Music-Genre Classification System based on Spectro-Temporal Features and Feature Selection

Shin-Cheol Lim, Jong-Seol Lee, Sei-Jin Jang, Soek-Pil Lee, and Moo Young Kim, *Senior Member, IEEE*

Abstract — *An automatic classification system of the music genres is proposed. Based on the timbre features such as mel-frequency cepstral coefficients, the spectro-temporal features are obtained to capture the temporal evolution and variation of the spectral characteristics of the music signal. Mean, variance, minimum, and maximum values of the timbre features are calculated. Modulation spectral flatness, crest, contrast, and valley are estimated for both original spectra and timbre-feature vectors. A support vector machine (SVM) is used as a classifier where an elaborated kernel function is defined. To reduce the computational complexity, an SVM ranker is applied for feature selection. Compared with the best algorithms submitted to the music information retrieval evaluation exchange (MIREX) contests, the proposed method provides higher accuracy at a lower feature dimension for the GTZAN and ISMIR2004 databases*[1].

Index Terms — Music genre classification, music information retrieval, modulation spectrum, feature selection, SVM.

I. Introduction

Digital music is widely available over the Internet and the storage devices. To provide an easy access to these vast music databases, music information retrieval (MIR) is an emerging research field. MIR can be applied to the consumer devices for music recommendation, music search, music summarization, query-by-singing/humming, and so on. Nowadays, text-based search engines with manual labeling have been replaced by automatic music retrieval systems [1], [2]. One of the essential parts in automatic music retrieval is classification of genres, moods, and composers. In this paper, a novel music-genre classification system is proposed based on spectro-temporal features and feature selection.

Features for music-genre classification are divided into top, mid, and low-level ones based on the feature characteristics [3]. Top-level one provides the semantic labels such as genre, mood, and artist labels defined by human. On the other hand, mid and low-level features are extracted from the audio signal. Mid-level features represent the properties of music such as rhythm, pitch, and harmonicity. Low-level features show the timbre structure and time-varying characteristics of the sound source [4]-[12]. Tzanetakis et al. proposed beat histogram, timbre features, and spectral roll-off/flux/centroid [4]. In [8] and [9], spectral flatness/crest features were proposed based on the modulation spectrum. Statistical and neural network approaches were proposed for classification such as Gaussian mixture model, hidden Markov model, multi-layer perceptron, and support vector machine (SVM) [4]-[12].

In [12], we proposed the SVM-based genre classification algorithm. In this paper, we use it as a baseline system and improve the performance in terms of classification accuracy and computational complexity. Based on basic timbre features, spectro-temporal features are calculated such as modulation spectral flatness/crest/contrast/valley features, mean/variance features, and min/max features. To reduce the computational complexity, feature-selection and dimension-reduction algorithms can be used such as principal component analysis (PCA), linear discriminant analysis (LDA), non-negative matrix factorization (NMF), best-first search, and SVM ranker methods [13]-[19]. More elaborated radial basis function (RBF) kernel for the SVM classifier is also designed.

Numerous MIR algorithms and systems are evaluated and compared annually under the controlled conditions from the music information retrieval evaluation exchange (MIREX) contest [20]-[21]. We compare the performance of the proposed genre-classification system with the best algorithms submitted to the MIREX contest.

The rest of this paper is organized as follows. In Section II, more detailed explanation of the proposed music-genre classification system is given. Section III and IV present the experimental results and conclusions, respectively.

II. Proposed Music-Genre Classification System

The proposed music-genre classification system is illustrated in Fig. 1. The training phase consists of timbre-feature extraction from the training dataset, spectro-temporal feature extraction based on the timbre features, generation of the feature-selection models, normalization of the selected features, and SVM genre modeling. The classification phase is composed of timbre and spectro-temporal feature extraction, feature selection based on the selection models, normalization, and classification based on the genre models. In this section, more detailed explanation on timbre features, spectro-temporal features, and feature selection methods is

[1] This work was supported by the Ministry of Knowledge Economy grant funded by the Korea government (No. 10037244).

S.-C. Lim and M. Y. Kim are with the Department of Information and Communication Engineering, Sejong University, Seoul, Korea (e-mail: en.shincheol@gmail.com and mooyoung@sejong.ac.kr).

J.-S. Lee and S.-J. Jang are with the Digital Media Research Center, Korea Electronics Technology Institute, Seongnam, Korea (e-mail: leejs@keti.re.kr and sjjang@keti.re.kr).

S.-P. Lee is with the Department of Digital Media Technology, Sangmyung University, Seoul, Korea (e-mail: esprit@smu.ac.kr).

Contributed Paper
Original manuscript received 08/29/12
Revised manuscript received 09/19/12
Current version published 12/28/12
Electronic version published 12/28/12.

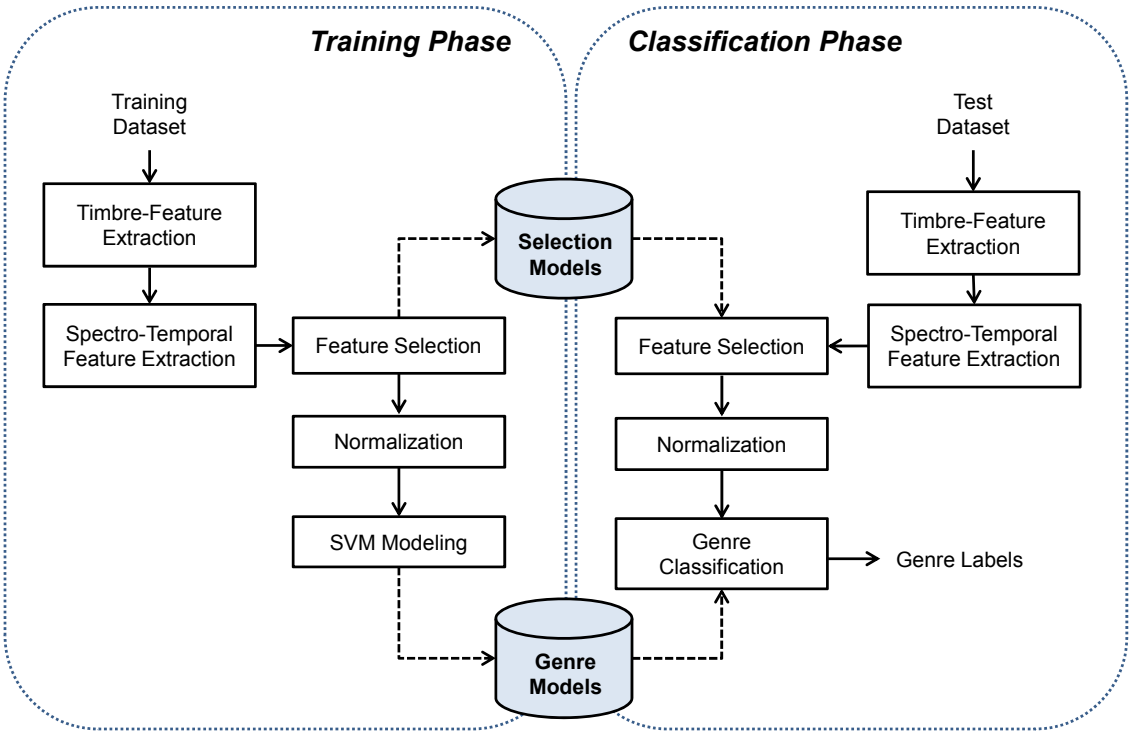

Fig. 1. Blockdiagram of the proposed automatic music-genre classification system.

given. Timbre features represent the spectral characteristics in a given analysis window, whereas spectro-temporal features describe the temporal evolution and variation of the timbre features in a set of analysis windows. For feature selection and dimension reduction, PCA, NMF, and SVM ranker are used. Genre modeling and classification are performed based on SVM.

A. Timbre Feature Set

Based on the short-time Fourier transform, timbre features are calculated for each frame of the music signal. The following timbre features are extracted for the genre classification system.

1) *Mel-frequency cepstral coefficients (MFCC)* are extracted based on the mel-scale band-pass filters to model the human auditory system [22]. Let *S(b)* denote the sum of spectra within a *b*-th mel-scale band ($0 \le b < B$) where B is the total number of bands. A k-th MFCC is computed by

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

where $0 \le k < L$.

2) *Deccorrelated filter bank (DFB)* was firstly proposed for speaker recognition [23]. We applied it for music genre classification in [24]. A k-th DFB is extracted using a high-pass filter as given by

$$DFB(k) = \log S(k+1) - \log S(k) \qquad (2)$$

where $0 \le k < B-1$. DFB considers variation of the amplitudes between neighboring bands.

TABLE I
FEATURE VECTORS OF THE PROPOSED GENRE CLASSIFICATION SYSTEM

Feature vectors				Dimension
Texture window	Mean		MFCC	13
			DFB	13
			OSC	16
	Variance		MFCC	13
			DFB	13
			OSC	16
	Max		MFCC	13
			DFB	13
			OSC	16
	Min		MFCC	13
			DFB	13
			OSC	16
Feature-based modulation spectrum	FMSFM		MFCC	13
			DFB	13
			OSC	16
	FMSCM		MFCC	13
			DFB	13
			OSC	16
	Mean	FMSC/FMSV	MFCC	26
			DFB	26
			OSC	32
	Var	FMSC/FMSV	MFCC	26
			DFB	26
			OSC	32
Octave-based modulation spectrum	MSFM		OBS	8
	MSCM		OBS	8
	Mean	MSC/MSV	OBS	16
	Var	MSC/MSV	OBS	16

3) *Octave-based spectral contrast (OSC)* utilizes the octave-scale band-pass filters to model the music signal [6]. The spectral amplitudes in a k-th octave-scale band, $\{x(k, 1), x(k, 2),$

$\cdots, x(k, N_k)\}$, are sorted as a descending order ($0 \le k < I$) where I is the total number of octave-scale bands, so we obtain the sorted spectral amplitudes, $\{x'(k, 1), x'(k, 2), \cdots, x'(k, N_k)\}$, where N_k is the number of amplitudes in a k-th band. The spectral peak $P(k)$ and valley $V(k)$ for a k-th octave-scale band are calculated by

$$P(k) = \log\left(\frac{1}{\alpha N_k} \sum_{n=0}^{\alpha N_k - 1} x'(k, n+1) \right) \tag{3}$$

$$V(k) = \log\left(\frac{1}{\alpha N_k} \sum_{n=0}^{\alpha N_k - 1} x'(k, N_k - n) \right) \tag{4}$$

where α is a neighborhood factor. Large values of the spectral peak and valley represent harmonic and non-harmonic components of the music signal, respectively. The spectral contrast is computed by

$$SC(k) = P(k) - V(k) \tag{5}$$

and the corresponding OSC is defined as $\{V(k), SC(k)\}$.

B. Spectro-Temporal Feature Set

Instead of directly using the timbre features, we summarize them based on the texture window [4] and the modulation spectrum [7] to generate the spectro-temporal features that represent the temporal evolution and variation of the spectral characteristics of the audio signal. All the feature vectors in the proposed system are shown in Table I.

A set of the consecutive timbre features, such as MFCC, DFB, and OSC, is selected for each texture window. The statistical mean and variance features are extracted by

$$\mu(k) = \frac{1}{P} \sum_{p=0}^{P-1} X(k, p) \tag{6}$$

$$\sigma^2(k) = \frac{1}{P} \sum_{p=0}^{P-1} \left(X(k, p) - \mu(k) \right)^2 \tag{7}$$

where $X(k, p)$ and P are a k-th component in a timbre-feature vector in a p-th frame and the number of timbre-feature vectors within a texture window, respectively. In this paper, additional statistical features, min and max, are calculated by

$$MIN(k) = \min_{0 \le p \le P-1} X(k, p) \tag{8}$$

$$MAX(k) = \max_{0 \le p \le P-1} X(k, p). \tag{9}$$

The min and max features represent the strength of energy in a texture window.

The modulation spectrum based methods include the modulation spectral flatness/crest measures (MSFM/MSCM) [8], [9], modulation spectral contrast/valley (MSC/MSV) [10], [11], and the modulation spectrum itself [25], [26]. Based on the idea in the modulation spectrum based methods, we proposed the feature-based MSFM/MSCM (FMSFM/FMSCM) [13]. For the timbre-feature vectors in a t-th texture window, $X_t(k, p)$, the feature-based modulation spectrum (FMS) is calculated by

$$FMS_t(k, m) = \sum_{p \in P_t} X_t(k, p) \exp(-j 2\pi m p / M) \tag{10}$$

where P_t is the number of timbre-feature vectors in a t-th texture window. The modulation Fourier-transform points are ranged by $0 \le m \le M-1$. Then, the average modulation spectrum for entire texture windows is calculated by

$$\overline{FMS}(k, m) = \frac{1}{T} \sum_{t=0}^{T-1} FMS_t(k, m) \tag{11}$$

where T is the number of texture windows for each music piece. Then, FMSFM and FMSCM are calculated as

$$FMSFM(k) = \frac{\sqrt[M/2]{\prod_{m=0}^{M/2-1} \overline{FMS}(k, m)}}{\frac{2}{M} \sum_{m=0}^{M/2-1} \overline{FMS}(k, m)} \tag{12}$$

and

$$FMSCM(k) = \frac{\max_{m=0,\ldots,M/2-1} \left(\overline{FMS}(k, m) \right)}{\frac{2}{M} \sum_{m=0}^{M/2-1} \overline{FMS}(k, m)} \tag{13}$$

respectively. Small and large values of FMSFM represent the flatness and peakiness of the average modulation spectrum, respectively. FMSCM has the opposite characteristics to FMSFM. If a k-th FMSFM component has a dominantly large value, the input musical piece has a recurrent pattern with a period that corresponds to the k-th modulation frequency. The feature-based MSV/MSC (FMSV/FMSC) are calculated by

$$FMSV(k, q) = \min_{m \in \theta_q} \left(\overline{FMS}(k, m) \right) \tag{14}$$

$$FMSC(k, q) = \max_{m \in \theta_q} \left(\overline{FMS}(k, m) \right) - FMSV(k, q) \tag{15}$$

where θq is a set of modulation-frequency bins in a q-th modulation band. Finally, the mean and variance of FMSV and FMSC in all modulation bands are calculated using (6) and (7).

To calculate the octave-based modulation spectrum (OMS), $X_t(k, p)$ in (10) is substituted by the octave-band sum (OBS) of the spectral components in a k-th octave-scale band and a p-th frame. Then, OMS is calculated into MSFM/MSCM and MSC/MSV as similar ways as (12)/(13) and (14)/(15),

respectively [8]-[10]. We note that FMS and OMS are calculated using timbre features and OBS, respectively.

C. Feature-selection and dimension-reduction algorithms

The feature-selection and dimension-reduction algorithms are applied to increase the recognition performance and decrease the computational complexity [13]-[19].

As a feature-selection method, a SVM ranker is used to evaluate the worth of each feature by using an SVM classifier [19]. Recursive feature elimination is performed by the square root of the weight, which is similar to the Fisher's discriminant criterion, given by the one-against-all SVM. Since eliminating one feature at a time becomes very sub-optimal, the iterative procedure with the pre-computed ranking criterion was proposed.

PCA is called a discrete version of Karhunen–Loève transform. For a set of K-dimensional data $X=[x_1, x_2, \cdots, x_K]^T$, let $\{\lambda_1,\lambda_2,\cdots,\lambda_R,\cdots,\lambda_K\}$ and $[\boldsymbol{w_1},\boldsymbol{w_2},\cdots,\boldsymbol{w_R},\cdots,\boldsymbol{w_K}]$ be the eigenvalues in a descending order and the corresponding orthonormal eigenvectors of $E[XX^T]$. To reduce the K-dimensional data to the R-dimensional space, reduced number of eigenvectors $W=[\boldsymbol{w_1},\boldsymbol{w_2},\cdots,\boldsymbol{w_R}]$ are applied as given by

$$Y = W^T X. \tag{16}$$

NMF is known as a source-separation algorithm for the multivariate data [27]. This algorithm was developed from positive matrix factorization [28]. Let D, K, and R denote the number of songs, the original feature-dimension, and the reduced feature-dimension, respectively. A non-negative matrix $V \in \mathbb{R}^{K\times D}$ is factorized into a basis matrix $W \in \mathbb{R}^{K\times R}$ and a weighting matrix $H \in \mathbb{R}^{R\times D}$ as given by

$$V \approx WH. \tag{17}$$

The distance between V and WH is decreased by an iterative multiplicative rule using a cost function such as Euclidean distance and Kullback-Leibler divergence. For a given W, dimension reduction can be performed as given by

$$\hat{V} = W^T V \tag{18}$$

where $\hat{V} \in \mathbb{R}^{R\times D}$.

III. EXPERIMENTAL RESULTS

We compare the proposed genre-classification system with the other benchmark methods on the GTZAN [4] and ISMIR2004 databases. GTZAN is composed of 10 genres such as blues (bl), classical (cl), country (co), disco (di), hiphop (hi), jazz (ja), metal (me), pop (po), reggae (re), and rock (ro). Each genre consists of 100 songs and each of them has a duration of 30s (22050Hz, 16bits, mono). ISMIR2004 contains 729 songs for training and test, respectively. It consists of 6 genres such as classical, electronic, jazz/blues, metal/punk, rock/pop, and world songs. Each song has a full duration (22050Hz, 16bits, mono).

TABLE II
CLASSIFICATION ACCURACY OF THE BASELINE SYSTEM AND THE PROPOSED GENRE CLASSIFICATION SYSTEMS

Systems	Accuracy (%) GTZAN	Accuracy (%) ISMIR2004	Dim.
Baseline system [12]	81.5	85.3	184
+Max/Min	82.7	85.7	268
+MSC/MSV/FMSC/FMSV	84.0	84.8	468
+Feature selection	85.0	86.3	160
+SVM RBF kernel	87.4	89.9	160

TABLE III
CONFUSION MATRIX OF THE PROPOSED SYSTEM FOR THE GTZAN DATABASE

	bl	cl	co	di	hi	ja	me	po	re	ro
bl	**91**	0	2	1	0	1	1	0	1	3
cl	0	**98**	0	0	0	1	0	0	0	1
co	0	0	**87**	2	0	2	1	5	2	1
di	1	0	2	**87**	3	1	0	2	1	3
hi	1	0	0	1	**85**	0	3	3	7	0
ja	2	3	1	1	0	**91**	0	0	1	1
me	1	0	0	0	1	0	**92**	0	0	6
po	0	0	9	2	3	0	0	**81**	2	3
re	4	1	2	2	6	0	0	2	**81**	2
ro	3	1	3	3	1	0	5	2	1	**81**

TABLE IV
CONFUSION MATRIX OF THE PROPOSED SYSTEM FOR THE ISMIR2004 DATABASE

	classical	electronic	jazz /blues	metal /punk	rock /pop	world
classical	**313**	1	0	0	0	6
electronic	1	**105**	0	0	4	4
jazz/blues	0	0	**23**	0	2	1
metal/punk	0	1	0	**37**	6	1
rock/pop	0	9	1	6	**81**	5
world	6	15	0	0	5	**96**

Timbre feature vectors, such as MFCC, DFB, and OSC, were extracted using a hamming window of around 92ms (2028 samples) with 50% overlap. For both MFCC and DFB, the number of mel-scale band-pass filters and the selected feature dimension were 14 and 13, respectively. For OSC, the following octave-scale band-pass filters were used: 0~100Hz, 100~200Hz, 200~400Hz, 400~800Hz, 800~160Hz, 1600~3200Hz, 3200~6400Hz, and 6400~11025Hz. The neighborhood factor α of OSC was set to be 0.02.

For each texture window of around 3s (65 analysis windows) with 50% overlap, we extracted mean, variance, max, and min features. Then, a statistical feature vector was calculated by averaging the mean, variance, max and min features over the entire texture windows. Modulation feature vectors, such as MSFM, MSCM, MSC/MSV, FMSFM, and FMSCM, were extracted based on the same texture window. FMSFM/FMSCM and FMSC/FMSV were obtained using various timbre feature vectors such as MFCC, DFB, and OSC.

A SVM was used as a classifier. The 10-fold cross validation was used for GTZAN. Training and test procedures for ISMIR2004 were performed based on the corresponding files in the database.

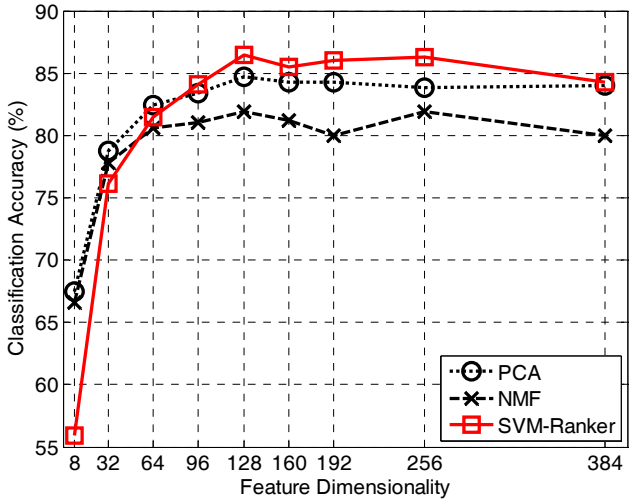

Fig. 2. Variation in Classification Accuracy of the Proposed System with PCA (dotted), NMF (dashed), and SVM ranker (solid). A linear SVM was used for the GTZAN database.

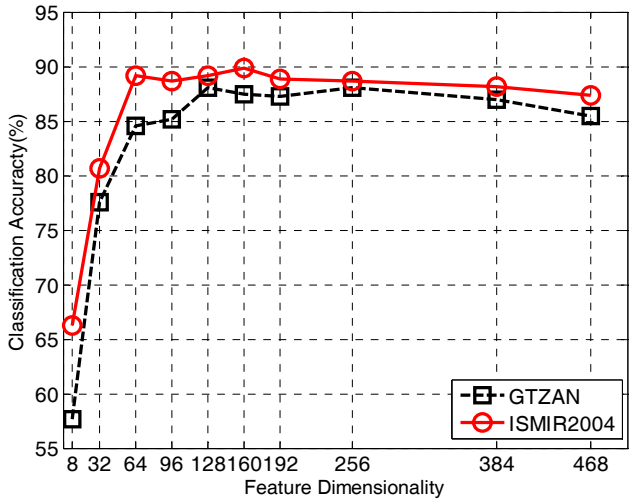

Fig. 3. Variation in Classification Accuracy of the Proposed System with SVM ranker for the GTZAN (dashed) and ISMIR2004 (solid) databases.

TABLE V
CLASSIFICATION ACCURACY OF THE GENRE-CLASSIFICATION ALGORITHMS FOR THE GTZAN AND ISMIR2004 DATASETS

Genre-classification methods	Accuracy (%) GTZAN	ISMIR 2004
Gaussian supervector [29] (MIREX2009: winner)	82.1	79.0
Block-level features [30] (MIREX2010: winner)	85.5	88.2
Gaussian supervector + visual features [31], [32] (MIREX2011: second place)	86.1	86.1
Proposed system	**87.4**	**89.9**

Table II shows the performance improvement in accuracy of the proposed methods compared with the baseline system in [12]. In [12], we proposed to use mean/variance of MFCC/DFB/OSC, MSFM/MSCM, and FMSFM/FMSCM as feature vectors. A linear SVM was used as a classifier. In the proposed system, by adding max/min, MSC/MSV, and FMSC/FMSV features, we obtained the increased accuracy of 84.0% and 84.8% for GTZAN and ISMIR2004, respectively. However, around 2.5 times higher dimensionality in the feature vector was required. To reduce the dimensionality, the SVM ranker was applied, which also gave the increased accuracy. By applying the RBF kernel to the proposed system, we obtained the accuracy improvement by 5.9% and 4.6% for GTZAN and ISMIR2004, respectively, compared with the baseline system. The feature dimensionality of the baseline system was 184, whereas that of the proposed system was 160.

Table III and IV represent the confusion matrices of the proposed genre-classification system for GTZAN and ISMIR2004, respectively. The columns and rows of each matrix represent original genre labels and their classified labels, respectively. In GTZAN, classical music was accurately classified, but pop, reggae, and rock genres yielded much lower accuracy. In ISMIR2004, 13.3% of metal/punk music was misclassified into rock/pop.

The feature-selection and dimension-reduction algorithms were applied to the full feature set. In this experiment, the SVM ranker was used for feature selection, while PCA and NMF were used for dimension reduction.

For dimension reduction, the weight matrix W was calculated after min-max normalization in the training stage. In the test stage, min-max normalization was also performed before applying W in (16) and (18).

In the SVM ranker, min-max normalization was performed to select the best components in the feature vector. However, the final feature set was determined in the original feature-vector domain without min-max normalization. For the SVM modeling, min-max normalization was performed to the selected components in the feature vector. In the test stage, min-max normalization was also applied to the final feature set.

Thus, the above feature-selection approach yielded much lower computational complexity. In the evaluation stage, only 160 dimensional feature vectors were calculated based on the feature-selection approach while 468 dimensional feature vectors were required for the dimension-reduction methods.

Fig. 2 shows the classification accuracy of the feature-selection and dimension-reduction algorithms for GTZAN. In this case, we used the linear SVM as a classifier. For higher dimensionality of the feature vectors, the SVM ranker produced better accuracy than PCA and NMF. The best accuracy was obtained for the SVM ranker with 128 dimensions. It yielded better performance than the full-dimensional feature vectors.

Fig. 3 shows the classification accuracy of the SVM classifier with the RBF kernel and the SVM ranker for GTZAN and ISMIR2004. The feature selection based on the SVM ranker was optimized for GTZAN, but the

performance trend was also similar in ISMIR2004. The best performances for GTZAN and ISMIR were 88.1% (128 and 256 dimensions) and 89.9% (160 dimensions), respectively. For the dimensionality higher than 64, the proposed method gave better accuracy than the baseline system of 184 dimensions [13].

As shown in Table V, the proposed method was compared with the best submissions to the MIREX contests in pop genre classification. In [29]-[32], classification accuracy for GTZAN and ISMIR2004 was reported. Since the accuracy for the winner of MIREX2011 for these databases was not published, the second best algorithm [31], [32] was selected for comparison. The proposed method provided the best accuracy for both GTZAN and ISMIR2004. For example, compared with the winner in MIREX2010, the proposed method gave 1.9% and 1.7% higher performance for GTZAN and ISMIR2004, respectively. They utilized the block-level feature vector such as spectral pattern, logarithmic fluctuation pattern, and correlation pattern. However, the feature-vector dimensionality was very high. In [30], for automatic tag prediction, they used 9448 dimensions as the original feature and around 7500 dimensions after dimension reduction. Thus, the proposed method produced the superior performance to the conventional methods in terms of classification accuracy and computational complexity.

IV. Conclusions

Based on the baseline system in [12], we propose the novel music-genre classification system that includes spectro-temporal features based on timber features, SVM ranker for feature selection, and RBF kernel estimation for SVM classification. Compared with the best algorithms in the MIREX contests, the proposed system provides higher classification accuracy for both GTZAN and ISMIR2004 databases. By utilizing the SVM ranker, the proposed system requires a much lower feature dimension. The required dimensions of the final feature sets for the proposed system and the MIREX2010 submission are 160 and around 7500, respectively. As a future work, we will focus on the design of visual features based on image processing and fusion techniques.

References

[1] X. Zhu, Y.-Y. Shi, H.-G. Kim, and K.-W. Eom, "An integrated music recommendation system," *IEEE Trans. Consumer Electron*., vol. 52, no. 3, pp. 917-925, 2006.

[2] K. Kim, K. R. Park, S.-J. Park, S.-P. Lee, and M. Y. Kim, "Robust query-by-singing/humming system against background noise environments," *IEEE Trans. Consumer Electron*., vol. 57, no. 2, pp. 720-725, 2011.

[3] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 303-319, 2011.

[4] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293-302, 2002.

[5] G. Tzanetakis, "MARSYAS submissions to MIREX 2009," in *Music Information Retrieval Evaluation eXchange* (*MIREX*), 2009.

[6] D. N. Jiang, L. Lu, H. J. Zhang, J. H. Tao, and L. H. Cai, "Music type classification by spectral contrast feature," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2002, pp. 113-116.

[7] S. Sukittanon, L. E. Atlas, and J. W. Pitton, "Modulation-scale analysis for content identification," *IEEE Trans. Signal Process.*, vol. 52, no. 10, pp. 3023-3035, 2004.

[8] D. Jang and C. D. Y, "Music genre classification using novel features and a weighted voting method," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2008, pp. 1377-1380.

[9] D. Jang and C. D. Y, "Music information retrieval using novel features and a weighted voting method," in *Proc. IEEE Int. Symposium on Industrial Electronics*, 2009, pp. 1341-1346.

[10] C-H. Lee, J-L. Shih, K-M. Yu, and J-M Su, "Automatic music genre classification using modulation spectral contrast feature," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2007, pp. 204-207.

[11] C.-H. Lee, J.-L. Shih, K.-M. Yu, and H.-S. Lin, "Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 670-682, 2009.

[12] S.-C. Lim, S.-J. Jang, S.-P. Lee, and M. Y. Kim, "Music genre/mood classification using a feature-based modulation spectrum," in *Proc. IEEE Int. Conf. Mobile IT Convergence*, 2011, pp. 133-136.

[13] C. A. de los Santos, "Nonlinear audio recurrence analysis with application to music genre classification," M.S. thesis, Univ. Pompeu Fabra, 2010.

[14] J. Serra, C. A. de los Santos, and R.G Andrzejak, "Nonlinear audio recurrence analysis with application to genre classification," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2011, pp. 169-172.

[15] B. Rocha, "Genre classification based on predominant melodic pitch contours," M.S. thesis, Univ. Pompeu Fabra, 2011.

[16] J. Salamon, B. Rocha, and E. Gómez, "Musical genre classification using melody features extracted from polyphonic music signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2012, pp. 81-84.

[17] E. Benetos and C. Kotropoulos, "Non-negative tensor factorization applied to music genre classification," *IEEE Audio, Speech, and Language Process.,* vol. 18, no. 8, pp. 1955-1967, 2010.

[18] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. thesis, University of Waikato, 1999.

[19] I. guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, no. 1-3, vol. 46, pp. 389-422, 2002.

[20] J. S. Downie, "The music information retrieval evaluation exchange (2005-2007): a window into music information retrieval research," *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247-255, 2008.

[21] J. S. Downie, A. F. Ehmann, M. Bay, and M. C. Jones, "The music information retrieval evaluation exchange: some observations and insights," *Advances in Music Information Retrieval*, vol. 274, pp. 93-115, 2010.

[22] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[23] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 5, pp. 1711-1723, 2007.

[24] S.-C. Lim, S.-J. Jang, S.-P. Lee and M. Y. Kim, "Music genre classification system using decorrelated filter bank," *The Journal of the Acoustical Society of Korea*, vol. 30, no. 2, pp. 100-106, 2010.

[25] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 576-588, 2010.

[26] I. Panagakis, E. Benetos, and C. Kotropoulos, "Music genre classification: a multilinear approach," in *Proc. Int. Symposium on Music Information Retrieval* (*ISMIR*), 2008, pp. 583-588.

[27] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Process. System*, vol. 13, pp. 556-562, 2001.

[28] P. Paatero, U. Tapper, R. Aalto, and M. Kulmala, "Matrix factorization methods for analysing diffusion battery data," *Journal of Aerosol Science*, vol. 22, pp. 273-276, 1991.
[29] C. Cao and M. Li, "ThinkIT's submission for MIREX 2009 audio music classification and similarity tasks," in *Music Information Retrieval Evaluation eXchange* (*MIREX*), 2009.
[30] K. Seyerlehner, M. Schedl, T. Pohle, and P. Knees, "Using block-level features for genre classification, tag classification and music similarity estimation," in *Music Information Retrieval Evaluation eXchange* (*MIREX*), 2010.
[31] M.-J. Wu and J.-M. Ren, "MIREX 2011 submission – combining visual and acoustic features for music genre classification," in *Music Information Retrieval Evaluation eXchange* (*MIREX*) *Audio Train/Test tasks*, 2011.
[32] M.-J. Wu, Z.-S. Chen, J.-S. R. Jang and Y.-H. Li and C.-H. Lu, "Combining visual and acoustic features for music genre classification," in *Proc. Int. Conf. Machine Learning and Application and Workshop*, 2011, pp. 124-129.

BIOGRAPHIES

Shin-Cheol Lim is a Master Student at the Dept. of Information and Communication Engineering, Sejong University, Seoul, Korea. He received a B.Sc. degree in Information and Communication Engineering from Sejong University, Seoul, Korea, in 2009. His research interests include speech and speaker recognition, speech enhancement, and music information retrieval.

Jong-Seol Lee received BS and MS degrees in Information and Communication Engineering from Chungbuk National University, Cheongju, South Korea, in 1996 and 2001, respectively. He is currently a member of Next-Generation Sound Supporting Center of Korea Electronics Technology Institute. His research interests include Database and music information retrieval.

Sei-Jin Jang received BS and MS degrees in Electronics Engineering from Kyungpook National University, Daegu, South Korea, in 1995 and 1997, respectively. From 1997 to 2002, he worked as a Senior Research Staff at Daewoo Electronics, Seoul, Korea. He is currently a head of Next-Generation Sound Supporting Center of Korea Electronics Technology Institute. His research interests include A/V signal processing and music information retrieval.

Seok-Pil Lee received BS and MS degrees in Electrical Engineering from Yonsei University, Seoul, South Korea, in 1990 and 1992, respectively. In 1997, he earned a PhD degree in Electrical and Electronics Engineering also at Yonsei University. From 1997 to 2002, he worked as a Senior Research Staff at Daewoo Electronics, Seoul, Korea. From 2002 to 2012, he worked as a head of Digital Media Research Center of Korea Electronics Technology Institute. He is currently an Associate Professor at the Dept. of Digital Media Technology, Sangmyung University. His research interests include A/V signal processing and the convergence of digital broadcast and telecommunication.

Moo Young Kim (M'96, SM'10) is an Associate Professor at the Dept. of Information and Communication Engineering, Sejong University, Seoul, Korea. He received an M.Sc. degree in electrical engineering from Yonsei University, Seoul, Korea, in 1995, and a Ph.D. degree in electrical engineering from KTH (the Royal Institute of Technology), Stockholm, Sweden, in 2004, respectively. From 1995 to 2000, he worked as a Member of Research Staff of the Human Computer Interaction Laboratory at Samsung Advanced Institute of Technology, Kiheung, Korea. From 2005 to 2006, he was a Senior Research Engineer of the Dept. Multimedia Technologies at Ericsson Research, Stockholm, Sweden. His research interests include speech and audio coding based on information theory, biometrics including speaker recognition, speech enhancement, joint source and channel coding, and music information retrieval.