Multimedia Systems (2006) 12(1): 55–67
DOI 10.1007/s00530-006-0034-0

REGULAR PAPER

M. Kashif Saeed Khan · Wasfi G. Al-Khatib

Machine-learning based classification of speech and music

Published online: 7 April 2006

Abstract The need to classify audio into categories such as speech or music is an important aspect of many multimedia document retrieval systems. In this paper, we investigate audio features that have not been previously used in music-speech classification, such as the mean and variance of the discrete wavelet transform, the variance of Mel-frequency cepstral coefficients, the root mean square of a lowpass signal, and the difference of the maximum and minimum zero-crossings. We, then, employ fuzzy C-means clustering to the problem of selecting a viable set of features that enables better classification accuracy. Three different classification frameworks have been studied: Multi-Layer Perceptron (MLP) Neural Networks, radial basis functions (RBF) Neural Networks, and Hidden Markov Model (HMM), and results of each framework have been reported and compared. Our extensive experimentation have identified a subset of features that contributes most to accurate classification, and have shown that MLP networks are the most suitable classification framework for the problem at hand.

Keywords Speech music classification · Audio signal processing · Audio features · Neural networks · Hidden Markov Models · Fuzzy c-means clustering

1 Introduction

The exponential growth of the Internet and the latest advances in networking and compression technologies have made huge amounts of audio data easily available. It is not unlikely that in the near future, on-line music services will overtake the usual distribution of audio stored on physical media. Currently, browsing and management of audio data rely mostly on textual information attached manually, which is an extremely time consuming task. Furthermore, this information is often incomplete or not available at all.

M. K. S. Khan · W. G. Al-Khatib (✉)
Information and Computer Science Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
E-mail: wasfi@ccse.kfupm.edu.sa

Audio signal classification applications are potentially far reaching and relevant. With the fast growth of multimedia repositories, in general, and audio data in specific, the development of technologies for spoken document indexing and retrieval is in full expansion. Audio data sources range from broadcast radio and television, to the huge volumes of recorded material in different forms, such as tapes and digital audio stored on the Web. Speech/music classification is an important task in multimedia indexing. It is usually the first step before any further processing on audio data. Some of the applications of speech/music classification include:

- Automatic speech recognition: Broadcast radio feed may contain music segments in between different programs. Identifying the "Speech" segments will give more reliable data to the Automatic Speech Recognizer (ASR), which contributes to the minimization of word error rates, out of vocabulary words, and eliminates unnecessary computations on non-speech data.
- Content-based indexing and retrieval: After speech/music classification process, one can give meaningful descriptions such as speech, music, silence, etc., to different segments of the audio data. Such indexing will support querying speech segments only, or music segments only from a multimedia database perspective.
- Speaker recognition: Extracting speech from the audio signals may enable speaker recognition techniques for identifying and tracking specific speakers for indexing or security purpose.
- Improving audio coding: Classifying audio data into speech, music, and silence can be useful in the process of decreasing the bit rate for silence segments and hence improve audio coding.
- Improving compression techniques: Some signal compression techniques are more suitable for speech signals, whereas other compression techniques may be more appropriate for music. By automatically determining the audio signal, the appropriate compression technique can be applied.

- Hearing instrument: Automatically adapting a hearing instrument for various listening situations (silence, speech, noise, music, wind, etc.) would free users from manually having to change the mode of the instrument using a push button located on the hearing instrument, sometimes a task that is problematic for many hearing instrument users.

The paper is organized as follows: We first survey up-to-date research work that has been carried out in music/speech classification in Sect. 2. Then, we present the approach we followed in selecting audio features to be considered for classification using fuzzy C-Means clustering in Sect. 3. Section 4 presents three different classification frameworks used to carry out the classification process along with the experimental setup. We follow this by presenting the experimental results in Sect. 5. Finally, a summary highlighting the interesting aspects of our work is presented and possible directions for future work are mentioned.

2 Related work

Many researchers have addressed problems related to speech/music classification, speech/music features, feature extraction, and classification frameworks. Scheirer et al. [1] and Saad et al. [2] have both examined the following five features intended to measure conceptually distinct properties of speech and music signals:

(1) Percentage of low energy frames,
(2) Roll off point of the spectrum,
(3) Spectral flux,
(4) Zero-crossing rate,
(5) Spectral centroid.

Scheirer et al. have additionally used the following eight features:

(1) Four Hz modulation energy
(2) Variance of the roll off point of the spectrum
(3) Variance of the spectral centroid
(4) Variance of the spectral flux
(5) Variance of the zero-crossing rate
(6) The cepstral residual
(7) Variance of the cepstral residual
(8) Pulse metric

Five of those are "variance" features, consisting of the variance in a 1-s window of an underlying measure which is calculated on a single frame. Scheirer et al. have used log transformations on all thirteen features. As a classification framework they have investigated four different classifiers:

- Multi-dimensional Gaussian maximum a posteriori (MAP) estimator,
- Gaussian mixture model (GMM) classification,
- Spatial partitioning scheme based on k-d trees and
- k Nearest-neighbor classifier

They have reported that the MAP Gaussian classifier does a much better job in rejecting music from the speech class than vice-versa, and among the four classifiers, the k nearest-neighbor classifier gave good results, around 92.2% accuracy. Saad et al. have proposed an algorithm in which speech/music classification is performed by using average percentage deviation. This value is calculated by finding the percentage deviation of each feature relative to the maximum deviation of that feature. If it is less than a particular threshold value, it is labeled as speech otherwise it is labeled as music. They have reported 94.25% accuracy.

Saunders [3] has described a technique for discriminating speech from music on broadcast FM radio based on the zero-crossing rate of the time domain waveform. His technique emphasized detecting certain characteristics of speech such as:

(1) limited bandwidth,
(2) alternate voiced and unvoiced sections, and
(3) energy variations between high and low levels.

It is indirectly using the amplitude, pitch and periodicity estimate of the waveform to carry out the detection process by using a multivariate Gaussian classifier. He has reported an average accuracy of 95%. Carey et al. [4] presents a comparison of several of the different features, some of them already used in [1, 3], and tested the same data by using Gaussian Mixture Models (GMM) as a classifier and Expectation Maximization (EM) algorithm for training. The following features were used for classification:

(1) Cepstral coefficients
(2) Delta cepstral coefficients
(3) Amplitude
(4) Delta amplitude
(5) Pitch
(6) Delta pitch
(7) Zero-crossing rate
(8) Delta zero-crossing rate

Separate experiments were carried out in combination of a feature and its derivative. The best performance resulted from using the cepstra and delta cepstra which gave an equal error rate (EER) of 1.2%. Parris et al. [5] have used cepstral coefficients, amplitude, and pitch features along with GMM and has reported an equal error rate of 0.7%. Chou and Gu [6] has proposed an approach for robust singing signal detection applied to applications of audio indexing in multimedia databases. The following set of features were being used by a GMM-based classifier:

(1) 4 Hz modulation energy
(2) Harmonic coefficients
(3) 4 Hz harmonic coefficients
(4) MFCC
(5) Log energy

In order to efficiently index the sound-track of multimedia documents, it is necessary to extract elementary and homogeneous acoustic segments. Pinquier et al. [7–9] have explored such prior partitioning which consists of detecting

audio signals as speech, music, speech-music, and other by using GMM. Speech detection has been carried out by considering cepstral coefficients, entropy modulation and 4 Hz modulation energy. Spectral coefficients, the number of segments and segment durations have been used to carry out music detection. They have reported 90.1% of accuracy. Harb and Chen [10] used first order sound spectrum's statistics as feature vectors, extracted using the Fast Fourier transform, and then used a neural network to estimate the probability of each mean/variance model. The NN used is a Multi-Layer Perceptron with the error back propagation training algorithm and the sigmoid function as an activation one. They have achieved 96% classification accuracy for context-dependent problems and 93% for context-independent ones. In [11], Harb and Chen have investigated audio for indexing purposes and proposed an algorithm that needs no training phase as is the case with GMM-based algorithms. It classifies audio signals into four classes: speech, music, silence, and other. Different features were used for different classes. For example, energy level and the zero crossing rate have been used to detect silence. To detect speech+music, the silence crossing rate and frequency tracking have been employed. Classification is achieved by thresholding these features. They have reported 90% classification accuracy. The possibility to discriminate between speech and music signals using features based on low frequency modulation has been investigated by Karnebäck [12]. Three different low frequency modulation parameters, 4 Hz amplitude and standard deviation, 4 Hz normalized amplitude, and 2–4 Hz normalized amplitude have been extracted and tested using GMMs. Classification accuracy of 93.6% have been reported. Wang et al. [13] present a simple and effective approach in which the proposed modified low energy ratio is first extracted as the only feature and then the system applies the Bayes MAP (Maximum A-posteriori Probability) classifier to decide the audio class. Around 97% of classification accuracy has been achieved. El-Maleh et al. [14] have focused on frame level narrow band speech/music discrimination by using four feature sets for experimentation:

(1) Line spectral frequencies (LSF)
(2) Differential LSF, the successive differences of LSF.
(3) LSF with the zero crossing count of the filtered input signal.
(4) LSF with Linear prediction zero crossing ratio, the ratio of the zero crossing count (ZCC) of the input and the ZCC of the output of the LP analysis filter.

They used two different classification algorithms: a quadratic Gaussian classifier and a k-nearest neighbor classifier. The k-nearest neighbor classifier gave the best results of 80.85%. Panagiotakis and Tziritas [15] have developed a system which first segments audio signals and then classifies them into one of the three main categories: speech, music, and silence. They have proposed an algorithm for classification based on RMS and Zero-Crossings. They have reported around 95% of classification accuracy.

One important note regarding previous related work is the absence of the ability to compare the different approaches, let alone determining the most accurate approach. The reason is that no standard set of audio data for the purpose of speech/music classification exists.

3 Selection of audio features using fuzzy C-means clustering

The first step in a classification problem is typically data reduction. The data reduction stage which is also called feature extraction, consists of discovering a few important facts about each class. The choice of features is critical as it greatly affects the accuracy of audio classification. The selected features must reflect the significant characteristics of each class of audio signals. In order to better discriminate different classes of audio, we consider features that are related to the temporal and spectral domains.

Typically, audio features are extracted at two levels: short-term frame-level and long-term clip-level. Here, a frame is defined as a group of adjacent samples lasting for 10–40 ms. The audio signal within such periods presumably remains stationary and short-term features both in the time-domain and in the frequency-domain can be extracted. For a feature to reveal the semantic meaning of an audio signal, we need to observe the temporal variations of frame features on a longer time scale, usually from 1 s to several tens of seconds. Such an interval is called an audio clip. An audio clip consists of a sequence of frames and clip-level features that characterize how frame-level features change over a clip. Clip boundaries may be the result of audio segmentation such that the content within each clip belongs to the same class. Fixed length clips, lasting for 2–3 s may also be used in determining clip boundaries.

In this section, we start by presenting audio features that have been considered. Then, we use cluster analysis to assist us in the process of selecting the smallest set of features that can possibly produce the highest classification accuracy.

3.1 Previously used audio features

Among the many features that have been used by other researchers for music/speech classification, we considered three features, namely the percentage of low energy frames, the spectral flux, and the linear predictive coefficients.

3.1.1 Percentage of "low energy" frames (%LEF)

This value measures the proportion of frames with root mean-squared (RMS) power less than 50% of the mean RMS power within a given period of time. According to [3] the energy distribution for speech is more left-skewed than that of music. The reason is that there are more quiet frames in speech as some pause between every word exists and hence the energy of the frame containing pauses is lower

than that in frames containing no pauses. This feature has been used in [1, 2, 16–21].

3.1.2 Spectral flux (SF)

This feature, also known as the delta spectrum magnitude, measures frame-to-frame spectral difference, and is computed according to the following equation:

$$\text{Spectral Flux} = \| \, |X_i| - |X_{i+1}| \, \| \, . \tag{1}$$

Thus, it characterizes the changes in the shape of the spectrum. Speech goes through more drastic frame-to-frame changes than music. The spectral flux value is higher for speech than it is for music [22–24]. This feature has been used in [1, 2, 18–25].

3.1.3 Linear predictive coefficients (LPC)

The basic idea behind linear prediction is that the next signal sample is predicted from a weighted sum of p previous samples, given as follows:

$$\hat{s}(n) = \sum_{i=1}^{p} a_i s(n-i), \tag{2}$$

where a_i represent the prediction coefficients, p is the predictor order, and $s(n-i)$ is a sample at time instance $n-i$. The prediction coefficients are determined by minimizing the mean squared error between the actual sample and the prediction. The prediction error signal, also called residual error, is given by

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i). \tag{3}$$

The prediction error is significantly higher for unvoiced speech than it is for voiced speech. The linear prediction coefficients considered in our work use the Levinson-Durbin recursion to solve the normal equations that arise from the least-squares formulation. This feature has been used in [19, 26–29].

3.2 Newly proposed audio features

This section presents audio features that we were first to propose and investigate for the purpose of music/speech classification, according to our knowledge.

3.2.1 Range of zero-crossings (R-ZC)

Zero-crossings count is a measure of the number of times that the audio signal amplitude passes through a value of zero, in a given time interval. Rather than using ZCC directly, we have used the difference of maximum and minimum zero-crossings as a feature vector. It is evident from Fig. 1a that it gives discriminating patterns for different classes of audio signal.

3.2.2 Mean (M-DWT) and variance (V-DWT) of the discrete wavelet transform

A serious drawback of using Fourier transform is that after transforming the audio signal into the frequency domain, the time information is lost. Wavelet analysis is capable of revealing aspects of data that other signal analysis techniques miss, which include trends, breakdown points, discontinuities in higher derivatives, and self-similarity. In wavelet analysis, a signal is split into an *approximation* and a *detail*. The approximation is then itself split into a second-level approximation and detail, and the process is repeated. For an n-level decomposition, there are $n+1$ possible ways to decompose or encode the signal. The approximations are the low-frequency components of the signal, whereas the details are the high-frequency components. Since we have only single dimensional data, we have used a single-level, 1-D 'Haar' wavelet transformation. We have investigated the statistical features of audio in the wavelet domain which are the mean (M-DWT), shown in Fig. 1b and the variance (V-DWT), shown in Fig. 1c. Lambrou et al. [30] have also used the two features, but for music genre classification. Delfs et al. [31] have used wavelet packet transform for the classification of piano sound.

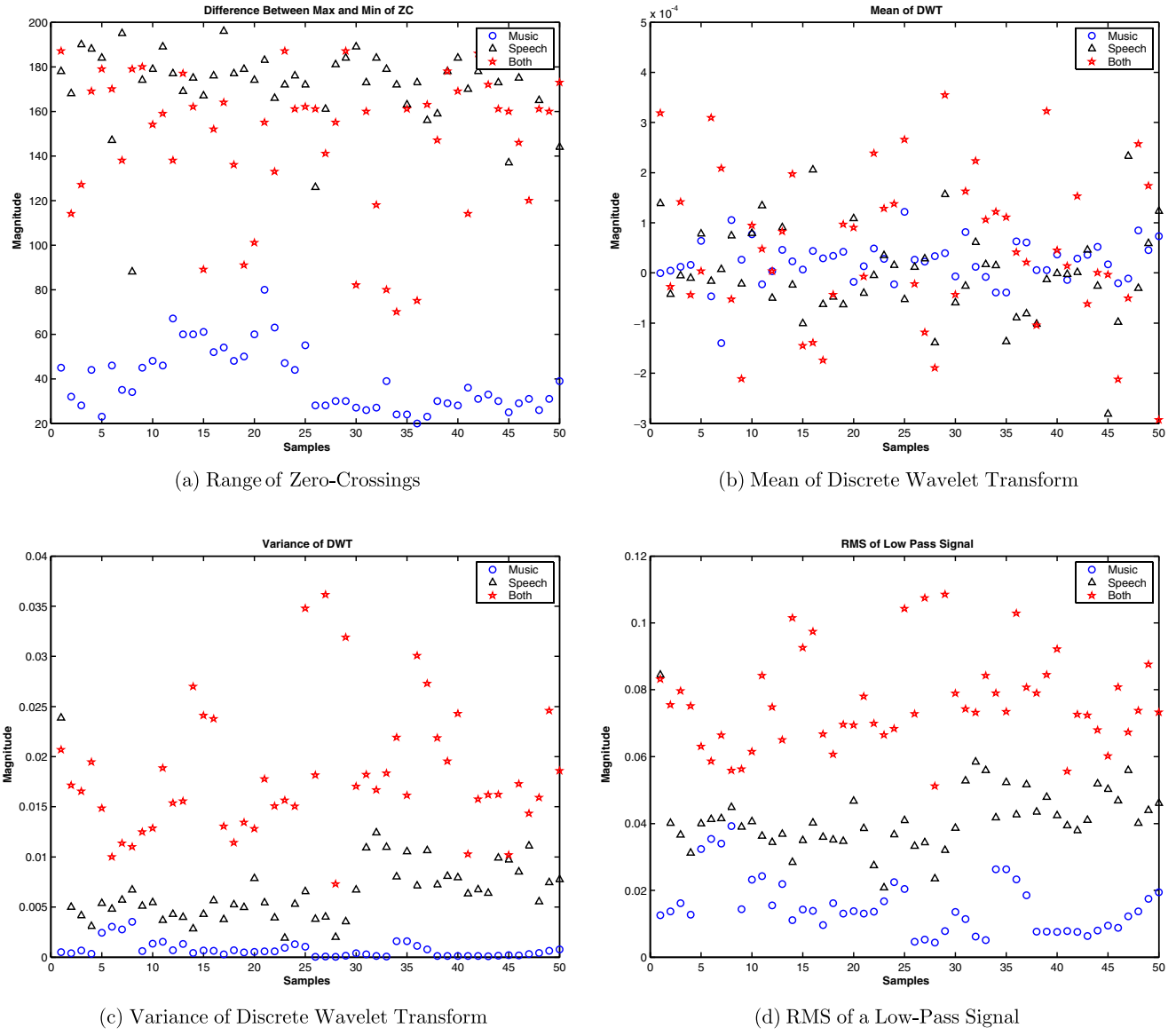
3.2.3 RMS of a lowpass signal (RMS-LPS)

Music signals have a wider bandwidth than speech extending up to 20 kHz. To limit the frequency band, we have applied a lowpass filter to filter out the high frequency contents. We have applied the Butterworth filter of 4th order with 1.1 kHz cutoff frequency. After that, we have taken the root mean square value of that lowpass response. The RMS value of a lowpass response for speech is higher than that of music as most of the speech contents are in the lower frequency band, as shown in Fig. 1d.

3.2.4 Variance of the mel frequency cepstral coefficients (V-12MFCC)

Primarily, MFCC have been used for their ability to imitate the behavior of a human ear. Psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency f, measured in Hz, a subjective pitch is measured on a scale called the *Mel* scale. The Mel-frequency scale is a linear frequency spacing below 1 kHz and a logarithmic spacing above 1 kHz. Filters spaced linearly at low frequency and logarithmic at high frequencies have been used to capture the phonetically important characteristics (voiced and unvoiced) of speech. The commonly used formula to approximately reflect the relation between the Mel-frequency and the physical frequency is given by

$$M(f) = 1125 \times \log_{10}\left(1 + \frac{f}{700}\right). \tag{4}$$



(a) Range of Zero-Crossings

(b) Mean of Discrete Wavelet Transform

(c) Variance of Discrete Wavelet Transform

(d) RMS of a Low-Pass Signal

Fig. 1 Discriminating abilities of R-ZC, M-DWT, V-DWT, and RMS-LPS, respectively

Although 12 coefficients are typically used for speech representation, Srinivasan et al. and Lippens et al. have found that the first five coefficients provide the best classification performance [17, 18]. Instead of using the MFCC coefficients, we have investigated the variance of the MFCC coefficients.

3.3 Fuzzy C-means clustering

In order to verify the "discriminating abilities" of each feature, researchers have used different techniques such as cluster analysis, distance measures, entropy analysis, and other related methods [26]. We have chosen to use cluster analysis in order to select the best combination of features that gives the highest classification accuracy. In particular, we have used the "*Fuzzy C-Means Clustering*". Fuzzy C-Means is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. This technique was originally introduced by James C. Bezdek in [32] as an improvement on earlier clustering methods. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters [33].

We have extracted RMS-LPS, M-DWT, V-DWT, SF, %LEF, and R-ZC at frame-level where each frame is of 20 ms duration. Each clip was of 3 s duration containing 150 frames in each clip. After extracting those features at frame-level, we have taken the mean of 150 values of each feature to get a single feature value for each clip. For instance, we get 150 values of M-DWT and V-DWT, each belonging to a single frame. After that, we take the mean of 150 M-DWTs and the mean of 150 V-DWTs to get a single M-DWT and

V-DWT feature vector for each clip. In case of R-ZC, we have first calculated the number of zero-crossings in each frame and then we subtracted the minimum zero-crossings from the maximum zero-crossings within a clip to get R-ZC. The 12 coefficients of both LPC and V-12MFCC were extracted at clip-level, where each clip was of 3 s duration. The feature vector of each clip consists of: a single value of RMS-LPS, M-DWT, V-DWT, SF, %LEF and R-ZC, 12 coefficients of LPC, and 12 coefficients of MFCC.

We have applied the Fuzzy C-Means clustering algorithm to find the contribution of each feature in classifying audio data into one of three different classes: *Music*, *Speech*, and *Speech+Music* (i.e., Speech with background Music). We have, then, applied the algorithm on all possible combinations of those features in order to determine the best combination. Since we are only examining the contribution of each feature, we have selected few audio samples from our audio database to extract those features. We have selected 50 audio samples of each class, i.e., 50 samples of Music, 50 samples of Speech, and 50 samples of Speech+Music. In the samples of both Speech and Speech+Music, the language was "English" and the speaker was "Male".

3.3.1 Individual feature contribution to classification

We have taken into account the percentage accuracy of each class to be above 80% in order to consider the feature's contribution for that class as significant. Tables 1 and 2 show that RMS-LPS, SF, and V-12MFCC are good features for classification of all three classes. It is obvious that the data samples cannot be perfect, i.e. there lies some ambiguities among the samples belonging to the same class. For example, in the samples of Speech+Music, varying volume of background music makes speech dominant or music dominant, which explains some of those ambiguities. Similarly, in the case of speech, one may have some noise in the background which could be mistaken by the classifier as background music.

While investigating the variance of MFCC, we have applied the Fuzzy C-Means clustering algorithm on each of the 12 coefficients as shown in Table 3. We have found that the first 4 coefficients give the same results when using the 12 coefficients, as shown in Table 4. In fact, we applied the same experiment for 5, 6, up to 11 coefficients without any notable change in the clustering accuracy. Therefore, the first 4 MFCC coefficients (V-4MFCC) seem to suffice for carrying out the audio classification.

Regarding the use of wavelets as features, there exist many families of wavelets that can be considered, like 'Haar wavelet', 'Daubechies wavelets', 'Meyer wavelet', 'Mex-

Table 1 Clustering results for RMS-LPS and SF

	Classes		
Cluster	Music (%)	Speech (%)	Speech+Music (%)
RMS of lowpass signal			
Music	92	6	0
Speech	8	92	12
Speech+Music	0	2	88
Total	100	100	100
Spectral flux			
Music	92	2	0
Speech	8	84	10
Speech+Music	0	14	90
Total	100	100	100

Table 2 Clustering results for V-12MFCC

	Classes		
Cluster	Music (%)	Speech (%)	Speech +Music (%)
Variance of MFCC (12 Coefficient)			
Music	98	0	6
Speech	0	86	2
Speech+Music	2	14	92
Total	100	100	100

Table 3 Clustering results for each coefficient of the variance of MFCC

	Classes		
Coefficient no.	Music (%)	Speech (%)	Speech+Music (%)
1	M = 80	M = 0	M = 20
	S = 0	S = 60	S = 40
	SM = 46	SM = 0	SM = 54
2	M = 96	M = 0	M = 4
	S = 0	S = 72	S = 28
	SM = 4	SM = 10	SM = 86
3	M = 92	M = 0	M = 8
	S = 0	S = 62	S = 38
	SM = 22	SM = 0	SM = 78
4	M = 92	M = 0	M = 8
	S = 0	S = 76	S = 24
	SM = 28	SM = 8	SM = 64
5	M = 98	M = 0	M = 2
	S = 2	S = 38	S = 60
	SM = 4	SM = 24	SM = 72
6	M = 90	M = 0	M = 10
	S = 0	S = 70	S = 30
	SM = 10	SM = 8	SM = 82
7	M = 98	M = 0	M = 2
	S = 2	S = 64	S = 34
	SM = 4	SM = 26	SM = 70
8	M = 100	M = 0	M = 0
	S = 0	S = 58	S = 42
	SM = 10	SM = 24	SM = 66
9	M = 96	M = 0	M = 4
	S = 0	S = 46	S = 54
	SM = 4	SM = 18	SM = 78
10	M = 100	M = 0	M = 0
	S = 2	S = 46	S = 52
	SM = 28	SM = 2	SM = 70
11	M = 94	M = 0	M = 6
	S = 4	S = 68	S = 28
	SM = 26	SM = 14	SM = 60
12	M = 94	M = 0	M = 6
	S = 0	S = 52	S = 48
	SM = 12	SM = 24	SM = 64

Table 4 Clustering results for V-4MFCC

Cluster	Classes		
	Music (%)	Speech (%)	Speech+Music (%)
Variance of MFCC (4 Coefficient)			
Music	98	0	6
Speech	0	86	2
Speech+Music	2	14	92
Total	100	100	100

ican hat wavelet', and others. We have investigated Haar wavelets, Meyer wavelets, and two types of daubechies wavelets DB2 and DB15 [34]. The results show that features extracted when using Meyer or DB15 wavelets do not contribute much to the process of classification. The results for the Haar wavelets, however, indicate that they performed more accurate clustering than that of DB2 wavelets, as shown in Tables 5 and 6. Hence, we have only considered Haar wavelets and discard the rest. From this point on, when we refer to discrete wavelet transform, we mean Haar discrete wavelet transform.

Tables 5, 6 and 7 further indicate that M-DWT and LPC do not contribute much to the process of audio classification. However, it could be possible that they may be useful when used with other features. Table 5 clearly shows that V-DWT is a good feature to classify Music and can be useful for Speech+Music as well. In addition, Table 7 shows that the percentage of low energy frames is a good feature to identify speech, whereas Table 8 shows that R-ZC can be useful to classify Music and Speech data but not Speech+Music data.

3.3.2 Contribution of sets of features to classification

After studying the individual contribution of each feature in the classification process, we elaborate on choosing a subset of those features that maximize classification accuracy and at the same time reduce computational time by choosing the smallest such set. Selection of a proper feature subset is not an easy task. For this reason, we applied the fuzzy C-means clustering algorithm on all possible sets of features. As it is not feasible to show the results for all the combinations of features, we have included results for sets of features that have given the highest clustering accuracy. Two different sets of features have given the highest clustering accuracy. Both sets consist of three features, two of which exist in both, viz R-ZC and V-12MFCC. The third feature in the first set is SF and the one in the second set is %LEF. Per class accuracy for both sets is the same as shown in Tables 9 and 10. After that, we repeated the same procedure including V-4MFCC instead of V-12MFCC. In this case, the highest clustering accuracy with minimum number of features was achieved with only V-4MFCC.

Table 5 Clustering results for the mean and variance of "Haar" discrete wavelet transform

Cluster	Classes		
	Music (%)	Speech (%)	Speech+Music (%)
Mean of Haar DWT			
Music	32	14	4
Speech	36	58	44
Speech+Music	32	28	52
Total	100	100	100
Variance of Haar DWT			
Music	100	48	0
Speech	0	50	32
Speech+Music	0	2	68
Total	100	100	100

Table 6 Clustering results for the mean and variance of "DB2" discrete wavelet transform

Cluster	Classes		
	Music (%)	Speech (%)	Speech +Music (%)
Mean of DB2 DWT			
Music	18	8	4
Speech	41	64	40
Speech+Music	40	28	56
Total	100	100	100
Variance of DB2 DWT			
Music	100	52	0
Speech	0	46	34
Speech+Music	0	2	66
Total	100	100	100

Table 7 Clustering results for linear predictor coefficients and the percentage of low energy frames

Cluster	Classes		
	Music (%)	Speech (%)	Speech+Music (%)
LPC			
Music	68	14	16
Speech	0	32	20
Speech+Music	32	54	64
Total	100	100	100
%LEF			
Music	66	0	4
Speech	4	96	52
Speech+Music	30	4	44
Total	100	100	100

Table 8 Clustering results for the range of zero crossings

Cluster	Classes		
	Music (%)	Speech (%)	Speech+Music (%)
R-ZC			
Music	100	0	6
Speech	0	90	62
Speech+Music	0	10	32
Total	100	100	100

Table 9 Clustering result for SF, R-ZC, and V-12MFCC

	Classes		
Cluster	Music (%)	Speech (%)	Speech+Music (%)
SF, R-ZC and V-12MFCC			
Music	98	0	4
Speech	0	86	2
Speech+Music	2	14	94
Total	100	100	100

Table 10 Clustering result for LEF, R-ZC, and V-12MFCC

	Classes		
Cluster	Music (%)	Speech (%)	Speech +Music (%)
%LEF, R-ZC and V-12MFCC			
Music	98	0	4
Speech	0	86	2
Speech+Music	2	14	94
Total	100	100	100

After applying the clustering technique and short listing the potential discriminative features, we apply a classification scheme, as further elaborated in the next section.

4 Classification frameworks

In order for a classifier to be successful, it has to possess certain characteristics. First, due to the variability of music and speech signals, the classifier must be able to generalize from a relatively little amount of training data. Second, the notion of speech and music may differ from one application to another. For example, speech with background music may be considered as speech or music depending on the relative volume of the music as compared to speech. Hence, the classifier must adapt to different situations in order to give accurate results. In addition, since audio data is composed of a large amount of information size-wise, a practical classifier must be fast and simple.

We have considered two major approaches to carry out the classification, Artificial Neural Networks (ANN) and Hidden Markov Models (HMM). Two types of feedforward neural network topologies have been investigated, the Multilayer Perceptron (MLP) and Radial Basis Functions (RBF). In this section, we will briefly introduce the three classification frameworks and then describe our experimental setup and show the results for each framework.

4.1 Multilayer perceptron (MLP)

The Multilayer Perceptron (MLP) network is probably the most often considered member of the ANN family in classification. The main reason for this is its ability to model simple as well as very complex functional relationships. An MLP network consists of an input layer of source nodes, one or more hidden layers of computation nodes, and an output layer of computation nodes, as shown in Fig. 2. The

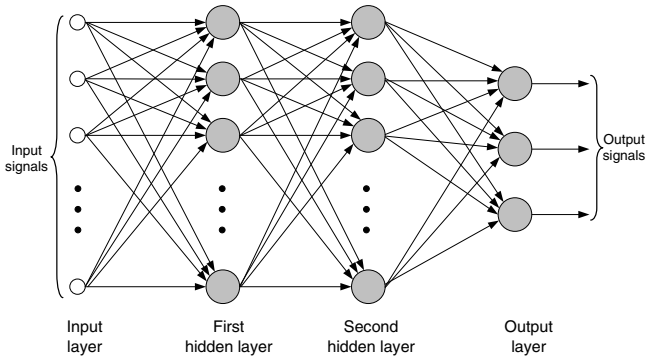


Fig. 2 Multilayer perceptron with two hidden layers

input signal propagates through the network in a forward direction, on a layer-by-layer basis. MLP networks successfully solve some difficult problems by training them in a *supervised* manner with a highly popular algorithm known as the *error back-propagation algorithm* or simply *back-propagation algorithm*. The back-propagation algorithm is based on the error-correction learning rule which requires pre-existing training patterns, and involves a forward propagation step followed by a backward propagation step.

We have employed an MLP network consisting of one hidden layer. The reason behind our choice of one hidden layer is the fact that continuous feedforward neural networks with a single hidden layer and a nonlinear sigmoidal activation function provide good approximations to arbitrary decision regions [35, 36]. Each neuron in the input layer corresponds to a feature value of the input feature vector. The output layer consists of three neurons, each corresponding to a class (Music, Speech, and Speech+Music). The MLP has been chosen to be *fully connected*, i.e., a neuron in any layer is connected to all neurons of the previous layer.

Prior to training, small random numbers have been generated to initialize weights on each communication link, called connection, between neurons. In addition, the input features have been normalized as neural networks risk saturation.[1] With regard to the number of neurons in the hidden layer, we have used 5 neurons and 10 neurons. After carrying out the classification process and comparing the results, we decided to work with only 5 neurons as the accuracy was not greatly affected by the increase in the number of neurons, although the processing time of the classifier increased substantially. Due to the nonlinear behavior of patterns, we have used a sigmoidal function as an activation function. We have used tan sigmoid function with output values between -1 and 1 for the hidden neurons, and log sigmoid function for the output neurons, with values ranging between 0 and 1.

4.2 Radial basis functions (RBF)

RBFs are feedforward network that are used in a wide variety of contexts such as function approximation, pattern

[1] Saturation refers to the situation where synaptic weights change slowly causing a very long training time if feature vectors contain values greater than 1.

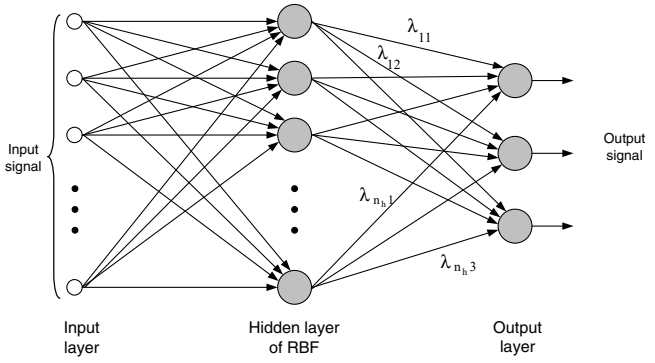


Fig. 3 A general RBF network

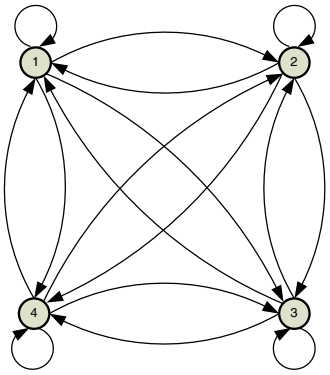


Fig. 4 An ergodic hidden Markov model

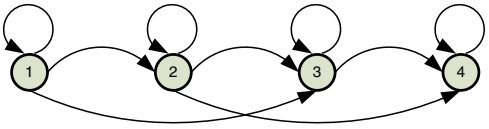


Fig. 5 A left-to-right hidden Markov model

recognition and time series prediction. Learning in RBF networks involves only one layer with lesser computations. This results in a reduction in the training time in contrast to MLP that uses back propagation algorithm to update the weights of all neurons. These features make RBF attractive in many practical problems.

The construction of an RBF network, in its most basic form, consists of three layers: the input layer of source nodes, the middle layer which is the only hidden layer in the network that applies a nonlinear transformation, and the output layer which is linear as shown in Fig. 3. Every input node is connected to all nodes of the hidden layer through unity weights (direct connection).

A *reduced RBF* classifier has been considered in our experimentation. A reduced RBF network is an RBF network in which the number of centers is less than the total number of input samples, as opposed to a *complete RBF*, where the number of centers is equal to that of the input samples. The number of centers used equals to 3, the number of classes. We have extracted the centers from the input features by using the Fuzzy C-Means clustering algorithm. As a learning algorithm we have used the average square error algorithm.

4.3 Hidden Markov models

Hidden Markov Models belong to a class of statistical models that employ the statistical properties of signals in carrying out recognition and/or classification. Other statistical models in this domain include Gaussian processes, Poisson processes, and Markov processes. According to Rabiner [37]:

> An HMM is a doubly stochastic process with an underlying stochastic process that is not observable (it is hidden), but can only be observed through another set of stochastic processes that produce the sequence of observed symbols.

A Markov chain or process is a sequence of events, usually called states, the probability of each of which is dependent only on events preceding it. A Hidden Markov Model (HMM) represents stochastic sequences as Markov chains where the states are not directly observed, but are associated with a probability density function. A general HMM is assumed to have a full state transition matrix, i.e. transitions can be made from any state in some way to any other state.

Such models are called *ergodic*, an example of which is shown in Fig. 4. In non-ergodic models, transitions can only be made to a state whose index is as large or larger than the index of the current state. Such models are called *left-to-right models*. Figure 5 shows an example left-to-right HMM.

In order to compare results obtained from using neural networks to those using statistical models like HMM's, we have trained three HMMs, one for each class. After computing the log-likelihood that a single sample generates, if the ith model, where $1 \leq i \leq 3$, gives the highest value, the sample is classified to belong to class i. This is called sequence classification.

4.4 Experimental setup

The experiments were carried out using a database of music, speech, and speech+music data. All speech and speech+music data were conversational and included examples from both genders. The following languages were represented: American English, Urdu, Japanese, Spanish, and Hebrew. The audio samples were extracted from documentaries and from different movies. There were approximately 2.25 h of speech, 2.72 h of music and 0.62 h of speech/music data distributed over 3-s audio files as shown in Table 11. Audio samples consist of 16-bit, 44.1 kHz, mono PCM wave files.

5 Experimental results

The three classifiers, MLP, RBF, and HMM have been applied on various sets of features. The results of each

Table 11 Audio data samples

Language	Number of samples		
	Music	Speech	Speech+Music
English	–	50	50
Urdu	–	1543	100
Japanese	–	427	336
Spanish	–	542	154
Hebrew	–	140	100
Total	3268	2702	740

experimentation on a set of features have been recorded in a table, where the accuracy for music is denoted by M, the accuracy for speech is denoted by S, and the accuracy for speech+music is denoted by S+M.

First, we have examined seven features, namely RMS-LPS, M-DWT, V-DWT, SF, %LEF, R-ZC, and LPC. Table 12 shows the percentage accuracies achieved by applying the three classifiers on this set.

MLP performance on English data has been consistently well, whereas HMM performed reasonably well, regardless of the language. RBF performed well in recognizing speech, but very poorly with speech+music data. HMM's outperformed the other two classifiers in recognizing speech+music in all languages except English and Urdu.

Next, we added the V-12MFCC to the previous set of features and ran the three classifiers to get the results shown in Table 13.

HMM outperformed RBF and MLP for English data and for "all" data, which consists of audio samples from all languages. MLP outperformed the overall classification accuracy of HMM and RBF for Urdu data. RBF continues exhibiting low accuracy on speech+music data.

When the Fuzzy C-Mean clustering algorithm was applied on individual features in Sect. 3.3.1, it was shown that M-DWT and %LEF do not seem to be good candidates for discriminating speech and music. Therefore, we have removed M-DWT and %LEF from the previous two sets and then applied the classification process on the new sets, which has given the results shown in Tables 14 and 15.

When the V-12MFCC was not used, it was evident that an overall degredation in performance has occured in the three classifiers. It seems that M-DWT and %LEF contribut to the speech+music classification accuracy most. However, when V-12MFC was added, Table 15 shows a significant increase of performance of the MLP classifier in all language categories. Such improvements were not evident in the other two classifiers.

Next, we consider using the variance of only the first four coefficients of MFCC, V-4MFCC, instead of V-12MFCC as the clustering results in Sect. 3.3.1 have indicated similar performance traits to V-12MFCC. Table 16 shows the classification performance after adding V-4MFCC to the original seven features (before removing M-DWT and %LEF), and Table 17 shows the performance after removing the two features.

One interesting result is the 100% accuracy achieved by MLP for English data after removing M-DWT and %LEF, and adding V-4MFCC to the original set of seven features. We can also notice a degradation in performance of classifying speech+music in MLP and HMM, which was reflected in the overall performance of the HMM framework, specifically.

It is clear that the classification results in this section confirm our findings of the clustering study of Sect. 3. This highlights the importance of carrying out clustering analysis before considering certain features for classification purposes.

Table 12 Classification results for RMS-LPS, M-DWT, V-DWT, SF, %LEF, R-ZC, and LPC

	Accuracy with MLP (%)				Accuracy with RBF (%)				Accuracy with HMM (%)			
Language	M	S	S+M	Total	M	S	S+M	Total	M	S	S+M	Total
English	100	93.33	93.33	95.55	100	100	30	76.67	100	84	73.33	85.78
Urdu	73.33	76.66	53.33	67.77	90	25	5	40	82	97.34	44	74.45
Japanese	89	85	31	68.33	86.57	94.03	0	60.20	58.40	76.60	44.20	59.73
Spanish	93.48	63.04	26.09	60.87	70	43.33	3.33	38.89	99.13	57.39	56.09	70.87
Hebrew	11.54	80.77	7.7	33.34	95	70	15	60	96.67	69.33	54.67	73.56
All	85.85	82.65	32.42	66.97	77.7	68.92	0	48.87	91.17	71.71	52.61	71.83

Table 13 Classification results for RMS-LPS, M-DWT, V-DWT, SF, %LEF, R-ZC, LPC, and V-12MFCC

Language	Accuracy with MLP (%)				Accuracy with RBF (%)				Accuracy with HMM (%)			
	M	S	S+M	Total	M	S	S+M	Total	M	S	S+M	Total
English	100	100	90	96.67	100	100	90	96.67	100	100	90.67	96.89
Urdu	95	90	100	95	25	65	0	30	55.33	100	64	73.11
Japanese	92.54	89.55	70.15	84.08	2.98	10.45	7.46	6.96	97	88	38.20	74.40
Spanish	93.33	86.67	23.33	67.78	80	30	23.33	44.44	95.65	34.78	78.26	69.56
Hebrew	95	65	75	78.33	90	0	25	38.33	91.33	70	28	63.11
All	86.48	57.43	60.81	68.24	76.35	31.76	0	36.04	97.66	52.79	80.81	77.09

Table 14 Classification results for RMS-LPS, V-DWT, SF, R-ZC, and LPC

	Accuracy with MLP (%)				Accuracy with RBF (%)				Accuracy with HMM (%)			
Language	M	S	S+M	Total	M	S	S+M	Total	M	S	S+M	Total
English	100	100	70	90	100	100	70	90	100	58.67	93.33	84
Urdu	70	95	25	63.33	55	75	30	53.33	34.67	96.67	52.67	61.33
Japanese	91.04	82.09	32.83	68.65	85.07	89.55	0	58.21	82.80	70.60	33.20	62.20
Spanish	93.33	0	26.67	40	76.67	36.67	0	37.78	94.35	33.48	72.61	66.81
Hebrew	100	75	0	58.33	95	70	0	55	93.33	51.33	58.67	67.78
All	88.51	27.70	62.84	59.68	33.78	33.78	16.90	28.15	87.48	57.21	49.48	64.72

Table 15 Classification results for RMS-LPS, V-DWT, SF, R-ZC, LPC, and V12-MFCC

	Accuracy with MLP (%)				Accuracy with RBF (%)				Accuracy with HMM (%)			
Language	M	S	S+M	Total	M	S	S+M	Total	M	S	S+M	Total
English	100	100	90	96.67	100	100	90	96.67	100	81.33	81.33	87.56
Urdu	80	100	95	91.67	25	60	0	28.33	47.33	98	42.67	62.67
Japanese	100	89.55	61.20	83.58	0	4.48	10.45	4.98	87.20	76	34.60	65.93
Spanish	93.33	83.33	43.34	73.33	80	26.67	23.33	43.33	94.78	47.83	67.39	70
Hebrew	95	70	75	80	90	0	45	45	66	57.33	42.67	55.33
All	86.48	37.16	78.38	67.34	71.62	23.65	0	31.76	91.62	49.01	62.25	67.63

Table 16 Classification results for RMS-LPS, M-DWT, V-DWT, SF, %LEF, R-ZC, LPC, and V-4MFCC

	Accuracy with MLP (%)				Accuracy with RBF (%)				Accuracy with HMM (%)			
Language	M	S	S+M	Total	M	S	S+M	Total	M	S	S+M	Total
English	100	100	93.33	97.78	100	100	90	96.67	100	66.67	84	83.56
Urdu	90	65	85	80	15	50	25	30	68.67	96.67	49.33	71.56
Japanese	82.1	95.52	71.64	83.09	2.98	5.97	11.94	6.96	73	84	31.20	62.73
Spanish	86.67	90	16.67	64.45	76.67	30	26.67	44.45	90.44	44.35	62.61	65.80
Hebrew	95	60	90	81.67	90	0	20	36.67	89.33	84.67	27.33	67.11
All	86.04	75.67	39.19	66.97	72.97	28.38	0	33.78	94.32	84.05	33.51	70.63

Table 17 Classification results for RMS-LPS, V-DWT, SF, R-ZC, LPC, and V-4MFCC

	Accuracy with MLP (%)				Accuracy with RBF (%)				Accuracy with HMM (%)			
Language	M	S	S+M	Total	M	S	S+M	Total	M	S	S+M	Total
English	100	100	100	100	100	100	90	96.67	100	89.33	85.33	91.56
Urdu	100	95	40	78.33	0	45	25	23.33	60.67	96.67	42	66.64
Japanese	95.52	85.07	71.64	84.08	0	2.98	11.94	4.97	94.80	79.80	25.60	66.73
Spanish	93.33	43.33	56.67	64.44	80	26.67	30	45.56	97.39	50	66.52	71.31
Hebrew	95	80	65	80	90	0	45	45	48	76.66	19.33	48
All	83.11	39.86	64.2	62.39	70.27	20.27	0	30.18	95.32	95.32	7.30	65.98

6 Conclusion and future work

Many techniques have been proposed in the literature for speech/music classification. In order to achieve acceptable performance, most of them require a large amount of training data, rendering them difficult for retraining and adaptation to new conditions. Other techniques are rather context oriented, as they have been tested on specific applications, such as speech/music classification in radio programs or in the context of broadcast news transcription. We have conducted extensive experimentation on a diverse set of audio data using three classification frameworks and introducing features that have not been used earlier for music-speech classification.

The results clearly show that RBF networks give satisfactory results only for the English language. Since RBF networks depend on the centers of clusters, the results indicate that for all languages, except English, the center of each cluster has not been correctly chosen by the classification algorithm.

Both, MLP networks and HMMs have given good results. A disadvantage of using HMMs is that it requires long training and testing time as compared to MLP. With 35 samples for training and 15 samples for testing, HMMs took close to 21 s for training and 1.3 s for testing, whereas MLPs

took 7.3 s for training and 0.046 s for testing. Also, HMMs need to be trained for each audio class separately, which requires more memory space. MLP is trained only once for all the audio classes simultaneously and the synaptic weights are stored once.

After investigating nine major audio features, one can conclude that applying an MLP classification framework on six of them, namely the range of zero-crossings, the variance of the Haar discrete wavelet transform, the root mean square of a lowpass signal, the spectral flux, the linear predictive coefficients, and the variance of four Mel frequency cepstral coefficients, has given the best results, achieving a 100% classification accuracy for English. As other languages have not achieved such accuracy, one must explore more audio features that behave similarly for different languages. Otherwise, one may need to have a closer look at different languages, closely studying their distinctive properties and the degree of similarity to music in order to justify the varying performance. There is, also, a need for an audio data benchmark that can be shared by researchers interested in music speech classification to facilitate more objective comparisons of various approaches.

Acknowledgements The authors acknowledge the support of King Fahd University of Petroleum and Minerals in the development of this work.

References

1. Scheirer, E., Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'97, IEEE), Vol. 2, pp. 1331–1334 (1997)
2. Saad, E.M., El-Adawy, M.I., Abu-El-Wafa, M.E., Wahba, A.A.: A multifeature speech/music discrimination system. In: Proceedings of the 19th National Radio Science Conference (NRSC'02, IEEE), pp. 208–213 (2002)
3. John Saunders: Real-time discrimination of broadcast speech/music. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96, IEEE), Vol. 2, pp. 993–996 (1996)
4. Carey, M.J., Parris, E.S., Lloyd-Thomas, H.: A comparison of features for speech, music discrimination. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'99, IEEE), Vol. 1, pp. 149–152 (1999)
5. Parris, E.S., Carey, M.J., Lloyd-Thomas, H.: Feature fusion for music detection. In: Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'99), pp. 2191–2194 (1999)
6. Chou, W., Gu, L.: Robust singing detection in speech/music discriminator design. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01, IEEE), Vol. 2, pp. 865–868 (2001)
7. Pinquier, J., Sénac, C., André-Obrecht, R.: Speech and music classification in audio documents. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02, IEEE), Vol. 4, pp. 4164–4164 (2002)
8. Pinquier, J., Rouas, J.-L., André-Obrecht, R.: Robust speech/music classification in audio documents. In: Proceedings of the 7th International Conference on Spoken Language (ICSLP'02), Vol. 3, pp. 2005–2008 (2002)
9. Pinquier, J., Rouas, J.L., André-Obrecht, R.: A fusion study in speech/music classification. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03, IEEE), Vol. 2, pp. II-17–II-20 (2003)
10. Harb, H., Chen, L.: Robust speech music discrimination using spectrum's first order statistics and neural networks. In: Proceedings of the 7th International Symposium on Signal Processing and its Applications, IEEE, Vol. 2, pp. 125–128 (2003)
11. Harb, H., Chen, L., Auloge, J.Y.: Speech/music/silence and gender detection algorithm. In: Proceedings of the 7th International Conference on Distributed Multimedia Systems (DMS'01), pp. 257–262 (2001)
12. Karnebäck, S.: Discrimination between speech and music based on a low frequency modulation feature. In: Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH'01), pp. 1891–1894 (2001)
13. Wang, W.Q., Gao, W., Ying, D.W.: A fast and robust speech/music discrimination approach. In: Proceedings of the Information, Communications & Signal Processing (ICICS-PCM'03, IEEE), Vol. 3, pp. 1325–1329 (2003)
14. El-Maleh, K., Klein, M., Petrucci, G., Kabal, P.: Speech/music discrimination for multimedia applications. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00, IEEE), Vol. 4, pp. 2445–2448 (2000)
15. Panagiotakis, C., Tziritas, G.: A speech/music discriminator based on rms and zero-crossings. IEEE Trans. Multimedia (2004)
16. Shao, X., Xu, C., Kankanhalli, M.S.: Applying neural network on content-based audio classification. In: Proceedings of the Fourth International Conference on Information, Communications and Signal Processing, IEEE, Vol. 3, pp. 1823–1825 (2003)
17. Lippens, S., Martens, J.P., De Mulder, T., Tzanetakis, G.: A comparison of human and automatic musical genre classification. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04, IEEE), Vol. 4, pp. IV-233–IV-236 (2004)
18. Srinivasan, S.H., Kankanhalli, M.: Harmonicity and dynamics-based features for audio. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04, IEEE), Vol. 4, pp. IV-321–IV-324 (2004)
19. Vesa Peltonen: Computational auditory scene recognition. Master's thesis, Department of Information Technology, Tampere University of Technology, Finland (2001)
20. Tzanetakis, G., Essl, G., Cook, P.: Automatic musical genre classification of audio signals. In: Proceedings of the International Symposium on Music Information Retrieval (ISMIR'01), pp. 205–210 (2001)
21. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Trans. Speech Audio Proc. **10**(5), 293–302 (2002)
22. Lu, L., Zhang, H.-J., Li, S.Z.: Content-based audio classification and segmentation by using support vector machines. ACM Mult. Sys. J. **8**(6), 482–492 (2003)
23. Bugatti, A., Flammini, A., Migliorati, P.: Audio classification in speech and music: A comparison between a statistical and a neural approach. EURASIP J. Appl. Sig. Proc. **4**, 372–378 (2002)
24. Lu, L., Jiang, H., Zhang, H.-J.: A robust audio classification and segmentation method. In: Proceedings of the 9th ACM International Conference on Multimedia (MM'01, ACM), pp. 203–211 (2001)
25. Lu, L., Zhang, H.-J., Jiang, H.: Content analysis for audio classification and segmentation. IEEE Trans. Speech Audio Proc. **10**(7), 504–516 (2002)
26. Beierholm, T., Baggenstoss, P.M.: Speech music discrimination using class-specific features. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04, IEEE), Vol. 2, pp. 379–382 (2004)
27. Hoyt, J.D., Wechsler, H.: Detection of human speech in structured noise. In: Proceedings of the International Conference on Neural Networks, IEEE, Vol. 7, pp. 4493–4496 (1994)
28. Li, D., Sethi, I.K., Dimitrova, N., McGee, T.: Classification of general audio data for content-based retrieval. Patt. Recog. Lett. **22**(5), 533–544 (2001)

29. Tzanetakis, G., Cook, P.: A framework for audio analysis based on classification and temporal segmentation. In: EUROMICRO Workshop on Music Technology and Audio Processing, IEEE, Vol. 2, pp. 61–67 (1999)
30. Lambrou, T., Kudumakis, P., Speller, R., Sandler, M., Linney, A.: Classification of audio signals using statistical features on time and wavelet transform domains. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98, IEEE), Vol. 6, pp. 3621–3624 (1998)
31. Delfs, C., Jondral, F.: Classification of transient time-varying signals using dft and wavelet packet based methods. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98, IEEE), Vol. 3, pp. 1569–1572 (1998)
32. Bezdek J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
33. Duda, R.O., Stork, D.G., Hart, P.E.: Pattern classification, 2nd edn. Wiley, New York (2001)
34. Kashif Saeed Khan, M.: Automatic classification of speech and music in digitized audio. Master's thesis, King Fahd University of Petroleum andMinerals, Dhahran, Saudi Arabia (2005)
35. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Math. Con. Sig. Sys. **2**(4), 303–314 (1989)
36. Mammone, R.J. (ed.): Artificial neural networks for speech and vision. Chapman & Hall Neural Computing, 1st edn. Chapman & Hall, London (1994)
37. Rabiner, L.R., Juang, B.H.: An introduction to hidden markov models. IEEE ASSP Magazine **3**(1), 4–16 (1986)