MUSICAL TYPICALITY: HOW MANY SIMILAR SONGS EXIST?

Tomoyasu Nakano[1] **Daichi Mochihashi**[2] **Kazuyoshi Yoshii**[3] **Masataka Goto**[1]
[1] National Institute of Advanced Industrial Science and Technology (AIST), Japan
[2] The Institute of Statistical Mathematics, Japan
[3] Kyoto University, Japan
[1] {t.nakano, m.goto}[at]aist.go.jp [2] daichi[at]ism.ac.jp
[3] yoshii[at]kuis.kyoto-u.ac.jp

ABSTRACT

We propose a method for estimating the musical "typicality" of a song from an information theoretic perspective. While musical similarity compares just two songs, musical typicality quantifies how many of the songs in a set are similar. It can be used not only to express the uniqueness of a song but also to recommend one that is representative of a set. Building on the type theory in information theory (Cover & Thomas 2006), we use a Bayesian generative model of musical features and compute the typicality of a song as the sum of the probabilities of the songs that share the type of the given song. To evaluate estimated results, we focused on vocal timbre which can be evaluated quantitatively by using the singer's gender. Estimated typicality is evaluated against the Pearson correlation coefficient between the computed typicality and the ratio of the number of male singers to female singers of a song-set. Our result shows that the proposed measure works more effectively to estimate musical typicality than the previous model based simply on generative probabilities.

1 INTRODUCTION

The amount of digital content that can be accessed has been increasing and will continue to do so in the future. This is desirable with regard to the diversity of the content, but is making it harder for listeners to select from this content. Furthermore, since the amount of similar content is also increasing, creators will be more concerned with the originality of their creations. All kinds of works are influenced by some existing content, and it is difficult to avoid an unconscious creation of content partly similar in some way to other content.

This paper focuses on *musical typicality* which reflects the number of songs having high similarity with the target song as shown in Figure 1. This definition of musical typicality is based on *central tendency*, which in cognitive psychology is one of the determinants of typicality [2]. Musical typicality can be used to recommend a unique or representative song for 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. And it can help listeners to understand the relationship between a song and such a song set. However, human ability with regard to typicality is limited. Judging similarity between two songs is a relatively simple task but is time-consuming, so judging the similarities of a million songs is impossible. Consequently, despite the coming of an era in which people other than professional creators 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 typicality proposed in this paper can help create an environment in which specialists and general users alike can know the answers to the questions "How often does this occur?" and "How many similar songs are there?".

Much previous work has focused on *musical similarity* because it is a central concept of MIR and is also important for purposes other than retrieval. For example, the use

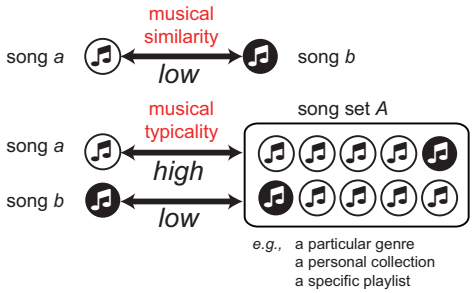

Figure 1. Musical similarity and typicality.

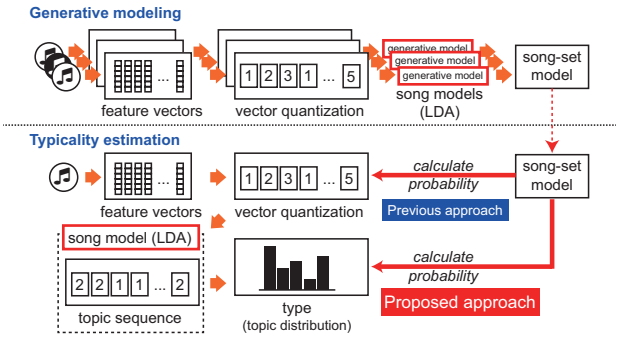

Figure 2. Estimation of music typicality represented by a discrete sequence based on the type theory. Both the previous and the proposed approach are illustrated.

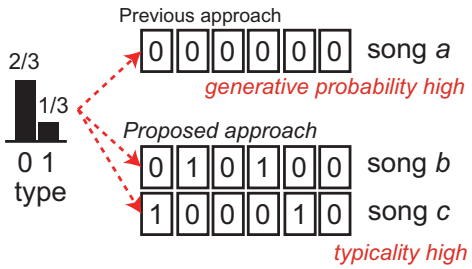

Figure 3. Examples of a song having high generative probability and songs having high typicality.

of similarity to automatically classify musical pieces (into genres, music styles, etc.) has been studied [10, 13], as has its use for music auto-tagging [17]. Each of these applications, however, is different from musical typicality: 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.).

Nakano *et al.* proposed a method for estimating musical typicality by using a generative model trained from the song set (Figure 2) [16] and showed its application to visualizing relationships between songs in a playlist [14]. Their method estimates acoustic features of the target song at each frame and represents the typicality of the target song represented as an average probability of each frame of the song calculated using the song-set model. However, we posit that the generative probability is not truely appropriate to represent typicality.

The method we propose here, in contrast, introduces the *type* from information theory for improving estimated musical typicality by a bag-of-features approach [16]. In this context, the type is same meaning with the unigram distribution. We first model musical features of songs by using a vector quantization method and latent Dirichlet allocation (LDA) [4]. We then estimate a song-set model from the song models. Finally, we compute the typicality of the target song by calculating the probability of a type of the musical sequence (quantized acoustic features) calculated using the song-set model (Figure 2).

2 METHOD

The key concept of the method in this paper is the *type* of a sequence on which we consider the typicality of a given music. Previous work have mentioned/used simple generative probabilities to compute musical similarity [1] or typicality [16] of a music and for singer identification [8]. However, simple generative probability will not conform to our notion of typicality. Imagine the simplest example in Figure 3: here, each song consists of alphabets of $\{0, 1\}$ and the stationary information source has a probability distribution on alphabets $Q(0) = 2/3,\ Q(1) = 1/3$.

Clearly, while the song "a" has the highest probability of generation, we can see that the sequences like "b" and "c" will occur more typically. This means that we should think about the *sum* of the probabilities of songs that are similar to the song to measure the typicality.

2.1 Type and the Typicality

Let us formalize our ideas from the viewpoint of information theory [5–7]. Let $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$ be a sequence of length n whose alphabet x comes from a set $\mathcal{X}$. We assume that $\mathbf{x}$ comes from a stationary memoryless information source, i.e. we can drop the order of symbols in $\mathbf{x}$ and regard $\mathbf{x}$ as a bag of words. Next, we introduce some definitions:

Definition 1 (type). Let $N(x|\mathbf{x})$ be the number of times that $x \in \mathcal{X}$ appeared in sequence $\mathbf{x}$. The *type* $P_\mathbf{x}$ of the sequence $\mathbf{x}$ is an empirical probability distribution of symbols in $\mathbf{x}$:

$$P_\mathbf{x} = \left\{ \frac{1}{n} N(x|\mathbf{x}) \,\middle|\, x \in \mathcal{X} \right\}. \tag{1}$$

We denote the space of all $P_\mathbf{x}$ as $\mathcal{P}_n$.

Definition 2. Let $P \in \mathcal{P}_n$. A set of sequences of length n that share the same type P is called a *type class* T^n of P:

$$T^n(P) = \{\mathbf{x} \in \mathcal{X}^n \mid P_\mathbf{x} = P\}. \tag{2}$$

Now let us denote the probability of a sequence $\mathbf{x}$ from an memoryless information whose symbol probabilities are $Q(x)$:

$$p(\mathbf{x}) = Q^n(\mathbf{x}) = \prod_{i=1}^{n} Q(x_i). \tag{3}$$

Given these definitions, the following simple theorems follow:

Theorem 1. The probability of a sequence $\mathbf{x}$ having type P from a stationary memoryless information source Q is expressed as follows:

$$Q^n(\mathbf{x}) = \exp\big[-n(H(P) + D(P||Q))\big] \tag{4}$$

Here, $H(P)$ and $D(P||Q)$ are an entropy of P and Kullback-Leibler divergence of P from Q, respectively.

$$H(P) = -\sum_{x \in \mathcal{X}} P(x) \log P(x) \tag{5}$$

$$D(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \tag{6}$$

Proof.

$$\begin{aligned} Q^n(\mathbf{x}) &= \prod_{i=1}^{n} Q(x_i) = \prod_x Q(x)^{N(x|\mathbf{x})} = \prod_x Q(x)^{nP(x)} \\ &= \prod_x \exp\big[nP(x) \log Q(x)\big] && (7) \\ &= \exp\Big[-n\Big(-\sum_x P(x) \log Q(x)\Big)\Big] && (8) \\ &= \exp\Big[-n\Big(H(P) + D(P||Q)\Big)\Big]. \quad \square && (9) \end{aligned}$$

Theorem 2 (lower and upper bounds)**.** For any type $P \in \mathcal{P}_n$,

$$\frac{1}{(n+1)^{|\mathcal{X}|-1}} \exp\{nH(P)\} \leq |T^n(P)| \leq \exp\{nH(P)\}. \tag{10}$$

Using the theorems above, the following important theorem can be proved.

Theorem 3. For any type $P \in \mathcal{P}_n$ and any probability distribution Q,

$$Q^n(T^n(P)) \doteq \exp\{-nD(P||Q)\}, \tag{11}$$

where $a_n \doteq b_n$ if $\lim_{n\to\infty}(1/n)\log(a_n/b_n) = 0$.

Proof. Using (4) and (10),

$$\begin{aligned} Q^n(T^n(P)) &= \textstyle\sum_{\mathbf{x}\in T^n(P)} Q^n(\mathbf{x}) \\ &= |T^n(P)| \exp(-n(H(P) + D(P||Q))) \\ &\doteq \exp(nH(P)) \cdot \exp(-n(H(P) + D(P||Q))) \\ &= \exp\{-nD(P||Q)\}\,. \quad \square \end{aligned} \tag{12}$$

This theorem says that the *sum* of the probabilities of sequences that share the same type P is given by an exponential of Kullback-Leibler divergence from the information source Q. While the equation (11) is usually used in information theory to formalize that such a probability exponentially decays with the length n, here we do not care for n but for the form of the function. Thus, we normalize (11) for a unit observation like the well-known *perplexity*, yielding the definition of typicality as follows:

Definition 3 (Typicality).

$$\text{Typicality}(P|Q) = \exp\{-D(P||Q)\} \tag{13}$$

where P is the type of a musical sequence and Q is a generative model of its musical features.

2.2 Generative modeling and Type

To evaluate the typicality estimation method, we compute the type of each song by modeling them in a way based on our previous work [16]. From polyphonic musical audio signals including a singing voice and sounds of various musical instruments, we first extract vocal timbre. We then model the timbre of songs by using a vector-quantization method and latent Dirichlet allocation (LDA) [4]. Finally, a song-set model Q is estimated by integrating all song models (Figure 2).

In addition, we use the expectation of Dirichlet topic distribution as a type P because the hyperparameter of the posterior Dirichlet distribution can be interpreted as the number of observations of the corresponding topic. In the other words, the P indicates mixing weights of the multiple topics.

2.2.1 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 [8, 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) [8].

We first use Goto's PreFEst [11] 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 resynthesized 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 [8] for details). Feature vectors of only the reliable frames are used in the following processes (model training and probability estimation).

2.2.2 Quantization

Vector quantization is applied using the k-means algorithm to convert acoustic feature vectors of each element to a symbolic time series representation. 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*.

2.2.3 Probabilistic generative model: song model

The observed data we consider for LDA are D independent songs $\boldsymbol{X} = \{\boldsymbol{X}_1, ..., \boldsymbol{X}_D\}$. A song $\boldsymbol{X}_d$ is N_d acoustic words $\boldsymbol{X}_d = \{\boldsymbol{x}_{d,1}, ..., \boldsymbol{x}_{d,N_d}\}$. The size of the acoustic words vocabulary equals to the number of clusters of the k-means algorithm, V. We consider a K-dimensional multinomial of latent topic proportions θ_d for each $\boldsymbol{X}_d$, and write $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_D)$.

Introducing latent topic assignments $\boldsymbol{Z}_d = \{\boldsymbol{z}_{d,1}, ..., \boldsymbol{z}_{d,N_d}\}$ for $\boldsymbol{X}_d$ and collectively write $\boldsymbol{Z} = \{\boldsymbol{Z}_1, ..., \boldsymbol{Z}_D\}$, the full joint distribution of our LDA model is given by

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\phi})p(\boldsymbol{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{\phi}) \tag{14}$$

where $\boldsymbol{\phi}$ indicates the emission distribution of each topic. The first two terms are likelihood functions, and the other two are prior distributions. The likelihood functions are defined as

$$p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\phi}) = \prod_{d=1}^{D}\prod_{n=1}^{N_d}\prod_{v=1}^{V}\left(\prod_{k=1}^{K}\phi_{k,v}^{z_{d,n,k}}\right)^{x_{d,n,v}} \tag{15}$$

and

$$p(\boldsymbol{Z}|\boldsymbol{\theta}) = \prod_{d=1}^{D}\prod_{n=1}^{N_d}\prod_{v=1}^{V}\theta_{d,k}^{z_{d,n,k}}. \tag{16}$$

We endow $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ conjugate Dirichlet priors:

$$p(\boldsymbol{\theta}) = \prod_{d=1}^{D}\text{Dir}(\theta_d|\alpha_0) \propto \prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{d,k}^{\alpha_0-1} \tag{17}$$

$$p(\boldsymbol{\phi}) = \prod_{k=1}^{K}\text{Dir}(\boldsymbol{\phi}_k|\beta_0) \propto \prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{k,v}^{\beta_0-1}\,. \tag{18}$$

where p($\boldsymbol{\theta}$) and p($\boldsymbol{\phi}$) are products of Dirichlet distributions and α_0, β_0 are their prior hyperparameters.

Finally, we obtain a type of each song $\boldsymbol{X}_d$ as an expectation of the Dirichlet posterior distribution of θ_d.

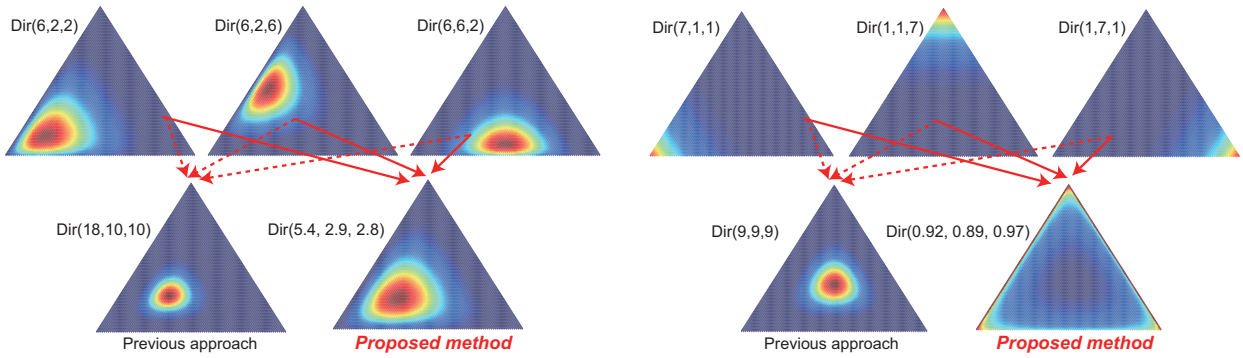

Figure 4. Song-set model estimation of previous approach and a model estimated by our proposed method.

2.3 Typicality over a set of Songs

Given the type of each song, we wish to compute the typicality of a song as compared to a set of other songs. In Section 2.1, we defined the typicality of a sequence of type P from an information source having distribution Q. Therefore, we need some way to estimate Q from the set of songs (i.e. types) beforehand. Actually, we do not have to estimate a single Q but compute an expectation around it:

$$\text{Typicality}(P|\Theta) = \mathbb{E}[\exp(-D(P||\theta))]_{\theta\sim\text{Dir}(\alpha)} \quad (19)$$

where $\Theta = \{\theta_1, \cdots, \theta_n\}$ is a set of types of other songs and $\text{Dir}(\alpha)$ is a prior Dirichlet distribution from which each $\theta_i \in \Theta$ is deemed to be generated.

In the previous work [16], we estimated the hyperparameter α by just summing the topic distributions $\theta_1, \cdots, \theta_n$. As shown in Figure 4, however, this could lead to an undesirable result and we employ a Bayesian formula to estimate α. This derivation is based on the following Dirichlet and Gamma distributions:

$$\text{Dir}(\theta|\alpha) = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1} \quad (20)$$

$$\text{Ga}(\alpha|a,b) = \frac{b^a}{\Gamma(a)} \alpha^{a-1} e^{-b\alpha} \quad (21)$$

Therefore,

$$p(\alpha|\Theta) \propto p(\Theta|\alpha)p(\alpha) \quad (22)$$

$$\propto \prod_k \alpha_k^{a-1} e^{-b\alpha_k} \cdot \prod_j \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k}, \quad (23)$$

which leads to

$$p(\alpha_k|\alpha_{k-1}, \Theta) \propto \alpha_k^{a-1} e^{-b\alpha_k} \cdot \prod_j \frac{\Gamma\left(\sum_k \alpha_k\right)}{\Gamma(\alpha_k)} \theta_k^{\alpha_k}. \quad (24)$$

Because we cannot expand $\Gamma(\sum_k \alpha_k)/\Gamma(\alpha_k)$, we make a following approximation with n being a nearest integer to $\sum_{j\neq k} \alpha_k$ [18]:

$$\frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\alpha_k)} = \frac{\Gamma(\alpha_k + \sum_{j\neq k} \alpha_j)}{\Gamma(\alpha_k)} \simeq \frac{\Gamma(\alpha_k + n)}{\Gamma(\alpha_k)} \quad (25)$$

$$= \alpha_k(\alpha_k + 1) \cdots (\alpha_k + n - 1) \quad (26)$$

$$= \prod_{i=0}^{n-1} \alpha_k(\alpha_k + i) \quad (27)$$

$$= \prod_{i=0}^{n-1} \sum_{y\in\{0,1\}} (\alpha_k)^{y_i} (i)^{1-y_i}. \quad (28)$$

Therefore, introducing auxiliary variables

$$y_i \sim \text{Bernoulli}\left(\frac{\alpha_k}{\alpha_k + i}\right), \quad (29)$$

we can make a following Gamma proposal for α_k:

$$p(\alpha_k|\alpha_{k-1}, \Theta) \quad (30)$$

$$\simeq \alpha_k^{a-1} e^{-b\alpha_k} \cdot \prod_j e^{\alpha_k \log \theta_{jk}} \cdot \prod_j \prod_{i=0}^{n-1} \alpha_k^{y_{ji}} \quad (31)$$

$$= \alpha_k^{a + \sum_j \sum_{i=0}^{n-1} y_{ji} - 1} \cdot e^{-\alpha_k (b - \sum_j \log \theta_{jk})} \quad (32)$$

$$= \text{Ga}(a + \sum_j \sum_{i=0}^{n-1} y_{ji}, b - \sum_j \log \theta_{jk}). \quad (33)$$

Because this is just a proposal, we further correct the bias using a Metropolis-Hastings algorithm with the exact likelihood formula (24).

2.4 Computing the Expectation

Once we obtain α from Θ, we can compute the expectation (19) analytically. Denoting $P = (p_1, \cdots, p_K)$ and writing $\mathbb{E}[]$ as $\langle\rangle$,

$$\text{Typicality}(P|\Theta) = \left\langle \exp(-D(P||\theta)) \right\rangle_{\theta\sim\text{Dir}(\alpha)} \quad (34)$$

$$= \left\langle \exp \sum_{k=1}^{K} p_k \log \frac{\theta_k}{p_k} \right\rangle_{\theta\sim\text{Dir}(\alpha)}$$

$$= \frac{1}{\exp(\sum_k p_k \log p_k)} \left\langle \exp \sum_k p_k \log \theta_k \right\rangle_{\theta\sim\text{Dir}(\alpha)}$$

$$= \exp(H(P)) \left\langle \prod_{k=1}^{K} \theta_k^{p_k} \right\rangle_{\theta\sim\text{Dir}(\alpha)}. \quad (35)$$

Here, the second term is

$$\left\langle \prod_{k=1}^{K} \theta_k^{p_k} \right\rangle_{\theta\sim\text{Dir}(\alpha)} = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \int \prod_k \theta_k^{\alpha_k - 1} \cdot \prod_k \theta_k^{p_k} d\theta$$

$$= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \int \prod_k \theta_k^{\alpha_k + p_k - 1} d\theta$$

$$= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(\alpha_k + p_k)}{\Gamma(\sum_k \alpha_k + p_k)}$$

$$= \frac{1}{\sum_k \alpha_k} \prod_k \frac{\Gamma(\alpha_k + p_k)}{\Gamma(\alpha_k)} . \quad (36)$$

Therefore, from (35) we finally obtain

$$\mathrm{Typicality}(P|\Theta) = \frac{\exp(H(P))}{\sum_k \alpha_k} \prod_k \frac{\Gamma(\alpha_k + p_k)}{\Gamma(\alpha_k)} . \quad (37)$$

3 EXPERIMENTS

The proposed methods were tested in an experiment evaluating the estimated typicality. To evaluate estimated results, we focused on vocal timbre which can be evaluated quantitatively by using the singer's gender.

3.1 Dataset

The song set used for the LDA-model-training and typicality estimation comprised 3,278 Japanese popular songs[1] 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 MDB.

The song set used for GMM training and k-means training to extract the acoustic features consisted of 100 popular songs from the RWC Music Database (RWC-MDB-P-2001) [9]. 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 this song set as the RWC MDB.

3.2 Experimental Settings

Conditions and parameters of the methods described in the METHODS section are described here in detail. Some conditions and each parameter value were based on previous work [15, 16].

3.2.1 Typicality estimation

The number of iterations of the Bayesian song-set model estimation described in Subsection 2.3 was 1000.

3.2.2 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. To estimate the features, the vocal sound was re-synthesized by using a sinusoidal model. The ΔF_0 was calculated every five frames (50 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 [3].

3.2.3 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.

[1] Note that some are Western popular songs and English in them .

3.2.4 Training the generative models

Training song models and song-set models of the vocal timbre by LDA, we used all of the 3,278 original recordings of the JPOP MDB.

The number of topics, K, was set to 100, and the model parameters of LDA were trained using the collapsed Gibbs sampler [12]. The hyperparameters of the Dirichlet distributions for topics and words were initially set to 1 and 0.1, respectively.

3.3 Four typicality measures

We evaluated the following four typicality computing conditions.

T1: computing the Euclidean distance
T2: computing the generative probability [16]
T3: computing the KL-divergence, equation (13)
T4: computing the KL-divergence, equation (37)

As a baseline method, under the T1 condition, one simple method used to estimate the typicality of vocal timbre calculated the Euclidean distance between mean feature vectors of a song and a song-set. Each mean vector was calculated from each song, using the reliable vocal frames, and was normalized by subtracting the mean and dividing by the standard deviation of all mean vectors.

Under the T2 condition, one typicality between a song and a set of songs is obtained by calculating a generative probability [16] of song P calculated using a song-set model of song Q. This typicality $p_g(P|Q)$ is defined as follows:

$$\log p_g(P|Q) = \frac{1}{N_P} \sum_{n=1}^{N_P} \log p(\mathbf{x}_{P,n}|\, \mathbb{E}[\boldsymbol{\theta}_Q], \mathbb{E}[\boldsymbol{\phi}]), \quad (38)$$

$$p(\mathbf{x}_{P,n}|\, \mathbb{E}[\boldsymbol{\theta}_Q], \mathbb{E}[\boldsymbol{\phi}]) = \sum_{k=1}^{K} \left(\mathbb{E}[\theta_{Q,k}] \cdot \mathbb{E}[\phi_{k,v}]\right), \quad (39)$$

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

The other two typicalities, under the T3 and T4 conditions, are calculated $\mathrm{Typicality}(P, Q)$ by using equations (13) and (37), respectively.

3.4 Experiment: musical typicality estimation

We evaluated the four typicality computing conditions (T1-T4) in combination with the following three song-set modeling conditions.

M1: computing a mean vector
M2: summing the Dirichlet hyperparameters [16]
M3: Bayesian estimation of the hyperparameters described in Subsection 2.3

We computed typicalities under five evaluation conditions T1+M1, T2+M2, T3+M2, T3+M3, and T4+M3.

Our typicality evaluation experiment used five hundred songs by a hundred singers (50 male and 50 female), each singer sung five songs. The songs are taken from the JPOP MDB and each of the songs included only one vocal. To

Evaluated conditions	First selection			Second selection			Third selection			Fourth selection			Fifth selection		
	ρ_{m}	ρ_{f}	ρ_{mf}	ρ_{m}	ρ_{f}	ρ_{mf}	ρ_{m}	ρ_{f}	ρ_{mf}	ρ_{m}	ρ_{f}	ρ_{mf}	ρ_{m}	ρ_{f}	ρ_{mf}
T1+M1	.855	.866	.855	.775	.821	.798	.935	.835	.882	.870	.876	.872	.914	.842	.876
T2+M2	.924	.930	.860	.905	.921	.866	.953	.918	.879	.925	.938	.875	.945	.910	.864
T3+M2	.921	.927	.861	.912	.921	.871	.951	.919	.880	.924	.935	.876	.944	.907	.865
T3+M3	.940	.961	.931	.910	.961	.926	.962	.955	.944	.936	.967	.934	.952	.950	.933
T4+M3	.936	.973	.942	.844	.973	.896	.968	.962	.952	.930	.976	.936	.970	.949	.939

Table 1. Pearson correlation coefficients of a hundred songs under the five evaluated conditions ("T4+M3" is the proposed method) and the underline means the highest value. The songs are randomly selected five times from five hundred songs.

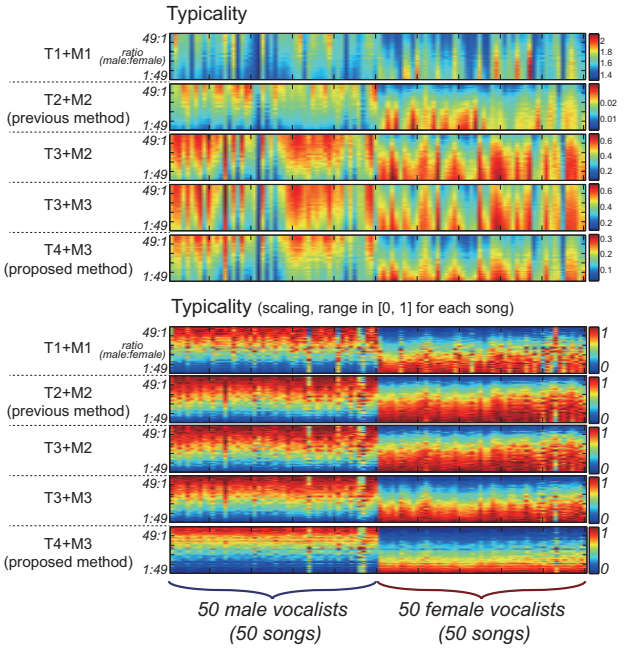

Figure 5. Estimated typicalities (first selection) and those scaled values for each of the five evaluation conditions.

estimate musical typicality, a hundred songs by different singers are randomly selected five times. Then, to integrate or estimate song-set models, fifty songs are randomly selected from the songs with different ratios of the number of male singers to female singers ($1:49, 2:48, ..., 49:1$). When a model with a high proportion of female songs is used, the typicality of songs sung by females is higher than the typicality of songs sung by males (and vice versa).

Estimated typicality was evaluated against the Pearson product-moment correlation coefficient between the computed typicality the ratio of the number of male singers to female singers with respect to song-set modeling. Before computing the coefficients, the typicality for each song was scaled to have values from 0 to 1 for evaluating smooth transition. Let ρ_{m}, ρ_{f}, and ρ_{mf} denote the coefficients under a set of songs consist of 50 songs by male singers, 50 songs by female singers, and all 100 songs, respectively.

The estimated typicalities and those scaled values are shown in Figure 5 for each of the five evaluation conditions. The Pearson's correlation coefficients are listed in Table 1. The results show that the proposed method achieved the highest value of the correlation coefficient (T4+M3). This means that the proposed method works better than the baseline method based on the Euclidean distance of mean vectors (T1+M1) and the previous method based on computing the generative probabilities (T2+M2). The results also show that estimated musical typicality by using the proposed method can reflect the ratio between the number of songs belonging to a class (*e.g.*, male singer) and the number of songs belonging to another class (*e.g.*, female singer).

4 CONCLUSION

We proposed a method for estimating musical typicality based on the type theory. Although this method is used for quantized acoustic features for vocal timbre in this paper, it can be used for other discrete sequence representations of music, such as quantized other acoustic features (*e.g.*, MFCCs to represent musical timbre/genre), lyrics and musical score. It can also be used with probabilistic representation instead of estimating musical similarities of all possible song-pairs by using a model trained from each song, for integrating or collaboration with other probabilistic approach as a unified framework.

Our definition of musical typicality was based on the central tendency [2] which is only the definition to be computed from the audio data; this is the reason to adopt it. In future work we expect to deal with two other definitions in cognitive psychology are *frequency of instantiation* and *ideals*. The frequency of instantiation is a perspective on social recognition, that is, things with a lot of exposure on media or in advertisements are typical, and ideals focuses on an ideal condition of the category, that is, things that are close to an ideal condition are typical.

Musical typicality can be used not only for music-listening support interface such as retrieving an uniqueness song or visualizing typicalities, but also to do this by developing a music-creation support interface enabling high/low typicality elements (*e.g.*, timbre and lyrics) to be used to increase originality or visualize typicality in order to avoid unwarranted accusations of plagiarism. 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 typicality to other content.

5 ACKNOWLEDGMENT

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

6 REFERENCES

[1] Jean-Julien Aucouturier and Francois Pachet. Music similarity measures: What's the use? In *Proc. ISMIR 2002*, pages 157–163, 2002.

[2] Lawrence 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):629–654, 1985.

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

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. New-York: Wiley, 2006.

[6] Imre Csiszár. The method of types. *IEEE Trans. on Information Theory*, 44(6):2505–2523, 1998.

[7] Imre Csiszár and János K orner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, 1981.

[8] Hiromasa Fujihara, Masataka Goto, Tetsuro Kitahara, and Hiroshi 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*, 18(3):638–648, 2010.

[9] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Popular, classical, and jazz music databases. In *Proc. ISMIR 2002*, pages 287–288, 2002.

[10] Masataka Goto and Keiji Hirata. Recent studies on music information processing. *Acoustical Science and Technology (edited by the Acoustical Society of Japan)*, 25(6):419–425, 2004.

[11] Masatala 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):311–329, 2004.

[12] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. In *Proc. Natl. Acad. Sci. USA (PNAS)*, volume 1, pages 5228–5235, 2004.

[13] Peter Knees and Markus Schedl. A survey of music similarity and recommendation from music context data. *ACM Trans. on Multimedia Computing, Communications and Applications*, 10(1):1–21, 2013.

[14] Tomoyasu Nakano, Jun Kato, Masahiro Hamasaki, and Masataka Goto. PlaylistPlayer: An interface using multiple criteria to change the playback order of a music playlist. In *Proc. ACM IUI 2016*, 2016.

[15] Tomoyasu Nakano, Kazuyoshi Yoshii, and Masataka Goto. Vocal timbre analysis using latent Dirichlet allocation and cross-gender vocal timbre similarity. In *Proc. ICASSP 2014*, pages 5239–5343, 2014.

[16] Tomoyasu Nakano, Kazuyoshi Yoshii, and Masataka Goto. Musical similarity and commonness estimation based on probabilistic generative models. In *Proc. IEEE ISM 2015*, 2015.

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

[18] Yee Whye Teh. A bayesian interpretation of interpolated kneser-ney. *Technical Report TRA2/06*, 2006.