




# Analysing Breathing Patterns in Reading and Spontaneous Speech

Gauri Deshpande<sup>1,2</sup> , Björn W. Schuller<sup>2,3</sup>, Pallavi Deshpande<sup>4</sup>,  
Anuradha Rajiv Joshi<sup>5</sup>, S. K. Oza<sup>4</sup>, and Sachin Patel<sup>1</sup>

<sup>1</sup> TCS Research, Pune, India

{gauri1.d,sachin.patel}@tcs.com

<sup>2</sup> University of Augsburg, Augsburg, Germany

schuller@ieee.org

<sup>3</sup> Imperial College London, London, UK

<sup>4</sup> Bharti Vidyapeeth (DTU) College of Engineering, Pune, India

{psdeshpande,skoza}@bvucoep.edu.in

<sup>5</sup> Department of Physiology, Bharti Vidyapeeth (DTU) Medical College, Pune, India

**Abstract.** This paper focuses on the time and phase-domain analysis of speech signals to extract breathing patterns. The speech signals under investigation fall into two categories: reading and spontaneous speaking. We introduce SBreathNet, a deep Long Short-Term Memory (LSTM) based regressive model, to extract breathing patterns from speech signals. SBreathNet is trained with speech collected from 100 individuals reading a phonetically balanced text and extracts the breathing patterns with an average Pearson correlation coefficient (r-value) of 0.61 with the true breathing signal captured using a respiratory belt. The average breaths-per-minute error (BPME) across 100 speakers is 2.50. The analysis is done using leave-one-speaker-out approach. Similarly, when SBreathNet is trained with spontaneous speech signals, it extracts the breathing patterns with an r-value of 0.41 and an average BPME of 3.9. By comparing the performance across speakers, speech categories, and speech-breathing categories, we aim to uncover the factors influencing SBreathNet's effectiveness when applied to these two types of speech signals.

**Keywords:** Speech analysis · Computational paralinguistics · Speech-breathing patterns · Health informatics

## 1 Introduction

Speech signals and breathing patterns are inherently interconnected as they both rely on the use of respiratory organs. Moreover, they are both susceptible to being influenced by various psychological and physiological factors. An illustrative instance is the noticeable alteration in an individual's voice when affected by a cough or cold. Similarly, as highlighted in the works of [1, 2], breathing patterns serve as an indicator for the presence of respiratory infection, COVID-19.

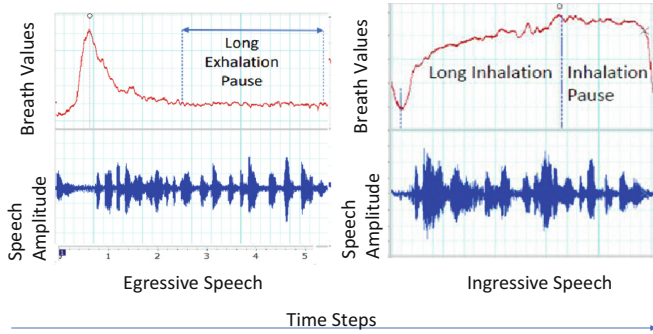
In particular, the authors of [1] reveal that breathing signals contain a greater wealth of information regarding COVID-19 infection compared to speech signals.

Among multiple ways of capturing breathing patterns, visual inspection is the simplest of all, but is prone to errors. All other techniques require a measurement instrument connected to the individual under observation. For example, in RIP, a transducer is connected over the chest area to convert the changes in lung volume into digital breathing patterns. The acquisition of such patterns to enable further analysis of the signals requires an instrument called a respiratory belt along with a data acquisition unit. As an individual needs to visit a clinic for an inspection of the breathing pattern, this is usually done only after the difficulty in breathing becomes severe. Consequently, our proposition is to utilise speech signals as a means to extract and analyse breathing patterns, enabling a deeper comprehension of an individual’s underlying physiological and psychological states. Speech signals can be conveniently captured using a smartphone microphone, even in non-clinical settings.

### 1.1 Previous Work

Various approaches have been employed to extract breathing patterns from speech signals. A common metric used to evaluate the performance of predictive models is the Pearson’s correlation coefficient, which measures the correlation between the predicted and true breathing patterns. Additionally, breathing parameters like breaths-per-minute (BPM) and tidal volume are also compared between the predicted and true patterns. Several speech features have been utilized to extract breathing patterns from speech. These include Mel Frequency Cepstral Coefficients (MFCCs), Root Mean Square Error, ZCR, and spectral slope, as described in [3]. Cepstrograms, as discussed in [4], and log mel-spectrograms, as explored in [5–8], have also been employed for this purpose. Furthermore, in their work, Nallanthighal et al. [7] have investigated the use of raw speech waveforms fed into deep neural networks.

In [5], simultaneous breathing and conversational speech is collected from 20 healthy subjects. A maximum Pearson correlation (r-value) of 0.47 is achieved with long-short term memory (LSTM) networks for a segment duration of 4 seconds (s). Further breathing parameters such as breathing rate and tidal volume are also calculated with an error rate of 4.3% and 1.8%, respectively. In [6], 40 healthy subjects’ data is analysed for the detection of breathing rate using LSTM models. The authors have compared mean squared error (MSE) with BerHu as the regression loss function (similar to Huber loss function – see [6]). They present the hypothesis that the breathing patterns have sudden peaks of inhalation followed by a gradually descending curve of exhalation which can be modelled using a BerHu loss function. They also present the results showing BerHu loss optimises the model better than MSE giving an r-value of 0.42. With the same approach, the authors of [7] have performed cross-corpus analysis and have achieved an r-value of 0.39 when training using Philips-Database and testing on the UCL-SBM database [9] and the r-value of 0.36 with the reversed datasets. The Computational Paralinguistics challenge (ComParE) organised



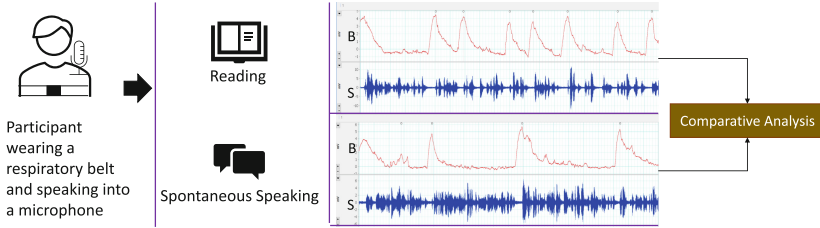
**Fig. 1.** Two broad categories of the speech-breathing patterns: speech during inhalation called ingressive and speech during exhalation called egressive speech-breathing.

at Interspeech 2020 [9] had in its Breathing Sub-Challenge a baseline Pearson correlation of  $r = 0.50$  on the development (16 speakers), and  $r = 0.73$  on the test data set (17 speakers). The winners of this challenge [10] reported  $r = 0.76$  between the breathing patterns derived from the speech signal and corresponding breathing values of the test set of 17 speakers.

Speech can be primarily categorised into two types: reading speech, which refers to speech produced while reading a paragraph, and spontaneous speech, which pertains to speech generated during natural, unplanned conversations or presentations. The breathing patterns generated during these two speech tasks differ. There have been relatively fewer comparisons made between models that extract breathing patterns from both reading and spontaneous speech. As reported in [11], spontaneous speech task exhibit grammatically inappropriate overlap of speech during exhale duration as compared to reading task. The spontaneous speech task also have longer exhalations overlapping with speech production than the passage reading task. In the study conducted in [7], a performance evaluation of deep learning models is performed on both reading and spontaneous speech tasks. The obtained  $r$ -values for the reading and spontaneous speech tasks are reported as 0.56 and 0.51 respectively. It is important to note that the speech data for these tasks is sourced from two separate databases with distinct speakers: the Philips read speech database and the UCL Speech Breath Monitoring (UCL-SBM) database, as mentioned in [9]. Analysing the differences and similarities in the captured breathing patterns of the same set of speakers between reading and spontaneous speech tasks would provide valuable insights.

The main contributions of this paper are as follows:

- We present a deep network: SBreathNet, trained with data from 100 speakers to extract breathing patterns from speech signals.
- We present additional insights with leave-one-speaker-out (LOSO) analysis on  $r$ -value and BPME metrics.



**Fig. 2.** Simultaneous speech (S) and breathing (B) is collected from the participants while they read and speak spontaneously. The speech signals are analysed for extracting breathing patterns from them. We present a comparison of the performance of the models extracting breathing patterns from reading and spontaneous speech signals.

- We augment the understanding of the speech-breathing patterns. As seen in Fig. 1, the left breathing pattern with a sudden inhalation peak followed by exhalation is called egressive speech-breathing. Here, the speech production happens during exhalation. The right side of Fig. 1 shows the breathing pattern with longer inhalation and a sudden drop during exhalation. This is called ingressive breathing pattern [12]. Here, the speech production happens during inhalation. In this paper, we introduce the impact of ingressive patterns on the model performance.
- We compare SBreathNet performance across 100 speakers, two speech categories (reading and spontaneous speaking), and two speech-breath categories: ingressesives and egressives.

## 2 Methodology

As shown in Fig. 2, speech and breathing patterns are captured simultaneously from the participants while they perform four different tasks: reading a phonetically balanced passage, speaking spontaneously, pronouncing vowels, and laughing out loudly. We present the analysis of speech-derived breathing patterns for the reading and spontaneous speaking tasks and compare their performances. This section explains the study design for capturing the data followed by speech representation techniques explored to build the model for extracting breathing patterns from speech signals.

### 2.1 Data Acquisition

ADInstruments’ respiratory belt transducer is used for recording the breathing patterns and a condenser microphone for recording the speech signals. ADInstruments PowerLab data acquisition system’s two channels are connected to these two recording devices to capture the time synchronised signals. The transducer is positioned on the chest (4 centimetres (cm) below the collarbone) and the head mounted microphone is placed at a distance of 4 cm from the mouth. A survey questionnaire is designed to capture the participants’ metadata comprising

personal and physiological information along with their anxiety level using the state and trait anxiety inventory scale. Personal information includes age group, gender, height, weight and if they have received any formal training of singing. We also ask them if they currently smoke or have smoked in the past. Physiological information includes the momentary pulse rate, and the blood pressure measured using Omron’s digital blood pressure monitoring machine.

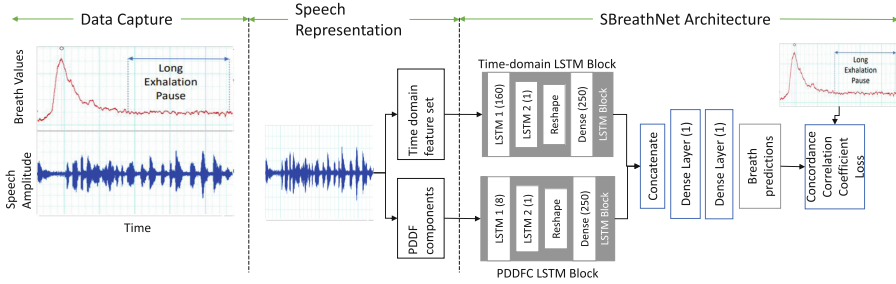
An approval from the ethical committee of Bharti Vidyapeeth Medical College is taken for execution of the data collection. An informed consent is taken from the participants for collection of data. The participants are seated in a chair and are given approximately 2 min time to relax before starting the experiment. They read the phonetically balanced sentences from the List 2, List 3, List 7, List 8, List 9 and List 10 of Harvard sentences. Harvard sentences are phonetically balanced sentences using specific phonemes at the same frequency as they appear in English [13]. Each participant took around two to three minutes to read these sentences. This activity is called as “Reading Task”. After this, the participants are asked to speak spontaneously about any topic they like. They are also given some pointers in the form of questions (such as, what are your hobbies, which is your favourite city and further on) to help them recall any incident they want to narrate. A timer of one minute is set such that they speak at-least for a minute. This is called as “Spontaneous Task”. This is followed by the “Vowels Task”, in which they pronounce five English vowels and 12 Devnagari vowels. At the end, each participant laughs out loudly (LoL) for around two to three seconds. This is called as “Vowels and LoL Task”.

## 2.2 Speech Representation Analysis

For the extraction of breathing patterns from speech, the significance of the low-level time-domain features is discussed in [14]. Using these features, an r-value of 0.57 is achieved between the speech and the predicted breathing patterns of the ComParE dataset [9]. Among other speech parameters used for understanding the respiratory problems such as COVID-19 from human voice, MFCCs and the phase-domain decomposed filter components (PDDFC) of speech signal are discussed to classify COVID-19 subjects from healthy subjects in [1]. We explore the time-domain features, MFCCs, and PDDFC for training an LSTM-based deep network, to extract the breathing patterns from the speech signals. It is observed that the combination of time-domain features along with PDDFC performs the best. Both the features are calculated for every speech frame of 20 milliseconds (ms). Time domain features form a feature vector of length 16 comprising of: ZCR, kurtosis, RMS, auto-correlation, and 10 time domain difference features [15] and PDDFC forms the feature vector of length 160.

## 2.3 SBreathNet: LSTM-Based Deep Model

As shown in the Fig. 3, the network architecture is trained using time domain features and PDDFC as input. The network is trained with a batch length of 250 corresponding to a duration of 5 s (A sample for every 20 ms is calculated, hence



**Fig. 3.** The speech signals that are captured with breathing patterns are represented using time and phase domain analysis. This representation is then fed to SBreathNet, LSTM-based deep architecture to extract breathing patterns from the speech signals.

250 samples =  $250 \times 20 \text{ ms} = 5000 \text{ ms}$ ). Both the inputs are passed separately to corresponding LSTM blocks comprising of two LSTMs and a dense layer. The outputs of these two LSTM blocks are concatenated and fed to two consecutive dense layers. This forms the output of the encoder network. The loss function calculates the concordance correlation coefficient (CCC) loss between true and predicted values. The network learns with a learning rate of 0.001 and with an Adam optimiser. The activation function of the last dense layer is the hyperbolic tangent (tanh) function. This causes the prediction values to range between  $-1$  to  $1$ . Figure 3 shows the number of nodes of each network layer in brackets.

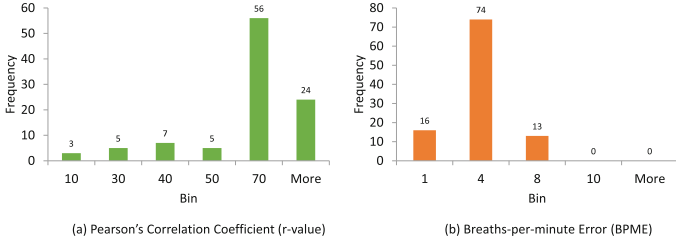
### 3 Results

With the SBreathNet model described in Sect. 2.3, the breathing patterns for both reading and spontaneous speech are extracted when trained with respective speech data. The predicted breathing patterns are then compared with the true breathing patterns captured with the respiratory belt to analyse the respective model performance. We present the separate analysis of reading and spontaneous speech, followed by comparing their performances and the challenges.

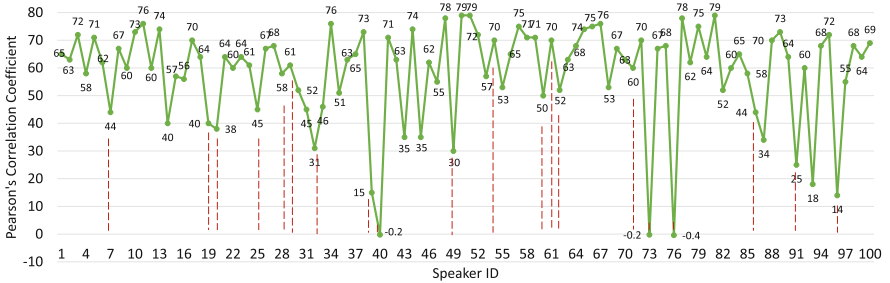
#### 3.1 Read Speech

An average r-value of 0.61 is achieved across the 100 speakers’ breathing patterns extracted from the speech signals captured while they read the phonetically balanced passage. We experimented with varying batch length values (the time-step value for the LSTM layer) of the network ranging from 1 s to 1 min to understand the impact of the time-series-encoding on the performance. The overall performance is achieved the best for 5 s based analysis. As seen in Fig. 4 (a), the number of speakers having an r-value above 0.50 is 80.

The BPME count for every speaker is calculated on the predictions obtained and compared with that of the true breathing pattern. The peak detection algorithm from scipy [16] is used for the detection of peaks keeping a distance as 100



**Fig. 4.** (a) Number of speakers belonging to six bins of r-value performance of the SBreathNet model trained with the speech captured during reading task. (b) Number of speakers belonging to five bins of breaths-per-minute error calculated between the true breathing patterns and the breathing patterns predicted using SBreathNet model for the reading task.



**Fig. 5.** Pearson's correlation coefficient is calculated between the true breathing patterns and the breathing patterns extracted using SBreathNet for the reading task. The predictions are obtained using Leave-one-speaker-out analysis.

points and a height as 0.2. Using the peak count, further, BPME is calculated for each speaker. An average BPME obtained is 2.50. Figure 4 (b) shows that 90% of the speakers have BMPE less than 4. Hence, SBreathNet can extract breathing patterns with an r-value above 0.50 for 80% speakers and a BPME below 4 for 90% speakers.

**Leave One Speaker Out Analysis:** Figure 5 visualises the LOSO performance of the SBreathNet architecture. As seen in the Figure, three speaker IDs: 40, 73, and 76 consistently have a negative r-value. As described before, varying batch-lengths from 1 s to 60 s are explored; also regularisation techniques are explored, however, the performance for these three speakers remains unchanged. The speaker-wise BPME for the 100 speakers for the predictions obtained using SBreathNet ranges between 0.3 to 7.5. It is observed that the change in BPME across the speakers is not synchronised with the r-value exhibited by them. This can be seen in the case of speakers with negative r-value of  $-0.40$  and  $-0.21$ , who have relatively low BPME of 3 and 2.1, respectively. This shows that SBreathNet captures the breathing event equally well for speakers with low r-value.

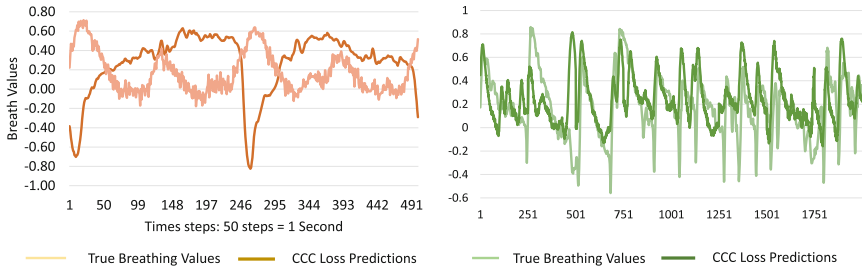
**Table 1.** Number of speakers belonging to each breathing pattern cluster and their corresponding performances. The performance is reported using r-value, BPME, and Centroid-R between the true and the predicted values.

Cluster	Speakers	R	BPME	Centroid-R
0	24	0.60	3.6	0.80
1	20	0.37	1.8	-0.30
2	16	0.68	2.4	0.74
3	26	0.66	2.2	0.68
4	14	0.65	2.6	0.90

**Clustering on True Breathing Patterns:** To further understand the performance exhibited in LOSO analysis, the true breathing patterns are studied using clustering technique. It is observed that an average breathing cycle duration lasts for around five seconds while a speaker reads loudly. The breathing patterns of the read-task are captured continuously for around 3–4 min from each speaker. Such breathing patterns are segmented into smaller breathlets of 5 s each giving around 35–45 such breathlets per speaker. With each breathlet as a data point, the elbow method indicates that five distinct clusters can be formed using a k-means clustering algorithm. On clustering the breathlets using k-means with random\_state defined as zero, the cluster centres show that four of the clusters represent four distinct locations of the inhalation peak in the five seconds duration. These locations are: 1) within first second, 2) between 2–4 s, 3) between 4–5 s, and 4) towards the end of the 5 s. For all these four clusters, the speakers speak during exhalation, hence these four clusters represent the egressive speech-breathing. The fifth cluster represents an inhalation that starts from the first second and the inhalation-pause lasts until five seconds. The speakers belonging to this cluster speak during inhalation, hence, this cluster represents the ingressive speech-breathing. This observation indicates that there are two broad categories of breathing data: ingressive and egressive as shown in Fig. 1. The red dotted lines in Fig. 5 are put against the speaker IDs that belong to cluster 1 and hence are ingressive speakers. It is observed that, the 14 out of 20 (70%) of the speakers exhibiting r-value below 0.50 (low-performers) are ingressive. This contributes to 70% of the total ingressive speakers. These results suggest that, ingressiveness considerably impacts the model performance.

Table 1 explains the average r-value (R) for the five clusters showing the least performance from ingressive cluster; cluster 1. The BPME for the five clusters is as given in Table 1. Once again, we observe a lack of synchronisation between the BPME and the r-values within the cluster. This is particularly evident in cluster 1, which exhibits the lowest average BPME of 1.8. Table 1 also provides an r-value between the mean 5 s breathlet of the five predicted clusters with the corresponding true ones (Centroid-R). For the four egressive clusters, the mean breathlets overlap well with the true ones.





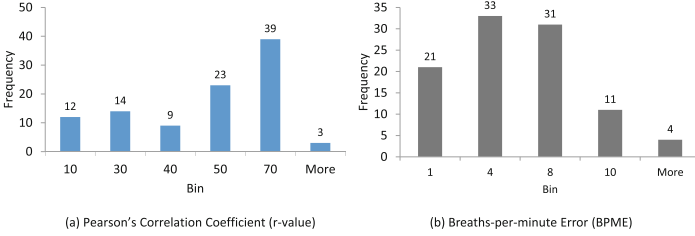
**Fig. 6.** Breathing predictions for speaker identity 76 (a) and 93 (b). 76 is ingressive and 93 is egressive speaker having an r-value of  $-0.40$  and  $0.21$  respectively.

**Ingressives and Egressives:** The average r-value of egressive speaker clusters (1, 3, 4, and 5) is  $0.65$  and that of ingressive speaker cluster is  $0.37$  using SBreathNet predictions. From the predicted breathing patterns of SBreathNet, it is observed that the ingressive pattern is apparent in four of the lowest performing ingressive speakers with the speaker IDs 40, 73, 76, and 96. Figure 6 (left) shows the 10 s prediction for speaker 76. As seen in the Figure, the breathing events are correctly identified resulting in predicting the BPME of only 1.2. However, the breathing pattern is inverted such that the inhalation and inhalation pause exhibited by true breathing patterns are not captured by the predictions. Instead, the predictions show an expiration for the corresponding time slot. This explains the absence of synchronisation between the r-value and the BPME across the speakers.

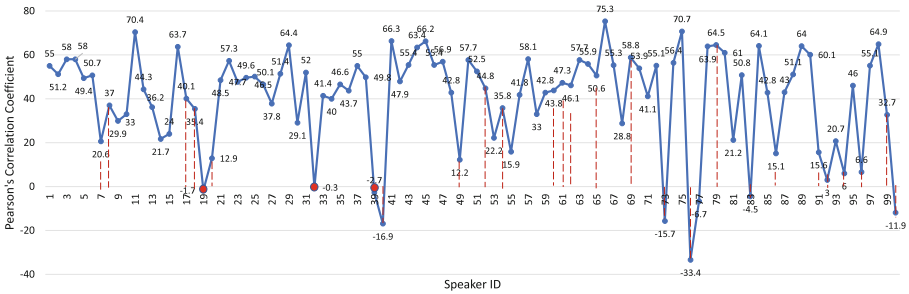
With the proposed model, 6 egressive speakers have a low performance such as speaker ID 93, who has an r-value of  $0.21$ . As seen in Fig. 6 (right), for the 20 s predictions of speaker 93, the peaks are correctly matched as well as the shape. However, the valleys are not matching between the predicted and true values. This is seen when the speakers exhale breath to a large extent resulting in deep valleys. Since the sound of such exhalation activity is not captured in speech or voice, it becomes difficult to trace them. This underscores the second challenge encountered by SBreathNet, which involves identifying the valleys of breathing patterns from speech signals.

### 3.2 Spontaneous Speech

This section describes the analysis being performed on the speech signals captured while the participants speak spontaneously. The results are obtained using the SBreathNet model trained on the spontaneous speech signal. The calculation of metrics such as the r-value and BPME remains the same as done for read speech. The average r-value, representing the breathing patterns extracted from the speech signals of the same 100 speakers during spontaneous speech, is found to be  $0.41$ . As illustrated in Fig. 7 (a), 42% of the speakers exhibit an r-value above  $0.50$ , while 65% of them have an r-value exceeding  $0.40$ . By experimenting with batch lengths ranging from 1 s to 60 s, we determine that the analysis based



**Fig. 7.** (a) Number of speakers belonging to six bins of r-value performance of the SBreathNet model trained with the speech captured during spontaneous speaking task. (b) Number of speakers belonging to five bins of breaths-per-minute error calculated between the true breathing patterns and the breathing patterns predicted using SBreathNet model for the spontaneous task.



**Fig. 8.** Pearson's correlation coefficient calculated between the true breathing patterns and the breathing patterns extracted using SBreathNet for the spontaneous task. The predictions are obtained using Leave-one-speaker-out analysis.

on a 5 s batch length produces the most optimal outcome. Figure 7 (b) displays the distribution of BPME across 5 bins. The data reveals that 54% of speakers have a BPME value below 4, whereas 15% of them exhibit a high BPME value surpassing 10.

**Leave One Speaker Out Analysis:** Based on the LOSO analysis, it is evident that there are eight speakers displaying negative r-values, ranging from  $-0.3$  to  $-33.4$ . Speaker ID 66 exhibits the highest correlation of 75.3%. When adjusting the batch length, the performance of the speakers with negative correlation varies, but consistently remains negative. Furthermore, the varying batch length not only impacts individual speaker performance but also affects the overall performance. Figure 8 illustrates the performance obtained using a batch length of 5 s, which yields the best overall results.

**Clustering on True Breathing Patterns:** To gain deeper insights into the LOSO performance, the true breathing patterns in spontaneous speech are

**Table 2.** Number of speakers belonging to each breathing pattern cluster and their corresponding performances. The performance is reported using r-value, BPME, and Centroid-R between the true and the predicted values of the spontaneous task.

Cluster	Speakers	R	BPME	Centroid-R
0	32	0.40	5.7	0.41
1	27	0.20	1.9	-0.27
2	17	0.50	4.0	0.71
3	12	0.54	4.2	0.56
4	12	0.56	4.2	0.19

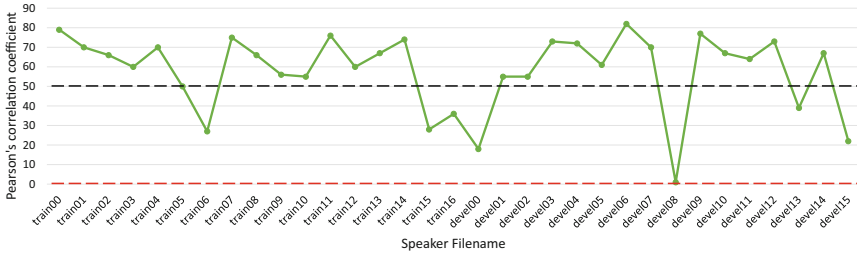
divided into breathlets of 5s each. This segmentation follows a similar approach to the analysis of breathing in read speech. When considering these actual 5s breathlets in spontaneous speech as individual data points, the elbow method suggests the presence of five clusters within the data. By utilising the k-means clustering algorithm, these five clusters are successfully identified and formed.

The two main categories of breathing patterns observed in Fig. 1 are also evident within the 5 clusters of spontaneous speech. Among these clusters, one exhibits an ingressive breathing pattern, while the remaining four demonstrate an egressive breathing pattern. Notably, all the eight speakers displaying negative correlation belong to the ingressive cluster. Among the 100 speakers, a total of 27 individuals exhibit ingressive breathing patterns. Within this group, 20 speakers (74%) possess an r-value lower than 0.40, indicating a weaker correlation. To highlight this, red dotted lines are delineated in Fig. 8 to represent the speakers belonging to the ingressive cluster, Cluster 1.

In addition, the performance of the SBreathNet model across the five clusters is illustrated in Table 2. Observing the table, it is evident that cluster 1 exhibits the lowest performance with an average correlation of 0.20. Nevertheless, it is noteworthy that the BPME for this cluster is also the lowest, suggesting accurate identification of breath events within this cluster. Furthermore, the Centroid-R value for the ingressive cluster is negative, indicating a negative correlation. Regarding the egressive clusters, Cluster 3 and Cluster 4 exhibit similar average r-values (R) compared to the other two egressive clusters. However, it is important to highlight that cluster 4 demonstrates notably low correlation in terms of mean breathlet (Centroid-R).

**Ingressives and Egressives:** According to the findings presented in Table 2, the average correlation of ingressive cluster is 0.20, while that of egressive clusters is 0.48. These results once again highlight the influence of ingressiveness on the model’s performance. However, even within egressive speakers, there are other factors that contribute to a decrease in the model’s performance. As previously discussed in the context of read speech, the presence of breathing valleys leads to a reduction in correlation since they are not captured in speech. This observation holds true for spontaneous speech as well. Additionally, spontaneous speech

presents the additional challenge of random breath events, which can result in a loss of synchronisation between speech signals and breathing patterns across both the speech-breathing categories of egressives and ingressives.



**Fig. 9.** Leave one speaker out performance of the deep LSTM model SBreathNet on the ComPaRe dataset.

### 3.3 Observations on ComPaRe Dataset

As described in [9], the authors have captured the speech and breathing patterns from 49 subjects following a similar protocol as described in Sect. 2.1. The authors mention that spontaneous speech is captured for 4 min per speaker. Training, validation and test partitions of the total data is provided in the challenge organised by the authors. The three partitions have 17, 16, and 17 speakers’ data respectively. For the presented discussion, we analyse the training and validation partition data. Figure 9 shows the speaker-wise performance obtained with SBreathNet architecture in the LOSO analysis. We used the same features as defined in Sect. 2.2 along with SBreahNet for this dataset. It is observed from the LOSO analysis of the 33 speakers that 6 speakers have r-value below 0.50. Only one speaker has the r-value as 0.0. On further analysis, it is seen that the breathing patterns of these six speakers follow ingressive pattern. Empirically, they have breathing values above the average value for more than half of the 5 s duration. The average r-value for egressive speakers is 0.67 and for ingressive speakers is 0.24.

The overall performance exhibited by SBreathNet is an r-value of 0.58 on the development partition of ComPaRe challenge dataset, which is comparable to the winners of this challenge [10] (0.64). Moreover, SBreathNet has 40K parameters as compared to 1.4M and 3.5M parameters of the models discussed in [10].

## 4 Discussion

In Sect. 3, the results obtained using time and phase domain features with SBreathNet architecture are presented for both read and spontaneous speech. It is observed that the performance for spontaneous speech is 0.2 lower compared to that of read speech. The correlation (r-value) between the predictions of read

and spontaneous speech across the 100 speakers is 0.59. This suggests that the predicted breathing patterns for the reading and spontaneous tasks of the same speaker have a similarity index of 59%.

Both categories of speech face similar challenges, including ingressiveness and deep valleys, which have a negative impact on the model’s performance. Speakers who exhibit ingressiveness during reading also tend to exhibit the same breathing pattern during spontaneous speech. However, it is not mandatory for the vice-versa to hold true. For instance, speakers with identity 9 and 100 show ingressive patterns while speaking spontaneously but not while reading. Among all the speakers, a total of 14 individuals showcase superior performance in terms of spontaneous speech compared to their performance in read speech. Likewise, 86 speakers demonstrate higher performance in the read speech task.

It is observed from the results that extracting breathing patterns for ingressive speech is difficult. To collect more data belonging to ingressive class, we need to understand such speaker characteristics. We asked further questions on physiological and psychological states to the speakers exhibiting ingressiveness in both reading and spontaneous task. The questions about their involvement in sports, yoga, swimming, if they were infected by COVID-19, about respiratory disorder in their family, the sleep quality, and their metabolic, physical and mental health were asked. We also discussed if they find themselves introvert, if they have stage fear and hence practise talking. None of the conditions are found uniform across all the speakers. For all of them, neither they nor anyone in their family have any respiratory disorders. 9 out of 20 common ingressive speakers reported that they are actively involved in sports activities related to athletics. 3 of them were infected by mild COVID-19 and were asymptomatic. The three ingressive speakers whose r-value is found negative in both reading and spontaneous task reported that they are introverts and had stage fear. They have practised speaking skills. This observation matches with the case study performed in [17]. The authors have found that a subject has used inspiratory speech for 6 years as a means of overcoming the communication problems of long-standing adductor spastic dysphonia. These observations show that not only physiological, but behavioural parameters also impact the breathing patterns of an individual.

## 5 Conclusion and Future Work

In this paper, we presented the novel SBreathNet architecture consuming time and phase domain features for the extraction of breathing patterns from speech signals during reading and spontaneous tasks. We performed LOSO analysis to understand the r-value between the predicted and the true breathing patterns for each speaker. The speaker-wise analysis helps in understanding the performance variation across speakers. This also reveals the impact of ingressiveness on the model performance. These observations are not evident from the overall performance of the model. We conclude that LOSO analysis is a strong analysis technique to understand the performance better and identify the challenges in

extracting breathing patterns from the speech signals. We have presented the impact of the ingressive speech on the model’s performance in extracting the breathing patterns accurately. Hence, in future work, we will have a focus on collecting more data and identifying ingressive speech.

During our discussion, we explored the performance of SBreathNet on two speech categories: reading and spontaneous speaking, involving a group of 100 speakers. It was evident that the model achieved superior results in reading speech compared to spontaneous speech. However, both categories encountered common challenges, such as ingressiveness and deep valleys in the breathing patterns. Moreover, the spontaneous speech category presented an additional obstacle: the randomness of breath events, including the start of inhalation and exhalation. We plan to extend our analysis to address these challenges in our future work.

## References

1. Deshpande, G., Schuller, B.W.: Covid-19 biomarkers in speech: on source and filter components. In: 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 800–803. IEEE (2021)
2. Brown, C., et al.: Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data. arXiv preprint [arXiv:2006.05919](https://arxiv.org/abs/2006.05919) (2020)
3. Ruinskiy, D., Lavner, Y.: An effective algorithm for automatic detection and exact demarcation of breath sounds in speech and song signals. *IEEE Trans. Audio Speech Lang. Process.* **15**(3), 838–850 (2007)
4. Routray, A.: Automatic measurement of speech breathing rate. In: Proceedings of the 27th European Signal Processing Conference (EUSIPCO), pp. 1–5. IEEE, A Coruña, Spain (2019)
5. Nallanthighal, V.S., Strik, H.: Deep sensing of breathing signal during conversational speech. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association, pp. 4110–4114. ISCA, Graz, Austria (2019)
6. Nallanthighal, V.S., Härmä, A., Strik, H.: Speech breathing estimation using deep learning methods. In: Proceedings of the 45th International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1140–1144. IEEE, Barcelona, Spain (2020)
7. Nallanthighal, V.S., Mostaani, Z., Härmä, A., Strik, H., Magimai-Doss, M.: Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings. *Neural Netw.* **141**, 211–224 (2021)
8. Mostaani, Z., Nallanthighal, V.S., Härmä, A., Strik, H., Magimai-Doss, M.: On the relationship between speech-based breathing signal prediction evaluation measures and breathing parameters estimation. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1345–1349. IEEE (2021)
9. Schuller, B.W., et al.: The INTERSPEECH computational paralinguistics challenge: elderly emotion, breathing & masks. In: Proceedings of the 21st Annual Conference of the International Speech Communication Association, pp. 2042–2046. ISCA, Shanghai, China (2020)

10. Markitantov, M., Dresvyanskiy, D., Mamontov, D., Kaya, H., Minker, W., Karpov, A.: Ensembling end-to-end deep models for computational paralinguistics tasks: compare 2020 mask and breathing sub-challenges. In: Proceedings of the 21st Annual Conference of the International Speech Communication Association, pp. 2072–2076. ISCA, Shanghai, China (2020)
11. Wang, Y.T., Green, J.R., Nip, I.S., Kent, R.D., Kent, J.F.: Breath group analysis for reading and spontaneous speech in healthy adults. *Folia Phoniatr. Logop.* **62**(6), 297–302 (2010)
12. Eklund, R.: Pulmonic ingressive phonation: diachronic and synchronic characteristics, distribution and function in animal and human sound production and in human speech. *J. Int. Phon. Assoc.* **38**(3), 235–324 (2008)
13. Rothauser, E.: IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust.* **17**, 225–246 (1969)
14. Deshpande, G., Schuller, B.W.: The DiCOVA 2021 challenge—an encoder-decoder approach for covid-19 recognition from coughing audio. In: Proceedings of the 21st Annual Conference of the International Speech Communication Association, pp. 931–935. ISCA (2021)
15. Deshpande, G., Viraraghavan, V.S., Gavas, R.: A successive difference feature for detecting emotional valence from speech. In: Proceedings of SMM19, Workshop on Speech, Music and Mind, vol. 2019, pp. 36–40 (2019)
16. Virtanen, P., et al.: Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods* **17**(3), 261–272 (2020)
17. Harrison, G.A., Troughear, R.H., Davis, P.J., Winkworth, A.L.: Inspiratory speech as a management option for spastic dysphonia: case study. *Ann. Otol. Rhinol. Laryngol.* **101**(5), 375–382 (1992)