A Fuzzy C-Means based GMM for Classifying Speech and Music Signals

R.Thiruvengatanadhan
Assistant Professor,
Department of Computer Science and Engineering
Annamalai University,
Annamalainagar,
Tamilnadu, India

P. Dhanalakshmi, Ph.D
Associate Professor,
Department of Computer Science and Engineering
Annamalai University,
Annamalainagar,
Tamilnadu,India

ABSTRACT

Gaussian Mixture Model (GMM) with Fuzzy c-means attempts to classify signals into speech and music. Feature extraction is done before classification. The classification accuracy mainly relays on the strength of the feature extraction techniques. Simple audio features such as Time domain and Frequency domain are adopted. The time domain features are Zero Crossing Rate (ZCR) and Short Time Energy (STE). The frequency domain features are Spectral Centroid (SC), Spectral Flux (SF), Spectral Roll-off (SR) and Spectral Entropy (SE) and Discrete Wavelet Transforms. The features thus extracted are used for classification. Commonly GMM uses Expectation Maximization (EM) algorithm to determine parameters. The proposed GMM makes use of fuzzy c-means algorithm. The fuzzy c-means algorithm is used to estimate the parameters of the GMM. Compute the probability density function and fix the Gaussian parameter. The proposed GMM model classifies the given input signal is either speech or music and compared with GMM using EM algorithm.

Keywords

Classification, Feature extraction, Discrete Wavelet Transform, Fuzzy c-means, Gaussian Mixture Model.

1. INTRODUCTION

Speech is transmitted through sound waves, which follow the basic principles of acoustics. The source of all sound is vibration. For sound to exist, a source (something put into vibration) and a medium (something to transmit the vibrations) are necessary. Important basic characteristics of waves are wavelength, amplitude, period, and frequency. Wavelength is the length of the repeating wave shape. Amplitude is the maximum displacement of the particles of the medium, which is determined by the energy of the wave. The time required to pass one wave at a given point is known to be as period. The number of waves passing at a time is termed as frequency [9]. Each and every complete vibration of a wave is called as cycle. Intensity and duration are other two physical properties of sound frequency are perceived as pitch whereas intensity is perceived as loudness [5].

Approaches in speech\music change point detection can be categorized into metric-based, model-based, decoder-guided, model-selection-based and hybrid approaches. Metric-based methods simply measure the difference between two consecutive audio clips that are shifted along the audio signal, and speech\music changes are identified at the maxima of the dissimilarity in terms of some distance metric, e.g. vector quantization distortion (VQD), KL distance and divergence shape distance (DSD). Model-based approaches are based on recognizing specific speakers via Gaussian mixture models (GMM) or hidden Markov Models (HMM). Decoder guided approach that segments a speech stream into male and female clips via a gender-dependent phone recognizer. In model-selection based methods, the segmentation problem is switched to a model selection problem between two nested competing models. Bayesian information criterion (BIC) is often adopted as the model selection criterion since it has some nice properties such as robustness, threshold-free and optimality. Recently, much effort has been devoted to hybrid methods that combine the merits from above different approaches to achieve better performance over single approaches.

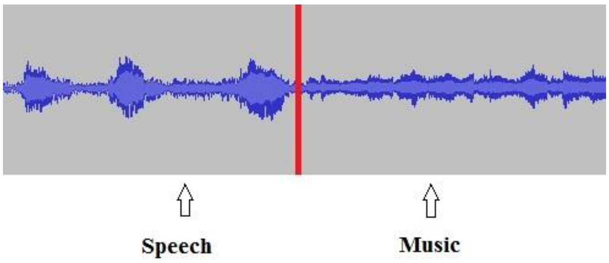

Fig. 1: Speech/Music classification system.

Music is an art form whose medium is sound and silence. Pitch, rhythm, dynamics, and the sonic qualities are common elements of timbre and texture. The term audio is used to indicate various kinds of audio signals, such as speech, music as well as more general sound signals combinations of audio recordings. However, the audio is usually treated as an opaque collection of bytes with only the most primitive fields attached, namely, file format, name, sampling rate, etc. Meaningful information can be extracted from digital audio waveforms in order to compare and classify the data efficiently. When such information is extracted, it can be stored as content description in a compact way [4].

A data descriptor is often called a feature vector and the process for extracting such feature vectors from audio is called audio feature extraction. Usually a variety of more or less complex descriptions can be extracted to feature one piece of audio data. The efficiency of a particular feature used for comparison and classification depends greatly on the application, the extraction process and the richness of the description itself. Digital analysis may discriminate whether an audio file contains speech, music or other audio entities. Great convenience will be provided for material searching and browsing in audio libraries to retrieve movie and sound clips [14], [12].

2. OUTLINE OF THE WORK

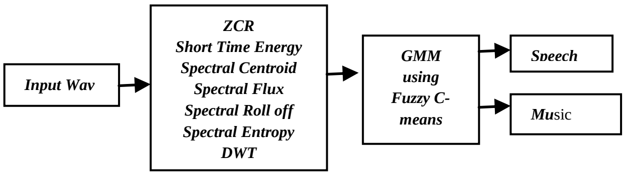

Fig. 2: Block Diagram for speech/music classification

In this paper, automatic audio feature extraction and classification approaches are presented. In order to discriminate the speech and music features such as Time domain features are Zero Crossing Rate (ZCR) and Short Time Energy (STE), the frequency domain features are Spectral Centroid (SC), Spectral Flux (SF), Spectral Roll-off (SR) and Spectral Entropy (SE) and Discrete Wavelet Transforms (DWT) are extracted to characterize the audio content. The GMM classification method is implemented using fuzzy c-means based clustering algorithm approach to fit the GMM parameters for classification. Experimental results show that the classification accuracy of GMM with Time domain and Frequency domain features can provide a better result. Figure 2 illustrates the block diagram of Speech/Music classification system.

3. ACOUSTIC FEATURES FOR AUDIO CLASSIFICATION

Acoustic feature extraction plays an important role in constructing an audio classification system. The aim is to select features which have large between class and small within class discriminative power. Discriminative power of features or feature sets tells how well they can discriminate different classes. Feature selection is usually done by examining the discriminating capability of the features.

3.1 Time Domain Features

3.1.1 Zero Crossing Rate

In case of discrete time signals, a zero crossing is said to occur if there is a sign difference between successive samples. The zero crossing rates (ZCR) are a simple measure of the frequency content of a signal. For narrow band signals, the average zero crossing rate gives a reasonable way to estimate the frequency content of the signal [10]. But for a broad band signal such as speech, it is much less accurate. However, by using a representation based on the short time average zero crossing rate, rough estimates of spectral properties can be obtained. In this expression, each pair of samples is checked to determine where zero crossings occur and then the average is computed over N consecutive samples. In Fig. 3 shows the Zero Crossing rate.

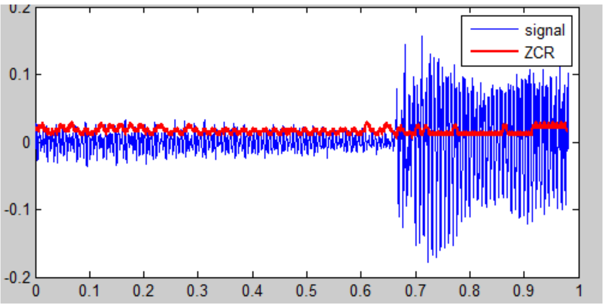

Fig. 3: Zero Crossing Rate

$$Zm = \sum_n |\text{sgn}\,[x(n)] - \text{sgn}\,[x(n-1)]|w(m-n) \quad (1)$$

Where the sgn function is

$$\text{sgn}[x(m)] = \begin{cases} 1, x(m) \geq 0 \\ -1, x(m) < 0 \end{cases} \quad (2)$$

and x(n) is the time domain signal for frame m.

Zero crossing rates proved to be useful in characterizing different audio signals and have been popularly used in speech/music classification problems. In general, speech signals are combined of alternating voices, sounds and unvoiced sounds in the syllable rate, in music signals does not have this kind of structure. Hence, compare to the speech signal, rate of zero crossing is greater than of music signals [6]. ZCR is a best discriminator between speech and music. Considering this, many systems have used ZCR for audio segmentation. A variation of the ZCR the high zero crossing rate ratios (HZCRR) are more discriminating than the exact value of ZCR.

3.1.2 Short Time Energy

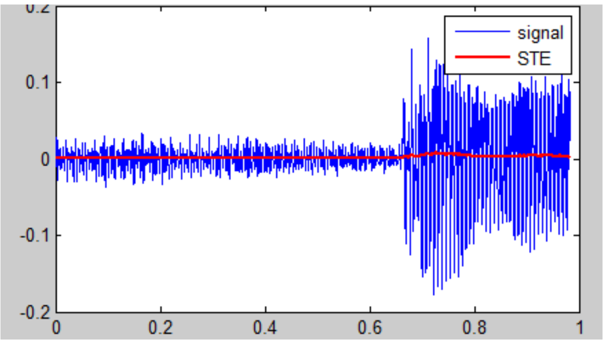

Fig. 4: Short Time Energy.

Short Time Energy (STE) is used in different audio classification problems. In speech signals, it provides a basis for distinguishing voiced speech segments from unvoiced ones. In case of very high quality speech, the short term energy features are used to distinguish speech from silence. In Fig. 4 shows the Short Time Energy. The energy E of a discrete time signal x(n) is defined by the expression [2].

$$E = \sum_{n=-\infty}^{\infty} x^2\,(n) \quad (3)$$

The amplitude of an audio signal varies with time. A convenient representation that reflects these amplitude variations is the short time energy of the signal. In general, the short time energy is defined as follows.

$$E_m = \sum_n [x(n)w(m-n)]^2 \quad (4)$$

The above expression can be rewritten as

$$E_m = \sum_n x\,(n)^2 h(m-n) \quad (5)$$

Where h(m) = w2(m). The term h(m) is interpreted as the impulse response of a linear filter. The choice of the impulse response, h(n) determines the nature of the short time energy representation. Short time energy of the audible sound is in general significantly higher than that of silence segments. In some of the systems, the Root Mean Square (RMS) of the amplitude is used as a feature for segmentation. It can be used as the measurement to distinguish audible sounds from silence when the SNR (signal to noise ratio) is high and its change pattern over time may reveal the rhythm and periodicity properties of sound. These are the major reasons for using STE in segmenting audio streams of various sounds and categories [1].

3.2 Frequency Domain Features

3.2.1 Spectral Centroid

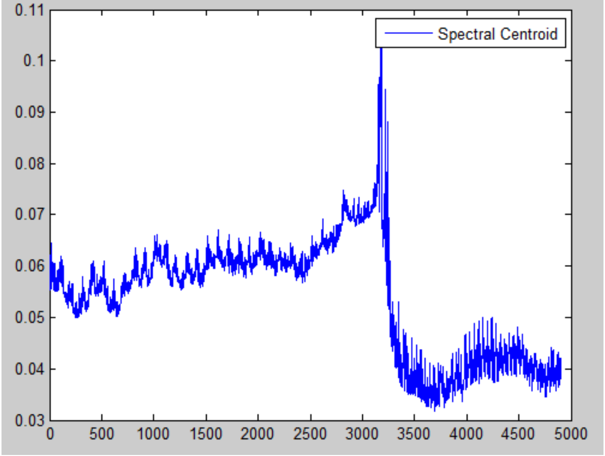

Fig. 5: Spectral Centroid.

The spectral centroid is a measure used in digital signal processing to characterize a spectrum. It shows where the center of mass of the spectrum. In Fig. 5 shows the Spectral Centroid. The centroid is calculated as the weighted mean of the frequencies present in the signal, which is determined using a Fourier transform, with their magnitudes as the weights

$$centroid = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \qquad (6)$$

Where *x(n)* represents the weighted frequency value, or magnitude, of bin number *n*, and *f(n)* represents the center frequency of that bin. This is a different statistic, the difference being essentially the same as the difference between unweight median and mean statistics. Both are measures of central tendency, in some situations they will exhibit some similarity of behavior. But since typical audio spectra are not normally distributed, the two measures will often give strongly different values. The spectral centroid is a high level predictor of the brightness of a sound, it is commonly used in digital audio processing for an automatic measure of musical timbre.

3.2.2 Spectral Flux

Spectrum flux (SF) is defined as the average variation value of spectrum between two adjacent frames in a given clip. Speech signals are composed of alternating voiced sounds and unvoiced sounds, while music signals do not have this kind of structure. Hence, for speech signal, its spectrum flux will be in general greater than that of music. The spectrum flux of environmental sounds is among the highest, and changes more dramatically than those of speech and music. This feature is useful for discriminating some strong periodicity environment sounds such as tone signal, from music signals [11].

$$\mathrm{SF} = \frac{1}{(\mathrm{N}-1)(\mathrm{k}-1)} \sum_{\mathrm{n}=1}^{\mathrm{N}-1} \sum_{\mathrm{k}=1}^{\mathrm{k}-1} [\log \mathrm{A}(\mathrm{n},\mathrm{k}) - \log \mathrm{A}(\mathrm{n}-1,\mathrm{k})]^2 \qquad (7)$$

Where A(n,k) is the discrete Fourier transform of the nth frame of input signal.

$$\mathrm{A}(\mathrm{n},\mathrm{k}) = \left| \sum_{\mathrm{m}=-\infty}^{\infty} \mathrm{x}(\mathrm{m}) \mathrm{w}(\mathrm{nL}-\mathrm{m}) \mathrm{e}^{\mathrm{j}\frac{2\pi}{\mathrm{L}}\mathrm{Km}} \right| \qquad (8)$$

x(m) is the original audio data, *w(m)* is the window function, L is the window length, K is the order of Discrete Fourier Transform (DFT), and N is the total number of frames. In a variation of the feature i.e. variance of the spectrum flux and variance of ZCR are used.

3.2.3 Spectral Roll off

As the Spectral Centroid, the Spectral Roll off is also a representation of the spectral shape of a sound and they are strongly correlated. It's defined as the frequency where 85% of the energy in the spectrum is below that frequency. If K is the bin that fulfills

$$\sum_{n=0}^{k} x(n) = 0.85 \sum_{n=0}^{N-1} x(n) \qquad (9)$$

Then the Spectral Roll off frequency is *f(K)*, where *x(n)* represents the magnitude of bin number *n*, and *f(n)* represents the center frequency of that bin.

3.2.4 Spectral Entropy

The spectral entropy is the quantitative measure of the spectral disorder. The entropy has been used to detect silence and voiced region of speech in voice activity detection. The discriminatory property of this feature gives rise to its use in speech recognition. The entropy can be used to capture the formants or the peakness of a distribution. Formants and their locations have been considered to be important for speech tracking [15].

$$E = -\sum_{i=0}^{L-1} n_i * \log_2(ni) \qquad (10)$$

3.3 Discrete Wavelet Transform

The Discrete Wavelet Transform (DWT), which is based on sub-band coding, is found to yield a fast computation of Wavelet Transform. It is easy to implement and reduces the computation time and resources required. The foundations of DWT go back to 1976 when techniques to decompose discrete time signals were devised [19]. Similar work was done in speech signal coding which was named as sub-band coding. In 1983, a technique similar to sub-band coding was developed which was named pyramidal coding. Later many improvements were made to these coding schemes which resulted in efficient multi-resolution analysis schemes. In DWT, a time-scale representation of the digital signal is obtained using digital filtering techniques. The signal to be analyzed is passed through filters with different cutoff frequencies at different scales. Filters are one of the most widely used signal processing functions.

The wavelet analysis process is to implement a wavelet prototype function, known as analyzing wavelet or mother wavelet. Coefficients in a linear combination of the wavelet function can be used in order to represent the development of the original signal in terms of a wavelet, data operations can be performed with the appropriate wavelet coefficients. Choose the best wavelets adapted to represent your data, also truncate the coefficients below a threshold [20].

Wavelets can be realized by iteration of filters with rescaling. The resolution of the signal, which is a measure of the amount of detail information in the signal, is determined by the filtering operations, and the scale is determined by up sampling and down sampling (sub sampling) operations [19]. The DWT is computed by successive low pass and high pass filtering of the discrete time-domain signal. This is called the Mallat algorithm or Mallat-tree decomposition. Its significance is in the manner it connects the continuous-time multi resolution to discrete-time filters. At each level, the high pass filter produces detail information d[n], while the low pass filter associated with scaling function produces coarse approximations, a[n]. The Discrete Wavelet Transform (DWT) is a special case of the WT that provides a compact representation of a signal in time and frequency that can be computed efficiently.

The DWT is defined by the following equation:

$$\mathrm{W(j,k)} = \sum\sum \mathrm{x(k)e^{-\frac{j}{2}}\Psi(2^{-j}n-k)} \qquad (11)$$

The $\Psi(\mathrm{t})$ is a time function with finite energy. The DWT can be analyzing using a fast algorithm related to multi rate filter banks.

4 GMM BASED CLASSIFICATION MODEL

There are many techniques for classifying audio samples into multiple classes. Classification algorithms are divided into supervised and unsupervised algorithms. In a supervised classification, a labeled set of training samples is used to train the algorithm whereas in the case of an unsupervised classification the data is grouped into some clusters without the use of labeled training set. Parametric and non-parametric classification is another way of categorizing classification algorithms. The functional form of the probability density of the feature vectors for each class is known in parametric methods. In non-parametric methods, on the other hand, no specific functional form is assumed in advance, instead, the probability density is rather approximated locally based on the training data. The Gaussian mixture model (GMM) is used in classifying different audio classes. The Gaussian classifier is an example of a parametric classifier [18]. It is an intuitive approach when the model consists of several Gaussian components, which can be seen to model acoustic features. In classification, each class is represented by a GMM and refers to its model. Once the GMM is trained, it can be used to predict which class a new sample probably belongs to. A variety of approaches to the problem of mixture decomposition have been proposed, many of which focus on maximum likelihood methods such as expectation maximization (EM) or maximum a posteriori estimation (MAP). Generally these methods consider separately the question of parameter estimation and system identification, that is to say a distinction is made between the determination of the number and functional form of components within a mixture and the estimation of the corresponding parameter values [14].

4.1 Expectation Maximization (EM)

Expectation maximization (EM) is seemingly the most popular technique used to determine the parameters of a mixture with an a priori given number of components [13]. This is a particular way of implementing maximum likelihood estimation for this problem. EM is of particular appeal for finite normal mixtures where closed-form expressions are possible such as in the following iterative algorithm by Dempster et al. (1977) The Expectation-maximization algorithm can be used to compute the parameters of a parametric mixture model distribution. It is an iterative algorithm with two steps: an expectation step and maximization step [7]. The expectation step with initial guesses for the parameters of our mixture model, ”partial membership” of each data point in each constituent distribution is computed by calculating expectation values for the membership variables of each data point [12]. That is, for each data point xi and distribution Y_i, the membership value $Y_{i,j}$ is

$$y_{i,j} = \frac{a_i f_Y(x_j;\theta_i)}{f_X(x_j)} \qquad (12)$$

The maximization step with expectation values in hand for group membership, plug-in estimates are recomputed for the distribution parameters. The mixing coefficients ai are the means of the membership values over the N data points.

$$a_i = \frac{1}{N}\sum_{j=1}^{N} y_{i,j} \qquad (13)$$

The component model parameters θ_i are also calculated by expectation maximization using data points x_j that have been weighted using the membership values. For example, if θ is a mean µ

$$\mu_i = \frac{\sum_j y_{i,j} x_j}{\sum_j y_{i,j}} \qquad (14)$$

With new estimates for a_i and the θ'_i, the expectation step is repeated to recomputed new membership values. The entire procedure is repeated until model parameters converge.

4.2. GMM with Fuzzy c-Means

One of the most widely used fuzzy clustering algorithms is the Fuzzy C-Means (FCM) Algorithm (Bezdek 1981). The FCM algorithm attempts to partition a finite collection of n elements X = x1, ..., xn into a collection of c fuzzy clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of c cluster centres C = c1, ..., cc and a partition matrix, W = wi, j [0, 1], i = 1, ..., n; j = 1, ..., c each element wij tells the degree to which element xi belongs to cluster cj . Like the k-means algorithm, the FCM aims to minimize an objective function [3]. The standard function is:

$$w_k(x) = \frac{1}{\sum_j \left(\frac{d\left(center_{k,x}\right)}{d\left(center_{j,x}\right)}\right)^{2} / (n-1)} \qquad (15)$$

which differs from the k-means objective function by the addition of the membership values u_{ij} and the fuzzifier m. The fuzzifier m determines the level of cluster fuzziness. A large m results in smaller memberships $\mathrm{w_{ij}}$ and hence, fuzzier clusters [8]. In the limit m = 1, the memberships wij converge to 0 or 1, which implies a crisp partitioning. In the absence of experimentation or domain knowledge, m is commonly set to 2. The basic FCM Algorithm, given n data points (x1, ..., xn) to be clustered, a number of c clusters with (c1, ..., cc) the center of the clusters, and m the level of cluster fuzziness with, Fuzzy c-means (FCM) is a data clustering technique in which a dataset is grouped into n clusters with every data point in the dataset belonging to every cluster to a certain degree [17]. For example, a certain data point that lies close to the center of a cluster will have a high degree of belonging or membership to that cluster and another data point that lies far away from the center of a cluster will have a low degree of belonging or membership to that cluster. In fuzzy clustering, each point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster [16]. Thus, points on the edge of a cluster, may be in the cluster to a lesser degree than points in the center of cluster. An overview and comparison of different fuzzy clustering algorithms is available. Any point x has a set of coefficients giving the degree of being in the k^{th} cluster $w_k(x)$. With fuzzy c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$C_k = \frac{\sum_x w_k(x)x}{\sum_x w_k(x)} \qquad (16)$$

The degree of belonging, $w_k(x)$, is related inversely to the distance from x to the cluster center as calculated on the previous pass. It also depends on a parameter m that controls how much weight is given to the closest center. The fuzzy c-means algorithm is very similar to the k-means algorithm:

```
1. Choose a number of clusters.
2. Assign randomly to each point coefficients for being in the clusters.
3. Repeat until the algorithm has converged (that is, the coefficients' change between two iterations is no more than , the given sensitivity threshold) :
      a) Compute the centroid for each cluster, using the formula above.
      b) For each point, compute its coefficients of being in the clusters, using the formula above.
```

The algorithm minimizes intra-cluster variance as well, but has the same problems as k-means; the minimum is a local minimum, and the results depend on the initial choice of weights. Using a mixture of Gaussians along with the expectation-maximization algorithm is a more statistically formalized method which includes some of these ideas: partial membership in classes.

5 EXPERIMENTAL RESULTS

5.1 Dataset

The speech and music audio data are recorded various sources namely 300 clips of speech and 300 clips of music. Each clip consists of audio data ranging from one second to about ten seconds, with a sampling rate of 8 kHz, 16-bits per sample, monophonic, and 128 kbps audio bit rate. The waveform audio format is converted into raw values i.e. 8000 sample values per second.

5.2 Feature Extraction

Six set of features and DWT feature is extracted from each frame of the audio by using the feature extraction techniques. Here the low level features both time domain and frequency domain features are taken. The time domain features are ZCR and STE, the frequency domain features are spectral centroid, spectral flux spectral roll-off and spectral entropy. DWT using multi rate filter banks feature will be calculated for the given wav file. The above process is continued for 600 wav files. The feature values for all the wav files will be stored separately for speech and music.

5.3 Classification based on GMM

GMM using EM algorithm to determine parameters. In this work Fuzzy c-means algorithm is used to determine the mean centers.

Training:

Step1:Determine mean centers using Fuzzy c-means algorithm.
Step2:Compute the distance matrix for each feature vector to the centroids.
Step3:Assign the feature vectors to the nearest centroids.
Step4:Grouping is done based on the minimum distance.
Step5:Compute Covariance matrix for the feature vectors belonging to the corresponding groups.
Step6:Compute probability density function for all the feature vectors.
Step7:Fit Gaussians using the centroids and covariance matrices.

Testing:
Step1: Assignment of the feature vectors is done based on the maximum likelihood selection.

The performance of the system for 2, 5 and 10

5.4 Performance measures

Sensitivity and Specificity are statistical measures used for studying the performance of classification. Sensitivity measures the proportion of actual positives which are correctly identified.

$$\text{Sensitivity } = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

Specificity measures the proportion of negatives which are correctly identified

$$\text{Specificity } = \frac{True\ Negative}{(False\ Positive + True\ Negative)}$$

Table 1 shows the sensitivity and specificity of speech and music for experiment conducted.

Table 1: The Sensitivity and Specificity

Performance Measures	Accuracy
Sensitivity	74 %
Specificity	88 %

Table 2: Performance of GMM for different Gaussian mixtures

GMM	2 mixtures	5 mixtures	10 mixtures
Speech	83 %	91 %	85 %
Music	82 %	91 %	83 %

The performance of different Gaussian mixtures are shown in Table 2. The distribution of the acoustic features is captured using GMM. The class to which the speech and music sample belongs is decided based on the highest output. Table 2 shows the performance of GMM for speech and music classification based on the number of mixtures. The number of Gaussian mixtures is increased from 2 to 10. The performance in terms of classification accuracy is studied. When the number of mixtures is 2, the performance is very low. When the mixtures are increased from 2 to 5, the classification performance slightly increases. When the number of mixtures varies from 5 to 10, there is no considerable increase in the performance and the maximum performance is achieved. There is no considerable increase in the performance when the number of mixtures is above 10. With GMM, the best performance is achieved with 5 Gaussian mixtures.

Experiments were conducted to test the performance of the system using EM algorithm. In this work, GMM modeled using Fuzzy c-means gave better performance compared to EM algorithm. Fig. 5 and 6 shows the performance of audio classification using GMM-EM and GMM-Fuzzy c-means for different duration respectively.

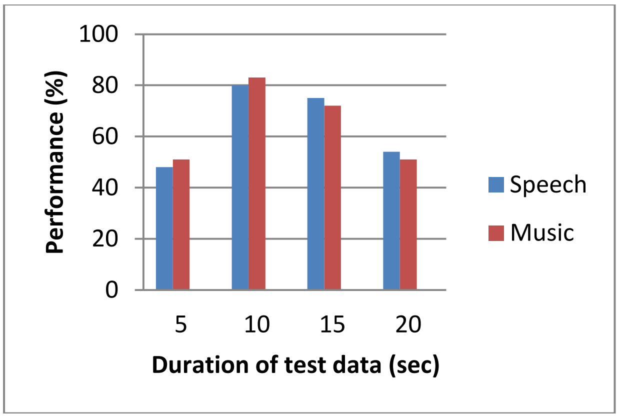

Fig. 5: Performance of audio classification for different duration of speech and music clips using GMM-EM

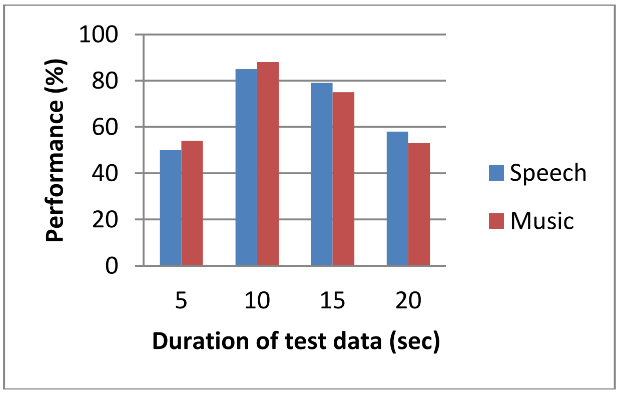

Fig. 6: Performance of audio classification for different duration of speech and music clips using GMM-Fuzzy c-means.

6 CONCLUSION

In this paper, six new feature vectors and Discrete Wavelet Transform features for the classification of speech and music files is presented. Further it is possible to improve the classification accuracy by using different types of domain based features together. First of all, we perform feature extraction technique to extract the features from the speech and music files for classification. The proposed classification method is implemented using fuzzy c-means based clustering algorithm approach to fit the GMM parameters for classification. The parameters are possible only due to mixture model of each sample is said to belong to a cluster only within certain probability. The average speech and music classification accuracy rate of the proposed method higher than GMM using EM algorithm. The overall accuracy of proposed method GMM using Fuzzy c-means is 91%. It shows that the proposed method can achieve better classification accuracy than other approaches. As the classification accuracy is high, this method can retrieve a data more effectively from a large database.

7 REFERENCES

[1] Arijit Ghosal BCD, Saha SK (2011) 'Speech/music classification using empirical mode decomposition', Second International Conference on Emerging Applications of Information Technology , pp 4952.

[2] Breebaart J, McKinney M(2003) ' Features for audio classification. ', IntConf on MIR

[3] Dat Tran TV, Wagner M (1998) ' Fuzzy Gaussian mixture models for speaker recognition',Proceedings of the International Conference on Spoken Language Processing , vol. 2, pp 759762.

[4] Changsheng Xu NCM, Shao X(2005) ' Automatic music classification and summarization. ',IEEE Trans Speech and Audio Processing , vol. 13, pp 441450.

[5] Chungsoo Lim Mokpo YWL, Chang JH (2012) ' New techniques for improving the practicality of an svm-based speech/music classifier.',Acoustics, Speech and Signal Processing (ICASSP) , pp 1657-1660.

[6] F Gouyon FP, Delerue O(2000) ' Classifying percussive sounds: a matter of zero crossing rate. ', Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00) Verona, Italy .

[7] H Watanabe SM, Kikuchi H (2010) ' Interval calculation of em algorithm for gmm parameter estimation',Proceedings of 2010 IEEE International Symposium ,pp 2686-2689.

[8] Joanna Czajkowska MB, Pietka E (2012) ' Kernelized fuzzy c-means method and gaussian mixture model in unsupervised cascade clustering', Information Technologies in Biomedicine , pp 58-66.

[9] Lim C, J-H(2012) ' Enhancing support vector machine-based speech/music classifica-tion using conditional maximum a posteriori criterion.',Signal Processing, IET ,vol. 64, pp 335-340.

[10] Panagiotakis C, Tziritas G (2005) ' A speech/music discriminator based on rms and zero-crossings.', IEEE Trans Multimedia .

[11] Peeters G(2004) ' A large set of audio features for sound description. ', tech rep, IRCAM .

[12] Redner R, Walker H (1984) 'Mixture densities, maximum likelihood and the emalgorithm. ', SIAM Review .

[13] Reynolds D (1993) ’A gaussian mixture modeling approach to text-independent speaer identification’,Intl. Technical Report 967 .

[14] Sourabh Ravindran KS, Anderson DV (2005) ’A physiologi-cally inspired method for audio classification. ’, Journal on Applied Signal Processing.vol. 9,pp. 1374-1381

[15] Toru Taniguchi MT, Shirai K(2008) ’ Detection of speech and music based on spectral tracking. ’, Speech Communication .vol. 50, pp. 547-563

[16] Tran D, Wagner M (1998) ’ Fuzzy gaussian mixture models for speaker recognition’, Special issue of the Australian Journal of Intelligent InformationProcessing Systems .vol. 5, No. 2, pp.293-300

[17] Tran D, Wagner M (1999) ’ Fuzzy approach to gaussian mixture models and gener-alised gaussian mixture models’, Proceedings of the Computation Intel-ligence Methods and Applications .pp 154-158

[18] Ziyou Xiong AD Regunathan Radhakrishnan, SHuang T(2004) ’ Effec-tive and efficient sports highlights extraction using the minimum description length criterion in selecting gmm structures. ’, IEEE Intl Conf Multimedia and Ex .pp. 1947-1950.

[19] Chun-Lin, Liu, A Tutorial of the Wavelet Transform, February 23, 2010

[20] Siwar Rekik, Driss Guerchi, Habib Hamam & Sid-Ahmed Selouani,” Audio Steganography Coding Using the Discrete Wavelet Transforms”, International Journal of Computer Science and Security, Volume .6 Issue .1, pp. 79-83, 2012.

8 AUTHOR’S DETAILS

R. Thiruvengatanadhan received his Bachelor's degree in Computer Science and Engineering from Annamalai University, Chidambaram in the year 2004. He received his M.E degree in Computer Science and Engineering from Annamalai University, Chidambaram. He is pursuing his Ph.D in Computer Science and Engineering from Annamalai University, Chidambaram. He joined the services of Annamalai University in the year 2006 as a faculty member and is presently serving as Assistant Professor in the Department of Computer Science &Engg. His research interests include audio signal processing, speech processing, image processing and pattern classification.

Dr. P. Dhanalakshmi received her Bachelor's degree inComputer Science and Engineering from Government College of Technology, Coimbatore in the year 1993. She received her M.Tech degree in Computer Applications from the reputed Indian Institute of Technology, New Delhi under the Quality Improvement Programme in the year 2003. She completed her Ph. D in Computer Science and Engineering from Annamalai University in the year 2011. She joined the services of Annamalai University in the year 1998 as a faculty member and is presently serving as Associate Professor in the Department of Computer Science &Engg. She has published 11 papers in international conferences and journals. She is guiding several students who are pursuing doctoral research. Her research interests include speech processing, image and video processing, pattern classification and neural networks.