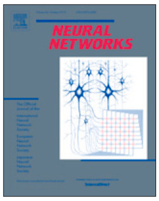
Neural Networks 141 (2021) 211–224

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



2021 Special Issue

Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings[☆,☆☆]

Venkata Srikanth Nallanthighal[a,b,*], Zohreh Mostaani[c,d], Aki Härmä[a], Helmer Strik[b], Mathew Magimai-Doss[c]

[a] *Philips Research, Eindhoven, The Netherlands*
[b] *Centre for Language Studies (CLS), Radboud University Nijmegen, The Netherlands*
[c] *Idiap Research Institute, Martigny, Switzerland*
[d] *Ecole polytechnique fédérale de Lausanne, Lausanne, Switzerland*

ARTICLE INFO

Article history:
Available online 5 April 2021

Keywords:
Speech breathing
Signal processing
Deep neural networks
Respiratory parameters
Speech technology

ABSTRACT

Respiration is an essential and primary mechanism for speech production. We first inhale and then produce speech while exhaling. When we run out of breath, we stop speaking and inhale. Though this process is involuntary, speech production involves a systematic outflow of air during exhalation characterized by linguistic content and prosodic factors of the utterance. Thus speech and respiration are closely related, and modeling this relationship makes sensing respiratory dynamics directly from the speech plausible, however is not well explored. In this article, we conduct a comprehensive study to explore techniques for sensing breathing signal and breathing parameters from speech using deep learning architectures and address the challenges involved in establishing the practical purpose of this technology. Estimating the breathing pattern from the speech would give us information about the respiratory parameters, thus enabling us to understand the respiratory health using one's speech.

1. Introduction

The COVID-19 pandemic demonstrated the necessity of remote digital health assessment tools like telehealth monitoring services for sustainable health care, particularly pertinent for elderly and vulnerable populations. Speech is a good indicator of the pathological condition of a person (Dibazar et al., 2002). Speech production is a complex process involving perfect synchrony among various systems like muscular, respiratory, cognitive, and autonomic nervous systems. Any lapse in these systems would significantly affect one's speech. The use of speech analytics has been gaining attention within the clinical and healthcare domains in recent years. This follows to the success of deep learning techniques in various speech technology applications (Miotto et al., 2017) and speech pathology (Cummins et al., 2018; Koolagudi et al., 2018), and has been replacing the conventional feature-based machine learning techniques (Cummins et al., 2017; Teixeira et al., 2013).

Speech and respiration are closely related. Speech is produced by organs evolved for the respiratory function of the body (MacLarnon et al., 1999). Breathing is a primary mechanism of speech generation. The special mechanism of using the respiratory system to produce the airflow necessary for phonation is termed as Speech Breathing (Von Euler, 1982). Speech breathing is implicated in many aspects of speech production, such as voice quality (Slifka, 2006), voice onset time (Hoit et al., 1993), and loudness (Huber et al., 2005). Vocalization mostly takes place during exhaling while inhaling is done in quick pauses in between utterances. We subconsciously exercise continuous breathing planning during the speech. Breathing planning is evident as we take in more air for a long continuous utterance (Włodarczak et al., 2017). Respiratory diseases such as Chronic Obstructive Pulmonary Disease (COPD), asthma, and respiratory infection are common in elderly populations. These conditions significantly influence the breathing capacity and thus influence breathing planning resulting in frequent breaks in an utterance in need of air. We can hear when a person has breathing difficulties, but the automatic detection is a complex task because the breathing

[☆] This document is partially supported by the Horizon H2020 Marie Skłodowska-Curie Actions Initial Training Network European Training Network project under grant agreement No. 766287 (TAPAS), Data Science Department, Philips Research, Eindhoven and Swiss National Science Foundation (SNSF) through the project Towards Integrated processing of Physiological and Speech signals (TIPS) grant number 200021_188754.

[☆☆] In this work we demonstrate that estimating breathing signal from speech is plausible using deep learning architectures.

[*] Corresponding author at: Philips Research, Eindhoven, The Netherlands.
E-mail address: srikanth.nallanthighal@philips.com (V.S. Nallanthighal).
URL: https://www.tapas-etn-eu.org/Members/srikanth--nallanthighal (V.S. Nallanthighal).

https://doi.org/10.1016/j.neunet.2021.03.029
0893-6080/

planning is based on linguistic and prosodic factors (Włodarczak et al., 2015).

The current research is related to the development of acoustic sensing technology for telehealth services related to respiratory conditions in particular. 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. This work is a continuation of our earlier work on sensing breathing signals and breathing parameters from speech (Nallanthighal et al., 2019, 2020).

In this paper, we present a comprehensive research study on deep learning architectures for estimating breathing patterns and breathing parameters directly from speech and address the following **challenges to establish the respiratory analysis of speech as a practical or clinical application.**

1. Firstly, we establish the possibility of estimating breathing signal from the speech signal in two ways: the spectral analysis and raw waveform analysis. We then propose a fusion method to enhance our system's performance.
2. To make the study more robust and reliable for practical or clinical purposes, we study two significant protocols: read speech and conversational (or spontaneous) speech. We perform the study on individual databases. We further perform a cross database analysis to assess how our systems generalize. This comprehensive study helps in applying this technology independent of databases.
3. For evaluating the estimated breathing signal from the neural network models, we use standard evaluation metrics for the regression problem: mean squared error and correlation with reference to actual breathing signal measured through respiratory inductive belts as discussed in Section 4.4. However, the practical utility would be justified by comparing the breathing parameters like breathing rate, tidal volume equivalent, and breath event sensitivity.
4. In some cases, breathing sounds are also audible in a speech recording. In such cases, it is easier to detect breath events. Ruinskiy and Lavner proposed an effective breath-event detection algorithm based on template matching for automatic detection and exact demarcation of breath sounds during speech (Ruinskiy et al., 2007). However, in the experiments reported in this paper, breathing sounds are not recorded or used in the analysis, which is ensured by recording speech at a distance from the speaker's mouth during data collection, as described in Section 4. The respiratory sensing from the speech is based on the composition of speech utterance, i.e., linguistic content and prosodic factors independent of breath sounds.

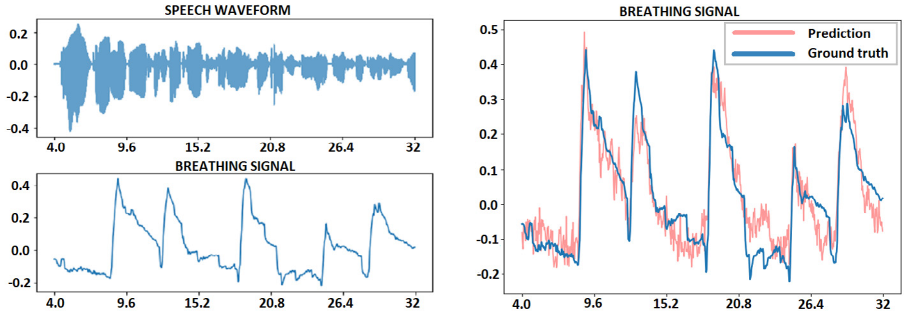
The remainder of the paper is organized as follows. We provide a background on the relation between speech and respiration and breathing parameter estimation from breathing signal in Section 2. In Section 3, we present the different approaches investigated for estimating breathing signal from speech, which includes neural networks based on spectral features input and raw speech input with various loss functions. Following this, in Section 4, we detail the experimental setup and databases used. In Section 5, we present results of different experimental studies. In Section 6, we present an analysis of the studied approaches. Finally, in Section 7, we conclude the paper.

2. Background

Very few studies are focused on the relationship of speech and respiration and the effect of speech on breathing pattern in the recent years. Breathing patterns provide medical doctors and speech therapists vital information about an individual's respiratory and speech planning (Székely et al., 2020), as well as cognitive and neurological health (Heck et al., 2017; Mitchell et al., 1996). J.D. Hoit and T.J. Hixon conducted early studies and explored different aspects of speech breathing like age (Hoit et al., 1987), body type (Hoit et al., 1986), gender (Hoit et al., 1989), and pathological conditions with neuromotor disorders to evaluate respiratory control in individuals (Solomon et al., 1993). Winkworth investigated the associations between linguistic factors and lung volumes in read speech and concluded that speech breathing is subject to a number of linguistic and prosodic influences (Winkworth et al., 1994). The amount of air breathed in and the volume of air in the lungs have been shown to be strongly influenced by the length and loudness of the intended utterance, whereas the expiratory duration is largely determined by the linguistic intent and this expiration can be modeled as a composition of phonemes with varying exhaustion flows for vowel and consonant phonemes (Klatt et al., 1968). 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 (Hammarsten et al., 2015). 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 (Henderson et al., 1965; Wang et al., 2010; Winkworth et al., 1994). The later could be due to the cognitive load during spontaneous speech (Mitchell et al., 1996).

When utilizing breathing for the purpose of the speech, the rate and volume of inhalation and the rate of exhalation during the utterance seem to be governed largely by the speech controlling system and its requirements with respect to phrasing, loudness, and articulation. Speech breathing demands more effort than normal quiet breathing. Quiet breathing encompasses relatively equal phases of inhalation and exhalation in terms of duration, amplitude, and velocity, whereas speech breathing is characterized by short inhalations to minimize interruptions to the speech flow and long exhalations due to higher resistance in the upper airway that prevent air from quickly flowing out (Puller, 1988). Thus during speech, the breathing rate is approximately halved compared to quiet breathing (Nallanthighal et al., 2019; Włodarczak et al., 2017). Wlodarczak et al. proposed that the relationship between speech and breathing is not one way and breathing can also shape the speech (Włodarczak et al., 2017). Thus, speech and breathing are closely related and gives an intuition for our hypothesis that speech breathing pattern can be sensed from the linguistic content and prosodic factors of the speech. This hypothesis has also been the basis for Breathing Sub-challenge of Interspeech 2020 ComParE challenge (Schuller et al., 2020).

Both the rib cage and the abdomen can be used to modulate alveolar pressure and airflow during speech. Some speakers exhibit the more vigorous use of rib cage over abdominal contributions, and some speakers show a relatively equal contribution from both the rib cage and the abdomen (Hixon et al., 1976). The chest wall has been treated as a two-part kinematic system comprised of the rib cage and diaphragm–abdomen in parallel with only one degree of freedom each (Konno et al., 1967), and wherein the volume displaced by each part is linearly related to the motions of points within it (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 (Konno et al., 1967).



(a) (b)

Fig. 1. (a) The speech waveform and corresponding breathing signal and (b) the predicted and ground truth for breathing signal.

Breathing signals are analyzed to get breathing rate and tidal volume equivalent, which are the essential respiratory parameters to detect a person's pathological condition. These parameters are compared for the actual and estimated sensor data to determine the accuracy of estimation.

1. *Breath event* is the event of inhalation, which marks the beginning of the breathing cycle. During speech, we observe a sharp peak during inhalation and gradual decline during exhalation, as shown in Fig. 1. This quasi-periodic pattern repeats over the course of a speech utterance. These peaks are determined using the Automatic Multiscale Peak Detection algorithm (AMPD) (Scholkmann et al., 2012), which is particularly relevant for detecting peaks in noisy periodic and quasi-periodic signals. Breath events of both actual and predicted breathing signals are compared to evaluate the overlap of breath events, which ensures better prediction. This overlap is evaluated by the sensitivity of breath events.
2. *Breathing rate* is average number of breath events per minute (Fuchs et al., 2015) and is computed by using Automatic Multiscale Peak Detection algorithm(AMPD) (Scholkmann et al., 2012).
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 (Konno et al., 1967). We normalize the average area under the curve per breath and use it as a tidal volume equivalent. This normalized tidal volume equivalent is used to compare actual and estimated breathing signals. We consistently use the term "tidal volume" to describe the above-mentioned speech tidal volume equivalent in this paper.

3. Approaches

In the previous section, we observed that there exists a relationship between speech and respiration. This suggests that breathing signal could be estimated from the speech signal. Estimating breathing signals from the speech signal can be formulated as a regression problem. Such a problem could be formulated as a feature vector-to-feature vector regression problem (Qi et al., 2019, 2020) or feature vector-to-signal regression problem (Xu et al., 2014) or signal-to-signal regression problem (Fu et al., 2017; Rethage et al., 2018; Sebastian et al., 2020). The feature vector-to-feature vector regression formulation presumes that there exists a mathematical model for the signals based on the regressed features. In the case of speech signal, such a model is based on source–system decomposition through short-term analysis (Makhoul, 1975; Oppenheim et al., 2004; Ou et al., 2012). However, defining such a feature-based model for breathing signal is not trivial. Feature vector-to-signal regression and signal-to-signal regression problems have largely focused on problems where the input and output signals are of the same kind. For instance, in speech enhancement the input is corrupted speech signal and the output is clean speech signal (Fu et al., 2017; Rethage et al., 2018). In the case of breathing signal estimation from speech signal, we are dealing with two different kinds of signals. Furthermore, although there exists a relationship between speech and respiration, this relationship has not been fully characterized. In the sense that we do not know exactly which properties of speech signal characterize breathing signal. So, in the present work, we approached breathing signal estimation in two different ways using deep learning:

1. apply short-term spectral processing on the input speech signal and then map the resulting representations to breathing signal. We refer to it as spectral features based approach. Here again we consider two sub-approaches. First, where the envelop of short-term spectrum is extracted and modeled with temporal context to estimate breathing signal. Second, where no such prior knowledge is applied and the short-term Fourier magnitude spectrum with temporal context is modeled to estimate breathing signal. This can be regarded as a feature vector-to-signal regression formulation.
2. learn to predict breathing signal directly from raw waveforms. We refer to it as raw waveform based approach. This can be regarded as a signal-to-signal regression formulation.

The motivation to use deep learning comes from the fact that we have less prior knowledge about the problem. Deep learning is capable of tackling lack of prior knowledge (Goodfellow et al., 2016).

3.1. Spectral features based approach

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 Mel-frequency scale, i.e., log Mel spectrogram. We use the spectrogram and log Mel spectrogram to represent the spectral features of the speech signals as inputs to neural networks.

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 frame size of 25 ms and stride of 10 ms (Sejdić et al., 2009). The Hamming window is applied to each frame and STFT 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.

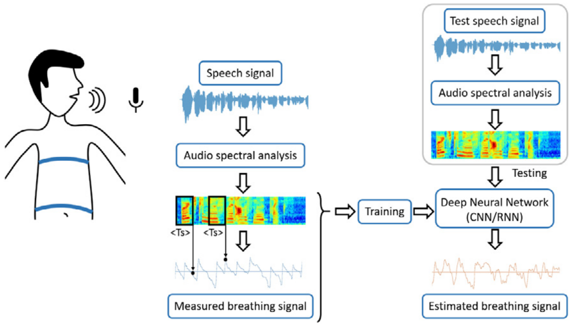


Fig. 2. Schematic diagram for estimating respiratory signal using Deep Neural Network Model based on spectral features.

CNN Model	RNN Model
Input: log Mel spectrogram or spectrogram m: frames in time window n : Mel filter banks	Input: log Mel spectrogram or spectrogram m: frames in time window n : Mel filter banks
Matrix $X_i(1 \times m \times n)$	Matrix $X_i(1 \times m \times n)$
1 × 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

Fig. 3. Deep neural network configurations of the spectral based methods for sensor value prediction.

Mel filter banks use Mel-frequency scaling, which is a perceptual scale to replicate human ear perception of sound (Stevens et al., 1937). 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

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 10 ms between windows for Philips database and a stride of 40 ms between windows for UCL-SBM database to train the neural network models as shown in Fig. 2. 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 (LSTM-RNN) models using the PyTorch software framework (Paszke et al., 2019). In the CNN model (Schmidhuber, 2015), the data is fed into a network of two convolutional layers with a single-channel and kernel 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 Fig. 3. Batch normalization is also applied on each convolution layer. This is followed by 3 fully connected layers with ReLU activation function. Adam optimizer (Kingma et al., 2015) with a learning rate of 0.001 is used as an optimization algorithm.

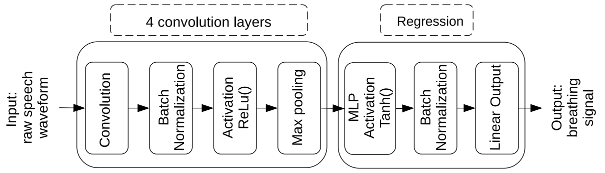


Fig. 4. An illustration of the end-to-end CNN model used in raw waveform based methods.

In the LSTM-RNN model, the data is fed into a network of two LSTM layers with 128 hidden units and a learning rate of 0.001. Adam optimizer is used as an optimization algorithm to update network weights iterative based on training data (Kingma et al., 2015). These hyperparameters for the network are best 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, lesser MSE, and comparable breathing parameters would be considered best for our study.

3.2. Raw speech waveform based approach

We used a CNN-based model to predict breathing signal values from raw speech waveform in this approach. The architecture has been originally proposed for speech recognition (Palaz et al., 2013) and later on has been studied on various speech processing tasks such as speaker recognition (Muckenhirn et al., 2018), gender recognition (Kabil et al., 2018), and more recently for depression detection (Dubagunta et al., 2019). Additionally, other studies have been performed to understand how the source and system-related information in the speech signal is modeled when using this approach (Muckenhirn et al., 2019, 2018; Palaz et al., 2019). The breathing signal can be related closely to the source of the speech signal and therefore, we chose to use this architecture in our experiments.

The CNN consists of four convolution and max-pooling layers followed by a fully connected layer (MLP), and finally, an output layer, as illustrated in Fig. 4. The number of filters in convolution layers is 128-256-512-512, with kernel sizes of 30-10-4-3 and kernel strides of 10-5-2-1. There are max-pooling layers with strides of 2-3-1-1 and a rectified linear unit (ReLu) as activation function after each layer. The MLP has one hidden layer with 10 units with hyperbolic tangent (Tanh) as activation. The output layer consists of one or two units with linear activation depending on the study. Batch normalization is also applied after each layer. Adam optimizer with learning rate of 0.001 is used for training. The system is implemented using Tensorflow (Abadi et al., 2016).

The input to the system is raw speech waveform, and the output of the neural network is a sample-by-sample prediction of the breathing signal and, when needed, the sensor gradient as well 4.2.1.1. In the latter case, the output of the network is two dimensional.

3.3. Fusion based approach

In this approach, we chose the best model from each of the previously mentioned approaches 3.1, 3.2, and we fuse the predicted breathing signal values by aligning the two signal and taking the mean of them. We expect the fused estimated breathing signal to be much closer to the actual breathing signal as it has the prospects of both spectral and raw waveform methods.

3.4. Regression loss functions

The loss function of a neural network model characterizes how well the model performs over training data. We investigate various loss functions in this study, starting with the standard regression loss functions to customized loss functions based on our problem of estimating breathing signal.

1. **Mean squared error loss function:** It is the most common regression loss function. It is a quadratic loss (L2 loss) and is the sum of squared distances between the target variable and predicted values.
2. **BerHu loss function:** In speech breathing patterns, the breath events are usually a sudden peak (inhalation) 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 (Zwald et al., 2012).

$$L_\delta(y, f(x)) = \begin{cases} (|y - f(x)| - 0.5\delta), & \text{if } |y - f(x)| \leq \delta \\ \delta * 0.5 * (y - f(x))^2, & \text{otherwise} \end{cases} \tag{3}$$

   We use BerHu loss as loss function for robust regression by integrating the advantages of both the L2 norm and L1 norm, thus penalizing the outliers (peaks) resulting in accelerated optimization of the model for estimating breathing signal.
3. **Correlation loss function:** We use *Pearson's correlation coefficient* as a measure of similarities between the predicted and actual breathing signals. It is, therefore, reasonable to define a loss function that optimizes this measure directly. We defined a custom Correlation loss function as following:

$$L_{Corr}(y, f(x)) = \frac{1}{1 + r(y, f(x))} - 0.5 \tag{4}$$

   where $r(y, f(x))$ is the *Pearson's correlation coefficient*. During training, the Correlation loss is computed by predicting the output signal values for a fixed number of consecutive samples and then calculating the *Pearson's correlation coefficient*, $r()$, between the predicted and the ground truth values. Hence the time dependency between consecutive samples is taken into account during training. We consider the number of consecutive samples, correlation window as a hyper-parameter.
4. **Correlation-MSE loss function:** The custom Correlation loss function removes all the scaling information and focuses on the cycles of the signal. It does not necessarily enforce the actual predicted signal values to get closer to the ground truth. To account for this aspect, we use a combination of Correlation and MSE loss in our training. The custom Correlation-MSE loss function is defined as the following:

$$L_{Corr-MSE}(y, f(x)) = L_{MSE}(y, f(x)) + L_{Corr}(y, f(x)) \tag{5}$$

   where $L_{MSE}(y, f(x))$ is the mean squared error loss and $L_{Corr}(y, f(x))$ is the custom Correlation loss defined in Eq. (4).

We investigate and compare the mean squared error loss function, and Berhu loss function for spectral features based methods, and the mean squared error loss function, customized Correlation, and Correlation-MSE loss functions for raw waveform based methods. The selection of appropriate loss functions has been reported based on repeated experimentation for significant performance improvement.

3.5. Hyper-parameters for models

The following are the essential parameters to be tuned for the design of our neural network models.

1. **Length of time window:** The window length for each speech input representation, i.e., raw waveform, spectrograms, and log Mel spectrograms, is crucial for estimating the breathing sensor value. We investigate speech inputs of fixed window length of 2s, 4s, and 8s for spectral representations and 2s, 3s, and 4s for raw waveform for better estimation.
2. **Length of correlation window:** It is defined as the number of consecutive points in the output of the system that is used to find the correlation between ground truth and predicted values during training. We investigated systems with correlation window sizes of 400, 512, and 1024.
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 investigated mapping with sensor value at the beginning and midpoint of the time window for Philips read speech database and found no significant difference in the estimation performance. Thus we extended the same endpoint mapping for models for both read speech and conversational speech protocols.

4. Experimental setup

4.1. Database and protocols

The study explores the respiratory analysis on the following two protocols: read speech and conversational speech. We use a speech database developed at Philips Research for read speech (Nallanthighal et al., 2019), and for conversational speech, we use the UCL Speech Breath Monitoring (UCL-SBM) database (Schuller et al., 2020).

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 2 kHz. These belts work on the principle of respiratory inductance plethysmography (RIP). They consist of a sinusoidal wire coil insulated inelastic. The belts' dynamic stretching creates waveforms due to changes in self-inductance and oscillatory frequency of the electronic signal. The electronics convert this change in frequency to a digital respiration waveform where the waveform's amplitude is proportional to the inspired breath volume. Thus the sum of the rib cage and abdomen expansions measured by the respiratory belt transducers is considered as the measure for the breathing signal. Earthworks microphone M23 is used for recording high-quality speech at 48 kHz. 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 (Fairbanks, 1960).

UCL Speech Breath Monitoring (UCL-SBM) database has been introduced in (Schuller et al., 2020). 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. However, we use a subset of this database: 17 speakers in Train subset and 16 speakers in Dev subset for training, validating, and testing as the respiratory signals of test subset speakers are not publicly available. For each speaker, a 4 min recording of speech with a sampling frequency of 16 kHz is provided. For speakers in Train and Dev sets, the breathing signal with a sampling frequency of 25 Hz is provided, which amounts to a sequence of 6000 values for each speaker. In this database, the respiratory signal is recorded from one (MLT1132, ADInstruments, Castle Hill, Australia) of the two piezoelectric respiratory belts worn by the subjects. The belt is positioned approximately four centimeters below the collarbone to record chest breathing and produces a linear voltage reading in response to changes in thoracic circumference associated with respiration.

4.2. Systems

Our experiments are based on the breathing signal estimation using neural network models in two different protocols: read speech protocol and conversational speech protocol, as described in Section 4.1.

In each protocol, performance of neural network models based on spectral features and raw waveform is compared with different cost functions (Section 3.4) and evaluated based on metrics of evaluation (Section 4.4). A fusion of two estimated breathing signals, each from neural network models based on spectral features and neural network models based on the raw waveform, is also reported. The same neural network models and system configurations are used for both read speech and conversational speech protocol. System configurations are explained in detail in this section.

4.2.1. Spectral features based systems

In spectral features based systems, for the Philips database of read speech protocol, we use 29 subjects for training, 3 subjects for validation, and 8 subjects for testing. Here the speech signal is sampled at 48 kHz, and the breathing signal is downsampled to 100 Hz. For the UCL-SBM database of conversational speech protocol, we use 15 subjects from the subset Train for training and the remaining 2 subjects from the Train subset for validation. All the 16 subjects of subset Dev are used for testing. Here the speech signal is sampled at 16 kHz, and the breathing signal's sampling frequency is 25 Hz. We use MSE and BerHu loss as described in Section 3.4 to train our systems.

Respiratory sensor as output. Spectral features of the speech (spectrograms or log Mel spectrogram as described in Section 3.1) of a fixed time window (2s, 4s, and 8s) 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, i.e., linguistic content and prosodic factors in that particular time window. Spectral features of speech and known respiratory sensor values are mapped with a stride of 10 ms between windows for Philips database and a stride of 40 ms between windows for UCL-SBM database to train neural network models, as defined in Section 3.1. These trained models are used to estimate the respiratory sensor values from a target speech signal in real-time to estimate the breathing signal. We compare the performance of the CNN and LSTM-RNN models using the spectral representations and fixed time window. The results of Table 1 suggest that log Mel spectrogram is a preferred input spectral feature representation of speech signal, and a fixed time window of 4s provides the least mean squared error loss and a high correlation for estimating the breathing signal. With this inference, we investigate all the models based on spectral features approach with input representation as log Mel spectrograms and fixed window length of 4s.

Table 1
Comparison of window lengths for spectrograms and log Mel spectrograms with MSE loss function for sensor vs speech signal model.

Models	*Window size*	*r*	MSE
CNN Model	2s	0.26	0.066
log Mel Spectrogram	**4s**	**0.472**	**0.034**
	8s	0.32	0.030
LSTM RNN Model	2s	0.36	0.026
Log Mel Spectrogram	**4s**	**0.476**	**0.019**
	8s	0.34	0.031
CNN Model	2s	0.29	0.096
Spectrogram	4s	0.27	0.016
	8s	0.24	0.062
LSTM RNN Model	2s	0.24	0.082
Spectrogram	4s	0.21	0.058
	8s	0.23	0.074

4.2.1.1. ***Respiratory sensor and sensor gradient as output.*** The gradient of the respiratory sensor signal over a fixed time window can be visualized as the net airflow equivalent (inhalation and exhalation) during that fixed time window. We investigated by mapping the spectral features of the speech of a fixed time window with the gradient of the sensor signal over a fixed time window. Log Mel spectrogram and known respiratory sensor gradient values are mapped with a stride of 10 ms between windows for Philips database and a stride of 40 ms between windows for UCL-SBM database to train deep neural network model so that the model understands the relationship of airflow over a speech utterance in a fixed time window. We used the same CNN and LSTM-RNN model architecture used for sensor and speech mapping models and observed a positive correlation close to 0.2, which explains its relevance but not sufficient to get good performance for estimating the breathing signal. However, by mapping speech signal of a fixed time window to both respiratory sensor value at the endpoint of the time window and respiratory sensor gradient over the fixed time window of 4s, we may achieve a better estimation of breathing pattern and enable the model to learn the breathing state and airflow relationship over a speech utterance. Based on this hypothesis, we investigate by training the models by mapping both the sensor and the sensor gradient for better estimation.

4.2.2. Raw speech waveform based systems

In raw speech waveform based systems, for the Philips database of the read speech protocol, similar to the systems using spectral features, we use 29 subjects for training, 3 subjects for validation, and 8 subjects for testing. However, we downsample the speech signal to 16 kHz and the breathing signal to 25 Hz. This is done for reducing the computational time and complexity for training the models based on raw speech waveform. For conversational speech protocol, similar to the systems using spectral features, we use 15 subjects of subset Train for training, the remaining 2 subjects of subset Train for validation, and 16 subjects of subset Dev for testing. Here the speech signal is sampled at 16 kHz, and the breathing signal's sampling frequency is 25 Hz. We train our systems with an input window length of 2s, 3s, and 4s and a correlation window length of 400, 512, and 1024 output samples. We use MSE, Correlation, and Correlation-MSE loss as described in Section 3.4 to train our systems.

Respiratory sensor as output. In this set of experiments, similar to the case for spectral features based systems, we map the raw speech waveform of input window length to sensor value at the endpoint of this window. We use a stride of 40 ms between windows to train the models for both Philips and UCL-SBM databases. The result are shown in Tables 3 and 4. The hyper-parameters for the best performance is variable based on the database and loss function. The details for choosing the best systems are explained in Section 5.

Respiratory sensor and sensor gradient as output. In this set of experiments, similar to the case for spectral features based systems, we map the raw speech waveform of input window length to both sensor value and the sensor gradient value over the input window length. We use a stride of 40 ms between windows to train the models for both Philips and UCL-SBM databases. We investigated if the gradient information will help our systems to learn the breathing pattern more accurately. The results are shown in Tables 3 and 4.

4.2.3. Fusion of the best systems

As mentioned in 3.3, we fuse the best system from spectral features based and raw speech waveform based approaches and investigate system performance. We perform this analysis for the systems trained by using only sensor values and systems trained by using both sensor and sensor gradient values on both read speech and conversational speech protocols. In case of Philips database where the output signal sampling frequency is different between the two approaches we downsampled the result from the spectral based method to 25 Hz, to be consistent with the output sampling frequency of raw waveform based database. The results are reported in Tables 3 and 4 and discussed in Section 5.

4.3. Cross database study

To investigate the generalization ability of the models trained with read speech and conversational speech protocols, we performed a cross database analysis. We consider the following scenarios:

1. Train on Philips read speech database and test on UCL-SBM conversational speech database.
2. Train on UCL-SBM conversational speech database and test on Philips read speech database.

It is interesting to obtain good performance in either of these scenarios which suggest that the models are learning the common aspects of the speech and breathing signal relationship independent of the speech protocol. The results are reported in Tables 3 and 4 and further discussed in Section 5.

4.4. Metrics for evaluation

Estimating breathing signal from speech is a regression problem. The following metrics are used to evaluate the estimation of the predicted breathing signal: *Pearson's correlation coefficient*, mean squared error, and breathing parameters. A higher Pearson's correlation ensures that the prediction follows the trend of the actual signal and a low mean squared error ensures that the prediction is in a dynamic range of the actual signal. Breathing parameters as described in Section 2, derived from actual and estimated breathing signal should be comparable with minimum error and high breath event sensitivity. Thus a model which can estimate the breathing signal with higher *Pearson's correlation coefficient*, lower mean squared error with respect to actual signal and comparable breathing parameters can be best considered.

Table 2
The hyper parameters of the reported systems for the raw waveform based approach. They have been chosen based on the *Pearson's correlation coefficient* of the systems trained using only respiratory sensor values as output.

Database	Loss function	Input window length (s)	Correlation window length (samples)
Philips	MSE	3	–
	Corr	4	400
	Corr-MSE	4	400
UCL-SBM	MSE	3	–
	Corr	2	1024
	Corr-MSE	3	1024

5. Results

The results for the experiments explained in Section 4 are presented here. The systems are trained using the Philips database of the read speech protocol and UCL-SBM database of the conversational speech protocol. We further perform a cross-database analysis to investigate the generalization abilities of our systems.

We trained our systems with the mentioned loss functions and hyper-parameters as discussed in Sections 3.4 and 3.5. In spectral based approaches, we found out that MSE and BerHu loss functions are performing better than correlation-based loss functions. As mentioned in Table 1, we observed that using log Mel spectrogram with input window length of 4s performs better compared to other combinations. In the rest of this paper, we report on the spectral based methods trained with log Mel spectrogram with a fixed window length of 4s and MSE and BerHu loss. We similarly trained our systems with different loss functions for the raw waveform based approach and observed that the MSE, Correlation, and Correlation-MSE loss perform better. Therefore we report the systems using these three loss functions. We trained our system with an input window of 2s, 3s, and 4s, and the correlation window size of 400, 512, and 1024 samples. The best performing system is different for Philips and UCL-SBM database and for different loss functions. In each case, we chose the system with the best performance in terms of correlation among those trained with only respiratory sensor values as an output and reported their performance. Table 2 summarizes the hyper-parameters of the reported systems for the raw waveform based methods on both Philips and UCL-SBM databases.

We are also reporting breathing parameters, including breathing rate error, breathing event sensitivity, and tidal volume error rate in each case. The important aspect while computing breathing parameters from the actual and estimated breathing signal is fixing the look-ahead window for computing the peaks in the signals using Automatic Multiscale Peak Detection algorithm (AMPD). We fix this look-ahead window to 2s which constitutes to 200 samples when the sampling rate of the breathing signal is 100 Hz and to 50 samples when the sampling rate is 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 of 2s for both true breathing signal and estimated breathing signal for a fair comparison. We observe that the estimated breathing rate is usually higher than the true breathing rate in Tables 3, 4, 5, and 6. This is due to the occurrence of intermittent peaks in the estimated breathing signal obtained from neural network models. Also, the estimated breathing signals usually are noisier than the true signal and there is high probability for the peak detection algorithm to find a false peak.

Table 3
Philips Database (read speech protocol): The r, MSE and breathing parameters for systems using spectral based, raw waveform based, and fusion based approaches.

Models	Loss Function	r	MSE	Breathing Parameters				
				Breathing Rate			Breath Event	Tidal Volume
				prediction (breaths/min)	true (breaths/min)	error (%)	Sensitivity	error (%)
Spectral Based Approach								
I/P: log Mel spec	MSE	0.476	0.019	10.42	9.84	5.89%	0.916	12.11%
O/P: Sensor	BerHu	0.482	0.039	10.98	9.84	11.58%	0.902	16.24%
Architecture: RNN								
I/P: log Mel spec	MSE	0.452	0.021	11.03	9.84	12.09%	0.882	11.62%
O/P: sensor & sensor gradient	BerHu	0.463	0.019	11.80	9.84	19.91%	0.842	14.72%
Architecture: RNN								
I/P: log Mel spec	MSE	0.472	0.034	10.85	9.84	10.26%	0.896	16.22%
O/P: sensor	BerHu	0.462	0.042	11.78	9.84	19.71%	0.821	18.84%
Architecture: CNN								
I/P: log Mel spec	MSE	0.437	0.051	11.52	9.84	17.68%	0.847	19.86%
O/P: sensor & sensor gradient	BerHu	0.413	0.063	12.10	9.84	22.96%	0.811	20.46%
Architecture: CNN								
Raw Waveform Based Approach								
I/P: raw waveform	MSE	0.47	0.0699	12.00	9.90	21.29%	0.929	18.47%
O/P: sensor	Corr	0.534	1.8627	11.81	9.90	19.35%	0.942	14.4%
Architecture: CNN	Corr-MSE	0.502	0.0616	11.24	9.90	13.55%	0.897	26.65%
I/P: raw waveform	MSE	0.45	0.068	12.51	9.90	26.45%	0.903	30.5%
O/P: sensor & sensor gradient	Corr	0.449	0.8857	12.45	9.90	25.81%	0.955	45.69%
Architecture: CNN	Corr-MSE	0.447	0.0874	11.30	9.90	14.19%	0.890	8.25%
Fusion Based Approach								
Model_1: spectral based	MSE							
Model_2: raw waveform based	Corr	0.562	1.024	10.68	9.84	8.53%	0.908	10.21%
O/P: sensor								
Model_1: spectral based	MSE							
Model_2: raw waveform based	Corr-MSE	0.480	0.071	11.18	9.84	13.61%	0.882	14.42%
O/P: sensor & sensor gradient								

5.1. Philips database study

Table 3 presents the performance of the systems trained and tested on the Philips database for the three approaches.

Looking into the systems trained with only sensor values, it is evident that the raw waveform based methods, on average, yields a slightly higher *Pearson's correlation coefficient* and a higher MSE. The MSE for systems trained with Correlation loss, however, is much higher than other methods. The spectral based methods yield lower breathing rate and tidal volume error than the raw waveform based methods; however, the breath events sensitivity is also slightly lower. In general, both spectral features based and raw waveform based approaches perform similarly, considering all the mentioned evaluation metrics. The spectral features based method with the highest *Pearson's correlation coefficient* is an RNN model trained with BerHu loss and $r = 0.482$. Even though comparable to other systems, the MSE and breathing parameters for this model are not the best. The next highest r value belongs to the RNN system trained with MSE loss function, which has lower MSE and performs better in terms of breathing parameters. Considering all the evaluation metrics, we chose the RNN model trained with MSE loss as the best model to be used for the fusion approach. In the case of raw waveform based methods, the system with the highest *Pearson's correlation coefficient* is trained with Correlation loss function with $r = 0.534$. This is the highest correlation among all the trained systems, not considering the fusion systems. The MSE for the same model is also the highest, with a value of 1.8627. It can be due to the nature of the defined Correlation loss function, which removes any scaling from the data and only focuses on the signal's repetitive temporal aspect. Unlike MSE loss functions, correlation-based loss functions do not decrease the distance between actual and predicted samples. The breathing parameters for this model are comparable or sometimes better than other systems. The reason for this can be because the peak detection algorithm is not sensitive to the actual peak value of the signal. Similarly, the tidal volume computation focuses on relative behavior of minimum and maximum value of the signal around a peak, but not their actual value. By combining Correlation and MSE loss, we acquire a system that performs comparable to the system with only correlation loss, but the MSE is much lower, and the breathing parameters, if not the best, are comparable to other systems. Adding MSE to the loss functions obligates the actual predicted values to get closer to the true values.

Looking into the systems trained with both sensor and sensor gradient, we observe a slight decrease in the *pearson's correlation coefficient* of all the systems and increase in the MSE in most of them. They behave differently with regard to breathing parameters however, it can be noted that overall we obtain a less performance using both sensor and sensor gradient values. For the systems trained with the spectral based approach, considering all the evaluation metrics, the system with the best performance is the LSTM-RNN trained with MSE loss. For raw waveform based approach, considering all the evaluation metrics, the system with the best performance is the CNN trained with Correlation loss for sensor output models and Correlation-MSE loss for sensor and sensor gradient output models. These two systems from each approach are used for the fusion based approach. The fusion system's performance obtained by aligning and averaging the predicted values for the chosen best systems from each approach is presented in the last section of Table 3. It can be seen that by only fusing the two systems in the output level, we can gain a boost in the performance of our system specially in terms

Table 4
UCL-SBM Database (conversational speech protocol): The r, MSE and breathing parameters for systems using spectral based, raw waveform based, and fusion based approaches.

Models	Loss Function	r	MSE	Breathing Parameters				
				Breathing Rate			Breath Event	Tidal Volume
				prediction (breaths/min)	true (breaths/min)	error (%)	Sensitivity	error (%)
Spectral Based Approach								
I/P: log Mel spec	MSE	0.482	0.039	10.44	9.35	11.65%	0.908	13.42%
O/P: Sensor	BerHu	0.448	0.018	10.71	9.35	14.42%	0.882	11.68%
Architecture: RNN								
I/P: log Mel spec	MSE	0.463	0.019	10.42	9.35	11.44%	0.871	08.74%
O/P: sensor & sensor gradient	BerHu	0.427	0.016	10.84	9.35	15.93%	0.840	10.55%
Architecture: RNN								
I/P: log Mel spec	MSE	0.437	0.042	11.11	9.35	18.82%	0.841	19.29%
O/P: sensor	BerHu	0.460	0.042	10.57	9.35	13.04%	0.864	18.62%
Architecture: CNN								
I/P: log Mel spec	MSE	0.411	0.036	10.62	9.35	12.24%	0.810	21.72%
O/P: sensor & sensor gradient	BerHu	0.413	0.063	11.44	9.35	22.31%	0.822	24.82%
Architecture: CNN								
Raw Waveform Based Approach								
I/P: raw waveform	MSE	0.411	0.0263	10.45	9.39	11.31%	0.887	28.13%
O/P: sensor	Corr	0.490	2.107	13.12	9.39	39.77%	0.982	12.94%
Architecture: CNN	Corr-MSE	0.463	0.0253	11.55	9.39	22.96%	0.933	18.25%
I/P: raw waveform	MSE	0.406	0.0268	09.20	9.39	02.00%	0.797	23.15%
O/P: sensor & sensor gradient	Corr	0.470	2.464	12.98	9.39	38.27%	0.957	08.36%
Architecture: CNN	Corr-MSE	0.459	0.0253	11.61	9.39	23.63%	0.928	17.92%
Fusion Based Approach								
Model_1: spectral based	MSE							
Model_2: raw waveform based	Corr	0.512	1.822	11.24	9.35	20.21%	0.942	12.33%
O/P: sensor								
Model_1: spectral based	MSE							
Model_2: raw waveform based	Corr-MSE	0.466	0.022	10.64	9.35	13.7%	0.862	14.44%
O/P: sensor & sensor gradient								

of correlation. We benefit from information learned by the two individual systems and obtain a system that performs better. This can be beneficial as we can train smaller and less complicated systems that are easier to train and still benefit from them.

Another point that can be seen in Table 3 is regarding the breathing rate for the ground truth and predicted values. The true breathing rate for the labels reported in the spectral based approach (9.84 breaths/minute) is slightly different from the raw waveform based approach (9.90 breaths/minute). The reason is that in our implementation of the spectral based methods, the first sample of breathing signal mapped to spectral features is the endpoint of the first 4s window. Thus the first 4s at the beginning of the breathing signal is dropped for each subject and this slight difference in the length of the breathing signal results in slight difference of breathing rates in the two approaches.

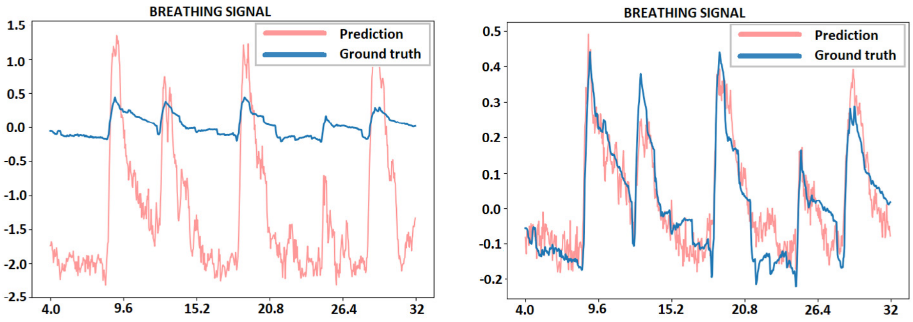
5.2. UCL-SBM database study

We performed a similar investigation using the UCL-SBM database. Table 4 presents the results for the spectral features, raw waveform, and fusion based methods. When looking to the systems trained with only sensor values, compared to the systems trained on the Philips database, on average, we obtain slightly lower *Pearson's correlation coefficient*, higher MSE, and higher error in the breathing parameters. We observe a similar trend among the performances of the systems compared to the ones trained on the Philips database. The *Pearson's correlation coefficient* for the raw waveform based methods are very similar to those of the spectral based methods. The MSE for the system trained with the Correlation loss is much higher than other systems, while the obtained *Pearson's correlation coefficient* is also the highest. Fig. 5 shows an example of the predicted and ground truth values for the breathing signal when trained using Correlation and Correlation-MSE loss. From Fig. 5 the reason for such a high MSE when using Correlation loss is evident. The system is not able to predict the appropriate dynamic range of the output and therefore, even though the predicted signal is following the same trend as the ground truth, their actual values are far from each other. For the systems trained with the spectral based approach, considering the evaluation metrics, the system with the best performance is the LSTM-RNN trained with MSE loss. For raw waveform based approach the system with the best performance is the CNN trained with Correlation loss for sensor output models and Correlation-MSE loss for sensor and sensor gradient output models. These two systems from each approach are used for the fusion.

Looking into the systems trained with both sensor and sensor gradient values we again do not see much improvements. From the last section of Table 4, fusing the best systems from two approaches, we obtain a system that performs better than the individual systems similar to Philips Database.

5.3. Cross database study

We performed a cross-database investigation for the systems trained with only sensor values. We used the systems trained on the Philips database to predict the breathing signal on UCL-SBM database and calculated the performance. Table 5 shows the result of our investigation. It can be seen that the spectral based approach methods are performing better compared to the



(a) (b)

Fig. 5. The predicted and ground truth for the breathing signal for a raw waveform based method trained on the UCL-SBM database using (a) Correlation and (b) Correlation-MSE loss functions.

Table 5
Train on Philips read speech database and test on UCL-SBM conversational speech database.

Models	Loss Function	*r*	MSE	Breathing Parameters				
				Breathing Rate			Breath Event	Tidal Volume
				prediction (breaths/min)	true (breaths/min)	error (%)	Sensitivity	error (%)
Spectral Based Approach	MSE	0.372	0.039	10.39	9.35	11.12%	0.872	15.64%
	BerHu	0.344	0.031	11.04	9.35	18.07%	0.820	14.20%
Raw Waveform Based Approach	MSE	0.353	0.0457	11.83	9.39	25.96%	0.895	29.48%
	Corr	0.284	2.2949	12.70	9.39	35.27%	0.933	18.57%
	Corr-MSE	0.299	0.0562	11.73	9.39	24.96%	0.867	3.20%
Fusion Based Approach	MSE,MSE	0.398	0.042	10.44	9.35	11.65%	0.868	14.62%

Table 6
Train on UCL-SBM conversational speech database and test on Philips read speech database.

Models	Loss Function	*r*	MSE	Breathing Parameters				
				Breathing Rate			Breath Event	Tidal Volume
				prediction (breaths/min)	true (breaths/min)	error (%)	Sensitivity	error (%)
Spectral Based Approach	MSE	0.364	0.049	10.42	9.84	5.89%	0.916	14.42%
	BerHu	0.331	0.071	11.24	9.84	14.22%	0.853	21.62%
Raw Waveform Based Approach	MSE	0.129	0.0727	9.13	9.9	7.74%	0.684	9.52%
	Corr	0.217	2.5687	13.15	9.9	32.9%	0.942	8.76%
	Corr-MSE	0.070	0.0922	12.32	9.9	24.52%	0.923	1.27%
Fusion Based Approach	MSE,Corr	0.347	2.346	10.66	9.84	8.3%	0.898	11.22%

raw waveform based methods, however their performance decreases compared to the performance when tested on the Philips database. We can gain even a better performance by fusing the systems from the two approaches and obtain a system with a correlation of 0.398, MSE of 0.042, and comparable breathing parameters.

Table 6 shows the result of testing the systems on the Philips database while trained on the UCL-SBM database. Once again, we observe a better performance from the spectral-based methods than the raw waveform based methods. The best system performs with a correlation of 0.364, MSE of 0.049, and comparable breathing parameters in the spectral based approach. The performance of the raw waveform based methods decreases drastically in this case. We observe that fusing the two systems seems to perform slightly worse than only the spectral-based model.

6. Analysis

This section presents an analysis of the studied approaches.

6.1. *MAE loss function*

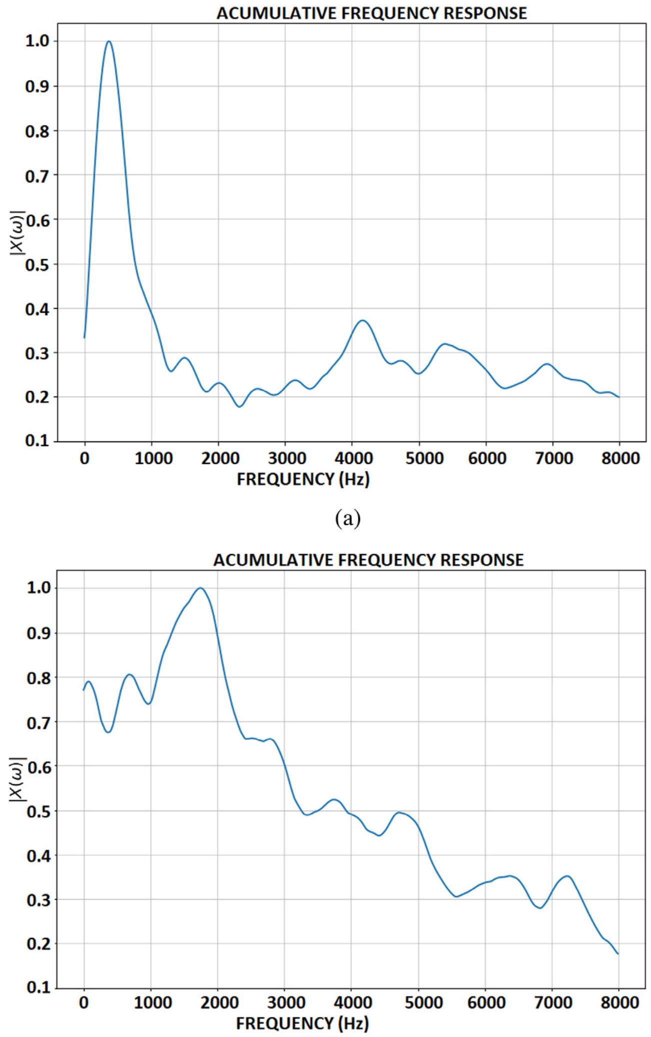
In this paper, we have investigated different loss functions (see Section 3.4). In the speech enhancement literature, the mean absolute error (MAE) loss function has also been proposed for regression (Qi et al., 2020). We conducted an analysis study using the MAE loss function on the Philips database. Table 7 presents the results of the study. It can be observed that MAE loss tends to yield comparable to those trained with MSE loss, see Table 3, both in terms of correlation coefficients and breathing parameter estimation. We do not observe a distinctive advantage as in the case of speech enhancement studies.

6.2. *Analysis of raw waveform based approach*

In the experimental studies, we observed a trend that spectral based approach typically yields better performance than the raw waveform based approach, particularly in terms of MSE and correlation. This is more evident in the cross database investigation. To understand the reason behind that, we analyzed the first

Table 7
MAE loss function for Philips Database (read speech protocol): The r, MSE and breathing parameters for systems spectral based systems and raw waveform based systems.

Models	Loss Function	r	MSE	Breathing Parameters				
				Breathing Rate			Breath Event	Tidal Volume
				prediction (breaths/min)	true (breaths/min)	error (%)	Sensitivity	error (%)
Spectral Based Approach								
RNN (I/P: log Mel spec, O/P: sensor)	MAE	0.448	0.027	11.15	9.84	13.31%	0.864	14.02%
RNN (I/P: log Mel spec, O/P: sensor & sensor gradient)	MAE	0.411	0.029	12.11	9.84	23.06%	0.862	17.27%
CNN (I/P: log Mel spec, O/P: sensor)	MAE	0.451	0.047	11.55	9.84	17.37%	0.844	23.24%
CNN (I/P: log Mel spec, O/P: sensor & sensor gradient)	MAE	0.419	0.081	12.03	9.84	22.25%	0.872	24.12%
Raw Waveform Based Approach								
CNN (I/P: raw waveform, O/P: sensor)	MAE	0.489	0.0572	11.68	9.90	18.06%	0.897	12.72%
CNN (I/P: raw waveform, O/P: sensor)	Corr-MAE	0.507	0.0592	11.94	9.90	20.65%	0.929	20.09%
CNN (I/P: raw waveform, O/P: sensor & sensor gradient)	MAE	0.431	0.063	11.3	9.90	14.19%	0.877	47.39%
CNN (I/P: raw waveform, O/P: sensor & sensor gradient)	Corr-MAE	0.486	0.0548	10.66	9.90	7.74%	0.858	12.01%



(a)

(b)

Fig. 6. The accumulative frequency response of the kernels for the first layer of the CNN model trained using Correlation-MSE loss trained on (a) Philips and (b) UCL-SBM database.

Table 8
Comparison between first 20 log Mel filterbank energies and 40 log Mel filterbank energies as input. The loss function used for all the systems is MSE. The abbreviations used for breathing parameters in the table are as following: BR = Breathing Rate, BE = Breathing Events, and TV = Tidal Volume.

Models	r	MSE	Breathing Parameters		
			BR error (%)	BE Sensitivity	TV error (%)
RNN-20	0.415	0.078	12.8%	0.882	17.2%
RNN-40	0.476	0.019	5.89%	0.916	12.11%
CNN-20	0.423	0.096	12.11%	0.858	15.6%
CNN-40	0.472	0.034	10.26%	0.896	16.22%

convolution layer filters by computing the cumulative frequency response. Fig. 6 shows the cumulative frequency response of the first convolution layer for CNN trained on Philips database and UCL-SBM database. It can be seen that in both cases, the filters are giving emphasis to low but different frequency regions. This somewhat explains the drop in r and MSE in the cross database study. The cumulative frequency responses also indicate the difference in performances between the spectral based approach and the raw waveform based approach. More precisely, log Mel frequency filter bank energies tend to characterize the spectral envelope covering the entire bandwidth. Raw waveform based approach is giving emphasis to the selective frequency region. To ascertain whether this difference impacts the performance, we trained spectral based approach systems with the first 20 log Mel filter bank energy as input with MSE loss. The mid-frequency of the 20th filter is 1949.99 Hz. This is close to the frequency region the CNNs in the raw waveform based approaches are emphasizing.

Table 8 compares the performance achieved with the first 20 log Mel filter bank energy as input (denoted as RNN-20 and CNN-20) with 40 filterbank energies as input (denoted as RNN-40 and CNN-40, results taken from Table 3 Line 1 and Line 5). We can observe that there is a clear drop in performance in both systems in terms of r and MSE. In the case of RNN architecture, there is a clear drop in performance in terms of breathing rate error, breathing event sensitivity, and tidal volume error. In the case of CNN architecture, there is not much change in breathing parameter estimation. This indicates that the spectral region modeled has an impact on the performance.

It is possible to guide the raw waveform based approach to model spectral envelop related information, similar to the spectral feature based approach. In phone classification studies, it has been consistently demonstrated that the raw waveform based approach is able to capture short-term spectral envelop information (Muckenhirn et al., 2019; Palaz et al., 2019). So, we could first pre-train the CNN on phone classification task and then adapt it for breathing signal estimation. This is part of our future work.

6.3. Comparison to other approaches

Sensing breathing signal estimation from speech signal is a relatively new problem. In that direction, the spectral based approaches dealt in this paper were the first to be explored (Nallanthighal et al., 2019, 2020). The Interspeech 2020 ComParE challenge devoted a sub-challenge on breathing signal estimation from speech signal (Schuller et al., 2020). The new methods development has happened in parallel with our work. A fair comparison to those methods based on the different metrics used in this paper is not feasible. The reason being that the sub-challenge compared the systems only based on r. Nevertheless, for the sake of completeness, we provide a comparison between the approaches investigated in this paper and the ComParE 2020 sub-challenge baseline and other systems developed as part of the challenge (Markitantov et al., 2020; Mendonça et al., 2020).

Table 9 provides the performance of various other deep learning techniques explored in this challenge, including the baseline paper (Schuller et al., 2020), winner of the challenge (Markitantov et al., 2020) as well as our proposed systems, which were tested on the Test subset of the UCL-SBM database. It is worth mentioning that our proposed systems in the table are coming from different runs during the ComParE challenge. So they are not exactly the same systems presented in Section 5.2. In the following, we denote our systems in this format: ANN type_input type_loss function. (2s) refers to 2 s long input. (4s, 1024) refers to 4 s long input with 1024 correlation window size. The system noted with (scaling) refers to the system where a mean subtraction and scaling between −1 and 1 are applied on the output of the CNN. Fusion_Raw refers to the system where the outputs of CNN_Raw_MSE and CNN_Raw_Corr are aligned through cross correlation and are averaged. Fusion_Raw_Spec refers to the system wherein the LSTM-RNN_Spec_MSE and CNN_Raw_Corr-MSE are combined with the same method as mentioned before. 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, Fusion-Raw, and Fusion_Raw_Spec yield performance competitive to the baseline CNN+LSTM RNN system. On the Test set, our approach Fusion_Raw_Spec is comparable to the End2End baseline and the BiLSTM system proposed in (Mendonça et al., 2020) and inferior to the winner of the challenge (Markitantov et al., 2020). It is worth mentioning that, in the Test set protocol, the neural networks were trained with training and development data. As a result, r is higher than the development set. As we have already observed in the different studies presented earlier, r by itself is not a complete indicator of breathing parameter estimation.

Table 9
Pearson's correlation coefficient r reported on the Dev set and the Test set. For the sake of clarity, our systems are denoted in the following format: *ANN type_input type_loss function.*

	Dev r	Test r
ComParE 2020 Breathing sub-challenge Baselines (Schuller et al., 2020)		
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 by Markitantov et al. in (Markitantov et al., 2020)		
1D CNN + LSTM (Raw signal)	0.607	0.744
ResNet18 + GRU (128 log Mel)	0.580	0.734
Fusion	**0.640**	**0.763**
Proposed Systems by Mendonça et al. in (Mendonça et al., 2020)		
BiLSTM Original	0.507	0.720
Our Proposed Systems		
CNN_Raw_MSE (2s)(scaling)	0.519	–
CNN_Raw_Corr (4s, 1024)(scaling)	0.514	–
CNN_Raw_Corr-MSE (4s, 512)(scaling)	0.532	0.628
CNN_Raw_Corr-MSE (4s, 512)	0.476	0.636
CNN_Spec_MSE	0.472	0.452
LSTM-RNN_Spec_MSE	0.448	–
Fusion_Raw	**0.552**	0.656
Fusion_Raw_Spec	0.541	**0.707**

7. Conclusion

Speech and respiration are closely related and therefore, it may be possible to use speech analysis for the sensing of the respiratory health of the speaker. In this article, we studied this interdependence and addressed the following challenges to establish respiratory analysis of speech as a practical or clinical application.

1. We explored the possibility of estimating breathing signal from the speech in two approaches: spectral analysis and raw waveform analysis, and established that the breathing signal could be reliably estimated directly from speech. We found that fusing the estimated breathing signals obtained from the best models trained on these two approaches is closer to the actual breathing signal.
2. To make the study more robust and reliable for practical or clinical purposes, we studied two significant protocols: read speech and conversational (or spontaneous) speech. We performed individual database studies and cross-database studies. The cross-database studies yielded promising performances in terms of breathing parameter estimation. This suggests that these methods can be applied independent of databases. However note that on individual databases results, the estimation is better than cross database.
3. For evaluating the estimated breathing signal from the neural network models, apart from the standard evaluation metrics for the regression problem: mean squared error and correlation with reference to actual breathing signal measured through respiratory inductive belts, we compared the breathing parameters like breathing rate, tidal volume equivalent, and breath event sensitivity. An extensive comparison in terms of these metrics was done in each protocol and reported. These parameters help in establishing the credibility for the practical application of the proposed methods.

Estimating breathing signal and parameters from the speech signal is an unobtrusive and potentially cost-effective option for long-term breathing monitoring in telehealth care applications. This technology could facilitate continuous breathing activity monitoring aiding a more thorough and adequate assessment for early recognition of abnormal breathing syndromes.

CRediT authorship contribution statement

Venkata Srikanth Nallanthighal: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft. **Zohreh Mostaani:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft. **Aki Härmä:** Conceptualization, Data curation, Formal anaylsis, Funding acquisition, Investigation, Project administration, Supervision, Writing - review & editing. **Helmer Strik:** Conceptualization, Formal analysis, Investigation, Supervision, Writing - review & editing. **Mathew Magimai-Doss:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing - review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Venkata Srikanth Nallanthighal

References

Abadi, Martín, Agarwal, Ashish, Barham, Paul, Brevdo, Eugene, Chen, Zhifeng, Citro, Craig, et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

Cummins, Nicholas, Baird, Alice, & Schuller, Björn W. (2018). Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, *151*, 41–54.

Cummins, Nicholas, Schmitt, Maximilian, Amiriparian, Shahin, Krajewski, Jarek, & Schuller, Bjorn (2017). "You sound ill, take the day off": Automatic recognition of speech affected by upper respiratory tract infection. In *2017 39th annual international conference of the IEEE engineering in medicine and biology society* (pp. 3806–3809). Seogwipo: IEEE.

Dibazar, Alireza A., Narayanan, S., & Berger, Theodore W. (2002). 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 (vol. 1)* (pp. 182–183). IEEE.

Dubagunta, S. Pavankumar, Vlasenko, Bogdan, & Magimai.-Doss, Mathew (2019). Learning voice source related information for depression detection. In *Proceedings of the IEEE international conference on acoustics, speech, and signal processing*.

Fairbanks, G. (1960). The rainbow passage. *Voice and Articulation Drillbook*, *2*.

Fu, Szu-Wei, Tsao, Yu, Lu, Xugang, & Kawai, Hisashi (2017). Raw waveform-based speech enhancement by fully convolutional networks.

Fuchs, Susanne, Reichel, Uwe D., & Rochet-Capellan, Amelie (2015). Changes in speech and breathing rate while speaking and biking. In *ICPhS 2015: 18th International Congress of Phonetic Sciences*.

Goodfellow, Ian, Bengio, Yoshua, & Courville, Aaron (2016). *Deep learning*. MIT press.

Hammarsten, Jonna, Harris, Roxanne, Henriksson, Nilla, Pano, Isabelle, Heldner, Mattias, & Włodarczak, Marcin (2015). Temporal aspects of breathing and turn-taking in Swedish multiparty conversations. In *Fonetik 2015* (pp. 47–50). Centre for Languages and Literature.

Heck, Detlef H, McAfee, Samuel S, Liu, Yu, Babajani-Feremi, Abbas, Rezaie, Roozbeh, Freeman, Walter J, et al. (2017). Breathing as a fundamental rhythm of brain function. *Frontiers in Neural Circuits*, *10*, 115.

Henderson, Alan, Goldman-Eisler, Frieda, & Skarbek, Andrew (1965). Temporal patterns of cognitive activity and breath control in speech. *Language and Speech*, *8*(4), 236–242.

Hixon, Thomas J., Mead, Jere, & Goldman, Michael D. (1976). Dynamics of the chest wall during speech production: Function of the thorax, rib cage, diaphragm, and abdomen. *Journal of Speech and Hearing Research*, *19*(2), 297–356.

Hoit, Jeannette D., & Hixon, Thomas J. (1986). Body type and speech breathing. *Journal of Speech, Language, and Hearing Research*, *29*(3), 313–324.

Hoit, Jeannette D., & Hixon, Thomas J. (1987). Age and speech breathing. *Journal of Speech, Language, and Hearing Research*, *30*(3), 351–366.

Hoit, Jeannette D, Hixon, Thomas J, Altman, Mary Ellen, & Morgan, Wayne J (1989). Speech breathing in women. *Journal of Speech, Language, and Hearing Research*, *32*(2), 353–365.

Hoit, Jeannette D., Solomon, Nancy Pearl, & Hixon, Thomas J. (1993). Effect of lung volume on voice onset time (VOT). *Journal of Speech, Language, and Hearing Research*, *36*(3), 516–520.

Huber, Jessica E., Chandrasekaran, Bharath, & Wolstencroft, John J. (2005). Changes to respiratory mechanisms during speech as a result of different cues to increase loudness. *Journal of Applied Physiology*, *98*(6), 2177–2184, PMID: 15705723.

Kabil, Selen Hande, Muckenhirn, Hannah, & Magimai-Doss, Mathew (2018). On learning to identify genders from raw speech signal using CNNs. In *Interspeech* (pp. 287–291).

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

Klatt, D. H., Stevens, K. N., & Mead, J. (1968). Studies of articulatory activity and airflow during speech*. *Annals of the New York Academy of Sciences*, *155*(1), 42–55.

Konno, K., & Mead, J. (1967). Measurement of the separate volume changes of rib cage and abdomen during breathing. *Journal of Applied Physiology*, *22*(3), 407–422, PMID: 4225383.

Koolagudi, Shashidhar G., Murthy, Y. V. Srinivasa, & Bhaskar, Siva P. (2018). Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition. *International Journal of Speech Technology*, *21*(1), 167–183.

MacLarnon, Ann, & P. Hewitt, Gwen (1999). The evolution of human speech: The role of enhanced breathing control. *American Journal of Physical Anthropology*, *109*, 341–363.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, *63*(4), 561–580.

Markitantov, Maxim, Dresvyanskiy, Denis, Mamontov, Danila, Kaya, Heysem, Minker, Wolfgang, & Karpov, Alexey (2020). Ensembling end-to-end deep models for computational paralinguistics tasks: ComParE 2020 mask and breathing sub-challenges. In *Proc. interspeech 2020* (pp. 2072–2076).

Mendonça, John, Teixeira, Francisco, Trancoso, Isabel, & Abad, Alberto (2020). Analyzing breath signals for the interspeech 2020 compare challenge. In *Proc. Interspeech 2020* (pp. 2077–2081).

Miotto, Riccardo, Wang, Fei, Wang, Shuang, Jiang, Xiaoqian, & Dudley, Joel T (2017). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246.

Mitchell, Heather L., Hoit, Jeannette D., & Watson, Peter J. (1996). Cognitive-linguistic demands and speech breathing. *Journal of Speech, Language, and Hearing Research*, *39*(1), 93–104.

Muckenhirn, Hannah, Abrol, Vinayak, Magimai-Doss, Mathew, & Marcel, Sébastien (2019). Understanding and visualizing raw waveform-based CNNs. In *Proc. interspeech 2019* (pp. 2345–2349).

Muckenhirn, Hannah, Doss, Mathew Magimai, & Marcell, Sébastien (2018). Towards directly modeling raw speech signal for speaker verification using CNNs. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 4884–4888). IEEE.

Nallanthighal, Venkata Srikanth, Härmä, Aki, & Strik, Helmer (2019). Deep sensing of breathing signal during conversational speech. In *Proc. interspeech 2019* (pp. 4110–4114).

Nallanthighal, V. S., Härmä, A., & Strik, H. (2020). Speech breathing estimation using deep learning methods. In *2020 IEEE international conference on acoustics, speech and signal processing* (pp. 1140–1144).

Oppenheim, A. V., & Schafer, R. W. (2004). From frequency to quefrency: a history of the cepstrum. *IEEE Signal Processing Magazine*, *21*(5), 95–106.

Ou, Zhijian, & Zhang, Yang (2012). Probabilistic acoustic tube: a probabilistic generative model of speech for speech analysis/synthesis. In *Artificial intelligence and statistics* (pp. 841–849). PMLR.

Palaz, Dimitri, Collobert, Ronan, & Magimai.-Doss, Mathew (2013. Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. In *Proceedings of interspeech* (pp. 1766–1770).

Palaz, Dimitri, Magimai.-Doss, Mathew, & Collobert, Ronan (2019). End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition. *Speech Communication*, *108*, 15–32.

Paszke, Adam, Gross, Sam, Massa, Francisco, Lerer, Adam, Bradbury, James, Chanan, Gregory, et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc..

Puller, Dennis (1988). Respiratory function in speech and song, by thomas j. hixon and collaborators, 433 pp, hard cover, college-hill press, Boston, Ma, 1987, $32.00. *The Laryngoscope*, *98*(6), 689.

Qi, Jun, Du, Jun, Siniscalchi, Sabato Marco, & Lee, Chin-Hui (2019). A theory on deep neural network based vector-to-vector regression with an illustration of its expressive power in speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(12), 1932–1943.

Qi, Jun, Du, Jun, Siniscalchi, Sabato Marco, Ma, Xiaoli, & Lee, Chin-Hui (2020). Analyzing upper bounds on mean absolute errors for deep neural network based vector-to-vector regression. *IEEE Transactions on Signal Processing*.

Rethage, D., Pons, J., & Serra, X. (2018). A wavenet for speech denoising. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 5069–5073).

Ruinskiy, D., & Lavner, Y. (2007). 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*, *15*(3), 838–850.

Schmidhuber, Jürgen (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Scholkmann, Felix, Boss, Jens, & Wolf, Martin (2012). An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals. *Algorithms*, *5*(4), 588–603.

Schuller, Björn W., Batliner, Anton, Bergler, Christian, Messner, Eva-Maria, Hamilton, Antonia, Amiriparian, Shahin, et al. (2020). The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks. In *Proc. Interspeech 2020* (pp. 2042–2046). Shanghai, China.

Sebastian, Jilt, Sur, Mriganka, Murthy, Hema A., & Magimai.-Doss, Mathew (2020). *Signal-to-signal neural networks for improved spike estimation from calcium imaging data*. Cold Spring Harbor Laboratory, bioRxiv.

Sejdić, Ervin, Djurović, Igor, & Jiang, Jin (2009). Time–frequency feature representation using energy concentration: An overview of recent advances. *Digital Signal Processing*, *19*(1), 153–183.

Slifka, Janet (2006). Some physiological correlates to regular and irregular phonation at the end of an utterance. *Journal of Voice*, *20*(2), 171–186.

Solomon, Nancy Pearl, & Hixon, Thomas J. (1993). Speech breathing in parkinson's disease. *Journal of Speech, Language, and Hearing Research*, *36*(2), 294–310.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185–190.

Székely, É., Henter, G. E., Beskow, J., & Gustafson, J. (2020). Breathing and speech planning in spontaneous speech synthesis. In *2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7649–7653).

Teixeira, João Paulo, Oliveira, Carla, & Lopes, Carla (2013). Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia Technology*, *9*, 1112–1122.

Von Euler, C. (1982). Some aspects of speech breathing physiology. In *Speech Motor Control* (pp. 95–103). Elsevier.

Wang, Yu-Tsai, Green, Jordan R, Nip, Ignatius SB, Kent, Ray D, & Kent, Jane Finley (2010). Breath group analysis for reading and spontaneous speech in healthy adults. *Folia Phoniatrica et Logopaedica*, *62*(6), 297–302.

Winkworth, Alison L, Davis, Pamela J, Ellis, Elizabeth, & Adams, Roger D (1994). Variability and consistency in speech breathing during reading: Lung volumes, speech intensity, and linguistic factors. *Journal of Speech, Language, and Hearing Research*, *37*(3), 535–556.

Włodarczak, Marcin, & Heldner, Mattias (2017). Respiratory constraints in verbal and non-verbal communication. *Frontiers in Psychology*, *8*.

Włodarczak, Marcin, Heldner, Mattias, & Edlund, Jens (2015). Breathing in conversation : An unwritten history. In *Linköping electronic conference proceedings*, *Proceedings of the 2nd european and the 5th nordic symposium on multimodal communication :* (110), (pp. 107–112). Stockholm University, Phonetics.

Xu, Yong, Du, Jun, Dai, Li-Rong, & Lee, Chin-Hui (2014). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(1), 7–19.

Zwald, Laurent, & Lambert-Lacroix, Sophie (2012). The berhu penalty and the grouped effect. arXiv preprint arXiv:1207.6868.