A MUSIC INFORMATION RETRIEVAL SYSTEM BASED ON SINGING VOICE TIMBRE

Hiromasa Fujihara and **Masataka Goto**

National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki 305-8568, Japan
{h.fujihara, m.goto} [at] aist.go.jp

ABSTRACT

We developed a music information retrieval system based on singing voice timbre, i.e., a system that can search for songs in a database that have similar vocal timbres. To achieve this, we developed a method for extracting feature vectors that represent characteristics of singing voices and calculating the vocal-timbre similarity between two songs by using a mutual information content of their feature vectors. We operated the system using 75 songs and confirmed that the system worked appropriately. According to the results of a subjective experiment, 80% of subjects judged that compared with a conventional method using MFCC, our method finds more appropriate songs that have similar vocal timbres.

1 INTRODUCTION

This paper describes a music information retrieval (MIR) system that searches for songs that have similar voice timbres of vocals to a query song presented by a user. By using this system, we can find a song by using its musical content in addition to traditional bibliographic information. This kind of retrieval is called content-based MIR, and our system, which focuses on singing voices as content, falls into this category.

There is a growing demand for such content-based MIR. Because of rapid and widespread diffusion of portable audio players and online music stores, many users can have an access to a large amount of music tracks and listen to any music they want anytime, anywhere. This trend has triggered a demand for an MIR that takes favorite songs from a large amount of music and uses them to discover new songs that the user has never heard before. When the query of the target song is not known and only vague information such as "preference" is available, the conventional method of searching for music that only uses bibliographic information is useless.

Although a number of studies on content-based MIR have been undertaken [2, 8, 1, 3, 9, 12, 10, 4, 11], they use low-level acoustic features, such as MFCC, spectral centroid, rolloff, and flux, for expressing musical content and do not use higher-level features such as the timbre of vocals.

We therefore developed an MIR system that can retrieve songs by using vocal timbres. To achieve this sys-
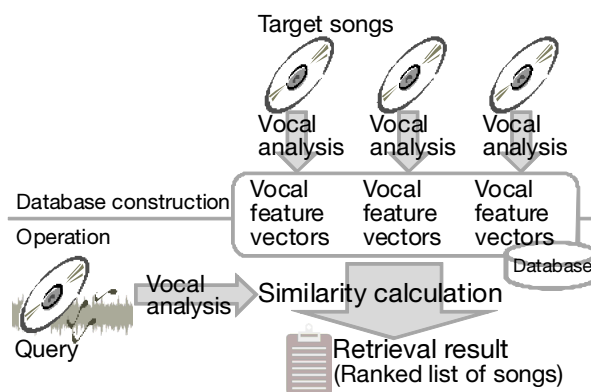
Figure 1. Overview of our MIR system.

tem, we have to extract, from the polyphonic audio signal of a song, vocal-based feature vector that represent the vocal characteristics and to calculate the vocal-timbre similarity between two songs by using the feature vectors. The vocal-based feature vectors were also used in our singer identification method proposed earlier [5]. We used the mutual information content as their similarity measure.

2 ARCHITECTURE OF THE SYSTEM

Among many songs registered in a database, our system searches for songs that have similar singing voice timbres to a song query given by a user. The overview of this system is shown in Figure 1. The system consists of a database construction (audio analysis) part and an operation (song retrieval) part. In the database construction part, target songs are stored in the database after being ripped or downloaded, and then each song is analyzed to extract feature vectors that represent vocal characteristics.

When the user enters a (favorite) song as a query to the system, the system analyses the query song and extracts feature vectors that also represent the query's vocal characteristics. The system then calculates the vocal-timbre similarity between the query song and each song in the database and shows the list of songs with high similarity.

3 IMPLEMENTATION OF THE SYSTEM

To implement the MIR system described in Section 2, we have to define the vocal-based feature vector and the vocal-timbre similarity measure.

3.1 Feature Extraction

To calculate feature vectors that represent vocal characteristics, we use a feature extraction method used in our singer identification method [5]. This method can reduce a negative influence of accompaniment sounds which are mixed with the singing voice in a song. This feature vector has a better representation of vocal characteristics than a feature vector like MFCC that just represents a mixture of accompaniment sounds and the singing voice.

This method consists of the following three parts: accompaniment sound reduction, feature extraction, and reliable frame selection. To reduce the negative influence of accompaniment sounds, the accompaniment sound reduction part first segregates and resynthesizes the singing voice from polyphonic audio signals on the basis of its harmonic structure. The feature extraction part then calculates feature vectors from the segregated singing voice. The reliable frame selection part selects reliable vocal regions (frames) from the feature vectors and removes unreliable regions that do not contain vocals or are highly influenced by accompaniment sounds.

3.1.1 Accompaniment Sound Reduction

For the accompaniment sound reduction part, we use a melody resynthesis technique that consists of the following three steps:

1. Estimating the fundamental frequency (F0) of the vocal melody using Goto's PreFEst [6].
2. Extracting the harmonic structure corresponding to the melody.
3. Resynthesizing the audio signal corresponding to the melody using sinusoidal synthesis.

We use Goto's PreFEst [6] for estimating the F0 of the melody. PreFEst estimates the most predominant F0 in frequency-range-limited sound mixtures. Since the melody line tends to have the most predominant harmonic structure in middle- and high-frequency regions, we can estimate the F0 of the melody by applying PreFEst with appropriate frequency-range limitation.

By using the estimated F0, we then extract the amplitude of the fundamental frequency component and harmonic components. For each component, we allow r cent error and extract the local maximum amplitude in the allowed area. The frequency $F_l^{(t)}$ and amplitude $A_l^{(t)}$ of the l th overtone ($l = 1, \ldots, L$) at time (t) can be represented as

$$
\begin{aligned}
F_l^{(t)} &= \underset{F}{\operatorname{argmax}} |S^{(t)}(F)| \\
&\quad (l\overline{F}^{(t)}(1 - 2^{\frac{r}{1200}}) \le F \le l\overline{F}^{(t)}(1 + 2^{\frac{r}{1200}})), \tag{1} \\
A_l^{(t)} &= |S^{(t)}(F_l)|, \tag{2}
\end{aligned}
$$

where $S^{(t)}(F)$ denotes the complex spectrum, and $\overline{F}^{(t)}$ denotes F0 estimated by the PreFEst. In our experiments, we set r to 20.

Finally, we use a sinusoidal model to resynthesize the audio signal of the melody by using the extracted harmonic structure, $F_l^{(t)}$ and $A_l^{(t)}$. Changes in phase are approximated using a quadratic function so that a frequency can change linearly. Changes in amplitude are also approximated using a linear function. The resynthesized audio signals, $s(k)$, are expressed as

$$
s(k) = \sum_{l=1}^{L} s_l(k), \tag{3}
$$

$$
s_l(k) = \left\{ (A_l^{(t+1)} - A_l^{(t)})\frac{k}{K} + A_l^{(t)} \right\} \sin(\theta_l(k)), \tag{4}
$$

$$
\theta_l(k) = \frac{\pi(F_l^{(t+1)} - F_l^{(t)})}{K}k^2 + 2\pi F_l^{(t)}k + \theta_{l,0}, \tag{5}
$$

where k represents a time in units of seconds and $k = 0$ corresponds to the time (t), K represents the duration between (t) and ($t+1$) in units of seconds, and $\theta_{l,0}$ means the initial phase.

3.1.2 Feature Extraction

From the resynthesized audio signals, we calculate feature vectors consisting of two features.

- LPC-derived mel cepstral coefficients (LPMCCs)
  It is known that the individual characteristics of speech signals are expressed in their spectral envelopes. We use LPMCCs as spectral feature because we have reported that, in the context of singer identification, LPMCCs represent vocal characteristics better than mel-frequency cepstral coefficients (MFCCs), which are widely used for music modeling [5].

- ΔF0s
  We use ΔF0s which represent the dynamics of F0's trajectory, because a singing voice tends to have temporal variations in its F0 in consequence of vibrato and such temporal information is expected to express the singer's characteristics.

3.1.3 Reliable Frame Selection

Because the F0 of the melody is simply estimated as the most predominant F0 in each frame [6], the resynthesized audio signals may contain both the vocal sound in singing sections and other instrument sounds in interlude sections. The feature vectors obtained from them therefore include unreliable regions (frames) where other accompaniment sounds are predominant. The reliable frame selection part removes such unreliable regions and uses only the reliable regions for calculating similarity. In order to achieve this, we introduce two kinds of Gaussian mixture models (GMMs), a vocal GMM λ_V and a non-vocal GMM λ_N. The vocal GMM λ_V is trained on feature vectors extracted from singing sections, and the non-vocal GMM λ_N is trained on those extracted from interlude sections. Given a feature vector $\mathbf{x}$, the likelihoods for the two GMMs, $p(\mathbf{x}|\lambda_V)$ and $p(\mathbf{x}|\lambda_N)$, represent how the feature vector $\mathbf{x}$ is like a vocal or a (non-vocal) instrument, respectively. We therefore determine whether the feature vector $\mathbf{x}$ is reliable or not by using the following equation:

$$
\log p(\mathbf{x}|\lambda_V) - \log p(\mathbf{x}|\lambda_N) \underset{\text{not-reliable}}{\overset{\text{reliable}}{\underset{<}{\gtrless}}} \eta, \tag{6}
$$

where η is a threshold.

It is difficult to decide a universal constant threshold for a variety of songs because if a threshold is too high for some songs, there are too few reliable frames to appropriately calculate the similarity. We therefore determine the threshold dependent on songs so that $\alpha\%$ of the whole frames in the song are selected as reliable frames. Note that most of the non-vocal frames are rejected in this selection step.

3.2 Similarity Calculation

We choose mutual information content to be the similarity measure between two songs. First, we model a probability distribution of the feature vectors for a song using GMM and estimate the parameters of the GMM for each song by using the EM algorithm. Then, we calculate the similarity between song X and song Y, $d_{\mathrm{CE}}(X, Y)$, by using the following equation:

$$
\begin{aligned}
d_{\mathrm{CE}}(X,Y) \quad = \quad & \log \prod_i \frac{\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}_i; \theta_X)}{\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}_i; \theta_Y)} \\
& + \log \prod_j \frac{\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{y}_j; \theta_Y)}{\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{y}_j; \theta_X)} \quad (7)
\end{aligned}
$$

where $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ represent feature vectors of reliable frames in song X and song Y, respectively, θ_X and θ_Y represent GMM parameters of song X and song Y, respectively, and $\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x}; \theta)$ represents the likelihood of GMM with parameter θ.

4 OPERATION OF THE SYSTEM

Figure 2 shows a screenshot of the system. For constructing the vocal and the non-vocal GMMs, we selected 25 songs from "RWC Music Database: Popular Music" (RWC-MDB-P-2001) [7]. In the system database, we registered other 75 songs from the RWC-MDB-P-2001, which were not used for constructing those GMMs. In Figure 2, a song "PROLOGUE" (RWC-MDB-P-2001 No.7) sung by a female singer "Tomomi Ogata" is given as a query. Given a query song, it took about 20 seconds to calculate similarities and output a ranked list of retrieved songs. As shown in Figure 2, the retrieval result consists of the rank, song title, artist name, and similarity.

In most of songs retrieved given various queries, vocal timbres of the top 10 songs were generally similar to that of each query song in our experience. For example, in Figure 2, the top 21 songs were sung by female singers, and vocal timbres of the top 15 songs shown in this figure were similar to the query song. Note that four songs from "Tomomi Ogata" who was the singer of the query took first, second, 10th, and 12th places. This is because the singing style of the 10th and 12th songs were different from that of the first and second songs and the query.

5 SUBJECTIVE EXPERIMENT

We conducted a subjective experiment to compare our system using the proposed vocal-based feature vector with a baseline system using the traditional MFCCs.

5.1 Experimental Procedure

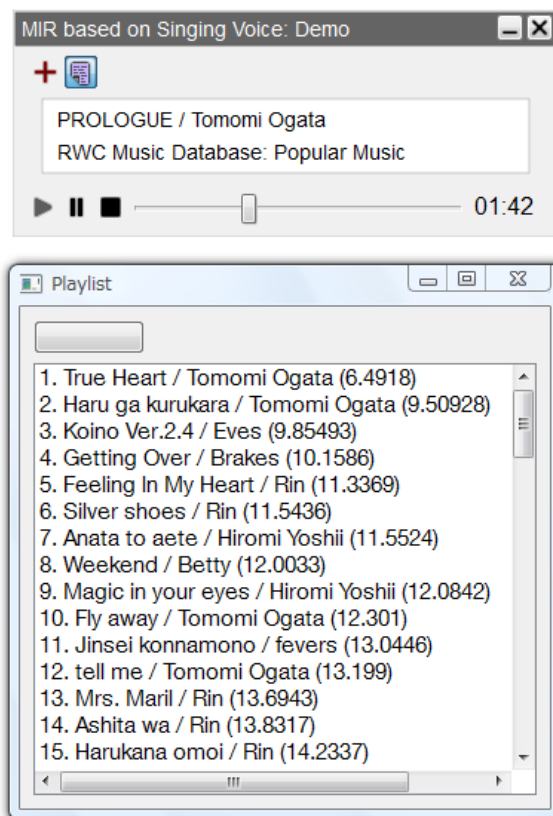Six university students (two male and four female) participated in this experiment. Those subjects had not re-

Figure 2. Screenshot of the system.

ceived any professional training in music. They first listened to a set of three songs — a query song (song X), the top ranked song retrieved by our system (song A/B), and the top ranked song retrieved by the baseline system (song B/A) —, and then judged which song was more similar to the query song (Figure 3). The subjects did not know which song was retrieved by our system and the song order of A and B is randomized. We allowed them to listen to those songs in any order as much as they like.

We selected ten query songs from the system database considering that these songs have a variety of genders and genres. For each query song, we asked the subjects the following questions:

Question 1 When comparing the singing voice timbres of song A and B, which song resembles song X?
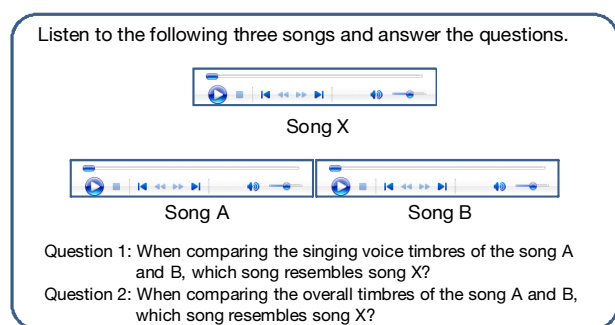
Question 2 When comparing the overall timbres of song A and B, which song resembles song X?

5.2 Results and Discussion

Figures 4 and 5 show the results of the experiment. 80% of the responses for 10 songs judged that the singing voice timbre by our method was more similar to that of the query song (Figure 4). On the other hand, 70% of the responses judged that the overall timbre by the baseline method was more similar to that of the query song (Figure 5). Accordingly, we confirmed that our method can reduce the influence of accompaniment sounds and find songs by using vocal timbres. We also found that our method can retrieve not only songs with similar vocal timbres (or by same singer) but also songs with similar singing styles.

Table 1. Query songs and the corresponding retrieved songs used for the subjective experiment: The three-digit number indicates the piece number of the RWC Music Database (RWC-MDB-P-2001). Given each query song, the top ranked song by the baseline method (MFCC) and the top ranked song by our method are shown on the same line.

Query song			Retrieved (top ranked) song	
Piece #	Gender	Language	MFCC	Our method
004	M	Japanese	031	082
010	F	Japanese	016	054
029	M	Japanese	017	012
035	F	Japanese	036	094
045	M	Japanese	090	042
053	F	Japanese	062	014
072	M	Japanese	071	076
077	F	Japanese	071	067
092	F	English	024	086
098	M	English	009	085

Listen to the following three songs and answer the questions.

Song X

Song A   Song B

Question 1: When comparing the singing voice timbres of the song A and B, which song resembles song X?

Question 2: When comparing the overall timbres of the song A and B, which song resembles song X?

Figure 3. Interface used for the subjective experiment.

For example, when the song RWC-MDB-P-2001 No.53 was used as a query, both our method and the baseline method retrieved the top ranked songs by the singer same with the query, but 5 out of 6 subjects judged that the song by our method was more similar to the query song in terms of the singing voice timbre.

6 CONCLUSION

We have described a vocal-timbre-based MIR system that uses a method for extracting vocal-based feature vectors from polyphonic audio signals and a method for measuring the vocal-timbre similarity between two songs. We tested the system on 75 songs and found that the system was useful for retrieving similar-vocal songs. Our experimental results with six subjects showed the effectiveness of our similarity measure.

We used the mutual information content as our similarity measure because of its symmetry and statistical simplicity. Although it was effective, it requires a high computational cost and a large storage because we have to use all the feature vectors to compute it. In the future, we plan to try other similarity measures, including the earth mover's distance (EMD) [3], to lower the computational cost. We also plan to integrate this system with other content-based MIR methods to give users a wider variety
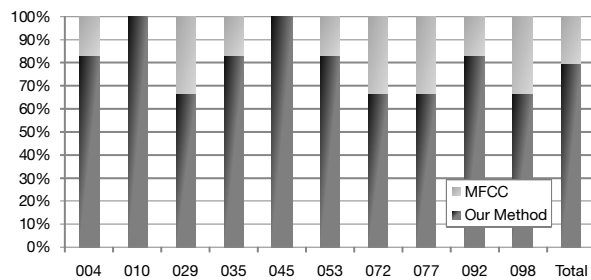
Figure 4. Evaluation result: Question 1: singing voice timbre.
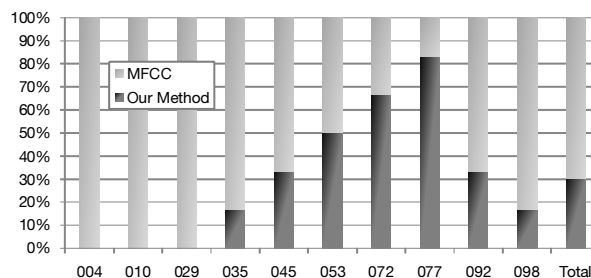
Figure 5. Evaluation result: Question 2: overall timbre.

of retrieval methods.

7 REFERENCES

[1] E. Allamanche, J. Herre, O. Hellmuth, T. Kastner, and C. Ertel. A multiple feature model for musical similarity retrieval. In *Proc. ISMIR2003*, pages 217–218, 2003

[2] J.-J. Aucouturier and F. Pachet. Music similarity measures: What's the use? In *Proc. ISMIR2002*, pages 157–163, 2002.

[3] A. Berenzweig, B. Logan, D. P. W. Ellis, and B. Whitman. A large-scale evaluation of acoustic and subjective music similarity measures. In *Proc. ISMIR2003*, pages 99–105, 2003.

[4] A. Flexer, F. Gouyon, S. Dixon, and G. Widmer. Probabilistic combination of features for music classification. In *Proc. ISMIR2006*, pages 111–114, 2006.

[5] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. Singer identification based on accompaniment sound reduction and reliable frame selection. In *Proc. ISMIR2005*, pages 329–336, 2005.

[6] 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):311–329, 2004.

[7] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC Music Database: Popular, classical, and jazz music databases. In *Proc. ISMIR2002*, pages 287–288, 2002.

[8] B. Logan. Content-based playlist generation: Exploratory experiments. In *Proc. ISMIR2002*, pages 295–296, 2002.

[9] M.F. McKinney and J. Breebaart, Features for audio and music classification. In *Proc. ISMIR2003*, pages 151–158, 2003.

[10] E. Pampalk, A. Flexer, and G. Widmer. Improvements of audio-based music similarity and genre classification. In *Proc. ISMIR2005*, pages 628–633, 2005.

[11] T. Pohle, P. Knees, M. Schedl, and G. Widmer. Independent component analysis for music similarity computation. In *Proc. ISMIR2006*, pages 228–233, 2006.

[12] G. Tanetakis, J. Gao, and P. Steenkiste. A scalable peer-to-peer system for music content and information retrieval. In *Proc. ISMIR2003*, pages 209–214, 2003.