Can Machine Learning Be Used to Recognize and Diagnose Coughs?

Charles Bales, Charles John, Hasan Farooq, Usama Masood, Muhammad Nabeel and Ali Imran
Dept. of Electrical and Computer Engineering, University of Oklahoma, Tulsa, USA
csbales@wpi.edu,{charlesj,hasan.farooq,usama.masood,muhmd.nabeel,ali.imran}@ou.edu

***Abstract*—5G is bringing new use cases to the forefront, one of the most prominent being machine learning empowered health care. Since respiratory infections are one of the notable modern medical concerns and coughs being a common symptom of this, a system for recognizing and diagnosing infections based on raw cough data would have a multitude of beneficial research and medical applications. In the literature, machine learning has been successfully used to detect cough events in controlled environments. In this work, we present a novel system that utilizes Convolutional Neural Networks (CNNs) to detect cough within environment audio and diagnose three potential illnesses (i.e., Bronchitis, Bronchiolitis, and Pertussis) based on their unique cough audio features. Our detection model achieves an *accuracy* of 90.17 % and a *specificity* of 89.73 %, whereas the diagnosis model achieves an *accuracy* of about 94.74 % and an *F1* score of 93.73 %. These results clearly show that our system is successfully able to detect and separate cough events from background noise. Moreover, our single diagnosis model is capable of distinguishing between different illnesses without the need of separate models.**

***Index Terms*—Cough detection, illness diagnosis, machine learning, Convolutional Neural Network, Mel-spectrogram.**

I. Introduction

A study conducted in 2016 estimated that 4.4 % of all deaths in that year, a number surpassing 2 million, were due to various lower respiratory tract infections in both young children and adults [1]. Early detection of potential respiratory tract infections can reduce the likelihood of severe complications at a later time. And of the sickness behaviors that act as early indicators, coughing is the most common. Thus, an automated system for the detection and preliminary diagnosis of cough events would thus serve as a useful tool in the medical field.

Research for cough event detection and diagnosis through machine learning categorization is a relatively recent effort and thus is not extensively developed. Additionally, there are a plethora of potential approaches that can be utilized to tackle the issue, with the wildly varying network model structures, cough sample sources, and audio file pre-processing making it difficult to identify a generally optimal strategy. This paper will therefore only focus on the use of a single network structure, a Convolutional Neural Network (CNN), to identify and attempt to diagnose recorded cough events using image recognition, a task at which this network structure is particularly adept [2].

Using a CNN for the specific purpose of audio categorization has been attempted in prior works several times, to varying levels of success [3], [4]. The publicly available ESC-50 dataset [5] serves as a popular choice when a large collection of environmental sounds is needed for this purpose. In [6], it is shown that a CNN model can be successfully applied to environmental sound classification. This has been further verified by achieving similar results when applying a CNN to the ESC-50 dataset [7].

Acknowledging the success of the CNN structure in categorizing audio into the fifty categories of the ESC-50 set, one of these being cough sounds, suggests that the network model may be capable of doing so with just coughs specifically. Indeed, this approach has been attempted in several more recent works. In [8], authors utilized several feature types, such as Perceptual Linear Predictive (PLP) power spectrum and cepstrum, in their experiments to test the efficacy of the CNN structure in correctly identifying cough audio with image recognition. They used five different features to convert the sound files into image inputs and achieved an *accuracy* of 99.65 % suggesting that image processing stands as a valid method for categorizing cough audio. Later, authors in [9] achieved a *specificity* of 92.7 % using cough sounds processed into 64 ms segmented frames as inputs for their implemented CNN and deemed this method comparable to other conventional approaches. Finally, a recent work [10] used spectrogram images of lengthy audio segments as the inputs for a CNN to differentiate between healthy and pathological respiratory sounds, and achieved an *accuracy* of 86 %.

Each of these approaches involve either a heavy pre-processing phase or inherent design restrictions. These include but are not limited to discarding frames of silent audio through the use of an RMS energy threshold to only categorize selected audio [9] and specifically selecting noise-reducing hardware for data collection [10]. What this paper seeks to accomplish is to significantly minimize recorded audio pre-processing before feeding data into a network model while maintaining acceptably high levels of *accuracy* for cough detection, among other metrics. A primary difference in the inputs used here compared to other approaches is that this training data does not involve categorizing isolated cough events but rather deciding whether a select audio clip input contains one or more coughs within its length. In brief, the main goal is to verify if a CNN model can recognize cough events by their distinct features while among typical environmental noise. If a cough event is detected, the audio is then sent in for further processing to be used for the diagnosis portion of this research.

The focus on low overhead for cough detection is for the sake of flexibility, allowing the detection model to potentially be run on devices with lower processing power than is typical, such as mobile devices. The categorization of audio inputs as different types of coughs requires a more complex machine learning model, making maintaining usefulness for mobile applications more difficult. As such, the minimization of processing power is less so a priority than model performance for cough sound diagnosis. In previous works, cough audio has been categorized for specific illnesses, such as in [11], where authors exploited several machine learning models to accurately diagnose croup using only features extracted from cough audio, such as Mel-frequency cepstral coefficients. Similarly, authors in [12] proposed a low-complexity algorithm capable of identifying pertussis without any false diagnosis. They were able to achieve a high success rate through just the analysis of cough audio, their model recognizing features unique to the whooping cough illness. These results suggest that coughs originating from specific infections or illnesses have a sufficient number of distinguishing features that machine learning models can use for categorization. Taking the next step of sorting through multiple potential causes simultaneously is what our work seeks to achieve.

Our main contributions can be summarized as follows:

- We propose a low-complexity CNN model for the detection of cough events (within typical outside environmental noise) from lengthy audio segments and then categorizing among three common cough-generating respiratory conditions.
- The dataset used to train our proposed model does not go through the typical level of pre-processing that other works use for cough detection.
- Our diagnosis model uses cough sound data to differentiate between several potential diagnoses that may be causing the cough symptoms, the novelty of this approach being that we attempt to sort through several different potential infection categories in the same model.

II. Cough Detection System

The effectiveness of a CNN typically scales with the size of the dataset used for training the model. Unfortunately, there is a particular scarcity of data in terms of cough audio. The results of prior works are difficult to compare with each other due to differing datasets used for model training, each with their own unique standardization. Cough data from a small sample of individuals [9], [13] will accurately identify coughs among these individuals and those similar, but may not be generally applicable to a larger populace. Studies with a large number of subjects and thus cough samples [10], [11] lead to accurate detection models, but datasets of this size for cough audio are typically collected in a controlled environment such as a hospital. Since the purpose of our CNN is to distinguish coughs within clips of environment sounds, controlled recording environments do not suit our purpose. Coughs cropped from YouTube audio [12] and other online sources, while numerous, vary wildly in audio quality. This variance, however, works towards the benefit of a more robust detection and diagnosis model that is able to identify relevant cough audio in less controlled conditions. This paper takes a similar approach to acquiring the data used for model training.

A. Detection Database Composition

The database used for training the CNN in cough detection is composed of various modified audio clips gathered from free online sources [5], [14]. Each of these audio files originally contained at least one cough event and are cropped to a length of five seconds with the full cough contained at some arbitrary point within. Audio that contained coughs in more than a five seconds period are separated into multiple files. The exact number of coughs per file is left intentionally non-standardized. This database is balanced by a collection of environmental and speech sounds, comprised primarily of audio from the ESC-50 database for environmental sound classification. In addition to ESC-50, this non-cough half of the database is supplemented by sound clips taken from unused portions of the original cough audio that did not contain cough events. All audio files used have a sampling rate of 44.1 kHz and are labeled according to their origin. Audacity® recording and editing software version 2.3.2[1] is utilized to ensure a consistent sampling rate, length, and mono waveform file format for each data item. The final data file count amounts to 993 cough items and 729 non-cough items.

The features we used for the detection model training take the form of Mel-spectrograms. The Mel scale is a pitch categorization where listeners judge changes in pitch to be equal in distance from one another along this scale. It is meant to make changes in frequency, such as with a spectrogram, more closely reflect audibles changes. There are several methods for converting the frequency scale to Mel. Here, we convert frequency f into Mel-scale m as

$$m = 2595 \times \log_{10}(1 + \frac{f}{700}) \quad . \tag{1}$$

For our detection database, the entire length of the input audio clips are converted into Mel-scaled frequency graphs using the Librosa [15] library Mel-spectrogram function. The resulting images of pixel size 432×288 are then converted to gray-scale to unify the intensity scaling and are then compiled to form the final database of cough and no cough environmental audio clips.

B. Detection CNN Structure

The relative success of using a CNN for cough detection from recorded audio via image recognition serves as the basis for the machine learning model used for this system. Even though our goal is to analyze a much larger time frame of five seconds, the model structure serves as a workable foundation.

Overview of our used CNN structure is shown in Fig. 1. Due to the large size of the plot, our modified CNN begins with a 2×2 max-pooling layer to lower the required overall model complexity before proceeding. Both of the following convolutional layers have 32 filters and a stride of $(1, 1)$, with

[1] Audacity® software is copyright © 1999-2019 Audacity Team.
The name Audacity® is a registered trademark of Dominic Mazzoni.

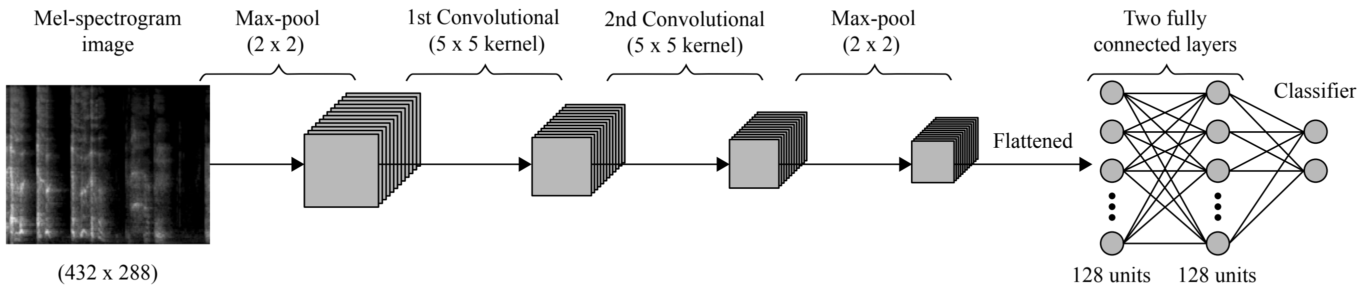


Fig. 1. Overview of the CNN structure used for cough detection.

ReLU activation following each. They both have a 5×5 filter size. After the convolutional layers is another 2×2 max-pooling layer, and then two fully connected layers with 128 neurons and ReLU activation immediately afterwards. These fully connected layers utilize a dropout of 0.5. The final layer is the softmax classification layer to distinguish between cough and not cough for the given input.

The number of convolutional and fully connected layers are kept low to minimize potential overfitting issues. The database file count is increased to assist in reducing overfitting as well. Since ReLU is the current standard for CNNs, it is used for the activation functions of this model, while Adam [16] is used as the optimizer due to its relatively better efficiency and flexibility. A binary log loss function completes the detection model.

C. Detection Model Training

The CNN is constructed in Python using Keras as the machine learning library. For the 1,722 database items, 993 cough items and 729 non-cough items, a validation split of 80 % training and 20 % validation is employed. The model, functioning with a batch size of 32 for 30 epochs, is run several times to assist in tuning the hyperparameters to find the optimal setup for the cough detection task. The validation set output, summarizing model *accuracy* and *specificity*, is used as the metric for our model success.

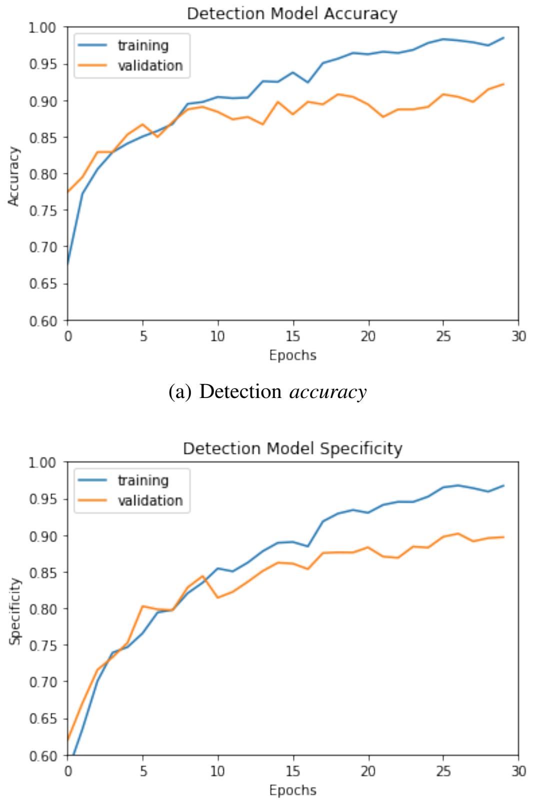
D. Detection Performance

The metrics being compared for the detection model are *accuracy* and *specificity*. *Accuracy* stands as a baseline metric for comparison, while *specificity* is important to this detection model in particular. The *accuracy* and *specificity* is calculated as:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad , \tag{2}$$

$$specificity = \frac{TN}{TN + FP} \quad . \tag{3}$$

Here, *TP*, *TN*, *FP*, and *FN* refer to True Positives, True Negatives, False Positives, and False Negatives, respectively. The *FP* is the value we wish to minimize the most to prevent any non-cough audio from being passed into the diagnosis model. Therefore, for this purpose, maximizing *specificity* is ideal.



(a) Detection *accuracy*

(b) Detection *specificity*

Fig. 2. Cough detection performance.

Figure 2 shows the performance of our detection model. The detection model, over the 30 training epochs, arrives at a relatively consistent result. It can be seen that we achieve a peak validation *accuracy* of about 90.17 % (see Fig. 2a) and a validation *specificity* of 89.73 % (see Fig. 2b). These results reflect the performance of system over data that comes from a comparatively huge swath of audio quality and cough volume and quantity. Much of the audio tested already contains ambient noise within due to the varied range of sources used for collecting the data. Due to the large file count, potential biases such as environmental consistencies in instances where coughs would typically occur are lessened, and will continue to diminish as the dataset increases in size. Hence, it is important to highlight that the performance of our proposed detection system will further improve, for example, in controlled environments, which usually were the focus of prior works.

III. Cough Diagnosis System

The data scarcity issues with the dataset of cough detection model are present for the diagnosis model, but compounded with the additional problem of labeled cough audio being considerably more scarce. Since the existing database for detection is comprised primarily of unlabeled cough audio with no identified illness relation, the same cannot be used as diagnosis training data. To train our cough diagnosis system, we collected cough sounds from 35 brochiolitis, 131 pertussis and 96 bronchitis patients. Obviously these are very small numbers of samples and more data is needed to make the solution more reliable. Even with small training data very promising accuracy has been observed on unseen test samples, as reported in diagnosis performance section.

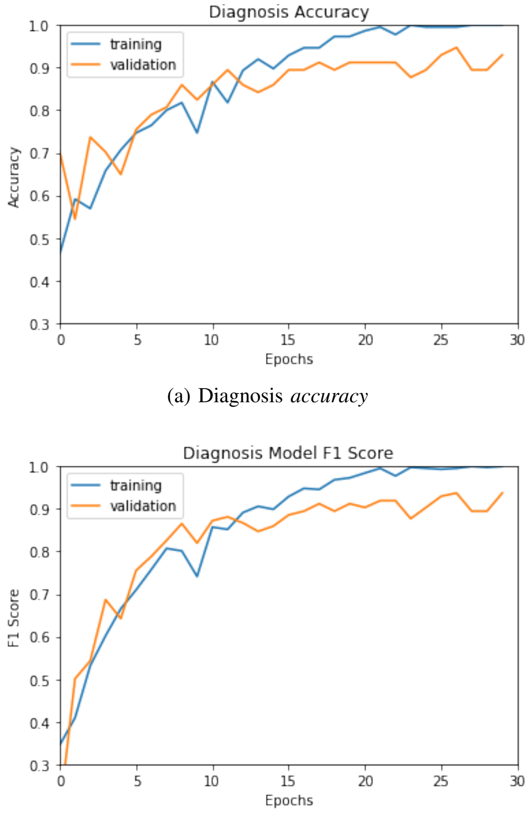
A. Diagnosis Database Composition

The audio specifically collected for diagnosis underwent a different process in being converted to the data used for training. Since the purpose of the diagnosis model is not to detect the presence of any number of coughs but to categorize according to the differences between specific cough events, the audio file inputs are much shorter in length. These sound files are cut to a single cough event using Audacity. With the longest of these files lasting two seconds, the rest are buffered to the same length with noiseless audio to maintain the time scaling for the spectrogram conversion process. Finally, these sound clips underwent the same image conversion process as the detection dataset. Using the Mel-spectrogram function of Librosa library, the two-second audio files are converted into 432×288 Mel-frequency graphs, and are then converted to grayscale. The resulting images made up the cough diagnosis dataset.

B. Diagnosis CNN Structure

The same type of machine learning categorization procedure used for binary detection can be applied to the labeled cough data originating from several different illnesses. However, due to the differences between two coughs of differing illnesses typically being less apparent than a cough and non-cough event, a more complex set of layering is required for this version of the CNN for Mel-spectrogram image analysis.

The diagnosis CNN begins identically in function when compared to the detection model (depicted in Fig. 1), but after the 2×2 max-pool, the two convolutional layers, and the 2×2 max-pooling layer that follows, there are an additional two convolutional layers and a 2×2 max-pooling layer. The network then continues into a similar structure of two fully-connected layers with an identical neuron count of 128, activation type, and dropout of 0.5. Softmax classification is again used for determining the predicted category label of each input.



(a) Diagnosis *accuracy*

(b) Diagnosis *F1*

Fig. 3. Cough diagnosis performance.

C. Diagnosis Model Training

The diagnosis CNN is constructed in a similar manner to the detection CNN, using the machine learning library Keras. The diagnosis database is comprised of 262 total cough sound items, distributed among Bronchiolitis, Pertussis and Bronchitis at counts of 35, 131, and 96, respectively. A validation split of 80 % training and 20 % validation is also employed for this neural network structure. The model is trained for 30 epochs, focusing on *accuracy* and *F1* score as metrics for diagnosis, with *F1* being chosen over *specificity* in this case due to the relative imbalance of the dataset. The *F1* score is calculated as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad , \tag{4}$$

where, $Precision = TP/(TP + FP)$ and $Recall = TP/(TP + FN)$.

D. Diagnosis Performance

Figure 3 shows the performance of our diagnosis model. The diagnosis model achieves a peak *accuracy* of 94.74 % (see Fig. 3a) and *F1* score of 93.73 % (see Fig. 3b). This *F1* score clearly shows that the ability of this machine learning model to distinguish between multiple illnesses based on various cough sounds. It is also important to note that in previous works, the focus is usually to diagnose a specific illness from cough sounds, whereas, as a novel approach, we distinguish between different types of coughs simultaneously in the same model.

As stated earlier, because of the unbalanced nature of the database used for training the diagnosis model, *F1* is used in addition to *accuracy* as sufficient metrics for making conclusions on the model performance. Since the data count is relatively low for the number of categories, more than one metric better interprets model performance. The potential bias mentioned for detection, environmental consistencies between categories unrelated to cough audio structure, is a relevant concern in this case. For example, individuals with a known case of bronchitis would typically be situated in hospitals more frequently than those with the flu or a cold. This issue would again be alleviated by a larger volume of data, or perhaps a controlled variance in making sure that one category of data is not overtly reliant on background sounds for its unique model training identifiers. Regardless, the achieved *accuracy* and *F1*, given the number of categories, is indicative of high level of success in diagnosing between specific illness-based coughs.

IV. Conclusion

Our work uses image recognition through a CNN to approach both the goal of cough event detection and of illness diagnosis by using cough audio converted to Mel-spectrograms. Our experiments show that the CNN is capable of accomplishing these tasks with a high level of *accuracy*. With only a small number of modifications, the low-complexity network model can be trained sufficiently for both ends. This is achieved in spite of limited available cough data, and the model performance should improve given a larger training dataset.

In future work, we focus on implementing the detection model on mobile devices for a more consistent data acquisition method via phone recording. Additionally, increasing the dataset size for diagnosis as well as affirmation of manual labeling *accuracy* through work with an experienced physician would stand to improve the diagnosis model for potentially a wider range of illnesses. The dataset used for detection would serve well as a community resource for similar future experiments.

Acknowledgement

This work is made possible by NSF Grant No. 1559483. The statements made herein are solely the responsibility of the authors.

References

[1] C. Troeger, B. Blacker, I. A. Khalil, P. C. Rao, J. Cao, S. R. M. Zimsen, S. B. Albertson, A. Deshpande, T. Farag, Z. Abebe *et al.*, "Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016," *The Lancet Infectious Diseases*, vol. 18, no. 11, pp. 1191–1210, Nov. 2018.

[2] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Columbus, OH: IEEE, June 2014, pp. 512–519.

[3] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN Architectures for Large-scale Audio Classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA: IEEE, Mar. 2017, pp. 131–135.

[4] T. Kim, J. Lee, and J. Nam, "Comparison and Analysis of SampleCNN Architectures for Audio Classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 285–297, May 2019.

[5] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *23rd ACM International Conference on Multimedia*, ser. MM 15. Brisbane, Australia: ACM, Oct. 2015, pp. 1015–1018.

[6] ——, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[7] Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep Convolutional Neural Network with Mixup for Environmental Sound Classification," in *Pattern Recognition and Computer Vision*, J.-H. Lai, C.-L. Liu, X. Chen, J. Zhou, T. Tan, N. Zheng, and H. Zha, Eds. Springer International Publishing, 2018, pp. 356–367.

[8] H.-H. Wang, J.-M. Liu, M. You, and G.-Z. Li, "Audio Signals Encoding for Cough Classification Using Convolutional Neural Networks: A Comparative Study," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Washington, D.C.: IEEE, Nov. 2015, pp. 442–445.

[9] J. Amoh and K. Odame, "Deep Neural Networks for Identifying Cough Sounds," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 5, pp. 1003–1011, Oct. 2016.

[10] M. Aykanat, Ö. Kiliç, B. Kurt, and S. Saryal, "Classification of Lung Sounds Using Convolutional Neural Networks," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, p. 65, Sep. 2017.

[11] R. V. Sharan, U. R. Abeyratne, V. R. Swarnkar, and P. Porter, "Automatic Croup Diagnosis Using Cough Sound Recognition," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 2, pp. 485–495, Feb. 2019.

[12] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas, "A Cough-Based Algorithm for Automatic Diagnosis of Pertussis," *PLOS ONE*, vol. 11, no. 9, p. e0162128, Sep. 2016.

[13] J. Liu, M. You, G. Li, Z. Wang, X. Xu, Z. Qiu, W. Xie, C. An, and S. Chen, "Cough Signal Recognition with Gammatone Cepstral Coefficients," in *IEEE China Summit and International Conference on Signal and Information Processing*. Beijing, China: IEEE, July 2013, pp. 160–164.

[14] F. Font, G. Roma, and X. Serra, "Freesound Technical Demo," in *21st ACM International Conference on Multimedia*, ser. MM 13. Barcelona, Spain: ACM, Oct. 2013, pp. 411–412.

[15] B. McFee, M. McVicar, S. Balke, V. Lostanlen, C. Thom, C. Raffel, D. Lee, K. Lee, O. Nieto, F. Zalkow, D. Ellis, E. Battenberg, R. Yamamoto, J. Moore, Z. Wei, R. Bittner, K. Choi, nullmighty-bofo, P. Friesch, F.-R. Stter, Thassilo, M. Vollrath, S. K. Golu, nehz, S. Waloschek, Seth, R. Naktinis, D. Repetto, C. F. Hawthorne, and C. Carr, "librosa/librosa: 0.6.3," Feb. 2019.

[16] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *International Conference on Learning Representations (ICLR)*, San Diego, CA, May 2015.