Music Information Retrieval from a Singing Voice Based on Verification of Recognized Hypotheses

Motoyuki Suzuki, Toru Hosoya, Akinori Ito and Shozo Makino
Graduate School of Engineering, Tohoku University
6-6-05, Aramaki-Aza-Aoba, Aoba-ku, Sendai, 980-8579, JAPAN
{moto, thosoya, aito, makino}@makino.ecei.tohoku.ac.jp

Abstract

Several music information retrieval (MIR) systems have been developed which retrieve musical pieces by the user's singing voice. All of these systems use only melody information for retrieval, although lyrics information is also useful for retrieval. In this paper, we propose an MIR system that uses both melody and lyrics information in the singing voice.

The MIR system verifies hypotheses output by a lyrics recognizer from a melodic point of view. Each hypothesis has time alignment information between the singing voice and recognized text, and the boundaries of each note can be estimated using the information. As a result, melody information is extracted from the singing voice. On the other hand, the melody information can be calculated from the musical score of the song because the recognized text must be a part of the lyrics of the song. The hypothesis is verified by calculating the similarity between the two types of melody information.

From the experimental results, the verification method increased the retrieval accuracy. Especially, it was very effective when the number of words in the user's singing voice was small. The proposed method increased the retrieval accuracy from 81.3% to 87.4% when the number of words was only three.

Keywords: MIR from singing voice, verification of recognized hypotheses, lyrics recognition.

1. Introduction

Recently, several music information retrieval (MIR) systems that use a user's singing voice as a retrieval key have been researched (for example, MIRACLE[1], SoundCompass[2], and our proposed method[3, 4]). These systems use melody information in the user's singing voice, however, the lyrics information is not taken into consideration.

Lyrics information is very useful for MIR systems. In a preliminary experiment, a retrieval key consisting of three Japanese letters narrowed hypotheses into five songs on average, and the average number of retrieved songs was 1.3 when five Japanese letters were used as a retrieval key. Note that 161 Japanese songs were used as the database, and a part of the correct lyrics was used as the retrieval key in this experiment.

In order to develop an MIR system that uses melody and lyrics information, the lyrics recognition method from a singing voice has been proposed[5]. This method uses a finite state automaton as a language model. It gave 77.4% word accuracy, and the retrieval accuracy given by the system achieved 85.9%.

Conventional MIR systems cannot use a singing voice as a retrieval key. One of the biggest problems is how to split the input singing voice into musical notes. Traditional MIR systems[6, 2, 7] assume that a user hums with plosive phonemes, such as phonemes /ta/ or /da/, because the hummed voice can be split into notes only using power information. On the other hand, recent MIR systems can split the input singing voice into musical notes, however, they often give inaccurate information. It is difficult for them to discriminate between one note with long duration and two notes with the same pitch because these systems split the input singing voice using pitch and power information. It is very hard to split the singing voice into musical notes without linguistic information.

The lyrics recognizer outputs several hypotheses as a recognition result. Each hypothesis has time alignment information between the singing voice and recognized text. A hypothesis can be verified using the time alignment information from a melodic point of view. In this paper, we propose a new MIR system using lyrics and melody information based on verification of recognized hypotheses.

2. Verification Method of Recognized Hypotheses

2.1. Overview of the System

Figure 1 shows an outline of the proposed MIR system. First, a user's singing voice is input to the lyrics recognizer, and the top N hypotheses with higher recognition score are output.

Each hypothesis h has the following information:

- Song name $S(h)$
- Recognized text $W(h)$
  It must be a part of the lyrics of the song $S(h)$.

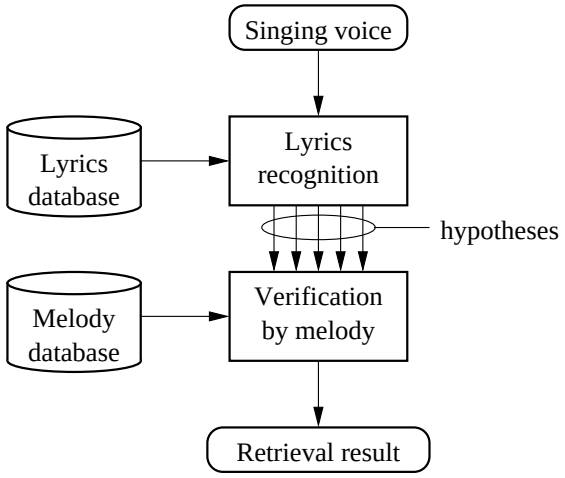

Figure 1. Outline of the MIR system using lyrics and melody

- Recognition score $R(h)$
- Time alignment information $F(h)$
  For all phonemes in the recognized text, frame numbers of the start frame and the end frame are output.

For a hypothesis h, the tune corresponding to the recognized text $W(h)$ can be obtained from the database because $W(h)$ must be a part of the lyrics of $S(h)$[5]. The melody information, which is defined as a relative pitch and relative IOI (Inter-Onset Interval) of each note, can be calculated from the tune. On the other hand, the melody information can be extracted using the estimated pitch sequence of the singing voice and $F(h)$. If the hypothesis h is correct, the two types of information should be similar. The verification score is defined as the similarity between the two types of information.

Finally, the total score is calculated from the recognition score and the verification score, and the hypothesis with the highest total score is output as a retrieval result.

2.2. Extraction of Melody Information from a Singing Voice

Relative pitch Δf_n and relative IOI Δt_n of a note n are extracted from the singing voice. In order to extract this information, boundaries between notes are estimated from time alignment information $F(h)$.

Figure 2 shows an example of the estimation procedure. For each song in the database, a correspondence table is made from the musical score of the song in advance. This table describes all of the correspondences between phonemes in the lyrics and notes in the musical score (for example, the i-th note of the song corresponds to phonemes from j to k).

When the singing voice and the hypothesis h are given, boundaries between notes are estimated from the time alignment information $F(h)$ and the correspondence table. The

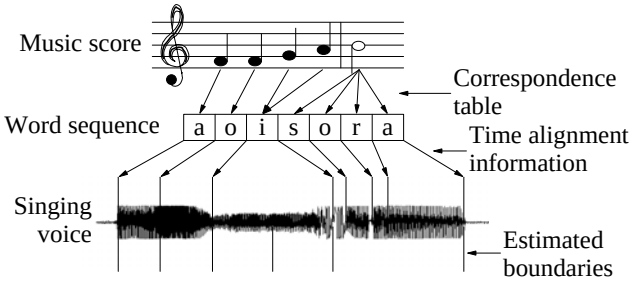

Figure 2. Example of estimation of boundaries between notes

phoneme sequence corresponding to the note n can be obtained from the correspondence table, and the start frame of n is obtained as the start frame of the first phoneme from $F(h)$. In the same way, the end frame of n is obtained as the end frame of the last phoneme.

After estimation of boundaries, pitch sequence is calculated by the praat[8] system frame-by-frame, and the pitch of the note is defined as the median of the pitch sequence corresponding to the note. IOI of the note is obtained as the duration between boundaries.

Finally, the pitch and IOI of the note n are translated into relative pitch Δf_n and relative IOI Δt_n using the two equations:

$$\Delta f_n = \log_2 \frac{f_{n+1}}{f_n} \quad (1)$$

$$\Delta t_n = \log_2 \frac{t_{n+1}}{t_n} \quad (2)$$

where, f_n and t_n are pitch and IOI of the n-th note respectively.

Note that estimated boundaries using the hypothesis are different from that using another hypothesis. Therefore, different melody information will be extracted using another hypothesis from the same singing voice.

2.3. Calculation of Verification Score

The verification score $V(h)$ corresponding to a hypothesis h is defined as the similarity between melody information extracted from the singing voice and the tune.

At first, relative pitch $\Delta \hat{f}_n$ and relative IOI $\Delta \hat{t}_n$ are calculated from the tune corresponding to the recognized text $W(h)$, and the verification score $V(h)$ is calculated by

$$V(h) = \frac{1}{N-1} \sum_{n=1}^{N-1} \Big\{ w_1 (|\Delta \hat{t}_n - \Delta t_n|) + (1 - w_1)(|\Delta \hat{f}_n - \Delta f_n|) \Big\} \quad (3)$$

where, N denotes the number of notes in the tune, and w_1 denotes a pre-defined weighting factor.

The total score $T(h)$ is calculated by Eq. (4) for each hypothesis h, and the final result H is selected by Eq. (5):

$$T(h) = w_2 R(h) - (1 - w_2) V(h) \quad (4)$$

$$H = \underset{h}{\operatorname{argmax}} T(h) \quad (5)$$

3. Experiments

In order to investigate the effectiveness of the proposed MIR system, several experiments were carried out.

HTK[9] was used as the recognizer, and a monophone HMM was used as an acoustic model. Note that a "monophone HMM" is a simple acoustic model, and it means that a phoneme is modeled by an HMM. The HMM was trained using a large amount of normally read speech, and the singing voice adaptation method[5] was carried out using 127 singing voice data sung by 6 male university students.

One hundred and ten singing voice data sung by another 6 male students were used as test data, and all of the test data were automatically segmented into words using the Viterbi algorithm[10]. It is assumed that a user sings a few words when the MIR system is used. Therefore, the number of words in a test sample can be controlled. There were 156 Japanese children's songs in the database.

The average word accuracy of the test data was 81.0%, and the lyrics recognizer output 1,000 hypotheses per test sample. In this hypotheses list, some similar hypotheses were output as another hypothesis. For example, both hypotheses h and $\bar{h}$ are in the hypotheses list as another hypothesis because $W(h)$ is a slightly different from $W(\bar{h})$, even though $S(h)$ is exactly the same as $S(\bar{h})$. The correct hypothesis was not included in the hypotheses list for 2.6% of test samples. This means that the maximum retrieval accuracy was limited to 97.4%.

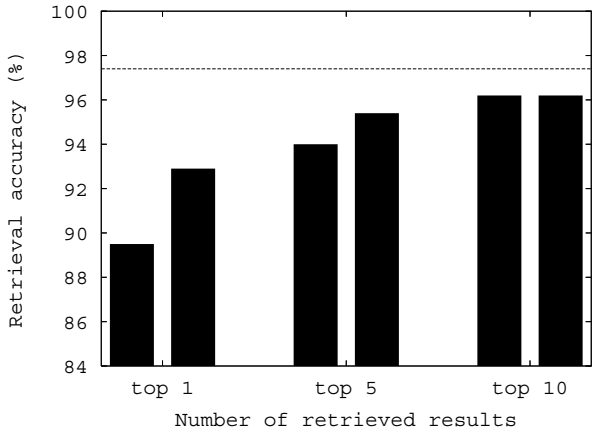

Figure 3. Retrieval accuracy using five words

Table 1. Relationship between the rank of the correct hypothesis and verification method

		After verification	
		Top 1	Others
Before verification	Top 1	753	8
	Others	37	52

3.1. Retrieval Accuracy for Fixed-length Input

In this section, the number of words in a test sample was fixed to five, and weighting factors w_1 and w_2 were set to optimum values *a posteriori*.

Figure 3 shows retrieval accuracy given by the before and after verification. In this figure, the left side of each number of retrieved results denotes the retrieval accuracy given by before verification, which is the same as the system proposed in [5], and the right side denotes that given by the proposed MIR system. The horizontal line denotes the upper limit of the retrieval accuracy.

This figure shows that the verification method was very effective in increasing retrieval accuracy. Especially, the retrieval accuracy of top 1 increased by 3.4 points, from 89.5% to 92.9%. However, the retrieval accuracy of top 10 was slightly improved. This result means that the hypotheses with higher (but not first-ranked) recognition score can be corrected.

Table 1 shows the relationship between the rank of the correct hypothesis and verification method. The numbers in this table indicate a number of samples, and the total number of test samples was 850.

In 753 test samples, which is 88.6% of the test samples, the correct hypothesis was ranked first before and after verification. The correct hypothesis became the first-rank by the verification in 37 samples. On the other hand, only 8 samples were corrupted by the verification method. This result showed that the verification method does not decrease the performance of lyrics recognition results for any samples, and several samples can be improved by the method.

3.2. Retrieval Accuracy for Variable Length Input

In this section, we investigate the relationship between the number of words in a singing voice and retrieval accuracy. The number of words was increased from 3 to 10. In this experiment, 152 singing voice data sung by 6 new males were added to the test samples in order to increase the statistical reliability of the experimental result. Other experimental conditions were the same as in the previous experiments.

Figure 4 shows the relationship between the number of words and retrieval accuracy. In this figure, the left side of each number of words denotes the retrieval accuracy given by before verification, and the right side denotes that given by the proposed MIR system.

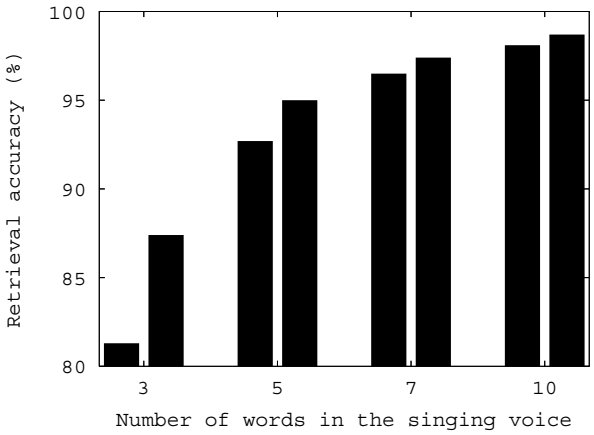

Figure 4. Retrieval accuracy using various number of words

This figure shows that the proposed MIR system gave higher accuracy for all conditions. Especially, the verification method was very effective when the number of words was small. There are many songs which have partially the same lyrics. If the number of words in the retrieval key is small, a lot of hypotheses are ranked at the same rank, and cannot be distinguished only using lyrics information. Melody information is very powerful in these situations. The χ^2-test showed that the difference between before and after verification is statistically significant when the number of words was set to 3 and 5.

3.3. Discussion: system performance when the lyrics are only partially known by a user

The proposed system assumes that the input singing voice consists of a part of the correct lyrics. If it includes a wrong word, the retrieval may fail.

This issue need to be addressed in future work, however, it is not fatal for the system. If a user knows several correct words in the lyrics, retrieval can still succeed because the proposed system gave about 87% retrieval accuracy with the query consisting of only three words. Moreover, the lyrics recognizer can correctly recognize a long query even if it includes several wrong words because of the grammatical restriction of FSA.

4. Conclusion

In this paper, we propose an MIR system that uses both melody and lyrics information in the singing voice.

The MIR system verifies hypotheses output by a lyrics recognizer from a melodic point of view. Each hypothesis has time alignment information between the singing voice and recognized text, and the boundaries of each note can be estimated using the information. As a result, melody information is extracted from the singing voice. On the other hand, the melody information can be calculated from the musical score of the song because the recognized text must be a part of the lyrics of the song. The hypothesis is verified by calculating the similarity between the two types of melody information.

From the experimental results, the verification method increased the retrieval accuracy. Especially, it was very effective when the number of words in the user's singing voice was small. The proposed method increased the retrieval accuracy from 81.3% to 87.4% when the number of words was only three.

References

[1] J. S. R. Jang, J. Chun, and M.-Y. Kao, "MIRACLE: A Music Information Retrieval System with Clustered Computing Engines," in *International Symposium on Music Information Retrieval*, 2001.

[2] N. Kosugi, Y. Nishihara, T. Sakata, M. Yamamoto, and K. Kushima, "A Practical Query-By-Humming System for a Large Music Database," in *ACM Multimedia 2000*, 2000, pp. 333–342.

[3] S.-P. Heo, M. Suzuki, A. Ito, and S. Makino, "Three dimensional continuous DP algorithm for multiple pitch candidates in music information retrieval system," in *Proc. ISMIR*, 2003, pp. 235–236.

[4] S.-P. Heo, M. Suzuki, A. Ito, S. Makino, and H.-Y. Chung, "Multiple pitch candidates based music information retrieval method for query-by-humming," in *Proc. AMR*, 2003, pp. 189–200.

[5] T. Hosoya, M. Suzuki, A. Ito, and S. Makino, "Lyrics recognition from a singing voice based on finite state automaton for music information retrieval," in *Proc. ISMIR*, 2005, pp. 532–535.

[6] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, "Query by humming: Musical information retrieval in an audio database," in *Proc. ACM Multimedia*, 1995, pp. 231–236.

[7] B. Liu, Y. Wu, and Y. Li, "A Linear Hidden Markov Model for Music Information Retrieval Based on Humming," in *Proc. ICASSP 2003*, 2003, vol. Vol. V, pp. 533–536.

[8] P. Boersma and D. Weenink, "praat," University of Amsterdam, http://www.fon.hum.uva.nl/praat/.

[9] Cambridge University Engineering Department, "Hidden Markov Model Toolkit," http://htk.eng.cam.ac.uk/.

[10] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm," *IEEE Trans. Information Theory*, vol. 13, no. 2, pp. 260–269, 4 1967.