INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Automatic Detection of Breath Using Voice Activity Detection and SVM Classifier with Application on News Reports

Mohamed Ismail Yasar Arafath K., Aurobinda Routray

Indian Institute of Technology Kharagpur, India
kmiyasar@gmail.com, aroutray@ee.iitkgp.ac.in

Abstract

Breath detection during speech has broad applications ranging from emotion recognition to detection of diseases. Most of the breath detection equipment are contact based. In the proposed method, we use a voice activity detector (VAD) to find the non-speech region and searches the breath only in this region since breath is a non-speech activity. This reduces the execution time. A support vector machine (SVM) classifier is used with radial basis function (RBF) kernel trained on the cepstrogram feature to detect the breaths in the non-speech regions. The classifier output is post-processed to join breathing segments which are closely spaced and remove small duration breaths. Speech breathing rate is calculated as the ratio of the number of breaths to the time between the first and last breath. The algorithm is tested on a student evaluation database. The algorithm yields an F1 Score of 94% and root mean square error (RMSE) of 7.08 breaths/min for the speech-breathing rate. The output has been validated using thermal videos. The breaths have been classified as full and partial detection based on the Intersection over Union (IOU). The algorithm is also tested on some news channel reports which gave a minimum F1 Score of 73%.

Index Terms: Breath detection, cepstrogram, speech-breathing, SVM, VAD.

1. Introduction

Breathing is an inevitable part of speaking. It prepares the air needed for our speech. Breathing during speech activity is called speech breathing. Unlike quiet breathing, the expiration time is almost ten times during speech breathing [1]. The breath sounds are normally audible in the speech recordings. Speech breathing parameters have wide applications from emotion detection [2], [3] to disease prediction [4].

The detection of breath during speech is usually accomplished using special hardwares like electromyography [5], chest pneumograph [1] along with speech recordings. These devices need to be connected to the human body which might affect the natural and spontaneous emotions of the speaker [2]. The normally used non-contact methods uses expensive thermal camera [6] or radar sensors. Goldman-Eisler in [2] found out the breaths by manually listening to the speech recordings.

There are very few work on prediction of physiological signals from speech [7]. These include heart rate [7], [8], respiratory sinus arrhythmia [7], skin conductance [9] . Reyes et al. in [10], found the quiet breathing phases using the smartphone camera and tracheal sound recordings. The same was found from the audio recordings by keeping the microphone near the nose by Abushakra and Faezipur in [11]. They used VAD to detect the breath regions because the other portions in the recording are mostly silence. Then Mel Frequency Cepstral Coefficients (MFCC) are used for identifying the inspiration and expiration phases in breaths. But during speech breathing, the breath sound is due to the inspiration and the expiration occurs as speech. The resemblance of the breath segments with some fricatives makes the breath detection during speech more difficult.

Speech breathing parameters has been used for the prediction of parkinsons disease by Hlavnička and colleagues [4]. They found out the breath segments from the speech using Linear Frequency Cepstral Coefficients (LFCC) along with zero-crossing rate, auto-correlation function, signal power, and minimum duration. Using these breaths, speech breathing parameters like respiration rate, latency of respiratory exchange, duration of pause intervals, etc. have been found out and used for prediction of the disease.

Breath sound detection has wide applications from production of high-quality songs and speeches [12] to cognitive studies [13]. The most commonly used features for breath detection from speech are MFCC [12], along with short-term energy, zero-crossing rate. This method detects weak fricatives in addition to the breath sounds [13]. MFCC and its variations has been used by [12], [14]. In addition, discrete wavelet transform has been used for breath detection by Igras and Ziólko in [15]. Fukuda et al., has used a VAD to divide the speech recording to speech segments containing breaths and then found out the potential breath candidates using gaussian mixture model. An SVM classifier has been used to verify the breath segments.

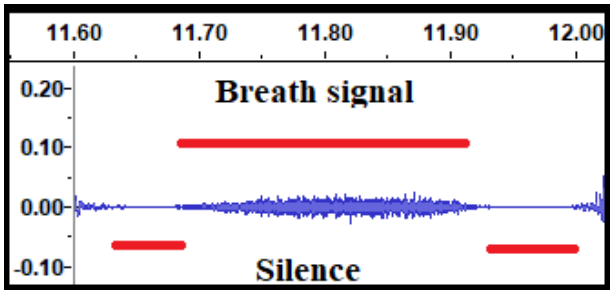
In our previous work [16], we have been searching the entire speech signal for the breath. The main contributions in the paper is an algorithm which uses a VAD to limit the search area of the speech recording and there by reducing the execution time. Later an SVM classifier has been used for detecting the breaths. We have used the Covarep VAD implementation [17]. The feature used in this algorithm is the cepstrogram matrix formed from the MFCC coefficients as used in [12]. The algorithm uses the breath duration to post-process the classifier output for better results. Due to the errors found during the perceptual recognition of breaths [18], the obtained output of the algorithm has been validated using thermal video along with the speech signals. Also, the algorithm has been tested on some YouTube videos in addition to the dataset used in our previous work [16].

The paper is organised as follows. Section 2 discusses how the true breaths have been identified from speech. The data sets used in the experiment is discussed in section 3. The proposed method is discussed in section 4. The results are shown in section 5. Section 6 discusses the conclusion and future work.

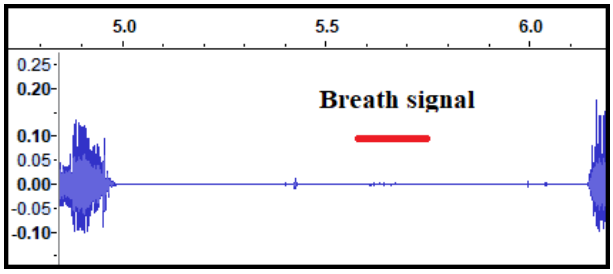
2. Identification of True Breath Segments

Breath signals are normally preceded and succeeded by silence regions as shown in Figure 1(a). But when speakers go out-of-breath during speaking, they take very sudden breaths. These breaths are different from the normal breaths. An example of

http://dx.doi.org/10.21437/Interspeech.2019-2434



(a) Normal Breath segment



(b) A low amplitude Breath segment

Figure 1: *Examples of the breath segments in speech [16].*

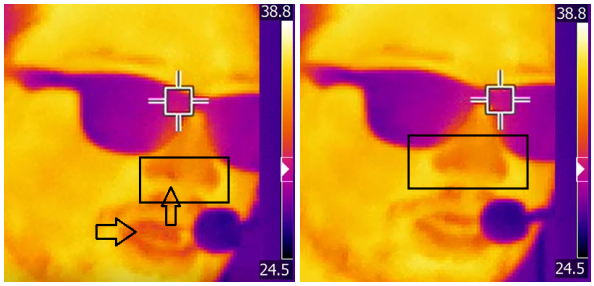
such a breath is given in Figure 1(b). Also some of the sounds of the fricatives closely resembles with the breath. So identifying the breath signals by merely seeing the breath waveform or by hearing might not give proper results.

Wang et al. in [18] points out that the judges have not been able to fully recognise the breaths from the speech by hearing. So some other measurements are required to validate the detection of breath signals. Lester and Hoit in [19] has shown that healthy adults inhale both through their mouth and nose during speaking. So, here we analyse the temperature near the mouth and nose in the thermal video to validate the breaths.

Thermal videos has been used for analyzing the breathing in [6], [20]. Here we have used the thermal videos to find the true breaths. The colormap shown in Figure 2 indicates that the red colour has less temperature than yellow. Figure 2(a) shows the thermal image during breath. The increased red colour in the mouth and nose regions indicate the decrease in temperature when compared to the yellowish colour in Figure 2(b). The decrease in temperature is due to the inhalation of the air. To identify the true breath locations, the thermal videos have been analysed first. Later the corresponding location in the speech waveform has been analysed to find the exact demarcation points. In case of the news reports, since we do not have the thermal signatures, the breath locations have been identified perceptually by listening to the speech recordings.

3. Dataset

This work uses the same dataset used in our previous work [16]. The data has been collected in various emotional situations: baseline recording and evaluation recording. Also the data collected is multimodal. The dataset contains speech, thermal video and normal video of the participants. The participants for the experiment have been a group of interns at Indian Institute of Technology Kharagpur. The participants have been asked to do a self introduction and to read a paragraph twice.



(a) During breath (b) During speech

Figure 2: *Thermal images taken during breath and speech [16].*

Diaphragmatic breathing [21] has been done by the participants to relax them before reading the paragraph for the second time. The evaluation recording have been performed in a such a way that the participants will be anxious. The real evaluation of the interns have been recorded and the participants have been asked to read a paragraph in between the evaluation. The recording has been performed in a closed room with very less noise.

The dataset contains speech recordings (at sampling rate 44.1 KHz) of the paragraph read along with thermal (at 6 frames per second) and normal video (at 60 frames per second) recordings. There are 47 recordings which include thirty two baseline and fifteen evaluation recordings. The audio recordings have an average length of 20.9 seconds. The breath sounds are audible in the recordings. The minimum observed breath duration is 120ms.

News reports In addition to the dataset, we have tested the proposed method on some news reports of Aljazeera taken from YouTube. Both the recordings used for the testing are closed room recordings. The press conference of New Zealand prime minister Jacinda Adern[1] has been taken as one test data. Only press briefing part of the video has been taken and the question and answer session has been removed. The recording contains sounds of camera flashes in addition to the speech, which can be considered as noise. The A news report from Aljazeera[2] has been cropped to include only the news room recordings. There are two anchors in this news report. The length of the first recording is 4 minutes 56 seconds and the second recording has 1 minute and 40 seconds. The music overlapped portions of both reports has also been removed.

4. Automatic Breath Detection

A speech recording can broadly be marked as speech and non-speech regions. The speech regions include speech segments and the remaining portions like silence, breath, etc, belongs to the non-speech regions. This is normally performed using voice activity detector. So the breath signals need only be checked in the non-speech regions. The proposed method uses this criteria and uses a VAD to find the non-speech regions. The flow chart of the proposed method is shown in Figure 3. Denoising algorithm has not been applied since the noise level in the speech recordings has been found to be low.

4.1. Voice activity detection

Voice activity detection is normally performed to detect the regions which has voice. This is useful in speech recognition ap-

[1] https://www.youtube.com/watch?v=krqjuv8AXA4
[2] https://www.youtube.com/watch?v=FjqxEbvig5U

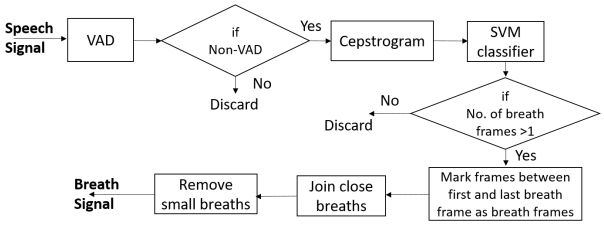
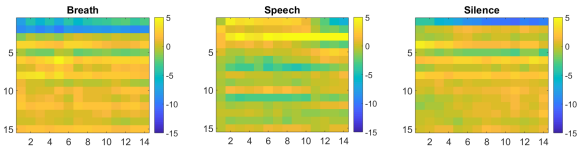


Figure 3: *Flowchart of proposed method*



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

Figure 4: *Cepstrogram of breath, speech and silence segments.*

plications. Here, we use the VAD to get the non-speech regions. We tried two available VAD implementations, namely free VAD from the Opensmile [22] and from the Covarep library. The Covarep VAD produced better results compared with the Opensmile VAD in our application. This VAD algorithm fuses the features of [23], [24], and [25]. The implementation is available with the Covarep library.

Not all the non-speech regions need to have a breath signal. So an SVM classifier has been used to find the breath in the non-speech regions. The non-voice regions identified are divided into frames of length FL with an overlap length of 10ms, where FL is taken as 80ms. The frames are classified as breath and non-breath frames using an SVM classifier trained using cepstrogram feature.

4.2. Cepstrogram

Cepstrogram feature has been calculated from the MFCC features. The pre-emphasized speech signal frames are split to sub-frames of 10ms with a hop size of 5ms. The DC removed MFCC for each of the sub-frames are stacked together as columns to form the cepstrogram matrix. The detailed steps are available in [12]. The cepstrogram of silence, non-breath and breath frames are given in Figure 4. The number of rows denote the MFCCs for each sub-frame and the columns are the number of sub-frames. The voicebox[3] toolbox has been used for the calculation of the MFCCs.

4.3. Classifier

An SVM classifier with an RBF kernel is used to identify if a non-voice region has breath or not. This is accomplished by the classification of each speech frame in the non-voice region as breath or non-breath. A non-voice region which has at least two breath frames is assumed to contain a breath.

The speech signals of 9 participants out of the 16 is used for training. The training set include both female and male participants. There are 131 breath segments and 154 non-breath segments in the training data. The cepstrogram matrix for the 3354 breath frames and 52179 non-breath frames from these segments has been calculated . The matrix has been reshaped to one-dimensional cepstrogram vector $\bar{X} \in \Re^{1\times 15N}$ and normalized before the training.

To overcome the miss-classification error and since breath is continuous process, in a non-speech region, all the frames between the first and last breath frames detected are considered as breath frames. A breath vector $Bv \in \Re^{1\times L}$, is made whose values are either 1 or 0 where L is the length of the speech input.

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

The output $B(n)$ is post-processed to join adjacent breaths and remove small breaths.

4.4. Post processing

It has been observed that if the threshold of VAD is not properly chosen it might detect a non-speech region as multiple. So we use heuristic to join the breath segment which are close.

Let bs_i and be_i be the start and end points of the i^{th} breath segment detected. The joining of close breaths is done as given in (2).

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

The breath segments can not be below a minimum length (here 100ms) and so are removed in those cases. For this, the starting and ending points of the breath segments are calculated again and the removal of smaller breaths are performed as in (3).

$$B(n) = 0 \text{ if } \begin{cases} n \in [bs_i \;\; be_i] \;\; \& \\ \\ be_i - bs_i < 0.1 \times \text{Sampling Frequency} \end{cases} \tag{3}$$

4.5. Speech-breathing rate

Speech breathing rate (sbr) is calculated as the ratio of number of breaths detected and the time between the first and last breath. Let there be N breaths and bs_1, bs_N be the starting points of first and last breath respectively, then

$$sbr = \frac{N-1}{bs_N - bs_1} \tag{4}$$

5. Result

The VAD output has been classified as non-speech and speech region based on the threshold value of 0.1. The method detected most of the breaths. The detected breaths have been compared with thermal validated true breaths. The IOU is calculated to classify the detected breaths as full or partial detection. A threshold of 0.5 has been used for the same. The method has been compared with [12]. The second method in [12] for false removal and breath demarcation has been implemented here. The result has been obtained by fine tuning various thresholds. The work has also been compared with our previous work [16] which do not use a VAD. The results have been provided in Table 1. Although our previous work better marginally in recall, the proposed method works faster.

The output of the proposed method is shown in Figure 5. The first graph shows the non-VAD regions detected. A threshold of 0.1 has been used to detect the non-VAD regions. The non-VAD regions which has at least two breath frames is shown

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

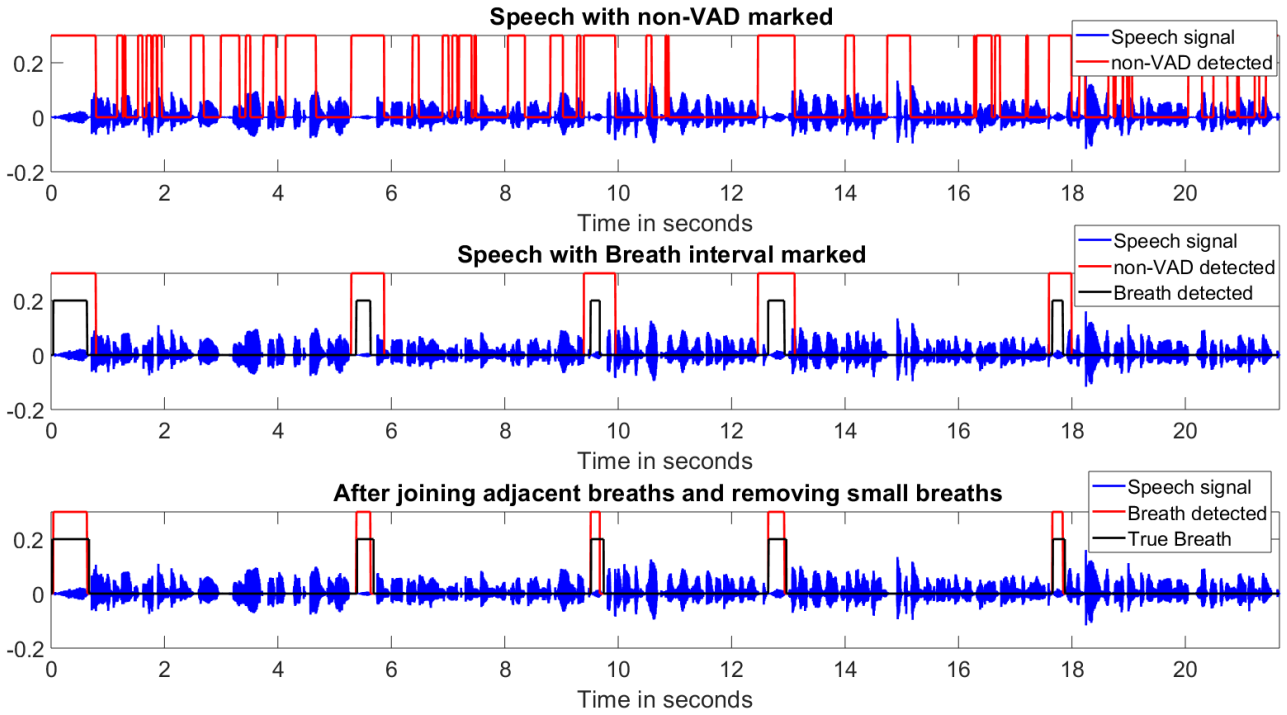


Figure 5: *Breath segments detected in the speech signal.*

Table 1: *Performance Evaluation of the Algorithm*

	Template Matching	**Previous method [16]**	**Proposed method**
Total Breaths	85	85	85
Full detection (IOU>0.5)	42	75	72
Partial detection (IOU<0.5)	24	7	8
False Detection	39	8	5
Not Detected	20	3	5
Precision	62.86	91.11	**94.11**
Recall	76.74	**96.47**	94.11
F1 Score	69.11	93.71	**94.11**
Execution Time	**9.87s**	42.38s	13.07s

Table 2: *Performance of Algorithm on News reports*
M1-Proposed Method, M2-Previous method [16]
NZ PM Speech- New Zealand Prime Minister Speech

	NZ PM Speech		**Aljazeera**	
	M1	**M2**	**M1**	**M2**
Total Breaths	81	81	31	31
Full detection (IOU>0.5)	50	28	18	16
Partial detection (IOU<0.5)	16	37	0	2
False Detection	15	33	0	0
Not Detected	15	22	13	13
Precision	**81.48**	66.33	**100.00**	**100.00**
Recall	**81.48**	74.71	**58.06**	58.06
F1 Score	**81.48**	70.27	**73.47**	**73.47**

in the second graph along with the breath detected in the non-VAD region. The output after joining close breaths and removing small breaths is shown in the third graph. The speech breathing rate calculated has an RMSE of 7.08 breaths/minute.

The algorithms have also been tested on news channel reports described in section 3. The ground truth has been identified by listening to the speech signals only, due to the absence of thermal videos. The results are given in Table 2. The performance of template matching method on news report has been found to be less.

6. Conclusion and Future Work

In this paper, we try to find out the breath segments from the speech signals. The proposed method has been compared with our previous work [16] and the template matching algorithm in [12]. The overall performance shows that the proposed method works better than the other methods. Although the performance of our previous work in terms of recall is better on the database, the execution time taken by the algorithm is very high compared to the proposed method. With the use of VAD, we can restrict the search region and hence, can reduce the execution time. The false detection is also reduced since the only the non-speech region is searched for breath.

In future, we are planning to develop a mobile-app for the same. Also, we are planning to use speech breathing parameters to detect stress/anxiety.

7. References

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

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

[3] E. Heim, P. H. Knapp, L. Vachon, G. G. Globus, and S. J. Nemetz, "Emotion, breathing and speech," *Journal of Psychosomatic Research*, vol. 12, no. 4, pp. 261–274, 1968.

[4] J. Hlavnička, R. Čmejla, T. Tykalová, K. Šonka, E. Ržička, and J. 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, p. 12, 2017.

[5] J. M. Clair-Auger, L. S. Gan, J. A. Norton, and C. 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.

[6] A. Basu, A. Routray, R. Mukherjee, and S. Shit, "Infrared imaging based hyperventilation monitoring through respiration rate estimation," *Infrared Physics & Technology*, vol. 77, pp. 382–390, 2016.

[7] 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.

[8] 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.

[9] B. Schuller, F. Friedmann, and F. 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.

[10] B. A. Reyes, N. Reljin, Y. Kong, Y. Nam, S. Ha, and K. 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.

[11] A. Abushakra and M. 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.

[12] 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.

[13] V. Rapcan, S. D'Arcy, and R. 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.

[14] T. Nakano, J. Ogata, M. Goto, and Y. Hiraga, "Analysis and automatic detection of breath sounds in unaccompanied singing voice," *Proc. of ICMPC 2008*, pp. 387–390, 2008.

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

[16] K. M. I. Y. Arafath and A. Routray, "Automatic measurement of speech breathing rate," *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019 (Forthcoming).

[17] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep - a collaborative voice analysis repository for speech technologies," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 960–964.

[18] Y.-T. Wang, J. R. Green, I. S. Nip, R. D. Kent, J. F. Kent, and C. Ullman, "Accuracy of perceptually based and acoustically based inspiratory loci in reading," *Behavior research methods*, vol. 42, no. 3, pp. 791–797, 2010.

[19] R. A. Lester and J. D. Hoit, "Nasal and oral inspiration during natural speech breathing," *Journal of Speech, Language, and Hearing Research*, 2014.

[20] R. Murthy and I. Pavlidis, "Non-contact monitoring of respiratory function using infrared imaging," *IEEE Engineering in Medicine and Biology Magazine*, vol. 25, pp. 57–57, 2006.

[21] Y.-F. Chen, X.-Y. Huang, C.-h. Chien, and J.-F. Cheng, "The effectiveness of diaphragmatic breathing relaxation training for reducing anxiety," *Perspectives in psychiatric care*, vol. 53, no. 4, pp. 329–336, 2017.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[23] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice activity detection: Merging source and filter-based information," *IEEE Signal Processing Letters*, vol. 23, no. 2, pp. 252–256, 2016.

[24] S. O. Sadjadi and J. H. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, 2013.

[25] T. Drugman, Y. Stylianou, L. Chen, X. Chen, and M. J. Gales, "Robust excitation-based features for automatic speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4664–4668.