Automatic Breathing Pattern Analysis from Reading-Speech Signals

Gauri Deshpande[1], Björn W. Schuller[2], Pallavi Deshpande[3], and Anuradha Rajiv Joshi[4]

***Abstract*— As the speech production mechanism is related to the breathing process, speech signals and breathing patterns impact each other. Breathing patterns are the physiological signals which help in understanding the psychological, physiological and cognitive states of an individual. Capturing such patterns relies on the availability of equipment such as respiratory belts, which are costly and uncomfortable to wear for long duration. In this paper, we attempt to extract the breathing patterns from speech signals, which are easily available and can be recorded using a smartphone's microphone. In the presented work, simultaneous speech and breath signals are captured from 100 Indians of the age group 20 to 25 years while they read a phonetically balanced passage in English language. We have identified five distinct breathing templates; following two broad speech-breath categories, exhibited by the speakers while they read the same passage. For one of the two categories, the time domain features with regression network can extract the breathing patterns from speech with a Pearson correlation coefficient of 0.70. By computational modelling, we distinguish these two breathing categories from speech with a classification accuracy of 79 %.**

I. INTRODUCTION

The kinematics and the physiology of the lung-thorax unit impacts breathing and the speech production mechanism of an individual. As explained in [1], the breathing patterns are an outcome of balancing the active forces generated by the respiratory muscles with the passive recoil forces generated by the lung-thorax unit. The main role of the respiratory system in speech is to provide the correct pressure drive to the larynx. As shown in Figure 1, a normal breathing cycle comprises of a rising curve reaching a peak value called inhalation, followed by an optional inspiratory-pause where the breath values remain almost at the peak value. The downward slope reaching to the minima indicates the exhalation phase followed by an optional expiratory-pause, where the breath values remain around the minimum value.

Winkworth et al. [2] discussed the inter-subject and intra-subject variability of lung volumes, speech intensity and linguistic influences of six healthy young women over 7 to 10 sessions of reading. However, they observed consistency in the speech-locations of inhalation, which they found to be correlated with the loudness and paragraph boundaries. Studies such as [3] explore the positive correlation between the depth of inhalation and duration of the following utterance. Similarly, the study with 20 subjects presented in [4] states that the inhalation time is the most consistent and sensitive measure for discriminating quite breathing from speech breathing. However, others such as [5] have found that the breath regulating mechanism is independent of phonation and pausing of speech of only two subjects in the study. Another view presented in [6] states that there exists a two-way relationship between both speech and respiratory signals impacting each other. Orlikoff et al. in [7] have compared the fundamental frequency (F0), electroglottographic and airflow measures of the phonation during inhalation and exhalation from 16 healthy men and women. They recorded 48.5 % higher airflow rate, 5.1 semitones higher F0 and higher EGG amplitude perturbation during inspiratory phonation. Nallanthighal et al. [8] studied the respiratory effort required for different phonemes and a group of phonemes by measuring the lung volume change and lung volume change rate. This study was conducted to establish the relationship between linguistic contents and respiratory effort, and hence, to demonstrate the possibility of extracting breathing patterns from speech signals.

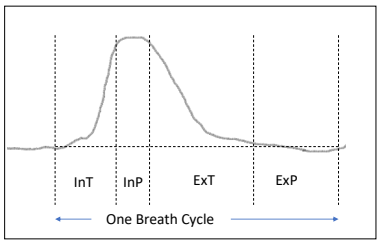

Fig. 1. A normal breathing cycle comprising of InT: Inhale Duration; InP: Inspiration Pause; ExT: Exhale Duration and ExP: Expiration Pause.

II. PREVIOUS WORK

Extraction of breathing patterns from the speech signal requires appropriate speech representation. The speech features used for this task include mel-frequency cepstral coefficients (MFCCs), energy, zero-crossing rate and spectral slope in [9], Cepstrograms in [10], and log mel-spectrograms in [11], [12], [13], [14]. The authors of [13] have also explored the use of the raw speech waveform fed to a deep network.

In [9], the breathing and speech data of 24 minutes with around 300 breathing events were collected from 14 singers. They define a breathing event as a segment present between

[1]Gauri Deshpande is working as Senior Scientist at TCS Research Pune, India and, pursuing PhD from University of Augsburg, Germany under the guidance of Prof. Björn W. Schuller. gauri1.d@tcs.com

[2]Björn W. Schuller is full professor and head of the chair of Embedded Intelligence for Health Care and Wellbeing, at the University of Augsburg, Germany. He is full professor of Artificial Intelligence and Head of GLAM - Group on Language, Audio & Music, at Imperial College London. schuller@ieee.org

[3]Pallavi Deshpande is full professor, at the Bharti Vidyapeeth (DTU) College of Engineering Pune, India. psdeshpane@bvucoep.edu.in

[4]Anuradha Rajiv Joshi is full professor, at the Bharti Vidyapeeth (DTU) Medical College, Department of Physiology Pune, India. anuradhajoshi30@gmail.com

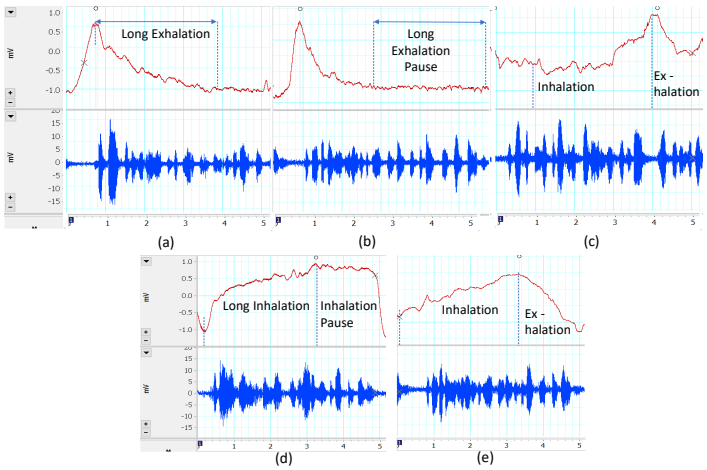

Fig. 2. Five Breathlets: a) Speech during a long exhalation period; b) Speech during exhalation and expiratory-pause; c) Speech during inhaling and exhaling in short duration and short amplitude; d) Speech during inhalation and inspiratory-pause; e) Speech during inhalation, reaching inhalation peak and continuing during exhalation. The x-axis represents time steps in seconds and y-axis represents breathing values (red) and speech amplitudes (blue).

two consecutive speech segments. The authors adopted the template matching algorithm for the detection of a breathing event followed by an edge detection algorithm for the identification of the breathing peak. Here, they have assumed the breathing signals to have static pre-defined templates. Similarly, in [10], after SVM based classification of breath events, the breath events are appropriately joined together and validated against the manual observations through listening audio and viewing thermal videos. In [11], simultaneous breathing and speech (spontaneous conversation and reading a phonetically balanced paragraph) 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 [12], 40 healthy subjects' data is analysed for the detection of breathing rate using LSTM models. The authors have compared MSE with BerHu as the regression loss function. 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 optimizes the model better than MSE giving an r-value of 0.42. With the same approach, the authors of [13] 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 [15] and the r-value of 0.36 with the reversed datasets. The Computational Paralinguistics challenge (ComParE) organised at Inerspeech 2020 [15] had a baseline Pearson correlation of r = 0.507 on the development, and r = 0.731 on the test data set. The winners of this challenge [16], reported r = 0.763 between the speech signal and corresponding breathing values of the test set.

III. Methodology

It had been a general assumption that the speech-breathing pattern of an individual follows a pre-defined template of sudden rise till the peak during inhalation and a gradually descending curve during exhalation. With the data from 100 individuals, we have an additional observation regarding the breathing patterns of the participants. This observation helps in understanding the ground truth in a better way and hence, enhancing the analysis. Further, with the focus on extracting these breathing patterns from the speech signal, it is important to understand the distinct templates of breathing co-occurring during speech production. We start with explaining the data collection protocol, followed by metadata details, pre-processing steps and prominent observations of the breathing patterns.

A. 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 synchronized 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 approximately 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 (STAI-6) 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 minutes 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 [17]. 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 atleast 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".

B. Data Details

1) Participant Details: Healthy individuals of the age group 18 to 23 years participated in the study. We collected data from 31 female and 69 male participants. All of them confirmed the absence of respiratory disorders such as COPD and asthma. Two of them have received formal training of singing and nine reported that they have either smoked in the past or currently smoke. The average height and weight of female participants is recorded as 160 cm (149 cm – 173 cm) and 53 kg (40 kg – 75 kg), respectively. For male participants, the average height of 170 cm (155 cm – 180 cm) and weight of 65 kg (50 kg – 98 kg) is recorded. The instantaneous pulse is found to range from 52 to 128.

2) Breathing Signal Analysis: Every speaker takes around four minutes to read the given passage and produces speech segments of a duration ranging from one second to seven seconds. We define a speech segment as the speech signal starting from a speech pause and ending at the start of next speech pause, where each pause has a duration of at least 200 ms. It is observed from the 100 participants' data that their breathing patterns vary significantly while they read the same passage. To statistically understand these differences, we extracted five parameters: a) Number of breath cycles while reading the entire passage, b) Inhalation time, c) Inspiratory-pause duration, d) Exhalation time and e) Expiratory-pause duration corresponding to the speech segments.

Based on these five parameters, we present corresponding five categories of a breathing cycle during read-speech in Figure 2. The figure depicts the five distinct Breathlets identified from the breathing patterns of 100 participants data; each Breathlet has 250 samples which corresponds to 5 s. Breathlet –a– represents the well known category of a breathing cycle in which the inhalation starts during speech pause, reaches the peak in a short time and the speech is produced during exhalation. Breathlet –b– is similar to –a– with the difference that the speech production happens during the expiratory pause as well. Breathlet –c– represents the random nature of a breathing curve with shorter amplitude range and longer breathing cycle. Breathlet –d– shows that the inhalation starts during the speech pause, however, the speech production happens during inhalation. Such Breathlets have long inhalation durations. The speaker continues to speak during the inspiratory pause period and has a quick exhalation. Breathlet –e– shows a similar duration for inhalation and exhalation. Also, the speech is produced during both, inhalation and exhalation.

TABLE I
DETAILS OF THE BREATHING CYCLE CATEGORIES AND THE NUMBER OF SPEAKERS BELONGING TO EACH CATEGORY.

#	Description	# Speakers
1	Short inhalation, long exhalation.	39
2	Short inhalation, moderate exhalation, long expiratory-pause.	41
3	Random inhalation and exhalation duration.	8
4	Long inhalation or inspiratory pause with short exhalation.	9
5	Similar inhalation and exhalation time.	3

Using KMeans algorithm on every five seconds of Breathing signal, five breathing pattern clusters are generated. With further analysis of the mean and variance of each cluster, we identified the Breathlets shown in Figure 2. It is observed that, each speaker can have one or more types of Breathlets appearing in their breathing pattern. Depending on the majority occurrence of the Breathlet type, we identified the number of speakers belonging to each of the five types of Breathlets. Table I explains each breath category with its description and provides the number of speakers belonging to each read-speech breath category. Note that, three of the nine participants who reported that they either smoked in the past or currently smoke, have Breathlets –a–, the other three have Breathlets –b– and the remaining three have one each of Breathlet –c–, –d– and –e–.

We also analyse the 33 speakers' data of the training and development partitions released at the ComParE challenge, Interspeech 2020 [15]. The Breathlet types defined in Figure 2 are present in this dataset as well. Six speakers' ('devel_00', 'devel_08', 'devel_13', 'devel_15', 'train_06', 'train_15') breathing patterns comprise of Breathlet types –c–, –d–, and –e– and 27 speakers' breathing patterns comprise of Breathlets –a– and –b–. Figure 3 shows sample breathing cycles from speakers 0 and 8 of the development partition, where the speech production is coinciding with inhalation.

C. Data Preprocesing

Both the signals, breathing and speech, are sampled at 40 kHz; speech is downsampled to 16 kHz and breathing to 50 Hz. Breathing values for every recording are divided by the maximum value to scale them in the range of -1 to $+1$.

1) Regression model to extract breathing patterns from Speech: We use a time-domain feature set for the extraction of breathing patterns from the speech signal. The feature set comprises of 16 frame-level features: root mean square, zero crossing rate, auto-correlation, kurtosis, and 10 time-domain-difference features [18]. The features are fed to a neural network with two LSTM layers, with eight and one nodes respectively. The LSTM network uses a custom loss function to calculate the Pearson correlation coefficient loss

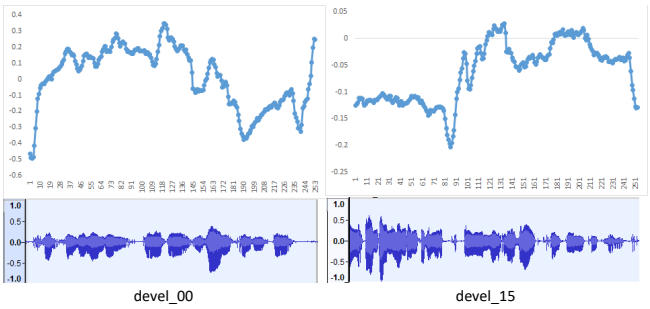

Fig. 3. Sample ingressive breathlets from the speakers' of the ComPaRe challenge dataset. Speech is produced during inhalation for these speakers.

value. We use Adam optimizer with a learning rate of 0.001 and a 'tanh' activation function at the last dense layer of the network.

2) Classifying Broad Categories of Breathing Patterns: For further analysis, we combine the Breathlets –a– and –b– in category 1 as they exhibit the similarity that speech production happens during exhalation. Similarly, Breathlets –c–, –d– and –e– are combined to form category 2. With the hypothesis that the mechanism of extracting the breathing patterns would differ for these two broad categories, we present an approach of classifying corresponding speech signals. 40 MFCCs are calculated for the speech signal of five seconds with hop-length of 250 speech samples. MFCCs are fed to an LSTM network for the binary classification between category 1 and 2.

IV. Results

We performed Leave One Speaker Out (LOSO) analysis for the speakers belonging to Breathlets –a– and –b– together comprising of 80 speakers. The average r-value for the validation set speaker is 0.65 (0.60 - 0.88) and the train set is 0.72 (0.60 - 0.80). The combined analysis of Breathlets –c–, –d– and –e– having 20 speakers with the same 16 features and the LSTM network yields an average r-value of 0.28 (0.00 - 0.40). For the ComParE dataset, with the same 16 features and the LSTM model, the 27 speakers having Breathlets –a– and –b– yields an average r-value of 0.68 (0.50 - 78). However, the six speakers' having Breathlets –c–, –d– and –e– gives an r-value of 0.00. Since the category 2 has only 20 subjects' data for Indian population and only 6 speakers from the ComParE challenge dataset, which is considerably less than the speaker count for category 1, we could not test the classification model on a considerable dataset. However, the early observations show that MFCCs are capable of classifying the two categories (both the categories are balanced) with a training accuracy of 0.79. More data and analysis is required to conclude on this.

V. Conclusion

It is important to understand the variations exhibited by the breathing patterns during read-speech. For the category 1 breathing patterns, where the speech is produced during exhalation, a simple two layered LSTM network along with time-domain features can extract the breathing patterns from the speech signal with an average r-value of around 0.70. However, the time-domain analysis is not suitable for category 2 Breathlets. In future work, we intend to understand the Breathlet categories better and classify them accurately using deep learning techniques.

References

[1] J. E. Huber and E. T. Stathopoulos, "Speech breathing across the life span and in disease," *The Handbook of Speech Production*, pp. 11–33, 2015.

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

[3] D. H. Whalen and J. M. Kinsella-Shaw, "Exploring the relationship of inspiration duration to utterance duration," *Phonetica*, vol. 54, no. 3-4, pp. 138–152, 1997.

[4] D. H. McFarland, "Respiratory markers of conversational interaction," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 1, pp. 127–128, 2001.

[5] D. Autesserre, Y. Nishinuma, and I. Guaitella, "Breathing, pausing, and speaking in dialogue." in *EUROSPEECH*, 1989, pp. 2433–2436.

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

[7] R. F. Orlikoff, R. J. Baken, and D. H. Kraus, "Acoustic and physiologic characteristics of inspiratory phonation," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1838–1845, 1997.

[8] V. S. Nallanthighal, A. Härmä, H. Strik, and M. M. Doss, "Phoneme based respiratory analysis of read speech," in *29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 191–195.

[9] D. Ruinskiy and Y. Lavner, "An effective algorithm for automatic detection and exact demarcation of breath sounds in speech and song signals," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 838–850, 2007.

[10] A. Routray and M. I. Y. Arafath K., "Automatic measurement of speech breathing rate," in *Proceedings of the 27th European Signal Processing Conference (EUSIPCO)*. A Coruña, Spain: IEEE, 2019, pp. 1–5.

[11] V. S. Nallanthighal and H. Strik, "Deep sensing of breathing signal during conversational speech," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association*. Graz, Austria: INTERSPEECH, 2019, pp. 4110–4114.

[12] V. S. Nallanthighal, A. Härmä, and H. Strik, "Speech breathing estimation using deep learning methods," in *Proceedings of the 45th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Barcelona, Spain: IEEE, 2020, pp. 1140–1144.

[13] V. S. Nallanthighal, Z. Mostaani, A. Härmä, H. Strik, and M. Magimai-Doss, "Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings," *Neural Networks*, vol. 141, pp. 211–224, 2021.

[14] Z. Mostaani, V. S. Nallanthighal, A. Härmä, H. Strik, and M. Magimai-Doss, "On the relationship between speech-based breathing signal prediction evaluation measures and breathing parameters estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1345–1349.

[15] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association, INTERSPEECH*. Shanghai, China: ISCA, 2020, pp. 2042–2046.

[16] M. Markitantov, D. Dresvyanskiy, D. Mamontov, H. Kaya, W. Minker, and A. Karpov, "Ensembling end-to-end deep models for computational paralinguistics tasks: Compare 2020 mask and breathing sub-challenges," in *Proceedings of INTERSPEECH*. Shanghai, China: ISCA, 2020, pp. 2072–2076.

[17] E. Rothauser, "Ieee recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[18] G. Deshpande and B. W. Schuller, "The dicova 2021 challenge—an encoder-decoder approach for covid-19 recognition from coughing audio," *Proceedings of Interspeech*, pp. 931–935, 2021.