iMobilAkou: The Role of Machine Listening to Detect Vehicle using Sound Acoustics.

Muddsair Sharif
University of Applied Sciences
Stuttgart, Germany
muddsair.sharif@hft-stuttgart.de

Mayur Hotwani
University of Applied Sciences
Stuttgart, Germany
92homa1mst@hft-stuttgart.de

Huseyin, Seker
Birmingham City Univeristy
Birmingham, United Kingdom
huseyin.seker@bcu.ac.uk

Gero Lückemeyer
University of Applied Sciences
Stuttgart, Germany
gero.lueckemeyer@hft-stuttgart.de

ABSTRACT

Machine Learning can work very well with image recognition, but it is used to recognize audio patterns. Machine listening identifies audio patterns of different entities like the car engine, human speaking, nature sounds, etc. The environmental sound classification plays an important role to encourage citizens to travel smartly within a city without creating unbearable noises. On the other hand, it also promotes the city council to maintain and predict a sustainable sound at rush hour with ins the city. The aim of this early-stage research is to present a methodology that will read the labeled audio files, extract features from them, feed features to a sequential model. Moreover, the model will have the ability to classify these audio files of vehicles based on their input feature(s) and then further categorize them as it either light-weight, medium-weight, heavy-weight, rail-bound or two-wheeled vehicle using the applications of machine listening and deep learning in the field of sound acoustics. Therefore, It will also classify unlabelled test data files on a pre-trained model. This research provides us the base model for the vehicle classification giving both advantages and disadvantages along with the possibility for future extensions.

CCS CONCEPTS

• **Software and its engineering** → **Integrated and visual development environments**; • **General and reference** → *Performance*; • **Computing methodologies** → **Concurrent computing methodologies**; **Massively parallel algorithms**.

KEYWORDS

Environmental sounds classification (ESC), Intelligent Personal Assistants (IPA), deep learning (DL), M4LAB, Machine Listening (ML), Sound Acoustic(SA), Message Passing Interface (MPI), Intelligent Mobility using Sound Acoustic (iMobilAkou).

1 INTRODUCTION

Environmental sounds classification (ESC) has been increasingly studied in recent years. Sound data contain more semantic information than visual data. Especially, ESC is of critical importance in many problems such as; **noise pollution analysis** [1, 2], **surveillance systems** [3], **context-aware applications** [4, 5], **environment monitoring** [6, 7], **Smart home use cases such as 360-degree safety and security capabilities** [8], **soundscape assessment** [9], and **smart city** [10, 11]. The classification of these sounds is more difficult than other sounds because there are too many parameters that generate noise in the ESC. The success rates of these methods are relatively low as compared to deep learning-based studies in recent years. Due to the low success rate, the researcher moves forward to introduce a deep learning model for ESC that has been frequently used in recent years in different fields [12–14].

Pollution by the exhaust is a large problem in today's cities. Even in the European Union, cities are struggling to meet healthy exhaust pollution limits. This kind of pollution largely originates from combustion engines powering today's traffic. Pollution by sound, however, also known as noise pollution, is currently much less regarded, though vehicles' sound emissions are proved to cause multiple diseases as well [15]. To control noise pollution and mobility in any part of a city, it is important to detect the type of vehicles operating daily in a particular area. Maintaining such a count of vehicles physically using human labor can be a cumbersome, error-prone, and labor-intensive task. Another way is to use sensors and devices to monitor the traffic movement of vehicles - which may become costly if widespread use is intended.

In the past two decades, the sound acoustic sciences have made significant advances in software for collecting and analyzing both archived and real-time systems. Despite these advances, large share of acoustic data collected from these devices remain unprocessed. Moreover, there are many challenges, including noise-based sensor data, the size of the data banks (i.e., terabytes and beyond), and the pervasive lack of standardization, resources, and systems.

Machine Listening can be progressive to monitor the traffic data received from the sound acoustics in order to control noise pollution and mobility in a certain area. Classification of vehicles by Machine Learning can be very useful in controlling the traffic in a specific area. By performing data analytics and visualization on the predicted classified data, city experts can use this data for various purposes like detecting and controlling noise pollution and easing the mobility of vehicles.

The aim of the research is to use machine listening technique over the collected data from acoustic device and categories them into *light-weight(s), medium- weight(s), heavy-weight(s), two-wheel(s), and rail-bound* vehicles and classify different type of vehicle *cars, motorcycles, trams, trucks, buses.* Although, such a process required various computations like fetching audio data, pre-processing over

fetched audio data, extracting features from the files, training the model over the extracted features, and predicting the output categories. The rest of the paper consists of the following section(s), The Architecture section which consists further two subsections i.e. signal-representation and pre-processing, then case-study section, followed by result and discussion as well as the conclusion and future research section.

2 ARCHITECTURE

The specific objective of this research is to present a state of the art approach that automatically detect/classify feature from the sound file and improved its classification/detection using deep learning technique. Although, several researchers have significantly applied different methodologies to obtain the proposed objective in the vision of sound acoustic to determine automatic sound classification. One distinct advantage of using this approach is to fetch feature(s) directly from the sound file and convert them into an array of code(s) knows as features array rather than plotting an image of the input wav file. Primarily, our approach presented in this paper consists of three major sections depicted in figure 1.

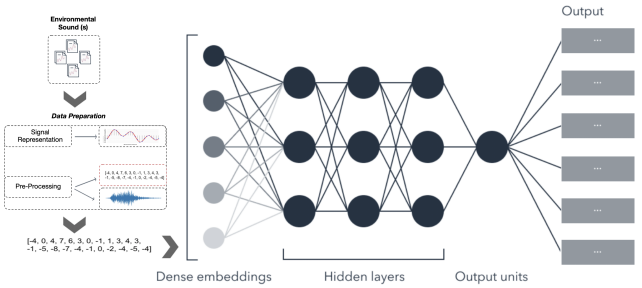

Figure 1: The Architecture

In the left-hand side of the figure 1, an audio sample or set of samples recorded from a device that contain(s) a sample wave of the environmental sound is given to a model as input in a computer-readable format which contains few seconds of sound. Moreover, this sound file passes through the data-preparation section which consists of signal representation and pre-processing. Although, the data-preparation section is thoroughly explained in the subsequent sub-section.

2.1 Signal-representation

The important step in audio signal processing is to convert the audio signal from the audio file into representations that are relevant to encode relevant audio information to be used further in the process of feature extraction and classification [16]. A data-set for the task(s) consists of a collection of wave files at a particular time interval that denotes the amplitude of the recorded signal over discrete-time samples. Moreover, the sample depth of the wave file figure-out the dynamic range of the signal. Typically, the wave signal representation is 16bit which means a sample amplitude values range is 65,536 maximum shows in figure 2 adjacent to the *signal-representation.*

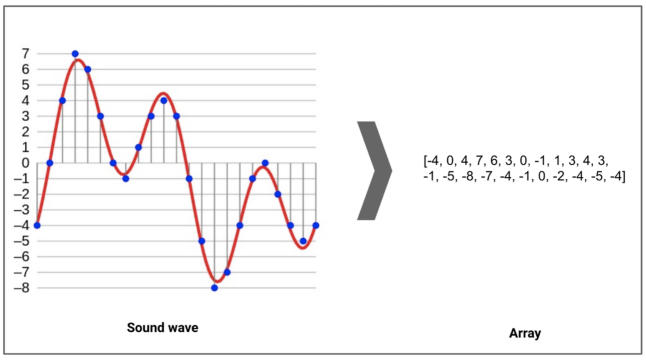

Figure 2: The signal representation

The above image shows how a sound excerpt is taken from a waveform and turned into a one dimensional array or vector of amplitude values.

2.2 Pre-processing

After having the data in place, we can apply machine listening to the data and visualize the audio for extracting features and then classifying the data using the classification model. To visualize the audio data, there are two alternatives that we are presenting here. Firstly, one way is to convert audio files to equivalent *image representation* and then feed the images to a CNN model [17]. Such representation can be a wave-plot shown in figure 2, which is a time vs amplitude plot, also called the *time-domain* representation of an audio signal. However, this wave-plot expresses the amplitude of the audio wave changing with time. In addition to this, the zero amplitude value denotes silence. Normally, such a type of wave-plot pattern occurs when sound waves move through the medium in high and low-pressure regions. These patterns are unique for different categories of sounds like *human, nature, vehicles, animals, etc.*. These patterns help the model to predict the correct class category. Secondly, to extract information directly from the audio file and then feed it to the CNN model. Feeding the features to the CNN model will lead us to classify features correctly. The feature extraction in general deals with fetching suitable information which can be supplied to the classification model to predict a particular output. For instance, features can be extracted from an image and used for face recognition, age recognition, etc. features in the case of an image can be relevant pixels that reveal useful information and hence discarding pixels that are not necessary for further classification. Moreover, the features can also be recovered from digital text documents and can be used for further text classification like *spam filtering, author identification, topic detection, etc.* [18]. In this case, also, the purpose of feature extraction remains the same, which is to remove irrelevant and redundant information and only capture the purpose information. In short, feature extraction finds applications in many fields and cannot be ignored when it comes to making accurate predictions. Nevertheless, the feature extraction can also work in the case of audio data which is our main point of implication. It deals with audio processing, music information retrieval, synthesis, and evaluation. It plays a major role in deciding the performance of the outcome of the model. Content-based classification is used to

identify the best contributing audio features from audio files [19]. Feature extraction becomes effective when the right features are being extracted which suits the problem statement in focus. Although image-based augmentation has been used extensively in previous research using different learning model technique but with the addition of sound acoustic revealed the attention of this approach that gets noticeable attention due to its direct digital conversion from wave-plot into a set of the encoded array which provides successive outcomes using different machine learning techniques.

It is now clear that before jumping to the classification of audio file(s), it is necessary to extract the best features first before moving forward. Due to the versatility, variability, and variety of the data set, it is not possible to extract the features manually. To ease out the process of feature extraction, there are plenty of already available audio extraction libraries and toolboxes. we have already evaluated some of many such libraries i.e. *librosa, PyAudioAnalysis, SurfBoard, Essentia, Aubio, and yaafe.* and comparison of these libraries which will further make it easy to select the best library for feature extraction. At this point, we are only giving the name of these libraries which we used to have a comparison in between concerning feature extraction and selected *librosa* which provide us the best feature selection according to our objective.

In the middle section of the figure 1 which identifies as the model-section, where we propose our model of the network that consists of numbers of layers i.e. input-layer, hidden-layer(s), and output-layer that has been developed in Keras tool which is simple, flexible and powerful deep learning application programming interface for creating neural networks. Moreover, important features of Keras includes creating deep-learning models and neural networks, training loop, callbacks, distributed training, and automatic support of GPUs and TPU's. Keras is built on the principle that "being able to go from idea to result with minimum possible delay is key to do good research" [20]. At this early stage of our research, we propose a generalized sequential model for the classification of vehicles. This model has one input layer, three hidden layers, and one output layer. The input layer is known as Dense embedding which extracts features from the sound file using the Libros audio extraction library discussed earlier. Input features can vary from one input feature to N number of features. The hidden layer (s) are responsible for converting the input layer features from one dimension to another by mapping every node to every other node in the network forming a deep and dense neural network. After the hidden layers, the final layer is the output layer which uses the softmax activation function. Each value in the output layer is mapped to known categories and the model is sequential and has exactly one output.

Once the model is created, the next step is to compile this created model which consists of different options that describe the compilation strategy. These options include optimizers such as *Gradient Descent, Stochastic Gradient Descent, Adagrad, AdaDelta, Adam* and so on., loss function, and metrics that describe the degree of evaluation. However, these methods are used to change the attributes of your network such as weight and learning rates to reduce the losses. we carry out comprehensive and consistent analysis and comparison of commonly used state-of-art approaches for sound classification via a deep learning approach in an urban acoustics environment. We apply this approach has been carried out on one of our use-cases whose explanation is given in the use-case section.

3 USE CASE

Big cities around the globe are facing certain challenges likewise noise-level pollution. This use case promotes the importance to reduce noise level pollution in a large cities. One could easily get annoyed by the noise of traffic, vehicles, especially in bigger cities like Stuttgart. Usually, the noise level is especially high in such cities that are situated inside a valley where sound can be reflected back by the hills. The city council of Stuttgart received a number of complaints from citizens regarding sound pollution. So the city council of Stuttgart has decided to control the sound pollution levels in this region that promote as a major part of the research which lies in the environmental sound classification category. An early stage of the research, the researcher from HfT promotes an automatic environmental sound classification-based strategy via sound acoustic to reduce the noise level via vehicle classification approach in the Stuttgart region.

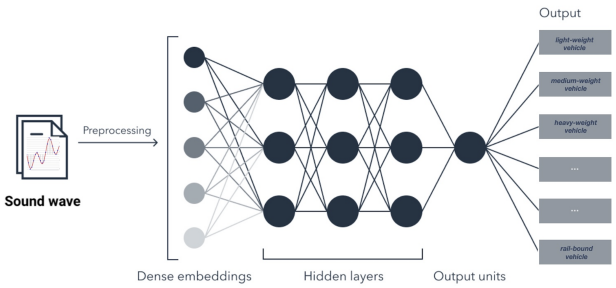

Figure 3: use-case: region-Stuttgart

Figure. 4 shows the generalized sequential model used for classification of vehicles. This model has one input layer, three hidden layers, and one output layer. The input layer consists of features extracted from the .wav audio files using the librosa audio extraction library discussed above in subsection **??**. Input features can vary from one input feature to N number of features. Moreover, the hidden layers are responsible for converting the input layer features from one dimension to other by mapping every node to every other node in the network forming a deep and dense neural network. After the hidden layers, the final layer is the output layer which uses the *softmax* activation function. Each value in the output layer is mapped to known categories and the model is sequential and has exactly one output.

The figure 4 shows a example scenario of testing the model described in the figure 1 using a .wav test file. The sound file is the sound of a car passing by on a highway. The *Car.wav* sound file is initially supplied to the model, after which the features from the audio file are extracted using librosa feature extraction methods. In addition to this, the extracted features are then provided as an input to the first hidden layer. The first hidden layer uses the *ReLU* activation function to decide which neurons should be active and which ones should be dead. The neurons highlighted with blue color are active. Only a few of the active neurons are passed to the

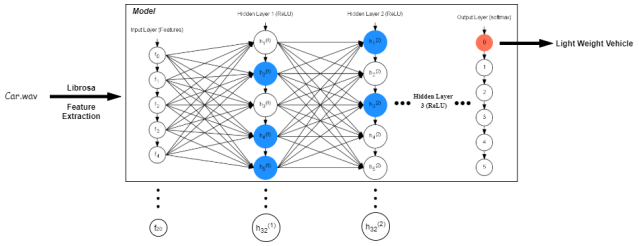

Figure 4: Testing Sequential Model using Single Sound File

next hidden layer. In the end, the output layer receives only one active neuron which is our final output. In the diagram, the model predicted the *Car.wav* file as the class, which belongs to the *Light Weight Vehicle* Category i.e. Car.

4 PERFORMANCE ANALYSIS AND DISCUSSION

The detail about how vehicle classification is done has been expressed thoroughly in the case-study section. Although, in this section, we provide the details about how the environment is used to train and test the model for classifying the vehicles. In addition to this, information related to training environment such as *operating system, device, and tools like IDE, libraries used to train the model* , are enlisted below

- **Device Operating System:** Microsoft Windows 10, 64-bit x64 based processor, 8gb RAM
- **Device Processor:** AMD Ryzen 5 3500U, 2.10 GHz
- **Programming Language:** Python [3.7]
- **IDE:** Pycharm 2020.2 Community Edition
- **Machine Learning Library:** Tensorflow 2.3.1
- **API for Tensorflow:** Keras 2.4.3
- **Feature Extraction Library:** Librosa
- **Visualization toolkit for ML:** Tensorboard 2.3.0
- **Data set Source:** FreeSound50k Data set

Training the model using FSD50K Dataset with 40K files takes about 5 hours 40 minutes and outputs a training accuracy of about 99.11 percent. Testing the model is quite straightforward and runs much faster. As test files are just a fraction of the percentage of the training files which means that if we split the training and test data as (80, 20), out of 40k files, only 8k will be used for testing. We have already discussed and decided on the alternatives for the process of achieving vehicle classification using deep learning which involves, pre-processing audio files and selecting an audio representation approach, feature extraction, tools and libraries to extract features, which features are available, which data set we use for training, and finally which model we use for classification.

Table 1: Comparison of Train/Test Accuracy in percentage on personal device

Training Model	**Training Model on**	**Train/Test Accuracy (%)**	**Total Time**
Augmented Dataset - 750 files	Personal Device	97.83/85.33	3 min 26 sec
Augmented Dataset - 2750 files	Personal Device	98.13/90.67	6 min 21 sec
FSD50K Dataset - 30K files	Personal Device	98.67/88.23	4 hr 35 min
FSD50K Dataset - 40K files	Personal Device	99.11/90.76	5 hr 40 min

Firstly, we will try to explain the difference between training or testing accuracy when the data set is used without augmentation vs with augmentation.Table 2 shows the comparison between vehicle data set with 250 files and before applying any data augmentation and vehicle data set with 750 files after using data augmentation. It is evident from these tables, that in the case of data augmentation the model gives better accuracy because it has more data to generalize and learn from it.

Table 2: Comparison on vehicle data-set before vs after data augmentation

2***Testing Model**	Before Augmentation **Train/Test Accuracy (%)**	After Augmentation **Train/Test Accuracy (%)**
Only MFCC with 10 segments	68/53	82/73
Only MFCC with 20 segments	81/33	97/85
Only MFCC with 40 segments	92/50	99/96
Only MFCC with 80 segments	99/57	99/97
Only MFCC with 100 segments	99/33	99/95
All Features W/O MFCC	64/47	73/60
Some features + MFCC	79/23	99/92
All Features + MFCC	92/47	99/94

Secondly, the Figure 5 gives an overview on how the model behaves for the different values of the noise factor. The noise factor is responsible for increasing the proportion of noise. It gives an illusion of influencing a weather condition such as rain. Although, there is a difference in that noise completely-overrides the sound of vehicle tires while the actual sound of rain does not. When we increase the noise factor, the model starts to depict less satisfactory results and with increasing noise factor, the model accuracy decreases further. Hence it is important to note that increasing in the noise factor should be in such a way that the original audio sound is still audible in the background.

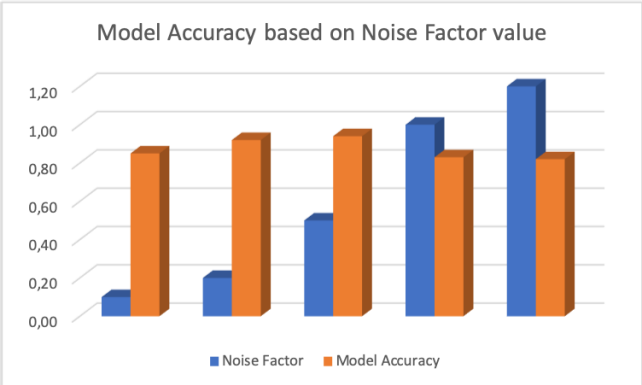

Figure 5: Model Accuracy based on Noise Factor value

thirdly, the data without augmentation is not sufficient and does not give satisfactory results. The model cannot predict an audio file with the sound of two or more vehicles passing by. If the data set is unbalanced, consisting of the unequal amount of files per category, the model will not predict expectantly when tested with the unlabelled or unseen test data. During data augmentation, using a large value for noise factor (larger than 1.0) would lead to a change in the semantic of the original audio file and thus the model test accuracy will also be comprised.

5 CONCLUSION AND FUTURE WORK

In this paper, we describe various approaches and alternatives and choose a suitable one for obtaining data, reading the data, extracting features from the data, and finally creating a model which leads to achieving the goal of classifying the vehicles into their respective categories. Demonstrating the behavior of the model on different noise values also provided the extent to which our model can predict accurately. We also showcased the technique of data augmentation which can be applied for data enrichment to achieve better results in case of data scarcity. Training time can be reduced to a large extent by using the parallel computing nodes on the HPC cluster. Also, using a balanced dataset for training, the model can predict even on the unlabelled test data. As a future extension, a clustering algorithm like k-means clustering can be used to separate skewed or noisy data and non-noisy data before starting the training process.

After classifying the data about vehicles into categories, city experts can extract which type of and how many vehicles pass by, which of these types produce the highest noise levels, etc. Finally using this extracted information, specialists from different areas can take suitable measures to control noise pollution in the region. For example, building architects can derive on how the building structure should be, other smart city experts can take appropriate measures to reduce noise pollution levels for that region.

6 ACKNOWLEDGMENTS

This publication is supported by M4_LAB. M4_LAB is a transfer project at the University of Applied Sciences - Stuttgart within the framework of the "Innovative Hochschule" initiative funded by of the Federal Ministry of Education and Research under the grant number 03IHS032A.

REFERENCES

[1] Ravi Katukam, Ajith Raj R, and Syed Asad Abbas. Face recognition using machine learning. *International Journal of Research and Analytical Reviews (IJR*, 7(2):1–2, 2020.
[2] Jayashree Padmanabhan and Melvin Jose Johnson Premkumar. Machine learning in automatic speech recognition:a survey. *IETE Technical Review*, 32(4):240–251, 2015.
[3] Nil Goksel-Canbek and Mehmet Emin Mutlu. On the track of artificial intelligence: Learning with intelligent personal assistants. *International Journal of Human Sciences*, 13(1):592–601, 2016.
[4] Romain Serizel, Victor Bisot, and Slim Essid. Machine listening techniques as a complement to video image analysis in forensics. *IEEE International Conference on Image Processing (ICIP), Phoenix, AZ*, pages 948–952, 2016.
[5] Eduardo Fonseca, Jordi Pons, and Xavier Favory, editors. *FREESOUND DATASETS: A PLATFORM FOR THE CREATION OF OPEN AUDIO DATASETS*, Suzhou, China, 2017. Proceedings of the 18th ISMIR Conference,.
[6] Bob L. Sturm. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. 2013.
[7] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, editors. *A Dataset and Taxonomy for Urban Sound Research*, Orlando, USA, 2014. Proceedings - 22nd ACM International Conference on Multimedia.
[8] Jort F. Gemmeke, Daniel P. W. Ellis, and Dylan Freedman, editors. *Audio Set: An ontology and human-labeled dataset for audio events*, 2017. ICASSP 2017 - 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
[9] Eduardo Fonseca, Jordi Pons, and Xavier Favory, editors. *FSD50K: an Open Dataset of Human-Labeled Sound Events*, 2020.
[10] Tom Ko, Vijayaditya Peddinti, and Daniel Povey. Audio augmentation for speech recognition. *International Speech Communication Association*, pages 3586–3589, 2015.
[11] J. Salamon and J. P. Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24 (3):279–283, 2017.
[12] David Moffat, David Ronan, and Joshua Reiss, editors. *An Evaluation of Audio Feature Extraction Toolboxes*, Trondheim, Norway, 2015. International Conference on Digital Audio Effects.
[13] G. Fox, J. A. Glazier, J. C. S. Kadupitiya, V. Jadhao, M. Kim, J. Qiu, J. P. Sluka, E. Somogyi, M. Marathe, A. Adiga, J. Chen, O. Beckstein, and S. Jha. Learning everywhere: Pervasive machine learning for effective high-performance computation. In *2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 422–429, 2019.
[14] A. Pisal, R. Sor, and K. S. Kinage. Facial feature extraction using hierarchical max(hmax) method. In *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pages 1–5, 2017.
[15] EU. Eu airquality report, 2020. URL https://www.eea.europa.eu/data-and-maps/dashboards/air-quality-statistics.
[16] Shuhui Qu, Juncheng Li, and Samarjit Das. Understanding audio pattern using convolutional neural network from raw waveforms. pages 1–2, 2016.
[17] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari. Sound classification using convolutional neural network and tensor deep stacking network. *IEEE Access*, 7:7717–7727, 2019.
[18] F. P. Shah and V. Patel. A review on feature selection and feature extraction for text classification. In *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 2264–2268, 2016.
[19] N. P. Patel and M. S. Patwardhan. Identification of most contributing features for audio classification. In *2013 International Conference on Cloud Ubiquitous Computing Emerging Technologies*, pages 219–223, 2013.
[20] *Librosa Documentation.* URL https://librosa.org/.