ON THE RELATIONSHIP BETWEEN SPEECH-BASED BREATHING SIGNAL PREDICTION EVALUATION MEASURES AND BREATHING PARAMETERS ESTIMATION

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] École polytechnique fédérale de Lausanne, Lausanne, Switzerland
[3]Philips Research, Eindhoven, The Netherlands
[4]Centre for Language Studies (CLS), Radboud University Nijmegen, The Netherlands

ABSTRACT

The respiratory system is one of the major components of the speech production system. Any alteration in breathing can result in changes in speech. Specific breathing characteristics, such as breathing rate and tidal volume, can indicate a person's pathological condition. More recently, neural network-based methods have started emerging for predicting the breathing signal from the speech signal. The neural networks are trained and evaluated with different objective measures, such as mean squared error (MSE) and Pearson's correlation. This paper investigates whether there is a systematic relationship between the different objective measures used for training and evaluating the neural network models and the end-goal, i.e. estimation of breathing parameters such as, breathing rate and tidal volume. Our investigations on two different data sets with two different neural network-based approaches show that there is no clear systematic relationship. In other words, obtaining a high Pearson's correlation on the evaluation set does not necessarily mean better breathing parameter estimation. Thus, indicating the need for developing other objective evaluation measures.

Index Terms— Respiratory parameters, Neural Networks, Speech breathing

1. INTRODUCTION

The respiratory system provides the necessary energy for producing speech by pushing air through vocal folds. This mechanism is called Speech Breathing [1]. Any changes in breathing can result in variations in speech. Pathological conditions, such as heart diseases, Parkinson's Disease (PD), and Chronic Obstructive Pulmonary Diseases (COPD), can significantly affect the breathing patterns. Measuring such variations can be used to investigate the underlying health condition of a person [2]. Breathing parameters such as breathing rate and tidal volume are used to quantify the variations in breathing. Estimating breathing parameters from speech therefore can be used as a non-intrusive diagnosis method.

Earlier studies on the relationship between breathing and speech show that the type of speech can affect breathing patterns [3, 4, 5], and breathing can shape the speech as well [6]. Speech breathing also impacts other speech aspects such as voice quality [7], voice onset time [8], and loudness [9].

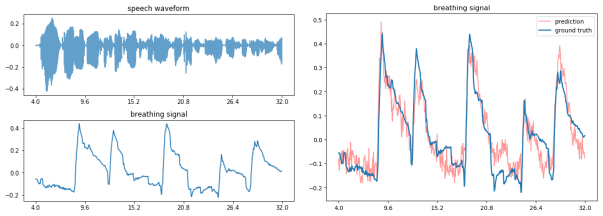
Predicting breathing patterns from speech signal has only recently gained more attention. Nallanthighal et al. used log Mel Spectrogram of speech to train a Convolutional Neural Network (CNN) and a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) to predict the breathing signal [10, 11]. More recently, as part of Interspeech 2020 ComParE challenge, Schuller et al., introduced several methods, including an End-to-End system consisting of a CNN combined with an LSTM-RNN to predict the breathing signal from speech [12]. In these works *Pearson's correlation coefficient* (r), which is generally used to determine the similarity between two signals, is used as quality measure for evaluating the performances of the systems. The other alternative measure studied in [10, 11] is the mean squared error (MSE).

As an end goal, the speech signal-based breathing parameter estimation can be regarded as a non-intrusive "instrument" for measuring breathing parameters. In other words, from the predicted signal we should be able to measure well breathing parameters such as breathing rate and tidal volume for application purposes. The neural network training and evaluation does not consider these aspects. So, a question that arises is: whether the evaluation measures used to evaluate the breathing signal prediction models are sufficient indicators for reliable breathing parameter estimation? We address this research question through single database and cross database studies using two deep learning-based approaches.

The remainder of the paper is organized as follows. Section 2 provides a background on estimation of breathing parameters from the breathing signal. Section 3 presents the experimental setup and Section 4 presents our result and analysis. Section 5 finally concludes with key observations.

2. BREATHING PARAMETERS ESTIMATION

Figure 1 illustrates speech signal with the ground truth breathing signal and a predicted breathing signal.



(a) (b)

Fig. 1: (a) Speech waveform and corresponding breathing signal and (b) Predicted and ground truth breathing signals.

Different breathing parameters are estimated from the breathing signal in the following manner:

1. *Breath event* is the event of inhalation, which marks the beginning of the breathing cycle. A breathing cycle during speech includes a sharp peak during inhalation and a gradual decline during exhalation which is repeated over the course of an utterance, as shown in Figure 1. The breath events are detected using the Automatic Multiscale Peak Detection algorithm (AMPD) [13]. The sensitivity of breath events is reported to evaluate the overlap of breath events between actual and predicted breathing signals.

2. *Breathing rate* is the average number of breath events per minute [14].

3. *Speech Tidal volume* is a measure of the amount of air a person inhales during a normal breath for speech. It gives information about the lung capacity of a person [15]. We use the normalized average area under the curve per breath as speech tidal volume for both actual and estimated breathing signals. The area under the curve per breath is the area between an estimated baseline and the actual signal values between each two local minimums. We consistently use the term "tidal volume" to describe the above-mentioned speech tidal volume equivalent in this paper.

The important aspect while computing breathing parameters is fixing a look-ahead window for computing the peaks in the signals using AMPD algorithm. We fix this window to 2s or 50 samples for the signal with sampling rate of 25 Hz. This 2s window is selected based on our observation that the minimum gap between two inhalation breath events is more than 2s. We use the same look-ahead window for both estimated and ground truth breathing signals. Figure 2 illustrates an example of a breathing signal with marked breathing events and estimated baseline.

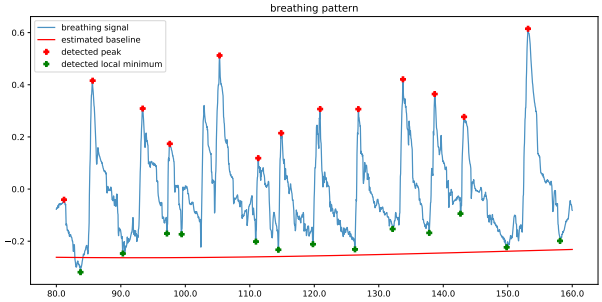


Fig. 2: An illustration of breathing pattern and the estimated breath events (peaks), local minimums, and baseline used for breathing parameter estimation.

3. EXPERIMENTAL SETUP

3.1. Dataset and evaluation protocol

We used the UCL Speech Breath Monitoring (UCL-SBM) database provided as part of Interspeech 2020 ComParE challenge [12]. This database includes recordings of conversational speech 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. Similar to the studies done in [12], all the systems were trained using the *Train* set, the first 15 speakers for training and the last 2 speakers for cross validation. We then tested our systems on the whole *Dev* set.

Beside training and evaluating the neural networks on a single database, we also conducted a cross database study where models trained on UCL-SBM database are tested on the Philips read speech database introduced in [11]. The Philips database includes the recordings from 40 healthy subjects (18 female and 22 male with age group ranging from 21 to 40 years old). The subjects are asked to read "The Rainbow Paragraph.", a widely used phonetically balanced paragraph [16]. The speech and breathing signal are recorded with sampling frequency of 48kHz and 2kHz respectively. For our experiments we downsampled the breathing signal to acquire the same sampling frequency as UCL-SBM database, i.e. 25Hz.

As per the ComParE challenge evaluation protocol, the performances are reported for a single signal made by concatenating the predicted breathing signal and the ground truth signal for all the files in the *Dev* set. The *Pearson's correlation coefficient* (r) is used as the measure of evaluation. We also computed mean squared error (MSE), as an alternative measure. We analyzed these measures w.r.t to the errors in estimation of breathing parameters presented in Section 2.

3.2. Approaches

We performed our studies using different systems obtained based on two different neural network-based approaches, as part of the Interspeech 2020 ComParE challenge investigation. The Pearson's correlation for the best baseline system on the *Test* set was **0.731** [12]. The approaches reported in this paper obtained a best Pearson's correlation of **0.707** on the *Test* set of the challenge [17]. We briefly present these systems in this section. More information can be found in [17].

3.2.1. Raw speech waveform modeling based approach

This approach takes inspiration from recent works on speech recognition [18] speaker recognition [19], and gender recognition [20]. As illustrated in Figure 3, in this approach the input to the neural network is raw speech waveform and the output is a sample-by-sample prediction of the breathing signal.

The CNN model consisted of four convolution layers followed by a hidden layer, and then finally an output layer. The number of filters in convolution layers were 128-256-512-512 with kernel sizes of 30-10-4-3 and kernel strides of 10-5-2-1. After each layer, there were max-pooling layers with strides of 2-3-1-1 and rectified linear unit (ReLu) as activation function. The MLP had one hidden layer with 10 hidden units with hyperbolic tangent (Tanh) as the activation function. Batch normalization was also applied after each layer. The output layer consisted of one unit with linear activation. Adam optimizer [21] with learning rate of 0.001 was used for training. The system was implemented using Tensorflow [22].

Three different loss functions had been used to train the CNNs, namely,

- Mean squared error loss denoted as MSE.
- A custom correlation loss defined as follows:

$$L_{Corr}(y, f(x)) = \frac{1}{1 + r(y, f(x))} - 0.5 \qquad (1)$$

  where $r(y, f(x))$ is the *Pearson's correlation coefficient*. This loss function is denoted as Corr. During training, the loss function was computed by predicting the breathing signal for a fixed number of consecutive input frames and calculating the correlation between the predicted and ground truth signals for that duration.

- a composite loss combining MSE and Corr, denoted as Corr-MSE,

The two hyper-parameters namely *input window length* and *correlation window length* were chosen after repeated experimentation. The former is the duration of past speech signal input to the system at each time frame and the latter is the number of consecutive samples used for calculating correlation based loss functions. The best system with MSE loss had input window length of 3s. The input window length for the best systems trained with Corr and Corr-MSE loss were 2s and 3s, respectively. The best correlation window length for both correlation based loss function was 1024 samples. The input frame shift was 40 ms, corresponding to the sampling rate of the output signal.

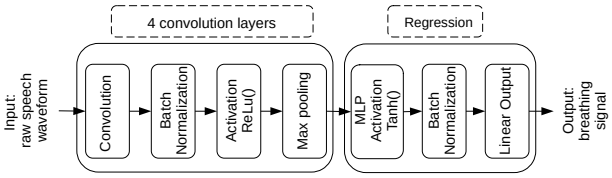


Fig. 3: An illustration of the raw waveform modeling based approach for breathing signal prediction.

3.2.2. Short-term speech spectral feature-based approach

In this approach, log Mel spectrogram is fed as input to an LSTM-RNN model and the breathing signal is predicted at the output in a sample-by-sample manner [10]. The LSTM-RNN model consisted of two LSTM layers with 128 hidden units. Adam optimiser with a learning rate of 0.001 was used to train the network. The system was implemented using PyTorch [23]. Two different loss functions had been used to train the network, namely,

- Mean squared error loss denoted as MSE.
- BerHu loss [24]:

$$L_{\delta}(y, f(x)) = \begin{cases} (|y - f(x)| - 0.5\delta), & \text{if } |y - f(x)| \leq \delta \\ \delta * 0.5 * (y - f(x))^2, & \text{otherwise} \end{cases} \qquad (2)$$

The motivation behind using BerHu loss was the fact that, in speech breathing patterns, usually a sudden peak (inhalation) is followed by a gradual descending curve (exhalation). Thus, for the model to estimate breathing patterns, the loss function should be more sensitive to peaks (outliers) and less sensitive for the rest, which can be achieved by using BerHu loss function.

4. RESULTS AND ANALYSIS

As mentioned earlier, all the neural network models were trained on ComParE challenge UCL-SBM training set. We conducted two breathing parameter estimation and evaluation studies, (a) single database study, where the breathing signal is estimated on the ComParE challenge *Dev* set of the UCL-SBM data and (b) cross database study, where the breathing signal is estimated on the whole Philips read speech database. The evaluation of the breathing parameters estimated from the predicted signal was carried out by comparing them to the

ground truth and measuring breathing rate (BR) error, breath event (BE) sensitivity and tidal value (TV) error. In the remainder of the section, for the sake of clarity, we will denote the systems in the following format: *ANN type_input type_loss function*.

4.1. Single database

Table 1 presents the result of the systems trained and tested on UCL-SBM data set.

Table 1: The r, MSE and breathing parameters for systems using raw waveform based and spectral based approaches. The abbreviations used for breathing parameters in the table are as following: BR=Breathing Rate, BE=Breathing Events, and TV=Tidal Volume.

loss Functions	r	MSE	Breathing Parameters		
			BR error (%)	BE Sensitivity	TV error (%)
Raw Waveform Based Approach					
MSE	0.411	0.0263	11.31%	0.887	28.13%
Corr	0.49	2.1068	39.77%	0.982	12.94%
Corr-MSE	0.463	0.0253	22.96%	0.933	18.25%
Spectral Based Approach					
MSE	0.482	0.039	11.65%	0.908	13.42%
BerHu	0.448	0.018	14.42%	0.882	11.68%

It can be seen that the systems trained with both approaches are performing closely in terms of correlation. This performance is comparable to the ComParE challenge best baseline system performance reported on the *Dev* set [12]. It can be observed that other metrics calculated for the systems are largely different. For example, the CNN_raw_Corr system with the correlation of 0.49 has a very high MSE and BR error compared to all the other systems, but the TV error for it is among the lowest. If we consider the systems trained with MSE loss, i.e. CNN_Raw_MSE, and LSTM-RNN_Spec_MSE, we observe that they both yield similar error for breathing rate (11.31% vs 11.65%) and MSE (0.026 vs 0.039), while the correlation and TV error for them are very different. Furthermore, if we compare CNN_Raw_Corr-MSE and LSTM-RNN_Spec_BerHu with similar correlation and MSE, we observe a large difference in breathing parameters, around 8% for BR error and 6% for TV error.

4.2. Cross database

Table 2 shows the performance of the systems trained on UCL-SBM database (models in Table 1) when tested on the whole 40 subjects of the Philips database.

It can be seen that the systems corresponding to the spectral based approach are yielding a higher correlation compared to the systems corresponding to the raw waveform modeling based approach, but they yield higher TV error. Furthermore, although the correlation values are considerably low compared to the single database study, the BR error, BE sensitivity and TV error are in similar ranges. Interestingly, the raw waveform based approach with MSE yields one of the lowest TV error, despite having a very low correlation.

It is evident from these results that r and MSE measures are not good indicators of BR error, BE sensitivity and TV error.

Table 2: Train on UCL-SBM conversational speech database and test on all Philips read speech database.

loss Functions	r	MSE	Breathing Parameters		
			BR error (%)	BE Sensitivity	TV error (%)
Raw Waveform Based Approach					
MSE	0.187	0.0737	10.91%	0.715	4.79%
Corr	0.298	1.9923	33.45%	0.993	5.14%
Corr-MSE	0.137	0.0928	19.81%	0.886	15.3%
Spectral Based Approach					
MSE	0.324	0.072	13.22%	0.862	11.56%
BerHu	0.292	0.108	16.82%	0.804	17.84%

5. CONCLUSION

Speech-based breathing signal prediction using neural networks is an emerging area of research. Currently, these models are evaluated based on Pearson's correlation between the predicted breathing signals and the ground truth signals. Breathing signal prediction and breathing parameter estimation studies conducted in this paper on two data sets with five different systems show that high Pearson's correlation measure or low MSE not necessarily indicates better breathing parameter estimation. Our future work will focus on investigating neural network training loss functions and evaluation measures that take into consideration breathing parameter estimation, in order to be consistent with the end-goal.

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.

7. REFERENCES

[1] C Von Euler, "Some aspects of speech breathing physiology," in *Speech motor control*, pp. 95–103. Elsevier, 1982.

[2] N.P. Solomon and T.J. Hixon, "Speech breathing in parkinson's disease," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 294–310, 1993.

[3] Wang et al., "Breath group analysis for reading and spontaneous speech in healthy adults," *Folia Phoniatrica et Logopaedica*, vol. 62, no. 6, pp. 297–302, 2010.

[4] Henderson et al., "Temporal patterns of cognitive activity and breath control in speech," *Language and Speech*, vol. 8, no. 4, pp. 236–242, 1965.

[5] Winkworth et al., "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.

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

[7] 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.

[8] Hoit et al., "Effect of lung volume on voice onset time (vot)," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 3, pp. 516–520, 1993.

[9] 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.

[10] V. S. Nallanthighal, A. Härmä, and H. Strik, "Deep sensing of breathing signal during conversational speech," in *Proc. of Interspeech*, 2019, pp. 4110–4114.

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

[12] Schuller et al., "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly emotion, Breathing & Masks," in *Proceedings of Interspeech*, Shanghai, China, 2020, pp. 2042–2046.

[13] F. Scholkmann, J. Boss, and M. Wolf, "An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals," *Algorithms*, vol. 5, no. 4, pp. 588–603, 2012.

[14] 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.

[15] 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.

[16] G. Fairbanks, "The rainbow passage," *Voice and articulation drillbook*, vol. 2, 1960.

[17] Z. Mostaani, V. S. Nallanthighal, A. Härmä, H. Strik, and M. Magimai.-Doss, "Estimating breathing pattern from raw speech waveform and short-term speech spectrum using neural networks," Idiap-RR Idiap-Internal-RR-40-2020, Idiap, 2020.

[18] D. Palaz, M. Magimai-Doss, and R. Collobert, "End-to-end acoustic modeling using convolutional neural networks for hmm-based automatic speech recognition," *Speech Communication*, vol. 108, pp. 15–32, 2019.

[19] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "Towards directly modeling raw speech signal for speaker verification using CNNS," in *Proc. of ICASSP*, 04 2018, pp. 4884–4888.

[20] 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.

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

[22] Abadi et al, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[23] Paszke et al., "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'Alché-Buc, E. Fox, and R. Garnett, Eds., pp. 8024–8035. Curran Associates, Inc., 2019.

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