International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-5, Issue-1)

Research Article | **January 2016**

A Survey on Speech Emotion Recognition in Human Computer Interaction

[1]Sanju D. Garle, [2]Angad Singh, [3]Girish Kapse

[1, 2] Department of Information Science & Technology, NRI Institute, Bhopal, Madhya Pradesh India

[3] Department of E&TC Engineering, Army Institute of Technology, Dighi, Pune, India

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 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

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. 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 Neighbour (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.

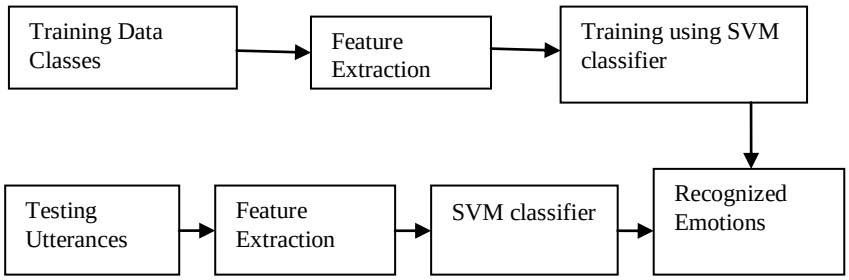

Fig. 1: Generalized System model for Emotion Recognition

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. 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) [4] and Support Vector Machine (SVM), Artificial Neural Network (ANN), Gaussian Mixtures Model (GMM). The paper reviews the mentioned classifiers. 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.

This review paper is organized as follows. In section II, a brief description about the speech emotion recognition system is given. Section III includes process of speech emotion feature extraction and Feature Selection. Section IV contains the speech emotional database description. In section V, information about the classifier selection, support vector machine classification, and . Finally, conclusion and future directions are given in section VI.

II. 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].

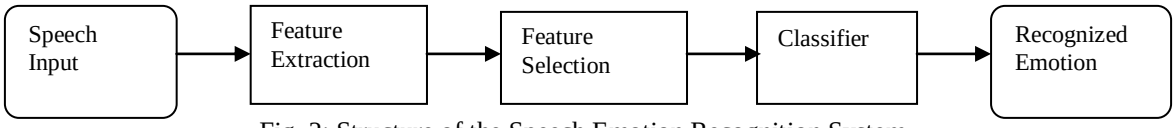

Fig. 2: Structure of the Speech Emotion Recognition System

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 [5]. The speech features can be classified 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, 10]. 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, 10]. 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 [5]. 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. 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].

A. Pitch: The pitch signal, or glottal waveform, has information about emotion because it depends on the tension and vibration of the vocal folds. Two features related to this signal are widely used, namely the pitch frequency (F0) and the glottal velocity volume.

B. Spectrum: The spectrum is characterized by these mentioned formants that model the spoken content. In fact, once the spectral envelope is estimated by using LPC method, further spectral features can be computed. Among them, there are the centroied, flux, roll-off, or even the ratio of spectral flatness. Furthermore, the long term average spectrum can be also obtained, a feature that gives important, general, spectral trends. Besides these specific features, other classical spectral features are also computed. Tools like Fast Fourier Transformation (FFT) easily allow us to get a glimpse over parameters such as phase, magnitude, intensity, or power coefficients in decibel scaling.

C. Mel-Frequency Cepstrum Coefficients (MFCCs): 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 [2]. Mel frequency cepstrum is an illustration of short term power spectrum of sound [10].

D. Zero Crossing Rate (ZCR): The Zero Crossing Rate counts how many times the speech signal changes its sign.

E. Spectral: the most important functional among the spectral statistical group is probably the computation of Discrete Cosine Transformation (DCT) coefficients, proven to be useful for speech recognition.

Specific selection of right features from the original data set will eliminate irrelevant features that might hinder the recognition rates. Consequently, it lowers down both the input dimensionality and the computational time of each experiment, since we will be dealing with a smaller, yet at least equally effective, set of data. It will be even easier to state generalized conclusions, because the remaining key features should be less correlated.

IV. SPEECH EMOTIONAL 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. This paper divides 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.

V. 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. The paper reviews the mentioned classifiers [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.

VI. CONCLUSION AND FUTURE DIRECTIONS

In this review, 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 with their performance through this review it is found that SVM having better classification compared to other classifier. From this review, In the future research we can 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 can be used to select the best(or relevant) feature subset. so that speed will increase and also reduce features hence improve accuracy. The performance of speech emotion recognition system is usually specified in terms of accuracy and speed.

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] 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] ChiuYingLay,NgHianJames."GenderClassificationfromSpeech",(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.

