VOL. 13, NO. 9, MAY 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

SPEAKER INDEPENDENT EMOTION RECOGNITION FROM SPEECH SIGNALS

B. Rajasekhar[1], M. Kamaraju[2] and V. Sumalatha[3]
[1]JNTUA, Anantapuramu, India
[2]Department of Electronics and Communication Engineering, Gudlavalleru Engineering College, Gudlavalleru, India
[3]Department of Electronics and Communication Engineering, JNTU College of Engineering, Anantapuramu, India
Email: surajb2000@gmail.com

ABSTRACT

Speech carries vast information about age, gender and the emotional state of the Speaker. Emotion Recognition is a difficult task of identifying a specific emotion from a speaker. In this work the effect of Discrete wavelet Transform (DWT), Cepstrum, Pitch and Mel-Frequency Cepstral Coefficients (MFCC) are considered in the detection of emotions and later the samples are trained and tested for recognizing the specific emotion. The data base considered is Telugu-Data Base which is prompted by two speakers male and female which contains four emotions Happy, Angry, Sad and Neutral. Various combinations of features are performed to recognize the corresponding emotion and these features are referred as Emotion-specific features. By considering these features combination recognition rate is increased. Features DWT, Cepstrum, MFCC and Pitch are used to extract the feature information. After feature extraction classification is performed by back-propagation neural network algorithm and later the performance is evaluated.

Keywords: emotion recognition, DWT, cepstrum, MFCC, neural network.

1. INTRODUCTION

A remarkable study is being done in the current years for improving human machine interaction on Speech Emotion Recognition (SER). The major issues in handling with analysis of emotions in speech are details of emotion, collection of data and feature representation [1]. The fundamental states of a speaker are emotions and the outgoing values of the fundamental states are emotion expressions [2]. The features which are used for investigation are prosody, voice quality and spectral features, these carry emotion correlates and these features are speaker and sound-specific [1]. Pattern recognition models like GMMs, SVMs, ANNs and HMMs are educated on the features extracted. Majority of the feature representations are speaker and sound-specific, hence the above said pattern recognition models require a phonetically/phonemically balanced data, covering several speakers. In realistic sense, it is complicated to gather such a database [1].

In several studies [3, 4, 5], it is found that there is some uncertainty among angry and happiness emotions. The results with respect to accuracy for emotion recognition systems designed and tested using simulated parallel corpus [4, 5, 6] is better when compared semi-natural or natural databases [7]. Fundamental frequency (*F*0) of emotional speech was the basis for previous studies [8].

The effect of Discrete Wavelet Transform (DWT) along with Cepstral coefficients in the recognition of emotions is performed and also a relative analysis of DWT, Cepstum, Mel-Frequency Cepstral Coefficients (MFCC) and pitch coefficients on emotion classification is done. Usage of compact feature vector, algorithm resulted improved recognition rates in identifying emotions from Telugu speech corpus with good accuracy. The method in [9] has produced a substantial decrease in the misclassification efficiency which is better than the algorithm said by [10], where only synthetically enlarged MFCC coefficients are considered as feature vector.

Feature extraction will play a significant role while extracting features [11, 12]. In order to extract features there are several analysis. In this work Spectral feature MFCC is considered for high accuracy, for

Prosody feature Pitch is considered and concatenating with other two different features Cepstrum and DWT and this combination is called as Emotion-Specific Feature Set.

2. EMOTION RECOGNITION

In identifying the Emotional state of a speaker the block diagram shown in Fig. 1 is considered. Steps to perform the emotion recognition are Speech Pre-Processing, Feature Extraction and Classification.

2.1. Speech pre-processing

Pre-Processing on a given speech signal is performed to avoid un-voiced segments in the speech signal which will improve the accuracy of feature extraction. This pre-processing is done by using an energy equation (1).

$$y(n) = \sum x^2 \quad (1)$$

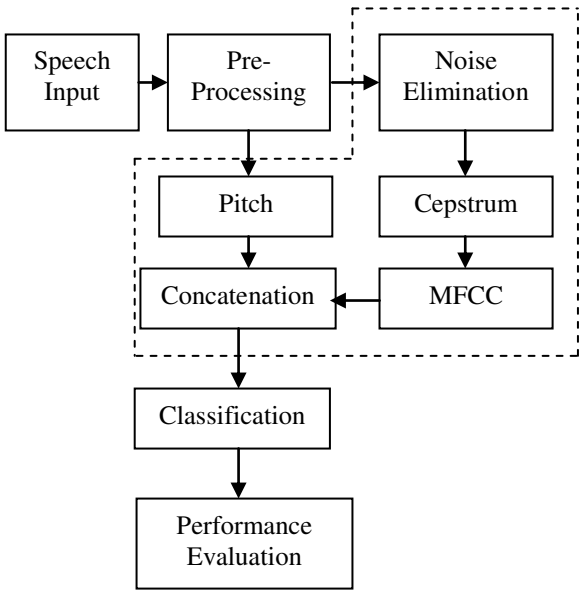

Figure-1. Block diagram of Emotion Recognition..

2.2. Feature extraction

The features extracted are from the pre-processed speech signal. The features extracted are DWT, Cepstrum, MFCC and Pitch. After the features are extracted different combinations of features are considered for classification.

2.2.1 Noise elimination using DWT

Input to Noise elimination is pre-processed speech signal and output is de-noised signal. Steps considered for eliminating the noise from the signal are apply DWT to Pre-Processed signal, assign a threshold limit using Soft Thresholding and apply Inverse Discrete Wavelet Transform (IDWT).

2.2.1.1 Discrete Wavelet Transform (DWT)

DWT plays a crucial role in evaluation of emotion performance. Here DWT is used to eliminate noise in speech signal. The type of wavelet used is 'db4' because of its high accuracy [13].

2.2.1.2 Assign threshold limit using soft thresholding and IDWT

Noise eliminating with the help of soft Thresholding is also referred as wavelet decomposition. To select threshold value equation (2) is considered

$$t = \sigma\sqrt{2}\log(N) \quad (2)$$

Where σ, standard deviation of noise and N, length of the signal. σ is obtained using equation (3)

$$\sigma = MAD/0.6745 \quad (3)$$

Where MAD is median of absolute values of wavelet coefficients.

Soft Thresholding is considered rather than hard Thresholding because it will reduce the noise without effecting the edges information. Hard Thresholding will shrink the wavelet coefficients because of which information will be lost. Hence, for this purpose soft Thresholding is treated as wavelet decomposition. The soft Thresholding limit is obtained by equation (4).

$$Y_{soft} = \begin{cases} sign(x)(|x| - \tau) & if\ |x| \geq \tau \\ 0 & if\ |x| \geq \tau \end{cases} \quad (4)$$

2.2.2 Cepstrum

Input to Cepstrum is a de-noised signal and output is Cepstrum coefficients. Cepstrum is used to find the information in between samples only; it will eliminate first few samples. Steps to calculate Cepstrum are first apply Fast Fourier Transform (FFT) to the de-noised signal, then apply natural logarithm on the FFT coefficients and then perform Inverse FFT.

2.2.3 MFCC

MFCC will give the information of first few coefficients only. So, the Cepstrum and MFCC are combined to get whole samples in the speech. MFCC processing steps are framing and Windowing, FFT, find absolute values, Mel-Scaled Filter Bank, apply Natural logarithm and apply Discrete Cosine Transform (DCT).

2.2.4 Pitch

Pitch is the fundamental frequency or the lowest frequency of the sound signal. In speech signal the pitch levels may be different, each emotion will have different intonations and it may affect the meaning of the speaker. Pitch Processing is applied on the pre-processed speech signal for obtaining the fundamental frequency and the steps performed are scaling the frequency, windowing, FFT, find absolute values and apply natural logarithm.

2.3 Classification

Back Propagation neural network algorithm is considered for classification of emotions. Back-propagation is a training method used for a multi layer neural network. The neural network training by back propagation has three stages namely feed-forward of input training pattern, back propagation of the associated error and adjustments of the weights.

3. IMPLEMENTATION RESULTS

The type of data base considered is Semi-natural data-base composed from both male and female speakers in telugu language, which consists of four emotions (angry, happy, neutral and sad), and it is named as Telugu data base. The sequence of training and testing Patterns will be same, for training sequence a total of 141 samples are considered with each emotion samples as, angry 41 samples, happy 25 samples, Neutral 40 samples, and sad 35 samples. To identify individual emotion and its performance single evaluation may not give accurate results. So, one and three iterations are considered to calculate accuracy with different feature sets.

Table-1. Emotion recognition rate with different Iteration and Nodes of the Network.

No. of iterations	**One**			**Three**		
No. of nodes	10	20	50	10	20	50
Emotion recognition rate (%)	77.64	82.7	93.96	74.87	82.91	92.33

Table-1 shows the increase in the emotion recognition rate by increasing the iterations and nodes of the network. Hence the further results are considered for three iterations with 50 nodes because of their better performance.

Table-2. Over all emotion recognition rate.

Feature set (s)	**Emotion recognition rate** (%)
DWT	64.74
Pitch	67.18
Cepstrum (CEPS)	63.16
MFCC	78.37
DWT+Pitch	66.72
DWT+CEPS	69.72
DWT+MFCC	88.64
DWT+MFCC+CEPS	89.17
DWT+MFCC+Pitch	92.87
DWT+Pitch+CEPS	95.98
DWT+CEPS+MFCC+Pitch	93.32

By considered different feature set combinations with three iterations and 50 nodes, it is observed that emotion recognition rate is low when individual feature set is considered where as for the combination of feature sets the recognition rate obtained is high as given in Table-2. It is also shown from Table 2, that the recognition rate is very high for the feature sets combinations DWT + Pitch + Cepstrum and DWT + Cepstrum + MFCC + Pitch.

Table-3. Recognition rate of each emotion.

Emotions	**Emotion recognition rate** (%) (DWT+Cepstrum+ MFCC+Pitch)	**Emotion recognition rate** (%) (DWT+Pitch+ Cepstrum)
Anger	94.59	95.40
Happy	92.62	95.82
Neutral	93.97	96.45
Sad	94.66	96.26

Table-3 gives the comparison of emotion recognition rate of each emotion for two feature set combinations. By observing Table-3 the individual emotion recognition rate for the feature set combination DWT+Pitch+Cepstrum is high when compared with DWT+Cepstrum+MFCC+Pitch. This is because complete set of MFCC coefficients are not considered to avoid the computation complexity.

4. CONCLUSIONS

Different combinations of feature sets are used to identify the corresponding emotion and these feature sets are referred as Emotion-Specific features. These are DWT, Cepstrum, MFCC and Pitch which are used to extract the features information. Artificial neural network is used for classification after feature extraction. It is concluded that by increasing number of nodes in the network and number of iterations the recognition rate is above 90%, the combinations of feature sets will give better emotion recognition rate than individual feature sets and the feature set combination DWT+Pitch+Cepstrum produced the individual emotion recognition rate above 95%.

REFERENCES

[1] P. Gangamohan, S.R. Kadiri, B. Yegnanarayana. 2016. Analysis of Emotional Speech - A Review, Springer International Publishing Switzerland, A. DOI 10.1007/978-3-319-31056-5_11.

[2] Darwin C. 1872. The expression of emotion in man and animals. Reprinted by University of Chicago Press, Murray, London, UK (1975).

[3] Ververidis D, Kotropoulos C. 2005. Emotional speech classification using Gaussian mixture models. In: International symposium on circuits and systems. Kobe, Japan. pp. 2871-2874.

[4] Grimm M, Kroschel K, Mower E, Narayanan S. 2007. Primitives-based evaluation and estimation of emotions in speech. Speech Commun. 49(10-11):787-800.

[5] Lugger M, Yang B. 2007. The relevance of voice quality features in speaker independent emotion recognition. In: ICASSP, vol 4. Honolulu, Hawaii, USA. pp. 17-20.

[6] Yeh L, Chi T. 2010. Spectro-temporal modulations for robust speech emotion recognition. In: INTERSPEECH. Chiba, Japan. pp. 789-792.

[7] Lee C, Mower E, Busso C, Lee S, Narayanan S. 2011. Emotion recognition using a hierarchical binary decision tree approach. Speech Commun. 53(9-10): 1162-1171.

[8] Wu S, Falk TH, Chan W. 2011. Automatic speech emotion recognition using modulation spectral features. Speech Commun. 53(5): 768-785.

[9] S Lalitha, D Geyasruti, R Narayanan, M Shravani. 2015. Emotion Detection Using MFCC and Cepstrum Features. In: Procedia Computer Science. 70(2015): 29-35.

[10] Inma Mohino-Herranz, Roberto Gil-Pita, Sagrario Alonso-Diaz and Manuel Rosam Zurera. 2014. MFCC Based Enlargement of the Training set for Emotion Recognition in Speech. International Journal (SIPIJ). 5(1).

[11] B Rajasekhar, M Kamaraju, V Sumalatha. 2016. Gender Driven Emotion Recognition System for Speech Signals Using Neural Networks. Proceedings of 6th International Advanced Computing Conference (IACC 2016), IEEECS.

[12] B Rajasekhar, M Kamaraju, V Sumalatha. 2017. FPGA Based Recognition of Emotions from Speech Signals. Proceedings of IEEE sponsored 3rd International Conference on Biosignals, Images and Instrumentation (ICBSII 2017). pp. 105-107.

[13] Firoz Shah.A, Vimal Krishnan V.R, Raji Sukumar. A, Athulya Jayakumar, Babu Anto. P. 2009. Speaker Independent Automatic Emotion Recognition from Speech:-A Comparison of MFCCs and Discrete Wavelet Transforms, 2009 International Conference on Advances in Recent Technologies in Communication and Computing. pp. 528-531.