Voice activity detection from the breathing pattern of the speaker

A G Ramakrishnan, Gireesh Krishnan
MILE Laboratory, Dept. of Electrical Engineering
Indian Institute of Science
Bengaluru, India
ramkiag@ee.iisc.ernet.in,

S Srivathsan
Dept. of Electrical and Electronics Engineering
National Institute of Technology Karnataka,
Surathkal, India
s.srivaths@gmail.com

***Abstract*—**

In this paper, we propose a method to perform voice activity detection using only the breathing signal of a speaker. Human breathing and speech production go hand in hand. Normal respiration and respiration during speech have a different profile. The former is generally symmetric as compared to an asymmetric profile in the case of respiration during speech. Impedance pneumography provides a mechanism to capture chest expansions and compressions due to breathing. We have recorded the breathing signal along with the speech audio for 44 subjects while they were speaking and quiet. We have classified cycles of breathing into two classes, namely during speech and normal, using the cycle-synchronous discrete cosine transform coefficients of the breathing signal with different classifiers. The best accuracy of 96.4% is obtained using the k-nearest neighbor classifier. From the classified breathing cycles, we determine the intervals when a subject is quiet and when he is speaking. We use the corresponding timeframes on the simultaneously recorded audio and achieve a good accuracy in voice activity detection. Compared to the earlier reported time resolution of 30 sec, we obtain a decision for every breathing cycle, which works out to an average resolution of about 3 sec.

Keywords: Voice activity detection, breathing pattern, impedance pneumography, cycle-synchronous DCT, speech, support vector machine.

I. INTRODUCTION

The simple act of breathing is something most of us take for granted. Human body is designed to do this without conscious thought. Breathing therapy is about making the unconscious act of breathing, conscious. Breathing is also central to speech, and the way we breathe when we speak is very different from the metabolic breathing. Breathing is an important aspect in speech production and professional singing. Researchers have however brought out that speech itself has effects on breathing. Rahman et al. [1] presented a mobile-based system to infer conversation episodes from the signal collected from an unobtrusively wearable respiratory inductive plethysmograph band worn around the user's chest. The signal was wirelessly transmitted to a mobile phone, where it was used to determine whether the wearer is speaking, listening, or quiet. They proposed a set of features computed from the respiration signal, and a classifier to distinguish among quiet, listening, and speaking events. Features extracted from the respiration signal include inhalation duration (standard deviation (SD)), exhalation duration (mean, median and 80th percentile (p80)), ratio of inhalation to exhalation durations (mean, median, SD and p80), stretch (mean, median, SD and p80), B-duration (mean, median and p80) and first difference of exhalation (mean). Stretch is the difference between the highest and the lowest amplitudes of the signal within a respiration cycle. B-duration is the time the signal continues to stay within 2.5% of the minimum amplitude. The recorded respiratory signal was broken down into windows of 30 seconds each. Each window was assigned to a class (quiet, listening or speaking event) if that event occurred for more than 66% of the total duration of the window. The classification was first carried out using a boosted decision tree. The classification accuracy was improved using the output of the tree given as the input to a hidden Markov model. In Wodarczaket al. [2], the focus is placed on the role of breathing as a perceptually salient turn-taking cue. The paper highlights that respiratory data could be particularly helpful for investigating mechanisms of turn management in a conversation. As turns are normally preceded by easily perceivable inhalations and followed by equally salient exhalations, they propose that these can be used as cues to understand turn-taking.

The discrete cosine transform (DCT) is a reversible transformation with an excellent energy compaction property and hence, has been used effectively to extract

features for classification in a number of applications. Peeta Basa Pati et al. [3] have reported an algorithm to identify the script of each word in a document image, wherein the low frequency DCT coefficients are chosen from the DCT matrix of each half of the word image to form a 36-dimensional feature vector. The performance of DCT features is compared with that of Gabor features for script identification in a multi-script document. Aparna et al. [4] brought out the importance of DCT features in optical character recognition to recognize printed Tamil text. Truncated DCT coefficients are used as a set of features and the nearest neighbor classifier is used for the classification of the symbols. Euclidean distance is used as a distance metric.

Ramakrishnan et al.[5] has investigated the efficacy of using global features alone (discrete Fourier transform, DCT), local features alone (preprocessed (x,y) coordinates) and a combination of both global and local features for recognizing online handwritten Tamil characters. Using support vector machine (SVM) with radial basis function kernel and the proposed features, the authors have reported an accuracy of more than 95% in a 155-class, online handwritten character recognition. In [6], Muralishankar et al. have proposed a novel algorithm for pitch modification. The linear prediction residual is obtained from pitch synchronous frames by inverse filtering the speech signal. Based on the desired factor of pitch modification, the dimension of the DCT coefficients of the residual is modified by truncating or zero padding, and then the inverse DCT is obtained. This period modified residual signal is then forward filtered to obtain the pitch modified speech.

In our study, we identify intervals of speech and silence by processing only the recorded breathing signal. We classify every cycle of breathing instead of a segment of specified duration. A cycle of breathing comprises inhalation of air into the lungs followed by exhalation of air out of the lungs. We use cycle-synchronous DCT based features for classification. These features were originally proposed by Ramakrishnan et al. for voice source characterization [7]. They showed that the DCT has the ability to capture the time-domain pulse shape of the voice source within its first few coefficients. Since the pulse shape of the voice source can be successfully exploited for speaker identification, its DCT has been successfully explored as a characterization of the voice source for speaker identification.

II. EXPERIMENTAL SETUP AND FEATURE EXTRACTION

Impedance pneumography is a non-intrusive technique to monitor a person's respiration rate, or breathing rate. This involves external electrodes which are applied on the thorax and abdomen. A low valued high-frequency ac current is injected into the tissue through the drive electrodes. The AC current causes a potential difference to develop across any two points between the drive electrodes. This potential difference is related to the impedance of the tissue between the voltage-sensing or receive electrodes. The equivalent impedance is defined as the ratio of the voltage difference between the two electrodes and the current that flows through the tissue. Changes in the electrical impedance of the lungs are mainly a result of the following two effects.

- During inspiration, there is an increase in the gas volume of the chest in relation to the fluid volume. This increase in volume of the gas causes the chest conductivity to decrease.
- During inspiration, the length of the conductance paths increases because of chest expansion.

These two effects together cause the electrical impedance to increase. There is a good correlation between this impedance change and the volume of the respired air. This relationship is approximately linear. The varying component of impedance generates a varying voltage component when a constant current is injected. This varying voltage component is the parameter used to determine the person's breathing rate.

In our study, we predict whether the person is speaking, by processing the recorded breathing pattern of the subject and confirm the same using the simultaneously recorded audio. The data has been collected only from male subjects (due to the locations of the pneumograph electrodes). Prior to the recording, the subjects are informed about the experimental details. All the metallic and electronic gadgets (e.g. rings, bracelets, mobile phones) in contact with their body are removed during the test. The subjects are asked to sit on a chair with their hands resting on their lap. Breathing signal is recorded from 44 volunteers aged between 17 and 20 years when they are speaking and when they are quiet. The data recorded is summarized in Table 1.

Each breathing cycle comprises an inhalation followed by an exhalation. To segregate each breathing cycle, the peaks and valleys of the respiratory data are to be identified.The peaks correspond to the end of inhalation and valleys correspond to the end of exhalation. Each breathing cycle therefore consists of two adjacent valleys with

one peak in between. The intervals between adjacent peaks and valleys give the durations of inhalation and exhalation. A few cycles of normal breathing signal, as well as those recorded during speaking are shown in Fig. 1.

Table 1- Details of data recorded for the study.

Sl. No	Phase of Speech	Details of recorded data
1	Quiet	Subjects were silent Time – 60-90 s
2	Speak	Subjects were asked to speak Time – 100-120 s
3	Mixed	Subjects were played an audio file, using a headphone, which contained cues for when to speak and when to be silent. This was a 5-minutes long recording with quiet and speak durations as listed below: 1- 20 sec:Silent 20- 40 sec:Speak 60- 80 sec : Silent 80-120 sec:Speak 120-135 sec:Silent 135-180 sec:Speak 180-205 sec:Silent 205-240 sec:Speak 240-270 sec:Silent 270-300 sec:Speak

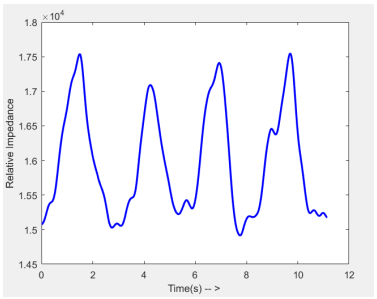


Figure-1 Sample segments of the breathing signal from a subject, when he was (a) quiet (b) speaking. The y-axis is the relative impedance and the x-axis is time in seconds. Note that the exhalations are longer during speaking.

As seen from the plots above, the breathing profile is nearly symmetric when the subject is quiet and asymmetric while speaking. It was shown in Ramakrishnan et al. [7] that since the pulse shape of the voice source can be successfully exploited for speaker identification, its DCT can be explored as an alternate characterization of the voice source for speaker identification. Since breathing profiles of speaking phase varied from that of quiet phases, we explored the option of using this approach on our data.

Discrete cosine transform linearly transforms the signal and expresses it in terms of a sum of cosine functions with related frequencies. For an N-point signal x(t), t = 0,1,2, ...,N-1, the DCT of the signal is given by the set of N real coefficients:

$$X(k) = c_k \sqrt{\frac{2}{N}} \sum_{t=0}^{N-1} x(t) cos\left[\frac{\pi k}{N}(t+\frac{1}{2})\right], k = 0, 1, 2..., N-1$$

$$c_k = \begin{cases} \frac{1}{\sqrt{2}}, & k = 0 \\ 1, & k > 0 \end{cases}$$

The first DCT coefficient is called the DC (direct current, zero frequency) coefficient and the remaining coefficients are called the AC (alternating current) coefficients. The advantage of DCT is that the energy of the original data gets concentrated in only a few coefficients of DCT depending on the correlation in the data. This means that most of the high frequency DCT coefficients may be zero or very small values.

To decide the dimension of the DCT features to be used for classification, we computed the DCT coefficients (except the DC coefficient) of each breathing cycle and found their total energy. Then, we found out how many coefficients of the DCT were required to capture 99 percent of the energy of that cycle. We averaged this over all the data and found that 70 DCT coefficients comprised 99 percent of energy for any cycle. Along with the 70 DCT features, we combine two features proposed by Rahman et al. [1], namely stretch (difference between the maximum and the minimum amplitudes) and inhalation to exhalation time ratio. So finally, we have a 72-dimensional feature vector. Since most of the high frequency DCT coefficients have a low amplitude or close to zero, we chose to experiment classification with only a few coefficients, while maintaining the same accuracies. Hence, the number of DCT coefficients was found again as mentioned above, which comprise 95% of the energy of each cycle on the average. We found out that 10 DCT coefficients were enough to capture 95 percent of the energy. These, together with stretch and inhalation to exhalation time ratio gave us a reduced 12-

dimensional feature vector. We compute these features for every breathing cycle delineated.

III. RESULTS AND DISCUSSION

The data collected from 44 subjects consists of respiratory signals when the subjects were speaking and quiet. The classification task at hand is to identify which cycles of the respiratory data belong to the subject's speaking, and which to being silent. We have trained four different classifiers using the DCT features of the training part of the 44 subjects' data.

We used 30% of data for validation to obtain classification accuracies. The confusion matrix obtained for the 72-dimensional feature vector is shown in Table 2.The accuracies are tabulated in Table3. The best accuracy (96.4%) is obtained using the kNN classifier with cosine distance metric, although the performances of the other classifiers are comparable.

We tried voice activity detection (VAD)using the results of the classifier with maximum accuracy i.e. the kNN classifier. The recorded audio and respiratory data are time synchronized using the starting time of the recoding. Using peak and valley information from the respiratory data, the breathing cycles are segmented. DCT features are extracted for each breathing cycle. These features are then fed to the classifier as input.

Table 2- Confusion matrix obtained using 72-dimensional features to classify breathing signal into normal (N) cycles and cycles during speech (DS).

Classifier	**True Class**	**10-fold cross validation (%)**		**Data 7:3 split (%)**	
		Predicted Class		Predicted Class	
		N	DS	N	DS
kNN	N	90.1	09.9	91.1	08.9
	DS	1.2	98.8	1.5	98.5
Linear SVM	N	92.9	7.1	93.5	06.5
	DS	11.1	88.9	13.6	86.4
Quadratic SVM	N	91.4	8.6	91.1	8.9
	DS	04.6	95.4	5.1	94.9
Cubic SVM	N	90.7	09.3	93.5	06.5
	DS	6.3	93.7	7.9	92.1
Gaussian SVM	N	68.3	31.7	67.8	32.2
	DS	0.3	99.7	0.0	100

Table 3- Accuracies obtained using 72-dimensional features in classifying breathing signal into normal cycles and during speech.

Classifier	**Accuracy with 10-fold cross validation (%)**	**Accuracy with data 70-30% (%)**
kNN	96.1	96.4
Linear SVM	90.1	88.4
Quadratic SVM	94.3	93.8
Cubic SVM	92.8	92.5
Gaussian SVM	90.7	90.7

Table 4- Confusion matrix obtained using truncated 12-dimensional features to classify the breathing signals into normal cycles (N) and cycles during speech (DS).

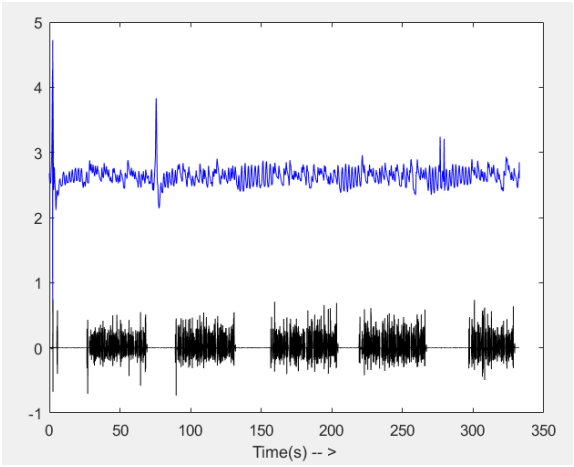
Classifier	**True Class**	**10-fold cross validation (%)**		**Accuracy (%) -data 7:3 split**	
		Predicted Class		Predicted Class	
		N	DS	N	DS
kNN	N	90.0	10.0	93.5	6.5
	DS	2.1	97.9	3.6	96.4
Linear SVM	N	93.6	6.4	95.3	4.7
	DS	12.0	88.0	12.8	87.2
Quadratic SVM	N	92.5	07.5	94.4	5.6
	DS	4.5	95.5	5.3	94.7
Cubic SVM	N	90.0	10.0	92.1	7.9
	DS	5.2	94.8	6.0	94.0
Gaussian SVM	N	85.0	15.0	84.6	15.4
	DS	1.4	98.6	2.1	97.9

Table 5-Accuracies obtained using truncated 12-dimensional features to classify the breathing signals into normal cycles and during speech.

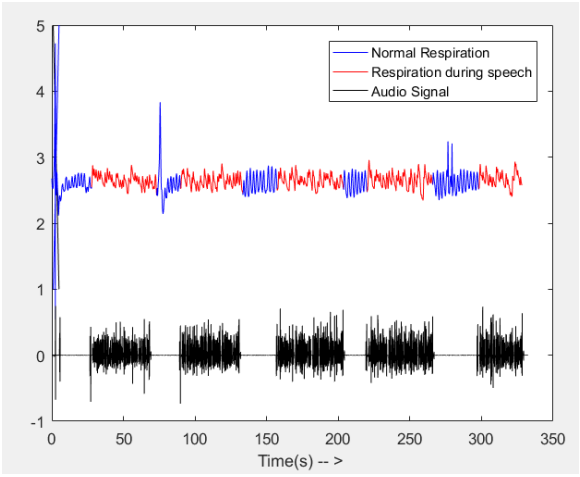
Classifier	**Accuracy with 10 fold cross-validation (%)**	**Accuracy with data 7:3 split (%)**
kNN	95.6	95.6
Linear SVM	89.6	89.5
Quadratic SVM	94.6	94.6
Cubic SVM	93.5	93.4
Gaussian SVM	94.7	94.1

From the classifier output, breathing cycles corresponding to the speaking and silent phases are identified. The time indices obtained from the breathing cycles of silent and speech phases are then used to label the audio signal as the speech and silent durations. Figure 2(a) shows the time-synchronized audio and respiratory data for one of the subjects, prior to classification. Figure 2(b) shows the same respiratory signal classified into cycles corresponding to the speech and silent phases of the subject.

The time-resolution of classification in our case is the same as the duration of every cycle of breathing. Thus, by computing the average duration of a breathing cycle, we obtain a resolution of approximately 2-3 seconds. However, in mConverse, the breathing patterns are classified using breathing windows of 30-second duration. Also in mConverse, the data was collected over a period of 2 days from 12 different subjects. Whereas, we collected data from 44 subjects and each recording was only 5 minutes long.



(a)



(b)

Figure-2 (a) The breathing and audio signals recorded together.(b) The classified breathing cycles, together with the corresponding audio data.

IV. CONCLUSION

The conventional approach in the field of VAD has been to detect speech and silent segments by processing the audio signals. We have attempted VAD by analyzing the nature of the breathing pattern of the speaker. From the breathing pattern of the speaker, we are able to identify when the speaker is quiet and when he is speaking. Using these timings, we identify the speech and silent segments in the audio signals with a fairly good accuracy of 96.4%.

Our approach may be potentially useful when VAD needs to be applied in a noisy environment. When speech is recorded in a noisy environment, the noise corrupts only the audio signal and the breathing data recorded simultaneously is not corrupted by the noise. If VAD is carried out using our approach, the time intervals during which the subject is speaking can be obtained from the respiratory data. This information can then be used to locate the noise only (otherwise silent) segments in the recorded audio signal. This facilitates the identification of the type and characteristics of the noise source [8] and any efficient noise cancellation algorithm [9] can be used to de-noise the corrupt voiced segments of speech.

Breathing is a continuous process. Human body is designed to do this without any conscious thought. The day to day activities that we perform influence our breathing pattern. We have restricted our study to analyze the breathing patterns present when a person is speaking and is silent. This approach could further be extended to many other fields, which have effects on breathing.

V. ACKNOWLEDGEMENT

The authors place on record their sincere gratitude to Dr. Sanjay Bharadwaj, Chief Architect, Skanray Technologies Pvt. Ltd., Mysore for lending their impedance pneumograph, with facilities for direct digitization of the data, for our research.

REFERENCES

1. Md. Mahbubur Rahman, Amin Ahsan Ali, Kurt Plarre, Mustafa al-Absiy, Emre Ertin and Santosh Kumar. mConverse: Inferring conversation episodes from respiratory measurements collected in the field. *Proc. 2nd Conf. Wireless Health,* Article No.10, *San Diego, California*,2011.

2. Marcin Wodarczak, Mattias Heldner and Jens Edlund. Breathing in conversation: an unwritten history. *Proc. 2nd European and 5th NordicSymp. Multimodal Commun.*, pp. 107-112, *Linkoping*, 2015.

3. Peeta Basa Pati and A. G. Ramakrishnan. Word level multi-script identification. *Pattern Recognition Letters*, 2008, Vol. 29, pp. 1218-1229.

4. K.G.Aparna and A.G.Ramakrishnan. A complete Tamil optical character recognition system. *Proc. Fifth IAPR Workshop on Document Analysis Systems DAS-02, Princeton, NJ*, August 19-21, 2002, pp. 53-57.

5. A.G. Ramakrishnan and Bhargava Urala. Global and local features for recognition of online handwritten numerals and Tamil characters. *ACM Proceedings. International Workshop on Multilingual OCR, (MOCR 2013)*, 24 Aug. 2013, *Washington DC, USA*.

6. R Muralishankar, A.G.Ramakrishnan and P Prathibha. Modification of pitch using DCT in the source domain. *Speech Communication*, 2004, Vol. 42/2, pp. 143-154.

7. A G Ramakrishnan, B Abhiram and S R Mahadeva Prasanna. Voice source characterization using pitch synchronous discrete cosine transform for speaker identification. *Journal of the Acoustical Society of America Express Letters*, Vol. 137, 2015.

8. K V Vijay Girish, A G Ramakrishnan and T V Ananthapadmanabha, "Cosine similarity based dictionary learning and source recovery for classification of diverse audio sources," Proc. 13th International IEEE India Conf. (INDICON), Bangalore, India, Dec. 16-18, 2016.

9. K V Vijay Girish, A G Ramakrishnan and T V Ananthapadmanabha, "Hierarchical classification of speaker and background noise and estimation of SNR using sparse representation," Proc. 17th Annual Conf. Inter. Speech Commun. Association (INTERSPEECH 2016), San Francisco, USA, Sept. 8-12, 2016.