COMPARATIVE ANALYSIS OF SPEAKER DEPENDENT, SPEAKER INDEPENDENT AND CROSS LANGUAGE EMOTION RECOGNITION FROM SPEECH USING SVM

E. Sarath Kumar Naik, PG Student [Embedded Systems]
Dept. of Electronics and Communication Engineering,
Gudlavalleru Engineering College,
Gudlavalleru, Andhra Pradesh, India.

K. Suvarna, Assistant Professor
Dept. of Electronics and Communication Engineering,
Gudlavalleru Engineering College,
Gudlavalleru, Andhra Pradesh, India.

***Abstract*—The main objective of this paper is to examine emotion recognition attainment in speaker dependent, speaker independent and cross language emotion recognition from speech. The utterance sample is observed with Mel-Frequency Cepstral Coefficients (MFCC's) for deriving acoustic features and then used to train Support Vector Machine (SVM) and eventually the calculated log likelihood from training is stored to database. It will identify the emotion of the speaker by inspecting the log value from the database. It is implemented in MATLAB 12b environment and showing 85.77% for speaker dependent, 55.51% for speaker independent and 49.83% for cross language results as correct consent.**

Keywords—Emotion Recognition; Cross language; MFCC; Speaker dependent emotion recognition; Speaker independent emotion recognition; Cross language emotion recognition.

I. INTRODUCTION

Speech is one of the natural process for mortals to communicate. Humanistic Utterance is an appropriate aspect for any personal. Emotions are indispensable for dispatching critical information; existence of emotion makes speech more natural. Speech signal carries speaker data also linguistic information. The abundance of information in articulation has excited several scientists to advance the system that naturally process the speech, this speech technology has abounding applications [1]. Speech signal accommodates acutely rich information that utilizes amplitude- modulated, time-modulated and frequency- modulated carriers to convey information regarding words, speaker identification, way of pronunciation, response, accent, the condition of health of the orator and utterance. All the above- mentioned information are transferred fundamentally within the classical telephone bandwidth of 4 kHz. The tone intensity above 4 kHz frequently conveys audio quality and sensation.

An important concern in speech emotion recognition is to regulate a set of important emotions to be categorize by an automated emotion recognizer. It is very ambitious to analyze all types of emotions present in speech. However, we will deal with six emotions in this task namely anger, disgust, surprise, happy, sad and fear.The exercise of speech emotion recognition is extremely stimulating, because it is not clear which speech features are most athletic in differentiating the emotions. The foremost spotlight has been accustomed to the spectral features, viz. Mel-Frequency Cepstral Coefficients (MFCC's) [2], [7], [9]. The capability of MFCCs has been scrutinized for speaker dependent, speaker independent and cross language emotion recognition.

Rest of the paper is organized as follows Section II describes the extraction of MFCCs features. Section III describes the classification model used for emotion recognition. Section IV describes the architecture of emotion recognition system. Section V describes experimental results and Section VI concludes the paper.

II. MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC'S) FEATURE EXTRACTION

In sound processing, the Mel-Frequency Cepstrum (MFC) may be a portrayal of the short-term power spectrum of a sound, supported a linear cosine remodel of a log power spectrum on a nonlinear Mel scale of frequency.

Before performing feature extraction the speech samples are to be collected. English and Telugu languages were used for conducting training and testing purpose. The collection of speech samples is to be done in a closed room and by using microphone so as to eliminate noise in the speech samples.

Mel-Frequency Cepstral Coefficients (MFCCs) are coefficients that generally improvise an MFC. They are

derived from a category of Cepstral illustration of the audio clip [2]. The primary step in any automatic speech recognition system is to extract options [7] i.e. describing the elements of the audio signal that are sensible for distinguishing the grammatical content and eliminate all the alternative junk that bears information like backdrop clamor, emotion etc [8]. The method to calculate Mel-frequency Cepstral Coefficients (MFCCs) is shown in Fig. 1.

Calculation of MFCCs involves the followings steps:

1. Frame the signal into short frames.
2. For every frame determine the periodogram estimate of the power spectrum.
3. Assign the Mel-filterbank to the power spectra; add the energy in every filter.
4. Take the exponent of all filterbank energies.
5. Take the DCT of the log filterbank energies.

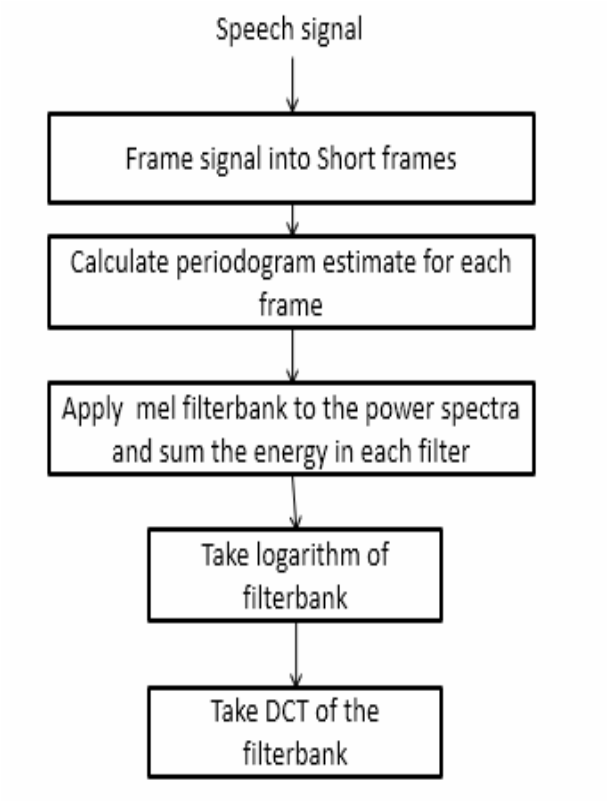

Fig. 1. Process to calculate Mel- frequency Cepstral coefficients (MFCCs)

III. Model Classification

Support Vector Machine

Support Vector Machine (SVM) could be a supervised machine learning algorithmic program which may be used for each classification and regression challenges. However, it is principally utilized in classification problems. During this algorithmic program, we tend to plot every information item as a degree in n- dimensional area with the worth of every feature being the value of a selected coordinate. Then, we tend to perform classification by finding the hyper-plane that differentiate the two categories very well.

The SVM approach is a high dimensional vector supervised learning methodology that is based on emotion assumptions. It predicts presence or absence of such specified feature of a category is not associated with the presence or absence of all different options. It is terribly easy to program and execute it, its parameters are easy to assume, learning or training is extremely quick and effective even on very giant databases and its accuracy is relatively higher as compared to the techniques [4].

Support vector machine for classification will be seen because the application of perceptron. Once classification problem is linearly dissociable, a hyper plane that produces two classes of dissociable, a hyper plane that produces two classes of dissociable data more near the plane is set up; sometimes the plane is named optimal separation hyper plane.

Relating to nonlinear problem, original data is mapped from a low dimensional space to the new data sets of a higher feature space through a nonlinear mapping. New knowledge sets are currently linearly dissociable in feature space, thus the classification in higher dimensional space is completed.

IV. Architecture Of Emotion Recognition System

System architecture used for emotion recognition is shown in Fig.2. Architecture has two phases; training phase shown in Fig 2(a) and testing phase shown in Fig 2(b).

In training phase, feature extraction is the first step. Mel-Frequency Cepstral Coefficients (MFCCs) are extracted from speech utterances of better-known emotions [2], [9]. As audio signal is consistently changing, therefore to modify things the speech signal is segmented into short frames as in short time scales the audio signal does not change much. Then calculate the power spectrum of every spectrum. This is often intended by the human cochlea that vibrates at completely different spots counting on the frequency of incoming sounds. Depending on the location in the cochlea that vibrates, completely different nerves fireplace informing the brain that bound frequencies are present. Our periodogram estimate performs the same job i.e., identifying which frequencies are within the frame. Once when getting filterbank energies, take the logarithm of them, and so then compute the DCT of the log filterbank energies [3], [6]. Emotion recognition model (SVM) is trained using the MFCC feature vectors.

In testing phase, feature vectors similar to the test utterances are given input to all or any trained models to seek out the emotion present in those utterances i.e. Mel- Frequency Cepstral Coefficients (MFCCs) are extracted from test samples and are trained using Support Vector Machine. After

successful feature extraction and model training, the obtained feature vectors are stored in database. Decision block selects the actual model having highest likelihood value [2]. As an example suppose for explicit speech sample, anger emotion model offers highest likelihood value compare to all or any other different emotional models [5], [6].

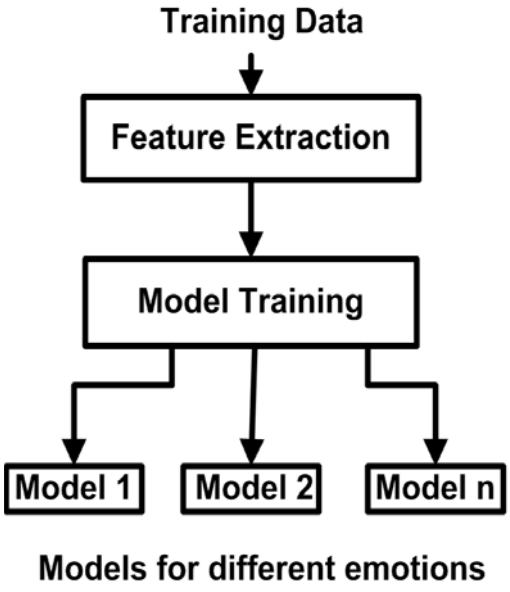

Fig. 2(a). Training Phase

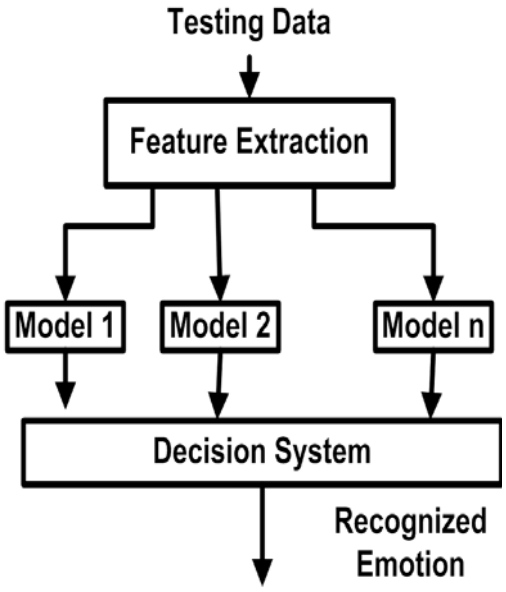

Fig. 2 (b). Testing Phase

Fig. 2. Architecture of emotion recognition system

V. RESULTS

In training phase, after performing feature extraction and Support Vector Machine (SVM) trained using Mel- Frequency Cepstral Coefficients (MFCCs) the feature vectors are stored to database. In testing phase, the test samples are gone through the same process as in testing face the obtained feature vectors are compared with the stored feature vectors in database and the decision block determine the emotion status of the user. The six emotions considered are Anger, Disgust, Surprise, Fear, Sad and Happy. The emotion statuses of the user are shown in fig.3.

Fig.3. Emotion statuses of user

Each emotion is repeated for 150 iterations, correct acceptance and false acceptance of each emotion is recorded and average of correct acceptance is taken which are then averaged to attain the single average value for speaker dependent, speaker independent and cross language.

A. *Speaker Dependent Emotion Recognition*

In speaker dependent emotion recognition, single speaker information for six emotions is employed for each training and testing purpose i.e. one speaker information is stored in database and same speaker data is once more used to test and train against the information stored in database. Table I shows the emotion recognition performance of speaker dependent.

TABLE I

PERFORMANCE OF SPEAKER DEPENDENT EMOTION RECOGNITION

	Anger	Disgust	Fear	Happy	Surprise	Sad	Avg %
Anger	132	2	1	9	0	6	
Disgust	6	128	5	1	10	0	
Fear	2	1	136	10	0	1	
Happy	3	0	1	140	2	4	85.77
Surprise	12	0	2	0	122	10	
Sad	2	15	6	9	4	114	

B. *Speaker Independent Emotion Recognition*

For efficient human-computer interaction, computer ought to have the capability to acknowledge mechanically the emotion present in speech vocalization spoken by any individual. In speaker independent emotion recognition, two totally different speaker information for six emotions is employed for training and testing purpose i.e. one speaker information is stored in database and another speaker information is used to train and test against the info stored in database. Table II shows the emotion recognition performance of speaker independent.

TABLE II

PERFORMANCE OF SPEAKER INDEPENDENT EMOTION RECOGNITION

	Anger	Disgust	Fear	Happy	Surprise	Sad	Avg %
Anger	89	1	7	16	11	26	
Disgust	2	78	10	25	22	13	
Fear	16	5	103	15	2	9	
Happy	43	12	6	76	5	8	55.51
Surprise	33	0	3	2	94	18	
Sad	7	51	17	5	10	60	

C. *Cross Language Emotion Recognition*

In cross language emotion recognition, two completely different speaker information of two different languages for six emotions is employed for training and testing purpose i.e. one speaker information of one language is stored in database another speaker information of another language is used to train and test against the information stored in database. Table III shows the emotion recognition performance of cross language.

TABLE III

PERFORMANCE OF CROSS LANGUAGE EMOTION RECOGNITION

	Anger	Disgust	Fear	Happy	Surprise	Sad	Avg %
Anger	80	16	2	31	18	3	
Disgust	28	72	13	7	9	21	
Fear	17	3	81	6	4	39	
Happy	53	8	4	78	3	4	49.83
Surprise	19	4	2	52	63	10	
Sad	3	5	47	9	11	75	

From Table I, II, III, it is clear that the testing emotion may not appear every time with correct acceptance i.e. while iterating a emotion speech for 150 times there may be a chance of obtaining incorrect emotion e.g. angry emotion is iterated for 150 times there is no need to get angry emotion as correct acceptance for all the time it may show any other emotion (Fear or Happy or Disgust or Surprise) as correct acceptance which is a false acceptance as in Table I, II, III.

VI. CONCLUSION

In this paper speaker dependent, speaker independent and cross language emotion recognition from speech using Support Vector Machine (SVM) has been analysed and compared. English and Telugu languages were used for conducting training and testing purpose. It is ascertained that emotion recognition performance depends on the speaker. Correct acceptance rate in speaker dependent is more when compared to speaker independent and cross language emotion recognition. Emotion recognition varies from language to language. Performance of Cross language emotion recognition is extremely less when compared to both Speaker dependent and Speaker independent.

ACKNOWLEDGMENT

We would like to thank our HOD Dr. M. Kamaraju, in department of Electronics and Communication Engineering in Gudlavalleru Engineering College for his great encouragement and facilities provided for this project.

REFRERNCES

[1] ManayBhayakar, JainathYadav and K. Sreenivas Rao, "Speaker Dependent, Speaker Independent and Cross Language Emotion Recognition from Speech using GMM and HMM", *in proc. IEEE National Conf., on communication (NCC)*, pp 1-5, 2013.

[2] Stephen A. Zahorian and Hongbing Hu, "A Spectral/temporal method for robust fundamental frequency tracking", in *Acoustical society of America*, pp. 4559-4571, 2008.

[3] S. koolagudi, R. Reddy, J. yadav, and K. Sreenivasa Rao, "IITKGP-SEHSC: Hindi speech corpus for emotion analysis", *in International Conference on Devices and Communications (ICDeCom)*, pp. 1-5, Feb. 2011.

[4] Hai-yan Yang, Xin-xing Jing, "Performance Test of Parameters for Speaker Recognition System based on SVM_VQ", *in proceedings IEEE International conference, on Machine Learning and Cybernetics*, pp 321-325, 2012.

[5] L. R. Rabiner, "A tutorial on hidden markov models and selected application in speech recognition", *proceedings of IEEE*, vol. 77, pp. 257– 286, Feb 1989.

[6] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. S. Rao, "IITKGP-SESC: Speech database for emotion analysis",*Springer-Verilog Berlin Heidelberg*, vol. 40, pp. 485–492, 2009.

[7] D. A. Reynolds and R. C. Rose, "Robust text independent speaker identification using Gaussian Mixture speaker model",*IEEE Trans. Speech Audio proc*., vol. 3, pp. 72–83, Jan 1995.

[8] H. R. Pfitzingera, N. Amirb, H. Mixdorffc, and J. Rosela, "Cross Language perception of hebrew and german authentic emotional speech", *in proceeding of 17th ICPhS*, Aug 2011.

[9] BiswajitNayak, MitaliMadhismita, Debendra Kumar Sahu, Rajendra Kumar Behra and Kamalakanta Shaw, "Speaker Dependent Emotion Recognition from Speech", *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 3, pp. 40-42, Nov 2013.

AUTHOR'S PROFILE

E. Sarath Kumar Naik is M. tech student in Gudlavalleru Engineering College, Andhra Pradesh, India. He has completed B. Tech from Prasad. V. Potluri. Siddhartha Institute of Technology, Andhra Pradesh, India, in 2014.

K. Suvarna is Assistant professor in ECE department of Gudlavalleru Engineering College. She has about 12 years of experience and presented papers in several International and National Conferences and published papers in International journals. Her area of interest includes Signal Processing, Antennas.