Emotion Recognition and Spoof Detection from Whispered Speech

Dawn Sivan
PG Scholar
College of Engineering Karunagappally
Karunagppally, Kerala, India
dawnsivan91@gmail.com

Gopakumar C.
Assistant Professor
College of Engineering Karunagappally
Karunagppally, Kerala, India
gopan.c.menon@gmail.com

***Abstract*— Speech Emotion Recognition is used to provide a more natural interaction for Human Computer Interface. Whispered speech is used when confidential matters are discussed. Usually the magnitude information of the spectrum dominates in the features of speech emotion recognition. This paper investigates the effectiveness of time, frequency and phase features for emotion recognition from whispered speech. The various features obtained from the speech sample are energy, entropy, statistical parameters, autocorrelation, zero crossing rate, modified group delay and all-pole group delay function. The resulting feature vector is used to train a Support Vector Machine. Here, a database containing whispered speech of different speakers with different emotions is created to demonstrate the potency of the proposed method.**

Keywords—whisper; speech; emotion recognition; spoof detection.

I. INTRODUCTION

Speech is a complex signal. It contains knowledge about the message, language, speaker, and emotions. Emotion is the mental state of an individual that, rather than through conscious effort arises spontaneously. Speech emotion recognition basically identifies the physical as well as emotional state of human being from his or her voice. It is devoted to increase the user-friendliness and provide a more natural interaction experience for human-computer interaction [1]. Emotion recognition from computer is still a challenging issue in the field of Human Computer Interface (HCI) when recognition is based solely on voice which is the basic mean of human communication. Such a system can be misused by modifying the speaker's voice and thereby introducing spoof. In order to protect a speaker verification system against a spoofing attack and make it more robust, one may implement a synthetic speech detection module that discriminates between original and disguised speech.

Whispered speech is a common communication approach in the libraries, some telephone conversations, meeting room, and so on where loud voice cannot be used. The research on whispered speech is gradually developing in recent years. Communication by whispered speech is of vital importance for patients with disabilities who are affected by disease of the vocal system such as functional aphonia or laryngeal disorders [1]. This work focuses on recognizing emotions and detecting spoof from whispered speech.

The most important preparation for automatic classification and recognition of emotions is to select a proper feature set as a description to the emotional speech.

II. LITERATURE REVIEW

Emotion recognition from speakers is obtained through processing methods including the isolation of the speech signal and extraction of elite features for the final classification [2]. In some cases, the process is helped by speech recognition systems, which contribute to classification using linguistic data. But emotional states do not have clear-cut bounds and often differ from person to person.

In [3], vector features are categorized as segmental (short-time) or supra-segmental (long-time) according to their temporal structure. Segmental features are estimated once for every small time frame. Supra-segmental features are determined over the entire utterance duration. These include spectral and prosodic features, such as pitch (or fundamental frequency), formants, energy, Mel Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstral Coefficients (LPCC), speaking rate, voice quality parameters, shimmer, jitter, while functionals include their statistics such as mean, maximum, minimum, change rate, kurtosis, skewness, zero-crossing rate, variance and so on.

Normally, the magnitude content is retained to derive acoustic features needed for speech emotion recognition [4]. The most frequently used acoustic features Mel-Frequency Cepstral Coefficients (MFCCs), which have been constantly proven useful in speech emotion recognition and speech recognition, are derived by applying a nonlinear Mel-scale filter bank on the power spectrum. [5] presented a minimum set of voice parameters for affective computing by using the first four MFCCs. Their work is based on an automatic extraction system which extracts an acoustic parameter set from an audio waveform without manual interaction or correction.

[6] gives a new spectral representation of speech signals through group delay functions. The group delay functions are noisy by themselves and difficult to interpret due to zeros that are close to the unit circle in the z-domain. These clutter the spectra. A new modified group delay function that reduced the effects of zeros close to the unit circle is used. The group delay function is undefined if the roots of the transfer function (poles or zeros) are on the unit circle. Since speech is the output of a stable system, roots that are close to the unit circle are only zeros. These zeros are due to either the use if an analysis window to truncate the speech signal or from noise. They are not of much interest in speech analysis, since in speech, zeros only occur in the production of nasals. The modified group delay function suppresses the zeros that are close to the unit circle. They show that the modified group delay can be used in conjunction with the standard MFCC-based feature in recognition.

[7] advocates the use of group delay functions derived from parametric all-pole models instead of their direct computation from the discrete Fourier transform. The presence of spurious high-amplitude spikes in the group delay function makes its processing difficult. In this paper, they tackle this problem by utilizing the group delay function from all-pole models of speech, formed by linear predictive analysis. Their goal was to utilize features from group delay functions of parametric all-pole models of speech signals. Feature extraction, or front-end, is a critical component in any speech processing system. It uses group delay functions for feature extraction in speaker recognition. Processing of group delay functions is not straightforward. The presence of zeros of the vocal tract system function (in the Z-transform representation) can lead to an ill-behaved group delay function. It can be seen that a zero (or dip in the spectrum) will lead to an indefinitely large value of the group delay function. These zeros can occur as a result of the excitation source, and can also be an artifact of short-term processing. Estimation of the group delay function at frequency bins near these zeros results in high amplitude peaks, hiding out the formant structure. The chirp group delay function avoids the zeros near the unit circle by evaluating the spectrum on a circle other than the unit circle. Linear prediction analysis of speech approximates the speech spectrum using an all-pole model. Considering the vocal tract as an all-pole filter allows an equivalent representation as a cascade of several second-order and first order all-pole filters. In this representation, the overall magnitude spectrum is the product of the magnitude spectra of the individual filters. The overall phase spectrum is a summation of the individual phase spectra. In spite of this lossy representation, the all-pole group delay function has valuable information which makes it suitable for speaker recognition. The high resolution property of the group delay spectrum may be helpful in capturing formant information in a more robust manner, compared to the magnitude spectrum. The higher order formants are more pronounced in the group delay spectrum, particularly in the low and high vocal effort conditions. Utilizing parametric all-pole models provide an effective mechanism to extract information from group delay functions, which otherwise suffer from signal processing difficulties. Thus, group delay features from all pole models can be used to effectively process phase information for speaker recognition.

[8] has demonstrated the usefulness of phase-based features for whispered speech emotion recognition. Such work made use of the modified group delay feature with a Fisher kernel and a linear kernel Support Vector Machine (SVM) for improved speech emotion recognition.

[1] looks into the phase-based features for whispered speech emotion recognition. In addition to the modified group delay feature, another phase-based feature derived by the group delay function of all-pole models is investigated. Besides, they propose a novel SVM-based speech emotion recognition framework that enjoys the benefit of SVMs. They adopt a linear kernel SVM to train the emotion recognition model with the resulting outer product vectors.

These aforementioned examples suggest that, incorporating the phase information along with the time and frequency parameters can extend the horizon for audio and speech signal processing beyond the current limit of phase independent solutions employed for long time by speech scientists. However, in the domain of speech emotion recognition, there exists very little research with respect to phase-based features. We extend [1] by including time and frequency features along with phase features for whispered speech emotion recognition and spoof detection.

III. METHODOLOGY

The proposed system generally consists of four major modules, as shown in Fig. 1, including a *denoising module, feature extraction module, normalization module* and *a linear kernel SVM module.*

This work employs wavelet denoising to make the signal free from noise. The time features like autocorrelation and zero crossing; frequency features such as energy, entropy, variance, standard deviation, maxima and minima; and the phase features namely the modified group delay and all pole group delay features are considered. To ease the way to build an efficient emotion classifier, SVMs, which have been found very powerful in a wide range of applications, are chosen as the back-end model. Besides, L2 normalization is considered to correct the values of the resulting feature vector to an appropriate range for the SVM classifier.

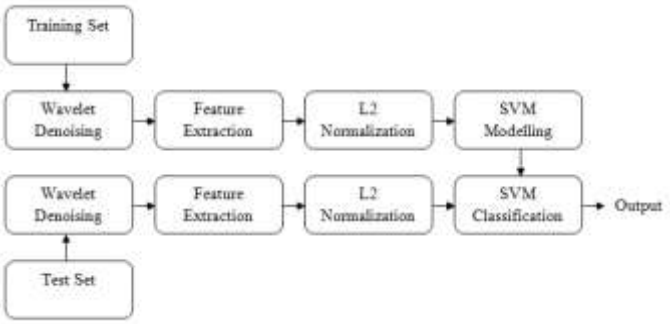

Fig. 1. Block scheme of the proposed speech emotion recognition system

A. Wavelet Denoising

A signal is affected by noise during acquisition or transmission, which results in a decrease in performance. Removing of the noise from the signal facilitate the processing. Hence, a denoising process is performed to remove the noise while retaining and not distorting the quality of processed signal. In this work, Symlet4 wavelets are used with a level of decomposition of $N = 2$.

B. Feature Extraction

This work exploits the frequency, time and phase features. The frequency features such as entropy, variance, standard deviation, maxima and minima are used. Entropy is a statistical measure of randomness. A voiced region of speech would induce low entropy since there are clear formants in the region. The entropy has been used to detect silence and voiced region of speech in voice activity detection. The discriminatory property of this feature gives rise to its use in speech recognition. The entropy can be used to capture the formants or the peakiness of a distribution. Formants and their locations have been considered to be important for speech tracking. The standard deviation is a measure of how far the signal fluctuates from the mean. The standard deviation is similar to the average deviation, except the averaging is done with power instead of amplitude. The variance represents the power of this fluctuation. Maxima and minima give the largest and smallest values of the signal.

The time features used are autocorrelation and short time zero crossing. Autocorrelation observes how similar the signal characteristics with respect to time. It is achieved by providing different time lag for the sequence and computing with the given sequence as reference. Zero Crossing Rate (ZCR) gives information about the number of zero-crossings present in a given signal. If the value is more in a given signal, then the signal is changing rapidly, and hence has high frequency information. If the value is less, then the signal changes slowly and has low frequency information. Thus ZCR gives indirect information about the frequency content of the signal.

The phase features used here are modified group delay and all pole phase modeling. The Fourier transform of a discrete time digital signal $x(n)$ can be computed in the polar form as:

$$X(\omega) = |X(\omega)|e^{j\phi(\omega)} \quad (1)$$

where $|X(\omega)|$ and $\phi(\omega)$ are the magnitude and phase spectrum.

The group delay function is derived as the negative derivative of the Fourier phase spectrum. It can be explicitly written as follows:

$$\tau_g(\omega) = -\frac{d(\phi(\omega))}{d\omega} \quad (2)$$

$$= \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \quad (3)$$

where ω is the angular frequency, $|X(\omega)|$ is the magnitude of the Fourier transform of $x(n)$, $Y(\omega)$ is the Fourier transform of the signal $y=nx(n)$ and the subscripts R and I indicate real and imaginary parts, respectively.

The features derived by the group delay function are discriminative and additive for recognition. But the function often leads to an erroneous representation of a given speech signal [1]. To overcome this, a modification of the group delay function is used.

It is computed as

$$\tau_m(\omega) = \frac{\tau_p(\omega)}{|\tau_p(\omega)|}|\tau_p(\omega)|^{\alpha} \quad (4)$$

where

$$\tau_p(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^{2\gamma}} \quad (5)$$

and $|S(\omega)|$ is a cepstrally smoothed form of $|X(\omega)|$. The two tuning parameters α and γ control the range dynamics of the MGD spectrum. Note that,

$$P(\omega) = X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega) \quad (6)$$

called the product spectra, includes information from both the magnitude and phase spectrum.

The speech signal is first being pre-emphasised and then framed by a Hamming window. Afterwards, the MGD features are computed. Finally, the DCT is applied on the MGD features so as to perform a decorrelation. In general, the first coefficient obtained by the DCT is excluded to avoid the effects of the average value.

The all pole group delay (APGD) is capable of interpreting properly the phase information. Unlike the MGD feature computing the group delay function directly from the signal, the APGD feature is rooted in group delay functions of parametric all-pole models of speech signals. Linear prediction analysis approximates the short-term power spectrum by means of all-pole models. It is formulated as

$$H(\omega) = \frac{G}{1 - \sum_{k=1}^{p} a(k) e^{-j\omega k}} \qquad (7)$$

where $a(k)$ are the prediction coefficients, G is the signal dependent gain, and p is the prediction order. The coefficients of $a(k)$ are calculated by minimizing the least square errors between the power spectrum of $H(\omega)$ and the speech power spectrum $|X(\omega^2)|$. Here, the gain G is simply set to 1 for normalization purpose.

C. Normalization

L2 normalization is investigated to reduce the dependence on the amounts of individual speaker information and improve the performance as long as classifiers (e. g., SVMs) involve dot-products. The L2 normalization is written as

$$l^2(x) = \frac{x}{\sqrt{\sum_{i=1}^{d} x_i^2}} \qquad (8)$$

It is easier to calculate derivatives of the L2 norm as it squares each vector component (compared to L1, which uses absolute values).

IV. RESULTS AND DISCUSSION

A. Database

Here, we created a database of the speech of six actors, both male and female, selected randomly, in different emotions, to evaluate the effectiveness of the proposed system. The emotions considered are anger, fear, happy, neutral and sad. The actors were requested to express three sentences each of all emotions. The speech inputs were recorded using free software, AUDACITY. As a result, the database contains 90 instances in total.

B. Experimental Setup

For the feature extraction, wavelet denoising and pre-emphasis are first conducted. Afterwards, frame windowing is performed using a Hamming window with a frame-length of 25 ms and a frame-shift of 10 ms. When computing the MGD features, we set the two tuning parameters and to 0.1 and 0.2 based on the preliminary experiments, which ends up in 36-dimensional MGD features including delta and acceleration coefficients. As for the APGD features, the order of the all-pole mode p is 30, and 18 APGD coefficients are kept. Delta and acceleration coefficients are appended to the APGD coefficients to form 54-dimensional APGD feature vectors. In addition to the two phase-based features, the most frequently used time and frequency features are also extracted. As for the basic supervised learner in the classification step, we use linear SVMs implemented in MATLAB R2013a.

TABLE I. CONFUSION MATRIX OF THE PROPOSED METHOD IN PERCENTAGE

Emotion	Anger	Fear	Happy	Neutral	Sad
Anger	**75**	4	8	5	8
Fear	0	**75**	0	12	13
Happy	4	0	**84**	8	4
Neutral	0	5	0	**79**	16
Sad	0	5	8	8	**79**

C. Results

We now investigate whether the proposed feature combination has lead to performance improvement. The corresponding confusion matrix for the feature combination on the created database is given in Table 1. Here, the following emotions are considered: Anger, Fear, Happy, Neutral and Sad.

The results show that combining the frequency features with the phase-based features together can improve the performance of emotion recognition. Moreover, using the time based features, speech spoof detection has been found to have an accuracy of 78.18%. To assess the overall performance of the whispered emotion recognition by the proposed method, parameters namely precision (positive predictive value), recall (sensitivity or true positive rate) and specificity (true negative rate) are compared to the existing method.

The corresponding values obtained for the proposed and existing methods are tabulated in Table 2 and represented graphically as in Figure 2.

TABLE II. PERFORMANCE COMPARISON OF PROPOSED METHOD WITH EXISTING METHOD

	Precision	Recall	Specificity
Existing	0.5500	0.5410	0.5000
Proposed	0.7833	0.7899	0.7851

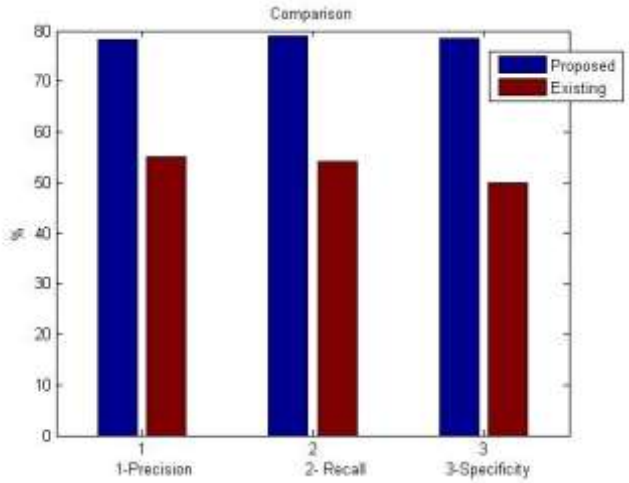

Fig. 2. Graphical Representation of Performance Characteristics of Proposed and Existing Method

V. CONCLUSION

Here, we focused on improving whispered speech emotion recognition along with spoof detection. We presented a novel framework by exploiting the effectiveness of the time, frequency and phase features. Experiments on a created whispered speech database were conducted, demonstrating that the present framework is competitive with or superior to other models using phase features alone.

Besides, future work can extend the proposed method to tasks such as speaker identification and verification.

References

[1] J. Deng, X. Xu, Z. Zhang, S. Fruhholz, and B. Schuller, "Exploitation of Phase-based Features for Whispered Speech Emotion Recognition", IEEE Access, Vol. 4, pp. 4299-4309, July 2016.

[2] C. N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," Artificial Intelligence Review, vol. 43, no. 2, pp. 155–177, Nov 2012.

[3] Schuller B. and Rigoll G., "Recognising Interest in conversational Speech–Comparing Bag of Frames and Supra-Segmental Features", Proceedings of INTERSPEECH, pp 1999–2002, 2009.

[4] P. Dumouchel, N. Dehak, Y. Attabi, R. Dehak, and N. Boufaden, "Cepstral and Long-term Features for Emotion Recognition" in Proc. INTERSPEECH, Brighton, UK, pp. 344–347, 2009.

[5] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190–202, July 2016.

[6] H. Murthy and V. Gadde, "The Modified Group Delay Function and Its Application to Phoneme Recognition," Proc. ICASSP, Hong Kong, China, pp. 68–71, May 2003.

[7] P. Rajan, T. Kinnunen, C. Hanilci, J. Pohjalainen, and P. Alku, "Using Group Delay Functions From All-Pole Models for Speaker Recognition" in Proc. INTERSPEECH, Lyon, France, pp. 2489–2493, 2013.

[8] J. Deng, X. Xu, Z. Zhang, S. Fruhholz, and B. Schuller, "Fisher kernels on Phase-based Features for Speech Emotion Recognition," in Proc. IWSDS, Saariselk¨a, Finland, pp. 1-6, 2016.