Journal of Computer Science 10 (1): 34-44, 2014
ISSN: 1549-3636

doi:10.3844/jcssp.2014.34.44 Published Online 10 (1) 2014 (http://www.thescipub.com/jcs.toc)

SPEECH/MUSIC CLASSIFICATION USING WAVELET BASED FEATURE EXTRACTION TECHNIQUES

Thiruvengatanadhan Ramalingam and P. Dhanalakshmi

Department of Computer Science and Engineering,
Faculty of Engineering and Technology, Annamalai University, Annamalainagar, Tamil Nadu, India

Received 2013-10-02, Revised 2013-10-08; Accepted 2013-11-08

ABSTRACT

Audio classification serves as the fundamental step towards the rapid growth in audio data volume. Due to the increasing size of the multimedia sources speech and music classification is one of the most important issues for multimedia information retrieval. In this work a speech/music discrimination system is developed which utilizes the Discrete Wavelet Transform (DWT) as the acoustic feature. Multi resolution analysis is the most significant statistical way to extract the features from the input signal and in this study, a method is deployed to model the extracted wavelet feature. Support Vector Machines (SVM) are based on the principle of structural risk minimization. SVM is applied to classify audio into their classes namely speech and music, by learning from training data. Then the proposed method extends the application of Gaussian Mixture Models (GMM) to estimate the probability density function using maximum likelihood decision methods. The system shows significant results with an accuracy of 94.5%.

Keywords: Audio Classification, Feature Extraction, Wavelet Transform, Support Vector Machine (SVM), Gaussian Mixture Model (GMM)

1. INTRODUCTION

The term audio is used to indicate all kinds of audio signals, such as speech, music as well as more general sound signals and their combinations. Multimedia databases or file systems can easily have thousands 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. 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. These compact descriptors are of great use not only in audio storage and retrieval applications, but also in efficient content-based segmentation, classification, recognition, indexing and browsing of data.

The music signal is a special class in the signal category that has its own characteristics different from the speech signal in many ways. First of all, music normally has a wide range frequency distribution among the audible range of human, from 0 to 20k Hz. The bandwidth of the speech signal is usually limited into 50 Hz to 7 k Hz and hence, the spectral centroids of music signal are higher than that of the speech. In addition, for considering time-domain characteristics, musical signal usually has a lower silence ratio except that it is sung by a singer or played on a solo instrument only. Compared to an ordinary speech signal, music has lower variability in zero-crossing rate [base]. Besides, music has normally more harmonic than other sound. Therefore, music has higher harmonic than speech. Music usually has regular beats that can be extracted to differentiate it from speech for the sake of the melody and background noise.

The problem of distinguishing speech signals from other audio signals (e.g., music) has become increasingly important as automatic speech recognition systems are applied to more real-world multimedia domains, such as the automatic transcription of broadcast news, in which

Corresponding Author: Thiruvengatanadhan Ramalingam, Department of Computer Science and Engineering,
Faculty of Engineering and Technology, Annamalai University, Annamalainagar, Tamil Nadu, India

speech is typically interspersed with segments of music and other background noise (Ghosal and Saha, 2011).

These Speech/music mixtures appear quite often in radio and television programmes. Movies, infotainment productions and commercials contain speech, music, sound effects and background sounds. Especially in commercials these signal classes appear often in a mixed and fast changing manner (Kim *et al*., 2012).

A variety of systems for audio segmentation or classification have been proposed in the past and many features such as Root Mean Square (RMS), Zero Crossing Rate (ZCR) (Khan *et al*., 2012), low frequency modulation (Golumbic *et al*., 2012), entropy and dynamism features (Krajewski *et al*., 2012) have been used.

The wavelets are suitable the tools for Speech/Music classification because they have ability to deal with non-stationary signals such as music and speech, analyze the signals in different scales and achieve variable time-frequency localization (Sumithra *et al*., 2011).

In this study, three different types of wavelet transform based features have been extracted. These characteristics are modeled using probability density function estimation which is called Gaussian Mixture Models (GMM). SVM is used to classify the audio signal into speech and music.

1.1. Related Work

During the recent years, there have been many studies on automatic audio classification and segmentation using several features and techniques. The most common problem in audio classification is speech/music classification, in which the highest accuracy has been achieved, especially when the segmentation information is known beforehand. An audio feature extraction and a multi-group classification scheme that focuses on identifying discriminatory time-frequency subspaces using the Local Discriminate Bases (LDB) technique has been described in (Mishra and Agrawal, 2012). For pure music and vocal music, a num-ber of features such as LPC and LPCC are extracted in (Nagavi and Bhajantri, 2012) to characterize the music content. Based on calculated features, a clustering algorithm is applied to structure the music content.

A new approach towards high performance speech/music discrimination on realistic tasks related to the automatic transcription of broadcast news is described in (Frikha and Hamida, 2012), in which an Artificial Neural Network (ANN) and HIDDEN Markov Model (HMM) are used. Subashini *et al*. (2012), a generic audio classification and segmentation approach for multimedia indexing and retrieval is described. A method is proposed in (Sporka *et al*., 2012) for speech/music discrimination based on root mean square and zero-crossings. The method proposed in (Jiang *et al*., 2013), investigates the feasibility of an audio-based context recognition system where simplistic low dimensional feature vectors are evaluated against more standard spectral features. Using discriminative training, competitive recognition accuracies are achieved with very low-order Hidden Markov models.

Feki *et al*. (2012) a speech/music discrimination system was proposed based on Mel-Frequency Cepstral Coefficient (MFCC) and GMM classifier. This system can be used to select the optimum coding scheme for the current frame of an input signal without knowing a priori whether it contains speech-like or music-like characteristics.

The classification of continuous general audio data for content-based retrieval was addressed in (Liu, 2010). The DWT is computed by successive low pass and high pass filtering of the discrete time-domain signal which extracts features that characterize their spectral change over time.

An approach given in (Theodorou *et al*., 2012) uses Support Vector Machine (SVM) for audio scene classification, which classifies audio clips into one of five classes: Pure speech, non-pure speech, music, environment sound and silence.

Audio classification techniques for speech recognition and audio segmentation, for unsupervised multi speaker change detection are proposed in (Abdolali and Sameti, 2012). Two new extended-time features: Variance of the Spectrum Flux (VSF) and Variance of the Zero-Crossing Rate (VZCR) are used to pre-classify the audio and supply weights to the output probabilities of the GMM networks. The classification is then implemented using weighted GMM networks.

1.2. Outline of the Work

In this study, automatic audio feature extraction and classification approaches are presented. In order to discriminate the speech and music features such as Discrete Wavelet Transform are extracted to characterize the audio content. Support Vector Machine (SVM) is applied to obtain the optimal class boundary between the classes by learning from training data. The performance of SVM is compared to GMM using maximum likelihood decision methods. Experimental results show that the classification accuracy of GMM with DWT features can provide a better result. **Figure 1** illustrates the block diagram of Speech/Music classification system.

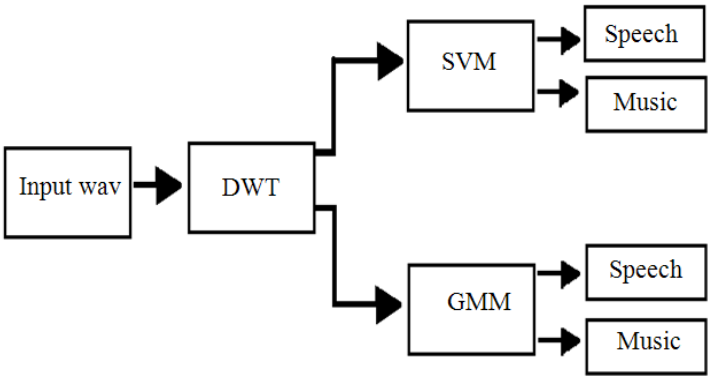

Fig. 1. Block Diagram for speech/music classification

2. FEATURES FOR SPEECH/MUSIC DISCRIMINATION

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 discriminative capability of the features. The performance of a set of features depends on the application. The design of descriptive features for a specific application is hence the main challenge in building audio classification systems.

In section 2, the related theoretical background on the features used for music/speech discrimination systems will be given briefly.

2.1. 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 (Liu, 2010). 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 (Rekik *et al*., 2012).

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 (Patil and Ruikar, 2012). The DWT is computed by successive low pass and high pass filtering of the discrete time-domain signal as shown in **Fig. 2**. 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. In the figure, the signal is denoted by the sequence x[n], where n is an integer. The low pass filter is denoted by G_0 while the high pass filter is denoted by H_0.

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 (1):

$$W(j,k) = \sum\sum x(k)e^{-\frac{j}{2}}\Psi(2^{-j}n-k) \quad (1)$$

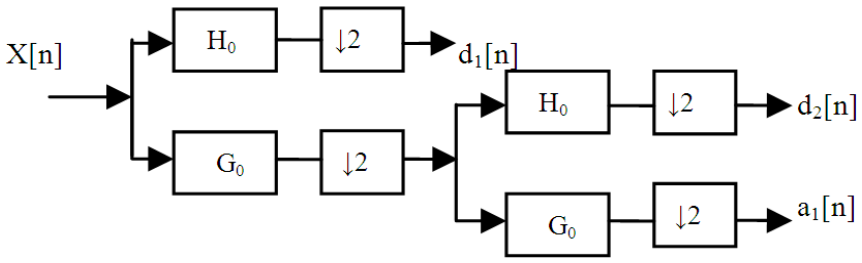

Fig. 2. Two level wavelet decomposition technique

where, Ψ(t) is a time function with finite energy and fast decay called the mother wavelet. The DWT analysis can be performed using a fast, pyramidal algorithm related to multi rate filter banks.

As a multi rate filter bank the DWT can be viewed as a constant Q filter bank with octave spacing between the centers of the filters. Each sub band contains half the samples of the neighboring higher frequency sub band. In the pyramidal algorithm the signal is analyzed at different frequency bands with different resolution by decomposing the signal into a coarse approximation and detail information. The coarse approximation is then further decomposed using the same wavelet decomposition step. This is achieved by successive high pass and low pass filtering of the time domain signal and is defined by the following Equation (2 and 3):

$$y_{high}[k] = \sum_n x[n]g[2k - n] \quad (2)$$

$$y_{low}[k] = \sum_n x[n]h[2k - n] \quad (3)$$

where, y_{high} [k], y_{low}[k] are the outputs of the high pass (g) and low pass (h) filters, respectively after sub sampling by 2. Because of the down sampling the number of resulting wavelet coefficients is exactly the same as the number of input points. A variety of different wavelet families have been proposed in the literature. In our implementation, the 4 coefficient wavelet family (DAUB4) proposed by Daubechies is used.

3. CLASSIFICATION MODEL

3.1. Support Vector Machine

SVM have the potential to handle very large feature spaces, because the training of SVM is carried out so that the dimension of classified vectors does not have as a distinct influence on the performance of SVM as it has in the conventional classifier (Lazouni *et al*., 2013). This will also benefit in classification of transient phenomena in power transformer, because the number of features to be the basis for classification of transient events may not have to be limited. Also SVM based classifiers are claimed to have good generalization properties compared to conventional classifiers, because in training the SVM classifier, the structural miscellaneous risk is to be minimized, whereas traditional classifiers are usually trained so that the empirical risk is minimized.

Support Vector Machine (SVM) is very effective method for general purpose pattern recognition. Given a set of points which belong to either of two classes, a SVM finds the hyperplane leaving the largest possible fraction of points of the same class on the same side, while maximizing the distance of either class from the hyper plane. SVMs perform pattern recognition between two classes by finding a decision surface that has maximum distance to the closest points in the training set which are termed support vectors. Principle of SVM is, where there are many possible linear classifiers that can separate the data, there is only one that maximizes the difference between. SVMs are particular classifiers that are based on the margin-maximization principle (Kapp *et al*., 2012). A powerful machine learning technique for data classification, SVM performs an implicit mapping of data into a higher (maybe infinite) dimensional feature space and then finds a linear separating hyper plane with the maximal margin to separate data in this higher dimensional space (Bhavsar and Panchal, 2012).

A SVM constructs a hyper plane or set of hyper planes in a high or infinite-dimensional space, which are used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class (Lim *et al*., 2012). To keep the computational load reasonable, the mappings used by SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function selected to suit the problem (Suresha *et al*., 2012).

SVM constructs a linear model to estimate the decision function using non-linear class boundaries based on support vectors. If the data are linearly separated, SVM trains linear machines for an optimal hyper plane that separates the data without error and into the maximum distance between the hyper plane and the closest training points. The training points that are closest to the optimal separating hyper plane are called support vectors.

Figure 3 shows the architecture of the SVM. SVM maps the input patterns into a higher dimensional feature space through some nonlinear mapping chosen a priori. A linear decision surface is then constructed in this high dimensional feature space. Thus, SVM is a linear classifier in the parameter space, but it becomes a non-linear classifier as a result of the non-linear mapping of the space of the input patterns into the high dimensional feature space.

For linearly separable data, SVM finds a separating hyper plane which separates the data with the largest margin. For linearly inseparable data, it maps the data in the input space into a high dimension space $x \in R^I \rightarrow \varnothing(x) \in R^H$ with kernel function $\varnothing(x)$, to find the separating hyper plane. An example for SVM kernel function $\varnothing(x)$ maps 2-Dimensional input space to higher 3-Dimensional feature space as shown in **Fig. 3**. SVM was originally developed for two class classification problems. The N class classification problem can be solved using N SVMs. Each SVM separates a single class from all the remaining classes (Lim and Lim, 2012).

SVM generally applies to linear boundaries. In the case where a linear boundary is in appropriate SVM can map the input vector into a high dimensional feature space. By choosing a non-linear mapping, the SVM constructs an optimal separating hyper plane in this higher dimensional space, as shown in **Fig. 4**. The function K is defined as the kernel function for generating the inner products to construct machines with different types of non-linear decision surfaces in the input space Equation (4):

$$K(x, x_i) = \varnothing(x).\varnothing(x)_i \quad (4)$$

The kernel function may be any of the symmetric functions that satisfy the Mercer's conditions (Brunner *et al.*, 2012). There are several SVM kernel functions are.

3.1.1. Gaussian Kernel

The Gaussian kernel is an example of radial basis function kernel Equation (5):

$$K(x, x_i) = \exp\left(-\frac{|x - x_i|^2}{2\sigma^2}\right) \quad (5)$$

Alternatively, it could also be implemented using Equation (6):

$$K(x, x_i) = \exp(-\gamma | x - x_i |^2) \quad (6)$$

The adjustable parameter sigma plays a major role in the performance of the kernel and should be carefully tuned to the problem at hand. If overestimated, the exponential will behave almost linearly and the higher-dimensional projection will start to lose its non-linear power. In the other hand, if underestimated, the function will lack regularization and the decision boundary will be highly sensitive to noise in training data.

3.1.2. Sigmoidal Kernel

Sigmoidal kernel functions which aren't strictly positive definite also have been shown to perform very well in practice. Despite its wide use, it is not positive semi-definite for certain values of its parameters Equation (7):

$$\tanh(\beta_0 x^T x_i + \beta_1) \quad (7)$$

where, x_i is support vectors, $\beta 0$, $\beta 1$ are constant values.

3.1.3. Polynomial KERNEL

The Polynomial kernel is a non-stationary kernel. Polynomial kernels are well suited for problems where all the training data is normalized Equation (8):

$$K(x, x_i) = (\alpha x^T x_i + c)^d \quad (8)$$

Adjustable parameters are the slope alpha, the constant term c and the polynomial degree d.

The dimension of the feature space vector $\varnothing(x)$ for the polynomial kernel of degree p and for the input pattern dimension of d is given by Equation (9):

$$\frac{(p+d)!}{p!d!} \quad (9)$$

For sigmoidal kernel and Gaussian kernel, the dimension of feature space vectors is shown to be infinite. Finding a suitable kernel for a given task is an open research problem. Given a set of audio corresponding to N categories for training, N SVMs are trained.

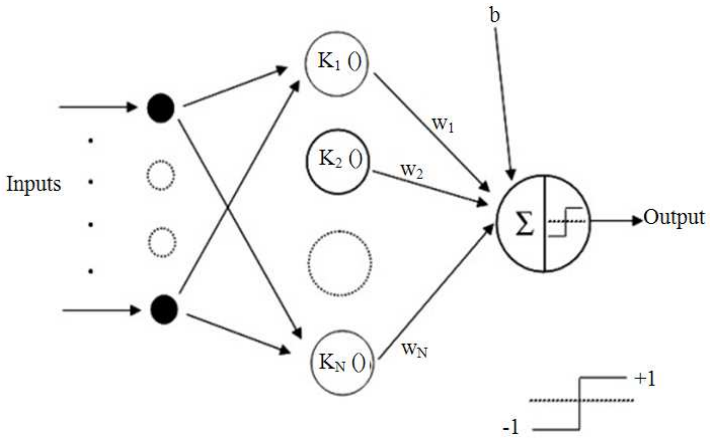

Fig. 3. Architecture of the SVM (Ns is the number of support vectors)

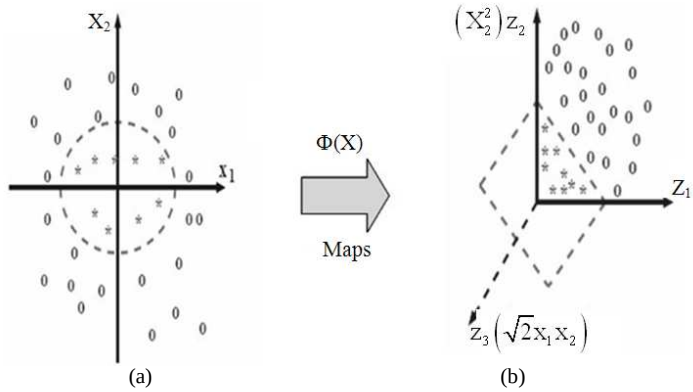

Fig. 4. An example for SVM kernel function Φ(x) maps 2-dimensional input space to higher 3-dimensional feature space. (a) Nonlinear problem. (b) Linear problem

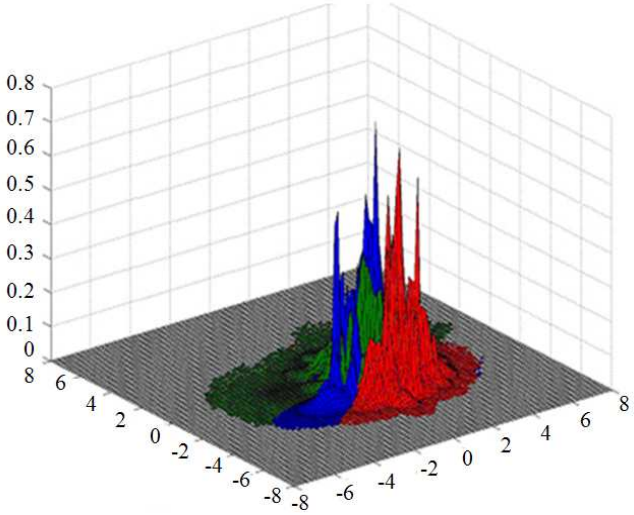

Fig. 5. Gaussian mixture models

Each SVM is trained to distinguish between one category and all other categories in the training set. During testing, the class label l of an audio x can be determined using Equation (10):

$$l = \begin{cases} n, & \text{if } d_n(x) + t > 0 \\ 0, & \text{if } d_n(x) + t \le 0 \end{cases} \quad (10)$$

where, $d_n(x) = \max\{d_i(x)\}_{i=1}^{N}$ and $d_i(x)$ is the distance from x to the SVM hyper plane corresponding to category i. The classification threshold is t and the class label l = 0 stands for unknown.

3.2. Gaussian Mixture Model

The Gaussian Mixture Model (GMM) is used in classifying different audio classes. The Gaussian classifier is an example of a parametric classifier. 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 (Xing *et al*., 2012).

The probability distribution of feature vectors is modeled by parametric or non-parametric methods. Models which assume the shape of probability density function are termed parametric. In non-parametric modeling, minimal or no assumptions are made regarding the probability distribution of feature vectors. The potential of Gaussian mixture models to represent an underlying set of acoustic classes by individual Gaussian components, in which the spectral shape of the acoustic class is parameterized by the mean vector and the covariance matrix, is significant.

Also, these models have the ability to form a smooth approximation to the arbitrarily-shaped observation densities in the absence of other information (Nidhyananthan and Kumari, 2013). With Gaussian mixture models, each sound is modeled as a mixture of several Gaussian clusters in the feature space. The basis for using GMM is that the distribution of feature vectors extracted from a class can be modeled by a mixture of Gaussian densities as shown in **Fig. 5**.

For a D dimensional feature vector x, the mixture density function for category s is defined as Equation (11):

$$P\left(\frac{x}{\lambda^s}\right) = \sum_{i=1}^{M} \alpha_i^s f_i^s(x) \quad (11)$$

The mixture density function is a weighted linear combination of m component uni-modal Gaussian densities $f_i^S(.)$. Each Gaussian density function $f_i^S(.)$ is parameterized by the mean vector is parameterized by the mean vector μ_i^S and the covariance matrix Σ_i^s using Equation 12:

$$f_i^s(x) = \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_i^s\right|}} \exp\left(-\frac{1}{2}\left(x-\mu_i^s\right)^T \left(\Sigma_i^s\right)^{-1}\left(x-\mu_i^s\right)\right) \quad (12)$$

where, $(\Sigma_i^S)^{-1}$ and $\left|\Sigma_i^S\right|$ denote the inverse and determinant of the covariance matrix Σ_i^s, respectively. The mixture weights $(\alpha_1^s, \alpha_2^s, ...\alpha_M^s,)$ satisfy the constraint $\sum_{i=1}^{M} \alpha_i^s = 1$. Collectively, the parameters of the model λ^s are denoted as $\lambda^s = \{\alpha_i^s, \mu_i^s, \Sigma_i^s\}$, i = 1,2,…M. The number of mixture components is chosen empirically for a given data set. The parameters of GMM are estimated using the iterative expectation-maximization algorithm.

The motivation for using Gaussian densities as the representation of audio features is the potential of GMMs to represent an underlying set of acoustic classes by individual Gaussian components in which the spectral shape of the acoustic class is parameterized by the mean vector and the covariance matrix. Also, GMMs have the ability to form a smooth approximation to the arbitrarily shaped observation densities in the absence of other information. With GMMs, each sound is modeled as a mixture of several Gaussian clusters in the feature space.

GMMs model the distribution of feature vectors. For each class, assume the existence of a probability density function expressible as a mixture of a number of multidimensional Gaussian distributions. The iterative Expectation Maximization (EM) algorithm is usually used to estimate the parameters for each Gaussian component and the mixture weights (Jothilakshmi and Kathiresan, 2012).

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 Posterior 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 (Watanabe *et al*., 2010). The E-step and M-step are repeated till the convergence

of the parameters. In most of the cases, number of iterations taken by the EM algorithm for convergence (Yangn *et al*., 2012). The parameters obtained after convergence are called optimal parameters. Bayesian classifier utilizes these optimal parameters for constructing the segmentation map. For every pixel it calculates posterior probabilities of classes.

3.2.1. 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. 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. 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 a maximization step (Watanabe *et al*., 2010). 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.

That is, for each data point x_i and distribution Y_i , the membership value $y_{i,j}$ is Equation (13):

$$y_{i,j} = \alpha_i f_y\left(x_j;\theta_i\right) / fx(x_j) \quad (13)$$

The maximization step with expectation values in hand for group membership, plug in estimates are recomputed for the distribution parameters. The mixing coefficients a_i are the means of the membership values over the N data points Equation (14):

$$\alpha_i = 1/N\sum_{j=1}^{N} y_{i,j} \quad (14)$$

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 Equation (15):

$$\mu_i = \Sigma_j y_{i,j} x_j / \Sigma_j y_{i,j} \quad (15)$$

with new estimates for the θ_{is}, the expectation step is repeated to recompute new membership values. The entire procedure is repeated until model parameters converge.

4. IMPLEMENTATION

4.1. Dataset

The broadcast audio data are recorded using a TV tuner card from various TV channels which comprise different 200 clips of speech, 360 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 (conversion from binary to ASCII) i.e., 8000 sample values per second. Silence segments are removed from the audio sequence for further processing.

4.2. Signal Pre-Processing

Audio signal has to be pre processed before extracting features. There is no added information in the difference of two channels that can be used for classification or segmentation. Therefore it is desirable to have a mono signal to simplify later processes. The algorithm checks the number of channels of the audio. If the signal has more than one channel, it is mixed down to mono. The amplitude of the signal is then normalized to the maximum amplitude of the whole file to remove any effects the overall amplitude level might have on the feature extraction (Mitra *et al*., 2012).

4.3. Feature Extraction

The feature is extracted from each frame of the audio by using the feature extraction techniques. Here the DWT features are taken. An input wav file is given to the feature extraction techniques. The feature values will be calculated for the given wav file. The above process is continued for 560 number of wav files. The feature values for all the wav files will be stored separately for speech and music.

4.4. Classification

When the feature extraction process is done the audio should be classified either as speech or music. In a more complex system more classes can be defined, such as silence or speech over music. The latter is often classed as speech in systems with only two basic classes. The extracted feature vector is used to classify whether the audio is speech or music. A method where the classification is based on the output of many frames together is proposed. In this method, based on the output the feature values are extracted from the speech/music wav file and it is appended with two categories. One

category is appended for speech wav and the other category is appended for the music wav. By using the feature values with appended value SVM training is carried out. As a result of the training data two model files will be created one for speech and the other for music. For testing the feature extraction is done on different speech and music wav files other than the speech and music wav files used in the training set. All the values would be used for testing, the SVM tests the features based on models created during the training. Each second consists of 100 frames and each frame is assigned a class by a SVM classifier. Then, a global decision is made based on the most frequently appearing class within that second.

4.5. Evaluation using SVM and GMM

A non-linear support vector classifier is used to discriminate the various categories. The N class classification problem can be solved using N SVMs. Each SVM separates a single class from all the remaining classes (one-vs-rest approach).

4.5.1. Training

For classification, the audio files other than the files used for training are tested. The extracted feature vector is used to classify whether the audio is speech or music. A method where the classification is based on the output of many frames together is proposed. Support vector machine is trained to distinguish acoustic features of a category from all other categories. Two SVMs are created for each acoustic feature for each category. For training, 100 feature vectors are extracted from all the two categories, for 1 second duration each. The same process is repeated for 4 secs, 6 secs and 8 secs. The training process analyzes audio training data to find an optimal way to classify audio frames into their respective classes. The derived support vectors are used to classify audio data. The training samples are loaded and two classes are created, for each category. The two categories will be trained with two class 0 and class 1 with 560 examples.

4.5.2. Testing

For testing, 100 acoustic feature vectors (1 sec of an audio file) are given as input to SVM model and the distance between each of the feature vectors and the SVM hyperplane is obtained. The average distance is calculated for each model. The average distance gives better performance than using distance for each feature vector. The category of the audio is decided based on the maximum distance. The same process is repeated for different features and the performance is studied. The testing sample is tested using the trained model and create a result. The result will show whether the audio is speech or music.

When **Table 1** is taken into consideration, it can be seen that wavelet based parameters have higher classification results than traditional features. The best performance has been obtained with Daubechies8 wavelet.

The choice of a Kernel depends on the problem at hand because it depends on what we are trying to model. The motivation behind the choice of a particular kernel can be very intuitive and straightforward depending on what kind of information we are expecting to extract about the data. The **Table 2** shows that the Gaussian kernel classification performance is greater than the other two kernels.

Gaussian mixtures for the two classes are modeled for the features extracted. For classification the feature vectors are extracted and each of the feature vector is given as input to the GMM model. The distribution of the acoustic features is captured using GMM. We have chosen a mixture of 2, 4, 5, 10 mixture models. The class to which the audio sample belongs is decided based on the highest output. Audio classification using GMM gives an accuracy of 95.9%. The performance of GMM for different mixtures as shown in **Fig. 6** shows that when the mixtures were increased from 5 to 10 there was no considerable increase in the performance. With GMM, the best performance was achieved with 10 Gaussian mixtures.

The performance of the system for 2, 5 and 10 Gaussian mixtures is shown in **Table 3**. 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 3** shows the performance of GMM for speech and music classification based on the number of mixtures.

The performance of the system using SVM and GMM for Speech/Music classification is given in **Table 4**.

Experiments were conducted to test the performance of SVM using gaussian, sigmoidal and polynomial kernel functions. SVM performs well with a lesser number of feature vectors. Using GMM, a better performance is achieved even if the size of feature vector is larger.

GMM best performance than SVM systems give equivalent results for each kind of category in **Fig. 6**.

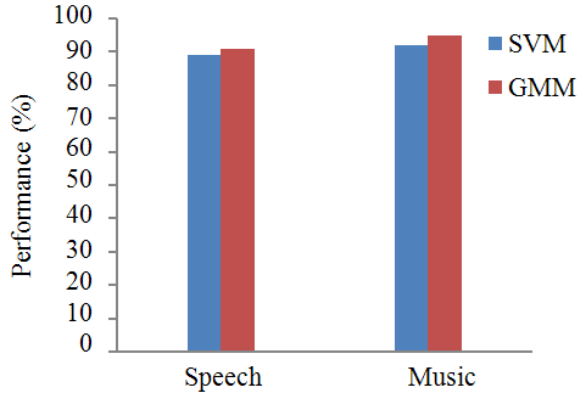

Fig. 6. Performance of speech/music classification using SVM and GMM

Table 1. Performance of classification for different Wavelet Transforms

Mother wavelet	Speech (%)	Music (%)	Overall (%)
Haar	94.3	90.4	91.8
Symlets2	93.7	88.5	90.6
Daubechies8	96.7	94.6	95.2

Table 2. Classification performance for different kernel function

Kernel function	Speech (%)	Music (%)
Gaussian	85.3	88.4
Sigmoidal	82.4	84.6
Polynomial	83.4	80.7

Table 3. Performance of GMM for different mixtures

GMM	2 mixtures (%)	5 mixtures (%)	10 mixtures (%)
Speech	92.4	92.5	93.4
Music	88.7	87.7	86.8

Table 4. Speech/Music classification performance using SVM and GMM

SVM (%)	GMM (%)
93.65	95.4

5. CONCLUSION

In this study a system for classifying the audio into speech and music using Discrete Wavelet Transform is presented. A nonlinear support vector machine learning algorithm is applied to obtain the optimal class boundary between the various classes namely speech and music by learning from training data using different kernel function and performance is studied. Experimental results show that the proposed audio classification scheme is very effective and the accuracy rate is 93.65%. The performance was compared to Gaussian Mixture Model which showed an accuracy of 95.4%. GMM using EM algorithm is used to estimate the parameters. The performance of GMM for different mixtures shows satisfactory results. The proposed feature extraction and classification models results in better accuracy overall 94.5% in speech/music classification. This work indicates that Support Vector Machines and Gaussian Mixture Model can be effectively used for audio classification. In future study other acoustic features namely Linear Prediction Coefficients, Linear Prediction Cepstral Coefficients can be extracted and the performance can be analysed and compared with the performance of Discrete Wavelet Transform features. Other pattern classification technique can also be studied to compare the performance with SVM and GMM. Even though by now some progress has been achieved, there are still remaining challenges and directions for further research, such as, extracting different features and developing better classification algorithms and integration of classifiers to reduce the classification errors.

6. REFERENCES

Abdolali, B. and H. Sameti, 2012. A novel method for speech segmentation based on speakers' characteristics. Signal Image Process. Int. J., 3: 65-78.

Bhavsar, H. and M.H. Panchal, 2012. A review on support vector machine for data classification. Int. J. Adv. Res. Comput. Eng. Technol., 1: 185-189.

Brunner, C., A. Fischer, K. Luig and T. Thies, 2012. Pairwise support vector machines and their application to large scale problems. J. Mach. Learn. Res., 13: 2279-2292.

Feki, I., A.B. Ammar and A.M. Alimi, 2012. New process to identify audio concepts based on binary classifiers framework. Int. J. Comput. Electr. Eng., 4: 515-518.

Frikha, M. and A.B. Hamida, 2012. A comparitive survey of ANN and hybrid HMM/ANN architectures for robust speech recognition. Am. J. Intell. Syst., 2: 1-8. DOI: 10.5923/j.ajis.20120201.01

Ghosal, A. and S.K. Saha, 2011. Speech/music classification using em-pirical mode decomposition. Proceedings of the Second International Conference on Emerging Applications of Information Technology, Feb. 19-20, IEEE Xplore Press, Kolkata, pp: 49-52. DOI: 10.1109/EAIT.2011.19

Golumbic, E.M.Z., N. Ding, S. Bickel, P. Lakatos and C.A. Schevon *et al*., 2012. Mechanisms underlying selective neuronal tracking of attended speech at a ''Cocktail Party", Neuron, 77: 980-991. DOI: 10.1016/j.neuron.2012.12.037

Jiang, Y.G., S. Bhattacharya, S.F. Chang and M. Shah, 2013. High-level event recognition in unconstrained videos. Int. J. Multimed. Inform. Retr., 2: 73-101. DOI: 10.1007/s13735-012-0024-2

Jothilakshmi, S. and N. Kathiresan, 2012. Automatic music genre classification for indian music. Proceedings of the International Conference on Software and Computer Applications, (SCA' 12), IACSIT Press, Singapore, pp: 55-59.

Kapp, M.N., R. Sabourin and P. Maupin, 2012. A dynamic model selection strategy for support vector machine classifiers. Applied Soft Comput., 12: 2550-2565. DOI: 10.1016/j.asoc.2012.04.001

Khan, A.U., L.P. Bhaiya and S.K. Banchhor, 2012. Hindi speaking person identification using zero crossing rate. Int. J. Soft Comput. Eng., 2: 1-4.

Kim, H.G., G.J. Jang, J.S. Park, J.H. Kim and Y.H. Oh, 2012. Speech Segregation based on pitch track correction and music-speech classification. Adv. Electr. Comput. Eng., 12: 15-20. DOI: 10.4316/AECE.2012.02003

Krajewski, J., S. Schnieder, D. Sommer, A. Batline and B. Schuller, 2012. Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech. Neurocomputing, 84: 65-75. DOI: 10.1016/j.neucom.2011.12.021

Lazouni, M.E.A., M.E.H. Daho, N. Settouti and M.A. Chikh, 2013. SVM computer aided diagnosis for anesthetic doctors. Int. J. Innov. Technol. Exp. Eng., 2: 235-240.

Lim, C. and J.H. Lim, 2012. Enhancing support vector machine-based speech/music classification using conditional maximum a posteriori criterion. IET Signal Proc., 6: 335-340. DOI: 10.1049/iet-spr.2011.0139

Lim, C., S.R. Lee, Y.W. Lee and J.H. Chang, 2012. New techniques for improving the practicality of an svm-based speech/music classifier. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Mar. 25-30, IEEE Xplore Press, Kyoto, pp: 1657-1660. DOI: 10.1109/ICASSP.2012.6288214

Liu, C.L., 2010. A tutorial of the wavelet transform.

Mishra, P. and S. Agrawal, 2012. Recognition of speaker using Mel frequency cepstral coefficient and vector quantization. Int. J. Sci. Eng. Technol. Res., 1: 12-17.

Mitra, V., H. Franco, M. Graciarena and A. Mandal, 2012. Normalized amplitude modulation features for large vocabulary noise-robust speech recognition. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Mar. 25-30, IEEE Xplore Press, Kyoto, pp: 4117-4120. DOI: 10.1109/ICASSP.2012.6288824

Nagavi, T.C. and N.U. Bhajantri, 2012. An Extensive Analysis of query by singing/humming system through query proportion. Int. J. Multimedia Applic., 4: 73-86.

Nidhyananthan, S.S. and R.S.S. Kumari, 2013. Language and text-independent speaker identification system using GMM. Wseas Trans. Signal Process., 4: 185-194.

Patil, V.D. and S.D. Ruikar, 2012. Wavelet-based image enhancement using nonlinear anisotropic diffusion. Int. J. Adv. Res. Comput. Sci. Software Eng., 2: 158-162.

Rekik, S., D. Guerchi, H. Hamam and S.A. Selouani, 2012. Audio steganography coding using the discrete wavelet transforms. Int. J. Comput. Sci. Security, 6: 79-83.

Sporka, A.J., O. Polacek and J. Havlik, 2012. Segmentation of speech and humming in vocal input. Radio Eng., 21: 923-929.

Subashini, K., S. Palanivel and V. Ramaligam, 2012. Audio-video based segmentation and classification using AANN. Int. J. Comput. Applic. Technol., 1: 53-56. DOI: 10.7753/IJCAT0102.1003

Sumithra, M.G., K.G. Thanuskodi and B. Deepa, 2011. A new robust hybrid approach to enhance speech in mobile communication systems. Am. J. Applied Sci., 8: 332-342. DOI: 10.3844/ajassp.2011.332.342

Suresha, M., N.A. Shilpa and B. Soumya, 2012. Apples grading based on SVM classifier. Proceedings on National Conferecne on Advanced Computing and Communications, (NCACC' 12).

Theodorou, T., I. Mporas and N. Fakotakis, 2012. Automatic sound classification of radio broadcast news. Int. J. Signal Process. Image Process. Patt. Recogn., 5: 37-48.

Watanabe, H., S. Muramatsu and H. Kikuchi, 2010. Interval calculation of EM algorithm for GMM parameter estimation. Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), May 30-Jun. 2, IEEE Xplore Press, Paris, pp: 2686-2689. DOI: 10.1109/ISCAS.2010.5537044

Xing, L., M. Zhu and J. Hu, 2012. A multi-semantic audio classification method based on tensor space. J. Inform. Comput. Sci., 9: 969-975.

Yangn, M.S., C.Y. Lai and C.Y. Lin, 2012. A robust EM clustering algorithm for gaussian mixture models. Patt. Recogn., 45: 3950-3961. DOI: 10.1016/j.patcog.2012.04.031