IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 11, Issue 6 (May. - Jun. 2013), PP 46-50
www.iosrjournals.org

Emotion Recognition Based On Audio Speech

Showkat Ahmad Dar[1], Zahid Khaki[2]
[1]*Department of Computer Science and Engineering, Islamic University of Science and Technology, Awantipora*
[2]*Department of Electronics and Communication Engineering, Islamic University of Science and Technology Jammu and Kashmir 192221, India*

Abstract: *Emotion recognition aims at automatically identifying the emotional or physical state of a human being from his or her voice. The emotional and physical states of a speaker are known as emotional aspects of speech and are included in the so called paralinguistic aspects. Although the emotional state does not alter the linguistic content, it is an important factor in human communication, because it provides feedback information in many applications as making a machine to recognize emotions from speech is not a new idea.*

This paper presents automatic text independent speaker emotion recognition system using the pattern classification methods such as the support vector mechanics (SVM) .Acoustic features are derived from the speech signal at the segmental level. The segmental features are the features extracted from short frames (10-30 ms) of the speech. Acoustic features are derived from the speech signal at the segmental level. Acoustic features are represented by Mel frequency cepstral coefficients .A 39 dimensional MFCC for each frame is used as acoustic feature vector. The DFT based cepstral coeffiecients are computed by computing IDFT (inverse DFT) of the log magnitude short time spectrum of speech signal. Mel wraped cepstrum is obtained by inserting an intermediate step of transforming the frequencies before computing the IDFT. The Mel scale is based on human perception of frequency of sound. SVMs are used to construct the optimal separating hyper plane for speech features .SVMs are used to build the models for each speaker and to compare with the test speaker's feature vectors.

Keywords: *Inverse discrete Fourier transforms(IDFT), linear prediction coefficients(LPC), linear prediction cepstral coefficients(LPCC), Support vector mechanics(SVM), Artificial Neural Networks(ANN), Hidden Markov Model(HMM), Gaussian Mixture Model(GMM).*

I. Introduction

Emotion form a significant part of human interaction, and providing computers with the ability to recognize and make use of such of such non-verbal information could open the way to new and exciting paradigms in human- computer interaction. The human face is a key element of non-verbal communication, with our facial emotion acting as a type of social semaphore.

Without the ability to express and recognize emotions , social interaction is less fulfilling and stimulating , with either or both parties often unable to fully understand the meaning of the other.

This paper deals with the text-independent Speaker emotion recognition using pattern classification techniques such as SVM (support vector machine).There are two major models for emotion recognition: Discriminative model and generative model. Discriminative model consists of Artificial Neural Networks (ANN) and Support Vector Machines (SVM), etc. Generative model consists of Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM), etc.Each of them can construct speaker models for speaker recognition task.

II. Speaker Verification And Identification

The speaker recognition system can be operated in either identification mode or verification mode. In speaker identification, the goal is to identify the speaker of utterances from a given population where as speaker verification involves validating the identity claim by a person. Speaker recognition systems can be classified into text dependent and text independent systems. Text dependent systems require the recitation of a predominant text, where as text independent systems accept speech utterances of unrestricted text.

1. Speaker Emotion Verification

1.1 Speaker identity exits in the physiological and behavioral characteristics of a speaker. The physiolocal characteristics correspond to the characteristics of the vocal tract system and that of the voice source .The behavioral characteristics are due to the manner in which speaker have learnt to use their speech production apparatus.

1.2 Speaker verification use the system or source features for capturing speaker specific information. Some of the system features used for speaker verification task are formats, linear prediction coefficients (LPC)

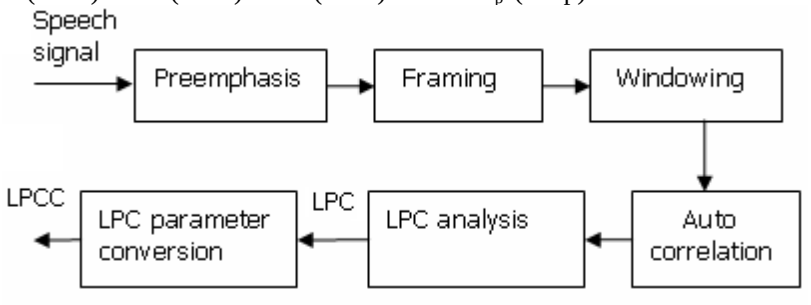
and linear prediction cepstral coefficients (LPCC). Source features such as Pitch, Intonation, and the linear prediction residual signal information for speaker verification.

III. Linear Prediction Coefficients

The theory of liner prediction (LP) is closely linked to modeling of the vocal tract system, and relies up on the fact that a particular speech sample may be predicted by a linear weight sum of the previous samples. The number of previous samples used for prediction is known as the order of prediction. The weights applied to each of the previous speech sample are known as linear prediction coefficients (LPC).They is calculated so as to minimize the prediction error.

A given speech sample at time n, s (n), can be approximated as a linear combination of the past p speech samples, such that

$$S(n) \approx a1s(n-1) + a2s(n-2) + a3s(n-3) + \cdots + a_p s(n-p)$$



3.1 Preprocessing

To extract the features from the speech signal, the signal must be preprocessed and divided into successive windows or analysis frames .so the following steps are performed before extracting the features. Preemphasis: The higher frequencies of the speech signal are generally weak. As a result there may not be high frequency energy present to extract features at the upper end of the frequency range.

3.2 Preemphasis

Preemphasis is used to boost the energy of the high frequency signals. The output of the preemphasis, $\hat{s}(n)$ is related the input s (n) by the difference equation

$$\hat{s}(n) = s(n) - \alpha s(n-1)$$

The typical value for α is 0.95

3.3 Frame blocking:

Speech analysis usually assumes that the signal properties change relatively slowly with time. This allows examination of a short time window of speech to extract parameters presumed to remain fixed for the duration of the window. Thus to model dynamic parameters, we must divide the signal into successive windows or analysis frames, so that the parameters can be calculated often enough to follow the relevant changes. The preemphasized speech signal, $\hat{s}(n)$ is blocked into frames of N samples (frame size), with adjacent frames being separated by M samples (frame shift). If we denote the ltframe of speech by xl (n), and there are L frames within the entire speech signal then

$$xl(n) = \hat{s}(Ml + n),\ 0 \le n \le N-1,\ 0 \le l \le N-1$$

3.4 Windowing: The next step in the processing is to window each individual Frame so as to minimize the signal discontinuities at the beginning and end of the frame. The window must be selected to taper the signal to zero at the beginning and end of each frame. If we define the window as w (n), $0 \le n \le N\text{-}1$, then the result of windowing the signal is

$$\tilde{x}l(n) = xl(n)w(n),\ 0 \le n \le N-1$$

IV. Linear Prediction Cepstral Coefficients

In many applications, Euclidean distance is used as a measure of similarity or dissimilarity between feature vectors. The sharp peak of the LP spectrum may produce large errors in a similarity test, even for a slight shift in the position of peaks .Hence the linear prediction coefficients are converted into cepstral coefficients using a recursive relation.Cepstral coefficients represents the long magnitude spectrum,and the first few model the smooth envelop of log spectrum .These coefficients can be obtained either from linear prediction coefficients or from inverse discrete fourire transform (IDFT).

IDFT (log (|DFT(x)|))

If x is LPC, the cepstral coefficients are known as Linear Prediction Cepstral Coefficients (LPCC).

V. Acoustic Feature Extraction

MFCC are widely used spectral features for speaker recognition , computation of the MFCC defers from the basic procedures described earlier, where the long magnitude spectrum is replaced with logarithm of Mel scale warped spectrum, prior to inverse Fourier transform operation .Hence the MFCC can be represented by the gross characteristics of the vocal tract system .*i.e.* ***IDFT (log|DFT(x)|))***

5.1 Process visualizaton

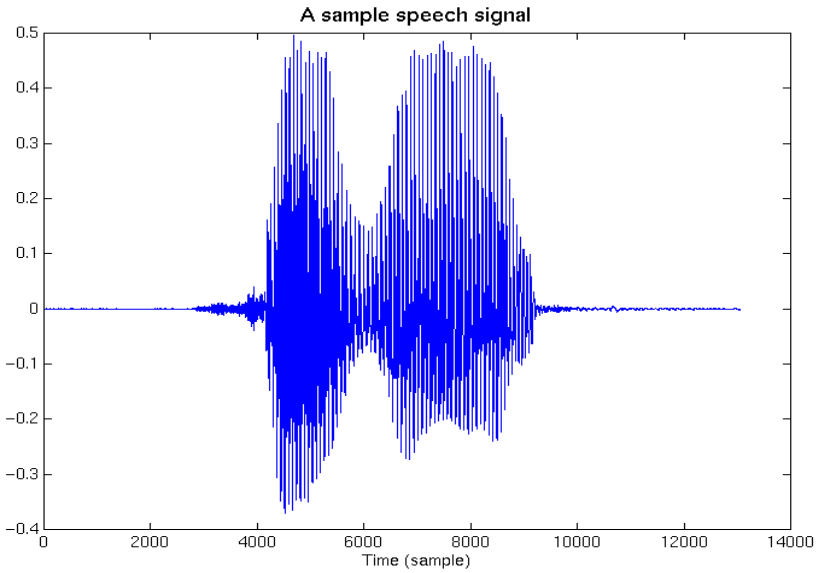


Figure 5.1 will serve as the audio signal intended for the analysis in this case. The next step is to frame the audio sample into portions of a predetermined size Figure 5.2 shows a frame belonging to a digital audio signal

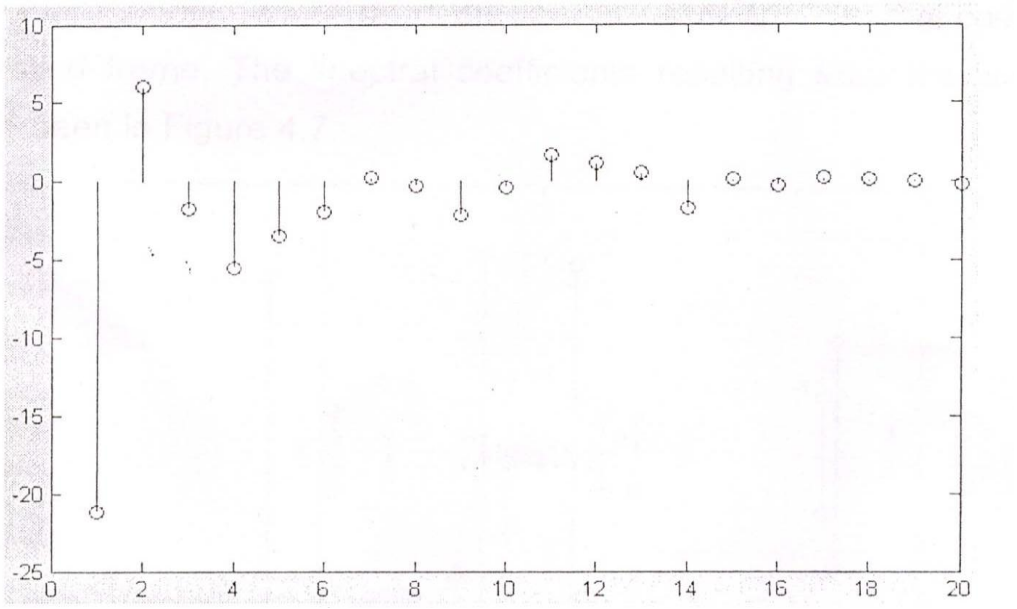


Figure 5.2

Next a window function is applied to the frame, in this case a Hamming window was used on the individual frames to smooth out the frame edges and reduce spectral distortion.

Then the spectral coefficients are processed with Mel-scale filter to convert these to the Mel scale. The filter bank used may be viewed in the Fig 5.3 the logarithms of these Mel spectral coefficients are then transformed to the frequency domain with the inverse DFT.

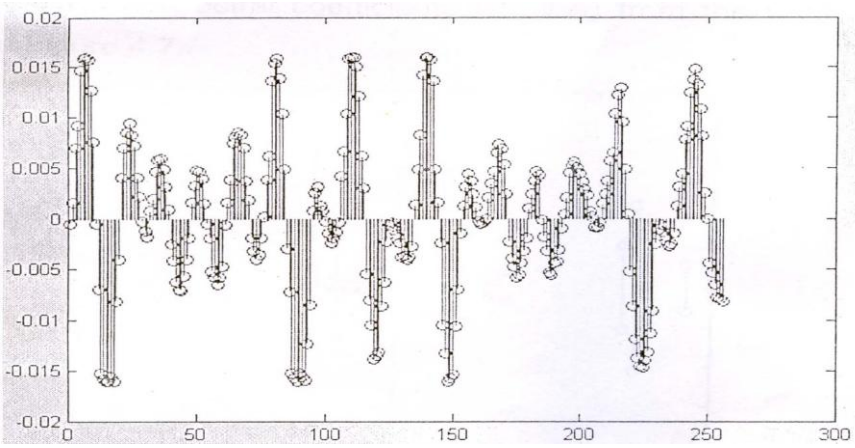


Fig 5.3 Audio Frame

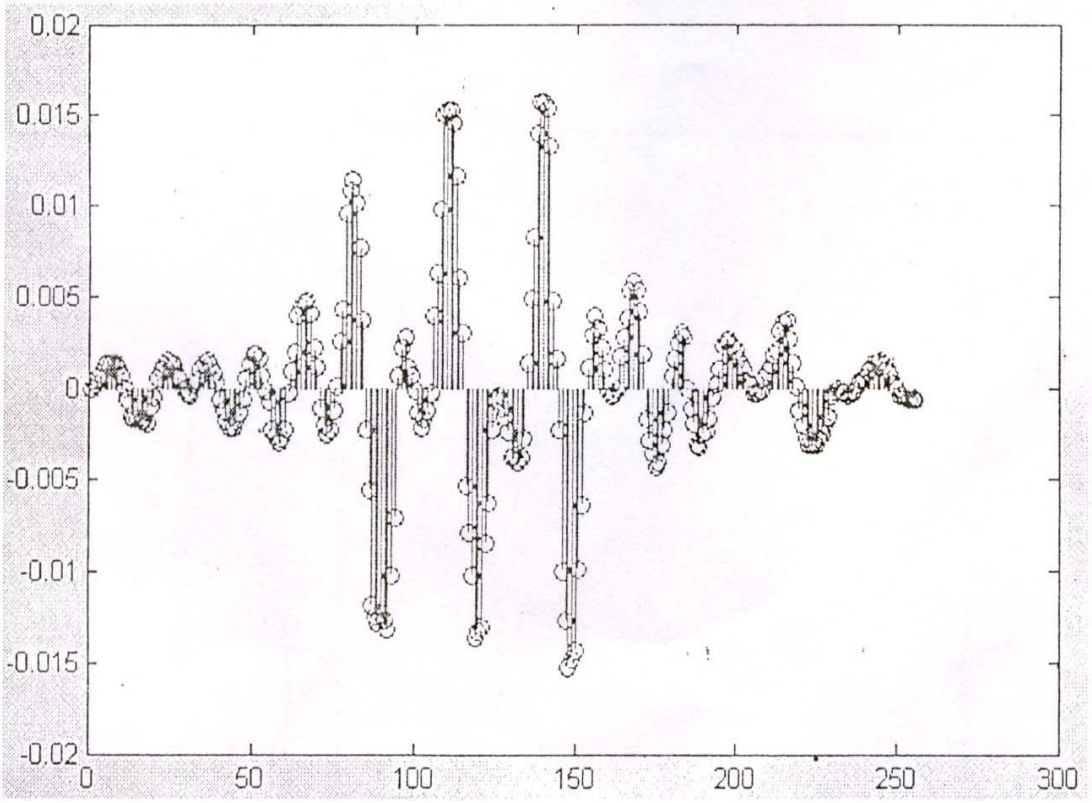


Fig 5.4 Audio window

VI. Conclusion

This paper allows for asynchrony in the audio and vocabulary states, while preserving the natural dependency of the audio signals. The audio sequences are treated separately and are no need for the problem. The advantage of the audio reorganization is confirmed by the experimental results. This model can be improved applied to a variety of human/machine system. In future work, we will improve current model to increase the efficiency, besides this, the research on recognition of emotion intensity will be performed through the analysis of audio feature s, which is different from the approach in. Also, we notice that some new dimensional reduction and the pattern classification methods like tensor based analysis proposed recently; we will carry out study on its application in emotion recognition field.

Reference

[1]. *Pongtep Angkititrakul and John H. L. Hansen*, **"Discrimination in–Set/out-of-set Speaker Recognition Systems"** IEEE TRANSACTIONS ON AUDIO , SPEECH , AND LANGUAGE PROCESSING, Vol.15, no.2, Feb.2007.

[2]. *Ran D. Zilca , Brain Kingsbury, Jiri Navrati, Ganesh N. Ramasamy* "PSEUDO PITCH SYNCHRONOUS ANALYSIS OF SPEECH WITH APPLICATION TO SPEAKER RECOGNITION", IEEE TRANSACTION ON AUDIO , SPEECH, AND LANGUAGE PROCESSING Vol.14, no. 2, Mar.2006.

[3]. **"***yildirim, S. Bulut, M., Lee , C.M., kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S.*, 2004. **An Acoustic Study Of Emotions Expressed In Speech**. In: Proc. Internat. Conf. On Spoken Language Processing (ICSLP 04), Korea, Vol.

[4]. *K. Sri Rama Murty and B. Yengnanarayana " **Combining Evidence from Residual Phase and MFCC Features for Speaker Emotion Recognition** IEEE signals processing letters, Vol.13,no. , jan.2006*

[5]. *Volunteers in Technical Assistance (VITA).*

[6]. Guillermo Garcia, Sung-kyo Jung, "**A Statistical Approach To Performance Evaluation Of Speaker Emotion Recognition Systems"** TECH/SSTP Lab., France Telecom R&D 22307 Lannion, France. W.B. Mikhel and Pravinkumar Premakanthan, "**An Improved Speaker Emotion Identification Technique Employing Multiple Representations of LPC**", in proceedings of IEEE , march 2000

[7]. *Sachin S. Kajarekar "***Four Weighing And Fusion : A Cepstral –SVM System For Speaker Emotion Recognition** VOL. 23, NO. 3, AUGUST 2008. **.**