# RECOGNITION OF EMOTIONS FROM AUDIO SIGNALS

Swapnali Tandel[1], Shital Patil[2], Vikrant Kadam[3], Srijita bhattacharjee[4]
[1,2,3]Student, PHCET, Rasayani, [4]Faculty PHCET, Rasayani

**Abstract**
**In this paper we have designed a system which is used to recognize human emotions from audio clip generated by speaker. In this system we have employed two statistical models such as SVM and HMM to classify emotions .In order to recognize emotions we extracted four acoustic features such as spectral centroid, spread, flatness and projection. This system is divided in to five different stages-audio preprocessing, feature extraction, segmentation, model training and classification. Audio preprocessing is used to remove noise present in the signal. In the feature extraction part, we extracted four acoustic features. Segmentation is used to divide audio clips in to voiced and unvoiced category. In training phase we trained models in order to classify emotions. For training purpose we generated audio database which consists of speech utterances belongs to ten different emotions. After training, in the stage of emotion verification we verified emotions by using our classifier SVM and HMM based on the category of emotions.**

**Index Terms: Speech Emotion Recognition, Support Vector Machines, Hidden Markov Models, Spectral Features**

## I. INTRODUCTION

Speech signal is to use to share information with each other. Information theory defined as speech is a source of message content or information. The nature of information communicated through speech is discrete. Speech Processing is an application of digital signal processing to process and analyze speech signal. Speech Processing is an application of Digital Signal Processing (DSP) to process and analyze audio signal. Its main objective is to transmit the speech signal efficiently. According to the World Health Organization's (WHO's) definition emotional well-being is the fundamental component of health. Research on well-being indicated that peoples with positive emotions like happiness, joy and contentment tend to work towards goal and such peoples attract other peoples with their positive attitude, energy and optimism [1]. There are various techniques are there in practice to identify human emotions, but here we are going to concentrate on recognition of emotions through speech. As speech signal is not only conveying massage but also it conveying emotions. Emotional changes are directly reflected in to the human speaking rate, tone, intonation and so many linguistic features. This technique helps doctor's to identify mental status of the patient by using Human Computer Interaction (HCI) [2].

Human emotions are mixtures of various psychological and physical factors. Although people may reveal their emotions in speech, speaking rates, stress, intonation and style of speaking also emotions are varying along person to person. To increase recognition rates, developing discriminant features from speech and selection robust classifier is very much important thing. For healthy life emotional wellbeing the social and behavioral sciences can play a very important role [10]. The Darwin's theory states that, each emotion induces some physiological and psychological changes in order to prepare peoples to do something, e.g., fear prepare us to run away from a danger, success makes peoples happy where as tension makes people sad and nervous. The emotion recognition technique is divided in to following stages: Audio preprocessing, feature extraction, segmentation, training and testing. Along with these stages emotion database is also an

essential part of speech recognition technique [6].

Classifiers that are used to recognize emotions from speech are Support Vector Machines (SVMs), Hidden Markov Models (HMMs), Artificial Neural Networks (ANNs), Gaussian Mixture Models (GMMs), K- Nearest Neighbors (KNNs). Here we are extracting spectral centroid, spectral flatness, spectral spread and spectral projection features from input speech signal in order to recognize emotion from speech. After feature extraction feature verification stage is there. We are forming database to classify 10 different emotions like happiness, sadness, fear, anger, boredom, disgust, excited, frustrated, surprise and neutral. To recognize these emotions we recorded sound clips belonging to various emotion categories. The sentences that we are using in our daily communication have been recorded in order to build database. Hence the database is created by own recorded audio clips [15].

## II. LITERATURE SURVEY

Features extracted from the audio signal have a great effect on the reliability of an emotion recognition system. Based on these features, the system is identifying emotions from speech. Naseem [4] was the first to develop a classification system based on compressive sensing/sparse representation for speaker identification. Their experiments were conducted using the TIMIT database and a sparse representation classifier. Lee [7] developed an emotional saliency detector by combining discourse information with acoustic features. Their experiment proved that the detector obtained satisfactory results by using a linear discriminant classifier. Very recently, Mower [3], [5] proposed an emotion recognition system based on three metrics of emotional expressions (dominance, valence and activation). Their system used the Support Vector Machine (SVM) to obtained emotion expression scores and then combined these scores together to identify emotions.

The advantage of spectral features over prosodic features is that they can obtain energetic information from the speech spectrum in frequency ranges of interest [2]. Nwea investigated the feasibility of using log frequency power coefficients (LFPCs) to model motions by comparing LFPCs with MFCCs and LPCCs[8]. Their experimental results proved that the average accuracy achieved by using LFPCs exceeded that achieved by using MFCCs and LPCCs, respectively. Paeschke also analyzed the pitch shape in expressive speech [12]. They invented different pitch features that might be helpful for speech processing, such as the steepness of rising and falling of the pitch, and direction of the pitch contour [9].

## III. SYSTEM OVERVIEW

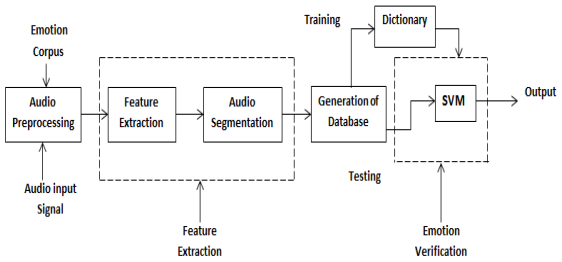


Fig. 1 Block Diagram of Speech Emotion Recognition

Fig. 1 shows system overview of speech emotion recognition using Support Vector Machines (SVM) and Hidden Markov Model (HMM). The proposed system is divided in to two sections: Extraction of audio features and Verification of emotions. The idea of proposed system is to segment sound clip by classifying into different classes as anger, happiness, sadness, frustrated, disgust, neutral, surprise, boredom excited and fear. It begins by extracting various audio power features. This extracted data is trained by SVM supervised learning models. This trained data is forwarded to classification models such as SVM [14]. Then it is classified into different classes of emotion as mentioned above. In proposed approach we use model based technique for effective results and because of better performance. In order to train classifier we are creating emotional database of 10 different emotions. We recorded various audio clips that we are going to use in our daily life communication [13].

## IV. FEATURE EXTRACTION

The features are calculated based on data and frequency of an audio file. A process of feature extraction is shown in figure 2:

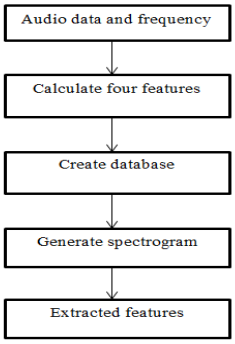


Fig.2 Feature Extraction Process

This module aims to extract or calculate features of every audio file which has to be verified. Audio power features are used in this system such as spectral centroid, spectral spread, spectral flatness and spectral projection. The features are calculated based on data and frequency of an audio file. Select audio data for feature extraction. Take audio data and its frequency as input of feature extraction process. Calculate features of frames using formula:

$$\frac{Abs\ (X) * 2}{i^{th}\ Block\ Length} \qquad (1)$$

Where, X is the values taken from the spectrogram. Create database of above mentioned four features. This database consists of matrix of 1:50 of each audio file. Generate the spectrogram of audio data as well as four audio power features as amplitude vs. time. It is generated based on audio data, window size, block length, and hop length. Finally, the features of audio data are extracted.

*A. Spectral Centroid*

Spectral centroid is defined as the measurement of the brightness of sound. This measurement is achieved by estimating the "center of gravity" using the magnitude and frequency information of Fourier Transform. The spectral centroid of individual frame is defined as the average frequency weighted by amplitudes, divided by sum of the amplitudes, or:

$$Spectral\ Centroid = \frac{\sum_{k=1}^{N} kF[k]}{\sum_{k=1}^{N} F[k]} \qquad (2)$$

Here,$F[k]$ is the amplitude corresponding to bin $k$ in DFT spectrum [13].

*B. Spectral Flatness*

Spectral flatness is defined as; it is measurement of the noisiness (sinusoidality) of a spectrum. For noisy signal its value is close to 1and for tonal signal it is close to 0[11].

The normalized log spectrum of time sequence is given by

$$V = V(\theta) = \log\{|E[\exp(j\theta)]|^2 / r_e(0) \qquad (3)$$

Where $r_e(0)$ denotes energy of time sequence, given by

$$r_e(0) = \sum_{n=-\infty}^{\infty} e_n{}^2 = \int_{-\pi}^{\pi} |E[\exp(j\theta)]^2 \frac{d\theta}{2\pi} \qquad (4)$$

*C. Spectral Spread*

It is the technique of spectral feature extraction in which transmitted signal is spread over a wide frequency band, which is much wider than that of the minimum bandwidth required to transmit the information to be sent. This is completed by taking a carrier of certain occupied bandwidth and power and 'Spreading' the same power over a wider occupied bandwidth [14].

*D. Spectral Projection*

The spectral projection is defined as the projection of spectrum on to a low dimensional subspace via reduced-rank spectral basis function; it is called as Spectral Projection. It is basically used to train Hidden Markov Models in order to apply uniformly to diverse source classification task with higher accuracy [15].

## V. SEGMENTATION

After feature extraction we have to apply each and every recorded audio clip to segmentation module. The main aim of segmentation is to identify boundaries between words or phonemes. Segmentation also helps for the purpose of speaker identification, speaker tracking etc. Segmentation is also divide input audio signal in to voiced and unvoiced based on the nature of audio clip and extracted features. Voiced sound is nothing but pure human voice whereas unvoiced sound is nothing but musical notes and pauses in the audio clip [12]. Here segmentation is used to divide voiced sounds from unvoiced. Only voiced sounds are considered for classification purpose unvoiced sounds are block.

## VI. TRAINING

It is the fourth stage in this technique. In this stage we have to train classifier models based on extracted features. For training purpose, we

created database of audio clips in 10 different emotions. We recorded total 200 audio clips, 20 clips per emotion. We selected random speakers and asked those speakers to generate audio clips in their own voice. We recorded audio clips in three different languages such as Marathi, Hindi and English. Then we extracted four acoustic features of those 200 audio clips and stored in database to train classifier models.

## VII. EMOTION VERIFICATION

In the stage of feature extraction we extracted 4 types of acoustic features like spectral centroid, flatness, spread and projection. Before that we have generated database of various speech samples that we are using in our daily life communication. Once database is created, we are selecting audio clip that we have to test and then we are calculating above mentioned 4 spectral features for each sound clip individually. At the same time we are training our classifiers that we have implemented in this technique. Here, we are using Support Vector Machines (SVM) and Hidden Markov Models (HMM). Then these classifiers are showing there result and classifying emotions according to their class like happy, sad, surprise etc.

## VIII. RESULT

For this experiment purpose we created a database of 10 different emotion classes. In this database we recorded audio clips that we are using in our day to day communication purpose. We are generated database by using various speakers and asked those speakers to generate speech utterances in their own language in different emotions. Here we are showing histogram image of emotion recognition technique with four spectral features and input audio clip which we have tested. The classification result for happy emotion and for the speech utterance "Yes, I liked it" is shown in below figure 3.

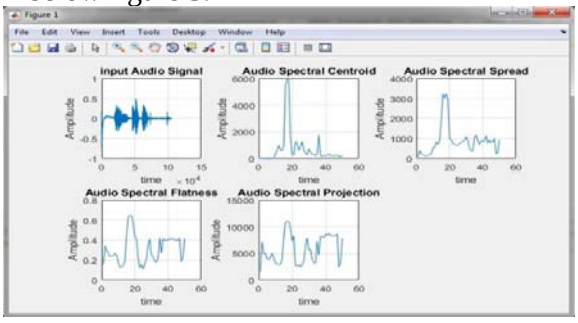


Fig.3 Histogram of Test Speech Utterance

figure 4 indicates that the input test utterance is belonging to happiness class.

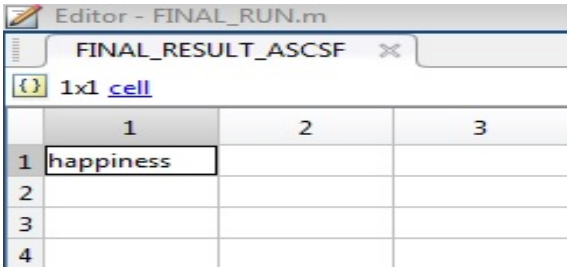


Fig.4 Classification Result

## IX. CONCLUSION

In this experiment we used Support Vector Machines (SVM) and Hidden Markov Model (HMM) classifiers to recognize human emotions. Database of ten different emotions has been created for this experiment purpose. From this recorded audio clips we calculated four spectral features like centroid, spread, flatness and projection. After extraction of audio features we are detecting the class of emotion by using SVM classifier. This technique is proven to be very effective for the calculation of human emotions with good recognition rate. This technique is useful for psychiatrics to determine the mental status of the user as well as it is also effective for E-learning to identify the mood of listener. We have implemented this technique using MATLAB software.

## REFERENCES

[1] A. Madan, M. Cebrian, S. Moturu, K. Farrahi, and A. S. Pentland, "Sensing the "Health State" of a community," IEEE Pervasive Comput., vol. 11, no. 4, pp. 36–45, Oct.–Dec. 2012.

[2] A. Tawari and M. M. Trivedi, "Speech emotion analysis: Exploring the role of context," IEEE Trans. Multimedia, vol. 12, no. 6, pp. 502–509, Oct. 2010.

[3] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 5, pp. 1057–1070, Jul. 2011.

[4] J. M. K. Kua, E. Ambikairajah, J. Epps, and R. Togneri, "Speaker verification using sparse representation classification," in Proc. IEEE Int.

Conf. Acoust., Speech, Signal Process., Prague, Czech Republic, May 22–27, 2011, pp. 4548–4551.

[5] D. Wu, T. D. Parsons, E. Mower, and S. Narayanan, "Speech emotion estimation in 3D space," in Proc. IEEE Int. Conf. Multimedia Expo, Singapore, Jul. 19–23, 2010, pp. 737–742.

[6] I. Luengo, E. Navas, and I. Hernáez, "Feature analysis and evaluation for automatic emotion identification in speech," IEEE Trans. Multimedia, vol. 12, no. 6, pp. 490–501, Oct. 2010.

[7] C. M. Lee and S. S. Narayanan,"Toward detecting emotions in spoken dialogs," IEEE Trans. Speech Audio Process., vol. 13, no. 2, pp. 293–303, Mar. 2005.

[8] T. L. Nwea, S. W. Foob, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.

[9] T. Taleb, D. Bottazzi, and N. Nasser, "A novel middleware solution to improve ubiquitous healthcare systems aided by affective information," IEEE Trans. Inf. Technol. Biomed., vol. 14, no. 2, pp. 335–349, Mar. 2010.

[10] "Special Area Exam Part II", April 28, 2001 Unjung Nam

[11] "A Spectral-Flatness Measure for Studying the Autocorrelation Method of Linear Prediction of Speech Analysis" Augustine H. Gray and John D. Markel

[12] "Audio Segmentation for Speech Recognition using Segment Features" David Rybach, Christian Gollan, Ralf Schl¨uter, Hermann Ney.

[13] "Speech Emotion Verification Using Emotion Variance Modeling and Discriminant Scale-Frequency Maps" Jia-Ching Wang, *Senior Member, IEEE*, Yu-Hao Chin, Bo-Wei Chen, *Member, IEEE*, Chang-Hong Lin, and Chung-Hsien Wu, *Senior Member, IEEE* IEEE/ACM Transactio ns On Audio, Speech, And Language Processing, Vol. 23, No. 10, October 2015

[14] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002

[15] "Audio Spectrum Projection Based on Several Basis Decomposition Algorithms Applied To General Sound Recognition and Audio Segmentation" Hyoung-Gook Kim, and Thomas.