Automatic Measurement of Speech Breathing Rate

Mohamed Ismail Yasar Arafath K.
Department of Electrical Engineering
Indian Institute of Technology, Kharagpur
India
kmiyasar@gmail.com

Aurobinda Routray
Department of Electrical Engineering
Indian Institute of Technology, Kharagpur
India
aroutray@ee.iitkgp.ac.in

***Abstract*—The speech breathing rate has been used for the early prediction of disease and detection of emotions. Most of the breath detection equipment are contact based. Here, we try to detect the speech breathing rate from speech recordings. Cepstrogram matrix is used as the feature for classifying the speech frames as breath or non-breath. The classifier used is the support vector machine (SVM) with a radial basis function (RBF) kernel. The classifier output is post-processed to join breathing segments which are closely spaced and remove breaths of small duration. The speech breathing rate is calculated from the breath to breath interval. The algorithm has been tested on a student evaluation database. When tested, the algorithm yields an F1 Score of 89% and root mean square error (RMSE) of 4.5 breaths/min for the speech-breathing rate. The breath segments have been validated by keenly listening to speech recordings and viewing thermal videos.**

***Index Terms*—Breath detection, cepstrogram, speech-breathing rate, SVM.**

I. Introduction

Speech breathing is usually referred to the breathing during speech activity. During speaking, the duration of expiration is almost ten times that of the quiet breathing. Also, the velocity at which air is inhaled is high compared to quiet breathing [1]. Hence, the breath sound is normally audible in the speech recordings. Speech breathing parameters have been used for disease prediction [2], [3] as well as in emotion detection [4], [5].

Measurement of speech breathing parameters is usually carried out using specialised devices like chest pneumograph [1] and surface electromyography [6] along with the microphone. These contact-based breath measurements need to connect external devices to the human body. These connections might inhibit the natural and spontaneous emotions of the speakers. The various non-contact methods use thermal camera [7] or radar sensors which are expensive. Speech breathing has also been analysed manually by listening to speech recordings [4]. This work tries to automate the calculation of the speech breathing rate from the recorded speech avoiding the requirement of any specialised hardware.

Prediction of physiological parameters from the speech has been a less explored area as stated by Jati et al., in [8]. The physiological parameters that have been predicted from speech include heart rate [9], [10], skin conductance [11], respiratory sinus arrhythmia [8]. Breath analysis from audio recordings have been performed in [12], [13]. Reyes et al., in [13] used the tracheal sound along with smartphone camera for finding the breathing phases. Abushakra and Faezipour in [12] used audio recorded near nose while the person was breathing. Here they have addressed the quiet breathing which is different from the speech breathing where the person being examined is speaking. They used voice activity detection (VAD) along with mel frequency cepstral coefficients (MFCC) for identifying the inspiration and expiration phases of breathing. The characteristics of the breath segment closely resemble with some unvoiced phonemes like fricatives which make the identification of breath segments from speech difficult.

Hlavnička et al. in [3] use linear frequency cepstral coefficients (LFCC) along with zero-crossing rate, auto-correlation function, signal power, and duration to classify the voiced regions. The LFCC was used for the detection of the unvoiced segments and later for the breath sound detection. This along with other parameters have been used for the prediction of the Parkinsons disease.

Breath sound detection is an important phase of speech breathing rate detection. It has been used to remove the breath segment in high-quality songs and speeches [14]. Ruinskiy and Lavner used MFCC [15] along with zero-crossing rate, short-term energy, spectral slope, and breath duration for the classification of the breath and non-breath segments of the recording. This method is found to detect weak fricatives in addition to the breath sounds [16]. The MFCC and its variations have been used in [16], [17] for the breath detection. Igras and Ziólko in [18] use discrete wavelet transform for breath detection.

In this paper, there are three contributions. Firstly, we have modified the existing breath detection algorithm in [14]. The feature used in this algorithm is the cepstrogram matrix formed from the MFCC coefficients. But unlike [14], we use an SVM classifier to classify the speech frames as breath or non-breath. The algorithm uses the breath duration features to post-process the classifier output for better results. Here, we have applied read speech only. Secondly, due to the errors found during the perceptual recognition of breaths [19], the obtained output of the algorithm has been validated using thermal video along with the speech signals. Lastly, we have designed an experiment for the data collection, which record thermal and normal video along with the speech in various emotional situations.

This paper is organised as follows. Section II describes how

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

the true breath segments are identified. The data collection method is discussed in section III. Section IV describes the proposed breath detection method.The results are shown in section V and section VI concludes the paper.

II. Identification of True Breath Signals

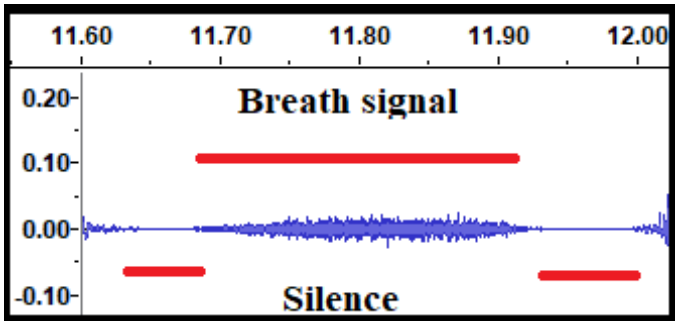
Speech breathing rate calculation from the speech recordings starts with the detection of the breath segments. A normal breath signal in a speech recording will be preceded and followed by silence region as shown in Figure 1(a). But all the breathing instances might not be like the one shown in Figure 1(a). Some speakers, when going out-of-breath, but don't want to stop in between a sentence, they take very sudden small duration breaths. Some of these breaths are very different from the normal breaths. An example is given in Figure 1(b). This speech segment has been confirmed as a breath event from the corresponding thermal video. The colour in the thermal image ranges from black to white. The white indicates the highest temperature and black the lowest. In Figure 2(a), the red region in the opening of the nose and the mouth indicate a decreased temperature due to the inhalation of air which is absent in the Figure 2(b). Figure 2(b) is the thermal image taken after this breath segment during the speech. The minimum duration of the breath is around 75 ms [14].

III. DATA COLLECTION

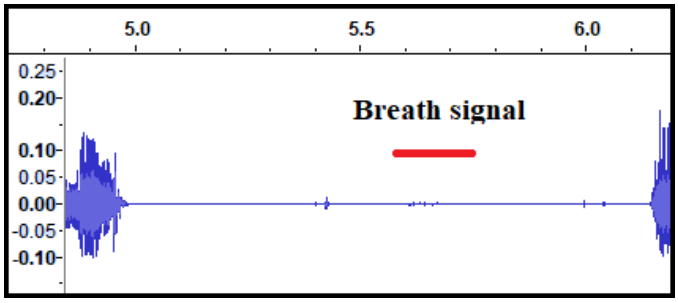
The data for the speech breathing rate measurement has been collected from a group of interns at the Indian Institute of Technology, Kharagpur. The dataset contains recordings in two situations: baseline recording and evaluation recording. For the baseline recording, the participants have been asked to do self-introduction and to read a paragraph twice. The recording room has been arranged to record their speech, thermal and normal video. Two psychologists have been part of the recording team and graded calmness of the subjects. The psychological state of the participants have been evaluated before and after the recording using short-State and Trait Anxiety Inventory (short-STAI) [20]. The participants have been asked to do diaphragmatic breathing [21] to relax them before they read the paragraph for the second time.

The referral is essential for the interns for getting admission to top universities and securing good jobs. The interns are generally under pressure to impress the professor, who will eventually act as a referee. Usually, the continuous evaluation of the interns are carried out and no special evaluation sessions are arranged. But to make the interns more anxious, a special evaluation session has been arranged. The students have been asked to read the same paragraph as in the baseline recording during the evaluation. The data in various emotional situations has been used to check whether the breaths in various emotions could be recognized.

The experiment has been approved by the ethical committee of the institute, and informed consent has been taken from the participants. There have been 16 participants for the experiment. The mean age of the participants is 21.22 years with a standard deviation of 2.1 years. This includes 5 female and 11 male participants. The microphone used for the recording is Ahuja HBM-50 headband microphone and audacity software has been used for the recording. The recording has been carried out on a Dell Inspiron laptop with a Core-i5 processor and 8GB RAM. The audio signals have been recorded in a 32-bit format at a sampling rate of 44.1 kHz. A Fluke Ti400 thermal camera has been used for recording. The thermal videos are recorded at 9 fps. A Nikon DSLR camera has been used for normal video recording. The videos are recorded at 59.94 fps. There are 47 recordings of the paragraph read, which include 32 baseline and 15 evaluation recordings, and 16 recordings of the self-introduction in the database. The average length of the audio signals in which the given paragraph has been read is 20.9 seconds. The breath sounds are audible in the speech recordings.

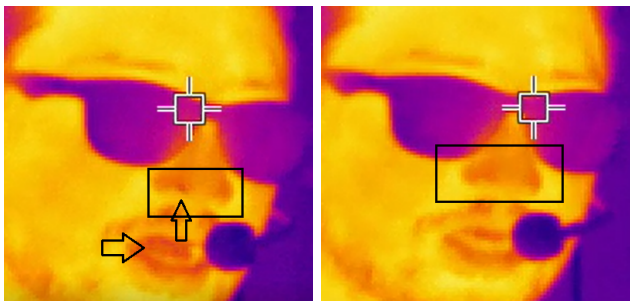


(a) Normal Breath segment



(b) A low amplitude Breath segment

Fig. 1. Examples of the breath segments in speech.



(a) During breath (b) During speech

Fig. 2. Thermal images taken during breath and speech.

IV. METHODOLOGY

The noise level in the speech recordings has been found to be low, and hence no denoising algorithm has been used. The speech signals used for training the SVM classifier has been manually divided into breath and non-breath segments. The non-breath segments include the speech, both voiced and unvoiced, and silence segments of the recordings. The frame size used for the algorithm has been selected as the minimum length of the breath signal FL_{min} as in [14]. The speech signals are divided into frames of length FL_{min} with a frame shift of 10 ms. The cepstrogram is calculated for each frame from the MFCCs as given in subsection IV-A. The cepstrogram of each speech frame is classified as breath and non-breath using the SVM classifier as given in subsection IV-B. Subsection IV-C describes the post-processing steps. The speech-breathing rate calculation is given in subsection IV-D.

A. Cepstrogram

The framed speech signals have been used for finding the cepstrogram. The procedure for calculating the cepstrogram is given below:

1) The pre-emphasis of the frame is carried out using a first-order difference filter.
   $H(z) = 1 - \alpha z^{-1}$ where α = 0.95.
2) The frames are divided into sub-frames of 10 ms duration with a hop size of 5 ms.
3) The MFCCs for each sub-frames are calculated and are concatenated as columns to form the cepstrogram matrix. The DC components are removed from each of the columns of the cepstrogram.
   The cepstrogram matrix X is given by

$$X = [V_1 \; V_2 \; \ldots \; V_N] \tag{1}$$

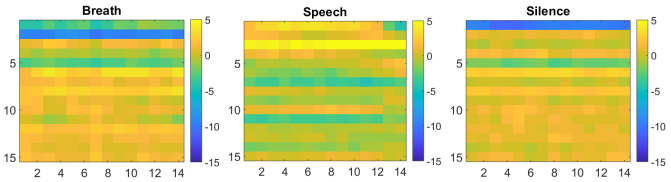
   where N is the number of sub-frames in a frame and $V_i \in \Re^{15\times 1}$ are the DC removed MFCC coefficients of the i^{th} sub-frame.

The MFCCs for each sub-frame has 15 values. The cepstrogram of silence, non-breath and breath frames are given in Figure 3. The voicebox[1] toolbox has been used for the calculation of the MFCCs.

B. Classifier

An SVM classifier with an RBF kernel has been used for the classification of each speech frame as breath or non-breath. The cepstrogram matrices of the frames of the manually separated breath and non-breath segments have been used as the features for SVM. The cepstrogram matrix has been reshaped into a one-dimensional vector $\bar{X} \in \Re^{1\times 15N}$ and used for the training of the classifier.The speech signals in the database have been divided into train and test data in the ratio of 40:60. The speech signals of 6 participants out of the 16 have been used for training. The training set include both female and male participants. There have been 85 breath segments and 108 non-breath segments in the training data.

[1]http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html



(a) Breath (b) Speech (c) Silence

Fig. 3. Cepstrogram of breath, speech and silence segments.

These segments produced 1718 breath frames and 35312 non-breath frames. The output of the classifier is a Breath vector $B(n) \in \Re^{1\times L}$, where L is the length of the speech input, whose values are either 1 or 0 and n is the index which varies from 1 to L.

$$B(n) = \begin{cases} 1 & \text{if } n \in \text{Breath frame,} \\ 0 & \text{if } n \in \text{Non-breath frame} \end{cases} \tag{2}$$

To overcome the miss-classification error, the output of the classifier output $B(n)$ has been post-processed.

C. Post processing

It has been observed that the classifier has erroneously classified some frames in the breath segment as non-breath. Breathing being a continuous process, in this circumstance, we use heuristic to join the closely spaced breath segments detected. The starting and ending index points of each breath frames has been found out. Let bs_i and be_i be the start and end index points of the i^{th} breath segment detected. The joining of close breaths has been done as given in (3).

$$B(n) = 1 \text{ if } \begin{cases} n \in [be_i \;\; bs_{i+1}] \;\; \& \\ bs_{i+1} - be_i < 2 \times FL_{min} \end{cases} \tag{3}$$

Also, the breath segments which do not have a minimum duration have been removed. The starting and ending points of the breath segments have been calculated again and the removal of smaller breaths have been performed as in (4).

$$B(n) = 0 \text{ if } \begin{cases} n \in [bs_i \;\; be_i] \;\; \& \\ be_i - bs_i < \frac{FL_{min}}{2} \end{cases} \tag{4}$$

D. Speech-breathing rate

The breath to breath interval (bb) has been calculated by finding the difference between the starting time of the adjacent breaths. The speech breathing rate (sbr) has been calculated as the inverse of the breath to breath interval (bb^{-1}). The average of the speech-breath rate has been calculated for each speech recording.

V. RESULT

The method has been able to detect most of the breath segments. The detected breaths have been validated using the thermal videos. The method has been compared with the method in [14]. In the edge detection stage the second method stated by Ruinskiy et al. has been used for the implementation.

Fig. 4. Breath segments detected in the speech signal.

TABLE I
PERFORMANCE EVALUATION OF THE ALGORITHM

	Ruinskiy and Lavner Method [14]	Proposed Method
Total Breaths	118	118
Full Detection (IOU≥0.5)	83	95
Partial Detection (IOU<0.5)	12	7
False Detection	46	8
Not Detected	23	16
Precision	67.38%	**92.73%**
Recall	80.51%	**86.44%**
F1 Score	73.36%	**89.47%**

TABLE II
PERFORMANCE OF PROPOSED METHOD ON BASELINE AND EVALUATION RECORDINGS

	Baseline recording	Evaluation recording
Total Breaths	75	43
Full Detection(IOU≥0.5)	63	32
Partial Detection (IOU<0.5)	1	6
False Detection	3	5
Not Detected	11	5
Precision	95.52	88.37
Recall	85.33	88.37
F1 Score	89.51	88.37

Though, this method is able to detect most of the breath segments, many false detections have been made. Also, the various threshold values have to be tuned to get good results. Both algorithms has been trained on the same data. The Intersection Over Union (IOU) is calculated for the breath segments detected. The breaths which has IOU greater than or equal to 0.5 is taken as full detection. The breaths which have IOU less than 0.5 are taken as partial detection. The recall, precision and F1 score have been calculated for both the methods.The results are given in Table I.

The performance of the proposed algorithm on the baseline and evaluation recordings is given in Table II. The result indicates that the breaths on baseline recordings is better detected than the evaluation. This might be due to the larger number of training samples in the baseline recording.

An example of the output has been shown in Figure 4. Here, the first graph shows the breaths detected after the classification. The second graph shows the output after the joining of the close breaths using (3). The output after the removal of small breaths using (4) has been given in the third graph. It can be observed that two adjacent breaths have been joined together and a breath of smaller duration has been removed. The algorithm could detect the starting edges of the breath with a difference less than 0.1 seconds. The RMSE of the speech breathing rate has been found to be 4.5 breaths/min.

VI. CONCLUSION

In this paper, we try to estimate the speech breathing rate from the speech recordings. The SVM classifier along with post-processing based on breath characteristics has been used

for breath detection. The algorithm has been found to be working better when compared with the algorithm in [14] on the created database.

In future, we are planning to develop a mobile-app for the same. We are planning to use speech breathing parameters to detect stress/anxiety. Also, the automatic measurement of speech breathing rate from spontaneous speech has to be addressed.

REFERENCES

[1] B Conrad and P Schönle, "Speech and respiration," *Archiv für Psychiatrie und Nervenkrankheiten*, vol. 226, no. 4, pp. 251–268, 1979.

[2] Nancy Pearl Solomon and Thomas J Hixon, "Speech breathing in parkinsons disease," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 294–310, 1993.

[3] Jan Hlavnička, Roman Čmejla, Tereza Tykalová, Karel Šonka, Evžen Ržička, and Jan Rusz, "Automated analysis of connected speech reveals early biomarkers of parkinsons disease in patients with rapid eye movement sleep behaviour disorder," *Scientific reports*, vol. 7, no. 1, pp. 12, 2017.

[4] Frieda Goldman-Eisler, "Speech-breathing activitya measure of tension and affect during interviews," *British Journal of Psychology*, vol. 46, no. 1, pp. 53–63, 1955.

[5] Edgar Heim, Peter H Knapp, Louis Vachon, Gordon G Globus, and S Joseph Nemetz, "Emotion, breathing and speech," *Journal of Psychosomatic Research*, vol. 12, no. 4, pp. 261–274, 1968.

[6] Joanna M Clair-Auger, Liu Shi Gan, Jonathan A Norton, and Carol A Boliek, "Simultaneous measurement of breathing kinematics and surface electromyography of chest wall muscles during maximum performance and speech tasks in children: Methodological considerations," *Folia Phoniatrica et Logopaedica*, vol. 67, no. 4, pp. 202–211, 2015.

[7] Anushree Basu, Aurobinda Routray, Rashmi Mukherjee, and Suprosanna Shit, "Infrared imaging based hyperventilation monitoring through respiration rate estimation," *Infrared Physics & Technology*, vol. 77, pp. 382–390, 2016.

[8] A. Jati, P. G. Williams, B. Baucom, and P. Georgiou, "Towards predicting physiology from speech during stressful conversations: Heart rate and respiratory sinus arrhythmia," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 4944–4948.

[9] D Skopin and S Baglikov, "Heartbeat feature extraction from vowel speech signal using 2d spectrum representation," in *Proc. the 4th Int. Conf. Information Technology*, 2009.

[10] Jennifer Smith, Andreas Tsiartas, Elizabeth Shriberg, Andreas Kathol, Adrian Willoughby, and Massimiliano de Zambotti, "Analysis and prediction of heart rate using speech features from natural speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 989–993.

[11] Björn Schuller, Felix Friedmann, and Florian Eyben, "Automatic recognition of physiological parameters in the human voice: Heart rate and skin conductance," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7219–7223.

[12] Ahmad Abushakra and Miad Faezipour, "Acoustic signal classification of breathing movements to virtually aid breath regulation," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 2, pp. 493–500, 2013.

[13] Bersain A Reyes, Natasa Reljin, Youngsun Kong, Yunyoung Nam, Sangho Ha, and Ki H Chon, "Towards the development of a mobile phonopneumogram: automatic breath-phase classification using smartphones," *Annals of biomedical engineering*, vol. 44, no. 9, pp. 2746–2759, 2016.

[14] Dima Ruinskiy and Yizhar 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.

[15] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, August 1980.

[16] Viliam Rapcan, Shona D'Arcy, and Richard B Reilly, "Automatic breath sound detection and removal for cognitive studies of speech and language," in *IET Irish Signals and Systems Conference (ISSC 2009)*. IET, 2009, pp. 1–6.

[17] Martin S Holmes, Shona D'arcy, Richard W Costello, and Richard B Reilly, "Acoustic analysis of inhaler sounds from community-dwelling asthmatic patients for automatic assessment of adherence," *IEEE journal of translational engineering in health and medicine*, vol. 2, pp. 1–10, 2014.

[18] Magdalena Igras and Bartosz Ziólko, "Wavelet method for breath detection in audio signals," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–6.

[19] Yu-Tsai Wang, Jordan R Green, Ignatius SB Nip, Ray D Kent, Jane Finley Kent, and Cara Ullman, "Accuracy of perceptually based and acoustically based inspiratory loci in reading," *Behavior research methods*, vol. 42, no. 3, pp. 791–797, 2010.

[20] Theresa M Marteau and Hilary Bekker, "The development of a six-item short-form of the state scale of the spielberger statetrait anxiety inventory (stai)," *British Journal of Clinical Psychology*, vol. 31, no. 3, pp. 301–306, 1992.

[21] Yu-Fen Chen, Xuan-Yi Huang, Ching-hui Chien, and Jui-Fen Cheng, "The effectiveness of diaphragmatic breathing relaxation training for reducing anxiety," *Perspectives in psychiatric care*, vol. 53, no. 4, pp. 329–336, 2017.