African Journal of Basic & Applied Sciences 8 (3): 136-144, 2016
ISSN 2079-2034

DOI: 10.5829/idosi.ajbas.2016.8.3.22803

Optimal Feature for Emotion Recognition from Speech

[1]C. Vijesh Joe and [2]S. Shinly Swarna Sugi

[1]Department of Information Technology, Karpagam College of Engineering, Coimbatore, India
[2]Department of Computer Science and Engineering, Karpagam College of Engineering, Coimbatore, India

Abstract: Emotion recognition plays an important role in various applications like from content based digital library search to psychological analysis. The motivation of this paper is to improve the quality of speech emotion recognition by separating the emotions clearly using the speech features with appropriate classifiers. At the time of finding the emotions from speech after feature extraction and model generation, the acoustical resemblance of certain emotions like happiness/angry, sad/boredom produce high correlations/likelihood so that they can't make it as unique while extracting the features. In this paper, the classification of seven emotional states anger, happy, sad, neutral, fear, disgust and boredom using the optimal features is selected by the Sequential Forward Selection (SFS) algorithm and Gaussian Mixture Model (GMM) classifier. Short term energy, MFCC and pitch contour are the features that has been considered in this work. Selecting the optimal features is used for recognizing the emotions with more accuracy.

Key words: Gaussian Mixture Model (GMM) · Mel Frequency Cepstral Coefficient (MFCC) · Emotional Recognition Model · Sequential Forward Selection (SFS) · Pitch Contour · Energy

INTRODUCTION

The emotion of speech is a kind of force on all sounds across the speech. In general, every speaker has their own speaker dependent style like characteristic articulation rate, intonation habit and loudness characteristic. The way of expressing the emotion and deducing it in a speech is dependent on the speaker's community, culture and language, gender, age, education, social status, health and physical engagements, etc. The emotions are majorly classified into happy, sad, anger, surprise, normal, fear and neutral. For example if a speaker in a silent room and in normal state the speech produced by him is 'neutral' [1-14]. Likewise due to certain changes in the environment and in the person the emotions arose in him which sets him in an emotional state.

In speech emotion recognition processing, one of the challenges is accurately recognizing the emotions from speech from the defined feature set. The ability to recognize.

Emotional states of a person perhaps the most important for successful inter-personal social interaction [15]. In general, the emotional speech recognition system can be characterized by the used features, the investigated emotional categories, the methods to collect speech utterances, the languages and the type of classifier used in the experiments. Emotions are analyzed using various spectral and prosody features but the accuracy of the recognition is limited due to acoustical resemblance of few emotions, like anger and happiness, sad and boredom. This paper work is mainly focused on improving the accuracy of recognition by selecting the optimum feature set among the available feature pool.

In order to select the optimum feature set, Sequential Forward Selection (SFS) algorithm has been used in which it sequentially select or add features from the feature pool to the optimal feature set based on the concept that addition of a particular feature is optimizing the defined objective function. The objective function of SFS algorithm is fixed to improve the accuracy in such a way to move the closely resembled emotions to move apart. Berlin Emotional Data base – EMO- DB as speech corpus was used to carry out this experiment. Also speech prosody features: energy and pitch contour and spectral features: Linear Prediction Coefficients (LPC), Mel Frequency Cepstral coefficients (MFCC) as features for machine learning. Gaussian Mixture model is a classifier

Corresponding Author: C. Vijesh Joe, Department of Information Technology, Karpagam College of Engineering, Coimbatore, India.

has been used in which this model is used as a parametric model of a probability distribution of speech features for machine learning. To generalize the results to be independent of data set, ten fold cross validation is done on corpus [6]. The experimental result shows improvement in the performance of emotion recognition system.

The organization of the paper is given below.

•Section 2 – Speech Signal Digitization
•Section 3 – Framing of Speech Signal
•Section 4 – Emotion Related Features of Speech
•Section 5 – Feature Selection
•Section 6 - Classifier
•Section 7 – Performance Analysis
•Section 8 - Conclusion

Speech Signal Digitization: Capturing the analog signal in digitized form is called as digitization. Here the discrete set of points is called samples [4]. The digitization of speech is done by sampling and quantization. Fig.2.1 represents the digitization of speech.

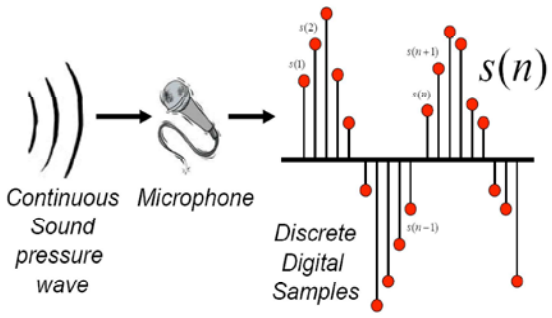


Fig. 2.1: Speech Signal Digitization [16-19]

Sampling: Sampling is the reduction of a continuous signal to a discrete signal i.e. the conversion of a sound wave (a continuous signal) to a sequence of samples (a discrete-time signal).

Quantization: Quantization is the process of mapping a large set of input values to a smaller set. Here representing real value of each amplitude as integer.

Framing of Speech Signal

Windowing: Speech processing systems divide the sampled speech signal into overlapping frames of size about 20-30 ms [13]. The N speech samples within each frame are processed and represented by a set of spectral features or linear prediction model of speech production.

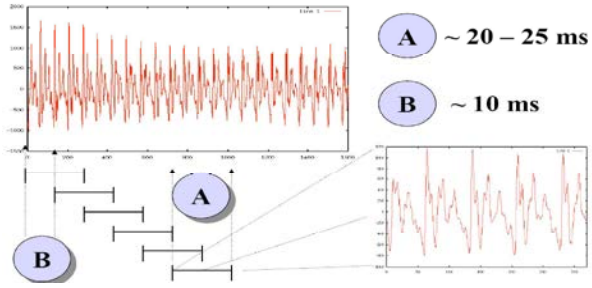


Fig. 3.1: Windowing function[17]

Frame Segmentation: The speech signal is generally considered as non-stationary, thus it is referred to as "quasi-stationary". To know the information about small regions of speech signal, segmentation is used to perform based on the frame size where the range should be within 10-25 milliseconds (ms). Also there is another term called Frame shift which represents the length of time between successive frames.

Common Window Shapes: Windowing is generally used to refer the term frame, which is used to analyze the short time interval of speech signal. To extract the information about the particular frame we generally move towards windowing technique.In speech signal processing many windowing techniques are available like Rectangular window, Hamming window, Barlett window and Hanning window. The advantage of the hamming window is smoothing of the curve and it introduces the least amount of impulse response of the window.
The formula for hamming window, h(t) is given by,

$$w(n) = 0.54 - 0.46\ \cos\left(\frac{2\pi n}{N-1}\right)$$

Where h (t) = signal*hamming window.

Time Domain and Frequency Domain Representation of Hamming Window: In time domain your model/system is evaluated according to the progression of it's state with time.

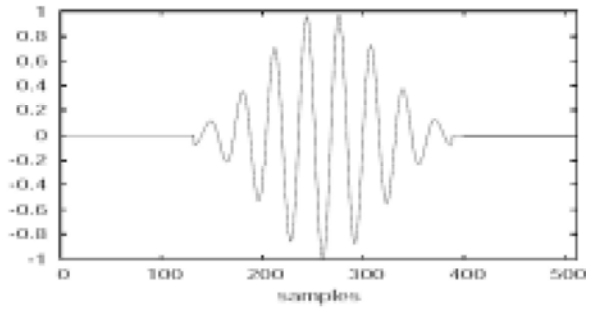
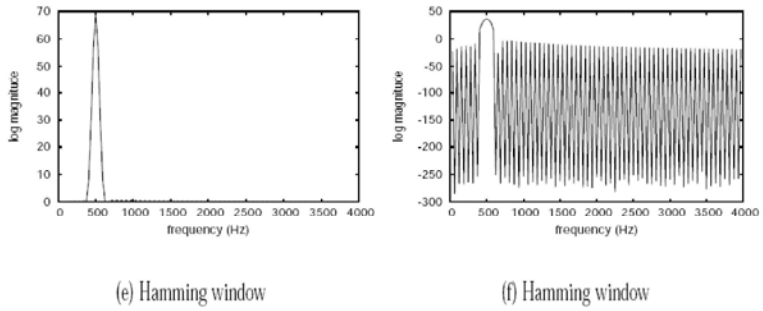


Fig. 3.2: Hamming window in Time Domain[18]

In Frequency domain your model/system is analyzed according to it's response for different frequencies.



(e) Hamming window (f) Hamming window

Fig. 3.2: Hamming window in Frequency

Speech Corpus: A speech corpus (or spoken corpus) is a database of speech audio files and text transcriptions. Speech corpus is very essential for emotion recognition from speech. The emotions that are recorded in database were acted, forced emotions. Spontaneous speech emotions would have been a better base for a perceptual investigation, but acquiring that is much harder. It does not to mention the ethical dilemma of forcing people to produce genuine emotions. In most social environment a grownup human being express the emotions aloud is highly impolite for most circumstances. So at the time of building an emotional database, one must make sure that the emotional information carried by the speech should not be affected with the possible emotional information of the semantic layer already present in the speech.

Cross Validation: In machine learning, fitting process optimizes the model parameters to make the model fit the training data as well as possible. If we then take an independent sample of validation data from the same population as the training data, it will generally turn out that the model does not fit the validation data as well as it fits the training data. This is called over fitting and is particularly likely to happen when the size of the training data set is small, or when the number of parameters in the model is large. Cross-validation is a way to predict the fit of a model to a hypothetical validation set when an explicit validation set is not available [10].

Cross-validation is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction and one wants to estimate how accurately a predictive model will perform in practice. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset called the training set and validating the analysis on the other subset called the validation set or testing set.

K-Fold Cross-Validation: In k-fold cross-validation, the original sample is randomly partitioned into k subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model and the remaining k - 1 subsamples are used as training data [18]. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds then can be averaged (or otherwise combined) to produce a single estimation. Over repeated random sub-sampling is that all observations are used for both training and validation and each observation is used for validation exactly once is The advantage of this method. 10-fold cross-validation is commonly used, but in general k remains an unfixed parameter.

Emotion Related Features of Speech: An important issue in the design of speech emotion recognition system is the extraction of suitable features that efficiently characterize different emotions [7]. There are many features evolved in emotion recognition. Some of the dominant features are listed below.

Signal Energy: The speech signal and its sampling frequency along with the frame size and frame shift are the inputs needed for computing the short term energy. Using the sampling frequency value, the number of samples for the given frame size and frame shift are computed. Also to compute short term energy, the input speech signal is considered in terms of 160 samples with a shift of 80 samples and the energy is computed for each frame. The short term energy values are then plotted as a function of time index.

This energy extraction is based on the segmentation where the speech signal is divided into frames and the energy for each frame has been calculated by the formula of,

$$E = \sum_{n=-\infty}^{\infty} s^2(n)$$

Figure 4.1 shows the extraction of energy for an anger wave. The peak shows the higher energy of an anger wave.

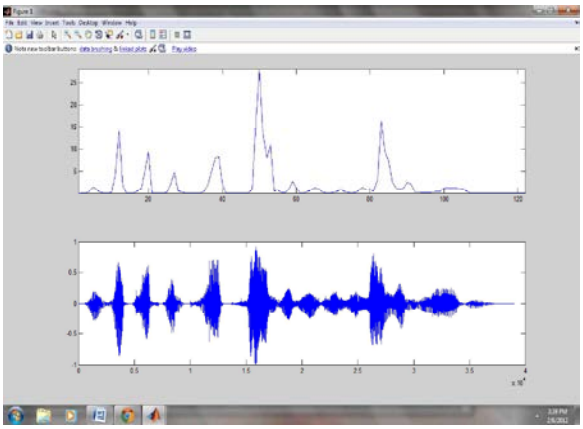


Fig. 4.1: Energy extraction of an emotional wave

Mel Frequency Cepstral Coefficient: MFCC are result from a transformation to a cepstrum space, in order to capture information of the time-varying spectral envelope [16]. A cepstrum can be obtained by applying a Fourier Transform on the log (f) plot, in order to separate in the frequency domain the slowly varying spectral envelope from the more rapidly varying spectral ?ne structure (separation of the source and ?lter spectrum)

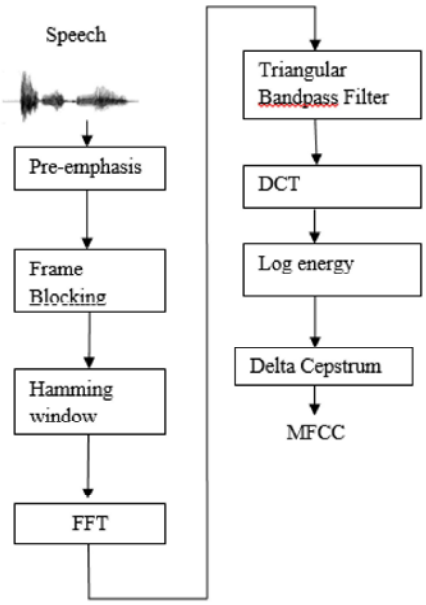


Fig. 4.2: MFCC Extraction from speech signal.

Pre-Emphasis: Boosting the energy in the high frequencies is generally refers Pre-Emphasis [11]. The spectrum for voiced segments has more energy at lower frequencies than higher frequencies referred as spectral tilt. The Spectral tilt is caused by the nature of the glottal pulse and boosting high-frequency energy gives more information to acoustic model which improves the phone recognition performance.

Mel Scale: A mel is a unit of pitch which defines a pair of sounds perceptually equidistant in pitch are separated by an equal number of mel. Mel-scale is approximately linear below 1kHz and logarithmic above 1 kHz.

Frame Blocking: The input speech is segmented into frames of 20-30ms with an overlap of 0.5 of the frame size. If the sample rate is 16 kHz and window size is 20 ms then 320 sample points will be the result. Twelve MFCC (Mel frequency cepstral coefficients) and one energy feature is used in this paper. With that twelve delta MFCC features, twelve double-delta MFCC features, one delta energy feature and one double-delta energy feature are added and so 39-dimensional features were obtained.

Hamming Window: To keep the continuity of the first and last points in the frame hamming window is generally used and so in order to observe clearly here in this paper hamming window is used [17].

Mel-Filter Bank Processing: Mel-Filter Bank Processing is a process of applying the bank of filters according Mel scale to the spectrum. Here each filter output is the sum of its filtered spectral components.

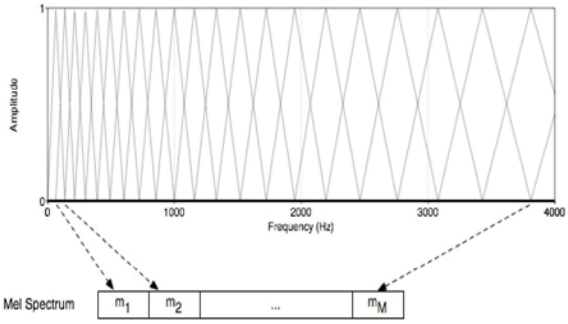


Fig. 4.3: Mel Filter Processing [17]

Log Energy Computation: Computing the logarithm of the square magnitude of the output of Mel-filter bank is said to be known as Log energy computation.

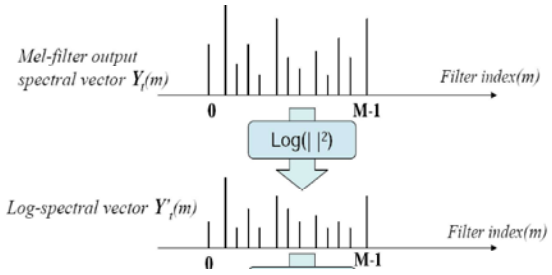


Fig. 4.4 Log energy computation [17]

Cepstrum: The Fourier transform of the logarithm of spectrum of a signal is said to be called as cepstrum. It is of many types like complex cepstrum, real cepstrum, power cepstrum and phase cepstrum. Among this power cepstrum is used for the analysis of human speech.

FFT: Spectral analysis shows that different timbers in speech signals corresponds to different energy distribution over frequencies. Therefore FFT is usually perform to obtain the magnitude frequency response of the system. It is also an efficient algorithm to compute DFT and its inverse. For N discrete frequency bands a complex number X[k] representing magnitude and phase of that frequency component in the original signal Discrete Fourier Transform (DFT) is given by

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi \frac{k}{N} n}.$$

DFT for Computing a Spectrum: A 25 ms Hamming-windowed signal and its spectrum as computed by DFT (plus other smoothing). Human hearing is not equally sensitive to all frequency bands. It is less sensitive at higher frequencies, roughly greater than 1000 Hz (i.e.) human perception of frequency is non-linear.

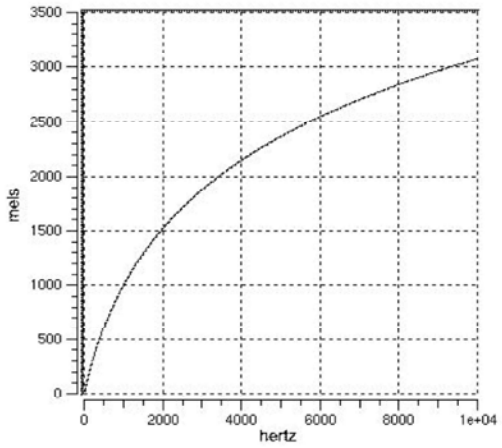


Fig. 4.5: DFT for computing spectrum

MFCC represents the short term power spectrum of a sound based on the linear cosine transform of a log power spectrum on a non-linear mel scale of frequency. Here 39 coefficients are taken from MFCC.

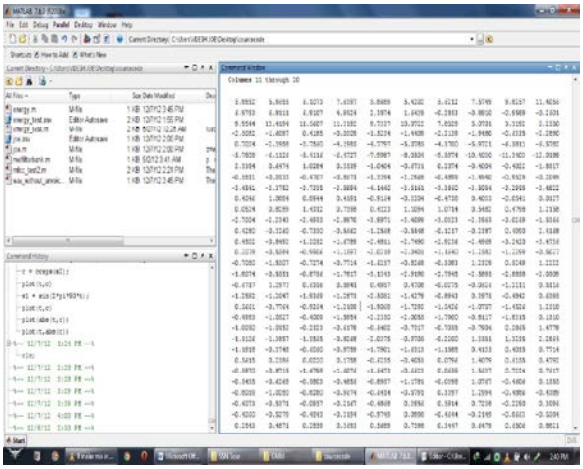


Fig. 4.6: Result of an MFCC extraction

Pitch Contour: Pitch contour is measured based on the pitch estimation. Pitch may be estimated in three methods like autocorrelation, cepstral and SIFT method. In this paper, pitch was calculated based on the autocorrelation method and the pitch period was computed by finding the time lag corresponds to the second largest peak from the central peak of autocorrelation sequence using simple peak picking algorithm. Using threshold silence and unvoiced sounds are eliminated from pitch. Finally smoothening the pitch using median filters resulted in pitch contour.

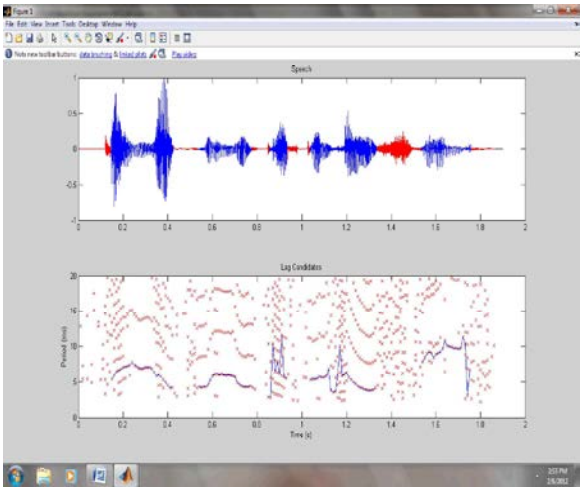


Fig. 4.7: Results of Pitch contour extraction.

Feature Selection

Sfs Algorithm: The algorithm starts with a null feature set and, for each step, the best feature that satisfies some criterion function is included with the current feature set, i. e., one step of the sequential forward selection (SFS) is performed [12]. The algorithm also verifies the possibility of improvement of the criterion if some feature is excluded. In this case, the worst feature (concerning the criterion) is eliminated from the set, that is, it is performed one step of sequential backward selection (SBS). Therefore, the SFFS proceeds dynamically increasing and decreasing the number of features until the desired is reached.

Classifier

Gaussian Mixture Model Classifier: GMM classifier is considered in this paper to classify the emotions. A Gaussian Mixture Model is a parametric probability density function represented as a weighted sum of Gaussian component densities [2]. They are commonly used as a parametric model of a probability distribution of continuous measurements or features. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation from a well-trained prior model [5]. These acoustic features have all been modeled by Gaussian mixture models, GMMs, on the frame level. The method has been tested on corpora in which the results indicate that using GMMs on the frame level is a feasible technique for emotion classification.

The parameters of GMM are estimated from training corpus using the K-Means algorithm iterative Expectation-Maximization (EM) algorithm. With the calculated GMM parameters, the test files are verified against different GMM model, the best suitable model is chosen as the class of test file. The accuracy of the Model will be calculated and compared for GMM with different feature sets. The feature which produces high accuracy in recognizing the emotion will be taken as the optimal feature set. The fusion of the optimal feature set and building a model for that will increases the accuracy [9]. The M-component Gaussian Mixture Model is a weighted sum of M component Gaussian densities as given by the equation,

$$p(x/\lambda) = \sum_{i=1}^{m} w_i g(x|\mu_i,\Sigma_i)$$

Each component density is a D-variate Gaussian function of the form,

$$g(x/\mu_i,\Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu_i)'\sum_i^{-1}(x-\mu_i)\right\}$$

with mean vector ìi, mixture weight wi and covariance matrix ∑i.. The mixture weights satisfy the constraint $\sum_{i=1}^{m} w_i = 1$. The complete Gaussian Mixture Model is parameterized by the mean vectors,

Mixture Weight:

$$w_i = \frac{1}{T}\sum_{t=1}^{T} Pr(i|x_t,\lambda)$$

Mean:

$$\mu_i = \frac{\sum_{t=1}^{T} Pr(i|x_t,\lambda)\, x_t}{\sum_{t=1}^{T} Pr(i|x_t,\lambda)}$$

Variance:

$$\sigma_i^2 = \frac{\sum_{t=1}^{T} Pr(i|x_t,\lambda) x_t^2}{\sum_{t=1}^{T} Pr(i|x_t,\lambda)} - \mu_i^2$$

The a posteriori probability for component i is given by

$$Pr(i|x_t,\lambda) = \frac{w_i g(x_t|\mu_i\Sigma_i)}{\sum_{k=1}^{M} w_k g(x_t|\mu_k,\Sigma_k)}$$

Likewise, for every feature the probability level has been calculated with respect to every emotion. (ie. for every feature of energy, mfcc, pitch contour) The probability level gets varies with respect to every feature. Based on this the feature selection and performance analyzation has been taken.

Key Features of GMM:

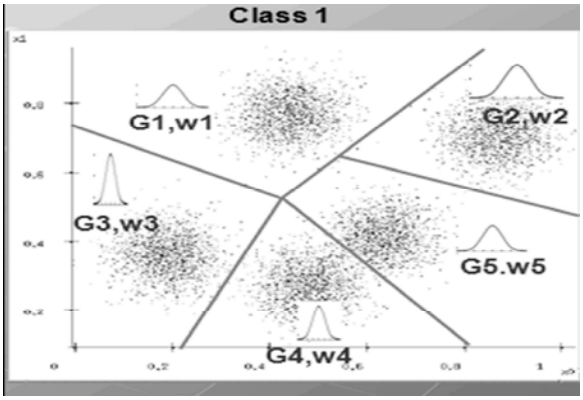


Fig. 6.1: Mixture components of GMM [17]

- Mixture Models are a type of density model which comprise a number of components usually Gaussian.
- These component functions are combined to provide a multi modal density
- Each component can also be associated with a form of 'a priori' probability: shows relative importance of each component.

Algorithms under GMM: There are some algorithms generally used for clustering in GMM. They are Maximum Likelihood, K-Means Clustering and Expectation Maximization. We have used K-means clustering algorithm here for this work.

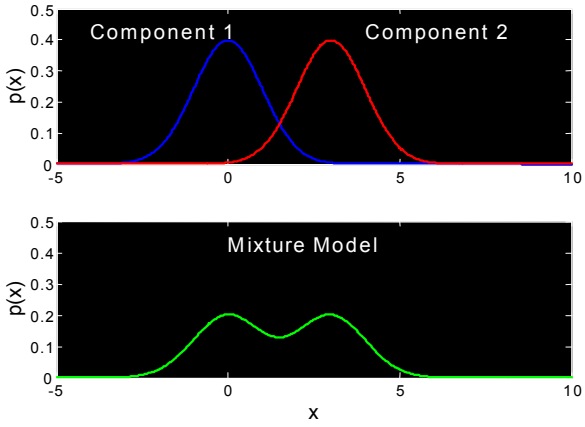


Fig. 6.2: Mixture model of two components[17]

Clustering: Clustering is defined as the process of organizing the objects into groups whose members are similar in some way" and therefore a collection of objects which are "similar" between them in one class and are "dissimilar" to the objects belonging to other clusters [3]. Organizing data into classes such that there is high intra-class similarity and low inter-class similarity and finding the class labels and the number of classes directly from the data (in contrast to classification) is more informally, finding natural groupings among objects.

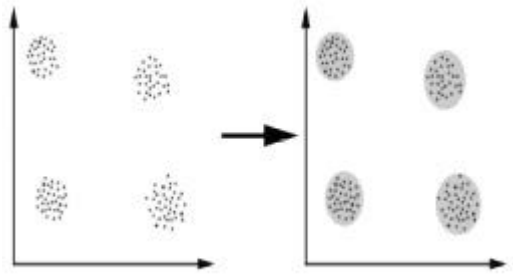


Fig. 6.3: Clustering is an unsupervised learning algorithm [17]

Performance Analysis

Cross-Validation: In this paper, ten fold cross validation has been performed. That means, 75 percent of partition is used for training and remaining 25 percent is used for testing purpose. Hence 10 rotations have been taken place. For every rotation, the performance in the accuracy of recognizing the emotion was taken place [8].

The results obtained using the feature MFCC, LPC, energy is better than result obtained from pitch contour. In overall, the sad emotion gives better accuracy and other emotions results are moderately better. The resemblance is noted among anger and happiness and also among disgust and sad. Optimal features found SFS (Sequential Forward Floating Selection) algorithm increases the accuracy level in emotion recognition.

Training: Under GMM, Training and testing has been performed. As we have followed 10 fold cross validation in this paper every single fold that consists of 90% of files were given for training. In the first extraction of features taken based on the segmentation of speech signal into frames [1]. Collection of those features was used for building the GMM model in order to recognize the emotion from speech. Mean covariance and mixtures are found in training. Using these parameters alone the testing has been performed.

Testing: Here in testing the remaining 10 % of files were given. One by one every file was tested. Like training in testing too initially the speech was segmented into frames. Through testing the accuracy in the recognization of the emotion has been noted. The following tables indicate how much probability that each emotion has been found at the time of testing using individual feature and combining feature. The results shows that the combining features give good accuracy than the individual features. For initial testing, three emotions was taken as in Table 4.1and for feature combination, all the emotions were taken as in Table 4.2.

Table 7.1: Results obtained with independent features

Emotion/Feature	MFCC%	Energy%	Pitch contour%
Anger	75	86	75
Sad	97	45	40
Neutral	90	54	39
Happy	84	65	48
Fear	76	45	61

Table 7.2: Results obtained with optimal feature

Feature/emotion	Fusion of features (MFCC,Pitch Condour, Energy)
Anger	76
Sad	67
Neutral	74
Happy	84
Fear	71

The above two Tabular columns clearly shows about the result obtained with independent features and result obtained with optimal features.

CONCLUSION

The emotions have been identified using GMM classifier using the fusion of optimal features MFCC and energy. The overall accuracy is fairly good. Generally HMM is used for speech processing where the states are interdependent. But in emotion the sequence of states is not maintained, thus GMM classifies faster than HMM. The results were analyzed using various features. But while combining these features and generating a model for this feature fusion shows very good accuracy than the individual features. Also this work provides a generic algorithm for feature fusion with the objective function of improvement in recognition accuracy. Thus it can be extended to any feature sets with little modification.

REFERENCES

1. Vijesh Joe C., 2013. Building and Evaluation of Tamil Emotional Speech Corpus, in the 5th National Conference on Signal Processing Communication and VLSI Design (NCSCV'13)on conducted by Anna University Regional Centre,Coimbatore.
2. Gaussian Mixture Models Douglas Reynolds, Lexington, MA 02140, USA dar@ll.mit.edu.
3. A Modified Fuzzy K-means Clustering using Expectation Maximization Sara Nasser, Rawan Alkhaldi, Gregory Vert Department of Computer Science and Engineering, 171, University of Nevada Reno, Reno NV 89557, USA.
4. A Speaker Independent Approach to the Classification of Emotional Vocal Expressions Hicham Atassi and Anna Esposito, 2008 20th IEEE International Conference on Tools with Artificial Intelligence.
5. Emotion recognition from Assamese speeches using MFCC features and GMM classifier, Aditya Bihar Kandali, IEEE, Aurobinda Routray, Member, IEEE and Tapan Kumar Basu,2009.
6. Calculation of Multivariate Normal Probabilities By Simulation, with Applications To Maximum Simulated Likelihood Estimation, Lorenzo Cappellari (University Cattolica, Milano and ISER) and Stephen P.Jenkins (ISER, University of Essex) ISER Working Paper,200-16.
7. Speech emotion recognition using both spectral and prosodic features,Chinese Academy of Sciences Beijing 100190,P.R.China,2009.
8. Emotional speech recognition: Resources, features and methods Dimitrios Ververidis and Constantine Kotropoulos, Artificial Intelligence and Information Analysis Laboratory, 541 24, Greece,2000.
9. Emotion Recognition and Evaluation of Mandarin Speech Using Weighted D-KNN Classification, Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh.
10. GMM super vector based SVM with spectral features for speech emotion recognition,Hao Hu, Ming-Xing Xu and Wei WuCenter.
11. Novel Hilbert Energy Spectrum Based Features for Speech Emotion Recognition, 2010 WASE International Conference on Information Engineering
12. Exploring the Benefits of Discretization of Acoustic Features for Speech Emotion Recognition T. Vogt and E. André, 2009.Proc. Int'l Speech Comm. Assoc., pp: 328-331.
13. Spectral Properties and Prosodic Parameters of Emotional Speech in Czech and Slovak Jiří Přibil and Anna Přibilová.
14. Emotion Aware System Based on Acoustic and Textual Features from Speech Yan-You Chen, Bo-Wei Chen, Jhing-Fa Wang, Yi-Cheng Chen, 2010 IEEE.
15. Feature Analysis for Emotion Recognition from Mandarin Speech Considering the Special Characteristics of Chinese Language Yi-hao Kao and Lin-shan Lee.
16. Emotion recognition in the next generation: an overview and Recent development, 2009.Bjorn schuller, 10th Annual Conference of International Speech Communication Association.

17. Speech Emotion Recognition based on Multi-output GMM and SVM,Fei Dong, Guobao Zhang, Yongming Huang, Haibin Liu School of Automation, China In proceeding of Chinese conference Pattern Recognition (CCPR) -2010.
18. Campbell J.P., 1997. Speaker Recognition: A Tutorial, 85.
19. Handan, China, Xuxiong Ling , Fuliang Zhang and Jianing Tong, 2009. Speech Emotion Recognition Based on Principal Component Analysis and Back Propagation Neural Network, SheguoWang Inf. & Electron. Eng. Inst., HeBei Univ. Of Eng.