Musical Similarity and Commonness Estimation based on Probabilistic Generative Models

Tomoyasu Nakano
National Institute of Advanced Industrial Science and Technology (AIST), Japan
t.nakano@aist.go.jp

Kazuyoshi Yoshii
Kyoto University, Japan
yoshii@kuis.kyoto-u.ac.jp

Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST), Japan
m.goto@aist.go.jp

***Abstract*—This paper proposes a novel concept we call *musical commonness*, which is the similarity of a song to a set of songs; in other words, its *typicality*. This commonness can be used to retrieve representative songs from a song set (*e.g.*, songs released in the 80s or 90s). Previous research on musical similarity has compared two songs but has not evaluated the similarity of a song to a set of songs. The methods presented here for estimating the similarity and commonness of polyphonic musical audio signals are based on a unified framework of probabilistic generative modeling of four musical elements (vocal timbre, musical timbre, rhythm, and chord progression). To estimate the commonness, we use a generative model trained from a song set instead of estimating musical similarities of all possible song-pairs by using a model trained from each song. In experimental evaluation, we used 3278 popular music songs. Estimated song-pair similarities are comparable to ratings by a musician at the 0.1% significance level for vocal and musical timbre, at the 1% level for rhythm, and the 5% level for chord progression. Results of commonness evaluation show that the higher the musical commonness is, the more similar a song is to songs of a song set.**

***Keywords*-musical similarity; musical commonness; typicality; latent Dirichlet allocation; variational Pitman-Yor language model;**

I. Introduction

The digitization of music and the distribution of content over the web have greatly increased the number of musical pieces that listeners can access but are also causing problems for both listeners and creators. Listeners find that selecting music is getting more difficult, and creators find that their creations can easily just disappear into obscurity. Musical similarity [1], [2] between two songs can help with these problems because it provides a basis for retrieving musical pieces that closely match a listener's favorites, and several similarity-based music information retrieval (MIR) systems [3]–[6] and music recommender systems [2], [7] have been proposed. None, however, has focused on the musical similarity of a song to a set of songs, such as those in a particular genre or personal collection, those on a specific playlist, or those released in a given year or a decade.

This paper focuses on musical similarity and *musical commonness* that can be used in MIR systems and music recommender systems. As shown in Figure 1, we define musical commonness as a similarity assessed by comparing a song with a set of songs. The more similar a song is to songs in that set, the higher its musical commonness. Our definition is based on *central tendency*, which, in cognitive psychology, is one of the determinants of *typicality* [8]. Musical commonness can be used to recommend a representative or introductory song for a song set, and it can help listeners understand the relationship between a song and a song set.

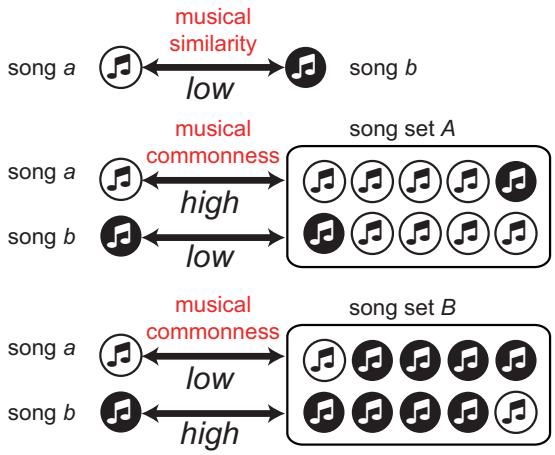

Figure 1. Musical similarity and commonness.

To estimate musical similarity and commonness, we propose a generative modeling of four musical elements: vocal timbre, musical timbre, rhythm, and chord progression (Figure 2). Two songs are considered to be similar if one has descriptions (*e.g.*, chord names) that have a high probability in a model of the other. This probabilistic approach has previously been used to compute similarity between two songs [9], [10]. To compute commonness for each element, a generative model is derived for a set of songs. A song is considered to be common to that set if the descriptions of the song have a high probability in the derived model.

II. Methods

From polyphonic musical audio signals including a singing voice and sounds of various musical instruments we first extract vocal timbre, musical timbre, and rhythm and

978-1-5090-0379-2/15 $31.00 © 2015 IEEE
DOI 10.1109/ISM.2015.102

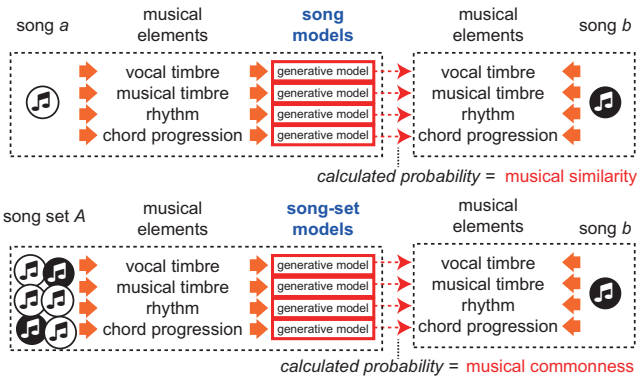

Figure 2. Musical similarity and commonness based on probabilistic generative modeling of four musical elements: vocal timbre, musical timbre, rhythm, and chord progression.

estimate chord progression. We then model the timbres and rhythm by using a vector quantization method and latent Dirichlet allocation (LDA) [11]. The chord progression is modeled by using a variable-order Markov process (up to a theoretically infinite order) called the variable-order Pitman-Yor language model (VPYLM) [12], [13].

When someone compares two pieces of music, they may feel that they share some factors that characterize their timbres, rhythms and chord progressions, even if they cannot articulate exactly what these factors are. We call these "*latent factors*" and would like to estimate them from low-level features. This is difficult to do for individual songs, but using the above methods (LDA and VPYLM) we can do so using many songs.

Finally, for each element we calculate two probabilities (Figure 2). One is for similarity estimation and is calculated by using a generative model trained from a musical piece (this model is called a *song model*). The other is for commonness estimation and is calculated by using a generative model trained from a set of musical pieces (this model is called a *song-set model*).

A. Similarity and commonness: Vocal timbre, musical timbre, and rhythm

The method used to train song models of vocal timbre, musical timbre, and rhythm is based on a previous work [14] on modeling vocal timbre. In addition, we propose a method to train song-set models under the LDA-based modeling.

1) Extracting acoustic features: Vocal timbre: We use the mel-frequency cepstral coefficients of the LPC spectrum of the vocal (LPMCCs) and the ΔF_0 of the vocal to represent vocal timbre because they are effective for identifying singers [10], [14].

We first use Goto's PreFEst [15] to estimate the F_0 of the predominant melody from an audio signal and then the F_0 is used to estimate the ΔF_0 and the LPMCCs of the vocal. To estimate the LPMCCs, the vocal sound is re-synthesized by using a sinusoidal model based on the estimated vocal F_0 and the harmonic structure estimated from the audio signal. At each frame the ΔF_0 and the LPMCCs are combined as a feature vector.

Then *reliable frames* (frames little influenced by accompaniment sound) are selected by using a vocal GMM and a non-vocal GMM (see [10] for details). Feature vectors of only the reliable frames are used in the following processes (model training and probability estimation).

2) Extracting acoustic features: Musical timbre: We use mel-frequency cepstral coefficients (MFCCs), their derivatives (ΔMFCCs), and Δpower to represent musical timbre, combining them as a feature vector. The MFCCs are musical timbre features used in music information retrieval [16], and this feature vector is often used in speech recognition.

3) Extracting acoustic features: Rhythm: To represent rhythm we use the fluctuation patterns (FPs) designed to describe the rhythmic signature of musical audio [16], [17]. They are features effective for music information retrieval [16] and for evaluating musical complexity with respect to tempo [18].

We first calculate the *specific loudness sensation* for each frequency band by using an auditory model (*i.e.*, the outer-ear model) and the Bark frequency scale. The FPs are then obtained by using a FFT to calculate the amplitude modulation of the loudness sensation and weighting its coefficients on the basis of a psychoacoustic model of the *fluctuation strength* (see [16], [17] for details). Finally, the number of vector dimensions of the FPs was reduced by using principle component analysis (PCA).

4) Quantization: All acoustic feature vectors of each element are converted to symbolic time series by using a vector quantization method called the k-means algorithm. In that algorithm the vectors are normalized by subtracting the mean and dividing by the standard deviation, and then the normalized vectors are quantized by prototype vectors (centroids) trained previously. Hereafter, we call the quantized symbolic time series *acoustic words*.

5) Probabilistic generative model: The observed data we consider for LDA are D independent songs $\vec{X} = \{\vec{X}_1, ..., \vec{X}_D\}$. A song $\vec{X}_d$ is N_d acoustic words $\vec{X}_d = \{\vec{x}_{d,1}, ..., \vec{x}_{d,N_d}\}$. The size of the acoustic words vocabulary is equivalent to the number of clusters of the k-means algorithm ($= V$), $\vec{x}_{d,n}$ is a V-dimensional "1-of-V" vector (a vector with one element containing a 1 and all other elements containing a 0). The latent variable of the observed $\vec{X}_d$ is $\vec{Z}_d = \{\vec{z}_{d,1}, ..., \vec{z}_{d,N_d}\}$. The number of topics is K, so $z_{d,n}$ indicates a K-dimensional 1-of-K vector. Hereafter, all latent variables of D songs are indicated $\vec{Z} = \{\vec{Z}_1, ..., \vec{Z}_D\}$.

The full joint distribution of the LDA model is given by

$$p(\vec{X}, \vec{Z}, \vec{\pi}, \vec{\phi}) = p(\vec{X}|\vec{Z}, \vec{\phi})p(\vec{Z}|\vec{\pi})p(\vec{\pi})p(\vec{\phi}) \qquad (1)$$

where $\vec{\pi}$ indicates the mixing weights of the multiple topics (D of the K-dimensional vector) and $\vec{\phi}$ indicates the

unigram probability of each topic (K of the V-dimensional vector). The first two terms are likelihood functions, and the other two are prior distributions. The likelihood functions are defined as

$$p(\vec{X}|\vec{Z},\vec{\phi}) = \prod_{d=1}^{D}\prod_{n=1}^{N_d}\prod_{v=1}^{V}\left(\prod_{k=1}^{K}\phi_{k,v}^{z_{d,n,k}}\right)^{x_{d,n,v}} \quad (2)$$

and

$$p(\vec{Z}|\vec{\pi}) = \prod_{d=1}^{D}\prod_{n=1}^{N_d}\prod_{v=1}^{V}\pi_{d,k}^{z_{d,n,k}}. \quad (3)$$

We then introduce conjugate priors as follows:

$$p(\vec{\pi}) = \prod_{d=1}^{D}\mathrm{Dir}(\vec{\pi}_d|\vec{\alpha}^{(0)}) = \prod_{d=1}^{D}C(\vec{\alpha}^{(0)})\prod_{k=1}^{K}\pi_{d,k}^{\alpha^{(0)}-1}, \quad (4)$$

$$p(\vec{\phi}) = \prod_{k=1}^{K}\mathrm{Dir}(\vec{\phi}_k|\vec{\beta}^{(0)}) = \prod_{k=1}^{K}C(\vec{\beta}^{(0)})\prod_{v=1}^{V}\phi_{k,v}^{\beta^{(0)}-1}, \quad (5)$$

where p($\vec{\pi}$) and p($\vec{\phi}$) are products of Dirichlet distributions, $\vec{\alpha}^{(0)}$ and $\vec{\beta}^{(0)}$ are hyperparameters of prior distributions (with no observation), and $C(\vec{\alpha}^{(0)})$ and $C(\vec{\beta}^{(0)})$ are normalization factors.

6) Similarity estimation: The similarity between song a and song b is represented as a probability of song b calculated using a song model of song a. This probability $p_g(b|a)$ is defined as follows:

$$\log p_g(b|a) = \frac{1}{N_b}\sum_{n=1}^{N_b}\log p(\vec{x}_{b,n}|\ \mathbb{E}[\vec{\pi}_a],\mathbb{E}[\vec{\phi}]), \quad (6)$$

$$p(\vec{x}_{b,n}|\ \mathbb{E}[\vec{\pi}_a],\mathbb{E}[\vec{\phi}]) = \sum_{k=1}^{K}\left(\mathbb{E}[\pi_{a,k}]\cdot\mathbb{E}[\phi_{k,v}]\right), \quad (7)$$

where $\mathbb{E}[\cdot]$ is the expectation of a Dirichlet distribution and v is the corresponding index (the word id) of the K-dimensional 1-of-K observation vector $\vec{x}_{b,n}$.

7) Commonness estimation: To estimate the commonness, we propose a method for obtaining a generative model from a song set without using the LDA-model-training process again. In this case, hyperparameters $\alpha_{d,k}$ of the posterior distribution can be interpreted as effective numbers of observations of the corresponding values of the 1-of-K observation vector $\vec{x}_{d,n}$.

This means that a song-set model of a song set A can be obtained by summing those hyperparameters $\vec{\alpha}_d = \{\alpha_{d,1},...,\alpha_{d,K}\}$. This model $\vec{\alpha}_A$ is defined as follows:

$$\vec{\alpha}_A = \sum_{d\in A}\left(\vec{\alpha}_d - \vec{\alpha}^{(0)}\right) + \vec{\alpha}^{(0)}, \quad (8)$$

where the prior ($\vec{\alpha}^{(0)}$) is added just once. Musical commonness between the song set A and the song a is represented as a probability of song a that is calculated using the song-set model of the song set A: $\log p_g(a|A)$.

B. Similarity and commonness: Chord progression

We first estimate key and chord progression by using modules of Songle [19], a web service for active music listening.

Before modeling, estimated results of chord progression are normalized. The root note is shifted so that the key will be /C/, flat notes (♭) are unified into sharp notes (♯), and the five variants of major chords with different bass notes are unified (they are dealt with as the same chord type). When same chord types continue, they are collected into a single occurrence (*e.g.*, /C C C/ into /C/).

1) Probabilistic generative model: For modeling of chord progression of a set of musical pieces, the VPYLM used as a song-set model is trained using a song set used to compute musical commonness. In the song modeling process, however, suitable training cannot be done using only a Bayesian model (VPYLM) because the amount of training data is not sufficient. To deal with this problem, we use as a song model a trigram model trained by maximum likelihood estimation.

2) Similarity and commonness estimation: Similarity and commonness are represented by using as the generative probability the inverse of the *perplexity* (average probability of each chord). To avoid the zero-frequency problem, chord similarity between two songs is estimated by calculating weighted mean probabilities of the song model and the song-set model. The weights are $(1-r)$ and r, respectively (r is set to 10^{-5}).

III. Experiments

The proposed methods were tested in experiments evaluating the estimated similarity (Experiment A) and the estimated commonness (Experiment B).

A. Dataset

The song set used for the model training, similarity estimation, and commonness estimation comprised 3278 Japanese popular songs that appeared on a popular music chart in Japan (http://www.oricon.co.jp/) and were placed in the top twenty on weekly charts appearing between 2000 and 2008. Here we refer to this song set as the JPOP music database (MDB). The twenty artists focused on for similarity evaluation are listed in Table I.

In addition, for GMM/k-means/PCA training to extract the acoustic features, 100 popular songs from the RWC Music Database (RWC-MDB-P-2001) [20] were also used. These 80 song in Japanese and 20 in English reflect styles of the Japanese popular songs (J-Pop) and Western popular songs in or before 2001. Here we refer this song set as the RWC MDB.

B. Experimental Settings

Conditions and parameters of the methods described in the METHODS section are described here in detail.

Table I
SINGERS OF THE 463 SONGS USED IN THE EXPERIMENTS.

ID	Artist name	Gender of vocalilst(s) (*: more than one singer)	Number of songs
A	Ayumi Hamasaki	female	33
B	B'z	male	28
C	Morning Musume	female*	28
D	Mai Kuraki	female	27
E	Kumi Koda	female	25
F	BoA	female	24
G	EXILE	male*	24
H	L'Arc-en-Ciel	male	24
I	Rina Aiuchi	female	24
J	w-inds.	male*	23
K	SOPHIA	male	22
L	Mika Nakashima	female	22
M	CHEMISTRY	male*	21
N	Gackt	male	21
O	GARNET CROW	female	20
P	TOKIO	male*	20
Q	Porno Graffitti	male	20
R	Ken Hirai	male	20
S	Every Little Thing	female	19
T	GLAY	male	19
Total		male 11 – female 9	463

1) Extracting acoustic features: For vocal timbre features, we targeted monaural 16-kHz digital recordings and extracted ΔF_0 and 12th-order LPMCCs every 10 ms. The feature vectors were extracted from each song, using as reliable vocal frames the top 15% of the feature frames. Using the 100 songs of the RWC MDB, a vocal GMM and a non-vocal GMM were trained by variational Bayesian inference [21]. We set the number of Gaussians to 32 and set the hyperparameter of a Dirichlet distribution over the mixing coefficients to 1.0[1].

For musical timbre features, we targeted monaural 16-kHz digital recordings and extracted Δpower, 12th-order MFCCs, and 12th-order ΔMFCCs every 10 ms. The feature vectors were extracted from 15% of the frames of each song and those frames were selected randomly.

For rhythm-based features, we targeted monaural 11.025-kHz digital recordings and extracted FPs by using the Music Analysis (MA) toolbox for Matlab [16]. A 1200-dimension FP vector was estimated every 3 seconds and the analysis frame length was 6 seconds. We then reduced the number of vector dimensions by using PCA based on the cumulative contribution ratio ($\leq 95\%$). A projection matrix for PCA was computed by using the 100 songs of the RWC MDB. Finally, a 78-dimensional projection matrix was obtained.

The conditions described above (*e.g.*, the 16- and 11.025-kHz sampling frequencies) were based on previous works.

2) Quantization: To quantize the vocal features, we set the number of clusters of the k-means algorithm to 100 and also trained used the 100 songs of the RWC MBD to train the centroids. This k is same number used in our previous work [14]. The number of clusters used to quantize the musical timbre and rhythm features was set to 64 in this evaluation.

3) Chord estimation: With Songle, chords are transcribed using 14 chord types: major, major 6th, major 7th, dominant 7th, minor, minor 7th, half-diminished, diminished, augmented, and five variants of major chords with different bass notes (/2, /3, /5, /b7, and /7). The resulting 168 chords (14 types $\times$ 12 root notes) and one "no chord" label are estimated (see [19] for details).

4) Training the generative models: Training song models and song-set models of the 4 musical elements by LDA and VPYLM, we used all of the 3278 original recordings of the JPOP MDB.

The number of topics K was set to 100, and the model parameters of LDA were trained by using the collapsed Gibbs sampler [22]. The hyperparameters of the Dirichlet distributions for topics and words were initially set to 1 and 0.1, respectively. The conditions were based on our previous work [14].

The number of chords used to model chord progression was 97: the 8 chord types (major, major 6th, major 7th, dominant 7th, minor, minor 7th, diminished, augmented) for each of the 12 different root notes, and one "no chord" label ($97 = 8 \times 12 + 1$).

5) Baseline methods: As baseline methods, simple methods were used to estimate the similarity and commonness.

The baseline methods used to estimate the similarity of vocal timbre, musical timbre, and rhythm calculated the Euclidean distance between mean feature vectors of two songs. In the baseline methods used to estimate the commonness of these elements, the mean feature vectors were calculated for a song-set were used to calculate the Euclidean distance from a target song. Each mean vector was normalized by subtracting the mean and dividing by the standard deviation.

To model chord progression, we used as a song model a unigram model trained by maximum likelihood estimation. The baseline modeling of chord progression of a set of musical pieces, the HPYLM n-gram model [23] as a song-set model (n is set to 1). To avoid the zero-frequency problem, chord similarity between two songs is also estimated by calculating weighted mean probabilities of the song model and the song-set model. The weights are $(1 - r)$ and r, respectively (r is set to 10^{-5}).

C. Experiment A: Similarity estimation

To evaluate musical similarity estimation based on probabilistic generative models, experiment A used all 3278 songs for modeling and the 463 songs by the artists listed in Table I (D_A). The 463 songs were sung by the twenty artists with the greatest number of songs in the modeling set. The evaluation set is very diverse: artists include solo singers and

[1]The trained GMMs were models in which the number of Gaussians was reduced, to 12 for the vocal GMM and to 27 for the non-vocal GMM.

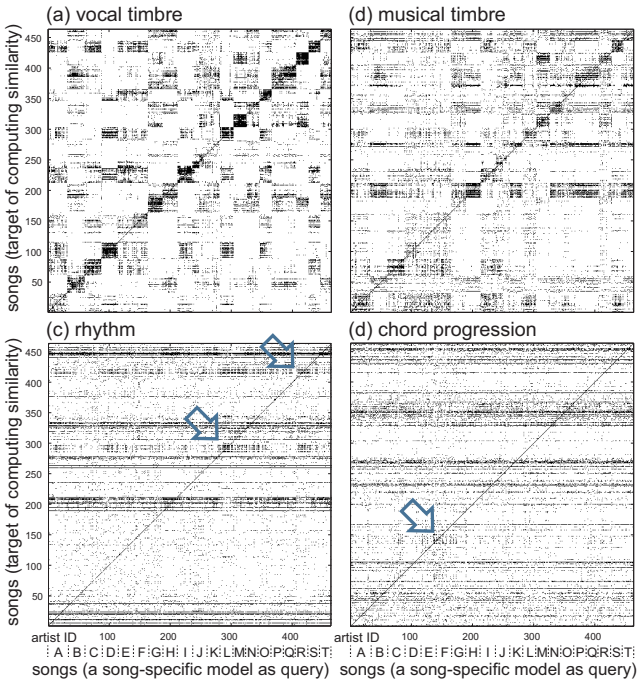

Figure 3. Similarities among all 463 songs.

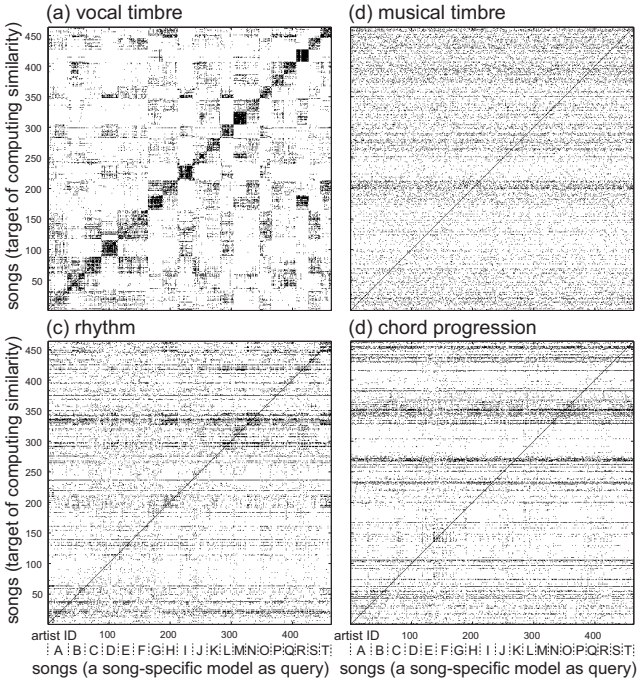

Figure 4. Baseline similarities among all 463 songs.

bands, and a balance of male and female vocalists (11 and 9, respectively).

1) Similarity matrix: We first estimated the similarities between the 463 songs with respect to the four musical elements. Figures 3 (a) through (d) show the similarity matrix for each of these elements, and Figure 4 shows the baseline results. In each figure the horizontal axis indicates query song models and the vertical axis indicates target songs of computing similarity.

The similarity matrix represents $214,369 = 463 \times 463$ pairs. In the matrices, only the 46 target songs (10% of D_A) having the highest similarities for each of the queries are colored black.

2) Comparing estimated similarities with human ratings: We next evaluated the song models by using human ratings. Twenty song pairs belonged to two groups, referred to as

Table II
THE TWENTY SONG PAIRS BELONG TO TWO GROUPS.

Group	Vocal timbre	Musical timbre	Rhythm	Chord progression
top10	L – O	B – A	K – G	D – I
	F – S	H – T	I – S	O – B
	J – I	Q – K	D – C	K – N
	A – D	M – R	E – Q	A – J
	B – Q	D – L	A – F	H – T
	M – R	S – I	O – H	P – C
	H – P	P – N	R – L	S – L
	E – C	O – G	N – T	F – E
	G – N	E – F	P – B	Q – G
	K – T	J – C	J – M	M – R
bottom10	F – E	G – J	P – O	O – P
	T – J	O – E	H – C	T – R
	H – D	C – B	G – S	B – M
	P – A	T – R	N – E	Q – N
	Q – L	Q – A	M – Q	J – A
	O – B	P – F	B – R	D – K
	G – S	I – M	K – F	S – G
	C – N	S – N	L – D	H – F
	M – K	H – L	T – J	I – C
	R – I	K – D	A – I	L – E

(L – O, for example, means a song of singer L and a song of singer O)

Table III
THE TWENTY SONG PAIRS BELONG TO TWO GROUPS (BASELINE).

Group	Vocal timbre	Musical timbre	Rhythm	Chord progression
top10	F – E	F – E	N – L	F – E
	A – C	M – I	O – E	M – I
	L – I	P – J	B – I	P – J
	N – K	B – S	R – G	B – S
	R – G	K – D	C – A	K – D
	T – Q	A – O	T – K	A – O
	B – H	H – N	S – P	H – N
	D – O	T – C	F – D	T – C
	M – J	G – R	M – H	G – R
	P – S	L – Q	Q – J	L – Q
bottom10	C – T	P – O	N – T	P – O
	H – S	S – T	L – H	S – T
	K – B	Q – B	O – B	Q – B
	P – F	H – D	E – A	H – D
	E – M	I – G	S – P	I – G
	N – D	M – K	C – F	M – K
	A – J	A – L	J – K	A – L
	L – Q	N – C	R – M	N – C
	G – I	R – E	G – I	R – E
	O – R	F – J	D – Q	F – J

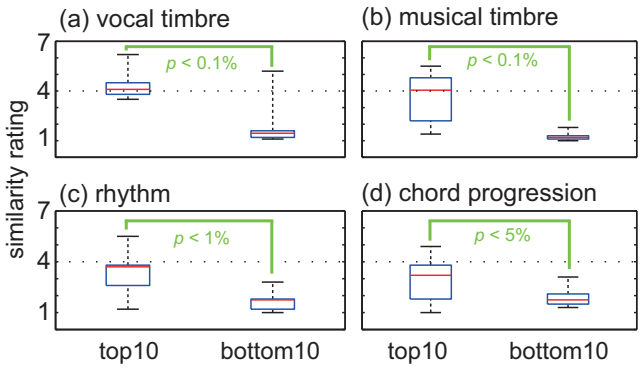

Figure 5. Box plots showing the statistics for the song-pair similarity ratings by a musician.

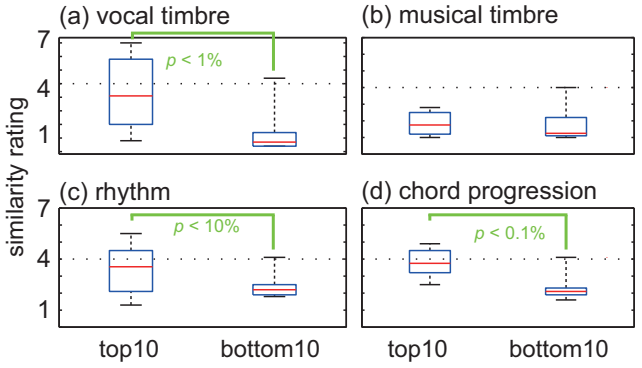

Figure 6. Box plots showing the statistics for the baseline song-pair similarity ratings.

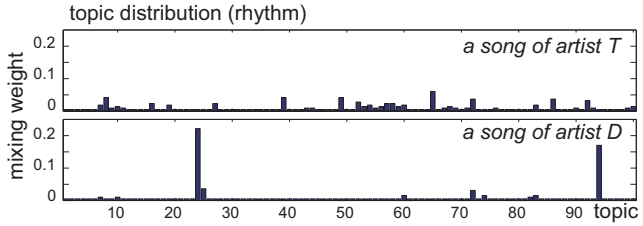

Figure 7. Top: topic distribution of a song that gets high similarity with most songs. Bottom: topic distribution of a song that gets low similarity with most songs.

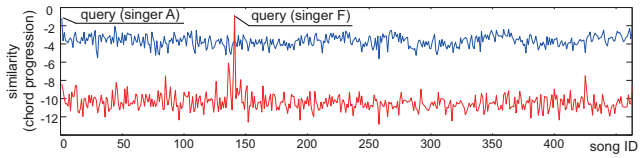

Figure 8. Blue: similarity between a query song and others with high similarity in most cases. Red: similarity between a query song and others with low similarity in most cases.

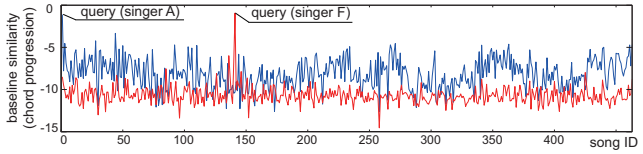

Figure 9. Blue: baseline similarity between a query song and others. Red: baseline similarity between a query song and others with low similarity in most cases.

the *top10* and *bottom10*. The top10 group included the ten song pairs having the highest similarities for each of the musical elements, under the selection restriction that there is no overlapping of singer names in the group. This means that this group comprises only pairs of songs sung by different singers. The bottom10 group includes the ten song pairs (also selected under the no-overlapping-name condition) having the lowest similarities for each of the musical elements. Table III shows the top10 and bottom10 groups.

A male musician who had experience with audio mixing/mastering and arrangement/composition of Japanese popular songs was asked to rate song-pair similarity on a 7-point scale ranging from 1 (not similar) to 7 (very similar). Rating to a precision of one decimal place (*e.g.*, 1.5) was allowed.

Figure 5 shows the results of the rating and Figure 6 shows the results of the rating based on the baseline results. The statistics of the ratings are shown by box plots indicating median values, 1/4 quantiles, 3/4 quantiles, minimum values, and maximum values. Testing the results by using Welch's *t*-test [24] revealed that the differences between the two groups were significant at the 0.1% level for vocal and musical timbre, the 1% level for rhythm, and the 5% level for chord progression (Figure 5).

3) Discussion: From the similarity matrix one sees that songs by the same artist have high similarity for vocal timbre and musical timbre. For rhythm and chord progression, on the other hand, some songs by the same artist have high similarity (indicated by arrows in Figures 3 (c) and (d)) but most do not. These results reflect musical characteristics qualitatively and can be understood intuitively.

On the similarity matrix for rhythm, horizontal lines can be seen. This means that there are songs that in most cases get high similarity regardless of which song is the query song. On the other hand, there are also songs that get low similarity with most query songs. LDA topic distributions for both kinds are shown in Figure 7. The former kind's is flat and has some topics having value, and the latter kind's has a few topics having value. On the similarity matrix for chord progression by using the trigram song models, there are query songs that get high similarity with all other songs (*e.g.*, a song of singer A) and there are query songs that get low similarity with all other (see, for example, Figure 8). In the baseline unigram setting, on the other hand, the query song of singer A has different similarities with all other songs (Figure 9).

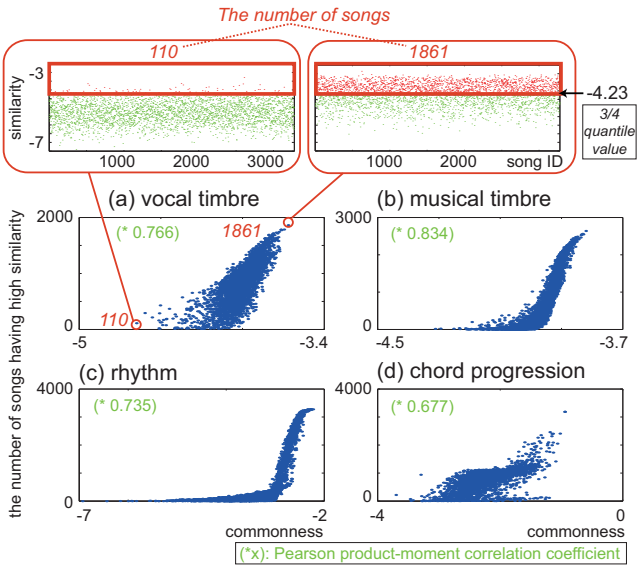

Figure 10. Relationships between estimated commonness of the four elements each song and the number of songs having high similarity with the song.

Table IV
PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENTS BETWEEN ESTIMATED COMMONNESS OF THE FOUR ELEMENTS EACH SONG AND THE NUMBER OF SONGS HAVING HIGH SIMILARITY WITH THE SONG.

element	correlation coefficients		
	condition	C	CB
vocal	S**	0.766	−0.175
musical timbre	S**	0.834	0.350
rhythm	S*	0.735	0.650
chord progression	S	0.670	0.886
vocal	SB*	0.137	0.960
musical timbre	SB	0.402	0.958
rhythm	SB	0.774	0.898
chord progression	SB**	0.759	0.846

Conditions
S: The number of songs having high similarity
SB: The number of songs having high similarity (baseline)
C: commonness
CB: commonness (baseline)

Estimated similarity is comparable to ratings by a musician
**: at the 0.1% significance level (Figure 5 and 6)
*: at the 1% significance level (Figure 5 and 6)

The comparison with the results of the expert ratings suggests that the proposed methods can estimate musical similarity appropriately. To improve the performance with regard to all elements, conditions such as those for extracting acoustic features, for quantization, for chord estimation, and for model training can be considered in future work.

D. Experiment B: Commonness estimation

To evaluate musical commonness estimation based on probabilistic generative models, experiment B also used the 3278 songs of the JPOP MDB to train the song-set models and for evaluating each musical element.

When evaluating the commonness estimation method, we first evaluated the number of songs having high similarity. For example, in Figure 1 the song a has many similar songs in the song set A. If a song having higher musical commonness and it is very similar to songs of a song set.

1) Relation between commonness and the number of songs having high similarity: Figure 10 shows the relationships comparing the estimated commonness of songs contained in the JPOP MDB to the number of songs having high similarity. We used as the threshold for deciding the similarity of an element to be high the 3/4 quantile value of all similarities among all possible song-pairs in the JPOP MDB (10,745,284 $= 3278 \times 3278$ song-pairs).

The Pearson product-moment correlation coefficients are also shown in each part of the figure, and Table IV. The reliability of the estimated similarity can be evaluated by using the results of Figure 5 and 6. The asterisk mark (*) and the double-asterisk mark (**) indicate the differences between the two groups (the top10 and bottom10 groups) significant at the 1% and 0.1% level, respectively.

Under conditions of the relatively reliable similarities ("vocal S**", "musical timbre S**", and "rhythm S*") the values of the correlation coefficient of the proposed method ("C": 0.766, 0.834, and 0.735) is bigger than the baseline method ("CB": −0.175, 0.350, and 0.650). The results suggest that the more similar a song is to songs of the song set, the higher its musical commonness in the proposed method. Although two coefficients of the condition "vocal SB*" and "chord progression SB**" are positive value ("C": 0.137 and 0.759), the value of the baseline method ("CB": 0.960 and 0.846) is bigger than those coefficients. The improvement of the correlation coefficients is a subject for future direction.

2) Application of commonness in terms of vocal timbre: Only the song-set models of vocal timbre can be evaluated quantitatively by using the singer's gender. These models are integrated song models with different ratios of the number of male singers to female singers.

To train song-set models, we used 14 songs by different solo singers (6 male and 8 female) from the JPOP MDB. We trained three types of song-set models: one trained by using all 14 songs, one trained by using one female song and all 6 male songs, and one is trained by using one male song and all 8 male songs.

Figure 11 shows the vocal timbre commonnesses based on the 3 different song-set models. When a model with a high proportion of female songs is used, the commonness of songs sung by females is higher than the commonness of songs sung by males (and vice versa). Figure 12 shows the statistics of the commonnesses are shown by box plots. The results suggest the commonnesses can be reflected vocal tract features.

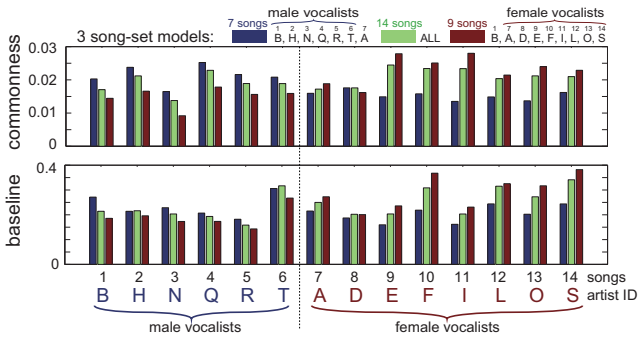

Figure 11. Vocal timbre commonness based on 3 different song-set models for 14 songs (6 male and 8 female).

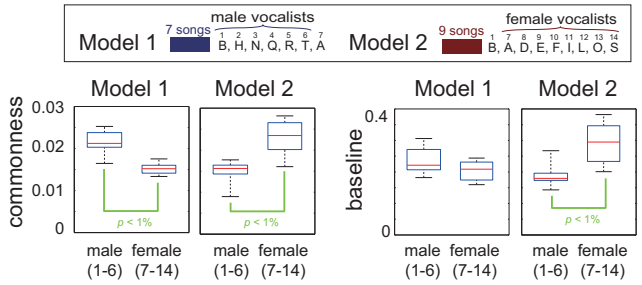

Figure 12. Box plots showing the statistics for the vocal timbre commonness (Figure 11).

IV. Conclusions and Future Work

This paper describes a musical similarity and commonness estimation method based on probabilistic generative models: LDA and the VPYLM. Four musical elements are modeled: vocal timbre, musical timbre, rhythm, and chord progression. The commonness can be estimated by using song-set models, which is easier than estimating the musical similarities of all possible pairs of songs.

The experimental results showed that our methods are appropriate for estimate musical similarity and commonness. The probability calculation can be applied not only to a musical piece but also to a part of a musical piece. This means that musical commonness is also useful to creators because a musical element that has high commonness is an established expression and can be used by anyone creating and publishing musical content (*e.g.*, a chord progression).

Since this paper focused on the above four elements, we plan to use melody (*e.g.*, F_0) as the next step. Future work will also include the integration of generative probabilities based on different models, calculating probabilities of parts of one song, investigating effective features, and developing an interface for music listening or creation by leveraging musical similarity and commonness.

Acknowledgment

This paper utilized the RWC Music Database (Popular Music). This work was supported in part by CREST, JST.

References

[1] M. Goto and K. Hirata, "Recent studies on music information processing," *Acoustical Science and Technology (edited by the Acoustical Society of Japan)*, vol. 25, no. 6, pp. 419–425, 2004.

[2] P. Knees and M. Schedl, "A survey of music similarity and recommendation from music context data," *ACM Trans. on Multimedia Computing, Communications and Applications*, vol. 10, no. 1, pp. 1–21, 2013.

[3] B. Pardo, Ed., *Special issue: Music information retrieval*, ser. Communications of the ACM, 2006, vol. 49, no. 8, pp. 28–58.

[4] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proc. IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[5] J. S. Downie, "The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research," *Acoust. Sci. & Tech.*, vol. 29, pp. 247–255, 2008.

[6] J. S. Downie, D. Byrd, and T. Crawford, "Ten years of ISMIR: Reflections on challenges and opportunities," in *Proc. ISMIR 2009*, 2009, pp. 13–18.

[7] Y. Song, S. Dixon, and M. Pearce, "A survey of music recommendation systems and future perspectives," in *Proc. CMMR 2012*, 2012, pp. 395–410.

[8] L. W. Barsalou, "Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories," *journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 11, no. 4, pp. 629–654, 1985.

[9] J.-J. Aucouturier and F. Pachet, "Music similarity measures: What's the use?" in *Proc. ISMIR 2002*, 2002, pp. 157–163.

[10] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity based music information retrieval," *IEEE Trans. on ASLP*, vol. 18, no. 3, pp. 638–648, 2010.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[12] D. Mochihashi and E. Sumita, "Infinite Markov model," in *Proc. NIPS2007*, 2007.

[13] K. Yoshii and M. Goto, "A vocabulary-free infinity-gram model for nonparametric bayesian chord progression analysis," in *Proc. ISMIR 2011*, 2014, pp. 645–650.

[14] T. Nakano, K. Yoshii, and M. Goto, "Vocal timbre analysis using latent Dirichlet allocation and cross-gender vocal timbre similarity," in *Proc. ICASSP 2014*, 2014, pp. 5239–5343.

[15] M. Goto, "A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[16] E. Pampalk, *Computational models of music similarity and their application to music information retrieval*. Ph.D. Dissertation, Vienna Inst. of Tech., 2006.

[17] E. Pampalk, A. Rauber, and D. Merkl, "Contentbased organization and visualization of music archives," in *Proc. ACMMM '02*, 2002, pp. 570–579.

[18] M. Mauch and M. Levy, "Structural change on multiple time scales as a correlate of musical complexity," in *Proc. ISMIR 2011*, 20011, pp. 489–494.

[19] M. Goto, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano, "Songle: A web service for active music listening improved by user contributions," in *Proc. ISMIR 2011*, 2011, pp. 311–316.

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

[21] C. M. Bishop, *Pattern recognition and machine learning*. Springer-Verlag New York, Inc., 2006.

[22] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proc. Natl. Acad. Sci. USA (PNAS)*, vol. 1, 2004, pp. 5228–5235.

[23] Y. W. Teh, "A hierarchical bayesian language model based on Pitman-Yor processes," in *Proc. COLING/ACL 2006*, 2006, pp. 985–992.

[24] B. L. Welch, "The significance of the difference between two means when the population variances are unequal," *Biometrika*, vol. 29, pp. 350–362, 1938.