IDIAP RESEARCH REPORT

ESTIMATING BREATHING PATTERN FROM RAW SPEECH WAVEFORM AND SHORT-TERM SPEECH SPECTRUM USING NEURAL NETWORKS

Zohreh Mostaani[a] Venkata Srikanth Nallanthighal
Aki Härmä Helmer Strik Mathew Magimai.-Doss

Idiap-RR-12-2024

OCTOBER 2024

[a]Idiap Research institute

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Estimating Breathing Pattern from Raw Speech Waveform and Short-term Speech Spectrum using Neural Networks

Zohreh Mostaani[1,2]*, Venkata Srikanth Nallanthighal*[3,4]*, Aki Härmä* [3]*, Helmer Strik*[4]*, Mathew Magimai Doss*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2] Ecole polytechnique fédérale de Lausanne, Lausanne, Switzerland
[3]Philips Research, Eindhoven, The Netherlands
[4]Centre for Language Studies (CLS), Radboud University Nijmegen, The Netherlands
{zohreh.mostaani, mathew}@idiap.ch, {srikanth.nallanthighal, aki.harma}@philips.com, w.strik@let.ru.nl

Abstract

Respiration process is integral part of speech production. Alternation in respiratory system and speech production system results in changes in speech. Therefore, speech signal, which can be acquired in a non-invasive manner, could be used to predict breathing patterns. There is a growing interest in that direction, which has gained further momentum with COVID-19 situation. In this paper, we investigate respiratory signal estimation through (a) raw waveform modeling and (b) modeling of short-term spectral features using deep learning techniques. Our investigations on ComParE 2020 Breathing sub-challenge showed that both the approaches perform well and yield systems competitive to the best performing CNN+RNN-LSTM baseline system. An analysis of the two investigated approaches revealed that raw waveform modeling-based approach yields better *Pearson's correlation coefficient*, but it is not able to predict well the dynamic range of the signals, when compared to spectral feature-based approach.

Index Terms: breathing patterns, convolutional neural networks, LSTM-RNN, speech technology

1. Introduction

The respiratory system, including diaphragm, chest cavity, and lungs, plays a very important part in producing speech. It provides the energy necessary to produce sounds by pushing air through vocal folds. It is therefore not surprising that the speech and breathing are related.

Very few studies focused on the effect of speech on breathing pattern. Hammarsten et al, investigated the inhalation duration and speech onset delay in different settings and reported that both of them are longer when speakers start to speak compared to when they are in the middle of a conversation [1]. In other works, the breathing pattern for read speech has been compared to spontaneous speech. They reported that a high percentage of the sentences in read speech is produced during one breath while the inhalations were short and frequent during spontaneous speech [2, 3, 4]. The latter could be due to the cognitive load during spontaneous speech [5]. Wlodarczak et al, proposed that the relationship between speech and breathing is not one way and breathing can also shape the speech [6].

Breathing patterns can reveal information about the underlying health condition of a speaker and therefore being able to estimate it from speech signal can have a wide range of applications. Some conditions such as heart diseases and Chronic obstructive Pulmonary Diseases (COPD) can affect the breathing pattern. Diagnosing such health conditions from speech can be very helpful in tele-health applications and as a non intrusive diagnosis method.

To the best of our knowledge there has not been many studies on breathing pattern estimation from speech. Nallanthighal et al used Log Mel Spectogram of speech to train a Convolutional Neural Network (CNN) and a Long short-term memory Recurrent Neural Network (RNN-LSTM) to predict breathing signal [7, 8]. They later used their method to detect mild dyspnea in speakers [9]. Schuller et al, used an End-to-End system consisting of CNN combined with RNN-LSTM to predict the breathing signal, as part of the Interspeech 2020 ComParE challenge [10].

In this paper, as part of the Interspeech 2020 ComParE Breathing sub-challenge, we investigate raw waveform modeling and short-term spectrum modeling methods for estimating the breathing signal from speech waveform, in the framework of deep learning. We study different loss functions to train the neural networks, namely, mean square error and Pearson's correlation-based. Experimental studies show that our methods yield systems competitive to the baseline CNN-RNN-LSTM approach.

The remainder of the paper is organized as follows. In Section 2, we introduce our methods. In Section 3, we present the experiment setup and results. In Section 4, we present an analysis of our systems. Finally we conclude in Section 5.

2. Methods investigated

This section presents the different neural networks-based methods modeling raw waveform and short-term spectral features for breathing pattern estimation.

2.1. CNN-based raw speech waveform

We adopted the convolutional neural network (CNN) based approach that was first proposed for speech recognition [11]. As illustrated in Fig. 1, the network architecture consists of a number of convolution layers followed by a hidden layer, and then finally an output layer. Our motivation behind adopting this approach was two folds. First, this approach has been studied on various speech processing tasks such as, speaker recognition [12], gender recognition [13], and depression detection [14]. Second, we have some good insight about how the source and system related information in the speech signal tends to get modeled in this approach [15, 12, 16].

Based on these prior insights, we borrowed a CNN architec-

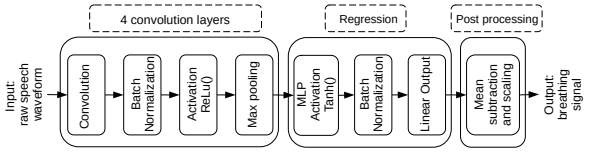


Figure 1: *An illustration of the end-to-end CNN model used with raw speech signal as input.*

ture that was used in a recent work for depression detection [14]. More precisely, the network has four convolution layers, max pooling layers and a fully connected layer (MLP) (see Fig. 1. The number of filters in convolution layers are 128-256-512-512 with kernel sizes of 30-10-4-3 and kernel strides of 10-5-2-1. After each layer, there are max-pooling layers with strides of 2-3-1-1 and rectified linear unit (ReLu) as activation function. The MLP has one hidden layer with 10 units with hyperbolic tangent (Tanh) as activation. Batch normalization is also applied after each layer. The output layer consists of one unit with linear activation. The input to the system is raw speech waveform and the output is a sample-by-sample prediction of the breathing signal. During test time, the output of the neural network is post-processed by removing the mean and scaling between -1 and +1.

We trained the CNN with standard **Mean Squared Error (MSE)** loss and a custom **correlation** loss. We used Adam optimizer [17] with weight decay of 0.001.

The custom correlation loss is defined as following:

$$L(y, f(x)) = \frac{1}{1 + r(y, f(x))} \quad (1)$$

where $r(y, f(x))$ is the *Pearson's correlation coefficient*.

During training, the correlation loss is computed by predicting the output for a fixed window of consecutive time points and then calculating the *Pearson's correlation coefficient*, $r()$, between the output signal and the target signal. The size of window is a hyper-parameter.

2.2. RNN and CNN trained on Short-term speech spectral features

In this approach, log Mel spectrograms are used as spectral features for speech [18]. Log Mel spectrogram of a speech signal with a fixed time window of 4 seconds is mapped with respiratory sensor value at the endpoint of the time window with a stride of 10ms between windows to train the Convolutional Neural Networks (CNN) and Long short-term memory Recurrent Neural Network (RNN-LSTM) models. These models will be used to estimate the respiratory sensor values of a speech signal.

In the CNN model [19], the data is fed into a network of two convolution layers with single channel and kernel size of 5 for filtering operation to extract local feature maps. Max pooling is deployed to reduce the dimensionality of feature maps while retaining the important information and ReLu activation function is applied to introduce non-linearity into the feature extraction process for each convolution layer. Batch normalisation is also applied on each convolution layer. This is followed by 3 fully connected layers with ReLU activation function. Adam optimiser [17] with a weight decay of 0.001 is used as an optimization algorithm.

In the RNN-LSTM model [20], the data is fed into a network of two LSTM layers with 128 hidden units and a learning rate of 0.001. Adam optimiser is used as an optimization algorithm to update the network weights iteratively based on training data [17]. **Mean Squared Error (MSE)** loss and a custom **correlation** loss as described in 2.1 are used as regression loss functions for both networks. These hyper-parameters for the network are best chosen for estimation after repeated experimentation.

3. Experiments

3.1. Dataset and evaluation protocol

The systems have been trained on a subset of the UCL Speech Breath Monitoring (UCL-SBM) database. This database has been introduced in [10]. It includes recordings from 49 speakers which are divided into three non-overlapping subsets; 17 speakers in *Train*, 16 speakers in *Dev*, and 16 speakers in *Test* subset. For each speaker a 4 minutes recording of speech with sampling frequency of 16KHz is provided. For speakers in *Train* and *Dev* set, the breathing signal with sampling frequency of 25 Hz is provided which amounts to a sequence of 6000 values for each speaker.

As done in the challenge paper [10], for the development studies, we used the *Train* set for training our system, the first 15 speakers for training and the last 2 speakers for cross validation. We then tested our system on the whole *Dev* set. For testing, the training and development data are pooled together to train the neural networks.

The performances of the systems are measured by concatenating the predicted breathing signals and the ground truth signals for all files to form one predicted signal and one ground truth signal, respectively, and computing *Pearson's correlation coefficient* r.

For the sake of clarity, our systems are denoted in the following format: *ANN type-input type-loss function*.

3.2. Raw waveform modeling based systems

In the development studies, we fine tuned two hyper-parameters:

1. duration of past speech signal input to the system at each time frame. For this, we varied the input speech duration from 2 seconds to 4 seconds in steps of one second.
2. correlation window size for computing the correlation loss. We considered window sizes of 400, 512 and 1024 samples. These window sizes were chosen taking into consideration that at least two breathing cycles at the output are covered.

The frame rate at the input was the sample rate at the output i.e., 40 ms.

Table 1 shows the result for the different systems trained. It can be seen that for both loss functions, the system performance for the input of 2 seconds is better or comparable to the systems with 3 seconds and 4 seconds input. There is no clear trend in performance w.r.t the size of correlation window. Longer correlation window size seems to be beneficial when modeling 4 seconds input.

3.3. Spectral feature based systems

The log Mel spectrogram, time window of 4 seconds and endpoint mapping are decided based on repeated experimentation as described in our previous work on estimating breathing signal from speech [7]. Using this setup, we train both CNN and

Table 1: *Pearson's correlation coefficient r obtained for raw waveform modeling systems on the Dev set.*

Models	Correlation window	Input window 2s	3s	4s
CNN-Raw-MSE	—	0.519	0.480	0.458
CNN-Raw-Correlation	400	0.496	0.492	0.473
	512	0.512	0.475	0.502
	1024	0.501	0.497	0.514

RNN-LSTM models as described in 2.2. For correlation loss, we used correlation window size of 1024 samples.

Table 2 presents the performance of spectral feature-based systems. It can be observed that CNN-based system performs better. Performance with correlation loss is inferior to MSE loss. When compared to raw waveform based systems, spectral feature-based system yields lower correlation.

Table 2: *Pearson's correlation coefficient r obtained for spectral feature based systems on the Dev set*

Models	r
CNN-Spec-MSE	0.472
CNN-Spec-Correlation	0.431
RNN-LSTM-Spec-MSE	0.441
RNN-LSTM-Spec-Correlation	0.420

3.4. Comparison to ComParE baseline systems

Table 3 compares the top two best performing raw waveform based systems and the spectral-based systems. For brevity, we only present the best performing baseline systems. (2s) refers to 2 seconds long input. (4s, 1024) refers to 4 seconds long input with 1024 correlation window size. Fusion-Raw refers to the system where the output of CNN-Raw-MSE and CNN-Raw-Correlation are aligned through cross correlation and are averaged. On the Dev set, our systems outperform low level descriptor based systems and bag-of-audio-words based systems. Raw waveform-based, CNN-Spec-MSE and Fusion-Raw yield performance competitive to the baseline CNN+LSTM RNN system.

Table 3: *Pearson's correlation coefficient r reported on the Dev set and the Test set for baseline systems and proposed systems.*

	Dev r	Test r
ComParE 2020 Breathing sub-challenge Baselines		
OPENSMILE: COMPARE functionals+SVM	0.244	0.442
OPENXBOW: COMPARE BoAW+SVM	0.226	0.366
End2End: CNN+LSTM RNN	0.507	**0.731**
Proposed Systems		
CNN-Raw-MSE (2s)	0.519	—
CNN-Raw-Correlation (4s, 1024)	0.514	—
CNN-Spec-MSE	0.472	—
RNN-LSTM-Spec-MSE	0.441	—
Fusion-Raw	**0.552**	0.656

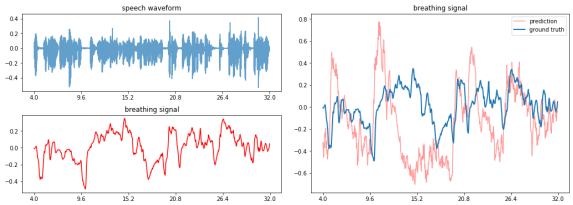
It is worth mentioning that as per the challenge guideline we have reported one test set result. Evaluation on test set is still on-going for other systems. As only five attempts are allowed, the test results for only a few of the systems will be updated after the challenge deadline.

4. Analysis

This sections presents an analysis of our systems.

4.1. A note on evaluation measure

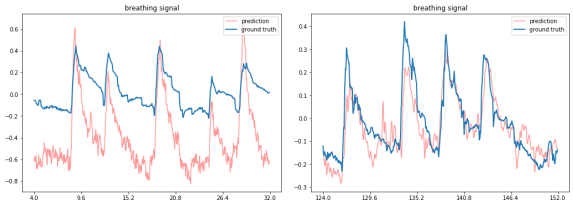
As noted earlier, the systems are evaluated by concatenating all the predicted and ground truth recordings from different files into a single file and comparing them. This may not completely reveal how the systems are performing on individual files. There may be differences across recordings due to reasons like sensor placement and body movements. Fig. 2 illustrates this aspect on recording "devel_00", where CNN-Raw-Correlation (4s, 1024) yields a r of -0.0013.



(a) (b)

Figure 2: *(a) The speech waveform and breathing signal and (b) the predicted and ground truth for breathing signal for recording "devel_00" where CNN-Raw-Correlation (4s, 1024) yields Pearson's correlation coefficient of -0.0013. It can be seen that the original breathing signal is very noisy which can contribute to lower performance.*

Furthermore, higher *Pearson's correlation coefficient* does not mean that all parts of the signal are predicted well. For instance, as shown in Fig. 3, although CNN-Raw-MSE yields better r than CNN-Spec-MSE, CNN-Spec-MSE is predicting well the different regions in the signal. It seems that CNN-Raw-MSE is predicting the peaks and cycles well.



(a) (b)

Figure 3: *The breathing signal and predicted breathing pattern for a file for (a) CNN-Raw-MSE model and (b) CNN-Spec-MSE model. The Pearson's correlation coefficient of the file for the first model is 0.807 and for the second model is 0.761.*

For better understanding, we compared the raw waveform modeling-based system and short term spectral feature-based system by computing,

1. r-average: Compute *Pearson's correlation coefficient* per recording and average it.

2. Compute MSE between the predicted signal and the ground truth signal.

Table 4 presents the results based on r, r-average and MSE. It can be seen that raw waveform modeling-based system and spectral feature-based system performs comparably in terms of r-average but MSE-wise spectral feature-based approach is performing better. This may be because the raw waveform modeling method is not able to predict well the dynamic range of the signals. To ascertain that, from each one of the files in the *Dev* set, we obtained the minimum and maximum value and used that to scale the CNN output as opposed to scaling between -1 and +1. CNN-Raw-MSE$_{dyn}$ and CNN-Raw-Correlation$_{dyn}$ denotes the system after applying dynamic range scaling based on original recording. It can be observed that MSE error reduces considerably and is comparable to spectral feature-based system. It is worth noting that r-average does not changes but there is considerable improvement in r. These results indicate that assessing the systems jointly in terms r, r-average and MSE could be more meaningful.

Table 4: *The r, r-average, and MSE for our systems with best reported r on the Dev set.*

Best Models	***r***	***r*-average**	**MSE**
CNN-Raw-MSE (2s)	0.519	0.547	0.149
CNN-Raw-Correlation (4s, 1024)	0.514	0.599	0.243
RNN-LSTM-Spec-MSE	0.441	0.549	**0.031**
CNN-Spec-MSE	0.472	0.567	0.033
Fusion-Raw	0.552	**0.603**	0.176
CNN-Raw-MSE$_{dyn}$ (2s)	0.589	0.547	0.039
CNN-Raw-Correlation$_{dyn}$ (4s, 1024)	**0.634**	0.599	0.038

4.2. Role of phase difference

Another factor that can affect the prediction is the phase difference between speech recording and breathing signal recording. Despite synchronization, this may arise simply due to sampling frequency differences. So, we measured performance of our systems by,

1. aligning the predicted signal and the ground truth signal by computing cross correlation between them and shifting one of the signals; and then
2. computing r and r-average.

Table 5 presents the results after aligning the predicted and ground truth signals. When compared to results presented in tables 3 and 4, there is improvement in terms of both r and r-average. This indicates that the networks are not predicting the output signal in-phase with the ground truth signal. The phase difference may matter when relating time precise events in the speech signal and the breathing signal.

5. Discussion and conclusions

This paper focused on estimating breathing pattern i.e. respiratory signal from speech signal using neural networks. Toward that, on ComParE 2020 Breathing sub-challenge, we investigated raw waveform modeling-based approach and spectral feature modeling-based approach. Our approaches, similar to the baseline CNN+RNN-LSTM approach, outperform low level descriptors and bag-of-audio words based approach. On the development set, our systems are comparable or better than CNN+RNN-LSTM. An analysis of raw waveform modeling-based approach and spectral feature modeling-based approach in terms of performance metric r-average and MSE revealed that both approaches yield systems that are not far in terms of *Pearson's correlation coefficient*, but spectral feature-based approach better models dynamic range of the respiratory signal. Our analysis also revealed that the network predictions are not necessarily in-phase with the ground truth respiratory signal.

Table 5: *The r and r-average calculated on the Dev set for our systems with output phase alignments*

Best Models	***r***	***r*-average**
CNN-Raw-MSE (2s)	0.565	0.614
CNN-Raw-Correlation (4s, 1024)	0.545	0.643
RNN-LSTM-Spec-MSE	0.491	0.621
CNN-Spec-MSE	0.524	0.628
Fusion-Raw	**0.589**	**0.656**

Our future work will focus along the following directions:

1. improving raw waveform modeling approach in terms of better prediction of dynamic range of the respiratory signal. One possible direction is to consider different loss functions such as, Berhu loss [21], which has been studied with spectral feature-based approach [8]. Our analysis indicates that raw waveform-based systems are predicting the peaks and cycles well, as a consequence yielding somewhat better Pearson's correlation, when compared to spectral-based approach. There may be benefit in fusing the outputs of these two approaches. We will investigate both directions.
2. analysing the raw waveform CNNs to understand what kind of spectral information is being learned for respiratory signal prediction. This can be done by analyzing the first convolution layer [15, 12] and through gradient propagation technique [16]. This would let us better understand the differences w.r.t spectral feature-based approach. Furthermore, it could provide insight to adapt the CNN architecture, which in the present work has been borrowed from speech processing task.

6. Acknowledgements

This work was partially supported by the Swiss National Science Foundation (SNSF) through the project Towards Integrated processing of Physiological and Speech signals (TIPS) grant number 200021_188754, and 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. The authors would like to thank S. Pavankumar Dubagunta for his help in setting up raw waveform CNN system and for the fruitful discussions.

7. References

[1] J. Hammarsten, R. Harris, N. Henriksson, I. Pano, M. Heldner, and M. Włodarczak, "Temporal aspects of breathing and turn-taking in swedish multiparty conversations," in *Fonetik 2015,*

Lund, Sweden, June 8-10 2015. Centre for Languages and Literature, 2015, pp. 47–50.

[2] Y.-T. Wang, J. R. Green, I. S. Nip, R. D. Kent, and J. F. Kent, "Breath group analysis for reading and spontaneous speech in healthy adults," *Folia Phoniatrica et Logopaedica*, vol. 62, no. 6, pp. 297–302, 2010.

[3] A. Henderson, F. Goldman-Eisler, and A. Skarbek, "Temporal patterns of cognitive activity and breath control in speech," *Language and Speech*, vol. 8, no. 4, pp. 236–242, 1965.

[4] A. L. Winkworth, P. J. Davis, E. Ellis, and R. D. Adams, "Variability and consistency in speech breathing during reading: Lung volumes, speech intensity, and linguistic factors," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 3, pp. 535–556, 1994.

[5] H. L. Mitchell, J. D. Hoit, and P. J. Watson, "Cognitive-linguistic demands and speech breathing," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 1, pp. 93–104, 1996.

[6] M. Włodarczak and M. Heldner, "Respiratory constraints in verbal and non-verbal communication," *Frontiers in psychology*, vol. 8, p. 708, 2017.

[7] V. S. Nallanthighal, A. Härmä, and H. Strik, "Deep sensing of breathing signal during conversational speech," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association, Interspeech*. Graz, Austria:[Sn], 2019, pp. 4110–4114.

[8] 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.

[9] S. Boelders, V. S. Nallanthighal, V. Menkovski, and A. Härmä, "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.

[10] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly emotion, Breathing & Masks," in *Proceedings of Interspeech*, Shanghai, China, September 2020, p. 5 pages, to appear.

[11] D. Palaz, R. Collobert, and M. M. Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," *arXiv preprint arXiv:1304.1018*, 2013.

[12] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "Towards directly modeling raw speech signal for speaker verification using cnns," 04 2018, pp. 4884–4888.

[13] S. H. Kabil, H. Muckenhirn, and M. Magimai-Doss, "On learning to identify genders from raw speech signal using cnns." in *Interspeech*, 2018, pp. 287–291.

[14] S. P. Dubagunta, B. Vlasenko, and M. Magimai.-Doss, "Learning voice source related information for depression detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[15] "End-to-end acoustic modeling using convolutional neural networks for hmm-based automatic speech recognition," *Speech Communication*, vol. 108, pp. 15 – 32, 2019.

[16] H. Muckenhirn, V. Abrol, M. M. Doss, and S. Marcel, "Understanding and visualizing raw waveform-based cnns," in *Proceedings of Interspeech*, no. CONF, 2019.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] K. Han, Y. He, D. Bagchi, E. Fosler-Lussier, and D. Wang, "Deep neural network based spectral feature mapping for robust speech recognition," in *INTERSPEECH*, 2015.

[19] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85 – 117, 2015.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] L. Zwald and S. Lambert-Lacroix, "The berhu penalty and the grouped effect," *arXiv preprint arXiv:1207.6868*, 2012.