Fusion of Static and Transitional Information of Cepstral and Spectral Features for Music Genre Classification

Chang-Hsing Lee, Jau-Ling Shih, Kun-Ming Yu, Hwai-San Lin, and Ming-Hui Wei
Department of Computer Science and Information Engineering, Chung Hua University
E-mail:{chlee, sjl, yu, m09502029, m09502064}@chu.edu.tw

Abstract
In this paper, an automatic music genre classification approach which integrates the features derived from static and transitional information of cepstral (MFCC) and spectral (OSC) features will be proposed. MFCC and OSC capture the characteristics of one audio frame. Therefore, the transitional information, including delta-MFCC, delta-OSC, delta-delta-MFCC, and delta-delta-OSC, are then extracted and combined with MFCC and OSC to improve the classification accuracy. Two information fusion techniques, including feature level fusion and decision level fusion, are developed to combine the extracted feature vectors. Experiments conducted on the music database employed in the ISMIR2004 Audio Description Contest have shown that the proposed approach can achieve a classification accuracy of 84.23%, which is better than the winner of the ISMIR2004 Music Genre Classification Contest.

1. Introduction
For the organization of a music retrieval system, music genre classification plays an important and preliminary role. Generally, a music genre classification problem is defined as the labeling of a music track to a proper genre. Therefore, an unlabeled music track can be placed in the appropriate section of a music database or an online music store. A number of supervised classification techniques have been developed to deal with this problem [1-8].

To determine the music genre of a music track, some discriminating audio features have to be extracted through content-based analysis of the music signal. Typically, short-term features which describe the timbral characteristics of audio signals are first extracted from every short time window (or frame) during which the audio signal is assumed to be stationary. The timbral characteristics generally exhibit the properties related to instrumentations or sound sources such as music, speech, features include zero rate (ZCR), spectral centroid, spectral bandwidth, spectral flux, spectral rolloff, Mel-frequency cepstral coefficients (MFCC), discrete wavelet transform coefficients [2, 9], octave-based spectral contrast (OSC) [3, 4], etc. Generally, the transitional information between short-term features of two neighboring frames, including delta and delta-delta values, have been proven useful in speech recognition system [10, 11].

Once the features are extracted from a music track, a classifier will be employed to determine the music genre of the given music track. Several supervised learning techniques, such as K-nearest neighbor (KNN) [1, 2], linear discriminant analysis (LDA) [2], Gaussian mixture models (GMM) [1, 2, 4], hidden Markov models (HMM) [12], Adaboost [13], and support vector machines (SVM) [2, 14], have been employed for audio classification.

In this paper, the static as well as transitional features of MFCC and OSC will be employed for music genre classification. These features include MFCC, delta-MFCC (ΔMFCC), delta-delta-MFCC (ΔΔMFCC), OSC, delta-OSC (ΔOSC), and delta-delta-OSC (ΔΔOSC).

2. The proposed music genre classification system
The proposed music genre classification system consists of two phases: the training phase and the classification phase. The training phase consists of two main modules: feature extraction and linear discriminant analysis (LDA) [15]. The classification phase consists of three modules: feature extraction, LDA transformation, and classification. A detailed description of each module will be described below.

2.1. Feature extraction
A feature set derived from MFCC and OSC will be used to capture both static and transitional information of music signals. These features will then be combined to improve the classification performance.

2.1.1. Mel-frequency cepstral coefficients (MFCC)
MFCC have been widely used for speech recognition due to their ability to represent the speech spectrum in a compact form. In fact, MFCC have been proven to be very effective in automatic speech recognition and in
2.1.2. Octave-based spectral contrast (OSC). OSC was developed to represent the spectral characteristics of a music signal [3, 4]. It considers the spectral peak and valley in each subband independently. In general, spectral peaks correspond to harmonic components and spectral valleys the non-harmonic components or noise in music signals. Therefore, the difference between spectral peaks and spectral valleys will reflect the spectral contrast distribution.

To compute the OSC features, FFT is first employed to obtain the spectrum of each audio frame. This spectrum is then divided into a number of subbands by using a set of Mel-scale band-pass filters. Let \( E(b) \), \( 0 \leq b < B \), denote the magnitude response of the \( b \)-th Mel-scale band-pass filters, where \( B \) is the total number of filters (\( B = 25 \) in this study). MFCC can be obtained by applying DCT on the logarithm of \( E(b) \):

\[
MFCC(l) = \sum_{b=0}^{L} \log_{10}(1 + E(b)) \cos(l \frac{\pi}{B} (b + 0.5)),
\]

where \( L \) is the length of MFCC feature vector (\( L = 20 \) in the study). Therefore, the MFCC feature vector for the \( t \)-th audio frame can be represented as follows:

\[
x_{t}^{MFCC} = [MFCC_{t}(0), \ldots, MFCC_{t}(L - 1)]^{T}.
\]

To compute the OSC features, FFT is first employed to obtain the spectrum of each audio frame. This spectrum is then divided into a number of subbands by using the set of octave scale filters shown in Table 1. Let \( \{M_{b,1}, M_{b,2}, \ldots, M_{b,N_b}\} \) denote the magnitude spectrum within the \( b \)-th subband, \( N_b \) is the number of FFT frequency bins in the \( b \)-th subband. Without loss of generality, let the magnitude spectrum be sorted in a decreasing order, that is, \( M_{b,1} \geq M_{b,2} \geq \cdots \geq M_{b,N_b} \). The spectral peak and spectral valley in the \( b \)-th subband are then estimated as follows:

\[
Peak(b) = \log\left(\frac{1}{\alpha N_b} \sum_{i=1}^{\alpha N_b} M_{b,i}\right),
\]

\[
Valley(b) = \log\left(\frac{1}{\alpha N_b} \sum_{i=1}^{\alpha N_b} M_{b,N_b-i+1}\right),
\]

where \( \alpha \) is a neighborhood factor (\( \alpha = 0.2 \) in this study).

The spectral contrast is given by the difference between the spectral peak and the spectral valley:

\[
SC(b) = Peak(b) - Valley(b).
\]

The feature vector of an audio frame consists of the spectral contrasts and the spectral valleys of all subbands. Thus, the OSC feature vector of the \( t \)-th audio frame can be represented as follows:

\[
x_{t}^{OSC} = [Valley_{t}(0), \ldots, Valley_{t}(B - 1), \ldots, Valley_{t}(B - 1)]^{T}.
\]

where \( B \) is the number of octave scale filters (\( B = 10 \) as shown in Table 1).

2.1.3. Delta and delta-delta features of MFCC and OSC. MFCC and OSC capture the cepstral and spectral characteristics of one audio frame. To describe the transitional information between adjacent audio frames, \( \Delta \text{MFCC}, \Delta \text{OSC}, \Delta \Delta \text{MFCC}, \) and \( \Delta \Delta \text{OSC} \) will be employed and combined with MFCC and OSC for audio classification.

The delta coefficients of MFCC and OSC can be derived by using the following equations:

\[
\Delta \text{MFCC}_{t}(l) = \text{MFCC}_{t}(l) - \text{MFCC}_{t-1}(l),
\]

\[
\Delta \text{OSC}_{t}(k) = \text{OSC}_{t}(k) - \text{OSC}_{t-1}(k),
\]

where \( 0 \leq l < L, 1 \leq t < N - 1 \), and \( 0 \leq k < 2B, 1 \leq t < N - 1 \), where \( \Delta \text{MFCC}_{t}(l) \) and \( \Delta \text{OSC}_{t}(k) \) denote respectively the \( l \)-th and \( k \)-th delta coefficients of MFCC and OSC of the \( t \)-th audio frame, \( N \) is the total number of frames.

Similarity, the delta-delta coefficients can be obtained by using the following equations:

\[
\Delta \Delta \text{MFCC}_{t}(l) = \Delta \text{MFCC}_{t}(l) - \Delta \text{MFCC}_{t-1}(l),
\]

\[
\Delta \Delta \text{OSC}_{t}(k) = \Delta \text{OSC}_{t}(k) - \Delta \text{OSC}_{t-1}(k),
\]

where \( \Delta \Delta \text{MFCC}_{t}(l) \) and \( \Delta \Delta \text{OSC}_{t}(k) \) denote respectively the \( l \)-th and \( k \)-th delta-delta coefficients of MFCC and OSC of the \( t \)-th audio frame.

To represent the whole music piece, the mean and standard deviation of each feature value (MFCC, \( \Delta \text{MFCC}, \Delta \Delta \text{MFCC}, \text{OSC}, \Delta \text{OSC}, \Delta \Delta \text{OSC} \) of all frames are calculated to form the feature vector:

\[
u_{t}^{MFCC} = \frac{1}{N} \sum_{l=0}^{N-1} \text{MFCC}_{t}(l),
\]

\[
\sigma_{t}^{MFCC} = \left( \frac{1}{N} \sum_{l=0}^{N-1} (\text{MFCC}_{t}(l) - \nu_{t}^{MFCC})^{2} \right)^{1/2},
\]

\[
u_{t}^{OSC} = \frac{1}{N} \sum_{k=0}^{N-1} \text{OSC}_{t}(k),
\]

\[
\sigma_{t}^{OSC} = \left( \frac{1}{N} \sum_{k=0}^{N-1} (\text{OSC}_{t}(k) - \nu_{t}^{OSC})^{2} \right)^{1/2},
\]

\[
u_{t}^{\Delta \text{MFCC}} = \frac{1}{N - 2} \sum_{l=1}^{N-2} \Delta \text{MFCC}_{t}(l),
\]

\[
\sigma_{t}^{\Delta \text{MFCC}} = \left( \frac{1}{N - 2} \sum_{l=1}^{N-2} (\Delta \text{MFCC}_{t}(l) - \nu_{t}^{\Delta \text{MFCC}})^{2} \right)^{1/2},
\]

\[
u_{t}^{\Delta \Delta \text{MFCC}} = \frac{1}{N - 2} \sum_{l=1}^{N-2} \Delta \Delta \text{MFCC}_{t}(l),
\]

\[
\sigma_{t}^{\Delta \Delta \text{MFCC}} = \left( \frac{1}{N - 2} \sum_{l=1}^{N-2} (\Delta \Delta \text{MFCC}_{t}(l) - \nu_{t}^{\Delta \Delta \text{MFCC}})^{2} \right)^{1/2},
\]

\[
u_{t}^{\Delta \text{OSC}} = \frac{1}{N - 2} \sum_{k=1}^{N-2} \Delta \text{OSC}_{t}(k),
\]

\[
\sigma_{t}^{\Delta \text{OSC}} = \left( \frac{1}{N - 2} \sum_{k=1}^{N-2} (\Delta \text{OSC}_{t}(k) - \nu_{t}^{\Delta \text{OSC}})^{2} \right)^{1/2},
\]

\[
u_{t}^{\Delta \Delta \text{OSC}} = \frac{1}{N - 2} \sum_{k=1}^{N-2} \Delta \Delta \text{OSC}_{t}(k),
\]

\[
\sigma_{t}^{\Delta \Delta \text{OSC}} = \left( \frac{1}{N - 2} \sum_{k=1}^{N-2} (\Delta \Delta \text{OSC}_{t}(k) - \nu_{t}^{\Delta \Delta \text{OSC}})^{2} \right)^{1/2}.
\]
Feature vector normalization. Since the dispersion is not identical for each feature value, a linear normalization will be employed to make the range of each feature value between 0 and 1:

\[ x'(m) = \frac{x(m) - x_{\text{min}}(m)}{x_{\text{max}}(m) - x_{\text{min}}(m)} \]

where \( x'(m) \) denotes the normalized \( m \)-th feature value, \( x_{\text{max}}(m) \) and \( x_{\text{min}}(m) \) denote respectively the maximum and minimum of the \( m \)-th feature values of all training music tracks. These reference values, \( x_{\text{max}}(m) \) and \( x_{\text{min}}(m) \), are computed during the training phase and are stored for later reference. In the classification phase, for actual normalization, each feature value extracted from the current music signal is modified using the reference maximum and minimum values to obtain its corresponding normalized values.

Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) [15] aims at improving the classification accuracy at a lower dimensional feature vector space. LDA deals with discrimination between various classes rather than representations of classes. The objective of LDA is to minimize the between-class distance while maximizing the within-class distance. In LDA, an optimal transformation matrix from an \( H \)-dimensional feature space to an \( h \)-dimensional space is determined, where \( h \leq H \). The most widely used transformation matrix is a linear mapping that maximizes the so-called Fisher criterion \( J_F \):

\[ J_F(A) = tr((A^T S_W A)^{-1}(A^T S_B A)) \]

where \( S_W \) and \( S_B \) denote the within-class scatter matrix and between-class scatter matrix, respectively. The within-class scatter matrix is defined as:

\[ S_W = \sum_{c=1}^{C} \sum_{n=1}^{N_c} (x_{c,n} - \bar{x}_c)(x_{c,n} - \bar{x}_c)^T \]

where \( x_{c,n} \) is the \( n \)-th feature vector labeled as class \( c \), \( \bar{x}_c \) is the mean vector of class \( c \), \( C \) is the total number of music classes, and \( N_c \) is the number of training vectors labeled as class \( c \). The between-class scatter matrix is given by:

\[ S_B = \sum_{c=1}^{C} N_c (\bar{x}_c - \bar{x})(\bar{x}_c - \bar{x})^T \]

where \( \bar{x} \) is the mean vector of all training vectors. From the definition of the Fisher criterion \( J_F \), we can see that LDA tries to find a transformation matrix that maximizes the ratio of between-class scatter to within-class scatter in a lower-dimensional space. The optimal solution is the \( H \times h \) transformation matrix, \( A_{LDA} \), given by:

\[ A_{LDA} = \arg\max_A \frac{tr(A^T S_B A)}{tr(A^T S_W A)} \]

The optimal transformation matrix \( A_{LDA} \) can be determined by finding the eigenvectors of \( S_W^{-1}S_B \). The \((C-1)\) columns of \( A_{LDA} \) are formed by the eigenvectors corresponding to the largest \((C-1)\) eigenvalues. The optimal transformation matrix \( A_{LDA} \) is employed to transform each normalized \( H \)-dimensional feature vector to a lower \( h \)-dimensional vector. Let \( x \) denote an \( H \)-dimensional feature vector, the reduced \( h \)-dimensional feature vector can be computed by:

\[ y = A_{LDA}^T x \]

Music genre classification phase

In the classification phase, the same linear normalization process is applied to each feature value. The normalized feature vector is then transformed to be a lower-dimensional feature vector by using the LDA transformation matrix \( A_{LDA} \). Let \( y \) denotes the LDA transformed feature vector. In this study, the nearest centroid classifier is used for music genre classification. For the \( c \)-th (\( 1 \leq c \leq C \)) music genre, the centroid of LDA transformed feature vectors of all training music tracks labeled as the \( c \)-th music genre is regarded as its representative feature vector:

\[ \bar{y}_c = \frac{1}{N_c} \sum_{n=1}^{N_c} y_{c,n} \]

where \( y_{c,n} \) denotes the LDA transformed feature vector of the \( n \)-th music track labeled as the \( c \)-th music genre, \( \bar{y}_c \) is the representative feature vector of the \( c \)-th music
music genre, and \( N_c \) is the number of training music tracks labeled as the \( c \)-th music genre. The distance between two feature vectors is measured by Euclidean distance. Thus, the subject code \( s \) that denotes the identified music genre is determined by finding the representative feature vector that has minimum Euclidean distance to \( y \):

\[
s = \arg \min_{y \in SC} d(y, \mathbf{y}).
\]

### 2.4. Information fusion

Information fusion techniques usually deal with the combination of different sources of information. The most widely used information fusion approaches can be divided into several categories: sensor data level fusion, feature level fusion, score fusion, and decision fusion [17, 18, 19]. In this study, both feature level fusion as well as decision level fusion approaches will be employed to combine the set of extracted audio feature vectors.

#### 2.4.1. Feature level fusion

For feature level fusion, all feature vectors are concatenated together to form a longer feature vector as follows:

\[
\mathbf{x} = [\mathbf{x}^{T}_{MFCC}, \mathbf{x}^{T}_{OSC}, \mathbf{x}^{T}_{MFCC}, \mathbf{x}^{T}_{OSC}, \mathbf{x}^{T}_{AOSC}, \mathbf{x}^{T}_{AOASC}]^T.
\]

LDA transformation is then applied to the combined feature vector for classification purpose.

#### 2.4.2. Decision level fusion

For decision level fusion, the summation rule and product rule will be employed to determine the classification result. For summation rule, each feature vector \((x_{MFCC}, x_{AMMFCC}, x_{AOSC}, x_{AOSC}, x_{AOASC}, x_{AOASC})\) will be first LDA transformed, the distance for each transformed feature vector will then be evaluated. Let \(d_{MFCC}(c), d_{AMMFCC}(c), d_{OSC}(c), d_{AOSC}(c), \) and \(d_{AOASC}(c)\) denote the distances between each transformed vector and the corresponding representative feature vector of the \( c \)-th class. The overall distance for the \( c \)-th music genre is defined as follows:

\[
d(c) = d_{MFCC}(c) + d_{AMMFCC}(c) + d_{OSC}(c) + d_{AOSC}(c) + d_{AOASC}(c), \quad 1 \leq c \leq C.
\]

For product rule, the overall distance is evaluated by multiplying the distance computed from each feature vector:

\[
d(c) = d_{MFCC}(c) \times d_{AMMFCC}(c) \times d_{OSC}(c) \times d_{AOSC}(c) \times d_{AOASC}(c), \quad 1 \leq c \leq C.
\]

### 3. Experimental results

The music database employed in the ISMIR2004 Audio Description Contest [20] is used for performance comparison. The database consists of 1458 music tracks in which 729 music tracks are used for training and the other 729 tracks for testing. The audio file format is 44.1 kHz, 128 kbps, 16 bits per sample, stereo MP3 files. In this study, each MP3 audio file is first converted into raw digital audio format before classification. These music tracks are classified into six classes (that is, \( C = 6 \)): Classical, Electronic, Jazz/Blue, Metal/Punk, Rock/Pop, and World. In summary, the music tracks used for training/testing include 320/320 tracks of Classical, 115/114 tracks of Electronic, 26/26 tracks of Jazz/Blue, 45/45 tracks of Metal/Punk, 101/102 tracks of Rock/Pop, and 122/122 tracks of World music genre.

Table 2 shows the average classification accuracy of different feature set with different information fusion methods. In this table, fusion1, fusion2, and fusion3 denote respectively feature level fusion, decision level fusion with summation rule, and decision level fusion with product rule. From this Table, we can see that for decision level fusion approach the performance will be improved when additionally transitional information is combined. But, for feature level fusion technique, we can see that when \( \Delta \)MFCC and \( \Delta \)OSC are combined with other feature vectors, no performance improvement can be obtained than other feature combinations. That is because \( \Delta \)MFCC and \( \Delta \)AMMFCC (similarly \( \Delta \)OSC and \( \Delta \)AOASC) present similar characteristics and the length of the feature dimension becomes too longer. However, for fusion2 and fusion3, each feature vector is independently treated and used for classification. Therefore, the length of the feature dimension will not become too long.

Table 3 shows the comparison with the results from the ISMIR2004 Music Genre Classification Contest. From this table, we can see that our proposed method (fusion2:MFCC+\( \Delta \)MFCC+\( \Delta \)AMMFCC+OSC+\( \Delta \)OSC+\( \Delta \)AOASC) achieves a classification accuracy of 84.23%, which is better than the winner of the contest.

Fig. 2 shows the confusion matrix of each feature set with different fusion approach. The confusion matrix demonstrates which tracks are correctly classified or not depending on the class. It is read as "<row>" is classified as <col>". For each row or column, the music genres are arranged in the order of Classical (C), Electronic (E), Jazz/Blue (J), Metal/Punk (M), Rock/Pop (R), and World (W). For instance, the number at the first column and last row represents the number of Classical tracks being classified as World music. A perfect matrix only contains numbers in the diagonal. Comparing Fig. 2(g), 2(h), and 2(i), we can see that the fusion2 and fusion3 can get a higher classification accuracy than fusion1 for Classical and Metal/Punk.

### 4. Conclusion
In this paper, we propose an automatic music genre classification approach by integrating the features derived from static and transitional information of cepstral (MFCC) and spectral (OSC) features. The extracted feature sets, including MFCC, ΔMFCC, ΔΔMFCC, OSC, ΔOSC, and ΔΔOSC, are employed and combined for audio classification. Two information fusion techniques, feature level fusion and decision level fusion, are developed to combine the extracted feature vectors. Experiments conducted on the music database employed in the ISMIR2004 Audio Description Contest have shown that the proposed approach can achieve a classification accuracy of 84.23%, which is better than the winner of the ISMIR2004 Music Genre Classification Contest.

5. Reference


<table>
<thead>
<tr>
<th>Filter number</th>
<th>Frequency interval (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>[0, 0]</td>
</tr>
<tr>
<td>1</td>
<td>(0, 100)</td>
</tr>
<tr>
<td>2</td>
<td>(100, 200)</td>
</tr>
<tr>
<td>3</td>
<td>(200, 400)</td>
</tr>
<tr>
<td>4</td>
<td>(400, 800)</td>
</tr>
<tr>
<td>5</td>
<td>(800, 1600)</td>
</tr>
<tr>
<td>6</td>
<td>(1600, 3200)</td>
</tr>
<tr>
<td>7</td>
<td>(3200, 6400)</td>
</tr>
<tr>
<td>8</td>
<td>(6400, 12800)</td>
</tr>
<tr>
<td>9</td>
<td>(12800, 22050]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>CA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC+OSC</td>
<td>79.70</td>
</tr>
<tr>
<td>MFCC+ΔMFCC+OSC+ΔOSC</td>
<td>83.26</td>
</tr>
<tr>
<td>MFCC+ΔMFCC+ΔMFCC+OSC+ΔOSC+ΔΔOSC</td>
<td>82.71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>CA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fusion1:MFCC+OSC</td>
<td>79.70</td>
</tr>
<tr>
<td>fusion2:MFCC+ΔMFCC+OSC+ΔOSC</td>
<td>82.85</td>
</tr>
<tr>
<td>fusion3:MFCC+ΔMFCC+ΔMFCC+OSC+ΔOSC+ΔΔOSC</td>
<td>81.76</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>CA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elias Pampalk (winner)</td>
<td>84.23</td>
</tr>
<tr>
<td>Kris West (2nd rank)</td>
<td>78.33</td>
</tr>
<tr>
<td>George Tzanetakis (3rd rank)</td>
<td>71.33</td>
</tr>
<tr>
<td>Thomas Lidy &amp; Andreas Rauber (4th rank)</td>
<td>70.37</td>
</tr>
<tr>
<td>Dan Ellis &amp; Brian Whitman (5th rank)</td>
<td>64.00</td>
</tr>
</tbody>
</table>
Fig. 1. Confusion matrices for different audio feature set: (a) fusion1:MFCC+OSC (b) fusion2:MFCC+OSC (c) fusion3:MFCC+OSC (d) fusion1:MFCC+ΔMFCC+OSC+ΔOSC (e) fusion2:MFCC+ΔMFCC+OSC+ΔOSC (f) fusion3:MFCC+ΔMFCC+OSC+ΔOSC (g) fusion1:MFCC+ΔΔMFCC+OSC+ΔΔOSC (h) fusion2:MFCC+ΔΔMFCC+OSC+ΔΔOSC (i) fusion3:MFCC+ΔΔMFCC+OSC+ΔΔOSC+ΔΔOSC