International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 10 (2015) pp. 26857-26870

http://www.ripublication.com

Computerized Methods For Audio Segmentation and Classification: Survey

[1]Mr. A. Muthumari*, [2]Dr.K.Mala,
[1]Assistant Professor, Department of Computer Science and Engineering
University College of Engineering
(A Constituent College of Anna University, Chennai),
Pullangudi, Ramanathapuram - 623 513. muthu_ru@yahoo.com
[2]Professor, Department of Computer Science and Engineering,
Mepco Schlenk Engineering College, Sivakasi – 626 005.
kmala@mepcoeng.ac.in

Abstract

Digital audio applications are part of our day to day life. The choice of the features is crucial to audio classifiers, as they should be able to discriminate the signal classes with certain margins. In the similarity measurement process, the signal features are typically compared to a set of features stored in a database, which have been generated through a training procedure. The audio data is segmented and classified different classes like silence and non- silence, the non silence is further segmented into speech, noise and other sounds. The audio classification is very useful for audio indexing, content-based audio retrieval and online audio distribution. In this paper, a state of art survey is carried out first by describing the various stages of audio segmentation and classification, and then it briefs the various audio processing methods and techniques such as pre-processing, segmentation feature extraction and classification for classifying the audio signal. This paper focuses on some of the existing computerized methods for audio segmentation, classification and presents a new direction in the automatic organization of audio collections.

Keywords: Audio, segmentation, classification, Feature Extraction, Audio retrieval, signal feature

Introduction

Audio processing has been widely used in many applications for the last decades. One of its major applications is the segmentation and classification of audio data. During the recent years, there have been many studies on audio segmentation and classification using several techniques and features. Traditionally, the most common

approach is speech/music classifications in which the highest accuracy has been achieved, especially when the segmentation information is known beforehand (i.e., manual segmentation). There are many digital audio databases on the World Wide Web nowadays; here audio segmentation and classification would be needed for audio searching and indexing. Recently, there has been a great deal of interest in monitoring and broadcast news programs, in this case classification of speech data in terms of speaker could help in efficient navigation through broadcast news archives. The segmentation a Dynamic Bayesian Network (DBN) that jointly infers onsets and end times of the most prominent sound events in the space, along with an extension of the algorithm for covering large spaces with distributed microphone arrays [1]. The recent research network of binary classifiers (NBC) framework is presented for content-based audio classification [2]. The performance evaluations of the framework over an 8-class benchmark audio database demonstrate its scalability and notable potential, as classification error rates of less than 9% are achieved. Some author's were done the audio scene classifications. That is SVM based classification pure speech, non- pure speech music, environment sounds and silence [3]. The testing database that is composed of about 4 hours audio data. A new approach towards high performance speech/music discrimination on realistic tasks related to the automatic transcription of broadcast news is described [4]. Computerized segmentation and classification system comprises of various audio processing methods such as Feature extraction and Classification methods. This survey addresses the various audio processing methods in segmentation and classification

Segmentation and classification of audio

Audio signals which include speech, music and environmental sounds are important types of media. The problem of distinguishing audio signals into these different audio types is thus becoming increasingly significant. A human listener can easily distinguish between different audio types by just listening to a short segment of an audio signal. However, solving this problem using computers has proven to be very difficult. Nevertheless, many systems with modest accuracy could still be implemented. Fig.1 illustrates block diagram of Audio Segmentation and Classification

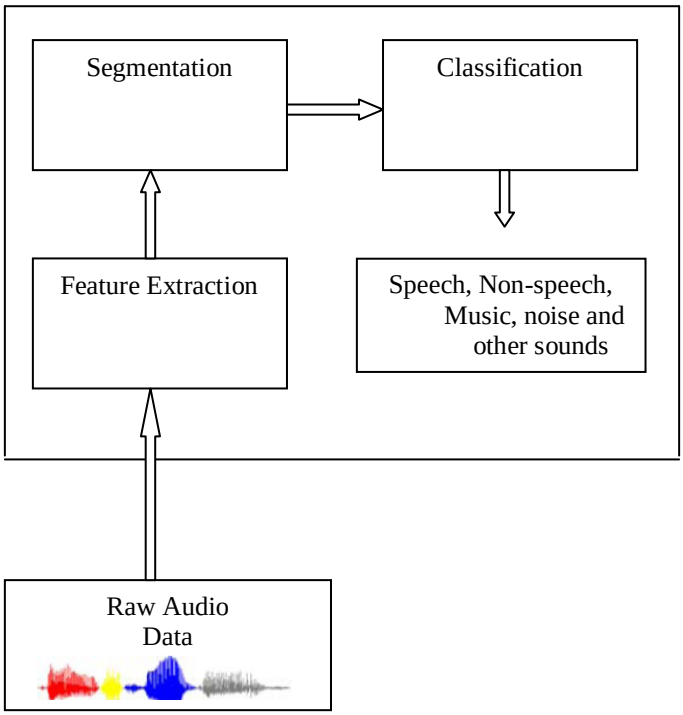

Figure 1: Block diagram of Audio Segmentation and Classification

Audio segmentation and classification have applications in wide areas. For instance, content based audio classification and retrieval is broadly used in the entertainment industry, audio archive management, commercial music usage, surveillance, etc. The audio classification is made up of two main sections: a signal processing section and a classified section. The signal processing part deals with the extraction of features from the audio signal. The various methods of time-frequency analysis developed for processing audio signals, in many cases originally developed for speech processing, are used. The classification part deals with classifying data based on the statistical information extracted from the signals.

A number of previous attempts have been made to classify audio files [5] [6]. In one sense the result of audio classification is subjective because the definition of the different classes is dependent on specific requirements and the subjective judgments of the person who interprets these requirements and constructs the definitions. Like segmentation, audio classification results can be quite different even for the same audio database. For example, music collections have been classified by genre [7] [8] [9], by mood or emotion [5] [10] [11] [12], or by instrumentation [13]. Experiments are reported in Homburg et al. [14] also supported the subjectivity of audio classification.

The sound to be classified may be not only from an audio database but other sources of sound. Some researchers have used news broadcasts as sound sources. Meinedo and Neto separated broadcast news into speech and non-speech, then further analyzed the speech and non-speech sounds separately [15]. The speech sounds were further divided by gender and the non-speech sounds were separated into noise and

music. Then they further presented a system for speech/non-speech classification, speaker segmentation, and speaker clustering, gender and background conditions classification for broadcast news [16]. The resulting classes were limited in part by the limited scope of the audio inputs. These approaches address the most well studied topic: speech/music classification. This type of classification is relatively straight forward and can achieve a high level of accuracy because the two classes are so broadly classified. There are many different approaches for speech/music classification [17] [18] [19]. Classification of speech, music and noise has also been reported [20] [21]. Similar approaches have been used to separate audio files into broad classes such as speech, music, environmental sound and silence [22] [23]. Especially when segmentation is performed beforehand, speech/music classification can be achieved with a high level of accuracy [23]. As the classes in these approaches are broadly defined the methods are limited in application and they are not suitable for databases containing a large number of narrow classes such as the Muscle Fish database.

Pre-processing

The aim of pre-processing is to remove silence from audio sequences. Silence is defined as a segment of imperceptible audio, including unnoticeable noise and very short clicks. Short time energy is used to detect silence. The short time energy function of a music signal is defined as where x(m) is the discrete time music signal, n is the time index of the short time energy and w(m) is a rectangle window, i.e., If the short time energy function is continuously lower than a certain set of threshold, the segmentation is indexed as silence. Silence segmentation will be removed from the audio sequence.

Audio Feature Extraction

Audio feature extraction plays an important role in analyzing and characterizing the audio content. Acoustic features representing the audio information can be extracted from the audio signal at the segmental level. The segmental features are the features extracted from a short (10–30 ms) segment of the audio signal. These features represent the short-time spectrum of the audio signal. The short-time spectrum envelope of the audio signal is attributed primarily to the shape of the vocal tract. The spectral information of the same sound uttered by two persons may differ due to changes in the shape of the individual's vocal tract system, and the manner of speech production. As discussed previously feature extraction plays an important role in segmentation classification of an audio signal. Hence it becomes all the more important to select those features that help the classification process more efficient. There are different types of features extraction methods, such as Zero-Crossing Rate (ZCR), High ZCR Ratio (HZCRR), Short-Time Energy (STE), Low STE Ratio (LSTER), Root Mean Square (RMS), Silence Frame Ratio (SFR), Sub-band Energy Distribution (SED), Spectrum Flux (SF), Spectral Centriod (SC), Spectral Spread (SS), Spectral Rolloff Frequency (SRF), Sub-band Periodicity (BP), Noise Frame Ratio (NFR), Linear Spectrum Pair (LSP), Linear Predictive Cepstral Coefficients

(LPCC) and Mel-frequency Cepstral Coefficients (MFCC). Now discuss some important audio feature extraction methods.

Short-Time Energy (STE)

The energy of a discrete time signals x is defined as

$$E= \sum_{m=\infty}^{\infty} x^2(m) \quad (1)$$

It gives a little information about the time dependent properties of the audio signals. In audio processing the energy of each analysis frame is important instead of the complete signal energy. This is known as short time energy. The feature selection and a combined supervised and unsupervised strategy for audio segmentation using the STE method [24] and here ESTER framework also used. The energy based segmentation using the experiment carried out by 'View4You' video indexing and retrieval system [25]. And find the performance of the metric based segmentation. As common approach in audio signal processing general audio features are applied on the segmentation part and in the classification Collective network of binary classifiers (CNBC) framework presented [2]. Basically, content-based classification and retrieval of the audio sound can be considered as a pattern recognition problem with two basic issues, feature selection and classification using selected features. Many methods have been proposed to solve this problem. One of the pioneering methods was introduced by the Music Fish group [26]. In this technique, an acoustical feature such as loudness, pitch, brightness, bandwidth, and the harmony is extracted from the audio signals.

Zero-Crossing Rate (ZCR)

In the case of discrete time signals, a zero crossing is said to occur if there is a significant difference between successive samples. The rate at which zero crossings happen is 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. But for a broadband signal such as speech, it is much less accurate [28]. However, by using a representation based on the short time average zero crossing rate, rough estimates of spectral properties can be obtained. The expression for the short time average zero crossing rates are shown below. 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.

$$Z_n = \sum_{m=-\infty}^{\infty} |\, \mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)] \,|\, w(n-w) \quad (2)$$

Where

sgn [x (n)] =1 x (n) ≥0

=-1 x (n) <0

And

w (n) =1/2N 0≤n≤N-1

=0 otherwise

Zero crossing rates have been proven to be useful in characterizing different audio signals and have been popularly used in speech/music classification problems. Variations of the zero crossing rates have also been used in some audio classification systems. It is suggested that a variation of the ZCR the high zero crossing rate ratio (HZCRR) [27], to be more discriminative than the exact value of ZCR.

The model for speech production suggested that the energy of voice speech is concentrated below about 3 kHz because of the spectrum falloff introduced by the glottal wave, whereas for unvoiced speech, most of the energy is found at higher frequencies. Since the higher frequencies imply high zero crossing rate, and low frequencies imply low zero-crossing rate, there is a strong correlation between zero-crossing rate and energy distribution with frequency. A reasonable generalization is that if the zero-crossing rate is high, the speech signal is unvoiced, while if the zero-crossing rate is low the speech signal is voiced.

There are a number of practical considerations in implementing a representation based on the short-time average zero-crossing rate. Although the basic algorithm for computing a zero-crossing rate requires only a comparison of signs of pairs of successive samples, special care must be taken in the sampling process. Clearly, the zero-crossing rate is strongly affected by DC offset in the analog-to digital converter. The application of the zero-crossing representation is as a simple intermediate step in obtaining a frequency domain representation of speech. The approach involves band pass filtering of speech signal in several contiguous frequency bands. Short-time energy and zero-crossing representations together give a representation that crudely reflects the spectral properties of the signal.

Linear Predictive Cepstral Coefficients (LPCC)

The cepstrum is a common transform used to gain information from a person's speech signal. It can be used to separate the excitation signal (which contains the words and the pitch) and the transfer function (which contains the voice quality). The cepstrum can be seen as information about the rate of change in the different spectrum bands. The cepstral coefficients are the coefficients of the Fourier transform representation of the logarithm magnitude spectrum. Cepstral coefficients of a sequence x are the coefficients of the inverse discrete Fourier transform (IDFT) of the log magnitude short-time spectrum [25]

$$\text{IDFT}\left(\log\left(|\text{DFT}(x)|\right)\right) \tag{3}$$

If x is LPC, the cepstral coefficients are known as Linear Prediction Cepstral Coefficients (LPCC).

Mel-frequency Cepstral Coefficients (MFCC)

After pre-processing, the spectral coefficients of the window frames are computed using fast Fourier transform. The results of the FFT will be information about the amount of energy at each frequency band. Human hearing is not equally sensitive at all frequency bands. It is less sensitive at higher frequencies roughly above 1000 Hz [29]. MFCC is extracted using this principle. The mapping of frequency in mel scale is linear below 1000Hz and logarithmic above 1000 Hz. So the band edges and center

frequencies of the filters are linear for low frequency and logarithmically increase with increasing frequency. We call these filters as mel-scale filters and collectively a mel-scale filter bank. As can be seen, the filters used are triangular and they are equally spaced along the mel-scale which is defined by

$$\text{Mel (f)} = 2595\log_{10}(1+\frac{f}{700}) \quad (4)$$

Each short term Fourier transform (STFT) magnitude coefficient is multiplied by the corresponding filter gain and the results are accumulated. Then DCT is applied to the log of the Mel spectral coefficients to obtain the mel frequency cepstral coefficients.

Audio Classification models

Unsupervised approaches

Unsupervised clustering algorithms take them advantage of the similarity measure to organize the music collection with clusters of similar titles. The simplest choice to measure distance between two feature vectors is, for example, to use a Euclidean distance or a cosine distance. However these distances will only make sense if the feature vectors are time-invariant. Otherwise two perceptually similar titles may be distant according to the measure if the similar features are time shifted. To build a time-invariant representation of a time series of feature vectors, one usually builds a statistical model of the distribution of the features and then uses the distance to compare these models directly.

Typical models include Gaussian and Gaussian mixtures (GMMs) (GMMs have been used to build song timbre models in [30], [31] and [32]). The Kullback-Leibler divergence or relative entropy is the natural way to evaluate the distance between probability distributions but it is not suited for GMMs.

K-means is probably the simplest and most popular clustering algorithm. It allows partitioning a set of vectors into K disjoint subsets. One of its weaknesses is that it requires the number of clusters (K) to be known in advance. Shao et al. [33] cluster their music collection with the Agglomerative Hierarchical Clustering, a clustering algorithm that starts with *N* singleton clusters (where *N* is the number of titles of the database) and that forms a sequence of clusters of successive merging.

The Self-Organizing Map (SOM) and the Growing Hierarchical Self-Organizing Map (GHSOM) are used to cluster data and organize them on a 2-dimensional space in such a way that similar feature vectors are grouped close together. SOMs are unsupervised artificial neural networks that map high dimensional input data onto lower dimensional output spaces while preserving the topological relationships between the input data items as faithfully as possible. GHSOMs are a special case of SOMs which make use of a hierarchical structure with multiple layers where each layer consists of a number of independent SOMs. Rauber et al. [34] use an output space of dimension 2 to allow a visual representation of a music collection with a GHSOM.

The supervised approach

The supervised approach to audio classification has been studied more extensively. As a first step, the system is trained with some manually labelled data, and then it is used to classify unlabeled data. The major interest of supervised classification compared to the expert system approach is that one does not need to explicitly describe musical genres: the classifier attempts to form automatically relationships between the features of the training set and the related categories. We describe here a number of commonly used supervised machine learning algorithms. Then we present the results obtained with these algorithms in literature.

- **K-Nearest Neighbor (KNN):** It is a non-parametric classifier based on the idea that a small number of neighbors influence the decision on a point. More precisely, for a given feature vector in the target set, the K closest vectors in the training set are selected (according to some distance measures) and the target feature vector is assigned the label of the most represented class in the K of neighbors (there is actually no other training than storing the features of the training set). KNNs are evaluated in the context of genre classification in [35].
- **Gaussian Mixture Models (GMM):** GMMs model the distribution of feature vectors. For each class, we assume the existence of a probability density function expressed 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. They can be used as classifiers by using a maximum likelihood criterion to find the model best suited to a particular song. They have been used to directly model musical genres in [35]. In [36], a treelike structure of GMMs is used to model the underlying genre taxonomy: a divide-and-conquer strategy is used to first classify items on a coarse level and then on successively finer levels. The classification decision is thus decomposed into a number of local routing or refinement decisions in the taxonomy. In addition, feature selection at every refinement level allows optimizing classification results. West and Cox [37] use a Maximal Classification Binary Tree built by forming a root node containing all the training data and then splitting that data into two child nodes by using single Gaussian classifier with Mahalanobis distance measurements. In order to split a node, all possible combinations of classes are formed and the combination of classes yielding the best split is chosen (notice that the creation of the tree is unsupervised whereas the classifiers used for splitting on each node are trained in a supervised manner).
- **Hidden Markov Model (HMM):** Hidden Markov Models can also be used for classification purposes. They have been extensively used in speech recognition because of their capacity to handle time series data. HMMs may be seen as a double embedded stochastic process, one process is not directly observable (hidden) and can only be observed through another stochastic process (observable) that produces the time set of observations.
- **Support Vector Machines (SVM):** SVMs are based on two properties: margin maximization (which allows for a good generalization of the classifier)

and nonlinear transformation of the feature space with kernels (as a data set is more easily separable in a high dimensional feature space). A Mixture of Experts solves a classification problem by using a number of classifiers to decompose it into a series of sub-problems. Not only does it reduce the complexity of each single task but it also improves the global accuracy by combining the results of the different classifiers (experts). Of course, the number of needed classifiers is increasing, yet, by having each of them handle a simpler problem, the overall required computational power is reduced.

- **Artificial Neural Network (ANN):** An ANN is composed of a large number of highly interconnected processing elements (neurons) jointly working to solve specific problems. The most widely used supervised ANN for pattern recognition is the Multi-Layer Perceptron (MLP). It is a very general model that can in principle approximate any non-linear function. MLPs have been used in [38] in the context of artist identification. Neural Networks as well as the other reviewed architectures (except HMMs), can only handle static patterns. This weakness is partly overcome in [38] by inputting a number of adjacent feature vectors into the network so that contextual information is taken into account: this strategy corresponds to the so called Feedforward Time-Delay Neural Network (TDNN). Other paradigms oriented towards the processing of temporal sequences have been proposed (Recurrent Networks such as the Elman-Network) but have not been used yet in the context of music genre classification. Soltau et al. [39] have introduced in the context of recognition of music genres an original method for explicit time modeling of temporal structure of music (ETM-NN): an MLP is trained to recognize music genres but, rather than considering its output, the activation of its hidden neurons is considered as a compact representation of the input feature vector (it is known indeed that the first half of a feed-forward network performs a specific nonlinear transformation of the input data into a space in which the discrimination should be simpler). Each hidden neuron can be seen as an abstract musical event – not necessarily related to an actual musical representation. The sequence of abstract events over time is then analyzed to build one single feature vector which is fed to a second network that implements the final decision about the genre of the musical piece.

Performance

In this section the analysis of performance results of the various existing methods were done. The performance measure of an algorithm was based on accuracy. The detailed results regarding the performance of various classification methods in existing systems are as follows.

The weighted SVM classifier the database included 600 files, with half utilized for training and the others for testing. The sampling rate is 16 kHz and each sample is 16 bits. For each audio class, half of the audio files are utilized for training and the others for testing. The frame size is 512 samples (32 ms), with 50% overlap in each of the two adjacent frames [40]. The accuracy rate is 94.8 %. A parallel Multi-layered Perceptron network is proposed which offers a classification accuracy of 84.4% to

distinguish between 6 different classes [41]. The performance was compared to Radial basis function neural network (RBFNN) which showed an accuracy of 93% [42]. Classification the feature vectors are extracted and each of the feature vector is given as input to the GMM model. The distributions of the acoustic features are captured using GMM. Here 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 92.9% [28]. Table I briefs the various classifiers along with their performance measures. Some researchers have developed sophisticated approaches for the classification of more diverse sounds. For example, Burred and Lerch extended the classification results to speech, background noise and 13 musical genres [43]. All of these approaches demonstrate the importance of audio classification accuracy in an audio processing system.

The performance of the system for SVM, NN, RBFNN and GMM for audio classification is given in Table 1.

Table 1: Audio classification performance using SVM, NN, RBFNN and GMM

Classifier Method	**Accuracy rate**
Weighted SVM	94.8 %
Neural Network	84.4%
Radial basis function neural network	93%
GMM	92.9%

New problems, relying on similar techniques, are emerging in the field of music information retrieval as new markets and applications eventually take off. Many of the previously introduced algorithms can be applied with minor changes to these new applications while results coming from innovative research fields can provide useful feedbacks on genre classification techniques. Briefs the various Paradigms and classification methods is given Table 2.

Table 2: Paradigms and classification methods

Unsupervised Clustering	**Supervised Classification**
1. No taxonomy: classification emerge from the data	1. Uses a taxonomy
2. Organization according to similarity between excerpts	2. The learning algorithm maps features to classes without describing rules explicitly
3. Typical clustering algorithms: KMeans, Agglomerative Hierarchical Clustering, Self-Organizing Map and Growing Hierarchical SOM	3. Typical supervised learning algorithms: KNN, Neural Networks, LDA, SVMs…

Research Directions

The audio information is vital role in real world problem. Though many algorithms and methods have been developed in segmentation and classification, the accuracy rate was up to 90-94% only. The audio analysis plays a vital role in audio analysis and enhancement, which comprises of, Segmentation, Feature extraction and Classification. Based on the survey, the Support Vector Machine based classifier shows a better accuracy rate compared to all the other classification methods. Despite of this, a better classifier result can be obtained by combining the Co-occurrence Based audio features such as Energy, contrast, correlation, Difference entropy etc. with the Support Vector Machine. The accuracy of SVM algorithm can be increased by selecting the features through the combination of neural network based searching procedure and by training them with it. The Accuracy of Neural network based classifier can also be increased by using suitable wavelet coefficients and also by selecting more suitable features for wavelet function

Conclusion

This paper presented a study of the existing audio segmentation and classification. The review of the existing methods demonstrates that the bottleneck in comparing the results of the published work is the difference in the parameters considered in forming the methods including the properties of training and test data sets and performance evaluation methods. The paper formulated a comparison of the performance of the existing reported approaches. Overall, systems that employ SVM based classification algorithms have been more popular and have demonstrated better performances. It can be concluded that the performance of several existing systems has been superior. In this article, we reviewed typical feature extraction techniques used in audio information retrieval for the different music elements; given these features the two main paradigms for audio classification were presented with their advantages and drawbacks. The exhibited performances can be improved by employing proper false positives reduction components.

References

[1] Gordon Wichern,Jiachen Xue, Harvey Thornburg, Brandon Mechtley and Andreas Spanias, “Segmentation, Indexing, and Retrieval for Environmental and Natural Sounds” in IEEE Transactions on audio, speech, and language processing, vol. 18, no. 3, march 2010.

[2] Toni Mäkinen, Serkan Kiranyaz, Moncef Gabbouj “Content-based Audio Classification using Collective Network of Binary Classifiers” in IEEE Finnish Centre of Excellence Program , pp. 116-123, 2011

[3] Hongchen JIANG, Junmei BAI, Shuwu ZHANG, Bo XU, “SVM-based Audio Scene Classification ” in IEEE Proc. of NLP-KE, 2005.

[4] C. Panagiotakis, and G. Tziritas, “A Speech/Music Discriminator Based on RMS and Zero-Crossings”, in IEEE Transactions on Multimedia, pp. 155-166, 2005.

[5] M. I. Mandel, G. E. Poliner and D. P. W. Ellis, "Support Vector Machine Active Learning for Music Retrieval," in Multimedia Systems, vol. 12, no. 1, pp. 3-13, Aug. 2006.

[6] G. Guo and S. Z. Li, "Content-based Audio Classification and Retrieval by Support Vector Machines," IEEE Trans. Neural Netw., vol. 14, no. 1, pp. 209- 215, Jan. 2003.

[7] Z. Cataltepe, Y. Yaslan and A. Sonmez, "Music Genre Classification Using MIDI and Audio Features," EURASIP J. on Applied Signal Process., vol. 2007, no. 1, pp. 150-150, Jan. 2007.

[8] S. Lippens, J. P. Martens, M. Leman, B. Baets, H. Meyer and G. Tzanetakis, "A Comparison of Human and Automatic Musical Genre Classification," in Proc. ICASSP'04, Montreal, Canada, vol.4, pp. 233-236, 2004

[9] M. F. McKinney and J. Breebaart, "Features for Audio and Music Classification," in Proc. ISMIR'03, Washington, D.C., USA, pp. 151-158, 2003

[10] Y. Feng, Y. Zhuang and Y. Pan, "Popular Music Retrieval by Detecting Mood," in Proc. SIGIR, Toronto, Canada, pp. 375-376 , 2003

[11] L. Lu, D. Liu and H. J. Zhang, "Automatic Mood Detection and Tracking of Music Audio Signals," IEEE Trans. Audio, Speech, and Language Process., vol. 14, no. 1, pp. 5-18, Jan. 2006.

[12] T. Pohle, E. Pampalk and G. Widmer, "Evaluation of Frequently Used Audio Features for Classification of Music into Perceptual Categories," presented at the 4th Int. Workshop on CBMI, Riga, Latvia, 2005.

[13] S. Essid, G. Richard and B. David, "Musical Instrument Recognition by Pairwise Classification Strategies," IEEE Trans. Audio, Speech, and Language Process., vol. 14, no. 4, pp. 1401-1412, July 2006.

[14] H. Homburg, I. Mierswa, B. Möller, K. Morik and M. Wurst, "A Benchmark Dataset for Audio Classification and Clustering," in Proc. ISMIR'05, London, UK, pp. 528-531, 2005

[15] H. Meinedo and J. Neto, "Audio Segmentation, Classification and Clustering in a Broadcast News Task," in Proc. ICASSP'03, Hong Kong, China, vol.2, pp. 5-8, 2003

[16] H. Meinedo and J. Neto, "A Stream-based Audio Segmentation, Classification and Clustering Pre-processing System for Broadcast News using ANN Models," in Proc. Interspeech'05-Eurospeech, Lisbon, Portugal, pp. 237-240

[17] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Music/Speech Discriminator," in Proc. ICASSP'97, Munich, Germany, 1997, vol.2, pp. 1331-1334, 2005

[18] K. El-Maleh, M. Klein, G. Petrucci and P. Kabal, "Speech/Music Discrimination for Multimedia Application," in Proc. ICASSP'00, Istanbul, Turkey, vol.4, pp. 2445-2448

[19] O. Turk, O. Sayli, H. Dutagaci and L. M. Arslan, "A Sound System Based on Subband Processing," in Proc. 3rd Process. Symp., Leuven, Belgium, pp. 641-644, 2002

[20] T. Zhang and C.-C. J. Kuo, "Hierarchical Classification of Audio Data for Archiving and Retrieving," in Proc. ICASSP'99, Phoenix, AZ, USA, 1999, vol.6, pp. 3001-3004, 2000

[21] S. Ravindran and D. V. Anderson, "Audio Classification and Scene Recognition and for Hearing Aids," in Proc. ISCAS'05, Kobe, Japan, vol.2, pp. 860- 863, 2005

[22] L. Lu, H. Jiang and H.-J. Zhang, "A Robust Audio Classification and Segmentation Method," in Proc. 9th ACM Int. Conf. Multimedia, Ottawa, Canada, vol.9, pp. 203-211, 2001

[23] S. Kiranyaz, A. F. Qureshi and M. Gabbouj, "A Generic Audio Classification and Segmentation Approach for Multimedia Indexing and Retrieval," IEEE Trans. Audio, Speech, and Language Process., vol. 14, no. 3, pp. 517-523, May 2006.

[24] Gaël Richard, Mathieu Ramona and Slim Essid "Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams" in IEEE ICASSP pp- 461- 464, 2007.

[25] Thomas Kemp Michael Schmidt Martin Westphal Alen Waibel"Stratrgies for automatic segmentation of audio data" in IEEE ICASSP, Vol. 3, pp- 1423- 1426, 2000.

[26] Wold,E.,Blum,T.,Keislar,D.,Wheaton,J.:Content-based classification, search, and retrieval of audio, IEEE Multimedia, pp.27-36, 1996.

[27] Lie Lu, Hao Jiang and HongJiang Zhang "A Robust Audio Classification and Segmentation Method" in proc. of the ninth ACM international conference on Multimedia, 2011.

[28] Yingying Zh, Zhong Ming, Qiang Huang" Automatic Audio Genre Classification Based on Support Vector Machine" in Third International Conference on Natural Computation, 2007

[29] P. Dhanalakshmi, S. Palanivel, V. Ramalingam "Classification of audio signals using AANN and GMM" in Elsevier Applied Soft Computing, 716–723, 2011

[30] J.J. Aucouturier, F. Pachet, "Music similarity measures: what's the use?" in Proc. of the 3rd Int. Symposium on Music Information Retrieval, 2002.

[31] E. Pampalk, A. Flexer, G. Widmer, "Improvements of audio based music similarity and genre classification?", in Proc. of the 6th Int. Symposium on Music Information Retrieval, London, UK, 2005.

[32] M. Mandel, D. Ellis, "Song-level features and support vector machines for music classification", in Proc. of the 6th Int. Symposium on Music Information Retrieval, London, UK, 2005.

[33] X. Shao, C. Xu, M. Kankanhalli, "Unsupervised classification of musical genre using hidden Markov model", in IEEE Int. Conf. of Multimedia Explore (ICME), Taibei, Taiwan, China, 2004.

[34] A. Rauber, E. Pampalk, D. Merkl, "Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarity", in Proc. of the 3rd Int. Conf. on Music Information Retrieval, Paris, France, 2002.

[35] G. Tzanetakis, P. Cook, "Musical genre classification of audio signals", in IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 5, July 2002.

[36] J.J. Burred, A. Lerch, "A hierarchical approach to automatic musical genre classification", in Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx), London, UK, 2003.

[37] K. West, S. Cox, "Finding an optimal segmentation for audio genre classification", in Proceedings of the 6th Int. Symposium on Music Information Retrieval, London, UK, 2005.

[38] A. Berenzweig, D. Ellis, S. Lawrence, "Using voice segments to improve artist classification of music", in Proc. of the AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio, 2002.

[39] H. Soltau, T. Schultz, M. Westphal, A. Waibel, "Recognition of music types", in Proc. IEEE Int .Conf. on Acoustics, Speech and Signal Processing (ICASSP), Seattle, USA, 1998.

[40] Wenjuan Pan, Yong Yao , Zhijing Liu andWeiyao Huang "Audio Classification in a Weighted SVM" in International Symposium on Communications and Information Technologies (ISCIT)468- 472, 2007.

[41] Vikramjit Mitra"A Neural Network Based Audio Content Classification" on international join conference, 2007

[42] P. Dhanalakshmi , S. Palanivel, V. Ramalingam "Classification of audio signals using SVM and RBFNN" in Elsevier Expert Systems with Applications, 1- 7, 2008

[43] J. J. Burred and A. Lerch, "A Hierarchical Approach to Automatic Musical Genre Classification," in Proc. DAFx-03, London, UK, pp. 8-11, 2003