QUERY-BY-EXAMPLE MUSIC RETRIEVAL APPROACH BASED ON MUSICAL GENRE SHIFT BY CHANGING INSTRUMENT VOLUME

Katsutoshi Itoyama†, Masataka Goto‡, Kazunori Komatani†, Tetsuya Ogata†, Hiroshi G. Okuno†*

† Graduate School of Informatics, Kyoto University * JSPS Research Fellow (DC1)
‡ National Institute of Advanced Industrial Science and Technology (AIST)
itoyama [at] kuis.kyoto-u.ac.jp

ABSTRACT

We describe a novel Query-by-Example (QBE) approach in Music Information Retrieval, which allows a user to customize query examples by directly modifying the volume of different instrument parts. The underlying hypothesis is that the musical genre *shifts* (changes) in relation to the volume balance of different instruments. On the basis of this hypothesis, we aim to clarify the relationship between the change of the volume balance of a query and the shift in the musical genre of retrieved similar pieces, and thus help instruct a user in generating alternative queries without choosing other pieces. Our QBE system first separates all instrument parts from the audio signal of a piece with the help of its musical score, and then lets a user remix those parts to change acoustic features that represent musical mood of the piece. The distribution of those features is modeled by the Gaussian Mixture Model for each musical piece, and the Earth Movers Distance between mixtures of different pieces is used as the degree of their mood similarity. Experimental results showed that the shift was actually caused by the volume change of vocal, guitar, and drums.

1. INTRODUCTION

One of promising approaches of Music Information Retrieval is the Query-by-Example (QBE) retrieval [1, 2, 3, 4, 5, 6, 7] where a user can receive a list of musical pieces ranked by their similarity to a musical piece (example) that the user gives as a query. Although this approach is powerful and useful, the user has to prepare or find examples of favorite pieces and it was sometimes difficult to control or change the retrieved pieces after seeing them because another appropriate example should be found and given to get better results. For example, even if a user feels that vocal or drum sounds are too strong in the retrieved pieces, it was difficult to find another piece that has weaker vocal or drum sounds while keeping the basic mood and timbre of the first piece. Because finding such music pieces is a matter of trial and error, we need more direct or convenient methods for QBE.

We solve this inefficiency by allowing a user to create new query examples for QBE by remixing existing musical pieces, i.e., changing the volume balance of the instruments. To obtain the desired retrieved results, the user can easily give alternative queries where the volume balance is changed from the original balance. For example, the above problem can be solved by customizing a query example so that the volume of the vocal or drum sounds are decreased. To remix an existing musical piece, we use an original sound source separation method that decomposes the audio signal of a musical piece into different instrument parts on the basis of its available musical score. To measure the similarity between the remixed query and each piece in a database, we use the Earth Movers Distance (EMD) between their Gaussian Mixture Models (GMMs). The GMM for each piece is obtained by modeling the distribution of original acoustic features consisting of intensity and timbre features.

The underlying hypothesis is that changing the volume balance of different instrument parts in a query causes the *musical genre shift* in the retrieved pieces — i.e., a user can change the genre of the retrieved pieces just by changing the volume balance of the query[1]. An existing study [8] has already suggested this hypothesis. Based on this hypothesis, our research focuses on clarifying the relationship between the volume change of different instrument parts and the shift in the musical genre in order to instruct a user in easily generating alternative queries. To clarify this relationship, we conducted two different experiments. The first experiment examined how much change of the volume of a single instrument part is needed to cause the musical genre shift through our QBE retrieval system. The second experiment examined how the volume change of two instrument parts (a two-instrument combination for the volume change) affects the shift in the genre. This relationship is explored by examining the genre distribution of the retrieved pieces. Experimental results showed that the desired musical genre shift in QBE results was easily caused just by changing the volume balance of different instruments in the query.

[1] We target genres which are mostly defined by organization and volume balance of musical instruments, such as classical music, jazz and rock. Although several genres are defined by specific rhythm-patterns and singing-style, e.g., a waltz and a hip-hop, we except them.

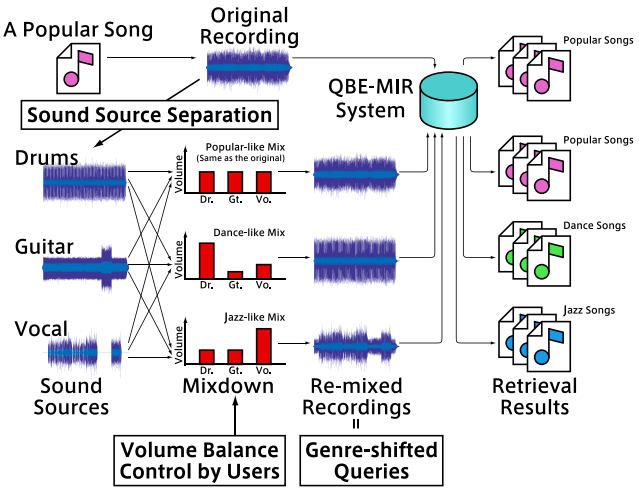


Figure 1: Overview of our QBE retrieval system based on genre shift. Controlling the volume balance causes the genre shift of a query song and our system returns songs that are similar to the genre-shifted query.

2. QUERY-BY-EXAMPLE RETRIEVAL BY GENRE-SHIFTED MUSICAL PIECES

In this section, we describe our QBE retrieval system that retrieves musical pieces based on the similarity of mood between musical pieces.

2.1. Musical Genre Shift

Our original term *"musical genre shift"* means the change of the musical genre of pieces on auditory features which is caused by changing the volume balance of musical instruments. For example, by boosting the vocal and reducing the guitar and drums of a popular song, auditory features extracted from the modified song are similar to the features of a jazz song. The instrumentation and volume balance of musical instruments are important in judging the musical genre. In fact, some musical genres are defined by the instrumentation. As shown in Figure 1, by automatically separating the original recording (audio signal) of a piece into musical instrument parts, a user can change their volume balance to cause the genre shift.

2.2. Acoustic Feature Extraction

Acoustic features that represent the musical mood are designed as shown in Table 1 upon existing studies of mood extraction [9]. These features extracted from the power spectrogram, $X(t,f)$, for each frame (100 frames per second). The spectrogram is calculated by short-time Fourier transform of the monauralized input audio signal, where t and f are the frame and frequency indices, respectively.

Table 1: Acoustic features representing musical mood.

Acoustic intensity features		
Dim.	Symbol	Description
1	$S_1(t)$	Overall intensity
2–8	$S_2(i,t)$	Intensity of each subband*
Acoustic timbre features		
Dim.	Symbol	Description
9	$S_3(t)$	Spectral centroid
10	$S_4(t)$	Spectral width
11	$S_5(t)$	Spectral rolloff
12	$S_6(t)$	Spectral flux
13–19	$S_7(i,t)$	Spectral peak of each subband*
20–26	$S_8(i,t)$	Spectral valley of each subband*
27–33	$S_9(i,t)$	Spectral contrast of each subband*

* 7-bands octave filterbank.

2.2.1. Acoustic Intensity Features

Overall intensity for each frame, $S_1(t)$, and intensity of each subband, $S_2(i,t)$, are defined as

$$S_1(t) = \sum_{f=1}^{F_N} X(t,f) \text{ and } S_2(i,t) = \sum_{f=F_L(i)}^{F_H(i)} X(t,f),$$

where F_N is the number of frequency bins of power spectrogram, $F_L(i)$ and $F_H(i)$ are the indices of lower and upper bounds for the i-th subband, respectively. The intensity of each subband helps to represent acoustic brightness. We use octave filterbanks that divide the power spectrogram into n octave subbands:

$$\left[1, \frac{F_N}{2^{n-1}}\right), \left[\frac{F_N}{2^{n-1}}, \frac{F_N}{2^{n-2}}\right), \ldots, \left[\frac{F_N}{2}, F_N\right],$$

where n is the number of subbands and set to 7 in our experiments.

2.2.2. Acoustic Timbre Features

Acoustic timbre features consist of spectral shape features and spectral contrast features, which are known to be effective in detecting musical moods [10, 9]. The spectral shape features are represented by spectral centroid $S_3(t)$, spectral width $S_4(t)$, spectral rolloff $S_5(t)$, and spectral flux $S_6(t)$ as follows:

$$S_3(t) = \frac{\sum_{f=1}^{F_N} X(t,f) f}{S_1(t)},$$

$$S_4(t) = \frac{\sum_{f=1}^{F_N} X(t,f)(f - S_3(t))^2}{S_1(t)},$$

$$\sum_{f=1}^{S_5(t)} X(t,f) = 0.95 S_1(t) \quad \text{and}$$

$$S_6(t) = \sum_{f=1}^{F_N} (\log X(t,f) - \log X(t-1,f))^2.$$

On the other hand, the spectral contrast features are obtained as follows. Let a vector,

$$(X(i,t,1), X(i,t,2), \ldots, X(i,t,F_N(i))),$$

be the power spectrogram in the t-th frame and i-th subband. By sorting these elements in descending order, we obtain another vector,

$$(X'(i,t,1), X'(i,t,2), \ldots, X'(i,t,F_N(i))),$$

where

$$X'(i,t,1) > X'(i,t,2) > \cdots > X'(i,t,F_N(i)),$$

and $F_N(i)$ is the number of the i-th subband frequency bins:

$$F_N(i) = F_H(i) - F_L(i).$$

Here, the spectral contrast features are represented by spectral peak $S_7(i,t)$, spectral valley $S_8(i,t)$, and spectral contrast $S_9(i,t)$ as follows:

$$S_7(i,t) = \log\left(\frac{\sum_{f=1}^{\beta F_N(i)} X'(i,t,f)}{\beta F_N(i)}\right),$$

$$S_8(i,t) = \log\left(\frac{\sum_{f=(1-\beta)F_N(i)}^{F_N(i)} X'(i,t,f)}{\beta F_N(i)}\right), \text{ and}$$

$$S_9(i,t) = S_7(i,t) - S_8(i,t),$$

where β is a parameter to extract stable peak and valley values, and set to 0.2 in our experiments.

2.3. Similarity calculation

Our QBE retrieval system needs to calculate the similarity between musical pieces, i.e., a query example and each piece in a database, on the basis of the whole mood of the piece. Musical genres usually depend on the whole mood, not detailed variations.

To model the mood of each piece, we use a Gaussian Mixture Model (GMM) that approximates the distribution of acoustic features. We set the number of the mixtures to 8. Although the dimension of the obtained acoustic features was 33, it was reduced to 9 by using the principal component analysis where the cumulative percentage of eigenvalues was 0.95.

To measure the similarity among feature distributions, we utilized Earth Movers Distance (EMD) [11]. The EMD is based on the minimal cost needed to transform one distribution into another one.

3. SOUND SOURCE SEPARATION USING INTEGRATED TONE MODEL

As mentioned in Section 1, musical audio signals should be separated into instrument parts beforehand to boost and reduce the volume of those parts. We describe our sound source separation method in this section.

Input and output of our method are described as follows:

Table 2: Parameters of integrated tone model.

Symbol	Description
$w_{kl}^{(J)}$	Overall amplitude
$w_{kl}^{(H)}$, $w_{kl}^{(I)}$	Relative amplitude of harmonic and inharmonic tone models
$u_{klm}^{(H)}$	Amplitude coefficient of temporal power envelope for harmonic tone model
$v_{kln}^{(H)}$	Relative amplitude of n-th harmonic component
$u_{klm}^{(I)}$	Amplitude coefficient of temporal power envelope for inharmonic tone model
$v_{kln}^{(I)}$	Relative amplitude of the n-th inharmonic component
τ_{kl}	Onset time
$\rho_{kl}^{(H)}$	Diffusion of temporal power envelope for harmonic tone model
$\rho_{kl}^{(I)}$	Diffusion of temporal power envelope for inharmonic tone model
$\omega_{kl}^{(H)}$	F0 of harmonic tone model
$\sigma_{kl}^{(H)}$	Diffusion of harmonic components along frequency axis
β, κ	Coefficients that determine the arrangement of the frequency structure of inharmonic model

Input Power spectrogram of a musical piece and its musical score (standard MIDI file).[2] We suppose the spectrogram and the score have already been aligned (synchronized) by using another method.

Output Decomposed spectrograms that correspond to each instrument.

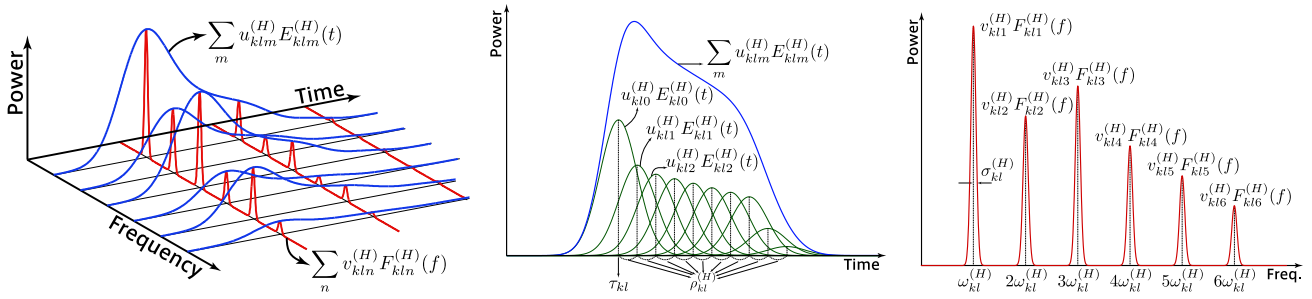
To separate the power spectrogram, we approximate the power spectrogram is additive. The musical audio signal corresponding to the decomposed power spectrogram is obtained by using the inverse short-time Fourier transform with the phase of the input spectrogram. In addition, we used an instrument sound database as training data for learning acoustic features of each musical instrument.

In this section, we first define the problem of separating sound sources and the integrated tone model. We then derive an iterative algorithm that consists of two steps: sound source separation and model parameter estimation.

3.1. Integrated tone model of harmonic and inharmonic models

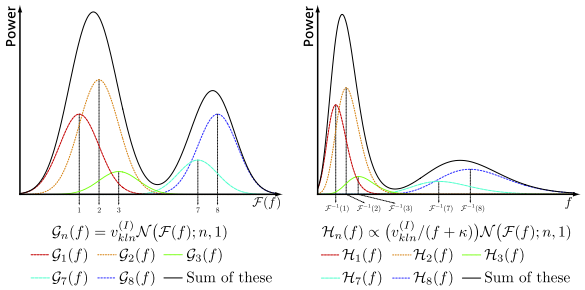
Separating the sound source means decomposing the input power spectrogram, $X(t,f)$, into a power spectrogram that corresponds to each musical note, where t and f are the time and the frequency, respectively. We assume that $X(t,f)$ includes K musical instruments and the k-th instrument performs L_k musical notes.

[2] Standard MIDI files for famous songs are often available thanks to Karaoke applications.



(a) Overview of harmonic tone model. (b) Temporal structure of harmonic tone model. (c) Frequency structure of harmonic tone model.

Figure 2: Overall, temporal and frequency structures of harmonic tone model. Harmonic tone model consists of a two-dimensional Gaussian Mixture Model, and is factorized into a pair of one-dimensional GMMs.



(a) Equally-spaced Gaussian kernels along the log-scale frequency, $\mathcal{F}(f)$. (b) Gaussian kernels obtained by changing the random variables of the kernels in (a).

Figure 3: Frequency structure of inharmonic tone model.

We use an integrated tone model, $J_{kl}(t,f)$, to represent the power spectrogram of the l-th musical note performed by the k-th musical instrument (*(k,l)-th note*). This tone model is defined as the sum of harmonic-structure tone models, $H_{kl}(t,f)$, and inharmonic-structure tone models, $I_{kl}(t,f)$, multiplied by the whole amplitude of the model, $w_{kl}^{(J)}$:

$$J_{kl}(t,f) = w_{kl}^{(J)}\left(w_{kl}^{(H)}H_{kl}(t,f) + w_{kl}^{(I)}I_{kl}(t,f)\right),$$

where $w_{kl}^{(J)}$ and $\left(w_{kl}^{(H)}, w_{kl}^{(I)}\right)$ satisfies the following constraints:

$$\sum_{k,l} w_{kl}^{(J)} = \iint X(t,f)\,dt\,df,$$
$$\forall k,l : w_{kl}^{(H)} + w_{kl}^{(I)} = 1$$

The harmonic tone model, $H_{kl}(t,f)$, is defined as a constrained two-dimensional Gaussian Mixture Model (GMM), which is a product of two one-dimensional GMMs, $\sum u_{klm}^{(H)}E_{klm}^{(H)}(t)$ and $\sum v_{kln}^{(H)}F_{kln}^{(H)}(f)$. This model is designed by referring to the harmonic-temporal-structured clustering (HTC) source model [12]. Analogously, the inharmonic tone model, $I_{kl}(t,f)$, is defined as a constrained two-dimensional GMM that is a product of two one-dimensional GMMs, $\sum u_{klm}^{(I)}E_{klm}^{(I)}(t)$ and $\sum v_{kln}^{(I)}F_{kln}^{(I)}(f)$. The temporal structures of these tone models, $E_{klm}^{(H)}(t)$ and $E_{klm}^{(I)}(t)$, are defined as an identical mathematical formula, but the frequency structures, $F_{kln}^{(H)}(f)$ and $F_{kln}^{(I)}(f)$, are defined as different forms. The definition of these models is as follows:

$$H_{kl}(t,f) = \sum_{m=0}^{M_H-1}\sum_{n=1}^{N_H} u_{klm}^{(H)}E_{klm}^{(H)}(t)v_{kln}^{(H)}F_{kln}^{(H)}(f),$$

$$I_{kl}(t,f) = \sum_{m=0}^{M_I-1}\sum_{n=1}^{N_I} u_{klm}^{(I)}E_{klm}^{(I)}(t)v_{kln}^{(I)}F_{kln}^{(I)}(f),$$

$$E_{klm}^{(H)}(t) = \frac{1}{\sqrt{2\pi}\rho_{kl}^{(H)}}\exp\left(-\frac{(t-\tau_{klm}^{(H)})^2}{2(\rho_{kl}^{(H)})^2}\right),$$

$$F_{kln}^{(H)}(f) = \frac{1}{\sqrt{2\pi}\sigma_{kl}^{(H)}}\exp\left(-\frac{(f-\omega_{kln}^{(H)})^2}{2(\sigma_{kl}^{(H)})^2}\right),$$

$$E_{klm}^{(I)}(t) = \frac{1}{\sqrt{2\pi}\rho_{kl}^{(I)}}\exp\left(-\frac{(t-\tau_{klm}^{(I)})^2}{2(\rho_{kl}^{(I)})^2}\right),$$

$$F_{kln}^{(H)}(f) = \frac{1}{\sqrt{2\pi}(f+\kappa)\log\beta}\exp\left(-\frac{(\mathcal{F}(f)-n)^2}{2}\right),$$

$$\tau_{klm}^{(H)} = \tau_{kl} + m\rho_{kl}^{(H)},$$

$$\omega_{kln}^{(H)} = n\omega_{kl}^{(H)},$$

$$\tau_{klm}^{(I)} = \tau_{kl} + m\rho_{kl}^{(I)}, \text{and}$$

$$\mathcal{F}(f) = \log\left(\frac{f}{\kappa}+1\right)\Big/\log\beta.$$

All parameters of $J_{kl}(t,f)$ are listed in Table 2. Here, M_H and N_H are the number of Gaussian kernels that represent temporal and frequency structures of the harmonic tone model, and M_I and N_I are the number of Gaussians that represent the ones of the inharmonic tone model, respectively. β and κ are coefficients that determine the arrangement of Gaussian kernels for the frequency structure of the

inharmonic model[3]. $u_{klm}^{(H)}$, $v_{kln}^{(H)}$, $u_{klm}^{(I)}$, and $v_{kln}^{(I)}$ satisfy the following conditions:

$$\forall k,l: \sum_m u_{klm}^{(H)} = 1, \quad \forall k,l: \sum_n v_{kln}^{(H)} = 1,$$
$$\forall k,l: \sum_m u_{klm}^{(I)} = 1, \text{and} \quad \forall k,l: \sum_n v_{kln}^{(I)} = 1.$$

As shown in Figure 3, function $F_{kln}^{(I)}(f)$ is derived by changing the variables of the following probability density function:

$$\mathcal{N}(g; n, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(g-n)^2}{2}\right),$$

from $g = \mathcal{F}(f)$ to f, i.e.,

$$F_{kln}^{(I)}(f) = \frac{dg}{df}\mathcal{N}\big(\mathcal{F}(f); n, 1\big) = \frac{1}{(f+\kappa)\log\beta}\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\mathcal{F}(f)-n)^2}{2}\right).$$

3.2. Iterative separation algorithm

The goal of this separation is to decompose $X(t,f)$ into (k,l)-th note by multiplying a spectrogram distribution function, $\Delta^{(J)}(k,l;t,f)$, which satisfies

$$\forall k,l,t,f: 0 \le \Delta^{(J)}(k,l;t,f) \le 1 \quad \text{and}$$
$$\forall t,f: \sum_{k,l} \Delta^{(J)}(k,l;t,f) = 1.$$

With $\Delta^{(J)}(k,l;t,f)$, the separated power spectrogram, $X_{kl}^{(J)}(t,f)$, is obtained as

$$X_{kl}^{(J)}(t,f) = \Delta^{(J)}(k,l;t,f)X(t,f).$$

Then, let $\Delta^{(H)}(m,n;k,l,t,f)$ and $\Delta^{(I)}(m,n;k,l,t,f)$ be spectrogram distribution functions that decompose $X_{kl}^{(J)}(t,f)$ into each Gaussian distribution of the harmonic and inharmonic models, respectively. These functions satisfy

$$\forall k,l,m,n,t,f: 0 \le \Delta^{(H)}(m,n;k,l,t,f) \le 1,$$
$$\forall k,l,m,n,t,f: 0 \le \Delta^{(I)}(m,n;k,l,t,f) \le 1, \quad \text{and}$$
$$\forall k,l,t,f: \sum_{m,n} \Delta^{(H)}(m,n;k,l,t,f) + \sum_{m,n} \Delta^{(I)}(m,n;k,l,t,f) = 1.$$

With these functions, the separated power spectrograms, $X_{klmn}^{(H)}(t,f)$ and $X_{klmn}^{(I)}(t,f)$, are obtained as

$$X_{klmn}^{(H)}(t,f) = \Delta^{(H)}(m,n;k,l,t,f)X_{kl}^{(J)}(t,f) \quad \text{and}$$
$$X_{klmn}^{(I)}(t,f) = \Delta^{(I)}(m,n;k,l,t,f)X_{kl}^{(J)}(t,f).$$

To evaluate the effectiveness of this separation, we use an objective function defined as the Kullback-Leibler (KL) divergence from $X_{klmn}^{(H)}(t,f)$ and $X_{klmn}^{(I)}(t,f)$ to the each Gaussian kernel of the harmonic and inharmonic models:

$$Q^{(\Delta)} = \sum_{k,l}\Bigg(\sum_{m,n}\iint X_{klmn}^{(H)}(t,f) \log\frac{X_{klmn}^{(H)}(t,f)}{u_{klm}^{(H)}v_{kln}^{(H)}E_{klm}^{(H)}(t)F_{kln}^{(H)}(f)}\,dt\,df + \sum_{m,n}\iint X_{klmn}^{(I)}(t,f) \log\frac{X_{klmn}^{(H)}(t,f)}{u_{klm}^{(H)}v_{kln}^{(H)}E_{klm}^{(H)}(t)F_{kln}^{(H)}(f)}\,dt\,df\Bigg).$$

By minimizing $Q^{(\Delta)}$ pertaining to the distribution functions, we obtain the spectrogram distribution function and decomposed spectrograms, i.e., separated sounds, on the basis of the parameters of the tone models.

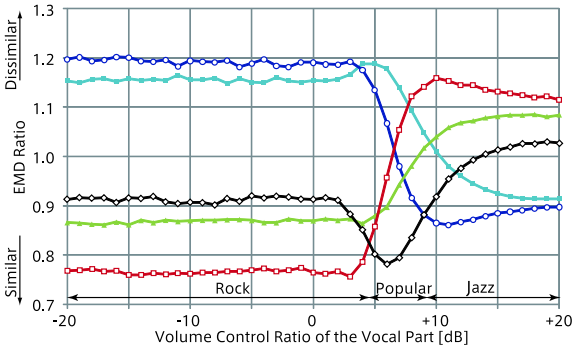
Once the input spectrogram is decomposed, the likeliest model parameters are calculated using a statistical estimation. We use prior distributions for several model parameters, $w_{kl}^{(H)}$ and $w_{kl}^{(I)}$, $u_{klm}^{(H)}$, $v_{kln}^{(H)}$, $u_{klm}^{(I)}$, and $v_{kln}^{(I)}$, to estimate robust parameters. These prior distributions are defined as Dirichlet distributions. An another objective function, $Q^{(\theta)}$, is used to estimate model parameters using the prior distributions, and is defined as

$$\begin{aligned} Q^{(\theta)} =& Q^{(\Delta)} + \log\mathcal{B}\Big(w_{kl}^{(H)}, w_{kl}^{(I)}; \alpha_{w_k}^{(H)}, \alpha_{w_k}^{(I)}\Big) \\ &+ \log\mathcal{D}\Big(\big\{u_{klm}^{(H)}\big\}; \big\{\alpha_{u_{km}}^{(H)}\big\}\Big) \\ &+ \log\mathcal{D}\Big(\big\{v_{kln}^{(H)}\big\}; \big\{\alpha_{v_{kn}}^{(H)}\big\}\Big) \\ &+ \log\mathcal{D}\Big(\big\{u_{klm}^{(I)}\big\}; \big\{\alpha_{u_{km}}^{(I)}\big\}\Big) \\ &+ \log\mathcal{D}\Big(\big\{v_{kln}^{(I)}\big\}; \big\{\alpha_{v_{kn}}^{(I)}\big\}\Big). \end{aligned}$$

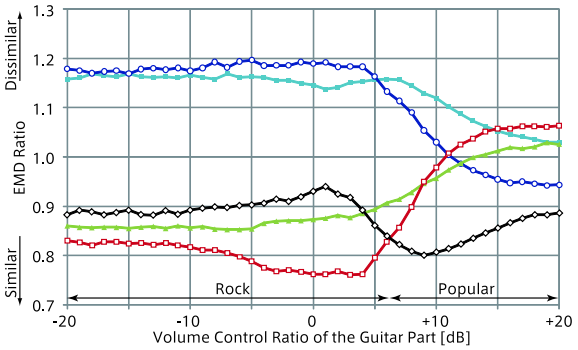
Here $\mathcal{B}(\cdot)$ and $\mathcal{D}(\cdot)$ are the probability density function of the Beta and Dirichlet distributions, and $\alpha_{w_k}^{(H)}$, $\alpha_{w_k}^{(I)}$, $\{\alpha_{u_{km}}^{(H)}\}$, $\{\alpha_{v_{kn}}^{(H)}\}$, $\{\alpha_{u_{km}}^{(I)}\}$, and $\{\alpha_{v_{kn}}^{(I)}\}$ are parameters of these distributions for each instrument. Note that this modification of the objective function has no effect on the calculation of the distribution function. The parameter update equations are described in Appendix A.

We obtain an iterative algorithm that consists of two steps: one is the step in which the distribution function is calculated while the model parameters are fixed, and the other is the step in which the parameters are updated under the distribution function. This iterative algorithm is equivalent to the Expectation-Maximization (EM) algorithm on the basis of the maximum A Posteriori estimation.

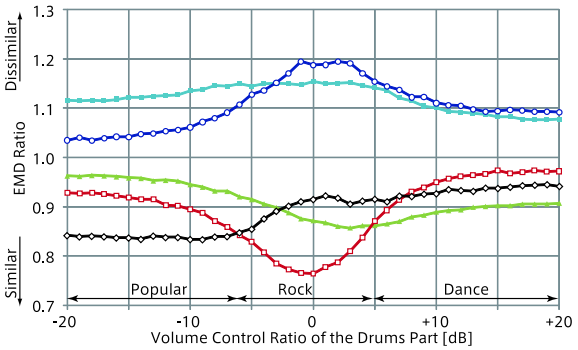
[3] If $1/(\log\beta)$ and κ are set to 1127 and 700, $\mathcal{F}(f)$ is equivalent to the mel scale of f Hz.



(a) Genre shift caused by changing the volume of vocal. Genre with the highest similarity changed from rock to popular and to jazz.



(b) Genre shift caused by changing the volume of guitar. Genre with the highest similarity changed from rock to popular.



(c) Genre shift caused by changing the volume of drums. Genre with the highest similarity changed from popular to rock and to dance.

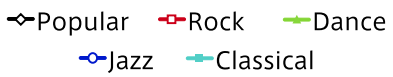
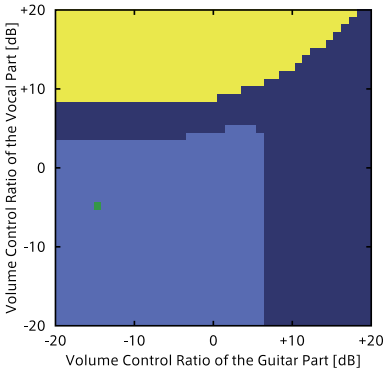
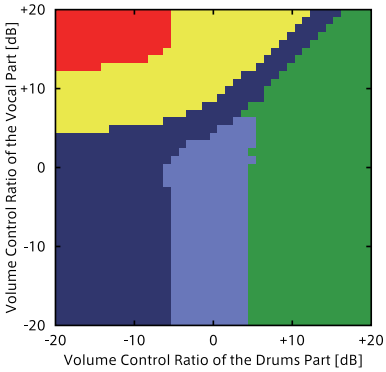


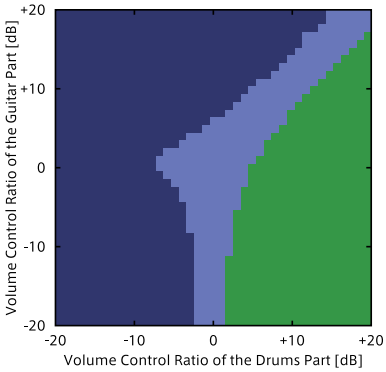
Figure 4: Ratio of average EMD per genre to average EMD of all genres while reducing or boosting the volume of single instrument part. Here, (a), (b), and (c) are for the vocal, guitar, and drums, respectively. Note that a smaller ratio of the EMD plotted in the lower area of the graph indicates higher similarity.



(a) Genre shift caused by changing the volume of vocal and guitar.



(b) Genre shift caused by changing the volume of vocal and drums.



(c) Genre shift caused by changing the volume of guitar and drums.

Figure 5: Genres that have the smallest EMD (the highest similarity) while reducing or boosting the volume of two instrument parts. The top, middle, and bottom are the cases of the vocal-guitar, vocal-drums, and guitar-drums, respectively.

4. EXPERIMENTAL EVALUATION

We conducted experiments to explore the relationship between instrument volume balances and genres. Given the query musical piece of which the volume balance is changed, the genres of the retrieved musical pieces are investigated.

Ten musical pieces were excerpted for the query from the *RWC Music Database: Popular Music* (RWC-MDB-P-2001 No. 1–10) [13]. The audio signals of these musical pieces were separated into each musical instrument part using the standard MIDI files which are provided as the AIST annotation [14]. The evaluation database consisted of other 50 musical pieces excerpted from the *RWC Music Database: Musical Genre* (RWC-MDB-G-2001). This excerpted database includes musical pieces in the following genres: Popular, Rock, Dance, Jazz, and Classical.

In the experiments, we reduced or boosted the volumes of three instrument parts — the vocal, guitar, and drums. To shift the genre of the musical piece by changing the volume of these parts, the part of an instrument should have sufficient duration[4]. Thus, the above three instrument parts were chosen because these parts satisfy the following two constraints:

1. be played in all 10 musical pieces for the query, and
2. be played for more than 60 % of the duration of each piece.

The EMDs were calculated between the acoustic feature distributions of each query song and each piece in the database, while reducing or boosting the volume of these musical instrument parts between -20 and $+20$ dB. Figure 4 shows the results of changing the volume of a single instrument part. The vertical axis is the relative ratio of the EMD averaged over the 10 pieces, which is defined as:

$$\text{EMD ratio} = \frac{\text{average EMD of each genre}}{\text{average EMD of all genres}} .$$

On the other hand, Figure 5 shows the results of simultaneously changing the volume of two instrument parts (instrument pairs).

4.1. Volume change of single instrument

The results in Figure 4 clearly show that the genre shift occurred by changing the volume of any instrument part. Note that the genre of the retrieved pieces at 0 dB (giving the original queries without any changes) is same for all the three figures (a), (b), and (c). Although we used 10 popular songs excerpted from the *RWC Music Database: Popular Music* for the queries, they are considered to be rock music at the genre with the highest similarity at 0 dB because those songs actually have the rock taste with strong guitar and drum sounds.

By increasing the volume of the vocal from -20 dB, the genre with the highest similarity shifted from rock (-20 to 4 dB), to popular (5 to 9 dB), and to jazz (10 to 20 dB) as shown in Figure 4 (a). By changing the volume of the guitar, the genre shifted from rock (-20 to 7 dB) to popular (8 to 20 dB) as shown in Figure 4 (b). Although it was commonly observed that the genre shifted from rock to popular in both cases of vocal and guitar, the genre shift to jazz occurred only in the case of vocal. These results indicate that the vocal and guitar would have different importance in the jazz music. By changing the volume of the drums, genres shifted from popular (-20 to -7 dB), to rock (-6 to 4 dB), and to dance (5 to 20 dB) as shown in Figure 4 (c). These results indicate the reasonable relationship between the instrument volume balance and the musical genre shift, and this relationship is consistent with typical images of musical genres.

4.2. Volume change of two instruments (pair)

The results in Figure 5 show that the genre shift depends on the volume change of two instrument parts. If one of the part is not changed (at 0 dB), the results are same with Figure 4.

Although the basic tendency in the genre shifts is similar to the single instrument experiment, classical music, which does not appear as the genre with the highest similarity in Figure 4, appears in Figure 5 (b) when the vocal part is boosted and the drums part is reduced. The similarity of rock music is decreased when we *separately* boosted one of the guitar or drums, but it is interesting that rock music can keep the highest similarity if both of the guitar and drums are boosted *together* as shown in Figure 5 (c). This result is well matched with the typical image of rock music, and suggests promising possibilities as a tool for customizing the query for QBE retrieval.

5. CONCLUSION

We have described musical genre shift by changing the volume of separated instrument parts and a QBE retrieval approach on the basis of the genre shift. This approach is important because it was not possible for a user to customize the QBE query in the past and the user always had to find different pieces to obtain different retrieved results. By using the genre shift based on our original sound sourse separation method, it becomes easy and intuitive to customize the QBE query just by changing the volume of instrument parts. Experiments results confirmed our hypothesis that the musical genre shifts in relation to the volume balance of instruments.

Although the current genre shift depends on only the volume balance, other factors such as rhythm patterns, sound effects, and chord progressions will also be useful to cause the shift if we can control them. In the future, we

[4] For example, the volume of an instrument that is performed for 5 seconds in a 5-minute musical piece may not affect the genre of the piece.

plan to pursue the promising approach proposed in this paper and develop a better QBE retrieval system that easily reflects user's intention and preference.

6. ACKNOWLEDGEMENTS

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research of Priority Areas, Primordial Knowledge Model Core of Global COE program and JST CrestMuse Project.

7. REFERENCES

[1] A. Rauber, E. Pampalk, and D. Merkl, "Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarity," in *Proc. ISMIR*, 2002, pp. 71–80.

[2] C. Yang, "The macsis acoustic indexing framework for music retrieval: An experimental study," in *Proc. ISMIR*, 2002, pp. 53–62.

[3] E. Allamanche, J. Herre, O. Hellmuth, T. Kastner, and C. Ertel, "A multiple feature model for musical similarity retrieval," in *Proc. ISMIR*, 2003.

[4] Y. Feng, Y. Zhuang, and Y. Pan, "Music information retrieval by detecting mood via computational media aesthetics," in *Proc. WI*, 2003, pp. 235–241.

[5] B. Thoshkahna and K. R. Ramakrishnan, "Projekt quebex: A query by example system for audio retrieval," in *Proc. ICME*, 2005, pp. 265–268.

[6] F. Vignoli and S. Pauws, "A music retrieval system based on user-driven similarity and its evaluation," in *Proc. ISMIR*, 2005, pp. 272–279.

[7] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Musical instrument recognizer "instrogram" and its application to music retrieval based on instrumentation similarity," in *Proc. ISM*, 2006, pp. 265–274.

[8] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H.G. Okuno, "Instrument equalizer for query-by-example retrieval: Improving sound source separation based on integrated harmonic and inharmonic models," in *Proc. ISMIR*, 2008, pp. 133–138.

[9] L. Lu, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 14, no. 1, pp. 5–18, Jan. 2006.

[10] D. N. Jiang, L. Lu, H. J. Zhang, J. H. Tao, and L. H. Cai, "Music type classification by spectral contrast features," in *Proc. ICME*, 2002, pp. 113–116.

[11] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Proc. ICCV*, 1998, pp. 59–66.

[12] H. Kameoka, T. Nishimoto, and S. Sagayama, "Harmonic-temporal structured clustering via deterministic annealing em algorithm for audio feature extraction," in *Proc. ISMIR*, 2005, pp. 115–122.

[13] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. ISMIR*, 2002, pp. 287–288.

[14] M. Goto, "AIST annotation for the RWC music database," in *Proc. ISMIR*, 2006, pp. 359–360.

A. PARAMETER UPDATE EQUATIONS

We describe the update equation for each parameter derived from the M-step of the EM algorithm by differentiating the cost function about each parameter. Let $X^{(H)}_{klmn}(t,f)$ and $X^{(I)}_{klmn}(t,f)$ be the decomposed power:

$$X^{(H)}_{klmn}(t,f) = \Delta^{(H)}(m,n;k,l,t,f)X^{(J)}_{kl}(t,f) \quad \text{and}$$
$$X^{(I)}_{klmn}(t,f) = \Delta^{(I)}(m,n;k,l,t,f)X^{(J)}_{kl}(t,f).$$

Summation or integration of the decomposed power over indices or variables are denoted by omitting these characters, e.g., $X^{(H)}_{kl}(t,f)$ and $X^{(H)}_{klm}$.

A.1. $w^{(J)}_{kl}$: overall amplitude

$$w^{(J)}_{kl} = X^{(H)}_{kl} + X^{(I)}_{kl}.$$

A.2. $w^{(H)}_{kl}$, $w^{(I)}_{kl}$: relative amplitude of harmonic and inharmonic tone models

$$w^{(H)}_{kl} = \frac{X^{(H)}_{kl} + \alpha^{(H)}_{w_k}}{X^{(H)}_{kl} + X^{(I)}_{kl} + \alpha^{(H)}_{w_k} + \alpha^{(I)}_{w_k}} \quad \text{and}$$
$$w^{(H)}_{kl} = \frac{X^{(I)}_{kl} + \alpha^{(I)}_{w_k}}{X^{(H)}_{kl} + X^{(I)}_{kl} + \alpha^{(H)}_{w_k} + \alpha^{(I)}_{w_k}}.$$

A.3. $u^{(H)}_{klm}$: amplitude coefficient of temporal power envelope for harmonic tone model

$$u^{(H)}_{klm} = \frac{X^{(H)}_{klm} + \alpha^{(H)}_{u_{km}}}{X^{(H)}_{kl} + \sum_m \alpha^{(H)}_{u_{km}}}.$$

A.4. $v^{(H)}_{kln}$: relative amplitude of n-th harmonic component

$$v^{(H)}_{kln} = \frac{X^{(H)}_{kln} + \alpha^{(H)}_{v_{kn}}}{X^{(H)}_{kl} + \sum_n \alpha^{(H)}_{v_{kn}}}.$$

A.5. $u^{(I)}_{klm}$: amplitude coefficient of temporal power envelope for inharmonic tone model

$$u^{(I)}_{klm} = \frac{X^{(I)}_{klm} + \alpha^{(I)}_{u_{km}}}{X^{(I)}_{kl} + \sum_m \alpha^{(I)}_{u_{km}}}.$$

A.6. $v^{(I)}_{kln}$: relative amplitude of n-th inharmonic component

$$v^{(I)}_{kln} = \frac{X^{(I)}_{kln} + \alpha^{(I)}_{v_{kn}}}{X^{(I)}_{kl} + \sum_n \alpha^{(I)}_{v_{kn}}}.$$

A.7. τ_{kl}: onset time

$$\tau_{kl} = \Bigg(\sum_m \int (t - m\rho^{(H)}_{kl}) X^{(H)}_{klm}(t)\, df + \sum_m \int (t - m\rho^{(I)}_{kl}) X^{(I)}_{klm}(t)\, df \Bigg) \Big/ \big(X^{(H)}_{kl} + X^{(I)}_{kl} \big).$$

A.8. $\omega^{(H)}_{kl}$: F0 of harmonic tone model

$$\omega^{(H)}_{kl} = \frac{\sum_n \int n f X^{(H)}_{kln}(f)\, df}{\sum_n n^2 X^{(H)}_{kln}}.$$

A.9. $\sigma^{(H)}_{kl}$: diffusion of harmonic components along frequency axis

$$\sigma^{(H)}_{kl} = \sqrt{\frac{\sum_n \int (f - n\omega^{(H)}_{kl})^2 X^{(H)}_{kln}(f)\, df}{X^{(H)}_{kl}}}.$$