# VOCAL TIMBRE ANALYSIS USING LATENT DIRICHLET ALLOCATION AND CROSS-GENDER VOCAL TIMBRE SIMILARITY

*Tomoyasu Nakano* *Kazuyoshi Yoshii* *Masataka Goto*

National Institute of Advanced Industrial Science and Technology (AIST), Japan

**ABSTRACT**

This paper presents a vocal timbre analysis method based on topic modeling using latent Dirichlet allocation (LDA). Although many works have focused on analyzing characteristics of singing voices, none have dealt with "latent" characteristics (topics) of vocal timbre, which are shared by multiple singing voices. In the work described in this paper, we first automatically extracted vocal timbre features from polyphonic musical audio signals including vocal sounds. The extracted features were used as observed data, and mixing weights of multiple topics were estimated by LDA. Finally, the semantics of each topic were visualized by using a word-cloud-based approach. Experimental results for a singer identification task using 36 songs sung by 12 singers showed that our method achieved a mean reciprocal rank of 0.86. We also proposed a method for estimating cross-gender vocal timbre similarity by generating pitch-shifted (frequency-warped) signals of every singing voice. Experimental results for a cross-gender singer retrieval task showed that our method discovered interesting similar pitch-shifted singers.

***Index Terms***— vocal timbre, cross-gender similarity, music information retrieval, latent Dirichlet allocation, word cloud

## 1. INTRODUCTION

The vocal (singing voice) is an important element of music in various musical genres, especially in popular music. Indeed, the vocal timbre and singing style can influence people's decision on which songs to listen to. In fact, several music information retrieval (MIR) systems based on vocal timbre similarity have been proposed [1–5]. When people listen to singing voices, they can feel that different vocal timbres and singing styles share some factors that characterize their timbres and styles. It is, however, not easy to define every factor even by singers themselves because such factors are *latent*. We call these shared factors *"latent topics"*. The aim of this study is to explore the latent topics of singing voices by deriving them from many singing voices sung by different singers. The latent topics are useful for MIR as well as singing analysis.

There are many reports of research on automatic estimation of singing characteristics from audio signals: characteristics such as voice category (*e.g.*, soprano or alto) [6, 7], gender [8–10], age [10], body size [10], race [10], vocal register [11], singing modeling ($F_0$, power, and spectral envelope) [12–19], breath sound [20,21], singing skill [6, 7, 22–25], enthusiasm [26], $F_0$ dynamics and musical genres [27], and the language of the lyrics [28–31] have been previously proposed. The above previous works, however, have not revealed latent topics that are shared by different singing voices.

To explore shared latent topics of voice timbres or singing styles, we propose a vocal timbre analysis method based on a topic modeling method called latent Dirichlet allocation (LDA) [32]. In LDA, each singing voice is represented as a weighted mixture of multiple topics shared by all the singing voices in our song database. The mixing weights of LDA can be used to compute singing voice similarity for MIR (Fig. 1, Ⓐ) and to visualize the semantics of each topic by using a word-cloud approach (Fig. 1, Ⓑ).

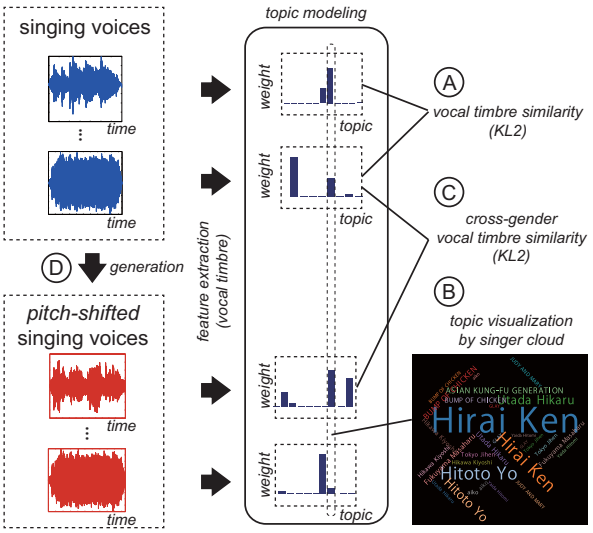


**Fig. 1**. Overview of topic modeling of singing voices: vocal timbre similarity, cross-gender vocal timbre similarity, and topic visualization by singer cloud.

Moreover, we also propose a method for estimating *cross-gender vocal timbre similarity* (Fig. 1, Ⓒ). For this estimation, pitch-shifted (frequency-warped) audio signals of all singing voices are automatically generated (Fig. 1, Ⓓ). For instance, by shifting up the pitch of a male singing voice, we are able to obtain a female-like singing voice. By using such pitch-shifted singing voices as queries for MIR based on the latent topics of singing voice timbres, we can find interesting cross-gender pairs of similar singing voices.

The remainder of this paper is structured as follows. Section 2 describes the proposed vocal timbre analysis method and cross-gender similarity estimation method. Section 3 describes two experiments we used to evaluate the methods. Section 4 concludes the paper by summarizing the key outcomes and discusses future work.

## 2. METHOD

This section describes a method of singing analysis by latent Dirichlet allocation (LDA) [32], and a method for estimating cross-gender vocal timbre similarity. We deal with vocal timbre features extracted from polyphonic musical audio signals including vocal sounds. The cross-gender similarity is computed after first generating pitch-shifted (frequency-warped) signals of all the target songs.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

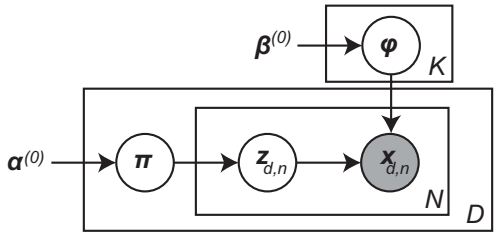


**Fig. 2**. Graphical representation of the latent Dirichlet allocation (LDA). First the finite sets of mixing weights $\boldsymbol{\pi}$ of the multiple topics and the unigram probabilities $\boldsymbol{\phi}$ of the singing words are stochastically generated according to Dirichlet prior distributions. Then one of $K$ topics is stochastically selected as a latent variable $\boldsymbol{z}_{d,n}$ according to a multinomial distribution defined by $\boldsymbol{\pi}$. Finally the singing word $\boldsymbol{x}_{d,n}$ is stochastically generated according to a multinomial distribution defined by $\boldsymbol{\phi}$.

There are previous works related to latent analysis of music, such as music retrieval based on LDA of lyrics and melodic features [33], chord estimation based on LDA [34, 35], combining document and music spaces by latent semantic analysis [36], music recommendation by social tag and latent semantic analysis [37], and music similarity based on the hierarchical Dirichlet process [38]. The self-organizing map (SOM) can be latent analysis, and SOM-based music clustering has been proposed [39]. Futhermore, there exist many research papers on acoustic analysis based on topic modeling (see, for example [40–43]). There are, however, none that dealt with singing features.

### 2.1. Feature extraction of vocal timbre

To extract vocal timbre features, we use modules of Songle [44], our Web service for active music listening. We first use Goto's PreFEst [45] to estimate the $F_0$ of the melody, and then LPMCC (mel-cepstral coefficients of LPC spectrum) of vocal and $\Delta F_0$ are estimated by using the $F_0$ and are combined them as a feature vector at each frame. Then *reliable frames* are selected as vocal by using a vocal GMM and a non-vocal GMM (see [3]). Finally, all feature vectors of the reliable frames are normalized by subtracting the mean and dividing by the standard deviation.

### 2.2. Converting vocal timbre features to symbolic information by using a $k$-means algorithm

LDA deals with symbolic information (*e.g.* text), not continuous feature values as described in subsection 2.1 This paper therefore propose that the vocal features are converted to symbolic time series by using a $k$-means algorithm. We call these symbolic representations of singing *singing words*.

### 2.3. LDA model formulation

The observed data we consider for LDA are $D$ independent singing voices $\boldsymbol{X} = \{\boldsymbol{X}_1, ..., \boldsymbol{X}_D\}$ already converted to symbolic time series as described in 2.2. A singing voice $\boldsymbol{X}_d$ is $N_d$ symbolic time series $\boldsymbol{X}_d = \{\boldsymbol{x}_{d,1}, ..., \boldsymbol{x}_{d,N_d}\}$ which are the reliable frames (see 2.1). The size of the singing words vocabulary is equivalent to the number of clusters of $k$-means algorithm ($= V$), $\boldsymbol{x}_{d,n}$ is a $V$-dimensional "1-of-$K$" vector (a vector with one element containing a 1 and all other elements containing a 0).

The latent variable of the observed singing voice $\boldsymbol{X}_d$ is $\boldsymbol{Z}_d = \{\boldsymbol{z}_{d,1}, ..., \boldsymbol{z}_{d,N_d}\}$. The number of topics is $K$, so $z_{d,n}$ indicates a $K$-dimensional 1-of-$K$ vector. Hereafter, all latent variables of singing voice $D$ are indicated $\boldsymbol{Z} = \{\boldsymbol{Z}_1, ..., \boldsymbol{Z}_D\}$.

Figure 2 shows a graphical representation of the LDA model used in this paper. The full joint distribution is given by

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

where $\boldsymbol{\pi}$ indicates the mixing weights of the multiple topics ($D$ of the $K$-dimensional vector) and $\boldsymbol{\phi}$ indicates the unigram probability of each topic ($K$ of the $V$-dimensional vector). The first two terms are likelihood functions, the other two terms are prior distributions. The likelihood functions themselves are defined as

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

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

We then introduce conjugate priors as follows:

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

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

where p($\boldsymbol{\pi}$) and p($\boldsymbol{\phi}$) are products of Dirichlet distributions. $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$ are hyperparameters; $C(\boldsymbol{\alpha}^{(0)})$ and $C(\boldsymbol{\beta}^{(0)})$ are normalization factors calculated as follows:

$$C(\boldsymbol{\eta}) = \frac{\Gamma(\hat{\eta})}{\Gamma(\eta_1) \cdots \Gamma(\eta_{|\eta|})}, \quad \hat{\eta} = \sum_{i=1}^{|\eta|} \eta_i \tag{6}$$

### 2.4. Singer identification by computing vocal timbre similarity

Similarity between two songs is defined in this paper as the inverse of the symmetric Kullback-Leibler distance (KL2) between two distributions, as follows:

$$d_{\mathrm{KL2}}(\boldsymbol{\pi}_A||\boldsymbol{\pi}_B) = \sum_{k=1}^{K} \pi_A(k) \log \frac{\pi_A(k)}{\pi_B(k)} + \sum_{k=1}^{K} \pi_B(k) \log \frac{\pi_B(k)}{\pi_A(k)}, \tag{7}$$

Here the mixing weights of a singing A is $\boldsymbol{\pi}_A$ and the mixing weights of a singing B is $\boldsymbol{\pi}_B$, and these are normalized to meet the probability criterion.

$$\sum_{k=1}^{K} \pi_A(k) = 1, \quad \sum_{k=1}^{K} \pi_B(k) = 1 \tag{8}$$

### 2.5. Topic visualization by using a word-cloud-based approach

The mixing weight of each song $\boldsymbol{\pi}$ is a $D$,$K$-dimensional vector ($D \times K$ matrix) which means that "$\boldsymbol{\pi}$ shows the predominant topics of each song $d$." The mixing weights can be useful for singer identification and cross-gender similarity estimation as described above in

**Table 1**. Singers of the 36 songs used in the experimental evaluation.

| ID | Singer name | Gender | # of songs |
|---|---|---|---|
| M1 | ASIAN KUNG-FU GENERATION | Male | 3 |
| M2 | BUMP OF CHICKEN | Male | 3 |
| M3 | Fukuyama Masaharu | Male | 3 |
| M4 | GLAY | Male | 3 |
| M5 | Hikawa Kiyoshi | Male | 3 |
| M6 | Hirai Ken | Male | 3 |
| F1 | aiko | Female | 3 |
| F2 | JUDY AND MARY | Female | 3 |
| F3 | Hitoto Yo | Female | 3 |
| F4 | Tokyo Jihen | Female | 3 |
| F5 | Utada Hikaru | Female | 3 |
| F6 | Yaida Hitomi | Female | 3 |

this Section. However, it is difficult to explain of semantic of each topic from the mixing weights.

This subsection considers the topic weights $\boldsymbol{\pi}$ as a $K$,$D$-dimensional vector. This means that "$\boldsymbol{\pi}$ shows the predominant songs for each topic $k$." It is utilized to interpret the semantics of each topic by showing a word cloud, which is one of word visualization methods frequently used on the web. We call this word cloud *singer cloud*. In the singer cloud, metadata of a singing (*e.g.* a singer's name or a song name) are visualized according to the mixing weights. In this paper, predominant singers of each topic are visualized with large size.

### 2.6. Cross-gender similarity by generating pitch-shifted signals

This paper describes a method for cross-gender similarity estimation. Pitch-shifted signals are generated by shifting them up/down the frequency axis according to the results of short-term frequency analysis. This shifting is equivalent to changing the shape of a singer's vocal tract.

All of these pitch-shifted signals are generated by using SoX[1].

## 3. EXPERIMENTAL EVALUATION

The proposed methods were tested in two experiments, one evaluating the singer identification and the other evaluating the cross-gender vocal timbre similarity estimation.

The songs used in these experiments were monaural 16-kHz digital recordings. The singers are listed in Table 1. We used 36 songs by 12 Japanese singers (6 male and 6 female), each singer sung 3 songs. Each of the songs included only one vocal. The songs were taken from commercial music CDs that appeared on a well-known popular music chart[2] in Japan and were placed in the top twenty on weekly charts appearing between 2000 and 2008.

Six recordings pitch-shifted by amounts ranging from $-3$ to $+3$ semitones were generated in 1-semitone steps. Since we also used the original recordings, we had 7 versions of each song and thus used $D = 252 (= 7 \times 3$ songs $\times 12$ singers ) songs for LDA.

Vocal features were extracted from each song (see 2.1), with the top 15% of feature frames used as reliable vocal frames. The number of clusters $V$ of the $k$-means algorithm was set to 100. The number of topics $K$ was set to 100, and the model parameters of LDA were trained by using the collapsed Gibbs sampler [46] with 1000 iterations. The hyperparameter $\boldsymbol{\alpha}^{(0)}$ was initially set to 1 and the hyperparameter $\boldsymbol{\beta}^{(0)}$ was initially set to 0.1.

[1] http://sox.sourceforge.net/

[2] http://www.oricon.co.jp/

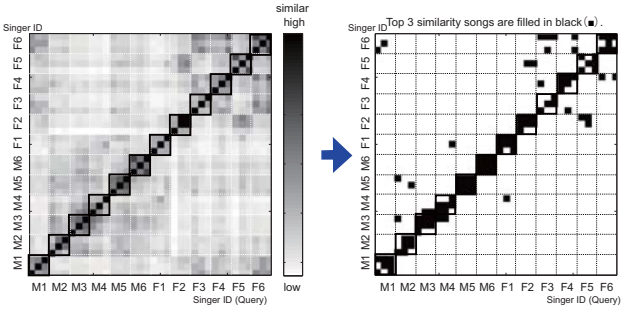


**Fig. 3**. A similarity matrix based on the mixing weights of topics.

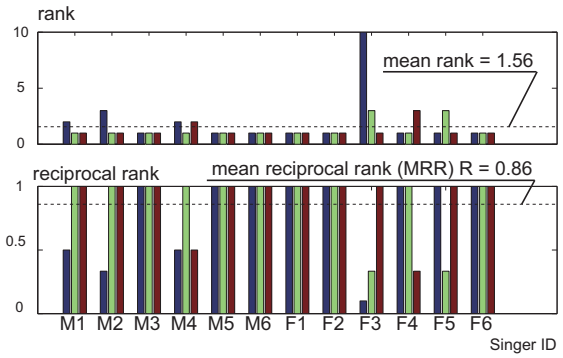


**Fig. 4**. The mean reciprocal rank and reciprocal ranks for all songs.

### 3.1. Experiment A: singer identification

To evaluate the singer identification using the LDA mixing weights $\boldsymbol{\pi}$, experiment A used only the $D_A = 36 (= 12 \times 3)$ songs without pitch-shifted signals. The left side of Fig. 3 shows a similarity matrix based on distance calculation using $\boldsymbol{\pi}$ (eq. 7). The right side of the figure shows that the similarities of top three similar songs of each song are filled with black color.

Figure 4 shows the mean reciprocal rank $R$ defined as follows:

$$R = \frac{1}{D_A} \sum_{d=1}^{D_A} \frac{1}{r_d} \times 100 \tag{9}$$

The mean reciprocal rank is the average of the reciprocal ranks of results for $D_A$ queries, where $r_d$ indicates the rank of song $d$ decided from the similarity. If a same singer's song has the highest similarity, the rank is 1.

These results suggest that songs by the same singer have similar topic weights, and the topic weights can be used to identify singers.

### 3.2. Experiment B: cross-gender similarity

To evaluate the cross-gender similarity estimation using the LDA mixing weights $\boldsymbol{\pi}$, experiment B used all 252 songs. Table 2 shows that a singer ID of the highest similarity song of each query and these values of pitch-shifted. The mixing weights of the 36 original songs without pitch-shifting were used as queries, and the retrieval targets were 245 songs ($= 252 - 7$: excluding 7 versions of oneself). Figure 5 shows numbers of singers who sang the highest similar song of each query. The mixing weights of the all 252 songs were used as queries.

**Table 2**. The highest similarity song of each query, and these values of pitch-shifted (experiment B). The "+1" means pitch-shifting up by 1 semitone. The underline means the most similar songs are sung by the opposite gender (M6 and F3).

| Queries | Most similar song for each query | | |
|---|---|---|---|
| (±0/×1) | query 1 | query 2 | query 3 |
| M1 | F4 (−3) | F4 (−3) | F6 (−3) |
| M2 | M1 (−2) | M3 (+1) | M3 (+1) |
| M3 | M2 (+1) | M2 (±0) | M6 (−1) |
| M4 | F6 (−3) | F5 (−3) | F1 (−3) |
| M5 | M3 (+2) | F1 (−3) | M2 (−1) |
| M6 | F3 (−3) | M3 (+1) | F3 (−3) |
| F1 | F6 (+1) | F5 (+1) | F5 (+2) |
| F2 | F6 (±0) | F6 (+1) | F6 (+1) |
| F3 | M6 (+3) | M6 (+3) | M6 (+3) |
| F4 | F5 (+3) | F4 (±0) | F6 (±0) |
| F5 | M6 (+3) | M6 (+2) | F2 (−2) |
| F6 | F2 (−2) | F5 (+2) | F4 (+1) |

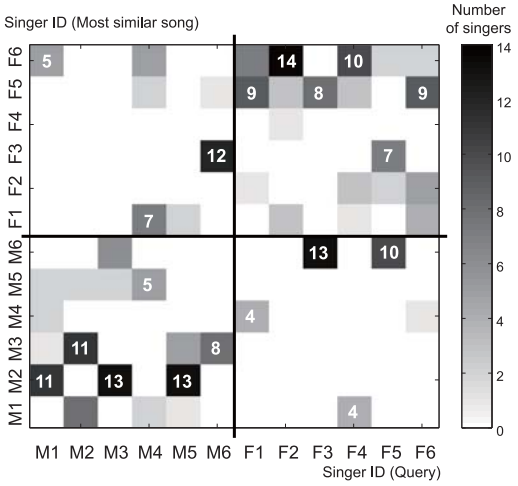


**Fig. 5**. Number of singers of the highest similarity song of each query (252 queries).

These results show that Hirai Ken (M6) and Hitoto Yo (F3) are similar when pitch-shifted by 3 semitones. In fact, they are well-known similar singers when pitch-shifted by 3 semitones. This suggests that the proposed method work well for the estimation of cross-gender similarity.

Figure 6 shows the mixing weights of a song "HitomiWoTojite" sung by Hirai Ken (M6) and its most similar song "MoraiNaki" sung by Hitoto Yo (F3) 3 semitones lower. The figure shows both song have high topic weights of topic 38 (the cluster number of the $k$-means algorithm).

### 3.3. Singer cloud

Figure 7 shows the singer clouds of topic 38 and 83. Topic 38 is high weight with both Hirai Ken (M6) and Hitoto Yo (F3), and topic 83 is high weight with only Hirai Ken (M6), as shown in Fig. 6. The size of each singer's name is defined by summing the same song's 7 mixing weights (*i.e.*, there are three names of each singer).

The results suggest that topic 38 has characteristics shared by Hirai Ken (M6), Hitoto Yo (F3) and Utada Hikaru (F5), and that topic 83 has characteristics shared by Hirai Ken (M6), Tokyo Jihen (F4) and GLAY (M4). Even though these two topics are shared by Hirai Ken, we found that they represent different factors of his singing voices.

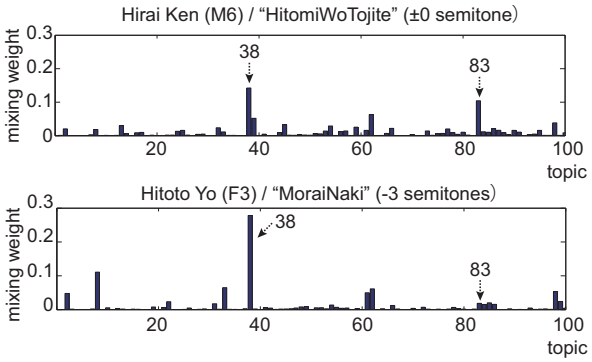


**Fig. 6**. Mixing weights of the similar song pair, Hirai Ken (M6) and Hitoto Yo (F3, 3 semitones lower). The topic 38 is high weight in both, and the topic 83 is high weight with only M6.

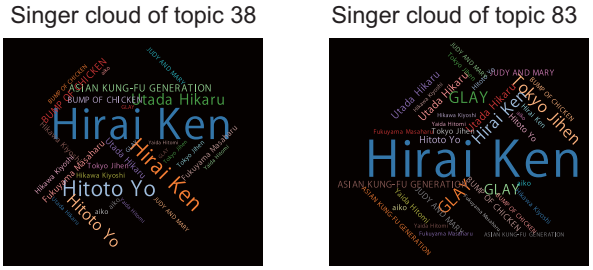


**Fig. 7**. Examples of topic visualization by the singer cloud. Topic 38 is high weight with both Hirai Ken (M6) and Hitoto Yo (F3), and topic 83 is high weight with only M6, as shown in Fig. 6.

## 4. CONCLUSIONS AND FUTURE WORK

This paper describes a vocal timbre analysis method based on latent Dirichlet allocation (LDA) where each song is represented as a weighted mixture of multiple topics that are shared by all singing voices. The paper also describes a method for estimating cross-gender vocal timbre similarity. While previous MIR works focused on retrieving only existing music, our MIR based on this cross-gender similarity can find songs whose pitch-shifted singing voices are similar to a query song. The experimental results showed that the mixing weights of LDA can be used for singer identification (see 3.1), cross-gender similarity estimation (see 3.2), and singer-cloud semantic visualization (see 3.3).

Since this paper focused on vocal timbre features, we plan to use $F_0$ information or other singing features as the next step. The future work will also include the use of a probabilistic topic model based on LDA [35, 47, 48] and a nonparametric Bayesian approach [48].

## 5. ACKNOWLEDGMENTS

This research was supported in part by OngaCrest, CREST, JST. The work reported in this paper used the Songle modules of Hiromasa Fujihara to estimate vocal LPMCC and $\Delta F_0$ from polyphonic audio signals. We thank Masahiro Hamasaki and Keisuke Ishida for their valuable advice to create the singer cloud.

## 6. REFERENCES


[1] A. Mesaros *et al.*, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in *Proc. of ISMIR 2007*, 2007.

[2] T. L. Nwe and H. Li, "Exploring vibrato-motivated acoustic features for singer identification," *IEEE Trans. on ASLP*, vol. 15, no. 2, pp. 519–530, 2007.

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

[4] W.-H. Tsai and H.-P. Lin, "Background music removal based on cepstrum transformation for popular singer identification," *IEEE Trans. on ASLP*, vol. 19, no. 5, pp. 1196–1205, 2011.

[5] M. Lagrange *et al.*, "Robust singer identification in polyphonic music using melody enhancement and uncertainty-based learning," in *Proc. of ISMIR 2012*, 2012.

[6] P. Żwan and B. Kostek, "System for automatic singing voice recognition," *J. Audio Eng. Soc*, vol. 56, no. 9, pp. 710–723, 2008.

[7] F. Maazouzi and H. Bahi, "Singing voice classification in commercial music productions," in *Proc. of ICICS*, 2011.

[8] B. Schuller *et al.*, "Vocalist gender recognition in recorded popular music," in *Proc. of ISMIR 2010*, 2010, pp. 613–618.

[9] F. Weninger *et al.*, "Combining monaural source separation with long short-term memory for increased robustness in vocalist gender recognition," in *Proc. of ICASSP 2011*, 2011, pp. 2196–2199.

[10] F. Weninger *et al.*, "Automatic assessment of singer traits in popular music: Gender, age, height and race," in *Proc. of ISMIR 2011*, 2011.

[11] K. Hirayama and K. Itou, "Discriminant analysis of the utterance state while singing," in *Proc. of ISSPIT 2012*, 2012, pp. 45–54.

[12] H. Mori *et al.*, "$F_0$ dynamics in singing: Evidence from the data of a baritone singer," *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 5, pp. 1068–1092, 2004.

[13] N. Minematsu *et al.*, "Prosodic analysis and modeling of nagauta singing to generate prosodic contours from standard scores," *IEICE Trans. Information and Systems*, vol. E87-D, no. 5, pp. 1093–1101, 2004.

[14] T. Saitou *et al.*, "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis," *Speech Communication*, vol. 46, pp. 405–417, 2005.

[15] Y. Ohishi *et al.*, "A stochastic representation of the dynamics of sung melody," in *Proc. ISMIR 2007*, 2007, pp. 371–372.

[16] E. Gómez and J. Bonada, "Automatic melodic transcription of flamenco singing," in *Proc. of CIM 08*, 2008.

[17] Y. Ohishi *et al.*, "A stochastic model of singing voice F0 contours for characterizing expressive dynamic components," in *Proc. of INTERSPEECH 2012*, 2012.

[18] S. W. Lee *et al.*, "Analysis for vibrato with arbitrary shape and its applications to music," in *Proc. of APSIPA ASC 2011*, 2011.

[19] R. Stables *et al.*, "Fundamental frequency modulation in singing voice synthesis," in *Lecture Notes in Computer Science*, 2012, vol. 7172, pp. 104–119.

[20] D. Ruinskiy and Y. Lavner, "An effective algorithm for automatic detection and exact demarcation of breath sounds in speech and song signals," *IEEE Trans. on ASLP*, vol. 15, pp. 838–850, 2007.

[21] T. Nakano *et al.*, "Analysis and automatic detection of breath sounds in unaccompanied singing voice," in *Proc. of ICMPC 10*, 2008.

[22] T. Nakano *et al.*, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Proc. of INTERSPEECH 2006*, 2006, pp. 1706–1709.

[23] C. Cao *et al.*, "An objective singing evaluation approach by relating acoustic measurements to perceptual ratings," in *Proc. of INTERSPEECH 2008*, 2008, pp. 2058–2061.

[24] Z. Jin *et al.*, "An automatic grading method for singing evaluation," in *Lecture Notes in Electrical Engineering*, 2012, vol. 128, pp. 691–696.

[25] W.-H. Tsai and H.-C. Lee, "Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features," *IEEE Trans. on ASLP*, vol. 20, no. 4, pp. 1233–1243, 2012.

[26] R. Daido *et al.*, "A system for evaluating singing enthusiasm for karaoke," in *Proc. of ISMIR 2011*, 2011, pp. 31–36.

[27] T. Kako and *et al.*, "Automatic identification for singing style based on sung melodic contour characterized in phase plane," in *Proc. ISMIR2009*, 2009, pp. 393–398.

[28] W.-H. Tsai and H.-M. Wang, "Towards automatic identification of singing language in popular music recordings," in *Proc. of ISMIR 2004*, 2004, pp. 568–576.

[29] J. Schwenninger *et al.*, "Language identification in vocal music," in *Proc. of ISMIR 2006*, 2006, pp. 377–379.

[30] V. Chandraskehar *et al.*, "Automatic language identification in music videos with low level audio and visual features," in *Proc. of ICASSP 2011*, 2011, pp. 5724–5727.

[31] M. Mehrabani and J. H. L. Hansen, "Language identification for singing," in *Proc. of ISMIR 2006*, 2006, pp. 4408–4411.

[32] D. M. Blei *et al.*, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[33] Eric Brochu and Nando de Freitas, ""name that song!": A probabilistic approach to querying on music and text," in *Proc. of NIPS2002*, 2002.

[34] D. J. Hu and L. K. Saul, "A probabilistic topic model for unsupervised learning of musical key-profiles," in *Proc. of ISMIR2009*, 2009.

[35] D. J. Hu and L. K. Saul, "A probabilistic topic model for music analysis," in *Proc. of NIPS-09*, 2009.

[36] R. Takahashi *et al.*, "Building and combining document and music spaces for music query-by-webpage system," in *Proc. of Interspeech 2008*, 2008, pp. 2020–2023.

[37] P. Symeonidis *et al.*, "Ternary semantic analysis of social tags for personalized music recommendation," in *Proc. of ISMIR 2008*, 2008.

[38] M. Hoffman *et al.*, "Content-based musical similarity computation using the hierarchical Dirichlet process," in *Proc. of ISMIR2008*, 2008.

[39] E. Pampalk, "Islands of music: Analysis, organization, and visualization of music archives," *Master's thesis, Vienna University of Technology*, 2001.

[40] P. Smaragdis *et al.*, "Topic models for audio mixture analysis," in *Proc. of the NIPS workshop on applications for topic models: text and beyond*, 2009.

[41] A. Mesaros *et al.*, "Latent semantic analysis in sound event detection," in *Proc. of EUSIPCO 2011*, 2011, pp. 1307–1311.

[42] S. Kim *et al.*, "Latent acoustic topic models for unstructured audio classification," *APSIPA Trans. on Signal and Information Processing*, vol. 1, pp. 1–15, 2012.

[43] K. Imoto *et al.*, "Acoustic scene analysis based on latent acoustic topic and event allocation," in *Proc. of MLSP 2013*, 2013.

[44] M. Goto *et al.*, "Songle: A web service for active music listening improved by user contributions," in *Proc. ofISMIR 2011*, 2011, pp. 311–316.

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

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

[47] S. Rogers *et al.*, "The latent process decomposition of cDNA microarray data sets," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 2, no. 2, pp. 143–156, 2005.

[48] K. Yoshii and M. Goto, "A nonparametric bayesian multipitch analyzer based on infinite latent harmonic allocation," *IEEE Trans. on ASLP*, vol. 20, no. 3, pp. 717–730, 2012.