International Journal of Recent Trends in Engineering & Research
ISSN (ONLINE) : 2455 – 1457
IMPACT FACTOR : 4.101

Speech Based Emotion Recognition

Ms. Preeti Chawaj[1], Prof. S. R. Khot[2]

[1]*Research Student, Department of Electronics and Telecommunication Engineering, D. Y. Patil College of engineering and Technology, Kolhapur, Maharashtra, India*
[2]*Associate Professor Department of Electronics and Telecommunication Engineering, D. Y. Patil College of engineering and Technology, Kolhapur, Maharashtra, India*

Abstract— This paper presents a method to identify the emotion of an audio segment with an intention to recognize human emotional/mental status. Four features namely energy, pitch, Formants, Mel frequency cepstral coefficients (MFCC) and their derivatives are used to recognize emotions such as fear, anger, happiness and sadness. PCA is used to reduce the feature dimensionality. Support vector machine is implemented to perform the emotional state classification. The overall recognition rate obtained is 84.99% using samples of Berlin emotional speech database.

Keywords—MFCC, Formants, Pitch, Energy, PCA, Support Vector Machine (SVM)

I. INTRODUCTION

Speech emotion recognition aims to automatically identify the current emotional state of a human being from his or her voice. Emotions plays an important role in human life and it is an important medium of expressing humans perspective or feelings and his or hers mental state to others. There are basically seven universal emotions which include anger, fear, happy, sad, surprise, bore-dome, neutral [1].

Speech emotion recognition is a latest topic of research since it has a wide range of applications. As we know the emotion recognition from the speech signal is an easy task for human beings because they have the natural ability to analyze the speech information. But for the machine it is difficult to analyze the emotion using speech information. The most important application of the speech emotion recognition system is to make the human machine interface more efficient. So that machine can detect what is said and who is said and also detect how it is said using speaker identification and speech emotion recognition techniques [2].

Other applications of the speech emotion recognition system are in intelligent toys, a call center it is used to timely detect customer dissatisfaction, lie detection. In gaming field, identifying player' emotion timely that is if the player is so interested in playing the game then system incorporates the high level and if the player is in idle state then the gaming system adjust itself with easy level. Another example includes, when there is an online group discussion, if students are interested in that topic, they will feel active, and shows their positive emotion. On the opposite, if they are not interested in it, they will show the contrary emotion. If we detect the correct emotion timely, and according to that give feedback to the teacher, it will help the teacher to adjust the teaching plan and boost the learning efficiency [3].

In the area of emotion recognition through speech signal, many more systems are adopted for recognizing emotional state of human being from audio or speech signal. To recognize these emotions, different intelligent systems have been developed by researchers and these different systems also explained with different features extracted and classifiers used for classification. Some of them are mentioned below:

Chaudhari [3]presented a system to recognize the human emotion like anger, happy, sad, surprise, neutral through speech signal using Hidden Markov Model and Support Vector Machine. Classification of the emotions has been done based Berlin emotional speech samples and the features extracted from these speech samples such as the prosodic and spectral features and the accuracy rate for HMM & SVM were 67.66% & 69.66% respectively.

Chandra Prakash [4]proposed the system to recognize emotions through speech using spectral features such as Mel-frequency cepstral coefficient and prosodic features like pitch, energy & classifier here used are K- Nearest Neighbor classifiers, Support Vector Machine Classifier and Gaussian mixture model classifier which detects six basic emotional states of speaker's such as anger, fear, disgust and neutral, , happiness, sadness by Berlin emotional speech database. Accuracy rate for KNN, GMM & SVM were 78.6%, 79%, 75% respectively. He explained about the importance of the classifier which leads to get optimum classification in this new feature space.

Nermine Ahmed Hendy [5] presented a system to recognize seven emotions like anger, bore, disgust, fear, happy, sad, neutral using Berlin emotional speech database. Features extracted here are statistics of pitch, formants, and energy contours, as well as spectral, perceptual and temporal features, jitter, and shimmer and the Artificial Neural Networks (ANN) was chosen as the classifier. Accuracy rate of 85% was achieved.

Dane [6] proposed a system using Gaussian mixture model and Hidden Markov model classifiers to identify five basic emotional states of speaker's such as anger, happiness, sad, surprise and neutral. Various extracted features include pitch, energy and Mel frequency cepstral coefficient. Based on these features, emotional classification and performance of classification are discussed. Gaussian mixture model was chosen which gets an accuracy of about 81%.

Anuja Bombatkar [7]focused on the speech emotion recognition system to recognize speaker's utterances into four emotional states such as anger, sadness, neutral, and happiness using K Nearest Neighbor (KNN) classifier. And the speech samples are chosen from Berlin emotional database which got the accuracy about 78% and Hindi speech database which got accuracy rate were 86.02% and features extracted from these utterances are energy, pitch, ZCC, entropy, Mel-Frequency cepstral coefficients (MFCC).

Recognition of state of emotions in speech using hopfield neural network (HFNN) technique is presented in [8]. Speech is recorded by 30 persons with five sentences. The ANN was modeled to classify two emotion states that are happiness and anger. Recognition rata of 60% were achieved.

The proposed work recognizes emotion of the speech signal by extracting the features such as energy, formants, pitch and Mel frequency cepstral coefficient (MFCC) along with their mean, maximum, minimum, standard deviation, range are also considered. To optimize these large features

principle component analysis(PCA) is used. Lastly support vector machine (SVM) classifier is employed to classify emotions (like happy, sad, surprise, angry) by comparing extracted features with the data from the training set.

II. FEATURE EXTRACTION

The feature extraction block is the most important part of the emotion recognition. An extraction stage involves the elaboration of the speech signal to obtain a certain number of variables, called features, useful for speech emotion recognition. These variables include energy, Formant frequency, pitch, Mel-frequency Cepstral Coefficients (MFCCs) and these are described one by one below.

A. Energy

One of the most important speech features which indicates emotion is energy. In order to obtain the statistics of energy features, we divide the speech signals into number of frames and for each frame, we extracts the value of energy by using the following equation.

$$E = \sum_{n=0}^{N-1} |x(n)|^2 \qquad (1)$$

Then we can calculate the statistics of energy for the whole speech sample by calculating the energy, such as mean value, max, min, standard deviation and range value of energy.
Corresponding result regarding energy with its statistics are listed in Table I:

Emotion	Samples	Mean	Max	Min	Std	Range
Angry	1	8.082691	61.72917	0.000176	11.89958	61.72899
	2	8.53472	65.65851	0.000109	14.58227	65.6584
	3	8.176914	59.91582	0.000757	10.6733	59.91506
	4	6.362147	69.70584	0.000186	11.60016	69.70565
	5	6.772347	54.54668	6.27E-05	11.80037	54.54661
Fear	1	9.265507	50.63826	0.00091	10.23664	50.63735
	2	9.973193	50.37497	0.221217	9.194314	50.15375
	3	9.909406	66.10916	0.001232	12.8845	66.10793
	4	11.06268	63.74874	0.041028	12.5295	63.70771
	5	11.91044	64.33928	0.0242	12.05667	64.315
Happy	1	3.638288	29.52386	5.04e-05	5.769346	29.52381

	2	3.286238	26.08828	2.44E-05	5.027867	26.08825
	3	3.753813	26.55184	0.000241	5.774164	26.5516
	4	3.014303	36.77465	0.000195	6.013607	36.77446
	5	4.017206	41.88668	0.000116	6.387368	41.88657
Sad	1	5.353829	31.25842	0.002868	6.799323	31.25555
	2	6.802053	33.18874	0.00907	6.552821	33.17967
	3	5.60396	34.27837	0.01513	6.233884	34.26324
	4	5.838151	38.58614	0.012946	7.558473	38.5732
	5	6.963093	40.23558	0.002383	8.349742	40.2332

Table. I Energy of different emotions with 5 samples

B. Formant Frequency

Formants are nothing but the spectral peaks of the sound spectrum of the speech signal. [12].

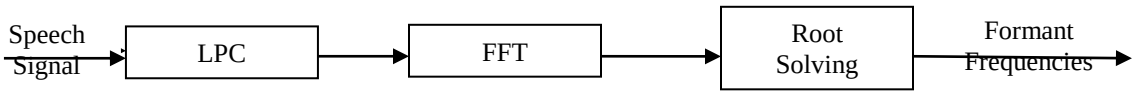

Fig.1 Block diagram of formant frequency detection using LPC

Formant Frequencies are extracted using LPC based Formants estimation method. The vocal tract is modeled as a linear filter with resonances. Graphically, the peaks of the vocal tract response of speech signal corresponding to its formant frequencies. If the vocal tract is modeled as a time-invariant, all-pole linear system, then each of the conjugate pair of poles that corresponds to a formant frequency or resonance frequency.

The LPC is based on a prediction of the current sample as a linear combination of past samples. Therefore the current sample is;

$$S(n) = \sum_{K=0}^{P-1} a_k \, s(n-k) + e(n) \qquad (2)$$

Where, P is order of the filter and a_k represents the linear prediction coefficients.

e(n) is the error in the model.

Z-transform of the Equation (1) is written as a linear filtering operation,

$$E(Z) = A_K(z).S(z) (3)$$

Where, E(Z) and S(Z) represents respectively, the Z-transform of the error signal and the speech signal.

Now if LPC predicted coefficients are known, and the residual error is available, then the speech signal can be reconstructed using synthesis filter.

$$H(Z) = \frac{S(Z)}{E(Z)} \qquad (4)$$

Where, H(Z) is the vocal tract filter.

The denominator of the transfer function is written as,

$$H(Z) = 1 + \sum_{k=0}^{P-1} a_k z^{-k} \quad = \prod_{k=0}^{P-1} (1 - c_k z^{-1}) \qquad (5)$$

Where, Ck are a set of complex numbers. Each complex conjugate pair of poles representing a resonance frequency

$$F_k = \left(\frac{Fs}{2 * pi}\right) \tan^{-1}\left(\frac{IM(c_k)}{RE(C_K)}\right) \qquad (6)$$

If the pole lies close to the unit circle, then the root represents a formant frequency.

$$R_k = (IM(c_k)^2 + RE(c_k)^2))^{\frac{1}{2}} \qquad (7)$$

Average formant frequencies for 20 samples for various emotions are mentioned in Table II:

Emotions	Formant Frequencies			
	F1	F2	F3	F4
Angry	906.1879	2596.973	4422.144	4531.959
Happy	666.5107	2426.665	4559.879	4802.78
Sad	822.4219	3290.17	3530.748	5005.17
Fear	666.1608	2944.612	4615.306	4912.929

Table. II Formant Frequencies of different emotions

C. MFCC

Mel frequency cepstral coefficients are [8],[9]coefficients that represent audio, based on perception. It is derived from the Fourier transform or the discrete cosine transform of the audio clip. Fig. 2 shows the MFCC feature extraction process, which involves the following steps:

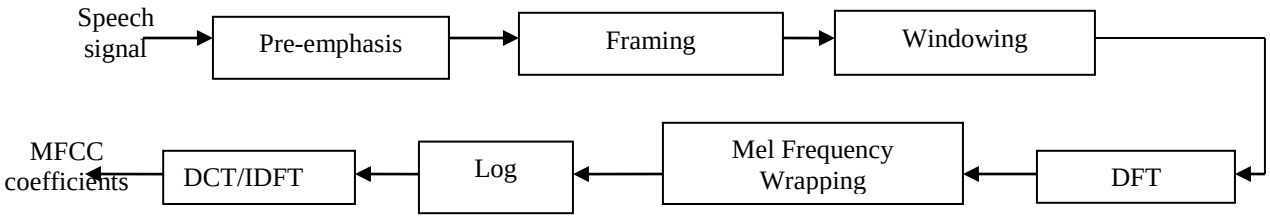

Fig. 2Computation steps of MFCC Coefficients.

1) Pre–emphasis

The pre- involves noise reduction, equalization, low-pass filtering. In speech processing voice has a low-pass behavior since pre-emphasis is applied by high-pass filtering the signal to obtain a uniform energy distribution spectrum. The first order FIR filter equation is used is

$$y[n] = x[n] - 0.95x[n] \quad (8)$$

2) Framing

An audio signal is constantly changing, so to simplify things we assume that on short time scales the audio signal doesn't change much. We frame the signal into 20-40ms frames. If the frame is much shorter we don't have enough samples to get a reliable spectral estimate, if it is longer the signal changes too much throughout the time.

3) Windowing

In order to remove edge effects, each frame is multiplied by Hamming window. Hamming window is represented by

$$w[n] = 0.54 - 0.46\cos\frac{2\pi n}{N-1} \quad (9)$$

Where $0 \leq N \leq n\text{-}1$

4) DFT

The next step is to calculate the power spectrum of each frame in order to identify which frequencies are present in the frame. FFT equation is represented as

$$x[K] = \sum_{n=0}^{N-1} x(n)e^{\frac{-j2\pi Kn}{N}} \quad (10)$$

5) Mel Filter-bank and Frequency wrapping

As the power spectrum cannot concern the difference between two closely spaced frequencies. For this reason, we take clumps of periodogram bins and sum them up to get an idea of how much energy exist in various frequency regions and this is performed by mel-filter bank. Usually in-between 20-40 triangular filters can be used but 26 filters are standard. 23 Mel triangular filters are shaped with 50% overlapping. From each filter, the spectrum is added to get one coefficient each; in this way we have considered the first 13 coefficients as our features. These frequencies are converted to Mel scale using following conversion formula. Mel-scale tells us exactly how to space our filter-bank and how wide to make them.

$$\text{mel(f)} = 2595 * \log10\left(1 + \frac{\text{f}}{700}\right) \qquad (11)$$

6) Log

Once we have the filter bank energies, we take the logarithm of them. The logarithm function converts multiplication into addition. This is motivated by human hearing. As we know that our ear can't hear loudness on linear scale. In order to double the perceived volume of a sound we have to add 8 times energy into it.

7) Discrete Cosine Transform

The final step is to compute the DCT of the log filter-bank energies. Because of our filter-banks got overlapped, the filter-bank energies got correlated with each other. DCT decorrelates those energies. Fig. 3 represents the 13 coefficients obtained for different emotions

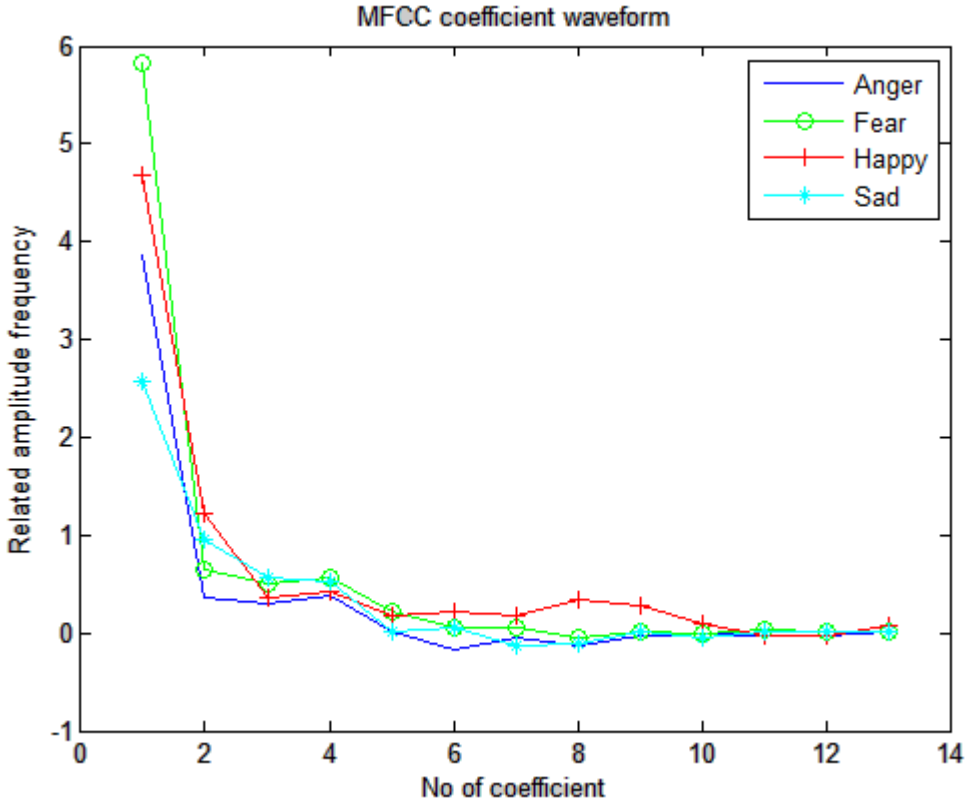

Fig. 3 Waveform of 13 MFCC coefficient for various emotions

D.Pitch

Speech signal exhibits a relative periodicity and its fundamental frequency, called pitch [10]. Many pitch estimation algorithms involves both time- domain and frequency-domain analysis. Pitch estimation using autocorrelation has been chosen because of its good applicability to audio signal and easy for implementation. The objective of this experiment is to estimate the pitch periods of a given speech signals by auto correlation method. The first step is to divide the given speech signal into 20-

40ms blocks of speech frames. The auto correlation sequence of each frame is then found by using the following equation.

$$\hat{R}(\tau) = \sum_{n=0}^{N-1-\tau} S(n)s(n+\tau)\tau \in (0,1 \dots \dots \dots N-1) \qquad (12)$$

Where, $R(\tau)$ in (1) is the autocorrelation of lag τ.

The autocorrelation function shows how much the signal correlates with itself, at different delays τ considering that, given a "sufficiently periodical" speech recording, its autocorrelation will present the highest value at delays corresponding to multiples of pitch periods:

$\tau_{\text{pitch}} = \arg\max \hat{R}(\tau)$(13)

The frequency of the pitch is computed as,

$$\rho_{\text{pitch}} = \frac{\text{Fs}}{\tau_{\text{pitch}}} (14)$$

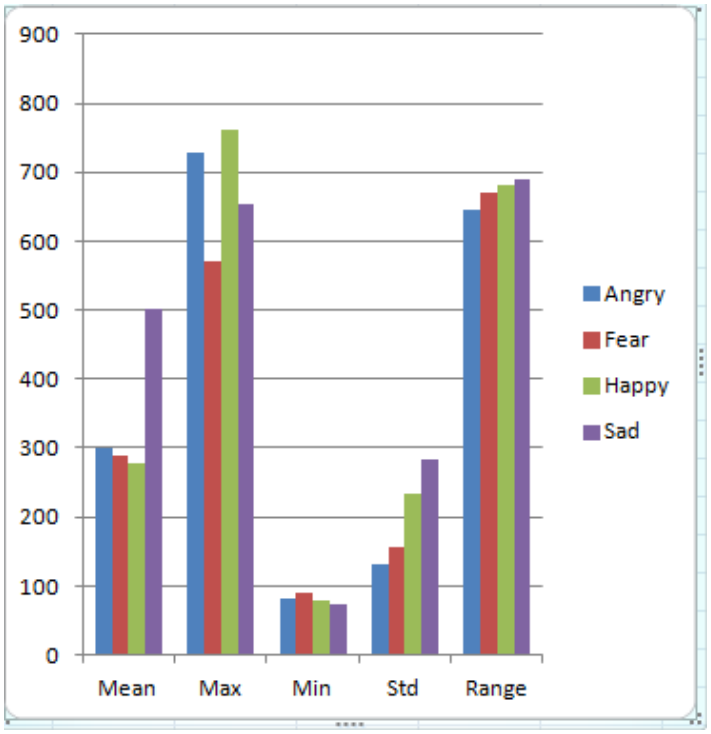

Table III. Pitch frequency for various emotions

III. PRINCIPLE COMPONENT ANALYSIS of FEATURE EXTRACTION

Principle component analysis (PCA) rotates the original data to a new coordinates, making the data as "flat" as possible. From given a set of two or more variables, PCA generates a new set of data with same number of variables, called principal components. Each principle component is a linear transformation of the entire original data set. The coefficients of the principle components are calculated so that the first principle component has maximum information with maximum variance. The second principle component is calculated to have the second most variance and which is uncorrelated with the first principle component.

Steps involved in the PCA computing is described as given below

Step1: get the feature vector
Step2: subtracts the mean

For PCA work properly, we need to subtract the mean from each of the data dimension

Mean = x1=$\sum_{i=1}^{n} xi$

Step3: Calculate covariance matrix
Step4: Calculate the eigenvectors and eigenvalues of the covariance matrix

Finding the eigenvectors and eigenvalues of the covariance matrix is equivalent of fitting that straight, principle component of the data set.

Step5: choosing components and forming a new vector

Reduced dimensionality comes into it. In general, once eigenvector is obtained from the covariance matrix, the next step is to order them by eigenvalues highest to lowest. We have two choices here. We can either form a feature vector with all of the eigenvectors or we can choose to leave out the smaller, less significant component and this results the component in order of significance and ignores the components of lesser significant.

Feature vector= (eig1, eig2,eig3……..eign).

Step5: deriving the new data set

Final data=Feature vector X mean adjusted data.

In our work, the features are selected as 20 X 40 out of 20 X 45.

IV. CLASSIFIER

The last step is classification. It assigns a label representing the recognized emotion by feature selection block and database. The main task of this stage is to choose an efficient method to provide accurate predicted results for emotion recognition. Each classifier has an initial phase in which it is trained to perform a correct classification and a following phase in which the classifier is tested.

The proposed work will emphasize on choosing the multiclass Support Vector Machine as a classifier because it is a simplest and widely used classifier. SVM is a supervised learning model it means we need dataset which has been labeled. The goal of a support vector machine is to find the optimal separating hyperplane. This hyperplane separates all data points of one class from those of the other class. The support vectors are the data points that are closest to the separating hyperplane. These points are on the boundary of the slab. The following figure illustrates these definitions, with + indicating data points of type 1 and – indicating data points of type –1.

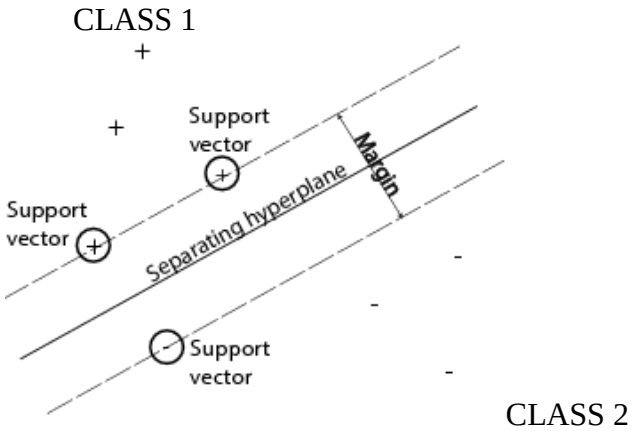

Fig .4 Describes the two class svm

The data for training is a set of points (vectors) xi along with their categories yi. For some dimension d, the xi ∈ Rd, and the yi = ±1. The equation of a hyperplane is

$$<w,x> + b = 0$$

Where, w is orientation of hyperplane and b is position of hyperplane.
<w,x> is the inner (dot) product of w and x,

The following problem defines the best separating hyperplane. Find w and b that minimize ||w|| such that for all data points (xi,yi),

$$yi(<w,xi> + b) \geq 1$$

The support vectors are the xi on the boundary, those for which yi(<w,xi> + b) = 1.
For mathematical convenience, the problem is usually given as the equivalent problem of minimizing <w,w>/2. This is a quadratic programming problem. The optimal solution w, b enables classification of a vector z as follows:

$$class(z) = sign(<w,z> + b)$$

But my aim is to classify 4 classes nothing but 4 emotions. For that we have to use one versus all multiclass SVM. It is an extension of two class classifier. This involves training of one classifier per class from total four classifier. For example in order to train class 1, it will assume 1st label as a +ve and the rest are –ve. Similarly for class 2 it will assume 2nd label as +ve and rest classes as –ve.

V. EXPERIMENTAL RESULTS

Berlin Emotional Speech Database (audio signals) comprises about 800 sentences (seven emotions x ten actors x ten sentences + some second versions). This available database consists of ten professional speakers (5 female and 5 male), each of them produce 10 German sentences (5 short and 5 longer) which could be used in day to day communication and are interpretable in all applied emotions. From these samples, most reliable samples were chosen for training and some samples are used for testing purpose. By choosing most accurate 20 samples, training can be done with the help of support vector machine. After completion of the training, 15 test samples are selected for testing purpose through which emotions has to be extracted. The Support vector machines again calculate the values of the features of the testing sample. Then on the basis of features extracted from the testing audio sample are then compared with the trained samples. Confusion matrix for different emotions is mentioned in Table IV:

Emotions	Recognized Emotions (%)			
	Angry	Happy	Sad	Fear
Angry	93.33	0	0	6.67
Happy	13.33	80	6.67	0
Sad	6.67	0	86.66	6.67
Fear	13.33	0	6.67	80

Table. IV. Confusion matrix for different emotion

VI. CONCLUSION

The propose system is able to recognize the human emotions that are anger, fear, sad and happy respectively using speech signal. The feature extraction is implemented by MFCC, pitch,

formant frequency, energy method. Later support vector machine (SVM) classifier is used to obtain the result. The experimental results detected for happy emotion is 80.025%, for fear it is 80%, for sad it is 86.66% and for fear it is 80%. Performance analysis shows the overall accuracy rate obtained is 84.99%. The rate obtained is comparatively higher as compare to other approaches. Lastly we conclude that the accurate recognition rate is based upon the right choice of extracted features, database and type of classifier.

REFERENCES

I.

II. Dipti. D. Joshi, M.B. "Recognition Of Emotion From Marathi Speech using MFCC And Dwt Algorithms", *ISSN*, Volume-2, Issue – 2, pp. 2278-5140, 2013.

III. Ashish B. Ingale, Dr. D. S. Chaudhari, "Speech emotion recognition using hidden Markov model and support vector machine", *IJAER, Vol. 1, Issue 3*, pp.316-318, April-June, 2012.

IV. Yixiong Pan, Peipei Shen and Liping Shen, "Speech Emotion Recognition using Support Vector Machine", *International Journal of Smart Home, Vol. 6, No. 2*, April 2012.

V. Chandra Praksah , Prof. V. B. Gaikwad, "Analysis Of Emotion Recognition System through Speech Signalusing KNN, GMM & SVM Classifier", *IJECS, Volume 4 Issue 6*,pp.12523-12528, June 2015.

VI. Nermine Ahmed Hendy and Hania Farag, "Emotion Recognition Using Neural Network: A Comparative Study",*ISSRI, Vol:7* , pp. 1149-1155, 2013.

VII. Akshay S. Dane Dr. S.L.Nalbalwar, "Emotion Recognition Through Speech Using Gaussian Mixture Model And Hidden Markov Model", *ISSN, Vol 3, Issue 4*, pp.742-746, April 2013.

VIII. Anuja Bombatkar, Gayatri Bhoyar, Khushbu Morjani, Shalaka Gautam, Vikas Gupta, "Emotion recognition using Speech Processing Using k-nearest neighbor algorithm", *IJERA, pp.68-71,* April 2014.

IX. Gyanendra Pratap & Rakesh Kumar Sharma, "Recognizing of emotions from speech by ANN", IJECE, vol.1, Issue I, Aug 2012.

X. N. Murali Krishna, P. V. Lakshmi, Y. Srinivas, J. Sirisha Devi, "Emotion Recognition using Dynamic Time Warping Technique for Isolated Words", *IJCSI* , Vol. 8, Issue 5, September 2011.

XI. Igor Bisio, Alessandro Delfino, Fabio lavagetto, Mario Marchese, and Andrea Sciarrone, "Gender Driven Emotion Recognition through Speech Signals for Ambient Intelligence Applications" , *IEEE* , *Vol. 1,* 21 January 2014.

XII. S. Demircan and H. Kahramanlı, "Feature Extraction from Speech Data for Emotion Recognition", *Journal of Advances in Computer Networks*, Vol. 2, No. 1, March 2014.

XIII. A. Khulage and Prof. B. V. Pathak, "Analysis of speech under stress using linear techniques and non-linear techniques for emotion recognition system"

