International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307, Volume-2, Issue-1, March 2012

# Speech Emotion Recognition

**Ashish B. Ingale, D. S. Chaudhari**

***Abstract*— In human machine interface application, emotion recognition from the speech signal has been research topic since many years. 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 is reviewed. The classifiers are used to differentiate emotions such as anger, happiness, sadness, surprise, neutral state, etc. 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, linear prediction cepstrum coefficient (LPCC), Mel frequency cepstrum coefficient (MFCC). The classification performance is based on extracted features. Inference about the performance and limitation of speech emotion recognition system based on the different classifiers are also discussed.**

***Keywords*— Classifier, Emotion recognition, Feature extraction, Feature Selection.**

## I. 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 [2].

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. In one of the research the spectrograms of real and acted emotional speech were studied and found similar recognition rate for both, which recommend that later one can be use for the speech emotion recognition system. In another research a correlation between emotion and speech features were present. Further humans and machine emotion recognition rate was Compared, in which same recognition rates were found for both. After this study a speech emotion recognition system using Hidden Markov Model was presented and achieved an accuracy of 70% for seven emotional states. In another study Support Vector Machine for speech motion recognition of the four different emotions with an accuracy of 73% was obtained [4, 5].

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. There may be two or more types of emotions, long term emotion and transient one, so it is not clear which type of emotion the recognizer will detect [1].

Emotion recognition from the speech information may be the speaker dependent or speaker independent. The different classifiers available are k-nearest neighbors (KNN), Hidden Markov Model (HMM) and Support Vector Machine (SVM), Artificial Neural Network (ANN), Gaussian Mixtures Model (GMM). The paper reviews the mentioned classifiers [4]. 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].

The paper is organized as follows: section two describes the overall structure of the speech emotion recognition system. The different features extracted in the feature extraction and the details about the feature selection are discussed in section three. Different classification schemes which could be use in the speech emotion recognition system describes in section four.

## II. SPEECH EMOTION RECOGNITION SYSTEM

Speech emotion recognition is nothing but the pattern recognition system. This shows that the stages that are present in the pattern recognition system are also present in the Speech emotion recognition system. The speech emotion recognition system contains five main modules emotional speech input, feature extraction, feature selection, classification, and recognized emotional output [2]. The

**Ashish B. Ingale**, Department of Electronics and Telecommunication Engineering, Government College of Engineering, Amravati, India, Mobile Phone No.: +919860305773, (e-mail: ingale.ashish7@gamil.co.in).
**D. S. Chaudhari**, Department of Electronics and Telecommunication Engineering, Government College of Engineering, Amravati, India, (e-mail: ddsscc@yahoo.com).

structure of the speech emotion recognition is as shown in Figure 1.

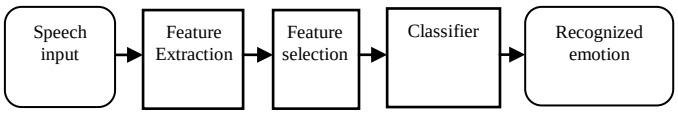


Figure 1. Structure of the Speech Emotion Recognition System.

The need to find out a set of the significant emotions to be classified by an automatic emotion recognizer is a main concern in speech emotion recognition system. 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].

## III. 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 [6]. The speech features can be divide into two main categories that is long term and short term features. The region of analysis of the speech signal used for the feature extraction is an important issue which is to be considering in the feature extraction. The speech signal is divided into the small intervals which are referred as a frame [1].

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, 6]. With the different emotional state, corresponding changes occurs in the speak rate, pitch, energy, and spectrum. Typically anger has a higher mean value and variance of pitch and mean value of energy. In the happy state there is an improvement in mean value, variation range and variance of pitch and mean value of energy. On the other hand the mean value, variation range and variance of pitch is decreases in sadness, also the energy is weak, speak rate is slow and decrease in spectrum of high frequency components. The feature of fear has a high mean value and variation range of pitch, improvement of spectrum in high frequency components. Therefore statistics of pitch, energy and some spectrum feature can be extracted to recognize emotions from speech [5, 6].

One of the main speech features which indicate emotion is energy and the study of energy is depends on short term energy and short term average amplitude [6]. As the arousal level of emotions is associated with the short term speech energy therefore it can be used in the field of emotion recognition. The pitch signal which is also referred as the glottal wave form is one more main feature which indicates emotion in speech. The pitch signal depends on the tension of the vocal folds and sub glottal air pressure, and it is produced from the vibration rate of the vocal cord. The pitch signal is characterize by the two features that is pitch frequency, and glottal air velocity at the vocal fold opening time instant. Number of harmonics present in the spectrum is directly get affected by the pitch frequency [7].

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 speech emotion recognition systems and the recognition rate of the MFCC is very good. In the low frequency region better frequency resolution and robustness to noise could be achieved with the help of MFCC rather than that for high frequency region [6]. Mel frequency cepstrum is an illustration of short term power spectrum of sound [4].

In feature extraction all of the basic speech feature extracted may not be helpful and essential for speech emotion recognition system. 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. Forward selection (FS) feature selection method could be used to select the best feature subset. In the initial stage forward selection initializes with the single best feature out of the whole feature set. The remaining features are further added which increases the classification accuracy. If the added number of features attained the preset number, the selection process should stop [10].

## IV. 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), etc. are the classifiers used in the speech emotion recognition system. Each classifier has some advantages and limitations over the others.

Only when the global features are extracted from the training utterances, Gaussian Mixture Model is more suitable for speech emotion recognition. All the training and testing equations are based on the supposition that all vectors are independent therefore GMM cannot form temporal structure

of the training data. For the best features a maximum accuracy of 78.77% could be achieved using GMM. In speaker independent recognition typical performance obtained of 75%, and that of 89.12% for speaker dependent recognition using GMM [1].

Other classifier that is used for the emotion classification is an artificial neural network (ANN), which is used due to its ability to find nonlinear boundaries separating the emotional states. Out of the many types, feed forward neural network is used most frequently in speech emotion recognition [7]. Multilayer perceptron layer neural networks are relatively common in speech emotion recognition as it is easy for implementation and it has well defined training algorithm [1].The ANN based classifiers may achieve a correct classification rate of 51.19% in speaker dependent recognition, and that of 52.87% for speaker independent recognition. According to the emotional state of the k utterances, the k-nearest neighbor classifier (k-NN) allocates an utterance to an emotional condition. The classifier can classify all the utterances in the design set properly, if 'k' equals to 1, however its performance on the test set will reduced. Utilizing the information of pitch and energy contours, the K-NN classifier attains an accurate classification rate of 64% for four emotional states. [7].

In speech recognition system like isolated word recognition and speech emotion recognition, hidden markov model is generally used; the main reason is its physical relation with the speech signals production mechanism. In speech emotion recognition system, HMM has achieved great success for modeling temporal information in the speech spectrum. The HMM is doubly stochastic process consist of first order markov chain whose states are buried from the observer [1]. For speech emotion recognition typically a single HMM is trained for each emotion and an unknown sample is classified according to the model which illustrate the derived feature sequence best [3]. HMM has the important advantage that the temporal dynamics of speech features can be caught second accessibility of the well established procedure for optimizing the recognition framework. The main problem in building the HMM based recognition model is the features selection process. Because it is not enough that features carries information about the emotional states, but it must fit the HMM structure as well. HMM provides better classification accuracies for speech emotion recognition as compared with the other classifiers [5]. HMM classifiers using prosody and formant features have considerably lower recall rates than that of the classifiers using spectral features [9]. The accuracy rate of the speech emotion recognition by using HMM classifier is observed as 76.12% for the speaker dependent in the previous study and that for the speaker independent it was 64.77% [1].

Transforming the original feature set to a high dimensional feature space by using the kernel function is the main thought behind the support vector machine (SVM) classifier, which leads to get optimum classification in this new feature space. The kernel functions like linear, polynomial, radial basis function (RBF) can be used in SVM model for large extent. In the main applications like pattern recognition and classification problems, SVM classifier are generally used, and because of that it is used in the speech emotion recognition system. SVM is having much better classification performance compared to other classifiers [1, 4]. The emotional states can be separated to huge margin by using SVM classifier. This margin is nothing but the width of the largest tube without any utterances, which can obtain around decision boundary. The support vectors can be known as the measurement vectors which define the boundaries of the margin. 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 are 75% and above 80% respectively [1, 7].

## V. CONCLUSION

Speech emotion recognition systems based on the several classifiers is illustrated. The important issues in speech emotion recognition system are the signal processing unit in which appropriate features are extracted from available speech signal and another is a classifier which recognizes emotions from the speech signal. The average accuracy of the most of the classifiers for speaker independent system is less than that for the speaker dependent.

Automatic emotion recognitions from the human speech are increasing now a day because it results in the better interactions between human and machine. To improve the emotion recognition process, combinations of the given methods can be derived. Also by extracting more effective features of speech, accuracy of the speech emotion recognition system can be enhanced.

## 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] I. Chiriacescu, "Automatic Emotion Analysis Based On Speech", M.Sc. THESIS Delft University of Technology, 2009.
[3] T. Vogt, E. Andre and J. Wagner, "Automatic Recognition of Emotions from Speech: A review of the literature and recommendations for practical realization", LNCS 4868, PP.75-91, 2008.
[4] S. Emerich, E. Lupu, A. Apatean, "Emotions Recognitions by Speech and Facial Expressions Analysis", 17th European Signal Processing Conference, 2009.
[5] A. Nogueiras, A. Moreno, A. Bonafonte, Jose B. Marino, "Speech Emotion Recognition Using Hidden Markov Model", Eurospeech, 2001.
[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] D. Ververidis and C. Kotropoulos, "Emotional Speech Recognition: Resources, Features and Methods", Elsevier Speech communication, vol. 48, no. 9, pp. 1162-1181, September, 2006.
[8] Z. Ciota, "Feature Extraction of Spoken Dialogs for Emotion Detection", ICSP, 2006.
[9] E. Bozkurt, E, Erzin, C. E. Erdem, A. Tanju Erdem, "Formant Position Based Weighted Spectral Features for Emotion Recognition", Science Direct Speech Communication, 2011.
[10] C. M. Lee, S. S. Narayanan, "Towards detecting emotions in spoken dialogs", IEEE transactions on speech and audio processing, Vol. 13, No. 2, March 2005.

**Ashish B. Ingale** received the B.E. degree in Electronics and Telecommunication Engineering from Sant Gadage Baba Amravati University, Amravati in 2008, and he is currently pursuing the M. Tech. degree in Electronic System and Communication (ESC) at Government College of Engineering Amravati. He has attended one day workshops on 'VLSI & EDA Tools & Technology in Education' and 'Cadence-OrCad EDA Technology' at Government College of Engineering Amravati.

**Devendra S. Chaudhari** obtained BE, ME, from Marathwada University, Aurangabad and PhD from Indian Institute of Technology, Bombay, Powai, Mumbai. He has been engaged in teaching, research for period of about 25 years and worked on DST-SERC sponsored Fast Track Project for Young Scientists. He has worked as Head Electronics and Telecommunication, Instrumentation, Electrical, Research and incharge Principal at Government Engineering Colleges. Presently he is working as Head, Department of Electronics and Telecommunication Engineering at Government College of Engineering, Amravati.

Dr. Chaudhari published research papers and presented papers in international conferences abroad at Seattle, USA and Austria, Europe. He worked as Chairman / Expert Member on different committees of All India Council for Technical Education, Directorate of Technical Education for Approval, Graduation, Inspection, Variation of Intake of diploma and degree Engineering Institutions. As a university recognized PhD research supervisor in Electronics and Computer Science Engineering he has been supervising research work since 2001. One research scholar received PhD under his supervision.

He has worked as Chairman / Member on different university and college level committees like Examination, Academic, Senate, Board of Studies, etc. he chaired one of the Technical sessions of International Conference held at Nagpur. He is fellow of IE, IETE and life member of ISTE, BMESI and member of IEEE (2007). He is recipient of Best Engineering College Teacher Award of ISTE, New Delhi, Gold Medal Award of IETE, New Delhi, Engineering Achievement Award of IE (I), Nashik. He has organized various Continuing Education Programmes and delivered Expert Lectures on research at different places. He has also worked as ISTE Visiting Professor and visiting faculty member at Asian Institute of Technology, Bangkok, Thailand. His present research and teaching interests are in the field of Biomedical Engineering, Digital Signal Processing and Analogue Integrated Circuits.