Speech Emotion Recognition System Using SVM AND LIBSVM

Sujata B. Wankhade
M.E(Final Year I.T)
Sipna C.O.E.T(Amravati), India
wankhadesujata@gmail.com

Pritish Tijare
Lecturer, Sipna C.O.E.T(Amravati), India
pritishtijare@rediffmail.com

Yashpalsing Chavhan
Lecturer, Anuradha Engineering College, Chikhli, India
yashpal.chavhan@yahoo.in

ABSTRACT
This paper introduces a approach to emotion recognition from speech signal using SVM as a classifier. Speech Emotion Recognition (SER) is a current research area in the field of Human Computer Interaction (HCI) with wide range of applications. The speech features such as, Mel Frequency cepstrum coefficients (MFCC) and Mel Energy Spectrum Dynamic Coefficients (MEDC) are extracted from speech utterance. The Support Vector Machine(SVM) is used as classifier to classify different emotional states such as anger, happiness, sadness, neutral, fear. The Berlin emotion database and Hindi emotion database are used for extracting the features from emotional speech .wav files. The recognition rates by implemented SVM are 62% and 71.66% for Berlin database and Hindi database respectively. The recognition rates by LIBSVM using RBF for Berlin database are 99.39% for cost value c=8. The recognition rates by LIBSVM using RBF kernel function for Hindi database are 78.33%.The accuracy rates by LIBSVM using Linear RBF kernel function for German independent files is 68.902%.

Categories and Subject Descriptors
I.5.0 [**Pattern Recognition**]: General.

General Terms
Performance, Experimentation, Human Factors.

Keywords
SVM, Lib-SVM, MFCC and MEDC Speech emotion, Emotion Recognition.

1. INTRODUCTION
Now a day vast researches done to increase Human computer Interaction (HCI) . Speech Emotion Recognition is a very recent research topic which helps to increase the Human Computer Interaction (HCI) field. As computers have become an integral part of our lives, the need has risen for a more natural communication interface between humans and computers. In today's HCI systems, machines can know who is speaking and what he or she is speaking of by the speaker and speech recognition system. When the machine are equipped with emotion recognition techniques, they can know how he or she is speaking, and they can react more appropriately and more naturally. To achieve this goal, a computer would have to be able to perceive its present situation and respond differently depending on that perception. Part of this process involves understanding a user's emotional state. To make the human-computer interaction more natural, it would be beneficial to give computers the ability to recognize emotional situations the same way as human does.

There are two ways for Automatic Emotion Recognition (AER) , i.e by speech signals or by facial expressions. In the field of HCI, speech is primary to the objectives of an emotion recognition system, as are facial expressions and gestures. Speech is considered as a powerful mode to communicate with intentions and emotions.In the recent years, a great deal of research has been done to recognize human emotion using speech information [1], [2]. Many researcher explored several classification methods including the Neural Network (NN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Maximum Likelihood Bayes classifier (MLC), Kernel Regression and K-nearest Neighbors (KNN), Support Vector Machine (SVM) [3], [4].

A support vector machine is a supervised learning algorithm developed over the past decade by Vapnik in 1995.The algorithm addresses the general problem of learning to discriminate between positive and negative members of a given class of n-dimensional vectors[12].The Support Vector Machine is used as a classifier for emotion recognition. The SVM is used for classification and regression purpose. It performs classification by constructing an N-dimensional hyperplanes that optimally separates the data into categories. The classification is achieved by a linear or nonlinear separating surface in the input feature space of the dataset. Its main idea is to transform the original input set to a high-dimensional feature space by using a kernel function, and then achieve optimum classification in this new feature space.

A Berlin Emotional database [5] and Hindi emotional speech signal files is used for feature extraction and training SVM. The Berlin database of emotional speech was recorded at the Technical University, Berlin. The database German contains speech with acted emotions in language. It contains 493 utterances of 10 professional actors five males and five females who spoke 10 sentences with emotionally neutral content in 7 different emotions. The emotions were wut (anger), langeweile (boredom), ekel (disgust), angust (fear), freude (happiness), trauer (sadness) and neutral emotional state.

There are various applications of Speech Emotion Recognition like Emotion Recognition software for call center it is a full-fledge prototype of an industrial solution for computerized call center and can help in detection of the emotional state in telephone call center conversations to provide feedback to an operator or a supervisor , psychiatric diagnosis, intelligent toys,

lie detection, learning environment, educational software, for monitoring purposes.

2. LITERATURE REVIEW / RELATED WORKS:

Under this point focus is on the literature available for speech emotions and recognition techniques.

2.1 Speech

Speech is the primary means of communication between human. Speech refers to the processes associated with the production and perception of sounds used in spoken language. A number of academic disciplines study speech and speech sounds, including acoustics, psychology, speech pathology, linguistics, cognitive science, communication studies, otolaryngology and computer science.

2.1.1 Speech production

Human speech is produced by vocal organs. Main energy source is the lungs with the diaphragm. When speaking, the air flow is forced through the glottis between the vocal cords and the larynx to the three main cavities of the vocal tract, the pharynx and the oral and nasal cavities. From the oral and nasal cavities the air flow exits through the mouth and nose, respectively. The V-shaped opening between the vocal cords, called the glottis, is the most important sound source in the vocal system. The vocal cords may act in several different ways during speech. The most important function is to modulate the air flow by rapidly opening and closing, causing buzzing sound from which vowels and voiced consonants are produced. The fundamental frequency of vibration depends on the mass and tension and is about 110 Hz, 200 Hz, and 300 Hz with men, women, and children, respectively. With stop consonants the vocal cords may act suddenly from a completely closed position in which they cut the air flow completely, to totally open position producing a light cough or a glottal stop. On the other hand, with unvoiced consonants, such as /s/ or /f/, they may be completely open. An intermediate position may also occur with for example phonemes like /h/.

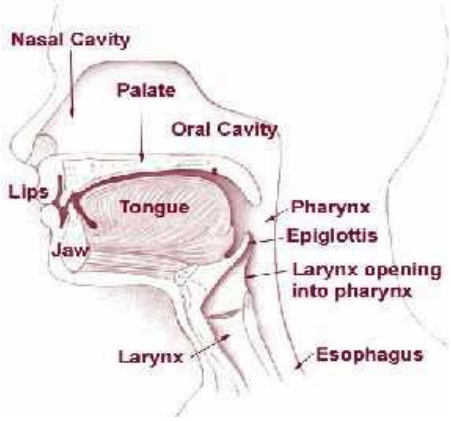


Figure 1: Speech production system [19]

The pharynx connects the larynx to the oral cavity. It has almost fixed dimensions, but its length may be changed slightly by raising or lowering the larynx at one end and the soft palate at the other end. The soft palate also isolates or connects the route from the nasal cavity to the pharynx. At the bottom of the pharynx are the epiglottis and false vocal cords to prevent food reaching the larynx and to isolate the esophagus acoustically from the vocal tract. The epiglottis, the false vocal cords and the vocal cords are closed during swallowing and open during normal breathing. The oral cavity is one of the most important parts of the vocal tract. Its size, shape and acoustics can be varied by the movements of the palate, the tongue, the lips, the cheeks and the teeth. Especially the tongue is very flexible, the tip and the edges can be moved independently and the entire tongue can move forward, backward, up and down. The lips control the size and shape of the mouth opening through which speech sound is radiated. Unlike the oral cavity, the nasal cavity has fixed dimensions and shape. Its length is about 12 cm and volume 60 cm3. The air stream to the nasal cavity is controlled by the soft palate.

2.1.2 Speech Perception

Speech perception refers to the processes by which humans are able to interpret and understand the sounds used in language. The study of speech perception is closely linked to the fields of phonetics and phonology in linguistics and cognitive psychology and perception in psychology. Research in speech perception seeks to understand how human listeners recognize speech sounds and use this information to understand spoken language. Speech research has applications in building computer systems that can recognize speech, as well as improving speech recognition for hearing- and language-impaired listeners.

The Fig. 3 shows how perception of consonant ‘d’ differs in syllables with different vowels. The process of perceiving speech begins at the level of the sound signal and the process of audition. After processing the initial auditory signal, speech sounds are further processed to extract acoustic cues and phonetic information. This speech information can then be used for higher-level language processes, such as word recognition.

Speech signals are usually considered as voiced or unvoiced. Voiced sounds consist of fundamental frequency (F0) and its harmonic components produced by vocal cords (vocal folds). The vocal tract modifies this excitation signal causing formant (pole) and sometimes anti formant (zero) frequencies. Each formant frequency has also an amplitude and bandwidth and it may be sometimes difficult to define some of these parameters correctly. The fundamental frequency and formant frequencies are probably

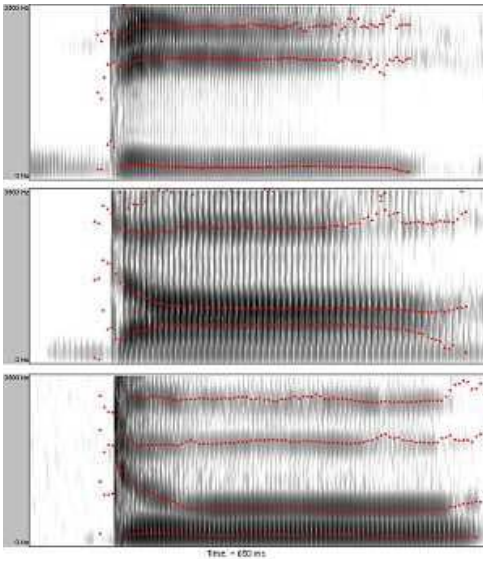


Figure 3: Speech perception of consonant "d" in different syllable (Reff:http://en.wikipedia.org/wiki/Speech perception)

the most important concepts in speech synthesis and also in speech processing. In unvoiced sounds, there is no fundamental frequency in excitation signal and therefore no harmonic structure either and the excitation can be considered as white noise. The airflow is forced through a vocal tract constriction which can occur in several places between glottis and mouth. Some sounds are produced with complete stoppage of airflow followed by a sudden release. Unvoiced sounds are also usually more silent and less steady than voiced ones. Whispering is the special case of speech. When whispering a voiced sound there is no fundamental frequency in the excitation and the first formant frequencies produced by vocal tract are perceived.

2.1.3 Speech Contents

In most languages the written text does not correspond to its pronunciation so that in order to describe correct pronunciation some kind of symbolic presentation is needed. Every language has a different phonetic alphabet and a different set of possible phonemes and their combinations. The number of phonetic symbols is between 20 and 60 in each language. A set of phonemes can be defined as the minimum number of symbols needed to describe every possible word in a language. In English there are about 40 phonemes, in German there are approximately 25 phonemes.

The information that is to be send through speech can be represented by con-catenation of elements from a finite set of symbols. The symbols from which every sound can be classified are called phonemes. Phonemes are abstract units and their pronunciation depends on contextual effects, speaker's characteristics, and motions.

During continuous speech, the articulatory movements depend on the preceding and the following phonemes. The articulators are in different position depending on the preceding one and they are preparing to the following phoneme in advance. This causes some variations on how the individual phoneme is pronounced. These variations are called allophones which are the subset of phonemes and the effect is known as coarticulation. For example, a word lice contains a light /l/ and small contains a dark /l/. These l's are the same phoneme but different allophones and have different vocal tract configurations. Another reason why the phonetic representation is not perfect is that the speech signal is always continuous and phonetic notation is always discrete. Different emotions and speaker characteristics are also impossible to describe with phonemes so the unit called phone is usually defined as an acoustic realization of a phoneme.

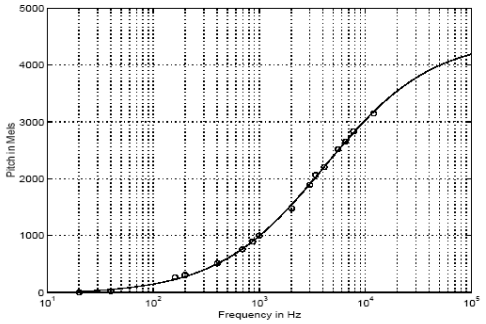
- A Syllable is a unit of organization for a sequence of speech sounds. For example, the word water is composed of two syllables: wa and ter. Syllables are often considered the phonological "building blocks" of words. Word that consists of a single syllable is called a monosyllable (like cat). Whereas a word consisting of two syllables is called a disyllable (like mon+key), word consisting of three syllables is called a trisyllable (like in+di+gent), A word consisting of morethan three syllables is called a polysyllable (like in+te+lli+gence).
- Prosody is the rhythm, stress, and intonation of speech. Prosody reflect the emotional state of a speaker; whether an utterance is a statement, a question, or a command; whether the speaker is being ironic or sarcastic; emphasis, contrast, and focus; and other elements of language which may not be encoded by grammar. Prosody includes pitch, intensity and durations. Sometimes, but not necessarily, voice quality and articularly features are also used. The following features take care of voice quality: formants means and band widths, harmonic to noise ratio, MFCC coefficients.
- Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). The perception of loudness is related to both the sound pressure level and duration of a sound.
- Pitch represents the perceived fundamental frequency of a sound. It is one of the three major auditory attributes of sounds along with loudness. The fundamental frequency (also called a natural frequency) of a periodic signal is the inverse of the pitch period length. The pitch period is in term, the smallest repeating unit of a signal. One pitch period thus describes the periodic signal completely.
- The significance of defining the pitch period as the smallest repeating unit canbe appreciated by noting that two or more concatenated pitch periods form a repeating pattern in the signal.
- A formant is a peak in the frequency spectrum of a sound caused by acoustic resonance. The Formant frequencies are distinguishing or meaningful frequency components of human speech.

2.1.4 Mel Scale

The mel scale [9] is proposed by Stevens, Volkman and Newman in 1937. It is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The mel is a unit of measure of perceived pitch or frequency of a tone. The reference point between this scale and normal frequency measurement is defined by equating a 1000 Hz tone, 40 dB above the listener's threshold, with a pitch of 1000 mels. Above about 500 Hz, larger and larger intervals are judged by listeners to produce equal pitch increments As a result, four octaves on the hertz scale above 500 Hz are judged to comprise about two octaves on the mel scale. The mel scale is approximately linear below 1 kHz and logarithmic above. The following formula is used to compute the mels for a given frequency f in Hz

$$mel = 1127 * \log(1 + f / 700)$$

.



.

Figure 3.3: Mel Scale [9]

2.2 Emotion

An Emotion [3][10] is a term for a mental and physiological state associated with a wide variety of feelings, thoughts, and behavior. Emotions are subjective experiences, or experienced from a individual point of view. Emotion is often associated with mood, temperament and personality. But in general emotions are short-term whereas moods are long-term and temperaments or personalities are very long-term. A particular mood may sustain for several days and temperament sustain for month or years. Human emotion can be of different types such as angry, happiness, sadness ,neutral, fear, disgust, surprise, shy, bored etc.

The study of Emotion is carried out from a long time. This study is helpful in the fields of Psychology, Sociology, Criminology, Physiology, etc. Researchers often use autonomic responses to measure emotion. One frequently used autonomic response is called the galvanic skin response. The galvanic skin response is an increase in the skins rate of electrical conductivity, which occurs when subjects sweat during emotional states. Researchers also use indicators such as blood pressure, muscle tension, heart rate, and respiration rate to measure emotion.

2.2.1 2.2.1 Theories of Emotion

Emotion is a complex, subjective experience accompanied by biological and behavioral changes. Emotion involves feeling, thinking, activation of the nervous system, physiological changes, and behavioral changes such as facial expressions.

Different theories exist regarding how and why people experience emotion. These include evolutionary theories, the James-Lange theory, the Cannon-Bard theory,Schacter and Singers two-factor theory, and cognitive appraisal.

3. EVOLUTIONARY THEORIES

In the 1870s, Charles Darwin proposed that emotions evolved because they had adaptive value. For example, fear evolved because it helped people to act in ways that enhanced their chances of survival. Darwin believed that facial expressions of emotion are innate (hard-wired). He pointed out that facial expressions allow people to quickly judge someone‘s hostility or friendliness and to communicate intentions to others.

Recent evolutionary theories of emotion also consider emotions to be innate responses to stimuli. Evolutionary theorists tend to downplay the influence of thought and learning on emotion, although they acknowledge that both can have an effect. Evolutionary theorists believe that all human cultures share several primary emotions, including happiness, contempt, surprise, disgust, anger, fear, and sadness. They believe that all other emotions result from blends and different intensities of these primary emotions. For example, terror is a more intense form of the primary emotion of fear.

4. THE JAMES-LANGE THEORY

In the 1880s, two theorists, psychologist William James and physiologist Carl Lange, independently proposed an idea that challenged commonsense beliefs about emotion. This idea, which came to be known as the James-Lange theory, is that people experience emotion because they perceive their bodie‘s physiological responses to external events. According to this theory, people don’t cry because they feel sad. Rather, people feel sad because they cry, and, like-wise, they feel happy because they smile. This theory suggests that different physiological states correspond to different experiences of emotion. James believed that after perceiving a stimulating event, an individual in-stantly and automatically experiences physiological changes (e.g., increased or decreased heart rate, changes in respiration rate, sweating). It is in thinking about and assessing these physiological changes that the individual assigns an emotion to them. we feel sad because we cry, angry because we strike, afraid because we tremble, and neither we cry, strike, nor tremble because we are sorry, angry, or fearful, as the case may be. lets take an example of a man being burglarized. Upon witnessing the burglar entering his home, his heart races, he breathes more rapidly, and his hands tremble and sweat. It is after cognitively assessing his physiological reactions to the situation that the man is able to assign the emotion of “frightened” to his experience. This theory suggests that different physiological states correspond to different experiences of emotion

5. SYSTEM IMPLEMENTATION

The importance of emotions in human-human interaction provides the basis for researchers in the engineering and

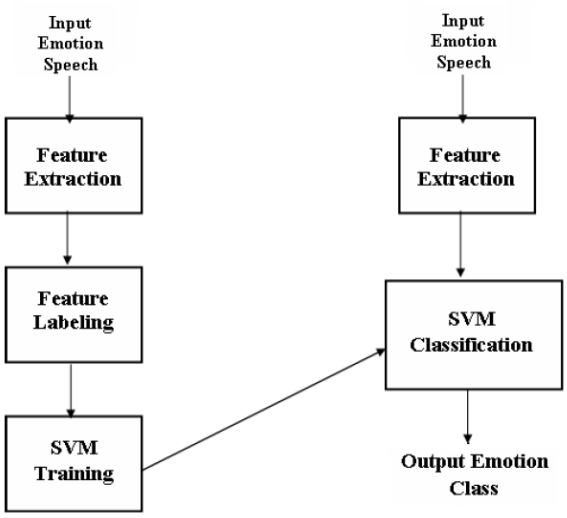


Figure 1. Speech Emotion Recognition System.

computer science communities to develop automatic ways for computers to recognize emotions. As shown in fig. 1 the input to the system is a .wav file from Berlin Emotion Database that contains emotional speech utterance from different emotional classes. After that features extraction process is carried out. In feature extraction process two features are extracted MFCC [6], [7] and MEDC [8]. After that the extracted features and their corresponding class labels are given as input to the LIBSVM classifier. The output of a classifier is a label of a particular emotion class. There are total five classes angry, sad, happy, neutral and fear. Each label represents corresponding emotion class.

5.1 Feature Extraction

In previous works several features are extracted for classifying speech affect such as energy, pitch, formants frequencies, etc. all these are prosodic features. In general prosodic features are primary indicator of speaker’s emotional state.

Here in feature extraction process two features are extracted Mel Frequency Cepstral Coefficient (MFCC) and Mel Energy

spectrum Dynamic coefficients (MEDC). Fig. 2 shows the MFCC feature extraction process. The MFCC technique makes use of two types of filters i.e linearly spaced filters and logarithmically spaced filters. O'Khalifa etal[11]had identified the main steps of MFCCfeature extraction process as shown in Fig. 2:

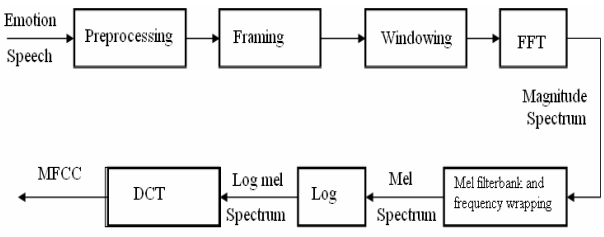


Figure 2. MFCC feature extraction

- Preprocessing: The continuous time signal (speech) is sampled at sampling frequency. At the first stage in MFCC feature extraction is to boost the amount of energy in the high frequencies. This preemphasis is done by using a filter.
- Framing: it is a process of segmenting the speech samples obtained from the analog to digital conversion (ADC), into the small frames with the time length within the range of 20-40 ms. Framing enables the non stationary speech signal to be segmented into quasi-stationary frames, and enables Fourier Transformation of the speech signal. It is because, speech signal is known to exhibit quasi-stationary behavior within the short time period of 20-40 ms.
- Windowing: Windowing step is meant to window each individual frame, in order to minimize the signal discontinuities at the beginning and the end of each frame.
- FFT: Fast Fourier Transform (FFT) algorithm is ideally used for evaluating the frequency spectrum of speech. FFT converts each frame of N samples from the time domain into the frequency domain.
- Mel Filterbank and Frequency wrapping: The mel filter bank [8] consists of overlapping triangular filters with the cutoff frequencies determined by the center frequencies of the two adjacent filters. The filters have linearly spaced centre frequencies and fixed bandwidth on the mel scale.
- Take Logarithm: The logarithm has the effect of changing multiplication into addition. Therefore, this step simply converts the multiplication of the magnitude in the Fourier transform into addition
- Take Discrete Cosine Transform: It is used to orthogonalise the filter energy vectors. Because of this orthogonalization step, the information of the filter energy vector is compacted into the first number of components and shortens the vector to number of components.

Another feature Mel Energy spectrum Dynamic coefficients (MEDC) is also extracted. It is extracted as follows:the magnitude spectrum of each speech utterance is estimated using FFT, then input to a bank of 12 filters equally spaced on the Mel frequency scale. The logarithm mean energies of the filter outputs are calculated $E_n(i)$, i= 1.....N. Then, the first and second differences of $E_n(i)$ are calculated. MEDC feature extraction process. The MEDC feature extraction process contains following steps shown in figure 3:

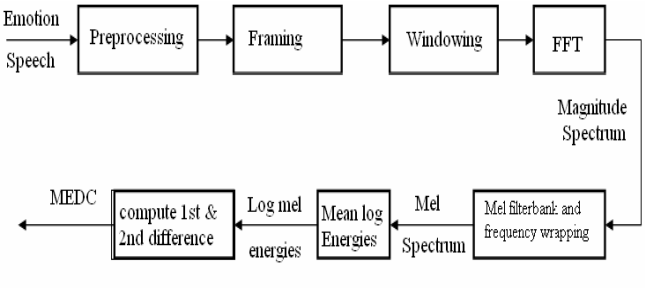


Figure 3. MEDC feature extraction

- Preprocessing, Framing, Windowing, FFT & Mel filterbank and Frequency wrapping processes of MEDC feature extraction are same as MFCC feature extraction.
- Take logarithmic mean of energies: In this process a mean log of every filter energies is calculated. This mean value represent energy of individual filter in a filterbank.
- Compute 1^{st} and 2^{nd} difference: The final Mel energy spectrum dynamics coefficients are then obtained by combining the first and second differences of filter energies.

5.2 Feature Labeling

In Feature labeling each extracted feature is stored in a database along with its class label. Though the SVM is binary classifier it can be also used for classifying multiple classes. Each feature is associated with its class label e.g. angry, happy, sad, neutral, fear.

5.3 SVM Training and Classification

The SVM is train according to labeled features. The SVM kernel functions are used in the training process of SVM. Binary classification can be viewed as the task of separating classes in feature space. SVM is a binary classifier, but it can also be used as a multiclass classifier. LIBSVM [9], [10] is a most widely used tool for SVM classification and regression developed by C. J. Lin. Radial Basis Function (RBF) kernel is used in training phase. Advantage of using RBF kernel is that it restricts training data to lie in specified boundaries. The RBF kernel nonlinearly maps samples into a higher dimensional space, so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. The RBF kernel has less numerical difficulties than polynomial kernel. The results of SVM classification are given in the form of confusion matrix tables.The confusion matrix represents the percentage of accurate classification and misclassification for the given class.

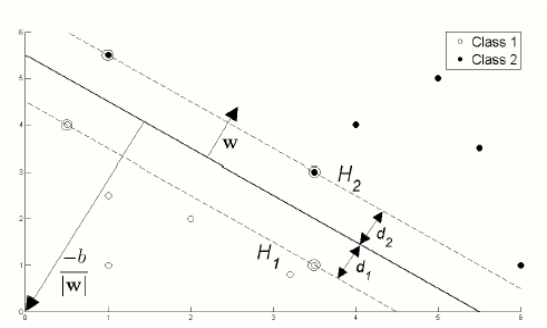


Figure4. Linear Classification using SVM

6. EXPERIMENTATION AND RESULTS

Hindi emotion database is created that contains a emotion dialogues from different bollywood movies Initially all files are cut in .mp3 format and then converted into .wav file format also Berlin Emotion database used which contains 406 speech files for five emotion classes. Emotion classes Anger, sad, happy, neutral, fear are having 127, 62, 71, 79 and 67 speech utterance respectively. The LIBSVM is trained on MFCC and MEDC feature vectors using RBF functions. The LIBSVM is used to test these feature vectors. The experimentation is carried out by varying cost values for RBF kernel.Gender independent files are taken for experimentation. The results of SVM classification are given in the form of confusion matrix tables.The confusion matrix represents the percentage of accurate classification and misclassification for the given class.

Table 1.shows confusion matrix implemented SVM for Hindi emotion utterance. The Angry emotion gives maximum 89.47% and Sad gives minimum 50% recognition rate. From table it is observed that maximum misclassification found in Sad and Fear. The overall recognition rate for Hindi emotion is 62%

Table 2.shows confusion matrix of implemented SVM for German emotion speech utterance using one-to-one multiclass method. It is observed that maximum misclassification is found in happy and fear emotions. The Angry emotion gives maximum

88.21% and Fear gives minimum 44.44% recognition rate. The overall recognition accuracy for German emotion is 62%

Table 3. shows confusion matrix for LIBSVM RBF classifier (Hindi) The LIBSVM using RBF kernel with cost value c=10 gives overall 77.66% recognition rate.

Table 4 shows Confusion matrix using LIBSVM RBF kernel gender independent.

Table4 Confusion matrix of the LINEAR LIBSVM classifier (German Gender Independent).

Table 5,6,7,8,9,10,11,12,13,14 shows confusion matrix for varying cost value from 1 to 10 for LIBSVM RBF kernel of German gender independent files.

Table 15 Accuracy Percentage recognition using variable cost value c = 1 to 10 for RBF Kernel from table 5 to 14

Table 1: Confusion matrix for SVM classifier (Hindi)

Emotion	Emotion Recognition (%)				
	Angry	**Sad**	**Happy**	**Neutral**	
Angry	89.47	0	5.26	5.26	
Sad	10	50	40	0	
Happy	14.28	7.14	78.57	0	
Neutral	41.17	0	5.88	58.82	

Table 2: Confusion matrix for SVM classifier (German)

Emotion	Emotion Recognition (%)				
	Angry	**Sad**	**Happy**	**Neutral**	**Fear**
Angry	88.24	0	7.84	0	3.92
Sad	0	64.0	0	36.0	0
Happy	39.28	0	42.85	10.71	17.14
Neutral	3.12	43.75	0	50.0	3.12
Fear	22.22	14.81	7.4	11.11	44.44

Table 3: Confusion matrix for LIBSVM RBF classifier (Hindi)

Emotion	Emotion Recognition (%)			
	Angry	**Sad**	**Happy**	**Neutral**
Angry	100	0	0	0
Sad	10	50	40	0
Happy	7.14	7.14	85.72	0
Neutral	35.29	0	0	64.7

Table 4. Confusion matrix of the LINEAR LIBSVM classifier (German Gender Independent)

Emotion	Emotion Recognition (%)				
	Angry	**Sad**	**Happy**	**Neutral**	**Fear**
Angry	**92.16**	**0**	**5.88**	**0**	**1.96**
Sad	**0**	**100**	**0**	**0**	**0**
Happy	**71.42**	**0**	**21.43**	**14.28**	**0**
Neutral	**0**	**15.62**	**0**	**31.28**	**15.62**
Fear	**17.85**	**3.57**	**25**	**0**	**46**

Table 5. Confusion matrix of the RBF LIBSVM classifier (German Gender Independent)default value of c

Emotion	Emotion Recognition (%)				
	Angry	**Sad**	**Happy**	**Neutral**	**Fear**
Angry	**94.12**	**0**	**5.88**	**0**	**0**
Sad	**0**	**100**	**0**	**0**	**0**
Happy	**42.85**	**0**	**57.58**	**3.57**	**0**
Neutral	**0**	**9.37**	**0**	**87.5**	**3.12**
Fear	**14.28**	**3.57**	**25**	**7.14**	**50**

Table 6. Confusion matrix of the RBF LIBSVM classifier (German Gender Independent)c=2

Emotion	Emotion Recognition (%)				
	Angry	**Sad**	**Happy**	**Neutral**	**Fear**
Angry	**96.08**	**0**	**3.92**	**0**	**0**
Sad	**0**	**100**	**0**	**0**	**0**

Happy	3.57	0	64.29	14.28	0
Neutral	0	6.25	0	93.75	0
Fear	17.85	7.14	21.42	0	53.58

Table 7. Confusion matrix of the RBF LIBSVM classifier (German Gender Independent)c=3

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	96.08	3.92	0	0	0
Sad	0	100	0	0	0
Happy	3.25	0	75	14.28	0
Neutral	0	6.25	0	93.75	0
Fear	14.28	0	21.42	3.57	60.72

Table 8. Confusion matrix of the RBF LIBSVM classifier (German Gender Independent)c=4

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	100	0	0	0	0
Sad	0	100	0	0	0
Happy	21.42	0	75	14.28	0
Neutral	0	6.25	3.125	90.625	0
Fear	10.71	0	25	7.14	57.15

Table 9. Confusion matrix of the RBF LIBSVM classifier (German Gender Independent)c=5

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	98.06	0	1.96	0	0
Sad	0	100	0	0	0
Happy	17.85	0	78.58	3.57	0
Neutral	0	0	3.125	96.875	0
Fear	7.14	0	28.57	3.57	60.72

Table 10. Confusion matrix of the RBF LIBSVM classifier (German Gender Independent)c=6

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	98.04	0	1.96	0	0
Sad	0	100	0	0	0
Happy	17.85	0	78.58	0	3.57
Neutral	0	0	3.125	96.875	0
Fear	7.14	0	25	0	67.86

Table 11. Confusion matrix of the RBF LIBSVM classifier (German Gender Independent)c=7

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	100	0	0	0	0
Sad	0	100	0	0	0
Happy	10.71	0	85.72	0	3.57
Neutral	0	3,125	0	96.87	0
Fear	7.14	0	25	0	67.86

Table 12. Confusion matrix of the RBF LIBSVM classifier (German Gender Independent)c=8

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	100	0	0	0	0
Sad	0	100	0	0	0
Happy	3.57	0	96.43	0	0
Neutral	0	0	0	100	0
Fear	0	0	0	0	100

Table 13. Confusion matrix of the RBF LIBSVM classifier (German Gender Independent)c=9

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	100	0	0	0	0
Sad	0	100	0	0	0
Happy	3.57	0	96.43	0	0
Neutral	0	0	0	100	0
Fear	0	0	0	0	100

Table 14. Confusion matrix of the RBF LIBSVM classifier (German Gender Independent)c=10

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	100	0	0	0	0
Sad	0	100	0	0	0
Happy	3.57	0	96.43	0	0
Neutral	0	0	0	100	0
Fear	0	0	0	0	100

Table 15. Percentage recognition using variable cost value c for RBF Kernel from table 5 to 14

Cost value	German Gender Independent Accuracy
C=1	79.2683%
C=2	83.5366%
C=3	86.585%
C=4	86.585%
C=5	88.414%
C=6	89.0244%
C=7	91.4634%
C=8	99.3902%
C=9	99.3902%
C=10	99.39.2%

7. CONCLUSION

In this paper Hindi emotion database is created and used that contains a emotion dialogues from different Bollywood movies Initially all files are cut in .mp3 format and then converted into (.wav) and Berlin emotion database of German language is used for feature extraction. MFCC and MEDC features are extracted from a speech files in .wav format. From experimentation and result it is proved that system is speaker and text independent. The recognition rates by implemented SVM are 62% and 71.66% for Berlin database and Hindi database respectively It is also observed that results from LIBSVM by using RBF kernel function are 99.39% for cost value c=8 . The recognition rates by LIBSVM using RBF function for Hindi database are 78.33% .The accuracy of Linear LIBSVM is 68.902%. Regarding LIBSVM using RBF kernels it is observed that by changing the parameters of a kernel functions better results can be obtain. It is observed that increasing value of cost in RBF Kernel function increases the accuracy rate at certain point and then remains constant.

8. REFERENCES

[1] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G., Emotion recognition in human-computer interaction, IEEE Signal Processing magazine, Vol. 18, No. 1, 32-80, Jan. 2001.

[2] D. Ververidis, and C. Kotropoulos, Automatic speech classification to five emotional states based on gender information, Proceedings of the EUSIPCO2004 Conference, Austria, 341-344, Sept. 2004.

[3] Christopher. J. C. Burges, A tutorial on support vector machines for pattern recognition, DataMining and Knowledge Discovery, 2(2):955-974, Kluwer Academic Publishers, Boston, 1998.

[4] Tristan Fletcher, Support Vector Machines Explained, unpublished.

[5] Burkhardt, Felix; Paeschke, Astrid; Rolfes, Miriam; Sendlmeier, Walter F.; Weiss, Benjamin A Database of German Emotional Speech. Proceedings of Interspeech, Lissabon, Portugal. 2005.

[6] Fuhai Li, Jinwen Ma, and Dezhi Huang, MFCC and SVM based recognition of Chinese vowels, Lecture Notes in Artificial Intelligence, vol.3802, 812-819, 2005

[7] M. D. Skowronski and J. G. Harris, Increased MFCC Filter Bandwidth for Noise-Robust Phoneme Recognition, *Proc. ICASSP-02*, Florida, May 2002.

[8] YL. Lin and G. Wei, Speech emotion recognition based on HMM and SVM, proceeding of fourth International conference on Machine Learning and Cybernetics,Guangzhou, 18-21 August 2005.

[9] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[10] C.W Hsu, C.-C. Chang, C.-J. Lin, A Practical Guide to Support Vector Classification, *Technical Report*, Department of Comptuer Science & Information Engineering, National Taiwan University, Taiwan.

[11] O.Khalifa,S.Khan,M.R.Islam, M.Faizal and D.Dol, 2004."Text Independent Automatic Speaker Recognition".3rd International Conference on Electrical & Computer Engineering, Dhaka, Bangladesh, pp.561-564.

[12] Ying Wang, Shoufu Du , Yongzhao Zhan, Adaptive and Optimal Classification of Speech Emotion Recognition Fourth International Conference on Natural Computation.

Author Biographies

Miss. Sujata B. Wankhade, student of second year Post Graduate Degree (M.E.) in Branch Information and Technology at Sipna College of Engg. And Technology. from S.G.B. Amravati University, Amravati in the year 2010-11.

Prof. P. A Tijare has completed M.E in Comp. Tech & Application and working as Assistant Professor I.T Dept. at Sipna C.O.E.T from last 7 years in SGB Amravati University, Amravati.