International Journal of Semantic Computing
Vol. 10, No. 1 (2016) 27–52

DOI: 10.1142/S1793351X1640002X

World Scientific
www.worldscientific.com

Musical Similarity and Commonness Estimation Based on Probabilistic Generative Models of Musical Elements

Tomoyasu Nakano[*,‡], Kazuyoshi Yoshii[†,§] and Masataka Goto[*,¶]

[*]*National Institute of Advanced Industrial Science and Technology (AIST)*
Ibaraki 305-8568, Japan

[†]*Kyoto University*
Kyoto 606-8501, Japan
[‡]*t.nakano@aist.go.jp*
[§]*yoshii@kuis.kyoto-u.ac.jp*
[¶]*m.goto@aist.go.jp*

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 set of songs (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 two song-sets: 3278 Japanese popular music songs and 415 English songs.Twenty estimated song-pair similarities for each element and each song-set were compared with ratings by a musician. The comparison with the results of the expert ratings suggests that the proposed methods can estimate musical similarity appropriately. Estimated musical commonnesses are evaluated on basis of the Pearson product-moment correlation coefficients between the estimated commonness of each song and the number of songs having high similarity with the song. Results of commonness evaluation show that a song having higher commonness is similar to songs of a song set.

Keywords: Musical similarity; musical commonness; typicality; latent Dirichlet allocation; variational Pitman-Yor language model.

1. 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

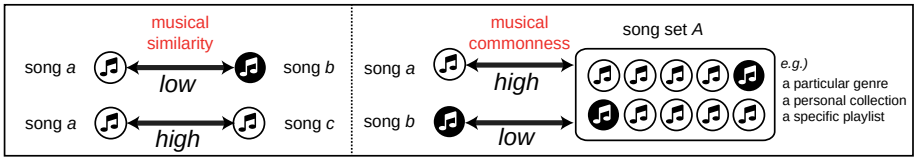


Fig. 1. Musical similarity and commonness.

obscurity. Musical similarity [1–3] 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 [1, 3–7] and music recommender systems [2, 8] 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 Fig. 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* [9]. Musical commonness can be used to recommend a representative or introductory song for a set of songs (e.g. songs released in the 80s), and it can help listeners understand the relationship between a song and such a song set.

To estimate musical similarity and commonness, we propose a generative modeling of four musical elements: vocal timbre, musical timbre, rhythm, and chord progression (Fig. 2). Previous works on music information retrieval have extracted various features[a] [1, 3, 5] including these four elements. We selected them to achieve diverse similarities and commonnesses via our estimation method. 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 mentioned/used to compute similarity between two songs [10, 11]. 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.

The following sections describe our approach and the experimental results of its evaluation. Section 2 presents acoustic features and probabilistic generative models and Sec. 3 describes estimation experiments and their evaluation. Section 4 considers our contribution in relation to previous works, Sec. 5 discusses the importance of musical commonness, and Sec. 6 concludes the paper, with directions for future work.

[a]For example, the following have all been used as features: singer voice, timbre, rhythm, onset, beat, tempo, melody, pitch, bass, harmony, tonality, chord, key, loudness, musical structure, and lyrics.

2. 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 estimate chord progression. We then model the timbres and rhythm by using a vector quantization method and latent Dirichlet allocation (LDA) [12]. 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) [13, 14].

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 (Fig. 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*).

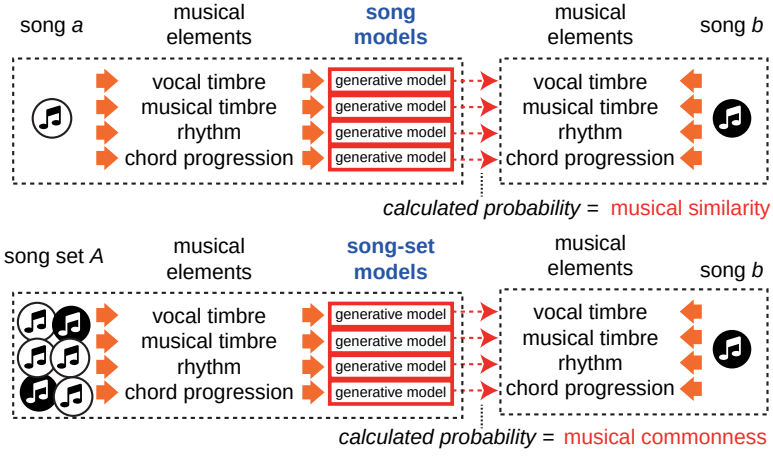


Fig. 2. Musical similarity and commonness based on probabilistic generative modeling of four musical elements: vocal timbre, musical timbre, rhythm, and chord progression.

2.1. *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 [15] on modeling vocal timbre. In addition, we propose a method to train song-set models under the LDA-based modeling.

2.1.1. *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 [11, 15]. In particular, the LPMCCs represent the characteristics of the singing voice well, since singer identification accuracy is greater when using LPMCCs than when using the standard mel-frequency cepstral coefficients (MFCCs) [11].

We first use Goto's PreFEst [16] 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 [11] for details). Feature vectors of only the reliable frames are used in the following processes (model training and probability estimation).

2.1.2. *Extracting acoustic features: Musical timbre*

We use mel-frequency cepstral coefficients (MFCCs) [17], their derivatives (ΔMFCCs), and Δpower to represent musical timbre, combining them as a feature vector. This combined feature vector is often used in speech recognition. The MFCCs are musical timbre features used in music information retrieval [18] and are robust to frame/hop sizes and lossy encoding provided that a minimum bitrate of approximately 160 Kbps is used [19].

2.1.3. *Extracting acoustic features: Rhythm*

To represent rhythm we use the fluctuation patterns (FPs) designed to describe the rhythmic signature of musical audio [18, 20]. They are features effective for music information retrieval [18] and for evaluating musical complexity with respect to tempo [21].

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 [18, 20] for details). Finally, the number of vector dimensions of the FPs was reduced by using principle component analysis (PCA).

2.1.4. *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*.

2.1.5. *Probabilistic generative model*

The observed data we consider for LDA are D independent songs $\vec{X} = \{\vec{X}_1, \ldots, \vec{X}_D\}$. A song $\vec{X}_d$ is N_d acoustic words $\vec{X}_d = \{\vec{x}_{d,1}, \ldots, \vec{x}_{d,N_d}\}$. 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 $\vec{Z}_d = \{\vec{z}_{d,1}, \ldots, \vec{z}_{d,N_d}\}$. 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 $\vec{Z} = \{\vec{Z}_1, \ldots, \vec{Z}_D\}$.

The full joint distribution of the LDA model is given by

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

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

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

and

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

We then introduce conjugate priors as follows:

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

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

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.

2.1.6. *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:

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

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

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 $\vec{x}_{b,n}$.

2.1.7. *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 $\vec{\alpha}_d = \{\alpha_{d,1}, \ldots, \alpha_{d,K}\}$. This model $\vec{\alpha}_A$ is defined as follows:

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

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: $\log p_g(a|A)$.

2.2. *Similarity and commonness: Chord progression*

We first estimate key and chord progression by using modules of Songle [22], 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/).

2.2.1. *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.

2.2.2. *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 (with r set to 10^{-5}).

3. Experiments

The proposed methods were tested in experiments evaluating the estimated similarity (Experiments A1 and A2) and the estimated commonness (Experiments B1 and B2).

3.1. *Dataset*

The song set used for model training, similarity estimation, and commonness estimation comprised 3278 Japanese popular songs[b] 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 (JPOP MDB). The twenty artists focused on for similarity evaluation are listed in Table 1.

Another song set used for model training, similarity estimation, and commonness estimation comprised 415 English songs performed by various types of artists (solo singers, male/female singers, bands, or groups). They were taken from commercial music CDs (*Billboard Top Rock 'n 'Roll Hits* 1968–1972, *Billboard Top Hits* 1975–1989, and *GRAMMY NOMINEES* 1996–2005). Here we refer to this song set as the English music database (ENG MDB). The twenty artists focused on for similarity evaluation are listed in Table 2.

The song set used for GMM/k-means/PCA training to extract the acoustic features consisted of 100 popular songs from the RWC Music Database (RWC-MDB-P-2001) [23]. These 80 songs 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 to this song set as the RWC MDB.

3.2. *Experimental settings*

Conditions and parameters of the methods described in Sec. 2 are described here in detail.

[b] Note that some are English songs in them.

Table 1. Singers of the 463 songs used in the experiments A1 and B1.

ID	Artist name	Gender of vocalist(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		11 male. 9 female	463

Table 2. Singers of the 62 songs used in experiments A2 and B2.

ID	Artist name	Gender of vocalist(s) (* more than one singer)	Number of songs
BO	Billy Ocean	male	4
ST	Sting	male	4
U2	U2	male	4
BB	Backstreet Boys	male*	3
BL	Blondie	female	3
BS	Britney Spears	female	3
CA	Christina Aguiler	female	3
DJ	Daryl Hall & John Oates	male*	3
EJ	Elton John	male	3
EM	Eminem	male	3
EC	Eric Clapton	male	3
KT	KC & The Sunshine Band	male*	3
PS	Pointer Sisters	female*	3
RK	R. Kelly	male	3
RM	Richard Marx	male	3
SC	Sheryl Crow	female	3
SS	Starship	male*	3
TH	Three Dog Night	male*	3
TT	Tommy James & The Shondells	male	3
AC	Ace Of Base	female*	2
Total		14 male. 6 female	62

3.2.1. *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 analysis frame length was 32 ms. To estimate the features, the vocal sound was re-synthesized by using a sinusoidal model with the frequency and amplitude of the lth overtone ($l = 1, \ldots, 20$). The ΔF_0 was calculated every five frames (50 ms), the order of LPC analysis was 25, the number of Mel-scaled filter banks was 15.

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 [24]. We set the number of Gaussians to 32 and set the hyperparameter of a Dirichlet distribution over the mixing coefficients to 1.0. 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.

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 Δ features were calculated every five frames (50 ms), the pre-emphasis coefficients for was 0.97, the number of Mel-scaled filter banks was 15, and the cepstral liftering coefficient was 22. 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 [18]. 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 work [15, 18].

3.2.2. *Quantization*

To quantize the vocal features, we set the number of clusters of the k-means algorithm to 100 and used the 100 songs of the RWC MDB to train the centroids. This k is same number used in our previous work [15]. The number of clusters used to quantize the musical timbre and rhythm features was set to 64 in this evaluation.

3.2.3. *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 [22] for details).

3.2.4. *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 and all of the 415 recordings of the ENG MDB.

The number of topics K was set to 100, and the model parameters of LDA were trained using the collapsed Gibbs sampler [25]. 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 [15].

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 ($97 = 8 \times 12 + 1$).

3.2.5. *Baseline methods*

The baseline methods used to estimate similarity and commonness were simple methods.

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 and 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.

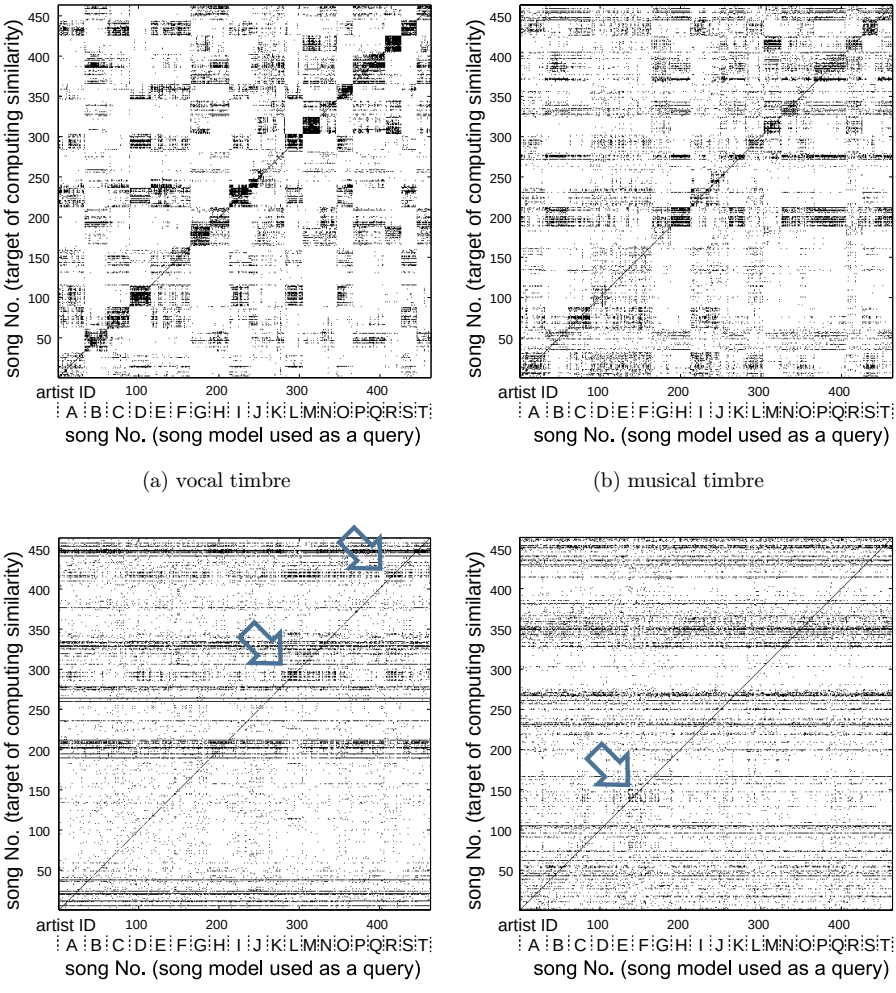
To model chord progression, we used as a song model a unigram model trained by maximum likelihood estimation. The baseline modeling of chord progression of a set of musical pieces, used as a song-set model the *HPYLM* n-gram model [26] (with n set to 1). To avoid the zero-frequency problem, chord similarity between two songs was also estimated by calculating weighted mean probabilities of the song model and the song-set model. The weights were $(1 - r)$ and r, respectively (with r set to 10^{-5}).

3.3. *Experiment A1: Similarity estimation (JPOP MDB)*

To evaluate musical similarity estimation based on probabilistic generative models, experiment A1 used all 3278 songs for modeling and estimated the similarities of the 463 songs by the artists listed in Table 1 ($D_{A1} = 463$). Those 463 songs were sung by the twenty artists with the greatest number of songs in the modeling set. The evaluation set was very diverse: artists include solo singers and bands, and a balance of male and female vocalists.

3.3.1. *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 Fig. 4 shows the baseline results. In each figure the horizontal



(a) vocal timbre (b) musical timbre

(c) rhythm (d) chord progression

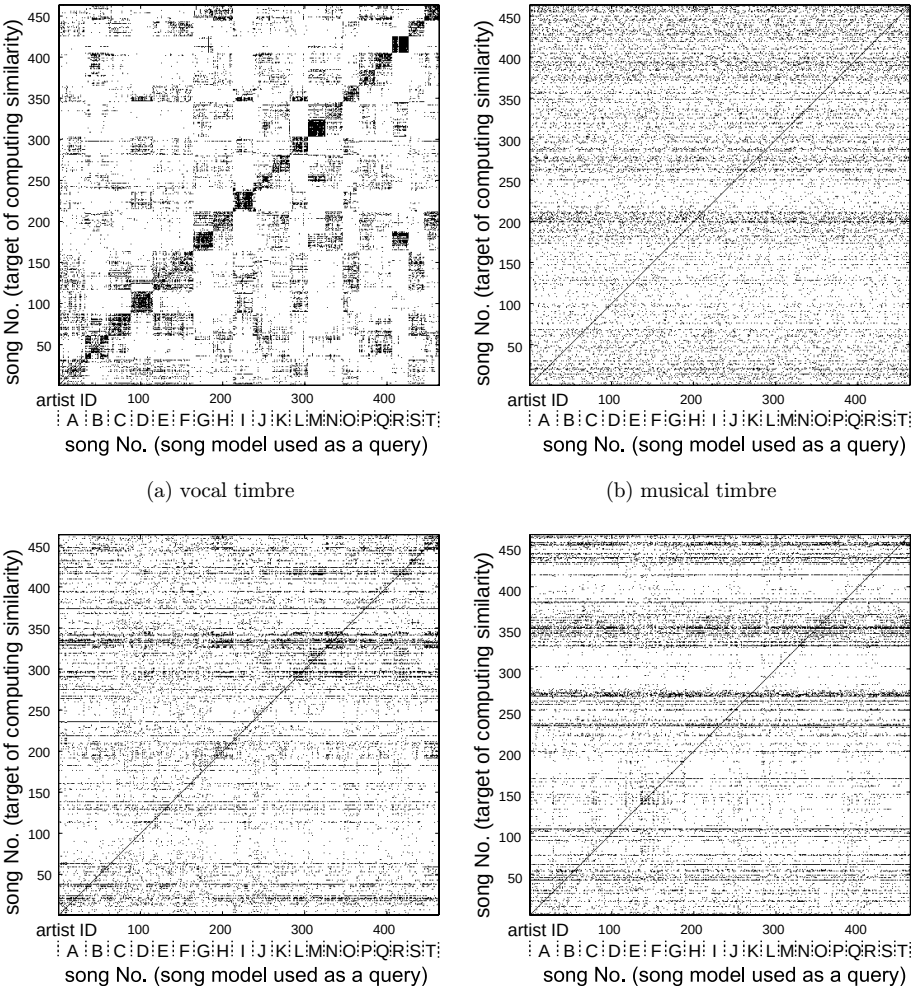
Fig. 3. Similarities among all 463 songs by the artists listed in Table 1 (similarities estimated using probabilistic generative models).

axis indicates song number (song model used as a query) and the vertical axis indicates target song number for similarity computation.

A similarity matrix represents 214,369 (463 × 463) pairs, and in each of the matrices only the 46 target songs (10% of D_{A1}) having the highest similarities for each of the queries are colored black.

3.3.2. *Comparing estimated similarities with expert human ratings*

We next evaluated the song models by using expert ratings. Twenty song pairs belonged to two groups, referred to as the *top10* and *bottom10*. The top10 group



(a) vocal timbre

(b) musical timbre

(c) rhythm

(d) chord progression

Fig. 4. Baseline similarities among all 463 songs by the artists listed in Table 1.

included the ten song pairs having the highest similarities for each of the musical elements, under the selection restriction that there was no overlapping of singer names in the group. This means that this group comprised only pairs of songs sung by different singers. The bottom10 group included the ten song pairs (also selected under the no-overlapping-name condition) having the lowest similarities for each of the musical elements. Table 3 shows the top10 and bottom10 groups based on the similarity estimated using the proposed methods and the baseline methods.

A music expert (a male musician) who was professionally-trained for music at his graduating school and had experience with audio mixing/mastering, writing lyrics, and arrangement/composition of Japanese popular songs was asked to rate song-pair

Table 3. The twenty song pairs belonged to two groups: Experiment A1 (JPOP MDB).

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

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

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 by the musician, and Fig. 6 shows the results of 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 [27] 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 (Fig. 5).

3.4. *Experiment A2: Similarity estimation (ENG MDB)*

To evaluate musical similarity estimation based on probabilistic generative models, experiment A2 used all 415 songs for modeling and estimated the similarities of the songs by the artists listed in Table 2 ($D_{A2} = 62$). Those 62 songs were sung by the twenty artists with the greatest number of songs in the modeling set.

We evaluated the song models by using expert ratings. As in experiment A1, twenty song pairs belonged to two groups, referred to as the *top10* and *bottom10*. Table 4 shows the top10 and bottom10 groups based on the similarity estimated

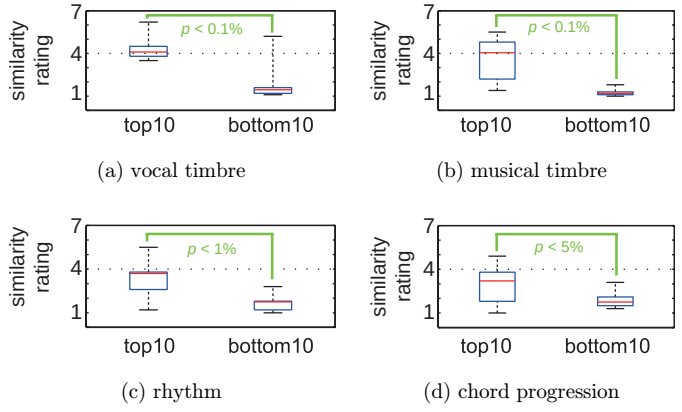


Fig. 5. Box plots showing the statistics for the song-pair similarity ratings by a musician: Experiment A1 (JPOP MDB).

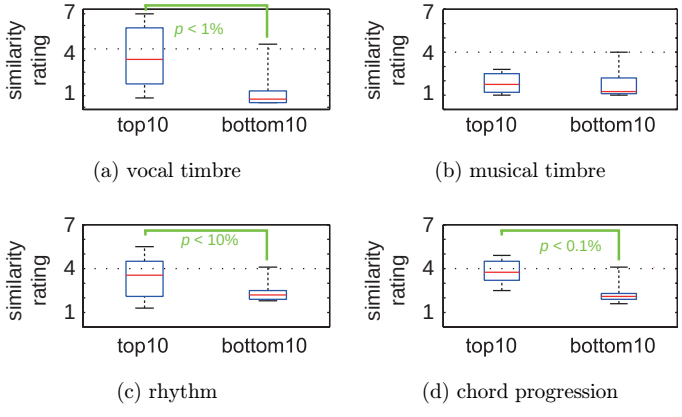


Fig. 6. Box plots showing the statistics for the baseline song-pair similarity ratings: Experiment A1 (JPOP MDB, baseline).

using the proposed methods and the baseline methods. Figure 7 shows the results of the rating by the musician, and Fig. 8 shows the results of rating based on the baseline results.

3.5. *Discussion for experiments A1 and A2*

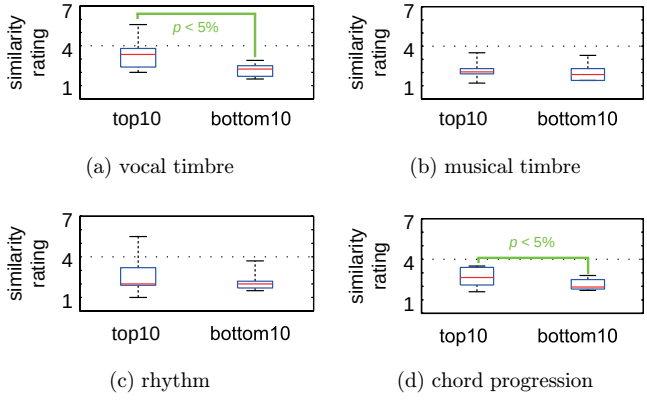
From the similarity matrices for the JPOP MDB one sees that songs by the same artist have high similarity for vocal timbre and musical timbre. For rhythm and chord progression, on the other hand, some songs by the same artist have high similarity (indicated by arrows in Figs. 3(c) and 3(d)) but most do not. These results reflect musical characteristics qualitatively and can be understood intuitively. Although the similarity matrices for the ENG MDB are not shown, they indicated a similar tendency.

Table 4. The twenty song pairs belonged to two groups: Experiment A2 (ENG MDB).

Proposed method				Baseline method			
Vocal timbre	Musical timbre	Rhythm	Chord progression	Vocal timbre	Musical timbre	Rhythm	Chord progression
top10							
EC – RK	BB – RM	RM – ST	PS – BL	BB – SC	EJ – U2	RM – ST	EC – BL
BL – TH	ST – EJ	BO – BB	U2 – TH	EJ – SS	DJ – KT	EC – KT	SC – TH
EM – EJ	U2 – TT	KT – U2	EC – SC	BO – KT	BO – TH	CA – RK	PS – TT
BB – TT	SS – DJ	EJ – CA	AC – SS	EC – RK	EC – SS	SS – U2	AC – SS
KT – DJ	SC – RK	DJ – EC	BB – KT	BL – DJ	BL – CA	EJ – TH	EJ – KT
PS – BS	BO – TH	AC – BS	RM – BO	PS – U2	RK – SC	DJ – TT	RM – BO
CA – AC	BS – CA	TH – PS	DJ – TT	AC – CA	PS – ST	BB – SC	BS – CA
U2 – ST	PS – BL	SC – RK	CA – EJ	BS – TT	RM – TT	BL – PS	ST – BB
SS – SC	EC – EM	SS – TT	RK – ST	RM – ST	BS – EM	AC – BS	EM – U2
BO – RM	AC – KT	EM – BL	EM – BS	EM – TH	AC – BB	BO – EM	RK – DJ
bottom10							
BS – BB	EJ – BB	RM – TT	SC – TH	SS – BS	SS – SC	RM – RK	RM – PS
U2 – SS	DJ – SC	TH – SC	RM – U2	PS – BB	RM – EJ	TT – EJ	TH – BO
SC – KT	TH – KT	RK – BL	CA – BB	SC – BO	TT – BB	SS – SC	SC – EC
TH – CA	RK – BO	SS – PS	BL – BO	KT – EJ	RK – PS	U2 – ST	DJ – BS
EM – BO	ST – PS	EJ – BO	EC – AC	U2 – CA	KT – EM	TH – BL	U2 – BB
TT – EC	U2 – BS	DJ – U2	ST – PS	EM – AC	EC – BO	KT – CA	BL – CA
RM – AC	BL – EC	EC – CA	SS – BS	TH – EC	TH – BL	EC – BO	SS – ST
EJ – PS	TT – EM	BB – BS	DJ – EM	ST – BL	ST – BS	PS – BB	AC – EM
DJ – ST	RM – CA	KT – AC	KT – RK	RK – DJ	U2 – DJ	DJ – BS	EJ – TT
RK – BL	SS – AC	ST – EM	EJ – TT	TT – RM	CA – AC	EM – AC	KT – RK

EC – RK, for example, means a song of singer EC (Eric Clapton) and a song of singer RK (R. Kelly)

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 Fig. 9. The former



(a) vocal timbre (b) musical timbre

(c) rhythm (d) chord progression

Fig. 7. Box plots showing the statistics for the song-pair similarity ratings by a musician: Experiment A2 (ENG MDB).

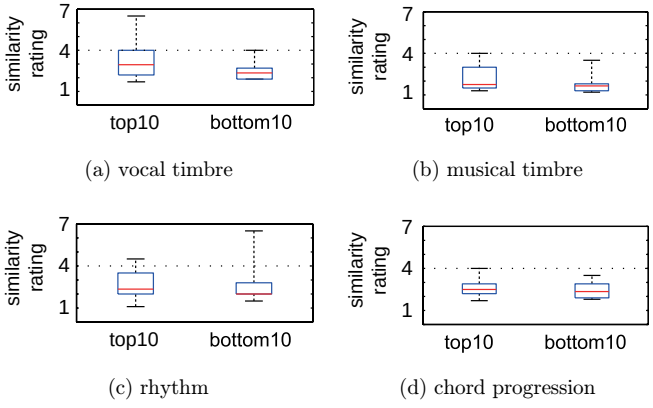


Fig. 8. Box plots showing the statistics for the baseline song-pair similarity ratings: Experiment A2 (ENG MDB).

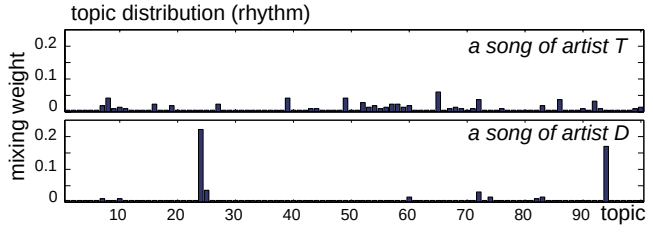


Fig. 9. Top: topic distribution of a song that gets high similarity with most songs. Bottom: topic distribution of a song that gets low similarity with most songs.

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, 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 songs (see, e.g. Fig. 10: Top). In the baseline unigram setting, on the other hand, the query song of singer A has different similarities with all other songs (Fig. 10: Bottom).

The comparison with the results of the expert ratings suggests that the proposed methods can estimate musical similarity appropriately. The musician was asked for the judgment (evaluation) criteria after the all ratings, and they were as follows:

vocal timbre (1) ringing based on the distribution of the harmonic overtone, (2) pitch (F_0, fundamental frequency), (3) degree of breathy voice.

musical timbre (1) composition of the musical instruments, (2) balance of loudness of each instruments, reverberation, and dynamics via the audio mixing/mastering. (3) music genre,

rhythm (1) rhythm pattern, (2) beat structure or degree of shuffle (swing), (3) music genre, (4) tempo.

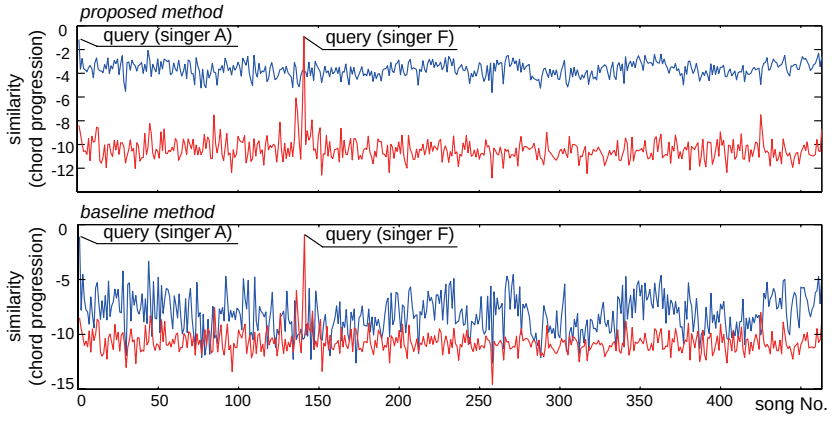


Fig. 10. Blue: similarity between a query song and others. Red: similarity between a query song and others with low similarity in most cases.

chord progression (1) the pattern of chord progression, (2) chords used in the songs.

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. Especially, in the results of the comparison of the estimated similarity with the expert ratings for the ENG MDB, there is no significant difference for musical timbre and rhythm (Fig. 7). The musician said that there are differences for the musical timbre and rhythm among the songs because of the released date of songs are wide-ranging (1968–2005).

3.6. *Experiment B1: Commonness estimation (JPOP MDB)*

To evaluate musical commonness estimation based on probabilistic generative models, experiment B1 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 Fig. 1 the song a has many similar songs in the song set A. If a song having higher (lower) commonness is very similar (is not similar) to songs of a song set.

Figure 11 shows the relationships between 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 10,745,284 (3278×3278) possible song-pairs in the JPOP MDB.

The Pearson product-moment correlation coefficients are shown in each part of the figure and are also listed in Table 5. The reliability of the estimated similarity can be evaluated by using the results shown in Figs. 5 and 6. The asterisk mark (*) and

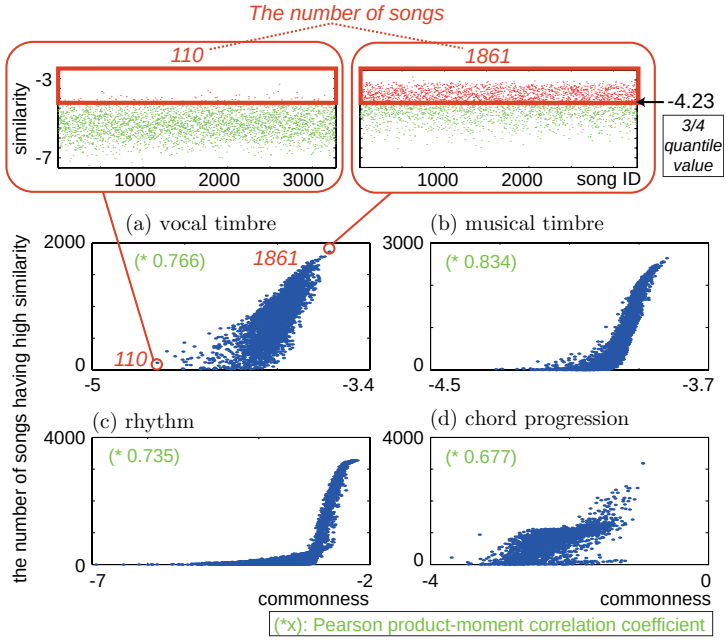


Fig. 11. Relationships between estimated commonness of the four elements of each song and the number of songs having high similarity with the song: Experiment B1 (JPOP MDB).

Table 5. 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: Experiment B1 (JPOP MDB). Conditions: S) The number of songs having high similarity, SB) The number of songs having high similarity (baseline), C) Commonness, CB) Commonness (baseline).

		Correlation coefficients	
Element	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

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

the double-asterisk mark (**) in Table 5 indicate differences between the top10 and bottom10 groups that are significant at the 1% and 0.1% levels, respectively.

Under conditions of the relatively reliable similarities ("vocal S**", "musical timbre S**", and "rhythm S*") the correlation coefficient of the proposed method ("C": 0.766, 0.834, and 0.735) are larger than those of 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 values ("C": 0.137 and 0.759), the corresponding coefficients for the baseline method ("CB": 0.960 and 0.846) are larger. The improvement of the correlation coefficients is a subject for future investigation.

3.7. *Experiment B2: Commonness estimation (ENG MDB)*

To evaluate musical commonness estimation based on probabilistic generative models, experiment B2 also used the 415 songs of the ENG MDB to train the song-set models and for evaluating each musical element. As in experiment B1, we evaluated the number of songs having high similarity.

Figure 12 shows the relationships between the estimated commonness of songs contained in the ENG 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 172,223 (415 × 415) possible song-pairs in the ENG MDB.

The Pearson product-moment correlation coefficients are shown in each part of the figure and are also listed in Table 6. The reliability of the estimated similarity can be evaluated by using the results shown in Figs. 7 and 8. The asterisks (*) in Table 6

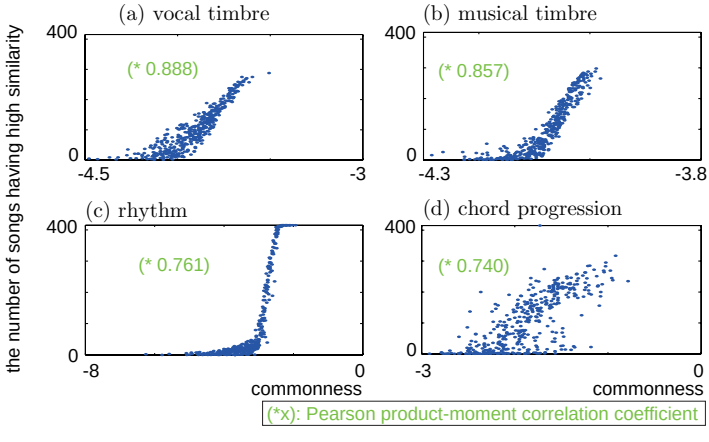


Fig. 12. Relationships between estimated commonness of the four elements each song and the number of songs having high similarity with the song: Experiment B2 (ENG MDB).

Table 6. Pearson product-moment correlation coefficients between estimated commonness of the four elements of each song and the number of songs having high similarity with the song: Experiment B2 (ENG MDB). Conditions: S) The number of songs having high similarity, SB) The number of songs having high similarity (baseline), C) Commonness, CB) Commonness (baseline).

		Correlation coefficients	
Element	Condition	C	CB
vocal	S*	0.888	−0.050
musical timbre	S	0.857	−0.421
rhythm	S	0.761	−0.820
chord progression	S*	0.740	0.707
vocal	SB	−0.172	0.959
musical timbre	SB	0.104	0.077
rhythm	SB	0.004	0.070
chord progression	SB	0.891	0.908

*Estimated similarity is comparable to ratings by a musician at the 5% significance level (Figs. 7 and 8).

indicate differences between the top10 and bottom10 groups that are significant at the 5% level.

Under conditions of the relatively reliable similarities ("vocal S*" and "chord progression S*") the values of the correlation coefficient of the proposed method ("C": 0.888 and 0.740) are bigger than the baseline method ("CB": −0.050 and 0.707). Moreover, the coefficients based on the proposed methods (row "S" and column "C") are all high (greater than 0.7). The results suggest that the similarity and commonness estimated using the proposed methods have a mutual relationship. This relationship is useful to use the commonness under the typicality definition. In other words, the commonness can be used instead of the number of songs having high similarity among the songs.

3.8. *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 13 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). In Fig. 14 the statistics of the commonnesses are shown by box plots. The results suggest the commonnesses can reflect vocal tract features.

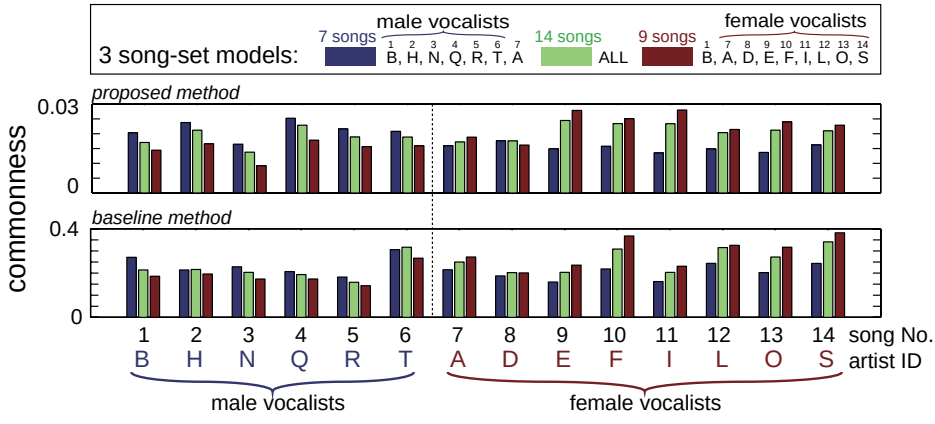


Fig. 13. Vocal timbre commonness based on 3 different song-set models for 14 songs (6 male and female).

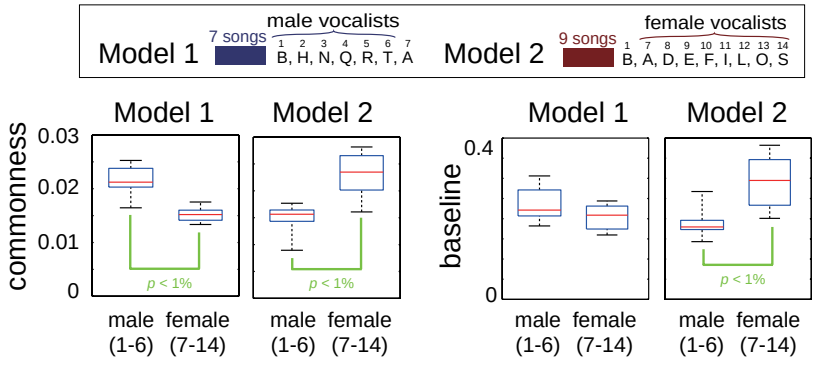


Fig. 14. Box plots showing the statistics for the vocal timbre commonness (Fig. 13).

3.9. *Applying the proposed method to other elements*

The proposed LDA-based method and the VPYLM-base method can be applied to various music-related characteristics. Lyrics, for example, are an important element of music in various genres, especially in popular music. Since the LDA and the VPYLM were originally proposed for text analysis, they can be used for lyrics modeling. In fact, there are three papers on work that used lyrics for LDA-based music retrieval [28–30].

Figure 15 shows that the results of the expert rating comparing with LDA-based estimated similarities and correlation coefficients between estimated commonness of the lyrics each song and the number of songs having high similarity with the song. This results are based on a set of lyrics of 1996 songs: 1896 Japanese popular lyrics a part of the JPOP MDB and 100 lyrics of the RWC MDB. 340 lyrics of the twenty artists with the greatest number of lyrics in the lyrics set are used to select the twenty

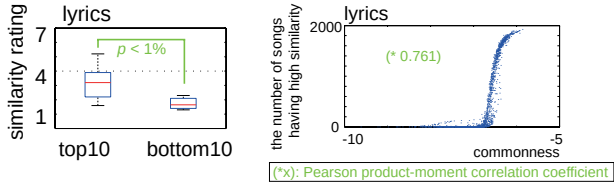


Fig. 15. Relationships between estimated commonness of the lyrics each song and the number of songs having high similarity with the song.

lyrics pairs, the *top10* and *bottom10*. The number of topics K was set to 100, and MeCab [31] was used for the morphological analysis of Japanese lyrics. 19,390 words (morphemes) is the vocabulary size in the 1996 lyrics. The results suggest that the proposed method can be applied to music lyrics.

As other characteristics, artist properties provide different type of relations among songs. The artist-level information obtained from web such as Wikipedia (`http://www.wikipedia.org/`) and its commonness (typicality) can be used to visualize the relations [32].

4. Related Studies

Musical similarity is a central concept of MIR and is also important for purposes other than retrieval. For example, the use of similarity to automatically classify musical pieces (into genres, music styles, etc.) is being researched [1, 2], and musical similarity can also be used for music auto-tagging [3]. However, each of these applications is different from the idea of musical commonness: *musical similarity* is usually defined by comparing two songs, *music classification* is defined by classifying a given song into one out of a set of categories (category models, centroids, etc.), and *music auto-tagging* is defined by comparing a given song to a set of tags (tag models, the closest neighbors, etc.). To the best of our knowledge, there is no research about the automatic estimation of *musical commonness*, defined as the typicality of a song with respect to a set of songs. Therefore, we think that music commonness is a novel concept which can be used to retrieve representative songs from a set of songs.

This paper has proposed a unified framework of probabilistic generative modeling to estimate musical similarity and commonness. To realize the framework, we have introduced latent analysis of music. There are previous works related to latent analysis of music, such as music retrieval based on LDA of lyrics and melodic features [28], lyrics retrieval based on LDA [29], assessing quality of lyrics topic model (LDA) [30], chord estimation based on LDA [33, 34], combining document and music spaces by latent semantic analysis [35], music recommendation by social tag and latent semantic analysis [36], and music similarity based on the hierarchical Dirichlet process [37]. In contrast to these previous reports, we showed that LDA and VPYLM can be combined to do musical similarity and commonness estimation using four musical elements (vocal timbre, musical timbre, rhythm, and chord progression) and lyrics.

5. Discussion

The contributions of this paper are 1) proposing the concept of musical commonness, 2) showing that a generative model trained from a song set can be used for commonness estimation (instead of estimating musical similarities of all possible song-pairs by using a model trained from each song), 3) showing how to evaluate the estimated commonness.

Described as in Sec. 1, the amount of digital content that can be accessed by people has been increasing and will continue to do so in the future. This is desirable but unfortunately makes it easier for the work of content creators to become buried within a huge amount of content, making it harder for viewers and listeners to select content. Furthermore, since the amount of similar content is also increasing, creators will be more concerned that their content might invite unwarranted suspicion of plagiarism. All kinds of works are influenced by existing content, and it is difficult to avoid the unconscious creation of content partly similar in some way to prior content.

However, human ability with regard to similarity is limited. Judging similarity between two songs one hears is a relatively simple task but takes time. One simply does not have enough time to search a million songs for similar content. Moreover, while humans are able to make accurate judgments based on past experience, their ability to judge "commonness" or "typicality" — the probability of an event's occurrence — is limited. When an uncommon event happens to be frequently observed recently, for example, people tend to wrongly assume that it is likely to occur. And when a frequent event happens to not be encountered, people tend to wrongly assume that it is rare. Consequently, with the coming of an "age of billions of creators" in which anyone can enjoy creating and sharing works, the monotonic increase in content means that there is a growing risk that one's work will be denounced as being similar to someone else's. This could make it difficult for people to freely create and share content.

The musical commonness proposed in this paper can help create an environment in which specialists and general users alike can know the answers to the questions "What is similar here?" and "How often does this occur?" Here we aim to make it possible for people to continue creating and sharing songs without worry. Furthermore, we want to make it easy for anyone to enjoy the music content creation process, and we want to do this by developing music-creation support technology enabling "high commonness" elements (such as chord progressions and conventional genre-dependent practices) to be used as knowledge common to mankind. We also want to promote a proactive approach to encountering and appreciating content by developing music-appreciation support technology that enables people to encounter new content in ways based on its similarity to other content.

We hope to contribute to the creation of a culture that can mutually coexist with past content while paying appropriate respect to it. This will become possible by supporting a new music culture that enables creators to take delight in finding their

content being reused, in much the same way that researchers take delight in finding their articles being cited. We feel that the value of content cannot be measured by the extent to which it is not similar to other content and that pursuing originality at all costs does not necessarily bring joy to people. Fundamentally, content has value by inducing an emotional and joyous response in people. We would like to make it a matter of common sense that content with emotional appeal and high-quality form has value. In fact, we would like to see conditions in which it is exactly the referring to many works that gives content its value, similar to the situation with academic papers. Through this approach, we aim to create a content culture that emphasizes emotionally touching experiences.

6. Conclusions and Future Work

This paper describes an approach to musical similarity and commonness estimation that is 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. And these methods are potentially applicable with other elements of music, such as lyrics. 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 (e.g. a chord progression) is an established expression and can be used by anyone creating and publishing musical content.

This paper showed the effectiveness of the proposed methods with song sets of different sizes. The JPOP MDB was used as a large song set (more than 1000 songs) and the ENG MDB was used as a medium-size song set (between 100 and 1000 songs). In [32] we used musical commonness for visualization and changing playback order with a small song set (less than 100 songs) as a personal music playlist. And the experimental results in Sec. 3.8 also show the effectiveness of the musical commonness with a small song set.

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 ofs 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.

Acknowledgments

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* **25**(6) (2004) 419–425.

[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* **10**(1) (2013) 1–21.

[3] M. Schedl, E. Gómez and J. Urbano, Music information retrieval: Recent developments and applications, *Foundations and Trends in Information Retrieval* **8**(2–3) (2014) 127–261.

[4] B. Pardo, Ed., Special issue: Music information retrieval, *Commun. ACM* **49**(8) (2006) 28–58.

[5] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes and M. Slaney, Content-based music information retrieval: Current directions and future challenges, in *Proc. IEEE* **96** (4) (2008) 668–696.

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

[7] J. S. Downie, D. Byrd and T. Crawford, Ten years of ISMIR: Reflections on challenges and opportunities, in *Proc. 10th Int. Society for Music Inf. Retrieval Conf.*, 2009, pp. 13–18.

[8] Y. Song, S. Dixon and M. Pearce, A survey of music recommendation systems and future perspectives, in *Proc. 9th Int. Symp. on Computer Music Modeling and Retrieval*, 2012, pp. 395–410.

[9] 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* **11**(4) (1985) 629–654.

[10] J.-J. Aucouturier and F. Pachet, Music similarity measures: What's the use? in *Proc. 3rd Int. Conf. Music Inf. Retrieval*, 2002, pp. 157–163.

[11] 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 Audio, Speech, and Language Processing* **18**(3) (2010) 638–648.

[12] D. M. Blei, A. Y. Ng and M. I. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research* **3** (2003) 993–1022.

[13] D. Mochihashi and E. Sumita, The infinite Markov model, in *Proc. Advances in Neural Information Processing Systems 20*, 2007, pp. 1017–1024.

[14] K. Yoshii and M. Goto, A vocabulary-free infinity-gram model for nonparametric bayesian chord progression analysis, in *Proc. 12th Int. Society for Music Inf. Retrieval Conf.*, 2014, pp. 645–650.

[15] T. Nakano, K. Yoshii and M. Goto, Vocal timbre analysis using latent Dirichlet allocation and cross-gender vocal timbre similarity, in *Proc. 39th Int. Conf. on Acoustics, Speech and Signal Processing*, 2014, pp. 5239–5343.

[16] 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* **43**(4) (2004) 311–329.

[17] B. Logan, Mel frequency cepstral coefficients for music modeling, in *Proc. Int. Symp. Music Inf. Retrieval*, 2000, pp. 1–11.

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

[19] J. Urbano, D. Bogdanov, P. Herrera, E. Gómez and X. Serra, What is the effect of audio quality on the robustness of MFCCs and chroma features? in *Proc. 15th Int. Society for Music Inf. Retrieval Conf.*, 2014, pp. 573–578.
[20] E. Pampalk, A. Rauber and D. Merkl, Contentbased organization and visualization of music archives, in *Proc. ACM Multimedia*, 2002, pp. 570–579.
[21] M. Mauch and M. Levy, Structural change on multiple time scales as a correlate of musical complexity, in *Proc. ISMIR*, 2011, pp. 489–494.
[22] 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. 12th Int. Society for Music Inf. Retrieval Conf.*, 2011, pp. 311–316.
[23] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, RWC music database: Popular, classical, and jazz music databases, in *Proc. 3rd Int. Conf. Music Inf. Retrieval*, 2002, pp. 287–288.
[24] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer-Verlag, New York, 2006).
[25] T. L. Griffiths and M. Steyvers, Finding scientific topics, in *Proc. Natl. Acad. Sci. USA*, Vol. 1 (2004), pp. 5228–5235.
[26] Y. W. Teh, A hierarchical bayesian language model based on Pitman-Yor processes, in *Proc. Joint Conf. of the Int. Committee on Computational Linguistics and the Association for Computational Linguistics*, 2006, pp. 985–992.
[27] B. L. Welch, The significance of the difference between two means when the population variances are unequal, *Biometrika* **29** (1938) 350–362.
[28] E. Brochu and N. de Freitas, "name that song!": A probabilistic approach to querying on music and text, in *Proc. of Advances in Neural Information Processing Systems*, 2002, pp. 1505–1512.
[29] S. Sasaki, K. Yoshii, T. Nakano, M. Goto and S. Morishima, LyricsRadar: A lyrics retrieval system based on latent topics of lyrics, in *Proc. 15th Int. Society for Music Inf. Retrieval Conf.*, 2014, pp. 585–590.
[30] L. Sterckx, T. Demeester, J. Deleu, L. Mertens and C. Develder, Assessing quality of unsupervised topics in song lyrics, in *Advances in Information Retrieval*, 2014, pp. 547–552.
[31] T. Kudo, MeCab: Yet another part-of-speech and morphological analyzer, http://mecab.sourceforge.net/.
[32] T. Nakano, J. Kato, M. Hamasaki and M. Goto, PlaylistPlayer: An interface using multiple criteria to change the playback order of a music playlist, in *Proc. 21st ACM Int. Conf. on Intelligent User Interfaces*, 2016, pp. 186–190.
[33] D. J. Hu and L. K. Saul, A probabilistic topic model for unsupervised learning of musical key-profiles, in *Proc. of 10th Int. Society for Music Inf. Retrieval Conf.*, 2009, pp. 441–446.
[34] D. J. Hu and L. K. Saul, A probabilistic topic model for music analysis, in *Proc. of 29th Annual Conference on Neural Information Processing Systems*, 2009, pp. 1–4.
[35] R. Takahashi, Y. Ohishi, N. Kitaoka and K. Takeda, Building and combining document and music spaces for music query-by-webpage system, in *Proc. of 9th Annual Conf. of the Int. Speech Communication Association*, 2008, pp. 2020–2023.
[36] P. Symeonidis, M. M. Ruxanda, A. Nanopoulos and Y. Manolopoulos, Ternary semantic analysis of social tags for personalized music recommendation, in *Proc. of 9th Int. Conf. on Music Inf. Retrieval*, 2008, pp. 219–224.
[37] M. Hoffman, D. Blei and P. Cook, Content-based musical similarity computation using the hierarchical dirichlet process, in *Proc. of 9th Int. Conf. on Music Inf. Retrieval*, 2008, pp. 349–354.