IJournals: International Journal of Software & Hardware Research in Engineering
ISSN-2347-4890
**Volume 4 Issue 2 February, 2016**

# An Improved and Fast Algorithm for Emotion Recognition in HCI using Active Feature Selection

**[1]Sanju Garle, [2]Angad Singh**

*[1,2] Department of Information Technology,*
*NIIST: NRI Institute of Information Science and Technology , Bhopal, India*
*sanju.amravati@gmail.com angada2007@gmail.com*

## ABSTRACT

*In Human Computer Interaction(HCI), emotion recognition from the speech signal has been research topic from many years Speech emotion recognition is one of the recent challenges in speech processing and Human Machine Interaction in order to address various operational needs for the real world applications. To identify the emotions from the speech signal, many systems have been developed. In this paper speech emotion recognition based on the previous technologies which uses different classifiers for the emotion recognition. The database for the speech emotion recognition system is the emotional speech samples and the features extracted from these speech samples are the energy, pitch, ZCR, intensity, linear prediction cepstrum coefficient (LPCC), Mel frequency cepstrum coefficient (MFCC). The classification performance is based on extracted features. The classifiers are used to differentiate emotions such as anger, disgust, fear, happy, neutral, sad, and surprise, etc. Multilevel SVM classifier which is used for identification of these seven discrete emotional states. The overall result of the conducted experiment is that the approach of using the Active Feature Selection achieved an average accuracy rate 85.25%.*

**Keywords: Classifier, Emotion recognition, Feature extraction, Active Feature Selection**

## 1. INTRODUCTION

There are many ways of communication but the speech signal is one of the fastest and most natural methods of communications between humans. Therefore the speech can be the fast and efficient method of interaction between human and machine also [1]. Humans have the natural ability to use all their available senses for maximum awareness of the received message. Through all the available senses people actually sense the emotional state of their communication partner. The emotional detection is natural for humans but it is very difficult task for machine. Therefore the purpose of emotion recognition system is to use emotion related knowledge in such a way that human machine communication will be improved. In speech emotion recognition, the emotions from the speech of male or female speakers are found out [1]. In the past century some speech features were studied which involved the fundamental frequencies, Mel frequency cepstrum coefficient (MFCC), linear prediction cepstrum coefficient (LPCC), etc., which form the basis for speech processing even today we have also considered some dynamic features like Mel-energy spectrum dynamic coefficients (MEDC) and have combined all the features to get a better result in emotion recognition. Many researches provide an in-depth insight into the wide range of algorithms for better classification options, such as Neural Networks (NN), Gaussian Mixture Model (GMM)[1], Hidden Markov Model (HMM)[4], Maximum Likelihood Bayesian Classifier (MLC), Kernel Regression and K-nearest Neighbors (KNN) and Support Vector Machine (SVM) [20], [21], [22], [23]. We have chosen Support vector machine for our research work as it gives better results in emotion recognition domain of various databases like BDES (Berlin Database of Emotional Speech) and MESC (Mandarin Emotional Speech Corpora). Fig.1. describes the generalized system model for speech emotion recognition.

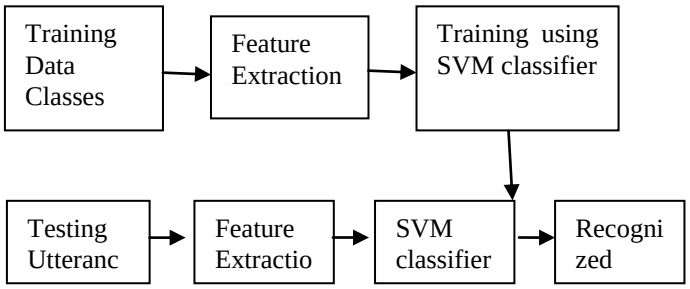


Fig.1.Generalized System model for Emotion Recognition

Emotion recognition from the speaker's speech is very difficult because of the following reasons: In



 

differentiating between various emotions which particular speech features are more useful is not clear. Because of the existence of the different sentences, speakers, speaking styles, speaking rates accosting variability was introduced, because of which speech features get directly affected. The same utterance may show different emotions. Each emotion may correspond to the different portions of the spoken utterance. Therefore it is very difficult to differentiate these portions of utterance. Another problem is that emotion expression is depending on the speaker and his or her culture and environment. As the culture and environment gets change the speaking style also gets change, which is another challenge in front of the speech emotion recognition system.

Automatic speech emotion recognition ASER basically aims at automatic identification of different human emotions or physical states through a human's voice. Emotion recognition system has various applications in the fields of security, learning, medicine, entertainment, etc. It can act as a feedback system for real life applications in the field of robotics, where robot will follow human commands by understanding the emotional

state of human. The successful recognition of emotions will open up new possibilities for development of an e-learning system with enhanced facilities in terms of student's interaction with machines. The idea can be incorporated in entertainment with the development of natural and interesting games with virtual reality experiences. It can also be used in the field of medicine for analysis and diagnosis of cognitive state of a human being. With the advancements in the field of human-machine interaction, calls for a user-friendly interface have become apparently very important for various speech oriented applications.

This paper is well organized as follows. In section 2, a brief description about the speech emotion recognition system is given. Section 3 includes process of feature extraction and Selection. Section 4 contains the speech emotional database description. In section 5, information about the classifier selection, we have chosen support vector machine classification, section 6 that contains implementation of Active Feature selection method, Conducted experiments and its results are provided in section 7 and finally conclusion and future directions are given in section 8.

## 2. SPEECH EMOTION RECOGNITION SYSTEM

Speech emotion recognition aims to automatically identify the emotional state of a human being from his or her voice. It is based on in-depth analysis of the generation mechanism of speech signal, extracting some features which contain emotional information from the speaker's voice, and taking appropriate pattern recognition methods to identify emotion.

The basic structure of the Speech Emotion Recognition System is shown in figure 2. The main concern in speech emotion recognition system is to find out a set of significant emotions to be classified by an automatic emotion recognizer. A typical set of emotions contains 300 emotional states. Therefore to classify such a great number of emotions is very complicated. According to "Palette theory" any emotion can be decomposed into primary emotions similar to the way that any color is a combination of some basic colors. Primary emotions are anger, disgust, fear, joy, sadness and surprise [1]. The evaluation of the speech emotion recognition system is based on the level of naturalness of the database which is used as an input to the speech emotion recognition system. If the inferior database is used as an input to the system then incorrect conclusion may be drawn. The database as an input to the speech emotion recognition system may contain the real world emotions or the acted ones. It is more practical to use database that is collected from the real life situations [1].

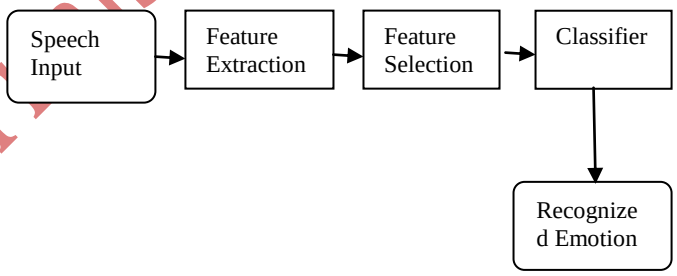


Fig.2: Structure of the Speech Emotion Recognition System

## 3. FEATURE EXTRACTION AND SELECTION

Any emotion from the speaker's speech is represented by the large number of parameters which is contained in the speech and the changes in these parameters will result in corresponding change in emotions. Therefore an extraction of these speech features which represents emotions is an important factor in speech emotion recognition system [5]. The prosodic features are known as the primary indicator of the speakers emotional states. Research on emotion of speech indicates that pitch, energy, duration, formant, Mel frequency cepstrum coefficient (MFCC), and linear prediction cepstrum coefficient (LPCC) are the important features [5, 10]. Linear prediction cepstrum coefficient (LPCC) gives the details about the characteristics of particular channel of any individual person and this channel characteristic will get change in accordance with the

different emotions, so by using these features one can extract the emotions in speech. The merits of using the LPCC is that it involves less computation, its algorithm is more efficient and it could describe the vowels in better manner. Mel frequency cepstrum coefficient (MFCC) is extensively used in speech recognition and the recognition rate of the MFCC is very good. Mel frequency cepstrum is an illustration of short term power spectrum of sound [10].

Fourteen features have been evaluated for use in the system. The features extracted to train our system are:

**Pitch***:* It is the main feature of an audio file. Sounds may be generally characterized by pitch, loudness, and quality. The perceived pitch of a sound is just the ear's response to frequency, i.e., for most practical purposes the pitch is just the frequency.

Pitch = frequency of sound.

**Standard Deviation***:* standard deviation (represented by the symbol sigma) shows how much variation or dispersion exists from the average (mean), or expected value. A low standard deviation indicates that the data points tend to be very close to the mean; high standard deviation indicates that the data points are spread out over a large range of values.

**Energy Intensity***:* This feature represents loudness of an audio signal, which is correlated to amplitude of signal.

**Energy Entropy***:* It expresses abrupt changes in the energy level of an audio signal. In order to calculate this feature, frames are further divided into K-sub windows of fixed duration.

**Shimmer***:* A frequent back and forth changes in amplitude (from soft to louder) in the voice. Shimmer Percent provides an evaluation of the variability of the peak-to-peak amplitude within the analyzed voice sample. It represents the relative period-to-period (very short-term) variability of the peak-to-peak amplitude.

**Jitter***:* It is defined as varying pitch in the voice, which causes a rough sound. Compare to shimmer, which describes varying loudness in the voice. Jitter is the undesired deviation from true periodicity of an assumed periodic signal. Jitter Percent provides an evaluation of the variability of the pitch period within the analyzed voice sample. It represents the relative period-to-period (very short-term) variability.

**Autocorrelation:** It is the cross-correlation of a signal with itself. Informally, it is the similarity between observations as a function of the time lag between them. It is a mathematical tool for finding repeating patterns,
such as the presence of a periodic signal obscured by noise, or identifying the missing fundamental frequency in a signal implied by its harmonic frequencies. It is often used in signal processing for analyzing functions or series of values, such as time domain signals.

**Noise to Harmonic ratio***:* Noise is an undesirable component that obscures a wanted signal. NHR is an average ratio of energy of the inharmonic components in the range 1500-4500 Hz to the harmonic components energy in the range 70-4500 Hz. It is a general evaluation of the noise presence in the analyzed signal (such as amplitude and frequency variations, turbulence noise, sub-harmonic components and/or voice breaks).

**Harmonic to Noise ratio***:* HNR represents the degree of acoustic periodicity, also called as Harmonicity object. Harmonicity is expressed in dB: if 99% of the energy of the signal is in the periodic part, and 1% is noise, the HNR is 10*log10(99/1) = 20 dB. A HNR of 0 dB means that there is equal energy in the harmonics and in the noise.

**Short Time Energy***:* The amplitude of the speech signal varies appreciably with time. In particular, the amplitude of unvoiced segment is generally much lower than the amplitude of voiced segments. Short Time energy provides a convenient representation that reflects these amplitude variations. The major significance of this is that it provides a basis for distinguishing voiced speech from unvoiced speech.[2]

**Zero Crossing Rate***:* It is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back. This feature has been used heavily in both speech recognition and music information retrieval, being a key feature to classify percussive sounds.

**Spectral Centroid***:* It is the weighted mean frequency. It indicates where the "center of mass" of the spectrum is. Because the spectral centroid is a good predictor of the "brightness" of a sound [5], it is widely used in digital audio and music processing as an automatic measure of music timbre.
$Ct= MtNn=1\ Mt\ n\ Nn=0\ n\ *n$ where $Mt(n)$
is magnitude of Fourier transform at frame t and frequency bin n. The centroid is a measure of spectral shape and higher centroid values correspond to "brighter" textures with more high frequencies.



 

**Spectral Rolloff***:* Spectral Rolloff point is defined as the *Nth* percentile of the power spectral distribution, where *N* is usually 85% or 95%. This measure is useful in distinguishing voiced speech from unvoiced: unvoiced speech has a high proportion of energy contained in the high-frequency range of the spectrum, where most of the energy for voiced speech and music is contained in lower bands.

$Mt\ n$ =0.85∗Rt$n$=1 $Mt(n)Nn$=1

Where Rt is the frequency below which 85% of the magnitude distribution is concentrated.

**Spectral Flux***:* It is a measure of how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against power spectrum for the previous frame. More precisely, it is usually calculated as the Euclidean distance between the two normalized spectra.

These features have been extracted for every uploaded wave file and then database of these features is prepared for each emotion in excel spreadsheet. When the spreadsheet is uploaded in MATLAB to train the system it is stored as .mat file to form four different clusters of emotion.

## 4. DATABASE DESCRIPTION

This database includes utterances belonging to seven basic Emotional states anger, disgust, fear, happy, neutral, sad and surprise. Each person recorded 140 short sentences (20 per emotion) of different lengths in his or her first language. This makes the database, a combination 4200 utterances, enrich in various modalities in terms of gender and languages. The speech samples were recorded with 16 bit depth and 44.1 KHz sampling frequency. In this paper we divides the emotion into seven categories anger, disgust, fear, happiness, neutral, sad and surprise, and tries to include all kinds of feelings in them. In order to obtain experiment utterances, some non-professionals have been invited to record their emotions, thus creating an emotional database. The design of the experiment is speaker independent and gender-independent. There seems to be two types of databases i.e. Create Database and Evaluate Database which are used within the emotion recognition research field. Firstly, databases made of acted emotions i.e. create database. These are built by asking actors to speak with a predefined emotion, and then each sample is manually labeled under its specific emotion.[19]
The second type of databases are the ones built from real-life systems (for example, call-centers, interviews, or meetings), thus containing authentic emotional speech.

## 5. CLASSIFIER SELECTION

In the speech emotion recognition system after calculation of the features, the best features are provided to the classifier. A classifier recognizes the emotion in the speaker's speech utterance. Various types of classifier have been proposed for the task of speech emotion recognition. Gaussian Mixtures Model (GMM), K-nearest neighbors (KNN), Hidden Markov Model (HMM) and Support Vector Machine (SVM), Artificial Neural Network (ANN),[1] etc. are the classifiers used in the speech emotion recognition system. Each classifier has some advantages and limitations over the others.[1][17] The application of the speech emotion recognition system include the psychiatric diagnosis, intelligent toys, lie detection, in the call centre conversations which is the most important application for the automated recognition of emotions from the speech, in car board system where information of the mental state of the driver may provide to the system to start his/her safety [1].
Classification process involves following steps:

1. Create training data set.
2. Identify class attribute and classes.
3. Identify useful attributes for classification (Relevance analysis).
4. Learn a model using training examples in Training set.
5. Use the model to classify the unknown data samples.

SVM is a supervised learning process comprising of two steps:

i. Learning (Training): Learn a model using the training data.
ii. Testing: Test the model using unseen test data to assess the model accuracy.

The input audio signal was divided into frames and all the features were calculated for each frame. Now, In order to draw one conclusion from all the features of several frames of the input signal, we need to consider some kind of statistics. Statistical features [16] like Mean, Standard Deviation, Max and Range were considered for each feature over all the frames, and a single feature vector was formed including all the statistical parameters, representing the input signal. Then, the normalized statistical feature vector was provided to the Support Vector Machine (SVM) classifier for training or testing.
SVM is having much better classification performance compared to other classifiers [1, 6]. The emotional states can be separated to huge margin by using SVM classifier.

An original SVM classifier was designed only for two class problems, but it can be use for more classes. Because of the structural risk minimization oriented training SVM is having high generalization capability. The accuracy of the SVM for the speaker independent and dependent classification is 75% to 80% for speech emotion recognition.

## 6. IMPLEMENTATION

Through this experiment, consider If all the extracted features gives as an input to the classifier this would not guarantee the best system performance which shows that there is a need to remove such a unusefull features from the base features. Therefore there is a need of systematic feature selection to reduce these features. For that we implement Active Feature selection (AFS) algorithm which is used to find most variance features and reduce the unusefull feature to select the best(or relevant) feature subset. So that speed is increase and also reduce features hence improve accuracy. Figure 3 shows the flow of Active Feature Selection method.

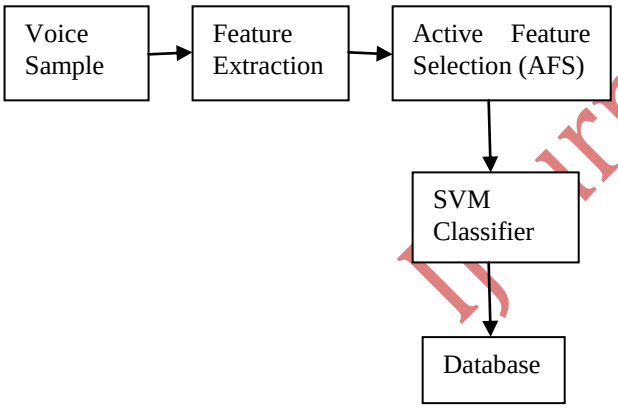


Fig.3 Active Feature Selection (AFS)

## 7. EXPERIMENT AND RESULT

In our experiment, we have taken speaker-independent and gender independent training models for SVM we have taken utterances of five speakers as training set and the other speaker's speech samples for testing purpose. The feature vector includes prosody features (like Pitch, ZCR, Short-term Energy, Long-entropy, intensity), quality features (three formant frequencies, Spectral Roll-off, Spectral flux, Spectral centroid), Mel Frequency Cepstral Coefficients, Linear Predictive Coding Coefficients, Mel-Energy spectrum Dynamic Coefficients. We have taken four statistics (mean, standard deviation, max and range) for each feature class in order to form a single feature vector for each utterance. A total of 570 speech samples for training and 120 speech samples are used for testing purpose. Our experiment shows average accuracy of 85% for emotion recognition with AFS (i.e. using Active Feature Selection) which is shown in tabular table format as follow.

| File Name | Emotion Name | Time Needed with AFS | Time Needed without AFS | Accu-racy with AFS | Accu-racy without AFS |
|---|---|---|---|---|---|
| 03a01Fa | Neutral | 0.1352 | 0.1387 | ✓ | ✓ |
| 03a01Nc | Disgust | 0.1334 | 0.1344 | ✓ | Fear |
| 03a01Wa | Anger | 0.1338 | 0.1342 | ✓ | ✓ |
| 03a02Fc | Happy | 0.1339 | 0.1338 | Anger | ✓ |
| 03a02Nc | Sad | 0.1342 | 0.1509 | ✓ | ✓ |
| 03a02Ta | Disgust | 0.1350 | 0.1454 | ✓ | ✓ |
| 03a02Wb | Surprise | 0.1490 | 0.1535 | ✓ | ✓ |
| 03a02Wc | Anger | 0.1351 | 0.1401 | ✓ | ✓ |
| 03a04Ad | Surprise | 0.1479 | 0.1508 | ✓ | ✓ |
| 03a04Fd | Happy | 0.1351 | 0.1357 | ✓ | Anger |

Accuracy with AFS: 9/10*100 = 90%

Accuracy without AFS: 8/100= 80%

Increase Speed with AFS in terms of Time is: 0.45

## 8. CONCLUSION AND FUTURE DIRECTIONS

In this paper we have discussed the technique of emotion recognition from human speech through feature extraction and selection of voice sample. We also presented the list of classifier through this experiment it is found that SVM having better classification compared to other classifier. We have implement Active Feature selection method for this only relevant feature will be found. consider, If all the extracted features gives as an input to the classifier this would not guarantee the best system performance which shows that there is a need to remove such a unusefull features from the base features. Therefore there is a need of systematic feature selection to reduce these features. Active Feature selection (AFS) method is used to select the best (or relevant) feature subset. So that speed is increase and also reduce features hence improve accuracy. The performance of speech emotion recognition system is usually specified in terms of accuracy and speed.

Despite of several researches in the field of emotion recognition a real time model for this application has not been developed yet. An implementable and robust real time model for these



 

applications can be a scope for future work. We are also targeting to implement our algorithm in a real time scenario in near future.

## 9. REFERENCES

[1] M. E. Ayadi, M. S. Kamel, F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases", Pattern Recognition 44, PP.572-587, 2011.

[2] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Commun.*, vol. 48, no. 9, pp. 1162–1181, Sep. 2006.

[3] I. Luengo and E. Navas, "Automatic Emotion Recognition using Parameters" pp. 493–496, 2005.

[4] [11] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," *2003 Int. Conf. Multimed. Expo.ICME '03. Proc. (Cat. No.03TH8698)*, vol. 1, pp. 1–4, 2003.

[5] M.A.Anusuya and S.K.Katti, "Speech Recognition by Machine: A Review", (IJCSIS) International Journal of Computer Science and Information Security, vol. 6, no. 3, pp.181-205, 2009.

[6]P.Shen, Z. Changjun, X. Chen, "Automatic Speech Emotion Recognition Using Support Vector Machine", International Conference On Electronic And Mechanical Engineering And Information Technology, 2011 .

[7] Rajesh Kumar Aggarwal and M. Dave, "Acoustic modeling problem for automatic speech recognition system: advances and refinements Part (Part II)", Int J Speech Technol, pp. 309– 320, 2011.

[8]S. Emerich, E. Lupu, A. Apatean, "Emotions Recognitions by Speech and Facial Expressions Analysis", 17th European Signal Processing Conference, 2009 .

[9] Chiu Ying Lay, Ng Hian James. "Gender Classification fromSpeech",(2005)Webreference:http://sg.geocities.com/nghianja/CS5240.doc

[10] Nobuo Sato and Yasunari Obuchi. "Emotion Recognition using MFCC"s" Information and Media Technologies 2(3):835-848 (2007) reprinted from: Journal of Natural Language Processing 14(4): 83-96 (2007)

[11 ] T L Nwe'; S W Foo L C De Silva, **"**Detection of Stress and Emotion in Speech Using Traditional And FFT Based Log Energy Features" 0-7803-8185-8/03 2003 IEEE ( 2003)

[12] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias and et al, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine*. *America*, vol. 18, pp. 32-80, January 2001.

[13] L. Devillers, C. Vaudable, and C. Chastagnol, "Real-life emotion-related states detection in call centers: a crosscorpora study," *in Proc. INTERSPEECH 2010. Chiba*, pp. 2350-2353, September 2010.

[14] D. Morrison and R. Wang, LC, "De Silva. Ensemble methods for spoken emotion recognition in call-centers. Speech Communication," *Speech Communication*. *Amsterdam*, vol. 49, pp. 98-112, February 2007.

[15] A. Batliner, K. Fischer, R. Huber, J. Spilker and E. Noth, "How to find trouble in communication," *Speech Communication*. *Amsterdam*, vol. 40, pp. 117-143, April2003.

[16] R. Cowie and E. Douglas-Cowie, "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," in *Proc. of ICSLP*, Philadelphia, Dec. 1998, pp. 1989–1992.

[17] N. Amir and S. Ron, "Towards an automatic classification of emotion in speech," in *Proc. of ICSLP*, Sydney, Dec. 1998, pp. 555–558.

[18] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura, "Acoustic nature and perceptual testing of corpora of emotionalspeech," in *Proc. of ICSLP*, Sydney, Dec. 1998, pp.225–228.

[19] C. Tchong, J. Toen, Z. Kacic, A. Moreno, and A. Nogueiras, "Emotional speech synthesis database recordings," Tech. Rep. IST–1999–No 10036–D2, INTERFACE Project, July 2000.

[20] Y. Pan, P. Shen, and L. Shen, "Speech Emotion Recognition Using Support Vector Machine," vol. 6, no. 2, pp. 101–108, 2012.

[21] M. Dumas, "Emotional Expression Recognition using Support Vector Machines."

[22] P. Shen and X. Chen, "Automatic Speech Emotion Recognition Using Support Vector Machine," pp. 621–625, 2011.

[23] B. Schuller, G. Rigoll, and M. Lang, "Machine - Belief Network Architecture," in *IEEE/ICASSP*, 2004, pp. 577–580.