INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Deep sensing of breathing signal during conversational speech

Venkata Srikanth Nallanthighal[1,2], *Aki Härmä* [1], *Helmer Strik* [2]

[1] Philips Research, Eindhoven, The Netherlands
[2] Centre for Language Studies (CLS), Radboud University Nijmegen

srikanth.nallanthighal@philips.com, aki.harma@philips.com, w.strik@let.ru.nl

Abstract

In this paper, we show the first results on the estimation of breathing signal from conversational speech using deep learning algorithms. Respiratory diseases such as COPD, asthma, and respiratory infections are common in the elderly population and patients in health care monitoring and medical alert services in general. In this work, we compare algorithms for the estimation of a known respiratory target signal, measured by respiratory belt transducers positioned across the rib cage and abdomen, from conversational speech. We demonstrate the estimation of the respiratory signal from speech using convolutional and recurrent neural networks. The estimated breathing pattern gives respiratory rate, breathing capacity and thus might provide indications of the pathological condition of the speaker. Evaluation of our model on our database of breathing signal and speech yielded a sensitivity of 91.2 % for breath event detection and a mean absolute error of 1.01 breaths per minute for breathing rate estimation.

Index Terms: breathing detection, pathological speech, speech technology, deep neural networks, respiratory diseases.

1. Introduction

The use of speech analytics has been gaining attention within the clinical and health-care domains in recent years. This follows to the success of deep learning techniques in various speech technology applications. Breathing activity during speech is an important indicator of a person's respiratory and cognitive health condition. Automatic detection and exact demarcation of breath sounds during speech is critical for developing health care services. Ruinskiy and Lavner proposed an effective breath-event detection algorithm based on template matching and the accurate detection of very short silence intervals (called edges) before and after breathing sounds [1]. However, this is limited to high signal noise ratio(SNR) conditions with pure speech. Breathing pattern during speech can be visualized as a systematic step by step process of exhaustion of air. This expiration can be modeled as a composition of phonemes with varying exhaustion flows for vowel and consonant phonemes [2]. Our hypothesis is that the inverse problem, i.e., modeling the breathing pattern using the composition of conversational speech is achievable using deep learning algorithms. This is the main focus of the current paper.

Speech and respiration are closely related. Speech is produced by organs evolved for the respiratory function of the body [3]. Breathing is a primary mechanism of voice generation maintaining a suitable level of subglottal pressure required for momentary production needs. Breathing is implicated in many aspects of speech production, such as voice quality [4], voice onset time [5] and loudness [6]. Vocalization mostly takes place during exhaling while inhaling is done in pauses between utterances. We perform continuously breathing planning during the speech so that we take in more air for a long continuous utterance [7]. When this process is disturbed due to respiratory or cognitive conditions, it influences our breathing planning and may lead to a break in an utterance because of the need for air. We can hear when a person has breathing difficulties but the automatic detection of this is a complex task for computers because the breathing planning is based on linguistic and prosodic factors [8].

The current research is related to the development of acoustic sensing technology for healthcare call centers. Speech recordings can be used for monitoring a wide range of health parameters ranging from respiratory diseases, heart conditions, neurodegenerative diseases to mental conditions. In this paper, the goal is to predict the breathing activity of a talker from the speech signal. The ultimate goal is to measure breathing signal and breathing parameters during telephone conversations of telehealth customers with call center respondents. Breathing monitoring from the speech conversations of these customers over multiple calls would give us the historical data of breathing parameters and would help us compare and understand person's pathological condition, decline or improvement over a period of time and also early detection of a condition.

In the experiments reported in this paper, we use data from healthy volunteers wearing a respiration belt over the ribcage and abdomen during the speech recordings. We compare different representations of speech content and different deep neural network models and discuss their differences in the task of predicting the respiratory belt signal which is here the breathing signal.

2. Approach

2.1. Breathing signal estimation

Spectral features of speech of a fixed time window are mapped with the respiratory sensor value at the endpoint of the time window. This is based on our hypothesis that the respiratory sensor value (breathing state) at the end of a time window is dependent on the composition of speech in that particular time window. The spectral features of speech and known respiratory sensor values are mapped to train deep neural network models. These trained models are used to estimate the respiratory sensor values from a target speech signal in real time to get the breathing signal as shown in Figure 1. For establishing a trained model, we need a dataset of ground truth breathing signals during speech. We designed the following experiment for creating this dataset.

2.2. Experiment and Dataset

Our speech database has been developed at Philips Research, Eindhoven, with the approval of the Internal Committee Biomedical Experiments, (ICBE). The data was collected using

http://dx.doi.org/10.21437/Interspeech.2019-1796

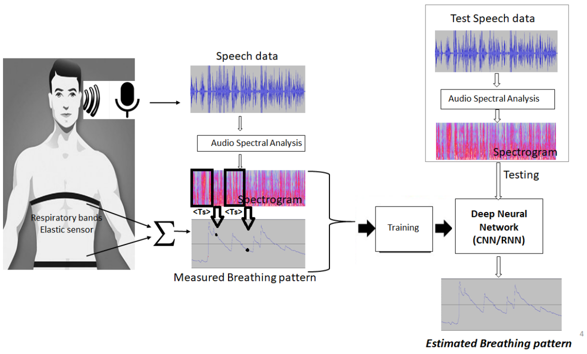

Figure 1: *Schematic diagram for estimating respiratory signal using Deep Neural Network Model.*

the following setup: two respiratory elastic transducer belts over the ribcage and abdomen to measure the changes in the cross-sectional area of ribcage and abdomen at the sample rate of 2 kHz; Earthworks microphone M23 for recording high-quality speech at 48 kHz.

20 healthy subject's data is collected under the following sessions each for approximately 5 minutes:

1. General conversation for spontaneous speech;
2. Reading a phonetically balanced script;
3. Normal breathing for reference breathing rate;
4. Prolonged vowel sound to estimate reference lung capacity; and,
5. Reading the script after exercise to simulate respiratory disease condition.

Respiratory belts are placed around the rib cage under the armpits and around the abdomen at the level of the umbilicus. 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.

The chest can be considered as a system of two compartments with only one degree of freedom each [9]. 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 [9]. Thus the sum of ribcage and abdomen expansions measured by the respiratory belt transducers is considered as the measure for the respiratory or breathing signal as shown in Figure 2.

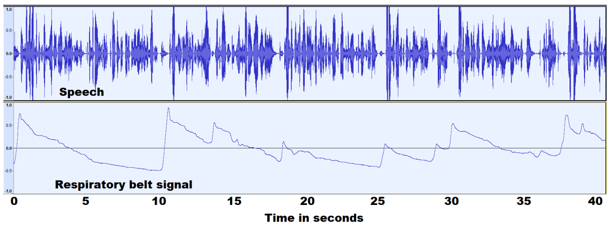

Figure 2: *Respiratory belt signal during speech recording*

3. Method

Spectrogram and log Mel spectrograms are used as spectral features for speech [10]. We investigate the following for the better estimation of breathing signal.

1. Length of time windows: the window length for each speech input, i.e., spectrograms, and log Mel spectrograms is crucial for estimating the breathing sensor value. We found that the average duration of a breath cycle during conversational speech is 7.5s from our database with a minimum of 2s to a maximum of 10s. Hence we investigate speech inputs of fixed window length of 2s, 4s, 8s for better estimation.
2. Spectral features of speech: spectrogram and log Mel spectrograms are considered as input spectral features of speech signals.
3. Mapping point of respiratory sensor: speech signal of a fixed time window is mapped with respiratory sensor value at the endpoint of the time window to train the models. We also investigated mapping with sensor value at the beginning and mid point of the time window and found no significant difference in the estimation.
4. Neural network models: convolutional Neural Network and Long short-term memory (LSTM) Recurrent Neural Network architecture are compared in this study.

3.1. Speech pre-processing for deep neural networks

Speech signals of the fixed time window length 2s, 4s and 8s are processed by a pre-emphasis filter to spectrally flatten the speech signals and boost higher frequencies. Spectral features are extracted from these windows.

3.2. Input Audio representations

1. Spectrogram: the spectrogram as a time-frequency representation of the speech signal of a time window is generated by a short-time Fourier transform (STFT) with short frames size 25ms and stride of 10ms [11]. The Hamming window is applied to each frame and Short-time Fourier transform is computed to get the power spectrum.
2. Log Mel Spectrogram: Mel filter banks (n=40) are applied to the power spectrum to get the Mel spectrum. Mel filter banks use Mel-frequency scaling, which is a perceptual scale to replicate human ear perception of sound [12]. It corresponds to better resolution at low frequencies and less at high frequencies.

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \tag{1}$$

$$f = 700(10^{\frac{m}{2595}} - 1) \tag{2}$$

where f is frequency in Hertz and m is Mel scale

Thus, we use the spectrogram and log Mel spectrogram to represent the spectral features of the speech signals as inputs to deep neural network.

Spectrogram and log Mel spectrogram of a speech signal of a fixed time window is mapped with respiratory sensor value at the endpoint of the time window with a stride of 10ms between windows to train the CNN and RNN models. These models will be used to estimate the respiratory sensor values of a speech signal in real time to get the breathing pattern.

3.3. Deep Neural Network models

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 3: *Deep neural network configurations for sensor value prediction*

In the current work the neural networks were implemented using the pytorch software framework [13]. In the CNN model [14], the data is fed into a network of two convolutional 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 rectified linear unit activation function is applied to introduce non-linearity into the feature extraction process for each convolutional layer as shown in Figure 3.

In the RNN-LSTM model [15], 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 network weights iterative based on training data [16]. Mean squared error is used as the regression loss function. These hyperparameters for the network are best chosen for estimation after repeated experimentation. Both CNN and RNN-LSTM networks are trained with the data of 15 specific subjects and tested on 5 other specific subjects.

4. Results

4.1. Mean squared error and Correlation

Spectrogram and log mel spectrograms are considered for comparison as input spectral features of speech for estimation of breathing pattern. Both CNN and RNN networks result in lesser mean squared error and higher correlation for actual and estimated breathing signals with log mel spectrogram compared to spectrograms for a fixed time window of 4s as shown in Table1.

Models	**Log Mel-Spectrogram**	**Spectrogram**
RNN (MSE)	0.0017	0.0058
RNN (Correlation)	0.47	0.21
CNN (MSE)	0.00229	0.016
CNN (Correlation)	0.41	0.27

Table 1: *comparison of spectrogram vs log Mel spectrogram with MSE and correlation for time window length of 4s.*

Hence log mel spectrograms are used as input representations in our study for estimating breathing parameters. We compare deep neural networks performance and breathing parameters for 2s, 4s and 8s time window lengths for estimating breathing signal and results are formulated in Figures 4-7. For comparing each subject's breathing rate estimation, we performed a leave one group (subject) out cross validation for 20 subjects using RNN-LSTM model with time window length 4s and the results are formulated in Figure 9.

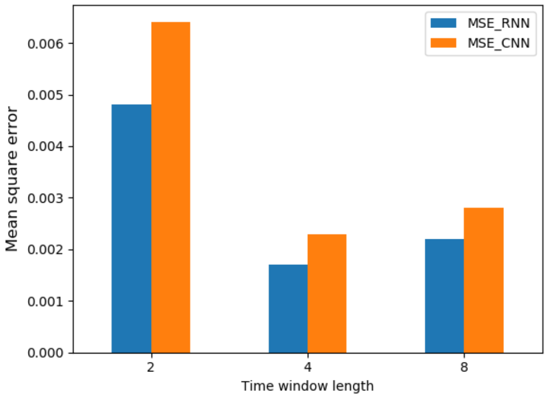

Figure 4: *Mean Squared Error for CNN and RNN-LSTM for time windows 2s, 4s, 8s of speech (log mel-spectrogram representation).*

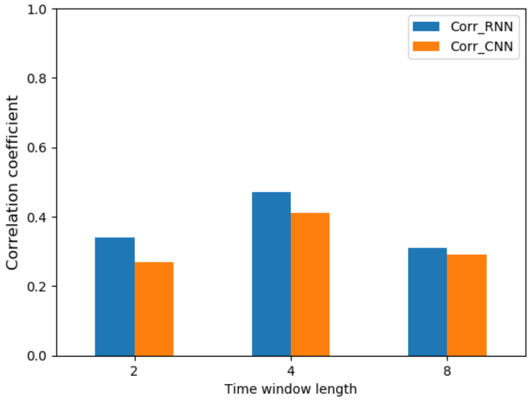

Figure 5: *Correlation for CNN and RNN-LSTM for time windows 2s, 4s, 8s of speech (log mel-spectrogram representation).*

4.2. Breathing parameters

Breathing signal is analyzed to get breathing rate and tidal volume which are the important respiratory parameters to detect the pathological condition of a person. These parameters are compared for the actual and estimated sensor data to determine the accuracy of estimation.

1. *Breathing rate* is average number of breaths per minute and is computed by using peak detection algorithm [17].
2. *Tidal volume* is a measure of the amount of air a person inhales during a normal breath. It gives information about the lung capacity of a person [9]. We normalise the average area under the curve per breath and use it to describe tidal volume. This normalised tidal volume equivalent is used for comparison for actual and estimated breathing signal.

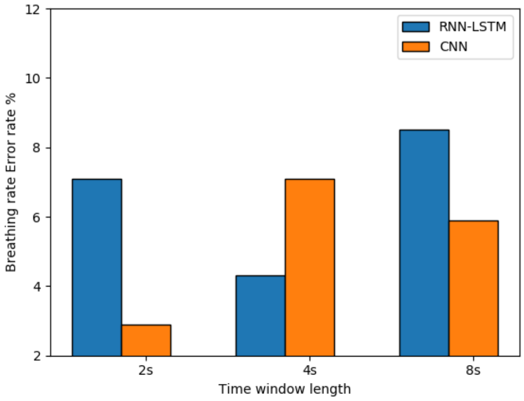

Figure 6: *Error percentage in Breathing rate estimation for CNN and RNN-LSTM for time windows 2s, 4s, 8s (log mel-spectrogram representation).*

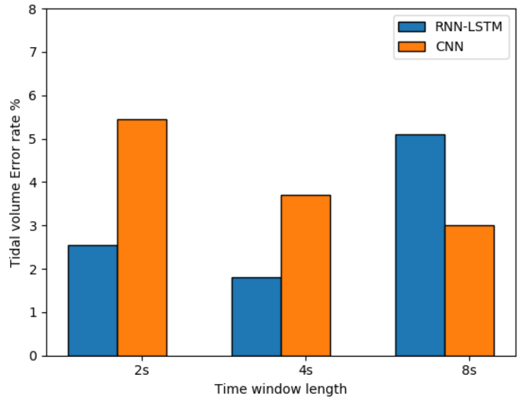

Figure 7: *Error percentage in Tidal Volume estimation for CNN and RNN-LSTM for time windows 2s, 4s, 8s (log mel-spectrogram representation).*

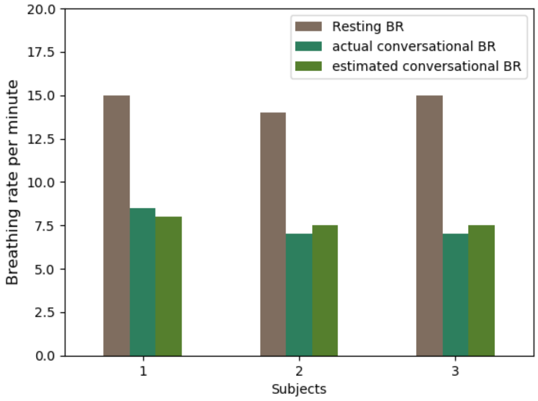

Figure 8: *Comparison of 3 subject's resting breathing rates with actual and estimated conversational breathing rate.*

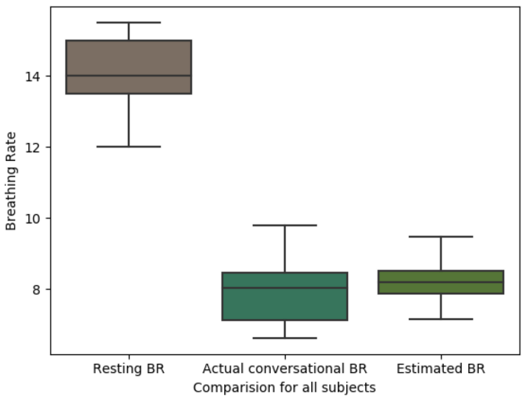

Figure 9: *Comparison of distribution of resting breathing rate, actual and estimated conversational breathing rate for all subjects using RNN-LSTM model.*

4.3. Summary

Our findings are summarised as follows:

1. Log Mel spectrogram representation gives better breathing signal estimation as input to the neural network than the spectrogram representation.
2. From Figures 4-7, 4s window of speech is optimum with lesser loss and higher correlation for estimating breathing signal parameters with 4.3 % error for breathing rate and 1.8% error for Tidal volume.
3. RNN-LSTM model performs better with a correlation of 0.47 and MSE loss of 0.0017 compared to CNN model with a correlation of 0.41 and MSE loss of 0.00229 for a 4s time window.
4. From Figures 8-9, we show that the breathing rate during conversational speech is nearly half the normal breathing rate. We got a sensitivity of 91.2 % for breath event detection and an mean absolute error of 1.018 breaths per minute for breathing rate estimation using our model.

5. Conclusions

In this paper, we propose a method to estimate breathing pattern during speech using deep learning models. We expect our proposed method can be used for estimating the breathing pattern of a customer during a call to the telehealth call center. This work can be extended by implementing different deep learning architectures like attention based models, multi-task learning with breathing rate as an auxiliary training parameter for better estimation. Breathing pattern would give us information about the respiration rate, breathing capacity and thus enable us to understand the pathological condition of a person using speech during conversations. This would help early and remote diagnosis for various health conditions.

6. Acknowledgements

This work was partially supported by the Horizon H2020 Marie Skodowska-Curie Actions Initial Training Network European Training Network project under grant agreement No. 766287 (TAPAS) and Data Science Department, Philips Research, Eindhoven.

7. References

[1] D. Ruinskiy and Y. Lavner, "An effective algorithm for automatic detection and exact demarcation of breath sounds in speech and song signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 838–850, March 2007.

[2] D. H. Klatt, K. N. Stevens, and J. Mead, "Studies of articulatory activity and airflow during speech*," *Annals of the New York Academy of Sciences*, vol. 155, no. 1, pp. 42–55, 1968.

[3] A. MacLarnon and G. P. Hewitt, "The evolution of human speech: The role of enhanced breathing control," *American journal of physical anthropology*, vol. 109, pp. 341–63, 07 1999.

[4] J. Slifka, "Some physiological correlates to regular and irregular phonation at the end of an utterance," *Journal of Voice*, vol. 20, no. 2, pp. 171 – 186, 2006.

[5] J. D. Hoit, N. P. Solomon, and T. J. Hixon, "Effect of lung volume on voice onset time (vot)," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 3, pp. 516–520, 1993.

[6] J. E. Huber, B. Chandrasekaran, and J. J. Wolstencroft, "Changes to respiratory mechanisms during speech as a result of different cues to increase loudness," *Journal of Applied Physiology*, vol. 98, no. 6, pp. 2177–2184, 2005, pMID: 15705723.

[7] M. Wodarczak and M. Heldner, "Respiratory Constraints in Verbal and Non-verbal Communication," *Frontiers in Psychology*, vol. 8, May 2017.

[8] M. Wodarczak, M. Heldner, and J. Edlund, "Breathing in conversation : An unwritten history," in *Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication :*, ser. Linkping Electronic Conference Proceedings, no. 110, 2015, pp. 107–112.

[9] 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.

[10] 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.

[11] E. Sejdi, I. Djurovi, and J. Jiang, "Timefrequency feature representation using energy concentration: An overview of recent advances," *Digital Signal Processing*, vol. 19, no. 1, pp. 153 – 183, 2009.

[12] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[13] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

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

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

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computing Research Repository (CoRR)*, vol. abs/1412.6980, 2015.

[17] S. Fuchs, U. D. Reichel, and A. Rochet-Capellan, "Changes in speech and breathing rate while speaking and biking," in *ICPhS 2015: 18th International Congress of Phonetic Sciences*, 2015. [Online]. Available: http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-25254-5