Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 38727, 8 pages
doi:10.1155/2007/38727

Research Article

Music Information Retrieval from a Singing Voice Using Lyrics and Melody Information

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

Received 1 December 2005; Revised 28 July 2006; Accepted 10 September 2006

Recommended by Masataka Goto

Recently, several music information retrieval (MIR) systems which retrieve musical pieces by the user's singing voice have been developed. All of these systems use only melody information for retrieval, although lyrics information is also useful for retrieval. In this paper, we propose a new MIR system that uses both lyrics and melody information. First, we propose a new lyrics recognition method. A finite state automaton (FSA) is used as recognition grammar, and about 86% retrieval accuracy was obtained. We also develop an algorithm for verifying a hypothesis output by a lyrics recognizer. Melody information is extracted from an input song using several pieces of information of the hypothesis, and a total score is calculated from the recognition score and the verification score. From the experimental results, 95.0% retrieval accuracy was obtained with a query consisting of five words.

1. INTRODUCTION

Recently, several music information retrieval (MIR) systems that use a user's singing voice as a retrieval key have been researched (e.g., MELDEX [1], SuperMBox [2], MIRACLE [3], SoundCompass [4], and our proposed method [5]). 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, several lyrics recognition systems have been developed. The lyrics recognition technique used in these systems is simply a large vocabulary continuous speech recognition (LVCSR) technique, based on an HMM (hidden Markov model), acoustic model [6], and a trigram language model. Ozeki et al. [7] performed lyrics recognition from the singing voice divided into phrases, and the word correct rate was about 59%. Sasou et al. [8] performed lyrics recognition using ARHMM-based speech analysis, and the word correct rate was about 70%. Moreover, we [9] performed lyrics recognition using an LVCSR system, and the word correct rate was about 61%. These results are considerably worse than the recognition performance for read speech.

Another problem is that it is difficult for conventional MIR systems to use a singing voice as a retrieval key. One of the biggest problems is how to split the input singing voice into musical notes [10]. Traditional MIR systems [4, 11, 12] 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. However, recent MIR systems do not need such an assumption. These systems split the input singing voice into musical notes using a continuity of pitch contour [13], neural networks [14], graphical model [15], and so on. Unfortunately, they often give inaccurate information. It is hard to split the singing voice into musical notes without linguistic information.

On the other hand, there are several works [10, 16] based on "frame-based" strategy. In this strategy, it is not needed to split the input singing voice into musical notes because an input query is matched with the database frame-by-frame. However, this algorithm needs much computation time [10].

The lyrics recognizer outputs several hypotheses as a recognition result. Each hypothesis has time alignment information between the singing voice and the recognized text.

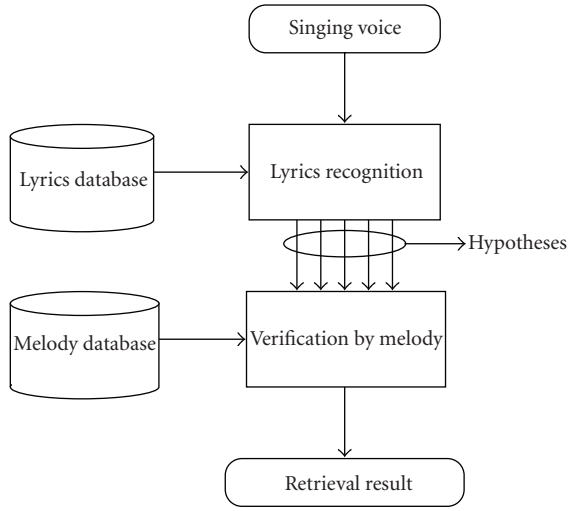

Figure 1: Outline of the MIR system using lyrics and melody.

It is easy to split the input singing voice into musical notes using time alignment information, and a hypothesis can be verified from a melodic point of view. In this paper, we propose a new MIR system using lyrics and melody information.

2. 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 the output.

Each hypothesis h has the following information.

(i) Song name $S(h)$.
(ii) Recognized text $W(h)$. It must be a part of the lyrics of the song $S(h)$ (the details are described in Section 3).
(iii) Recognition score $R(h)$.
(iv) Time alignment information $F(h)$. For all phonemes in the recognized text, frame numbers of the start frame, and the end frame are the 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 $S(h)$. 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, both types of information should be similar. The verification score is defined as the similarity of both 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.

In the system, the lyrics recognition step and the verification step are carried out separately. In general, a system consisting of one step gives higher performance than a system consisting of two or more steps because the system consisting of one step can search the optimum hypothesis for all models. If the system only has one step which uses both lyrics and melody information, the retrieval performance may increase. However, it is difficult to use melody information in the lyrics recognition step.

The recognition score is calculated by the lyrics recognizer frame-by-frame. If pitch information is included in the lyrics recognition, the pitch contour should be modeled by HMM. However, there are two major problems. The first problem is that pitch cannot be calculated for several frames corresponding to unvoiced consonants, short pause, and so on. However, pitch information is needed for *all* frames in order to calculate pitch score frame-by-frame. Therefore, pitch information of such a "pitchless" frame should be given appropriately.

The second problem is that a huge amount of singing voice is needed for modeling of pitch contour. A pitch contour of singing voice cannot be represented by a step function, even though the pitch contour of a tune can be represented. This means that the HMM corresponding to pitch contour should be trained using a huge amount of singing voice. Moreover, singer adaptation may be needed because a pitch contour may be different depending on a singer. Therefore, it is very difficult to make the pitch HMM.

3. LYRICS RECOGNITION BASED ON A FINITE STATE AUTOMATON

3.1. Introduction

An LVCSR system performs speech recognition using two kinds of models—an acoustic model and a language model. An HMM [6] is the most popular acoustic model, and it models the acoustic feature of phonemes. On the other hand, bigram or trigram models are often used as language models. A trigram model describes the probabilities of three contiguous words. In other words, it only considers a part of the input word sequence. One reason why an LVCSR system uses a trigram model is that a trigram model has high coverage over an unknown set of speech inputs.

Thinking of a song input for music information retrieval, it seems reasonable to assume that the input song is a part of a song in the song database. This is a very strong constraint compared with ordinary speech recognition. To introduce this constraint into our lyrics recognition system, we used a finite state automaton (FSA) that accepts only a part of the lyrics in the database. By using this FSA as a language model for the speech recognizer, the recognition results are assured to be a part of the lyrics in the database. This strategy is not only useful for improving the accuracy of lyrics recognition, but also very helpful for retrieving a musical piece, because the musical piece is naturally determined by simply finding the part among the database that strictly matches the recognizer outputs.

3.2. An FSA for recognition

Figure 2 shows an example of a finite state automaton used for lyrics recognition. In Figure 2, "<s>" denotes the start

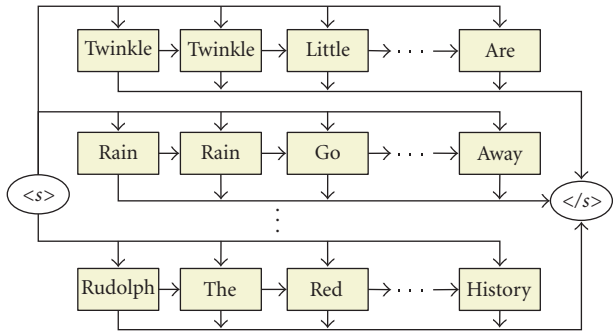

Figure 2: Automaton expression of the grammar.

Table 1: Experimental conditions.

Test query	850 singing voices sung by 6 males consisting of five words
Acoustic model	Monophone HMM trained from read speech
Database	Japanese children's songs 238 songs

symbol, and "</s>" denotes the end symbol. The rectangles in the figure stand for words and the arrows are possible transitions. One row in Figure 2 stands for the lyrics corresponding to one song.

In this FSA, transition from the start symbol to any word is allowed, but only two transitions from the word are allowed: the transition to the next word and the transition to the end symbol. As a result, this FSA only accepts a part of the lyrics that starts from any word and ends at any word in the lyrics.

When lyrics are recognized using this FSA, the song name can be determined as well as the lyrics by searching the transition path of the automaton.

3.3. Lyrics recognition experiment

A lyrics recognition experiment was carried out using the FSA as a language model. Table 1 shows the experimental conditions. The test queries were singing voice samples, each of which consisted of five words. The singers were six male university students, and 110 choruses were collected as song data. The test queries were generated from the whole song data by automatically segmenting the song into words. It is thought that typically people sing a few words when using MIR systems. Therefore, we decided on a test query length of five words. Segmentation and the recognition were performed by HTK [17]. The total number of test queries was 850. The acoustic model was a monophone HMM trained from normal read speech.

Table 2 shows the result of word recognition rates (word correct rate and word accuracy) and error rates. In the table, "trigram" denotes the results using a trigram language model trained from lyrics in the database. The word correct rate (Corr) and word accuracy (Acc) in Table 2 are calculated as follows:

Table 2: Word recognition/error rate [%].

Grammar	Corr	Acc	Sub	Ins	Del
FSA	75.9	64.5	19.9	4.2	11.4
Trigram	58.3	48.2	31.7	10.0	10.1

Table 3: Retrieval accuracy [%].

Retrieval key	Top 1	Top 5	Top 10
Recognition results	76.0	83.9	83.9
Correct lyrics	99.7	100.0	100.0

$$\begin{aligned} \text{Corr} &= \frac{N - D - S}{N}, \\ \text{Acc} &= \frac{N - D - S - I}{N}, \end{aligned} \tag{1}$$

where N denotes the number of words in the correct lyrics, D denotes the number of deletion error (Del) words, S denotes the number of substitution error (Sub) words, and I denotes the number of insertion error (Ins) words. The recognition results of the proposed method outperformed the conventional trigram language model. Especially, the substitution and insertion error rate was decreased by FSA because the recognized word sequence is restricted within a part of the lyrics in the database.

Table 3 shows the results of retrieval accuracy up to the top 10 candidates. Basically, the retrieval accuracy of the top R candidate is the probability of the correct result being listed within the top R list generated by the system. The retrieval accuracy of the top R candidate $A(R)$ was calculated as follows:

$$\begin{aligned} A(R) &= \frac{1}{Q} \sum_{i=1}^{Q} T_i(R), \\ T_i(R) &= \begin{cases} 1, & r(i) + n_i(r(i)) - 1 \le R, \\ 0, & r(i) > R, \\ \dfrac{R - r(i) + 1}{n_i(r(i))}, & \text{otherwise}, \end{cases} \end{aligned} \tag{2}$$

where Q denotes the number of queries, $r(i)$ denotes the rank of the correct song in the ith query, $n_i(x)$ denotes the number of songs in the xth place in the ith query, and $T_i(R)$ means the probability that the correct song appears in the top Rth candidates of the ith query.

In Table 3, "recognition results" is the retrieval accuracy using recognized lyrics and "correct lyrics" is the retrieval accuracy using the correct lyrics. Note that the retrieval accuracy of the top result from the "correct lyrics" was not 100% because several songs contained the same five words of lyrics.

In our results, about 84% retrieval accuracy was obtained by the proposed method. As the retrieval accuracy itself is better than that of the query-by-humming-based system [18], this is a promising result.

Table 4: Word recognition/error rate [%].

Adaptation	Corr	Acc	Sub	Ins	Del
Before	75.9	64.5	19.9	4.2	11.4
After	83.2	72.7	13.8	3.1	10.5

Table 5: Retrieval accuracy [%].

Adaptation	Top 1	Top 5	Top 10
Before	76.0	83.9	83.9
After	82.7	88.5	88.5

3.4. Singing voice adaptation

As the acoustic model used in the last experiment was trained from read speech, it may not properly model the singing voice. The acoustical characteristics of singing voice are different from those of read speech [19]. Especially, a high-pitched voice and prolonged notes degrade the accuracy of speech recognition [7]. In order to improve the acoustic model for modeling the singing voice, we tried to adapt the HMM to the singing voice using a speaker adaptation technology.

Speaker adaptation is a method to customize an acoustic model for a specific user. The recognizer uses a small amount of the speech of the user, and the acoustic model is modified so that the probability of generating the user's speech becomes higher. In this paper, we exploited the speaker adaptation method to modify the acoustic model for the singing voice. As we do not want to adapt the acoustic model to a specific user, we used several users' voice data for the adaptation.

In the following experiment, the MLLR (maximum likelihood linear regression) method [20] was used as an adaptation algorithm. One hundred twenty-seven choruses sung by 6 males were used as the adaptation data. These 6 singers were different from those who sang the test queries. Other experimental conditions were the same as those shown in Table 1.

Table 4 shows the word recognition rates before and after adaptation. These results show that the adaptation improved the word correct rate by more than 7 points. Table 5 shows the retrieval accuracy results. These results prove the effectiveness of the adaptation.

As a result, the singing voice adaptation method is very effective. In other words, the acoustical characteristics of singing voice are very different from those of read speech. We point out that the adapted HMM can be used for any singer because the proposed adaptation method did not adapt the HMM to a specific singer.

3.5. Improvement of the FSA: consideration of Japanese phrase structure

The FSA used in the above experiments accepts any word sequences which are a subsequence of the lyrics in the database. However, no user begins to sing from any word in the lyrics and finishes singing at any word. As the language of the texts in these experiments is Japanese, the constraints of Japanese phrase structure can be exploited.

A Japanese sentence can be regarded as a sequence of "*bunsetsu*." A "*bunsetsu*" is a linguistic structure similar to a phrase in English. One "*bunsetsu*" is composed of one content word followed by zero or more particles or suffixes.

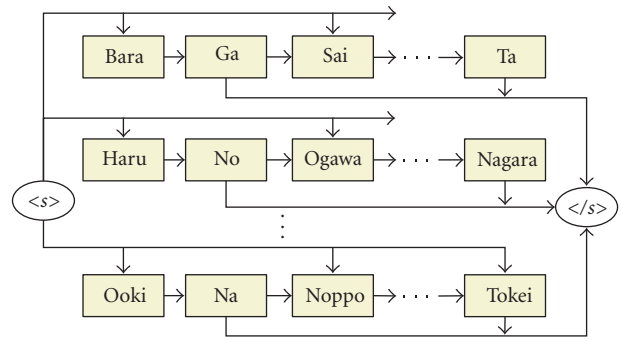

Figure 3: Example of improved grammar.

In Japanese, singing from a particle or a suffix hardly ever occurs. For example, in the following sentence:

Bara	*Ga*	*Sai*	*Ta*	···
Rose	(subject)	Bloom	(past)	

"*bara ga*" and "*sai ta*" are "*bunsetsu*", and a user hardly ever begins to sing from "*ga*" or "*ta*." Therefore, we changed the FSA described in Section 3.2 as follows.

(1) Omit all transitions from the start symbol "<s>" to any particles or suffixes.
(2) Omit all transitions from the start or middle word of a "*bunsetsu*" to the end symbol "</s>."

An example of the improved FSA is shown in Figure 3.

The lyrics recognition experiment was carried out using the improved FSA. The adapted HMM described in Section 3.4 was used for the acoustic model, and the other experimental conditions were the same as those shown in Table 1.

The results are shown in Tables 6 and 7. Both word recognition rates and retrieval accuracy improved compared with that of the original FSA. The word correct rate and the retrieval accuracy of the first rank were about 86%. These results showed the effectiveness of the proposed constraints.

In this section, the Japanese phrase structure is used for effective constraints. However, this does not mean that the proposed FSA cannot apply to other languages. If a target language has phrase-like structure, the FSA can represent the structure of the target language.

4. VERIFICATION OF HYPOTHESIS USING MELODY INFORMATION

The lyrics recognizer outputs many hypotheses, and the tune corresponding to the recognized text can be obtained from the database. The melody information, which is defined as a

Table 6: Word recognition/error rate [%].

FSA	Corr	Acc	Sub	Ins	Del
Original	83.2	72.7	13.8	3.1	10.5
Improved	86.0	77.4	10.6	3.4	8.6

Table 7: Retrieval accuracy [%].

FSA	Top 1	Top 5	Top 10
Original	82.7	88.5	88.5
Improved	85.9	91.3	91.3

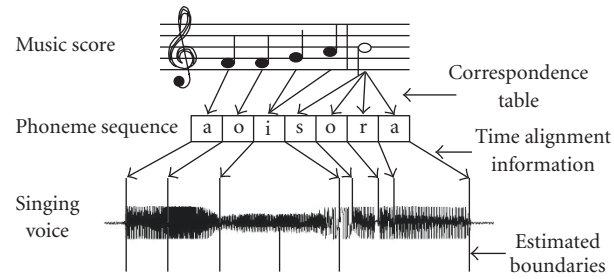

Figure 4: Example of estimation of boundaries between notes.

relative pitch and relative IOI 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 time alignment information. The verification score is defined as the similarity of both types of information.

Note that the lyrics recognizer with FSA is needed to verify hypotheses. If a general LVCSR system with trigram language model is used as a lyrics recognizer, the tune corresponding to the recognized text cannot be obtained because the recognized text may not correspond to the part of the correct lyrics.

4.1. Extraction of melody information

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 4 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 (e.g., the ith 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 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 [21] 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 following two equations:

$$\Delta f_n = \log_2 \frac{f_{n+1}}{f_n}, \qquad \Delta t_n = \log_2 \frac{t_{n+1}}{t_n}, \tag{3}$$

where, f_n and t_n are pitch and IOI of the nth note, respectively.

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

4.2. Calculation of verification score

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.

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

$$V(h) = \frac{1}{N-1} \sum_{n=1}^{N-1} \{ w_1 (| \Delta \hat{t}_n - \Delta t_n |) + (1 - w_1) (| \Delta \hat{f}_n - \Delta f_n |) \}, \tag{4}$$

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

Total score $T(h)$ is calculated by (5) for each hypothesis h, and the final result H is selected by (6):

$$T(h) = w_2 R(h) - (1 - w_2) V(h), \tag{5}$$

$$H = \operatorname*{argmax}_h T(h). \tag{6}$$

4.3. Experiments

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

The number of songs in the database was 156, and other experimental conditions were the same as in previous experiments described in Section 3. The average word accuracy of the test queries was 81.0%, and 1 000 hypotheses were output from the lyrics recognizer for a test query. In these hypotheses' list, some similar hypotheses were output as another hypotheses. For example, both hypotheses h and $\overline{h}$ are in the hypotheses' list as another hypotheses because $W(h)$ is slightly different from $W(\overline{h})$, even though $S(h)$ is exactly the same as $S(\overline{h})$. The correct hypothesis was not included in the hypotheses' list for 2.6% of test queries. This means that the maximum retrieval accuracy was limited to 97.4%.

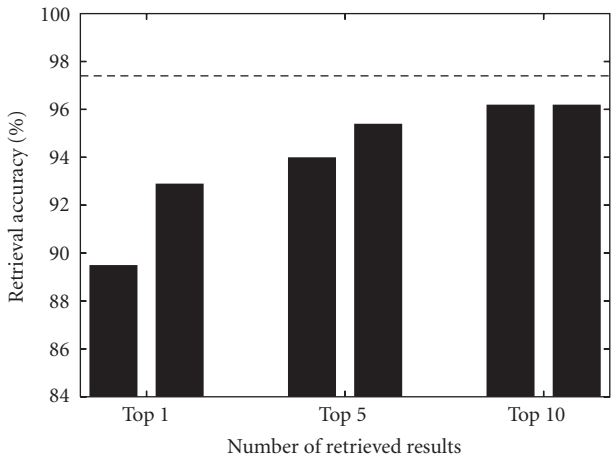

Figure 5: Retrieval accuracy using five words.

4.3.1. Retrieval accuracy for fixed-length query

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

Figure 5 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 before verification, which is the same as the system proposed in Section 3, 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 the top 10 was slightly improved. This result means that the hypothesis with higher (but not first-ranked) recognition score can be corrected by the verification method.

Table 8 shows the relationship between the rank of the correct hypothesis and verification method. The numbers in this table indicate the number of queries, and the total number of test queries was 850.

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

4.3.2. Retrieval accuracy for variable length query

In this section, we investigate the relationship between the number of words in a test query and retrieval accuracy. The number of words in a query was increased from 3 to 10. In this experiment, 152 song data sung by 6 new males were added to the test queries in order to increase the statistical reliability of the experimental result. The total number of test queries is shown in Table 9. Other experimental conditions were the same as in the previous experiments.

Table 8: 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

Table 9: Number of test queries.

Number of words	3	5	7	10
Number of queries	2240	1959	1929	1791

Figure 6 shows the relationship between the number of words in a query and retrieval accuracy. In this figure, the left side of each number of words denotes the retrieval accuracy given before verification, and the right side denotes that given by the proposed MIR system.

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 using only 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.

5. DISCUSSION

5.1. 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 needs 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.

5.2. Scalability of the system

In this paper, the proposed system was examined using a very small database. When the system is applied to practical use, a large database is used in the system. In this situation, following two problems will be occurred.

The first problem is computation time in the lyrics recognition step. When the number of songs in the database

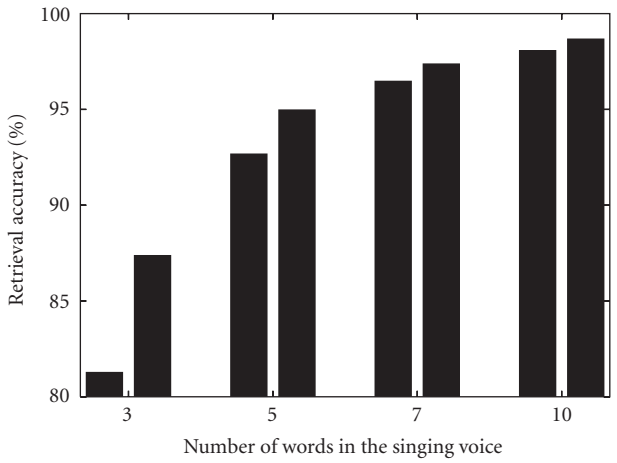

Figure 6: Retrieval accuracy using various number of words.

increases, the FSA becomes lager. Therefore, the lyrics recognition needs much calculation time and memory. In order to solve this problem, a preselection algorithm would be needed before lyrics recognition. This issue needs to be addressed in future work.

The second problem is deterioration of the recognition performance. There are many songs which have similar lyrics in the large database. It causes deterioration of the recognition performance. However, these misrecognition can be corrected by using melody information. As a result, the retrieval accuracy is slightly decreased.

6. CONCLUSION

We proposed an MIR system that uses both melody and lyrics information in the singing voice.

First, we tried to recognize lyrics in users' singing voice. To exploit the constraints of the input song, we used an FSA that accepts only a part of word sequences in the song database. From the experimental results, the proposed method proved to be effective, and a retrieval accuracy of about 86% was obtained.

We also proposed an algorithm for verifying a hypothesis output by the lyrics recognizer. Melody information is extracted from an input song using several pieces of information of the hypothesis, and a total score is calculated from the recognition score and the verification score. From the experimental results, the proposed method showed high performance, and 95.0% retrieval accuracy was obtained with a query consisting of five words.

The proposed system would be applied to a practical situation in our future work.

REFERENCES

[1] R. J. McNab, L. A. Smith, D. Bainbridge, and I. H. Witten, "The New Zealand digital library MELody inDEX," *D-Lib Magazine*, vol. 3, no. 5, pp. 4–15, 1997.

[2] J.-S. R. Jang, H.-R. Lee, and J.-C. Chen, "Super MBox: an efficient/effective content-based music retrieval system," in *Proceedings of the 9th ACM International Conference on Multimedia (ACM Multimedia '01)*, pp. 636–637, Ottawa, Ontario, Canada, September-October 2001.

[3] J.-S. R. Jang, J.-C. Chen, and M.-Y. Kao, "MIRACLE: a music information retrieval system with clustered computing engines," in *Proceedings of the 2nd Annual International Symposium on Music Information Retrieval (ISMIR '2002)*, Bloomington, Ind, USA, October 2001.

[4] N. Kosugi, Y. Nishihara, T. Sakata, M. Yamamuro, and K. Kushima, "A practical query-by-humming system for a large music database," in *Proceedings of the 8th ACM International Conference on Multimedia (ACM Multimedia '00)*, pp. 333–342, Los Angeles, Calif, USA, October-November 2000.

[5] S.-P. Heo, M. Suzuki, A. Ito, and S. Makino, "An effective music information retrieval method using three-dimensional continuous DP," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 633–639, 2006.

[6] L. R. Rabiner and B.-H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[7] H. Ozeki, T. Kamata, M. Goto, and S. Hayamizu, "The influence of vocal pitch on lyrics recognition of sung melodies," in *Proceedings of Autumn Meeting of the Acoustical Society of Japan*, pp. 637–638, September 2003.

[8] A. Sasou, M. Goto, S. Hayamizu, and K. Tanaka, "An autoregressive, non-stationary excited signal parameter estimation method and an evaluation of a singing-voice recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '05)*, vol. 1, pp. 237–240, Philadelphia, Pa, USA, March 2005.

[9] T. Hosoya, M. Suzuki, A. Ito, and S. Makino, "Song retrieval system using the lyrics recognized vocal," in *Proceedings of Autumn Meeting of the Acoustical Society of Japan*, pp. 811–812, September 2004.

[10] N. Hu and R. B. Dannenberg, "A comparison of melodic database retrieval techniques using sung queries," in *Proceedings of the ACM International Conference on Digital Libraries*, pp. 301–307, Association for Computing Machinery, Portland, Ore, USA, July 2002.

[11] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, "Query by humming: musical information retrieval in an audio database," in *Proceedings of the 3rd ACM International Conference on Multimedia (ACM Multimedia '95)*, pp. 231–236, San Francisco, Calif, USA, November 1995.

[12] B. Liu, Y. Wu, and Y. Li, "A linear hidden Markov model for music information retrieval based on humming," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, vol. 5, pp. 533–536, Hong Kong, April 2003.

[13] W. Birmingham, B. Pardo, C. Meek, and J. Shifrin, "The MusArt music-retrieval system: an overview," *D-Lib Magazine*, vol. 8, no. 2, 2002.

[14] C. J. Meek and W. P. Birmingham, "A comprehensive trainable error model for sung music queries," *Journal of Artificial Intelligence Research*, vol. 22, pp. 57–91, 2004.

[15] C. Raphael, "A graphical model for recognizing sung melodies," in *Proceedings of 6th International Conference on Music Information Retrieval (ISMIR '05)*, pp. 658–663, London, UK, September 2005.

[16] M. Mellody, M. A. Bartsch, and G. H. Wakefield, "Analysis of vowels in sung queries for a music information retrieval system," *Journal of Intelligent Information Systems*, vol. 21, no. 1, pp. 35–52, 2003.

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

[18] A. Ito, S.-P. Heo, M. Suzuki, and S. Makino, "Comparison of features for DP-matching based query-by-humming system," in *Proceedings of 5th International Conference on Music Information Retrieval (ISMIR '04)*, pp. 297–302, Barcelona, Spain, October 2004.

[19] A. Loscos, P. Cano, and J. Bonada, "Low-delay singing voice alignment to text," in *Proceedings of International Computer Music Conference (ICMC '99)*, Beijing, China, October 1999.

[20] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

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

Motoyuki Suzuki was born in Chiba, Japan, in 1970. He received the B.E., M.E., and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1993, 1995, and 2004, respectively. Since 1996, he has worked with the Computer Center and the Information Synergy Center, Tohoku University, as a Research Associate. From 2006 to 2007, he worked with the Centre for Speech Technology Research, University of Edinburgh, UK, as a Visiting Researcher. He is now a Research Associate of Graduate School of Engineering, Tohoku University. His interests include spoken language processing, music information retrieval, and pattern recognition using statistical modeling. He is a Member of the Institute of Electronic, Information, and Communication Engineering, the Acoustical Society of Japan, and the Information Processing Society of Japan.

Toru Hosoya was born in Gunma, Japan, in 1981. He received the B.E. and M.E. degrees from Tohoku University, Sendai, Japan, in 2004 and 2006, respectively. From 2003 to 2006, he had researched about music information retrieval from singing voice, in Tohoku University. He is now a System Engineer in NEC Corporation, Japan.

Akinori Ito was born in Yamagata, Japan, in 1963. He received the B.E., M.E., and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1984, 1986, and 1992, respectively. Since 1992, he has worked with Research Center for Information Sciences and Education Center for Information Processing, Tohoku University. He joined the Faculty of Engineering, Yamagata University, from 1995 to 2002. From 1998 to 1999, he worked with College of Engineering, Boston University, MA, USA, as a Visiting Scholar. He is now an Associate Professor of Graduate School of Engineering, Tohoku University. He has engaged in spoken language processing, statistical text processing, and audio signal processing. He is a Member of the Institute of Electronic, Information, and Communication Engineering, the Acoustical Society of Japan, the Information Processing Society of Japan, and the IEEE.

Shozo Makino was born in Osaka, Japan, on January 3, 1947. He received the B.E., M.E., and Dr. Eng. degrees from Tohoku University, Sendai, Japan, in 1969, 1971, and 1974, respectively. Since 1974, he has been working with the Research Institute of Electrical Communication, Research Center for Applied Information Sciences, Graduate School of Information Science, Computer Center, and Information Synergy Center, as a Research Associate, an Associate Professor, and a Professor. He is now a Professor of Graduate School of Engineering, Tohoku University. He has been engaged in spoken language processing, CALL system, autonomous robot system, speech corpus, music information processing, image recognition and understanding, natural language processing, semantic web search, and digital signal processing.