Emotion Recognition using Dynamic Time Warping Technique for Isolated Words

N. Murali Krishna[1] **P.V. Lakshmi[2]** **Y. Srinivas[3]** **J. Sirisha Devi[4]**

1. Dept of CSE, GITAM University
Vishakhapatnam, 530045, Andhra Pradesh, India

2. Dept of IT, GITAM University
Vishakhapatnam, 530045, Andhra Pradesh, India

3. Dept of IT, GITAM University
Vishakhapatnam, 530045, Andhra Pradesh, India

4. Dept of CSE, CMRCET, JNTU (H)
Hyderabad, 501401, Andhra Pradesh, India

Abstract

Emotion recognition helps to recognize the internal expressions of the individuals from the speech database. In this paper, Dynamic time warping (DTW) technique is utilized to recognize speaker independent Emotion recognition based on 39 MFCC features. A large audio of around 960 samples of isolated words of five different emotions are collected and recorded at 20 to 300 KHz sampling frequency. Training and test templates are generated using 39 MFCC features. In the proposed work, we have extracted the MFCC coefficients from the speech database and DTW is used to store a prototypical version of each word in the vocabulary and compute incoming emotion with each word. For the classification of emotions SVM is used. The experimental results are provided using MFCC, Delta Coefficients (ΔMFCC) and Delta Delta Coefficients(ΔΔMFCC) . It is proposed that higher recognition rates can be achieved using MFCC features with DTW which is useful for different time varying speech utterances.

Keywords: Dynamic Time warping (DTW), MFCC (Mel frequency cepstral coefficient), Feature extraction, SVM

I. Introduction

Emotion recognition(ER) has made great strides with the development of digital signal processing hardware and software. But despite of all these advances, machines cannot match the performance of their human counterparts in terms of accuracy and speed, especially in case of speaker independent emotion recognition. Emotion identification [1] provides useful information in other sound source identification applications, such as speaker recognition and speech recognition. Here our approach is to classify emotions using MFCC features [6] and DTW. Recognition accuracy for MFCC feature is considered as it mimics the human ear perception. So ER recognition using MFCC features is illustrated in this paper.

A central topic in our paper is the emotion recognition using isolated words. In section-2 we describe the basics of the proposed system, in section-3 we discuss regarding feature extraction, in section-4 we present the classification procedure with DTW and section -5, concludes the paper.

2. Proposed Work:

2.1 Database collection:

In this paper we ,have considered five Emotions, namely Sad, Happy, Angry, Surprise and Neutral . We have collected 150 isolated words of each person with different time varying constraint at 20 to 300 KHz sampling frequency. Total 960samples are collected for our experimentation. Out of 960samples 480 samples are considered for training and 480 samples are considered as testing samples. Training templates are generated using 39 MFCC features.

2.2 Principal of Emotion recognition:

The two basic tasks in emotion recognition [3], [5] are pre-processing of speech signals and then classification part. In pre-processing, we analyse the speech signal before extracting the required features from it. Different operations are performed on the input speech signal such as removal of silence part, reemphasis, segmentation and framing, windowing, Mel Cestrum analysis and recognition (Matching) of the isolated words. The two phases of emotion recognition algorithms are testing and training phase. The block schematic is given in the figure 1.

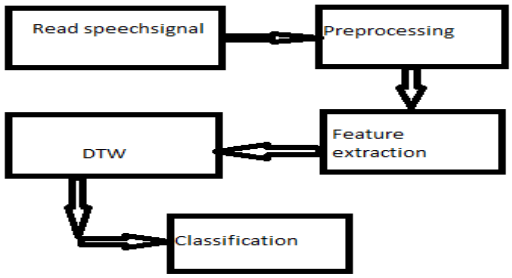

Fig 1: Block Diagram- schematic of emotion recognition

2.3 Removal of silence part:

By reducing the dimensions of feature vector we can improve the time complexity, which can be done by the removal of the silence part of speech. One of the best techniques to remove silence part is considering the Energy feature. Energy of each frame is calculated. Based on threshold value of energy the silence part is removed. The energy of each frame is given by

$$E(i) = \frac{1}{N}\sum_{n=1}^{N}|xi(n)|2 \quad (1)$$

We used Silence part removal algorithm which says divide the signal into number of frames, calculate energy for each frame, calculate threshold using median and compare energy of each frame with threshold. If energy of frame is greater than threshold then consider it otherwise it is silence part of signal and eliminates it. Some of the results of silence part removal for some speech signals are shown in figures- 2a, 2b, 2c, 2d.

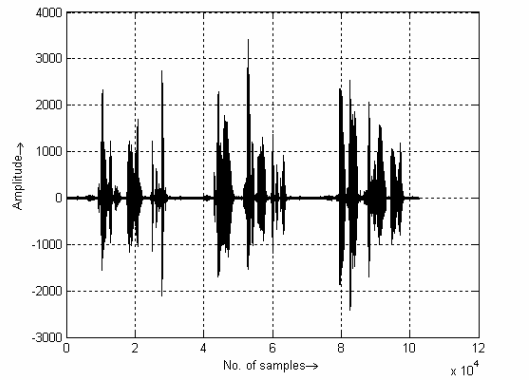

Figure-2a Speech signal in sad emotion before silence removal

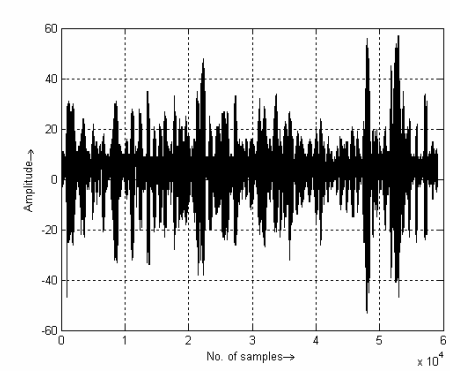

Figure-2b Speech signal in sad emotion after silence removal

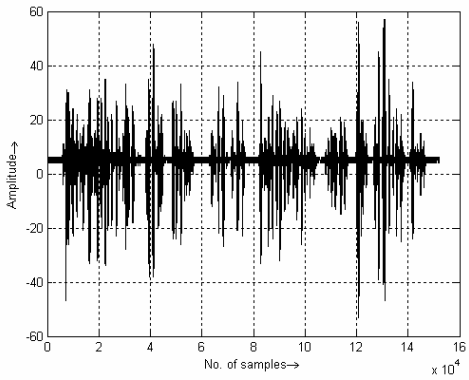

Figure-2c Speech signal in happy emotion after silence removal

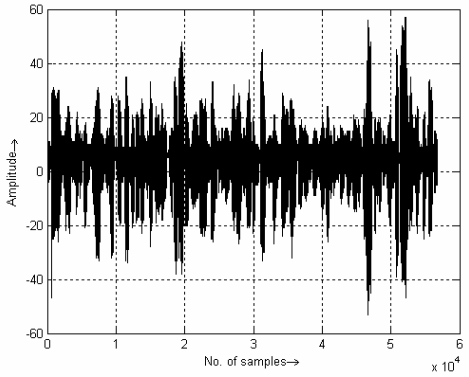

Figure-2d Speech signal in happy emotion after silence removal

3. Feature extraction:

This is the vital part of emotion recognition based on which the remaining part of classification and recognition depends on of acoustic model of speech signals. MFCC feature works better for Emotion recognition than LPC, LPCC with different experimentation [11], [12], [13]. There is no standard number of MFCC coefficients for recognizing the sound in any literature. MFCC is the way of representing the spectral information of a sound in compact form. In our paper we carried out the experimentation on MFCC to finalize the number of coefficients. It is proposed that 8-14 number of MFCC coefficients is sufficient to recognize the emotions [4], [10]. The algorithm for getting MFCC feature is as follows:

1) pre-emphasis,
2) hamming windowing,
3) FFT to obtain power spectrum,
4) log of FFT,
5) Mel filter bank,
6) DCT for decorrelation,
7) Δ MFCC (optional),
8) and ΔΔMFCC coefficients.

The mathematical details of each step are briefly described below.

Step 1: Pre–emphasis

This process will increase the energy of signal at higher frequency. It enables the passing of each speech signal through a first order FIR filter which emphasizes higher frequencies. The first order FIR filter equation is used is

$$Y[n] = X[n] - 0.95\,X[n] \quad (2)$$

Step 2: Framing

Each speech signal is divided into frames of 36 ms(milliseconds) and most of spectral characteristics remain the same in this duration, with 50 % of overlapping.

Step 3: Windowing

To remove edge effects, each frame is shaped with hamming window. Hamming window works better than other windows. The hamming window is represented by

$$w(n) = 0.54 - 0.46\cos\frac{2\pi n}{n-1} \quad (3)$$

where $0 \leq N \leq n\text{-}1$

Step 4: Fast Fourier Transformation (FFT)

FFT is used to get log magnitude spectrum to determine MFCC. We have used 1024 point to get better frequency resolution.

Step 5: Mel Filter Bank Processing

The 20 Mel triangular filters are designed with 50% overlapping .From each filter the spectrum are added to get one coefficient each, in this way we have considered the first 13 coefficients as our features. These frequencies are converted to Mel scale using following conversion formula.

$$f(mel) = 2595 * \log 10\,(1 + \frac{f}{700}) \quad (4)$$

we have considered 13 MFCC coefficients because, of the fact it gives better recognition accuracy than other coefficients.

Step 6: Discrete Cosine Transformation(DCT)

DCT of Each Mel frequency Ceptral are taken for de-correlation and energy compaction is called as MFCC. The set of coefficient are called MFCC Acoustic Vectors. Therefore, each input speech signal is transformed into a sequence of MFCC Acoustic Vector from which reference templates are generated.

Step 7: Delta Energy and Delta Spectrum

Features related to the change in cepstral features over time are represented by 13 delta features (12 cepstral features plus one energy feature), and 13 double delta or acceleration features. Each of the 13 delta features represents the change between frames, while each of the 13 double delta features represents the change between frames in the corresponding delta features. In similar fashion all the total 39 MFCC feature are calculated for every frame which constitute feature vector. Mel filter bank generated is shown in figure 3.

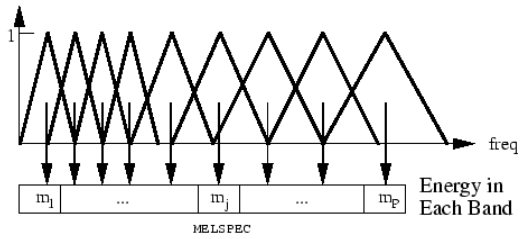

Figure 3: Mel filter bank.

4. Classification

After generating reference template i.e., generating the Training Set, the classification is done using SVM classifier [2]. For best matching, each input signal feature vector is calculated using MFCC as shown in fig 1. For every word DTW score is determined with reference template (Test sample). The best matching sample is classified using SVM classifier

4.1 Dynamic time Warping (DTW):

DTW [7], [8] is one of the approaches to emotion recognition to store a prototypical version of each word in the vocabulary and compute incoming emotion with each word. First a template of sequence of feature vectors is taken. The template is a single utterance of the word selected to be typical by some process. Then the comparison can be achieved by pair-wise comparison of the feature vectors. DTW achieves its goal by finding an optimal match between two sequences of features vectors which allows for stretched and compressed sections of the sequence. The DTW scores are calculated using above algorithm. Based on score SVM[4] is applied to find minimum DTW distortion. Based on minimum DTW distance the emotions are recognized. The DTW works better for different time varying /duration signal of same speech signal. This is an advantage over usual Euclidian distortion measure.

5. Conclusions:

We have conducted experiments on five human emotions which are happy, sad, angry, surprise and neutral using MFCC and DTW algorithms. Algorithms are implemented in MATLAB [9] environment, to identify the emotions. The results obtained after classification are tabulated and shown in Table 1 ,against each emotion.

	Recognition accuracy in % for individual emotion				
Features	Happy	Sad	Angry	Surprise	Neutral
MFCC(13)	83	82	85	81	78
MFCC + ΔMFCC (26)	84	86	91	87	82
MFCC+ ΔMFCC + ΔΔMFCC(39)	92	94	91	89	90

Table 1: Recognition accuracy rates

The Recognition accuracy for MFCC,ΔMFCC,ΔΔMFCC are obtained and presented in the Table- 1.The results are considered for 39 features (MFCC+ΔMFCC+ΔΔMFCC).

From the above Table-1, it can be seen thqat the emotions are clearly recognized and the recognition rates are above 78%, this is recognition rates are due to the consideration of using 39 MFCC features . Dynamic time warping is simply time alignment method which works better for time varying signals. Every distortion measure should be based on DTW for better recognition accuracy. Here recognition system considers only one specific measure of a sound i.e. the MFCCs, and yet still achieves quite accurate results. Also the results can be further improved using Hidden Markov Model (HMM), Gaussian Mixture Model (GMM).

References

[1] E. Mower, M. Matari´c, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," IEEE Trans. on Audio, Speech, and Language Processing, Accepted for Publication.

[2] B. Schuller, G. Rigoll and M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing, vol.1, pp. I-577-580, 17-21 May, 2004.

[3] V.A. Petrushin. Emotion recognition in speech signal: Experimentalstudy, development, and application. ICSLP, 2000.Conf. on Machine Learning and Cybernetics, vol. 8, pp. 4898–4901, Aug. 2005.

[4] E. Osuna, R. Freund, and E Girosi, "Support vector machines: Trainingand applications," A.I. Memo 1602, MIT A. I. Lab., 1997.

[5] O. Kwon, K. Chan, J. Hao, and T. Lee, "Emotion recognition by speechsignals," in Proc. 8th Eur. Conf. Speech Commun. Technol., 2003, pp.32–35.

[6] Ezzat T., Tomaso Poggio T. Discriminative Word-Spotting Using Ordered Spectro-Temporal Patch Features, Interspeech 2008

[7] E. Keogh and M. Pazzani, Scaling up Dynamic Time Warping for Datamining Applications, In Proc. of the Sixth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, pp.285-289, Boston, Massachuseetts, 2000.

[8] H. Sakoe, and S. Chiba, Dynamic Programming Algorithm Optimization for Spoken Word Recognition, In IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-26, 1978.

[9] Introduction to Computer Programming With MATLAB Department of Phonetics and Linguistics, Univ. College Landon [Online]. Available:www.phon.ucl.ac.uk/courses/spsci/matlab/, 2004.

[10] YI-LIN LI, Gang Wei, "Speech emotion recognition based on HMM and SVM", Proceedings of the FourthInternational Conference on Machine Learning andCybernetics, Vol.8, 18-21 Aug. 2005, pp.4898 – 4901

[11] Zhongzhe, X,Dellandrea, E.: Weibei D eal.Features Extraction andSelection for Emotional Speech Classification IEEE.(2005)411-416

[12] Ververidis, D., Kotrropoulos, K.: Emotional speech recognition:Resource, features, and methods. Speech Communication 48(2006) 1162-1181

[13] Prof .Sujata Pathak, Prof .Arun Kulkarni (2011) ,"Recognising emotions from speech".