Multimedia Systems (2022) 28:779–791
https://doi.org/10.1007/s00530-021-00886-3

REGULAR PAPER

Music genre classification based on auditory image, spectral and acoustic features

Xin Cai[1] · Hongjuan Zhang[1]

Received: 23 June 2021 / Accepted: 30 December 2021 / Published online: 10 January 2022

Abstract

Music genre is one of the conventional ways to describe music content, and also is one of the important labels of music information retrieval. Therefore, the effective and precise music genre classification method becomes an urgent need for realizing automatic organization of large music archives. Inspired by the fact that humans have a better automatic recognizing music genre ability, which may attribute to our auditory system, even for the participants with little musical literacy. In this paper, a novel classification framework incorporating the auditory image feature with traditional acoustic features and spectral feature is proposed to improve the classification accuracy. In detail, auditory image feature is extracted based on the auditory image model which simulates the auditory system of the human ear and has also been successfully used in other fields apart from music genre classification to our best knowledge. Moreover, the logarithmic frequency spectrogram rather than linear is adopted to extract the spectral feature to capture the information about the low-frequency part adequately. These above two features and the traditional acoustic feature are evaluated, compared, respectively, and fused finally based on the GTZAN, GTZAN-NEW, ISMIR2004 and Homburg datasets. Experimental results show that the proposed method owns the higher classification accuracy and the better stability than many state-of-the-art classification methods.

Keywords Music genre classification · Auditory image feature · Spectral feature · Acoustic feature · Feature fusion

1 Introduction

With the development of the information age and multimedia technologies, online music databases have grown significantly and digital music can be widely available in various ways. Thus, it is increasingly difficult for users to query and retrieval in the massive music database, this trend has put forward higher requirements for retrieval of music information. Since the genre is one of the important musical labels [18], the classification of music genre is also crucial for the retrieval of music information. Moreover, due to the vast amount of existing collections, it is very necessary to develop the effective automatic Music Genre Classification (MGC) system [6].

A representative MGC system commonly consists of two main parts: (1) feature extraction and processing part; (2) classification part. Among them, the feature extraction and processing part includes feature extraction from different domains such as time signal, spectrum, and feature processing by aggregating features from short time frame to long time window, outlier processing, normalization, etc. The classification part, on the other hand, includes the training of the classification model as well as its validation. It is noted that the first part is the most important factor that decisively affects the performance of MGC method. Therefore, there are many works have focused on how to extract the suitable feature for the classification problem of music genre. For example, Tzanetakis and Cook [54] presented three sets of features that representing timbral texture, rhythmic content and pitch content and created the benchmark dataset, i.e. the GTZAN dataset, which is widely used in the field of MGC. After this, another classification method has been proposed by Li et al. [30] on the basis of the Daubechies Wavelet Coefficient Histograms (DWCH) which improves the performance when combined with timbre features. Moreover, the psycho-acoustic-based feature set reflecting the sensation of

Communicated by P. Pala.

✉ Hongjuan Zhang
zhanghongjuan@shu.edu.cn

[1] Department of Mathematics, Shanghai University, Shanghai 200444, People's Republic of China

human loudness was presented in [31] and evaluated in [51]. In addition, Lee et al. [28] and Lim et al. [32] adopted modulation spectral to capture the spectra evolution and time-varying information about music signal, and has acquired good classification results using these acoustic features.

Over the last decade, another feature, i.e., the spectrogram of audio signal, has also been proved its validity on MGC tasks. Moreover, due to the similarities between the image and spectrogram, the related algorithm for image processing can be similarly utilized on the MGC problem. For example, Costa et al. [11] originally explored Gray Level Co-occurrence Matrix (GLCM) from the spectrogram image for classifying music genres. On this basis, some typical texture descriptors in image processing, such as Local Binary Pattern (LBP) [13], Gabor filters [14], Local Phase Quantization (LPQ) [14], were gradually presented for MGC. Furthermore, Nanni et al. [38, 39] considered the complementary related to spectral features and traditional musical acoustic features, and performed well in classification. Motivated by the excellent performance in image recognition and Natural Language Processing (NLP) of deep learning, [57, 59] apply Convolutional Neural Network (CNN) and deep attention, respectively, to capture information of spectrogram on MGC tasks.

It is worth noting that majority of existing methods in the field of MGC, such as audio content-based [18], symbolic content-based [10], lyric-based [8, 53], meta-data-based [34], and hybrid approaches [9]. To the best of our knowledge, most methods on MGC tasks systematically consider music features based on music content and obtain better classification results than other methods [1, 5, 22, 23, 49, 50]. Audio content-based methods aim to explore the music features of the audio signal, including acoustic features, spectral features, and their combinatorial features. Although they have achieved good performance, there is still a lot of room for improvement. In fact, a test about the human's capability for the music genre recognition was conducted, which showed that college students with little to moderate musical training could distinguish 70% among 10 genres only based on 300 ms music clips [48]. And in another similar test, human listeners could only achieve 76% accurate in classifying 6 music genres after listening to 30 s music clips [33]. Both studies indicate that humans have a better classification ability, which may attribute to the cochlea's physiological structure, even for the participants with little musical literacy, or even for the classification of many shorter music clips. Therefore, a kind of promising feature, which is inspired by the human auditory system, has been proposed recently. Moreover, some related computational models have been built for mimicking the audio processing mechanism of the cochlea in the following works. Specifically, Panagakia and Kotropoulos [45] originally proposed a classification framework that considered the auditory cortex model using cortical representation (CR) features [37] for MGC problem. And Nonaka et al. [40] and Hyder et al. [25] utilized another auditory image model (AIM) [2] to the snoring classification (SC) and Acoustic Scene Classification (ASC) respectively.

Motivated by these above-mentioned works, here we will focus this kind of promising feature and further explore its combination with the acoustic and spectral features to extract suitable features of the music signal. It is noted that this idea considers both to mimic the audio processing of the cochlea and to utilize the musical knowledge owned by humans, which will be a more comprehensive perspective than many existing approaches. Figure 1 shows the proposed classification system, which consists of three parts: feature extraction and processing, classification and fusion. In the feature extraction and processing stage, we will extract three different feature sets including the auditory image feature set based on the AIM, the spectral feature set from the logarithmic spectrogram, and the acoustic feature set based on timbre feature and psycho-acoustic model. Then, the support vector machine (SVM) classifier with radial basis function (RBF) kernel is adopted in the classification stage. And in the last stage, the final decision is obtained by fusing the classification score according to sum-rule.

In summary, the main contributions of the proposed system are the following: (1) Propose a novel auditory-inspired feature set based on the auditory image processed by AIM. (2) Employ the logarithmic frequency spectrogram to extract

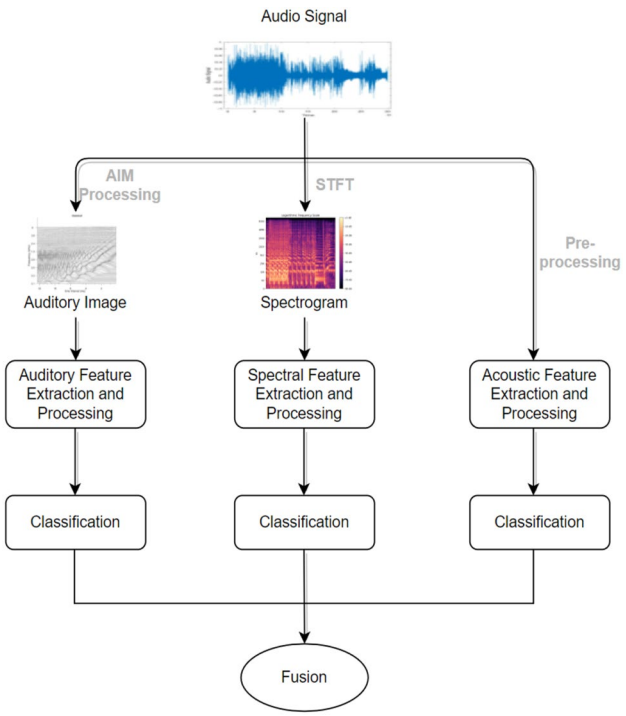


Fig. 1 Flowchart of the proposed music genre classification system

spectral features rather than linear. (3) Investigate and compare the effectiveness of auditory image features, spectral features and acoustic features and their combination.

The remainder of this paper is organized as follows. Section 2 describes the auditory image model and lists the texture descriptors we have adopted. Section 3 presents the spectral features which are extracted from the spectrogram image. The acoustic features are detailed in Sect. 4. Datasets and experimental results are presented in Sect. 5. Finally, we offer our conclusion in Sect. 6.

2 Proposed approach: auditory image features

In this section, we will focus on auditory image feature set extraction based on the auditory image model inspired by the auditory system [2], which mimics the initial perception of sounds heard by the human and is very effective for sounds recognition.

2.1 AIM process

The AIM is a time-domain functional model that simulates human auditory processing [47]. Its processing stage is divided into five principal modules: Pre-Cochlear Processing (PCP), Basilar Membrane Motion (BMM), Neural Activity Pattern (NAP), The Identification of The Strobe Points (STROBE) and Stabilized Auditory Image (SAI) [2]. An illustration of the AIM processing stage and the corresponding physiological structure is shown in Fig. 2.

In the following, we will introduce its five modules, respectively. During the PCP stage, the input audio signal is filtered by the bandpass filter, which simulates the effect of sound from outer ear up to the oval windows of the cochlea [19–21]. In the BMM stage, the one-dimensional audio signal is converted into multi-channel waveform signals on different frequency bands according to the selection process of cells at different positions on the cochlear basilar membrane for different frequencies. And the dynamic compressive gammachirp filter-bank (dcgc) with 50 filter channels is used to simulate the amplitude and time delay of the audio signal on the basilar membrane at different positions of the human ear [26]. In the next stage, the half-wave rectification and low-pass filtering (hl) are performed in the NAP module that mimics the inner hair cells in the cochlea, and convert the physical information of the BMM module into the auditory nerve information of the cochlea. The STROBE stage aims to find the strobe points to construct the auditory image without distorting the temporal fine-structure information [46]. Followed by the SAI stage, the output of the NAP is converted into an auditory image after the strobe points have been determined at the STROBE stage. At this point, the time dimension of the NAP is converted into the time-interval dimension of the SAI by strobed temporal integration, which is based on the sound perception principle of the human ear. Remarkably, the temporal integration mechanism of AIM generates a stabilized auditory image, which is characterized as a two-dimensional sound signal depiction that the vertical axis is frequency, while the horizontal axis represents the time interval relative to the strobe-times [36].

In this study, we will propose a novel auditory feature based on the AIM. Since AIM is a sophisticated computational model, in which an overlapping 35 ms SAI frame is generated every 5 ms during the AIM processing, we want to use middle 10 s segment of each audio signal to extract auditory feature to reduce the computational burden. Specifically, the 10 s audio clip is converted into an auditory image via the AIM processing firstly. Secondly, the value of each filter channel is regarded as 'pixel', and then the auditory image of SAI frame is converted into 2D grayscale image. Finally, texture descriptions are extracted from grayscale image as auditory image features, hereinafter referred to as AIM features. And in this work, the AIM processing is implemented in MATLAB with AIM2006 (modules: gm2002, dcgc, hl, sf2003, ti2003) [40].

2.2 Texture descriptions

The texture feature describes the recurring local patterns and arrangement rules in the image, which helps to distinguish different images [56]. Therefore, we use the following texture descriptions in this work.

– Local Binary Pattern (LBP): The LBP operator is an efficient texture descriptor with rotation invariance and gray

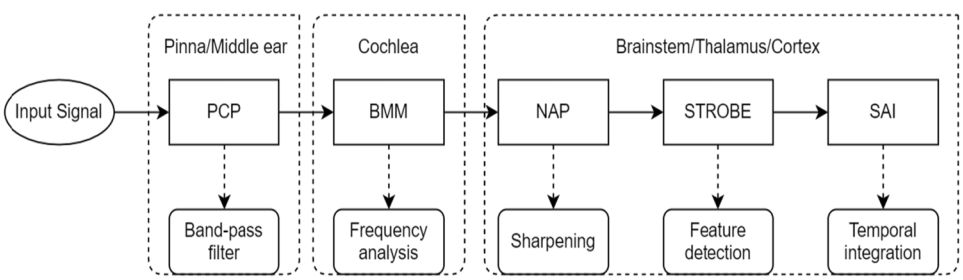


Fig. 2 Flowchart of signal processing with AIM

invariance, describes a local region as a binary pattern. The original LBP operator was later extended to uniform patterns [42]. In this work, the multi-scale uniform local binary pattern with different radii and sample point aims to achieve multi-resolution analysis. The final description is obtained from the concatenation of LBP with $(R = 1, P = 8)$ and $(R = 2, P = 16)$.

- LBP Histogram Fourier (LBPHF): The LBPHF [60] features are computed by the discrete Fourier transform of the feature vector of the uniform LBP histograms. The multi-scale LBPHF descriptor is extracted from a series of LBPHF with $(R = 1, P = 8)$ and $(R = 2, P = 16)$.
- Rotation Invariant Co-occurrence Among Adjacent LBPs (RICLBP): The LBP features were extended to co-occurrence among adjacent LBPs (CoALBP) by introducing the concept of co-occurrence among LBPs to enhance the descriptive ability of LBP. RICLBP [41] is obtained rotation invariance by introducing rotation equivalence class to CoALBPs, so that RICLBP has not only rotation invariance but also a high descriptive ability. Concatenation of RICLBP with $(R = 1, P = 8)$,$(R = 2, P = 8)$,$(R = 4, P = 8)$ as RICLBP feature.
- Dense Sampling Based LBP (DSLBP): The DSLBP [58] description introduces a simple form of up-sampling to calculate more neighboring pixels, so that owning more stable LBP codes and carrying out richer information. The final description is obtained from the concatenation of DSLBP with $(R = 1, P = 8)$ and $(R = 2, P = 16)$.
- Local Phase Quantization (LPQ): LPQ [43] is a blur invariance texture analysis method. The LPQ operator quantifies the phase of the local image windows using short-time Fourier transform. The multi-scale LPQ with radius 3 and 5.

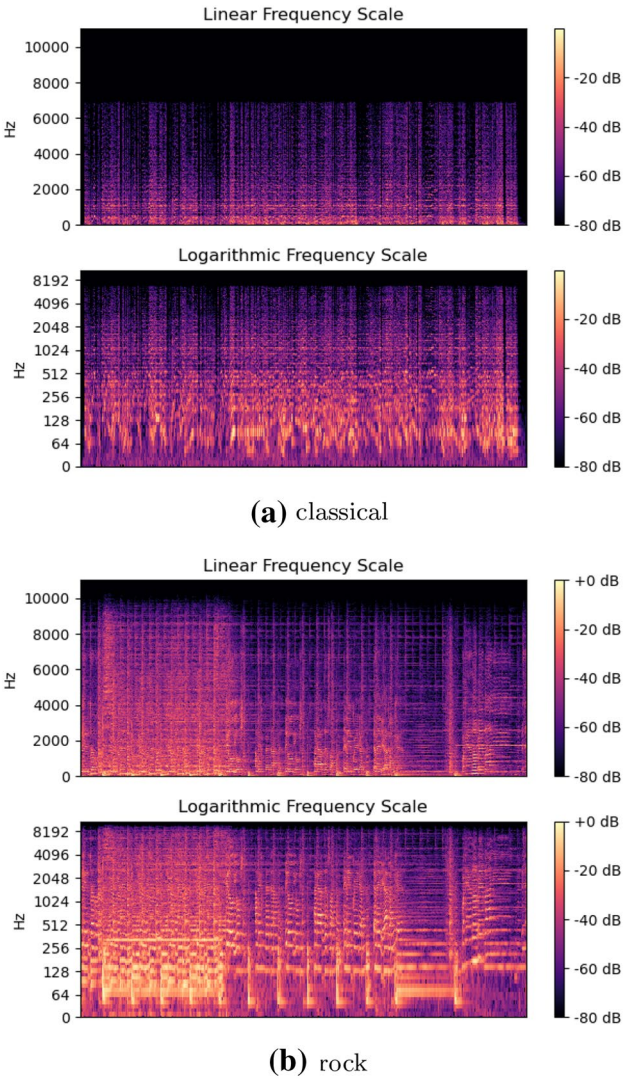


Fig. 3 Spectrogram on different frequency scale of different music genre

3 Proposed approach: spectral features

The spectral feature set is described in this section, which is extracted from the spectrogram on a logarithmic frequency scale for MGC.

In the following, we will first give the reason why we use a logarithmic frequency spectrogram instead of the linear.

Figure 3 illustrates an example of spectrograms on different frequency scales taken from different genres of music pieces, with linear at the top and logarithmic at the bottom. Figure 3a shows spectrograms of classical music piece in the ISMIR2004 dataset, while Fig. 3b are spectrograms of rock music genre in the GTZAN dataset. In the spectrogram image, the vertical axis shows the frequency (from 0 to 10 kHz) and the horizontal axis shows the time of the audio clip. In addition, the intensity of each point in the image represents the decibel of the signal and can also be thought of as the logarithmic scale of the amplitude.

In this example, we can observe that most of the information is concentrated at the bottom of the spectrogram under a linear frequency scale, but in the logarithmic frequency spectrogram the low frequency part is pulled up a lot. This means that the logarithmic spectrogram has ability to capture low frequency information than the linear. Moreover, it is the fact that the ear's perception of pitch is usually proportional to the logarithm of frequency rather than frequency itself. Therefore, it is more practical and meaningful to express the frequency scale of the spectrogram by logarithmic frequency scale.

In this work, we first adopt a signal segmentation strategy [12] to divide the original audio signal into three 10-s sub-signals, which are taken in the beginning, middle and

end of each original audio, respectively. Then the segment signal is converted into a spectrogram image according to a logarithmic frequency scale, and this spectrogram image is regarded as a texture image for the texture descriptor extraction. Finally, the spectral feature is performed by calculating the following texture descriptions: LBP, LBPHF, RICLBP, DSLBP and LPQ. It is noted that, although both spectral and AIM features are ultimately calculated by texture descriptors, their differences are that spectral feature extraction is performed on the spectrogram image, whereas AIM feature extraction is made in the auditory image calculated by the AIM.

4 Proposed approach: acoustic features

In this section, we will focus on acoustic feature set consisting of Mel-Frequency Cepstral Coefficients (MFCC) and psycho-acoustic music descriptors [31].

MFCC is inarguably the single most important feature type that was widely used for both speech recognition and music classification. It can model the subjective frequency content of audio signals, and produce quite good classification result compared to many other methods that use more sophisticated collections of features [29, 35]. MFCC is extracted based on the mel-scale band-pass filters. Let $S(b), 0 \leq b \leq B$, denote the sum of spectra in bth mel-scale band, where B is the total number of the bands. A kth MFCC is computed by applying the Discrete Fourier Transform (DCT) on the logarithm of $S(b)$ as follows:

$$\mathrm{MFCC}(k) = \sum_{b=0}^{B-1} \log S(b) \cos\left(k \frac{\pi}{B}(b + 0.5)\right). \quad (1)$$

In audio analysis, the first 13 coefficients of the MFCC extracted from each audio frame are usually taken, and we also take the first 13 coefficients in this work. Moreover, final MFCC feature is calculated by averaging the mean, variance, max, min of each of the entire frames of the audio signal, due to the frame-level timbre feature it belongs. And psycho-acoustic music descriptor describes the rhythmic structure on the different frequency bank and reflects the human sensation of loudness. The following psycho-acoustic descriptors [31] are employed in this study.

- Statistical Spectrum Descriptor (SSD): Statistical measures describe the audio content in terms of energy, change in beat and rhythm over the specific critical frequency band. The rhythmic content of a piece of audio is described by calculating the following statistical moments on each of the 24 critical bands: mean, median, variance, skewness, kurtosis, min-value and max-value.
- Rhythm Histogram (RH): The magnitudes of each modulated frequency bin for the 24 critical bands are summed up to form a 'rhythm energy' histogram for each modulated frequency. The histogram contains 60 bins, reflecting the modulation frequency between 0 and 10 Hz. For a given piece of audio, the feature set of the rhythm histogram is the median of the histogram of each 6 s segment processed.
- Modulation Frequency Variance Description (MVD): The MVD descriptor measures the variations of a specific modulation frequency over the critical frequency bands. The MVD vector is obtained by calculating the statistical moment of each modulation frequency over the 24 critical frequency bands: mean, median, variance, skewness, kurtosis, min-value and max-value. The MVD descriptor of a piece of audio is calculated based on the mean of the MVDs taken from every 6 s segments.
- Temporal Statistical Spectrum Description (TSSD): This statistical spectrum descriptor is extracted by calculating the statistical moments of audio clips in different time positions within a piece of audio. This descriptor incorporates the temporal information from the SSD that captures timber variations and changes in rhythm for all the critical frequency bands.
- Temporal Rhythm Histogram (TRH): The TRH descriptor captures the changes of rhythmic over time by extracting individual rhythmic histograms from various segments in a piece of audio.

5 Experimental evaluation

This section describes four datasets in detail, and evaluates the performance of the proposed MGC system by conducting experiments on these four datasets for which the audio files are publicly available.

5.1 Datasets

The proposed system of MGC is evaluated on four datasets using classification accuracy as evaluation indicator. In particular, the GTZAN, ISMIR2004 and Homburg dataset are commonly used manually annotated benchmark datasets in music classification domain.

- GTZAN: The GTZAN dataset was created by Tzanetakis and Cook [54], and consists of the following 10 genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock. Each genre is composed of 100 music clips of 30 s duration and each item is stored in WAV format having sampled frequency of 22.05 kHz, 16 bits and mono audio file. The dataset is evaluated based on a randomized ten-fold cross-validation protocol.

- GTZAN-NEW: Due to the availability of the GTZAN dataset, it is widely used in the area of music classification. Note that, the dataset contains some faults such as repetitions, mislabels, and distortions [52]. To address these faults, the GTZAN dataset was processed by Foleis et al. [16] based on the findings in [52]. In this work, we refer to this dataset as GTZAN-NEW. The dataset consists of 948 30 s music clips divided into 10 genres, and the distribution of music clips in each genre is as follows: blues (100), classical (100), country (100), disco (94), hip-hop (98), jazz (87), metal (91), pop (90), reggae (88), rock (100). The results reported below are the average classification accuracy achieved using the tenfold cross-validation protocol.
- ISMIR2004 [4]: The ISMIR2004 dataset contains 1458 full songs distributed 6 genres as follows: classical (640), electronic (229), jazz-blues (52), metal-punk (90), rock-pop (203), world (244). Each song has a full duration and stored as a 44.1 kHz, 16 bits and stereo MP3 file. The training and test sets for the dataset have been specified in the corresponding files. In addition, we deleted audio files with a duration less than 30 s to use the signal segmentation strategy to compare with other classification methods. Therefore, the training set and test set in our experiments are both 724 audio files. This research first converts each stereo MP3 file into a mono WAV file, and then performs feature extraction. Moreover, audio files of ISMIR2004 dataset are down-sampled to 22.05 kHz before AIM feature extraction to reduce the amount of calculation.
- Homburg [24]: The Homburg dataset contains 1886 music clips from 1463 different artists and each clip is associated with a 10 s audio sample. These clips are unequally distributed over 9 genres: alternative (145), blues (120), electronic (113), folk-country (222), funk-soul-rnb (47), jazz (319), pop (116), rap-hiphop (300), rock (504). Audio samples are encoded in MP3 format with a sampling rate of 44.1 KHZ and a bit-rate of 128 mbit/s. In our experiments, we processed the dataset in the same way as the ISMIR2004 dataset, i.e., first converting the audio in stereo MP3 format to mono WAV. In addition, the audio was down-sampled to 22.05 kHz before extracting the auditory features.

5.2 Classifiers

In this work, three classifiers are employed, that is, SVM [3], K-Nearest Neighbor (KNN) [15] and Sparse Representation-based Classification (SRC) [55]. In particular, SVM and KNN have been commonly used in music classification domain, and Panngakis et al. [44] has successfully employed SRC in MGC tasks.

- SVM is a kind of binary classifier based on the principle of large margin. It gives the instance labels of the two categories by finding the optimal separation hyperplane which maximize the distance between the nearest instance in the training set and the hyperplane. Aiming at the multi-class problem of MGC, the one-versus-the-rest strategy of SVM with RBF kernel is adopted.
- KNN is a simple algorithm that uses the labels of K nearest instances in the training set to predict the label of testing instance.
- SRC aims to seek for the sparsest representation of the test samples with respect to the training samples via l_1 norm minimization. The test samples are classified according to the representation coefficient and assigned to the object class with the smallest residual.

5.3 Experimental results

The first experiment aims at comparing the performance obtained by three different classifiers, i.e. SVM, KNN and SRC, on the GTZAN dataset. In this experiment, we consider training different classifiers with acoustic and spectral features independently, and their average classification accuracy and standard deviation of the tenfold cross-validation for all experiments are shown in Table 1. Note that the prefix 'S' denote the spectral feature, for example, 'S-LBP' is the LBP texture descriptor of spectral feature set. These three classification methods pairwise relationships in terms of accuracy were investigated by conducting the Student's t test [7] for 10 repeated evaluation steps. SVM performed significantly better than SRC and KNN with a p value less than 0.05.

Table 1 Classification accuracy (%) and standard deviations on the GTZAN dataset using different classifiers

Feature	Classifiers		
	SVM	KNN	SRC
MFCC	67.7 ± 3.47	65.0 ± 4.34	66.9 ± 4.41
SSD	$\mathbf{74.3 \pm 1.90}$	65.2 ± 2.44	73.5 ± 3.88
RH	43.5 ± 2.69	37.6 ± 3.72	41.5 ± 3.88
MVD	50.6 ± 3.04	38.6 ± 3.32	43.4 ± 4.18
TSSD	60.2 ± 3.22	55.7 ± 4.00	58.7 ± 3.85
TRH	44.2 ± 3.09	28.2 ± 2.18	40.2 ± 3.06
S-LBP	77.9 ± 3.73	65.3 ± 4.08	62.7 ± 3.85
S-LBPHF	70.0 ± 3.95	60.7 ± 3.72	67.4 ± 4.03
S-RICLBP	75.8 ± 3.43	58.8 ± 3.12	61.2 ± 3.16
S-DSLBP	$\mathbf{81.5 \pm 2.73}$	69.0 ± 3.44	59.4 ± 2.76
S-LPQ	81.2 ± 3.66	69.3 ± 3.47	72.3 ± 2.53

The best results for acoustic and spectral features are in bold

Table 2 Evaluation indicators and standard deviations on the GTZAN dataset using SVM classifier

Feature	Evaluation indicators			
	Accuracy(%)	Precision	Recall	F1-score
MFCC	67.7 ± 3.47	0.6848 ± 0.0445	0.6812 ± 0.0454	0.6829 ± 0.0438
SSD	**74.3 ± 1.90**	**0.7402 ± 0.0213**	**0.7413 ± 0.0189**	**0.7407 ± 0.0195**
RH	43.5 ± 2.69	0.4300 ± 0.0337	0.4285 ± 0.0360	0.4291 ± 0.0338
MVD	50.6 ± 3.04	0.5020 ± 0.0307	0.4957 ± 0.0411	0.4994 ± 0.0345
TSSD	60.2 ± 3.22	0.6091 ± 0.0310	0.6116 ± 0.0333	0.6120 ± 0.0313
TRH	44.2 ± 3.09	0.4529 ± 0.0330	0.4343 ± 0.0300	0.4529 ± 0.0291
S-LBP	77.9 ± 3.73	0.7796 ± 0.0408	0.7795 ± 0.0418	0.7794 ± 0.0401
S-LBPHF	70.0 ± 3.95	0.7052 ± 0.0311	0.6968 ± 0.0404	0.7009 ± 0.0401
S-RICLBP	75.8 ± 3.43	0.7591 ± 0.0311	0.7556 ± 0.0354	0.7573 ± 0.0330
S-DSLBP	**81.5 ± 2.73**	0.8163 ± 0.0264	0.8166 ± 0.0297	0.8164 ± 0.0278
S-LPQ	81.2 ± 3.66	**0.8179 ± 0.0319**	**0.8187 ± 0.0300**	**0.8182 ± 0.0301**

The best results for acoustic and spectral features are in bold

Additionally, we further evaluate the classification models using precision, recall, and F1 score, all results are reported in Table 2. Overall, SVM with high accuracy, precision, recall, and F1-score seems to outperform other methods. Therefore, we use SVM classifier for all subsequent experiments. And the SVM classifier with RBF kernel will be employed in all subsequent experiments to eliminate the differences caused by different classifiers. Meanwhile, we will set the same parameters (cost penalty $C = 1000$, gamma $\gamma = 0.1$) just like [38] for all experiments to avoid overfitting. And min-max normalization will be used for features prior to SVM training.

In the following experiment, we will investigate how the classification accuracy is affected by different frequency spectrograms, which include logarithmic and linear frequency spectrogram. Note that this experiment is based on spectral features which are extracted from spectrogram on ISMIR2004 dataset. Specifically, due to the use of the signal segmentation strategy, the final classification decision is obtained by the sum-rule. The sum-rule can be viewed to be computing the average a posteriori probability for each class over all the classifier outputs,

$$SR(v) = \arg\max_{k=1}^{c} \sum_{i=1}^{n} P(\omega_k | l_i(v)), \tag{2}$$

where v denote the pattern to be classified, c is the total number of classes, n is the number of classifiers, $P(\omega_k|l_i(v))$ represents the estimation of probability of pattern v belonging to the class ω_k according to ith classifier.

Table 3 shows the classification accuracy of texture descriptors extracted respectively from the linear frequency spectrogram based on three different sub-windowing strategies and from the logarithmic frequency spectrogram using global sub-windowing strategy. Among then, global scale denote that features are extracted from the whole spectrogram, linear scale represents the spectrogram image is divided into 10 linear zones of equal sized sub-windows and each sub-window will extract different features, and mel scale means the spectrogram image is divided into 15 sub-windows based on the frequency bands according to the human perception. From the results of Table 3, we can see that the advantage of the logarithmic frequency spectrogram

Table 3 Classification accuracy (%) on the ISMIR2004 dataset based on the linear frequency spectrogram and logarithmic frequency spectrogram

Feature	Linear spectrogram [38]			Logarithmic spectrogram
	Global	Linear	Mel	Global
S-LBP	80.5	81.1	81.4	84.4
S-LBPHF	76.7	81.1	80.7	80.7
S-RICLBP	77.3	78.8	79.4	83.7
S-DSLBP	80.2	80.5	80.6	84.5
S-LPQ	78.3	80.6	80.5	**85.5**
S1=S-LBP+S-LBPHF+S-LPQ	**82.9**	80.9	82.0	83.6
S2=S-LBP+S-LBPHF+S-RICLBP	81.9	80.8	80.9	83.8
S3=S-LBP+S-LPQ	81.9	81.9	80.9	84.3

The best results for linear and logarithmic spectrogram are in bold

is greater than the linear for the MGC tasks. In particular, for the linear frequency spectrogram, the accuracy of S1 is the highest, that is 82.9%, but still lower than the corresponding result of the logarithmic, that is 83.6%. And, the S-LPQ texture descriptor extracted from logarithmic frequency spectrogram achieve 85.5% accuracy, which is the best one among all results. The results show that that the logarithmic frequency spectrogram is closer to the human perception of pitch than the linear frequency spectrogram, and the features extracted from the logarithmic frequency spectrogram can describe the music genres more effectively than those extracted from the linear.

Table 4 Classification accuracy(%) and standard deviations of single feature on all the datasets

Feature	Dataset			
	GTZAN	GTZAN-NEW	ISMIR2004	Homburg
MFCC	67.7 ± 3.47	64.5 ± 3.95	70.1	42.6 ± 4.68
SSD	**74.3 ± 1.90**	**72.2 ± 3.88**	81.3	**45.4 ± 3.03**
RH	43.5 ± 2.69	37.6 ± 2.51	64.2	38.0 ± 4.85
MVD	50.6 ± 3.04	45.7 ± 3.02	74.1	41.9 ± 2.13
TSSD	60.2 ± 3.22	56.3 ± 4.66	**82.0**	38.6 ± 2.36
TRH	44.2 ± 3.09	40.1 ± 3.62	68.0	39.3 ± 2.44
S-LBP	77.9 ± 3.73	76.5 ± 3.69	84.4	58.8 ± 2.42
S-LBPHF	70.0 ± 3.95	67.6 ± 3.45	80.7	52.7 ± 3.18
S-RICLBP	75.8 ± 3.43	72.8 ± 3.89	83.7	56.8 ± 3.47
S-DSLBP	**81.5 ± 2.73**	**79.1 ± 3.03**	84.5	**61.0 ± 3.44**
S-LPQ	81.2 ± 3.66	80.2 ± 3.46	**85.5**	59.8 ± 3.09
M-LBP	64.6 ± 3.38	60.2 ± 2.85	71.0	51.9 ± 2.91
M-LBPHF	62.0 ± 3.90	58.5 ± 3.88	**73.8**	50.3 ± 3.00
M-RICLBP	65.5 ± 3.71	61.8 ± 2.57	73.1	51.0 ± 2.20
M-DSLBP	62.8 ± 3.71	58.4 ± 3.67	69.2	52.9 ± 3.56
M-LPQ	**66.8 ± 2.09**	**62.5 ± 2.27**	72.0	**54.2 ± 3.02**

The best results for three feature sets for each dataset are in bold

To verify the validity of the three kinds of features in the four datasets, all the single feature are coupled with SVM classifier in the next experiment. Table 4 presents the comparison results, which utilize spectral feature extracted from the logarithmic frequency spectrogram. Note that the prefix 'M' represents the single feature of AIM feature. From the view of the single feature, S-DSLBP or S-LPQ outperforms other single features on different datasets. In detail, the S-DSLBP texture descriptor performs well both on the GTZAN, GTZAN-NEW and Homburg dataset, and on the ISMIR2004 dataset the S-LPQ texture descriptor achieves highest accuracy. In addition, SSD and MFCC are both superior to other stand-alone acoustic feature in terms of their performance and stability. To reduce the amount of data processing, the extraction of the AIM feature is performed with the middle 10 s part of each music piece. It can be observed from Table 4 that the performance of M-LPQ on the GTZAN, GTZAN-NEW and Homburg dataset is better than other stand-alone AIM features. And on the ISMIR2004 dataset, the best classification accuracy is achieved by adopting the M-LBPHF descriptors. It is worth noting that we only process and classify the 10 s down sampled music clips

Table 5 Classification accuracy (%) and standard deviations of feature combinations on all the datasets

Feature	Dataset			
	GTZAN	GTZAN-NEW	ISMIR2004	Homburg
A1=SSD+MFCC	79.1 ± 3.10	76.6 ± 3.58	85.7	49.1 ± 3.54
A2=SSD+RH+MFCC	**81.3 ± 2.87**	**80**.8 ± 3.82	86.8	54.0 ± 3.19
A3=SSD+MVD+MFCC	78.4 ± 3.32	77.3 ± 3.91	**87.5**	**55.1 ± 3.13**
S1=S-LBP+S-LBPHF+S-LPQ	78.1 ± 3.36	77.0 ± 3.01	83.6	60.2 ± 2.50
S2=S-LBP+S-LBPHF+S-RICLBP	75.8 ± 3.06	73.8 ± 3.95	83.8	60.5 ± 3.33
S3=S-LBP+S-LPQ	80.9 ± 3.33	78.2 ± 3.45	84.3	**61.3 ± 3.24**
S4=S-LBP+S-RICLBP	77.8 ± 2.86	74.5 ± 3.59	**84.8**	61.1 ± 3.11
S5=S-DSLBP+S-LPQ	**80.8 ± 3.25**	**78.6 ± 3.26**	84.0	60.8 ± 2.38
S6=S-LBP+S-RICLBP+S-LPQ	78.4 ± 3.69	76.4 ± 2.69	84.0	59.5 ± 3.00
S7=S-LBP+S-DSLBP+S-LPQ	78.8 ± 3.68	77.5 ± 3.42	83.8	60.6 ± 3.56
F1=A2+S-LPQ	90.4 ± 3.47	86.1 ± 3.52	90.9	63.5 ± 3.49
F2=A2+S3	90.3 ± 2.05	86.0 ± 3.55	90.3	63.9 ± 2.74
F3=A2+S5	90.4 ± 2.26	85.7 ± 2.79	**91.2**	64.2 ± 3.74
F4=A2+S5+M-LBP	91.7 ± 3.32	**86.2 ± 2.25**	90.5	64.6 ± 1.41
F5=A2+S5+M-LBPHF	91.1 ± 2.77	85.4 ± 2.85	90.5	64.7 ± 2.65
F6=A2+S5+M-RICLBP	**91.8 ± 3.02**	86.0 ± 2.34	90.9	**65.2 ± 2.31**
F7=A2+S5+M-DSLBP	91.7 ± 2.49	86.0 ± 3.75	90.5	**65.2 ± 2.87**
F8=A2+S5+M-LPQ	91.7 ± 2.44	85.9 ± 1.60	90.5	65.0 ± 2.67

The best results for three feature combinations for each dataset are in bold

Table 6 Evaluation indicators and standard deviations of test results on all the datasets

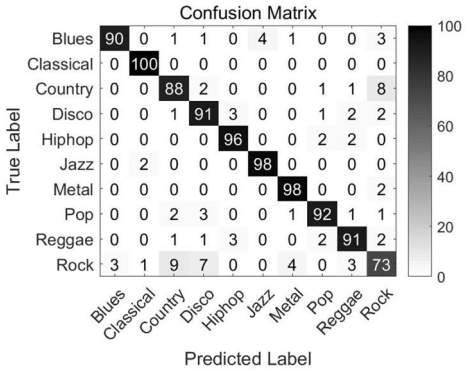
Dataset	Feature	Evaluation indicators			
		Accuracy (%)	Precision	Recall	F1-score
GTZAN	F1=A2+S-LPQ	90.4 ± 3.47	0.9022 ± 0.0336	0.9036 ± 0.0296	0.9028 ± 0.0309
	F2=A2+S3	90.3 ± 2.05	0.8988 ± 0.0205	0.8994 ± 0.0220	0.8991 ± 0.0206
	F3=A2+S5	90.4 ± 3.07	0.9042 ± 0.0325	0.9040 ± 0.0316	0.9042 ± 0.0313
	F4=A2+S5+M-LBP	91.7 ± 3.32	0.9159 ± 0.0323	0.9184 ± 0.0337	0.9172 ± 0.0326
	F5=A2+S5+M-LBPHF	91.1 ± 2.77	0.9150 ± 0.0227	0.9119 ± 0.0273	0.9134 ± 0.0244
	F6=A2+S5+M-RICLBP	**91.8 ± 3.02**	**0.9239 ± 0.0310**	0.9160 ± 0.0334	0.9194 ± 0.0319
	F7=A2+S5+M-DSLBP	91.7 ± 2.49	0.9184 ± 0.0228	0.9200 ± 0.0253	0.9192 ± 0.0238
	F8=A2+S5+M-LPQ	91.7 ± 2.44	0.9229 ± 0.0242	**0.9206 ± 0.0227**	**0.9222 ± 0.0232**
GTZAN-NEW	F1=A2+S-LPQ	86.1 ± 3.52	0.8594 ± 0.0330	0.8634 ± 0.0317	0.8613 ± 0.0317
	F2=A2+S3	86.0 ± 3.55	0.8585 ± 0.0380	0.8526 ± 0.0436	0.8555 ± 0.0405
	F3=A2+S5	85.7 ± 2.79	0.8579 ± 0.0268	0.8535 ± 0.0304	0.8566 ± 0.0279
	F4=A2+S5+M-LBP	**86.2 ± 2.25**	0.8597 ± 0.0205	**0.8681 ± 0.0256**	**0.8639 ± 0.0223**
	F5=A2+S5+M-LBPHF	85.4 ± 2.85	0.8545 ± 0.0318	0.8589 ± 0.0351	0.8567 ± 0.0322
	F6=A2+S5+M-RICLBP	86.0 ± 2.34	0.8661 ± 0.0251	0.8577 ± 0.0176	0.8618 ± 0.0202
	F7=A2+S5+M-DSLBP	86.0 ± 3.75	0.8551 ± 0.0436	0.8577 ± 0.0431	0.8563 ± 0.0429
	F8=A2+S5+M-LPQ	85.9 ± 1.60	**0.8637 ± 0.0176**	0.8617 ± 0.0144	0.8626 ± 0.0152
ISMIR2004	F1=A2+S-LPQ	90.9	0.8395	0.8994	0.8684
	F2=A2+S3	90.3	0.8325	0.9042	0.8669
	F3=A2+S5	**91.2**	**0.8486**	0.9119	0.8791
	F4=A2+S5+M-LBP	90.5	0.8435	0.9216	0.8809
	F5=A2+S5+M-LBPHF	90.5	0.8392	0.9209	0.8782
	F6=A2+S5+M-RICLBP	90.9	0.8431	**0.9239**	**0.8816**
	F7=A2+S5+M-DSLBP	90.5	0.8433	0.9154	0.8779
	F8=A2+S5+M-LPQ	90.5	0.8458	0.9170	0.8799
Homburg	F1=A2+S-LPQ	63.5 ± 3.49	0.4444 ± 0.0325	0.4727 ± 0.0753	0.4572 ± 0.0508
	F2=A2+S3	63.9 ± 2.74	0.4487 ± 0.0268	0.4800 ± 0.0545	0.4630 ± 0.0366
	F3=A2+S5	64.2 ± 3.74	0.4530 ± 0.0270	0.5148 ± 0.0874	0.4791 ± 0.0470
	F4=A2+S5+M-LBP	64.6 ± 1.41	0.4570 ± 0.0169	0.4916 ± 0.0514	0.4728 ± 0.0296
	F5=A2+S5+M-LBPHF	64.7 ± 2.65	0.4574 ± 0.0143	0.4957 ± 0.0619	0.4742 ± 0.0337
	F6=A2+S5+M-RICLBP	**65.2 ± 2.31**	**0.5289 ± 0.0374**	0.4932 ± 0.0220	0.4932 ± 0.0220
	F7=A2+S5+M-DSLBP	**65.2 ± 2.87**	0.4622 ± 0.0296	0.5008 ± 0.0612	0.4799 ± 0.0407
	F8=A2+S5+M-LPQ	65.0 ± 2.67	0.4656 ± 0.0168	**0.5209 ± 0.0317**	**0.4948 ± 0.0188**

The best results for three feature combinations for each dataset are in bold

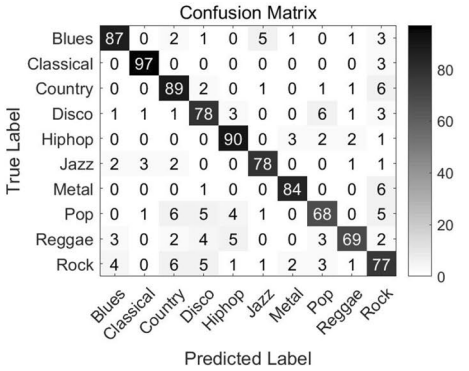
of all complete music recordings on the ISMIR2004 dataset, but still achieve good classification results.

For improving the results of those three kinds of feature set, we will evaluate the classification performances under the different fusions about AIM, spectral and acoustic features on the GTZAN, GTZAN-NEW, ISMIR2004 and Homburg datasets. The experimental results of the different feature fusions are summarized in Table 5. It's important to note that A1–A3 are the acoustic feature combination of direct concatenate feature vectors, while S1–S7 denote the spectral feature combination fused by sum-rule [27]. Moreover, the symbolic representation in Table 5 that F1, F2 and F3 are combinations of spectral and acoustic features, while F4, F5, F6 and F7 are combined with AIM, spectral and acoustic features. And the way they're combined is also the sum-rule.

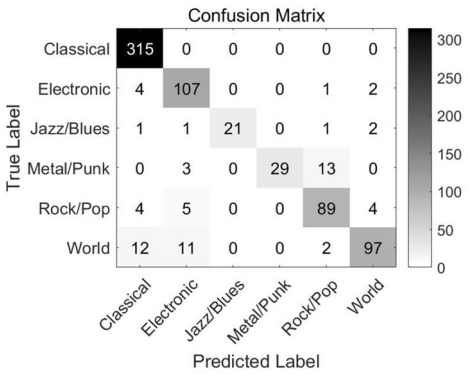
- A1 denotes the combined feature vector by concatenating the feature vectors of SSD, MFCC.
- A2 denotes the combined feature vector by concatenating the feature vectors of SSD, RH, MFCC.
- A3 denotes the combined feature vector by concatenating the feature vectors of SSD, MVD, MFCC.
- S1 denotes the sum-rule among S-LBP, S-LBPHF and S-LPQ.
- S2 denotes the sum-rule among S-LBP, S-LBPHF and S-RICLBP.



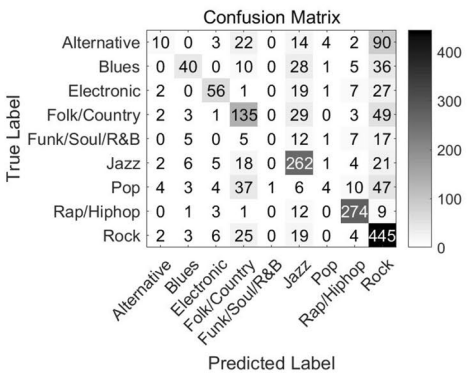
(a) GTZAN



(b) GTZAN-NEW



(c) ISMIR2004



(d) Homburg

Fig. 4 Confusion matrices of the proposed system on different datasets

– S3 denotes the sum-rule among S-LBP, S-LBPHF and S-RICLBP.
– S4 denotes the sum-rule among S-LBP and S-RICLBP.
– S5 denotes the sum-rule among S-DSLBP and S-LPQ.
– S6 denotes the sum-rule among S-LBP, S-RICLBP and S-LPQ.
– S7 denotes the sum-rule among S-LBP, S-DSLBP and S-LPQ.

By inspecting Table 5, S4 or S5 is superior to other feature combinations in the same feature set. Especially, S4 gets the best result on ISMIR2004 datasets, that is 84.8%. Moreover, we find that the performance of the fusion of two or three different feature sets is stable and greatly improve the accuracy of MGC with single feature. On the GTZAN dataset, the best MGC accuracy is obtained by the fusion of A2, S5 and the AIM features M-RICLBP (F6), which reaches 91.8%. Furthermore, based on F4 that combines A2, S5 and AIM feature M-LBP, the system achieves its best classification result on the GTZAN-NEW dataset, that is 86.2%. Similar to the GTZAN and GTZAN-NEW datasets, the combination of features F6 and F7 fusing the three feature sets achieves the highest classification results on the Homburg dataset, that is 65.2%. On the ISMIR2004 dataset, the highest accuracy of the proposed method is 91.2% (F3), which is the best classification accuracy as far as we know. From this table, we can also see that on the ISMIR2004 dataset, the performance of the fusion of three types of features is slightly inferior to that of the ensemble based on spectral and acoustic features. The reason behind this could be that we have performed a down-sampling pre-processing on the all raw audio of the ISMIR2004 dataset for reducing the computational cost caused by the AIM's complexity. Therefore, if further improvement of the AIM could be achieved, the AIM features of the whole audio could be extracted and then all information could be retained, so that the performance of the proposed system could be improved.

Due to the distribution of the ISMIR2004 and Homburg dataset is unbalanced, the values of accuracy, precision, recall and F1-score are used to evaluate the performance of the model. In this experiment, instead of training the classifier with respect to each single feature, consider representative combinations of features. And the mean and standard deviation of the evaluation indicators for ten repetitions of the classification model are shown in Table 6. On the GTZAN dataset, all values reach above 90%. And the values are around 90% on both GTZAN-NEW and ISMIR2004 datasets. In addition, since nearly 30% of the audio chips in the Homburg dataset belong to the rock genre, the training and test samples are unevenly distributed in terms of genre, but their F1-scores also come

Table 7 Comparison with the state-of-the-art classification systems

Method	Classifier	Dataset/accuracy(%)		
		GTZAN	ISMIR2004	Homburg
F3=A2+S5	SVM	90.4 ± 3.07	**91.2**	64.2 ± 3.74
F4=A2+S5+M-LBP	SVM	91.7 ± 3.32	90.5	64.6 ± 1.41
F6=A2+S5+M-RICLBP	SVM	**91.8 ± 3.02**	90.9	**65.2 ± 2.31**
Raw audio signal [1]	CNN	80.9[1]	–	–
Mel-spectrogram [57]	CNN	90.7	–	–
Mel-spectrogram [59]	BRNN + PCNNA	90.0	–	–
Spectral+Acoustic [39]	SVM+AdaBoost	90.6	90.9	–
Spectral+Acoustic [38]	SVM+AdaBoost	89.8	90.2	–
CR [45]	JSLRR	89.40	85.45	63.46
Spectro-temporal features [32]	SVM	87.4	89.9	–
GSV+GF [56]	SVM	86.1	86.1	–
LBP [11]	SVM	–	80.65	–
Feature combination [17]	SVM(SG)	90.9	–	–
Modulation spectral analysis [28]	LDA	90.6	86.83	–
DWCH+timbral features [30]	SVM	80	–	–
Marsyas features [54]	GMM	61	–	–

The best classification results under three datasets are in bold

1 The experimental protocol uses a stratified threefold on the GTZAN dataset, where each track of each fold is split into 21 short 5 s segments

to about 50%. This shows that the prediction results of the model were reliable. By observing the results in Table 6, we can see that F4 and F6 achieve the best classification results on the GTZAN and GTZAN-NEW datasets. In addition, F3 obtains the highest accuracy and precision on the ISMIR dataset, while F6 obtains the best recall and F1-score. Moreover, F6 gets the best accuracy and precision on the Homburg dataset. Based on the above findings, F3, F4 and F6 performed the best results in all datasets. Therefore, we further investigated the pairwise relationship of three combinations of features (F3, F4, F6) on accuracy using the Student's *t* test. The results showed that the accuracy of the F6 training classifier significantly outperformed F3 and F4 with p values less than 0.05. Furthermore, the confusion matrices of the genre classification system for the four datasets are shown in Fig. 4. The columns and rows of each matrix represent the original true genre labels and their predicted labels, respectively. In GTZAN and GTZAN-NEW dataset, classical music was accurately classified, but rock music is the least accurately classified and easily misclassified to country and disco music. In the ISMIR 2004 dataset, 28.8% of metal-punk music is misclassified as rock pop, and other genres in the Homburg dataset are also misclassified as rock music.

To further explain the advantage of the performance of the proposed MGC system, we also compare the proposed system with the state-of-the-art systems on the GTZAN, ISMIR2004 and Homburg datasets, whose results are shown in Table 7. According to Table 7, the combination of features based on AIM, spectral and acoustic features obtains the best performance on the GTZAN and Homburg datasets with 91.8% and 65.2%, respectively. And on the ISMIR2004 dataset, the fusion of spectral features and acoustic features also has the highest classification accuracy, which is 91.2%. Experimental results prove that the proposed MGC system can improve the classification performance, whether it is the fusion of acoustic and spectral features or the fusion of the above three kinds of feature sets.

6 Conclusion

Inspired by the auditory system and the spectrogram feature, this work proposes a novel music genre classification framework that integrates the auditory image, spectral and acoustic features. In particular, auditory image feature is extracted based on the auditory image model to simulate the human cochlea's processing. And spectral feature is calculated from the logarithmic frequency spectrogram other than linear one. Acoustic feature is composed of traditional musical features MFCC and psycho-acoustic music descriptors to extract appropriate semantic information from music. To evaluate the classification accuracy of the proposed MGC system, we conducted a set of experiments on different benchmark datasets, such as GTZAN, GTZAN-NEW, ISMIR2004 and Homburg datasets. We find that three type features fusion, that is auditory image,

spectral and acoustic feature, or the combination of spectral and acoustic feature has more discriminating compare to the traditional music features for MGC tasks. Moreover, the logarithmic frequency spectrogram have better classification accuracy indeed than those based on linear since the logarithmic spectrogram has the ability to capture low frequency information. All experimental results show that the proposed method owns the higher classification accuracy and the better stability than many state-of-the-art classification methods.

Acknowledgements We thank all the referees and the editorial board members for their insightful comments and suggestions, which improved our paper significantly. This study was funded by the National Natural Science Foundation of China under the Grants No. 11501351.

References

1. Allamy, S., Koerich, A.L.: 1D CNN Architectures for Music Genre Classification. arXiv preprint arXiv:210507302 (2021)
2. Bleeck, S., Ives, T., Patterson, R.: Aim-mat: the auditory image model in matlab. Acta Acust. Acust. **90**, 781–787 (2004)
3. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp 144–152 (1992). https://doi.org/10.1145/130385.130401
4. Cano, P., Gômez, E., Gouyon, F., Herrera, P., Koppenberger, M., Ong, B., Serra, X., Streich, S., Wack, N.: ISMIR 2004 Audio Description Contest. Technical Report. Music Technology Group, Bracelona (2006)
5. Castillo, J.R., Flores, M.J.: Web-based music genre classification for timeline song visualization and analysis. IEEE Access **9**, 18801–18816 (2021). https://doi.org/10.1109/ACCESS.2021.3053864
6. Chaki, J.: Pattern analysis based acoustic signal processing: a survey of the state-of-art. Int. J. Speech Technol. (2020). https://doi.org/10.1007/s10772-020-09681-3
7. Chan, W.C., Liang, P.H., Shih, Y.P., Yang, U.C., Chang Lin, W., Hsu, C.N.: Learning to predict expression efficacy of vectors in recombinant protein production. BMC Bioinform. **11**(1), 1–12 (2010)
8. Çoban, Ö., Özyer, G.T.: Music genre classification from turkish lyrics. In: 2016 24th Signal Processing and Communication Application Conference (SIU), pp 101–104 (2016). https://doi.org/10.1109/SIU.2016.7495686
9. Çoban, Ö.: Turkish music genre classification using audio and lyrics features. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi **21**(2), 322–331 (2017)
10. Corrêa, D.C., Rodrigues, F.A.: A survey on symbolic data-based music genre classification. Expert Syst. Appl. **60**, 190–210 (2016). https://doi.org/10.1016/j.eswa.2016.04.008
11. Costa, Y., Oliveira, L., Koerich, A., Gouyon, F.: Music genre recognition using spectrograms. In: 2011 18th International Conference on Systems, Signals and Image Processing, pp 1–4 (2011)
12. Costa, C.H.L., Valle, J.D., Koerich, A.L., Koerich, R.L.: Automatic classification of audio data. IEEE Trans. Syst. Man Cybernet. **1**, 562–567 (2004). https://doi.org/10.1109/ICSMC.2004.1398359
13. Costa, Y., Oliveira, L., Koerich, A., Gouyon, F., Martins, J.: Music genre classification using lbp textural features. Signal Process. **92**(11), 2723–2737 (2012). https://doi.org/10.1016/j.sigpro.2012.04.023
14. Costa, Y., Oliveira, L., Koerich, A., Gouyon, F.: Music genre recognition using gabor filters and lpq texture descriptors. Progress Pattern Recogn. Image Anal. Comput. Vis. Appl. **8259**, 67–74 (2013). https://doi.org/10.1007/978-3-642-41827-3_9
15. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**(1), 21–27 (1967). https://doi.org/10.1109/TIT.1967.1053964
16. Foleis, J.H., Tavares, T.F.: Texture selection for automatic music genre classification. Appl. Soft Comput. **89**, 106–127 (2020). https://doi.org/10.1016/j.asoc.2020.106127
17. Fu, Z., Lu, G., Ting, K., Zhang, D.: On feature combination for music classification. In: Structural, Syntactic, and Statistical Pattern Recognition, pp 453–462 (2010)
18. Fu, Z., Lu, G., Ting, K.M., Zhang, D.: A survey of audio-based music classification and annotation. IEEE Trans. Multimedia **13**(2), 303–319 (2011). https://doi.org/10.1109/TMM.2010.2098858
19. Glasberg, B., Moore, B.: Derivation of auditory filter shapes from notched-noise data. Hear. Res. **47**(1), 103–138 (1990). https://doi.org/10.1016/0378-5955(90)90170-T
20. Glasberg, B., Moore, B.: Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise. J. Acoust. Soc. Am. **108**(5), 2318–2328 (2000). https://doi.org/10.1121/1.1315291
21. Glasberg, B., Moore, B.: A model of loudness applicable to time-varying sounds. J. Audio Eng. Soc. **50**, 331–342 (2002)
22. Gogate, M., Dashtipour, K., Hussain, A.: Visual Speech In Real Noisy Environments (VISION): A Novel Benchmark Dataset and Deep Learning-Based Baseline System. In: Proceeding Interspeech 2020, pp 4521–4525 (2020b). https://doi.org/10.21437/Interspeech.2020-2935
23. Gogate, M., Dashtipour, K., Adeel, A., Hussain, A.: Cochleanet: a robust language-independent audio-visual model for speech enhancement. Inf. Fus. **63**, 273–285 (2020). https://doi.org/10.1016/j.inffus.2020.04.001
24. Homburg, H., Mierswa, I., Möller, B., Morik, K., Wurst, M.: A benchmark dataset for audio classification and clustering. ISMIR **2005**, 528–531 (2005)
25. Hyder, R., Ghaffarzadegan, S., Feng, Z., Hansen, J., Hasan, T.: Acoustic Scene Classification using a CNN-Supervector System Trained with Auditory and Spectrogram Image Features. pp. 3073–3077 (2017). https://doi.org/10.21437/Interspeech.2017-431
26. Irino, T., Patterson, R.: A dynamic compressive gammachirp auditory filterbank. IEEE Trans. Audio Speech Lang. Process. **14**(6), 2222–2232 (2006). https://doi.org/10.1109/TASL.2006.874669
27. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**(3), 226–239 (1998). https://doi.org/10.1109/34.667881
28. Lee, C.H., Shih, J.L., Yu, K.M., Lin, H.S.: Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. IEEE Trans. Multimedia **11**, 670–682 (2009). https://doi.org/10.1109/TMM.2009.2017635
29. Li, T.L., Chan, A.B.: Genre classification and the invariance of mfcc features to key and tempo. In: International Conference on MultiMedia Modeling, Springer, pp 317–327 (2011)
30. Li, T., Ogihara, M.: Toward intelligent music information retrieval. IEEE Trans. Multimedia **8**(3), 564–574 (2006). https://doi.org/10.1109/TMM.2006.870730
31. Lidy, T., Rauber, A.: Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In: Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR 2005), pp 34–41 (2005)
32. Lim, S., Lee, J., Jang, S., Lee, S., Kim, M.Y.: Music-genre classification system based on spectro-temporal features and feature

selection. IEEE Trans. Consum. Electron. **58**(4), 1262–1268 (2012). https://doi.org/10.1109/TCE.2012.6414994
33. Martens, J.P., Leman, M., Baets, B., Meyer, H.: A comparison of human and automatic musical genre classification. IEEE Int. Conf. Acoustics Speech Signal Process. **4**, 233–236 (2004)
34. McKay, C., Fujinaga, I.: Improving automatic music classification performance by extracting features from different types of data. In: Proceedings of the International Conference on Multimedia Information Retrieval. pp. 257–266 (2010). https://doi.org/10.1145/1743384.1743430
35. Mitrović, D., Zeppelzauer, M., Breiteneder, C.: Features for content-based audio retrieval. In: Advances in Computers: Improving the Web, vol 78, Elsevier. pp .71–150 (2010). https://doi.org/10.1016/S0065-2458(10)78003-7
36. Muller, F., Mertins, A.: On using the auditory image model and invariant-integration for noise robust automatic speech recognition. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4905–4908 (2012). https://doi.org/10.1109/ICASSP.2012.6289019
37. Munkong, R., Juang, B.: Auditory perception and cognition. IEEE Signal Process. Mag. **25**(3), 98–117 (2008). https://doi.org/10.1109/MSP.2008.918418
38. Nanni, L., Costa, Y., Lumini, A., Kim, M.Y., Baek, S.R.: Combining visual and acoustic features for music genre classification. Expert Syst. Appl. **45**, 108–117 (2016). https://doi.org/10.1016/j.eswa.2015.09.018
39. Nanni, L., Costa, Y., Lucio, D., Silla, C., Brahnam, S.: Combining visual and acoustic features for audio classification tasks. Pattern Recogn. Lett. **88**, 49–56 (2017). https://doi.org/10.1016/j.patrec.2017.01.013
40. Nonaka, R., Emoto, T., Abeyratne, U.R., Jinnouchi, O., Kawata, I., Ohnishi, H., Akutagawa, M., Konaka, S., Kinouchi, Y.: Automatic snore sound extraction from sleep sound recordings via auditory image modeling. Biomed. Signal Process. Control **27**, 7–14 (2016). https://doi.org/10.1016/j.bspc.2015.12.009
41. Nosaka, R., Suryanto, C.H., Fukui, K.: Rotation invariant co-occurrence among adjacent lbps. In: Park, J.I., Kim, J. (eds.) Computer Vision - ACCV 2012 Workshops, pp. 15–25. Springer, Heidelberg (2013)
42. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002). https://doi.org/10.1109/TPAMI.2002.1017623
43. Ojansivu, V., Heikkilä, J.: Blur insensitive texture classification using local phase quantization. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) Image and Signal Processing, pp. 236–243. Springer, Heidelberg (2008)
44. Panagakis, Y., Kotropoulos, C.L., Arce, G.R.: Music genre classification using locality preserving non-negative tensor factorization and sparse representations. In: ISMIR, pp 249–254 (2009)
45. Panagakis, Y., Kotropoulos, C.L., Arce, G.R.: Music genre classification via joint sparse low-rank representation of audio features. IEEE/ACM Trans. Audio Speech Language Process. **22**(12), 1905–1917 (2014). https://doi.org/10.1109/TASLP.2014.2355774
46. Patterson, R., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., Allerhand, M.: Complex sounds and auditory images. In: Cazals, Y., Horner, K., Demany, L. (eds) Auditory Physiology and Perception, Pergamon. pp. 429–446 (1992). https://doi.org/10.1016/B978-0-08-041847-6.50054-X
47. Patterson, R.D., Allerhand, M.H., Giguère, C.: Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform. Acoust. Soc. Am. J. **98**(4), 1890–1894 (1995). https://doi.org/10.1121/1.414456
48. Perrot, D., Gjerdigen, R.: Scanning the dial: an exploration of factors in the identification of musical style. In: Proceedings of the 1999 Society for Music Perception and Cognition, p 88 (1999)
49. Qiu, L., Li, S., Sung, Y.: 3D-DCDAE: Unsupervised music latent representations learning method based on a deep 3d convolutional denoising autoencoder for music genre classification. Mathematics **9**(18), 2274 (2021). https://doi.org/10.3390/math9182274
50. Qiu, L., Li, S., Sung, Y.: DBTMPE: Deep bidirectional transformers-based masked predictive encoder approach for music genre classification. Mathematics **9**(5), 530 (2021). https://doi.org/10.3390/math9050530
51. Schindler, A., Rauber, A.: An audio-visual approach to music genre classification through affective color features. In: Hanbury A, Kazai G, Rauber A, Fuhr N (eds) Advances in Information Retrieval. pp. 61–67 (2015). https://doi.org/10.1007/978-3-319-16354-3_8
52. Sturm, B.L.: The GTZAN dataset: its contents, its faults, their effects on evaluation, and its future use. CoRR abs/1306.1461:1–29 (2013)
53. Tsaptsinos, A.: Lyrics-based music genre classification using a hierarchical attention network. CoRR abs/1707.04678 (2017)
54. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Trans. Speech Audio Process. **10**(5), 293–302 (2002). https://doi.org/10.1109/TSA.2002.800560
55. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2), 210–227 (2009). https://doi.org/10.1109/TPAMI.2008.79
56. Wu, M., Chen, Z., Jang, J.R., Ren, J., Li, Y., Lu, C.: Combining visual and acoustic features for music genre classification. In: 2011 10th International Conference on Machine Learning and Applications and Workshops, vol 2, pp. 124–129 (2011). https://doi.org/10.1109/ICMLA.2011.48
57. Yang, H., Zhang, W.Q.: Music genre classification using duplicated convolutional layers in neural networks. In: Proc. Interspeech 2019, pp. 3382–3386 (2019). https://doi.org/10.21437/Interspeech.2019-1298
58. Ylioinas, J., Hadid, A., Guo, Y., Pietikäinen, M.: Efficient image appearance description using dense sampling based local binary patterns. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) Computer Vision - ACCV 2012, pp. 375–388. Springer, Heidelberg (2013)
59. Yu, Y., Luo, S., Liu, S., Qiao, H., Liu, Y., Feng, L.: Deep attention based music genre classification. Neurocomputing **372**, 84–91 (2020). https://doi.org/10.1016/j.neucom.2019.09.054
60. Zhao, G., Ahonen, T., Matas, J., Pietikainen, M.: Rotation-invariant image and video description with local binary pattern features. IEEE Trans. Image Process. **21**(4), 1465–1477 (2012). https://doi.org/10.1109/TIP.2011.2175739