A Survey on Speech Emotion Recognition Using MFCC and Different classifier

Supriya B.Jagtap[1], Dr.K.R.Desai[2], Ms. J. K. Patil[3]

[1]P.G. Scholor,Dept. of Electronics and Telecommunication Engineering,
Bharati Vidyapeeth's College of Engineering, Kolhapur, (India)
[2]Prof. and H.O.D, Dept. of Electronics and Telecommunication Engineering,
Bharati Vidyapeeth's College of Engineering, Kolhapur, (India)
[3]AssociateProf.,Dept. of Electronics and Telecommunication Engineering,
Bharati Vidyapeeth's College of Engineering, Kolhapur, (India)

ABSTRACT

In this paper methodology for emotion recognition from speech signal is presented. Speech emotion recognition means extracting the emotional state of speaker and detecting the actual intension of the speaker through his or her speech. The goal is to recognize the emotions likeHappiness, Anger, Boredom, Sadness, Surprise, Fear and Neutral. This Paper presents survey of three methods forSpeech emotion recognition, Application of features like energy, formant, Mel frequency cepstral coefficient (MFCC)and differentclassifiers such as Support Vector Machine (SVM), Binary SVM,K-Nearest Neighbors approach(KNN),Radial Basis Function(RBF), Random Decision Forest(RBF)and Gaussian mixture Model(GMM) are discussed.In addition to the mentioned techniques it gives an outline of the areas where emotion recognition could be utilized such as healthcare, psychology, smart phones, marketing, call centers and cognitive science.

Keywords:Emotion recognition, speech features, classification methods, speech database

I.INTRODUCTION

Emotions play a vital role in human communication. Speechis one of the most natural forms of communication between human and computer. With the visitationof technology in the recent years, more intelligent interaction between humans and machines is desired [1]. The importance of recognizing emotions from human speech has grown with the increasing role of spoken language interfaces in human-computer interaction applications. The goal of speech emotion recognition system is to understandemotions which are present in speech and synthesizing actual intention of the person.And recognize the emotions likeHappiness, Anger, Boredom, Sadness, Surprise, Fear and Neutral.

From a recent scenario, for human emotion recognition through speech signal a wide ranging research is made and this researches using different speech information and signal. Many researchersuse different classifiers or also develop own classifier [1]. For recognition of emotional state researchers use different classifier such as,Support Vector Machine (SVM), Gaussian Mixture Model (GMM), Binary SVM,K-Nearest Neighbors approach(KNN),Radial Basis Function(RBF), Gradient boostingand Random Decision Forest(RBF)[3]. All this classifier used for a various application and it gives a better performance rate. Emotion recognition mostly used to develop wide range of applications such as stress management for healthcare center, psychology, marketing, call centers and cognitive science areas and identifying person's emotionstate at the moment and making appropriate treatment can enhance the motivation of person.

Speech emotion recognition system has three main methods which is used to detecting the humans emotion, in first method extractinga feature by using MFCC which is one of the spectral features, in second method using a different classifier and third method is the database which is present in any form and thenrecognize the emotion at the output. This is a basic system used for speech emotion recognition. Instead of using this system we find out actual intension and emotion of person. And this system is suitable to be used over Hospitals, Healthcare centers, and Smartphone platforms and able to recognize different emotional states of human.

In the next section, the previous related work on speech emotion recognition systems is explained Subsequent, Section 2 provides an insight on the database that is used for implementing the system and provides the framework along with the approach used for featureextraction using Mel Frequency Cepstral Coefficient (MFCC) explained in detail. In next section, the different classifiers are discussed.

II.SPEECH EMOTION RECOGNITION FRAMEWORK

Speech emotion recognition system is basic object recognition system. This shows that the stages involved in the object recognition system are also present in the Speech emotion recognition system. There are five main modules in speech emotion recognition system consist emotional speech input, feature extraction, feature selection, classification and recognized emotional output [2]

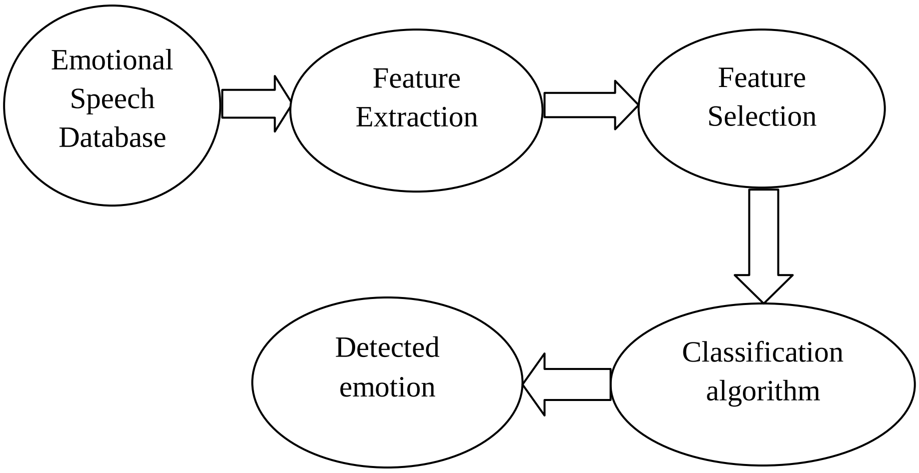

Fig.1 Basic block diagram of speech emotion recognition

2.1. Speech: The act of speaking; expression or communication of thoughts and feelings by spoken words.

Emotions: It is a change in physical and psychological feeling which influences behavior and thought of humans.

2.2. Database: Database is very important part of speech emotion recognizer; the main role is to check quality, naturalness and noise level of speech signal from database used in performance. Database can be divided into three categories a) Actor (Simulated) b) Elicited (Induced) c) Natural emotional speech databases.

1) Simulated Speech Databases: It is collected be recording artists or actors expressing linguistic neutral sentences in different emotions. In this category databases is standardized and result can be easily compared.

2) Elicited Database: Elicited Databaseis recorded by artificially creating emotional situations without the knowledge of the speaker.

3) Natural Databases: Natural Databaseis created by collecting the real world conversations such as Call centre conversation [5].

In some decade this standard database has been used, such as Berlin database, speech under simulation and actual stress (SUSAS) database, hanbat database.

2.3. Feature Extraction and selection:

The system extractsthe best features from the audio signal. Different features represent different speech information in highly overlapped manner. Speech features are divided in some categories: Prosody features, Vocal Tract Features and Excitation Source Features.

1) Prosody Features: The most popular prosody features includes, Fundamental Frequency (F$_{0)}$, Energy, Duration, Formants.

2) Vocal Tract Features: Vocal Tract Features are obtained by analyzing the characteristics of Vocal tract, which are well reflected in frequency domain analysis of speech signal.

3) Excitation source features: are obtained from speech signal after suppressing vocal tract characteristicEnergy and pitch are basic features of speech signals.MFCC reduces the computational complexity of the approach, gives better ability to extract the features and can be find the different parameters like

- Pitch of speech: The actual frequency calculated in Hertz (Hz) and pitch frequency measured on the Mel Frequency Scale. Using equation number…. (1) Mel frequency calculated.
- Energy: The value of energy can be obtained by calculating mean value, local maxima, local minima, variance in each of the speech signal [4].

Extraction ofMel Frequency Cepstral Coefficient (MFCC):

In speech recognition the Mel frequency cepstral coefficients are most widely used feature. It represents the short term power spectrum of sound, based on liner cosinetransform of a logpower spectrum on a non-linear Mel scale of frequency. MFCC based on the human auditory perception with regard to frequencies [3]. Fig.2 shown block diagram for MFCC feature extraction [4].

- Pre-emphasis: Pre-emphasis is required to increase signal energy. In this process, speech signal is passed through a filter which increases the energy of signal.
- Framing: divide the speech signal into frames' usually by applying a window function at fixed intervals that is signal with time length 20 to 40ms.

- Windowing: After framing process, the windowing process is performed. Windowing function reduces the signal discontinuities at the start and end of each frame. In this process, frame is shifted with a 10ms span. That means each frame contains some overlapping portion of previous frame.

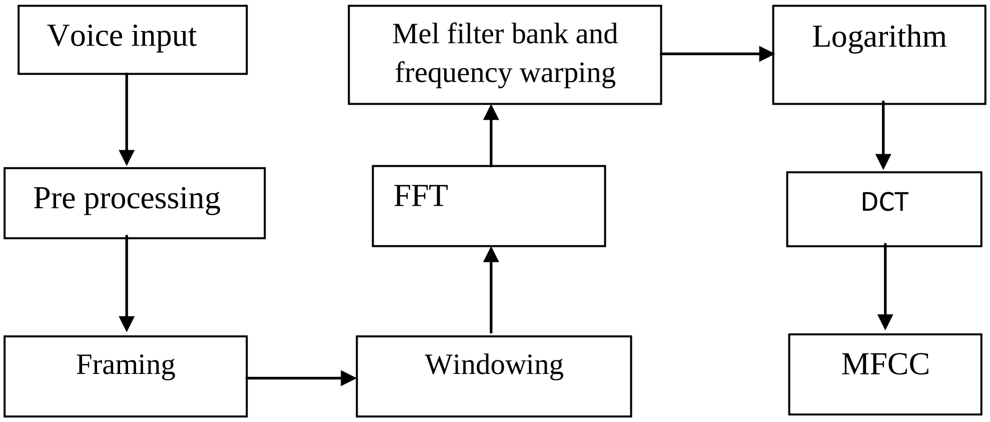

Fig-2 Block diagram for MFCC Feature extraction

- Fast Fourier Transform (FFT): FFT algorithm is used for converting the n samples from time domain to frequency domain. FFT is used to generate the frequency spectrum of each frame.
- Mel Filter Bank: Mel scale filter bank: This is a set of 20-30 triangular filtersapplied to each frame. The Mel Scale Filter Bank identifies howmuch energy exists in a particular frame. The mathematicalequation to convert the normal frequency *f*to the Mel scale *m*is as follows,

$Mel (F) = 2595\times \log10 (1 + f / 700)$………………. (1)

- Cepstrum: the obtained Mel Frequency Cepstrum is converted back to time domain with the help of DCT algorithm [4].

2.4. Speech emotion classification using different classifier:

Support Vector Machine (SVM)classifier: Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. It provides very simple method for linear classification. But performance in case of non linearityseparable data largely depends on the choice of kernel.In a large database learning or training is very fast and effective and its accuracy is comparatively better in comparison to the other techniques [3, 6,10].

K-nearest neighbor (KNN) classifier:KNN is a type of lazy learning. This method is used for classification and regression.KNN is the simple classification method which is nearer neighbors contributes more to the average than the more distant ones. KNN is used when we do not have any prior knowledge of data distribution. It is simplest algorithm than other machine algorithm. This method is based on Euclidean distance between the features of test sample and train sample [6].

Gaussian Mixture Model (GMM) Classifier:A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities of feature vector,~*x*, which is dimensional continuous valued data vector, by the linear combination of multivariate Gaussian distribution [6].GMM parameters are calculate approximate values from training data.

Radial Basis Function (RBF): A radial basis function (RBF) is a real value function whose value depends only on the distance from the origin. Sums of radial basis functions are typically used to approximate. This approximation process can also be interpreted as a simple kind of neural network [7].

Random Forests (RDF): To increase computational power, Random Forest uses an ensemble of learning methods. It isused for regression, classification and other tasks. Random Forests works by constructing a large number of decision trees at the time of training and it outputs the mean prediction or mode of the class of the individual trees. A random decision forest prevents decision trees from over fitting the training data [3].

Gradient boosting:it is a technique used for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods, and it generalizes them by allowing optimization of an arbitrary differentiable lossfunction [3]**.**

The rest of paper is organized as follows section 3 gives brief review of work done by earlier researches in the field of MFCC using different classifiers with their results. And section 4 presents a concluding remark based on study presented in section 2.

III.LITERATURE SURVEY

In the last decades, there has been a lot of work in speech emotion recognition which can be mainly described as a wide variety of complex classifiers and features extracted by using MFCC to guarantee a better accuracy (8).Previous work in this area included use of various classifiers like SVM,K-Nearest Neighbor (KNN), Gaussian Mixture Model (GMM),Radial Basis Function (RBF)andBinary SVMClassifier etc. are discussed.

Mohan Ghai et al. [3] developed a system using MFCC and tested on Berlin database. Proposed work is done by using a different classifier namely Support Vector Machine (SVM), Random Decision Forest (RDF) and Gradient Boosting. It identifies seven classes of emotion.SVM achieved an average accuracy of 55.89%, Gradient boosting achieved 65.23% and random decision forest classifier achieved 81.05%. Out of this three classifiers highest accuracy is obtained in a random decision forest classifier (RDF) 81.05% accuracy is achieved.

Milton et al. [7] developed a system using MFCC. In this system they use a Berlin emotion database. It consists of 535 acted emotions in German language with 7 different emotions andit is a multi-speaker database. In this proposed work they use Three Stage SVM classifier. And this classifier is also compared with some other classifier which recognizesseven classes of emotions like: Happiness, Angry, Disgust, Fear, Boredom, Sadness, and Neutral. This author achieved average accuracy of Three Stage SVM 68% and compared with SVM using Radial Basis Function (RBF), Linear Kernel i.e.55.4%, 65% and 68%. Out of these three classifiers highest accuracy is obtained in a SVM 68% accuracy is achieved.

Chandra prakash, and prof.V.B.Gaikwad [8],developed a system by using MFCC and tested on Berlin database. In this paper emotion is recognized through speech using spectral features like pitch, energy and study is carried out using K-Nearest Neighbor(KNN),Support Vector Machine(SVM), and Gaussian Mixture Model(GMM) which is used for detection of six emotional states of speaker such asHappiness, Angry, Disgust, Fear, Sadness, and Neutral. This author achieved an average accuracy of 78% for KNN, 84% for GMM, and 73.50% for SVM. Out of this three classifiers highest accuracy is obtained in a Gaussian Mixture Model (GMM) 84% accuracy is achieved.

N.Ratan Kanth and S.Saraswathi [9], in this paper author used recording of professional artists from All India Radio (AIR) as a database. They use features related to pitch, loudness, frequency, and energy etc.are considered. Feature selection is done using correlation based feature selection (CFS) and best first search (BFS) for selection o best features. in this paper author uses a two classifiers four binary SVM and multiclass SVM. Binary SVM achieved an average accuracy of 77.78% and multiclass SVM achieved 58.7%. Out of thesetwo classifiers highest accuracy is obtained in a binary SVM 77.78% accuracy is achieved.

Table 1 gives brief comparison of above all methods with respect to classifiers used, accuracy achieved and number of emotions detected.

Table.1:comparison of speech emotion recognition systemsusing different classifiers

Name of author	Classifier	Accuracy in %	Emotion Classified
Mohan Ghai et al.[3]	RDF	81.05	7
	Gradient Boosting	65.23	
	SVM	55.89	
Milton et al. [7]	SVM Using RBF	55.40	6
	SVM Using Linear kernel	65.00	
	3-stage SVM	6800	

Chandra prakash, and Prof. V. B. Gaikwad[8]	KNN	78.00	7
	GMM	84.00	
	SVM	73.00	
N. Ratan Kanth and S. Saraswathi[9]	Binary SVM	77.78	4
	Multiclass SVM	58.07	

IV.CONCLUSION

In this paper, we studiedmostly preferred speech emotion recognition methods, methodology for featureextraction and different classifier used for emotion classification.Wediscussedabout MFCCtechnique and there different characteristicof speech signal such as pitch, energy, Loudness and frequency. Mel Frequency Cepstral Coefficient (MFCC) reduces the frequency information of speech signal into small number of coefficient which is easy to compute and extract the features. In this survey paper, we discussed some classifiers. It is found that accuracy of the system depends on proper database of speech samples.

REFERENCES

[1]Ritu D.Shah1, Dr. Anil. C. Suthar,"Speech Emotion Recognition Based on SVM Using MATLAB", DOI: 10.15680/IJIRCCE.2016. 0403004,International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 3, March 2016

5 – 8887) Volume 69– No.9, May 2013

[2]Akshay S. Utane, S.L. Nalbalwar, "Emotion Recognition through Speech", International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA

[3]Mohan Ghai, Shamit Lal, Shivam Dugga l and Shrey Manik, "Emotion Recognition On Speech Signals Using Machine Learning", 978-1-5090-6399-4/17/$31.00_c 2017 IEEE,2017 International Conference On Big Data Analytics and computational Intelligence (ICBDAC)

[4]Sandeep Pathak1, Vaishali Kolhe, "A Survey on Emotion Recognition from Speech Signal", International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 5, Issue 7

[5]Bryan E. Mart´ınez and Jaime Cerda Jacobo, "An improved characterization methodology to efficiently deal with the speech emotion recognition problem", 2017 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC 2017). Ixtapa, Mexico

[6]Saikat Basu, Jaybrata Chakraborty, Arnab Bag and Md. Aftabuddin," "A Review on Emotion Recognition using Speech", 978-1-5090-5297-4/17/$31.00 ©2017 IEEE, International Conference on Inventive Communication and Computational Technologies (ICICCT 2017)

[7] A. Milton, S. Sharmy Roy,S. Tamil Selvi, "SVM Scheme for Speech Emotion Recognition using MFCC Feature", International Journal of Computer Applications

[8]Chandra prakash, prof. V. B. Gaikwad, "analysis of emotion recognition system through speech signal using KNN, GMM AND SVM Classifier", IJECS VOLUME.4

[9]N. Ratna Kanth, S. Saraswathi, "Affect Recognition in Telugu Emotional Speech Using MFCCFeatures**",** International Journal of Computer & Mathematical Sciences IJCMSISSN 2347 – 8527Volume 6

[10]Miss. Sneha S. Mane, Dr. K. R. Desai, "Intelligent Facial Emotion Recognition using modified-PSO ",*International Journal of Engineering and Techniques - Volume 3 Issue 3, May-June 2017*

[11]P.Vijayalakshmi, A. Anny Leema,"Real-time Speech Emotion Recognition Using Support Vector Machine", International Journal of System and Software Engineering Volume 2 Issue 1