Navigating Descriptive Sub-Representations of Musical Timbre

Spyridon Stasis
Digital Media Technology Lab
Birmingham City University
Birmingham, United Kingdom
spyridon.stasis@bcu.ac.uk

Jason Hockman
Digital Media Technology Lab
Birmingham City University
Birmingham, United Kingdom
jason.hockman@bcu.ac.uk

Ryan Stables
Digital Media Technology Lab
Birmingham City University
Birmingham, United Kingdom
ryan.stables@bcu.ac.uk

ABSTRACT

Musicians, audio engineers and producers often make use of common timbral adjectives to describe musical signals and transformations. However, the subjective nature of these terms, and the variability with respect to musical context often leads to inconsistencies in their definition. In this study, a model is proposed for controlling an equaliser by navigating clusters of datapoints, which represent grouped parameter settings with the same timbral description. The associated interface allows users to identify the nearest cluster to their current parameter setting and recommends changes based on its relationship to a cluster centroid. To do this, we apply dimensionality reduction to a dataset of equaliser curves described as warm and bright using a stacked autoencoder, then group the entries using an agglomerative clustering algorithm with a coherence-based distance criterion. To test the efficacy of the system, we implement listening tests and show that subjects are able to match datapoints to their respective sub-representations with 93.75% mean accuracy.

Author Keywords

audio equalisation, intelligent music production, digital audio effects, neural networks, semantic audio

ACM Classification

H.5.5 [Information Interfaces and Presentation] Sound and Music Computing, H.5.2 [Information Interfaces and Presentation] User Interfaces

1. MUSICAL TIMBRE

The study of perceived musical timbre has been widely researched [9, 10], often in relation to a vocabulary of descriptive terms [25, 21]. These adjectives are commonly used to describe the timbre of musical signals at various points in the production process, from performance using musical instruments [8], to the application of audio effects [5, 27], and audio mastering [24]. A common methodology is to identify underlying correlations between statistical audio features and specific descriptive terms [1], often when measured in some reduced dimensionality subspace.

A recent trend in timbral research has been to incorporate crowd-sourced data, collected using a medium such as the web [5, 23, 22], or through digital audio workstation plugins [26]. This allows for large databases of adjectives to be used in music production, thus bridging the gap between amateurs and experienced producers by identifying common statistical properties of terms within a shared vocabulary. This contributes to a range of semantically driven audio processing techniques such as spectral morphing [3, 30, 4], additive synthesis [31], and psychoacoustic timbral modification [11, 18].

1.1 Audio Equalisation and Existing Interfaces

In this study, we explore the use of descriptive terms in the process of audio equalisation. Equalisation [29] is one of the most widely used audio effects in sound engineering, allowing a user to apply gain to multiple frequency bands concurrently. Semantic audio applications tend to address the problem of creative equalisation, such as enabling a user to match or retrieve a desired spectral envelope. However, corrective equalisation is also an active area of research, such as detecting and attenuating problematic frequencies in a live environment [6].

As equalisation is such a prominent aspect of the sound production process, a number of systems have been developed to facilitate the creative process. SubjEQt [16] and 2DEQ [19] for example, both manipulate parameter spaces of an equaliser by allowing the user to navigate a reduced dimensionality space. Here, the parameter-space representations of a vocabulary of adjectives is either determined by the researcher, or a small number of external samples. On a larger scale, SocialEQ [5], retrieves timbral descriptions by crowdsourcing user-inputs from a web interface. Here, the system is able to learn equalisation curves from a subject's evaluation of a set of audio samples. Finally, the SAFE EQ [27, 28] presents crowdsourced audio effects data through a plugin interface with a reduced-dimensionality parameter space, allowing users to navigate a subset of adjectives using a 2-dimensional controller.

In all of the aforementioned systems, the assumption is that subjects exhibit significant agreement towards a statistical representation of a set of descriptive terms. This however tends not to be the case in some instances [7]. In this study, we show that terms often have various interpretations or sub-representations, leading to several clusters being attributed to a single term. This could be due to an external factor such as the instrument, genre or mood of the audio signal, or to multiple collective interpretations of a descriptor. We use these sub-representations to provide an interface that allows users to navigate the various clusters of a single term, whilst providing recommendations based on their proximity to a cluster's centroid. The system provides musicians and audio engineers with alternative parameter settings for given tasks. This allows experienced users to explore new creative directions in music production, and

NIME'17, May 15-19, 2017, Aalborg University Copenhagen, Denmark.

provides novice users with an interface to intuitively control complex parameter spaces.

2. METHODOLOGY

To build the interface, we first collect a dataset of adjectives, which describe sounds processed using an equaliser. We then build a model which reduces the dimensionality of the parameter space, then applies clustering to identify sub-representations of individual descriptors. Using the model, we develop an interface that allows users to navigate the space, based on recommended cluster centroids. Finally, we evaluate the system using subjective listening tests.

2.1 Dataset

To build the system, we extract the two most common timbral adjectives (*warm* and *bright*) from a large set of annotated audio effect settings, collected through the SAFE Project[1] [26]. In total 900 instances were collected from a Digital Audio Workstation plugin, across a range of musical instruments and genres. These terms are amongst the most widely researched in the field [3, 4, 31].

Each entry into the dataset comprises 13 parameter states, taken from a parametric EQ, a wide range of spectral and temporal audio features taken before and after the processing has been applied, user metadata, and a description of the sound transformation. The equaliser used to collect the data cascades five biquad filters in series. These are two shelving filters for low and high frequency bands, and three peaking filters. The peaking filters allow for the parametrisation of gain, centre frequency and bandwidth, while the shelving filters allow for adjustment of the gain and centre frequency values. For this study, we discard audio features and focus on modelling the equalisation curves via their respective filter parameters.

2.2 Objective Evaluation

To optimise the parameters in the model and to evaluate the effectiveness of various aspects, we apply objective metrics to the data. To identify the degree to which natural clusters exist within the low-dimensional representation of the dataset, we measure spatial randomness in the data using the Hopkins Statistic [17, 13]. Here, p points (*Set 1*), distributed randomly in the low-dimensional space, are generated and p points (*Set 2*) from the existing dataset are sampled. The distance of the points is then calculated in both *Set 1* and *Set 2* to their nearest neighbour, where dr are the nearest neighbour distances of the sampled points and di are the nearest neighbour distances of the generated points. Calculation of the Hopkins statistic is shown in Equation 1:

$$H = \frac{\sum_{i=1}^{p} di}{\sum_{i=1}^{p} dr + \sum_{i=1}^{p} di} \tag{1}$$

Similarly, to evaluate the parameterisation of sub-representations after hierarchical clustering has been applied, we measure the cophenetic correlation. This can be performed by evaluating the cophenetic distance matrices of the clustering algorithm, since dendrograms are graphical representation of a cophenetic matrix [15, 20]. To calculate the cophenetic correlation let $x(i, j)$ be the Euclidean distance between points i and j, and $d(i, j)$ the distance of the two points in the dendrogram, i.e height at which i and j are first joined. Then by letting x and d be the average values of $x(i, j)$ and $d(i, j)$ respectively, the cophenetic correlation c is calculated using Equation 2:

$$c = \frac{\sum_{i<j}(x(i,j) - x)(d(i,j) - d)}{\sqrt{[\sum_{i<j}(x(i,j) - x)^2][\sum_{i<j}(d(i,j) - d)^2]}} \tag{2}$$

2.3 Subjective Evaluation

To evaluate the extent to which the system is able to find coherent sub-representations, we implement a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) test, using the Web Audio Evaluation Toolbox [14]. Subjects were asked to rank the acoustic similarity of a reference sound to other versions of the same sound, processed with randomised equalisation curves from clusters found by the system. Our hypothesis is that sounds which are processed with curves from the same sub-representation should be perceptually more similar than curves from different clusters. For each trial, two recordings of electric guitars were used, one performed in a blues style and one performed in a metal style, and after processing, the audio samples were normalised in order to prevent variation due to differences in overall level. The number of samples presented at any one time during the experiment is based on the resulting number of clusters for each term (*warm* and *bright*). A similarity matrix of distance vectors from each reference audio sample is constructed. In total, 20 subjects participated in the listening tests, with varying levels (0-5 years) of professional audio production experience. All subjects had normal hearing and were aged between 18-40.

3. MODEL OVERVIEW

The model, as shown in Figure 1, projects a high-dimensional dataset (EQ parameters) into a low-dimensional space using a Stacked Autoencoder (SAe). Clustering is applied to the data to identify salient sub-representations within each descriptor using the agglomerative clustering algorithm described in Section 3.1. Users are then able to navigate the space, where the system will recommend parameter settings by maximising the coherence between the user-input and each of the cluster centroids. Finally, a 13-dimensional parameter array is reconstructed from the 2-dimensional input.

The SAe model [12] is used (as in [27, 28]) to project data onto a low-dimensional subspace by learning a low-dimensional representation of a vector of input features using a neural network architecture. This allows us to model complex relationships between and high-and-low dimensional spaces, and to approximate parameter settings from low-dimensional inputs using the SAe's decoder layers. We implement the system using the Theano Python Library [2]. To accurately train the model, the EQ data is normalised and rescaled during the decoding process.

3.1 Coherence-based Hierarchical Clustering

We implement a variation of the hierarchical clustering algorithm, which aims to cluster data-points based on their spectral representations. Once the instances have been separated into clusters, new input parameters will be measured against the resulting cluster centroids using the same criterion, with the system then providing guidance in achieving the sub-representation most closely related to the input. This method allows the system to take into account not only the relationship in low-dimensional space, but also the high-dimensional parameters of the EQ.

In order to apply hierarchical clustering to the frequency bins of the EQ curve, a modified distance metric based on coherence is implemented. As defined in Eq 3, $G_{ab}(f)$ is the cross-spectral density between the equalisation curve a

[1] http://www.semanticaudio.co.uk

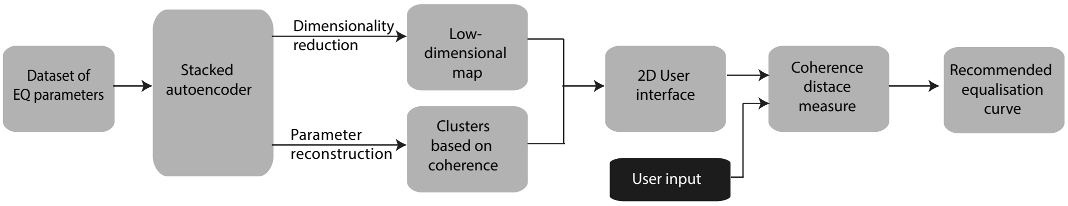


Figure 1: Block Diagram of the Proposed Model

and b, where $G_{aa}(f)$ and $G_{bb}(f)$ is the spectral density of the a and b curves respectively. A distance matrix is constructed by finding the pairwise coherence measurements between all frequency bins of the reconstructed EQ parameters, the clusters are computed sequentially, and applied to the low-dimensional map. Using this method it is expected that the resulting groups will share characteristics based on a criterion that identifies the level of similarity between frequency distributions, rather than the distribution of the points in the low-dimensional space. The resulting clusters are depicted in Figures 2 and 3.

$$C_{ab}(f) = \frac{|G_{ab}(f)|^2}{G_{aa}(f)G_{bb}(f)} \tag{3}$$

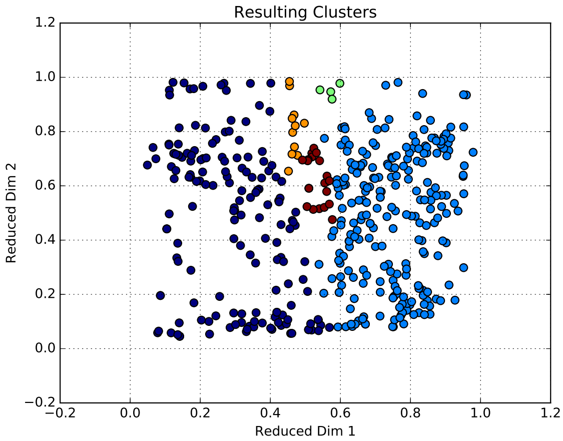


Figure 2: Resulting Clusters for the *Bright* descriptor

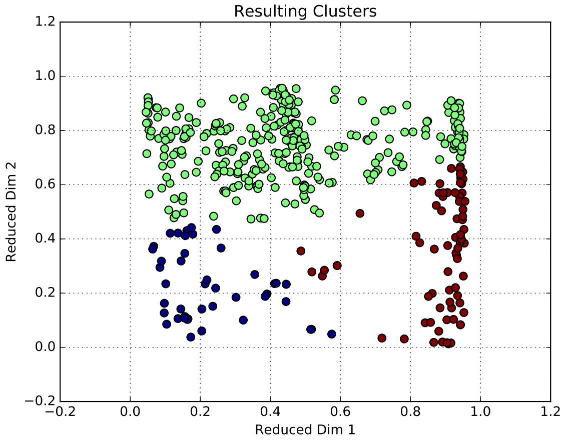


Figure 3: Resulting Clusters for the *Warm* descriptor

4. INTERFACE

The interface (shown in Figure 4) incorporates a 2 dimensional subspace with a 5-band parametric EQ[2], where settings are recommended to the user (the white line) based on the proximity of the existing curve (the blue line) to the cluster centroid (shown in 2-dimensions on the right hand side of the figure) using the coherence metric discussed in section 3.1. The low-dimensional map is computed when the interface is initialised, and the recommendation will change when the coherence with the current cluster centroid is no longer minimal.

Equalisation curve recommendations are presented through the high-and-low-dimensional interfaces concurrently, however due to the nonlinear mapping between the two spaces, the relationship between the two may be unclear. This is due to several correlated parameters being loaded onto a small number dimensions. This can make controlling the sound timbre intuitive as users can navigate the clusters in 2-dimensions, and the parameters will quantise to the nearest cluster centroid. Users of the effect can control the number of clusters selected by the algorithm, using a single threshold parameter. A high threshold will result in a low number of clusters, while a low threshold will provide more sub-representations.

5. RESULTS

5.1 Clustering Tendency

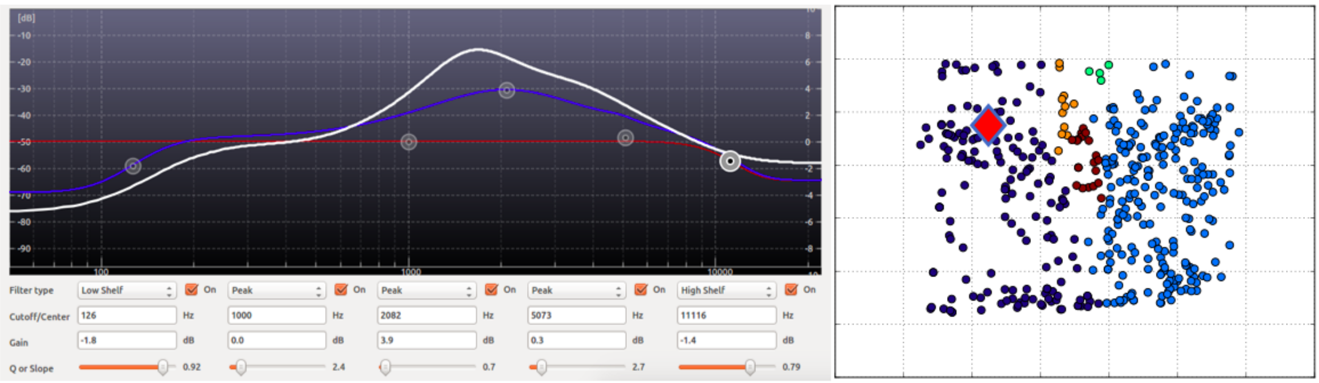
The Hopkins Statistic will produce a result from 0 to 1, where 1 is indicative of a highly clustered dataset, and 0 suggests the data points are uniformly distributed [17]. For our data, the equalisation curves described as *bright* exhibit a score of 0.544 with a standard deviation of 0.027 and the equalisation curves described as *warm* exhibit a score of 0.561 with a standard deviation of 0.043. This suggests that natural clusters may exist in the data, although clusters may be diffused and boundaries may be overlapping.

5.2 Parameter Selection

We construct linkage matrices based on the frequency bins of each point in the low-dimensional space and evaluate their cophenetic correlation. Table 1 shows that for the *bright* descriptor, the *average* linkage criterion is the most suitable, with *centroid*, *ward* and *complete* linkages exhibiting slightly less accurate measurements. Similarly, *median*, *weighted* and *single* linkage score perform less favourably. This is similar to the *warm* descriptor, with the *average* linkage criterion scoring the highest and the *centroid* method performing less favourably. Similarly, *weighted*, *complete* and *median* linkages display significantly less accurate scores and *single* linkage achieves a significantly lower score. These results are also in accordance with the results of [20], where *average* and *centroid* linkages were the most appropriate for a variety of artificially generated datasets.

To select the model's hyperparameters, we apply a hierarchical clustering process, iterating the number of clusters

[2] Based on PyEQ: https://github.com/tmwoz/pyEQ

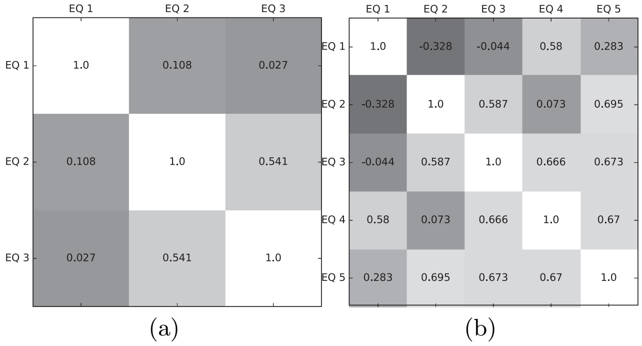


Filter type	Low Shelf	☑ On	Peak	☑ On	Peak	☑ On	Peak	☑ On	High Shelf	☑ On
Cutoff/Center	126	Hz	1000	Hz	2082	Hz	5073	Hz	11116	Hz
Gain	-1.8	dB	0.0	dB	3.9	dB	0.3	dB	-1.4	dB
Q or Slope		0.92		2.4		0.7		2.7		0.79

Figure 4: The user interface of the equaliser, with the high-dimensional parameter space on the left, and the low-dimensional space on the right. The red diamond on the left illustrates the position within the clustered sub-representation and the white line on the frequency analyser represents the recommended equaliser curve corresponding with the closest centroid.

Linkage	*Bright* Cophenetic Correlation	*Warm* Cophenetic Correlation
Single	0.320214	0.522341
Complete	0.862697	0.685418
Average	**0.878668**	**0.812211**
Weighted	0.800794	0.703954
Centroid	0.870976	0.811222
Median	0.846046	0.682768
Ward	0.862679	0.675793

Table 1: Comparison of the cophenetic correlation for both descriptors across seven different linkage options



	EQ 1	EQ 2	EQ 3
EQ 1	1.0	0.108	0.027
EQ 2	0.108	1.0	0.541
EQ 3	0.027	0.541	1.0

(a)

	EQ 1	EQ 2	EQ 3	EQ 4	EQ 5
EQ 1	1.0	-0.328	-0.044	0.58	0.283
EQ 2	-0.328	1.0	0.587	0.073	0.695
EQ 3	-0.044	0.587	1.0	0.666	0.673
EQ 4	0.58	0.073	0.666	1.0	0.67
EQ 5	0.283	0.695	0.673	0.67	1.0

(b)

Figure 5: Coherence distance Matrix for (a) *Warm* and (b) *Bright* descriptors showing coherence between clustered sub-representations

(N_{cl}) each time from 2-100. On each iteration, the coherence between the cluster centroids is measured and N_{cl} is accepted once the coherence between 2 clusters exceeds a threshold. We chose this method as an alternative to the commonly used inconsistency approach to cluster identification as it produces a varying value for N_{cl} when the global depth parameter is varied. The threshold is empirically set to +0.7, which generally signifies a strong positive correlation. Sequentially a value of 1 is assigned if a pair has a coherence greater than +0.7, and 0 otherwise. This process results in a percentage measurement of the correlation between all centroids.

The optimal value of N_{cl} for each descriptor is then considered to be the highest number of clusters in which all pairwise correlations fall below +0.7. We find that, after the optimisation process, $N_{cl} = 5$ for equaliser curves described as *bright*, and $N_{cl} = 3$ for equaliser curves described as *warm*. Their corresponding coherence matrices are illustrated in Figures 5a and 5b, and the resulting sub-representations are presented as mean equalisation curves in Figures 6a and 6b.

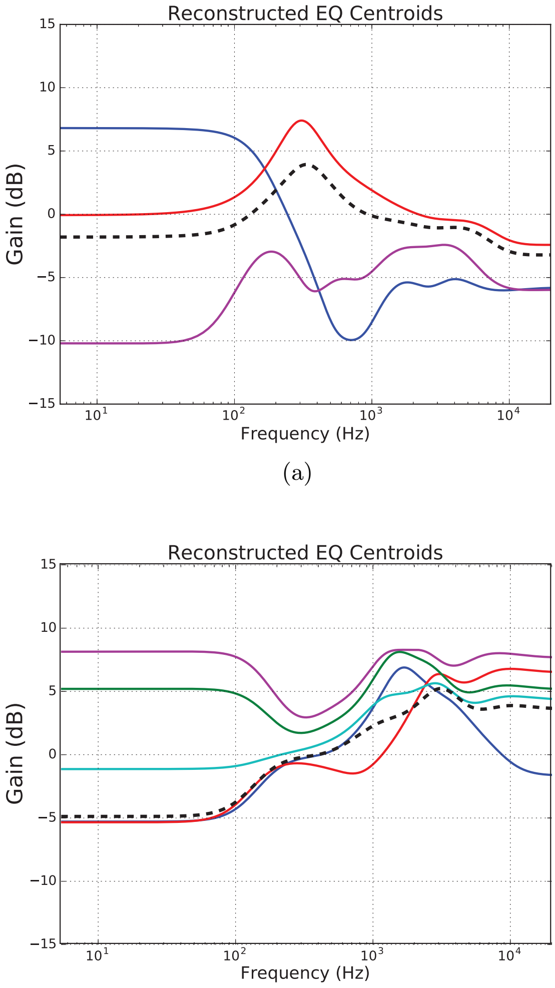
5.3 Subjective Evaluation

The reconstructed equaliser curves are applied to 2 audio samples, a blues guitar and metal guitar, 5 of which are bright and 3 of which are warm (i.e. one from each sub-representation). This results in a total of 16 samples, which were presented to each participant. The listening tests show that all participants are able to identify the reference. In the case of *warm* for both audio samples, subjects have predominantly allocated the audio sample to the predicted cluster. For the *bright* descriptor, the subjects allocate equaliser curves to the predicted cluster centroids for the metal guitar sample, with a small number of misclassifications (see Figure 7b). However, for the blues guitar sample, cluster differentiation appears to be slightly harder, with subjects allocating the correct sample in four out of the five tests, bringing the total percentage of correct identification to 93.75%. While this is not a bad score, the ratings that subjects provided display a high degree of similarity, in most cases, between at least three samples, as is displayed in Figure 7a.

This result is not unexpected, as can be deduced from Figure 5b, where comparisons of equaliser curves are able to achieve scores between +0.66 to +0.695, while the *warm* descriptor displays more diverse sub-representations, as displayed in Figure 5a. These figures show that the *warm* equaliser curves are sufficiently different for subjects to detect differences, but this may not hold true for the *bright* sub-representations.

To measure the confidence of the test participants, the standard deviation of subjects' responses is calculated. For the *warm* descriptor samples (Figures 7c and 7d), the correct response also displays the lowest standard deviation value, meaning that the subjects were more confident of their responses. However, the same behaviour is not displayed for the *bright* descriptor. For the blues guitar (Figure 7a), subjects exhibited the lowest standard deviation values, on the correct answers, for two of the five tests, the same holds true for the metal samples (Figure 7b). Therefore it can be said that the subjects displayed less certainty for their choices on the *bright* descriptor tests.

To evaluate the extent to which the perceived distances between the sub-representations is preserved in the cluster-



(a)

(b)

Figure 6: Resulting sub-representations for (a) *Warm* and (b) *Bright* descriptors showing coherence between clustered sub-representations.

ing procedure, we measure the Spearman (rank) correlation between the coherence distance matrix and the subjective responses. For the *warm* descriptor, a correlation of 0.83 is exhibited and for the *bright* descriptor a correlation of 0.79 is exhibited. This suggests that the interface is able to cluster equalisation curves effectively, preserving a perceptually relevant distance within and between clusters.

6. CONCLUSION

We present a system for the navigation of sub-representations within timbral descriptions of equaliser curves. From the listening test results it is clear that the interface is able to cluster samples into perceptually relevant groups, and can provide recommended settings based on a user's input. We evaluate the system using two descriptors, *warm* and *bright*, and find that a different number of clusters form for each term. For the *warm* descriptor, the subjects are always able to identify the correct sub-representation when presented with equaliser curves from each cluster, and are consistent in their responses across two audio samples from different genres. However, for the bright descriptor, subjects are able to correctly allocate the equaliser curve to the corresponding sub-representation with 90% accuracy (93.75% across both terms). In this case, the number of correct allocations varies between the audio samples, and they are not consistent in their responses. This could be because the *warm* descriptor is has a lower number of clusters (3), allowing for more variability between sub-representations, whereas the *bright* descriptor has 5, which may potentially be too high for significant perceived variance. The proposed system is able to aid the creative process of music production, by providing alternative representations of described musical timbre. This may lower boundaries to entry for novice users, and will provide new expressive functionality for expert users.

7. ACKNOWLEDGMENTS

The work of the first author is supported by the Alexander S. Onassis Public Benefit Foundation

8. REFERENCES

[1] J. W. Beauchamp. Synthesis by spectral amplitude and" brightness" matching of analyzed musical instrument tones. *Journal of the Audio Engineering Society*, 30(6):396–406, 1982.

[2] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: A cpu and gpu math compiler in python, 2010.

[3] T. Brookes and D. Williams. Perceptually-motivated audio morphing: Brightness, 2007.

[4] T. Brookes and D. Williams. Perceptually-motivated audio morphing: Warmth, 2010.

[5] M. Cartwright and B. Pardo. Social-EQ: Crowdsourcing an equalization descriptor map, 2013.

[6] A. Clifford and J. Reiss. Proximity effect detection for directional microphones. In *Audio Engineering Society Convention 131*. Audio Engineering Society, 2011.

[7] G. Darke. Assessment of timbre using verbal attributes, 2005.

[8] A. C. Disley, D. M. Howard, and A. D. Hunt. Timbral description of musical instruments. In *International Conference on Music Perception and Cognition*, pages 61–68, 2006.

[9] J. M. Grey. *An exploration of musical timbre*. Number 2. Dept. of Music, Stanford University, 1975.

[10] J. M. Grey. Multidimensional perceptual scaling of musical timbres. *the Journal of the Acoustical Society of America*, 61(5):1270–1277, 1977.

[11] S. Hafezi and J. D. Reiss. Autonomous multitrack equalization based on masking reduction. *Journal of the Audio Engineering Society*, 63(5):312–323, 2015.

[12] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

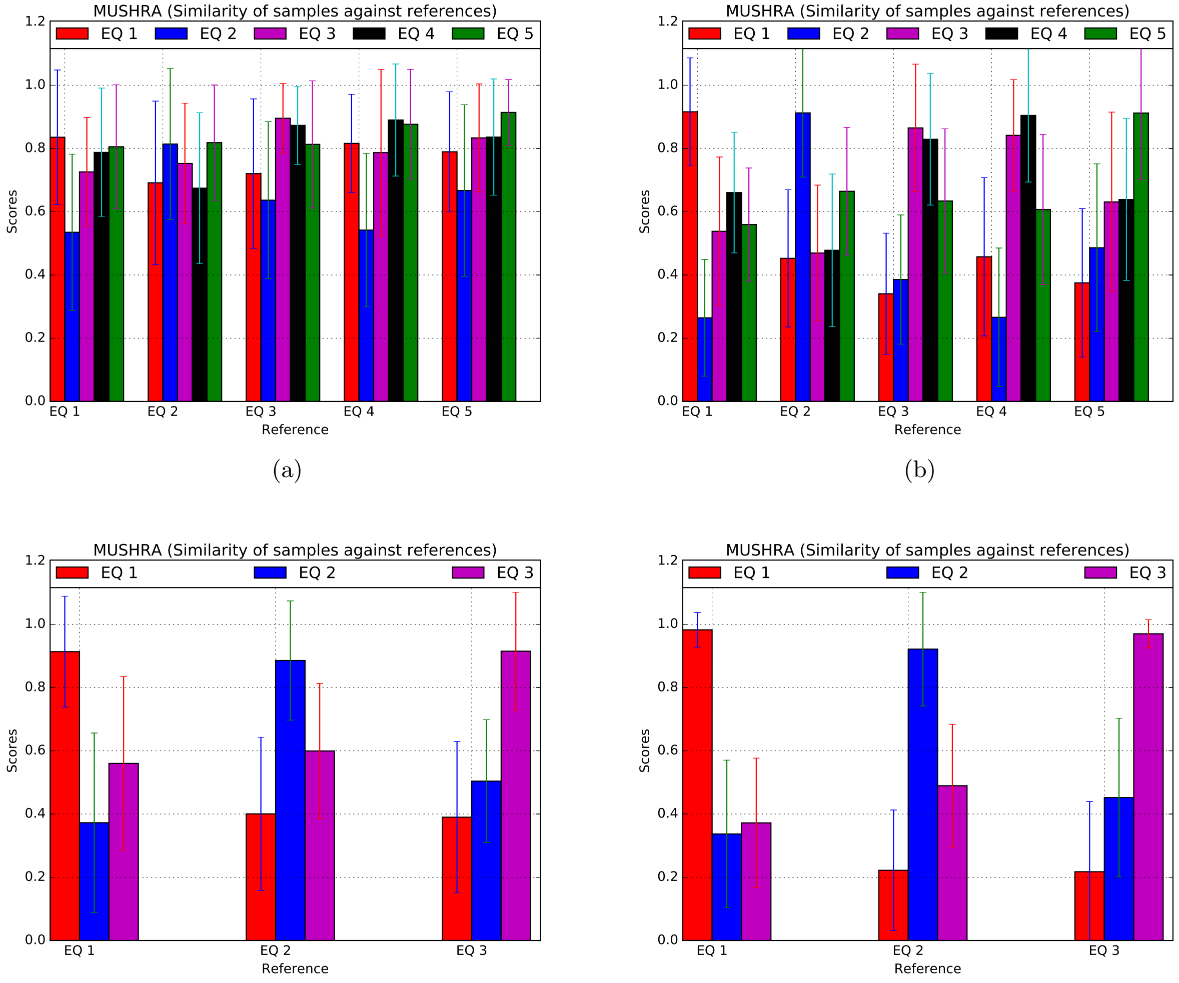
[13] B. Hopkins and J. G. Skellam. A new method for determining the type of distribution of plant individuals. *Annals of Botany*, 18(2):213–227, 1954.

[14] N. Jillings, D. Moffat, B. De Man, J. D. Reiss, and R. Stables. Web audio evaluation tool: A framework for subjective assessment of audio. 2016.

[15] F.-J. Lapointe and P. Legendre. Comparison tests for dendrograms: A comparative evaluation. *Journal of Classification*, 12(2):265–282, 1995.

[16] S. Mecklenburg and J. Loviscach. subjEQt: Controlling an equalizer through subjective terms, 2006.

[17] T. Pang-Ning, M. Steinbach, V. Kumar, et al. Introduction to data mining, 2006.



(a)

(b)

(c)

(d)

Figure 7: Bar Plot MUSHRA results for (a) Blues *Bright*, (b) Metal *Bright*, (c) Blues *Warm* and (d) Metal *Warm* sub-representations, where the error bars represent standard deviation.

[18] E. Perez-Gonzalez and J. Reiss. Automatic equalization of multichannel audio using cross-adaptive methods, 2009.

[19] A. T. Sabin and B. Pardo. 2DEQ: an intuitive audio equalizer, 2009.

[20] S. Saraçli, N. Doğan, and İ. Doğan. Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, 2013(1):1–8, 2013.

[21] M. Sarkar, B. Vercoe, and Y. Yang. Words that describe timbre: A study of auditory perception through language, 2007.

[22] P. Seetharaman and B. Pardo. Audealize: Crowdsourced audio production tools.

[23] P. Seetharaman and B. Pardo. Reverbalize: a crowdsourced reverberation controller. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 739–740. ACM, 2014.

[24] E. Skovenborg. Development of semantic scales for music mastering. In *Audio Engineering Society Convention 141*, Sep 2016.

[25] R. Stables, B. De Man, S. Enderby, J. D. Reiss, G. Fazekas, and T. Wilmering. Semantic description of timbral transformations in music production. In *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, pages 337–341, New York, NY, USA, 2016. ACM.

[26] R. Stables, S. Enderby, B. Man, G. Fazekas, J. D. Reiss, et al. Safe: A system for the extraction and retrieval of semantic audio descriptors. 2014.

[27] S. Stasis, R. Stables, and J. Hockman. A model for adaptive reduced-dimensionality equalisation, 2015.

[28] S. Stasis, R. Stables, and J. Hockman. Semantically controlled adaptive equalisation in reduced dimensionality parameter space. *Applied Sciences*, 6(4):116, 2016.

[29] V. Valimaki and J. Reiss. All about equalization: Solutions and frontiers. *Appl. Sci*, 2016.

[30] D. Williams and T. Brookes. Perceptually-motivated audio morphing: softness, 2009.

[31] A. Zacharakis and J. Reiss. An additive synthesis technique for independent modification of the auditory perceptions of brightness and warmth, 2011.