Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105(Online)
Volume **1**, Number **2**, May **2016**

Robust Music Genre Classification Based on Sparse Representation and Wavelet Packet Transform with Discrete Trigonometric Transform

Shih-Hao Chen[1], Sung-Yuan Ko[1], and Shi-Huang Chen[2]

[1]Dept. of Information Engineering, I-Shou University, Taiwan, email: a9988123@yahoo.com.tw
[2]Dept. of CSIE, Shu-Te University, Taiwan, email: shchen@stu.edu.tw

Received October 2015; revised May 2016

Abstract. *This paper proposes a robust method for the application of music genre classification. The proposed method first uses moving average filter and Butterworth low-pass filter to partly eliminate the effect of fluctuation in short-term signal. Then, it makes use of the sparse representation based classification (SRC) and wavelet packet transform (WPT) with discrete trigonometric transforms (DTTs) to accurately classify and increase classification performance. Sparse representation based classification has been widely used for music genre classification via the primal-dual algorithm for linear programming to search the most compact representation of the signal in the digital domain. To investigate its performance, the proposed method is validated by comparison with various discrete cosine transform types and classification methods. Experimental results show that the accuracy of DCT-II orthogonal is clearly better than that of DCT-II non orthogonal via SRC classifier. Specifically, the best classification result with the odd orthogonal DCT-II is 89.7%, which is significantly better than the 86.69% accuracy rate obtained by the even orthogonal DCT-II both on the ISMIR 2004 Genre dataset. It is shown that the proposed method greatly improves the performances of previous music genre classification algorithms.*

Keywords: Best basis algorithm, Wavelet packet transform, Music genre classification, Sparse representation based classification.

1. **Introduction.** Due to the rapid growth and development of digital music content, automatic music genre classification has been a challenging task in the filed of Music Information Retrieval (MIR) [1]. Since a typical multimedia database often contains millions of audio clips, it is very difficult to manage such a large music database. It follows from previous researches [2][3] that audio signal usually carries evidence information in its genre. Hence the need to automatically recognize to which class a musical genre belong makes the automatic analysis of music signals and content-based musical information retrieval (MIR) an emerging research area. In general, an automatic music analysis is to make use of several characteristics that can capture the information about music content. Among these characteristics, music genre information is regarded as a principal one. Musical genres are the main top-level descriptors used by music dealers and librarians to organize their music collections [3]. It can be used to describe music as well as to structure music database [4]. However, musical genres have no strict definitions, as their boundaries vary with the public, marketing, historical, and cultural factors. Another problem is that most of current musical genre annotation is still performed manually [3][5]. The automatic musical genre classification is still one of the most important parts of MIR [1]. Many researchers have studied or proposed methods capable of automatically extracting music information by using a computational approach to structure and organize the musical genres [6].

Most of the music genre classification algorithms resort to the so-called bag-of-features approach [2], which models the audio signals via the long-term statistical distribution of their local spectral features.

In general, the most popular features used in recent studies could be roughly categorized into short-term and long-term features [3]. The short-term features, which can represent the spectrum of music, include spectral centroid, spectral roll-off, mel-frequency cepstral coefficient (MFCC), and etc. The long-term features, which can characterize either the variation of spectral shape or beat information, include low-energy [4], and beat histogram, and etc [3][7]. Most music classification systems so far are based on pattern recognition techniques to recognize the classes of music genre defined in the taxonomy. Once the features are extracted from an audio clip, a classifier will be employed to determine the genre of the given an audio clip. Several statistical techniques, such as neural networks, hidden Markov models (HMM), Gaussian mixture models (GMM), K-nearest neighbors (KNN) [3], sparse representation based classification (SRC), and support vector machines (SVM), have been employed for automatic musical genre classification.

On the other hand, various content-based analysis methods of music signal are proposed for music genre classification. Among these techniques, SRC, which were introduced by Wright et al. in [8], have been regarded as a new learning algorithm for various applications, such as face recognition [8] and image classification. The sparse representation is computed by the l_1-regularized least square method. To investigate its performance, the proposed method is validated by comparison with various discrete cosine transform types and classification methods. Experimental results show that the accuracy of DCT- II orthogonal is clearly better than that of DCT- II non orthogonal via SRC classifier. Specifically, the best classification result with the DCT- II odd orthogonal is significantly better than the Type II even orthogonal on the ISMIR 2004 Genre dataset. By using topology preserving non-negative matrix factorization (TPNMF) and SRC, instead of the 2D auditory temporal modulations and SRC, Y. Panagakis and C. Kotropoulos [9] managed to significantly improve the previous work [10] on classification performance. This paper compares the results of Y. Panagakis and C. Kotropouloss method [9] and builds a more robust music genre classification system by incorporating additional wavelet packet transform (WPT) with best cosine transform and the best wavelet packet basis via best basis algorithm (BBA). The application of a wavelet package transform can generate a wavelet decomposition that offers a richer signal analysis. The best basis algorithm is obtained by minimizing the Shannon entropy. The method proposed in this paper uses the Top-Down search strategy with cost function to select the best basis of WPT. In contrast to the conventional methods, it can be attributed to better feature extraction and classification accuracy. Experiments are carried out using the ISMIR2004 GENRE database with 6 types of music genres and about 1458 music clips. Experimental results show that the use of proposed method can obtain significant improvements in music genre classification accuracy. The average music genre classification accuracy rate of the proposed method can achieve 89.7%.

The rest of the paper is organized as follows. The proposed music genre classification system is presented in Sections II, including moving average filter and Butterworth low-pass filter, introduction to the wavelet package transform, wavelet package analysis with best basis algorithm, feature extraction, discrete trigonometric transform, and the introduction to the sparse representation based classification. Experimental results are described in Sections III. Finally, conclusions are given in Sections IV.

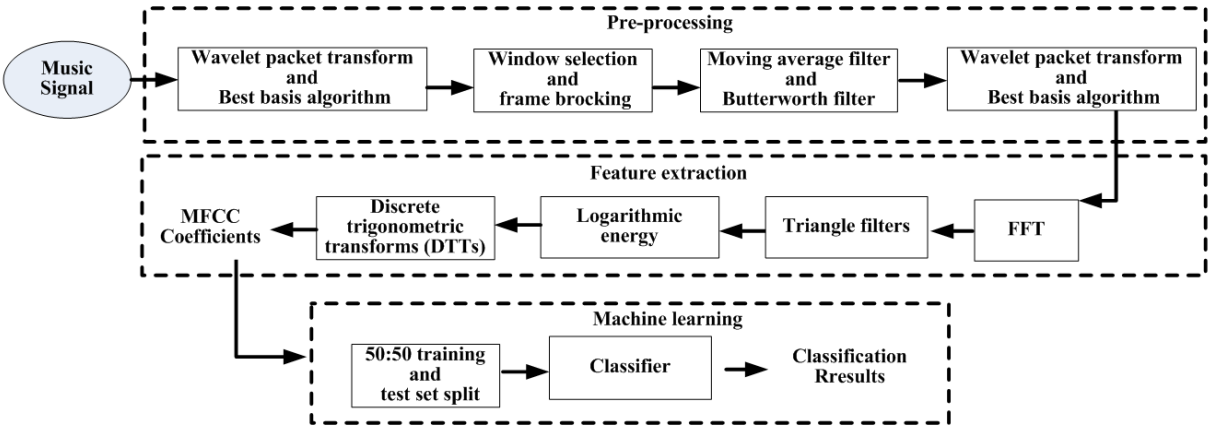
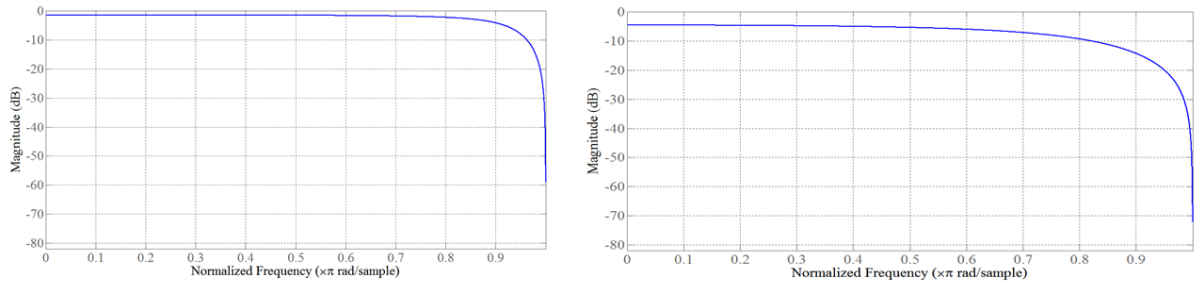


FIGURE 1. The proposed genre classification system calculation flow diagram

2. **Proposed Music Genre Classification System.** The proposed genre classification system consists of three phases: (1) pre-processing phase, (2) feature extraction phase, and (3) the machine learning phase. The pre-processing phase is composed of moving average filter / butterworth low-pass filter, frame blocking and window function selection, wavelet package transform with best basis algorithm. The feature extraction phase consists of fast Fourier transformation (FFT), triangle filters, logarithmic energy and discrete trigonometric transforms (DTTs). The machine learning phase is composed of 50:50 training and test set split and classifier. Fig. 1 shows the flow diagram of the proposed genre classification system. A detailed description of each module will be described below

2.1. **Moving average filter and Butterworth low-pass filter.** The moving average filter and Butterworth low-pass filter are the two commonly used methods in the field of digital signal processing. Butterworth low-pass filter discussed here is determined by the cutoff frequency C and the order of filter F. There are four examples shown in Fig. 2. The horizontal axis shows the normalized frequency (For example, assume that data sampling rate is 44100 Hz, design a 3th-order low-pass Butterworth filter with cutoff frequency of 8000 Hz, which corresponds to a normalized value of 0.3628), whereas the other axe indicates the magnitude (dB). This paper also applied MF-point moving average filter to an audio sound to reduce random noise while retaining a sharp step response. The MF-point moving average filter is depicted in Fig. 3



(a) C=20000 and F=3

(b) C=16000 and F=3

(c) C=12000 and F=3

(d) C=8000 and F=3

FIGURE 2. Butterworth low-pass filter with Cutoff frequency (C=8000, 12000, 16000 and 20000 Hz)

2.2. **Introduction to the Wavelet Package Transform.** A wavelet packet transform (WPT), which was first introduced by Coifman et al. [11], is shown in Fig. 4, where $h(k)$ and $g(k)$ are the analysis low-pass and high-pass filters, respectively. In addition, the symbol $\downarrow 2$ denotes the down-sampling by 2. The equations of WPT filtering operations is described as

$$a_i(k) = \sum_n h(n-2k)a_{i+1}(n) \tag{1}$$

$$d_i(k) = \sum_n g(n-2k)a_{i+1}(n) \tag{2}$$

where $a_i(k)$ and $d_i(k)$ are called the approximation and detail coefficients of the wavelet decomposition of ,$a_{i+1}(n)$ respectively.

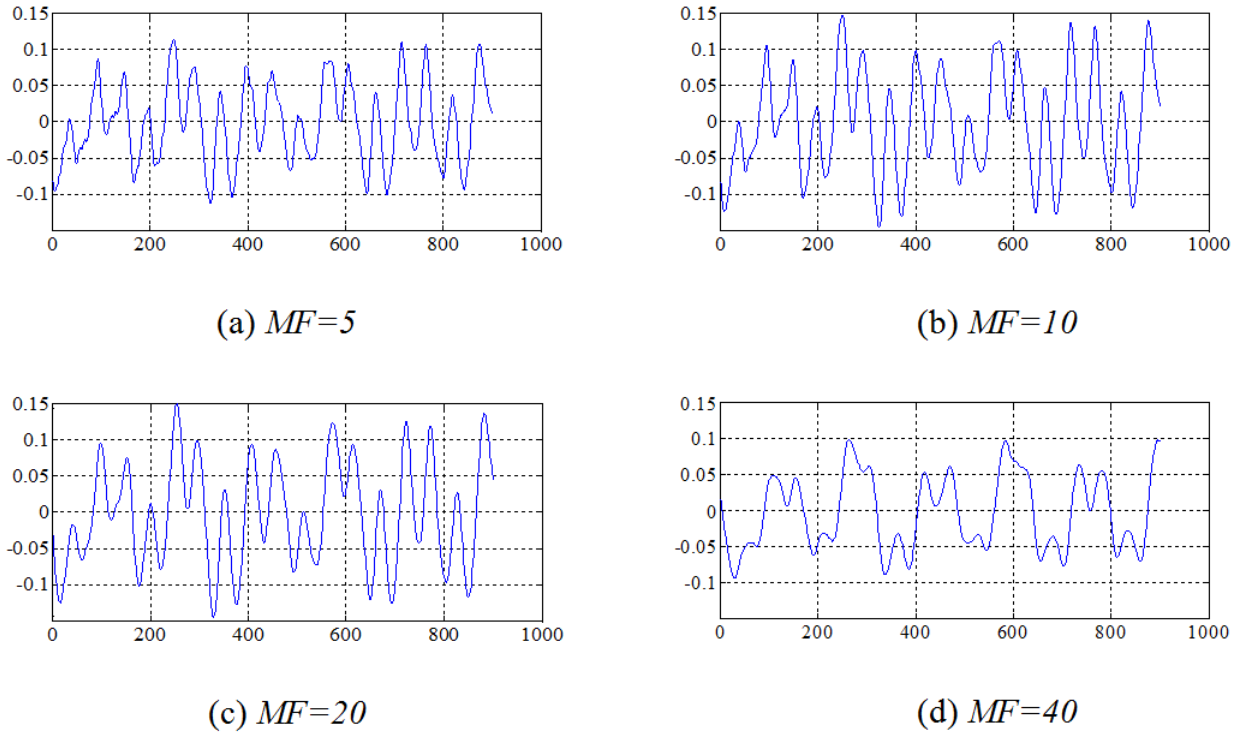


FIGURE 3. MF-point moving average filter to an audio sound (MF =5, MF =10, MF =20 and MF =40)

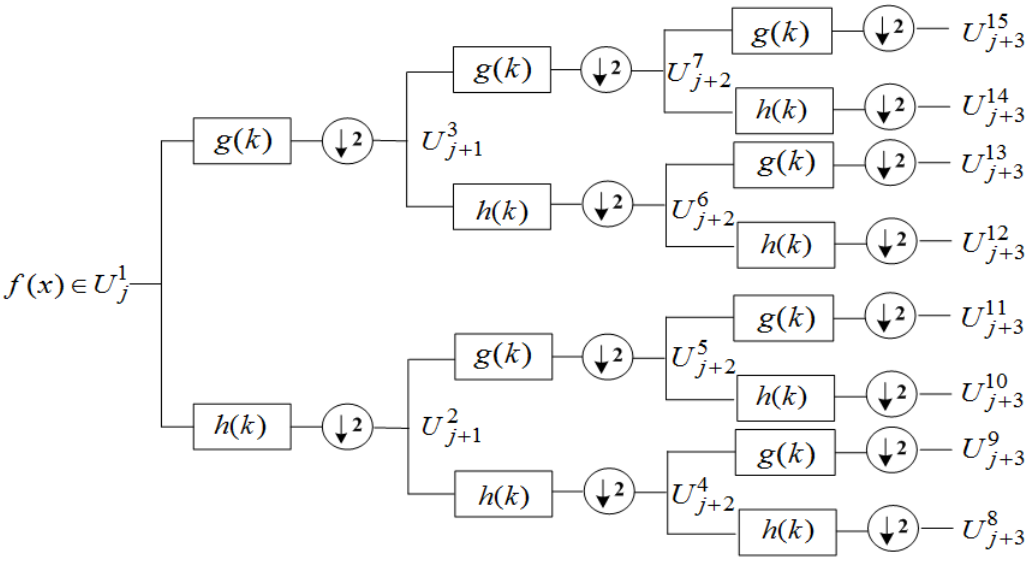


FIGURE 4. Three-level wavelet packet transform.

Since wavelet packet transform is a generalization of the dyadic wavelet transform (DWT), it is regarded as a more effective tool than the Fourier transform for audio processing. WPT provides good spectral and temporal resolutions through the filter bank structure in arbitrary regions of the time-frequency plane. WPT can easily transform discrete signal from the time domain into time frequency domain. The transformation product is a set of coefficients that represents the spectrum analysis as well as the spectral behavior of the signal. Therefore, the wavelet package transform is able to provide an optimal representation for music.

2.3. **Wavelet Package Analysis with Best Basis Algorithm.** The basic idea of wavelet package transform (WPT) is to concentrate energy of signal into part of trees, so it is important to find the best wavelet packet basis via best basis algorithm (BBA). An example of the wavelet packet tree with three-level decomposition is shown in Fig. 5.

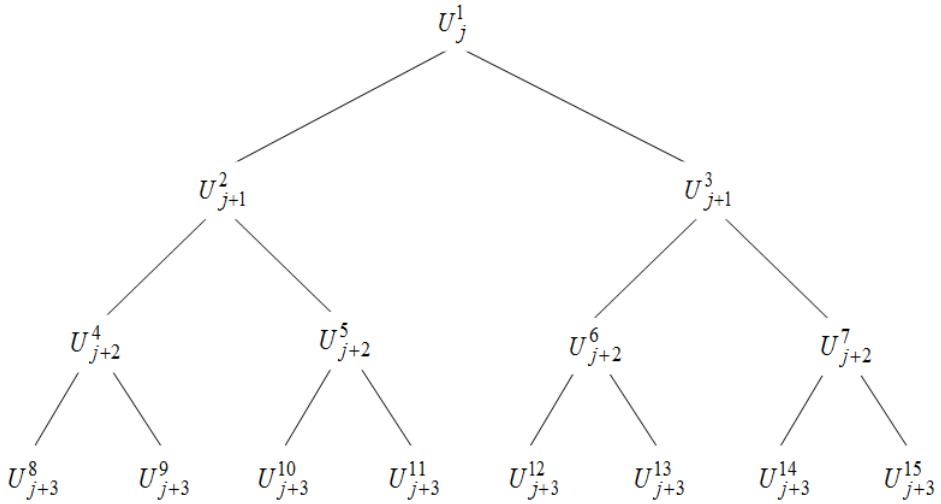


FIGURE 5. Three-level Wavelet Packet Tree Decomposition

The best basis algorithm is one of the important issue of the wavelet packet analysis. The basic idea of optimal wavelet packet decomposition based on cost function, namely, the Shannon entropy, is introduced to find the best wavelet packet (WP) base in music genre classification. Based on the above mentioned observations, the optimal basis is picked up by optimizing the information cost function. The algorithm proposed in this paper uses the top-down tree search strategy with cost function to select the best basis using basis selection method [11], [12]. This could be done by adopting Shannon entropy, a new method based on BBA is presented to minimizing the information cost function.

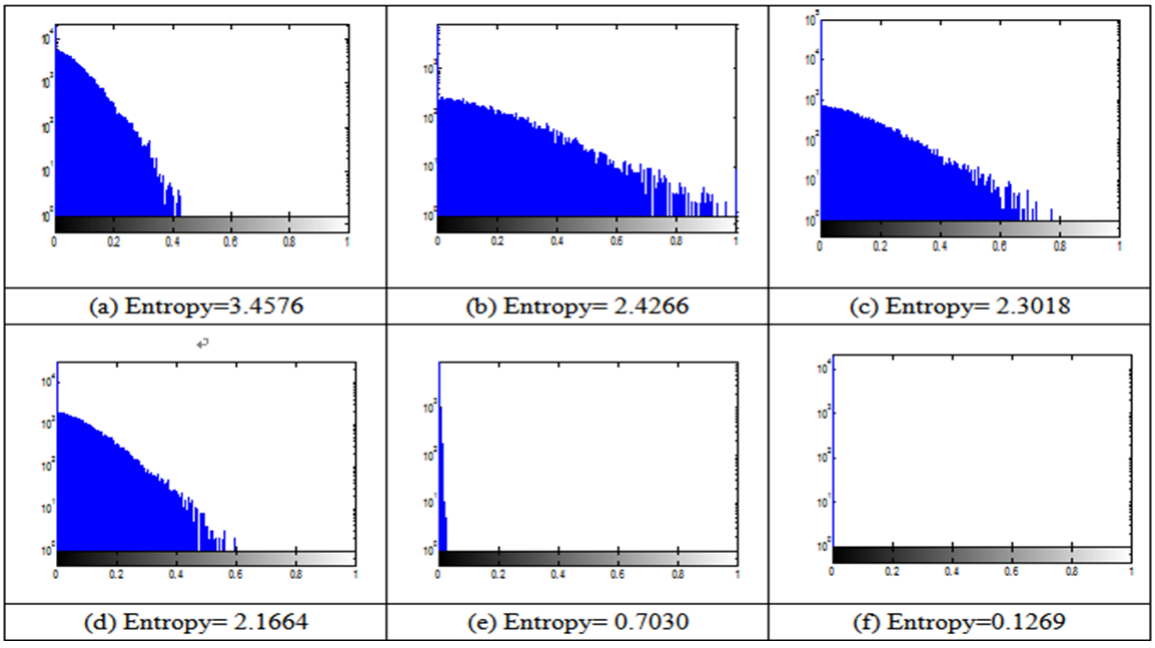


FIGURE 6. Six types of entropy were invited for evaluation.

A one dimension orthogonal wavelet packet base can be described by a binary-tree with the root node U_0^1 , the nodes without any child node are called the leaf nodes, and except the leaf nodes, each node U_j^n

has two child U_{j+1}^{2n} and U_{j+1}^{2n+1}. The binary-tree structure assures a simple algorithm for selecting the best wavelet packet base. For a given music signal, one can perform J-level full wavelet packet decomposition, and the wavelet packet coefficients at the node U_j^n can be represented as

$$U_j^n = U_{j+1}^{2n} \oplus U_{j+1}^{2n+1} \tag{3}$$

where $n = 1, \cdots, 2^J - 1$ and $j = 0, 1, \cdots, J$. For each node, its cost function can be calculated by

$$H = \sum_{i=1}^{N} P(a_i) \cdot I[P(a_i)] = -\sum_{i=1}^{N} P(a_i) \log_2 P(a_i) \tag{4}$$

$$I[P(a_i)] = -\log_2 P(a_i) \tag{5}$$

where $\{a_i\}$, $1 \le i \le N$,defined to be the histogram for the intensity music and is the number of bins in the histogram.

The entropy discussed here is implemented by using a two stage process. First, a histogram is estimated and thereafter the entropy could be calculated. Six types of entropy were used for evaluation, which are listed in Fig. 6. This collection of entropy value is designed roughly to provide ideas and templates to selecting the "best basis" for decisions.

In the selection method, the entropy value based on the Bottom-up binary tree scheme is used for further comparison. Fig. 7 shows the entropy value at each node of a three-level wavelet packet tree decomposition and the optimal base is indicated. From Fig. 7 we know that the best basis algorithm can be implemented by an optimal base procedure. Therefore, the optimal base procedure is shown as below.

Algorithm for optimal base procedure

---

1. Determine the wavelet decomposition level j($j = 1,2,\ldots$)

2. Compute entropy value of each node by histogram technique in completely decomposition tree.

3. Compare $\mathrm{E}(parent)_j (j = 1,2,\ldots)$ *and* $\mathrm{E}(chold)_{j+1} (j = 1,2,\ldots)$. If $\mathrm{E}(chold)_{j+1} + \mathrm{E}(chold)_{j+2} < \mathrm{E}(parent)_j$, then $\mathrm{E}(chold)$ will be considered as a leaf node of a tree.

4. Repeat the steps 2-3 for each $\mathrm{E}(chold)$ considering it as current node. Otherwise $\mathrm{E}(parent)$ acts as a leaf node of a tree.

Notice that H is determined according to the input music signals. The entropy in music signals will cause low entropy when less information it contains. On the contrary, bigger entropy mean more information. Another interesting observation is that the high entropy is associated with increasing number of wavelet packet decomposition. Specifically, the proposed method using the wavelet packet decomposition performs best at depth 1 with db8 wavelet packets. As a rule of thumb, this paper concludes that the entropy is close to zero can lead to poor performance in music classification.

2.4. **Feature Extraction.** Feature selection is one of the important and frequently used techniques in audio processing for music content analysis. These features should reflect the acoustic characteristics of different kinds of music signals. Among these features, mel-frequency cepstral coefficient (MFCC) and log energy are commonly used for speech recognition, music classification, and other audio/speech related applications [13-14]. The detailed procedure is given in the following.

1) MFCC: Let $s(n), n = 1, .., N$, be a music signal frame that is pre-emphasized to increase the acoustic power at higher frequencies. In order to reduce the effects of spectral leakage and to minimize waveform distortion caused by ringing effect, the proposed method multiplies each frame by a window. The window

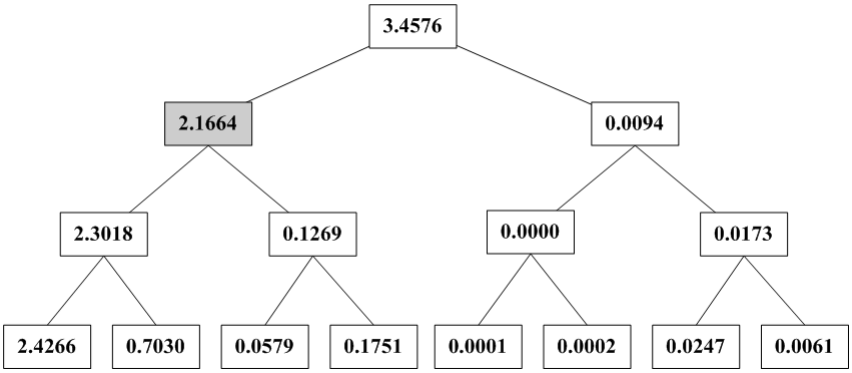


FIGURE 7. The entropy value of the three-level wavelet packet tree decomposition at each node is given and the optimal base is indicated

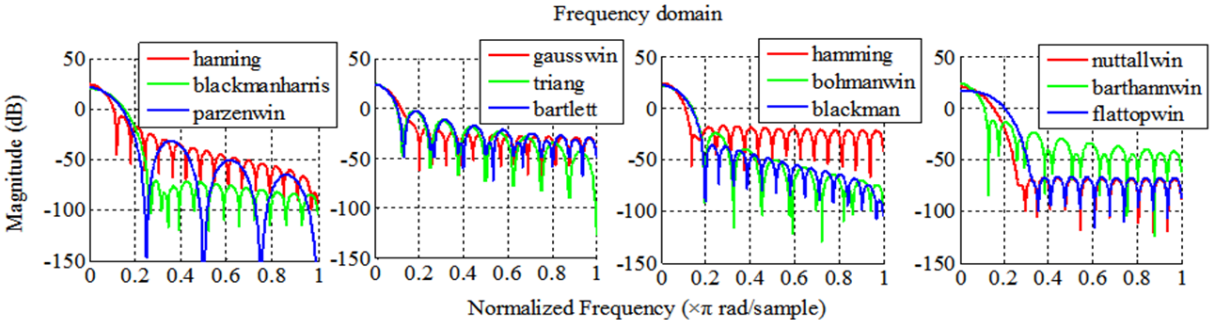


FIGURE 8. Comparison of 12 window functions of Behaviors

discussed here is implemented via a series of cosine function. As shown in Fig. 8, 12 window functions have their own unique characteristic. Each of these characteristics has various amplitudes and shapes [15]. Usually, there are two parameters that could control the trade-off between main-lobe width and side-lobe area. Ideally, a windowing function would produce a narrow main-lobe and low level side-lobes. In other words, as the main-lobe narrows, the frequency resolution increases. Finally, the time domain signal, *s(n)*, is transferred into frequency domain by an M point discrete Fourier transform (DFT). The resulting energy spectrum can be represented as

$$|S(k)|^2 = \left| \sum_{n=1}^{M} s(n) \cdot e^{\left(\frac{-j2\pi nk}{M}\right)} \right|^2 \tag{6}$$

where $1 \leq k \leq M$. Then, according to the previous psychophysical studies, human perception of the frequency content of sounds follows a subjectively defined nonlinear scale called the "mel" scale [16]. It can be defined as,

$$f_{\text{mel}} = 1125 \ln(1 + \frac{f}{700}) \tag{7}$$

where f is the actual frequency in Hz. Next, the triangular filter banks, whose frequency bands are linearly spaced in the mel scale defined in (7), are imposed on the spectrum obtained in (6). The outputs $\left\{e\left(i\right)_{i=1\sim Q}\right\}$ of the mel-scaled band-pass filters can be calculated by a weighted summation between respective filter response $H_i\left(k\right), i = 1\tilde{}Q$, Q is the number of triangular band-pass filters in the bank and the energy spectrum $\left|S\left(k\right)\right|^2$ as

$$e(i) = \sum_{k=1}^{M/2} |S(k)|^2 \cdot H_i(k) \tag{8}$$

where k denotes the coefficient index in the M-point DFT and Hi(k) is defined as

$$H_i(k) = \begin{cases} 0, & \text{for} \quad k < f_{b(i-1)} \\ \frac{(k - f_{b(i-1)})}{(f_{b(i)} - f_{b(i-1)})}, & \text{for} \quad f_{b(i-1)} \le k < f_{b(i)} \\ \frac{(f_{b(i+1)} - k)}{(f_{b(i+1)} - f_{b(i)})}, & \text{for} \quad f_{b(i)} \le k < f_{b(i+1)} \\ 0, & \text{for} \quad k > f_{b(i+1)} \end{cases} \tag{9}$$

In (9), $f_{b(i)}$ are the boundary points of the filters and are depended on the sampling frequency F_s and the number of points M in DFT. That is

$$f_{b(i)} = \left(\frac{F_s}{M}\right) \cdot f_{\text{mel}}^{-1}\left(f_{\text{mel}}\left(f_{\text{low}}\right) + i\frac{f_{\text{mel}}\left(f_{\text{high}}\right) - f_{\text{mel}}\left(f_{\text{low}}\right)}{M+1}\right) \tag{10}$$

Here, f_{low} and f_{high} are respectively the low and high boundary frequencies for the entire filter bank. f_{mel}^{-1} is the inverse to (7) transformation, formulated as

$$f_{\text{mel}}^{-1} = 700\left[e^{\left(\frac{f_{\text{mel}}}{1125}\right)} - 1\right] \tag{11}$$

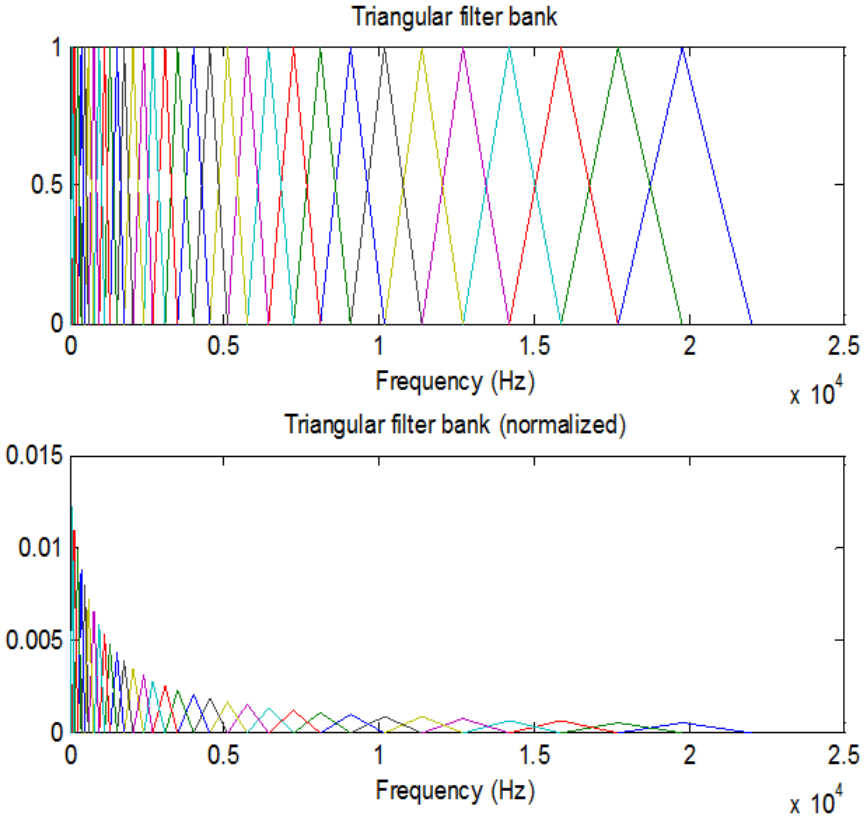


FIGURE 9. Original (upper one) and normalized (lower one) mel-space triangular filter bank (Q=32)

Fig. 9 shows the original as well as normalized mel space triangular filter bank with Q = 32. Finally, discrete cosine transform (DCT) is taken on the log filter bank energies $\{log\left[e\left(i\right)\right]\}_{i=1}^{Q}$ and the MFCC coefficients C_m can be written as,

$$C_m = A\sum_{p=0}^{Q-1} \log[e(p+1)\cdot T_{dct}; for\, m = 0, \cdots, L-1 \tag{12}$$

where L is the desired number of mel-scale cepstral coefficients, A is the scale factor of the discrete cosine transform, and T_{dct} is a trigonometric function (i.e., sin(x) and cos(x)). In Section II-E, this paper describes the four common kernel matrixes for the discrete cosine transform. Here A, and T_{dct} are also given in Section II-E. 2) log energy: The log energy is usually cooperated with MFCC for applications

of speaker recognition and audio segmentations [17]. The definition of log energy used in this paper is defined in (13).

$$E = \log\left(\sum_{n=0}^{N-1} s[n]^2\right) \tag{13}$$

where N is the number of music samples in a frame. Comparing Fig. 10(a)-(e), one can find that the amplitude distribution of different triangular filter banks can be visually differentiated. In this experiment, five set of triangular filter banks (Q) are estimated and then compared with the signal components which are processed by the discrete Fourier transform (DFT). It is clear that the amplitude envelope describes an envelope of the spectrum in the frequency domain. Note that MFCCs were calculated with Q triangular filters (20, 50, 100, etc.). Thus, the performance using triangular filter bank with Q = 300 would outperform their corresponding performance using triangular filter bank with Q = 20 to 200.

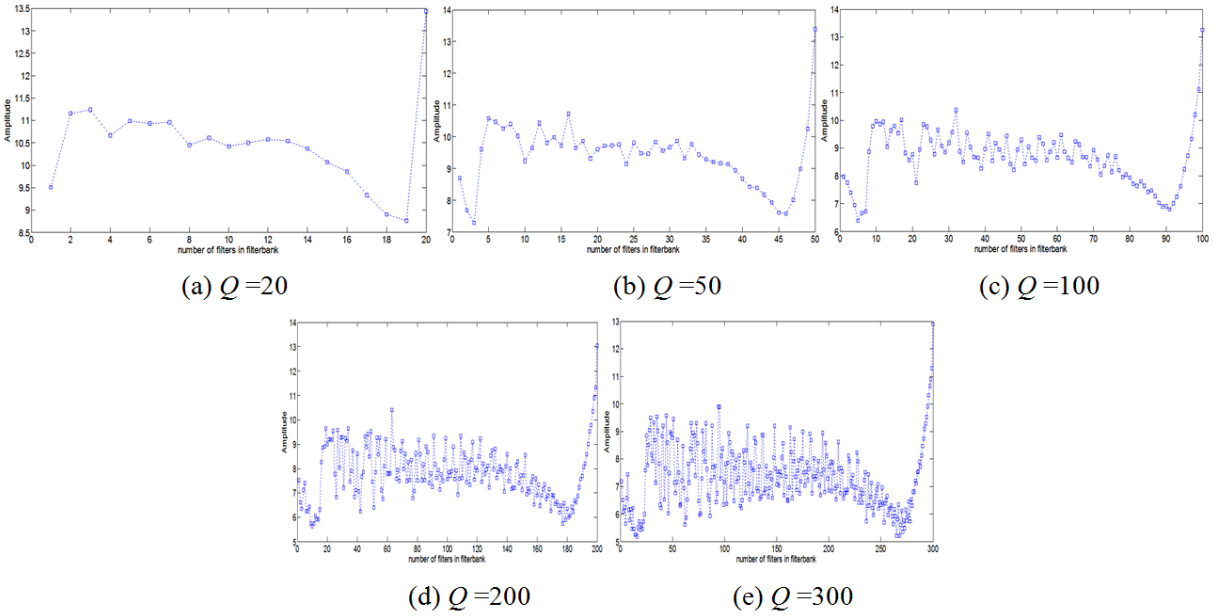


FIGURE 10. Comparison of five set of triangular filter banks (Q) for estimating the signal components processed by the discrete Fourier transform (DFT).

2.5. **Discrete Trigonometric Transform.** The discrete cosine transform (DCT) is a powerful technique which can be used to convert a signal into elementary frequency components. The discrete cosine transform discussed here is implemented via the 4 members of the family of discrete trigonometric transforms (DTTs). Among these members, DCT-II, which were categorized by Wang and are tabulated in [18]-[20], have been played an important role in audio and speech processing. In contrast with conventional methods using discrete Fourier transform (DFT), discrete Hartley transform (DHT) and discrete wavelet transform (DWT) calculated from diagonal matrices have left- and right-multiply the DCT kernel matrix [21], respectively. There are four common kernel matrixes $T_{dct} = \{Xnon\_e,\ Xnon\_o,\ X_e,\ X_o\}$ for the discrete cosine transform, which can be computed as follow:

1) Even extension of discrete cosine transform matrix using non-orthogonal and A=2.

$$X_{non\_e} = A\cos\left[m \cdot \left(\frac{2p+1}{2}\right) \cdot \frac{\pi}{Q}\right] \tag{14}$$

2) Odd extension of discrete cosine transform matrix using non-orthogonal and A=2

$$X_{non\_o} = A\cos\left[m \cdot \left(\frac{2p+1}{2Q-1}\right) \cdot \pi\right] R \tag{15}$$

where R is right diagonal matrices. In order to amplify the signal components, function (16) and (17) can be obtained by modifying scale factor on the DCT kernel matrix as follow:

3) Even extension of discrete cosine transform matrix using orthogonal and $A = \sqrt{\frac{2}{Q}}$.

$$X_e = AL\cos\left[m \cdot \left(\frac{2p+1}{2}\right) \cdot \frac{\pi}{Q}\right] \tag{16}$$

4) Odd extension of discrete cosine transform matrix using orthogonal and $A = 2/\sqrt{2Q-1}$

$$X_o = AL\cos\left[m \cdot \left(\frac{2p+1}{2Q-1}\right) \cdot \pi\right] R \tag{17}$$

Here, L and R are left and right diagonal matrices can be defined as follow:

$$L = \begin{bmatrix} l_0 & 0 & \cdots & 0 \\ 0 & l_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l_{z-1} \end{bmatrix} \; and \quad R = \begin{bmatrix} r_0 & 0 & \cdots & 0 \\ 0 & r_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{z-1} \end{bmatrix} \tag{18}$$

where subscript z is number of filters in filter bank. The only thing to note here is, the scaling factors l_0 and r_{z-1} are equal to $1/\sqrt{2}$

2.6. **Introduction to the Sparse Representation based Classification.** Consider a matrix of training samples, e.g., $A = [A_1, A_2, \cdots, A_N]$ consists of the audio chips from N classes, where $A_i = [a_{i,1}, a_{i,2}, \cdots a_{i,n_i}] \in R^{m \times n_i}$. For a test sample $y \in R^m$, the problem of spares representation is to find a column vector $c_i = [c_{i,1}, c_{i,2}, \cdots, c_{i,n_i}]^T$ such that

$$y = \sum_{j=1}^{n_i} a_{i,j} c_{i,j} = A_i c_i \tag{19}$$

for some scalars $c_{i,j} \in R, j = 1, 2, \cdots n_i$.

Then the linear representation of y can be rewritten in terms of all training samples as

$$y = Ac \tag{20}$$

where $c = [0, \cdots, 0, c_{i,1}, c_{i,2}, \cdots, c_{i,n_i}, 0, \cdots, 0]^T \in R^n$ is a coefficients vector whose elements are zero except those associated with the i-th class.

Due to the system y=Ac is typically underdetermined, therefore its solution is not unique. The following l_0-optimization problem can be resolved by choosing the minimum l_0-norm solution. The problem of sparse representation can be converted into

$$\hat{c}_0 = \arg(\min \|c\|_0) \; subject \; to \; Ac = y \tag{21}$$

where $\|\bullet\|_0$ denotes the l_0-norm of a vector, which counts the number of nonzero entries in a vector. The problem of finding the solution to sparse representation is NP-hard due to its nature of combinational optimization. The above linear programming problem can be solved in [22]. It has been proved that if the solution $\hat{c}_0$ is sparse enough, then the solution of the l_0-minimization problem (21) is equal to the solution to the following -l_1 minimization problem:

$$\hat{c}_1 = \arg(\min \|c\|_1) \; subject \; to \; Ac = y \tag{22}$$

Or alternatively, solve

$$\hat{c}_1 = \arg(\min \|c\|_1) \; subject \; to \; \|Ax - y\|_2 \leq \varepsilon \tag{23}$$

where the error tolerances $\varepsilon > 0$

The l_1-minimization algorithm can be implemented by a primal-dual interior point method called l_1-magic [23]. Therefore, the SRC procedure in [8] is shown as below.

TABLE 1. The ISMIR2004 GENRE database used in the experiments listing classes and number of titles per class

Classes	Number of tracks for training	Number of tracks for testing
Classical	320	320
Electronic	115	114
Jazz/Blues	26	26
Metal/Punk	45	45
Rock/Pop	101	102
World	122	122

Algorithm1. The sparse representation-based Classification (SRC) Algorithm

1. Input: a matrix of training samples $A=[a_1,a_2,\cdots a_n]\in R^{m\times n}$ for n classes, a test sample $y\in R^m$.
2. Normalize the columns of A to have unit ℓ_2-norm.
3. Solve the ℓ_1-norm minimizaton problem:

$$\hat{c}_1 = \arg\min\|c\|_1 \quad \text{subject to} \quad Ac = y$$

4. Compute the residuals by:

$$r_i(y)=\|y-A\delta_i(\hat{c}_1)\|_2$$

for $i=1,\ldots,n$, where δ_i is the characteristic function which selects the coefficients associated with the i-th class.

5.Output the identity by:

identity(y) = $\arg\min_i r_i(y)$, where identity function stands for finding the class label of y.

3. Experimental results.

3.1. **Datasets.** In the following experiments, a public music database named ISMIR2004 GENRE [24] is utilized to evaluate classification performances. The ISMIR2004 GENRE database consists of 1458 music tracks in which 729 music tracks are used for training and the other 729 tracks are applied to testing, the pieces being unequally distributed over 6 genres, as shown in Table I. The sampling rate of the audio file is 44.1 kHz with16-bit resolution.

3.2. **Classification Results.** Fig. 11 shows the average classification accuracy implemented using 12 types of window functions. Based on the above results, the proposed method chooses the function among the 12 windows to determine the main-lobe width and side-lobe area with empirical analysis. Note that the triangular (Bartlett) window is applied to minimize the signal discontinuities at the borders of each frame in this paper. In addition, to investigate the importance of various discrete cosine transform types, four types of DCT is used for music genre classification and accuracy comparison.

As shown in Fig. 12, the accuracy of Type II orthogonal DCT [19] is clearly better than that of Type II non orthogonal DCT [21]. Specifically, the best classification result with the Type II odd orthogonal DTC is 89.7% , which is significantly better than the 86.69% accuracy rate of the Type II even orthogonal DTC. Note that [19] and [21] apply the same feature set to achieve music genre classification, but method

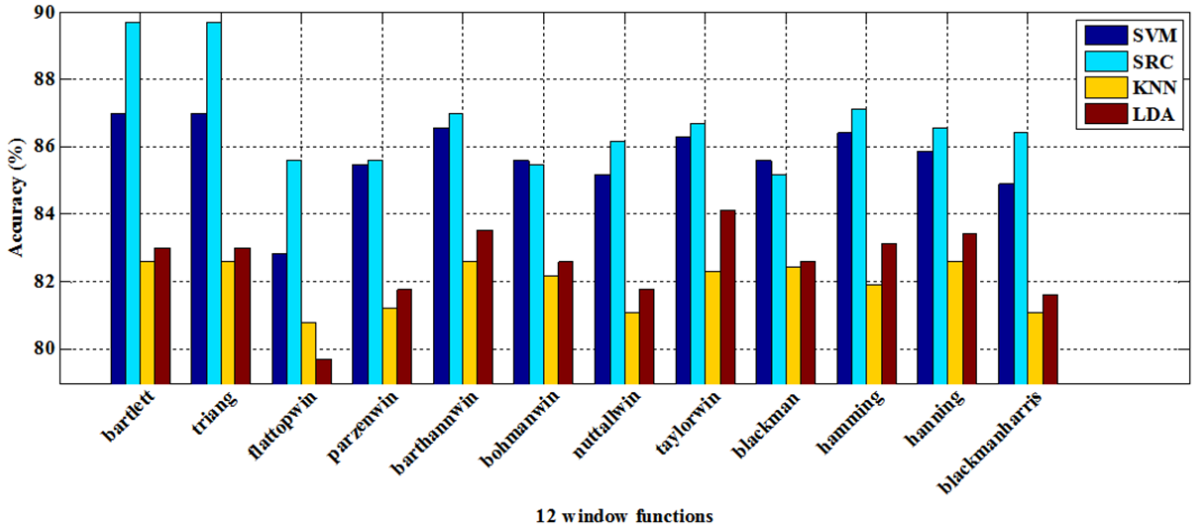


FIGURE 11. Comparison of 12 window functions for average classification accuracy

[19] outperforms method [21] by using orthogonal instead of non-orthogonal DTC. However, Type II even orthogonal DTC is the most commonly used one.

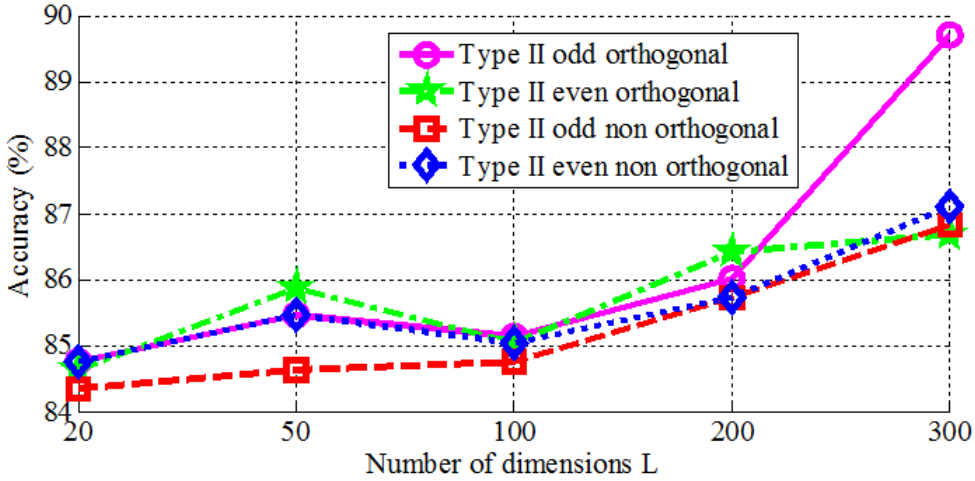


FIGURE 12. Comparison of the importance of various discrete cosine transform for music genre classification

Furthermore, several statistical techniques are tested in order to demonstrate the superiority of the proposed approach. Four statistical techniques (SVM, SRC, KNN and LDA) are compared over a range of pre-selected values of the cutoff frequency C and the order of filter F, whereas number of dimensions L in MFCC varies from 20 to 300. The comparisons of classification accuracy rate on the ISMIR 2004 Genre dataset for various features dimension (20, 50, 100, 200 and 300) as well as different classifiers (SVM, SRC, KNN and LDA) are shown in Fig. 13 (a)-(d). It is obvious that the proposed method using SRC classifier is significantly better than the other three methods.

These four figures correspond to the use of different frame sizes FS. (a) FS= 5944.3 ms (262144 samples) and overlap= 185.8 ms (8192 samples). (b) FS= 2972.2 ms (131072samples) and overlap= 2972.2 ms (131072samples). (c) FS= 2972.2 ms (131072samples) and overlap= 1486.1 ms (65536 samples). (d) FS= 1486.1 ms (65536 samples) and overlap= 1486.1 ms (65536 samples).

It reconfirms the common belief that given the same feature set, the choice of the classifier and frame size is important. Another interesting observation is that the classification accuracy rate could be improved when one adopt long-term analysis for audio signals, as shown in Fig. 13 (a).

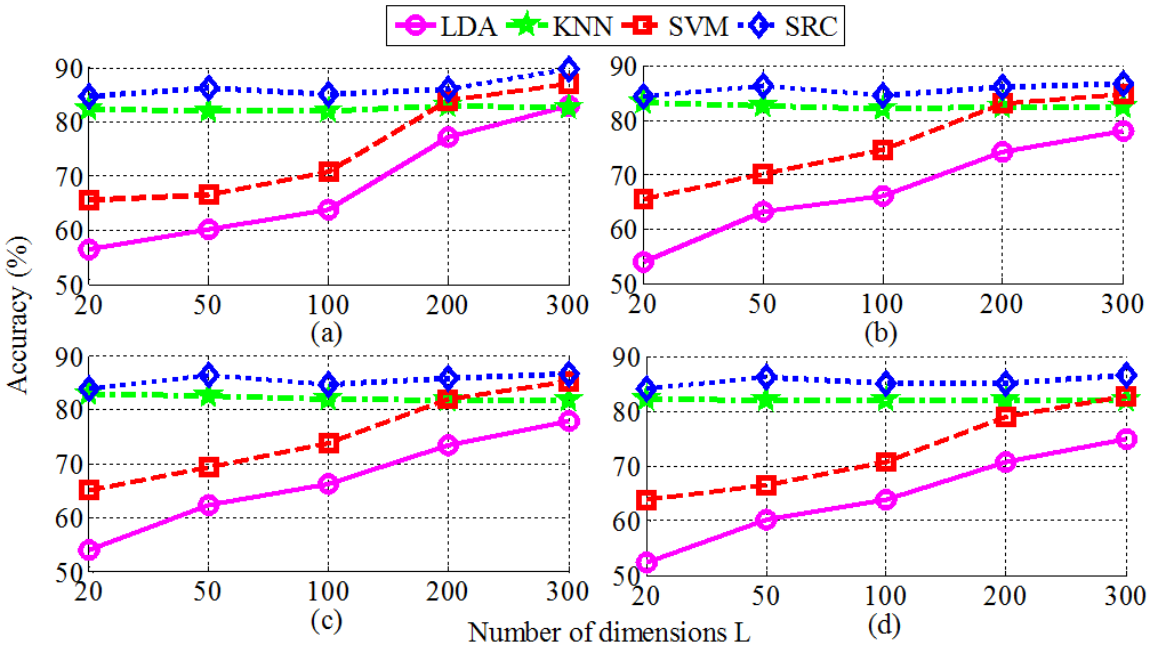


FIGURE 13. Classification accuracy of different features dimension as well as different classifiers on the ISMIR2004 Genre dataset

TABLE 2. Average classification accuracy using a MF-point moving average filter

F \ C	1	3	5	10
8000	87.38	87.52	87.52	87.24
12000	87.52	87.79	87.38	87.24
16000	87.38	87.52	87.52	88.34
20000	87.93	88.07	87.38	87.24

(a) *MF=5.*

F \ C	1	3	5	10
8000	86.56	87.65	87.38	87.11
12000	87.24	87.52	88.20	87.52
16000	87.93	87.11	87.52	87.24
20000	87.11	88.20	87.65	87.24

(b) *MF=10*

F \ C	1	3	5	10
8000	88.07	87.52	87.52	88.20
12000	88.20	87.52	87.38	87.52
16000	88.89	88.34	88.75	88.75
20000	89.16	89.71	89.03	88.07

(c) *MF=20*

F \ C	1	3	5	10
8000	86.56	86.69	87.24	87.24
12000	86.15	87.11	86.42	86.56
16000	86.28	87.24	86.83	86.15
20000	86.69	86.69	86.69	87.11

(d) *MF=40*

Tableshows the moving average filter MF of the proposed methods could perform the best result with MF=20 for the pre-selected values C = 20000 and F = 3. Thus, this paper experiments on the ISMIR 2004 Genre dataset by using 50:50 training and test set split techniques to evaluate various genre classification systems. The best classification accuracy rate of 89.7% was obtained under the condition that feature extraction by wavelet package transform and classification by SRC.

Tablecompares our proposed approach with other approaches [2], [10], [25], [26], [27], [28], [29], [30], [31], [32], [33] on the ISMIR 2004 Genre dataset with the same experimental setup. It is clear that the achieved classification accuracy rate of 89.7% outperforms all previously reported rates as shown in Table .

Finally, Table shows detailed SRC performance in musical genre classification in form of confusion matrix. The row indexes of the confusion matrix correspond to predicted genre and the column indexes correspond to the actual genre. One could observe that the diagonal elements present the correctly classified observations for each case, and the off-diagonal elements show the number of misclassifications. Note that a perfect matrix only contains numbers in the diagonal.

TABLE 3. Best results obtained on the ISMIR 2004 Genre classification contest (50:50 training and test set split)

Authors	CA
D. Ellis & B. Whitman	64.00%
T. Lidy & A. Rauber	70.37%
G. Tzanetakis	71.33%
K. West	78.33%
T. Lidy & A. Rauber [26]	79.70%
I. Panagakis et al. [34]	80.95%
Bergstra et al. [35]	82.30%
Pampalk et al. [36]	82.30%
Holzapfel et al. [37]	83.50%
E. Pampalk	84.07%
Y. Song et al. [29]	84.77%
C.Rusu [38]	85.59%
Chang-Hsing et al. [32]	86.83%
Our approac	89.71%
Y. Panagakis et al. [10]	93.56%
Y. Panagakis et al. [33]	94.93%

TABLE 4. Genre confusion matrix on the ISMIR 2004 Genre classification (50:50 training and test set split)

	Classical	Jazz/Blue	Electronic	Rock/Pop	Metal/Punk	World
Classical	320	0	0	0	0	0
Jazz/Blue	0	26	0	0	0	0
Electronic	14	1	70	14	3	12
Rock/Pop	5	1	13	71	3	9
Metal/Punk	0	0	0	0	45	0
World	0	0	0	0	0	122

4. **Conclusions.** In this paper, sparse representation based classification (SRC) and wavelet packet transform (WPT) with discrete trigonometric transforms (DTTs) are applied to the task of music genre classification. The music genre features used in the proposed method includes MFCC and log energy, which can represent the time-varying behavior of music. To investigate its performance, the proposed method is validated by comparison with various discrete cosine transform types and classification methods. The average music genre classification accuracy rate of the proposed method is 89.7% on the ISMIR2004 Genre dataset. Numerical experiments show that sparse representation approach can match the best performance achieved by moving average filter, Butterworth low-pass filter, and wavelet packet transform (WPT) with discrete trigonometric transforms (DTTs). There are two directions that need to be explored in the future. The first direction is to investigate how to improve the computational efficiency for sparse representation approach. The second direction of our future work is to investigate how to improve the accuracy of the average music genre classification.

REFERENCES

[1] T. Li, and G. Tzanetakis, Factors in automatic musical genre classification of audio signals, *Applications of Signal Processing to Audio and Acoustics*, pp. 143-146, 2003.

[2] N. Scaringella, G. Zoia, and D. Mlynek, Automatic genre classification of music content: A survey, *IEEE Signal Proc. Magazine*, vol. 23, no.2, pp. 133141, 2006.

[3] G. Tzanetakis and P. Cook, Musical genre classification of audio signals, *IEEE Trans. Speech Audio Proc.*., Vol. 10, no. 5, pp. 293302, Jul. 2002.

[4] Nicolas Scaringella, Giorgio Zoia, and Daniel Mlynek, Automatic genre classification of music content: A survey, *IEEE Signal Proc. Magazine*, vol. 23, no. 2, pp. 133-141, 2006.

[5] A. Flexer, A closer look on artist filters for musical genre classification, *Proc. International Conference on Music Information Retrieval*, 2007.

[6] A. Meng, P. Ahrendt, J. Larsen, L. K. Hansen, Temporal Feature Integration for Music Genre Classification, *IEEE Trans. on Speech and Audio Processing*, vol. 15, no. 5, pp. 1654 - 1664, 2007.

[7] D. W. Jang, M. H. Jin, C. D. Yoo, Music genre classification using novel features and a weighted voting method, *Proc. IEEE International Conference on Multimedia and Expo*, pp. 1377-1380, 2008.

[8] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Yi Ma, Robust Face recognition via sparse representation, *Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210227, 2009.

[9] Y. Panagakis and C. Kotropoulos, Music genre classification via topology preserving non-negative tensor factorization and sparse representations, *IEEE Trans. Audio, Speech, and Language Processing*, pp. 249 - 252, 2010.

[10] Y. Panagakis, C. Kotropoulos, and G. R. Arce, Music genre classification via sparse representations of auditory temporal modulations, *Proc. the 17th European Signal Processing Conference*, pp. 1-5, 2009.

[11] R. R. Coifman and M. V. Wickerhauser, Entropy-based algorithms for best basic selection, *IEEE Trans. on Information Theory*, vol. 38, no.2, pp.713-718, 1992.

[12] Ashraf A. Kassim, Niu Yan and Dornoosh Zonoobi. 2008. Wavelet packet transform basis selection method for set partitioning in hierarchical trees, Journal of Electronic Imaging, Vol 17, no.3, pp.033007- 1-033007-9, Jul-Sep 2008.
A. A. Kassim, N. Yan, and D. Zonoobi, Wavelet packet transform basis selection method for set partitioning in hierarchical trees, *Journal of Electronic Imaging*, vol.17, no.3, pp. 033007-033007, 2008.

[13] C.C. Lin, S.H. Chen, T. K. Truong, and Y. Chang, Audio Classification and Categorization Based on Wavelets and Support Vector Machine, *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 644-651, 2005.

[14] L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, *Prentice Hall*, 1993.

[15] A. H. Nuttall, Some Windows with Very Good Sidelobe Behavior, *IEEE Trans. on Acoustics Speech and Signal Processing*, vol. ASSP-29, no. 1, pp. 84-91, 1981.

[16] S. B. Davis and P. Mermelstein, Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences, *IEEE Trans. on Acoust. Speech Signal Processing*, vol. ASSP-28, no. 4, pp. 357-365, 1980.

[17] C. P. Chen and J. A. Bilmes, MVA Processing of Speech Features, *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 1, pp.257-270, 2007.

[18] Z. Wang, Fast algorithms for the discrete W transform and the discrete Fourier transform, *IEEE Trans. on Acoust. Speech Signal Processing*, vol. ASSP-32, no.4, pp. 803-816, 1984.

[19] Z. Wang and B. R. Hunt, The discrete W transform, *Applied Mathematics and Computation*, vol. 16, pp. 19-48, 1985.

[20] K. R. Rao and P. Yip, Discrete Cosine Transform: Algorithms, Advantages, Applications, *Academic Press*, 1990.

[21] S. A. Martucci, Symmetric convolution and the discrete sine and cosine transforms, *IEEE Trans. on Signal Processing*, vol. 42, pp. 1038-1051, 1994.

[22] D. L. Donoho, and X. Huo, Uncertainty principles and ideal atomic decomposition, *IEEE Trans. on Information Theory*, vol. 47, no. 7, pp. 28452862, 2001.

[23] E. Cand‘es and J. Romberg, l1-magic: a collection of MATLAB routines for solving the convex optimization programs central to compressive sampling, http://users.ece.gatech.edu/ justin/l1magic/

[24] The International Society for Music Information Retrieval , 2004. ISMIR2004 Audio Description Contest - Genre/ Artist ID Classification and Artist Similarity , http://ismir2004.ismir.net/genre_contest/

[25] A. Holzapfel and Y. Stylianou, Musical genre classification using nonnegative matrix factorization-based features, *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 424434, 2008.

[26] T. Lidy and A. Rauber, Evaluation of feature extractors and psycho-acoustic transformations for music genre classification, *Proc. 6th Int. Symp. Music Information Retrieval*, London, pp. 3441, 2005.

[27] E. Benetos and C. Kotropoulos, A tensor-based approach for automatic music genre classification, *Proc. 16th European Signal Processing Conf.*, Switzerland, 2008.

[28] T. Lidy, A. Rauber, A. Pertusa, and J.M. Inesta, Combining audio and symbolic descriptors for music classification from audio, Music Information Retrieval Evaluation Exchange, 2007.
[29] Y. Song and C. Zhang, Content-based information fusion for semi-supervised music genre classification, *IEEE Trans. on Multimedia*, vol. 10, no. 1, pp. 145152, 2008.
[30] T. Li and M. Ogihara, Toward intelligent music information retrieval, *IEEE Trans. on Multimedia*, vol. 8, no. 3, pp. 564573, 2006.
[31] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kegl, Aggregate features and Adaboost for music classification, *Machine Learning*, vol. 65, no. 23, pp. 473484, 2006.
[32] C. H. Lee, J. L. Shih, K. M. Yu, and H. S. Lin, Automatic Music Genre Classification Based Modulation Spectral Analysis of Spectral and Cepstral Features, *IEEE Trans. on Multimedia*, vol. 11, no. 4, 2009.
[33] Y. Panagakis and C. Kotropoulos, Music genre classification via topology preserving non-negative tensor factorization and sparse representations, *IEEE Trans. on Audio, Speech, and Language Processing*, pp. 249 - 252, 2010.
[34] I. Panagakis, E. Benetos and C. Kotropoulos, Music genre classication: A multilinear approach, *Proc.9th Int. Symp. Music Information Retrieval*, Philadelphia-USA, pp. 583-588, 2008.
[35] J. Bergstra, N. Casagrande, D. Erhan, D. Eck and B. Kegl, Aggregate features and AdaBoost for music classification, *Machine Learning*, vol. 65, no. 2-3, pp. 473-484, 2006.
[36] E. Pampalk, A. Flexer and G. Widmer, Improvements of audio-based music similarity and genre classification, *Proc. 6th Int. Symp. Music Information Retrieval*, pp. 628-633, London-UK, 2005.
[37] A. Holzapfel and Y. Stylianou, Musical genre classification using nonnegative matrix factorization-based features, *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 424-434, 2008.
[38] C. Rusu, Classification of music genres using sparse representation in overcomplete dictionaries, *CEAI*, vol.13, no. 1, pp. 35-42, 2011.