A New Feature for Musical Genre Classification of MPEG-4 TwinVQ Audio Data

Michihiro KOBAYAKAWA ♥ Takaya MORITA ♦
Mamoru HOSHI ♠ Tadashi OHMORI ♣

This paper proposes a new musical feature to classify MPEG-4 TwinVQ compressed data into musical genre without decoding to audio signals. To extract the musical feature, we use the LSP (Line Spectrum Pair) parameters directly extracted from a bitstream without any computation, and the Discrete Wavelet Transform (DWT). We experimented on 2,196 **compressed music data collected from** 10 **musical genres and evaluated the performance of the musical feature for musical genre classification. The maximum of average correct ratio for musical genre classification was** 81.7%**. Experiment showed that the musical feature had very good performance for musical genre classification in the compressed domain of MPEG-4 TwinVQ audio compression.**

1. Introduction

Audio compression techniques are widely used. We can obtain compressed music data through music distribution systems that contain hundreds of thousands of the compressed music data. As compressed music data is generated, stored and distributed continuously, methods for effectively managing compressed music data are required.

In addition to a compression function, functions for describing the contents of compressed music data are required in order to effectively use the compressed data. Compression techniques, such as MPEG-7 or MPEG-21, are required to have these functions.

To retrieve a piece of music, conventional music retrieval systems usually use keywords such as the title, artist, musical genre, etc. It is difficult to describe musical genre of a piece of music. Several researchers proposed a content-based musical genre classification from audio signal [4, 9, 13, 19]. We propose a new music feature for content-based musical genre classification in the compressed domain of MPEG-4 TwinVQ audio [11].

The rest of this paper is organized as follows. Section 2 shows the related works of musical genre classification and music information retrieval in the compressed domain. Section 3 briefly introduces MPEG-4 TwinVQ audio compression. Section 4 proposes a musical feature using LSP parameter directly extracted from a bitstream of the TwinVQ audio data and the Discrete Wavelet Transform (DWT). Section 5 describes experiments on 2,196 compressed data and discusses the performance of the musical feature for musical genre classification. Finally, Section 6 concludes this paper.

2. Related Work

We briefly describe previous works on signal-based musical genre classification and on music information retrieval in the compressed domain MPEG audio standard.

The following basic features are often used for analyzing audio contents:

- energy [7, 12, 20],
- normalized subband energy [12],
- (F_s, F_t)-subband energy between F_s-th subband and F_t-th subband,
- spectral centroid [12, 20],
- root mean squared subband vector, mean subband value, rolloff point, low energy, spectral flux [20],
- cepstrum [15],
- Mel Frequency Cepstrum Coefficients (MFCC) [18].

They are based on the subband values or calculated in the polyphase filterbank analysis.

First, we show several works on musical genre classification in the signal domain.

Tzanetakis *et.al* extracted nine features as the "musical surface features"and rhythm features. The musical surface features are based on the centroid, rolloff, flux, zerocrossing and low energy, and the rhythm features are based on the Discrete Wavelet Transform [18].

Li *et.al* proposed the Daubechie Wavelet Coefficient Histograms for musical genre classification and evaluated the performance using several classifiers [4].

Tzanetakis [17], Mandel *et.al* [8], Lidy *et.al* [6], Guaus *et.al* [1] joined in MIREX2007 Evaluation Task for audio

♥ Non-member Graduate School of Information Systems, the University of Electro-Communications kobayakawa@computer.org, kobayakawa@acm.org
♦ Non-member Graduate School of Information Systems, the University of Electro-Communications morita@hol.is.uec.ac.jp
♠ Non-member Graduate School of Information Systems, the University of Electro-Communications hoshi@is.uec.ac.jp
♣ Member Graduate School of Information Systems, the University of Electro-Communications omori@is.uec.ac.jp

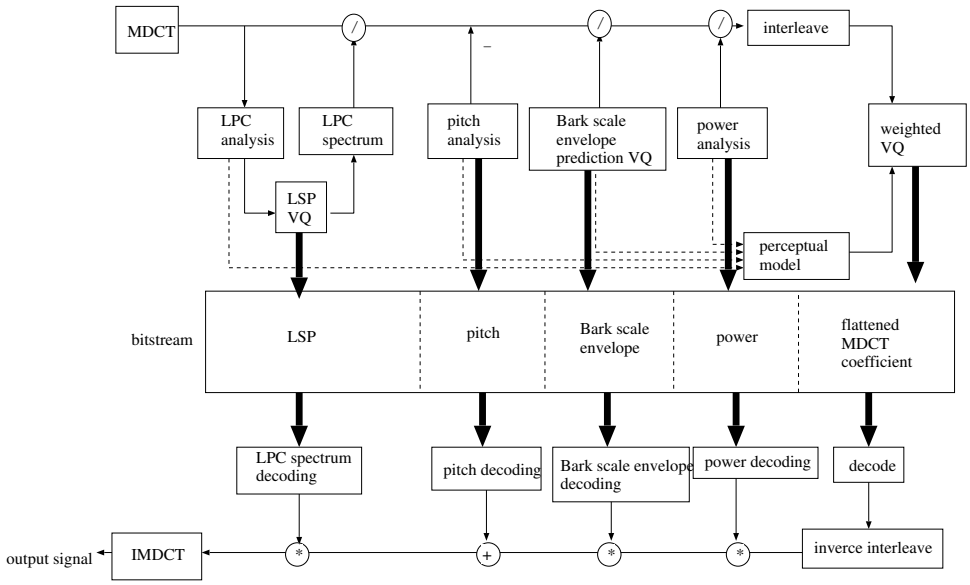

Figure. 1 Architecture of MPEG-4 TwinVQ audio compression.

genre classification [9].

Tzanetakis used the spectral centroid, rolloff, flux, MFCC as the musical features, and used a linear support vector machine as a classifier. Mandel *et.al* used the spectral features based on MFCCs and used support vector machines. Lidy *et.al* extracted rhythm pattern and symbolic feature from audio signals and used the support vector machines. Guaus *et.al* used a set of timbre descriptors and rhythm descriptors , and used support vector machines.

The summary results of MIREX 2007 Evaluation Task for audio genre classification are shown at "http://www.music-ir.org/mirex2007/index.php/Audio_Genre_Classification_Results".

Next, we show works on music information retrieval in the compressed domain.

Liu *et al.* proposed a method for classifying the encoded data to identify a singer. They used the energy of a frame for classification [7].

Pye proposed a method for managing music data [15]. They classified the MP3 music data into several genres and labeled them artist name using cepstrum.

Nakajima *et al.* proposed a method for classifying the encoded audio data [12]. They discriminated between short pieces of silence and pieces of non-silence by using energy, and then classified pieces of non-silence into music, speech, and applause by using the normalized subband energy and center frequency of subband energy on *1 second* of MPEG audio data.

Tzanetakis *et al.* proposed a method for segmenting the encoded data into short pieces and classifying them into music and speech [20].

Table. 1 Genre and the number of compressed data.

genre	# of compressed data
baroque	223
bossanova	119
dance	176
hiphop	196
jazz	431
march	91
oldies	435
rock	200
tango	220
waltz	99
total	2,196

3. TwinVQ audio compression

This section briefly describes the architecture of MPEG-4 TwinVQ audio compression. The details of TwinVQ audio compression are referred to [11] or the URL http://www.twinvq.org/. A sequence of original signals of a piece of music is divided into frames. The MDCT (Modified Discrete Cosine Transform) is then applied to the signals of each frame. For the MDCT coefficients of each frame, five analyses, namely, LPC analysis, Pitch analysis, Barkscale envelope analysis, Power analysis and Weighted VQ analysis, are carried out in this order (Figure 1). In LPC analysis, the LSP (Line Spectrum Pair) parameters are stored in the bitstream. They are used for describing the power envelope of MDCT coefficients of each frame.

Table. 2 Average correct ratio of the (D, L) musical feature vector $\boldsymbol{f}_{D,L}$ for the test data (%).

	L									
D	1	2	3	4	5	6	7	8	9	10
1	33.6	48.6	50.7	52.3	53.9	54.4	54.5	58.6	58.8	59.1
2	55.4	65.0	67.5	67.8	69.0	68.5	69.2	71.9	72.2	71.7
3	59.2	67.8	69.7	70.4	71.7	71.1	71.6	75.4	75.1	75.1
4	62.4	70.2	71.8	72.6	73.9	74.0	74.9	77.5	77.4	77.7
5	63.7	70.4	73.2	73.6	74.6	74.7	76.1	78.4	78.2	78.0
6	65.8	71.9	74.4	74.9	76.4	77.3	77.9	80.9	80.4	80.5
7	67.1	73.6	76.3	76.0	77.9	78.0	78.6	80.7	81.1	81.5
8	67.8	74.0	76.4	76.7	78.1	78.6	78.9	81.2	81.6	81.0
9	67.9	75.2	77.3	77.7	78.7	78.7	79.5	81.7	81.5	81.5
10	69.3	75.5	77.4	77.6	78.7	78.6	79.5	81.5	81.6	81.5
11	69.4	76.0	77.7	77.7	78.5	78.5	79.0	81.0	81.2	80.8
12	69.7	76.6	78.0	77.7	78.3	78.4	78.8	80.7	80.8	80.5
13	70.4	76.5	77.8	78.1	78.4	79.3	79.3	80.9	80.9	81.2
14	70.9	76.8	78.2	78.3	79.0	79.4	79.0	80.6	80.5	80.0
15	71.1	77.2	78.2	78.2	79.1	79.2	79.0	80.1	80.4	79.6
16	71.2	77.0	78.7	78.3	79.3	79.0	78.3	80.2	80.1	78.9
17	71.6	77.2	79.0	78.5	79.4	78.6	78.1	80.0	79.8	78.2
18	71.9	77.7	79.2	78.8	79.5	78.8	78.5	80.5	79.7	78.5
19	72.4	78.4	79.4	79.3	79.4	79.7	78.5	80.2	79.7	78.8
20	72.1	78.4	79.4	79.0	79.4	79.1	78.2	80.2	79.4	78.4

4. Musical Feature

This section describes a musical feature for musical genre classification of MPEG-4 TwinVQ audio data. The musical feature is based on the sequence of LSP parameters obtained from the bitstream and on multi-resolution analysis by DWT.

We define the LSP parameter vector of the m-th frame by

$$\boldsymbol{\omega}_m^{LSP} = (\omega_{m,1}^{LSP}, \ldots, \omega_{m,d}^{LSP}, \ldots, \omega_{m,20}^{LSP})^T, \quad (1)$$

where $\omega_{m,d}^{LSP}$ is the d-th LSP parameter of the m-th frame, and T denotes the transposition. The sequence $\boldsymbol{\Omega}^{LSP}$ consisting of N LSP parameter vectors is represented by $20 \times N$ matrix

$$\boldsymbol{\Omega}^{LSP} = (\boldsymbol{\omega}_1^{LSP}, \boldsymbol{\omega}_2^{LSP}, \ldots, \boldsymbol{\omega}_N^{LSP}), \quad (2)$$

where N is the number of frames of TwinVQ audio data. The d-th row of the matrix $\boldsymbol{\Omega}^{LSP}$ is expressed as

$$\boldsymbol{\omega}_d^{LSP} = \omega_{1,d}^{LSP}, \ldots, \omega_{N,d}^{LSP}, \quad (3)$$

and is called "the d-th sequence of length N".

Then, by using DWT with the base Haar, we decompose the d-th sequence $\boldsymbol{\omega}_d$ and then obtain the detailed wavelet coefficients $c_{d,l,t}$ $(l = 1, \ldots, L;\ t = 1, \ldots, \lfloor \frac{N}{2^l} \rfloor)$ up to level L.

We calculate the mean spectrum $p_{d,l}$ of level l $(l = 1, \ldots, L)$ defined by the average of the power of the detailed wavelet coefficients $c_{d,l,t}$ of level l,

$$p_{d,l} = \frac{\sum \{c_{d,l,t}\}^2}{\frac{N}{2^l}},$$

and the standard deviation $\sigma_{d,l}$

$$\sigma_{d,l} = \sqrt{\frac{1}{\frac{N}{2^l}} \sum_{i=1}^{\frac{N}{2^l}} (c_{d,l,i}^2 - p_{d,l})^2}.$$

We then define the spectrum vector $\boldsymbol{p}_{d,L}$ by the spectrum up to level L as

$$\boldsymbol{p}_{d,L} \equiv (p_{d,1}, p_{d,2}, \ldots, p_{d,L}), \quad d = 1, \ldots, 20,$$

and the standard deviation vector up to level L as

$$\boldsymbol{\sigma}_{d,L} \equiv (\sigma_{d,1}, \sigma_{d,2}, \ldots, \sigma_{d,L}).$$

We define the vector $\boldsymbol{f}_{D,L}$ of the compressed music data up to order D and wavelet decomposition level L as

$$\boldsymbol{f}_{D,L} \equiv (\boldsymbol{p}_{1,L}, , \ldots, \boldsymbol{p}_{D,L}, \boldsymbol{\sigma}_{1,L}, \ldots, \boldsymbol{\sigma}_{D,L}),$$

$$1 \leq D \leq 20, \quad 1 \leq L \leq 10.$$

and call it "the (D, L) musical feature vector".

5. Experiments

This section evaluates the performance of the musical feature for genre classification.

We can directly obtain the LSP parameter vector (20 dimensional) of each frame. The time of extracting the LSP parameter vector was only $0.9\mu sec$ on average (variance 5.4×10^{-23}, minimum $0.1\mu sec$, maximum $1.4\mu sec$) (OS: Linux, CPU: AMD Athron 64 X2 Dual Core Processor 4600+, Memory: 3GB). For example, the computation time of length of $10,000$ frames (play time is 232.2 *sec*) is only $90msec$ on average.

5.1 Data set and Classifier

Data set

We experimented on $2,196$ compressed music data of 10 musical genres such as "baroque", "bossanova", "dance", "hiphop", "jazz", "march", "oldies", "rock", "tango" and "waltz".

We collected the pieces of music from the collected works of "baroque", "bossanova", "dance", "oldies" and "tango", and from some albums of march and waltz. In addition, we selected the pieces of music of hiphop and rock from our library consisting of $20,000$ pieces of music. Table 1 shows the number of pieces of music of each musical genre.

Classifier

We used the Discriminant Analysis (DA) as a classifier (Figure 2) to evaluate the performance of the (D, L) musical feature vector. The DA is a simple and widely used standard statical classifier.

5.2 Musical Genre Classification

This section evaluates the classification performance of the proposed musical feature.

We applied the MPEG-4 TwinVQ encoder to each signal of piece of music and obtained $2,196$ compressed music data, then compute the musical feature vector for each compressed music data.

The musical genre classification was done as follows. Firstly, as training data we randomly selected 50 percent of the compressed music data of the same genre and used the rest as the test data (50%).

The system executed discriminant analysis to the set of the training data, and obtained the discriminant space and the centroid vector for each musical genre.

To classify the test data, the system maps the musical feature vector of the test data into the discriminant space and then compute the distance between the test musical feature vector and centroid points, and returns the genre of the nearest centroid as the musical genre of the test data.

For each set of the (D, L) musical features, the process above was repeated three times and the average accuracy of classification was obtained. We show the performance of the (D, L) musical feature in the form of a confusion matrix on percentage (*e.g.*, Table 3 for the $(9, 8)$ musical feature vector). The rows correspond to the actual genre and the columns to classified genre. For example, the cell of row "dance", column "hiphop" with value 2.27 means that 2.27% of the dance music was wrongly classified as hiphop music.

The average of diagonal elements of the confusion matrix for the (D, L) musical feature shows the total performance of the (D, L) musical feature for musical genre classification. The average is called the "average correct ratio".

We examine how the number D of LSP parameters and the wavelet decomposition level L affect on the performance of musical genre classification.

Table 2 shows the average correct ratio of $\boldsymbol{f}_{D,L}$ for $D = 1 \sim 20$, $L = 1 \sim 10$.

The average correct ratios for $6 \leq D \leq 14$ and $8 \leq L \leq 10$ were greater than or equal to 80.0%. The maximum ratio was 81.7% (for $\boldsymbol{f}_{9,8}$).

Hereafter, we examine the performance of musical genre classification using the $(9, 8)$ musical feature vector $\boldsymbol{f}_{9,8}$. The results of musical genre classification using the vectors $\boldsymbol{f}_{9,8}$ are shown in Table 3 (confusion matrix). All diagonal elements in Table 3 were greater than or equal to 70.0, except waltz (67.35%). The correct ratios for five genres: baroque, hiphpo, jazz, march and rock were greater than or equal to 86.36. Waltz music had the minimum (worst) correct ratio 67.35. Waltz music was misclassified as baroque (12.24%), and jazz, march and tango (6.12% each). The misclassifications are similar to what a human would do.

The results of experiments showed that the musical genre classification by using the (9, 8) musical feature vector of LSP parameter had very good performance.

6. Conclusion

This paper proposed a new musical feature for musical genre classification of MPEG-4 TwinVQ compressed music data. The key idea was to use the LSP parameters directly extracted from bitstream in the MPEG-4 TwinVQ audio data without any computation. We applied the Discrete Wavelet Transform to a sequence of the LSP parameter and obtained the (D, L) musical feature vector. To evaluate the performance of the musical feature, we experimented musical genre classification on $2,196$ compressed music data of 10 musical genres. We got the maximum average correct ratio 81.7% for the $(9, 8)$ music feature vector based on the LSP parameter. Although the used classifier (DA) was very simple, the proposed music feature had very good performance for musical genre classification.

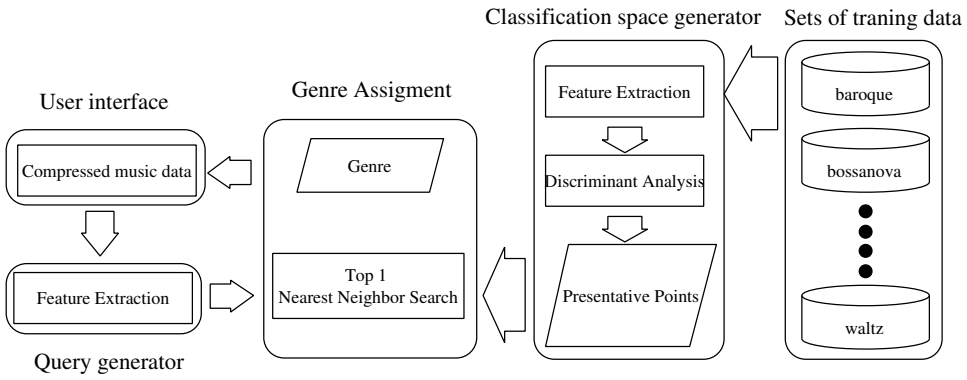

Figure. 2 Musical genre classification using the discriminant analysis.

Table. 3 The accuracy of genre classification using the $(9, 8)$ musical feature vector $\boldsymbol{f}_{9,8}$ for the test data (%).

	classified as									
genre	baroque	bossa	dance	hiphop	jazz	march	oldies	rock	tango	waltz
baroque	86.36	0.00	0.00	0.00	0.00	6.36	1.82	0.00	1.82	3.64
bossa	0.00	71.19	1.69	1.69	11.86	1.69	10.17	0.00	1.69	0.00
dance	4.55	4.55	79.55	2.27	3.41	0.00	3.41	2.27	0.00	0.00
hiphop	0.00	2.00	5.00	91.00	1.00	0.00	1.00	0.00	0.00	0.00
jazz	0.93	3.72	0.47	0.00	85.58	0.93	2.79	0.47	4.65	0.47
march	2.22	0.00	0.00	0.00	0.00	88.89	0.00	0.00	2.22	6.67
oldies	1.32	1.32	3.08	0.44	7.05	0.88	78.41	0.88	5.73	0.88
rock	2.00	2.00	4.00	0.00	0.00	0.00	2.00	90.00	0.00	0.00
tango	3.00	0.00	0.00	0.00	10.00	1.00	8.00	0.00	78.00	0.00
waltz	12.24	0.00	0.00	0.00	6.12	6.12	2.04	0.00	6.12	67.35

We can obtain the approximate values of the LPC coefficients and the LPC cepstrum from the LSP parameters with some computations, and compute the musical feature vectors using the approximate the LPC coefficients and the LPC cepstrum in the same way as described in section 4. The performance of them are found in [10].

In this paper, we focused on the musical feature using the LSP parameters, because they can be directly extracted from the bitstream without any computation and are fundamental information of the MPEG-4 TwinVQ compressed data.

The contribution of this paper is to present a new musical feature to classify MPEG-4 TwinVQ compressed data into musical genre without decoding to audio signals.

[References]

[1] E. Guaus and P. Herrera. A Basic System For Music Genre Classification. *http://www.music-ir.org/mirex/2007/abs/GC_guaus.pdf*, 2007.

[2] J.-S. R. Jang and H.-R. Lee. Hierarchical Filtering Method for Content-based Music Retrieval via Acoustic Input. In *Proceedings of the Ninth ACM International Conference on Multimedia (MM2001)*, pages 401–410, 2001.

[3] N. Kosugi, Y. Nishihara, T. Saketa, M. Yamamuro, and K. Kushima. A Practical Query-By-Humming System for a Large Music Database. In *Proceedings of the Eighth ACM International Conference on Multimedia (MM2000)*, pages 333–342, 2000.

[4] T. Li, M. Ogihara and Q. Li. A comparative study on content-based music genre classification. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR2003)*, pages 282–289, 2003.

[5] T. Lidy and A. Runber. Evaluation of Feature Extractions and Psycho-Acoustic Transformations for Music Genre Classification. In *Proceedings of International Symposium on Music Information Retrieval (ISMIR2005)*, pages 34–41, 2005.

[6] T. Lidy, A. Runber, A. Pertusa, and J. M. Inesta. Combining Audio and Symbolic Descriptors for Music Classification from Audio. *http://www.music-ir.org/mirex/2007/abs/AI_CC_GC_MC_AS_lidy.pdf*, 2007.

[7] C.-C. Liu and P.-H. Tsai. Content-based Retrieval of MP3 Music Objects. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM2002)*, pages 506–511, 2002.

[8] M. Mandel and D. P. W. Ellis. LABROSA'S Audio Music Similarity and Classification Submissions. *http://www.music-ir.org/mirex/2007/abs/AI_CC_GC_MC_AS_mandel.pdf*, 2007.

[9] MIREX2007. http://www.music-ir.org/mirex/2007/index.php/Audio_Genre_Classification.

[10] T. Morita, M. Kobayakawa, M. Hoshi, and T. Ohmori. Automatic Musical Genre Classification using TwinVQ. In *Proceedings of the Nineteenth Data Engineering WorkShop 2008 (DEWS2008)*, E8-1, 2008 (in Japanese).

[11] T. Moriya, N. Iwakami, K. Ikeda, and S. Miki. Extension and Complexity of TwinVQ Audio Coder. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP1996)*, pages 1029–1032, 1996.

[12] Y. Nakajima, Y. Lu, M. Suguro, A. Yoneyama, H. Yanagihara, and A. Kurematsu. A Fast Audio Classification From MPEG Coded Data. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP1999)*, pages 3005–3008, 1999.

[13] N. M. Norowi , S. Doraisamy and R. Wirza. Factors Affecting Automatic Genre classification: An investigation Incorporating Non-Western Musical Forms. In *Proceedings of International Symposium on Music Information Retrieval (ISMIR2005)*, pages 13–20, 2005.

[14] S. Pfeiffer, S. Fischer, and W. Effelsberg. Automatic Audio Content Analysis. In *Proceedings of the Fourth ACM International Conference on Multimedia*, pages 21–30, 1996.

[15] D. Pye. Content-based Methods for Management of Digital Music. In *Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing (ICASSP2000)*, pages 2437–2440, 2000.

[16] J. Shifrin, B. Pardo, C. Meek, and W. Birmingham. HMM-Based Musical Query Retrieval. In *Proceedings of Joint Conference on Digital Libraries (JCDL2002)*, pages 295–300, 2002.

[17] G. Tzanetakis. MARSYS Submissions to MIREX207. *http://www.music-ir.org/mirex/2007/abs/AI_CC_GC_MC_AS_tzanetakis*, 2007.

[18] G. Tzanetakis, G. Essl and P. Cook. Automatic Musical Genre Classification of Audio Signals. In *Proceedings of International Symposium on Music Information Retrieval (ISMIR2001)*, http://ismir2001.ismir.net/pdf/tzanetakis.pdf, 2001.

[19] G. Tzanetakis and P. Cook. Music genre classification of audio signals. In *IEEE Transaction on Speech and Audio Processing*, Vol. 10, No. 5, pages 293–302, 2002.

[20] G. Tzanetakis and P. Cook. Sound Analysis using MPEG Compressed Audio. In *Proceedings of IEEE International Conference Acoustic, Speech, and Signal Processing (ICASSP2000)*, pages 761–764, 2000.

Michihiro KOBAYAKAWA

He received Dr. E. degree from the University of Electro-Communications in 2001. From 2001 to 2008, he was on Graduate School of Information Systems, The University of Electro-Communications. He is an Associate Professor of Okinawa National College of Technology, Okinawa, Japan. His interests include a content-based multimedia retrieval, and algorithms and data structures for multimedia data retrieval. He is a member of ACM, IEEE CS and IEICE.

Takaya MORITA

He received the degree of Master of Engineering from the University of Electro-Communications in 2008. He works for Hitachi Corporation Ltd. His research interests include content-based music information retrieval and music retrieval.

Mamoru HOSHI

He received Dr. E degree in mathematical engineering from the University of Tokyo in 1985. He is a Professor of Graduate School of Information Systems, The University of Electro-Communications. His research interests include algorithms and data structures for searching, multimedia retrieval, and random number generation. He is a member of ACM, IEEE CS, IEEE IT, IEICE, IPSJ, and SITA.

Tadashi OHMORI

He is an Associate Professor in Graduate School of Information Systems, The University of Electro-Communications. He received his Dr. Eng. degree from The University of Tokyo, 1990. His research interests include parallel database machines, transaction processing, and Web mining.