International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 06, June-2015

Emotion Speech Recognition using MFCC and SVM

Shambhavi S. S
Department of E&TC
DYPSOEA
Pune,India

Dr.V. N Nitnaware
Department of E&TC
DYPSOEA
Pune,India

***Abstract*— Recognizing basic emotion through speech is the process of recognizing the intellectual state. Emotion identification through speech is an area which increasingly attracting attention within the engineers in the field of pattern recognition. Emotions play an extremely important role in human life. It is important medium of expressing humans viewpoint or feelings and his or hers mental state to others. Humans have natural ability to recognize emotions through speech information. Emotional computing has gained enormous research interest in the development of Human Computer Interaction over the past ten years. With the increasing power of emotion recognition, an logical computer system can provide a more friendly and effective way to communicate with users in areas such as video surveillance, interactive entertainment, intelligent automobile system and medical diagnosis.**

Here our approach is to classify emotions using Mel Frequency Cepstral Coefficients features and Support Vector Machine classifiers. Recognition accuracy for these feature is considered as it mimics the human ear perception. So emotion recognition using these features are illustrated.

Keywords—Emotion Recognition,MFCC(MelFrequency Cepstrum Coefficients),Pre processing,Feature extraction,SVM(Support Vector Machine)

I. INTRODUCTION

The speech signal is the fastest and the most natural method of communication between humans. This fact has motivated researchers to think of speech as a fast and efficient method of interaction between human and machine. However, this requires that the machine should have the sufficient intelligence to recog-nize human voices. Since the late fifties, there has been tremendous research on speech recognition, which refers to the process of converting the human speech into a sequence of words. However, despite the great progress made in speech recognition, wearestill far from having a natural interaction between man and machine because the machine does not understand the emotional state of the speaker. This has introduced a relatively recent research field, namely speech emotion recognition, which is defined as extracting the emotional state of a speaker from his or her speech. It is believed that speech emotion recognition can be used to extract useful semantics from speech, and hence, improves the performance.

The word emotion describes a short-term, consciously perceived, valenced state, either positive or negative.

II.PROPOSED METHODOLOGY

2.1 Why is it required ?

The main objective of employing (SER) Speech Emotion Recognition is to adapt the system response upon detecting frustration or annoyance in the speakers voice.

2.2 Emotion Speech Recognition is challenging task

1.It is not clear which speech features are more powerful in distinguishing between the emotions.

2.The acoustic variability introduced by the existence of different sentences, speakers, speaking styles and speaking rates adds another obstacle because these properties directly affect most of the common extracted speech features such as pitch and energy contours.

3.we recognize emotions by both speech and facial expressions, we also can recognize emotions by recognizing the spoken. All these three techniques have been emulated in computer systems and robotics, in non-biological emotion recognition systems we either determine the emotions by deciphering the facial expressions of the subject or we try to recognize emotions by speech and lastly there have been attempts to classify emotionally expressive words and recognize emotions from them.

But emotional recognition in humans happens in a very different way. Emotions in humans are a result of evolutionary development, hence different parts of the brain are involved in processing different kind of emotions, in healthy individuals, Neural mapping and study of limbic systems have allowed us insight on how the neural network in

our brain works with various chemicals like dopamine and serotonin to recognize and create emotional responses.Each emotion produced in speech is represented by different pitch, loudness and rate of the speech.

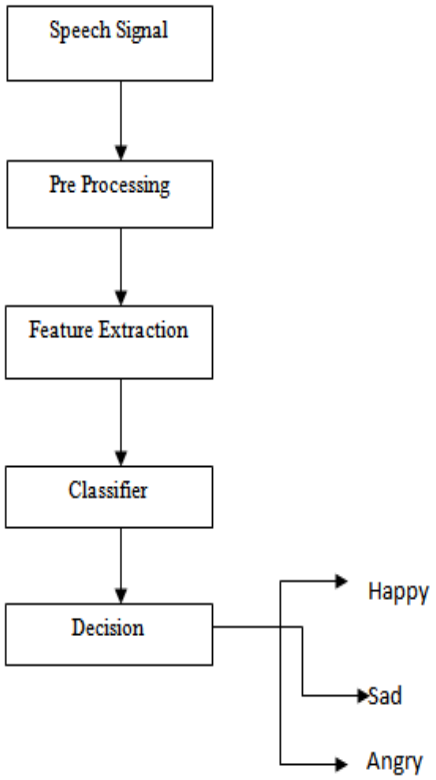

Fig 2.2.Flow chart of implementation of the proposed system.

Steps of proposed method

a) Pre processing
b) Framing
c) Windowing
d) FFT
e) Feature extraction
f) Classifier

a) Preprocessing

The term pre processing refers to all operations ,required to be performed on the time samples of speech signal before extracting features.For example due to recording environmental differences, sort of energy normalization has to be done to all utterences. From the whole utterance a short signal is taken removing silent parts as they do not carry any information. Signals are estimated to their energy to make it normalize. Speech signals are divided into frames of desired length and are analyzed. In this stage first the signal is denoised by soft thresholding the coefficients and since silence parts of the signal do not carry any information, are removed by thresholding the energy of the signal.

b) Framing

The pre-emphasized speech signal is then blocked into frames of N sample points with adjacent frames being separated by M (lower than N). The first frame is composed of the first N sample points. The second frames begin the Mth sample points after first frame and overlaps it by N-M sample points and so on. This process continues till all are accommodated within one or more frames. In our work, the frame length N = 256(10ms). There is overlap between two adjacent frames to ensure stationary between frames.

c) Windowing

Hamming window is applied to each frame to remove discontinuities in signal and ensure continuity between first and last data points. Each individual frame is windowed inorder to minimize the signal discontinuities at the beginning and the end of each frame.

d) FFT

It converts each frame from time domain signals into frequency domain and obtain frequency response of each frame.

e) Feature extraction

It involves extracting important information associated with the given speech and removing all the remaining useless information.Features such as energy,pitch,power and MFCC are extracted.

Pitch

The term pitch refers to the ear's perception of tone height.Pitch is grounded by human perception. It is a very obvious property of speech, also for non-experts, and it is often erroneously considered to be most important for emotion perception. Generally, a rise in pitch is an indicator for higher arousal, but also the course of the pitch contour reveals information on affect. Pitch can be calculated from the time or the frequency domain.Pitch does not exist for the unvoiced parts of the speech signal.

Energy

Loudness is the strength of a sound as perceived by the human ear. It is hard to measure directly, therefore the signal energy is often used as a related feature. Energy can be calculated from the spectrum after a Fourier transformation of the original signal.

Again, like pitch, high energy roughly correlates with high arousal, but also variations of the energy curve give hints on the speaker's emotion.

MFCC

Mel-frequency cepstral coefficients (MFCCs) are a parametric representation of the speech signal, that is commonly used in automatic speech recognition, but they have proved to be successful for other purposes as well, among them speaker identification and emotion recognition.They are claimed to be robust of all the features for any speech tasks.

A mel is a unit of measure of perceived pitch or frequency of a tone. Through the mapping onto the Mel-scale, which is an adaptation of the Hertz-scale for frequency to the human sense of hearing, MFCCs enable a signal representation that is closer to human perception.They are calculated by applying a Mel-scale filter bank to the Fourier transform of a windowed signal. Subsequently, a DCT

(discrete cosine transform) transforms the logarithmised spectrum into a cepstrum.

Mel filter banks consists of overlapping triangular flters with the cut off frequencies determined by the center frequencies of the two adjacent filters.The filters have linearly spaced center frequencies and fixed band width on the mel scale.The logatithm has the effect of changing multiplication into addition.It converts the multiplication of the magnitude in the FT into addition.

f) SVM Classifier

The identification of emotion-related speech features is extremely challenging task.Support Vector Machine is used as a classifier to classify different emotional states such as anger,sadness,fear,happy,boredom.SVM is simple and efficient algorithm which has a very good classification performance compared to other classifiers. SVM are the popular learning method for classification ,regression and other learning tasks. SVM has a better classification performance on a small amount of training samples. But we are lacking in guidelines on choosing a better kernel with optimized parameters of SVM.There is no uniform pattern used to the choice of SVM with its parameters and kernel function with its parameters.The paper proposed methods about selecting optimized parameters and kernel function of SVM.

The process of the system is as follows:

STEP1: Extracting speech emotion feature from utterances.

STEP2: The main task in optimized process is to improve the classification accuracy rate of the SVM.

STEP3: After optimizing process, the system trains an optimized model used to classify.

STEP4: The system gives a classification result (class label or recognition rate) about test samples.

The major principle of SVM is to establish a hyperplane as the decision surface maximizing the margin of separation between negative and positive samples. Thus SVM is designed for twoclass pattern classification. Multiple pattern classification problems can be solved using a combination of binary support vector machines.

III.APPLICATIONS

1. Beneficial to the orators.
2. Call centers
3. To improve ones emotional states according to various situations.
4. In human robotic interface.
5. In intelligent spoken tutoring systems.

IV.RESULTS

Finally the main emotions such as happy or joy,anger and neutral are classified for a particular incoming speech signal.

This is done using GUI matlab.

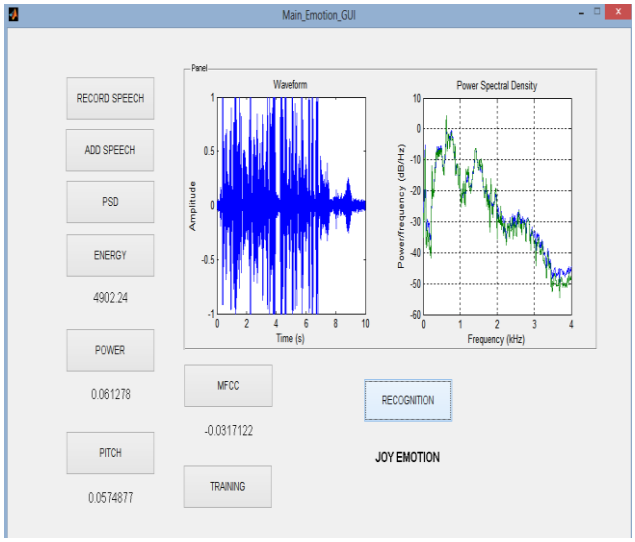

Fig4.1Matlab GUI for emotion recognition using speech

V.CONCLUSION

As technology evolves, interest in human like machines increases. Technological devices are spreading and user satisfaction increases importance. A natural interface which responds according to user needs has become possible with affective computing. The key issue of affective computing is emotions. Any research which is related with detection, recognition or generating an emotion is affective computing. User satisfaction or un-satisfaction could be detected with any emotion recognition system. Besides detection of user satisfaction, such systems could be used to detect anger or frustration. In such cases, user could be restrained like driving a car. In emotion detection tasks, speech or face emotion detections are the most popular ones. Easy access to face or speech data made them very popular. Speech carries a rich set of data. In human to human communication, via speech information is conveyed. Acoustic part of speech carries important info about emotions.MFCC are used for the feature extraction .Algorithm with the SVM's overall performance is tested.

VI.AKNOWLEDGEMENT

I take this opportunity to express my deep heartfelt gratitude to all those people who have helped me in the successful completion of the paper.First and foremost, I would like to express my sincere gratitude towards my guide Dr.V.N Nitnaware for providing excellent guidance, encouragement.Without his valuable guidance, this work would never have been a successful one.I would like to express my sincere gratitude to our Head of the Department of Electronics & Communication Engineering, Prof.Santhosh Bari for his guidance and inspiration.I would like to thank our Principal Dr.V.N. Nitnaware for providing all the facilities and a proper environment to work in the college campus.

VII.REFERENCES

[1] Mehrdad J. Gangeh, AliGhodsi,Mohamed S. Kamel,"Multiview Supervised Dictionary Learning in Speech Emotion Recognition," IEEE Transaction on audio, speech, and language processing.

[2] Shikha Gupta1, Jafreezal Jaafar2, Wan Fatimah wan Ahmad3 and Arpit Bansal4 J. Clerk Maxwell, "Feature extraction using mfcc" Signal & Image Processing : An International Journal (SIPIJ) Vol.4, No.4, August 2013

[3] N.Murali Krishna1,P.V. Lakshmi2,Y. Srinivas3 J.Sirisha Devi4," Emotion Recognition using Dynamic Time Warping Technique for Isolated Words," IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011

[4] Aastha Joshi," Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Volume 3, Issue 8, August 2013.

[5] Eslam Mansour mohammed1, Mohammed Sharaf Sayed2," LPC and MFCC Performance Evaluation with Artificial Neural Network for Spoken Language Identification ," International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 6, No. 3, June,2013