models. The system was demonstrated using both types of model and evaluated by means of a subjective listening test. The single-band algorithms were based on the EBU Tech 3341 (2016) short-term loudness meter, both with and without a perceptually-motivated frequency weighting, known as the K-filter. The other type of model was an efficient adaptation of the excitation pattern (EP) based loudness model of Glasberg and Moore (2002). For each model, the system was configured to automatically mix the component tracks of a blues, classical, gospel, barbershop quartet and rock multitrack recording.

A MUSHRA style listening test was conducted in which listeners were asked, for each song, to rate the automatic mixes and those produced by four human mixing engineers in terms of achieving equal loudness across tracks, in addition to comparing overall mix quality. The results indicate that the model of Glasberg and Moore (2002) may not be applicable to complex musical sounds, as songs involving instruments with different spectral profiles were generally rated poorly compared to other mixes. In particular, this model generated perceptually marked deviations from the equal track-loudness target balance in the blues and rock songs. In comparison, the simpler energy-based measures performed fairly well relative to the mixes generated by the human engineers. Further tests are warranted to assess whether the K-filter should be revised, especially considering the performance of the automatic mixer when driven by an RMS-level detector based on a 3 s integration time. In addition, shorter temporal integration and an adaptive gating mechanism, such as those employed by the ITU-R BS.1770 (2015) algorithm, should also be compared with the single-band algorithms assessed here.

Some insight into the use of partial loudness models applied to multitrack audio was given, but the incorporation of inter-track masking calculations using current EP procedures is not expected to improve predictions of relative loudness; further experiments and modelling work is needed. Therefore, the added complexity of current partial loudness models is unlikely to enhance the subjective quality of the mixes they generate when incorporated into an automatic mixing device. In the next two chapters, the results of controlled loudness-matching experiments involving musical sounds are presented and subjective reference datasets are produced for the systematic evaluation of these loudness models, amongst others, in order to better quantify their applicability to musical sounds.

Chapter 6

Measuring the Relative Loudness of Musical Sounds

One of the main considerations in this thesis is the perception of musical sounds. In particular, the ability to predict the perceived loudness of complex sounds is of primary importance, especially in fields of audio engineering. In Chapter 2 the topic of loudness perception was introduced and a number of physical properties of sound were highlighted as influential factors on sensory magnitude. A perceptual phenomenon is typically measured using controlled psychophysical experiments involving synthetic stimuli. Early studies concentrated on developing a linear scale of loudness by measuring the perceived magnitude of pure tones as a function of intensity (Richardson and Ross 1930; Fletcher and Munson 1933; Stevens 1955, 1956b). Equal-loudness contours summarise the sensitivity of the ear as a function of frequency at different listening levels (ISO 226 2003). Additional influential variables include stimulus duration (Munson 1947; Buus et al. 1997; Epstein and Florentine 2006), spectral bandwidth (Fletcher and Munson 1933; Zwicker et al. 1957; Zwicker and Scharf 1965), component phase in multitone complexes (Gockel et al. 2002), binaural loudness (Zwicker and Zwicker 1991; Glasberg and Moore 2010) and partial masking (Scharf 1964; Gockel et al. 2003).

Spectral loudness summation (SLS) refers to the ability of the auditory system to aggregate loudness across different critical bands. For example, a broadband noise is perceived louder than a narrowband noise when both have the same overall intensity (Zwicker et al. 1957). The ear also integrates energy over time. It is well-known that the loudness of a tone burst decreases for durations less than 100 ms (Fastl and Zwicker 2007) and that the amount of temporal integration is level dependent (Buus et al. 1997). Some investigators have also studied the loudness of amplitude-modulated sounds (Zhang and Zeng 1997; Moore et al. 1999c), frequency-modulated tones (Zwicker 1974), and tone sequences with different spectro-temporal patterns (Zwicker 1969). Although these factors play an important role in the perception of everyday sounds, they are commonly studied independently using synthetic stimuli.

State-of-the-art psychophysical excitation pattern (EP) loudness models are grounded in auditory theory and have been developed to describe the following phenomena (Chalupper and Fastl 2002):

1. The effect of level and frequency of narrowband sounds.

2. The influence of spectral bandwidth.

3. The dependence of loudness on stimulus duration.

4. The effect of amplitude- and frequency-modulated tones.

Dynamic loudness models such as those presented in Chapter 3 are therefore evaluated against empirical data that reveal the above phenomena. For example, the ISO 226 (2003) equal-loudness contours are often used as ground-truth data for evaluating a model's ability to account for the sensitivity of the ear as a function of frequency and sound intensity (Glasberg and Moore 2006). Rennies et al. (2010) compared the dynamic models of Chalupper and Fastl (2002) and Glasberg and Moore (2002) by assessing their predictions of temporal effects reported in key psychophysical studies. However, in order for a model to be of practical and widespread interest, it remains to be seen how such models generalise beyond the scientific setting to sounds encountered in the natural world. In order to evaluate, adapt and propose computational models applicable to natural sounds, appropriate subjective datasets must be developed. Few perceptual studies have targeted the loudness of sounds encountered daily such as music and speech. For these types of sounds, it is likely that both physical and cognitive factors govern the final percept. It is, therefore, useful to conduct listening experiments to assess the generalisability of these algorithms.

Between 2003 and 2006, a number of researchers carried out loudness-matching experiments involving programme material typically used in radio and television broadcasts. This work formed the basis of the ITU-R BS.1770 (2015), in which an algorithm for estimating the perceived loudness of music and speech was presented. The recommendation is essentially based on the published work of Soulodre and associates (Soulodre et al. 2003; Soulodre and Norcross 2003; Soulodre 2004; Soulodre and Lavoi 2005). The experimental procedure used by Soulodre et al. (2003) required subjects to adjust the level of various monophonic broadcast sequences to a segment of a female speech recording. This setup was used at five different sites around the world, collectively involving 97 participants and 48 sound excerpts to formulate a subjective dataset. In a later paper, Soulodre (2004) evaluated ten loudness models against the empirical data and discovered that a simple equivalent continuous sound level ($\mathrm{L_{eq}}$) measure with a revised low-frequency B curve (RLB) frequency-weighting performed the highest across a number of metrics. The $\mathrm{L_{eq}}$ metric became the basis of ITU-R BS.1770 which has been subsequently used in many broadcast recommendations and specifications (EBU Tech 3341 2011; ATSC A/85 2013; EBU R 128 2014).

Skovenborg (Skovenborg and Nielsen 2004a; Skovenborg et al. 2004) showed how a balanced pair-matching experiment design is more robust compared to a fixed-reference design (as used by Soulodre et al. (2003)) in that no one segment is distinguished as the reference. This avoids the problem of reference selection which may introduce bias effects and poor between-subject agreement when a range of different sounds are involved. The goal of the listening test was to obtain a set of equal-loudness matches involving 145 segments of music and speech via the method of adjustment (MOA). Similar to Soulodre et al. (2003), however, the stimuli were extracted from audio sequences commonly used in broadcasting and were not instrument specific. The results were modelled using linear regression to test various hypotheses and estimate subjective bias terms as a result of the measurement procedure. In a second paper (Skovenborg and Nielsen 2004b), the regression coefficients corresponding to the estimated relative loudness levels were used to evaluate 12 different loudness models. Although the $\mathrm{L_{eq}}$ suggested by Soulodre et al. (2003) did outperform several loudness meters, including Zwicker's dynamic multiband model, two new algorithms developed by Skovenborg and Nielsen were most successful in predicting the experimental data. The first was a single-band model named LARM (see Chapter 3) and the second a multiband loudness model employing an octave-band filter bank. It is interesting to note that both studies (Skovenborg and Nielsen 2004b; Soulodre 2004) reported single-band models to be the most robust of loudness metrics.

Some efforts have been made to evaluate the performance of loudness models when applied to single instruments, as opposed to speech and mastered music. Pestana et al. (2013), for example, obtained loudness matches of recordings of single instruments taken from five multitrack projects, using either the lead vocal or bass drum as a fixed reference. The experimental data were used to assess the predictions of the ITU-R algorithm. The author proposed a number of alterations to the algorithm to improve the fit. This suggests that single-band loudness models may not be directly applicable to sounds with frequency-localised power spectra and distinct temporal profiles, rather than the broadcast material used to assess such models in previous work. Terrell (2013, Ch. 6) also conducted an experiment involving recordings of single instruments. In that study, subjects were required to estimate loudness ratios of paired musical sounds taken from four recordings: vocal, piano, double-bass and hand-drum. For example, subjects were asked 'how loud is the piano compared to the vocal?' All pairs were tested both in isolation, e.g. piano to vocal, and when the comparison sound was presented simultaneously with the remaining instruments, e.g. piano to vocal, with vocal presented in the mix. This measurement technique is referred to as magnitude estimation (Stevens 1975). Terrell showed that the loudness ratios were dependent on listening level, and found the results from the solo and simultaneous conditions to be similar, indicating minimal masking in the second experiment. The data were then modelled using Glasberg and Moore's (2002) dynamic loudness model, with an extension to explicitly weight the time-varying loudness function in accordance with signal dynamics. This adaptation, he named the dynamic sound-stream bias (DSSB), showed higher correlation to the experimental data compared to the global-loudness descriptors suggested by Glasberg and Moore (2002).

In this chapter, the design and results of different loudness-matching experiments involving musical sounds are presented. The primary goal was to establish a subjective reference dataset targeting single instruments commonly found in rock and blues genres. The resulting data is used in the next chapter for evaluating existing loudness models as well as establishing new loudness descriptors applicable to time-varying musical sounds.

6.1 Measuring loudness

Loudness is defined as the perceptual strength of a sound. Although it is easy for a listener to compare the loudness of two sounds, e.g. sound A is louder than sound B, it is not so clear as to how this comparison can be measured quantitatively—by what amount is sound A louder than sound B? Approaches to loudness measurement broadly fall into two categories: equal-loudness matching and scaling procedures. The choice of paradigm is of course dependent on the experimenter's hypothesis, but factors such as measurement bias, precision and efficiency are all important considerations. Marks and Florentine (2011) state that the selected approach should satisfy two basic requirements of internal consistency. The first is that if the measured loudness of sound A is greater than that of sound B, then sound A is louder than sound B, and sound B is softer than sound A. When the loudness of both sounds are measured as equal, the system must assign them the same loudness value. The second principle is that loudness equalities must be transitive. This means that if sound A is as loud as sound B, and sound B is as loud as sound C, then sound A must be as loud as sound C.

In the following sections, a number of key procedures employed in modern science for the measurement of loudness are detailed (for a historical overview of loudness measurement see Marks and Florentine (2011)).

6.1.1 Equal-loudness matching

Equal-loudness matching is used to measure the point of subjective equality (PSE) between the loudness of two or more stimuli. All other dimensions of sound such as pitch and timbre are to be ignored by the listener. Typically, the subject is presented with a fixed sound (reference stimulus) and a variable sound (comparison or test stimulus), that alternate in time, and the intensity level of the latter is varied, either directly or indirectly, to obtain the point of equal loudness. Measurements are often represented as differences in root mean square (RMS) level between the two sounds when both are judged to be equally loud. Three loudness-matching procedures are discussed here: the MOA, the method of constant stimuli, and the adaptive procedure. As noted by Marks and Florentine (2011), the method of constant stimuli is no longer used as it is highly inefficient and has been replaced by modern adaptive procedures.

The method of adjustment

The MOA permits the listener to freely adjust the level of the variable sound until its loudness matches that of the fixed sound. Once the subject is satisfied with their adjustment, the final level of the variable sound is recorded. This procedure is simple, intuitive and efficient, but also error-prone. When the sounds are presented in succession, the listener tends to judge the second sound as louder or softer than the first. This order effect can be reduced by repeating the match with the stimuli presented in reverse order. Additional bias effects can be caused by the initial level of the variable stimulus, the position of the rotary dial (or slider) (Stevens 1955) and by stimulus persistence (Gelfand 2010) which occurs when the listener continues to adjust the level beyond the 'true' matching level. To minimise such phenomena, the starting level of the variable stimulus can be randomised or balanced across subjects. For example, the level of the variable stimulus could be set 10 dB higher than the level of the fixed stimulus in one trial, and 10 dB lower in another. Furthermore, the listener may be instructed to adjust the variable stimulus alternately louder and softer than the fixed stimulus to 'bracket' or 'home-in' on the PSE. A final and important source of bias is the overestimation of the loudness of the reference stimulus (Marks and Florentine 2011). As such, the listener tends to set the level of the variable sound too high, though this effect is most significant at low-medium listening levels. In most cases, the best strategy is to have the listener adjust both sounds and average the results.

The method of constant stimuli

The method of constant stimuli is a classic psychophysical technique used to measure perceptual thresholds. For equal-loudness measurement the experimenter is interested in measuring the threshold corresponding to the PSE. The two sounds to be matched in loudness are presented in alternation and the subject is asked which of the two sounds is louder. This is repeated many times with the test sound set to various predefined intensities, with the presentation order randomised. The proportion of 'louder' responses at each intensity is recorded and a curve is fit to the data to form a psychometric function. With sufficient data, numerous thresholds can be measured. The PSE, for example, is taken as the intensity where the variable sound was perceived louder than the fixed sound on 50% of the trials. This approach has the advantage of minimising bias due to presentation order, the direction of adjustment, and ensures that the variable sound is presented at levels both above and below that of the fixed sound. The experimenter should take care to select a sufficient range of levels which includes the expected PSE. Thus, pilot experiments are necessary. The pitfall is the large number of trials required to make a single measurement. Example data from a simulated experiment is given in Table 6.1. It can be seen that the subject was sure that

SPL (dB)	% Louder
70	100
68	95
66	85
64	70
62	55
60	35
58	20
56	10
54	8
52	5
50	0

Table 6.1: Percentage of time the variable sound was perceived louder than the fixed sound at different intensity levels.

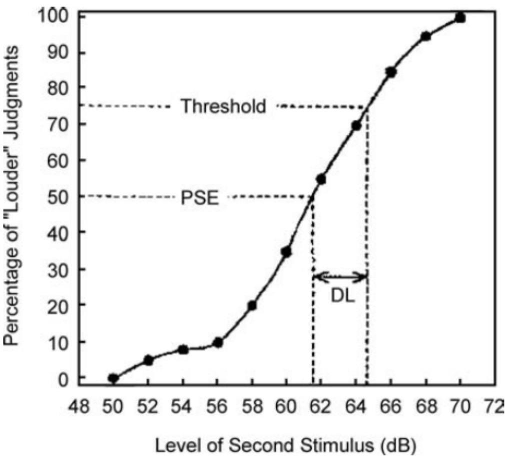


Figure 6.1: Psychometric function corresponding to the data in Table 6.1.
Image from Gelfand (2010)

the variable sound was louder than the fixed sound at a sound pressure level (SPL) of 70 dB. At 50 dB SPL the variable sound was softer. The psychometric function for these data is plotted Figure 6.1. The 50% point corresponds to the level at which the variable sound was perceived both louder and softer than the fixed sound an equal number of times. This level (here 61.5 dB) is taken as the PSE. Also shown in the figure is the 75% point which is generally accepted as the threshold for 'louder than'. The difference between these two levels (from equally loud to variable is louder) is called the difference limen (DL) for intensity (see Section 2.3.3).

Adaptive procedures

Adaptive procedures are by far the most complex yet commonly used loudness-matching paradigm because they are designed to minimise bias effects, maximise precision, and are more efficient than the method of constant stimuli. The listener is presented with two stimuli in succession, separated by a short inter-stimulus gap. As with the method of constant stimuli, the listener is then asked which of the two sounds is louder. Unlike the method of constant stimuli, the level of the variable sound in the next trial is based on the subject's previous responses—the level adjustments are adaptive. A set of rules allocated by the experimenter are used in the level adjustment calculation to target a specific threshold measurement. Thus, no time is 'wasted' sampling the psychometric function at points other than the one of interest. The order of presentation should be randomised and the duration of the pause separating the two sounds should also be considered.

Most of the constraints built into the adaptive system are based on so called 'reversal points'. A reversal occurs when the subject's response changes from 'louder' to 'softer' or vice versa. The most frequently used adaptive procedure to determine the PSE is the simple up-down or staircase method. This involves increasing the level of the variable sound until the subject judges it to be louder than the fixed sound, at which point a reversal occurs. The level is then decreased on the subsequent trial and continues to decrease with every 'louder' response. When a 'softer' response occurs (or no response) the direction of adjustment is reversed. This series of adjustments creates an 'adaptive track' that takes the form of a stair-case when plotted over trial number. The step size is determined by the experimenter and often decreases as the track converges to the PSE (Levitt 1971). For example, Buus et al. (1999) employed a 5 dB step size until the fourth reversal, after which it was reduced to 2 dB. A track typically ends after a predetermined number of reversal points has occurred. The precision of the measurement increases with the number of trials but consequently requires subjects to make many judgements which can lead to fatigue and thus increase measurement variability (Marks and Florentine 2011). The level at threshold is taken as the average of the levels at the reversal points, typically excluding those recorded using the largest step-size. An example of an adaptive track according to the simple up-down rule using two step sizes is given in Figure 6.2.

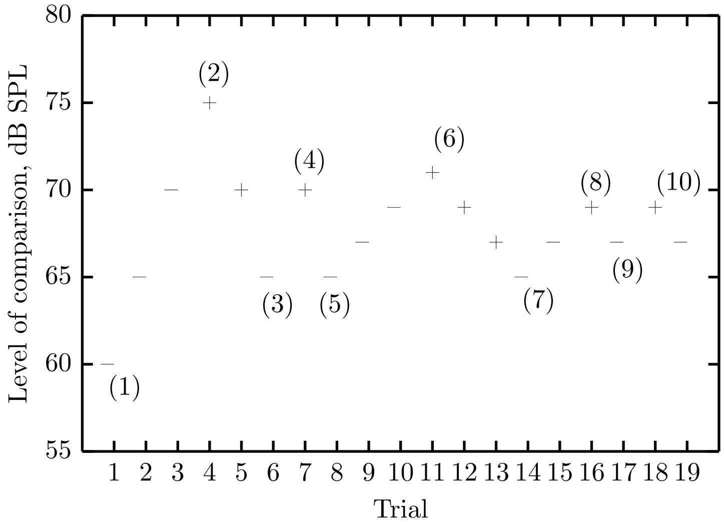


Figure 6.2: Example of an adaptive track obtained from a loudness-matching experiment. The step size was 5 dB until four reversals and 2 dB thereafter. The track ended after a total of 10 reversals. A ‘louder’ response is marked as +, and a ‘softer’ response marked as -. Numbered labels refer to a set of ascending or descending trials known as a run.

One major disadvantage of the staircase procedure is that subjects are able to identify the adaptive mechanism and therefore control the level of the variable sound indirectly. For example, a ‘louder’ response leads to an attenuation of the variable sound, whereas a ‘softer’ response leads to a boost. Memory effects can also occur in which the subject compares the loudness of the variable sound across successive trials and thus may begin to ignore the loudness of the fixed stimulus. Shuffling presentation order can help to minimise this effect, as can interleaving several adaptive tracks (Marks and Florentine 2011). The level of the fixed stimulus may be varied across concurrent tracks such that variations in overall loudness take place. Mixing tracks in which both sounds are varied in level can also be used to reduced bias (Buus et al. 1997). This is likely to force the subject to base their judgement on the perception within a trial, but comes at the cost of extending exposure time making loudness judgements tedious.

6.1.2 Scaling procedures

In contrast to methods based on loudness matching, where measures of equivalence and rank order are possible, scaling procedures target quantitative scales of loudness (Marks and Florentine 2011). These scales provide additional information such as how *much* louder one sound is relative to another. Here, two scaling methods are introduced: magnitude estimation and magnitude production.

Magnitude estimation

In magnitude estimation the subject maps the perceived loudness of a sound to a set of numbers of their choosing. In the simplest case, the subject is presented with a stimulus and asked to assign a numerical value denoting its loudness. Although the method is simple and efficient, it is susceptible to sequential effects where previous rankings influence future rankings (Stevens 1955). Another variant of this procedure is ratio magnitude estimation in which the participant is asked to judge the ratio relations between stimuli. For example, if one sound is assigned the number 10, and a second sound is reported to be half as loud as the first, a response of five would be given. Another source of bias is the over- and under-estimation of loudness when the variable sound is presented at high and low intensities, respectively. In addition, because the method is unbounded, the set of numbers used by a single subject may be in high disagreement with those of the group, thus skewing the distribution of the collective ratios (Stevens 1955). Rather than using the arithmetic mean to describe the central tendency, the geometric mean is used as this metric is less affected by skewed distributions and thus preserves ratio relations better than the arithmetic mean. An interesting aspect of magnitude estimation is that responses are a product of two internal processes (Marks and Florentine 2011). The first is the transformation from acoustic signal to global-loudness percept, the second involves a mapping from sensory magnitude to numerical representation. Only if the latter is linear can valid measurements of loudness be made.

Magnitude production

In this procedure the subject hears a stimulus and is asked to adjust its level such that the loudness matches a target number assigned by the experimenter. It can be seen as the reverse task of magnitude estimation; rather than matching numbers to stimuli, the subject matches stimuli to numbers. When measuring relative loudness, the subject is presented the two sounds in alternation and instructed to adjust the level of the variable sound such that the loudness associated with it is a fraction of the loudness of the fixed sound. As with the MOA, biasses due to the type of potentiometer, direction of adjustment, fixed stimulus and presentation order exist.

Stevens (1975) points out that in all matching experiments there is a tendency for subjects to shorten the range over whichever variable they are allowed to adjust. In this respect, the results from magnitude estimation and magnitude production can be averaged to compensate for the biases they introduce. It is therefore assumed that the biases introduced by the two methods are opposite but equal in magnitude.

6.2 Experiment 1: preliminary loudness assessment

The purpose of this experiment was two fold:

1. To assess whether the transitivity of loudness holds for short excerpts of musical sounds.
2. To assess the practicality of employing an adaptive loudness-matching procedure for the purpose of developing a reference subjective dataset.

The decision to use equal-loudness matching was based on the primary goal of obtaining subjective data for assessing the quality of predictive models originating from both psychoacoustic and engineering fields. For example, single-band algorithms such as the ITU-R BS.1770 (2015) do not measure loudness in terms of the true definition being perceptual intensity, but rather integrate signal power in a way that provides a level reading that is useful for normalising the loudness of programme material. Thus, loudness-matching data are needed to evaluate these algorithms. Although level differences can be obtained from scaling data, they do not specifically target accurate measurement of the PSE, and are therefore generally less reliable (Rennies et al. 2013). An adaptive equal-loudness matching procedure was chosen for this experiment because they are now established as the 'gold standard' for measuring specific points on the psychometric function, owing to their ability to reduce both bias effects and within-subject variability using a relatively efficient indirect perceptual measurement.

6.2.1 Method

Stimuli

Because this experiment was primarily an exploratory investigation, only four sounds were used. The stimuli were extracted from a multitrack recording of 'Good Time' by the funk-rock band Louis Cressy Band, downloaded from the Free Multitrack Download Library (Cambridge Music Technology 2013). Four 3.6 s segments were taken from the multitrack session. Each segment represented one of the four musical instruments played by the band: bass, drum kit, electric guitar and voice. The stimuli were calibrated such that the RMS levels were approximately 80 dB SPL at the eardrum after accounting for the average magnitude response and sensitivity of a pair of BeyerDynamic DT990 (600 Ohms) headphones as measured on a GRAS 43AG ear and cheek simulator which comprises an IEC 60711 compliant ear simulator.

Participants

Eight students from the Sound Engineering and Production degree course at Birmingham City University participated in the study. All were aged between 20 and 25 with no known hearing impairment. All subjects participated in an initial practice session to familiarise themselves with the software and task. The main experiment lasted approximately 1.5 hours.

Apparatus

The experiment was performed on a Macintosh computer connected to a Metric Halo ULN-2 Firewire audio interface operating at a sampling frequency of 44100 Hz. The outputs of the interface were connected to a TDT HB7 headphone buffer driving both capsules of a BeyerDynamic DT990 (600 ohms) headset. The HB7 buffer was configured in AC coupling mode to remove DC bias. For routine calibration, the output of the headphone buffer was fed to multi-meter with the headphones connected in parallel. Listeners were seated in an isolated vocal recording booth inside the Toft recording studio at Birmingham City University.

Procedure

The goal of the loudness assessment was to measure the point of equal loudness between every pairwise combination of the four sounds, using a two-interval forced-choice paradigm. The listener heard two segments presented in a random order and separated by 600 ms of silence. For a given trial the task was to indicate which of the two segments was the loudest by pressing one of two keypad buttons representing each sound (first or second). A button press was only accepted during an 800 ms silent interval which occurred after playback of the second excerpt had ended. For each stimulus pair, one of the sounds was variable in level and the other was fixed in level. The variable sound was set to ± 10 dB about the initial calibration level and adjusted using an up-down procedure. If the button corresponding to the variable sound was pressed (denoting the variable sound is louder), its level was reduced and increased otherwise. An initial step size of 5 dB was used which reduced to 2 dB after two reversals. After a total of 11 reversals, the adaptive track ended and the level of the variable stimulus at the point of equal loudness was obtained by averaging the levels of the final four reversals. Loudness matches were obtained for all pairwise combinations of the four sources, giving 10 pairs, four of which involved same-segment comparisons, e.g. bass and bass. For the six pairs comprising different segments, matches were obtained with both segments varied in level by interleaving the two adaptive tracks. That is, for a given set of concurrent tracks, subjects would perform matches with both segments serving as the variable stimulus. This was done to force the listeners to focus on the loudness of the segments for the current trial and reduce bias effects caused by the constant presentation level of the fixed stimulus. This gave six interleaved tracks for the different-segment pairs, which were repeated twice by each subject. For each of the four same-segment pairs, two identical tracks were interleaved. This gave a total of 16 interleaved tracks formulating 32 matches per participant. The order of these 16 interleaved tracks was randomised on an individual basis, and each adaptive track was selected at random on a given run.

6.2.2 Results

The loudness-matching data are expressed as a set of level differences between the gains of the variable and fixed stimulus, which themselves are relative to the calibration levels. Let $\Delta L_{i,j}$ denote this level difference, where i is the index of the fixed stimulus, and j is the index of the variable stimulus. For example, $\Delta L_{i,j} = 5$ means that that stimulus j was set 5 dB higher than stimulus i at the point of equal loudness. For simplicity, these level differences are referred to as matching gains.[1]

Subject consistency and between-subject agreement

The subjects matched the relative loudness of four segments by listening to all pairwise combinations with every sound serving as the fixed stimulus. Because each stimulus pair was matched with both stimuli varied in level, i.e. as (i, j) and (j, i), it was possible to measure the consistency of the subjects. This was done by taking Ginis mean difference (GMD) between the four matches associated with each stimulus pair (i, j), with the (j, i) level differences negated for (assumed) symmetry. Note that this measure of consistency include any potential bias effects. The GMD statistic is defined as:

$$\mathrm{GMD} = \frac{1}{N(N-1)} \sum_{n=1}^{N} \sum_{m=1}^{N} |y_n - y_m|, \tag{6.1}$$

where N is the number of observations (y) in a given condition. In this case, N is the number of matches (four), and y is the set of matching gains associated with a given pair and subject. The mean GMD across subjects was 2.1 dB (standard deviation (SD) = 0.6 dB). For pairs involving the same stimulus, e.g. bass and bass, the average GMD was 1.2 dB (SD = 0.5) and, for pairs involving different segments, the mean GMD was 2.8 dB (SD = 0.9). The former gives an indication of the intensity DL for the four segments. The latter indicates greater variability in the reproducibility of the subjects when comparing different sounds.

After averaging replicate pair matches for each subject, between-subject agreement was measured using the SD of the matches associated with each unique pair (ten in total). The average SD across the ten pairs was 1.8 dB (SD = 1.2). The average SD for the four same-segment pairs was

[1] The justification for this representation will be apparent in Section 6.2.2. Essentially, one can think of $\Delta L_{i,j}$ as the difference in loudness levels of segments i and j.

0.8 dB (SD = 0.3), and 2.5 dB (SD = 0.9) for the remaining six pairs involving different segments. The mean GMD of subjective gains was also computed for each pair. For the same-segment pairs, the average GMD was 0.9 dB (SD = 0.3) and, for different-segment pairs, the average GMD was 3.0 dB (SD = 1.1). The within- and between-subject variability for the ten pairs, as measured using the GMD statistic, are summarised in Table 6.2. It is interesting to note that matches involving the bass segment are associated with the highest variance, which exceeds 3 dB for both sources of variation.

Pair	Within	Between
G/G	0.8	1.0
B/B	1.2	0.5
V/V	1.3	1.3
D/D	1.4	0.8
D/V	1.9	2.3
D/G	2.0	2.3
G/V	2.1	1.5
B/V	3.3	3.6
B/D	3.5	4.3
B/G	3.7	3.8
Mean (SD)	2.1 (1.0)	2.1 (1.3)

Table 6.2: Average within- and between-subject GMD for the ten segment pairs matched by eight listeners. Rows are organised in ascending order according to the within-subject GMD.

Bias effects

When a listener matches the loudness of two identical sounds, one expects an average gain of zero. Any value that is systematically different suggests a bias effect, potentially caused by an under- or over-estimation of the loudness of the fixed-stimulus. Bias effects associated with the fixed stimulus can also be investigated by averaging the levels obtained with the two segments varied. For example, if $\Delta L_{i,j} = 6$ and $\Delta L_{j,i} = -4$, the average is 1 dB, which implies that the variable stimulus was set 1 dB above the 'true' matching level. To explore this, the gains were averaged for each of the ten pairs and input to a one-way repeated-measures analysis of variance (ANOVA).[2] The ANOVA showed that the bias terms were reliably different across the ten pairs ($\chi^2(9) = 20.9$, $p = 0.01$) but the distribution of the fixed-effect residuals were non-normal. Therefore, a one-way Friedman rank sum test was performed which showed no statistically significant effect ($\chi^2(9) = 14.5$, $p = 0.1$).

Figure 6.3 shows the distribution of these terms for each pair. Although the results were non-conclusive, it can be seen that estimated bias for pairs bass/bass (B/B) and drums/voice (D/V) are positive for all but one subject, and that the two distributions have low variance compared to the other pairs. This suggests the presence of a small bias effect which may be dependent on the pair of sounds. The overall bias across subjects was 0.50 dB (95% CI = [0.01, 0.98]), which was found to be statistically significantly different from zero ($t(7) = 2.42$, $p = 0.046$), although further study with a larger sample size is recommended to gauge whether this effect is stimulus dependent.

A linear model of the matching data

Skovenborg et al. (2004) suggested a set of relative loudness levels (in dB) underlying the subjective matching gains. For example, if $\Delta L_{i,j} = 5$, segment j was set 5 dB higher than segment i, and

[2] The signs of the matching gains were preserved.

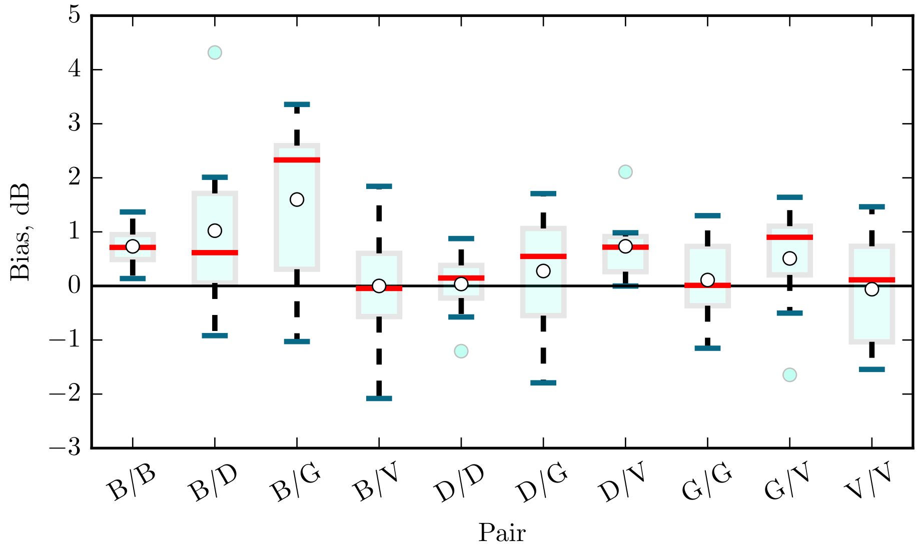


Figure 6.3: Boxplots of the subject gains per stimulus pair, using the raw matching data. Each boxplot comprises eight values (one gain per subject). The red horizontal line is the median and the white circle is the mean.

therefore the matching gains represent a difference in loudness level between the two sounds:

$$\Delta L_{i,j} = l_i - l_j, \tag{6.2}$$

with l_i and l_j the relative loudness level of segment i and j, respectively. Skovenborg and Nielsen (2004a) showed that, under the assumption of local linearity and loudness transitivity, the vector of relative loudness levels can be estimated using a linear regression model of the form:[3]

$$\begin{bmatrix} \Delta L_{0,1} \\ \Delta L_{1,0} \\ \Delta L_{0,3} \\ \Delta L_{3,0} \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 1 & 0 & 0 & -1 \\ -1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \times \begin{bmatrix} l_0 \\ l_1 \\ l_2 \\ l_3 \end{bmatrix} + \begin{bmatrix} e_0 \\ e_1 \\ e_2 \\ e_3 \\ \vdots \end{bmatrix}, \tag{6.3}$$

or using matrix notation:

$$\Delta L = Xl + e. \tag{6.4}$$

In this experiment, the per subject design matrix X is a 24 (matches) × 4 (segments) matrix which includes all different-segment pair matches. The system is overdetermined because there are more unique loudness matches than there are parameters (six pairs and four segments). The columns are coded using categorical variables, where 1 denotes the fixed stimulus and -1 denotes the variable stimulus. A zero value indicates that the segment was not present in the pair. Note that in practice, the column representing the first segment is replaced with a vector of ones, representing the model intercept, in order to establish linear independence amongst the columns, and thus obtain a unique least squares solution for the vector l. Stated another way, when no segment is selected to represent the intercept, the matrix is rank deficient (in this case $rank(X) = 3$). The final column e is a vector of error terms assumed to be normally distributed with zero mean and constant variance.

[3] The example design matrix is given for pairs bass/drums and bass/voice with each sound serving as the fixed and variable stimulus.

In a balanced experiment where all sounds serve as both the fixed and variable stimulus with the same absolute frequency and no bias effects are present, the expected value of the raw matching data is zero. A systematic bias b pertaining to the loudness of the fixed stimulus that is consistent across all pairs would yield:

$$\Delta L_{i,j} = (l_i + b) - l_j. \tag{6.5}$$

If b is positive, e.g. 0.5 dB as found in the previous section, a systematic error is present which could be attributable to an overestimation in the loudness of the fixed stimulus. With the formulation above, this bias shows up in the intercept, i.e. $l_0 \neq 0$, and indicates that the grand mean is statistically different from zero.[4] The intercept term can therefore be used to test for an overall bias effect. The remaining parameters l_1, l_2, and l_3 denote the loudness levels of segments 1, 2 and 3 (here: drums, guitar and voice), respectively, relative to the loudness of segment 0 (bass). For example, if $l_1 = 4$, then segment 1 is 4 dB louder than segment 0. Note that the choice of reference does not affect model efficacy and can be thought of as the baseline category.

To check if the model fits the observed data well, the regression model was first fit to each subject individually. A common goodness-of-fit measure is the coefficient of determination R^2 which measures the proportion of variation in the dependent variable (the matching gains) accounted for by the variation in the predictors (the loudness levels) (Chatterjee and Hadi 2006). The coefficient of determination is defined as

$$R^2 = 1 - \frac{\sum(y_n - \hat{y}_n)^2}{\sum(y_n - \bar{y})^2}, \tag{6.6}$$

where the numerator in the second term is the sum of squared residuals and the denominator is the total sum of squares in the dependent variable. In addition to R^2, the root mean square error (RMSE) between the predicted and fitted values were measured for each subject.

Subject	SE	R^2	RMSE	F-value
1	0.79	0.85	2.05	37.54
2	0.82	0.74	2.11	18.45
3	0.53	0.70	1.36	15.31
4	1.15	0.89	2.98	52.98
5	1.31	0.80	3.38	27.30
6	0.65	0.92	1.69	79.53
7	0.54	0.92	1.39	77.38
8	0.70	0.91	1.80	66.20
Mean (SD)	0.73 (0.17)	0.84 (0.09)	2.09 (0.73)	

Table 6.3: Summary of the individual regressions performed for each subject. All F-values were significant at level $\alpha = 0.0001$. SE is the standard error of the regression coefficients (relative loudness levels).

Table 6.3 summarises the individual fits by showing the SEs of the three regression coefficients, the R^2, the RMSE, and the F-value used for the F-test. All regressions were statistically significant ($p < 0.0001$), which suggests a reliable association between the predictors and the matching data, at the subject level. Under the assumption that the most consistent subject would fit the model exactly, the variation in the two goodness-of-fit metrics suggest that some subjects are more consistent than others. When measured in terms of RMSE, the more inconsistent the subject, i.e. those with larger RMSE, the greater the level of uncertainty in the estimates of relative loudness.

Figure 6.4 shows the parameter estimates from the individual fits, where the drum, guitar and vocal coefficients denote their loudness relative to the bass. The intercepts are positive (0–2 dB) for

[4] Additional, stimulus-specific bias terms can also be estimated by appended columns to the design matrix, but this was not done here.

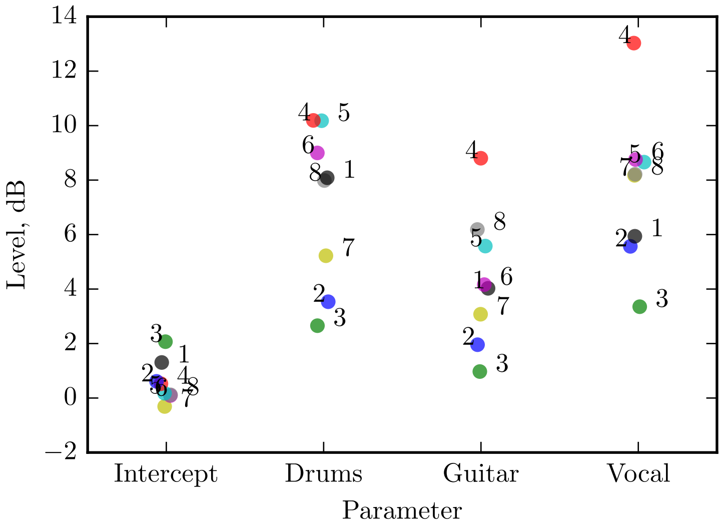


Figure 6.4: Parameter estimates derived from the per subject fits (labelled 1–8).

all but one subject, suggesting a systematic bias effect, in agreement with the previous analysis. The loudness levels are also positive, which in conjunction with the SEs reported in Table 6.3, indicates that the drums, guitar and vocal were all louder than the bass sample. There is also large variation in the relative levels, e.g. subject three perceives an average relative loudness of about 2 dB, compared to a difference of 11 dB for subject four. A single linear regression model does not accommodate such sources of variance, which would otherwise would violate the assumption of independence. Such concerns are better dealt with by mixed-effects models as they allow the model parameters to vary around the estimated population values, by explicitly incorporating random effects (see Pinheiro and Bates (2000), Winter (2011), and Field et al. (2012)).

To demonstrate this, a linear model was fit to the data with and without random effects.[5] Fitting a single fixed-effects only model, i.e. a standard regression model, to the entire dataset produces the results shown in Table 6.4. The estimates correspond to the average of those displayed in Figure 6.4, all of which are statistically significant, especially so for the coefficients (relative levels).[6]

	Estimate	SE	t-value	p-value
Intercept	0.7	0.2	2.9	0.004
Drums	7.2	0.4	17.5	0
Guitar	4.7	0.4	11.4	0
Vocal	7.9	0.4	19.2	0

Table 6.4: Summary of the single regression analysis applied to the combined matching data from all subjects.

Table 6.5 shows the results when random intercepts and slopes are included in the model; that is, when the intercept and slopes are allowed to vary for each of the eight subjects.[7] This time there is an SD associated with each estimate, because the model has now estimated the between-subject variability for each term. Furthermore, the SEs are roughly twice as large as those reported by the one-size fits all model, and are closer to those obtained by calculating the uncertainty of the

[5] The function `lme4` from the R package `lme4` (Bates et al. 2015) was used to fit the mixed-effects model.
[6] The p-values of zero reflect the limit of floating-point precision.
[7] P-values are not computed by the function `lmer`.

	Estimate (SD)	SE	t-value
Intercept	0.7 (0.4)	0.2	3.0
Drums	7.2 (2.8)	1.0	6.9
Guitar	4.7 (2.5)	0.9	5.1
Vocal	7.9 (2.7)	1.0	7.8

Table 6.5: Summary of the multi-level regression analysis applied to the combined matching data from all subjects.

parameters derived from the individual fits (SEs of 0.3, 1.1, 1.0 and 1.1 dB). This is because the fixed-effects only model sees every observation as being independent, which is not the case. Using a likelihood ratio test, the inclusion of random slopes was found to improve the fit over the intercept-only model ($\chi^2(3) = 49.9$, $p = 0$), which was then further improved by allowing the slopes to be correlated ($\chi^2(3) = 22.6$, $p = 0.0005$). To compute the significance of the fixed-effects, a likelihood ratio test can be performed using only the intercept (grand mean) to predict the data, i.e. denote the null model. Indeed, the inclusion of the fixed-effects improved the fit over the intercept-only model ($\chi^2(3) = 22.4$, $p = 0.00005$), which indicates that the stimuli were not equally loud when presented at an equal RMS level.

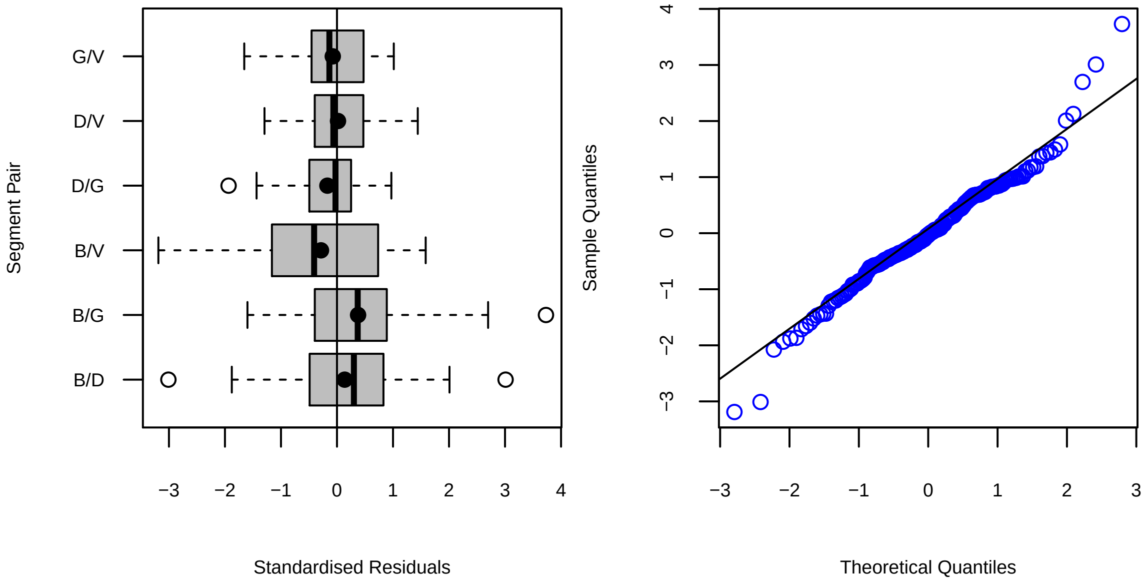


Figure 6.5: Standardised residuals by segment pair (left) and Q-Q plot of standardised residuals (right) for the mixed-effects model applied to the loudness-matching data from experiment 1. The solid circle and vertical line shown on each boxplot denotes the mean and median of the conditional residuals, respectively.

Figure 6.5 shows a plot of the standardised residuals for each segment pair and a Q-Q plot of the residuals from this model. The plot on the left shows that the means and medians of the residuals are centred about zero, although there is some notable deviation and increased variance for pairs involving the bass segment. This is to be expected considering that the subjects were less consistent in these conditions (see Table 6.2). There are also a few notable cases with large residuals, leading to a deviation from the straight line (and therefore normality) in the Q-Q plot. The magnitude of eight of the 192 (4.2%) standardised residuals exceeded 2.0, and five (2.6%) were above 2.5. For an inaccurate model, one would expect about 5% of these values to exceed 2.0, and 1 % to exceed 2.5 (Field et al. 2012), so there are no major concerns in this respect. Four data points from two subjects (S4, S5) were, however, associated with large standardised residuals: S4

$= \{3.0, 3.7\}$ and S5 $= \{3.2, 3.0\}$. S5 generally had a notably larger residual variance compared to the remaining seven participants. These four cases were associated with matches in which the bass samples served as the variable stimulus. Two of these residuals corresponded to gains as large as 18.8 dB applied by two separate subjects when adjusting the bass guitar to match the loudness of the electric guitar (S4) and when matched to drums (S5). Removal of the four residuals led to a change of 0.1 dB for one of the coefficients. Furthermore, the residuals from the original fit did show some deviation from normality (see Q-Q plot in Figure 6.5), which was confirmed by a Shapiro-Wilk test ($W = 0.98$, $p = 0.002$). However, Q-Q plots of the residuals from the second model did show less severe signs of non-normality which was further supported by the Shapiro-Wilk test ($W = 0.99$, $p = 0.39$). The fixed-effects remained statistically significant ($\chi^2(3) = 21.5$, $p = 0.00008$).

Overall, one can conclude that the linear model does explain variation in the matching data and can be used to estimate the relative loudness of the sounds involved. The statistical significance of the fixed-effects indicates that the stimuli were not equally loud when RMS normalised—the relative loudness of at least one segment was reliably different from zero. This is confirmed by the estimates shown in Table 6.5 which shows that the drums, guitar and voice samples were perceived by the listeners to be louder than the bass. The intercept term representing the average gain was reliably different from zero (mean = 0.69, 95% CI = [0.2, 1.1] dB) indicating a systematic effect in the level-matching procedure.

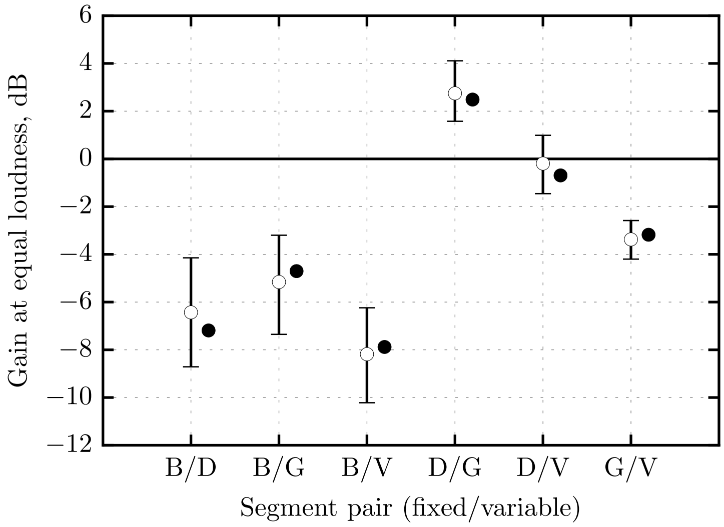


Figure 6.6: Mean gain of the variable stimulus required to match the loudness of the fixed stimulus for different segment pairs as measured experimentally (white circles) and derived from the linear model (closed circles). Error bars show the 95% bootstrapped between-subject CI of each mean. Gains of 0 dB correspond to equal RMS levels, as highlighted by the solid horizontal line.

Figure 6.6 shows the mean gain at the point of equal loudness for all pairs involving different sounds (white circles). The horizontal line at 0 dB represents equal RMS level = equal loudness. Note that drums and voice were the only two segments that were equally loud when presented at an RMS level of 80 dB SPL. The black circles show the fitted means derived from the regression model after setting the intercept to zero, i.e. discounting any bias. The predicted means are consistent with the matching data: the RMSE and maximum deviation between the six average gains and those derived from the model was 0.45 dB and 0.76 dB, respectively. Discrepancies are largest for pairs involving the bass segment, although these pairs are associated with greater uncertainty as shown by the CIs. Overall, the model fit indicates that subjects were internally consistent when matching different pairs made up from the four sounds, i.e. the assumption of loudness transitivity appears to be valid.

6.2.3 Summary

When matching the loudness of short musical sounds, listeners make a conscious effort to judge differences in loudness as accurately as possible. This task is made easier when both segments are identical, because the task is not confounded by other aspects of perception such as pitch and timbre; the listener can focus solely on detecting differences in signal intensity. As with subject consistency, between-subject agreement was better for pairs involving the same sounds, i.e. in a discrimination task. For pairs made up from different segments, listeners were less consistent for those pairs involving the bass guitar. Additional tests should be conducted to identify potential causes for this, such as the method of reproduction, e.g. headphones. The results from all matches suggested a positive bias in the equal-loudness matching procedure. When averaged over all matches, the CI indicated that this bias effect could be up to 1 dB. Further precautions, such as interleaving loudness matches from different segment pairs within a run, might aid to reduce this effect. These findings also highlight the importance of having both sounds in a pair serving as the reference segment. Further research is be required to identify if bias effects pertain to specific combinations of sounds.

A linear model of the subjective adjustments was used to inspect for transitive relations between the loudnesses of the stimuli. That is, if sound A is 5 dB louder than sound B, and sound B is 3 dB louder than sound C, then sound A must be 8 dB louder than sound C. The model was able to reliably explain systematic variation in the matching data and therefore predict the means of the individual matches. Again, this model revealed a systematic effect whereby listeners placed the variable sound slightly above the point of equal loudness. Importantly, the average estimates produced by the model revealed that the stimuli were not equally loud when RMS normalised. This shows, at least for the sounds used in this study, that equal RMS does not lead to equal loudness. Gain corrections of up to 8 dB were required to loudness-align the stimuli, which is considered large enough to justify the move away from RMS based loudness estimators for applications involving musical instruments.

In this experiment, loudness matches were obtained using an adaptive procedure, which are today seen as the gold standard for measuring the PSE. It is important to note that the duration of this experiment was roughly 1.5 hours per subject for 32 loudness matches involving segments of 3.6 s in length. Roughly half of these were required to achieve symmetry in the matching data, i.e. matches with both segments varied in level. On average, a single match lasted around 2.5 minutes, which, when interleaving tracks, means that the listener could be listening to sequences of audio for up to 5 minutes. Some participants reported the task to be laborious, finding it difficult to concentrate after repeatedly matching the same pair of segments for long periods of time. This may explain the few outlying matching gains (as large as 18 dB) in the dataset. Such experimental procedures may therefore not be appropriate when measuring the global loudness of musical sounds, especially in large scale experiments where the goal is to obtain a reference dataset for a wide variety of sounds. Alternative procedures suited to such requirements are investigated in the next section.

6.3 Experiment 2: main loudness assessment

In this section the stimuli, participants and experimental procedure used to establish a set of loudness matches involving sixteen musical sounds is described. The aim of this experiment was to formulate a subjective dataset suitable for evaluating loudness models. It is similar in form to the previous experiment, except that a larger sample of sounds and subjects was used.

6.3.1 Method

Stimuli

16 sounds belonging to one of four categories of musical instruments were used for the experiment. The four groups were: bass, guitar, percussion and singing voice. These categories were selected based on common instruments played by musicians of rock and blues music, and provide a heterogeneous collection of sounds characteristic of those experienced by music producers when mixing multitrack recordings. The sixteen sounds were extracted from the Apple Loops library, included with the digital audio workstation (DAW) Logic Pro 8, a diverse catalogue of high-quality recordings of musical instruments, soundscapes and sound effects. The average stimulus duration of the selected audio excerpts was 4.1 seconds (SD = 0.4), sufficient to allow for a good impression of global loudness. Moore et al. (2003) used 2.1 s segments of speech for their loudness-matching experiment which is shorter than the duration of sounds used here, but the spectro-temporal behaviour of musical sounds is generally more varied compared to speech, and therefore greater exposure time is needed to gauge representative attributes of such sounds. Terrell (2013, Chapter.7) also conducted a loudness assessment involving musical stimuli also lasting about 4 seconds. Furthermore, the findings from experiment 1 showed that listeners were able to reliably estimate the loudness of sound segments similar to those used in this study.

ID	Contents	Duration (s)	RMS (dB SPL)	Peak (dB SPL)	Centroid (Hz)	Spread (Hz)	Q
B_0	Electric bass guitar	4.8	82.2	91.8	97	49	2.0
B_1	Electric bass guitar	4.3	80.2	93.2	117	47	2.5
B_2	Electric bass guitar	4.0	75.5	91.4	309	705	0.4
B_3	Electric bass guitar	3.3	78.8	90.7	133	67	2.0
G_0	Electric guitar	4.4	70.7	95.3	1732	1888	0.9
G_1	Electric guitar	4.8	77.0	89.5	2211	1425	1.6
G_2	Electric guitar	4.0	72.6	94.9	904	1176	0.8
G_3	Acoustic guitar	4.0	72.1	88.9	525	1142	0.5
P_0	Bass drum	4.3	75.6	97.4	107	189	0.6
P_1	Full kit	3.8	74.3	95.9	1903	3323	0.6
P_2	Shaker	4.3	66.0	90.8	11512	3675	3.1
P_3	Snare drum	4.1	69.7	93.1	1268	2098	0.6
V_0	Group vocal	3.1	74.9	88.8	2063	2118	1.0
V_1	Male vocal with reverb	4.2	75.8	91.7	2110	2057	1.0
V_2	Female vocal	4.5	76.3	93.4	2411	3295	0.7
V_3	Male vocal	4.1	73.8	89.3	1502	2486	0.6
	Mean (and SD)	4.1 (0.4)	74.7 (3.9)	92.3 (2.6)	1806 (2639)	1608 (1176)	1.2 (0.8)

Table 6.6: Details of the stimuli used in experiment 2, organised by instrument category: bass, guitar, percussion, and singing voice.

An overview of the stimuli is given in Table 6.6. The SPLs refer to the calibrated levels derived from a preliminary listening task (see Section 6.3.1). The RMS and peak levels have been approximated by taking into account the average sensitivity and frequency response of the headphones used throughout the experiment. Also listed is the median of the spectral centroids, spectral spreads and Qs (centroid / spread) as calculated from the power spectrograms of the stimuli.

Choice of method

Since the objective of this study was to establish a subjective dataset for assessing the predictive quality of loudness models when applied to musical sounds, a loudness-matching paradigm was used. As discussed in Section 6.1, there are a number of different experimental procedures for measuring loudness. Those that adapt to the listener's response over time are commonly employed in psychophysical experiments involving laboratory stimuli, e.g. Zhang and Zeng (1997) and Buus et al. (1999), because of their precision and ability to target specific points on the psychometric function. However, because probabilistic sampling requires many trials for a single match, these approaches are often very time-consuming, especially when the stimulus duration exceeds one or two seconds. This became evident when conducting the previous experiment which lasted around 1.5 hours per subject to match all pairwise combinations of four sounds with replication (a total of 32 matches).

To further investigate the practicalities of different measurement designs, a number of short pilot experiments were conducted. The time required to match the loudness of each of six pairs formed from three short musical segments (2 to 4 s duration) was measured on two participants using two procedures. The first was the MOA and the second was a two-interval forced-choice procedure, similar to the one used in experiment 1. In both methods the fixed and variable sounds where presented in succession with a 0.5 s inter-stimulus gap. For the MOA, the listener had direct control of the level of the variable stimulus and could submit the final matching level at any point, terminating the repeating sequence. A 5 dB step size was employed for the first two reversals which was then reduced to 2 dB. The adaptive track ended after a total of 10 reversals. The average time to complete a single loudness match was 47 s for the MOA and 140 s for the adaptive procedure. Additional variations of these two paradigms were implemented and tested by the author. For example, interleaving multiple adaptive tracks, as done in experiment 1, can aid to reduce bias effects (Florentine et al. 1996). However, when the author performed loudness matching using all six concurrent tracks, the complete experiment time grew to 17 minutes which became tiresome and made the exercise laborious.

Adaptive procedures are perhaps better suited to experiments involving stimuli of very short-duration, and, if a large number of matches are required, regular breaks should be scheduled. In addition, participants from an audio engineering background tend to prefer more direct measurements as they are familiar with process of balancing acoustic signals. Feedback from the two subjects used in these short pilot experiments supported this by reporting that they preferred to be actively involved in the task, rather than passively responding to the computer for long periods of time. In addition, the MOA is suitable for relatively long stimuli because the subject can adjust the level of the variable stimulus during playback (Skovenborg et al. 2004).

MOA with constraints

In order to accommodate an experiment design involving many pairs of 4 s sound excerpts, it was decided to obtain measurements of subjective loudness equality using a variation of the MOA which incorporated constraints to minimise bias effects, at the cost of being slightly less time effective (average match time approximately 64 s). This procedure is now described.

Given a pair of sounds, the subject could listen to either one by pressing the computer keyboard buttons labelled '1' or '2'. Each time a button was pressed, current playback was terminated and the sound mapped to the activated button was triggered. At no point could both segments be auditioned simultaneously. The level of the first sound (presented on the left of the screen) was always fixed and the subject varied the level of the second sound (presented on the right) to determine the point of equal loudness. The subject could increase or decrease the level of the variable stimulus by pressing the 'up' or 'down' arrow key, respectively. When satisfied with their adjustment, the participant was told to press the 'Accept' button.

The participant was instructed to 'bracket' the point of equal loudness by making the variable stimulus both louder and softer than the fixed stimulus several times. This aids the subject to 'home in' on the PSE. To ensure that listeners were bracketing the point of equal loudness, the following constraints were built into the software. First, the step size for the change in level was 3 dB until two reversals had occurred, after which it was reduced to 1 dB.[8] The final step size was 0.5 dB after a total of four reversals had been registered. Exponential averaging was applied to the level adjustments using a time constant of 22 ms. The 'Accept' button was only registered if these constraints had been satisfied, otherwise, a 'Keep bracketing' message was displayed momentarily.

The starting levels of the stimuli were derived from a preliminary experiment (see Section 6.3.1). These levels served to calibrate the sounds for approximate equal loudness at a comfortable listening level. As in experiment 1, any additional gain applied during the experiment was measured relative to these starting levels.

Finally, because the listener could trigger the two stimuli at will, it is unlikely that the sounds were auditioned over their entire duration throughout a trial. One fix is to implement an alternating playback loop at the cost of increasing the duration of the experiment. Instead, a compromise was made by forcing the subject to listen to a single playback sequence upon pressing the 'Accept' button, in which the two segments were presented successively in full duration. This sequence reoccurred if the level of the variable stimulus was readjusted, otherwise a second press of the 'Accept' button was required to confirm the decision. Once confirmed, the relative gains were stored.

To minimise bias effects pertaining to presentation order and/or influence of the fixed stimulus, each segment in a given pair served as a reference, i.e. matches were obtained as (i, j) and (j, i). A total of four matches were obtained for each segment pair: two with each segment varied. The level of the variable sound was set alternately above and below its calibrated level by setting the relative gain to a random value from a ± 10 dB range. This source of randomisation was introduced to minimise effects caused by the initial level of the variable sound. A pair of segments was selected at random from one of 40 unique pairs (see Section 6.3.1) and if the two replicate matches were complete, the role of each segment was reversed, i.e. the fixed stimulus became the variable stimulus and vice versa. With the roles reversed, the relative gain of the (new) reference sound was set to the average gain of the (previous) variable sound as previously set by the listener. This design gave a total of four loudness matches per segment pair, and with 40 segment pairs, this totals 160 matches per subject.

For each loudness match, the difference between the relative gains of the variable and fixed sounds was calculated. An example of estimating the difference in relative gain for a pair of segments pair following the experimental procedure outlined above is given as follows:

- Fixed, variable segment indices: (i, j).
  Initial gains: $L_j = 8$, $L_i = 0$.
  Final gains: $L_j = 3$, $L_i = 0$.
  Level difference: $L_j - L_i = 3$.

- Fixed, variable segment indices: (i, j).
  Initial gains: $L_j = -5$, $L_i = 0$.

[8] Again, a reversal is defined as a change in the direction of the variable stimulus level.

Final gains: $L_j = 2$, $L_i = 0$.
Level difference: $L_j - L_i = 2$.

- Fixed, variable segment indices: (j, i).
  Initial gains: $L_i = 10$, $L_j = 2.5$.
  Final gains: $L_i = -1$, $L_j = 2.5$.
  Level difference: $L_i - L_j = -3.5$.

- Fixed, variable segment indices: (j, i).
  Initial levels: $L_i = -3$, $L_j = 2.5$.
  Final gains: $L_i = -0.5$, $L_j = 2.5$.
  Level difference: $L_i - L_j = -3$.

- Estimate of unbiased matching gain required for equal loudness between fixed segment i and variable segment j: $\Delta L_{i,j} = (3 + 2 + 3.5 + 3)/4 = 2.875 \text{ dB}$.

Balanced pair-matching design

Skovenborg and co-workers (Skovenborg and Nielsen 2004a; Skovenborg et al. 2004) define a balanced pair-matching experiment design as one in which the frequency of occurrence of each sound segment is the same, with no segment distinguished as the reference. Their mathematical specifications indicate a preference for each segment to be paired with another segment only once. For example, an experiment involving three sound segments yields the pair combinations (0, 1), (2, 0) and (1, 2). In this example, the pair order is also balanced in that every segment serves as both the fixed and variable stimulus once. Notice that this design makes use of redundancy because each stimulus occurs multiple times but in different pairs. Skovenborg argues that this design compares favourably against the fixed-reference design where all stimuli are matched against a single reference stimulus. The benefits of the balanced pair-matching experiment design are summarised as follows:

- The choice of reference stimulus is no issue.
- Any bias introduced by a particular reference is avoided—or at least spread out over all segments when using a linear model of the data.
- Variance associated with individual adjustments can be evened out through the inherent redundancy.
- All adjustments can be used to estimate the relative loudness of each segment via a linear model and so systematic errors can be estimated and thus minimised.

Requirements

Skovenborg and Nielsen (2004a) specify the formal requirements of a balanced pair-matching design. Given N sound segments, there are a total of N^2 possible pairs, which can be expressed as a complete experiment design matrix D of size $N \times N$. Element $D(i, j)$ means that segment i is fixed in level (denoted stimulus A) and segment j is variable (denoted stimulus B). To determine if a given design is balanced, the authors specified the following variables:

$$countA(i) = \sum_{j=1}^{N} D(i,j), \tag{6.7}$$

$$countB(j) = \sum_{i=1}^{N} D(i,j), \tag{6.8}$$

$$countAB(i) = countA(i) + countB(i) \quad \text{and} \tag{6.9}$$

$$pairCount(i,j) = D(i,j) + D(j,i), \tag{6.10}$$

where *countA* and *countB* determine the number of times each segment occurs as fixed stimulus A and variable stimulus B, respectively. Accordingly, *countAB* gives the total frequency of a segment, which might include duplicate pairs in the case of $D(i,j) = D(j,i) = 1$. Finally, *pairCount* counts the total number of times a specific *pair* occurs in the design.

The authors qualify a design as balanced if the following criteria are met:

$$countMaxDiff \leq 1; \quad countRange = 0; \quad pairCountRange = 0; \tag{6.11}$$

where

$$countMaxDiff = \max(|countA - countB|), \tag{6.12}$$

$$countRange = \max(countAB) - \min(countAB) \quad \text{and} \tag{6.13}$$

$$pairCountRange = \max_{i \neq j}(pairCount(i,j)) - \min_{i \neq j}(pairCount(i,j)). \tag{6.14}$$

When $countMaxDiff = 0$, each segment occurs the same number of times as segment A and B. When $countMaxDiff = 1$, a given segment is matched once more either as stimulus A or B. The *countRange* variable compares the maximum and the minimum absolute frequency of the segments. For example, when $countRange = 0$, every segment occurs the same number of times. In the case that $pairCountRange = 0$, all $N(N-1)/2$ pairs are matched and there are no duplicates. To allow for a subset of the total unique matches, the authors relaxed the requirements to:

$$countMaxDiff \leq 2; \quad countRange \leq 1; \quad pairCountRange \leq 1. \tag{6.15}$$

One problem is that a balanced design, according to Equations 6.11 and 6.15, is no longer balanced if the design is made symmetric. This is demonstrated in Figure 6.7, where subplot (a) which shows a balanced experiment design for $N = 6$ according to the relaxed requirements of Equation 6.15. The design is classified as balanced because each segment occurs three times in three different pairs, and each segment serves as stimulus A and B at least once. However, when the design is made symmetric, as shown in subplot (b), the balanced pair-matching requirements are no longer satisfied, because although each segment is matched six times, only three of these are distinct. In essence, Equation 6.15 favours unique matches over replicates. Note also that this design has perfect symmetry, in that each segment serves as the variable and fixed stimulus the same number of times.

To accommodate symmetrical designs, such as the one shown in Figure 6.7(b), or balanced pair-matching designs involving repeated matches, i.e. $D(i,j) = 2$, the problem can be reformulated

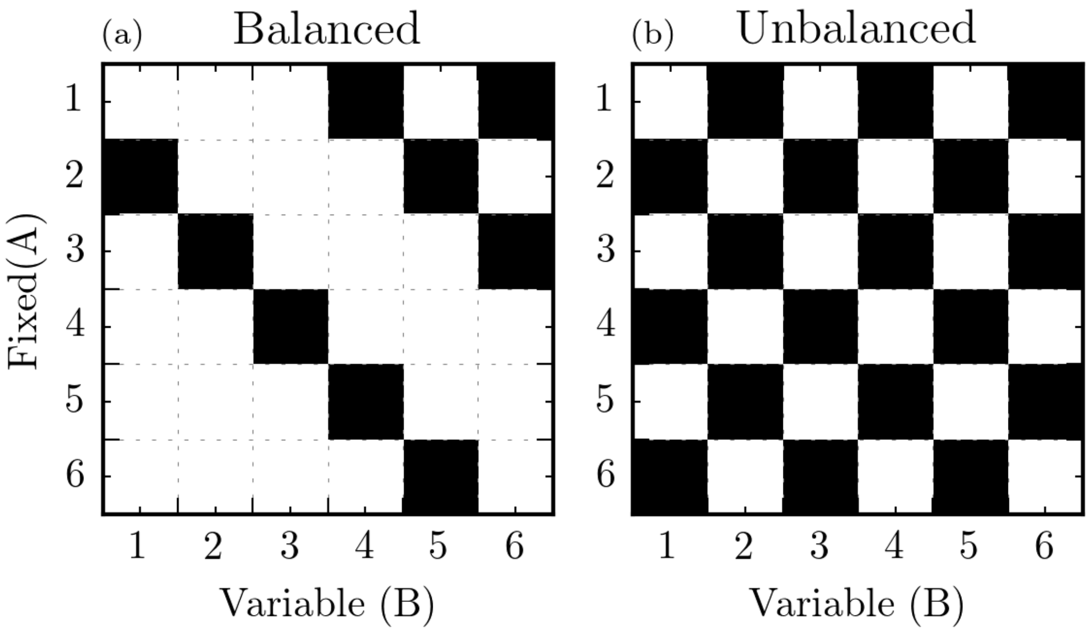


Figure 6.7: (a) Balanced and (b) unbalanced experiment design according to Equation 6.15.

here as follows:

$$pairCount(i,j) = D(i,j) + D(j,i), \tag{6.16}$$

$$count_0(i) = \sum_{j=1}^{N} pairCount(i,j), \tag{6.17}$$

$$count_1(i) = \sum_{1\leq j\neq i}^{N} 1, \quad \text{if } pairCount(i,j) > 0 \quad \text{and} \tag{6.18}$$

$$connectionCount = \sum_{i,j\in K} pairCount(i,j), \tag{6.19}$$

where K is a set of 2D indices specifying the upper triangle of the matrix $pairCount$. $count_0$ specifies the absolute frequency of a segment. For example, if $D(i,i) = 1$, i.e. segment i is matched against itself, then $count_0(i)$ is incremented by two. In contrast, $count_1(i)$ denotes the number of *other* segments that segment i is matched against, and ignores duplicate pairs. The number of unique matches, where $i \neq j$, is measured by the variable *connectionCount*. For a connected design in which every segment occurs the same number of times, and every segment is matched to the same number of other segments, the requirements of the balanced pair-matching experiment design can be modified as

$$rangeCount_0 = 0; \quad rangeCount_1 = 0; \quad connectionCount \geq N(N-1)/2; \tag{6.20}$$

where

$$rangeCount_0 = \max(count_0) - \min(count_0) \quad \text{and} \tag{6.21}$$

$$rangeCount_1 = \max(count_1) - \min(count_1). \tag{6.22}$$

Note that $countMaxDiff$ is not included in these specifications, as empirical investigations indicate that it is not needed for capturing the intercept or bias associated with choice of fixed stimulus. However, this variable may easily be incorporated as in Equation 6.11. To accommodate more flexible designs, e.g. for a user-specified number of matches, the following requirements can be used instead

$$rangeCount_0 \leq 1; \quad rangeCount_1 \leq 1; \quad connectionCount \geq N-1. \tag{6.23}$$

Measure	Asymmetric	Symmetric
$countA$	1,2	3
$countB$	1,2	3
$countAB$	3	6
$pairCount(i \neq j)$	0,1	0,2
$countMaxDiff$	1	0
$countRange$	0	0
$pairCountRange$	1	2
$count_0$	3	6
$count_1$	3	3
$rangeCount_0$	0	0
$rangeCount_1$	0	0
$connectionCount$	9	9

Table 6.7: Balanced-pair matching design measures given by Skovenborg and Nielsen (2004a) (top half) and those suggested in this chapter (bottom half) for the two experiment designs shown in Figure 6.7.

Table 6.7 shows the measures reported by Skovenborg and Nielsen (2004a) (top half) and those presented here (bottom half) for the two experiment designs shown in Figure 6.7. The symmetric design is unbalanced according to Equation 6.15 because the *pairCountRange* variable exceeds one. In contrast, both designs are balanced according to Equation 6.23. Note that $connectionCount = 9$ for both designs, because there are nine unique matches.

If the experimenter wishes to maximise the number of unique matches and minimise pair replication, as preferred by Skovenborg et al. (2004), then the following criteria can be used:

$$rangeCount_0 \leq 1; \quad rangeCount_1 \leq 1; \quad connectionCount \geq \min(N(N-1)/2, nMatches); \tag{6.24}$$

where $nMatches$ is the desired number of matches, which should be at least $N-1$ to ensure the N segments are connected. Although not used in this study, this final set of requirements should be used for generating random balanced designs when the goal is to maximise diversity in the listening experiment. These requirements then make it apparent that the design favours segment-pair diversity over symmetry, and is thus more in line with Equation 6.15. Algorithms developed by the author suggest that, for a given number of matches, random experimental designs are more easily generated using the above criteria compared to those given by Equation 6.15, primarily because $countMaxDiff$ is dropped from the requirements.

Design matrix

In experiment 1, it was possible to have subjects match every possible pair of segments because of the small number of sounds. The question now is whether a full balanced experiment design involving $N = 16$ sound segments can be afforded. This gives a total of $N(N-1)/2 = 120$ pairs, denoted $nTotalPairs$. Recall that in this experiment, where within-subject consistency is

[9] Four matches were obtained per pair (two repeated matches were measured with each segment varied in level).

considered and symmetry in the matching design matrix is accounted for, this would require a total of $4 \times 120 = 480$ matches per subject.[9] As discussed, pilot experiments using the selected loudness-matching procedure revealed that a single match of two four second segments lasted approximately one minute on average. Thus, with 480 matches, one can expect an experiment duration of eight hours per subject which is impractical. Instead, Skovenborg and Nielsen (2004a) suggested taking a subset of the total pairs from the full experiment design in order to estimate the relative loudness levels of *all* segments. Skovenborg's investigations led to the suggestion that the deviation from the best estimate of the relative loudness levels of all segments (derived from the full balanced-pair experiment) decreases linearly as a function log number of pairs. The reduction in mean absolute error from the minimum fixed-reference design involving $N-1$ matches can be represented as a percentage:[10]

$$r = 1 - \frac{\log(connectionCount/(N-1))}{\log(N/2)}. \tag{6.25}$$

Thus, $r = 0$ is equivalent to a fixed-reference design with just $N-1$ adjustments, and hence no reduction in error, and $r = 1$ is the full experiment involving $N(N-1)/2$ adjustments. $r = 0.5$ corresponds to the 50% error reduction point, involving

$$nPairs_{50\%} = \lceil (N-1)e^{\log(\frac{N}{2})0.5} \rceil \tag{6.26}$$

unique pairs. Equation 6.26 gives the number of pairs required, on average, to achieve a mean absolute deviation from the best estimates of the segment levels that is half-way between the deviation resulting from the minimum fixed-reference design and that of the complete experiment (Skovenborg and Nielsen 2004a). Given $N = 16$, the 50% error reduction point was estimated using Equation 6.26 as 43 unique pairs. However, if each of the 16 segments occur in five distinct pairs, a total of 40 distinct pairs may be used which is equivalent to the 47% error reduction point. It is important to highlight that the design used here incorporated additional redundancy in that both every segment pair was matched four times. Of course a larger dataset could have been obtained by simply selecting random pairs of sounds from a larger collection of stimuli, i.e. 40 distinct pairs made up from 80 segments. However, by scaling the experiment down to sixteen sounds, the benefits of the balanced pair-matching design can be reaped. Furthermore, balanced pair-matching design allows for checks of internal consistency in the loudness-matching data.

Figure 6.8 shows how the experiment design can be expressed as connected graph, where the vertices depict the stimuli and edges show connections between them. Notice that the graph is undirected (the matrix is symmetric) because both orders of every pair have been included in the design. The general idea is that a highly connected graph spreads the variance across all conditions (Skovenborg and Nielsen 2004a). The experiment design is indeed balanced according to the revised requirements of Equation 6.23. Stimuli were organised such that every segment occurs five times and was paired with at least one segment from each of the four instrument categories. For example, segment B_0 was matched against segments B_1, P_0, G_0, V_0 and V_3. With this configuration, every instrument category is paired against the other categories with nearly the same frequency.

Subjects

14 male listeners aged between 21 and 28 (median = 23) from the Sound Engineering and Production degree course at Birmingham City University participated in the experiment. Before starting the experiment, all subjects familiarised themselves with the 16 sound segments set to the cal-

[10] Note that Skovenborg and Nielsen defined the minimum experiment as one involving N adjustments, but here $N-1$ is used since this reference-to-reference match is essentially redundant.

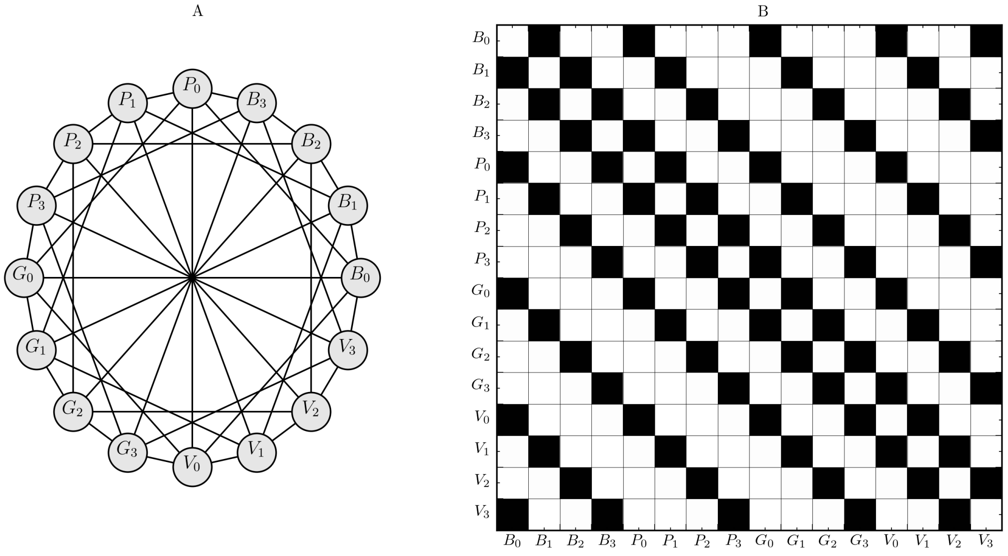

A
B
P_0
P_1
P_2
P_3
B_3
B_2
B_1
B_0
G_0
G_1
G_2
G_3
V_0
V_1
V_2
V_3
B_0 B_1 B_2 B_3 P_0 P_1 P_2 P_3 G_0 G_1 G_2 G_3 V_0 V_1 V_2 V_3

Figure 6.8: Balanced experiment design represented as an undirected graph (A) and the corresponding symmetric adjacency matrix (B).

ibrated sound levels (see below). A 10-minute training stage was then initiated which allowed subjects to become acquainted with the user interface and gain experience with the loudness-matching procedure. Daily assessments required each subject to perform loudness matching in two separate 20 minute sessions, which were time managed by the software. On average, each subject participated in about four of these 50 minute assessments.

Apparatus

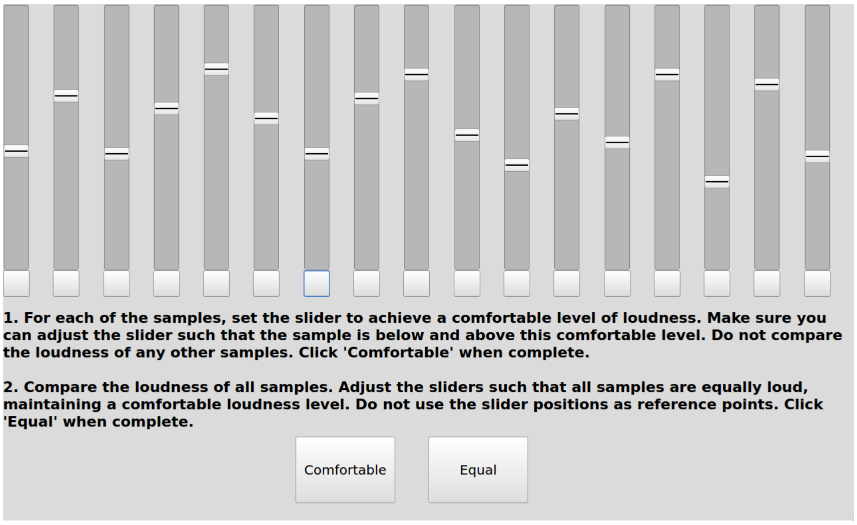
The listening equipment was the same as used in experiment 1 (refer to Section 6.2.1). The stimuli were sampled at 44.1 kHz, monophonic, and presented diotically over headphones. All listening assessments took place in the semi-anechoic chamber at Birmingham City University.

Presentation level

Because the relationship between loudness and intensity is nonlinear, the results from a set of loudness matches obtained at two different presentation levels will differ. Given that RMS normalisation does not guarantee equal loudness (as determined in experiment 1), it was decided to run a small-scale listening assessment such that the stimuli were approximately loudness normalised and presented at a comfortable level. Although Bech and Zacharov (2006) recommend using a dynamic loudness metric to normalise the loudnesses of heterogeneous sounds, such as the ones used in this study, it is not clear which metrics apply to single musical instruments.

Nine sound-engineering students participated in the initial loudness calibration test which consisted of two stages. Six of the subjects used in this listening test also participated in the main experiment. First, each subject was presented with a set of sliders and buttons (one per stimulus) and asked to adjust the slider position in order to obtain a comfortable listening level independently of all other sounds. Playback was triggered by pressing one of seventeen buttons corresponding to each sound (a 1 kHz tone was included). If a button was pressed during playback, the audio was terminated and playback of the desired sound would commence. Once complete, the fader settings were stored and the subject moved on to the second stage where the task was to simply 'tweak' the sliders such that all sources were equally loud but still comfortable to listen to. Once satisfied, a second set of levels were stored. This procedure was repeated twice for each subject. This two-stage process was simple for the subjects who were all experienced with balancing signals via sliders.

The stimuli listed in Table 6.6 were used to formulate the listening material with the addition of a 1 kHz pure tone to yield an approximate mapping to the phon scale. Each slider could be used to attenuate the source according to a 50 dB range and a resolution of 0.25 dB. A random offset between -10 and 0 dB was applied independently to each slider, in each of the two trials. The slider gain was applied to the peak normalised stimuli, which were presented in random order. Figure 6.9 shows the graphical user interface as used by the subjects.



1. For each of the samples, set the slider to achieve a comfortable level of loudness. Make sure you can adjust the slider such that the sample is below and above this comfortable level. Do not compare the loudness of any other samples. Click 'Comfortable' when complete.

2. Compare the loudness of all samples. Adjust the sliders such that all samples are equally loud, maintaining a comfortable loudness level. Do not use the slider positions as reference points. Click 'Equal' when complete.

Figure 6.9: Test interface for the two-stage loudness balance experiment. The subject is able to control the stimulus levels using the sliders.

Results

Data points corresponding to the pure tone were discarded as the majority of subjects reported the tone to be too uncomfortable to listen to when set as loud as the musical stimuli. A few participants also reported that it was difficult to compare the loudness of the sinusoid to the other sounds. This is in agreement with Skovenborg et al. (2004) who found a high between-subject disagreement for segment pairs involving tones. Consequently, in most cases, inclusion of the 1 kHz tone biased the derived levels towards a lower presentation level. For each subject, repeated data points were averaged and a set of equal-loudness gains were obtained. These are a vector of relative signal gains (in dB) that can be used to loudness normalise the stimuli. The subjective equal-loudness gains were then averaged to estimate a single representative set of levels. The calibration levels for the main experiment were obtained by offsetting this gains such that the mean gain was equal to the overall average gain set by the subjects in stage one of the experiment (excluding the pure tone).

6.3.2 Results

Each subject performed a total of 160 loudness matches involving 40 distinct pairs formed from 16 musical sounds. Loudness matches were obtained twice with each segment serving variable stimulus. The dependent variable ΔL represents the difference in relative signal gain between the variable and fixed stimulus at the point of equal loudness. Along with the matching level, the response time of each match was recorded. In this section, an analysis of both variables is made, with particular focus on the adjustments, before reporting the results of the statistical model.

Response time

The duration of every loudness match was recorded for each subject. Excluding two obvious outliers in which two matches lasted over 5 minutes, the mean matching time averaged across the 14 participants was 61.0 s (SD = 14.4). The total duration of the experiment was around 38 hours, amounting to approximately 2.7 hours per subject to match the loudness of 40 distinct segment pairs with replication.[11]

The response times were averaged across the forty pairs for each subject and replication, such that time-order effects could be investigated. That is, after hearing and matching the loudness of a pair of sounds once, did the listener respond faster on subsequent matches involving the same pair? A repeated-measures ANOVA with response time as the dependent variable and replication number (1–4) as a numerical (not categorical) independent variable was performed to address this question. The ANOVA was performed using a mixed-effects model with a correlated random intercept and slope for each subject as this significantly improved the fit over the random intercept-only model ($\chi^2(2) = 9.0$, $p = 0.01$) and the random slope-only model ($\chi^2(2) = 66.1$, $p = 0$).[12] The estimated slope describing the relationship between response time and replication number was -3.8 s (SE = 0.7) and was found to be a statistically significant predictor when compared against the intercept only model ($\chi^2(1) = 16.7$, $p = 0.00004$). The negative slope means that the response time decreased, on average, by 3.8 s per reoccurring pair, i.e. the subjects became faster at loudness matching as their experience developed. Since the random intercept and slope model were statistically significant, overall response time varied across the subjects (SD = 17.4 s), as did the decrease in response time (SD = 2.0 s). Note that pair replicates three and four correspond to cases where the role of stimuli in a given pair were reversed. Boxplots of the data, along with the estimated regression line, are shown in Figure 6.10.

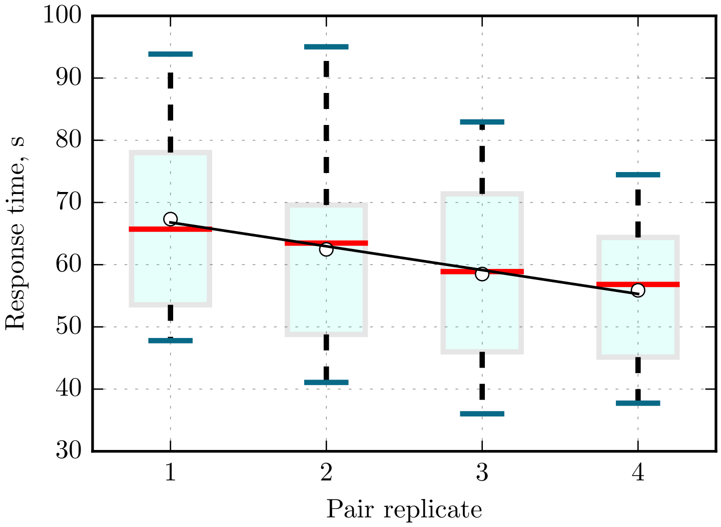


Figure 6.10: Boxplots showing the distribution of response times for a given set of matches. The red line and white circle in each boxplot corresponds to the median and mean of the response times, respectively. The black solid line corresponds to the line of best fit as derived from the ANOVA.

[11] Two matches as (i, j) and two matches as (j, i).

[12] There are two degrees of freedom because the random intercept-slope model also estimates a correlation term.

The matching gains

Subject consistency

Each subject matched the loudness of 40 pairs twice with each stimulus varied in level. The GMD statistic was used to quantify the consistency of the subjects across the four matches associated with each pair. The average GMD was 2.0 dB (between-subject SD = 0.4; between-pair SD = 0.5). The within-subject SD associated with each pair of sounds was also calculated. The average SD was 1.6 dB (between-subject SD = 0.3; between-pair SD = 0.4). These measures given an indication of the consistency and typical resolution of the subjects when judgements the loudness of single instruments.

Lower quartile		Upper quartile	
Pair	GMD	Pair	GMD
B_0/B_1	1.3 (0.9)	B_3/P_3	2.1 (0.9)
V_0/V_1	1.3 (0.8)	P_1/P_2	2.2 (1.3)
G_1/G_2	1.4 (0.5)	B_0/G_0	2.3 (0.9)
G_2/G_3	1.5 (0.8)	P_3/G_3	2.3 (1.4)
V_1/V_2	1.5 (0.6)	P_0/G_0	2.4 (0.7)
G_1/V_1	1.5 (0.6)	G_2/V_2	2.5 (1.1)
G_3/V_3	1.6 (0.7)	P_0/V_0	3.0 (1.9)
B_2/V_2	1.6 (0.6)	P_2/V_2	3.1 (1.9)
B_0/V_3	1.6 (1.1)	P_2/G_2	3.1 (2.2)
B_1/P_1	1.6 (0.8)	P_2/P_3	3.3 (1.3)

Table 6.8: Average GMD across subjects (and SD) for segment pairs occurring in the lower and upper quartiles of the GMD distribution.

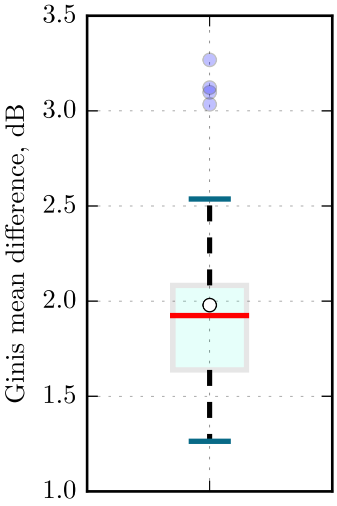


Figure 6.11: Boxplot of the mean GMDs for the 40 segment pairs.

To identify segment pairs that were matched with varying degrees of consistency, the GMDs were averaged over subjects yielding a single consistency measure for each of the forty pairs. Table 6.8 lists the mean (and SD) of the GMDs associated with segment pairs falling in the lower and upper quartiles of the distribution of values, as shown in Figure 6.11. Those in the lower quartile have GMDs ranging between 1 and 2 dB, and those in the upper quartile vary between 2 and 3.5 dB. Of the ten pairs constituting the lower quartile, five consisted of segments from the same instrument category. Both B_0/B_1 and V_0/V_1 were the most consistently matched pairs, with GMDs of just 1.3 dB. It is interesting to note that the segments involved in these pairs share similar crest factors and spectral distributions (see Table 6.6), although this cannot be said of all pairs found in the lower quartile.

Of the ten most consistently matched pairs, only one contained a percussive sample, yet eight of the ten least consistently matched pairs contained percussive samples. Note also that the GMDs associated with pairs involving percussive samples tend to have higher SDs, indicating that there is greater variation in the consistency of the subjects for sounds which are generally more difficult to judge. The least consistently matched pair of sounds was the shaker P_2 and the snare drum P_3, with a mean GMD of 3.3 dB. Furthermore, four of the five pairs involving the shaker appear in the upper quartile, as are three of the pairs involving the snare drum. A few participants expressed difficulty in gauging the loudness of the shaker recording because the shaker 'gets progressively louder towards the end of the stimulus.' Therefore, participants may have applied different temporal weights to this segment throughout the experiment. A final observation is that segments G_2 and V_2 both occur in the lower quartile twice, with average GMDs of about 1.5 dB, but when paired together the mean GMD increased to 2.5 dB, putting them in the upper quartile. This shows (indicatively), that subject consistency is very much dependent on the pairing of sounds, and as such it is very difficult to distinguish a suitable single instrument for use as a reference stimulus in loudness-matching experiments.

Between-subject agreement

Between-subject agreement was measured using the SD and GMD of the 40 unique loudness matches (after averaging replicates) submitted by the 14 subjects. The SD indicates the average spread of the matching levels for a given pair, whilst the GMD describes the average difference in matching level submitted by any two subjects. The average inter-subject SD was 2.5 dB (SD = 0.8) and the median also 2.5 dB. The average GMD across the 40 pairs was 2.9 dB (SD = 1.0), the median was 2.8 dB, and the 95% percentile was 4.4 dB.

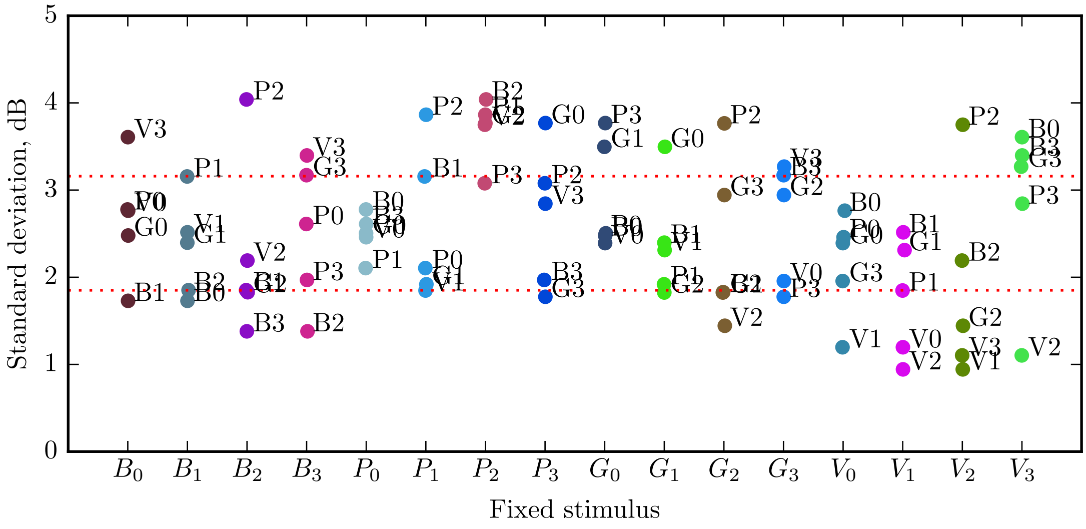


Figure 6.12: Sample SDs of the loudness matches for the forty pairs, arranged by stimulus. The lower and upper red dotted lines are the 25th and 75th percentiles of the SDs, respectively.

A scatter plot of the SDs associated with the 40 pairs is shown in Figure 6.12. The SDs are arranged by stimulus to display the between-subject agreement associated with specific segment combinations. As a result, the plot shows 80 points, but only 40 are unique. The two red dotted horizontal lines represent the interquartile range of the 40 SDs. It can be seen that 75% of the matches had SDs less than 3.2 dB and 25% were less than 1.9 dB. As expected from measures of intra-subject variation, matches involving the shaker P_2 are associated with large inter-subject variation (mean SD across five pairs = 3.7 dB, SD = 0.4). Given that the other sounds paired with this segment generally have lower SDs when paired with other sounds, it seems likely that the listeners were both uncertain and in disagreement with one another when judging the loudness of this particular segment. Interestingly, pairs with high subject agreement are comprised of segments from the same instrument category (notably bass and voice). The three pairs with the lowest SDs involved voice samples. Sample V_3, a recording of male singer, had an SD of 1.1 dB when paired with V_2 (female singer), yet when paired with bass, percussion and guitar segments, the variation increased to around 3–4 dB. The high consensus amongst listeners for pairs of voice samples might be explained partly by the voice being a common everyday sound, and partly by the mixing experience of the subjects where vocal stems tend to take priority. However, such trends are not apparent for the guitar or percussive samples, most likely due to greater variation in the physical characteristics of the selected clips, e.g. shaker versus drum kit. From the percussion class, subjects found the bass drum P_0 and drum kit P_1 easier to match compared to other percussive combinations (SD = 2.1 dB), likely because these two segments share similar timbral qualities.

A linear model of the matching gains

In experiment 1, a mixed-effects model was used to estimate the relative loudness of the stimuli and test for differences in loudness between them. It should be noted that even if the levels underlying a loudness-matching experiment are not reliably different from zero, giving rise to a non-significant statistical fit, the variability of those estimates are still important, especially if the goal is to use them for evaluating different loudness models.

A mixed-effects model was fit to the group data, where the intercepts and slopes were allowed to vary across subjects (as in experiment 1). The design matrix consisted of 160 rows per subject, which included repeated matches, and 16 columns representing the different segments. As before, the first column was replaced with a vector of ones, denoting the intercept or average gain in the experiment. The mixed-effects model failed to converge when allowing for correlated random slopes, likely due to the large number of parameters to be estimated from the data.[13] The model was therefore fit without correlation. The random slopes were highly significant ($\chi^{15} = 1298.6$, $p = 0$), indicating a reliable source of between-subject variability in the parameters. Furthermore, the fixed-effects were also statistically significant ($\chi^{15} = 45.2$, $p = 0.00007$), supporting the hypothesis that the stimuli were *not* equally loud (at the group level) when presented at the calibration levels derived from the pilot experiment. A Q-Q plot of the standardised residuals revealed distinct heavy-tails, meaning the assumption of normally distributed within-subject errors could not be made (also confirmed by the Shapiro-Wilk test of normality: $W = 0.99$, $p = 2 \times 10^{-7}$). Furthermore, when no correlation of slopes was used, the distribution of the random effects was found to depend on the choice of reference stimulus (used for the intercept), thus influencing the distribution of the coefficients across subjects.

Due to uncertainties with the mixed-modelling approach, it was decided to obtain per subject estimates from individual regressions and investigate how well these parameters summarise the observed data. The average R^2 from the single-subject fits was 68.7% (SD = 10.7%) and the mean RMSE was 1.8 dB (SD = 0.4). A permutation test was run to identify if the segment levels (predictors) accounted for the explained variation beyond chance level. To do this, the rows of the design matrix were shuffled and the per subject models were re-fit to the data. This is consistent with the null hypothesis that if the stimuli were judged equally loud, the coefficients would not capture any variation based on differences in stimulus loudness. Using 10 000 unique permutations, which included the original specification of design matrix, the average of the R^2 and RMSE permutation distributions was 14.7% (SD = 3.0) and 3.1 dB (SD = 0.06), respectively.

Figure 6.13 shows the permutation distributions for the subject-level R^2 statistic. The original test statistics are shown by the red diamonds—clearly the experiment design matrix captures the true structure of the matches submitted by the subjects. Since the permutation distributions only included a test statistic as large as those from the original fits once (in all cases), it can be concluded that the original observed statistics are unlikely to occur under the null hypothesis ($p = 0.0001$,

[13] Again, the function `lmer` was used to model the data.

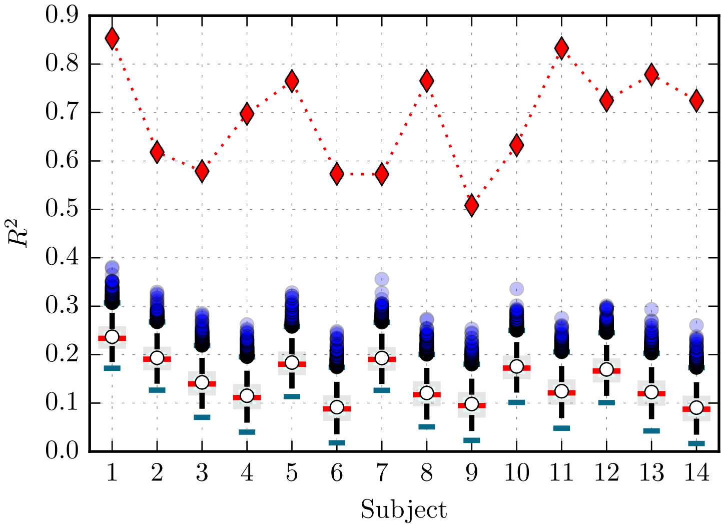


Figure 6.13: Boxplots showing the permutation distribution of the R^2 statistic for each subject. Red diamonds connected by dotted lines show the observed R^2 value according to the experiment design.

95% CI = [0.00002, 0.00057]) and therefore that the relative loudness estimates do indeed capture systematic variation in matching data.

Residual and influence plots were used to determine potential influential points for each individual. In particular, large studentised residuals, high-leverages and outlying Cooks distances were used to identify potential problematic observations. Inspection of the residuals by pair replicate (1–4), suggested a systematic effect which may have been caused by the initial loudness of the variable stimulus, potentially combined with a time-order effect (see Figure 6.10).

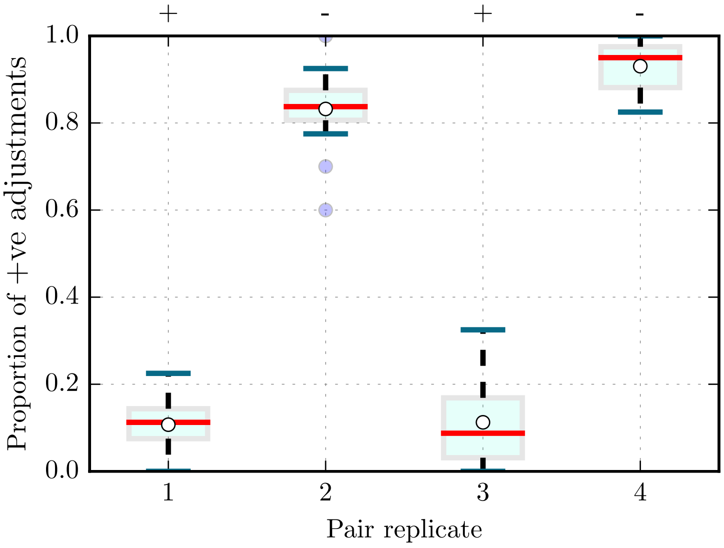


Figure 6.14: Proportion of positive level adjustments submitted by the 14 subjects for pair replicates 1–4. The sign (+/-) of the random level offset applied to the variable stimulus is labelled on the upper abscissa.

Figure 6.14 shows a boxplot, for each replicate match, of the average proportions of the matches submitted by the subjects that were classified as a positive level adjustment. That is, if the subject incremented the level of the variable stimulus more times than they decremented the level, the match was classified as a positive adjustment. Match replicates one and two correspond to segment pairs matched as (i, j), and replicates three and four denote those matched as (j, i). It can be seen that the subjects applied more attenuation in the first and third rounds, and a greater boost in level in the second and fourth rounds. These differences are attributable to the random level offset that alternated between positive and negative values for each consecutive same-segment pair match (as shown by the upper abscissa). That is, a positive level offset caused the listeners to reduce the level of the variable sound in order to match the loudness of the reference, and a negative offset led them to increase the level. If subjects were over- or under-adjusting the level of the variable stimulus, a non-zero difference in the matching level between the first and second replicates, and the third and fourth replicates, would be expected.

This effect was incorporated into the regression by adding a new column to the design matrix in which replicates one and three were denoted 1 (positive level offset), and replicates two and four were denoted -1 (negative level offset). Thus, the matching gains are predicted using the model (assuming i and j are non-zero):

$$\Delta L_{i,j} = l_0 + l_i - l_j + v \times \mathrm{sgn}(\zeta) \tag{6.27}$$

where the intercept l_0 captures the bias associated with the loudness of the fixed stimulus, and v is the bias associated with the initial loudness of the variable stimulus, as determined by the random level offset ζ. With this parameter included, the average *adjusted* R^2 value, which punishes raw R^2 for the addition of predictors, increased for the majority of subjects, but the overall effect was small (95% CI = [0.5, 2.3]% change). The addition of this variable did reduce a portion of the systematic pattern portrayed by the residuals, but a slight asymmetry between match rounds one and three was still observed, suggesting the presence of a temporal effect.

A conservative approach to data removal was taken: roughly 1% of the observations were removed from each fit. After removing a few data points per subject, the average R^2 across subjects was 72.8% (SD = 9.2) and the mean RMSE was 1.7 dB (SD = 0.3). These measures contrast to an average R^2 and RMSE of 88.4% (SD = 5.4) and 1.1 dB (SD = 0.2), when using the 80 matching gains, obtained after averaging replicate matches whilst maintaining both adjustment orders, to predict the same data. The remaining unexplained variation measured using the mean data is presumably attributable to pure within-subject error. When pair symmetry was ignored, i.e. averaging over all four pair replicates to arrive at 40 data points per subject, the mean R^2 was 76.9% (SD = 9.2) and the mean RMSE was 1.5 dB (SD = 0.3), respectively. Thus, the regression model performs on par with the averaging typically used to summarise loudness-matching data, e.g. that of Moore et al. (2003). However, it is important to stress that the relative loudness estimates are based on all matches submitted by the subjects, thus evening out variance and biasses associated with individual matches.

With the subjects parameters extracted from their respective models, the average intercept was 0.2 dB (95% CI = [-0.1, 0.5]), so the presence of a bias caused by the fixed or variable stimulus cannot be inferred at the group level, although the CI is indicative of a small positive bias. The average value for the term capturing the effect of the random level offset was also 0.2 dB, and reliably different from zero (95% CI = [0.1, 0.4]), suggesting that when the variable stimulus was initially louder than the reference, subjects set the final level slightly above the true matching level.

Figure 6.15 compares the experimentally measured gains at the point of equal loudness for the forty distinct pairs with those predicted using the parameters obtain from the regression models. Both raw and fitted matching levels were averaged over both adjustment orders. The correlation

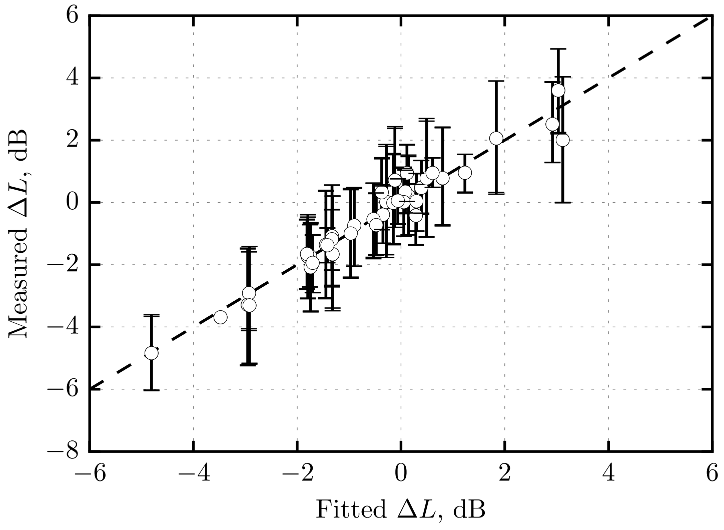


Figure 6.15: Observed difference in the matching gain between two sound segments at the point of equal loudness plotted against the fitted values derived from the single-subject regression models. Error bars show the bootstrapped 95% between-subject CI of each mean.

between the 40 measured matching gains and those fitted by the linear model was 0.98 with an RMSE of 0.4 dB. The dashed line in Figure 6.15 indicates where points would occur if there were no errors in the experimental data and the model fit were perfect. The close agreement between the measured and fitted values indicates that subjects were internally consistent when matching different pairs made up from the sixteen sounds, i.e. the assumption of transitivity appears to be valid. The figure also suggests that the majority of pairs were equally loud to within ±2 dB of the calibrated levels obtained from the initial loudness balancing experiment. A small proportion of pairs appear to be larger than this, implying that the loudness of only a few segments were not as well approximated by the balancing task. Eight of the nine segment pairs with measured average gains exceeding 2 dB contained percussive samples, suggesting that these samples were less accurately equalised by the balancing procedure. The largest average gain was -4.7 dB corresponding to segment pair P_0/V_0. The largest gain required for equal loudness estimated using the model parameters was -4.8 dB, corresponding to segment pair P_2/V_0 (which was not measured directly).

Estimating the equal-loudness gains

The relative loudness levels extracted from the matching data can be used to loudness normalise the 16 sounds. In the previous section the segment levels were made relative to the loudness of the first segment (serving as the intercept). To remove any bias, the intercept is set to zero. A set of equal-loudness gains (in dB) is then obtained by negating the relative loudness vector and then subtracting the mean. This transformation yields the best estimate of relative gains for aligning the loudness of the sounds. By centring the stimuli on zero, the choice of reference stimulus becomes a non-issue, and the gains are not biassed towards particularly quiet or loud sounds. A similar representation was proposed by Skovenborg et al. (2004) for evaluating the predictions of loudness models.

The white circles in Figure 6.16 show the equal-loudness gains estimated using the model parameters obtained from the main experiment. The 95% between-subject CIs were calculated

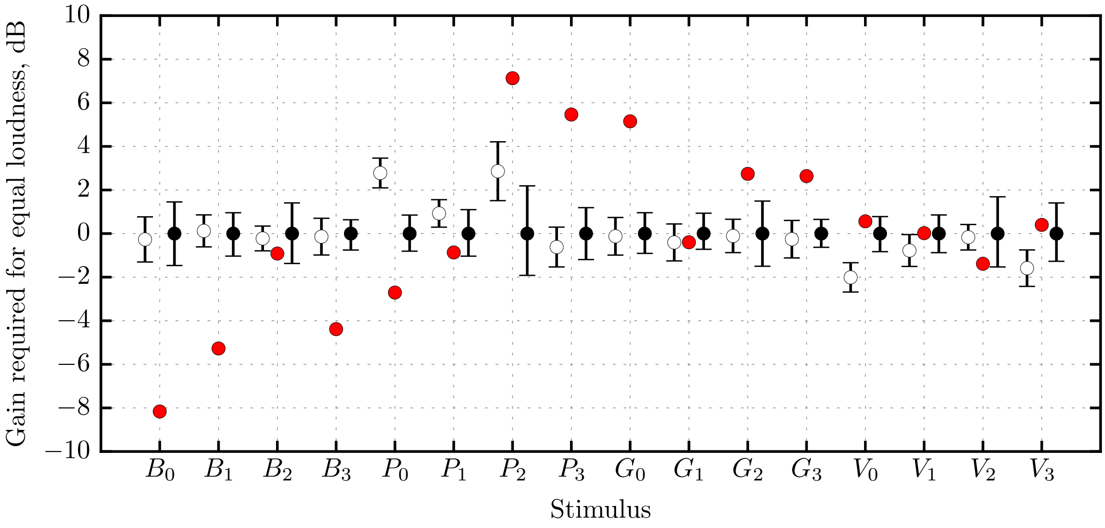


Figure 6.16: Open circles show the gains required for equal loudness estimated using a linear model of the matching data from the main experiment. Closed circles show the average equal-loudness gains derived from the pilot experiment, which represent the level used to loudness calibration the stimuli for the main experiment. For this reason they are centred on 0 dB. Error bars show the between-subject bootstrapped 95% CI of each mean. Red circles show the gain required for RMS level alignment.

using the percentile bootstrap method. As expected from Figure 6.15, the majority of segments appear to be equally loud as indicated by the strong overlap between the CIs. Segments P_0 and P_2 were notably softer than the other segments, therefore requiring a boost in level.

The restricted spread of mean gain (-3 to +4 dB according to the CIs) is a consequence of the initial loudness-balancing task which was used to calibrate the stimuli for the main experiment. This is demonstrated by the black circles which show the equal-loudness gains derived from the pilot experiment. For most sounds, there is a strong overlap between the intervals, indicating that both procedures led to similar relative loudness judgements.[14] Note, however, that the two means are not fully independent, as six of the nine subjects used for the pilot experiment also participated in the main assessment. Except for pairs involving segments P_0, P_2, V_0 and V_3, the relative loudness of all other pairs is similar when compared across the two experimental procedures. Although these findings are only indicative, given that the group estimates from both procedures do not deviate substantially from one another (considering their uncertainties), the simple and efficient loudness-matching procedure used for the pilot experiment is worthy of further investigation.

The red circles show the gains required to equate the RMS levels of the stimuli, with the headphone response taken into account. In general, there is a poor alignment of the gains, indicating that RMS normalisation is not appropriate for loudness aligning the types of sounds used in this study.

6.4 Discussion

The experiments presented in this chapter were designed to capture the relative loudness of sounds typically experienced by audio engineers and music producers. This class of sounds is commonly referred to as multitrack audio, examples of which include: drum elements, guitar riffs, basslines,

[14] The intervals obtained from the pilot experiment are in some cases wider than those estimated from the pilot experiment due to lower number of subjects and (most likely) a less controlled experiment design.

piano sequences, synthesised sounds and the singing voice. Most previous studies of loudness perception that have involved everyday sounds typically involve programme material such as mastered songs, isolated speech, and dialogue with background sounds. As noted by Skovenborg et al. (2004), when dealing with real-world sounds, there is a high degree of subjectiveness which can affect measurement accuracy beyond what may be reported in traditional psychoacoustic studies where abstract sounds are used to measure specific properties of auditory perception. It is therefore important to employ an experimental procedure that attempts to minimise such influences as best as possible.

Adaptive methods are based on probabilistic sampling and target a point on the psychometric function as efficiently as possible. For loudness matching, this threshold corresponds to the 50% point (the PSE). Such procedures are common in auditory research where specific hypotheses are tested using artificial sounds that have been synthesised to target the research question at hand. An adaptive paradigm was used in the first experiment conducted here to measure the level differences needed to equate the loudness of four musical sounds. When the duration of the experimental stimuli exceeds a few seconds, adaptive systems take a long time to converge, making the measurement tiresome for the listener, especially when matching-tracks are interleaved and/or the test corpus is vast. When the purpose of the listening assessment is to formulate a dataset for assessing the accuracy of computational hearing models, researchers often turn to the MOA for its simplicity and efficiency (Jones and Torick 1982; Moore et al. 2003; Soulodre et al. 2003; Skovenborg et al. 2004; Soulodre and Lavoi 2005; Pestana and Alvaro 2012; Ma et al. 2014).

In the second experiment presented here, the MOA was used to measure the relative loudness of a larger sample of sounds more efficiently, but constraints were incorporated to minimise biasses introduced by having the listener directly set the level. These rules were inspired by those used in more time-consuming indirect measurement procedures, and encourage the listener to bracket the point of equal loudness by setting the variable stimulus both softer and louder than the fixed sound. The average time to measure a single match involving 4 s sound segments was roughly 1 m, compared to 2.5 m for the two-interval forced-choice method used in experiment one. Skovenborg et al. (2004) used the direct MOA (with no constraints), and reported a median response time of 14 s, which partly explains why a larger test corpus could be accommodated in that assessment. In this study, the time to equate the loudness of two of sounds systematically decreased by roughly 4 s after completing each set of 40 matches once (on average). Unfortunately the experiment was designed such that all segment pairs were matched in the reverse adjustment order after completing those in the first adjustment order. Thus, any temporal effects such as fatigue, experience and boredom would have had a greater impact on the matches in the second adjustment order. A simpler experiment design without the segment-pair selection constraint (as used here) would have allowed for more sophisticating counterbalancing schemes.

In the first experiment involving four sounds, a positive bias effect of up to 1 dB was identified, suggesting that the listeners applied too much gain when matching the loudness of paired stimuli. Such effects have been observed in previous loudness-matching studies (Florentine et al. 1996; Moore et al. 2003) and might be caused by a tendency of the listeners to overestimate the loudness of the fixed stimulus (Marks and Florentine 2011). This phenomenon was not observed in the second experiment, although the uncertainty of the estimated bias was indicative of a similar sized effect. In the second experiment, a small bias effect was measured indicating that the listeners set the level of the variable sound slightly above or below the true matching level depending on whether the initial loudness was higher or lower, respectively, than that of the fixed sound. This finding was also reported by Soulodre et al. (2003), where listeners used the MOA to match the loudness of music and speech. They found a positive difference between the matching gains measured with a high and low initial offset applied to the level of the test stimulus. This highlights the importance of counterbalancing (or at least randomising) the initial level offset applied to the variable sound. It therefore seems that the bracketing constraints incorporated into the procedure used here were not sufficient to completely eliminate this effect. Note that Skovenborg et al. (2004) found an opposite (negative) bias effect to the one measured here, but they attributed this to the overall direction of the rotary knob used to set the level, rather than the initial presentation level, although the two are indeed closely related. Thus, the listeners in that assessment tended to overadjust the level of the variable sound, rather than underadjust. This opposite effect might be related to the fact that the subjects in that experiment manipulated loudness using a continuous rotary knob.

In the loudness-matching experiment of Soulodre et al. (2003), various items of broadcast audio content were matched in loudness to a sample of female speech. Within-subject variability was measured using the absolute difference between two replicated matches performed by the subjects. The only difference between the two conditions was the starting level of the variable stimulus (high or low). The mean difference was 1.2 dB, which contrasts with an average GMD of 2.8 dB measured in the first study conducted here, and 2 dB in the second. Note that these measures of within-subject consistency include both bias factors and noise. The first experiment conducted here comprised a smaller sample of subjects and only four sounds, which may explain the larger variance. Soulodre et al. (2003) reported between-subject variability using the SEs of the matches which ranged between 0.3 (min) and 0.6 dB (max). Given that 24 subjects were used, this corresponds to between-subject SDs of 1.5 dB to 2.9 dB. For this study, the average SD in experiments one and two was 2.5 dB. The maximum SD from experiment two was about 4 dB, and therefore higher than the maximum reported by Soulodre et al. (2003). The authors in that study identified that subjects were more consistent with one another when matching the loudness of samples involving the human voice, which agrees with the findings here in that voice-to-voice matches had the lowest inter-subject variability.

Skovenborg et al. (2004) conducted a balanced-pair loudness-matching assessment involved 145 segments of music and speech, using linear regression to estimate the optimal loudness levels. The median SD of same-pair adjustments was 1.7 dB, which contrasts with a median of 2.5 dB from experiment two of this study. Between-subject GMD statistics were also reported, with 50% of the pairs having GMD values less than 2.1 dB, and 95% less than 4.2 dB. These values are slightly lower than the GMD percentile statistics measured in experiment one of this work (median: 2.8 dB; 95% percentile: 4.4 dB). Skovenborg et al. (2004) found a large inter-subject disagreement for pairs involving a 1 kHz pure tone. When conducting the pilot loudness balancing task administered in preparation for experiment two of this project, subjects reported that they found the 1 kHz tone difficult to compare to other excerpts and that it influenced their perception when determining an overall comfortable listening level. Since no replicate matches were made by Skovenborg et al. (2004), within-subject variability was based on the RMSE from the subject-level regressions. These errors ranged from 1.5 dB to 1.9 dB. In this work, the mean RMSE in experiments one and two was 2.1 dB and 1.7 dB, respectively, both of which had a slightly wider error spread (1.4–3.4 dB and 1.2–2.3 dB, respectively). It is important to emphasise that the R^2 statistic depends on the variation in the matching gains. For example, the average R^2 in experiment one was 84% (SD = 10), compared to 73% (SD = 9) in experiment two. However, the stimuli in the first experiment were calibrated using RMS-level normalisation, but those in the second were set roughly equally loud prior to the main assessment. Thus, the matching gains in the first experiment were of a greater magnitude compared to those in the second experiment due to more pronounced inter-stimulus loudness differences at the reference levels.

Moore et al. (2003) investigated the effects of dynamic range compression on the loudness of speech using equal-loudness matching. The authors found that, when presented at the same RMS level, compressed speech was perceived louder than uncompressed speech, by up to 3 dB. Furthermore, this phenomenon was found to increase with SPL. They used the MOA with similar constraints as used in this study, and had listeners vary both the compressed and uncompressed segments. A negative bias effect of -0.4 dB was detected, indicating the listeners underestimated the loudness of the fixed stimulus. This contrasts with the small *positive* fixed-stimulus bias measured here. SEs of the matching data were in the region of 0.25 dB, giving a very low mean between-subject SD of 0.6 dB. The SDs displayed in Figure 6.12 are indeed small (around 1 dB) for the three voice-to-voice loudness matches.

Jones and Torick (1982) conducted a series of listening tests to inform the design of a loudness meter applicable to broadcasting. In one experiment, 30 subjects from a broad age group matched the loudness of 12 processed segments using the MOA and a noise stimulus as the fixed reference. Between-subject variation was measured using the interquartile range (IQR) of the adjustments. For the middle age group (24–64), the IQR was around 2.5 dB on average. A second experiment was then conducted in which 11 of the most consistent listeners from the first experiment matched the loudness of eight unprocessed segments. The IQRs in that experiment varied from 1 to 3.5 dB. In the study presented here, the average IQR of the different-segment pair matches in experiment one and two was 3.1 dB (SD = 1.3) and 3.1 dB (SD = 1.4), respectively.

In general, it can be concluded that there was slightly more intra- and inter-subject variability in the measurements made here compared to those reported in previous work. Since the methodologies are quite similar, and the participants all had an audio engineering background and no reported hearing impairment, the likely cause for these differences is the type of stimuli. A few researchers have measured the loudness of segments taken from multitrack recordings. Ma et al. (2014) performed a *partial-loudness* matching experiment in which 12 listeners were asked to match the loudness of a segment against itself when combined with 3–4 other segments. This was done for four songs using the MOA. Each segment was varied both in silence and in the mix to counteract minimise variable-stimulus bias. Indeed, the authors reported an effect in which subjects over-attenuated the solo stem when the mixed stem was fixed in level. Assuming that the error bars in their Figure 2 are standard deviations, the spread of the matching gains varied between 1 and 3 dB. These are higher than the average SD for the same-segment pairs reported in the first experiment here (0.8 dB), which is likely caused by the extra variance introduced by masking and the presence of other sounds. Pestana and Alvaro (2012) carried out a loudness-matching assessment involving five songs with 9–11 tracks each. Due to the large number of samples, tracks within each song were matched with reference to the bass drum and the vocal track (two fixed-references). Because the results were presented as errors with respect to the ITU-R BS.1770 (2012) loudness algorithm, and grouped by source type, it is difficult to draw quantitative comparisons with the results presented here. The author did comment that subjects found the task difficult, and reported a strong bias depending on which segment was used as the reference.

Fixed-reference designs are common in loudness perception studies, especially when it is necessary to have subjects match the loudness of a large test corpus. The disadvantage of these designs have been covered in great detail by Skovenborg (Skovenborg and Nielsen 2004a; Skovenborg et al. 2004). In experiment one of this study, the bass sample was judged most inconsistently by the subjects, which might be explained by strong differences in frequency content between the bass and the other higher-frequency instruments, but could also be attributable to low-frequency resonances of the headphone, making loudness judgement difficult. In the second experiment, listeners were least consistent for pairs involving the shaker and snare drum, where within-subject deviations between replicated matches were around 3 dB on average. The balanced pair-matching design helps to minimise such uncertainty by having all sounds serve as the reference, incorporating redundancy and employing a statistical model to estimate the best fitting segment levels.

The loudness-matching data measured in the first experiment was subjected to a mixed-effects regression model which showed that stimuli were not equally loud when presented at the same RMS level. Large gains of up to 8 dB were needed to correct for inter-stimulus loudness differences. In experiment two, the relative loudness levels estimated from the matching data were transformed to a set of equal-loudness gains which was then compared against those derived from a preliminary loudness balancing task in which subjects matched the loudness of all 16 sounds by using a standard mixer interface. The two gain vectors were very similar, but systematic differences of up to 2 dB were identified for a few sounds. In both cases, RMS-level normalisation was found to be inadequate for loudness alignment. This suggests that simple loudness metrics which have been found to perform well on broadcast material, such as the L_{eq}, may not be robust when applied to recordings of isolated musical instruments.

6.5 Summary

Loudness-matching experiments have been conducted using musical sounds as stimuli and audio engineers as listeners. In a first experiment, an adaptive loudness-matching procedure was used to measure the point of subjective equality between all pairwise combinations of a drum set, bass guitar, electric guitar and a male vocal. Both intra- and inter-subject variability were analysed, and a small bias effect of 0.7 dB pertaining to the fixed stimulus was identified. A linear model of the matching data was used to extract an equal-loudness gain balance, which defines the set of relative-gains required to loudness normalise the stimuli. The derived balance was markedly different from that of an equal RMS-level gain balance, suggesting that more sophisticated models of perception are needed to accurately measure the loudness of musical instruments.

A second experiment involving 16 musical sounds of the same instrumentation as used in the first study was conducted. For efficiency, the method of adjustment was used to measure the relative loudness of paired sounds according to a balanced pair-matching experiment design, the specifications of which were reformulated to allow for greater flexibility. In particular, redundancy was incorporated into the experiment by having listeners repeat loudness matches with each segment matched to a number of other segments and also having each segment serve as both the fixed and variable stimulus. It was found that both within- and between-subject variability depended on the paring of sounds, with subject agreement being highest when matching vocals to vocals. Linear modelling was used to derive the equal-loudness gain balance, in which a systematic effect relating to the initial presentation level of the variable stimulus was measured with an effect size of 0.2 dB. An additional effect was observed in which the duration of the matching task systematically decreased over time.

The listening assessment conducted in Chapter 5 suggested that sophisticated psychophysical loudness models grounded in auditory theory may not be suitable for multitrack material. Furthermore, models based on a time-varying level measurement were found to perform relatively well for the task of automatic music mixing. Considering that experiment 1 of this chapter showed RMS level to also be a poor predictor of loudness, it is unclear as to which metrics are most applicable for musical sounds. This problem is addressed in the next chapter where the predictions of several state-of-the-art loudness algorithms are compared against the equal-loudness gain balances obtained from four independent listening experiments, including those presented here.

Chapter 7

Evaluation of Loudness Models Applied to Musical Sounds

Multitrack audio describes a type of sound material commonly worked upon by mixing engineers in music production. In the majority of cases, the mixing session involves recordings of single instruments such as: electric, acoustic and bass guitar; singing voice (lead and backing vocals); piano, keyboard and synthesisers; and percussive instruments such as shakers, snare drums, cymbals and full drum kits. Few studies have assessed the predictive capabilities of loudness models when applied to such musical sounds. Instead, most previous work has concentrated on modelling the perceived loudness of commercial programme material, such as mastered music (complete songs), and speech with and without background sound effects. This chapter addresses this problem by evaluating the performance of state-of-the-art dynamic loudness models using musical content.

7.1 The subjective reference datasets

In the previous chapter, two equal-loudness-matching experiments were performed in which subjects matched the loudness of musical sounds classified by four instruments common in blues and rock genres: percussion, bass, guitar and vocals. In the first experiment, eight audio engineering students matched the loudness of all possible pairwise combinations of four samples taken from a multitrack rock song: full drum kit, bass guitar, electric guitar and lead vocals. This full experiment design was conducted in preparation of a second, larger experiment involving sixteen sounds (four sounds from each category). In this second experiment, a subset of the 120 total possible pairs were loudness matched by fourteen subjects. Forty unique segment-pairs were matched using a balanced experiment design; the absolute frequency of occurrence was the same for every segment and each segment was matched against five other segments, four of which belonged to a different stimulus category. Additional redundancy was incorporated by having both sounds in a pair varied in level, and all matches were repeated twice, such that within-subject consistency could be quantified. In both experiments, no single sound was selected as a reference for all matches. This was done deliberately to minimise the uncertainty or potential bias associated with a specific sound (Skovenborg and Nielsen 2004a). A set of gains for aligning the loudness of the stimuli (here termed 'equal-loudness gains') were derived from the empirical data by means of linear regression. These gains were derived for each listener of the two experiments.

The primary dataset used for evaluating the loudness models is here called $\mathrm{A}_{16,\,14}$, where the subscript 16 refers to the number of segments involved in the experiment, and the subscript 14 corresponds is the number of participants. Because model parameters were fitted to this dataset, as discussed later, it was decided to draw upon additional data to give some insight into the generalisability of the models. Therefore, the data from the preliminary experiment, denoted $\mathrm{B}_{4,\ 8}$, was also used, as well as two additional datasets which are described in the following sections. The musical sounds used in these final two experiments were not specifically drawn from the same instrument categories as used in the first two experiments, allowing for a broader performance evaluation. These experiments reflect the author's initial explorations of different experimental procedures suitable for equal-loudness matching of short audio excerpts.

Dataset $\mathbf{C}_{12,\ 13}$

Stimuli

13 subjects matched the loudness of 11 musical sounds to a reference sound. Of the 12 segments in total, the following 10 were extracted from the sound quality assessment material (SQAM) database (EBU Tech 3253 2008): claves (track 26), cymbal (track 31), acoustic guitar (track 58), double bass (track 11), flute (track 13), triangle roll (track 32), piano (track 60), vocal quartet (track 48), female soprano vocal (track 44) and bass vocal (track 47). Two additional segments—a bass drum and a bass guitar riff—were taken from a multitrack rock song. The 12 segments were selected to give a varied timbral range; representative excerpts between three and five seconds were taken from each recording.

The reference stimulus levels used to calibrate the main experiment were set by two individuals trained in critical listening and had previous loudness matching experience. These individuals balanced the stimuli for equal loudness whilst preserving a comfortable listening level. The average root mean square (RMS) and peak sound pressure level (SPL) at the eardrum was estimated to be 79.8 dB (standard deviation (SD) = 2.7) and 95.1 dB (SD = 4.2), respectively, based on the average sensitivity and frequency response of the BeyerDynamic DT990 headphones as measured on a GRAS 43AG ear and cheek simulator.

Subjects

Of the 13 participants, nine were graduates from the Sound Engineering and Production course at Birmingham City University, and two were third year undergraduate students on the same course. Of the remaining two subjects, one was a graduate from the Sound and Multimedia degree at the same university (the only female), and the other was an experienced music producer and PhD graduate from a music technology course at McGill University. 12 of the subjects were between 20 and 29 years of age and the oldest, most experienced subject was 38. Five participants had previous loudness matching experience.

Procedure

A fixed-reference loudness-matching experiment was conducted using the method of adjustment (MOA). A 4.8 s segment from a recording of a male singing voice (track 47 from the SQAM database) was selected as the reference sound because the loudness of this segment was judged, by the author, to be relatively steady throughout its duration and thus give a robust overall loudness impression. As discussed previously, the balanced-pair-matching design, in which no single reference is chosen as the reference sound, would have been preferable, but a fixed-reference design was chosen for efficiency in this experiment. BeyerDynamic DT990 headphones, as used in the authors other assessments, were used for reproduction. The RMS level of the reference stimulus at the eardrum was approximately 80 dB SPL.

The remaining 11 segments were paired with the reference. On a given trial, a stimulus pair was selected at random and the subject could trigger either the first or second sound. The level of the former was fixed at 0 dB (with respect to the calibration level) and the latter was made variable after offsetting the gain by a random value within a range of ± 10 dB. The listener was instructed to make the variable sound as loud as the fixed sound by using the up and down arrow keys to increase or decrease the level of the variable sound, respectively. Prior to submitting the final matching level, the listener was told to bracket the point of equal loudness. A submission constraint was built into the system to force the listener to do this; the level increment was 3 dB for four reversals, after which it was reduced to 0.5 dB. When happy with their adjustment, the subject clicked the on-screen 'Accept' button, which triggered a sequence in which the two sounds, separated by a short interval, were played in full duration. The subject could then either submit the matching level and load the next segment pair or readjust the level of the variable sound for the current trial. A readjustment would always require the listener to re-accept the current match and audition the confirmation sequence. Finally, equal-loudness matches were obtained with both segments in each pair varied in level, the choice of which was also selected at random. The 11 segment pairs were matched in both presentation orders giving a total of 22 matches per subject. Experiment durations were between 15 and 40 minutes. Six of the subjects were available for re-testing and so repeated measurements were made based on the consistency of the first round of matches.

Results

A conservative approach to data removal was taken and, although some subjects were very inconsistent on particular segment-pairs, their average estimate was similar to the group mean so these cases were not removed. A total of four data points were discounted based on the matches submitted by three subjects that were inconsistent with repeated matches. Within-subject consistency was measured using the Ginis mean difference (GMD), accounting for differences in sign when the roles of the stimuli reversed. Note that some participants submitted multiple adjustments for each segment pair with both segments varied. Within-subject consistency ranged between 1 and 3.3 dB, with a mean GMD of 2.3 dB (SD = 0.6) and median of 2.2 dB. Between-subject variation was measured using the GMD and SD of each segment-pair after averaging the gains submitted by each subject. The average GMD across the 11 pairs was 2.6 dB (SD = 0.9) and the average SD was 2.3 dB (SD = 0.8 dB). Subject agreement was best for two vocal excerpts (female soprano GMD = 1.1 dB, vocal quartet GMD = 1.5 dB), which was expected because the reference segment was also a vocal with identical lyrics and musical phrasing. There was considerable disagreement as to the loudness of the bass guitar (GMD = 4.2 dB), followed by the claves recording (GMD=3.6 dB).

A series of single subject-level regressions were then used to summarise the matching data (refer to Section 6.2.2). The average intercept across subjects was -0.62 dB (95% CI = [-0.9, -0.4]), meaning that the subjects were, on average, setting the level of the variable stimulus below the true matching level. This suggests that the listeners perceived a stimulus as being softer when it was fixed in level. Unfortunately, the effect of the random level offset could not be estimated for all subjects due to the design matrix being singular. However, this bias is expected to be relatively small, especially when averaging across subjects. The relative-loudness estimates obtained from the regressions were transformed to a set of gains for loudness normalisation as shown in Figure 7.1. The gains have a narrow spread, which is expected given that the stimuli were first loudness aligned by two listeners, prior to the main assessment. It can be seen that the CIs are notably large (>1 dB)

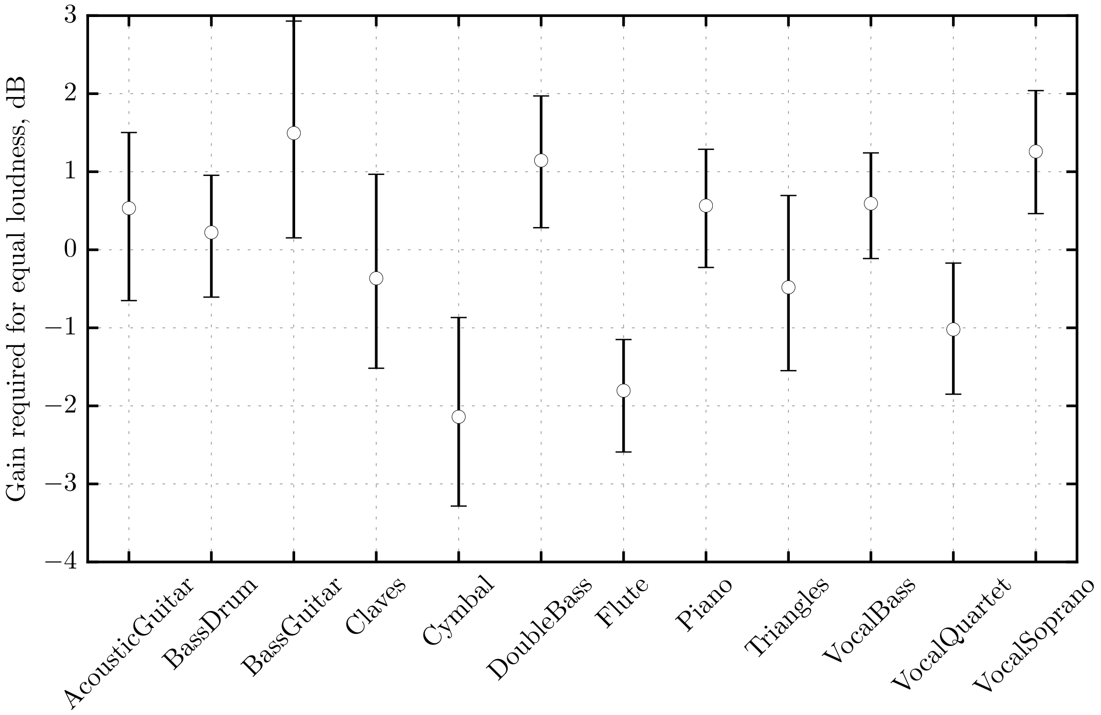


Figure 7.1: Gains required for equal loudness as derived from the fixed-reference loudness matching experiment. Error bars show the 95% between-subject CI of each mean (white circles).

for the bass guitar, claves and cymbal, which is attributed to the greater variability in the matches associated with these segments.

Dataset $D_{10,4}$

Six listeners performed a loudness matching task involving the following ten sounds: accordion, acoustic guitar, cello, clarinet, piano, bass guitar, electric guitar, drum kit (bass, snare and hi-hat), female country vocal and male rap vocal. The reference stimulus levels were left as the peak-normalised audio derived from the original full-length recordings and the overall listening level was set such that all segments were presented at a comfortable loudness. The average RMS level was 79.0 dB SPL (SD = 4.2 dB) and the average peak level was 94.8 dB SPL (SD = 1.6), after accounting for the sound transmission through headphones. Initially, the experiment was configured as a fixed-reference design, whereby participants would loudness match every segment with respect to a single reference segment. However, a few subjects were available for further testing over a weeks period, which allowed for a greater collection of loudness matches. For example, two listeners were able to complete every possible pairwise comparison (90 matches), whilst a third matched all segments using both the accordion and acoustic guitar as reference. Thus, the degree of within-subject consistency was evened out to a greater extent for those participants who were 'sampled' more frequently. All subjects completed the original fixed-reference experiment design and matches were obtained with the level of every sound adjusted.

Loudness matches were obtained using a two-interval forced-choice paradigm. On a given trial, the listener heard two sounds separated by a 600 ms inter-stimulus gap. After hearing both sounds, the subject responded as to which of the two sounds was the loudest. If the sound corresponding to the variable stimulus was selected, its level was reduced, and increased otherwise. An initial step size of 5 dB was used for the first two reversals, after which it was reduced to 2 dB. The final matching level (level of the variable at the point of equal loudness) was taken as the average of the levels at the final four reversal points. Two loudness tracks were interleaved for every pair of sounds, such that both sounds served as the variable stimulus. An interleaved track was selected at random, as was stimulus presentation order. The order of these tracks was different for each subject.

After completing the experiment, a set of loudness normalisation gains were obtained by fitting a linear model to each individual's data. The root mean square errors (RMSEs) were markedly large (>4 dB) for two of the subjects, and their R^2 values were around 55% lower than the remaining four subjects. This reflects both poor internal consistency and large within-subject consistency. For this reason only four of the most consistent listeners' data were used, where the model RMSEs ranged from 1.5 dB to 2.6 dB. The mean within-subject consistency, as measured using the GMD metric on the raw matches submitted by these four subjects, was 1.4, 1.4, 2.1 and 2.5 dB. Although this is a small sample size, the equal-loudness gains for these four subjects were in high agreement with one another (RMSE between each subject's gains and the group medians $= \{0.7, 0.8, 1.1, 1.4\}$, correlation with median $= \{0.90, 0.90, 0.96, 0.98\}$. For the purpose of model evaluation, this dataset was treated with caution and therefore assessed differently to datasets A-C, as outlined in Section 7.4.4.

7.2 The loudness models

Five published algorithms were evaluated: the model of Glasberg and Moore (2002), the model of Chalupper and Fastl (2002), the model of Chen and Hu (2012), LARM (Skovenborg and Nielsen 2004b), and the EBU R 128 meter (EBU R 128 2014; EBU Tech 3341 2016). In this study, these models are referred to as GM02, CF02, CH12, LARM and EBU, respectively. The first three are multiband excitation pattern (EP) based loudness models whereas the latter are single-band loudness algorithms (for a review see Chapter 3). Each model takes a time-domain signal as input and outputs a loudness-time series. In this work, the time series output by the three multiband models and the LARM algorithm has been termed the short-term loudness (STL). The GM02 and CH12 also provide a second series based on an integration over a longer time period: the long-term loudness (LTL). In this assessment, the release time used to compute the LTL of the GM02 was reduced from 2 s to 0.2 s, following the revision proposed by Moore et al. (2003). The EBU model, founded on the ITU-R BS.1770 (2015) algorithm, provides three loudness-time functions: the momentary, short-term and integrated loudness. In this study, the momentary loudness (ML) was used because it captures the most dynamic properties of the signal. Furthermore, the EBU integrated loudness is simply a gated average of the ML. When measured over the entire duration of the audio, the final reading is called the *programme loudness* (EBU R 128 2014) and is identical to the ITU-R BS.1770 (2015) measure. An overview of the loudness models evaluated in this study, including their associated global-loudness descriptors (see Section 7.2.1) as per the original publication, is given in Table 7.1. The final model (LML) shown in Table 7.1 is a new loudness estimation algorithm proposed in Section 7.2.2.

Model	Structure	Output	Global-loudness descriptor
GM02	1. Outer/Middle ear 2. Multi-resolution spectrogram 3. Excitation transformation (Cams) 4. Specific loudness transformation 5. Spectral summation 6. Temporal integration	STL LTL Unit: Sones / Phons Time step: 1 ms	Peak STL Mean LTL
CF02	1. Highpass filter 2. Excitation transformation (Barks) 3. Outer/Middle ear 4. Loudness calculation 5. forward masking 6. Spectral masking 7. Spectral summation 8. Temporal integration	STL Unit: Sones / Phons Time step: 2 ms	Peak STL 95th percentile of STL
CH12	1. Multi-resolution spectrogram 2. Outer/Middle ear 3. Excitation transformation (Cams) 4. Spectral summation 5. Temporal integration	STL LTL Unit: Sone / Phon Time step: 1 ms	Peak STL Mean LTL
LARM	1. revised low-frequency B curve (RLB) highpass filter 2. Amplitude envelope extraction 3. Moving power mean using a 0.5 window	STL Unit: dB Time step: 1 ms	Envelope power mean
EBU	1. K-filter 2. Mean power in 400 ms overlapping blocks	ML Unit: LU FS Time step: 1 ms	Programme loudness
LML	1. Multi-resolution spectrogram 2. Outer/Middle ear 3. EP (Cams) 4. Spectral summation 5. Temporal integration	STL Unit: dB Time step: 2 ms	Percentile of STL

Table 7.1: Overview of the loudness models evaluated.

7.2.1 Global-loudness descriptor

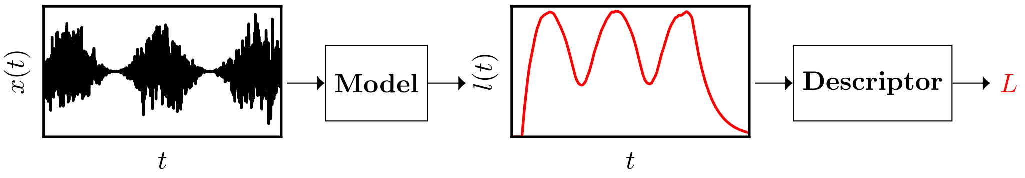


Figure 7.2: Flow chart of the loudness extraction process. $x(t)$ is the stimulus waveform, $l(t)$ is the loudness-time series, and L is the overall loudness impression (scalar).

In order to quantify the overall loudness impression of a stimulus, developers generally propose a simple statistic to apply to the loudness-time function, as depicted in Figure 7.2. For example, Glasberg and Moore (2002) and Moore et al. (2003) suggested the peak STL for very short duration sounds, and the mean LTL for fluctuating sounds such as noise, amplitude modulated tones, and speech. Each of these measures generate a single loudness value that describes the global loudness of the sound in question.

Note that the LARM algorithm was originally proposed for global-loudness estimation only, rather than outputting a time series (Skovenborg and Nielsen 2004b). Global loudness is calculated by taking the power mean of the amplitude envelope. The structure outlined in Table 7.1 represents an adapted implementation, based on the loudness meter prototype suggested by Skovenborg and Nielsen (2007), such that a temporal profile could be extracted. Although exact details of this meter were not specified, a good visual match to the STL shown in Figure 7 of that paper was obtained by setting the release time constant of the asymmetrical lowpass filter to 1.5 s and employing a 0.5 s analysis window to compute a running power mean.[1] Thus, in addition to the original measure of global loudness, the STL obtained using a sliding rectangular window was also used in the evaluation.

7.2.2 A linear multiband loudness model

Existing multiband loudness models are generally complex in design and produce estimates of perceived loudness in units of sones. This means that optimisation routines are required to establish a map between the loudness domain (sones) to the level domain (dB) (see Section 7.3). In addition, these models have been developed to predict empirical data derived from psychophysical studies involving abstract sounds such as pure tones and noise. In contrast, computationally cheaper single-band models were developed to predict the relative loudness of music and speech presented at moderate listening levels. However, these latter models cannot account for the dependence of loudness on spectral bandwidth, nor do they facilitate the modelling of time-frequency masking. Furthermore, multiband models are more attractive for multi-source applications because they offer auditory representations of sounds from which inter-source masking measurements can be made.

It was decided to develop a new model that amalgamates design principles from both approaches in order to better understand which algorithmic features are important, or indeed problematic, and if a combined approach is superior for practical loudness estimation. For example, can the frequency-weighted energy measurement approach be improved upon by incorporating a simple model of spectral loudness summation (SLS)? If so, how does the performance (and parameters) of this algorithm compare to existing EP approaches?

The new LML model is similar in architecture to the GM02 but incorporates the RLB filter, as

[1] The authors suggested window lengths of 0.5 s and 2.5 s for the STL and LTL, respectively.

used by the EBU and LARM algorithms, and is linear with respect to listening level. This latter aspect helped to simplify the fitting procedure, but also has the advantage of making the LML practical to work with by engineers. Furthermore, the purpose of this work was not to establish a complete psychoacoustic model that accounts for a broad range of empirical data, but rather to gain deeper insight into the practical effects introduced by key parameters when measuring the loudness of real-world sounds. The RLB filter is a simplified transfer function to account for the reduced sensitivity of the ear at low frequencies, although other transfer functions could be used. A rounded exponential (roex) auditory filter bank followed by a compressive function is used to model SLS. Finally, two asymmetrical lowpass filters are employed to smooth the instantaneous loudness (IL), generating the STL and LTL. The outline of the LML algorithm is as follows:

1. Three parallel short-time Fourier transforms (STFTs) using Hann windows of length 64 ms, 32 ms and 16 ms for frequency bands 10–935 Hz, 935–2095 Hz and 2095–16 000 Hz, respectively. The fast Fourier transform (FFT) buffer size for all segments is based on the largest window size (next power of two). The STFT time step is 2 ms.
2. Compressed power spectrum according to a criterion of 0.3 Cams.
3. RLB frequency weighting.
4. Outer-ear filter, e.g. free field or headphones.
5. Bank of roex filters using an 1 Cam increment. The roex filters are fixed in shape and are made asymmetrical by assuming an input level of 70 dB/equivalent rectangular bandwidth (ERB).
6. Intensity compression per critical band determined by the exponent α.
7. Summation of specific loudness (SL).
8. Temporal integration of IL to obtain the STL using an asymmetrical lowpass filter (an exponential moving average (EMA) filter with separate attack τ_a and release τ_r time constants). A second EMA is used to measure the LTL.
9. Compute global loudness by taking a percentile of the STL and converting to decibels.

The configuration of the multi-resolution STFT was based on three Hann windows, with the largest fixed at 64 ms, and finding the frequency band limits such that the main lobe bandwidth in each frequency band was less than the ERBs of the auditory filters at those frequencies. Note that this configuration gives finer spectral resolution than that used by Glasberg and Moore (2002), which might be attributed to the use of the equivalent noise bandwidth to determine the band edges, rather than the bandwidth of the main lobe.[2] However, Chen and Hu (2012) found the spectral broadening introduced by the multi-resolution STFT proposed by Glasberg and Moore resulted in SLS for pure tones (refer to Section 3.3.4). Therefore, greater frequency resolution was used in the algorithm presented here.

The outer-ear filter was bypassed for the assessment carried out in this chapter because the stimuli were all filtered according to the transfer function of the headphones used in the listening experiments (see Section 7.2.3). Following Glasberg and Moore (2002), the attack and release time of the STL integrator was set to 22 ms and 50 ms, respectively. The same parameters of the LTL integrator were set to 0.1 s and 0.2 s (Moore et al. 2003). Two parameters of the LML were

[2] Glasberg and Moore (2002) did not specify how they arrived at their specification for the multi-resolution STFT.

then allowed to vary: the amount of critical-band compression α, and the percentile P used to measure the overall loudness impression from the STL. The values of the these two parameters were obtained by fitting the model to subjective dataset A, using Scipy's `minimize` function (Jones et al. 2001), which was configured to Sequential Least Squares Programming (Kraft 1988). The fitting procedure involved minimising the variance of the loudness estimates (in dB) when all stimuli were loudness aligned according to the average normalisation gains as measured experimentally. Appropriate constraints were set on the parameters and the optimisation routine was run a number of times with different initial guesses. The best fitting values were $\alpha = 0.56$ and $P = 97$, giving an RMSE of 1.70 dB. This compressive exponent equates to a reduction of 1.3 dB per 3 dB increase in critical-band intensity. Had $\alpha = 1$, the complexity of the LML would have been completely unnecessary. Interestingly, the best case RMSE when fitted the same model with α constrained to 1, was 2.4 dB (with $P = 96$), indicating that spectral processing improved the overall fit by 1.1 dB.

7.2.3 Model configuration

For a given dataset, the stimuli were calibrated using the same reference levels as used in the listening assessment. Subsequent alterations were then made according to the type of loudness model:

- Multiband models: These models require a calibration gain to map normalised floating-point values to pressure values in Pascals, which requires knowledge of the presentation sound pressure level. The sensitivity of the BeyerDynamic DT990 headphones used in the experiments was measured using a GRAS ear and cheek simulator (IEC 60711 ear simulator, KB0066 Kemar large ears and 40AO microphone). The sound level in the left and right capsule was measured as 98.4 and 99.0 dB SPL, respectively, in response to a 1 kHz 1V RMS pure tone. The RMS voltage of a full-scale 1 kHz tone was measured throughout the listening experiments and the average value was used to calculate the calibration gain based on the average sensitivity of the headphones (98.7 dB SPL at 1 kHz, 1V RMS). For example, if a 1 kHz tone (RMS $\approx$ -3 dB full scale (FS)) produced an output voltage of 2 V at the headphone terminals, the RMS presentation level at the eardrum was estimated to be 104.7 dB SPL. It was determined that a full-scale 1 kHz tone produced an RMS level of 110.2 dB SPL.
- Single-band models: These algorithms generally operate independently of absolute sound level. Note that the EBU programme loudness *is* dependent on digital level (LU FS) due to first-stage absolute gating (see Section 3.2.3). It was therefore decided to offset the stimuli (after applying the initial reference gains) such that the highest peak across the stimuli was 0 dB FS.

After setting the stimulus levels for both the mutliband and single-band algorithms, the stimuli were filtered using a 4096$^{\text{th}}$ order finite impulse response (FIR) filter ($f_s = 44100$), simulating the transmission through the BeyerDynamic DT990 headphones as measured on a GRAS ear and cheek simulator. The RMSE between the magnitude response of the digital filter and the target headphone values was 0.02 dB for frequencies between 20 Hz and 20 kHz. The maximum discrepancy was 0.86 dB. The FIR filter was normalised to have unity gain at 1 kHz. Figure 7.3 shows the transfer function of the headphones (measured) and FIR filter (modelled).

The default outer-ear filter was bypassed for the GM02 and CH12 and the middle-ear transfer functions were left active. The CF02 uses an initial highpass filter followed by a frequency-domain weighting function to model the transmission response of the outer and middle ear (Chalupper and Fastl 2002). The weighting function modelling the sound transmission from the free field

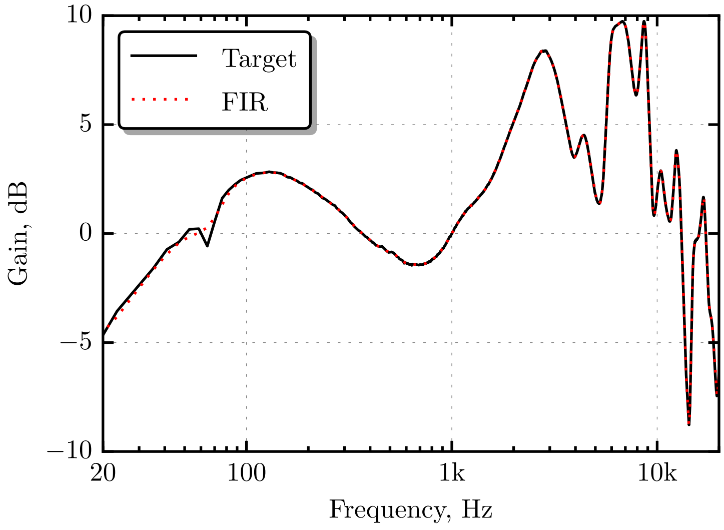


Figure 7.3: Average magnitude response of the BeyerDynamic DT990 headphones (target) and modelled response using a 4096th order FIR filter.

to the cochlea was removed under the assumption that the main influence of the middle ear is accounted for by the highpass filter. This approach was also adopted by GENESIS Acoustics for their Matlab implementation of Zwicker's model (on which the CF02 is based), in order to accommodate dummy-head microphones (Genesis Acoustics 2010). Finally, the RLB-weighting filter was left active for the EBU, LARM and LML algorithms. The EBU meter employs the K-weighting filter, which comprises a high-frequency shelving filter and a second-stage highpass filter (the RLB) (ITU-R BS.1770 2015). However, the former was removed because the purpose is to model the 'acoustic effects of the head' (ITU-R BS.1770 2015, p.3), which does not apply here.

Global-loudness descriptors

Each loudness model outputs at least one loudness-time function. This function then needs to be transformed to a scalar value describing the overall loudness of the input stimulus, i.e. establish an appropriate estimate of global loudness (see Figure 7.2). This transformation takes the form of a statistic applied to the loudness values extracted from the entire stimulus. For example, Moore et al. (2003) used the mean LTL output by the GM02 for predicting the overall loudness of speech. In contrast, Zwicker suggested taking the peak (Zwicker 1977) or 95th percentile of the STL (Fastl and Zwicker 2007). More recently, Terrell (2013, Chapter 7) found that, for the GM02, a linear combination of peak STL and mean LTL worked best for musical sounds when the weights were stimulus dependent. The general idea is that peak STL is most appropriate for very transient signals and mean LTL is better suited for sounds with a more steady dynamic profile. Terrell called this global-loudness descriptor the dynamic sound-stream bias (DSSB).

In this study, two approaches to measuring the global loudness of a sound segment were used for each model: the reference descriptor proposed by the model developer(s), and the percentile descriptor. The percentile descriptor, denoted P_X, is the loudness value at which X percent of values fall below throughout the time course of the stimulus. For example, P_{100} denotes the 100th percentile of the loudness-time function, which is equal to the peak value. The percentile was chosen for each model using an optimisation procedure that minimised the variance of the objective loudness levels when all stimuli were loudness aligned according to the subjective equal-loudness gains derived from dataset A.

The mean LTL was used as the reference descriptor the GM02, CH12 and LML, and the percentile descriptor was based on the more dynamic STL-time series. The suggested summary statistic for the CF02 is the peak loudness (Rennies et al. 2010), but it was also decided to include the mean STL as a measure of overall loudness.[3] For the EBU model, programme loudness was used as the reference statistic, in accordance with ITU-R BS.1770 (2015). The percentile descriptor was based on the ML for this model. The reference measure of global loudness for LARM was obtained by calculating the power mean of the stimulus amplitude envelope with an exponent of 1.5 (Skovenborg and Nielsen 2004b). Percentile loudness was calculated directly from the STL extracted using the adapted version of LARM (see Section 7.2.1). To summarise, the global loudness of a sound segment was measured using two types of summary statistic for each model: a variant of the mean, denoted μ, and a percentile measure, denoted P_X.

Baseline loudness metric

To give some insight into the practical advantages of the employing any one of the aforementioned loudness models, a baseline metric was established. Specifically, the RMS level measurement was used to represent the most simple device to measure relative loudness. This corresponds to the equivalent continuous sound level ($\mathrm{L_{eq}}$) measurement when the averaging time is set to the duration of the signal (see Equation 3.8).

7.3 Evaluation procedure

Equal-loudness matching data are generally represented by a set of relative signal gains or RMS level differences associated with paired stimuli. Each measurement defines the level adjustment needed to match the loudness of two sounds. A loudness model can therefore be evaluated by simulating the loudness matching experiment and comparing the objective adjustments with the subjective adjustments. For single-band algorithms, the level adjustment associated with a pair of sounds can be predicted directly by taking the difference between the two loudness readings. In contrast, most multiband models calculate loudness in sones, which unfortunately makes the prediction of matching data more involved. This is because the sone scale is not proportional to sound intensity (nor intensity level). Consequently, iterative methods are often necessary to find level adjustment needed for equal loudness. For example, Moore et al. (2003) incremented the gain of the variable stimulus in steps of 1 dB until the meter reading (in sones) bracketed the loudness of the fixed stimulus and then used interpolation to estimate the matching gain.

In this assessment, all matches from a given experiment were used to obtain a set of N segment relative gains that can be used to loudness normalise the stimuli. Thus, rather than deal with paired adjustments, the subjective gains can be compared with a model's predicted gains. For single-band models, the predicted equal-loudness gains g_m was obtained by transforming the vector of loudness levels output by the device l_m:

$$g_m = \bar{l}_m - l_m, \tag{7.1}$$

where $\bar{l}_m$ is the mean loudness level in dB.[4] This normalisation strategy is essentially the same as the one used by Skovenborg et al. (2004) when evaluating the performance of different loudness models.

[3] The CF02 does not generate multiple time functions.

[4] Here it is expected that the stimuli are input at the same reference levels as used in the actual listening assessment.

To simplify the assessment of multiband models, some researchers, e.g. Skovenborg and Nielsen (2004b), employ the single-band estimation approach (Equation 7.1) by transforming the absolute loudness readings in sones into loudness levels in phons. In essence, a linear relationship between sound level and loudness level is assumed. Although this approach is certainly convenient, the compression performed in the SL domain, which defines the slope of (log) loudness versus signal level, is both stimulus and level dependent. Appendix B evaluates the error introduced when predicting level differences using the sone-to-phon transformation with both the GM02 and CH12 loudness models. The results show that predicting level changes directly from the loudness levels is only practical for small increments in level (< 1 dB). To improve this method, the following procedure is proposed:

1. For each segment, compute the overall loudness (in sones) using an appropriate global-loudness descriptor, e.g. mean or peak loudness.
2. Convert the loudness values to decibels.
3. Calculate the average level and set as the target.
4. For each segment, iterate the gain such that measured loudness level is within a given tolerance, e.g. 0.01 dB, of the target level obtained from step 3.
5. Centre the levels on zero to obtain g_m.

When evaluating the nonlinear models on dataset C, the target level in step 3 was replaced with the level of the reference stimulus (male vocal) to simulate the structure of the fixed-reference loudness matching experiment. However, both approaches were tested and the differences in error, across all metrics, were negligible, e.g. the RMSE differed by less than 0.1 dB.

7.3.1 Testing the procedure

The original matching data used to establish dataset A was formed of 40 distinct loudness matches describing the relative loudness of 16 musical sounds. Using least squares regression, 16 relative loudness levels were estimated allowing the loudness of every sound to be compared against the loudness of every other. In order to apply the statistical model to the matching data, an assumption of local linearity in the matching data was made. Goodness-of-fit statistics and residual plots were used to judge the appropriateness of this model. Similarly, it is necessary to validate the proposed algorithm for the nonlinear loudness models. This was done by using the GM02 to simulate the complete experiment design for the 16 segments making up dataset A, and then estimating the *simulated* level adjustments via three different approximations. The full-experiment design consisted of 120 distinct pairs, with every segmented varied in level, giving a total of 240 pairs. For each stimulus pair, the gain of the variable stimulus at the point of equal loudness was estimated using the iterative procedure described in Section 3.3.6.

Two simulations were run, one using the mean LTL as a measure of overall loudness, and the other using the peak STL. Linear regression was then applied to the simulated matching data (240 matches) in order to estimate the relative loudness levels from which the data could be predicted under linearity, i.e. calculating all pairwise differences of the 16 levels to predict the 240 gain adjustments. The second approach to estimating the 16 relative levels was using the iterative method outlined in the previous section. The third method was based on the aforementioned sone-to-phon approximation. Note that the latter two techniques do not require the full 240 matches to be computed.

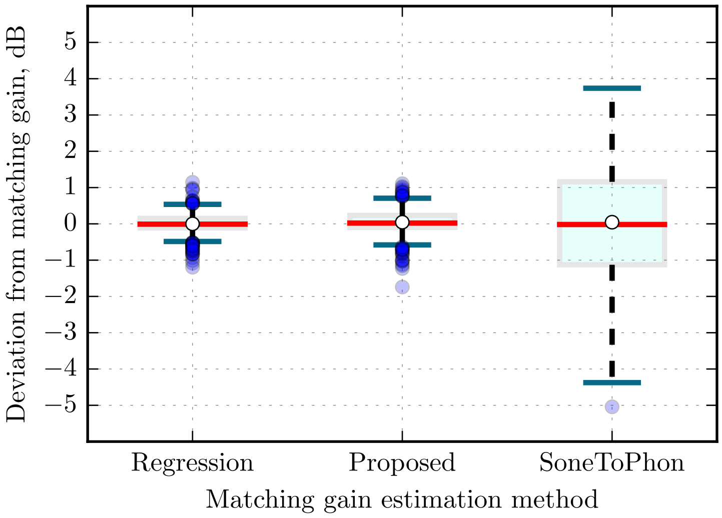


Figure 7.4: Distribution of pair-matching prediction errors for the three methods. The 240 gains obtained using each global-loudness descriptor are treated as a single dataset, resulting in 480 data points per boxplot.

Figure 7.4 shows the distribution of errors between the matching data obtained from the simulation and those predicted using the 16 levels obtained using either linear regression, the proposed iterative process and the sone-to-phon approximation (from left to right). The data for mean LTL and peak STL were combined for each boxplot. It can be seen that for the first two methods, the majority of the predicted gains are within 1 dB of the target gains. In contrast, the sone-to-phon approach is considerable less accurate, with discrepancies as large as 5 dB. This method assumes that the loudness growth functions of the stimuli are consistent with that of the 1 kHz pure tone.

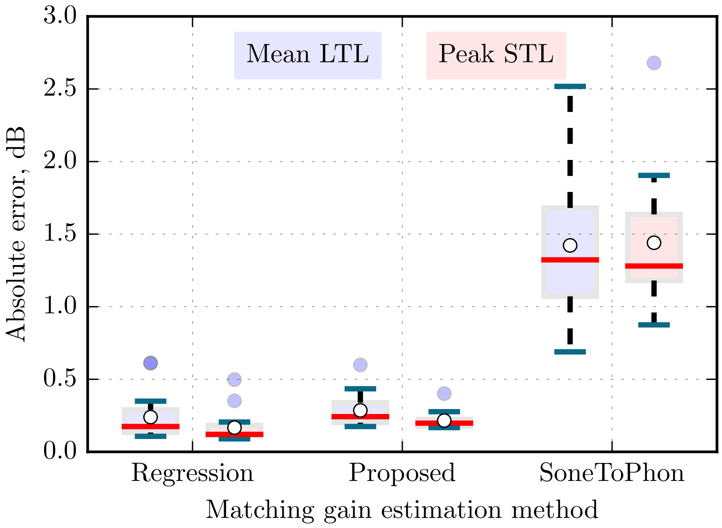


Figure 7.5: Distribution of mean absolute stimulus errors for three methods, according to mean LTL or peak STL. Ground truth data corresponds to the gains obtained from the simulated experiment involving all pairs of stimuli matched in both presentation orders.

For each method and feature, the 240 absolute errors were averaged by variable stimulus, giving 16 data points; these data are shown in Figure 7.5. The blue boxplots show the distributions of these averaged errors when using the mean LTL to determine overall loudness and the pink boxplots show those for the peak STL. Predicting the simulated data using the sone-to-phon transformation yields the largest overall error (median = 1.4 dB, interquartile range (IQR) = 0.3). Predicting the matching data using the parameters from the regression model resulted in an overall median stimulus error of 0.1 dB (IQR = 0.1), reflecting the nonlinear behaviour of the model. Similarly, the proposed iterative procedure yields an average error of 0.2 dB (IQR = 0.1).

To formally compare the distributions, an analysis of variance (ANOVA) was performed on the log transformed absolute errors to correct for normality, with stimulus as the random factor, and method and descriptor as the independent variables. The ANOVA showed a statistically significant effect of method ($F(2, 75) = 314.2$, $p < 0.0001$) and descriptor ($F(1, 75) = 13.3$, $p = 0.0005$) and a statistically significant interaction ($F(2, 75) = 4.6$, $p = 0.01$). Comparing the average (log) absolute error across methods showed that the sone-to-phon approach performed reliably worse compared to the other two methods for both LTL and STL descriptors ($p < 0.0001$, Tukey corrected for multiple comparisons). The proposed iterative approach produced an average (log) error that was reliably higher than the estimates generated by the regression method for peak STL only ($p = 0.0001$). Note, however that errors are all very small for these two approaches: the geometric means were 0.15 dB and 0.21 dB for the regression and proposed technique (averaged over both descriptors), respectively. Thus, at least for this set of stimuli, both methods are comparable in terms error magnitude. Note, however, that the iterative procedure significantly reduces the computational cost of estimating the levels compared to the full-pair method: the complexity of the iterative method is N compared to N^2 associated with the process of simulating the full-experiment design. The latter also requires a statistical model to extract the best fitting levels.

It is important to note that the spread of gains obtained from the simulated experiment were much larger than those obtained in the actual subjective experiment. For example, the largest matching gain from the simulated experiment was 28.2 dB when using mean LTL and 19.5 dB when using peak STL, compared to a maximum of 4–5 dB as measured experimentally. If the predictions of the model were more in line with the subjective values, one would expect lower error in the linear approximations because the nonlinearities associated with level dependency are less extreme. In what follows, model predictions were made using the proposed iterative procedure.

7.3.2 The error metrics

The equal-loudness gains predicted by each model were compared against the subjective gains. Thus, for any experiment design involving N segments (including the reference segment in the case of a fixed-reference design), there are N relative gains (in dB). Let g_i denote the mean subjective gain of segment i, and let $\hat{g}_i$ be the prediction of the model. The error between the predicted and subjective gain is then

$$e_i = \hat{g}_i - g_i. \tag{7.2}$$

To account for the uncertainty in the subjective data, the epsilon-insensitive error was taken (see Baykaner (2014) and Francombe et al. (2015)):

$$e_i^* = \min\left(0, \hat{g}_i - \mathrm{CI}_{\mathrm{L},i}\right) + \max\left(0, \hat{g}_i - \mathrm{CI}_{\mathrm{U},i}\right) \tag{7.3}$$

where $\mathrm{CI}_{\mathrm{L},i}$ and $\mathrm{CI}_{\mathrm{U},i}$ is the lower and upper boundary of the 95% between-subject confidence interval corresponding to the gain of segment i, respectively. This measure describes the additional error after taking into account the sampling noise associated with the gains. In effect, model predictions that fall inside the CI are not classified as erroneous ($e^* = 0$) but, instead, perfectly valid estimates given the variation in the data. In this study, CIs for the segment gains were estimated using the percentile bootstrap method (see Appendix C).

To quantify the error over all segments, the RMSE was calculated:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2}. \tag{7.4}$$

The epsilon-insensitive root mean square error (RMSE*) denotes the same statistic applied to epsilon-insensitive errors e^*. Worst-case prediction error was quantified using the maximum absolute value of these two errors, denoted $\max(e)$ and $\max(e^*)$, respectively. In addition to errors based on euclidean distance, the correlation between the subjective and predicted equal-loudness gains was calculated using Pearson's correlation coefficient r. This measure quantifies the strength of linear relationship between two variables and ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation); a value of zero indicates no correlation.

Francombe et al. (2015) proposed the outlier ratio as a summary of the number of predictions falling outside of the uncertainty of the subjective values, relative to the total number of predictions. Whilst perfectly valid, imagine data with very little uncertainty, perhaps due to a very large sample size. As the sample size increases, the outlier ratio approaches one for a non-perfect model, even if the between-subject variation was comparable to the same study conducted with fewer subjects. It would be therefore be useful to establish an evaluation measure that summarises model discrepancies relative to the between-subject variance, rather than sampling uncertainty. One approach it to count the proportion of predictions classified as potential outliers considering the variability in the empirical data:

$$\mathrm{SOR} = \frac{1}{N}\sum_{i=1}^{N} c_i \tag{7.5}$$

with

$$c_i = \begin{cases} 1: & \text{if } \hat{g}_i > Q_{3,i} + k \times \mathrm{IQR}_i \\ 1: & \text{if } \hat{g}_i < Q_{1,i} - k \times \mathrm{IQR}_i \\ 0, & \text{otherwise,} \end{cases}$$

where $Q_{1,i}$, $Q_{3,i}$ and IQR_i denote the first quartile, third quartile and IQR, respectively, of the subjective gains corresponding to segment i, and k is a factor commonly set to 1.5. If a model prediction stands out from the bulk of the subjective values, the outlier count c is incremented. The SOR is then the number of outliers as a proportion of the total number of predictions. Skovenborg and Nielsen (2004b) proposed the mean subjective deviation (MSD) statistic, which is also based on the IQR of the subjective gains:

$$\mathrm{MSD} = \frac{1}{N}\sum_{i=1}^{N} \frac{|\hat{g}_i - m_i|}{\mathrm{IQR}_i}, \tag{7.6}$$

where m_i is the median of the subjective gains for segment i. This statistic quantifies the discrepancy between the model prediction and the median of the gains, normalised by a robust measure of between-subject variability.

The final performance metric, proposed here, is the mean normalised percentile rank deviation

(MNPRD):

$$\text{MNPRD} = \frac{1}{N}\sum_{i=1}^{N} 1 - \frac{|50 - R_i|}{50}, \tag{7.7}$$

where R_i is the percentile rank of the model prediction, given the subjective gains for that segment. The MNPRD ranges between 0.5 and 1.0 for values falling within the IQR of the distribution, with higher values assigned to predictions closer to the median. As the predicted gain moves towards the tails of the distribution, MNPRD approaches zero. Note that the MNPRD is not based on the absolute discrepancy between the prediction and median gain. For example, if the subjective gains were positively skewed, such that values falling in the second quartile were closer to the median than those in the upper quartile, an error of, say, -1 dB would result in a lower (worse) MNPRD compared to an error of +1 dB; the statistic is more tolerant of positive errors because this is a characteristic of the underlying distribution.

7.4 Results

In this section, the evaluation results are presented for the four subjective datasets: $\text{A}_{16,\ 14}$, $\text{B}_{4,\ 8}$, $\text{C}_{12,\ 13}$ and $\text{D}_{10,\ 4}$. Recall that the first and second subscript denote the number of sounds and the number of participants involved in an experiment, respectively. After assessing the performance of the algorithms against the subjective data, an additional analysis was performed to inspect systematic differences in behaviour of the models when applied to a large corpus of multitrack audio. The main purpose of this analysis was to give some insight into inter-model differences when applied to a diverse selection of material and relate these comparisons to the main evaluation.

GM02	CF02	CH12	LML	LARM	EBU
100	100	100	97	80	86

Table 7.2: Optimal percentiles for measuring global loudness of the sound segments in dataset A.

The optimal percentiles for measuring the global loudness of dataset A are given in Table 7.2. Observe that the 100th percentile of the loudness-time series, equivalent to peak loudness, was optimal for the first three multiband models. A slightly lower than maximum loudness worked best for the LML. The percentiles were lowest for the two single-band algorithms, which is partly explained by the greater temporal smoothing compared to the other multiband models.

7.4.1 Dataset $\mathbf{A}_{16,\ 14}$

Boxplots of the absolute errors between the predicted and subjective gains are shown in Figure 7.6 for each model. Except for the baseline L_{eq} metric, the errors obtained using the two global-loudness descriptors are shown for each model: one using a form of temporal averaging of the loudness-time series (denoted μ), and another using the best fitting percentile P_X. The boxplots indicate that the algorithms performed best using the percentile descriptor, although there appears to be little practical difference between the two approaches for LARM and EBU—both the median and spread of the two distributions are similar. Better performance is to be expected for the percentile descriptor because it was derived using the same dataset, although peak loudness (P_{100}) has been used by previous researchers to determine the overall loudness using multiband models (Fastl and Zwicker 2007; Rennies et al. 2010). In contrast, the mean loudness descriptor introduced a few large outliers for the three published multiband models, with poor overall performance for the CF02 and GM02 where the median errors were around 6 dB. Surprisingly, the error distributions,

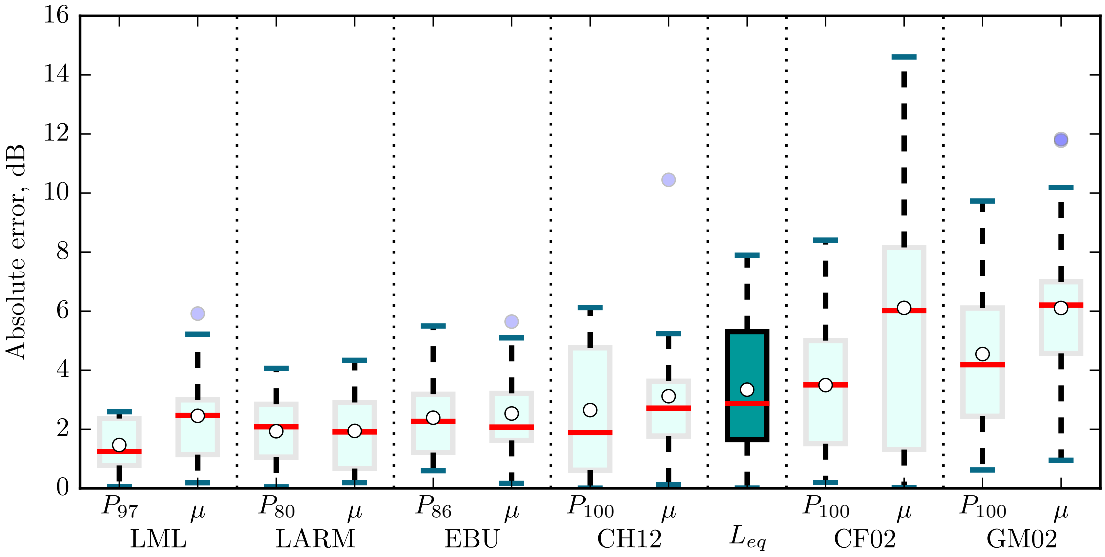


Figure 7.6: Boxplots of the absolute error for each model and descriptor on dataset $A_{16,\,14}$. Subscript μ denotes the global loudness measured using on a form of temporal averaging. Subscript P_X denotes the global loudness measured by taking the X^{th} percentile of the loudness-time function. The dark green boxplot is the $\mathrm{L_{eq}}$ metric, serving as a baseline metric. The models are sorted by ascending RMSE*, from left to right.

shown in Figure 7.6, of the best performing versions of these two models (with P_{100}) are comparable to that of the simple $\mathrm{L_{eq}}$ metric. Although the LML performed the best overall using P_{97}, it should be noted that this algorithm was biassed towards superior performance because an additional parameter was determined using the same dataset.

Model	Descriptor	RMSE*	RMSE	MSD	MNPRD	r	max(e^*)	SOR
LML	P_{97}	**1.03**	**1.70**	**1.00**	**0.37**	0.49	**1.84**	**0.06**
	μ	2.21	2.93	1.70	0.24	0.65	4.53	0.19
LARM	P_{80}	1.58	2.25	1.22	0.31	0.32	3.13	0.13
	μ	1.75	2.35	1.26	0.29	-0.0	3.41	0.13
EBU	P_{86}	2.08	2.79	1.52	0.23	0.45	4.5	0.31
	μ	2.19	2.94	1.56	0.2	0.47	4.65	0.31
CH12	P_{100}	2.72	3.4	1.8	0.3	0.5	4.79	0.31
	μ	3.16	3.86	1.76	0.16	**0.79**	9.07	0.38
$\mathrm{L_{eq}}$	μ	3.3	4.0	2.27	0.18	0.1	6.9	0.38
CF02	P_{100}	3.52	4.22	2.22	0.19	0.61	7.03	0.56
	μ	6.95	7.61	4.1	0.2	0.77	13.93	0.69
GM02	P_{100}	4.59	5.3	2.85	0.08	0.57	8.71	0.56
	μ	6.12	6.86	3.77	0.07	0.77	11.09	0.81

Table 7.3: Performance of the loudness models on dataset $A_{16,\,14}$. Models sorted by ascending RMSE*, from top to bottom. Values in bold red highlight the best ranking model for a given metric.

Table 7.3 shows the evaluation measures of seven statistics, with bold red text highlighting the best ranking model for a given metric. Note that RMSE*s are lower than their respective RMSEs because the former metric accounts for uncertainty in the subjective data. The $\mathrm{LML_P}$ performed best for all but the correlation coefficient, with distances from the confidence limits of the subjective gains averaging around 1 dB and all predictions within 2 dB of the confidence limits. One of the sixteen LML gains was flagged as an outlier by the SOR metric, but in most cases the predictions fell within 1 IQR of the median gain, as expressed by the MSD metric. As with the other multiband models, the LML did not perform as well when using the mean loudness descriptor. Excluding the MNPRD and r measures, both LARM and EBU single models outperformed the other multiband procedures. Surprisingly, the simple $\mathrm{L_{eq}}$ measure resulted in lower RMS and maximum errors than the best performing variants of the GM02 and CF02.

From observing Table 7.3, the correlation coefficient r does not tell the same story as the other metrics: the correlations are highest (best) for the three published multiband methods when using the mean loudness descriptor, yet the performance of these variants can be seen as inadequate when judged by metrics based on euclidean distance. This suggests the trend of the predicted adjustments were in line with the subjective values, but scaled such that the absolute error is inflated. LARM, for example, produced a much narrower spread of gains compared to the GM02—closer to the spread of the subjective gains—yet the GM02 amplified variations in loudness between certain sounds. In additional, no correlation of practical importance was found between the gains predicted by LARM_μ and those submitted by the subjects. A strong correlation should not be seen as favourable in the presence of large deviations in gain; it may simply indicate the need for a correction factor applied to the model predictions. Furthermore, the strength of the correlation is dependent on the variance in the subjective values; if all sounds were deemed equally loud by the group of subjects, the mean ratings would be randomly distributed about zero, a non-zero correlation would be highly unlikely.

Model	Descriptor	Worst 4 predictions (e^*)
LML	P_{97}	G3(1.4), P1(-1.7), V2(-1.8), V3(1.7)
	μ	G1(-3.2), P2(4.5), P3(4.4), V2(-2.2)
LARM	P_{80}	B0(-2.1), B1(-2.5), G3(3.1), P3(2.1)
	μ	G3(3.4), P0(-3.4), V0(2.2), V3(2.8)
EBU	P_{86}	B0(-4.5), B1(-4.2), G3(2.6), P2(2.4)
	μ	B0(-4.6), B1(-4.1), G3(2.4), P2(3.7)
CH12	P_{100}	B0(4.8), P1(-4.4), P2(4.7), P3(-4.2)
	μ	B0(3.6), G1(-4.4), P0(3.1), P2(9.1)
$\mathrm{L_{eq}}$	μ	B0(-6.9), B1(-4.7), P0(-4.8), P3(5.2)
CF02	P_{100}	B0(4.5), B3(6.8), P1(-4.4), P2(7.0)
	μ	G1(-10.8), P0(13.9), P2(12.0), P3(10.1)
GM02	P_{100}	B0(8.7), B1(5.9), B3(7.7), P2(7.3)
	μ	G1(-9.3), P0(11.1), P2(10.5), V1(-6.6)

Table 7.4: Four segments with the largest epsilon-insensitive errors (shown in brackets) per model/descriptor combination. B, G, P and V denote stimulus classes: bass guitar, electric/acoustic guitar, percussion and voice, respectively.

Table 7.4 shows, for each model variant, the signed epsilon-insensitive error e^* for specific segments falling in the upper 20% of the error distribution, i.e. the worst four predictions. A positive error indicates that the model underestimated the relative loudness of a particular segment, compensating by applying too much gain. In contrast, a negative error indicates an overestimation in relative loudness. Note that the choice of descriptor had no impact on the selection of the four segments for the EBU model. As shown in Figure 7.6, the peak loudness descriptor caused the most significant change in the error distribution for the CF02 and GM02, resulting in a different set of worst-case predictions. In the case of the mean loudness descriptor, the predicted relative loudness of segment P_2 (shaker) was too low for the three published multiband procedures, and the magnitude of these discrepancies decreased when using the peak to predict the subjective loudness judgement. The $\mathrm{GM02_P}$ had difficulties with three of the four bass samples, and all three models assign too much gain to the bass guitar segments.

7.4.2 Dataset $\mathrm{B}_{4,\,8}$

Dataset B was measured in an independent experiment in which eight subjects matched the loudness of four musical sounds taken from a rock multitrack recording: bass guitar, full drum kit, electric guitar and male singing voice. The predictions of all seven loudness models, configured as in the previous section, were compared with the subjective values—no further alterations were made to the algorithms.

Model	Descriptor	RMSE*	RMSE	MSD	MNPRD	r	max(e^*)	SOR
CH12	P_{100}	**0.35**	**0.95**	0.57	**0.44**	0.95	**0.57**	**0.0**
	μ	1.04	1.55	1.41	0.38	0.9	2.06	0.25
CF02	P_{100}	0.74	1.43	**0.54**	0.38	0.97	1.42	**0.0**
	μ	2.97	3.57	2.9	0.31	0.79	5.05	0.5
LML	P_{97}	0.75	1.49	0.86	0.38	0.91	1.22	**0.0**
	μ	1.35	2.01	1.76	0.25	0.76	2.23	0.5
GM02	P_{100}	1.18	1.91	0.74	0.38	**0.98**	1.84	**0.0**
	μ	3.84	4.53	3.36	0.31	0.86	5.83	0.5
LARM	μ	1.59	2.36	1.36	0.19	0.66	2.51	0.25
	P_{80}	1.95	2.77	1.43	0.25	0.48	3.32	0.5
EBU	P_{86}	1.72	2.57	1.23	0.13	0.74	3.01	0.25
	μ	1.9	2.76	1.61	0.13	0.96	2.99	0.5
L_{eq}	μ	2.33	3.17	1.59	0.13	0.04	3.5	0.25

Table 7.5: Performance of the loudness models on dataset $\mathrm{B}_{4,\,8}$. Models sorted by ascending RMSE*, from top to bottom. Values in bold red highlight the best ranking model for a given metric.

Table 7.5 summarises the performance of the models, sorted by RMSE*. Although the rankings are different from those in Table 7.3, the RMS and worst-case errors are fairly consistent for the LML, EBU and LARM. For example, the RMSE* for LARM$_\mu$ was 1.8 dB for dataset A and 1.6 dB for dataset B. Thus, reducing the evaluation to just four sounds did not cause any significant change in predictive quality. The LML$_{\mathrm{P}}$ gave a low RMSE* of 0.8 dB, and the maximum deviation from the confidence limits was 1.2 dB, with no predictions considered as outliers when compared with the subjects. This suggests that the LML generalises well beyond the training data for these types of sounds. As with dataset A, the choice of descriptor was not crucial for the EBU model.

The CH12, CF02 and GM02 all performed remarkably better in this evaluation, although the rank order, with respect to one another, is the same as in the previous evaluation. The CH12$_{\mathrm{P}}$ performed exceptionally well: its RMSE* was just 0.4 dB, with a max(e^*) of 0.6 dB. The largest discrepancies for all four multiband models, using the percentile measure, were not considered as outliers when considering the inter-subject variance, and MSDs were all less than one. Using the mean of the loudness-time function to predict global loudness degraded prediction accuracy for all multiband methods, most notably for the CF02 and GM02 where maximum error increased by about 3.8 dB, and two of their four predictions classified as outliers.

Inspection of the predicted gains revealed that those generated by the multiband models, configured with the peak loudness descriptor, were in excellent alignment with the subjective gains, whereas the EBU and LARM overestimated the relative loudness of the bass guitar (the largest discrepancy). In the case of the EBU$_{\mathrm{P}}$, this explains the single outlier (value = 1/4) shown in the final column of Table 7.5.

For all but the L_{eq}, the correlation was considerably higher for this dataset, which is attributed to the greater spread in the segment gains as measured experimentally. The correlation coefficient was essentially zero for the L_{eq}, which can be explained by the RMS normalisation used to calibrate the stimuli. In other words, the $\mathrm{L_{eq}}$ measure did very little to the four segments since they were RMS normalised, after accounting for the response of the headphones, prior to the main experiment. Had the transfer function of the headphones been taken into account by the $\mathrm{L_{eq}}$ metric, r would be undefined. In contrast, the gains generated by the $\mathrm{GM02_P}$ were highly correlated with those of the listeners ($r = 0.98$).

7.4.3 Dataset $\mathrm{C_{12,\,13}}$

Model	Descriptor	RMSE*	RMSE	MSD	MNPRD	r	max(e^*)	SOR
LML	P_{97}	**2.02**	2.82	1.17	0.35	**0.61**	5.50	**0.17**
	μ	2.19	2.80	**1.10**	**0.45**	0.40	**4.55**	0.25
EBU	μ	2.05	**2.75**	1.23	0.33	0.27	5.35	0.25
	P_{86}	2.25	2.96	1.4	0.36	0.27	4.98	0.33
LARM	P_{80}	2.09	2.85	1.36	0.32	0.47	5.14	0.25
	μ	3.23	3.9	1.46	0.41	0.27	8.78	**0.17**
$\mathrm{L_{eq}}$	μ	3.38	4.07	1.68	0.26	-0.05	10.17	0.25
CH12	μ	3.81	4.59	2.03	0.14	0.4	9.37	0.58
	P_{100}	4.3	5.1	2.29	0.12	0.47	10.72	0.42
CF02	P_{100}	4.35	5.21	2.49	0.08	0.54	7.78	0.67
	μ	7.05	7.8	3.94	0.04	0.3	18.2	0.67
GM02	P_{100}	5.94	6.85	3.34	0.05	0.56	12.3	0.83
	μ	6.99	7.87	4.15	0.03	0.4	14.03	1.0

Table 7.6: Performance of the loudness models on dataset $\mathrm{C_{12,\,13}}$. Models sorted by ascending RMSE*, from top to bottom. Values in bold red highlight the best ranking model for a given metric.

This dataset was derived from a loudness-matching experiment involving thirteen audio engineers and twelve sound segments. The performance of the loudness models is summarised in Table 7.6. In contrast to the previous evaluations, all RMSEs were greater than 2 dB, and max(e^*) values range between 4 and 12 dB for the best performing versions of each algorithm. None of the SORs were zero, so taken together the inflated distance measures can be attributed to a few segments causing problems for the algorithms. For the best performing CH12, CF02 and GM02 adaptations, notably large RMSE*s of 4–6 dB were measured, with a large proportion of predictions classified as outliers. In terms of average performance, the predictions of these three multiband models were inferior to the simple $\mathrm{L_{eq}}$ metric. As in the previous evaluations, switching from the percentile to the mean global-loudness descriptor degraded the performance of the CF02 and GM02, but had less of an impact on the EBU and LML. The $\mathrm{GM02}_{\mu}$ was the worst performing algorithm, with every prediction considered an outlier. Of the three published multiband models, the $\mathrm{GM02_P}$ gave the highest correlation but ranked lowest according to the other metrics.

The best performing variants of the LML, EBU and LARM were very similar in terms of the two RMSE metrics. The EBU performed slightly better using the mean descriptor (programme loudness), and the LML and LARM using percentile loudness. The predictions of these models were roughly 1–1.4 times the IQR from median of the subjective gains, with two/three of their predictions considered outliers.

Figure 7.7 shows the distribution of absolute segment errors for each model/descriptor pairing. When using the mean loudness descriptor, all but two of the gains predicted by LARM were within 2.3 dB of the subjective gains, but their outliers were higher than those measured on the predictions of the EBU or LML. As shown in Figure 7.7, the median errors corresponding to the three published multiband methods all lie above the median of the $\mathrm{L_{eq}}$ metric, with the GM02

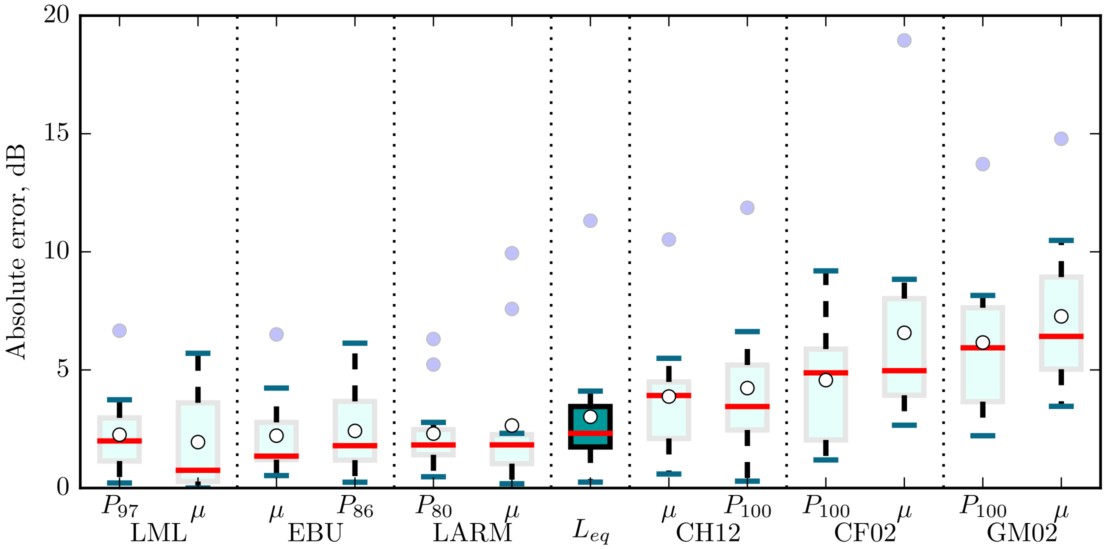


Figure 7.7: Boxplots of the absolute error for each model and descriptor on dataset $C_{12,\ 13}$. Subscript μ denotes the global-loudness descriptor using a form of temporal averaging. Subscript P_X denotes the global-loudness measured by taking the X^{th} percentile of the loudness-time function. The dark green boxplot is the L_{eq} metric, serving as a baseline metric. The models are sorted by ascending RMSE*, from left to right.

performing the worst overall. The boxplots suggest that all of the models had difficulties with particular sounds—most notably the CF02$_\mu$ which gave the largest segment error. These outlying prediction errors are listed in Table 7.7.

Model	Descriptor	Worst two predictions (e^*)
LML	P_{97}	Bass (2.3), Triangles (5.5)
	μ	Kick (4.2), Triangles (4.6)
EBU	μ	Kick (3.5), Triangles (5.3)
	P_{86}	Kick (3.7), Triangles (5.0)
LARM	P_{80}	Claves (-4.1), Triangles (5.1)
	μ	Claves (-8.8), Triangles (6.4)
L_{eq}	μ	Cymbal (2.9), Triangles (10.1)
CH12	μ	Q.Vocal (-4.7), Triangles (9.3)
	P_{100}	Bass (5.2), Triangles (10.7)
CF02	P_{100}	Triangles (7.1), Bass (7.7)
	μ	B.Vocal (-8.1), Kick (18.2)
GM02	P_{100}	Piano (-7.3), Bass (12.3)
	μ	Bass (9.0), Kick (14.0)

Table 7.7: Two segments with the largest epsilon-insensitive errors (shown in brackets) per model/descriptor combination.

As with dataset A, the choice of descriptor led to a different selection of segments for the multiband methods. When using the average loudness to predict the overall loudness judgement, the kick drum was problematic for the CF02 and GM02, with the former setting the gain 18 dB *beyond* the average subjective adjustment. Using the maximum loudness instead of the average reduced this discrepancy because of the greater weight assigned to the transient, although both algorithms still underestimated the contribution of the bass guitar by about 8–12 dB (with respect to the 95% CI). LARM$_{\text{P}}$ (incorrectly) set the level of the claves 4.1 dB lower than the average listener, which is likely attributed to the fast time constant of 1.45 ms used for temporal integration,

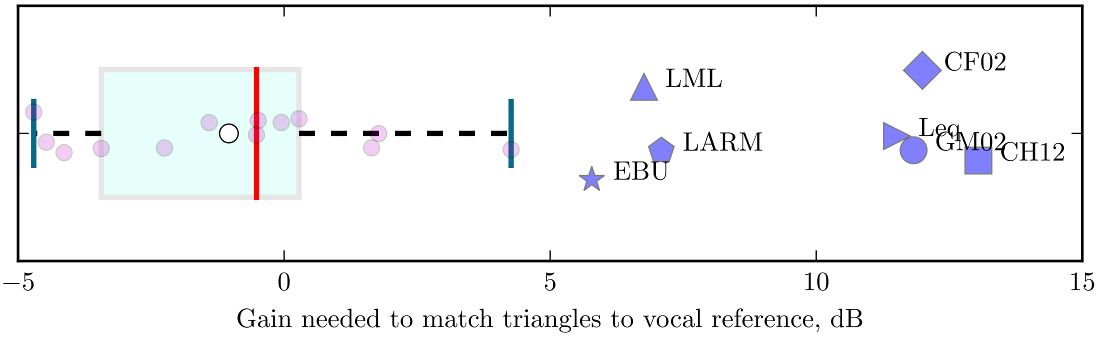


Figure 7.8: Boxplot of the loudness matches submitted by the 13 subjects when matching triangles to the male vocal. Red vertical line shows the median, white circle shows the mean, and purple circles show individual data points. The blue markers show the predictions of the algorithms using the percentile descriptor.

leading to a large loudness estimate for this segment. Interestingly, all models underestimated the relative loudness of the triangles. Figure 7.8 shows the level adjustments applied by the subjects when matching the triangles to the vocal (the reference stimulus), along with the model predictions. Although there is fairly wide spread of subjective adjustments, it can be seen that the algorithms were off by 5–15 dB, with the three published multiband models and the L_{eq} applying the highest level correction.

7.4.4 Dataset $D_{10,\ 4}$

Dataset D was formed using only four participants, and although within-subject variability was reduced by incorporating redundancy into the experiment design, and the gains were highly correlated across subjects, the results of this assessment should be viewed only as indicative. Statistics based on sampling error were not used, i.e. those requiring a CI, and the median was used to quantify central tendency of the gains. RMS and maximum errors were thus based on the absolute deviation from the median of the subjective gains. Since the agreement between the four subjects was high, the algorithms were punished harder by metrics that consider inter-listener variability, e.g. MSD, MNPRD and SOR. The evaluation results using the remaining statistics are given in Table 7.8.

As with dataset B, the correlations were high ($r > 0.80$), which is partly explained by the peak normalisation used to set the reference levels of the stimuli. Had an approximate equal-loudness normalisation been carried out, as was done in experiments A and C, the correlations would have been lower. The correlations were highest for the LML, EBU and L_{eq}, indicating that these algorithms captured the trend of the variance in the gains. However, the average deviation from the medians of the subjective gains was only 1.3 dB and 1.5 dB for the EBU_P and LML_P, respectively, compared to the slightly higher RMSE of 1.9 dB for the L_{eq}. The maximum discrepancies (max(e)) were also less severe for these two models (2.6 dB). The EBU performed notably better using the percentile descriptor, which contrasts with the previous three assessments where differences in error caused by descriptor were negligible.

Maximum error was markedly large for the GM02 and CF02, especially when using the mean loudness descriptor. At best, their RMSEs were roughly twice as large as that of the EBU. Although half of the GM02 predictions were considered, by the SOR metric, to be outliers, this result is somewhat unfair due to the small number of subjects. However, the maximum errors

Model	Descriptor	RMSE	MSD	MNPRD	r	max(e)	SOR
EBU	P_{86}	**1.32**	2.19	**0.35**	0.93	2.61	0.3
	μ	1.9	2.58	0.25	**0.94**	3.99	0.4
LML	P_{97}	1.46	**1.74**	0.25	0.84	**2.55**	0.3
	μ	2.39	2.75	0.45	0.91	5.32	0.5
CH12	μ	1.5	2.04	0.25	0.86	2.65	**0.2**
	P_{100}	1.93	1.74	0.3	0.83	4.5	0.3
LARM	μ	1.63	1.79	0.15	0.81	3.2	0.2
	P_{80}	1.72	2.52	0.3	0.83	3.07	0.4
L_{eq}	μ	1.85	2.26	0.25	0.91	3.9	0.4
GM02	P_{100}	2.67	2.63	0.15	0.59	6.6	0.5
	μ	4.58	4.69	0.4	0.86	9.22	0.5
CF02	P_{100}	3.03	3.07	0.15	0.55	7.72	0.6
	μ	5.58	5.46	**0.35**	0.89	12.61	0.6

Table 7.8: Performance of the loudness models on dataset $\mathrm{D}_{10,\,4}$. Models sorted by ascending RMSE, from top to bottom. Values in bold red highlight the best ranking model for a given metric.

were very large compared to the better performing estimators. These two multiband models gave the lowest correlation when determining global loudness using the peak value, whilst using the average loudness resulted in a high correlation but markedly high absolute error. In contrast, the CH12 performed better with the mean loudness descriptor when assessed using the RMSE and max(e) metrics, although not in terms of MSD and MNPRD. This is indicative of a few outlying predictions introduced by the peak loudness descriptor.

Since the GM02 and CF02 models performed about 1–2 dB worse than the simple L_{eq}, it is of interest to know where their pitfalls lie, especially considering that their maximum errors were far greater. When using the mean loudness descriptor, the two segments (and signed error) with the largest deviations from the median of the four subjects were: drums (+9.2 dB) and electric bass (+6.7 dB) for the $\mathrm{GM02}_{\mu}$, and drums (+12.6 dB) and accordion (-7.6 dB) for the $\mathrm{CF02}_{\mu}$. Switching to the peak loudness descriptor caused these errors to drop to -2.8 dB (drums) and +6.6 dB (bass) for the $\mathrm{GM02}_{\mathrm{P}}$. For the $\mathrm{CF02}_{\mathrm{P}}$, the error for the drums and accordion decreased to -2.8 dB and -2.3 dB, respectively, but the error for the bass guitar increased from +6.0 dB to +7.7 dB, being the largest discrepancy. Thus, though measuring global loudness using the peak value greatly improved the overall fit of both algorithms, the relative loudness of the electric bass was severely miscalculated irrespective of descriptor.

7.4.5 Overall performance

Due to the relatively small sample of sounds used in these experiments, it is difficult to gauge, from a single dataset, an estimate of *typical* model performance for a general collection of musical sounds. Indeed, CIs for the error statistics generated using the sound segments as random variables were not judged reliable by the author after inspecting the bootstrap distributions.[5] Instead, this section compares the errors across datasets to draw empirical inferences on expected behaviour.

The top three ranking models in each dataset, as measured using four performance metrics, are shown in Table 7.9. In all but three conditions, the LML ranked third or higher. The EBU and LARM also ranked highly for datasets A, C and D, but the multiband models ranked higher on dataset B and were generally superior according to the correlation coefficient.

Figure 7.9 shows, for each of the four datasets, scatter plots of max(e), RMSE and MSD

[5] This was not always the case but depended on both the dataset and the error metric.

Dataset	RMSE	MSD	r	max(e)
A	LML$_{\mathrm{P}}$	LML$_{\mathrm{P}}$	CH12$_{\mu}$	LML$_{\mathrm{P}}$
	LARM$_{\mathrm{P}}$	LARM$_{\mathrm{P}}$	CF02$_{\mu}$	LARM$_{\mathrm{P}}$
	EBU$_{\mathrm{P}}$	EBU$_{\mathrm{P}}$	GM02$_{\mu}$	EBU$_{\mathrm{P}}$
B	CH12$_{\mathrm{P}}$	CF02$_{\mathrm{P}}$	GM02$_{\mathrm{P}}$	CH12$_{\mathrm{P}}$
	CF02$_{\mathrm{P}}$	CH12$_{\mathrm{P}}$	GF02$_{\mathrm{P}}$	LML$_{\mathrm{P}}$
	LML$_{\mathrm{P}}$	GM02$_{\mathrm{P}}$	EBU$_{\mu}$	CF02$_{\mathrm{P}}$
C	EBU$_{\mu}$	LML$_{\mu}$	LML$_{\mathrm{P}}$	LML$_{\mu}$
	LML$_{\mu}$	EBU$_{\mu}$	GM02$_{\mathrm{P}}$	EBU$_{\mathrm{P}}$
	LARM$_{\mathrm{P}}$	LARM$_{\mathrm{P}}$	CF02$_{\mathrm{P}}$	LARM$_{\mathrm{P}}$
D	EBU$_{\mathrm{P}}$	LML$_{\mathrm{P}}$	EBU$_{\mu}$	LML$_{\mathrm{P}}$
	LML$_{\mathrm{P}}$	CH12$_{\mathrm{P}}$	LML$_{\mu}$	EBU$_{\mathrm{P}}$
	CH12$_{\mu}$	LARM$_{\mu}$	LML$_{\mu}$	L$_{\mathrm{eq}}$

Table 7.9: Top three ranking models in each dataset, according to four evaluation metrics. Models are ordered from top to bottom by ascending rank.

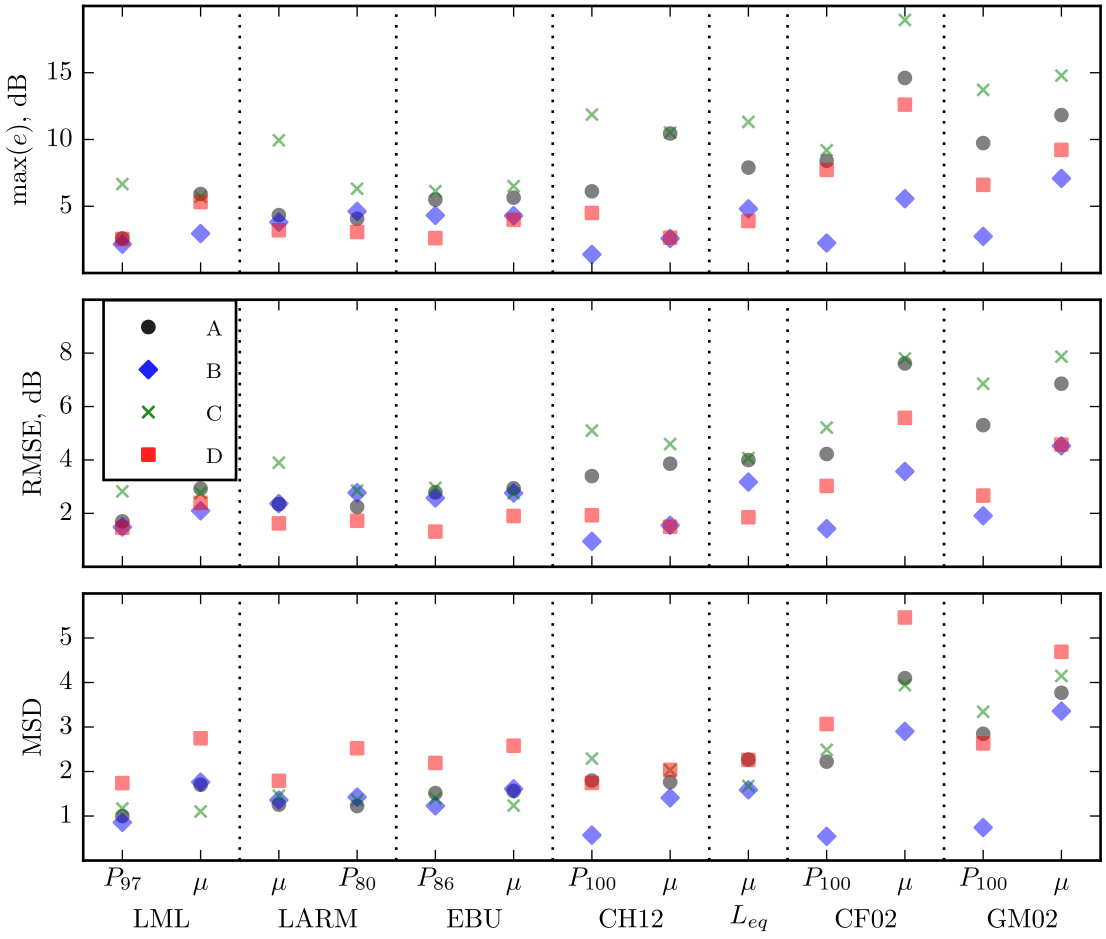


Figure 7.9: Scatter plot of max(e) (top), RMSE (middle) and the MSD (bottom) as measured on datasets A-D for each model/descriptor combination. From left to right, the models are ordered by ascending MSD of the pooled data.

across the different model/descriptor combinations. Recall that the MSD is based on the absolute deviation between the model prediction and the median of the subjective gains, normalised by the IQR. The MSD values were generally highest for dataset D which is attributed to the high agreement amongst the four participants. Thus, the models receive a larger penalty compared to the other datasets, despite the RMSEs being comparable (or indeed lower in the case of dataset D).

Worst-case error was generally the largest on dataset C across all models. As discussed, specific segments from this dataset, notably the triangles, were more problematic for the algorithms.

The LML and EBU models were fairly consistent across both descriptors used to quantify global loudness. Conversely, the CF02 and GM02 gave notably large absolute errors when using mean loudness, especially for datasets A and D where maximum errors exceeded 10 dB. Excluding dataset B, these two models failed to outperform the simple L_{eq} metric, especially when comparing the MSD values. The four multiband models performed best on dataset B.

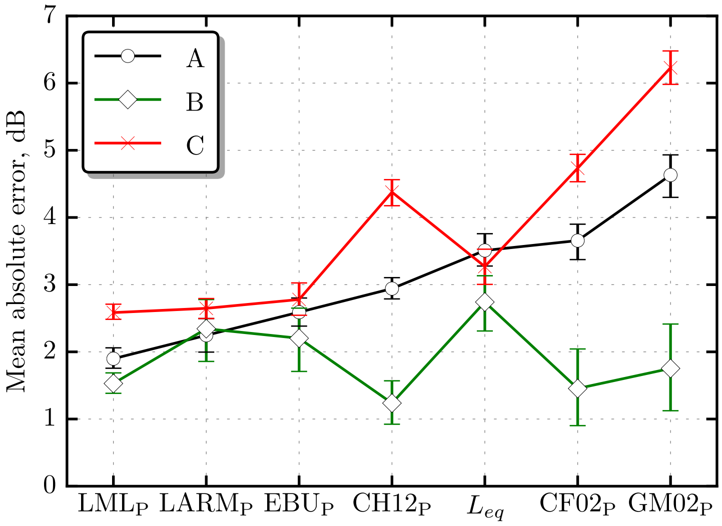


Figure 7.10: Mean absolute error across subjects by model for datasets A (black), B (green) and C (red). Error bars show the bootstrapped 95% within-subject CI of each mean. Excluding the L_{eq}, the models were configured with the percentile loudness descriptor.

Figure 7.10 shows the mean absolute error (MAE) between the subjective gains and those predicted by the models configured with the percentile loudness descriptor. The 95% within-subject CIs were computed using the errors from individual subjects, and thus give an indication of the differences in error if these experiments were repeated using a different set of subjects (but the same collection of sounds).[6] Results for dataset D are not shown as only four subjects contributed to that assessment. It can be seen that the CIs are widest for dataset B, which can be attributed to fewer subjects than in the remaining assessments. For these data, there was strong overlap between the four multiband procedures, and the CH12 performed most favourably. The single-band methods performed better than the three published multiband models on the remaining datasets, especially dataset C, where differences of 1 dB and above can be observed from the Figure.

The plotted mean errors corresponding to the LML, LARM and EBU do not appear to be reliably different for dataset C, nor are these differences of practical importance. The LML outperformed the two single-band models on datasets A and B, which is to be expected given that the LML was tuned to the former and both collections included bass, guitar, drums and vocals. Note that the two single-band methods were at most about 1 dB worse than the LML for these two cases. The GM02 gave the largest average error on datasets A and C, 2.5 dB higher (best case) than the simple L_{eq} measure for the latter dataset. The CF02 performed similarly to the L_{eq} on dataset A, but was around 1 dB worse on dataset C. Note that although these two models were most successful at predicting the relative loudness of dataset B, which comprised only four

[6] Bootstrap distributions for the within-subject CIs were approximately normally distributed.

sounds, they ranked bottom on dataset D.

7.4.6 Inter-model differences

To further explore the performance of the different algorithms, a pairwise analysis of the gains required to loudness normalise a large corpus of musical sounds was performed. The stimuli consisted of 110 segments extracted from 26 different multitrack recordings, all but one was taken from the Free Multitrack Download Library (Cambridge Music Technology 2013). Project genres included rock, blues, jazz, classical, hip-hop, dance and electronic music. Additional segments were taken from a string quartet recording captured by the author. A wide selection of tracks were taken from the 26 sessions to give a diverse yet representative corpus of material commonly experienced by music producers and recording engineers. A total of 110 tracks were automatically segmented into 4–6 s excerpts using an adaptive algorithm that detected activity in the recordings, which were then verified by the author through listening. The collection consisted of bass (14 samples), brass (4), guitar (14), percussion (25), strings (16), voice (15), and synthesised sounds and a few miscellaneous instruments such as bells and accordion (22). The segmentation procedure applied 0.1 s ramps to the onset and offset of each clip, which was then RMS normalised to 73 dB SPL. This level was judged by three experienced listeners to give a comfortable overall loudness when listening back to a pink noise reproduced over a single loudspeaker position 1.2 m away from the centre of the head. The RMS normalised segments were also comfortable to listen to at this level, despite obvious fluctuations in loudness across the segments.

The models were configured as done throughout this chapter, but this time using the default presentation transfer function. This corresponded to a free-field (frontal incidence) transfer function for the published multiband models, the RLB curve for LARM, and the K-filter for the EBU and LML. This was done to simulate the performance of the models in studio environments where music is reproduced via loudspeaker positioned roughly 1–2 m in front of the head.[7]

Equal-loudness gains were obtained for the 110 sounds using each algorithm configured with both the mean and percentile descriptors. Since the segments were already RMS normalised, the gains can be viewed as deviations from an $\mathrm{L_{eq}}$ alignment. Figure 7.11(a) shows the distribution of RMS levels after loudness equalisation according to each procedure, using the percentile descriptor. The spread of levels is markedly wide for the three published multiband procedures, indicating a greater sensitivity to the physical characteristics of the stimuli. Recall that these algorithms were configured using an outer-ear filter that modifies the input spectrum quite drastically at higher frequencies compared to filters employed by the other algorithms. The RLB mainly affects frequencies below 100 Hz, whilst the highshelf constituting the second stage of the K-filter gives a 4 dB boost above 3 kHz - modest when considering the complex form of the free field to eardrum transfer function as measured acoustically. The top 5% of all positive gains within each model were predominately associated with bass sounds (guitars and synthesisers), demonstrating a common strategy across the algorithms to attenuate low frequencies and in most cases boost frequencies in the 1–10 kHz region when measuring loudness. This was especially apparent for the GM02, where gains large as 32 dB were applied to bass and kick drum segments, resulting in very high reproduction levels. The EBU gave the narrowest spread of levels, with 75% of the segments (for both descriptors) within 2.5 dB of the input level. Note also that the estimated probability density of the EBU levels is more uniform compared to the remaining densities. The EBU, followed by LARM, was therefore the most consistent with a simple energy measurement, i.e. the $\mathrm{L_{eq}}$. The behaviour of the LML appears to lie somewhere between the other multiband models and single-

[7] For modelling convenience, exact speaker placement and reproduction cross-talk was not taken into account.

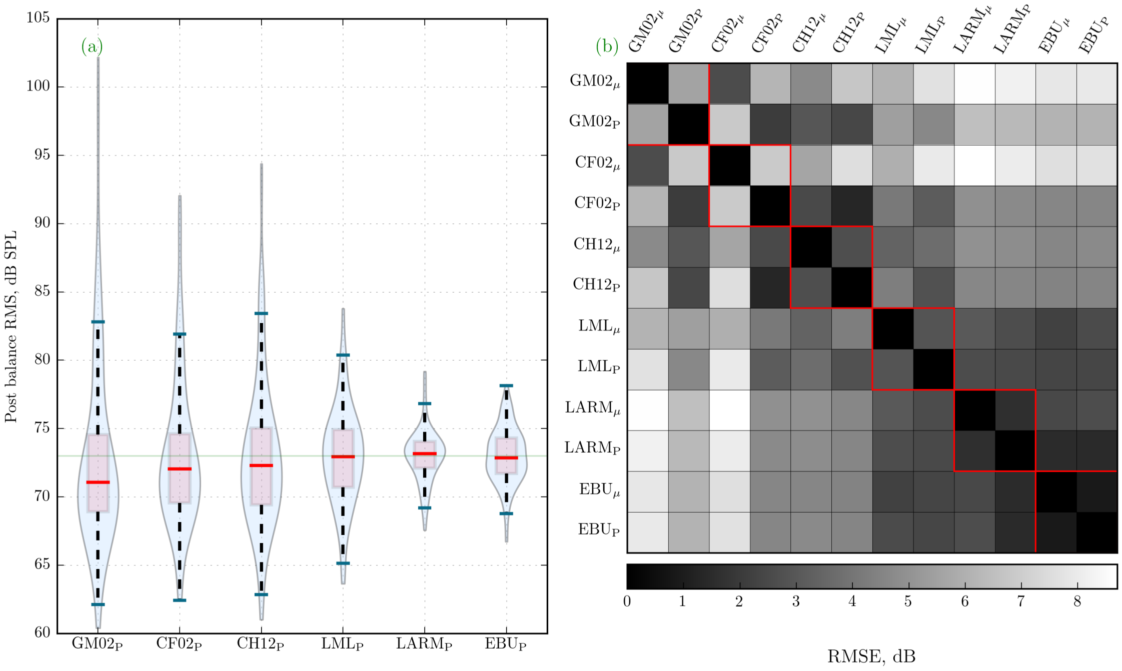


Figure 7.11: (a) Violin plots of the stimulus RMS levels after loudness equalisation, and (b) RMSE matrix for assessing the similarity of the gains. The green horizontal line in (a) shows the input level for all segments. Red squares group together same model comparisons.

band models.

Figure 7.11(b) shows the RMSE of the pairwise comparisons of the predicted equal-loudness gains. In terms of percentile loudness, the two models that agreed the most was the CF02 and CH12 (RMSE = 1.3 dB, 95% CI = [1.0, 1.6]), followed by LARM and the EBU (RMSE = 1.4 dB, 95% CI = [1.2, 1.6]). The two models that produced the most contrasting equal-loudness gains was the GM02 and LARM (RMSE = 6.3 dB, 95% CI = [5.2, 7.5]), followed by the GM02 and EBU (RMSE = 6.1 dB, 95% CI = [4.8, 7.3]). These four effects were expected given the different architectures of the algorithms. The similarity matrix reveals that the GM02 and CF02 yield notably different gains when compared to the two single-band methods, especially when using the mean loudness descriptor. Consistent with the previous assessments presented in this chapter, these two algorithms perform markedly different depending on the global-loudness descriptor.

Model	EBU	LARM	CH12	LML	GM02	CF02
RMSE (dB)	0.86	1.60	2.67	2.89	5.56	6.87
95% CI (dB)	[0.75, 0.97]	[1.34, 1.82]	[2.2, 3.12]	[2.53, 3.22]	[4.68, 6.41]	[5.67, 8.15]

Table 7.10: RMSE, with bootstrapped 95% CI, between the equal-loudness gains obtained using mean loudness and percentile loudness descriptors. See Table 7.2 for percentile values.

Table 7.10 gives, for each model, the exact RMSE, with bootstrapped 95% CI, between the gains obtained using the two descriptors. The choice of mean or percentile measurement was indeed less crucial for the EBU model, where average adjustments differed by 0.9 dB. In contrast, the choice of descriptor had the largest effect on the CF02 and GM02 gains, consistent with the results from the previous assessments involving subjective data. Peculiarly, the CH12 was less influenced by descriptor compared to the GM02, despite both having similar spectral decomposition and temporal integration stages. This discrepancy is primarily attributed to the different exponential smoothing times. The GM02 uses attack and release times of 22 ms and 50 ms for the STL, respectively, and

100 ms and 200 ms for the LTL. Recall that the LTL decay constant was originally 2 s, which was revised in subsequent work (Moore et al. 2003). In contrast, the CH12 uses STL attack and release times of 16 ms and 32 ms, respectively, and LTL time constants of 100 ms and 2 s. As a result, the LTL-time series output by the CH12 emphasises peaks in the STL more so than the GM02, and therefore, the mean LTL tends to track the peak STL more closely in the case of CH12.

7.5 Discussion

7.5.1 Overview

In music production, audio engineers work with a specific class of sounds called multitrack audio. In general, this includes individual musical instruments such as those comprising a four piece rock band. In the field of Intelligent Music Production (IMP), tools are developed to aid the mixing engineer with the process of manipulating these individual elements to construct a single sonic mixture. It is therefore imperative that these devices listen to the incoming sound streams in a manner consistent with human perception. Since the most basic task in music production is the level balance, this study has assessed the ability of different computational methods to measure the perceived loudness of musical sounds.

The predictive quality of seven algorithms was assessed using four independent subjective datasets. Three of the these were published multiband models: the GM02 (Glasberg and Moore 2002), the CF02 (Chalupper and Fastl 2002) and the CH12 (Chen and Hu 2012). These devices are peripheral ear models that employ an auditory filter bank to transform the spectrum into an EP from which loudness can be computed. The predicted loudness of a sound is given in sones, meaning that iterative methods are required to map the loudness domain (sones) to the level domain (dB) commonly measured in subjective experiments. Three heuristic single-band models, which are based on measures of frequency-weighted energy, were also assessed: the EBU (EBU R 128 2014; ITU-R BS.1770 2015; EBU Tech 3341 2016), LARM (Skovenborg and Nielsen 2004b) and the L_{eq} metric. In addition to these established procedures, a new linear multiband loudness (LML) model was proposed which extends the simple frequency-weighting energy measurement by employing an auditory filter bank to incorporate spectral effects on loudness.

Except for the L_{eq}, which was included as a baseline metric, the algorithms were configured to use the transfer function of the BeyerDynamic DT990 headphones used in the four listening assessments; this was done to establish a consistent and fair evaluation. For example, the GM02, CF02 and CH12 all require the method of stimulus presentation to be specified. The EBU algorithm assumes loudspeaker reproduction, incorporating a highshelf filter to boost the level in the upper-frequency range to simulate the acoustic influence of the head. The EBU, LARM and LML models all make use of the RLB filter to approximate the ears reduced sensitivity to lower frequencies. Therefore, filters based on low-frequency sensitivity or the transmission of sound through the middle ear were left intact, but outer-ear filtering was configured to simulate the headphone reproduction used in the listening tests.

The four subjective datasets were derived from loudness-matching experiments involving short-duration (3–5 s) segments of multitrack audio. Dataset A was established using a balanced pair-matching experiment involving 14 subjects and 16 sounds composed of bass, electric, and acoustic guitars; lead and group vocals; and percussive instruments. Redundancy was incorporated in the experiment design to reduce both within-subject inconsistency and systematic errors associated with the experimental method. Dataset B involves a bass guitar, electric guitar, drum kit and a singing voice sample, all taken from a multitrack recording. Eight listeners matched the loudness of all pairwise combinations of the stimuli. In dataset C, 12 subjects matched 11 sound segments to a singing voice recording. For dataset D, four subjects matched the loudness of 10 musical sounds. Redundancy was also used here in that multiple sounds served as the reference stimulus. The samples used in the last two experiments were not all from the same instrument categories characterising the first two. For all datasets, a regression analysis was conducted to estimate a set of relative loudness levels, from which a set of gains for loudness normalising the stimuli was determined. The loudness models were evaluated by assessing their ability to reproduce the subjective equal-loudness gains.

When conducting a loudness-matching experiment, it is assumed that the listener is able to judge the overall loudness of a given sound. A loudness meter outputs a loudness-time function, from which it is necessary to determine a single value representative of this global-loudness percept. In most cases developers suggest a statistic that quantifies a form of central tendency, such as the arithmetic mean (Glasberg and Moore 2002), gated average (ITU-R BS.1770 2015), or power mean (Skovenborg and Nielsen 2004b). These global-loudness descriptors were all used in the evaluation, and if no metric based on a time average was given, the arithmetic mean of the time series was used. In addition, a percentile measurement was established for each procedure (excluding the L_{eq}), using dataset A to determine the optimal percentile value. For the three published multiband models, peak loudness was best (100th percentile). For the LML, EBU and LARM, the 97th, 86th, and 80th percentiles were selected, respectively. Furthermore, the amount of critical-band compression employed by the LML was determined using a constrained optimisation routine with dataset A as the training data.

The algorithms were evaluated using distance metrics, percentile ranking and correlation measures. Variability in the subjective data was also taken into account when comparing model predictions against the relative gains submitted by the subjects. Statistical methods were also drawn upon to take into account the variability in the results. It was found that the correlation coefficient r strongly depended on the spread of the subjective gains, a finding also highlighted by Skovenborg and Nielsen (2004b). In many cases, models with high r values were also found to have large absolute errors, making comparisons between the loudness models difficult based on r alone. However, high correlations point towards a rescaling of a model's estimates in order to better align them with the subjective gains.

7.5.2 Global-loudness descriptor

One of the main findings was that both the GM02 and CF02 performed poorly when using mean loudness to estimate the overall loudness impression. These two configurations always resulted in considerable maximum errors. For example, in the evaluation on dataset A, the largest deviation from the limits of the 95% CIs was 13.9 dB for the CF02 and 11.1 dB for the GM02. Even higher discrepancies were measured on dataset C using these two models. Note that the authors of the GM02 suggested taking the mean LTL for estimating the global loudness of a fluctuating sounds such as speech (Glasberg and Moore 2002; Moore et al. 2003), an approach that been used for loudness estimation in IMP applications (Ward et al. 2012; Terrell et al. 2013; Wichern et al. 2015). However, whilst the mean loudness would likely work well for stimuli of the same type, e.g. speech (Moore et al. 2003), it fails to capture the saliency of transients in more dynamic material. Peak loudness was therefore found to be superior for determining the global loudness of the more complex musical sounds evaluated here. This is consistent with the suggestions of Zwicker (Zwicker 1977; Fastl and Zwicker 2007) and those of more recent work (Chalupper and Fastl 2002).

The LML performed best using percentile loudness, especially for the assessments involving percussive material. The choice of loudness descriptor was less important for the single-band methods, and further research is warranted to address this issue for the CH12. These findings were supported by an inter-model analysis using a wide range of multitrack material which revealed that the choice of loudness descriptor had a large effect on the equal-loudness gains generated by the GM02 and CF02 (see Table 7.10). Compared to these two algorithms, the CH12 performed more consistently over both descriptors, which is explained by the sustained release of its LTL integrator.

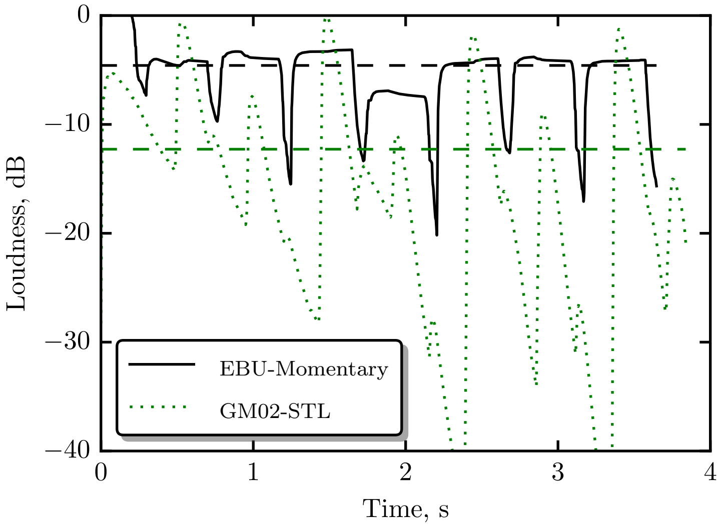


Figure 7.12: Peak normalised loudness-time series output by the EBU and GM02 models in response to a drum kit. Two time functions are output by the GM02: the STL (green dotted line) and the LTL (not shown). The average LTL is shown by the green dashed horizontal line. The black solid line shows the EBU ML, and the black dashed line shows the programme loudness.

The choice of descriptor was less crucial (<1 dB) for the EBU model, because (a) the selected percentile lies between the median and peak value meaning percentile loudness was close to the average, and (b) the ML fluctuates over time to a lesser degree compared to the other models because of differences in temporal integration. This is demonstrated in Figure 7.12 which shows the loudness-time functions output by the EBU and GM02. To facilitate the comparison, the STL and LTL generated by the GM02 were converted to phons, and then peak normalised by a common factor. The green dashed horizontal line shows the mean LTL output by the GM02. The black dashed horizontal line shows the programme loudness output by the EBU. Notice that level difference between the average and peak loudness is greatest for the GM02 (roughly 13 dB). It can be seen that the STL of the GM02 shows greater temporal variation compared to the ML of the EBU; the temporal profile of the drums is more clearly defined by the former model. The GM02 uses a 22 ms attack time and a 50 ms release time, which is then further smoothed by the LTL integrator (100 ms attack and 200 ms release). In contrast, the EBU makes use of a 400 ms sliding window which is then gated and averaged over to arrive at the programme loudness. Thus, the spread of the STL distribution tends to be far greater than that of the ML, leading to greater differences between the centre of the distribution and its maximum.

7.5.3 General performance

Overall, the CF02 and GM02 performed the worst on the larger datasets (A, C and D), with RMSEs ranging between about 2.5 and 8 dB, and maximum errors as high as 18 dB. Compared to these two procedures, the CH12 gave lower RMS and maximum errors across all four datasets (considering the best performing variants of each), ranking first on dataset B. For datasets A, C and D, the simple L_{eq} performed at least as good—if not better—than the more complex GM02 and CF02 methods. Further testing is warranted to better gauge the predictive quality of the CH12, but this model does seem the most promising of the three published EP based procedures. In general, when dataset B is ignored, it is difficult to make a case for the GM02, CF02 or CH12, for the types of sounds considered here, especially considering the simplicity of LARM and the EBU. These two heuristic algorithms were generally consistent across the four datasets, and may therefore be considered more robust to different real-world stimuli.

The LML performed well in all four datasets when using the 97th percentile of the STL to determine global loudness. Although this model was (more) biassed towards superior performance on dataset A, it performed at least as good as the better performing algorithms in the other evaluations, indicating good generalisability. Unlike the single-band methods, the performance of the LML did not deteriorate on dataset B. This suggests that all four multiband procedures, particularly the LML, do offer some benefit over the single-band approach, but it is difficult to reconcile the findings across all four datasets for the three published methods. With that exception, the LML, LARM and EBU models appear to be the most robust loudness models applicable to multitrack material, with percentile loudness working well across all three algorithms. Although these findings suggest that the EBU model performs slightly better by taking the 86th percentile of the ML, further research is needed for comparison with the integrated programme loudness approach (ITU-R BS.1770 2015). However, in conjunction with the results given in Table 7.10, informal listening tests suggest only minor perceptible differences between the two approaches.

The seven algorithms underestimated the loudness of the triangles relative to the loudness of vocal as used as the reference stimulus in dataset C. After listening to the loudness-normalised stimuli for each algorithm, the author believes that the subjective adjustment may have been biassed by the *annoyance* of the strong tonal resonances of the triangles. Indeed, four of the algorithms did apply excessive gain to the triangles for equal loudness (see Figure 7.8), but the adjustment predicted by the LARM, LML, and EBU models should not be considered widely different from the subjective adjustment. Figure 7.13 shows the temporal waveforms and spectrograms of the two segments at the point of equal loudness as determined by the average listener in the assessment. It can be seen that the RMS level of the triangles is 7.9 dB lower than the level of the reference. The spectrogram corresponding to the triangles reveals very distinct spectral lines concentrated in the upper-frequency region, producing a rich tonal sound that may have been perceived as sharp or piercing by the listeners.

7.5.4 Spectral effects

From analysing the individual segment errors across the different datasets, the GM02, CF02, and to a lesser extent the CH12, appear to over-boost the level of bass instruments, especially when using the peak loudness descriptor. This finding is consistent with the subjective assessment carried out in Chapter 5, where the GM02 was used to automatically balance the loudness of five multitrack recordings: the automatic mixing device always produced a bass heavy mix for songs involving bass guitar and kick drum. When comparing the equal-loudness gains generated by the algorithms for a larger corpus of audio files (Section 7.4.6 in this study), large differences in gain correction were identified for bass instruments. It was initially believed that the pre-cochlear filters of the multiband methods would explain such discrepancies, but further investigation indicates that the modelling of SLS plays an important role.

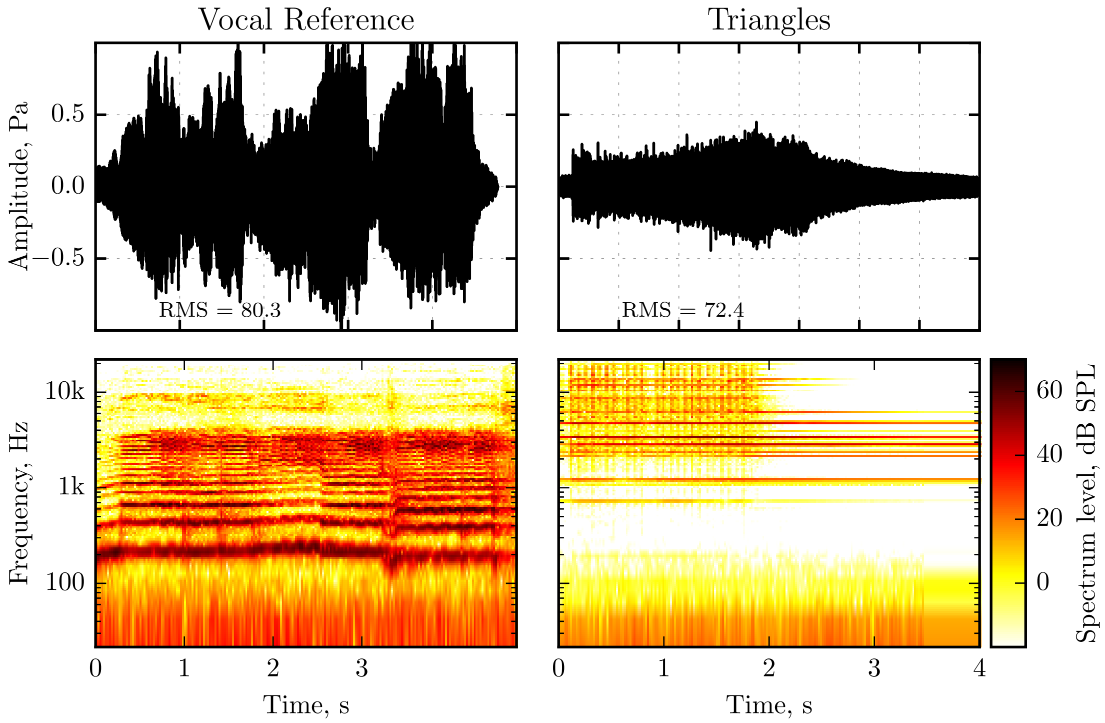


Figure 7.13: Waveforms and power spectrograms of the triangles and vocal reference at the point of equal loudness. RMS levels (dB SPL) are shown in the top subplots.

In dataset A, the gain applied to bass segment B_0 by all three models was 5–10 dB above the average subjective gain. In contrast, the gains applied by the best variants of the EBU and LARM algorithms were between 3 and 5 dB *below* the subjective adjustment. The GM02 displayed notable problems with three of the four bass segments (B_0, B_1 and B_3). Although the LML performed exceptionally well on this dataset, further informal listening tests carried out by the author indicate that this algorithm also shares a similar tendency as the GM02 to over-boost low-frequency instruments, especially when the amount of critical-band compression, as defined by the exponent α, is increased, i.e. when $\alpha \to 0$. These observations are demonstrated in Figure 7.14 which shows the temporal waveforms and excitograms of the four bass guitar samples which were computed using the GM02 with the outer- and middle-ear filter bypassed. The RMS level of each segment was set to 70 dB SPL. It can be seen that the excitogram of sample B_2 (third along) is broader than the other samples whose spectra are concentrated below about 10 Cams (444 Hz).

The equal-loudness gains for these four samples, as computed using the four multiband models are displayed in Table 7.11. Two sets of gains are shown for the LML, one for each α value. In all cases, the sound field to cochlea transmission filters were bypassed, and global loudness was measured using the 95^{th} percentile of the STL. Recall that the LML was configured with the same time-constants as the GM02, and therefore, any differences in gain were primarily based on differences in the transformation from excitation to SL.

All models measured sample B_2 as the loudest, resulting in a negative level adjustment, and the largest level differences (for equal loudness) was associated with sample B_0 and B_2. This occurred because sample B_0 is relatively tonal, with localised energy in select critical bands, whereas sample B_2 excites a wider area over the perceptual-frequency scale. Consequently, the amount of predicted SLS was greatest for sample B_2, leading to an attenuation of signal level. The LML configured with $\alpha = 0.56$ corresponds to the model used throughout this study and was the most conservative with

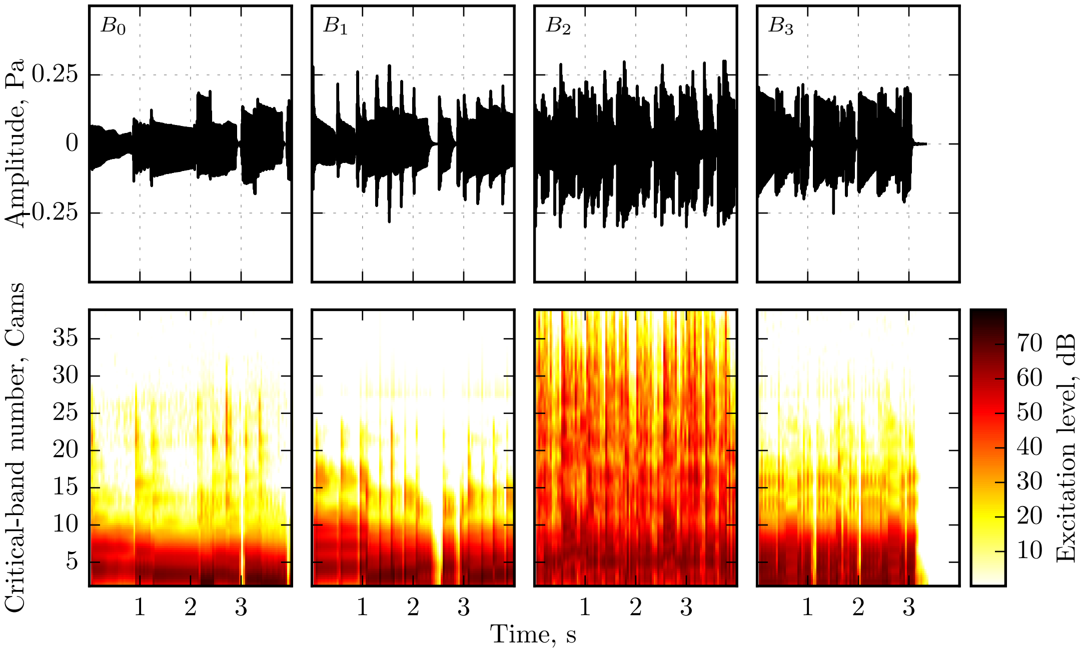


Figure 7.14: Temporal waveforms (top) and excitograms (bottom) corresponding to the four bass guitar samples of dataset A. The excitograms were computed using the GM02 with the outer- and middle-ear filter bypassed. The RMS level of each segment was 70 dB SPL.

Model	B_0	B_1	B_2	B_3
LML ($\alpha = 0.56$)	1.4	0.2	-2.0	0.5
LML ($\alpha = 0.22$)	6.2	2.7	-10.4	1.4
GM02	6.7	2.0	-10.6	1.8
CF02	4.7	1.7	-9.2	2.1
CH12	5.9	2.1	-9.8	1.8

Table 7.11: Equal-loudness gains (dB) generated by the multiband models for the four bass guitar samples of dataset A. The input level of each segment was 70 dB SPL and the pre-cochlear filters were bypassed on all models.

respect to level adjustment. Reducing α to 0.22, however, introduced greater critical-band intensity compression, and so the amount of SLS increased, resulting in a higher relative loudness of sample B_2. As a result of this, the equal-loudness gains generated by the LML were comparable with those of the GM02, as were the gains of the CH12 and CF02. Note that when the pre-cochlear filters were reintroduced, the spread in the levels was found to further increase, being most pronounced for the three published models. For example, when using the BeyerDynamic DT990 and middle-ear transfer functions, the GM02 predicted a very large matching gain of 20.0 dB associated with segment pair B_0/B_2, whereas the LML predicted 9.9 dB using the same pre-cochlear filter.[8]

In short, these findings suggest that current multiband algorithms overestimate the loudness of broadband sounds relative to narrowband sounds as a result of too much SLS. A second, more pronounced example of this effect is shown in Figure 7.15 which shows the excitograms of a bass guitar and vocal segment from dataset C. This time the excitograms were generated using the same configuration of the GM02 as used to predict the data in the main assessment. It can be seen that the vocal, which served as the reference stimulus, excites a broader area on the critical-band scale compared to the bass guitar, implying greater SLS. The bottom subplot is similar to

[8] Recall that, during the main assessments, the LML was configured with RLB filter and not the middle-ear filter proposed by (Glasberg and Moore 2002).

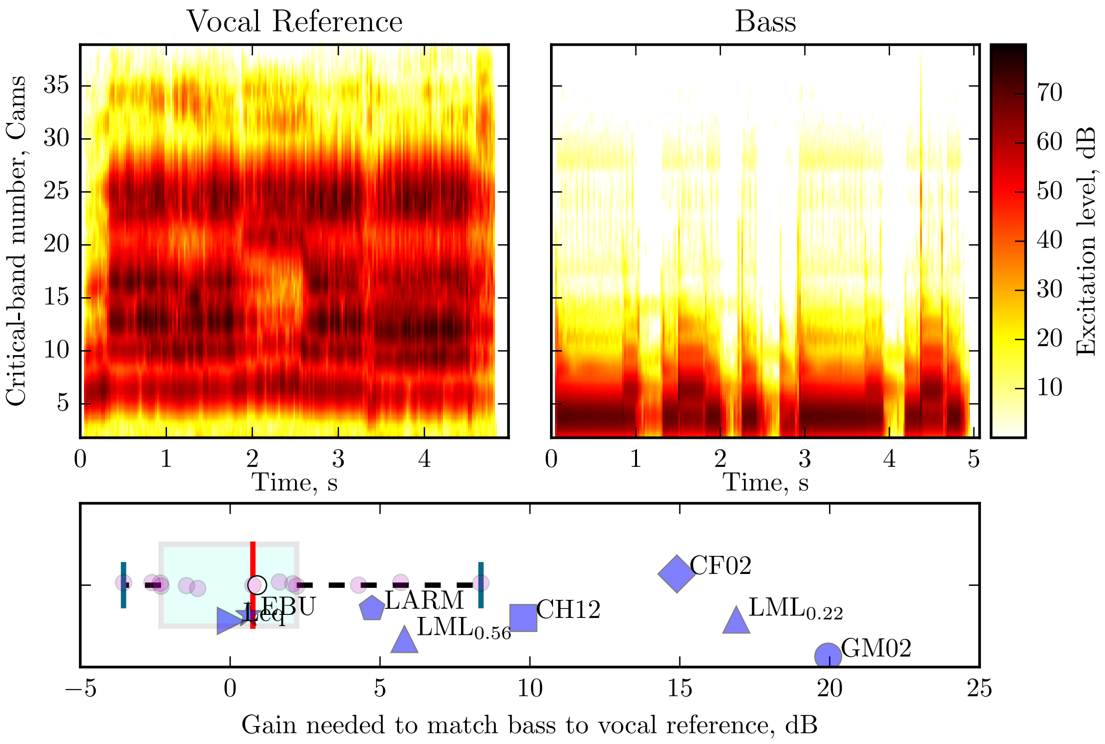


Figure 7.15: GM02 generated excitograms (top) of the bass guitar and vocal segments in dataset C. The bottom subplot the subjective adjustments (purple circles) required to match the loudness of the two samples, supported by a boxplot of the experimental matches. The labelled markers show the predictions of the algorithms. The numerical subscript of the LML label reflects the compressive exponent α used to model SLS.

Figure 7.8 but shows the experimentally measured matching gains (with the corresponding boxplot) between the bass guitar and vocal, and those predicted by the algorithms. A positive gain means that the bass had to be increased in level in order to match the loudness of the vocal. The predictions of the EBU and L_{eq} models fell close to the median/mean of the data, which was close to 0 dB, suggesting that equal RMS approximates equal loudness for these stimuli. As expected, the multiband models predicted a greater positive adjustment compared to the single-band methods, which can be attributed to broader EP of the vocal. This was most extreme, and erroneous when considering the subjective adjustments, for the CF02, GM02, and the LML with $\alpha = 0.22$. Therefore, even though the LML did not make use of the same middle-ear filter as employed by the GM02, increasing the amount of critical-band compression resulted in an alignment of their relative loudness estimates. The CH12 predicted a lower matching gain compared to the CF02 and GM02, which is believed to be attributed to the faster growth in loudness of the bass guitar as the sound level is increased. In other words, this model predicts less SLS at high intensities.

High-resolution long-term excitation- and SL-patterns of these two segments are shown in subplots (a) and (b) of Figure 7.16, respectively. The auditory patterns were calculated according to ANSI S3.4 (2007) from the long-term power spectra. Note that this stationary loudness procedure uses the same auditory peripheral model as employed by the GM02. The stimuli were input to the model at their reference levels, which were roughly 80 dB SPL. It can be seen that the EP of the bass guitar (black) is highly tonal compared to that of the vocal (red), with the vast majority of energy localised in the auditory filter centred at 3.5 Cams. The total excitation level has been annotated for the two EPs, along with the level difference ΔL (in green). The total excitation

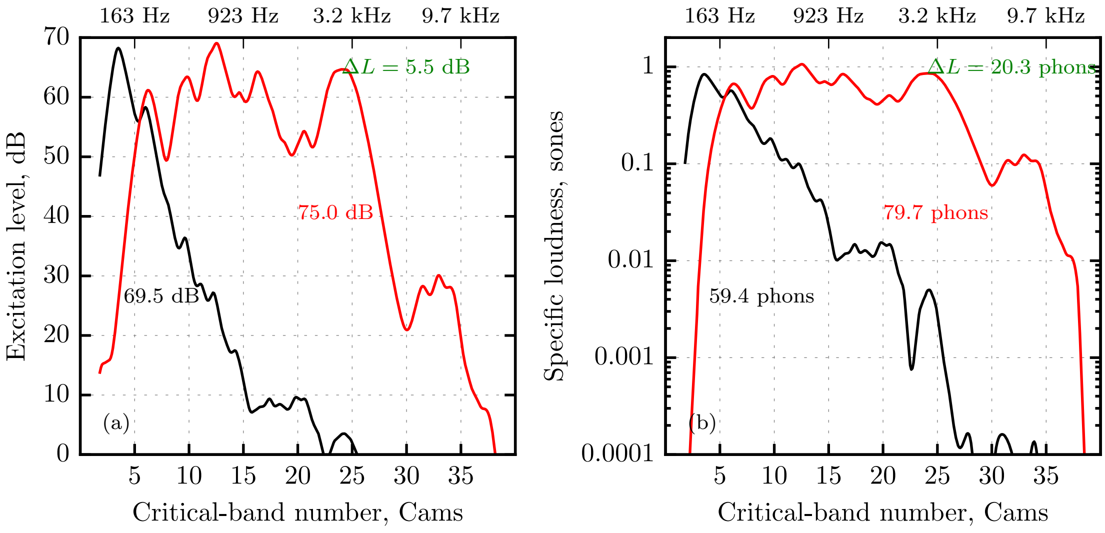


Figure 7.16: (a) EPs and (b) SL patterns of the bass guitar (black lines) and vocal (red lines) segments of dataset C. The signals were filtered using the transfer function of the headphones as used in the listening test, and then power spectral densities were computed using Welch's method with a window size of 8192 and hop size of 256 samples. The auditory patterns were then calculating according to ANSI S3.4 (2007).

of the vocal was 5.5 dB higher than that of the bass, attributed to the transmission loss of the middle-ear filter at low frequencies.[9] The SL patterns shown in subplot (b) reflect the compressive nonlinearity employed by the algorithm to transform the EP to a loudness density. This time, the area under the curves are measured in sones and have been transformed and displayed as a loudness level in phons. As a result of this nonlinearity, the (approximate) level difference required to equate the loudness of these two spectra was far greater than that needed to equate overall excitation. As a result, the stationary model predicted a greater loudness of the vocal. This confirms that the majority of the matching gain applied to the bass guitar by the GM02 is attributed to the SL transformation, and not to differences in stimulus energy or indeed the total amount of excitation.

Figure 7.17 shows a set of predicted equal-loudness contours to highlight differences in frequency sensitivity between the algorithms. The contours were derived by sampling the audible frequency range using 30 frequencies logarithmically spaced between 30 Hz and 10 kHz. A 1 kHz pure tone presented at an effective eardrum level of 75 dB SPL was used to establish a target loudness. The algorithms were configured with the same headphone transfer function as used in the listening experiments. Note that (a) there is a single contour representing LARM and EBU since they both use the RLB filter,[10] and (b) no contour is shown for the L_{eq} because the contour is simply the inverse of the response shown in Figure 7.3. The trend of the contours produced by the three published multiband models are very similar, and generally reveal greater level-boost at lower frequencies compared to the remaining models. The contours corresponding to the LML, EBU and LARM are fairly consistent with one another across the frequency range, with small deviations below about 30 Hz, attributed to LML's frequency-domain processing. As expected, the GM02 assigned the most gain below 1 kHz, with significant deviations from the other methods occurring near 100 Hz (roughly 10 dB higher than the CH12 and LML).

Note that whilst the plotted contours corresponding to the GM02 and CH12 are comparable,

[9] This was verified by removing the middle-ear filter.

[10] Recall that the first stage highshelf of the K-filter was bypassed.

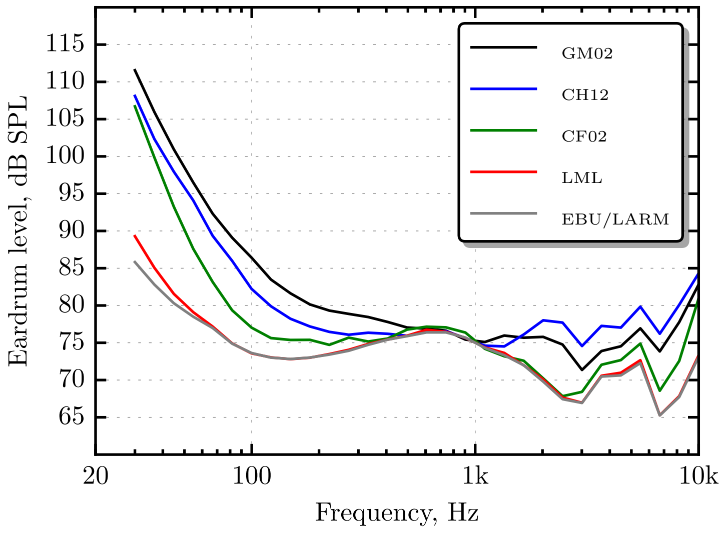


Figure 7.17: Equal-loudness contour predictions for the different loudness models evaluated in this study. Each contour shows the signal gain required to match the loudness of a pure tone of a given frequency to that of a 1 kHz tone presented at an effective eardrum level of 75 dB SPL.

the CH12 performed better in the each of the four assessments. This further suggests that the SL transformation strongly influences the predictions of these algorithms. In addition to the bandpass shaped frequency sensitivity of these EP based models, this gives some insight into why they tend to place narrowband sounds—especially in the low-frequency region—at higher reproduction levels when matching musical sounds of varied spectral content.

7.5.5 Related work

Music and speech

Two of the largest evaluations of loudness models applied to music and speech were performed by Soulodre and co-workers (Soulodre et al. 2003; Soulodre and Norcross 2003; Soulodre 2004), and Skovenborg and Nielsen (2004b). These studies contributed to the efforts of the International Telecommunication Union Radiocommunication Sector (ITU-R) who were investigating audio meters suitable for use in digital sound production (ITU-R SRG-3 2002). Soulodre et al. (2003) conducted a fixed-reference loudness-matching experiment involving 48 audio segments of broadcast audio, primarily speech presented with and without background sounds (music, environmental sounds and other talkers). A few items were mastered songs with and without a lead singer. The 48 segments were matched twice by 25 listeners, once with a positive level offset applied to the variable stimulus and again with a negative level offset applied. Stimuli were presented over a single loudspeaker positioned directly in front of the listener. A segment of female speech presented at 60 dB SPL(A) was used as the reference signal. Soulodre and Norcross (2003) compared the predictions of seven variants of the $\mathrm{L_{eq}}$ metric against the subjective matches. Each $\mathrm{L_{eq}}$ had a specific frequency weighting: flat; A, B and C weightings; an inverse absolute threshold contour; a variant of the B-weighting curve; and the RLB curve as proposed by the authors. They concluded that the $\mathrm{L_{eq}}$(RLB) performed best, but did acknowledge that the $\mathrm{L_{eq}}$(broadband) and $\mathrm{L_{eq}}$(C) variants were comparable in performance. For example, both the $\mathrm{L_{eq}}$ and the $\mathrm{L_{eq}}$(RLB) had an RMSE of 1.35 dB and a maximum error of 4.3 dB, respectively. The MSD was just 0.43 for the $\mathrm{L_{eq}}$(RLB), and 0.45 for the $\mathrm{L_{eq}}$. The A and absolute-threshold weightings, both characterised by significant low-frequency roll-off, resulted in maximum errors as high as 7–8 dB, and RMSEs near 3 dB. In a second paper, Soulodre (2004) pooled together the results from experiments with the same procedure (stimuli, setup and loudness-matching method) conducted at five separate sites around the world, giving a total of 97 listeners. 10 meters, submitted by several audio companies and research institutions, were assessed on their ability to predict the relative loudness of the test items.[11] The results of the evaluation were consistent with the previous study: $\mathrm{L_{eq}}$(RLB) ranked first, and the simple $\mathrm{L_{eq}}$ second.

Under criticism from algorithm proponents that the selected audio segments were not sufficiently representative, two further validation tests were conducted (ITU-R BS.1770 2015). The second dataset was roughly equally split into new monophonic segments of speech and music, and the third was a mixture of mono, dual mono, stereo and multichannel sequences (Soulodre and Lavoi 2005). The correlation between the subjective matches and the predictions of the $\mathrm{L_{eq}}$(RLB) was very high in all three datasets ($0.979 \leq r \leq 0.985$), and the ITU-R concluded that the multichannel algorithm performed very well over the 336 sequences of the subjective database covering a broad range of programme material. Soulodre and Lavoi (2005) found that summing the different channels using a dummy head and a sphere microphone gave practically the same goodness-of-fit measures, and, as a result, suggested that the speaker-to-eardrum transfer function be modelled as a rigid sphere. This led to the highshelf 'pre-filter' used in the first stage of the K-filter, prior to the RLB weighting (ITU-R BS.1770 2015).

Around the same time as Souldre's investigations, Skovenborg and Nielsen (2004b) assessed the performance of 12 loudness models on two subjective datasets derived from two separate listening experiments. Unlike Soulodre's larger study, the two different datasets were independent in that the subjects, stimuli and setup were different, although the experimental procedure was the same. The first 'music' dataset consisted of 145 items, extracted from rock and pop music, jazz music, classical music and recordings of speech. The second 'speech' dataset featured 11 rock and pop songs and 37 segments of speech. A balanced pair-matching procedure was used to obtain a set of relative loudness estimates for each dataset. A similar set of evaluation metrics to those used in this chapter were employed to assess the performance of 12 different algorithms: a multiband model called HEIMDAL and single-band algorithm called LARM (both developed by the authors); $\mathrm{L_{eq}}$, $\mathrm{L_{eq}}$(A-D), $\mathrm{L_{eq}}$(M) and $\mathrm{L_{eq}}$(RLB); the peak programme meter (PPM); and two implementations of Zwicker's loudness model (Zwicker 1960, 1977; Zwicker et al. 1991; Fastl and Zwicker 2007). In the case of the latter model, Skovenborg and Nielsen found the 95% percentile of the loudness-time function achieved the highest accuracy compared to the 50^{th} and 75^{th} percentiles. This agrees with the findings reported in this chapter: the multiband procedures founded on Zwicker's algorithm perform better when greater weight is assigned to loudness maxima.

For both the music and speech datasets HEIMDAL ranked first, LARM second, and $\mathrm{L_{eq}}$(RLB) third. The algorithms were classified into four groups based on the 95% CI of the errors: group 1 consisted of HEIMDAL and LARM; group 2 of $\mathrm{L_{eq}}$(RLB), $\mathrm{L_{eq}}$(C), and $\mathrm{L_{eq}}$; group 3 of $\mathrm{L_{eq}}$(B), PPM and the two Zwicker models; and group 4 of $\mathrm{L_{eq}}$(D), $\mathrm{L_{eq}}$(A) and $\mathrm{L_{eq}}$(M). Class 1 was deemed reliably better than classes 3 and 4, and class 2 was found to be reliably better than class 4. The largest RMSE across the two datasets was 0.75 dB for HEIMDAL and 0.96 dB for LARM. 95% of the absolute errors ($\mathrm{E_{95}}$) were below 1.5 dB and 2 dB for HEIMDAL and LARM, respectively. In the two evaluations, the predictions of these algorithms fell, on average, within 1 IQR of the subjective levels. The RMSE of the $\mathrm{L_{eq}}$(RLB) and $\mathrm{L_{eq}}$ meter was around 1 dB on both datasets, but the MSD was largest in the music evaluation, being about 1.2. $\mathrm{E_{95}}$ was about 0.5 dB larger given the musical stimuli, averaging around 2.3 dB. For this same dataset, Zwicker's model scored

[11] Note that the identities of the majority of models submitted by the different organisations were anonymous.

Metric	This study				Soulodre	Skovenborg & Nielsen	
	A	B	C	D		Music	Speech
MAE	3.3	2.7	3.0	1.4	0.9	1.0	0.8
Max E	7.9	4.8	11.3	3.9	3.6	**6.0**	**2.7**
MSD	2.3	1.6	1.7	2.3	0.45	1.2	0.7
Max SD	7.1	2.6	4.5	6.8	-	**5.8**	**2.5**

Table 7.12: Performance of the L_{eq} metric as measured in this study and by Soulodre (2004) and Skovenborg and Nielsen (2004b). Values in bold were obtained through visual inspection.

an RMSE of 2.1 dB, with an E_{95} of 3.6 dB and an MSD of 2.3.

Although Soulodre and Skovenborg assessed the models on a larger number of sounds than used here, it is believed that the type of audio employed in this study is more difficult for the algorithms to measure. This is demonstrated in Table 7.12 which shows, from top to bottom, the mean and maximum absolute error, and the mean and max Subjective Deviation measures of the L_{eq} on the four datasets used in this study, and those reported by the two aforementioned studies. Firstly, the MAEs are lower for the three studies involving speech and commercial music, although the small sample size used in the experiments conducted here does mean that outliers have greater influence over the measures. However, the maximum error for Skovenborg and Nielsen's speech database, was only 2.7 dB. Similarly, Soulodre reported a comparatively low maximum error of 3.6 dB, which focused primarily on dialogue mixed with background sounds. The worst-case predictions become increasingly large as the sound material becomes more complex, with poor performance on datasets A and C where maximum errors of 7.9 and 11.3 dB are observed, respectively.

More interesting are the measures of Subjective Deviation, which takes into account the spread of the subjective levels. Soulodre obtained a low MSD of 0.45, which Skovenborg and Nielsen (2004b) attributed to a greater spread in the subjective values, which effectively reduces the Subjective Deviation through IQR normalisation. Since both the MAE and MSD values are larger in the four assessments conducted here, it seems plausible that the degraded performance of the L_{eq} can be attributed to a more difficult selection of sound segments. Had the spread of subjective levels been significantly wider in this study, the MSD would compensate for the greater absolute deviations. Since this is not the case, and the finding that the L_{eq} was outperformed by a number of other models assessed in this investigation, one can infer that these differences are attributable to the type of audio, rather than the consistency or agreement of the participants.

Metric	This study				Skovenborg & Nielsen	
	A	B	C	D	Music	Speech
MAE	1.9	1.9	2.6	1.0	0.7	0.6
Max E	4.3	3.8	9.9	2.2	**3.1**	**2.1**
MSD	1.3	1.4	1.5	1.6	1.0	0.5
Max SD	2.7	2.1	6.1	5.0	**4.8**	**2.1**

Table 7.13: Performance of the LARM model as measured in this study and by Skovenborg and Nielsen (2004b). Values in bold were obtained through visual inspection.

In agreement with the results reported here, Skovenborg and Nielsen found LARM to outperform the L_{eq} across several metrics on both datasets.[12] Table 7.13 shows that LARM performed remarkably well on speech, with a MAE as low as 0.6 dB and the largest error being 2.1 dB. The statistics indicate that the performance of LARM deteriorates as the segments become more varied in content. The absolute errors measured here are indeed greater than those reported by

[12] Although LARM was not found to be statistically better than the L_{eq}, judging by the 95% CIs, LARM was as least as good as the L_{eq}.

Skovenborg and Nielsen, but when normalised by the between-subject variability, the errors are not too dissimilar from their music assessment. Thus, the multitrack material was more difficult for LARM, certainly when compared to its superior performance on the speech material, but the uncertainty of the subjective levels was also slightly greater in these assessments, as discussed in Chapter 6.

Metric	This study				Skovenborg & Nielsen	
	A	B	C	D	Music	Speech
MAE	3.5	1.1	4.6	2.3	1.7	1.2
Max E	8.4	2.2	9.2	7.7	>**5.0**	**3.3**
MSD	2.2	0.5	2.5	3.1	2.3	1.0
Max SD	5.3	0.9	4.7	8.5	>**6.0**	**4.5**

Table 7.14: Performance of Zwicker's model measured in this study (using the CF02) and by Skovenborg and Nielsen (2004b). Values in bold were obtained through visual inspection.

Finally, Table 7.14 compares a few measures of goodness-of-fit for Zwicker's loudness model. The comparison was made using the results obtained with the CF02 (peak loudness) to determine overall loudness. The CF02 model is a refinement of the classic procedure of Zwicker (1960, 1977), but in principle should not give markedly different results when predicting data measured using normal-hearing subjects. This model performed well on dataset B, though this did only involve four sounds. The MAE is largest for the remaining three assessments carried out here, although this is likely due to outliers, as indicated by the maximum errors. The MSD values obtained here are comparable to those reported by Skovenborg and Nielsen in their music concentrated assessment. This suggests a wider spread in the subjective values, which is expected considering the type of sounds involved. Note that, according to the information presented in these tables, the Zwicker-type models are inferior to both LARM and the $\mathrm{L_{eq}}$ in assessments A, C, D, and those conducted by Skovenborg and Nielsen (2004b).

Seefeldt et al. (2004) of Dolby Laboratories extended a Zwicker-type psychoacoustic model, developed as part of the ITU-R study, and compared its performance against the $\mathrm{L_{eq}}$ and $\mathrm{L_{eq}}$(RLB) measures. The essence of this algorithm is a simplified version of the GM02. The authors modified the SL transformation to better predict the loudness growth of wideband and narrowband noise. This resonates with the findings presented here that the multiband procedures require a modification to the SLS stage. Seefeldt et al. (2004) assessed their revised psychoacoustic model, the $\mathrm{L_{eq}}$ and the $\mathrm{L_{eq}}$(RLB) using an 'ITU-R type' subjective dataset, and an 'expanded' dataset involving 51 segments taken from a more varied sound collection. The authors noted that although most segments involved speech, the expanded dataset involved some solo instruments with narrowband spectra. The key results from this study are summarised in Table 7.15.

Model	$\mathrm{Type_{ITU\text{-}R}}$			$\mathrm{Type_{Expanded}}$		
	MAE	Max E	MSD	MAE	Max E	MSD
$\mathrm{L_{eq}}$	1.2	3.6	0.7	1.8	11.3	1.0
$\mathrm{L_{eq}}$(RLB)	1.2	3.6	0.6	1.7	10.5	1.0
Modified Psychoacoustic	1.1	3.4	0.6	1.1	3.7	0.7

Table 7.15: Performance of three loudness models on two datasets reported by Seefeldt et al. (2004). $\mathrm{Type_{ITU\text{-}R}}$ refers to the subjective dataset most similar to the first dataset used to develop ITU-R BS.1770 (2015), and $\mathrm{Type_{Expanded}}$ is the extended dataset.

The errors are largest for the expanded dataset, which can be attributed to the more diverse segment selection. As reported in previous work, the $\mathrm{L_{eq}}$ and $\mathrm{L_{eq}}$(RLB) metrics are consistent with one another. The modified psychoacoustic model is superior on the expanded dataset, especially in terms of maximum error. This result is interesting because it suggests that the Zwicker-type methods have potential to outperform simple single-band models, a result not in line with the findings of the assessment conducted in this chapter, nor with those reported in other work. However, a few things are worth highlighting:

1. The L_{eq} and L_{eq}(RLB) perform on par with one another. This suggests that very few segments (if any) had spectra concentrated below 100 Hz, since this is the region where the RLB filter is most effective.

2. If the previous statement is true, then the modified psychoacoustic model might not have suffered from issues at low frequencies as discovered here.

3. Seefeldt et al. (2004) used a long-term average EP to speed up the loudness normalisation process, and thus determined global loudness using a single representative auditory pattern. This means that temporal information is averaged out. In this chapter, peak loudness was found to be superior compared to the average loudness. Again, this indicates that the stimuli were probably easier for the algorithms to measure, being comparatively more broadband and stationary than the musical sounds employed here.

4. The authors refined the SL transformation, indicating that the predictive quality of their psychoacoustic model was most likely superior to the GM02 and CF02 assessed in this study.

Taken together, the model of Seefeldt et al. (2004) would probably not perform consistently at this standard when applied to single instruments, despite the impressive performance measured using the expanded dataset. Note that in a subsequent paper, Seefeldt and Lyman (2006) compared various implementations of both Dolby's modified psychoacoustic model and the ITU-R BS.1770 (2015), using multichannel audio. Their results showed the average performance of both algorithms to be practically indistinguishable, but only five mastered items were used, the spectra of which were relatively broadband. A related investigation was recently carried out by Francombe et al. (2015) who collected loudness-matches of eight different reproduction methods, e.g. headphones and 5-channel surround sound, using seven programme items, in order to assess seven models of loudness. A fixed-reference loudness-matching design was used where the 5-channel surround served as the reference spatial reproduction. In addition to RMS based measures and variants of the ITU-R BS.1770 (2012) algorithm, the multiband procedures of Fastl and Zwicker (2007) and Glasberg and Moore (2002) were assessed in terms of their ability to match the loudness of the *same* programme item under different reproductions. The RMSE* was 0.39, 0.66 and 0.69 dB for the GM02, Zwicker model and ITU-R BS.1770 (2012), respectively. These were considered the better performing models, though only the mean-channel RMS approach led to very large discrepancies (RMSE* of 6.3 dB). Correlations between the subjective and predicted matches were generally very high (>0.85) which can be explained by the large spread in level adjustments required to loudness normalise the different reproductions. It was concluded that GM02 performed the best overall, but that most of the published loudness models were also consistent with the subjective data. Considering the low magnitude of these errors, and the contrasting ranking of the models with those presented in this chapter, these findings indicate that established objective models of loudness are most consistent with human listeners when the input stimuli share similar spectro-temporal qualities.

Finally, Moore et al. (2003) measured the effects of dynamic range compression on the perception of speech, using their model—the GM02—to predict the results. In short, they were able to predict the empirical data by using the average LTL. They found a shorter LTL release time of

200 ms (used in this chapter) to improve the fit over the 2 s time constant originally proposed (Glasberg and Moore 2002). Their results showed that compression of speech increases loudness for a fixed RMS level and that this effect increases with compression ratio. Furthermore, the level differences needed to match the loudness of uncompressed and compressed speech was found to increase with listening level by as much as 3 dB. It is not clear whether single-band loudness models, such as the ITU-R BS.1770 (2015), would be able to reproduce their main effect of compression, since this strongly depends on the temporal smoothing applied to IL calculation. Current single-band models, and indeed the LML algorithm presented here, would not, however, be able to reproduce the level-dependent effects of dynamic compression, as this requires a nonlinear input-output function.

Multitrack audio

Driven by the fact that state-of-the-art loudness algorithms were either founded on empirical psychoacoustic studies involving laboratory sounds, or developed for the purpose of normalising the loudness of full-range mastered audio content, a few researchers have applied loudness models to multitrack audio. For example, Pestana (Pestana and Alvaro 2012; Pestana et al. 2013) examined the accuracy of the ITU-R BS.1770 (2015) algorithm (equivalent to the EBU programme loudness assessed here) when applied to recordings of individual instruments such as drums, guitars, bass, piano and vocals. No error metrics were reported so a quantitative comparison to the results obtained here was not possible. Instead, Pestana looked at the deviations between the predicted and subjective level adjustment of a stimulus with respect to a reference stimulus, being either a kick drum or a vocal. Differences between the model prediction and the group average was typically between -6 and 3 dB, and the author concluded ‘We have seen that the evaluation of individual tracks in a multitrack context is a much more specific task [than broadband mixes], where the algorithm fails to agree with human subjects’ (Pestana et al. 2013, p.8). Pestana found a medium-sized negative correlation between log centroid and signed prediction error, suggesting that the algorithm underestimates the perceived loudness of stimuli centred at high frequencies. After experimenting with various parameter settings, Pestana (2013) proposed a reduction of the window size used to compute the ML from 400 ms to 280 ms, and an increase in the pre-filters highshelf gain from 4 dB to 10 dB, i.e. a 280 ms/+10 dB combination. An informal plot of signed errors (taken from all datasets used in this chapter) versus log stimulus centroid did not reveal this trend, although further research involving more high-frequency content is warranted to give proper insight. Note, however, that the pre-filter was bypassed in the study conducted here and replaced with the transfer function of the headphones.[13] This response, shown in Figure 7.3, is characterised by a boost between 1 and 10 kHz, with a notably large gain of 10 dB in the 6–7 kHz region. In contrast, the highshelf forming the first stage of the K-filter adds only 4 dB above 1 kHz, which might explain why the findings here are not entirely in agreement with those reported by Pestana.

To investigate this further, the four evaluations conducted here were re-run using the EBU model with a 280 ms window size, but any small improvements were not consistent across the four datasets. However, when using the 280 ms/+10 dB refinement suggest by Pestana (2013)—bypassing the headphone transmission—a few notable improvement were observed. The results are summarised in Table 7.16, which includes the original ITU-R specification of a +4 dB highshelf at 1.6 kHz. In general, the modification suggested by Pestana et al. (2013), outperforms the standard algorithm in all but the final dataset, with the greatest improvement seen on dataset

[13] The RLB filter was still active.

Model	RMSE				MSD			
	A	B	C	D	A	B	C	D
EBU (400/DT990/M)	2.94	2.76	**2.75**	1.90	1.56	1.61	**1.23**	2.58
EBU (400/DT990/P86)	2.79	2.57	2.96	**1.32**	1.52	1.23	1.40	**2.19**
EBU (400/+4/M)	2.54	2.67	3.28	1.94	1.47	1.33	1.54	2.57
EBU (400/+4/P86)	2.66	2.49	3.47	1.49	3.47	1.07	1.69	2.42
EBU (280/+10/M)	**2.34**	**1.12**	3.12	2.62	1.23	0.83	1.59	3.92
EBU (280/+10/P86)	2.58	1.15	3.56	2.21	**1.23**	**0.55**	1.81	3.33

Table 7.16: Comparison of different variants of the EBU loudness model. The configuration is given in brackets as: window size (ms) / highshelf gain (dB) or Beyer Dynamic DT990 headphones / global-loudness descriptor. Bold red font denotes the minimum of each column.

B. Presumably the consistent 10 dB boot above 3.5 kHz captured the influence of high-frequency instruments on loudness in the first two datasets, such as the shaker and drum elements. Further informal listening tests suggest that the loudness algorithms assessed here tend to underestimate the relative loudness of very high-frequency stimuli, and so boosting frequencies above 5 kHz seems to reduce this discrepancy. This likely explains the improved performance on dataset A which included more percussive content. Notice that the +4 dB booth was not optimal in any dataset for either statistic, although the errors are comparable to the headphone model. In short, these results do not provide conclusive evidence to support an exact refinement of the algorithm, but instruments with concentrated high-frequency content do appear to be problematic for current loudness models. This effect may be attributed to psychoacoustic sharpness rather than the assumed transmission characteristic the outer and middle ear.

Terrell (2013, Chapter.7) evaluated and extended the GM02 to better estimate subjective loudness ratios of paired musical sounds presented in both solo and simultaneous (masking) conditions. A variant of the GM02 which accounts for masking (Glasberg and Moore 2005) was employed, but the fundamental architecture is the same as the original model—exactly so in the case of solo conditions (refer to Section 3.3.5 for more details). Empirical data were obtained by asking listeners to judge loudness ratios between paired sounds, presented in quiet or mixed with other sounds ('in the mix'). Four segments of duration 4 s were used: voice, piano, hand drum and double bass. With the subjective ratio data collected, the GM02 was used to predict the data using the peak STL and the mean LTL. Terrell proposed the dynamic sound-stream bias (DSSB), a global-loudness descriptor based on a linear combination of peak and mean loudness by means of a signal-dependent weight. Terrell's modelling results showed peak STL to work better than mean LTL, in agreement with this work, and signal-specific weighting to outperform both approaches. The optimal coefficient of each sound was found to correlate with signal crest-factor, such that greater weight was assigned to peak loudness for transient signals. Unfortunately, because only four sounds were used in his evaluation, the generalisability of the DSSB is questionable. Furthermore, the author of this thesis initially attempted to fit the DSSB to the subjective matching data, but no relationship of sufficient magnitude was found between crest-factor and the sound-stream bias coefficient. This means that although the DSSB would likely outperform more traditional measures of global loudness on the training dataset, it is not clear how the descriptor can be applied to others, at least without more extensive research.

Similar to Terrell, Ma et al. (2014) also assessed the ability of the GM02 to predict the loudness of musical instruments when mixed with others, i.e. predict partial loudness. 16 sound segments of length 10 s were taken from four different multitrack projects and used for the listening assessment. 12 listeners matched the loudness of each segment in isolation with the loudness of the same segment presented in the mix. The authors used the mean LTL to estimate the overall loudness, and found a good fit to the data (level differences within 2 dB) after reducing sensitivity of the masking calculation. Note that, in this case, the mean LTL descriptor worked well because subjects were listening for changes in loudness on the same sound source. Thus, whilst the mean LTL would fail to predict the relative loudness of, for example, a bass guitar and a snare drum, the accuracy of this descriptor would be less of a concern when estimating the loudness of the *same segment* under different masking conditions. This resonates with the earlier discussion of the findings presented by Francombe et al. (2015) and Moore et al. (2003), two studies which assessed loudness models using the same two sources but with different processing applied.

Spectral loudness summation

Fastl et al. (2009) concluded that the DIN 45631 (1991) and the ANSI S3.4 (2007) stationary loudness models capture the well-established loudness growth function of a 1 kHz pure tone. The authors found, however, that the ANSI S3.4 estimated the loudness of pink noise to be around 5 dB higher than the estimates of the DIN 45631. As covered in Chapter 3, the GM02 is the dynamic realisation of ANSI S3.4, whereas the CF02 is a dynamic model derived from DIN 45631. The spectral stages of these two models are based on different principles: the ANSI S3.4 uses narrower level-dependent auditory filters based on the Cam frequency scale; the DIN 45631 uses auditory filters positioned on the Bark scale and an alternative model of spectral masking. Because the critical-bands defined by ANSI S3.4 are narrower in bandwidth, it requires more bands (about 41 vs 24) to cover the audible range. As a result, this model predicts *greater* SLS compared to DIN 45631. Schlittenlacher et al. (2012b) also observed strong differences in SL around 3 kHz between the two standards, and suggested that ANSI S3.4 overestimates the amount of SL compared to subjective results.

To obtain ground truth data for assessing the standardised models when predicting SLS, Schlittenlacher et al. (2011) run additional loudness-matching experiments involving pink noise and a 1 kHz pure tone. Their results showed that the predictions of the DIN 45631 fell within the IQR of the subjective data over 30–70 dB SPL. In contrast, the ANSI S3.4 overestimated the loudness of the pink noise relative to the subjective matches, although the predictions showed a good correspondence to the trend of the matches as a function of level. The two models were then evaluated on loudness-matching data involving eight technical sounds, such as those generated by a notebook fan and a hair dryer. Both models overestimated the loudness of the stimuli (relative to the 1 kHz tone) with the miscalculation being most severe for ANSI S3.4.

Further tests were conducted by Schlittenlacher and Ellermeier (2015) to investigate differences in SLS between the two standards. This time they used bandpass filtered pink noise with spectra concentrated in three different frequency bands: 0.125–1 kHz, 0.5–2 kHz and 1.25–5 kHz. Two experimental procedures were used: the MOA using a 1 kHz pure tone as the reference stimulus, and magnitude estimation via cross-modality matching with line length. For the lowest frequency band, the predictions of the two models were comparable , although the DIN 45631 captured the subjective data slightly better. For the 0.5–2 kHz noise band, both standards predicted too high a loudness compared to the empirical data, although the DIN 45631 showed better agreement with the data at lower levels. The data from both MOA and magnitude estimation procedures revealed the worst in the standards' predictive quality for the highest frequency band, where the overestimation in loudness of pink noise was most pronounced. Again, the DIN 45631 performed slightly better than ANSI S3.4. The authors concluded that SLS appears to be significantly lower in human subjects compared to current models. Note that the subjective loudness matches from this study were used in Chapter 3 (see Figure 3.19) to compare the predictions of these two standards in addition to the stationary loudness model of Chen et al. (2011)—the stationary counterpart to the CH12. The simulations run in that chapter confirmed the observations of Schlittenlacher and Ellermeier (2015), and it was found that the predictions of the CH12 were curvilinear, deviating from the subjective data most at mid sound levels, similarly to ANSI S3.4. The CH12 showed a notable decrease in SLS as the level of the tone increased beyond about 70 dB SPL. Consequently, this model predicts less SLS than the two standardised procedures at higher sound levels, particularly at high frequencies.

Meunier et al. (2000) compared the predictions of variants of Zwicker's (similar to DIN 45631) and Moore's (similar ANSI S3.4) stationary loudness models with subjective data derived from loudness-matching experiments involving bands of noise and 24 relatively steady environmental sounds. A 90 Hz wide noise centred at 1 kHz was used as the reference stimulus. They reported that the predictions of the two algorithms generally lay within the variability of the data, but Zwicker's model performed best for the different noise bands. Most notable was that Moore's algorithm overestimated the loudness of the band limited noise centred at 3 kHz region, setting the matching level below that of the listeners. Zwicker's model also performed better for the technical stimuli, with the predictions of Moore's method generally falling within the upper quartile of matching data associated with each sound. The authors concluded that Zwicker's method underestimates the loudness of sounds below 70 phons, whereas Moore's method overestimates the loudness of sounds above 70 phons.

As discussed, Rennies et al. (2013) evaluated the CF02 and GM02 using speech-like signals and technical sounds. They found that when a narrowband signal (50 Hz wide noise centred at 1 kHz) is compared to a broadband signal (noise from a jet engine), these two algorithms significantly underestimated the loudness of the former compared to the subjective matches. In particular, the CF02 predicted that the narrowband noise had to be about 9 dB higher than the equal-loudness gain submitted by the subjects, whereas the GM02 was around 14 dB higher. In another study, Rennies and Verhey (2009) reported that these two models behaved most differently for tones modulated at high frequencies, which was attributed to the reduced SLS in the model of CF02.

The findings from the above studies indicate that the GM02/ANSI S3.4 and the CF02/DIN 45631 do not correctly predict the amount of SLS introduced by broadband sounds, and that this miscalculation is frequency dependent. The discrepancy appears to be greatest for the GM02, especially above 1 kHz and at higher presentation levels. Although the experiments in this study did not directly measure subjective SLS, the stimulus prediction errors indicated that the multiband algorithms tended to over-boost the level of narrowband low-frequency instruments to match the loudness of broader, higher-frequency sounds.

The consistently good performance of the LML suggests that current EP based loudness models, especially the GM02, can be improved if the amount of SLS is reduced. Interestingly, further tests have shown that an implementation of the LML for stationary sounds only was found to capture the data of Schlittenlacher and Ellermeier (2015) remarkably well for a linear model, providing additional evidence in support of reducing SLS. The fact that the CH12 always ranked higher than both the GM02 and CF02 in the four assessments conducted in this chapter further supports this theory, since this model generally predicts lower level differences between narrowband and broadband sounds at higher sound levels. In this respect, it is not clear how the multiband models would rank relative to one another had the listening experiments been conducted at lower listening levels. It is expected that the behaviour of the CH12 would operate in a more compressive region and thus approximation the behaviour of the GM02, considering that their peripheral filtering stages are derived from similar psychophysical methods.

An important question is whether the degree of compression, and thus SLS, should vary with spectral bandwidth (and frequency), such that loudness growth functions of pure tones can still be correctly predicted. For example, although the LML employs a larger compressive exponent than the one used by GM02, thus better predicting the relative loudness between narrowband and broadband sounds, it does not estimate the well-known doubling of loudness per 10 dB increase in sound level for a 1 kHz pure tone. Indeed, Seefeldt et al. (2004, p. 4) found their psychoacoustic model, which uses the same auditory filter bank as the GM02, to possess similar problems as found in this work when applied to programme audio:

> ‘However, on the expanded development database, the model performs poorly for several signals. Most of these signals possess spectra which might be classified as narrowband in comparison to the majority of the signals in the database.’

The authors proposed that two SL functions, one suitable for narrowband sounds and another for wideband sounds, could be interpolated between based on the predicted bandwidth of the input. Alternative methods for reducing the calculated loudness of broadband sounds relative to the loudness of narrowband sounds include reducing the number of critical bands by broadening the ERB, expanding the upper slopes of tonal excitation patterns, or transforming the perceptual-frequency scale *after* estimating the SL pattern (Rennies. J. et al. 2009; Schlittenlacher et al. 2012a; Schlittenlacher et al. 2012b). As noted by Schlittenlacher et al. (2012b), this latter approach may reflect higher stages of auditory processing.

Global-loudness descriptors

In general, there is less ambiguity as to how global loudness should be calculated when using a single-band estimator. Both LARM and ITU-R BS.1770 (2015), for example, were originally developed to give a single reading corresponding to the overall loudness judgement of broadcast material, rather than generating a loudness-time function. In contrast, the dynamic multiband models were designed to predict the temporal integration of loudness as measured using tone bursts rather than real audio content. As such, it is less certain as to how the overall-loudness impression of real dynamic sounds should be quantified given the time series output by these models.

The evaluations presented here generally confirm previous suggestions (Zwicker 1977; Fastl and Zwicker 2007) that the peak loudness—or values representative of loudness maxima—is preferable to a simple average for fluctuating sounds. This is especially the case when the decay time of the temporal integrator is comparable to the attack time, since the loudness-time function does not accentuate peaks, but instead fluctuates markedly over time, particularly for transient signals. In this case, the loudness of dynamic material is underestimated.

As previously mentioned, Glasberg and Moore (2002) suggest using the mean LTL for fluctuating sounds, which was found to work well in their study investigating the loudness of compressed speech. In contrast, this feature did not perform well on the sound material used in this study, which is consistent with the work of Terrell (2013, Chapter.7). Rennies et al. (2010) found the LTL generated by the GM02 to be too slow to describe the dependence of loudness of repeated noise bursts on repetition rate, and that peak STL is superior. However, in a subsequent paper, Rennies et al. (2013) concluded from their loudness-matching data that the loudness of speech is primarily determined from its long-term spectrum, and that temporal modulations have little effect on loudness. As such, the authors found the mean LTL to be the best predictor when modelling their results. Interestingly, they also found the mean LTL to outperform the peak STL when predicting the relative loudness of a snare drum and of a machine gun to a broadband noise reference, contradicting the findings reported here. It should be noted, however, that their implementation of the GM02 was based on a 2 s release time, rather than the revised 200 ms (Moore et al. 2003) as used in this work. Faster release times de-emphasise loudness maxima, and hence why the average loudness did not perform well here. This is supported by the smaller difference between the equal-loudness gains obtained using the CH12 with peak and mean loudness, compared to the differences observed when using the GM02 and CF02. In contrast to the other models, the CH12 makes use of the longer 2 s decay time.

Hots et al. (2013) found the mean LTL of the GM02 to be superior to the peak STL of both the GM02 and CF02 when predicting level differences between a 1.5 kHz tone and a bandpass filtered noise centred at the same frequency. Again, the assumption is that the long-term spectrum is the primary source of loudness perception, and that temporal fluctuations have minimal contribution. In a recent paper, Zorilă et al. (2016) compared the ability of the GM02 to equate the loudness of speech subjected to different dynamic range processing. They compared the mean and peak LTL (not STL) descriptors with the model configured with a 2 s and 200 ms release time. The authors concluded that the peak LTL was optimal for both configurations, with slightly better performance using the original 2 s release time. The authors also showed that dynamic aspects of the stimuli did influence the perceived loudness, and thus stationary models did not work as well.

Although the results are not conclusive, the following points are made:

1. For short-duration sounds, such as a tone burst or single transient, the maximum of the STL should be used.

2. A release-time constant of 2 s is preferred for generating the LTL.

3. Due to the longer warm-up time, the mean LTL is dependent on stimulus duration, and as a result, a simple average may not be representative for segments of less than a few seconds in duration. For this reason, segments involving multiple transients presented at low repetition rates may not be accurately represented by peak LTL.

4. The peak LTL appears to be both more robust and representative of overall loudness compared to the average value.

5. A percentile measure of the STL with a (modified) longer release time may avoid the need for two loudness-time functions altogether, the choice of which is currently stimulus specific. Furthermore, a lower than maximum value may be more robust across different types of sound material.

In the next section, a new, robust approach to estimating the loudness of an arbitrary signal using any loudness estimation algorithm is presented based on the findings of this study, although the key parameters are easily adapted in light of new data.

7.6 Global-loudness estimation of multitrack audio

The audio used to form the four datasets in this study are all of relatively short duration, which is important for obtaining reliable subjective data. For these signals, the overall loudness impression can be estimated by applying statistical functions to the loudness-time series extracted from the sound in question. In this chapter, high percentiles of the loudness-time function were found to work best for these short segments, especially for the multiband procedures where peak loudness was optimal. These measures may not, however, be appropriate when applied to sounds that are less dynamically stable over time, such as unprocessed multitrack recordings, where sudden fluctuations in amplitude or indeed silence can strongly influence the objective measurement.

In order to determine unbiased and robust feature statistics from audio signals of long duration, periods of silence and background noise must be excluded as best as possible from the measurement. The ITU-R algorithm achieves this using an adaptive gate to remove silent and soft periods from the time integration. An alternative offline approach, proposed here, that is applicable to any loudness-time function, is based on the dynamic gate introduced in Chapter 5. A Schmitt trigger, defined by Equation 5.8, is employed to segment the recording into regions of musical activity. The input to the dual comparator is a power envelope, obtained by squaring the input signal and smoothing it using a bi-directional one-pole filter with a 20 ms time constant. This is further smoothed by a second filter with a 1 s time constant to obtain two thresholds; an upper and lower dynamic threshold is determined by scaling the second envelope by -6 dB and -16 dB, respectively. This gives a 10 dB window to reduce fluctuations in gate state. The standard gating protocol is then used to get the onset and offset times of each region. These parameter values were found to work well on a variety of instrument recordings with little noise contamination, but are easily adapted. An example of the segmentation procedure applied to a snare drum and a lead vocal is shown in Figure 7.18.

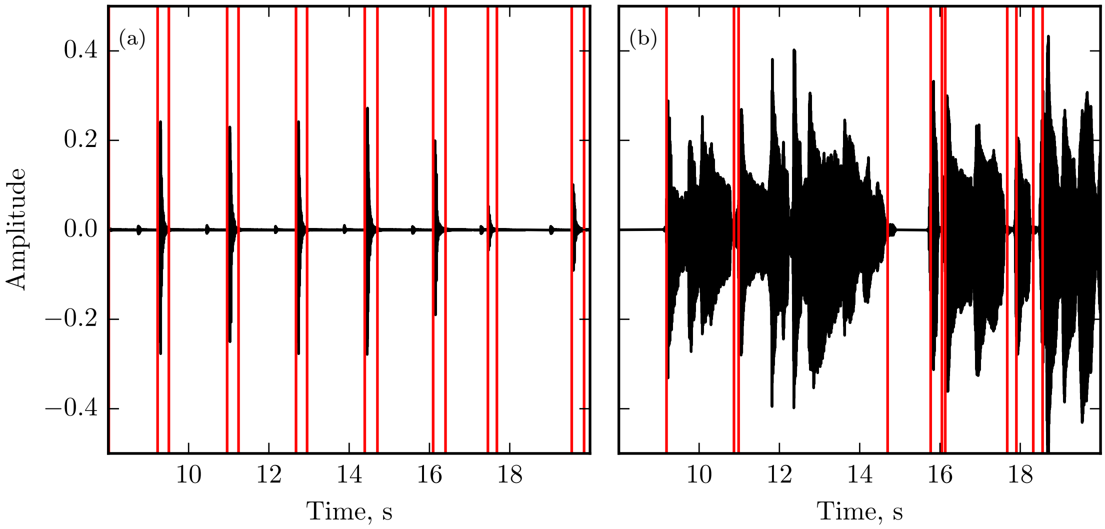


Figure 7.18: Example of the segmentation algorithm applied to (a) snare drum and (b) lead vocal.

After running the segmentation procedure and extracting the loudness-time series, analysis frames coinciding with inactive regions are discarded. Knowing that the louder portions of a sound are perceptually more salient than softer regions, all frames falling below the median loudness are also discarded. The characteristic of the resulting distribution is rich with information. In particular, the following observations were identified when analysing the STL distributions of different multitrack recordings extracted using the GM02 and CH12 models:

- The presence of outliers in the upper tail of the distribution of gated loudness values indicates the need for dynamic range compression to control for rapid increases in loudness that may not be an intentional or desired part of the performance.

- A heavily positively skewed distribution suggests the need for level automation and possibly dynamic range compression. In many cases this may simply reflect a dynamic performance rather than instability. In any case, the global-loudness estimate fluctuates drastically over time and is therefore relatively uncertain.

To arrive at the global-loudness prediction, it is here suggested that the 90th percentile of the *gated* loudness values is used. This generally gives similar (to within 0.5 phon) estimates as the 95th percentile of the raw (ungated) time series of stable recordings. For models such as LARM or the EBU, the same procedure can be applied to the fastest changing time series, e.g. the ML in the case of the EBU.

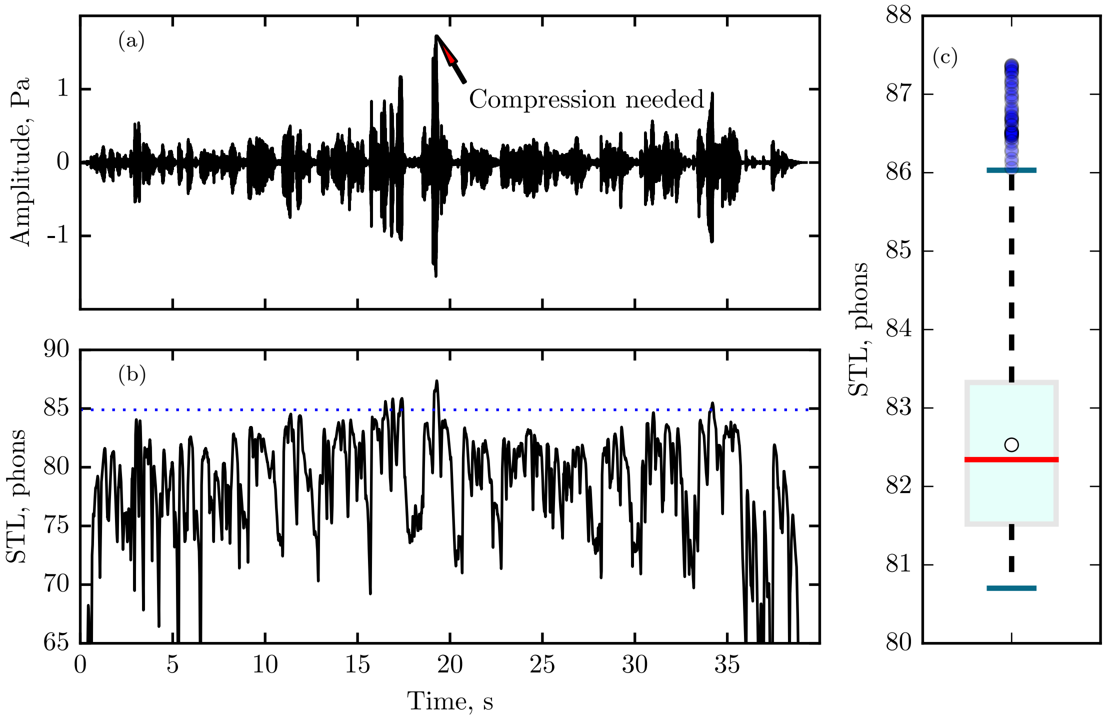


Figure 7.19: (a) Temporal waveform, (b) CH12 STL and (c) boxplot of gated loudness levels extracted from a recording of a female vocalist. The horizontal line in the subplot (b) denotes the estimated global-loudness level.

The above procedure is demonstrated in Figure 7.19, which shows (a) the temporal waveform of a female singing voice, and (b) the STL as measured using the CH12. It can be seen that the maximum loudness at time 19 s (annotated in subplot (a)) would not be representative of this recording, which is introduced as the vocalist momentarily sings closely into the microphone. The proposed method overcomes this by taking the 90th percentile of the gated values, where gating is necessary to remove background noise. Subplot (c) is a boxplot of the gated loudness levels extracted from this recording. It can be seen that the distribution is positively skewed with many outliers in the upper tail, which is attributed to the increase in level around time 15–20 s.

The procedure can be extended to establish a robust moving estimate of global loudness, a recommendation made in the concluding remarks of Chapter 5 for the purpose of real-time loudness metering. The implementation is based on a moving 3 s window, populated on every sample determined relevant by the adaptive gate.[14] When a new estimate is needed, e.g. every R samples, the medium value in the buffer is used as a threshold for the current population, from which the 90th percentile is taken. The resulting time series is a running estimate of the loudness impression

[14] For real-time processing, the gate is based on causal filtering.

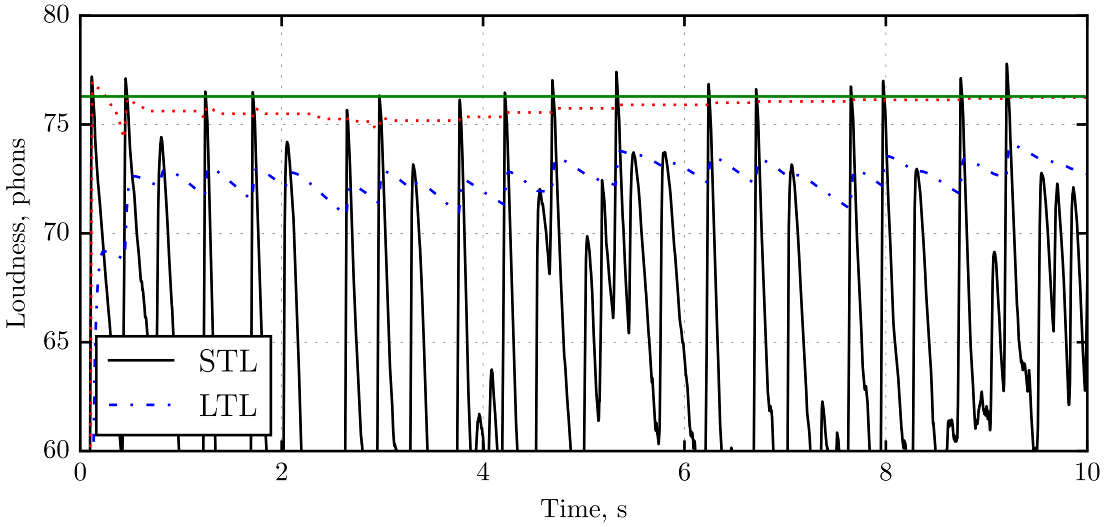


Figure 7.20: STL- and LTL-time functions, generated by the CH12, of a kick drum recording. The green horizontal line denotes the estimated global-loudness level using the proposed noncausal measurement, and the red dotted line shows the time-varying loudness impression using the online percentile tracking method.

at the macro-time scale. Figure 7.20 illustrates this by showing the STL, LTL, global and time-varying loudness impression of a kick drum excerpt taken from a rock multitrack recording. It can be seen that the time-varying percentile (dotted red line) gives a good correspondence to the robust estimate of global loudness (green horizontal line), both of which emphasise the peaks in the STL. In addition to the aforementioned measures of skew and relative frequency of outliers of the gated loudness distribution, the variation in the moving-percentile measure is also useful for describing the uncertainty in the estimate of global loudness. Notice that the running percentile captures the peak of the first transient, despite the 3 s buffer being only partially filled. This is due to the medium thresholding which excludes very quiet frames. In contrast, the 100 ms attack time used for LTL integrator means that the resulting time function takes slightly longer time to 'warm-up', and loudness peaks are less prominent compared to the percentile readings. Because the percentile tracking procedure is informed by the adaptive signal gate, and makes use of a relatively long analysis window, it is robust to sudden deviations in loudness and maintains representative readings after periods of silence. Such properties make this measurement extremely useful for live metering and related applications.

7.7 Summary

Loudness-matching data from a number of different listening tasks involving musical sounds and audio engineers were used to assess the predictive quality of state-of-the-art loudness models. The evaluation framework was designed to respect the functionality of the different implementations, yet bring them closer to the experimental conditions experienced by the listeners such that a fair and systematic comparison could be made. Both nonlinear EP loudness models, founded on empirical ground-up psychophysics, and heuristic single-band devices targeting broadcasting, were included in the assessment. The models were assessed in terms of their ability to loudness normalise musical sounds commonly experienced in music production. Different statistical analyses were carried out to compare the algorithms against the subjective data. The main findings are summarised below.

1. Compared to the multiband loudness models, single-band models require considerably less processing power and manual configuration to align the loudness of a given set of sounds.

2. None of the algorithms were as accurate as an average group of listeners, in the sense that predictions often fell outside of the confidence limits of the data. Furthermore, model predictions were often classified as outliers when considering the model as a subject of the listening assessment.

3. Mean loudness was found to be problematic for the GM02 and CF02. This descriptor is less of a concern when the time-smoothing filter is highly asymmetrical, accentuating loudness maximum over time. As such, peak STL drastically reduced large discrepancies.

4. These two models generally place low-frequency instrumentation at higher reproduction levels than necessary, although further research is required to gauge average performance. Such behaviour appears to be primarily attributable to excessive spectral loudness summation (SLS), as well as the parameters that determine frequency sensitivity, such as the assumed middle-ear transmission filter.

5. Of the three published multiband methods, the most recent algorithm, the CH12, ranked first in each of the four assessments. Worst-case error was also less severe for this model.

6. The LARM and EBU models performed relatively well overall, and should be preferred over more complex multiband procedures until further research (using a broad selection of multitrack material matched by experienced listeners) can justify their added complexity.

7. The EBU model performed consistently on a range of different material and can be considered conservative in terms of gain correction applied to RMS level aligned stimuli. The experiments indicate that further refinements to the pre-filter and possible incorporation of spectral information may improve its predictive quality.

8. The L_{eq}, with no frequency weighting, did not behave consistently on single instruments, and may only be useful for aligning the loudness of broadband sounds, such as complete music mixes.

9. The new linear multiband loudness (LML) model gave the best performance overall, but its parameters should be refined using a larger corpus of sounds, which may—from a top-down perspective—give insight into how current psychoacoustic models should be adapted to improve prediction accuracy of everyday sounds.

10. The assessments carried out here indicate that the modelling of SLS does offer an improvement over the single-band approach when measuring the loudness of musical sounds.

11. Supervised algorithms that take into account higher levels of perception such as context, instrument type and perceptual annoyance are most likely needed, but more empirical data is needed to establish rich training information. The LML could be extended towards this goal, given that both temporal and spectral information is extracted efficiently from the input source.

This chapter concluded by presenting a new method for estimating the global loudness of an arbitrary stimulus, which was then adapted for real-time loudness metering. The only requirement is that the selected loudness model extracts a loudness-time function that reflects the momentary loudness as perceived by the listener.

Chapter 8

Conclusions

The objective measurement of sensory magnitude has been the subject of scientific research for over a century, demonstrating great attention from scholars and engineers alike. The work presented in this thesis contributes to this extensive body of scientific literature through investigating current approaches to loudness estimation and their practical application in audio engineering. In particular, the research direction has steered towards applications in music production, assessing the behaviour of these perceptual models when applied to the types of sounds commonly experienced by musicians and music producers. This is important because, as computational power increases, developers of Intelligent Music Production (IMP) systems are turning towards more complex auditory approaches to perceptual measurement, under the assumption that they generalise beyond the laboratory to real musical sounds. For completeness, the primary objectives of this thesis are reintroduced:

1. To evaluate the performance of state-of-the-art loudness models when applied to musical sounds.
2. To develop flexible, computationally efficient auditory models for calculating loudness.
3. To establish new validated procedures for measuring the loudness of single instruments.
4. To develop a real-time mixing framework that accommodates both linear and nonlinear loudness estimators to perform perceptually-motivated music mixing.

The following section reviews the work conducted in each chapter of this thesis and how it relates to the above objectives. The main findings are then summarised, specific contributions are stated, and recommendations for further research are given.

8.1 Summary of research

This section summarises the content of each chapter and highlights the associations to thesis objectives. Key findings from the thesis as a whole are then presented.

8.1.1 Chapter 2: The Hearing System

An overview of the physiology and psychology of hearing was conducted in order to establish the necessary background for all subsequent work in this thesis, especially when discussing mathematical approaches to loudness calculation. The peripheral ear comprises three components: the outer, middle and inner ear. The first two determine the transmission of sound from outside of the head to the cochlea, the frequency response of which is bandpass in shape. Sound at the oval window, the entrance to the inner ear, is then decomposed into different frequency bands, the bandwidths of which increase with centre frequency and become progressively asymmetrical with increasing sound level. This complex decomposition of sound is a result of the active mechanical filtering taking place within the cochlea.

Key findings from a number of psychophysical experiments were summarised and it was shown how perceptual responses to sounds are correlated with physiological measurements. Attention was paid to simultaneous masking, where the concept of the critical band was introduced to better understand current psychophysical theories of frequency selectivity and the internal tonotopic coding of sound.

Loudness perception was shown to depend on a number of variables, including stimulus level, duration, frequency and bandwidth. It was found that a complete model of loudness must simulate the basic processing performed by the peripheral hearing system. For example, an auditory filter bank is necessary to predict the observation that for sounds of constant intensity, loudness increases if the bandwidth exceeds that of the ear's critical bands. The phenomenon is known as spectral loudness summation (SLS) and is believed to be attributed to the spatial compressive encoding of sound.

8.1.2 Chapter 3: Loudness Models

A historical overview of loudness models was presented, with a particular focus on psychoacoustic models founded on the concept of the critical band and the modelling of internal representations of sound. Loudness models applicable to both stationary and time-varying sounds were introduced and detailed implementation-level documentation of the current state was presented. This work was necessary to identify key parameters of the different algorithms and also establish a deeper understanding of their architecture and computational resources—a prerequisite for objective 2.

A comparison of three excitation pattern (EP) based loudness models was conducted to give insight into the predictive quality and behaviour of more loudness calculation procedures. Two of the models, ANSI S3.4 (2007) and DIN 45631 (1991), are standardised procedures, and the third (Chen et al. 2011) is a relatively recent approach that stems from more recent psychophysical measures of frequency selectivity. This comparative study was necessary to establish a foundation for objective 1. For example, how should these level-dependent models be configured to simulate psychoacoustic experiments? How well do they perform on the stimuli they were designed to predict subjective attributes of? Do these models share common characteristics when estimating the loudness of stationary sounds?

In combination with observations reported in the literature, the study lead to the following conclusions:

- A model of peripheral auditory filtering is necessary to describe a range of loudness phenomena as measured in the laboratory, notably frequency dependencies of loudness and SLS.
- All three models show a tendency to overestimate the loudness of broadband sounds, especially those with spectra concentrated above 1 kHz.
- Unlike ANSI S3.4 and the model of Chen et al. (2011), the DIN 45631 fails to adequately capture the revised equal-loudness contours (ISO 226 2003).
- In contrast, the Chen et al. (2011) model provides the best overall account of the frequency sensitivity of the ear.

- Modifications to the frequency-dependent critical-band compression used by ANSI S3.4 can improve the fit to the subjective equal-loudness contours at higher frequencies.

After comparing the predictions of the three procedures, the chapter concluded by investigating different approaches to the modelling of nonsimultaneous SLS, an important yet under-researched area of loudness perception, particular for fluctuating sounds. It was shown how current dynamic models of loudness (Glasberg and Moore 2002) can be extended to better describe the dependence of SLS on the temporal structure of sound by incorporating temporal smoothing in the excitation- and specific loudness (SL)-time patterns. Results were compared with the predictions of existing methods incorporating a temporal masking stage (Chalupper and Fastl 2002), which themselves were only indicative of trend in the data. The influence of certain signal processing techniques, such as multi-resolution time-frequency decomposition, on model predictions were also highlighted.

8.1.3 Chapter 4: Efficient Loudness Estimation

This chapter aimed to establish efficient methods for computing auditory patterns and estimating the loudness of time-varying sounds. This work therefore directly targets objective 2, and was a fundamental step towards incorporating psychoacoustic hearing models into practical, real-world audio engineering applications (objective 4). The dynamic EP based loudness models of Glasberg and Moore (2002) and Chen and Hu (2012) were profiled, analysed, optimised and extended to realise a real-time binaural loudness meter. In both cases, it was shown that the computational workhorse corresponded to the transformation from power spectrum to EP.

A number of signal-processing techniques were incorporated into the models, primarily to target the initial bottleneck, which were then treated as approximations to the published procedures. This allowed for errors in both the measured loudness and the intermediate auditory features (excitation and SL patterns) to be measured. Musical signals and several error metrics were used to evaluate the accuracy of the approximated features. A novel aspect of this work was the measurement of error in the predicted *global loudness* of each stimulus, such that perceptually relevant effect sizes could be interpreted in the application domain.

It was shown that filter spacings greater than 0.1 Cams could be combined with a compressed spectrum to significantly reduce the complexity of the excitation transformation, whilst allowing for systematic control over the error introduced. For the first time, the hopping Goertzel discrete Fourier transform (HGDFT) proposed by Cassidy and Smith III (2007) was fully integrated into the model of Glasberg and Moore (2002) as an alternative to the parallel fast Fourier transform (FFT) approach to computing multi-resolution spectra. The primary advantage of this modification was a huge reduction in the number of frequency components (69 vs 960) required to estimate EPs. The effects of EP interpolation, hop size, frequency-domain outer/middle-ear weighting, and sound level were all quantified and statistically analysed to better understand model specific effects.

The derived parameterised versions of the models achieve extremely high speedup factors and are able to estimate the global loudness of musical sounds within tolerable error bounds. The effect of key parameters on the modelled auditory features was, however, found to depend strongly on the algorithm, resulting in a different set of recommendations for achieving fast, low-error feature extraction. An association between the normalised root mean square error (nRMSE) in the loudness-time function and discrepancies in predicted global loudness was made, which is useful for future investigations into the optimisation of dynamic loudness models.

8.1.4 Chapter 5: Loudness Driven Automatic Music Mixing

The development of efficient multiband loudness estimators was imperative if auditory models are to be embedded into music production applications where multiple concurrent tracks are to be processed in a small time window. This was demonstrated by developing a flexible real-time music mixing framework in which the levels of a multitrack recording were controlled by both single band and multiband loudness models (objective 4). The goal of this work was to identify if a more complete model of loudness perception grounded in auditory theory could be used to improve the subjective quality of automatically computer-generated music mixes.

To do this, five songs of different genres were automatically mixed by the developed system, configured using three measures of loudness: a 3 s RMS-level detector, a variant of the EBU Tech 3341 (2016) short-term loudness (STL), and the long-term loudness (LTL) generated by the real-time version of Glasberg and Moore's (2002) algorithm developed in Chapter 4. Using a variation of the multiple stimuli with hidden reference and anchor (MUSHRA) methodology (ITU BS.1534 2015), subjective assessments were conducted to compare the quality of the computer-generated mixes with those of four mix engineers. Based on the inherent ranking of the mixes performed by the listeners, it was demonstrated how nonparametric statistical methods can be employed as an alternative to traditional approaches to analysing the output of a MUSHRA experiment which are not always applicable.

The subjective evaluation showed that the model of Glasberg and Moore (2002) is inferior to simplistic measures of loudness when tasked with loudness-balancing musical instruments. More specifically, for songs involving a more diverse class of instruments covering a wide frequency range, equal track-loudness mixes generated using this algorithm were far from ideal, receiving bad-poor quality ratings that were comparable to those assigned to the anchor stimulus. In contrast, both RMS- and EBU-based mixes were judged similarly by the listeners, with the two approaches performing relatively well when the automatic mixes of five songs were compared with those of four human engineers.

An analysis of the relative-loudness balances of the mixes of ten mixing engineers was carried out to gain further insight into inter-model differences. Mix balance representations were found to be model specific, none of which supported the equal-track loudness hypothesis commonly assumed by current automatic mixers. The data suggested that the multiband loudness model of Glasberg and Moore (2002) underestimates the relative loudness of low-frequency instruments in a mix, and that this observation is not circumvented by modelling inter-track masking, which itself appears to introduce further complications within the context of a real music application.

The findings from this chapter are relevant for understanding the predictive behaviour of the state-of-the-art loudness algorithms when embedded in music applications, therefore providing an additional step towards accomplishing objective 1. In short, this study achieved two things:

1. A successfully working real-time mixing device that accommodates all current dynamic loudness models.
2. The subjective evaluation of an EP based loudness model embedded in a music-mixing application.

8.1.5 Chapter 6: Measuring the Relative Loudness of Musical Sounds

A number of procedures for measuring subjective loudness were reviewed and two loudness-matching experiments involving musical sounds were conducted. An indirect adaptive loudness-matching procedure and a modified version of the direct method of adjustment (MOA) were used in two independent experiments to obtain loudness matches of single instruments typically found in rock and blues music. The first experiment required listeners to compare the loudness of all pairs of four musical sounds, including same-stimulus pairs. The second experiment involved 16 musical sounds and utilised a balanced pair-matching design, the requirements of which were reviewed and reformulated to accommodate specific experimental properties. This experiment design minimises the bias and uncertainty inherent with the fixed-reference loudness-matching design, and does not require the listener to match the loudness of every stimulus pair.

Intra- and inter-subject variability was measured to quantify the uncertainty of audio engineers when making loudness judgements of single instruments. These measures indicated that specific segment-pairs were more easily matched than others, especially those involving the singing voice. In the second, larger experiment, it was shown that subjects became systematically faster when matching pairs of sounds over time.

It was demonstrated how linear regression models of varying sophistication can be used to extract the relative loudness levels, and thus the equal-loudness gains, of the sound segments. By doing so, the entire set of M loudness matches associated with N sound segments, with $M >> N$, are reduced to a set of N relative levels, thus reducing parameter variance. Furthermore, it was demonstrated how bias effects associated with the experimental procedure can be estimated.

The results showed that RMS level normalisation is inadequate for equalising the loudness of different musical sounds. Second, the regression models supported transitive relations in the listeners' loudness judgements, a necessary assumption for deriving the equal-loudness gains. This chapter therefore validated both the measurement procedure and statistical analysis required to establish ground truth data needed to evaluate current loudness models (objective 1) and subsequently establish predictors of global loudness of multitrack audio (objective 2).

8.1.6 Chapter 7: Evaluation of Loudness Models Applied to Musical Sounds

This chapter aimed to address objective 1 directly by evaluating the performance of five published loudness models and a baseline metric—the equivalent continuous sound level (L_{eq}). The subjective reference data was derived from four independent loudness-matching experiments involving musical sounds as stimuli and audio engineers as participants. The algorithms were assessed by their ability to loudness equalise each set of musical sounds. Different approaches to predicting the experimental data were reviewed and an appropriate methodology was developed. Several performance metrics were presented, two of which newly proposed, for measuring both absolute error and error relative to the between-subject variability.

The assessment of the published models concluded that:

1. Two single-band algorithms, ITU-R BS.1770 (2015) and LARM (Skovenborg and Nielsen 2004b), were generally superior to the multiband algorithms and performed consistently across the four datasets, with RMSEs ranging between 1 and 3 dB.

2. Of the multiband procedures tested, the most recent model proposed by Chen and Hu (2012) worked best, followed by the model of Chalupper and Fastl (2002) and then Glasberg and Moore (2002).

3. The procedures of Chalupper and Fastl (2002) and Glasberg and Moore (2002) revealed markedly large worst-case errors (8–15 dB), and, except for the smallest dataset, were outperformed by the simple L_{eq} metric.

4. The five algorithms were unable to predict the relative loudness of specific sounds, demonstrating the need for further refinements to current approaches to real-world loudness estimation.

A new linear multiband loudness (LML) model was proposed which combined the revised low-frequency B curve (RLB) filter, as used by the ITU-R BS.1770 (2015) algorithm, with a simplified EP model similar to that of Glasberg and Moore (2002). The transformation from excitation to SL was implemented according to a simple power-law function of critical-band intensity, based on a compressive exponent of 0.56. The effect is a reduction in the amount of SLS predicted compared to the three published multiband methods assessed, which was shown to yield considerably better performance, especially in terms of maximum error.

When deriving model predictions, two different approaches to estimating the global loudness (a single quantity) of a fluctuating sound was presented: one based on a form of temporal average, and another using a percentile of the loudness-time function. Both approaches worked similarly well for the two energy-based models, although further analysis revealed the choice of decision variable to have a greater influence over the equal-loudness gains predicted by LARM. Peak loudness was optimal for the three published multiband methods. In contrast, average loudness tended to significantly degrade the predictive quality of models employing a more symmetrical temporal integration stage with attack and release time constants below about 200 ms.

Finally, a new robust measure of the overall loudness of an arbitrary stimulus, both for offline and online processing, was presented. The algorithm is particularly suited to measuring the loudness of multitrack audio and has the advantage of working independently of the underlying loudness algorithm. Both the proposed LML model and the new global-loudness descriptor therefore address objective 3 of this thesis.

8.1.7 Key findings

The summaries of work above highlight a number of outcomes that are instrumental to achieving the main aim of this thesis:

1. Psychoacoustic excitation-pattern based models of loudness can be highly simplified without sacrificing measurement accuracy for many applications.

2. The optimised algorithms can be incorporated into IMP systems to optimise the music mixing objectives.

3. Equal-loudness mixing is not a realistic objective for automatic mixers, but stable, satisfactory music mixes can be realised using simple energy-based loudness predictors.

4. Current dynamic single-band loudness models are robust to different types of musical sounds, but their limited parameters prohibit further development into loudness modelling and thus these models do not extend to applications that require consideration of inter-source masking or perceptually-motivated objectives.

5. In their current form, dynamic EP based loudness models do not generalise well to musical sounds with marked differences in spectra.

6. A model of SLS is necessary to predict the loudness of complex sounds encountered in the real-world. The model should incorporate temporal persistence to better capture the loudness of fluctuating sounds.

7. The performance of multiband loudness models on musical stimuli can be significantly improved by reducing the predicted amount of SLS.

8.2 Contributions

The research presented in this thesis has resulted in a number of original contributions, which are summarised below.

An efficient multiband loudness estimator

A number of different approaches to optimising psychoacoustic models of loudness have been assessed using musical sounds (Chapter 4). The main contributions of this work are:

- A low-cost accurate digital implementation of the outer- and middle-ear transmission filter given in ANSI S3.4 (2007).
- A complete adaptation and evaluation of the HGDFT used as a frontend to a standardised loudness calculation procedure.
- Parameterised and modified versions of two multiband loudness models for real-time binaural loudness estimation, supplemented with detailed evidence of expected performance.

Perceptual studies focussing on the perceived loudness of musical sounds

Methodological approaches to loudness measurement and model evaluation were assessed in Chapters 6 and 7, leading to:

- A validated methodology for measuring and deriving subjective loudness data of musical sounds, combined with a framework for evaluating objective predictors.
- The derivation of subjective loudness-matching data for musical sounds from four independent loudness-matching experiments.

Evaluation of loudness models

- A quantified assessment of psychoacoustic multiband models for stationary loudness estimation (Chapter 3).
- A comparison of state-of-the-art loudness models applied to musical sounds (Chapters 5 and 7)
- The identification of key issues with current loudness algorithms, primarily multiband procedures, when applied to single instruments (Chapter 7).

Loudness measurement for music production

- A mixing framework that accommodates any loudness feature and can realise arbitrary target relative-loudness balances in real-time (Chapters 5).
- A new hybrid model for loudness measurement suited to multitrack audio (Chapter 7).
- The clarification of evidence concerning the decision variable to quantify the global loudness of fluctuating sounds (Chapter 7).

- A new robust and adaptive algorithm for measuring the global loudness of an arbitrary signal for both offline and online applications. The approach is not confined to any particular loudness estimator and is controlled with a single parameter (Chapter 7).

8.3 Critique and further work

8.3.1 Spectral loudness summation

One of the primary advantages of the current EP based loudness models is their ability to predict the perceived magnitude of a sound, rather than giving a level reading as done by current single-band methods used in broadcasting. Thus, in order to measure the absolute strength of a sound, or indeed the quantitative relations of the instruments comprising a mix, models capable of outputting loudness as a *ratio variable* are needed. Only models designed to predict the loudness growth functions of stimuli, as measured experimentally using magnitude estimation and magnitude production, are truly fit for predicting the *extent* to which the loudness of one sound compares to others.

Although the work conducted here has not directly evaluated loudness models using a ground-up psychoacoustic approach, a connection has been made between poor real-world performance and known shortcomings of current EP based loudness models (Fastl et al. 2009; Rennies. J. et al. 2009; Rennies and Verhey 2009; Heeren et al. 2011; Schlittenlacher et al. 2011; Hots et al. 2013; Rennies et al. 2013; Schlittenlacher and Ellermeier 2015). In particular, the findings here provide indirect evidence supporting the idea that current models of loudness designed to predict the loudness growth of the pure tone are not consistently accurate when measuring the relative loudness of narrowband and broadband sounds. That is, these procedures do not seem to accurately combine the contributory loudnesses of multiple critical bands.

It is here argued that the over-sensitivity of these models to the spectral distribution of a sound is the primary obstacle limiting their generalisation to externally valid sound material.

To tackle this problem, more experiments are needed to measure the SLS of *stationary* sounds with different spectral profiles across frequency, in isolation of confounding variables such as musical context. Ideally, both loudness matches and magnitude estimates should be measured experimentally to assess models in terms of predicting both loudness growth and between-stimulus loudness relations. Such subjective data would help to test different modelling strategies with the goal of better understanding the underlying mechanism(s) responsible for SLS. In addition, the sounds used in the subjective assessments of this thesis could be used as starting point towards synthesising the necessary stimuli needed to investigate this issue systematically. Is the observed tendency of these models to overestimate the level difference, at the point of equal-loudness, between a narrowband low-frequency instrument and a broader-frequency instrument still observed when musical context and temporal information are eradicated? Does this effect depend on the frequency separation between the two sounds?

8.3.2 Global-loudness estimation

There remains some uncertainty as to how the global loudness of an arbitrary stimulus should be computed given its loudness-time function. This is less of a concern for the energy-based methods since they are not designed to predict the loudness of very short sounds, although this does raise further questions as to their applicability to impulsive sounds or use as a momentary loudness meter. From the review of approaches and the empirical findings presented in Chapter 7, it is suggested that the more complex EP models of loudness should employ an extended release time when generating the STL, and a percentile—not peak—of the time series should represent the decision variable. It is difficult to suggest exact parameter values, but the 95th percentile appears a reliable first choice for unknown cases. As for the release time, further work is needed. The gated percentile approach to estimating the global-loudness percept, as presented at the end of Chapter 7, may also be assessed in more low-level psychoacoustic studies. Further study should also investigate signal-adaptive techniques similar to the one proposed by Terrell (2013, Chapter.7), but using amplitude-modulated tones and bandlimited noise with a wide range of modulation frequencies/bandwidths, rather than musical signals, to allow for a more systematic investigation.

8.3.3 Loudness estimation of multitrack audio

Of the existing algorithms available, ITU-R BS.1770 (2015) and LARM (Skovenborg and Nielsen 2004b) appear to provide the most consistent and stable approaches for loudness-matching musical sounds. The evaluation of models did reveal that all approaches were not always consistent with the subjective adjustments, with certain sounds, e.g. triangles and shaker, causing specific problems. Difficulties in estimating the loudness of sounds with very high-frequency content, such as percussive instruments, have also been reported in other work (Mansbridge et al. 2012b; Pestana and Alvaro 2012; Pestana et al. 2013), although previously suggested modifications to the ITU-R BS.1770 (2015) were only indicative of an improvement in this study. It is plausible that musical context has a strong influence over the subjective judgements of such sounds, so signal classification or similarity measures of the different auditory representations may be needed to completely eliminate residual variance attributed to perceived annoyance and subjective preference. Therefore, a more extensive set of loudness-matching experiments is warranted to address the severity of these issues. Ideally, the material should encompass a broader range of musical instruments than used in this work, and random balanced-pair-matching designs should be used in combination with the MOA to maximise the size of the dataset and minimise bias effects. It is important to stress, however, that more low-level psychoacoustic studies should be the primary focus for the development of multiband methods.

A new hybrid approach to loudness estimation has been presented in this thesis - the LML model. This model showed consistently good performance when predicting the relative loudness of different instruments, and emphasises the limitation of models that operate entirely in the time-domain, such as the two aforementioned single-band algorithms. Further tests are required to determine the generalisability of the LML under different types of reproduction, most importantly loudspeaker.

This highlights an important drawback of the subjective reference data established in this thesis: all experimental data were measured using the same headphones, and stimuli were generally presented at sound levels in the region of 70–80 dB SPL (RMS). Ideally additional experiments should be conducted to investigate two hypotheses:

1. The predictive quality of a loudness model depends on the listening level used in the subjective assessment.

2. The performance of loudness models are significantly better under specific reproductions.

Testing these hypotheses may reveal bias effects in the subjective measurements, rather than specific problems with the models themselves. The general behaviour of the models across the different assessments conducted in this thesis, in addition to the correlations made with previous work, does, however, suggest that the main findings are indeed generalisable, though not all observations may be consistent under significantly different listening conditions.

8.3.4 Partial loudness

As highlighted in the fourth point listed in Section 8.1.7, auditory masking is an important aspect of loudness perception when multiple sounds are competing to be heard. Methods to predict the partial loudness of time-varying sounds were presented in Chapter 3, but their usefulness in real-world conditions is questionable, as elaborated on in Chapter 5. Of course, the prediction of partial loudness requires a quantitative model of auditory masking. Since current psychoacoustic models of loudness are relatively crude with respect to more elaborate masking models, current implementations are limited to rather simple measures of energetic masking. Irrespective of scope, work should first concentrate on solving the SLS problem and then validate the refined models on real-world sounds, *before* considering the influence of competing sounds. A more sophisticated set of rules based on gestalt principles are needed to account for masking effects of musical sounds, e.g. to handle ambiguous cases involving correlated sounds. Thus, a cognitive layer is likely needed to establish appropriate, source-specific masking functions.

8.3.5 Intelligent Music Production

More work in IMP is needed to collect relative-loudness balance data from experienced music engineers to better quantify best practises. A number of approaches have been presented in this thesis to extract this information. However, one of the main questions is how these empirical templates should be applied to unseen multitrack recordings with correlated sources or indeed independent tracks that constitute a single element in the template, e.g. the kick drum of a drum kit. Therefore, the mixing device needs to integrate instrument grouping principles to optimally map the template to the mixing parameters.

The problem of partial loudness is of particular interest to the automatic mixing community, and no remedy is provided here, although a word of caution is. Instead, measures of inter-track masking should be researched, and used to supplement other auditory predictors, such as loudness, to optimise the mix. Though the target loudness balance would not be precisely realised (under masking), a first approximation may be sufficient if a subsequent system targets the minimisation of inter-track masking using different auditory criteria. This was the justification for the automatic mixing system presented by Ward et al. (2012)

8.3.6 Adaptive efficient loudness modelling

A number of different approaches to optimising the loudness model of Glasberg and Moore (Moore and Glasberg 1997; Moore et al. 1997; Glasberg and Moore 2002) have been presented over the years. This thesis has systematically investigated a myriad of techniques in detail and demonstrated how they can be successfully integrated into related algorithms (Chen et al. 2011; Chen and Hu 2012) and applied to real-word problems. A selection of parameter sets were provided with measures of expected performance relative to the reference algorithms. These parameters were found to depend on both loudness model and stimulus type. Thus, there is no one-size fits all approximation to fast, low-resource, high-accuracy loudness estimation. An interesting approach is to employ adaptive auditory filtering schemes like those proposed by Krishnamoorthi (2011), although these still generally require frame-by-frame calculations and do not consider target error thresholds or the effect of listening level. A useful extension for multitrack music applications would be to employ a single loudness model for all tracks, but use an activity detection function, such as the one proposed in Chapter 4, to determine which sounds should be processed at each time step. The avoidance of audible but quiet signals would have no practical influence on global loudness so long as the loudness-time functions were updated when salient musical activity occurs for a given sound. This would significantly reduce the processing power required by more sophisticated loudness algorithms, especially those that consider masking, when embedded in a multitrack or multichannel audio application.

Appendix A

Validation of ANSI S3.4 2007

This appendix provides the validation results of the author's implementation of the ANSI S3.4 (2007) loudness calculation procedure for steady-state sounds. Annex A of this standard provides examples of model predictions for various spectra. Some of these examples have been reproduced using the software developed as part of this thesis for the purpose of validating the implementation.[1] The procedure was implemented as detailed in Section 3.3.3. An example of estimating the loudness of a 1 kHz pure tone presented at 40 dB SPL binaurally in the free field with frontal incidence is given below:

```python
import loudness as ln

inputBank = ln.SignalBank()
nSources = nEars = nChannels = nSamples = fs = 1
inputBank.initialize(nSources, nEars, nChannels, nSamples, fs)

# Set the centre frequency of the first channel
inputBank.setCentreFreq(0, 1000)
# Set the intensity (40 dB SPL) in normalised units
inputBank.setSample(0, 0, 0, 0, 10 ** (40 / 10))

model = ln.StationaryLoudnessANSIS342007()
model.initialize(inputBank)

model.process(inputBank)
outputBank = model.getOutput('Loudness')

print("Loudness in sones {0:.2f}".format(outputBank.getSample()))
```

[1] https://github.com/deeuu/loudness

Tables A.1-A.9 show the validation results for a range of stimuli: pure tones, filtered noise, multiple tones, and tone plus noise. Each table gives a parameter, a target loudness as published in the standard, a measured loudness as output by the author's implementation, and a relative error. The relative error, expressed as a percentage, is calculated according to:

$$Error = 100 \times \frac{|target - measured|}{target}. \tag{A.1}$$

With exception of the 0.0002 sone discrepancy at 5 dB SPL for a 1 kHz pure tone (Table A.1), the predictions of the implementation are less than 0.2% of the published values and the majority of the (rounded) predictions are identical to those given in the standard. Such small deviations could be attributable to differences in the interpolation used to derive the outer- and middle-ear transfer function or due to the approximations used for estimating the specific loudness (SL) transformation parameter values.

It is not clear as to the reason for the 2.27% error shown in Table A.1. To investigate this matter further, this condition (1 kHz tone presented in the free field at 5 dB SPL) was checked using an implementation of the standardised model by Genesis Acoustics as part of an open-source toolbox for Matlab (Genesis Acoustics 2010). The overall loudness predicted by the Matlab version was 0.0090 sones which agrees with the measured value listed in Table A.1. A detailed comparison between the source code of ANSI S3.4 and the author's implementation would be required to identify the true cause of this discrepancy. Unfortunately, only a compiled program is provided with the standard.

Finally, predictions of the free-field absolute threshold predictions were validated by inputting a pure tone at each frequency specified in ISO 389-7 (2005) and varying the intensity level until a loudness level of 2.2 phons ± 0.031 was measured.[2] This value corresponds to the loudness level at threshold as per ANSI S3.4 (2007). Thus, the SPL required to achieve a loudness level of 2.2 phons gives an estimate of the level at threshold. Figure A.1 shows the International Organization for Standardization (ISO) thresholds (black line) and those predicted by the author's implementation of the ANSI S3.4 (2007) (dotted red line). The maximum deviation over 20 Hz-14 kHz was 0.19 dB. This is in good agreement with ANSI S3.4 in which all predictions are reported to be within 0.2 dB of the ISO thresholds for frequencies between 50 Hz and 12.5 kHz.

[2] The error bounds are based on the accuracy of the sone-to-phon conversion and tolerance used in the iterative procedure.

SPL (dB)	Target	Measured	Error (%)
0	0.0011	0.0011	0.00
1	0.0018	0.0018	0.00
2	0.0028	0.0028	0.00
3	0.0044	0.0044	0.00
4	0.0065	0.0065	0.00
5	0.0088	0.0090	2.27
8	0.017	0.017	0.00
10	0.029	0.029	0.00
15	0.070	0.070	0.00
20	0.142	0.142	0.00
25	0.255	0.255	0.00
30	0.422	0.422	0.00
35	0.662	0.662	0.00
40	0.997	0.997	0.00
45	1.461	1.461	0.00
50	2.098	2.098	0.00
55	2.970	2.970	0.00
60	4.166	4.166	0.00
65	5.813	5.813	0.00
70	8.102	8.102	0.00
75	11.326	11.326	0.00
80	15.980	15.981	0.01
85	22.929	22.930	0.00
90	33.216	33.216	0.00
95	48.242	48.243	0.00
100	70.362	70.363	0.00
105	103.274	103.275	0.00
110	152.776	152.776	0.00
115	227.855	227.856	0.00
120	341.982	341.982	0.00

Table A.1: Validation results for a 1 kHz pure tone presented binaurally in the free field at different SPLs.

SPL (dB)	Target	Measured	Error (%)
20	0.35	0.35	0.00
40	1.8	1.8	0.00
60	7.1	7.1	0.00
80	27.5	27.5	0.00

Table A.2: Validation results for a 3 kHz pure tone presented binaurally in free field at different SPLs.

Bandwidth (Hz)	Target	Measured	Error (%)
100	4.25	4.25	0.00
1000	14.29	14.28	0.07

Table A.3: Validation results for a band of white noise centred at 1 kHz, presented binaurally in the free field, with a spectrum level of 40 dB SPL and a bandwidth of 100 Hz and 1 kHz.

Bandwidth (Hz)	Target	Measured	Error (%)
100	4.25	4.25	0.00
1000	8.02	8.01	0.12

Table A.4: Validation results for a band of white noise centred at 1 kHz with a spectrum level of 40 dB and 30 dB for bandwidths of 100 Hz and 1 kHz respectively.

Band SPL (dB)	Target	Measured	Error (%)
0	0.071	0.071	0.00
10	0.67	0.67	0.00
20	2.51	2.51	0.00
30	6.26	6.25	0.16
40	12.7	12.7	0.00
50	23.3	23.3	0.00

Table A.5: Validation results for a broadband noise, presented binaurally in free field at different third octave band levels.

Component SPL (dB)	Target	Measured	Error (%)
60	6.35	6.35	0.00

Table A.6: Validation results for a complex tone comprised of pure tones at 1500, 1600 and 1700 Hz with component SPLs of 60 dB, presented binaurally in the free field.

Component SPL (dB)	Target	Measured	Error (%)
60	12.62	12.63	0.08

Table A.7: Validation results for a complex tone comprised of pure tones at 1000, 1600 and 2400 Hz each with an SPL of 60 dB, presented binaurally in free field.

Component SPL (dB)	Target	Measured	Error (%)
30	1.99	1.99	0.00

Table A.8: Validation results for a complex tone comprised of pure tones equally spaced at 100 Hz in the interval [100, 1000] Hz. The SPL of each component was 30 dB, and the complex tone was presented binaurally in free field.

Tone SPL, Spectrum level SPL (dB)	Target	Measured	Error (%)
40	5.14	5.14	0.00

Table A.9: Validation results for a 1 kHz sinusoid combined with a 100 Hz wide band of white noise noise centred at 1 kHz, presented binaurally in free field.

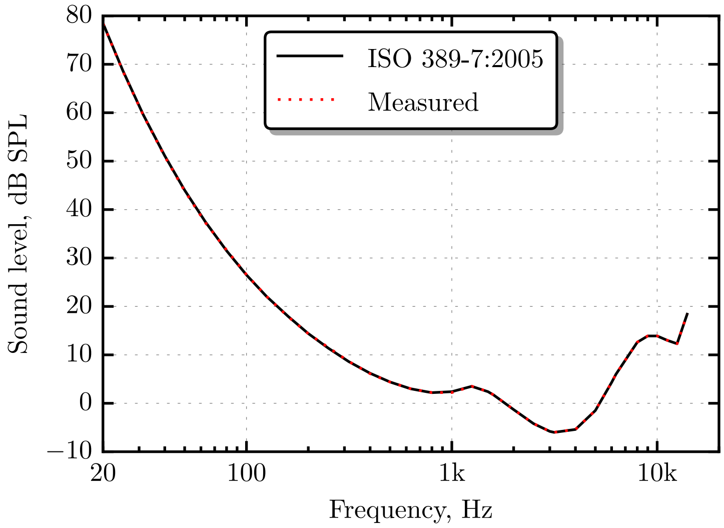


Figure A.1: Absolute thresholds as given in ISO 389-7 (2005) and predictions of the loudness model.

Appendix B

Level change prediction using sone-to-phon mapping

To give some insight into the accuracy of predicting stimulus level differences using an assumed linear relationship between sound level and (objective) loudness level, a small experiment was conducted in which the sound level of 30 real-world sounds was varied and then predicted using the loudness models of Glasberg and Moore (2002) (the GM02) and Chen and Hu (2012) (the CH12). Short-segments of 30 stereo sounds were extracted from the sound quality assessment material (SQAM) database (EBU Tech 3253 2008) and peak calibrated to a random level between 84 and 94 dB SPL, giving a spread of listening levels across the sources. Average long-term loudness (LTL) and peak short-term loudness (STL) global-loudness descriptors (in sones) were then extracted using the two models. For this test, the GM02 was configured with the revised LTL release time constant of 200 ms (Moore et al. 2003). Gain offsets between -5 and 5 dB in steps of 0.5 dB were then applied to the stimuli and the process was repeated. The measures of global loudness were each converted to a loudness level (in phons), denoted L_p, using a model-specific sone-to-phon transformation as well as the simple sone-to-phon relation (Fastl and Zwicker 2007)

$$L_p = 40 + 10 \log_2 \Psi, \tag{B.1}$$

where Ψ is the loudness of the sound in quiet, in sones. Let $L_{p,0}$ denote the loudness level of a given stimulus at the input level. Let $L_{p,X}$ be the loudness level of the same stimulus measured after applying a gain of X dB, relative to the input level. The change in loudness level introduced by the change in signal level was then measured using:

$$\Delta L_p = L_{p,X} - L_{p,0}. \tag{B.2}$$

If $\Delta L_p = X$, then changes in signal level yield equivalent changes in loudness level. To quantify the accuracy of predicting the changes in signal level from the model loudness estimates, the relative error was calculated:

$$\epsilon = \frac{\Delta L_p - X}{X}. \tag{B.3}$$

The root mean square (RMS) and maximum absolute relative errors for both the peak STL and mean LTL measures of global loudness are given, as a percentage, in Table B.1. The overall average RMSE is around 19% for the GM02 and 41% for the CH12. Therefore, a change of 5 dB in signal level is, on average, associated with a prediction error of approximately 1 dB for the GM02 and 2 dB for the CH12. Compared to the GM02, the CH12 performs notably worse when using

Equation B.1, which is not surprising considering the models predict markedly different rates of growth in loudness at higher levels (refer to Section 3.3.6). The maximum relative errors are also more pronounced for the CH12, especially when using Equation B.1.

Descriptor	GM02		CH12	
	SoneToPhon	Equation B.1	SoneToPhon	Equation B.1
Peak STL	24.1 (30.5)	18.8 (23.3)	38.2 (56.1)	56.5 (156.8)
Mean LTL	18.5 (29.37)	15.9 (22.4)	38.9 (54.2)	32.3 (89.0)

Table B.1: RMS (and maximum) relative error (%) in the level change predictions for peak STL and mean LTL using the model-specific sone-to-phon transformations (SoneToPhon) and the approximation given by Equation B.1.

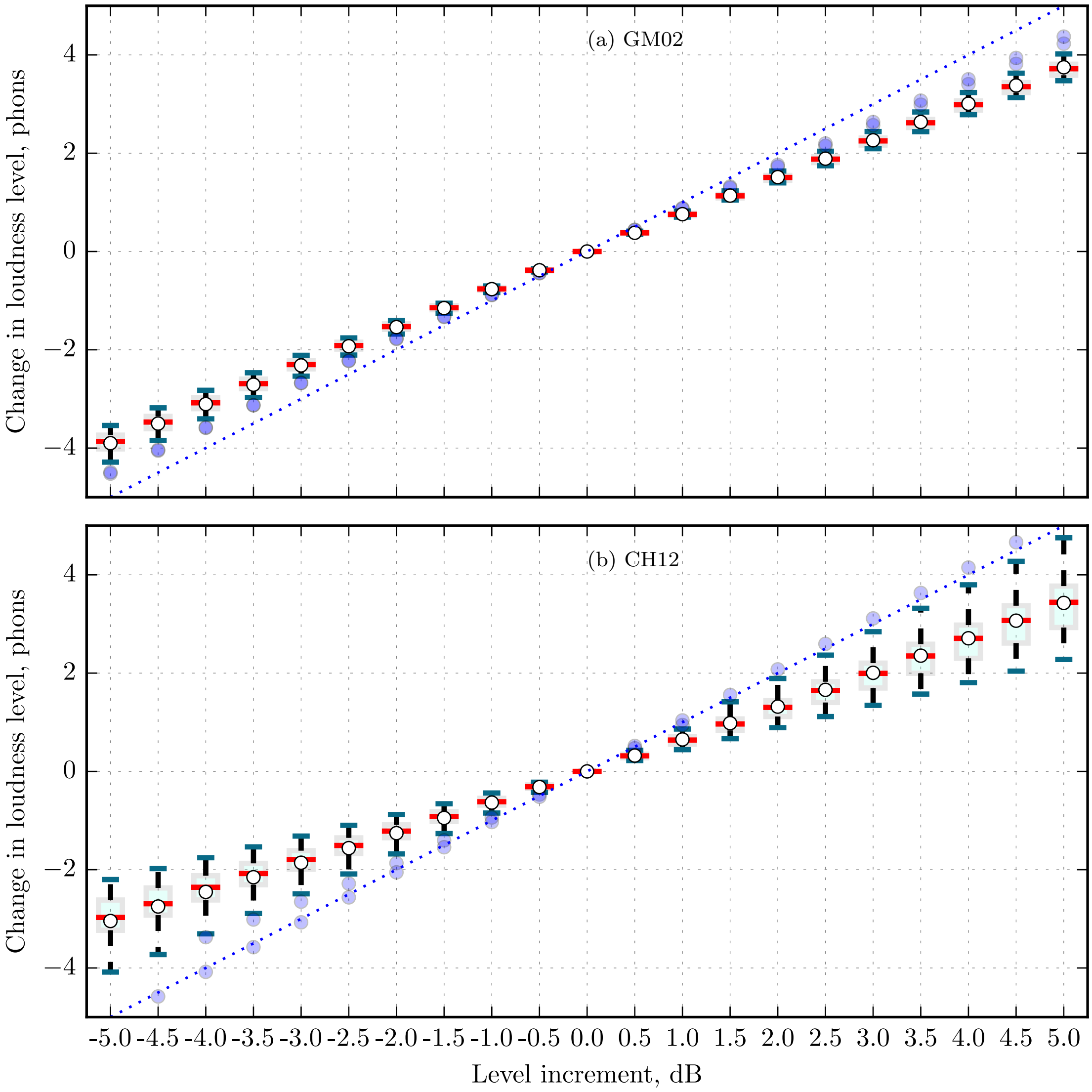


Figure B.1: Boxplots of the loudness level differences introduced by a given increment in signal level for (a) the GM02 and (b) the CH12. The loudness level estimates were made using the model-specific sone-to-phon transformation applied to the peak STL measure. The dotted straight line reflects a perfect relationship between signal gain (dB) and loudness level (phons).

Figure B.1 shows boxplots of the changes in the loudness levels introduced by a given increment in signal level. Loudness levels were determined using the model-specific formula applied to the peak STL. It can be seen that changes in loudness level are less than the corresponding changes in signal level, and thus the means of the boxplots deviate from the dotted straight line which indicates a perfect fit between the two variables. Note that the discrepancies are greatest for the CH12. Although regression lines could be fitted to these two datasets to improve predictive quality, it should be emphasised that the slopes are stimulus dependent, as reflected by the unequal spreads of the boxplots. To conclude, the linear approximation of level change by a sone-to-phon transformation is only accurate (in terms of dB) for relatively small differences and thus iterative procedures should be preferred when estimating level adjustments.

Appendix C

Between- and within-subject confidence intervals

The purpose of this appendix is to show how confidence intervals (CIs) for a given statistic can be calculated using different approaches and demonstrate their application within the context of the work presented in this thesis.

Consider an experiment in which ten people were asked to adjust the level of two test sounds, A and B, such that they are equal in loudness to that of a third fixed-reference sound, C. This is a fixed-reference loudness-matching experiment where sound C serves as the reference stimulus. It is a repeated-measures or within-subject experiment because both conditions were measured on the same ten subjects. The data for this hypothetical study are given in Table C.1. The marginal means suggest that the listeners increased the level of sound A and sound B in order to establish equal loudness with sound C. This suggests that sound C is louder than the two test sounds. There are also clear differences between the subjects. For example, subject 3 perceived sound C as being around 2 dB louder than the test stimuli, compared to subject 10 who perceived an average difference of 8 dB.

Subject	A	B	Mean
1	2.3	4.0	3.2
2	6.1	8.5	7.3
3	1.0	3.5	2.2
4	1.7	3.3	2.5
5	7.8	10.0	8.9
6	6.8	9.2	8.0
7	5.3	8.0	6.7
8	2.6	4.5	3.5
9	2.9	4.6	3.8
10	7.4	9.1	8.3
Mean (SD)	4.4 (2.5)	6.5 (2.7)	5.4 (2.6)

Table C.1: Hypothetical data from a fixed-reference loudness-matching experiment. SD is the standard deviation of the mean adjustment level.

In order to draw population inferences from this experiment, it is necessary to estimate the uncertainty of the sample means attributable to sampling error. That is, if this experiment were to be repeated many times, the summary statistics given in Table C.1 would fluctuate each time. This uncertainty can be estimated according to the sampling distribution of the sample means. First the between-subject variation in each condition can be expressed using the sample SD:

$$\mathrm{SD}_j = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_{i,j} - \bar{x_j})^2}, \tag{C.1}$$

where N is the number of observations, x_i is the i^{th} data point in condition j and $\bar{x}_j$ is the average. However, the SD only quantifies the inter-subject variation in our sample (it is however an unbiased approximation of the variation in the population). Instead, it is necessary to estimate the variation in the sample means. Assuming that multiple samples cannot be afforded, the standard error (SE) is used as an approximation:

$$\mathrm{SE} = \frac{\mathrm{SD}}{\sqrt{N}}. \tag{C.2}$$

According to the central limit theorem, for large samples (>30), the sampling distribution is centred on the population mean and has a SD approximately equal to the SE. With that said, the SE can be used to calculate a CI for the mean, which describes a range of values within which we expect to include the population mean a certain percentage of the time. For example, if the above experiment were to be repeated a further 19 times, and a CI was computed for each sample mean, then 19 out of the 20 would be expected to contain the true mean. The wider the interval, the more probable the true value has been captured, at the cost of being less certain as to the exact value of the parameter. Note that a 95% CI is not a region within which we are 95% confident that the population value will fall. This is because the population parameter is fixed, meaning the parameter value either is or isn't in the interval.

For large samples, these limits are computed using the SE of the mean:

$$\mathrm{CI} = \bar{x} \pm z_{\mathrm{cp}} \times \mathrm{SE}, \tag{C.3}$$

where z is a score from a normal distribution with a mean of zero and variance of 1, with the subscript cp denoting the cumulative probability for the z-value which is derived from the desired significance level α:

$$\mathrm{cp} = 1 - \frac{\alpha}{2}. \tag{C.4}$$

For example, the popular 95% CI is calculated by setting $\alpha = (1 - 0.95) = 0.05$, which gives a cumulative probability of 0.975. This gives a z-score of approximately 1.96 and therefore

$$\mathrm{CI}_{95} = \bar{x} \pm 1.96 \times \mathrm{SE}. \tag{C.5}$$

Referring back to Table C.1, it can be seen that ten people participated in the experiment. The z interval is based on the shape of the sampling distribution being normal, which is only applicable in large samples. For small samples (<30), the sampling distribution has a t-distribution, which approaches normality as the sample size increases. Therefore, for small samples, Equation C.3 is replaced with

$$\mathrm{CI} = \bar{x} \pm t_{\mathrm{df,cp}} \times \mathrm{SE}, \tag{C.6}$$

where t is the critical value from a t-distribution based on a $N - 1$ degrees of freedom (df).

For the example data, where $N = 10$, the SD for the mean of sounds A and B are 2.5 dB and 2.7 dB, respectively, giving SEs of 0.8 dB and 0.9 dB. The 95% CIs are calculated using Equation C.6, with df $= (N - 1) = 9$ and $cp = 0.975$. The required t-value is 2.26, and so the estimated intervals are 95% CI = [2.6, 6.2] (sound A), and 95% CI = [4.5, 8.4] (sound B). Figure C.1(a) shows the matching gain applied by each subject in each condition (connected by lines) as well as the sample means with their respective 95% *between-subject* CIs (plot b). The

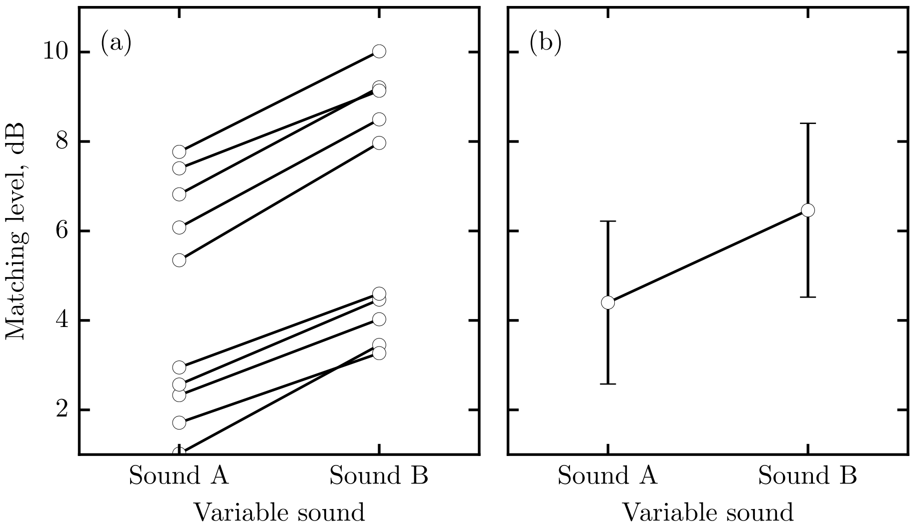


Figure C.1: (a) Individual data points for the example loudness matching data, and (b) sample means with between-subject 95% CIs.

intervals express a range of plausible values, given the data, for population means. In this case, the two intervals are positive, which leads to the inference that both test sounds (A and B) are indeed softer than sound C, when presented at their reference levels. The magnitude of these differences (the absolute effect size) is entirely captured by lower and upper confidence limits.

When two CIs do not overlap, one can infer that the two means are reliably different at the specified significance level. However, when the two intervals do overlap, the mean difference may or may not be statistically significant. From comparing the CIs shown in subplot (b), it is not clear whether sound A is systematically louder than sound B, i.e. whether the difference in means is not entirely due to chance. However, subplot (a) paints a very different picture: every subject applied more gain to sound B compared to sound A, suggesting that the mean of sound B is systematically larger than the mean of sound A. A paired t-test reveals that the two dependent means are indeed reliably different ($t(9) = 16.5$, $p = 5 \times 10^{-8}$). This effect is not apparent from looking at the between-subject CIs because they are contaminated by the subject differences show in the final column of Table C.1. If the purpose of error bars in figures is to reveal significant condition differences or reveal trends in the data, then within-subject CIs should be used (Loftus and Masson 1994; Cousineau 2005; O'Brien and Cousineau 2014). This is a CI computed *after* removing inter-subject differences. Note that such a procedure can only be performed if repeated-measures have been made on each subject.

Following Cousineau (2005), the within-subject CI is calculated by standardising the data. The mean of each subject $\bar{x_i}$ is subtracted from their respective condition measurements $x_{i,j}$ and their data are centred on the grand mean $\bar{x}$:

$$y_{i,j} = x_{i,j} - \bar{x_i} + \bar{x}, \tag{C.7}$$

where y_i is the aligned data for subject i. Equation C.3 or C.6 is then applied to each condition (column) to get the within-subject intervals. Table C.2 shows the standardised dataset for the running example presented in this appendix. Notice that although the condition means remain the same, the SDs are now lower. This is because the between-subject variance has been eliminated, as shown by the final column where the individual means are equal to the grand mean. For these

Subject	A	B	Mean
1	4.6	6.3	5.4
2	4.2	6.6	5.4
3	4.2	6.6	5.4
4	4.7	6.2	5.4
5	4.3	6.6	5.4
6	4.2	6.6	5.4
7	4.1	6.7	5.4
8	4.5	6.4	5.4
9	4.6	6.3	5.4
10	4.6	6.3	5.4
Mean (SD)	4.4 (0.2)	6.5 (0.2)	5.4 (0)

Table C.2: Aligned data for the example fixed-reference loudness-matching experiment.

data, this source of variability corresponds to inter-subject differences in the perception of the reference stimulus (sound C).

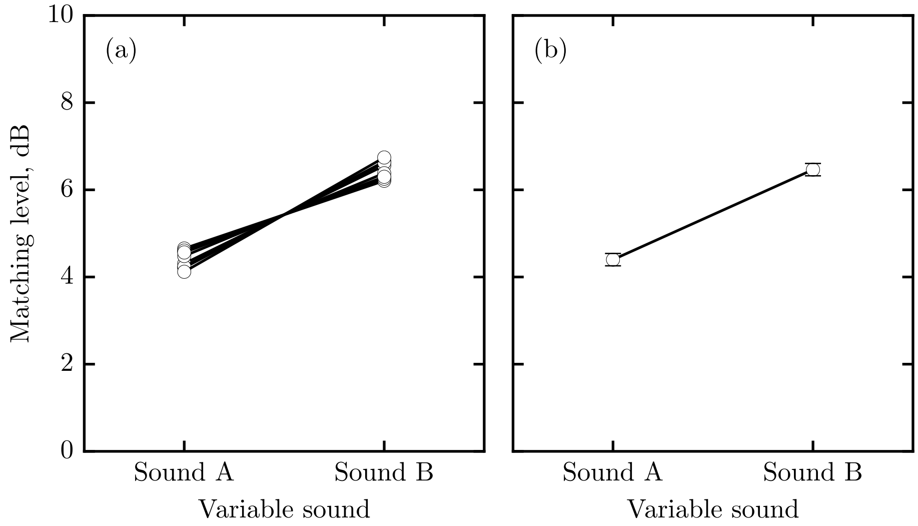


Figure C.2: (a) Individual data points for the aligned loudness matching data, and (b) sample means with within-subject 95% CIs.

Figure C.2 shows the aligned data and the condition means with within-subject 95% CIs. Again, the data are the same as used for Figure C.1 but with subject variability removed. The effect of the alignment procedure is clear: the variance within each condition is substantially reduced which leads to tighter error bars. The removal of subject variability leaves only two sources of variability: the condition variance and the subject × condition interaction variance (Cousineau 2005).

To summarise, the between-subject CI provides information about the value of the population mean and must be based on both inter-subject variance and interaction variance (Cousineau 2005). In contrast, the within-subject CI provides information about trend in a set of population means by eliminating person variability. In the example above, the paired t-test, which harnesses the power of a repeated-measured design, was highly significant, yet the between-subject CIs did not reflect this because this additional power was not taken into account. When the within-subject CI was used, the plot of sample means reflected the inference drawn from the statistical test.

The hypothetical data presented here involves just two conditions. In more complex within-subject experiment designs, such as a factorial, the researcher typically wants to convey the results of the statistical tests performed, e.g. an N-way repeated-measures analysis of variance (ANOVA), or use the plots to identify trends prior to further analyses. In any case, one problem with the standardisation procedure proposed by Cousineau (2005) is that the SEs of the normalised data are smaller than those in true population (Morey 2008). To counteract this issue, Morey (2008) introduced a design-dependent correction factor to adjust the SEs. O'Brien and Cousineau (2014) showed how to incorporate the correction factor in the data such that correct SEs can be computed directly. Using Equation C.7 to obtain the standardised data y_i for the i^{th} subject, the corrected and aligned data w are obtained using

$$w_{i,j} = \sqrt{\frac{J}{J-1}} \times (y_{i,j} - \bar{y}_j) + \bar{y}_j, \tag{C.8}$$

where $\bar{y}_j$ is the mean of the j^{th} condition and the first term is the correction factor where J is given by

$$J = \prod_{f=1}^{F} L_f, \tag{C.9}$$

where F is the number of experimental factors and L_f is the number levels in factor f. For the two-measure design used for the hypothetical experiment, the correction factor is $\sqrt{2} \approx 1.41$, which means that the standard errors are increased by about 41%. The correction decreases as the number of conditions increases, e.g. for a 2×3 factorial design, $J = 6$, which gives a correction factor of $\sqrt{6/5} \approx 1.095$.

C.0.1 Bootstrapping

In the previous section, CIs were computed to estimate the uncertainty of the mean from one sample of the population. CIs are based on the sampling distribution of the statistic (here the mean). When z or t intervals are computed, as given by Equations C.3 and C.6, an assumption is made about the shape of the sampling distribution. For large samples, it is assumed that the sampling distribution is normal, and in small samples, the shape is defined by a t-distribution given the degrees of freedom. In most cases, a normality check of the sample data is used to validate this: if the data are normal, the population distribution of the variable is normal and therefore so is the sampling distribution. But what if the sample data do not follow a normal distribution? In large samples, the central limit theorem is sometimes used to escape this problem since the sampling distribution tends to approach normality as the sample size increases (regardless of the population distribution).

An alternative and more flexible approach is bootstrapping, which estimates the sampling distribution directly from sample data (rather than through theory). Following Hesterberg et al. (2003), the general procedure is:

1. Rather than repeating an experiment by taking many random samples of size N from the population, draw many resamples of size N by repeatedly sampling with replacement from one random sample. Sampling with replacement means using all observations each time a sample is drawn, even if that observation has already been selected.

2. For each resample, compute the statistic of interest and store the result. The resulting set of values is called bootstrap distribution, and provides information about a statistics sampling distribution. The bootstrap distribution is typically formed using 1000–10000 resamples.

3. From the bootstrap distribution, compute a CI using the t-interval or the percentile method.

In effect, the bootstrap distribution stands in for the population, and many samples are drawn. The shape of the distribution approximates the shape and spread of the sampling distribution of the chosen statistic, and tends to be centred on the original statistic value rather than the true parameter value (Hesterberg et al. 2003). The shape of the bootstrap distribution can be used to check for normality of the sampling distribution. Similarly, the SD of the bootstrap distribution estimates the SD of the sampling distribution, and is therefore termed the bootstrap SE. The bootstrap SE measures the variability in the statistic under random sampling. Finally, the distance between the original statistic value and the centre (mean) of the bootstrap distribution determines the bias. The bootstrap bias is used to estimate the bias of a statistic, which is the difference between the mean of its (unknown) sampling distribution and the true population value.

If the bootstrap distribution is normal and has little bias, a CI for the sample mean ($\bar{x}$) can be obtained using the bootstrap SE ($\mathrm{SE}_{\mathrm{boot}}$) and the appropriate t-value:

$$\mathrm{CI}_{\mathrm{boot}} = \bar{x} \pm t_{\mathrm{df,cp}} \times \mathrm{SE}_{\mathrm{boot}}, \tag{C.10}$$

where the degrees of freedom corresponds to the size of the original sample minus one. Note that in practice, the sample mean can be replaced with a variety of statistics since the bootstrap procedure is based on resampling rather than statistical theory (which might not say anything about the sampling distribution). If the bootstrap distribution is non-normal or reveals large bias, the t-interval should not be used.

The bootstrap percentile method can be employed to calculate a CI without t-values, and should be used when the bootstrap bias is small. The percentile approach has the advantage that it does not ignore skewness in the bootstrap distribution and so gives asymmetric CIs. To get a $P\%$ bootstrap percentile CI, calculate the $(100 - P)/2$ and $P + (100 - P)/2$ percentiles of the bootstrap distribution which estimate the lower and upper limits of the interval, respectively. When bootstrap t and bootstrap percentile intervals do not agree, neither should be used (see Hesterberg et al. (2003) for examples and additional bootstrap techniques).

Abbreviations

ANOVA analysis of variance

ANSI American National Standards Institute

ART Aligned Rank Transform

CF characteristic frequency

CF02 Chalupper and Fastl's 2002 dynamic loudness model

CH12 Chen and Hu's 2012 dynamic loudness model

CI confidence interval

CR complexity reduction

DAW digital audio workstation

DFT discrete Fourier transform

DIN German Institute for Standardization

DL difference limen

DSSB dynamic sound-stream bias

EBU European Broadcast Union

EMA exponential moving average

EP excitation pattern

ERB equivalent rectangular bandwidth

RMSE* epsilon-insensitive root mean square error

FFT fast Fourier transform

FIR finite impulse response

FS full scale

GM02 Glasberg and Moore's 2002 dynamic loudness model

GMD Ginis mean difference

HGDFT hopping Goertzel discrete Fourier transform

HRTF head related transfer function

IIR infinite impulse response

IL instantaneous loudness

ILD interaural level difference

IMP Intelligent Music Production

IPL instantaneous partial loudness

IQR interquartile range

ISO International Organization for Standardization

ITD interaural time difference

ITM inter-track masking

ITU-R International Telecommunication Union Radiocommunication Sector

$\mathbf{L_{eq}}$ equivalent continuous sound level

LKFS loudness K-weighted relative to Full Scale

LML linear multiband loudness

LTL long-term loudness

LTPL long-term partial loudness

LU Loudness Units

LUFS Loudness Units relative to Full Scale

LUT lookup table

MAE mean absolute error

MAF minimal audible field

MAP minimal audible pressure

ML momentary loudness

MNPRD mean normalised percentile rank deviation

MOA method of adjustment

MSD mean subjective deviation

MUSHRA multiple stimuli with hidden reference and anchor

nMAE normalised mean absolute error

nPE normalised peak error

nRMSE normalised root mean square error

PC personal computer

PE peak error

PPM peak programme meter

PSE point of subjective equality

PSL partial specific loudness

PSTH post-stimulus-time histogram

PTC psychophysical tuning curve

RLB revised low-frequency B curve

RMS root mean square

RMSE root mean square error

roex rounded exponential

SD standard deviation

SDFT sliding discrete Fourier transform

SE standard error

SL specific loudness

SLS spectral loudness summation

SOR subjective outlier ratio

SPL sound pressure level

SQAM sound quality assessment material

SRG Special Rapporteur Group

STFT short-time Fourier transform

STL short-term loudness

STPL short-term partial loudness

UEN uniform exciting noise

WS weighted spectrum

Bibliography

Aibara, R., Welsh, J. T., Puria, S., and Goode, R. L. (2001). "Human middle-ear sound transfer function and cochlear input impedance." In: *Hearing Research* 152, pp. 100–109.

Aichinger, P. (2009). "Investigation of psychoacoustic principles for automatic mixdown algorithms". Masters thesis. University of Music and Performing Arts Graz.

Aichinger, P., Sontacchi, A., and Schneider-Stickler, B. (2011). "Describing the transparency of mixdowns: the masked-to-unmasked-ratio". In: *Proceedings of the 130th Audio Engineering Society Convention.*

ANSI S1.11 (1986). *Specification for octave-band and fractional-octave-band analog and digital filters.* Tech. rep. American National Standards Institute.

ANSI S3.4 (2005). *Procedure for the computation of loudness of steady sounds.* Tech. rep. American National Standards Institute.

— (2007). *Procedure for the computation of loudness of steady sounds.* Tech. rep. American National Standards Institute.

ATSC A/85 (2013). *Recommended practice: techniques for establishing and maintaining audio loudness for digital television.* Tech. rep. Advanced Television Systems Committee.

Baker, R. J., Rosen, S., and Darling, A. M. (1998). "An efficient characterisation of human auditory filtering across level and frequency that is also physiologically reasonable". In: *Psychophysical and Physiological Advances in Hearing.* Ed. by Palmer, A. R., Rees, A., Summerfield, A. Q., and Meddis, R. London: Whurr Publishers, pp. 81–88.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). "Fitting linear mixed-effects models using lme4". In: *Journal of Statistical Software* 67.1.

Bauch, H. (1956). "Die bedeutung der frequenzgruppe für die lautheit von klängen". In: *Acta Acoustica united with Acoustica* 6.1, pp. 40–45.

Baykaner, K. R. (2014). "Predicting the perceptual acceptability of auditory interference for the optimisation of sound zones". PhD thesis. University of Surrey.

Bech, S. and Zacharov, N. (2006). *Perceptual audio evaluation - theory, method and application.* Chichester: Wiley.

Békésky, G. V. (1960). *Experiments in hearing.* Ed. by Wever, E. New York: McGraw-Hill.

Beranek, L. L., Marshall, J. L., CudWorth, A. L., and Peterson, A. P. G. (1951). "Calculation and measurement of the loudness of sounds". In: *Journal of the Acoustical Society of America* 23.3, pp. 261–269.

Beranek, L. L. and Peterson, A. P. G. (1948). "Determination of the loudness of noise from simple measurements". In: *Journal of the Acoustical Society of America* 20.4, pp. 592–592.

Blauert, J. (1996). *Special hearing: the psychophysics of human sound localization.* Cambridge, Massachusetts: MIT University Press Group Ltd.

Boer, E. de and Jongh, H. R. de (1978). “On cochlear encoding: potentialities and limitations of the reverse-correlation technique”. In: *Journal of the Acoustical Society of America* 63.1, pp. 115–135.

Boer, E. de and Kuyper, P. (1968). “Triggered correlation”. In: *IEEE transactions on bio-medical engineering* 15.3, pp. 169–179.

Bradter, C. and Hobohm, K. (2008). “Loudness calculation for individual acoustical objects within complex temporally variable sounds”. In: *Proceedings of the 124th Audio Engineering Society Convention.*

Bregman, A. S. (1994). *Auditory scene analysis: the perceptual organization of sound.* Cambridge, Massachusetts: MIT Press.

Brown, G. B. (2004). *Smoothing, forecasting and prediction of discrete time series.* New York: Dover Publications Inc.

BS 4198 (1967). *Method for calculating loudness.* Tech. rep. British Standards.

BS 61672 (2003). *Electroacoustics – sound level meters – part 1: specifications.* Tech. rep. British Standards.

— (2013). *Electroacoustics – sound level meters – part 1: specifications.* Tech. rep. British Standards.

Burdiel, E., Vetter, L., Simpson, A. J. R., Terrell, M. J., McPherson, A., and Sandler, M. B. (2012). “Real-time implementation of Glasberg and Moore’s loudness model for time-varying sounds”. In: *Proceedings of the 133rd Audio Engineering Society Convention.*

Buus, S. and Florentine, M. (2002). “Growth of loudness in listeners with cochlear hearing losses: recruitment reconsidered”. In: *Journal of the Association for Research in Otolaryngology* 3, pp. 120–139.

Buus, S., Florentine, M., and Poulsen, T. (1999). “Temporal integration of loudness in listeners with hearing losses of primarily cochlear origin.” In: *Journal of the Acoustical Society of America* 105.6, pp. 3464–80.

— (1997). “Temporal integration of loudness, loudness discimination, and the form of the loudness function”. In: *Journal of the Acoustical Society of America* 101.2, pp. 669–680.

Cabrera, D., Ferguson, S., and Schubert, E. (2007). “‘Psysound3’: software for acoustical and pyschoacoustical analysis of sound recordings”. In: *Proceedings of the 13th International Conference on Auditory Display.*

Cambridge Music Technology (2013). *The ‘mixing secrets’ free multitrack download library.* URL: `http://www.cambridge-mt.com/ms-mtk.htm`.

Camerer, F. (2010). *On the way to loudness nirvana - audio levelling with EBU R 128.* Tech. rep. European Broadcasting Union.

Cassidy, R. J. and Smith III, J. O. (2007). “Efficient time-varying loudness estimation via the hopping Goertzel DFT”. In: *Proceedings of the 50th Midwest Symposium on Circuits and Systems.*

Chalupper, J. and Fastl, H. (2002). “Dynamic loudness model (DLM) for normal and hearing-impaired listeners”. In: *Acta Acoustica united with Acoustica* 88, pp. 378–386.

Charbonneau, J., Novak, C., and Ule, H. (2009). “Loudness prediction model comparison using the equal loudness contours”. In: *Canadian Acoustics* 37.3, pp. 64–65.

Chatterjee, S. and Hadi, A. S. (2006). *Regression analysis by example.* 4th ed. New Jersey: Wiley.

Chen, Z. and Hu, G. (2012). “A revised method of calculating auditory excitation patterns and loudness for time-varying sounds”. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.*

Chen, Z., Hu, G., Glasberg, B. R., and Moore, B. C. J. (2011). "A new method of calculating auditory excitation patterns and loudness for steady sounds". In: *Hearing Research* 282.1-2, pp. 204–215.

Cohen, J. R. (1989). "Application of an auditory model to speech recognition." In: *Journal of the Acoustical Society of America* 85.6, pp. 2623–2629.

Cousineau, D. (2005). "Confidence intervals in within-subject designs: a simpler solution to Loftus and Masson's method". In: *Tutorials in Quantitative Methods for Psychology* 1.1, pp. 42–45.

Dannenberg, R. B. (2007). "An intelligent multi-track audio editor". In: *Proceedings of the 2007 International Computer Music Conference*. Vol. 2, pp. 89–94.

De Man, B., Leonard, B., King, R., and Reiss, J. D. (2014). "An analysis and evaluation of audio features for multitrack music mixtures". In: *Proceedings of the 15th International Society for Music Information Retrieval Conference.*

De Man, B., McNally, K., and Reiss, J. D. (2017). "Perceptual evaluation and analysis of reverberation in multitrack music production". In: *Journal of the Audio Engineering Society* 65.1/2, pp. 108–116.

DIN 45406 (1966). *Peak programme meter for electroacoustic wide band transmission.* Tech. rep. German Institute for Standardization.

DIN 45631 (1967). *Procedure for calculating loudness level.* Tech. rep. German Institute for Standardization.

— (1991). *Procedure for calculating loudness level and loudness.* Tech. rep. German Institute for Standardization.

DIN 45631/A1 (2010). *Calculation of loudness level and loudness from the sound spectrum – Zwicker method – amendment 1: calculation of the loudness of time-variant sound.* Tech. rep. German Institute for Standardization.

Dugan, D. (1975). "Automatic microphone mixing". In: *Journal of the Audio Engineering Society* 23.6, pp. 442–449.

Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (2003). "Note on informational masking". In: *Journal of the Acoustical Society of America* 113.6, pp. 2984–2987.

EBU R 128 (2010). *Loudness normalisation and permitted maximum level of audio levels.* Tech. rep. European Broadcast Union.

— (2014). *Loudness normalisation and permitted maximum level of audio signals.* Tech. rep. European Broadcast Union.

EBU Tech 3253 (2008). *Sound quality assessment material recordings for subjective tests.* Tech. rep. European Broadcast Union.

EBU Tech 3341 (2011). *Loudness metering: 'EBU mode' metering to supplement loudness normalisation in accordance with EBU R 128.* Tech. rep. European Broadcast Union.

— (2016). *Loudness metering: 'EBU mode' metering to supplement loudness normalisation in accordance with EBU R 128.* Tech. rep. European Broadcast Union.

EBU Tech 3342 (2016). *Loudness range: A measure to supplement EBU R 128 loudness normalization.* Tech. rep. European Broadcast Union.

Edmonds, B. A. and Culling, J. F. (2009). "Interaural correlation and the binaural summation of loudness." In: *Journal of the Acoustical Society of America* 125.6, pp. 3865–70.

Elliott, L. L. (1967). "Development of auditory narrow-band frequency contours". In: *Journal of the Acoustical Society of America* 42.1, pp. 143–153.

Epstein, M. and Florentine, M. (2006). "Loudness of brief tones measured by magnitude estimation and loudness matching". In: *Journal of the Acoustical Society of America* 119.4, p. 1943.

Epstein, M. and Florentine, M. (2005). “A test of the equal-loudness-ratio hypothesis using cross-modality matching functions.” In: *Journal of the Acoustical Society of America* 118.2, pp. 907–913.

Fastl, H. (1975). “Loudness and masking patterns of narrow noise bands”. In: *Acta Acoustica united with Acoustica* 33.4, pp. 266–271.

Fastl, H., Volk, F., and Straubinger, M. (2009). “Standards for calculating loudness of stationary or time-varying sounds”. In: *Proceedings of the 38th International Congress on Noise Control Engineering.*

Fastl, H. and Zwicker, E. (1987). “Loudness level at 400 Hz: psychoacoustical measurement and calculation by ISO 532b”. In: *Fortschritte der Akustik, DAGA* 87, pp. 189–193.

— (2007). *Psychoacoustics: facts and models.* 3rd ed. Berlin Heidelberg: Springer.

Fechner, G. T. (1860). *Elemente der psychophysik.* Leipzig: Breitkopf und Härtel.

Feddersen, W. E. (1957). “Localization of high-frequency tones”. In: *Journal of the Acoustical Society of America* 29.9, p. 988.

Field, A., Jeremy, M., and Field, Z. (2012). *Discovering statistics using R.* Londonzw: Sage.

Fletcher, H. (1940). “Auditory patterns”. In: *Reviews of Modern Physics* 12.1, pp. 47–65.

Fletcher, H. and Munson, W. A. (1933). “Loudness its defintion, measurement and calculation”. In: *Journal of the Acoustical Society of America* 5.1, pp. 82–108.

— (1937). “Relation between loudness and masking”. In: *Journal of the Acoustical Society of America* 9.1.

Florentine, M. (2011). “Loudness”. In: *Loudness.* Ed. by Florentine, M., Popper, A. N., and Fay, R. R. Vol. 37. Springer Handbook of Auditory Research. New York: Springer. Chap. 1, pp. 1–16.

Florentine, M., Buus, S., and Bonding, P. (1978). “Loudness of complex sounds as a function of the standard stimulus and the number of components.” In: *Journal of the Acoustical Society of America* 64.4, pp. 1036–40.

Florentine, M., Buus, S., and Poulsen, T. (1996). “Temporal integration of loudness as a function of level.” In: *Journal of the Acoustical Society of America* 99.3, pp. 1633–1644.

Florentine, M. and Epstein, M. (2006). “To honor Stevens and repeal his law (for the auditory system)”. In: *Fechner Day.* Ed. by Kornbrot, D. E., Msetfi, R. M., and MacRae, A. Vol. 22.

Ford, J., Cartwright, M., and Pardo, B. (2015). “MixViz: a tool to visualize masking in audio mixes”. In: *Proceedings of the 139th Audio Engineering Society Convention.*

Francart, T. and Mcdermott, H. J. (2012). “Development of a loudness normalisation strategy for combined cochlear implant and acoustic stimulation”. In: *Hearing Research* 294.1-2, pp. 114–124.

Francombe, J., Brookes, T., Mason, R., and Melchior, F. (2015). “Loudness matching multichannel audio programme material with listeners and predictive models”. In: *Proceedings of the 139th Audio Engineering Society Convention.*

Frigo, M. and Johnson, S. G. (2005). “The design and implementation of FFTW3”. In: *Proceedings of the IEEE* 93.2, pp. 216–231.

Gelfand, S. A. (2010). *Hearing: an introduction to psychological and physiological acoustics.* 5th ed. Essex: Informa Healthcare.

Genesis Acoustics (2010). *Loudness online.* Online. URL: `http://genesis-acoustics.com/en/loudness_online-32.html`.

Ghitza, O. (1994). “Auditory models and human performance in tasks related to speech coding and speech recognition”. In: *IEEE Transactions on Speech and Audio Processing* 2.1, pp. 115–132.

Gigerenzer, G. and Strube, G. (1983). “Are there limits to binaural additivity of loudness?” In: *The Journal of experimental psychology: Human perception and performance* 9.1, pp. 126–36.

Glasberg, B. R. and Moore, B. C. J. (2002). “A model of loudness applicable to time-varying sounds”. In: *Journal of the Audio Engineering Society* 50.5, pp. 331–342.

— (1990). “Derivation of auditory filter shapes from notched-noise data”. In: *Hearing Research* 47.1-2, pp. 103–138.

— (2005). “Development and evaluation of a model for predicting the audibility of time-varying sounds in the presence of background sounds”. In: *Journal of the Audio Engineering Society* 53.10, pp. 906–918.

Glasberg, B. R. and Moore, B. C. J. (2000). “Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise”. In: *Journal of the Acoustical Society of America* 108.5 Pt 1, pp. 2318–2328.

Glasberg, B. R. and Moore, B. C. J. (2006). “Prediction of absolute thresholds and equal-loudness contours using a modified loudness model”. In: *Journal of the Acoustical Society of America* 120.2, p. 585.

— (2010). “The loudness of sounds whose spectra differ at the two ears”. In: *Journal of the Acoustical Society of America* 127.4, pp. 2433–2440.

Glasberg, B. R., Moore, B. C. J., and Stone, M. A. (1999). “Modelling changes in frequency selectivity with level”. In: *Psychophysics, Physiology and Models of Hearing.* Ed. by Dau, T., Hohmann, V., and Kollmeier, B. Oldenburg: World Scientific, pp. 143–154.

Gockel, H., Moore, B. C. J., and Patterson, R. D. (2003). “Asymmetry of masking between complex tones and noise: partial loudness”. In: *Journal of the Acoustical Society of America* 114.1, pp. 349–360.

— (2002). “Influence of component phase on the loudness of complex tones”. In: *Acta Acoustica united with Acoustica* 88.3, pp. 369–377.

Gonzalez, E. P. and Reiss, J. D. (2009). “Automatic gain and fader control for live mixing”. In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*

— (2008). “Improved control for selective minimization of masking using interchannel dependancy effects”. In: *Proceedings of the 11th International Conference on Digital Audio Effects.*

Grimm, G., Hohmann, V., and Verhey, J. L. (2002). “Loudness of fluctuating sounds”. In: *Acta Acustica united with Acustica* 88.3, pp. 312–323.

Hafezi, S. and Reiss, J. D. (2015). “Autonomous multitrack equalization based on masking reduction”. In: *Journal of the Audio Engineering Society* 63.5, pp. 312–323.

Härmä, A. (1998). *Temporal masking effects: single incidents.* Tech. rep. Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing.

Härmä, A., Karjalainen, M., Savioja, L., Välimäki, V., Laine, U. K., and Huopaniemi., J. (2000). “Frequency-warped signal processing for audio applications”. In: *Proceedings of the 109th Audio Engineering Society Convention.*

Harris, F. J. (1978). “On the use of windows for harmonic analysis with the discrete Fourier transform”. In: *Proceedings of the IEEE* 66.1, pp. 51–83.

Hartmann, W. (1998). *Signals, sound and sensation.* Corrected fifth printing 2005. New York: Springer.

Hass, H. (1951). “Uber den einfluss eines einfachechos an die horsamkeit con sprache”. In: *Acta Acoustica united with Acoustica* 1, pp. 49–58.

Heeren, W., Rennies, J., and Verhey, J. L. (2011). “Spectral loudness summation of nonsimultaneous tone pulses”. In: *Journal of the Acoustical Society of America* 130.6, pp. 3905–3915.

Hellman, R. P. (1985). “Perceived magnitude of two-tone-noise complexes: loudness, annoyance, and noisiness”. In: *Journal of the Acoustical Society of America* 77.4, pp. 1497–1504.

Hesterberg, T. C., Monaghan, S., Moore, D. S., Epstein, R., and Rucker, J. (2003). “Bootstrap methods and permutation tests”. In: *The Practice of Business Statistics*. Ed. by Moore David S.and McCabe, G. P., Duckworth, W. M., and Sclove, S. L. New York: W. H. Freeman. Chap. 18 (Companion chapter).

Hots, J., Rennies, J., and Verhey, J. L. (2013). “Loudness of sounds with a subcritical bandwidth: a challenge to current loudness models?” In: *Journal of the Acoustical Society of America* 134.4, pp. 334–339.

Houtsma, A. J. M. (1980). “Intensity perception XI. experimental results on the relation of intensity resolution to loudness matching”. In: *Journal of the Acoustical Society of America* 68.3, pp. 807–813.

Hu, Y. and Loizou, P. (2004). “Incorporating a psychoacoustical model in frequency domain speech enhancement”. In: *IEEE Signal Processing Letters* 11.2, pp. 270–273.

IRCAM (2003). *Listen HRTF database*. Online. URL: `http://recherche.ircam.fr/equipes/salles/listen/index.html`.

Irino, T. and Patterson, R. D. (2001). “A compressive gammachirp auditory filter for both physiological and psychophysical data”. In: *Journal of the Acoustical Society of America* 109.5, pp. 2002–2008.

Irino, T. and Patterson, R. D. (1997). “A time-domain, level-dependent auditory filter: the gammachirp”. In: *Journal of the Acoustical Society of America* 101.1, pp. 412–419.

ISO 131 (1959). *Expression of the physical and subjective magnitudes of sound*. Tech. rep. International Organisation for Standardization.

ISO 226 (1987). *Acoustics – normal equal-loudness-level contours*. Tech. rep. Internation Organization for Standardization.

— (2003). *Acoustics – normal equal-loudness-level contours*. Tech. rep. International Organization for Standardization.

ISO 389-7 (2005). *Acoustics – reference zero for the calibration of audiometric equipment – part 7: reference threshold of hearing under free-field and diffuse-field listening conditions*. Tech. rep. International Organization for Standardization.

ISO 532 (1966). *Acoustics – method for calculating loudness level*. Tech. rep. International Organization for Standardization.

— (1975). *Acoustics – method for calculating loudness level*. Tech. rep. International Organization for Standardization.

ISO 532-2 (2014). *Acoustics – method for calculating loudness – part 2: Moore-Glasberg method (draft)*. Tech. rep. International Organization for Standardization.

ISO/IEC 11172-3 (1993). *Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 3: Audio*. Tech. rep. International Organisation for Standardization.

ITU BS.1534 (2015). *Method for the subjective assessment of intermediate quality level of coding systems*. Tech. rep. 3. International Telecommunication Union.

ITU-R BS.1770 (2006). *Algorithms to measure audio programme loudness and true-peak audio level*. Tech. rep. International Telecommunications Union.

— (2012). *Algorithms to measure audio programme loudness and true-peak audio level*. Tech. rep. 3. International Telecommunication Union.

— (2015). *Algorithms to measure audio programme loudness and true-peak audio level*. Tech. rep. 4. International Telecommunication Union.

ITU-R BS.1771 (2006). *Requirements for loudness and true-peak indicating meters bs series'*. Tech. rep. International Telecommunication Union.

ITU-R SRG-3 (2002). *Status report (2) - document 6p/145-e.* Tech. rep. International Telecommunications Union.

Jacobsen, E. and Lyon, R. (2003). “The sliding DFT”. In: *Signal Processing Magazine, IEEE*, pp. 74–80.

Jacobsen, E. and Lyons, R. (2004). “Tips & tricks”. In: *Signal Processing Magazine, IEEE*, pp. 1–2.

Jain, R. and Chlamtac, I. (1985). “The P^2 algorithm for dynamic calculation of quantiles and histograms without storing observations”. In: *Communications of the ACM* 28.10, pp. 1076–1085.

Jesteadt, W. and Leibold, L. J. (2011). “Loudness in the laboratory, part 1: steady-state sounds”. In: *Loudness*. Ed. by Florentine, M., Popper, A. N., and Fay, R. R. Vol. 37. Springer Handbook of Auditory Research. New York: Springer. Chap. 5, pp. 17–56.

Johannesma, P. I. M. (1972). “The pre-response stimulus ensemble of neurons in the cochlear nucleus”. In: *Proceedings of the Symposium on Hearing Theory*, pp. 58–69.

Jones, B. L. and Torick, E. L. (1982). “A new loudness indicator for use in broadcasting”. In: *Proceedings of the 71st Audio Engineering Society Convention.*

Jones, E., Oliphant, T., and Peterson, P. (2001). *SciPy: Open source scientific tools for Python.* Comp. software. URL: http://www.scipy.org/.

Julstrom, S. and Tichy, T. (1983). “Direction-sensitive gating: a new approach to automatic mixing”. In: *Proceedings of the 73rd Audio Engineering Society Convention.*

Karjalainen, M. (1985). “A new auditory model for the evaluation of sound quality of audio systems”. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.*

Klapuri, A. (1999). “Sound onset detection by applying psychoacoustic knowledge”. In: *Proceedings of The IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP).*

Kleinschmidt, M., Tchorz, J., and Kollmeier, B. (2001). “Combining speech enhancement and auditory feature extraction for robust speech recognition”. In: *Speech Communication* 34.1-2, pp. 75–91.

Kraft, D. (1988). *A software package for sequential quadratic programming.* Tech. rep. DFVLR-FB 88-28. DLR German Aerospace Center: DLR German Aerospace Center – Institute for Flight Mechanics.

Krishnamoorthi, H. (2011). “Incorporating auditory models in speech/audio appllications”. PhD thesis. Tempe, Arizona: Arizona State University.

Krishnamoorthi, H., Berisha, V., and Spanias, A. (2008). “A low-complexity loudness estimation algorithm”. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing.*

Krishnamoorthi, H., Spanias, A., and Berisha, V. (2009). “A frequency/detector pruning approach for loudness estimation”. In: *IEEE Signal Processing Letters* 16.11.

Leibold, L. J., Tan, H., Khaddam, S., and Jesteadt, W. (2007). “Contributions of individual components to the overall loudness of a multitone complex.” In: *Journal of the Acoustical Society of America* 121.5, pp. 2822–2831.

Levitt, H. (1971). “Transformed up down methods in psychoacoustics”. In: *Journal of the Acoustical Society of America* 49.2, pp. 467–477.

Loftus, G. R. and Masson, M. E. J. (1994). “Using confidence intervals in within-subject designs”. In: *Psychonomic Bulletin & Review* 1.4, pp. 476–490.

Lokki, T. and Karjalainen, M. (2002). “Analysis of room responses, motivated by auditory perception”. In: *Journal of New Music Research* 31.2, pp. 163–169.

Lopez, S. V. (2010). "Content-based processing for masking minimization in multi-track recordings". Masters thesis. Barcelona, Spain: Pompeu Fabra University.

Lopez, S. V. and Janer, J. (2010). "Quantifying masking in multi-track recordings". In: *Sound and Music Computing Conference.*

Lopez-Poveda, E. A. and Meddis, R. (2001). "A human nonlinear cochlear filterbank". In: *Journal of the Acoustical Society of America* 110.6, pp. 3107–3118.

Lydolf, M. and Møller, H. (2000). "Measurements of equal-loudness contours between 20 Hz and 1 kHz". In: *Proceedings of the 29th International Congress and Exhibition on Noise Control Engineering (InterNoise).*

Ma, Z., Reiss, J. D., and Black, D. (2014). "Partial loudness in multitrack mixing". In: *Proceedings of the 53rd Audio Engineering Society International Conference on Semantic Audio.*

Maddams, J. A., Finn, S., and Reiss, J. D. (2012). "An autonomous method for multi-track dynamic range compression". In: *Proceedings of the 15th International Conference on Digital Audio Effects.*

Mansbridge, S., Finn, S., and Reiss, J. D. (2012a). "An autonomous system for multi-track stereo pan positioning". In: *Proceedings of the 133rd Audio Engineering Society Convention.*

— (2012b). "Implementation and evaluation of autonomous multi-track fader control". In: *Proceedings of the 132nd Audio Engineering Society Convention.*

Marks, L. E. and Florentine, M. (2011). "Measurement of loudness, part 1: methods, problems, and pitfalls". In: *Loudness.* Ed. by Florentine, M., Popper, A. N., and Fay, R. R. 1st ed. Vol. 37. Springer Handbook of Auditory Research. New York: Springer. Chap. 2, pp. 17–57.

Marozeau, J. (2011). "Models of loudness". In: *Loudness.* Ed. by Florentine, M., Popper, A., and Fay, R. 1st ed. Vol. 37. Springer Handbook of Auditory Research. New York: Springer. Chap. 10, pp. 261–284.

Matelijan, I. (2014). *ARTA program for impulse response measurement and real time analysis of spectrum and frequncy response.* User Manual.

McCartney, J. (1996). "Supercollider: a new real time synthesis language". In: *Proceedings of the International Computer Music Conference.*

McDermott, H. J., McKay, C. M., Richardson, L. M., and Henshall, K. R. (2003). "Application of loudness models to sound processing for cochlear implants." In: *Journal of the Acoustical Society of America* 114.1 (4 Pt 1), pp. 2190–2197.

McDermott, H. and Varsavsky, A. (2009). "Better fitting of cochlear implants: modeling loudness for acoustic and electric stimuli". In: *Journal of Neural Engineering* 6.6, pp. 1–8.

McGill, W. J. and Goldberg, J. P. (1968). "A study of the near-miss involving weber's law and pure-tone intensity discrimination". In: *Perception & Psychophysics* 4.2, pp. 105–109.

Meddis, R. (1988). "Simulation of auditory-neural re uptake". In: *Journal of the Acoustical Society of America* 83.3, pp. 1056–1063.

— (1986). "Simulation of mechanical to neural transduction in the auditory receptor". In: *Journal of the Acoustical Society of America* 79.3, pp. 702–711.

Meddis, R., Lopez-Poveda, E. A., Popper, A. N., and Fay, R. R. (2010). *Computational models of the auditory system.* Vol. 16. 5. New York: Springer.

Meddis, R., O'Mard, L. P., and Lopez-poveda, E. A. (2001). "A computational algorithm for computing nonlinear auditory frequency selectivity." In: *Journal of the Acoustical Society of America* 109.6, pp. 2852–2861.

Meunier, S., Marchioni, A., and Rabau, G. (2000). "Subjective evaluation of loudness models using synsthesised and environmental sounds". In: *Proceedings of the 29th International Congress and Exhibition on Noise Control Engineering.*

Mintz, F. (1952). “A loudness chart for octave-band data on complex sounds”. In: *Journal of the Acoustical Society of America* 24.1, p. 80.

Moore, B. C. J. (2013). *An introduction to the psychology of hearing.* 6th ed. Leiden: Brill.

— (2014). “Development and current status of the “Cambridge” loudness models”. In: *Trends in Hearing* 18, pp. 1–29.

— (1995). *Hearing.* 2nd ed. Handbook of perception and cognition. London: Academic Press, Elsevier.

— (1996). “Masking in the human auditory system”. In: *Collected Papers on Digital Audio Bit-Rate Reduction*, pp. 9–19.

— (2000). “Use of a loudness model for hearing aid fitting. IV. Fitting hearing aids with multi-channel compression so as to restore ‘normal’ loudness for speech at different levels”. In: *British Journal of Audiology* 34.3, pp. 165–177.

Moore, B. C. J., Alcantara, J. I., Stone, M. A., and Glasberg, B. R. (1999a). “Use of a loudness model for hearing aid fitting: II. Hearing aids with multi-channel compression”. In: *British Journal of Audiology* 33.3, pp. 157–170.

Moore, B. C. J. and Glasberg, B. R. (2004). “A revised model of loudness perception applied to cochlear hearing loss”. In: *Hearing Research* 188.1, pp. 70–88.

— (1997). “A revision of Zwicker’s loudness model”. In: *Acta Acoustica united with Acoustica* 2.2, pp. 335–345.

Moore, B. C. J. and Glasberg, B. R. (1987). “Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns”. In: *Hearing research* 28.2-3, pp. 209–225.

Moore, B. C. J. and Glasberg, B. R. (2007). “Modeling binaural loudness”. In: *Journal of the Acoustical Society of America* 121.3, pp. 1604–1612.

— (1983). “Suggested formulae for calculating auditory-filter bandwidths and excitation patterns”. In: *Journal of the Acoustical Society of America* 74.3, pp. 750–753.

Moore, B. C. J. and Glasberg, B. R. (1998). “Use of a loudness model for hearing-aid fitting. I. Linear hearing aids.” In: *British journal of audiology* 32.5 (5), pp. 317–335.

Moore, B. C. J., Glasberg, B. R., and Baer, T. (1997). “A model for the prediction of thresholds, loudness, and partial loudness”. In: *Journal of the Audio Engineering Society* 45.4, pp. 224–240.

Moore, B. C. J., Glasberg, B. R., and Stone, M. A. (1999b). “Use of a loudness model for hearing aid fitting: III. A general method for deriving initial fittings for hearing aids with multi-channel compression”. In: *British Journal of Audiology* 33.4, pp. 241–258.

— (2003). “Why are commercials so loud? - perception and modeling of the loudness of amplitude-compressed speech”. In: *Journal of the Audio Engineering Society* 51.12, pp. 1123–1132.

Moore, B. C. J., Launer, S., Vickers, D. A., and Baer, T. (1998). “Loudness of modulated sounds as a function of modulation rate, modulation depth, modulation waveform and overall level”. In: *Pyschophysical and Physiological Advances in Hearing.* Ed. by Palmer, A. R., Rees, A., Summerfield, A. Q., and Meddis, E. Whurr, London, pp. 465–472.

Moore, B. C. J., Vickers, D. A., Baer, T., and Launer, S. (1999c). “Factors affecting the loudness of modulated sounds”. In: *Journal of the Acoustical Society of America* 105.5, pp. 2757–72.

Moore, B. C., Glasberg, B. R., Plack, C. J., and Biswas, A. K. (1988). “The shape of the ear’s temporal window”. In: *Journal of the Acoustical Society of America* 83.3, pp. 1102–16.

Moorer, J. A. (2000). “Audio in the new millennium.” In: *Journal of the Audio Engineering Society* 48, pp. 490–498.

Morey, R. D. (2008). “Confidence intervals from normalized data: a correction to Cousineau (2005)”. In: *Tutorials in Quantitative Methods for Psychology* 4.2, pp. 61–64.

Munson, W. A. (1947). “The growth of auditory sensation”. In: *Journal of the Acoustical Society of America* 19, pp. 584–591.

Nemenyi, P. (1963). “Distribution-free multiple comparisons”. PhD thesis. New Jersey: Princeton University.

Neunaber, B. (2008). *Parameter quantization in direct-form recursive audio filters.* Tech. rep. QSC Audio Products.

O’Brien, F. and Cousineau, D. (2014). “Representing error bars in within-subject designs in typical software packages”. In: *The Quantitative Methods for Psychology* 10.1, pp. 56–67.

O’Connor, K. N. and Puria, S. (2006). “Middle ear cavity and ear canal pressure-driven stapes velocity responses in human cadaveric temporal bones.” In: *Journal of the Acoustical Society of America* 120.3, pp. 1517–1528.

OProfile (2015). *Home page.* online. URL: `http://oprofile.sourceforge.net/news/`.

Oxenham, A. J. and Plack, C. J. (2000). “Effects of masker frequency and duration in forward masking: further evidence for the influence of peripheral nonlinearity.” In: *Hearing research* 150.1-2, pp. 258–266.

Oxenham, A. J. and Moore, B. C. J. (1994). “Modeling the additivity of nonsimultaneous masking”. In: *Hearing Research* 80.1, pp. 105–118.

Paillard, B., Mabilleau, P., Morissette, S., and Soumagne, J. (1992). “PERCEVAL: perceptual evaluation of the quality of audio signals”. In: *Journal of the Audio Engineering Society* 40.1/2, pp. 21–31.

Patterson, R. D. (1976). “Auditory filter shapes derived with noise stimuli”. In: *Journal of the Acoustical Society of America* 59.3, pp. 640–654.

Patterson, R. D. and Moore, B. C. J. (1986). “Auditory filters and excitation patterns as representations of frequency resolution”. In: *Frequency selectivity in hearing.* Ed. by Moore, B. C. J. London: Academic Press, Elsevier, pp. 123–177.

Patterson, R. D. and Nimmo-Smith, I. (1980). “Off-frequency listening and auditory-filter asymmetry”. In: *Journal of the Acoustical Society of America* 67.1, pp. 229–245.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1987). *Annex B of the SVOS final report: an efficient auditory filterbank based on the gammatone function.* Tech. rep.

Pestana, P. D. (2013). “Automatic mixing systems using adaptive audio effects”. PhD thesis. Catholic University of Portugal.

Pestana, P. D. and Alvaro, B. (2012). “Accuracy of ITU-R BS.1770 algorithm in evaluating multitrack material”. In: *Proceedings of the 133rd Audio Engineering Society Convention.*

Pestana, P. D., Reiss, J. D., and Barbosa, A. (2013). “Loudness measurement of multitrack audio content using modifications of itu-r bs.1770”. In: *Proceedings of the 134th Audio Engineering Society Convention.*

Pestana, P. and Reiss, J. D. (2014). “Intelligent audio production strategies informed by best practices”. In: *Proceedings of the 53rd Audio Engineering Society International Conference on Semantic Audio.*

Pickles, J. (2008). *An introduction to the physiology of hearing.* 3rd ed. Bradford: Emerald.

Pinheiro, J. C. and Bates, D. B. (2000). *Mixed-effects models in S and S-Plus.* New York: Springer.

Plack, C. J. and Moore, B. C. J. (1990). “Temporal window shape as a function of frequency and level”. In: *Journal of the Acoustical Society of America* 87.5, pp. 2178–2187.

Plack, C. J. and Oxenham, A. J. (2002). “Linear and nonlinear processes in temporal masking”. In: *Acta Acoustica united with Acoustica* 88, pp. 348–358.

Plack, C. J. and Oxenham, A. J. (1998). “Basilar-membrane nonlinearity and the growth of forward masking”. In: *Journal of the Acoustical Society of America* 103.3, pp. 1598–1608.

Pohlert, T. (2016). “The pairwise multiple comparison of mean ranks package (pmcmr)”. In: *R package*, pp. 1–9. URL: `http://cran.r-project.org/package=PMCMR`.

Puria, S., Peake, W., and Rosowski, J. (1997). “Sound-pressure measurements in the cochlear vestibule of human-cadaver ears”. In: *Journal of the Acoustical Society of America* 101.5, pp. 2754–2770.

Rayleigh, L. (1907). “Our perception of sound duration”. In: *Philosophical Magazine* 13.74.

Reddy, T. A. (2011). *Applied data analysis and modeling for energy engineering and scientists.* New York: Springer.

Reiss, J. D. (2011). “Intelligent systems for mixing multichannel audio”. In: *Proceedings of the 17th International Conference on Digital Signal Processing.*

Rennies. J., Verhey, J. L., Chalupper, J., and Fastl, H. (2009). “Modeling temporal effects of spectral loudness summation”. In: *Acta Acoustica united with Acoustica* 95, pp. 1112–1122.

Rennies, J. and Verhey, J. L. (2009). “Comparison of loudness models for artificial and environmental sounds”. In: *Proceedings of the International Conference on Acoustics (NAG/DAGA).*

Rennies, J., Verhey, J. L., and Fastl, H. (2010). “Comparison of loudness models for time-varying sounds”. In: *Acta Acoustica united with Acoustica* 96, pp. 383–396.

Rennies, J., Verhey, J. L., Appell, J. E., and Kollmeier, B. (2013). “Loudness of complex time-varying sounds? A challenge for current loudness models”. In: *Proceedings of Meetings on Acoustics.* Vol. 19, pp. 050189–050189.

Richardson, L. F. and Ross, J. S. (1930). “Loudness and telephone current”. In: *Journal of General Psychology* 3, pp. 288–306.

Riesz, R. R. (1928). “Differential intensity sensitivity of the ear for pure tones”. In: *Psychological review* 31, pp. 867–875.

Robinson, D. W. and Dadson, R. S. (1956). “A re-determination of the equal-loudness relations for pure tones”. In: *British Journal of Applied Physics* 7, pp. 166–181.

Robinson, D. J. M. (2002). “Perceptual model for assessment of coded audio”. PhD thesis. Department of Electronic Systems Engineering, University of Essex.

Robles, L. and Ruggero, M. (2001). “Mechanics of the mammalian cochlea”. In: *Psychological review* 81, pp. 1305–1352.

Ruggero, M. A., Rich, N. C., Recio, A., and Narayan, S. S. (1997a). “Basilar-membrane responses to tones at the base of the chinchilla cochlear”. In: *Journal of the Acoustical Society of America* 101.4, pp. 2151–2163.

Ruggero, M. A., Rich, N. C., Recio, A., Narayan, S. S., and Robles, L. (1997b). “Basilar-membrane responses to tones at the base of the chinchilla cochlea.” In: *Journal of the Acoustical Society of America* 101.4, pp. 2151–63.

Scharf, B. (1961). “Complex sounds and critical bands”. In: *Psychological Bulletin* 58.3, pp. 205–217.

— (1969). “Dichotic summation of loudness”. In: *Journal of the Acoustical Society of America* 45.5, pp. 1193–1205.

Scharf, B. (1959). “Loudness of complex sounds as a function of the number of components”. In: *Journal of the Acoustical Society of America* 31.6, p. 783.

— (1964). “Partial masking”. In: *Acta Acoustica united with Acoustica* 14.

Scheuren, J. (2014). “ISO 532 – living and working with alternative loudness standards”. In: *Proceedings of the 43rd International Congress on Noise Control Engineering.*

Schlittenlacher, J. and Ellermeier, W. (2015). “Spectral loudness summation: shortcomings of current standards”. In: *Journal of the Acoustical Society of America* 137, EL26–EL31.

Schlittenlacher, J., Ellermeier, W., and Hashimoto, T. (2012a). "Loudness model extension improving predictions for broadband sounds". In: *Proceedings of the 41st International Congress on Noise Control Engineering*.

Schlittenlacher, J., Hashimoto, T., and Fastl, H. (2011). "Loudness of pink noise and stationary technical sounds". In: *Proceedings of the 40th International Congress on Noise Control Engineering*.

Schlittenlacher, J., Fastl, H., Hashimoto, T., Kuwano, S., and Namba, S. (2012b). "Differences of loudness algorithms across the frequency spectrum". In: *38th German Annual Conference on Acoustics (DAGA 2012)*.

Schofield, D. (1985). *Visualisations of speech based on a model of the peripheral auditory system.* Tech. rep. UK National Physical Laboratory Report 62/85, HMSO.

Scott, H. H. (1957). "Historical development of the sound level meter". In: *Journal of the Acoustical Society of America* 29.12, pp. 1331–1333.

Seefeldt, A. (2007). "Loudness domain signal processing". In: *Proceedings of the 123rd Audio Engineering Society Convention*.

Seefeldt, A., Crockett, B., and Smithers, M. (2004). "A new objective measure of perceived loudness". In: *Proceedings of the 117th Audio Engineering Society*.

Seefeldt, A. and Lyman, S. (2006). "A comparison of various multichannel loudness measurement techniques". In: *Proceedings of the 121st Audio Engineering Society Convention*.

Senior, M. (2011). *Mixing secrets for the small studio*. Massachusetts: Focal Press.

Shaw, E. A. (1974). "Transformation of sound pressure level from the free field to the eardrum presented in the horizontal plane." In: *Journal of the Acoustical Society of America* 56.6, pp. 1848–1861.

Shaw, E. A. and Vaillancourt, M. M. (1985). "Transformation of sound-pressure level from the free field to the eardrum presented in numerical form". In: *Journal of the Acoustical Society of America* 78.3, pp. 1120–1122.

Shin, J. W. and Kim, N. S. (2007). "Perceptual reinforcement of speech signal based on partial specific loudness". In: *IEEE Signal Processing Letters* 14.11, pp. 887–890.

Simpson, A. J. R. and Reiss, J. D. (2011). "The effect of loudness overflow on equal-loudness-level contours". In: *Proceedings of the 130th Audio Engineering Society Convention*.

Skovenborg, E. and Nielsen, S. H. (2004a). "Evaluation of designs for loudness-matching experiments". In: *Proceedings of the International Conference on the Subjective and Objective Assessment of Sound (SOAS)*.

— (2004b). "Evaluation of different loudness models with music and speech material". In: *Proceedings of the 117th Audio Engineering Society Convention*.

Skovenborg, E. and Nielsen, S. H. (2007). "Real-time visualisations of loudness along different time scales". In: *Proceedings of the 10th International Conference on Digital Audio Effects*.

Skovenborg, E., Quesnel, R., and Nielsen, S. H. (2004). "Loudness assessment of music and speech". In: *Proceedings of the 116th Audio Engineering Society Convention*.

Slaney, M. (1993). *An efficient implementation of the auditory filter bank*. Apple computer technical report 35. Advanced Technology Group, Apple Computer, Inc.

Smith III, J. O. (2008). *Introduction to digital filters with audio applications*. 2nd ed. W3K Publishing. URL: `www.w3k.org/books`.

— (2011). *Spectral audio signal processing*. W3K Publishing. URL: `www.w3k.org/books`.

Soulodre, G. A. (2004). "Evaluation of objective loudness meters". In: *Proceedings of the 116th Audio Engineering Society Convention*.

Soulodre, G. A. and Lavoi, M. C. (2005). "Stereo and multichannel loudness perception and metering". In: *Proceedings of the 119th Audio Engineering Society Convention.*

Soulodre, G. A., Lavoie, M. C., and Norcross, S. G. (2003). "The subjective loudness of typical program material". In: *Proceedings of the 115th Audio Engineering Society Convention.*

Soulodre, G. A. and Norcross, S. G. (2003). "Objective measures of loudness". In: *Proceedings of the 115th Audio Engineering Society Convention.*

Sporer, T. (1996). "Evaluating small impairments with the mean opinion scale-reliable or just a guess?" In: *Proceedings of the 101st Audio Engineering Society Convention.*

Sporer, T., Liebetrau, J., and Schneider, S. (2009). "Statistics of MUSHRA revisited". In: *Proceedings of the 127th Audio Engineering Society International Conference.*

Steinberg, J. C. (1937). "The dependence of hearing impairment on sound intensity". In: *Journal of the Acoustical Society of America* 9.1, p. 11.

Stevens, S. S. (1936). "A scale for the measurement of a psychological magnitude: loudness". In: *Psychological review* 43.5, pp. 405–416.

— (1957a). "Calculating loudness". In: *Noise Control* 3.11, pp. 11–22.

— (1956a). "Calculation of the loudness of complex noise". In: *Journal of the Acoustical Society of America* 28.5, pp. 807–832.

— (1957b). "On the psychophysical law". In: *The Psychological Review* 64.3, pp. 153–181.

— (1966). "Power-group transformations under glare, masking and recruitment." In: *Journal of the Acoustical Society of America* 39, pp. 725–735.

— (1961a). "Procedure for calculating loudness: mark VI". In: *Journal of the Acoustical Society of America* 33.11, pp. 1577–1585.

— (1975). *Psychophysics: introduction to its perceptual, neural, and social prospects.* New York: Wiley.

— (1956b). "The direct estimation of sensory magnitudes: loudness". In: *The American Journal of Psychology* 69.1, pp. 1–25.

— (1955). "The measurement of loudness". In: *Journal of the Acoustical Society of America* 27.5, pp. 815–829.

— (1961b). "To honor Fechner and repeal his law: a power function, not a log function, describes the operating characteristic of a sensory system". In: *Science* 133.3446, pp. 80–86.

Stone, M. A., Moore, B. C. J., and Glasberg, B. R. (1997). "A real-time DSP-based loudness meter". In: *Contributions to Psychological Acoustics.* Ed. by Schick, A. and Klatte, M. Oldenburg: Bibliotheks- und Informationssystem der Universität Oldenburg, pp. 587–601.

Supper, B., Brookes, T., and Rumsey, F. (2006). "An auditory onset detection algorithm for improved automatic source localization". In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.3, pp. 1008–1017.

Suzuki, Y. and Takeshima, H. (2004). "Equal-loudness-level contours for pure tones." In: *Journal of the Acoustical Society of America* 116.2, pp. 918–933.

Tan, C.-T., Moore, B. C. J., Zacharov, N., and Mattila, V.-V. (2004). "Predicting the perceived quality of nonlinearly distorted music and speech signals". In: *Journal of the Audio Engineering Society* 52.7/8, pp. 699–711.

Terhardt, E. (1979). "Calculating virtual pitch". In: *Hearing research* 1, pp. 155–182.

Terrell, M. J. (2013). "Perceptual mixing for musical production". PhD thesis. Queen Mary University of London.

Terrell, M. J. and Reiss, J. D. (2009). "Acoustic monitor mixing for live musical performance". In: *Journal of the Audio Engineering Society* 57.11, pp. 927–936.

Terrell, M. J., Simpson, A. J. R., and Sandler, M. B. (2013). “A perceptual audio mixing device”. In: *Proceedings of the 34th Audio Engineering Society Convention.*

— (2014). “The mathematics of mixing”. In: *Journal of the Audio Engineering Society* 62.1, pp. 4–13.

Texas Instruments (2013). *LM3916 dot/bar display driver.* Tech. rep. Texas Instruments.

Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., and Colomes, C. (2000). “PEAQ - the ITU standard for objective measurement of perceived audio quality”. In: *Journal of the Audio Engineering Society* 48.1/2, pp. 3–29.

Tuomi, O. and Zacharov, N. (2000). “A real-time binaural loudness meter”. In: *Proceedings of the 139th meeting of the Acoustical Society of America.*

Verhey, J. L. and Kollmeier, B. (2002). “Spectral loudness summation as a function of duration”. In: *Journal of the Acoustical Society of America* 111.3, pp. 1349–1358.

Viemeister, N. F. and Bacon, S. P. (1988). “Intensity discrimination, increment detection, and magnitude estimation for 1-kHz tones”. In: *Journal of the Acoustical Society of America* 84, pp. 172–178.

Wallach, H., Newman, E. B., and Rosenzweig, M. R. (1949). “The precedence effect in sound localization”. In: *The American journal of psychology* 62.3, pp. 315–336.

Ward, D., Athwal, C., and Kokuer, M. (2013). “An efficient time-varying loudness model”. In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*

Ward, D., Enderby, S., Athwal, C., and Reiss, J. D. (2015). “Real-time excitation based binaural loudness meters”. In: *Proceedings of the 18th International International Conference on Digital Audio Effects.*

Ward, D. and Reiss, J. D. (2016). “Loudness algorithms for automatic mixing”. In: *Proceedings of the 2nd Audio Engineering Society Workshop on Intelligent Music Production.*

Ward, D., Reiss, J. D., and Athwal, C. (2012). “Multi-track mixing using a model of loudness and partial loudness”. In: *Proceedings of the 133rd Audio Engineering Society Convention.*

White, P. (2008). “Automation for the people”. In: *Sound on Sound Magazine.*

Wichern, G., Wishnick, A. S., Lukin, A., and Robertson, H. (2015). “Comparison of loudness features for automatic level adjustment in mixing”. In: *Proceedings of the 139th Audio Engineering Society Convention.*

Widmann, U., Lippold, R., and Fastl, H. (1998). “A computer program simulating post-masking for applications in sound analysis systems”. In: *Proceedings of the INTER-NOISE and NOISE-CON Congress and Conference.*

Wilcox, R. (2012). *Introduction to robust estimation and hypothesis testing (statistical modeling and decision science).* 3rd ed. Massachusetts: Academic Press, Elsevier.

Wilson, A. and Fazenda, B. M. (2015). “Navigating the mix-space: theoretical and practical level-balancing technique in multitrack music mixtures”. In: *Proceedings of the 12th Sound and Music Conference.*

Winter, B. (2011). *A very basic tutorial for performing linear mixed effects analyses.* Tech. rep. Tutorial 2.

Wobbrock, J. O., Findlater, L., Gergle, D., and Higgins, J. J. (2011). “The aligned rank transform for nonparametric factorial analyses using only anova procedures”. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems.*

Yost, W. A. (2007). *Fundamentals of hearing: an introduction.* 5th ed. Massachusetts: Academic Press, Elsevier.

Zacharov, N. (1998). “An overview of multichannel level alignment”. In: *Proceedings of the 15th Audio Engineering Society International Conference on Audio, Acoustics & Small Spaces.*

Zhang, C. and Zeng, F.-g. (1997). "Loudness of dynamic stimuli in acoustic and electric hearing". In: *Journal of the Acoustical Society of America* 102.5, pp. 2925–2934.

Zorilă, T.-C., Stylianou, Y., Flanagan, S., and Moore, B. C. J. (2016). "Effectiveness of a loudness model for time-varying sounds in equating the loudness of sentences subjected to different forms of signal processing." In: *Journal of the Acoustical Society of America* 140.1, pp. 402–408.

Zwicker, E. (1984). "Dependence of post-masking on masker duration and its relation to temporal effects in loudness". In: *Journal of the Acoustical Society of America* 75.1, pp. 219–223.

Zwicker, E. (1969). "Der einfluss der zeitlichen struktur von tönen auf die addition von teillautheiten (influence of the temporal structure of tones on the addition of partial loudnesses)". In: *Acta Acoustica united with Acoustica* 21, pp. 16–25.

Zwicker, E. (1960). "Ein verfahren zur berechnung der lautstarke (a procedure for calculating loudness)". In: *Acta Acoustica united with Acoustica* 10, pp. 304–308.

— (1977). "Procedure for calculataing loudness of temporally variable sounds". In: *Journal of the Acoustical Society of America* 62.3, pp. 675–682.

— (1961). "Subdivision the frequency range over which the human". In: *Journal of the Acoustical Society of America* 33.2, p. 248.

— (1958). "Über psychologische und methodische grundlagen der lautheit (on psychological and methodological foundations of the loudness)". In: *Acustica united with Acustica* 8, pp. 237–258.

Zwicker, E., Fastl, H., Widmann, U., Kurakata, K., Kuwano, S., and Germtuo, E. R. (1991). "Program for calculating loudness according to DIN 45631 (ISO 532B)". In: *Journal of the Acoustical Society of America* 12.1, pp. 39–42.

Zwicker, E. and Feldtkeller, R. (1955). "Uber die lautstarke von gleichformigen gerauschen (on the loudness of continuous noises)". In: *Acta Acoustica united with Acoustica* 5, pp. 303–316.

Zwicker, E., Flottorp, G., and Stevens, S. S. (1957). "Critical band width in loudness summation". In: *Journal of the Acoustical Society of America* 29.5, pp. 548–557.

Zwicker, E. and Scharf, B. (1965). "A model of loudness summation". In: *Psychological review* 72.1, pp. 3–26.

Zwicker, E., Terhardt, E., and Paulus, E. (1979). "Automatic speech recognition using pyschoacoustic models". In: *Journal of the Acoustical Society of America* 65.2, pp. 487–498.

Zwicker, E. and Zwicker, U. T. (1991). "Dependence of binaural loudness summation on interaural level differences, spectral distribution, and temporal distribution". In: *Journal of the Acoustical Society of America* 89, pp. 756–764.

Zwicker, E. (1974). "Sensation and Measurement". In: ed. by Moskowitz, H. R., Scharf, B., and Stevens, J. C. Dordrecht: Springer. Chap. Loudness and excitation patterns of strongly frequency modulated tones, pp. 325–335.

Zwislocki, J. J. (1965). "Analysis of some auditory characteristics". In: *Handbook of Mathematical Psychology.* Ed. by Luce, R. D., Bush, R., and Galanter, E. Vol. 3. New York: Wiley, pp. 1–97.

Zwislocki, J. J. and Hellman, R. P. (1960). "On the "psychophysical law"". In: *Journal of the Acoustical Society of America* 32.924.