Detection of Coughing and Respiratory Sensing in Conversational Speech

Venkata Srikanth Nallanthighal[1,2], Aki Harm[1], Ronald Rietman[1] and Helmer Strik[2]
[1]*Philips Research, Eindhoven, The Netherlands*
[2]*Centre for Language Studies (CLS), Radboud University Nijmegen, The Netherlands*

Keywords: Speech Breathing, Signal Processing, Deep Neural Networks, Respiratory Parameters, Speech Technology.

Abstract: Coughing and shortness of breath are typical symptoms in people suffering from COPD, asthma, and COVID-19 conditions. Separate studies have shown that coughing and respiratory health parameters, respectively, can be sensed from a conversational speech recording using deep learning techniques. This paper looks into joint sensing of coughing events and the breathing pattern during natural speech. We introduce an algorithm and demonstrate its performance in realistic recordings. We observed sensitivity of 92.4% and 91.6% for cough detection and breath event detection, respectively.
Clinical Relevance: Joint sensing of coughing events and respiratory parameters gives a more holistic picture of the respiratory health of a patient which can be very useful for future telehealth services.

1 INTRODUCTION

The importance of respiratory sensing and cough monitoring needs little justification, especially in home monitoring of patients suffering from respiratory conditions. The COVID-19 pandemic demonstrated the necessity of remote digital health assessment tools for telehealth services. This is particularly pertinent for elderly and vulnerable populations who already have a chronic disease. Post-covid patients may suffer from respiratory symptoms for several months after a severe form of COVID-19 disease and separate telehealth services are being developed for these patients.

Speech is a good indicator of the pathological condition of a person (Alireza A. Dibazar, et al., 2002), especially for respiratory conditions like COPD, asthma, and COVID-19. These conditions significantly influence the breathing capacity and cause vocal respiratory symptoms such as coughing or wheezing.

We can hear when a person has breathing difficulties or coughing, but the automatic detection is a complex task. The breathing planning is complex process based on linguistic and prosodic factors (Marcin Włodarczak, et al., 2015) and the detection of breathing events from continuous speech is difficult.

We have demonstrated recently that it is possible to use neural networks to estimate the breathing parameters from a speech audio signal (Venkata Srikanth Nallanthighal, et al., 2019; V. S. Nallanthighal, et al., 2020). Neural networks have also been used successfully for the detection of cough sounds from non-speech recordings (A. C. den Brinker, et al., 2021; Justice Amoh and Kofi Odame, 2016; Yusuf A Amrulloh, et al., 2015; Hui-Hui Wang, et al., 2015). However, the current authors are not aware of an earlier studies on the simultaneous detection of coughs and respiratory behavior in conversational speech recordings. Obviously, speaking, breathing, and coughing are all functions of the lungs, and tightly interlinked, which makes the simultaneous sensing a challenging problem. One of the key questions is related to the causal relations and independence of different sensing targets. The results of this paper seem to indicate that holistic sensing of different aspects of respiratory health from a free speech data is possible using modern machine learning techniques.

2 BACKGROUND

The current research is related to the development of acoustic sensing technology for telehealth call services especially for patients with respiratory

Nallanthighal, V., Harm, A., Rietman, R. and Strik, H.
Detection of Coughing and Respiratory Sensing in Conversational Speech.
DOI: 10.5220/0010709700003123
In *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2022) - Volume 5: HEALTHINF*, pages 289-295
ISBN: 978-989-758-552-4; ISSN: 2184-4305

symptoms. In this scenario a nurse, or an automated agent, has therapeutic conversations with the patient, collects information, and answers questions about the care or the symptoms. Breathing monitoring from telehealth customers' speech conversations over multiple calls would give us the historical data of breathing parameters and help us compare and understand a person's pathological condition, decline, or improvement over time and early detection of a condition.

The breathing, speech, and coughing, are all functions closely related to the lungs, and influenced by the condition of the respiratory system. Based on physiological considerations, cough sounds are often considered as consisting of four different phases : inspiratory, contractive, compressive, and expulsive. The inspiratory phase is initiated by breathing in and is terminated by the closure of the glottis. In the contractive phase, groups of respiratory muscles contract, leading to a marked elevation of alveolar, pleural, and subglottic airway pressures. In the expulsive phase, the glottis opens quickly followed by rapid exhalation of air under a large pressure gradient.The rapid movement of air expelled from the lung generates the cough sounds with contributions coming from different areas of the respiratory system. The mechanism of cough sound production shares some similarities to that of speech production. The current authors are not aware of a previous work where the goal is simultaneous sensing of respiratory parameters such as respiratory rate and coughing activity and intensity from a conversational recording.

In this paper we study joint detection of cough events and sensing of respiratory health parameters, such as breathing rate and tidal volume, from a conversational speech recording. We give an overview of the algorithms for cough detection and respiratory signal estimation and compare those to the corresponding state-of-the-art systems. Next, we evaluate their performance in a speech data where the goal is a joint sensing of the respiratory parameters, and detection of coughing during speech. The first results presented in this paper show that the two models can work relatively independently. In the discussion, we propose that joint sensing of breathing, speaking, and coughing can be potentially modeled using a holistic cardio-pulmonary speech model that also takes the linguistic context into account.

Chronic obstructive pulmonary disease (COPD) is a progressive respiratory disease characterized by chronic inflammation of the lung airways which results in airflow limitation. This results in frequent shortness of breath (SOB) and coughing. SOB can be detected from one's speech (Sander Boelders, et al., 2020). Coughing is a prominent indicator of several problems such as COPD, and it is also the main reason for why patients seek medical advice. Frequent COPD exacerbations are associated with a high mortality and heavy use of healthcare resources. COPD patients with chronic cough and shortness of breath may represent a target population for whom specific respiratory sensing and cough monitoring strategies should be developed.

3 METHODS

3.1 Respiratory Sensing from Speech

Both the rib cage and the abdomen can be used to modulate alveolar pressure and airflow during speech. Some speakers exhibit the stronger use of rib cage over abdominal contributions, and some speakers show a relatively equal contribution from both the rib cage and abdomen (Thomas J. Hixon, et al., 1976). When a known air volume is inhaled and measured with a spirometer, a volume-motion relationship can be established as the sum of the abdominal and rib cage displacements (K. Konno and J. Mead, 1967).

The Philips read speech database was collected at Philips Research, Eindhoven, The Netherlands in 2019, with the approval of the Internal Committee Biomedical Experiments (ICBE) of Philips Research. The data was collected using the following setup: two respiratory elastic transducer belts over the ribcage under the armpits and around the abdomen at the umbilicus level to measure the changes in the cross-sectional area of ribcage and abdomen at the sample rate of 2kHz. These belts work on the principle of respiratory inductance plethysmography (RIP). Earthworks microphone M23 is used for recording high-quality speech at 48kHz. The microphone is placed at a distance of one meter from the speaker, and the data collection is conducted in a specialized audio room for noise-free and echo-free recordings. 40 healthy subjects with no respiratory conditions(18 female and 22 male with age group ranging from 21 to 40 years old) are asked to read "The Rainbow Paragraph", a widely used phonetically balanced paragraph (G. Fairbanks, 1960).

In our data collection two respiratory belts were placed around the rib cage under the armpits and around the abdomen at the level of the umbilicus, respectively. These belts work on the principle of respiratory inductance plethysmography (RIP). They consist of a sinusoidal wire coil insulated in elastic.

Dynamic stretching of the belts creates waveforms due to change in self-inductance and oscillatory frequency of the electronic signal and the electronics convert this change in frequency to a digital respiration waveform where the amplitude of the waveform is proportional to the inspired breath volume. Thus the sum of rib cage and abdomen expansions measured by the respiratory belt transducers is considered as the measure for the breathing signal.

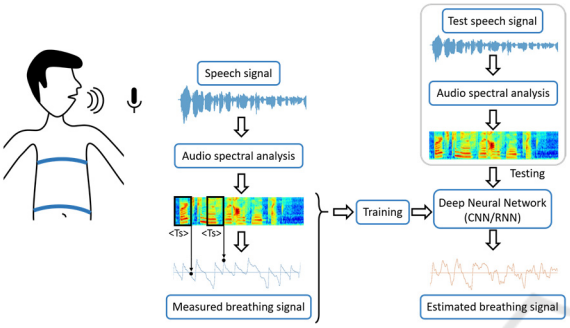

Figure 1: Schematic diagram for estimating respiratory signal using Deep Neural Network Model based on spectral features.

Estimating breathing signals from the speech signal is a regression problem. The breathing signal is a quasi-periodic signal whose characteristics are dependent on the prosodic and linguistic content of the speech signal. This information of speech can be modeled using spectral features. Spectral features are based on a time-frequency decomposition of the speech signal. In this paper we use a linear spectrogram computed using short-time Fourier transform, and a nonlinear spectrogram with logarithmic magnitude values on a Melfrequency scale, i.e., log Mel spectrogram.

The spectrogram and log Mel spectrogram of a speech signal of a fixed time window are mapped with respiratory sensor value at the endpoint of the time window with a stride of 10ms between windows for Philips database to train the neural network models as shown in Figure 1. These models will estimate the respiratory sensor values of a speech signal in real time to get the breathing pattern.

Using spectral features as an input representation of speech signal, we implement convolutional neural network (CNN) and Long short-term memory recurrent neural network (RNNLSTM) models using the PyTorch software framework (Adam Paszke, 2019). In the CNN model (Jürgen Schmidhuber, 2015), the data is fed into a network of two convolutional layers with a single-channel and kernel

CNN Model Input: log Mel spectrogram or spectrogram m: frames in time window n : Mel filter banks	RNN Model Input: log Mel spectrogram or spectrogram m: frames in time window n : Mel filter banks
Matrix X_i(1 x m x n)	Matrix X_i(1 x m x n)
1 x conv3-1;s1 Maxpooling 3x3	LSTM model
1x conv5-1;s1 Maxpooling 3x3	Layers =2
3 Fully Connected layers	Hidden size= 128
OUTPUT: sensor value	OUTPUT: sensor value

Figure 2: Deep neural network configurations of the spectral based methods for sensor value prediction.

size of 3 and 5 respectively for filtering operation to extract local feature maps. Max pooling is deployed to reduce the dimensionality of feature maps while retaining the vital information. The rectified linear unit activation function is applied to introduce non-linearity into the feature extraction process for each convolutional layer, as shown in Figure 2. Batch normalisation is also applied on each convolution layer. This is followed by 3 fully connected layers with ReLU activation function. The Adam optimiser (Diederik P. Kingma and Jimmy Ba, 2015) with a weight decay of 0.001 is used as an optimization algorithm.

In the RNN-LSTM model, the data is fed into a network of two LSTM layers with 128 hidden units and a learning rate of 0.001. The Adam optimizer is used as an optimization algorithm to update network weights iteratively based on training data (Diederik P. Kingma and Jimmy Ba, 2015). These hyperparameters for the network were chosen for estimation after repeated experimentation.

As estimating breathing pattern from speech using neural networks is a regression problem, we use the following two metrics for evaluation and comparison: correlation and mean squared error(MSE) of estimated breathing signal and the actual respiratory sensor signal. Also, we compare the breathing parameters derived from the estimated and actual breathing signals. The model that estimates breathing signals with a higher correlation, lower MSE, and comparable breathing parameters would be considered best for our study.

Table 1: Philips Database (read speech protocol): The *r*, MSE and breathing parameters for systems using spectral based approach.

Models	Loss Function	*r*	MSE	Breathing Parameters				
				Breathing Rate			Breath Event Sensitivity	Tidal Volume error (%)
				prediction (breaths/min)	true (breaths/min)	error (%)		
Spectral Based Approach								
I/P: log Mel spec O/P: sensor Architecture: RNN	MSE BerHu	0.476 0.482	0.019 0.039	10.42 10.98	9.84 9.84	5.89% 11.58%	0.916 0.902	12.11% 16.24%
I/P: log Mel spec O/P: sensor Architecture: CNN	MSE BerHu	0.472 0.462	0.034 0.042	10.85 11.78	9.84 9.84	10.26% 19.71%	0.896 0.821	16.22% 18.84%

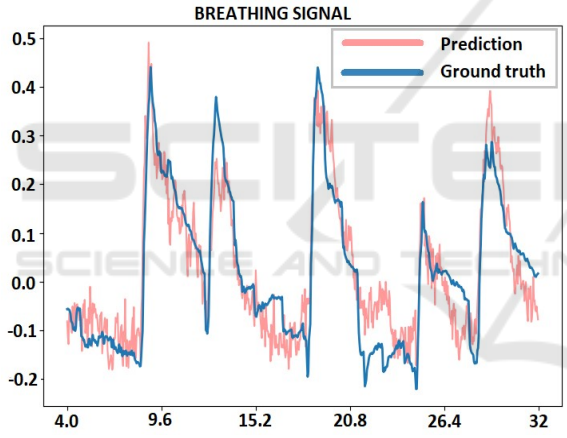

Figure 3: The predicted and ground truth for breathing signal Table I presents the performance of the systems trained and tested on the Philips database.

3.2 Cough Monitoring

Cough is an important clinical symptom in respiratory diseases, yet there is not any gold standard to assess it. The most common approach to addressing these challenges is to find features of the acoustic signal that offer good discrimination between coughs and non-cough sounds. Barry et al,. used Linear Predictive Coding and Mel-frequency Cepstral Coefficients (MFCC) to model the sound of coughs and used a Probabilistic Neural Network to classify time windows as containing a cough or not (Samantha J. Barry, et al., 2006). Other researchers have explored other features based on adaptations of speech recognition features (Thomas Drugman, et al. 2011) and custom-designed handcrafted features (Eric C. Larson, et al., 2011). In this paper, we explore advanced machine learning algorithms like XGBoost models and deep learning algorithms for cough detection.

The data of cough sounds is collected from a recent trial of COPD patients in their home environment. The data consist of one second audio snippets of night-time recordings in the vicinity of COPD patients, where the monitoring period ran over a period of 90 days. Part of this data has been annotated and is used in our study.

1) XGBoost Models:
In classical machine learning, a model is trained on features that are extracted from the signal. For each sound snippet, 12 MFCCs are extracted around the peak of the transient. Additionally, the sound levels just before and after the peak are extracted, as well as the number of acoustic transients in the 60 second interval centered on the time of the recorded snippet. An XGBoost (Tianqi Chen and Carlos Guestrin, 2016) model was trained on the features of the annotated snippets.

2) Deep Learning Models:
Convolutional and recurrent neural networks are used for cough detection. In the CNN, the input is a fixedsized image, a segment of the logMel spectrogram, and the output is a single

label. The RNN takes in a sequence of spectral frames and outputs a sequence of labels.

The CNN architecture was inspired by the popular LeNet-5 architecture (Yann LeCun, et al., 2015) which yielded state-of-the-art performance on the MNIST handwritten digits dataset. Compared to other well-known architectures such as the AlexNet, LeNet-5 is a much smaller network and more suitable for smaller datasets.

The RNN model consists of 6 layer encoder-decoder architecture, which allows the network to process and classify input sequences of arbitrary length. The encoder is made up of 3 layers: 2 bidirectional recurrent layers with 128 and 64 units respectively and a unidirectional layer with 32 recurrent units. Our encoder is set up to process sequences of up to a certain maximum length we set depending on the experiment (see experiment section below). All recurrent neurons in the encoder are Gated Recurrent Units (GRU), which can identify longterm dependencies in a sequence of input data. The last layer of the encoder outputs a fixed representation (the 32 activations) which is then used to initialize the decoder. The decoder is a single recurrent layer of 64 Long Short Term Memory (LSTM) units, combined with an attention mechanism. The attention mechanism enables the network to focus on salient parts of the input features and ultimately results in improved classification performance. Currently, our decoder is set up to output a single label for each input sequence. Following the decoder, we have a fully connected layer with 256 ReLU neurons. Finally, the classification layer outputs a class label using the softmax function.

Table 2: Comparison of cough detection models.

Models	AUC scores	standard deviation
XGBoost Model	0.9040	0.0740
CNN (LeNet-5)	0.9116	0.0576
RNN(seq2seq)	0.9141	0.0502

4 EXPERIMENTS ON JOINT SENSING

The goal of joint sensing of respiration and cough events is to verify if the models developed independently for respiratory sensing and cough detection can work together. The hypothesis to support this is that the two models are trained on two different datasets specific for respiratory sensing and cough detection, respectively.

4.1 Joint Data Set

We collected a database of speech recordings with coughs of 10 healthy volunteers(6 Male and 4 Female) who were asked to read ”The Rainbow Paragraph”, a widely used phonetically balanced paragraph (G. Fairbanks, 1960). The participants were asked to voluntarily cough at random places while reading the paragraph. Each recording is about 2 minutes long with at least five coughs. Cough events and breathing events were annotated manually by the authors for each of the ten subjects. We found and manually annotated a total of 54 cough events and 103 breathing events for the ten recordings. This database is used for testing our models on cough monitoring and respiratory detection simultaneously.

4.2 Results

1) Performance of Respiratory Sensing: The pre-trained deep learning model using RNN architecture is used to compute the estimated respiratory signal for each of the ten recordings. From these estimated respiratory signals, respiratory breathing events are identified using an Automatic Multiscale Peak Detection Algorithm (AMPD) (Felix Scholkmann, et al.,2012). These breath events computed from the estimated respiratory signal are compared with the manually annotated ground truth breath events, and the results are formulated in Table III

2) Accuracy in Cough Detection: The RNN model described in the previous section is used for cough detection for each of the ten recordings. A total of 54 cough events are present in these ten recordings. Actual cough events are compared against the detected cough events to measure precision and sensitivity and are reported in Table III.

Table 3: Joint detection of cough events and respiratory events.

Modality	Total events	Precision	Sensitivity
Cough detection	54 events	94,2%	92.4%
Breath event detection	103 events	89.6%	91.6%

It is observed that joint sensing of cough and respiratory events is plausible by using pre-trained

models for individual tasks. We observed a precision of 94.2% and sensitivity of 92.4% for the 54 cough events and a precision of 89.6% and sensitivity of 91.6% for breath event detection.

5 CONCLUSIONS

An extensive study on methods for using speech as a modality for respiratory sensing and cough monitoring is presented in this paper. These strategies are essential for patients suffering from respiratory conditions, especially in remote monitoring services. Our results evaluated on joint datasets of 10 healthy volunteers conclude that joint sensing of coughs and respiratory parameters is possible by training deep learning models on separate datasets specific to respiratory sensing and cough detection. However, evaluation of this strategy on speech recordings of patients suffering from respiratory conditions is essential and is the future scope of our work.

ACKNOWLEDGMENT

This work was partially supported by the Horizon H2020 Marie Skłodowska-Curie Actions Initial Training Network

European Training Network project under grant agreement No. 766287 (TAPAS) and Data Science Department, Philips Research, Eindhoven.

REFERENCES

Alireza A. Dibazar, S. Narayanan, and Theodore W. Berger, “Feature analysis for automatic detection of pathological speech,” in *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society, Engineering in Medicine and Biology*. IEEE, 2002, vol. 1, pp. 182–183.

Marcin Włodarczak, Mattias Heldner, and Jens Edlund, “Breathing in conversation: An unwritten history,” in *Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication:* 2015, number 110 in Linkoping Electronic Conference Proceedings,¨ pp. 107–112.

Venkata Srikanth Nallanthighal, Aki Harm¨a, and Helmer Strik, “Deep¨ Sensing of Breathing Signal During Conversational Speech,” in *Proc. Interspeech 2019*, 2019, pp. 4110–4114.

V. S. Nallanthighal, A. Härmä, and H. Strik, “Speech breathing estimation using deep learning methods,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1140–1144.

A. C. den Brinker, M. Coman, O. Ouweltjes, M. G. Crooks, S. Thackray-Nocera, and A. H. Morice, “Performance requirements for cough classifiers in real-world applications,” in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 96–100.

Justice Amoh and Kofi Odame, “Deep neural networks for identifying cough sounds,” *IEEE transactions on biomedical circuits and systems*, vol. 10, no. 5, pp. 1003–1011, 2016.

Yusuf A. Amrulloh, Udantha R. Abeyratne, Vinayak Swarnkar, Rina Triasih, and Amalia Setyati, “Automatic cough segmentation from non-contact sound recordings in pediatric wards,” *Biomedical Signal Processing and Control*, vol. 21, pp. 126–136, 2015.

Hui-Hui Wang, Jia-Ming Liu, Mingyu You, and Guo-Zheng Li, “Audio signals encoding for cough classification using convolutional neural networks: A comparative study,” in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2015, pp. 442–445.

Sander Boelders, Venkata Srikanth Nallanthighal, Vlado Menkovski, and Aki Harm¨ a,¨ “Detection of mild dyspnea from pairs of speech recordings,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4102–4106.

Thomas J. Hixon, Jere Mead, and Michael D. Goldman, “Dynamics of the chest wall during speech production: Function of the thorax, rib cage, diaphragm, and abdomen,” *Journal of speech and hearing research*, vol. 19, no. 2, pp. 297–356, 1976.

K. Konno and J. Mead, “Measurement of the separate volume changes of rib cage and abdomen during breathing,” *Journal of Applied Physiology*, vol. 22, no. 3, pp. 407–422, 1967, PMID: 4225383.

G. Fairbanks, “The rainbow passage,” *Voice and articulation drillbook*, vol. 2, 1960.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia

Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche-Buc, E. Fox, and R. Garnett,´ Eds., pp. 8024–8035. Curran Associates, Inc., 2019.

Jurgen Schmidhuber, “Deep learning in neural networks: An overview,”¨ *Neural Networks*, vol. 61, pp. 85 – 117, 2015.

Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *Computing Research Repository (CoRR)*, vol. abs/1412.6980, 2015.

Samantha J. Barry, Adrie D. Dane, Alyn H Morice, and Anthony D Walmsley, "The automatic recognition and counting of cough," *Cough*, vol. 2, no. 1, pp. 1–9, 2006.

Thomas Drugman, Jerome Urbain, and Thierry Dutoit, "Assessment of audio features for automatic cough detection," in *2011 19th European Signal Processing Conference*. IEEE, 2011, pp. 1289–1293.

Eric C. Larson, TienJui Lee, Sean Liu, Margaret Rosenfeld, and Shwetak 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.

Tianqi Chen and Carlos Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016, KDD '16, pp. 785–794, ACM.

Yann LeCun et al., "Lenet-5, convolutional neural networks," *URL: http://yann. lecun. com/exdb/lenet*, vol. 20, no. 5, pp. 14, 2015.

Felix Scholkmann, Jens Boss, and Martin Wolf, "An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals," *Algorithms*, vol. 5, no. 4, pp. 588–603, 2012.