SS+CEDNet: A Speech Privacy Aware Cough Detection Pipeline by Separating Sources

K. M. Naimul Hassan, Mohammad Ariful Haque
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka-1205, Bangladesh
naimul.hassan273@gmail.com, arifulhoque@eee.buet.ac.bd

***Abstract*—Cough is one of the most distinguishable symptoms for Influenza-like-illness (ILI) and Severe Acute Respiratory Infection (SARI). Considering the recent worldwide COVID-19 pandemic, many types of research are ongoing all around the world for the accurate detection of cough events. But background speech events make it difficult for the algorithms to detect cough events and the performance of the models drops significantly. At the same time, speech privacy is not preserved in the traditional cough detection models. In this paper, we are proposing a pipeline, named SS+CEDNet, to overcome these problems. The pipeline consists of a Source Separation (SS) and a Cough Event Detection (CED) model. The SS model at first separates the cough and speech sources. Finally, the separated cough source is passed through the CED model to detect cough events. The pipeline not only preserves speech privacy by separating the sources but also shows a better cough detection accuracy.**

***Index Terms*—cough, source separation, detection, privacy, audio**

I. Introduction

Since coughing is a natural and protective reflex, most of the time it is ignored. But cough is one of the most common symptoms for which patients worldwide seek medical attention. It is a major symptom of Influenza-like-illness (ILI) and Severe Acute Respiratory Infection (SARI). So, cough detection is a task that has caught attention of many researchers around the world.

Cough detection from audio recordings is one of the most common ways of detecting cough. There are several studies exploring audio-based cough recognition algorithms. The authors in [1]–[3] have used Mel-frequency cepstral coefficient (MFCC) along with Hidden Markov Model (HMM) to train cough recognition models. The authors in [4], [5] have used Spectrogram based feature to train the cough recognition models. Acoustic features such as: LPC [6], Hilbert Marginal Spectrum [7] or Gammatone Cepstral Coefficient [8] have also been used to train both static (e.g., Random Forest, Support Vector Machine) and temporal models (e.g., Hidden Markov Model, Recurrent Neural Network). There are also some recent studies where Convolutional Neural Network (CNN) is explored [4], [9]. One of the major concern of implementing these algorithms is the privacy of speech events in the background of the cough events. Cough event detection with speech in the background itself is a challenging problem. The authors in [9] have used data augmentation and trained their CNN model with cough data with background speech. It is shown that the model does not perform well in real scenario if it is not trained with background speech. So, adding background speech is required. But it will increase the size of the dataset and increasing data is expensive. To detect cough in a speech privacy preserving way, they have trained a cough detection model and a speech detection model separately. During their implementation, in case of storing data, if the cough detection model detects cough event for a segment and the speech detection model does not detect any background speech for that segment, then that segment is stored. If both the cough and speech detection model detect cough and speech for the same segment, it is not stored. This indicates a loss of cough data. The motivation of this research is to develop a pipeline which will improve the performance of the cough detection models and also will be able to detect cough events preserving the speech privacy without any loss of data. The idea is to separate the cough and speech sources from the mixed segment at first and then use the cough detection model on the separated sources. Audio source separation models are vastly used to encounter speech and music source separation problem. The authors in [10] have developed a novel waveform-to-waveform model, with a U-Net structure and bidirectional LSTM to separate music sources. A fully convolutional time-domain audio separation network (Conv-TasNet), a deep learning framework for end-to-end time-domain speech separation is developed in [11]. The authors in [12] used a combination of convolutional and recurrent neural networks. A multi-scale neural network which is an adaptation of U-Net is proposed in [13] for source separation. To our best knowledge, source separation models have not been applied for separating cough sound sources yet.

In this paper, a pipeline named as SS+CEDNet is proposed where a Source Separation (SS) model is used to separate the two (cough and speech) sources and Cough Event Detection (CED) model is used later. The performance of the pipeline is evaluated in terms of various evaluation metrics.

II. Problem Description

Let Y be an audio signal consisting of two sound sources - cough and speech. For source separation purpose, we assume that speech is in the background and cough is in the foreground. Y has a time duration of T seconds. If it

978-1-6654-0156-2/22/$31.00 ©2022 IEEE

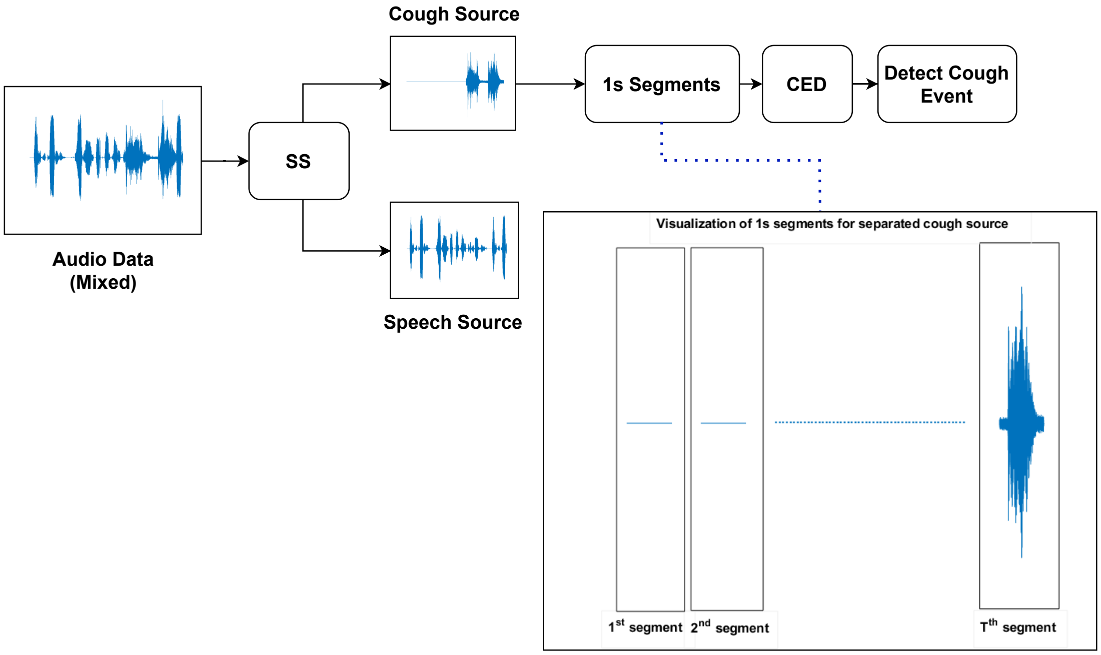

Fig. 1. Block Diagram of SS+CEDNet.

is segmented into 1s audio clips, then we will get a total number of T clips. So, Y now can be defined as the set of 1s segments, $Y = \{y_1, y_2, y_3, ..., y_T\}$, where y_i is the i^{th} number of segment $(i \in \{1, 2, 3, ..., T\})$. The goal is to develop an algorithm or a model which will predict properly if the y_i segment is a cough event or not and also will preserve the privacy by separating the speech source from the background.

III. Proposed System

The proposed speech privacy aware pipeline, named SS+CEDNet, consists of one SS model and one CED model. The block diagram of SS+CEDNet is shown in Fig. 1. At first, the input audio data, Y, is fed into the SS model. The SS model separates the sources and we get two output signals. One of the outputs contains the separated speech sound and the other contains the separated cough sound. Each of the outputs has a time duration same as the input audio signal. Then, the separated cough source is segmented into 1s clips. Finally, these 1s segments are passed through the CED model and the CED model predicts if the segment contains a cough event or not. Thus by separating the speech source earlier by the SS model, it is possible to detect cough events by the CED model maintaining the speech privacy.

A. Source Separation (SS)

In this study, we have used the Wave-U-Net architecture [13] as the SS model. The Wave-U-Net is a convolutional neural network for separating sound sources working directly on raw audio data. It is an adaptation of the U-Net architecture [14] to the one-dimensional time domain for performing end-to-end source separation. Through a series of downsampling and upsampling blocks involving 1D convolutions combined with a downsampling/upsampling process, features are computed on multiple scales/levels of abstraction and time resolution, and then combined to make a prediction. The model architecture of Wave-U-Net used in this study is shown in Fig. 2. The depth of the model was set to be 6 (6 upsampling blocks and 6 downsampling blocks) considering both performance and training time. If the input audio data has a duration of T s, then two audio data each having the same duration, T are extracted from Wave-U-Net. Each of the output audio data contains only one sound source (either cough or speech).

B. Cough Event Detection (CED)

After the sources are separated from SS model, the separated cough source is segmented into 1s clips. These 1s segments are used as the inputs for the CED model to detect cough events. Multiple CED algorithms, both deep learning and traditional machine learning approaches, were experimented in this study such as: Yet Another MobileNet (YAMNet) [15], Random Forest, Support Vector Machine (SVM) and Naive Bayes. Cough sound events (only cough/ cough+speech) and non-cough sound events (only speech) were considered as class 1 and class 0 respectively.

1) Deep Neural Network Approach: YAMNet is a pre-trained deep neural network which is able predict audio events from 521 classes, such as laughter, barking, or a siren. It employs the MobileNetV1 [16] depthwise-separable convolution architecture and can use an audio waveform as input. It can make independent predictions for each of the 521 audio events from the AudioSet corpus [17]. YAMNet can be used as a high-level feature extractor: the 1,024-

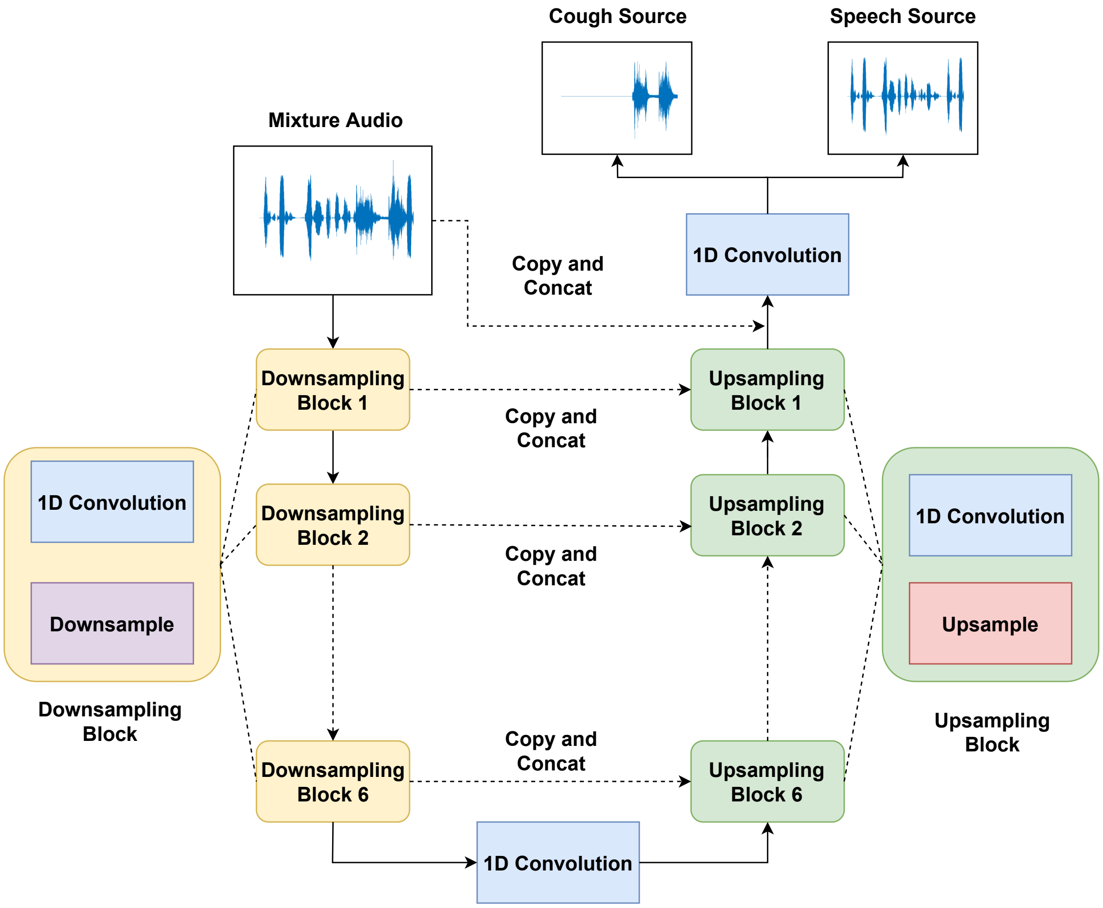

Fig. 2. Model Architecture of Wave-U-Net.

dimensional embedding output. In this experiment, the base (YAMNet) model's input features were used and they were fed into a shallower model consisting of two hidden layers. This network was trained on a small amount of data for classifying cough events.

2) Traditional Machine Learning Algorithms: Traditional machine learning algorithms such as: Random Forest (RF), Support Vector Machine (SVM) and Naive Bayes (NB) can be used to detect the cough events from different time and frequency domain based features extracted from the raw audio data. In this study, Short-time Fourier transform (STFT), mel-frequency cepstral coefficients (MFCC), zero crossing rate (ZCR), chromagram, melspectrogram, spectral contrast and tonnetz were extracted from each of the 1s segments and these features were used as the input of RF, SVM and NB.

IV. Experiment

In this section, the experimental setup for this study is discussed.

A. Dataset

1) Audio Data for SS Model: A dataset of soundscapes, where each soundscape was created by combining and transforming a set of existing audio files, was used to train and test the SS model. The existing audio files were taken from TIMIT [18], MUSAN [19] and Audio Set [17] datasets. Speech audio events were taken from TIMIT and MUSAN datasets. Cough audio events were taken from Audio Set corpus. Then the soundscapes were generated by the help of Scaper [20], a python library. Each of the soundscapes contain speech in the background and cough in the foreground. Each of the soundscapes generated by combining TIMIT speech and Audio Set cough data has a time duration of 5s. In case of soundscape combining MUSAN speech and Audio Set cough data, the time duration is 10s. The soundscapes were generated in such a way that the SNR is uniformly distributed between -10 to 25 dB (cough and speech events were considered as signal and noise respectively). A total number of 2400 soundscapes were generated for the experiment. The generated dataset was split into train, validation and test sets in the ratio of 80:20:20. A sample visualization of a generated soundscape along with the corresponding sources (speech and cough) is shown in Fig. 3.

2) Audio Data for CED Model: Speech data from TIMIT, MUSAN and cough data from Audio Set were used to train and test the CED models for binary classification. A total number of 8692 speech data and 6157 cough data were used

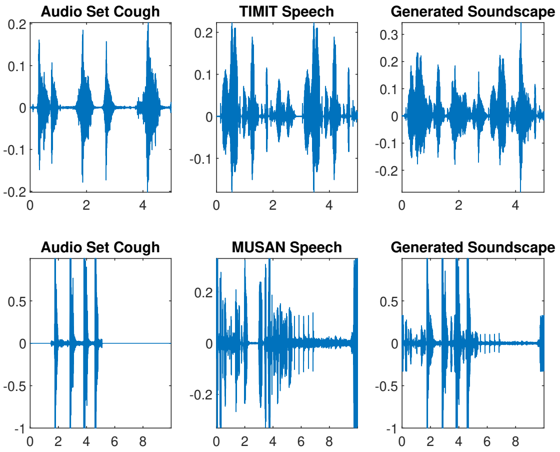

Fig. 3. Visualization of a Generated Soundscape

in the experiment and the length of each of the audio data is 1s. The dataset was split into train, validation and test sets in the ratio of 60:15:20.

B. SS Model Training

The SS model, Wave-U-Net was developed using the PyTorch framework. NVIDIA GeForce GTX 1650 GPU was used as the hardware accelerator for training the model. The loss function for the training algorithm was *L1 (Least Absolute Deviations)*. *Adam* was used as the optimizer algorithm with an initial learning rate of 0.001. The learning rate was decreased by a factor of 10 whenever the validation loss did not decrease or started increasing for consecutive epochs. The minimum learning rate set was 10^{-10}. The training was stopped early if there was no significant improvement of the validation loss for consecutive epochs. The summary of the training is provided in Table I.

TABLE I
SUMMARY OF THE TRAINING CONFIGURATIONS OF WAVEUNET

Accelerator	GPU
Loss function	L1
Optimizer	Adam
Initial learning rate	0.001
No. of epochs	2000
Execution time	12h 43m

C. CED Model Training

1) YAMNet Transfer Learning: YAMNet transfer learning model was developed using the Keras and Tensorflow frameworks. Kaggle's GPU was used as the hardware accelerator for training the model. Since the task is a binary classification, the loss function for the training algorithm was *Binary Cross-entropy*. *Adam* was used as the optimizer algorithm with an initial learning rate of 0.01 and similar to Wave-U-Net training, the learning rate was decreased whenever the validation loss did not decrease or started increasing for consecutive epochs. The minimum learning rate set was 10^{-15}. Finally, the training was stopped early if there was no significant improvement of the validation loss for consecutive epochs. The summary of the training is provided in Table II.

TABLE II
SUMMARY OF THE TRAINING CONFIGURATIONS OF YAMNET TRANSFER LEARNING

Accelerator	GPU
Loss function	Binary Cross-entropy
Optimizer	Adam
Initial learning rate	0.01
No. of epochs	20
Execution time	350s

2) RF, SVM and NB: As mentioned in section III-B2, STFT, MFCC, ZCR, chromagram, melspectrogram, spectral contrast and tonnetz were extracted from each of the 1s audio clips and these features were used to train and test RF, SVM and NB. The size of the feature vector was 193.

V. PERFORMANCE EVALUATION

The achieved results of the proposed approach along with the evaluation metrics are discussed in this section. At first, the individual performances of SS and CED models are discussed and finally the performance of the SS+CEDNet pipeline is discussed in detail.

A. Evaluation Metrics

1) Evaluation Metrics for SS: In this study, Source-to-Distortion Ratio (SDR) was used to evaluate the performance of the SS Model (Wave-U-Net). SDR can be defined as,

$$SDR = 10log_{10}(\frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2}) \qquad (1)$$

where s_{target} is the true source, and e_{interf}, e_{noise} and e_{artif} are error terms for interference, noise, and added artifacts, respectively [21].

Since there are two sources- cough and speech, the performance of SS task was evaluated using the following metrics:

- Cough SDR
- Speech SDR
- Overall SDR

2) Evaluation Metrics for CED: Since, the CED task is a binary classification, metrics used for the CED models are-

- Precision
- Recall
- F1 Score

B. Individual Performance Analysis of SS and CED Models

1) Performance of the SS Model: The performance of the SS model is shown in Table III. Higher SDR indicates a better source separation and an overall SDR of 11.87 indicates a near perfect source separation task. For better understanding, a visualization of the output of the SS model for a test sample along with the ground truth sources is shown in Fig. 4.

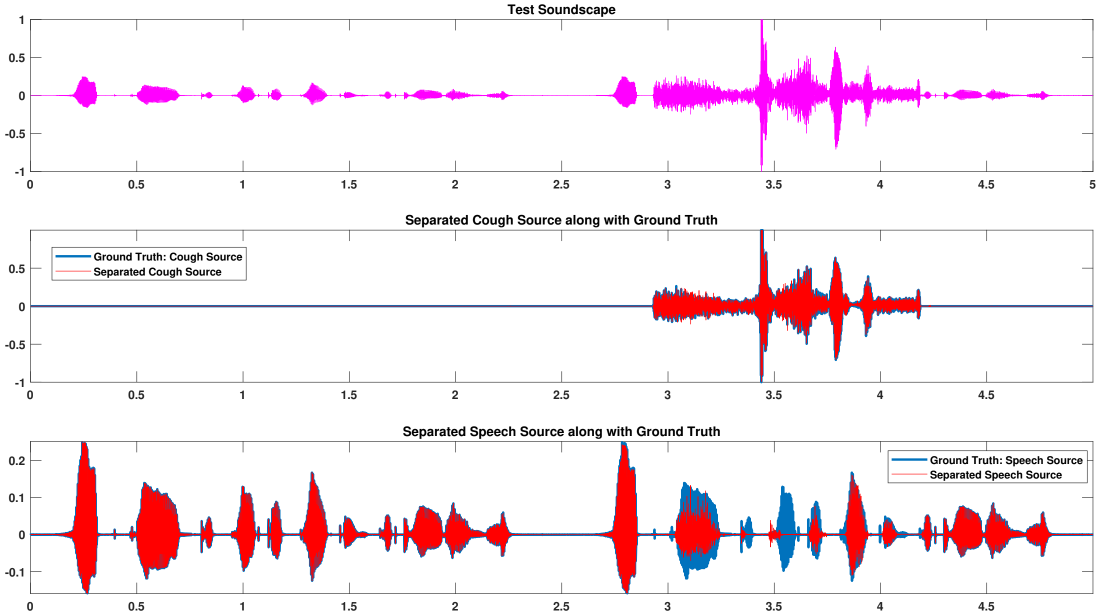

Fig. 4. Visualization of the Output of the SS model for a Test Sample along with the Ground Truth Sources

TABLE III
PERFORMANCE OF THE SS MODEL

Source	SDR
Cough	13.02
Speech	10.73
Overall	11.87

2) Performance of the CED Models: All the CED models are evaluated in terms of precision, recall and F1 score. The summary of the results are shown in Table IV.

TABLE IV
PERFORMANCE OF THE CED MODELS

Model	Precision (%)	Recall (%)	F1 Score (%)
YAMNet	99.28	99.92	99.60
RF	99.83	99.95	99.89
SVM	99.91	99.95	99.93
NB	94.71	98.98	96.80

From the results, it can be clearly understood that all the CED models are trained and evaluated properly on the single source audio data.

C. Performance analysis of SS+CEDNet

Now, we will discuss the performance of our proposed pipeline, SS+CEDNet. A total number of 400 soundscapes each containing both speech and cough events were used for the evaluation. First of all, the soundscapes were passed through the SS model for source separation. Then the separated cough source was segmented into 1s clips. These 1s clips were finally used for cough detection by the CED models.

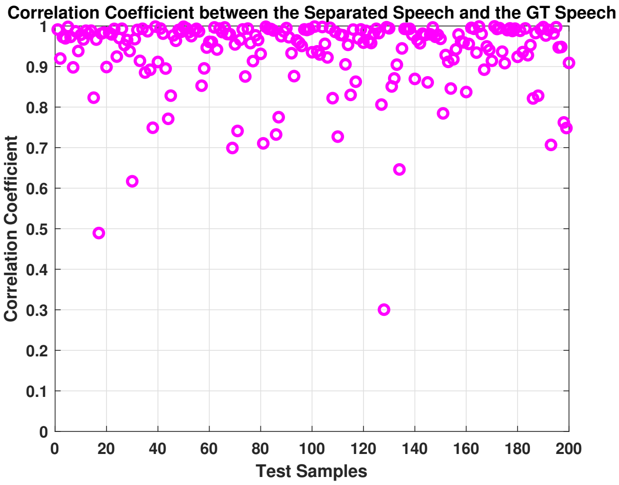

Fig. 5. Scatter Plot of Correlation Coefficients between the Separated Speech and the Ground Truth (GT) Speech sources

The SS model is implemented to preserve speech privacy by separating the speech source from the mixed audio. For a quantitative analysis of how well the model is able to separate the speech source, we have calculated the correlation coefficient between the ground truth speech source and the SS model's separated speech source. An average correlation coefficient of 0.94 (out of 1) is achieved from the test dataset. This is an indication that the SS model is able to separate the speech source quite well and thus it preserves the speech privacy. A scatter plot of the correlation co-efficients for randomly selected 200 samples is shown in Fig. 5. To compare the performance of the proposed pipeline and to check the effectiveness of source separation on cough detection, the soundscapes were segmented into 1s clips without any source separation and were directly passed through the CED models. The result comparison is shown in Table V and for a better quantitative understanding, the relative improvement of the F1 score is also shown in Table VI. The pipeline shows a maximum relative F1 score improvement of 13.8%.

TABLE V
PERFORMANCE ANALYSIS OF SS+CEDNET

Model	Precision (%)	Recall (%)	F1 Score (%)
YAMNet	94	81	87
Wave-U-Net+YAMNet	*98*	*99*	*99*
RF	96	87	92
Wave-U-Net+RF	*98*	*93*	*96*
SVM	92	91	92
Wave-U-Net+SVM	*96*	*93*	*96*
NB	81	90	85
Wave-U-Net+NB	*83*	*96*	*89*

TABLE VI
RELATIVE F1 SCORE IMPROVEMENT OF SS+CEDNET

Model	Relative F1 Improvement (%)
Wave-U-Net+YAMNet	13.8
Wave-U-Net+RF	4.34
Wave-U-Net+SVM	4.34
Wave-U-Net+NB	4.7

It is evident that the proposed pipeline, in other words, source separation prior to cough event detection shows promising performance by helping the CED models to detect cough events more precisely and also to secure speech privacy (by separating speech source).

VI. CONCLUSION

A pipeline called SS+CEDNet, consisting of sound source separation and cough event detection models, is proposed for detecting cough events in a privacy preserving way. The SS+CEDNet is able to preserve the speech privacy by separating speech source, and since speech is separated from the background, the efficiency of the CED models have also increased significantly. The proposed pipeline can be useful to implement speech privacy preserving cough detection or deploy cough count devices in public places like hospitals.

REFERENCES

[1] 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.

[2] S.-y. Takahashi, T. Morimoto, S. Maeda, and N. Tsuruta, "Cough detection in spoken dialogue system for home health care," in *Eighth International Conference on Spoken Language Processing*, 2004.

[3] C. Zhu, L. Tian, X. Li, H. Mo, and Z. Zheng, "Recognition of cough using features improved by sub-band energy transformation," in *2013 6th International Conference on Biomedical Engineering and Informatics*. IEEE, 2013, pp. 251–255.

[4] 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.

[5] 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 *Proceedings of the 13th international conference on Ubiquitous computing*, 2011, pp. 375–384.

[6] S. J. Barry, A. D. Dane, A. H. Morice, and A. D. Walmsley, "The automatic recognition and counting of cough," *Cough*, vol. 2, no. 1, pp. 1–9, 2006.

[7] S. Le and W. Hu, "Cough sound recognition based on hilbert marginal spectrum," in *2013 6th International Congress on Image and Signal Processing (CISP)*, vol. 3. IEEE, 2013, pp. 1346–1350.

[8] J.-M. Liu, M. You, G.-Z. Li, Z. Wang, X. Xu, Z. Qiu, W. Xie, C. An, and S. Chen, "Cough signal recognition with gammatone cepstral coefficients," in *2013 IEEE China Summit and International Conference on Signal and Information Processing*. IEEE, 2013, pp. 160–164.

[9] F. Al Hossain, A. A. Lover, G. A. Corey, N. G. Reich, and T. Rahman, "Flusense: a contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–28, 2020.

[10] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[11] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[12] N. Takahashi, N. Goswami, and Y. Mitsufuji, "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 106–110.

[13] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[15] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. d. C. Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards learning a universal non-semantic representation of speech," *arXiv preprint arXiv:2002.12764*, 2020.

[16] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[17] 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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[18] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," 1993. [Online]. Available: https://hdl.handle.net/11272.1/AB2/SWVENO

[19] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[20] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 344–348.

[21] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.