APPLYING SPEECH DERIVED BREATHING PATTERNS TO AUTOMATICALLY CLASSIFY HUMAN CONFIDENCE

Gauri Deshpande[1,2], *Yagna Gudipalli*[1], *Sachin Patel*[1], *Björn W. Schuller*[2,3]

[1]TCS Research Pune, India
[2]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[4]Bharti Vidyapeeth (DTU) College of Engineering Pune, India
[5]Bharti Vidyapeeth (DTU) Medical College, Department of Physiology Pune, India

ABSTRACT

Non-verbal expressions of speech are used to understand a spectrum of human behaviour parameters; one of them being confidence. Several speech representation techniques, from hand-crafted features to auto-encoder representations, are explored for mining such information. We introduce a deep network trained with 100 speakers' data for the extraction of breathing patterns from the speech signals. This network gives an average Pearson's correlation coefficient of 0.61 and a breaths-per-minute error of 2.5 across 100 speakers. In this paper, we propose the novel use of speech-derived breathing patterns as the feature set for the binary classification of confidence levels. The classification model trained with the data from 51 interview candidates gives an average AUC of $76\,\%$ in classifying the confident speakers from the non-confident ones using breathing patterns as the feature set. On comparing this performance with that of Mel frequency cepstral coefficients and auto-encoder representations, we observe an absolute improvement of $8\,\%$ and $5\,\%$ respectively.

***Index Terms*—speech-breathing, affective computing, time-series analysis, computational paralinguistics, human confidence classification**

I. INTRODUCTION

In the context of this paper, human confidence (or self-confidence) is the confidence felt and expressed by an individual in a one-on-one discussion with an interviewer. As per the DeGroot–Friedkin model explained by Jia et al. in [1], an individual's self-confidence varies in a discussion having a sequence of topics. The state of confidence and the breathing process have an impact on each other. As seen in [2], breathing practises have helped pregnant women gain confidence while they experienced labour pain. Similarly, in [3], individuals with high self-rated apprehension are found to have more pauses, longer breath groups, and more interjections in their speech. To capture the breathing patterns, a dedicated sensor called the respiratory belt is connected to the chest, and its transducer converts the breathing-based chest movements into time-series breathing patterns. However, this is an intrusive mechanism and depends on expensive equipment. Hence, we propose the use of speech, which has the benefit of a non-intrusive approach to capturing the data.

II. PREVIOUS WORK

For the extraction of breathing patterns from speech, a variety of speech features are explored, such as Mel-frequency cepstral coefficients (MFCCs), energy, zero-crossing rate, and spectral slope in [4], cepstrograms in [5], and log Mel-spectrograms in [6]–[9]. The authors of [8] have also explored the use of the raw speech waveform fed to a deep network. In [6], 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. In [7], 40 healthy subjects' data is analysed for the detection of breathing rate using LSTM models, giving an r-value of 0.42. The Computational Paralinguistics challengE (ComParE) Breathing Sub-Challenge organised at Interspeech 2020 [10] had a baseline Pearson correlation of $r = 0.50$ on the development, and $r = 0.73$ on the test data set. The winners of this challenge [11] reported $r = 0.76$ between the speech signal and the corresponding breathing values of the test set. In all these studies, less than 50 subjects participated, and a maximum r-value of 0.76 is achieved between the breathing patterns and speech signals.

For the detection of confidence expressions from speech signals, Jiang and Pell analysed the impact of human confidence levels on the speech acoustics in [12], [13], and [14]. Speech parameters such as fundamental frequency, amplitude, speech rate, duration, and harmonic-to-noise ratio are found useful in classifying the confidence levels with an accuracy of 0.62 for the speaker-independent analysis. Joshua et al. in [15] validated the influence of vocal speed, intonation, and pitch on the perception of confidence expressed on more than 300 students' speech data. Specifically, increased speech rate, falling intonation, and lowered pitch

are found to indicate high speaker confidence. Sabu et al. in [16] have studied the confidence expressions among 195 children of age group $10-14$ years while reading a paragraph. The authors report an accuracy of $65\,\%$ for three-class classification and $82\,\%$ for binary classification (high and medium combined as high class) using acoustic features such as pause, pitch, and speech rate using a random forest regressor. This analysis is suitable for a specific context of evaluating the students' comfort with the language and not for assessing the self-efficacy of a speaker while responding spontaneously to an unknown scenario or question.

In this paper, we propose the novel approach of using speech-derived breathing patterns (SDBP) as a feature set for the classification of confident speakers from non-confident ones. MFCCs are the most widely used feature sets across all the analysis associated with speech. Auto-encoder-based representation is the most recent technique that has been adopted across multiple use cases of speech analyses. Hence, we compare the performance of SDBP with theirs.

III. DATA ACQUISITION

Figure 1 illustrates the separate collection of data for both tasks: deriving breathing patterns and capturing individuals' confidence expressions from speech.

III-A. Speech-Breathing Data Collection

We appointed 100 students (69 male;31 female) of the age group 18 to 23 years to participate in our study. ADInstruments' respiratory belt transducer is used for recording the ground truth breathing patterns, and a condenser microphone is used for recording the simultaneous speech signals. The PowerLab data acquisition system's two channels are connected to these two recording devices to capture the time-synchronised signals. The transducer is positioned on the chest (4 centimetres (cm) below the collarbone), and the head-mounted mic is placed at a distance of approximately 4 cm from the mouth. The participants are seated in a chair and given approximately 2 minutes to relax before starting the experiment. They read the List 2, List 3, List 7, List 8, List 9, and List 10 of the Harvard sentences. Harvard sentences are phonetically balanced sentences using specific phonemes at the same frequency as they appear in English [17]. Each participant takes around $2-3$ minutes to read these sentences. Both the signals, breathing and speech, are sampled at 40 kHz; speech is downsampled to 16 kHz and breathing to 50 Hz. Breathing values are divided by the maximum value to scale them in the range of -1 to $+1$.

III-B. Speech-Confidence Data Collection

A study is designed to collect data from 51 individuals in the age group $22-30$ years. The data collection happens over a video phone call with a sampling rate of 16 kHz. The candidates are briefed about the data collection procedure. We get their consent to record their audio-visual responses. An interview session with a candidate

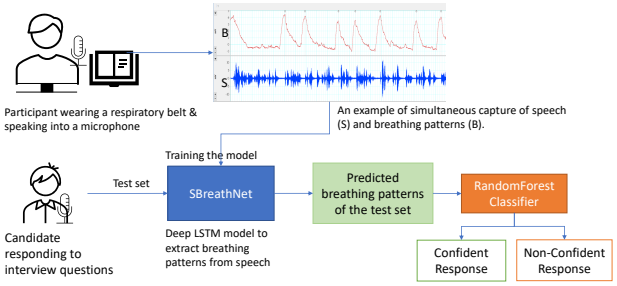

Fig. 1. An overview of the approach presented in this paper. SBreathNet: the deep regression model for the extraction of breathing patterns from speech signals. SDBP is used as a feature set for the classification of confidence classes.

comprises five questions. The questions are selected to induce varying levels of confidence, such as a question to "Describe yourself" (question number 1) to capture a confident response and a question about "What would you do in an unimaginable situation" (question numbers 4 and 5) to capture non-confident responses. The candidates do not know the questions before they participate in the session, and hence, spontaneous responses are captured from them. All the responses are labelled by the speakers themselves and three more researchers in two categories of confidence: – confident or non-confident–. The group of annotators has two females and one male; they are all in the age range of 30 to 40; conduct behavioural studies as their profession, and are Indians. The final label is calculated using a majority voting approach; there is one label for every audio-visual response. 51 individuals participated; hence, we get a total of (51 X 5) 255 responses, each with a duration ranging between $10-30$ seconds. For all the responses, at least two researchers' labels match those given by the candidates themselves. Hence, for all the responses, we get a majority vote of three out of four. There are 37 ($14\,\%$ of the total responses and $36\,\%$ of the responses to questions 4 and 5) non-confident responses and 218 confident responses. These statistics reflect the difficulty in capturing non-confident responses from the candidates in a study setup.

IV. METHODOLOGY

As shown in Figure 1, the deep regression model that extracts the breathing patterns from the speech signals is referred to as SBreathNet. This network is trained on the data described in Section III-A and infers the breathing patterns for the data mentioned in Section III-B. The SDBPs are then used as a feature set for the classification of confident and non-confident speech using a RandomForest classifier.

IV-A. Speech Representations

As depicted in Figure 2, four different speech representation techniques – MFCCs, time-domain features, phase-domain decomposed speech components, and the raw speech

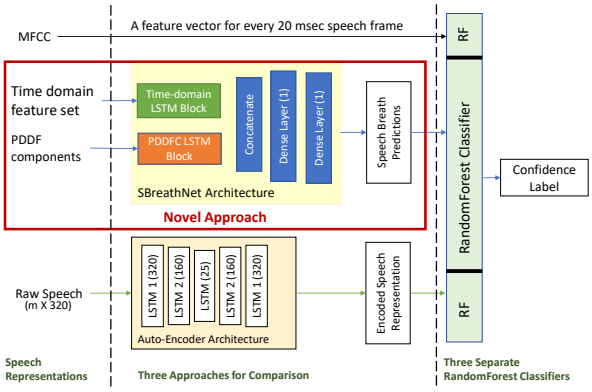

Fig. 2. Four different speech representation techniques: MFCCs, time-domain features, phase decomposed components, and raw speech frames, are used across three approaches for the classification of confident and non-confident responses. The center of the diagram explains the novel approach of using the deep regression network (SBreathNet) to extract breathing patterns from the speech signal. The auto-encoder architecture provides the representation for performance comparison. With the three approaches, confident and non-confident classifications are achieved using respective RandomForest (RF) classifiers.

frames – are used in three different approaches for the classification of confident and non-confident responses.

For every 20 miliseconds (ms) of speech frame, an MFCC feature vector of length 13 is calculated. Each $1X13$ MFCC feature vector is treated as a sample while feeding the RandomForest classifier. The ground truth for every speech response of around $3-4$ minutes is extended to each speech frame for the classification purpose.

For the extraction of breathing patterns from speech, the significance of the low-level time-domain features is discussed in [18]. Using these features, a Pearson correlation coefficient of 0.57 is achieved between the speech and the predicted breathing patterns of the ComParE dataset [10]. Among other speech parameters used for understanding the respiratory problems such as COVID-19 from human voice, MFCCs and the phase-domain decomposed filter components (PDDFC) of the speech signal are discussed for the classification of COVID-19 subjects from healthy subjects in [19]. We explored the time-domain features, MFCCs, and PDDFC for training SBreathNet. It is observed that the combination of time-domain features with PDDFC performs the best. Both features are calculated for every speech frame of 20 ms. Time domain features form a feature vector of length 16 comprising of: ZCR, kurtosis, RMS, auto-correlation, and 10 time domain difference features and PDDFC forms the feature vector of length 160.

To obtain an auto-encoder representation, raw speech frames of 40 ms duration are fed to the auto-encoder.

IV-B. Model Architectures

The centre of Figure 2 shows the SBreathNet architecture for the extraction of breathing patterns from speech signals. This network is trained using time domain features and PDDFC as input with a batch length of 250 corresponding to a duration of 5 seconds (a sample for every 20 ms is calculated; hence, 250 samples $= 250X20$ ms $= 5000$ ms). Both inputs are passed separately to corresponding LSTM blocks consisting of two LSTMs and a dense layer. The output of these two LSTM blocks are concatenated and fed to two consecutive dense layers. This forms the output of the SBreathNet. The loss function calculates the concordance correlation coefficient (CCC) loss between the true and predicted values. The network learns with a learning rate of 0.001 and with an Adam optimizer. The activation function of the last dense layer is a hyperbolic tangent (tanh) function. This causes the prediction values to range between $-1 and 1$.

The bottom part of Figure 2 shows the auto-encoder network architecture. The raw frames of duration 40 ms are normalised and are passed as an input to the auto-encoder network. With multiple configuration trials, we utilize four LSTM layers to capture the time-series nature of speech, followed by a final dense layer. Figure 2 illustrates the architecture. The input data has a dimension of mx320, where m denotes the number of 40 ms speech frames present in each response. After experimenting with several node sizes in the bottleneck layer, 25 nodes are found to perform the best in regenerating the input by the auto-encoder. The LSTM layers in the auto-encoder are fine-tuned to have a learning rate of 0.001 with an Adam optimizer. The loss function used calculates the Pearson's correlation coefficient (r) between the input and the re-generated output of the auto-encoder and returns '1-r' as the loss value. A batch size of one is used with a batch length of 25 to encode the speech of one second ($25X40$ ms) in one batch.

IV-C. Confidence Classification using a RandomForest classifier

All three feature sets – MFCCs, SDBPs, and auto-encoder representations – are fed to the RandomForest Classifier as shown in Figure 2. In the first approach, an MFCC vector represents a 20 ms speech frame, which is then fed to the RandomForest classifier. In the second approach, we get a breathing pattern of 5 seconds predicted for every 5 seconds of speech. These 5 seconds SDBPs are then fed to a RandomForest classifier as feature sets. The third approach presents an auto-encoder representation for every 1 second of speech frame. A batch size of one is used with a batch length of 25 to encode the speech of one second ($25X40$ ms) in one batch. These 1 second representations are then fed to the RandomForest classifier. The Random Forest algorithm is built with 100 trees and a maximum depth of 7.

Table I. Fold-wise performance of MFCCs, auto-encoder, and SDBPs using a RandomForest classifier.

	SDBPs			Auto-Encoder Representation			MFCC features		
#	AUC	Accuracy	Precision	AUC	Accuracy	Precision	AUC	Accuracy	Precision
1	57.8	67.2	55.2	53.3	68.1	55.5	48.9	61.8	48.7
2	95.2	95.0	94.8	95.2	95.3	95.3	93.4	93.2	93.0
3	93.4	93.8	93.6	93.6	94.4	94.7	91.0	91.6	91.3
4	63.0	65.7	63.6	56.7	63.3	60.2	53.1	59.1	54.4
5	68.4	70.4	77.5	53.9	57.4	61.6	51.7	55.3	55.2
Average	**75.6**	78.4	76.9	70.5	75.7	73.5	67.6	72.2	68.5

V. RESULTS

This section presents the results of the SBreathNet regression model. Further, we explain the confidence classification performance of the RandomForest classifier using SDBPs as the feature set and compare it with that of the MFCCs and auto-encoder representations.

V-A. Regression

The performance of the regression model, SBreathNet, is calculated using Pearson's correlation coefficient (r-value) as metric and three different loss functions: CCC, Huber loss, and MSE. An average r-value of 0.61, 0.55, and 0.55 is achieved across the 100 speakers with the loss functions CCC, Huber, and MSE respectively. The breaths-per-minute (BPM) count for every speaker is calculated on the predictions obtained with the three loss functions and compared with that of the true breathing pattern. A peak detection algorithm from scipy [20] is used for detection of peaks keeping a distance as 100 points and the height as 0.2. Using the peak count, further the BPM is calculated for each speaker. An average BPM error obtained is 2.50, 2.95, and 2.65 for the CCC, Huber, and MSE loss functions.

V-B. Classification

We use speaker independent training and validation partitions to improve the generalising capability of the RandomForest model. Speaker-independent analysis indicates that the speakers in the training and validation partitions are different and unseen. The results for all the models are calculated over five folds. The distribution of the data across these five folds is as shown in Table II; each fold is balanced for only training partition by performing augmentation by repetition.

Table II. Duration (in minutes) of confident and non-confident responses in the train and validation partitions in each fold of the 5-fold cross-validation.

	Train		Validation	
#	Confident	Non-confident	Confident	Non-confident
1	29	19	11	4
2	38	23	10	7
3	35	23	12	7
4	32	18	8	5
5	33	18	7	6

As seen in Table I, the SDBPs exhibit a highest AUC of 75.6 % averaged across five folds of the data. When compared with MFCCs and the auto-encoder representation based classification, SDBPs outperform in all other metrics as well. Specifically, SDBPs exhibit an AUC that is higher than that of MFCCs and the auto-encoder representation by an absolute value of around 8 % and 5 % respectively. The SDBPs when fused with the auto-encoder representation gives 71.7 % AUC across the five folds, which is an average of the performance exhibited by the two feature sets individually. This strengthens the contribution of SDBPs as the feature set for confidence level classification.

VI. DISCUSSION

To further understand the classification performance of the SDBPs, we have calculated the average representation for confident and non-confident classes. As seen in Figure 3, the depth of the breathing pattern is found to differ between the confident and non-confident speakers. Non-confident speakers exhibit deep breaths as they also take longer pauses while speaking. However, the confident speakers are found to have shallow breaths. From the average breaths per minute calculated for both the classes the confident class is found to have an absolute increment of 2 breaths per minute on an average when compared with the non-confident class.

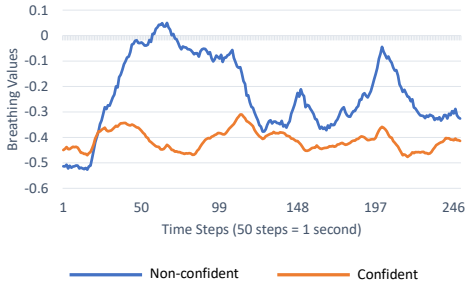

Fig. 3. The average breathing patterns for the confident and non-confident classes.

VII. CONCLUSION AND FUTURE WORK

We conclude that speech-derived breathing patterns not only perform better in automatically classifying confident and non-confident speech responses, but also help in understanding the rationale. We presented an empirical evidence of enhancement in performance by using the proposed feature

set over MFCCs and auto-encoder representations. In future work, we intend to extend this analysis to other behavioural parameters such as emotions, stress, and anxiety.

VIII. REFERENCES

[1] P. Jia, A. MirTabatabaei, N. E. Friedkin, and F. Bullo, "Opinion dynamics and the evolution of social power in influence networks," *SIAM review*, vol. 57, no. 3, pp. 367–397, 2015.

[2] V. Campbell and M. Nolan, "'it definitely made a difference': A grounded theory study of yoga for pregnancy and women's self-efficacy for labour," *Midwifery*, vol. 68, pp. 74–83, 2019.

[3] "Acoustic characteristics of public speaking: Anxiety and practice effects," *Speech Communication*, vol. 53, no. 6, pp. 867–876, 2011.

[4] 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.

[5] 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), pp. 1–5, IEEE, 2019.

[6] 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), pp. 4110–4114, INTERSPEECH Communication Association (ISCA), 2019.

[7] 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), pp. 1140–1144, IEEE, 2020.

[8] 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.

[9] 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 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1345–1349, IEEE, 2021.

[10] 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*, (Shanghai, China), pp. 2042–2046, INTERSPEECH Communication Association (ISCA), 2020.

[11] 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 the 21st Annual Conference of the International Speech Communication Association*, (Shanghai, China), pp. 2072–2076, INTERSPEECH Communication Association (ISCA), 2020.

[12] X. Jiang and M. D. Pell, "Encoding and decoding confidence information in speech," in *Proceedings of the 7th international conference in speech prosody (social and linguistic speech prosody)*, vol. 5762579, 2014.

[13] X. Jiang and M. D. Pell, "The sound of confidence and doubt," *Speech Communication*, vol. 88, pp. 106–126, 2017.

[14] X. Jiang and M. D. Pell, "Predicting confidence and doubt in accented speakers: Human perception and machine learning experiments," in *Proceedings of Speech Prosody*, pp. 269–273, 2018.

[15] J. J. Guyer, L. R. Fabrigar, and T. I. Vaughan-Johnston, "Speech rate, intonation, and pitch: Investigating the bias and cue effects of vocal confidence on persuasion," *Personality and Social Psychology Bulletin*, vol. 45, no. 3, pp. 389–405, 2019.

[16] K. Sabu and P. Rao, "Automatic prediction of confidence level from children's oral reading recordings.," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association*, (Shanghai, China), pp. 3141–3145, INTERSPEECH Communication Association (ISCA), 2020.

[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.," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association*, pp. 931–935, INTERSPEECH Communication Association (ISCA), 2021.

[19] G. Deshpande and B. W. Schuller, "Covid-19 biomarkers in speech: on source and filter components," in *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 800–803, IEEE, 2021.

[20] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, *et al.*, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.