A Comparative Study of Features for Acoustic Cough Detection Using Deep Architectures*

Igor D. S. Miranda[1], Andreas H. Diacon[2] and Thomas R. Niesler[1]

Abstract— **Automatic cough detection is key to tracking the condition of patients suffering from tuberculosis. We evaluate various acoustic features for performing cough detection using deep architectures. As most previous studies have adopted features designed for speech recognition, we assess the suitability of these techniques as well as their respective extraction parameters. Short-time Fourier transform (STFT), mel-frequency cepstral coefficients (MFCC) and mel-scaled filter banks (MFB) were evaluated using deep neural networks, convolutional neural networks and long-short term models. We find experimentally that, by regarding each cough sound as a single input feature instead of multiple shorter features, better performance can be achieved. Longer analysis windows also provide enhancement in contrast to the classic 25 ms frame. Although MFCC performance is improved by sinusoidal liftering, STFT and MFB lead to better results. Using MFB and the optimum segment and frame lengths, an improvement exceeding 7% in the area under the receiver operating characteristic curve across all classifiers is achieved.**

I. INTRODUCTION

Persistent coughing is a usual symptom of many respiratory diseases and has been widely used by physicians as a parameter to guide medical evaluation and track progress of treatment. For instance, the World Health Organization [1] advises that individuals with cough episodes for two weeks or more should be tested for Tuberculosis. Since coughing information is generally provided by the patients, several researches have attempted to develop automatic methods to detect, classify and count cough sounds in order to devise objective parameters for patient assessment.

A wide range of digital signal processing and machine learning techniques has been applied to differentiate coughs from other sounds. In terms of acoustic features, most of this previous work has adopted techniques from automatic speech recognition, especially mel-frequency cepstral coefficients (MFCC) and their variations, partitioning the sounds into segments and frames using the scheme illustrated in Fig. 1 [2]–[12]. MFCCs have also been combined with other features, such as F0, formant frequencies, spectral flatness and zero crossing rate, which some authors have found to increase system performance [13]–[15].

In the field speech recognition, recent studies have shown improved performance when the classic quefrency components used to compute MFCCs are replaced by mel-scaled filter banks (MFB), still using the partitioning scheme of Fig. 1 [16]–[18]. These authors highlight the spectro-temporal locality as relevant information for sound classification, especially when deep learning architectures are used. The same approach has also been applied to environmental sound classification [19], [20].

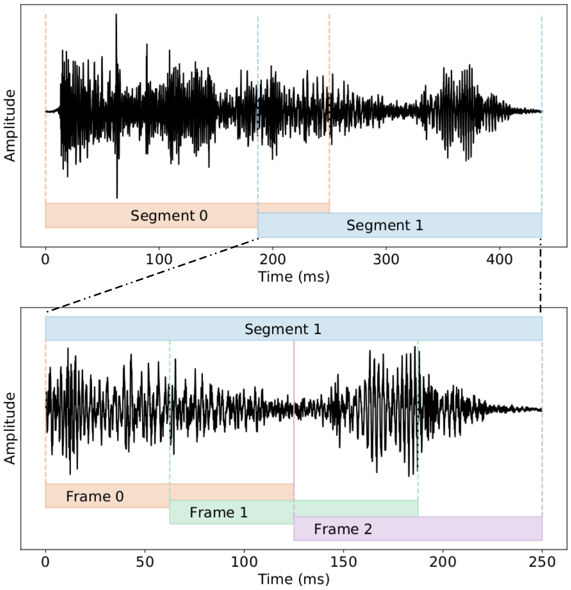


Fig. 1. A typical partitioning scheme used to extract spectral or cepstral features from a cough audio signal. In the top panel, segments are taken from the signal. In the bottom panel, a segment is divided into frames to compute multiple FFTs. The features computed from all frames in a segment constitute a single input vector for classification.

Other authors have reported improved cough detection using related feature extraction strategies. Larson at al [21] apply principal component analysis (PCA) to sound spectrograms, which has the benefit of preserving patient privacy. Amoh and Odame [22] use the short-time Fourier transform (STFT) as input vectors to convolutional neural networks (CNN) and long-short term models (LSTM), assuming that higher level features could be learnt automatically by these deep architectures.

It has however not been conclusively established which acoustic features are most effective for cough detection when using deep models as classifiers. Therefore, this study aims to assess the relative performance achieved by cough detectors when using STFT, MFB and MFCCs as acoustic features and deep neural networks (DNN), CNN and LSTM as classifiers.

*The study was funded by the Lungenliga St. Gallen-Appenzell, Switzerland. We thank the NVIDIA corporation for the donation of GPU equipment used for this research.

[1] Igor Miranda and Thomas Niesler are with Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa trn@sun.ac.za

[2] Andreas Diacon is with TASK Applied Science, Cape Town, South Africa ahd@sun.ac.za

Additionally, we will attempt to determine the best segment and frame lengths for the partitioning schemes used by these feature extraction techniques. The results presented in this paper form part of a larger and ongoing project aimed at the assessment of tuberculosis patients through automatic cough monitoring.

II. METHODS

A. Data Acquisition

Our dataset was compiled from two publically available sources. The first is the Audio Set provided by Google, which consists of extracts from 1.8 million Youtube videos that were manually labeled according to an ontology of 632 audio event categories [23]. The second is the Freesound audio database, which is not explicitly transcribed but is accompanied by titles and descriptions provided by the users who uploaded them.

The Audio Set is a selection of 10-second video clips with one or more labels indicating what the human annotators could hear. As these labels were assigned to the entire video, we performed a second annotation step to identify where each sound event starts and ends. This process was accomplished semi-automatically, by first using an energy detector to determine the boundaries of possible sound events and then labeling each individual event manually. This step also served to confirm the correctness of the provided labels. The same procedure was applied to the Freesound data.

Since we ultimately aim to deploy our cough detector in hospital wards or in domestic environments, we included sound classes other than coughing, specifically speech, sneezing, domestic/home sounds and throat clearing. The domestic/home category bundles together sounds labelled in the Audio Set that relate to daily home activities, such as door slams, collisions between objects, toilet flushing and running engines. From the Freesound database, all audio files that have the words ”cough”, ”sneeze” or ”throat” in their titles or descriptions were considered for inclusion.

The final dataset consists of 7781 audio excerpts extracted from 3132 files, corresponding to an average of 2.5 excerpts per file. Since it is likely that different files were uploaded by different users, a meaningful diversity has been achieved in terms of sources, background noises, distortion, reverberation and so forth, which is important for good generalization. The dataset composition is shown in Table 1. Note that 1151 files were used to collect coughs, which suggests a high number of individuals for this particular category in our dataset. In comparison, other studies considering coughing generally contain recordings from fewer than 20 individuals [2], [11], [21], [22].

B. Acoustic Features

We evaluate the cough detection performance of three audio features: STFT, MFB and MFCC. These features were chosen because they have been widely and successfully used in speech recognition systems before and since the advent of deep architectures. In particular, as the end-to-end approach becomes more prominent, less-engineered features such as STFT and MFB have become a common choice. However, MFCCs are still employed in many studies since they continue to provide reasonable performance in conjunction with current machine learning models.

TABLE I. Composition of the audio dataset.

Category	**Files**	**Audio Excerpts**
Cough	1151	3114
Speech	501	2326
Sneeze	778	1013
Throat Clearing	304	301
Domestic/Home Sounds	398	1027
Total	3132	7781

In our preliminary experiments, the following sinusoidal liftering over the MFCCs, designed for noisy speech recognition [24], [25], showed an improvement in cough detection.

$$w_i = 1 + \frac{M}{2} sin\left(\frac{\pi i}{M}\right) \tag{1}$$

In Equation 1, w_i, $i = 1, 2, ..., M$, are the lifter weights applied to the M cepstral coefficients. Both raw MFCCs and liftered-MFCCs (L-MFCCs) were included in our analysis.

All features mentioned above are extracted from audio segments that are further divided into a sequence of consecutive overlapping frames, as depicted in Fig. 1. Features extracted from all frames within each segment are regarded as a single classifier input vector.

The choice of segment and frame durations in this partitioning scheme remains an open question for cough detection. Two approaches were considered for experimental evaluation: multiple short segments per sound event or a single long segment. Frame lengths optimal for speech recognition have been adopted in several studies, but the specific characteristics of coughs may mean that different values would be better suited. Both segment and frame durations will be subject to the analysis presented in Section III.

C. Experimental Setup

Features were evaluated with DNN, CNN and LSTM classifiers, using the area under the receiver operating characteristic curve (AUC) as a performance measure. The architectures of the classifiers were inspired by previous work on small-footprint keyword spotting [26]–[28].

The DNN implementation consisted of three hidden layers with 128 ReLu units per layer. The CNN had one convolutional layer and two 128-unit hidden layers. Filter structures of 5x5x32 and 2x2x1 were used for the convolutional and the max-pooling layers, respectively, padding their inputs to maintain the same dimensions. Two 832-unit layers were adopted for the LSTM implementation. As preliminary results did not show improvement using multiclass classifiers, a two-class softmax layer was chosen as output for all models.

Librosa and Keras with a Tensorflow backend were used to extract features and train classifiers. Training was performed using the Adam optimizer with categorical cross-entropy as a loss function and early stopping. Three sets of experiments were performed. In the first, a hyperparameter search was used to find the segment and frame durations achieving the

TABLE II. Top five cross-validated results of grid search for each feature type and segment/frame length (ms) using DNNs.

STFT		**MFB**		**MFCC**		**L-MFCC**	
Seg. / Frame	AUC	Seg. / Frame	AUC	Seg. / Frame	AUC	Seg. / Frame	AUC
640 / 32	0.943	640 / 32	0.955	640 / 128	0.917	640 / 128	0.935
640 / 24	0.942	640 / 128	0.953	800 / 128	0.915	800 / 128	0.933
800 / 32	0.939	800 / 128	0.953	480 / 128	0.915	640 / 32	0.932
640 / 128	0.939	800 / 32	0.952	960 / 128	0.913	640 / 64	0.932
800 / 128	0.939	640 / 24	0.952	640 / 64	0.913	640 / 24	0.931

TABLE III. Top five cross-validated results of grid search for each feature type and segment/frame length (ms) using CNNs.

STFT		**MFB**		**MFCC**		**L-MFCC**	
Seg. / Frame	AUC	Seg. / Frame	AUC	Seg. / Frame	AUC	Seg. / Frame	AUC
640 / 24	0.959	800 / 64	0.973	640 / 64	0.933	640 / 64	0.955
640 / 64	0.958	640 / 64	0.973	640 / 24	0.931	640 / 128	0.951
960 / 24	0.954	640 / 128	0.972	640 / 32	0.930	800 / 128	0.949
640 / 32	0.954	960 / 128	0.971	800 / 64	0.930	640 / 24	0.948
800 / 32	0.953	960 / 64	0.971	480 / 64	0.929	800 / 64	0.948

TABLE IV. Top five cross-validated results of grid search for each feature type and segment/frame length (ms) using LSTMs.

STFT		**MFB**		**MFCC**		**L-MFCC**	
Seg. / Frame	AUC	Seg. / Frame	AUC	Seg. / Frame	AUC	Seg. / Frame	AUC
800 / 64	0.950	640 / 64	0.956	640 / 64	0.938	480 / 64	0.934
640 / 64	0.949	800 / 64	0.955	480 / 64	0.936	640 / 128	0.928
960 / 64	0.948	960 / 128	0.952	640 / 128	0.935	480 / 128	0.928
480 / 32	0.946	800 / 128	0.952	480 / 24	0.932	800 / 128	0.922
480 / 64	0.941	640 / 128	0.952	800 / 128	0.930	640 / 64	0.917

highest performance for each type of feature. The second experiment optimised the number of MFB filter banks, while the third experiment considered whether MFCC derivatives provide any performance improvement, as they do in speech recognition. Experimental results were compared with a baseline system using MFCCs with derivatives and a 25 ms frame shifted by 10 ms.

To ensure the robustness of our results, a stratified cross-validation scheme was applied to all experiments. Firstly, the dataset was divided into training and test sets using a 80/20 ratio. Then, using only the training set, features were evaluated using 10-fold cross validation, in which the folds were randomly selected and AUC was used as a performance criterion. The best parameters found for each of the three experiments were applied to the test set. Both train/test and inner 10-fold divisions were performed while considering not only the class stratification but also that sounds from the same source should be assigned to the same set and fold.

III. RESULTS AND DISCUSSION

A. Evaluation of Features and Segment/Frame Durations

A grid search was performed in order to evaluate the chosen features extracted with different segment and frame durations for each classifier described in Section II-C. As segment lengths, values of 160, 320, 480, 640, 800 and 960 ms were considered while for frame lengths values of 24, 32, 64 and 128 ms were used. Within each segment, frames were extracted with a 50% overlap.

Segments were extracted from the audio event with a 25% overlap until a maximum length of 1 s was reached, which correspond to the maximum cough duration [29]. For segment lengths greater than 500 ms, a single segment was extracted. All features were computed using the logarithm of the squared magnitude of the spectral coefficients, to which 40-dimensional mel scaling was applied for MFB, MFCC and L-MFCC. For both types of MFCCs, the first 13 cepstral coefficients were used as features.

The five best results for DNN, CNN and LSTM classifiers are shown in Tables II, III and IV, respectively. Although there is no single best configuration of segment and frame durations, larger values generally exhibit better performance across all considered features and classifiers. Some configurations with short segments also performed well when using a LSTM classifier, producing the best outcome for L-MFCC.

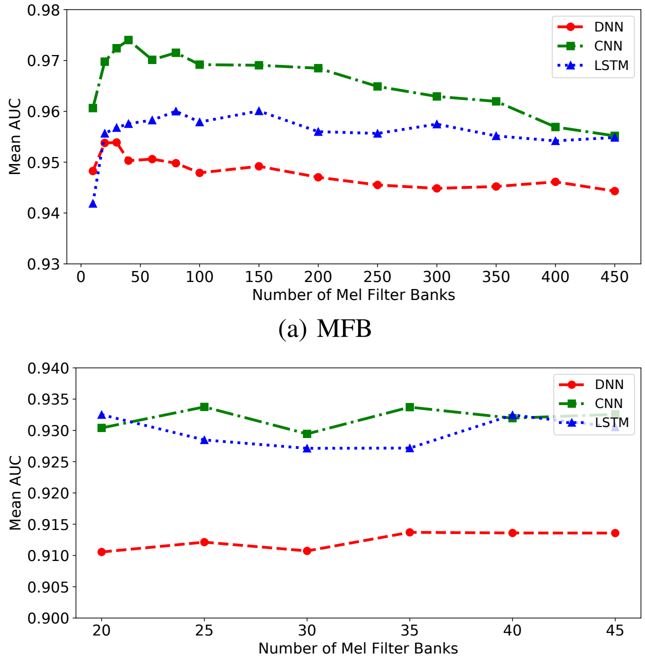
The success of long frame lengths is an important observation, as it differs from established practice in automatic speech recognition systems and also from most previous cough detectors, which almost always use 25 ms frames. This is consistent with the fact that deep architectures have the ability to learn from low level features, which in this case may be the sound envelope.

From the Tables II, III and IV, lengths of 640 ms and 64 ms for segments and frames are reasonable choices as they provided improved performance across all considered features and classifiers. These two parameter values will be utilized in the remaining experiments, unless otherwise specified.

It is also evident from Tables II, III and IV that, for all classifiers, the best performance was achieved with MFB features. This is consistent with recent progress in speech recognition using deep architectures.

B. Evaluation of Mel Filter Bank Dimension

Mel filter banks with between 25 and 40 filters have been commonly used for both MFCC and MFB extraction. In a second experiment, we investigated the dependence of



(a) MFB

(b) MFCC

Fig. 2. Cough classification performance as a function of the filter bank dimension for (a) MFB features and (b) MFCC features.

performance on the number of filters. For MFB, filter banks sizes between 10 and 450 have been considered. For MFCCs, filter bank dimensions between 20 and 45 were considered for the computation of 13 cepstral coefficients. Results are depicted in Fig. 2.a for MFB and in Fig. 2.b for MFCCs.

For MFB, best results were obtained with 30, 40 and 150 filter banks for DNN, CNN and LSTM, respectively. For MFCCs, DNN achieved best performance with 40 filter banks whereas CNN and LSTM performed best with 30. Fig. 2 confirms that, as in speech recognition, 40 dimensional mel filter bank are a reasonable choice for cough detection.

C. *Evaluation of MFCC Derivatives*

Experiments were performed to evaluate whether or not appended MFCC derivatives provide improvements as commonly assumed. This evaluation was performed with 40-dimensional mel filter banks and 13 cepstral coefficients, and then computing first and second derivatives. For CNN, the MFCCs and derivatives were organized in a three-dimensional fashion as proposed for speech [19]. The results for MFCC and L-MFCC are presented in Table V. Note that, in most cases, derivatives do not improve classification performance. Indeed, deteriorated performance is usually observed.

D. *Test Set Results*

Table VI presents results for each classifier on the test set for all features considered in this study and the baseline features. Again, the four studied features used 40-dimensional mel filter banks and 13 cepstral coefficients.

TABLE V. Cough classification performance in terms of AUC when using MFCCs with and without derivatives.

Feature	**DNN**	**CNN**	**LSTM**
MFCC	0.914	0.933	0.932
MFCC+Δ+2Δ	0.904	0.920	0.905
L-MFCC	0.934	0.954	0.931
L-MFCC+Δ+2Δ	0.931	0.945	0.940

TABLE VI. Cough classification performance on the test set.

	DNN		**CNN**		**LSTM**	
Feature	Acc.	AUC	Acc.	AUC	Acc.	AUC
Baseline	0.792	0.865	0.843	0.915	0.813	0.863
MFCC	0.805	0.881	0.853	0.925	0.847	0.919
L-MFCC	0.857	0.927	0.876	0.944	0.845	0.918
STFT	0.869	0.932	0.877	0.946	0.873	0.938
MFB	0.883	0.940	0.912	0.965	0.866	0.912

We see that similar improvements are observed on the test set in comparison with the experiments reported in previous sections, except when using LSTMs. This may be explained by the fact that audio input sequences for this application are short and have highly variable length, increasing the variance of LSTM training, due to its architecture and learning procedure.

IV. CONCLUSION

We have presented a comparative evaluation of different acoustic features for automatic cough detection using deep architectures. MFCC, liftered-MFCC, STFT and MFB features were compared using DNN, CNN and LSTM classifiers. To conduct this evaluation, we compiled a dataset with a substantial number of coughing individuals in comparison with previous studies.

Our experiments indicate that using a single long audio segment during feature computation provides better performance than the subdivision into smaller overlapped segments. Additionally, better performance was achieved using frames longer than the standard 25 ms window. Although these values may be fine-tuned for a specific architecture, 640 ms segments and 64 ms frames perform well across all classifiers and features considered. In agreement with experience in the field of speech recognition, 40-dimensional mel filter banks also provided good results for both MFCCs and MFB.

STFT and MFB features provided the best cough detection accuracies, adding to a growing corpus of research showing that less-engineered features may provide better performance with deep architectures.

Although MFCCs have not performed as well as the other candidates considered, many systems will continue using these legacy features. In this case, it is advantageous to apply sinusoidal liftering as it provides a small improvement for DNN and CNN classifiers, but not for LSTM. MFCC first and second derivatives did not enhance performance and can be omitted, thereby reducing feature dimensionality.

References

[1] World Health Organization (WHO), "Early detection of tuberculosis: an overview of approaches, guidelines and tools," 2011.

[2] S. Matos, S. S. Birring, I. D. Pavord, and H. Evans, "Detection of cough signals in continuous audio recordings using hidden Markov models," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 6, pp. 1078–1083, 2006.

[3] S.-H. Shin, T. Hashimoto, and S. Hatano, "Automatic detection system for cough sounds as a symptom of abnormal health condition," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 486–493, 2009.

[4] Z. Sun, A. Purohit, K. Yang, N. Pattan, D. Siewiorek, A. Smailagic, I. Lane, and P. Zhang, "Coughloc: Location-aware indoor acoustic sensing for non-intrusive cough detection," in *Proc. International Workshop on Emerging Mobile Sensing Technologies, Systems, and Applications*, 2011.

[5] B. H. Tracey, G. Comina, S. Larson, M. Bravard, J. W. López, and R. H. Gilman, "Cough detection algorithm for monitoring patient recovery from pulmonary tuberculosis," in *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2011, pp. 6017–6020.

[6] J.-M. Liu, M. You, Z. Wang, G.-Z. Li, X. Xu, and Z. Qiu, "Cough detection using deep neural networks," in *Proc. IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2014, pp. 560–563.

[7] B. Ferdousi, S. F. Ahsanullah, K. Abdullah-Al-Mamun, and M. N. Huda, "Cough detection using speech analysis," in *Proc. 18th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2015, pp. 60–64.

[8] J.-M. Liu, M. You, Z. Wang, G.-Z. Li, X. Xu, and Z. Qiu, "Cough event classification by pretrained deep neural network," *BMC Medical Informatics and Decision Making*, vol. 15, no. 4, p. S2, 2015.

[9] M. You, Z. Liu, C. Chen, J. Liu, X.-H. Xu, and Z.-M. Qiu, "Cough detection by ensembling multiple frequency subband features," *Biomedical Signal Processing and Control*, vol. 33, pp. 132–140, 2017.

[10] J. Monge-Álvarez, C. Hoyos-Barceló, K. Dahal, and P. Casaseca-dela Higuera, "Audio-cough event detection based on moment theory," *Applied Acoustics*, vol. 135, pp. 124–135, 2018.

[11] J. Monge-Alvarez, C. Hoyos-Barcelo, P. Lesso, and P. Casaseca-dela Higuera, "Robust detection of audio-cough events using local Hu moments," *IEEE Journal of Biomedical and Health Informatics*.

[12] L. Di Perna, G. Spina, S. Thackray-Nocera, M. G. Crooks, A. H. Morice, P. Soda, and A. C. den Brinker, "An automated and unobtrusive system for cough detection," in *Proc. IEEE Life Sciences Conference (LSC)*, 2017, pp. 190–193.

[13] T. Drugman, J. Urbain, and T. Dutoit, "Assessment of audio features for automatic cough detection," in *Proc. 19th European Signal Processing Conference*. IEEE, 2011, pp. 1289–1293.

[14] V. Swarnkar, U. Abeyratne, Y. Amrulloh, C. Hukins, R. Triasih, and A. Setyati, "Neural network based algorithm for automatic identification of cough sounds," in *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 1764–1767.

[15] Y. A. Amrulloh, U. R. Abeyratne, V. Swarnkar, R. Triasih, and A. Setyati, "Automatic cough segmentation from non-contact sound recordings in pediatric wards," *Biomedical Signal Processing and Control*, vol. 21, pp. 126–136, 2015.

[16] C. V. Cotton and D. P. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 69–72.

[17] A. Mohamed, "Deep neural network acoustic models for ASR," Ph.D. dissertation, Graduate Department of Computer Science, University of Toronto, 2014.

[18] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting spectro-temporal locality in deep learning based acoustic event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 26, 2015.

[19] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[20] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[21] E. C. Larson, T. Lee, S. Liu, M. Rosenfeld, and S. N. Patel, "Accurate and privacy preserving cough sensing using a low-cost microphone," in *Proc. 13th International Conference on Ubiquitous Computing*. ACM, 2011, pp. 375–384.

[22] 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, 2016.

[23] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[24] B.-H. Juang, L. Rabiner, and J. Wilpon, "On the use of bandpass liftering in speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 11, 1986, pp. 765–768.

[25] K. K. Paliwal, "Decorrelated and liftered filter-bank energies for robust speech recognition," in *Proc. Sixth European Conference on Speech Communication and Technology*, 1999.

[26] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks." in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 14, 2014, pp. 4087–4091.

[27] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Proc. International Speech Communication Association (Interspeech)*, 2015, pp. 1478–1482.

[28] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 4580–4584.

[29] J. Korpáš, J. Sadloňová, and M. Vrabec, "Analysis of the cough sound: an overview," *Pulmonary Pharmacology*, vol. 9, no. 5-6, pp. 261–268, 1996.