2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-1179-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/EMBC46164.2021.9630082

Investigating the Relationship between Cough Detection and Sampling Frequency for Wearable Devices

Mahmoud Abdelkhalek[1,3], Jinyi Qiu[1,3], Michelle Hernandez[2,3], Alper Bozkurt[1,3] and Edgar Lobaton[1,3]

***Abstract*— Cough detection can provide an important marker to monitor chronic respiratory conditions. However, manual techniques which require human expertise to count coughs are both expensive and time-consuming. Recent Automatic Cough Detection Algorithms (ACDAs) have shown promise to meet clinical monitoring requirements, but only in recent years they have made their way to non-clinical settings due to the required portability of sensing technologies and the extended duration of data recording. More precisely, these ACDAs operate at high sampling frequencies, which leads to high power consumption and computing requirements, making these difficult to implement on a wearable device. Additionally, reproducibility of their performance is essential. Unfortunately, as the majority of ACDAs were developed using private clinical data, it is difficult to reproduce their results. We, hereby, present an ACDA that meets clinical monitoring requirements and reliably operates at a low sampling frequency. This ACDA is implemented using a convolutional neural network (CNN), and publicly available data. It achieves a sensitivity of 92.7%, a specificity of 92.3%, and an accuracy of 92.5% using a sampling frequency of just 750 Hz. We also show that a low sampling frequency allows us to preserve patients' privacy by obfuscating their speech, and we analyze the trade-off between speech obfuscation for privacy and cough detection accuracy.**

Clinical relevance—This paper presents a new cough detection technique and preliminary analysis on the trade-off between detection accuracy and obfuscation of speech for privacy. These findings indicate that, using a publicly available dataset, we can sample signals at 750 Hz while still maintaining a sensitivity above 90%, suggested to be sufficient for clinical monitoring [1].

I. Introduction

A cough is a significant symptom that lowers quality of life [2], [3] and it is the most common reason to seek medical attention [4]. The frequency of cough is used as a clinical marker to diagnose and monitor chronic respiratory conditions, most commonly asthma. Approximately 25% of chronic cough diagnoses have been associated with cough variant-asthma [5] and standardized questionnaires, such as the Asthma Control Test, assess the presence of nocturnal coughs. In particular, a recent study suggested that nocturnal coughs and sleep quality are useful indicators to enable early detection of asthma attacks [6].

Recording coughs can also be used for the diagnosis of other medical problems. For example, J. Laguarta et al. [7] achieved a sensitivity of 98.5% and a specificity of 94.2% for subjects diagnosed with COVID–19 by only using audio recordings of forced coughs. Hence, the continuous monitoring of coughs using a wearable device has great potential to enable the early detection of respiratory ailments and reduce healthcare costs.

Several commercial-off-the-shelf and research prototype hardware have been used in recent clinical studies for cough detection. These include:

- LifeShirt (Vivometrics Inc., Ventura, CA, U.S.A.), which combines a respiratory inductance plethysmography system with an accelerometer and a unidirectional microphone on the upper-chest [8];
- LR102 (Logan Sinclair Research), which combines electromyography with a contact sound transducer [9];
- Pulmotrack-CC (KarmelSonix, Haifa, Israel), which combines a pneumogram belt with two contact microphones [10] and
- The Cayetano Cough Monitor, Leicester Cough Monitor, and VitaloJAK (Vitalograph, UK), which use contact and lapel microphones attached to the upper sternum for simple audio measurements [11] [12] [13].

More recently, devices using only accelerometers were introduced by academic research teams and startup companies for lower power consumption, reduced hardware cost, and much smaller form factors. These include ADAMM (Healthcare Originals, Rochester, NY, USA) [14], Rogers Lab (Northwestern University, Evanston, IL, USA) [15] and Sonasure (Quvium Ltd, UK) [16]. Additionally, other research studies in the literature focused on using patients' smartphone microphones [7], [17], [18].

A minimum cough detection sensitivity of 90% is recommended in the literature for sufficient clinical monitoring [1]. Several studies reported satisfying this requirement, but using controlled data that was collected in a clinical setting. For example, S. Birring et al. [12] implemented an ACDA using Gaussian Mixture Models and Hidden Markov Models, achieving an average sensitivity of 91%. F. Barata et al. [19] made use of a CNN to achieve an average sensitivity of 89.1%. However, neither of these studies make their data publicly available, and so their results cannot be verified.

*This work was supported by National Science Foundation (NSF) under award CNS-1552828, IIS-1915599, IIS-1915169 and EEC-1160483 (ERC for ASSIST).

[1]Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695, USA; Corresponding author: edgar.lobaton@ncsu.edu

[2]Department of Pediatrics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27559, USA

[3]M.A. developed the methodology, and planned and performed the numerical experiments. J.Q. identified the models, verified their architectures and contributed with the evaluation. M.H., A.B. and E.L. contributed with the main conceptual ideas, background search and medical relevance. E.L. supervised the research efforts. All authors discussed the results and contributed to the final manuscript.

Furthermore, concerns related to patient privacy were neglected in these studies. Conversely, the authors in [20] developed an ACDA that aims to preserve patients' privacy by obfuscating their speech. More specifically, they reduced the dimensionality of magnitude spectrograms using principal component analysis, from which a subset of principal components was used for training a classifier. This subset was also used for the reconstruction of audio, showing that the intelligibility of speech decreases as the number of chosen eigenvectors decreases. However, it is possible that other dimensionality reduction techniques could have reduced the intelligibility of speech more efficiently.

Our main contributions are:

- Development of an ACDA that can achieve a sensitivity of 92.7%, a specificity of 92.3%, and an accuracy of 92.50% using a sampling frequency of 750 Hz.
- Modelling the relationship between audio window lengths and the classification performance of the ACDA at a sampling rate of 16 kHz.
- Modelling the relationship between sampling rate and the level of speech obfuscation for a window length of 1.5 seconds.
- Development and release of an open-access audio dataset consisting solely of coughs to further the development of reproducible ACDAs.

Figure 1 provides an overview of our proposed methodology in the scope of this paper. First, an embedded system device capable of performing some standard signal processing (e.g., applying a linear filter) is used to collect the data. This data is then transmitted to a data aggregator (e.g., a smartphone) for recording and processing. Our goals are to identify cough instances accurately and to analyze the effect of the filtering operation on the obfuscation of the signal for privacy purposes. We initially focus on a downsampling filter for simplicity, interpretability of the analysis, and reducing the data streaming rate and power consumption of the embedded system. More complex filtering will be applied in our future work. The paper is organized as follows: Section II provides an overview of the methodology for cough detection, section III reports the results of our analysis, and section IV concludes with possible extensions of this work.

II. Methodology

A. Overview

We focus on the case in which a given audio segment first passed through a downsampling filter h_θ with parame θ, which can be thought of as the downsampling rate. filtered audio segment is then passed through the functio with parameters ϕ. The purpose of this function is to as a label $\hat{y}$ to the filtered audio segment such that $\hat{y} \in \{0$ where a label of 0 corresponds to an audio segment contains human speech and a label of 1 corresponds t audio segment that contains a cough.

We will treat θ as a hyper-parameter for our anal to examine its effect on performance as a function of

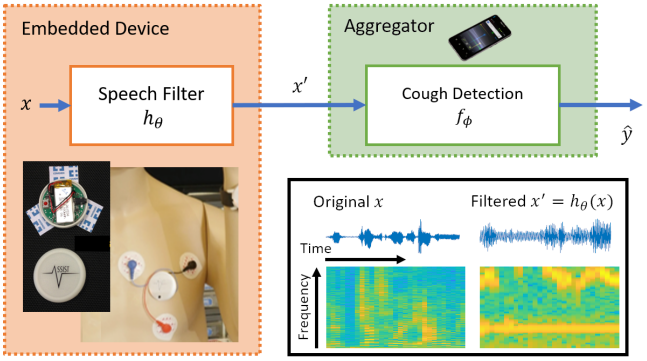

Fig. 1: Overview of our proposed solution. Acoustic signal x is captured using an embedded wearable device (e.g., [21]). The signal is processed on the device obtaining the filtered signal $x' = h_\theta(x)$. An aggregator (e.g., a smartphone) received the filtered signal and uses it to come up with a prediction $\hat{y} = f_\phi(h_\theta(x))$ for cough or non-cough. The signals on the bottom-right show samples of x and x'.

downsampling rate. Given that $L_y(\hat{y})$ is the classification loss, we aim to find

$$\phi^* = \underset{\phi}{\operatorname{argmin}} \sum_x L_y(f_\phi(h_\theta(x))) \quad (1)$$

where the summation is done over all training data points.

B. Implementation

If the audio segment x is to be resampled from the sampling frequency F_s^{old} to F_s^{new}, then to account for rational resampling rates, the filter h_θ consists of three stages: upsampling, anti-aliasing, and downsampling. If $\frac{[F_s^{\text{new}}]}{[F_s^{\text{old}}]}$ is a reduced fraction, then the upsampling stage involves inserting $[F_s^{\text{new}}]-1$ zeros between consecutive samples in the audio segment x. Next, the anti-aliasing stage is implemented using a windowed sinc low-pass filter with a cut-off frequency equal to $\min\left(\frac{F_s^{\text{old}}}{2}, \frac{F_s^{\text{new}}}{2}\right)$ Hz, while the downsampling stage involves selecting every $[F_s^{\text{old}}]$-th sample of the output of the anti-aliasing stage and discarding the other samples. The resampling operation was implemented using the software package given in [22].

Figure 2 shows the implementation of f_ϕ. The proposed architecture combines a pre-trained CNN for feature extraction [23], which we refer to as FENet, with a logistic regression

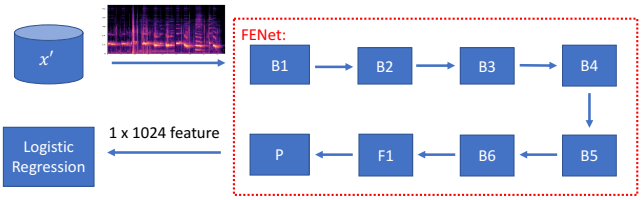

Fig. 2: Implementation of f_ϕ. The Mel-Spectrogram of x' is used as an input, and the FENet and a logistic regression classifier are combined to provide the desired output.

classifier to predict $\hat{y}$. We first obtain the Mel-Spectrogram [24] of x' and use it as an input. Then, blocks B1 to B5 each consist of two cascaded copies of a convolutional layer followed by batch normalization and then a ReLU activation, with 16, 32, 64, 128, and 256 feature maps respectively. Block B6 consists of a convolutional layer with 512 feature maps followed by batch normalization, a rectified linear unit (ReLU) activation function, defined as $\text{ReLU}(x) = \max(x, 0)$, and then max-pooling. Block F1 consists of a convolutional layer with 1024 feature maps followed by batch normalization and a ReLU activation. Finally, block P is a global pooling layer that averages all pixels in each of the 1024 feature maps resulting in a 1024-dimensional feature vector. FENet was pre-trained on Google's AudioSet [25], which contains weakly-labelled (e.g. a 1 second audio window labelled as y may contain only 0.2 s of audio labelled as y) audio examples of 527 sound events. We chose this CNN architecture because of its promising performance on environmental sounds (including human and non-human sounds), as discussed further in [23]. The logistic regression classifier outputs a prediction $\hat{y}$ based on a 1024-dimensional feature vector that is output from FENet. Both FENet and the logistic regression classifier were implemented in PyTorch [26].

C. Data Collection and Augmentation

We collected 269 audio files in WAV format labelled as "cough" from the ESC-50 [27] and FSDKaggle2018 [28] datasets, and 28,539 audio files in FLAC format labelled as "speech" from the LibriSpeech [29] dataset, which were later converted to WAV format during training and testing. All audio files ranged from 5 s to 30 s in length. The audio files extracted from the LibriSpeech were strongly labelled with each file consisted of only speech audio. The audio files extracted from the ESC-50 and FSDKaggle2018 datasets were weakly labelled; these also contained audio with other labels in addition to cough audio. Also, there was no guarantee that these files contained only 1 cough, and could contain no coughs or more than 1 cough. To avoid mixing the training data with the testing data, we split this dataset into training (80%), validation (10%), and testing (10%) subsets before processing and augmentation.

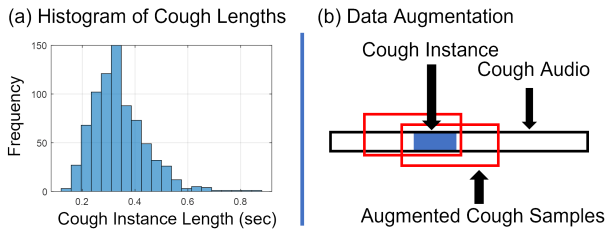

Fig. 3: (a) Histogram of the cough instance lengths. (b) Overview of the random sampling of fixed-length cough windows. The filled blue rectangle represents a cough instance within an audio file. The red rectangles are randomly sampled fixed-length window around this cough instance.

Given the 269 audio files labelled as "cough", we manually extracted the timestamps of 768 individual coughs that ranged from 120 ms to 880 ms in length. Figure 3 (a) illustrates the histogram of the cough instances captured. We are making these extracted coughs publicly available in [30]. We used the cough template given in [1] to determine what a cough looks like while extracting these timestamps. Additionally, because our proposed solution requires input signals of fixed length, and because we had much more speech audio than cough audio available, we extracted a large number of fixed-length windows around these variable-length coughs via uniform random sampling. Figure 3 (b) shows what this extraction and augmentation procedure looks like.

To extract windows from speech audio files, we uniformly randomly sampled fixed-length windows from these audio files and then used the frame admission protocol given in [31] to accept windows that met certain criteria, such as its energy content and entropy. In total, for a window length of 1.5 s, the random sampling and augmentation process yielded 113,574 speech windows and 94,019 cough windows.

III. Results and Discussion

In this section, we discuss the metrics used for evaluation, and the impact of window size and filtering on performance.

A. Evaluation Metrics

We used the Virtual Speech Quality Objective Listener (ViSQOL) [32] as a metric to evaluate the degradation of speech quality caused by our filters. We made use of publicly available implementations by the authors [33].

ViSQOL compares the original signal x and target signal z to provide a speech quality score $S_{\text{VIS}}(x, z)$ using the Mean Opinion Score (MOS) [34]. This score ranges between 1 (bad) and 5 (excellent). We observed that if $z = x$, the score will be 5. If z is random white noise, we observed an average score of 1.03 with standard deviation of 0.11. When evaluating performance, we report the following average score:

$$\bar{S}_{\text{VIS}} = \sum_{x} S_{\text{VIS}}(x, x'), \tag{2}$$

where the summation is over all x samples in the testing set.

To evaluate the performance of the classifier f_ϕ, we used three metrics: Sensitivity (sens), specificity (spec), and accuracy (acc). These metrics were computed using the number of true positives (TP), the number of true negatives (TN), the number of false positives (FP), and the number of false negatives (FN) predicted by the classifier f_ϕ:

$$\text{sens} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$\text{spec} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$
$$\text{acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

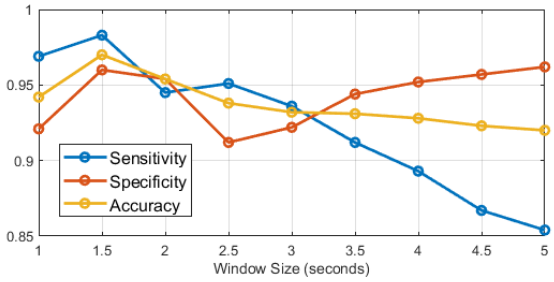

Fig. 4: Performance for various window sizes at 16 kHz sampling rate. The trends show a higher performance using a 1.5 s window.

B. Impact of Window Size on Classification Performance

We investigated the effect that the length of each x has on the performance of the classifier f_ϕ. As discussed in section II-C, the maximum length of a cough was observed to be 880 ms. To compare results with other studies, such as [35], we evaluated our classifier using window lengths of 1 s up to 5 s in increments of 0.5 s, as shown in Figure 4. A window length of 1.5 s consistently outperformed other window lengths in terms of sensitivity, specificity, and accuracy. Therefore, the length of x was chosen to be 1.5 s.

Interestingly, beyond a window length of 3 s, the specificity and sensitivity of the classifier start to diverge. This could be due to the fact that it is more difficult for the classifier to distinguish a cough from background noise in longer windows, and hence long cough windows look more like speech windows. The classifier, then, labels a test window as speech more often.

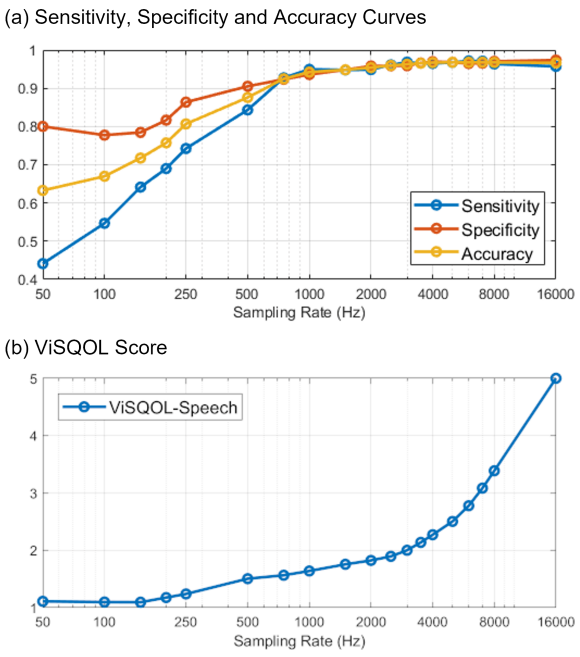

Fig. 5: Performance on the test set as a function of sampling frequency using a window length of 1.5 s. (a) Sensitivity, specificity and accuracy curves. Due to variability in the runs and to enhance visualization, we smoothed this plot by weighted average between neighboring data points. (b) ViSQOL score for the speech data points.

C. Impact of Filtering on Performance and Signal Quality

As shown in Figure 5 (a), the ability of the classifier f_ϕ to discriminate between cough audio and speech audio is almost unchanged from approximately 1 kHz up to 16 kHz. Moreover, a sampling frequency of 750 Hz can be used while still maintaining a sensitivity above 90%. This means that we can decrease the data transmission rate between the embedded system and the data aggregator, shown in Figure 1, and the data storage on the aggregator by approximately 95% without a substantial loss of performance.

Additionally, we were able to reduce the quality of speech by approximately 68%, from a ViSQOL score of 5 at 16 kHz to a ViSQOL score of 1.564 at 750 Hz, as shown in Figure 5 (b), thereby allowing us to obfuscate human speech to preserve patient privacy without sacrificing classification performance. Future work will focus on reducing the quality of speech such that the ViSQOL score is closer to 1.

Although the highest power of speech lies between 85-255 Hz, which is within our ranges of down sampling, human speech recognition requires the overtones to be present [36]. While signals sampled at 2.5 kHz were recognizable to us, we observed that the same signal sampled at 2 kHz was not.

IV. Conclusion and Future Work

In this work, we demonstrated the effectiveness of a classifier f_ϕ applied to cough and speech audio using various sampling frequencies and window sizes. This, in turn, allowed us to effectively perform obfuscation of human speech for preserving patient privacy. However, to generalize the detection of cough among multiple classes of audio recordings, including speech, we hypothesize that it is beneficial to treat this as an anomaly detection problem. More specifically, given that enough cough samples are provided, we can learn an accurate model of the distribution of coughs and treat any other audio as an outlier of such distribution.

Furthermore, in our study, the fact that humans cannot recognize speech does not necessarily mean a machine cannot recognize it. There have been studies aiming to connect human speech recognition and Automatic (i.e., machine) Speech Recognition (ASR) [37]. Meanwhile, ASR systems have evolved from using hand-crafted features [38] to more data-driven models through deep learning [39], [40]. In particular, models such as Deep Speech2 [41] are capable of recognizing speech in noisy environments. As part of our future work, we plan to investigate the impact of filtering on the performance of ASR systems, and consider more complex filters that aim to further obfuscate speech by reducing their performance.

Other directions for future work include:

- Improving the ACDA to discriminate coughs from more diverse classes of audio.

- Augmenting the ACDA with the ability to classify different kinds of coughs, such as a deep cough originating from the throat or a lighter cough.
- Conducting clinical trials using the improved ACDA.
- Investigating whether speech obfuscation interferes with the ability to discriminate coughs from other classes of audio.
- A more comprehensive comparison between the performance of the improved ACDA and other state-of-the-art ACDAs.

REFERENCES

[1] J. Smith and A. Woodcock, "New developments in the objective assessment of cough," *Lung*, vol. 186 Suppl 1, S48–54, 2008.

[2] L. Polley, N. Yaman, *et al.*, "Impact of Cough Across Different Chronic Respiratory Diseases: Comparison of Two Cough-Specific Health-Related Quality of Life Questionnaires," *Chest*, vol. 134, no. 2, pp. 295–302, Aug. 1, 2008.

[3] W. Ma, L. Yu, *et al.*, "Changes in health-related quality of life and clinical implications in Chinese patients with chronic cough," *Cough*, vol. 5, no. 1, p. 7, Sep. 25, 2009.

[4] P. V. Dicpinigaitis, "Clinical perspective—cough: An unmet need," *Current Opinion in Pharmacology*, Respiratory • Musculoskeletal, vol. 22, pp. 24–28, Jun. 1, 2015.

[5] W. M. Corrao, "Pearls and pitfalls in the diagnosis of cough variant asthma.," in *Allergy & Asthma Proceedings*, vol. 39, 2018.

[6] P. Tinschert, F. Rassouli, *et al.*, "Nocturnal cough and sleep quality to assess asthma control and predict attacks," *Journal of asthma and allergy*, vol. 13, p. 669, 2020.

[7] J. Laguarta, F. Hueto, *et al.*, "COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 275–281, 2020.

[8] M. A. Coyle, D. B. Keenan, *et al.*, "Evaluation of an ambulatory system for the quantification of cough frequency in patients with chronic obstructive pulmonary disease," *Cough*, vol. 1, no. 1, pp. 1–7, 2005.

[9] S. Leconte, G. Liistro, *et al.*, "The objective assessment of cough frequency: Accuracy of the lr102 device," *Cough*, vol. 7, no. 1, pp. 1–8, 2011.

[10] E. Vizel, M. Yigla, *et al.*, "Validation of an ambulatory cough detection and counting application using voluntary cough under different conditions," *Cough*, vol. 6, no. 1, pp. 1–8, 2010.

[11] A. Proaño, M. A. Bravard, *et al.*, "Protocol for studying cough frequency in people with pulmonary tuberculosis," *BMJ open*, vol. 6, no. 4, 2016.

[12] S. Birring, T. Fleming, *et al.*, "The leicester cough monitor: Preliminary validation of an automated cough detection system in chronic cough," *European Respiratory Journal*, vol. 31, no. 5, pp. 1013–1018, 2008.

[13] K. McGuinness, K. Holt, *et al.*, "P159 validation of the vitalojak™ 24 hour ambulatory cough monitor," *Thorax*, vol. 67, no. Suppl 2, A131–A131, 2012.

[14] M. Sterling, H. Rhee, *et al.*, "Automated cough assessment on a mobile platform," *Journal of medical engineering*, vol. 2014, 2014.

[15] K. Lee, X. Ni, *et al.*, "Mechano-acoustic sensing of physiological processes and body motions via a soft wireless device placed at the suprasternal notch," *Nature biomedical engineering*, vol. 4, no. 2, pp. 148–158, 2020.

[16] S. P. Schmidt, *Cough detection, analysis, and communication platform*, US Patent 10,820,832, Nov. 2020.

[17] L. Orlandic, T. Teijeiro, *et al.*, "The coughvid crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms," *arXiv preprint arXiv:2009.11644*, 2020.

[18] L. Kvapilova, V. Boza, *et al.*, "Continuous sound collection using smartphones and machine learning to measure cough," *Digital biomarkers*, vol. 3, no. 3, pp. 166–175, 2019.

[19] F. Barata, K. Kipfer, *et al.*, "Towards Device-Agnostic Mobile Cough Detection with Convolutional Neural Networks," in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, Jun. 2019, pp. 1–11.

[20] E. C. Larson, T. Lee, *et al.*, "Accurate and privacy preserving cough sensing using a low-cost microphone," in *Proceedings of the 13th International Conference on Ubiquitous Computing*, ser. UbiComp '11, New York, NY, USA: Association for Computing Machinery, Sep. 17, 2011, pp. 375–384.

[21] J. Dieffenderfer, H. Goodell, *et al.*, "Low-power wearable systems for continuous monitoring of environment and health for chronic respiratory disease," *IEEE journal of biomedical and health informatics*, vol. 20, no. 5, pp. 1251–1264, 2016.

[22] Torchaudio Contributors, *Torchaudio.transforms.resample*, version 0.8.0, Mar. 4, 2021.

[23] A. Kumar, M. Khadkevich, *et al.*, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 326–330.

[24] M. Huzaifah, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," *arXiv:1706.07156 [cs]*, Jun. 21, 2017.

[25] J. F. Gemmeke, D. P. W. Ellis, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[26] A. Paszke, S. Gross, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, *et al.*, Eds., Curran Associates, Inc., 2019, pp. 8024–8035.

[27] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*, Brisbane, Australia: ACM Press, Oct. 13, 2015, pp. 1015–1018.

[28] E. Fonseca, M. Plakal, *et al.*, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," *arXiv preprint arXiv:1807.09902*, 2018.

[29] V. Panayotov, G. Chen, *et al.*, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 5206–5210.

[30] M. Abdelkhalek, J. Qiu, *et al.*, *Coughs: Esc-50 and fsdkaggle2018*, version 1.0.0, Available at https://doi.org/10.5281/zenodo.5136592, Zenodo, Jul. 2021.

[31] H. Lu, W. Pan, *et al.*, "Soundsense: Scalable sound sensing for people-centric applications on mobile phones," in *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '09, Kraków, Poland: Association for Computing Machinery, 2009, pp. 165–178.

[32] M. Chinen, F. S. Lim, *et al.*, "Visqol v3: An open source production ready objective speech and audio metric," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2020, pp. 1–6.

[33] M. Chinen, F. S. C. Lim, *et al.*, *Google/visqol*, https://github.com/google/visqol, Accessed: 2021-03-23.

[34] R. C. Streijl, S. Winkler, *et al.*, "Mean opinion score (mos) revisited: Methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[35] J. Kim, "Urban sound tagging using multi-channel audio feature with convolutional neural networks," in *Detection and Classification of Acoustic Scenes and Events 2020*, Nov. 2020.

[36] R. J. Baken and R. F. Orlikoff, *Clinical measurement of speech and voice*. Cengage Learning, 2000.

[37] O. Scharenborg, "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," *Speech Communication*, vol. 49, no. 5, pp. 336–347, 2007.

[38] D. Dimitriadis, P. Maragos, *et al.*, "Robust am-fm features for speech recognition," *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 621–624, 2005.

[39] A. Coates, A. Ng, *et al.*, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings, 2011, pp. 215–223.

[40] D. Yu, M. L. Seltzer, *et al.*, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605*, 2013.

[41] D. Amodei, S. Ananthanarayanan, *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, PMLR, 2016, pp. 173–182.