AUTOMATIC DISCRIMINATION BETWEEN SINGING AND SPEAKING VOICES FOR A FLEXIBLE MUSIC RETRIEVAL SYSTEM

Yasunori OHISHI† *Masataka GOTO*†† *Katunobu ITOU*††† *Kazuya TAKEDA*†

†Graduate School of Information Science, Nagoya University
††National Institute of Advanced Industrial Science and Technology (AIST)
†††Faculty of Computer and Information Sciences, Hosei University

1. ABSTRACT

This paper describes a music retrieval system that enables a user to retrieve a song by two different methods: by singing its melody or by saying its title. To allow the user to use those methods seamlessly without changing a voice input mode, a method of automatically discriminating between singing and speaking voices is indispensable. We therefore first investigated measures that characterize differences between singing and speaking voices. From subjective experiments, we found that even short term characteristics such as the spectral envelope represented as MFCC can be used as a discrimination cue, while the temporal structure is the most important cue when longer signals are given. According to these results, we developed the automatic method of discriminating between singing and speaking voices by combining two measures: MFCC and an F0 (voice pitch) contour. Based on this method, we built the music retrieval system that can accept both singing voices for the melody and speaking voices for the title.

2. MUSIC RETRIEVAL SYSTEM

We propose a music retrieval system that retrieves a song by query-by-humming for singing voice or by dictating its title by automatic speech recognition (ASR) for speaking voice. The main modules of this system are described below.

2.1. Speech Discriminator

The role of this module is automatic discrimination between singing and speaking voices. From subjective experiments, approximately one second is enough for humans to discriminate between singing and speaking voices. Even with a 200-ms signal, discrimination accuracy is more than 70%. This suggests that not only temporal characteristics corresponding to rhythm and melody but also such short-term features as spectral envelopes carry discriminative cues. Moreover, to examine how listeners distinguish between these two voices, we conducted subjective experiments with singing and speaking voice stimuli whose voice quality and prosody were systematically distorted by using signal processing techniques. The experimental results suggest that spectral and prosodic (temporal) cues complementarily contributed to perceptual judgments [1]. Therefore, we propose an automatic vocal style discriminator by using two different measures — short-term and long-term feature measures. The short-term feature measure exploits the spectral envelope represented by using Mel-Frequency Cepstrum Coefficients (MFCC) and their derivatives (ΔMFCC). The long-term feature measure exploits the dynamics of F0 extracted from voice signals.

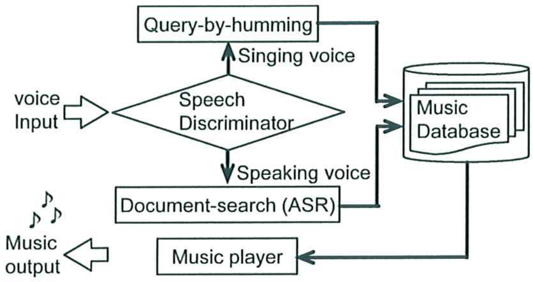

Fig. 1. Flowchart of the proposed music retrieval system

2.1.1. *Short-term spectral feature measure*

To measure a spectral envelope, Mel-Frequency Cepstrum Coefficients (MFCC) and their derivatives (ΔMFCC), which are successfully used for envelope extraction in speech recognition applications, were used. Every 10 ms, MFCC are calculated for 25-ms hamming windowed frames. MFCC is used up to the 12th coefficients. ΔMFCC is calculated as regression parameters over five frames.

2.1.2. *Long-term feature measure*

F0 is estimated by using the predominant-F0 estimation method of Goto et al. [2]. Using this method, we determined the F0 value for every 10 ms. ΔF0 is calculated by five-point regression, as in the MFCC case.

2.1.3. *Training the discriminative model*

The distribution of feature vector (MFCC and ΔMFCC, ΔF0) are represented by 16-mixture Gaussian Mixture Models (GMM) trained on the training set using the EM algorithm for both singing and speaking voice signals. The variances of distributions were modeled by a diagonal covariance matrix. Discrimination was performed through the maximum likelihood principle:

$$\hat{d} = \underset{d=\text{singing, speaking}}{\text{argmax}} \frac{1}{N} \sum_{t=1}^{T} \log p(\mathbf{x}_n; \Lambda_d), \qquad (1)$$

where $\mathbf{x}_t$ is the tth feature vector, T is input signal length and $\Lambda_d (d = \text{singing, speaking})$ are GMM parameters for the distribution of feature vectors. Function p calculates posterior probability by using GMM parameters.

2.2. Query-By-Humming

Query-by-Humming enables one to retrieve the title of a musical piece by humming or singing its melody using sounds like "la-la-la...". In other words, humming or singing a melody becomes the search key for finding a musical piece with that melody. Fundamentally, a reference pitch pattern from a query and an input pitch pattern from a database are extracted and fed into the matching method. We implemented a music retrieval method by a humming query based on start frame feature dependent continuous dynamic programming [3].

2.3. Document-search (ASR)

This module retrieves a song based on its title recognized by ASR for speaking voices. The speech recognition module used an open source Large Vocabulary Continuous Speech Recognition (LVCSR) engine, *Julian-3.4.2* [4]. We used a gender independent acoustic model of Phonetically-Tied Mixture (PTM) with 3,000 states (129 codebooks) and 64 Gaussians and created some recognition grammars like " Please listen to <Song Title>". The dictionary for speech recognition contains 142 words including 33 artists names, 100 music titles and command words used for retrieval.

3. EVALUATION OF PROPOSED SYSTEM

We prepared 100 WAV files (*"RWC Music Database: Popular Music"* (RWC-MDB-P-2001) [5]) as a music database. Therefore, the prototype of this system can retrieve a song from 100 music songs. And we used an annotation database[6] labeled melodies of the music database manually to match with a reference pattern from a query for query-by-humming.

3.1. Singing voice database

We used 7,500 voice samples excerpted from the "AIST Humming Database"[7]. Those samples, each about 12.0 seconds long, consist of 3,750 samples of singing voices and 3,750 samples of humming voices recorded from 75 subjects (37 males, 38 females). At an arbitrary tempo without musical accompaniment, each subject sang two excerpts from the chorus and the first verses of 25 songs in different genres (50 sound samples). The songs were selected from the music database as above. Singing voices (1,875 samples) that sang the first verses are used for training GMM of singing voice. Singing and humming voices (a total of 3,750 samples) that sang the chorus are used as queries for the proposed system.

3.2. Speaking voice database

We used 3,750 voice samples excerpted from the "AIST Humming Database"[7]. Those samples, each about 7.0 seconds long, consist of speaking voices recorded from 75 subjects (37 males, 38 females). Each subject read the lyrics of two excerpts from the chorus and the first verses of same 25 songs as singing voice database. Speaking voices (1,875 samples) that read the lyrics of the first verses are used for training GMM of speaking voice. Also, we prepared 60 utterances for the title and artist name of the target song and the command words, e.g., "please listen to <Song title>" recorded from 4 males and 2 females. We used those utterances as queries for the proposed system.

Table 1. Discrimination and search rates for singing, humming and speaking voices: Search rate means the average rate including correct songs in the top 10 of 100 songs.

	Singing	Humming	Speaking
Discrimination rate	96.2%	98.0%	100%
Search rate	50.8%	52.1%	96.7%

3.3. Experimental Results

Table 1 shows discrimination and search rates for each voice with a large voice database as described above. The average discrimination rate between singing (including humming) and speaking voices is 98.1%. It can be seen that two measures (short-term spectral and long-term feature measures) can effectively capture the signal features that discriminate between singing and speaking voices. The average search rates of correct songs in the top 10 of 100 songs by query-by-humming and ASR for song titles are 51.5% and 96.7% respectively. Especially, the retrieval rate by query-by-humming is low. Some singers could retrieve a song title by singing its melody correctly, but others could not. And, some songs could be retrieved easily by all singers because of melody simplicity, but others could not.

4. CONCLUSION

We proposed a music retrieval system that enables a user to retrieve a song by singing its melody or by saying its title. Firstly, this system discriminates between singing and speaking voices automatically for an input voice, and retrieves a song by query-by-humming for singing voice or by dictating the song title by ASR for speaking voice. Experimental results show that our system is able to discriminate between singing and speaking voices with 98.1%. The average retrieval rates of correct songs in the top 10 of 100 songs by query-by-humming and ASR for song titles are 50.5% and 96.7% respectively. In the future, we plan to propose a new query-by-humming method adaptable to all sorts of music (melody) and singing ability.

5. REFERENCES

[1] Y. Ohishi, M. Goto, K. Itou, and K. Takeda, "On the human capability and acoustic cues for discriminating the singing and the speaking voices," in *Proc. ICMPC 2006*, pp. 1831–1837.

[2] M. Goto, K. Itou, and S. Hayamizu, "A real-time filled pause detection system for spontaneous speech recognition," in *Proc. Eurospeech 1999*, pp. 227–230.

[3] T. Nishimura et al., "Music signal spotting retrieval by a humming query using start frame feature dependent continuous dynamic programming," *The Special Interest Group Notes of IPSJ (MUS)*, vol. 2001, no. 103, pp. 57–62, (in Japanese).

[4] T. Kawahara et al., "Recent progress of open-source LVCSR engine julius and japanese model repository; software of continuous speech recognition consortium," in *Proc. ICSLP 2004*, vol. 2, pp. 3069–3072.

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

[6] M. Goto, "Aist annotation for the RWC music database," in *Proc. ISMIR 2006*.

[7] M. Goto and T. Nishimura, "Aist humming database:music database for singing research," *The Special Interest Group Notes of IPSJ (MUS)*, vol. 2005, no. 82, pp. 7–12, (in Japanese).