LYRICS RECOGNITION FROM A SINGING VOICE BASED ON FINITE STATE AUTOMATON FOR MUSIC INFORMATION RETRIEVAL

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

ABSTRACT

Recently, 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 the melody information for retrieval. Although the lyrics information is useful for retrieval, there have been few attempts to exploit lyrics in the user's input. In order to develop a MIR system that uses lyrics and melody information, lyrics recognition is needed. Lyrics recognition from a singing voice is achieved by similar technology to that of speech recognition. The difference between lyrics recognition and general speech recognition is that the input lyrics are a part of the lyrics of songs in a database. To exploit linguistic constraints maximally, we described the recognition grammar using a finite state automaton (FSA) that accepts only lyrics in the database. In addition, we carried out a "singing voice adaptation" using a speaker adaptation technique. In our experimental results, about 86% retrieval accuracy was obtained.

Keywords: MIR, lyrics recognition, FSA

1 INTRODUCTION

Recently, several music information retrieval (MIR) systems that use the user's singing voice as a retrieval key have been researched, some examples are MiDiLiB (University of Bonn), MELDEX (Rodger J. McNab, Lloyd A. Smith, David Bainbridge and Ian H. Witten, 1997), Themefinder (Stanford University), TuneServer (University of Karlsruhe), SuperMBox (J.S.Roger Jang, H.Lee and J.Chen, 2001), SoundCompass (Naoko Kosugi, Yuichi Nishihara, Tetsuo Sakata, Masashi Yamamoto and Kazuhiko Kushima, 2000), MIRACLE (J.S.Roger Jang, Jiang-Chun, Ming-Yang Kao, 2001), etc.

These systems use melody information in the user's singing voice. In these systems, the lyrics sung by the voice are not taken into consideration. We are attempting to develop a MIR system that uses melody and lyrics information in the user's singing voice. Figure 1 shows an outline of the MIR system using lyrics and melody information. First of all, lyrics and melody are acquired from the user's song input. Then the system retrieves the musical piece by using one of the recognized lyrics and melody. Finally, the two results are integrated into the retrieval result. Because it performs the retrieval by using both the lyrics and melody, the accuracy of retrieval is expected to be better than using only the melody.

We have been trying to research lyrics recognition from the user's singing voice as a first step towards the realization of this system. The lyrics recognition technique used in several conventional works is simply a large vocabulary continuous speech recognition (LVCSR) technique, based on an HMM acoustic model and a trigram language model. Ozeki et al. performed lyrics recognition from the singing voice divided into phrases, and the word correct rate was about 59% (Hironao Ozeki, Takayuki Kamata, Masataka Goto and Satoru Hayamizu, 2003). Moreover, we performed lyrics recognition us-

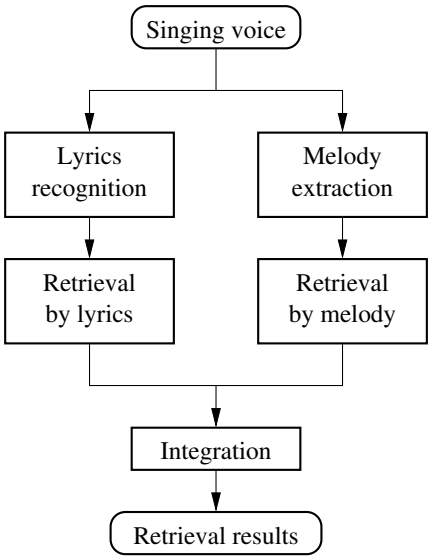

Figure 1: Outline of the MIR system using lyrics and melody

ing an LVCSR system, and the word correct rate was about 61% (Toru Hosoya, Motoyuki Suzuki, Akinori Ito and Shozo Makino, 2004). These results are considerably worse compared with the recognition performance for read speech.

When a user sings a song as the means of retrieval, it is natural to assume that the sung lyrics are a part of the database. This assumption means that the lyrics recognition for MIR can exploit stronger linguistic constraints than general speech recognition. To achieve this, we used a finite state automaton (FSA) that accepts any subsequences of the lyrics in the database as a language model for lyrics recognition.

2 LYRICS RECOGNITION BASED ON A FINITE STATE AUTOMATON

2.1 Introduction

A large vocabulary continuous speech recognition (LVCSR) system performs speech recognition using two kinds of models — an acoustic model, and a language model. An HMM (Hidden Markov Model) is the most popular acoustic model. An HMM models the acoustic feature of phonemes. On the other hand, bigram or trigram models are often used as language models. A trigram model describes 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 make the assumption that the input song is a part of one of the songs 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 in improving the accuracy of lyrics recognition, but also very helpful to the retrieval of a musical piece, because the musical piece is naturally determined by simply finding the part among the database that strictly matches the recognizer outputs.

2.2 An FSA for recognition

Figure 2 shows an example of the finite state automaton used for lyrics recognition. In Fig. 2, "<s>" is the start symbol, and "</s>" is the end symbol. The rectangles in the figure stand for words and the arrows are possible transitions. One row in Fig. 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.

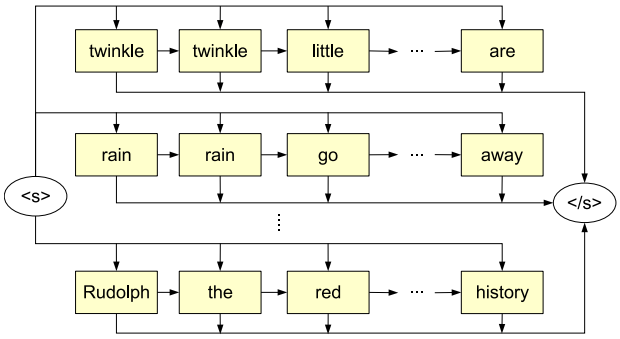

Figure 2: Automaton expression of the grammar

Table 1: Experimental conditions

Recognition engine	HTK
Test data	Singing voice of five words
Acoustic model	monophone HMM trained from read speech
Database	Japanese children's songs 238 songs

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.

2.3 Experiment

The lyrics recognition experiment was carried out using the FSA as a recognition grammar. Table 1 shows the experimental conditions. The test data were singing voice samples, each of which consisted of five words. The singers were five male university students. These song data were generated from the whole song data by automatically segmenting the song into words. It is thought that people sing a few words when people use MIR systems. Therefore, we decided on a test data length of five words. Segmentation and the recognition were performed by HTK (Cambridge University Engineering Department). The acoustic model was a monophone HMM trained from normal read speech. The recognition result is shown in Table 2 and 3.

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:

$$\text{Corr} = \frac{N - D - S}{N}$$

$$\text{Acc} = \frac{N - D - S - I}{N}$$

where N is the number of words in the correct lyrics, D is the number of deletion error words, S is the number of substitution error words, and I is the number of insertion error words. The recognition results of the proposed method outperformed the conventional trigram language model.

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

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 to be listed within the top-R list generated by the system. The retrieval accuracy of the top-R candidate $A(R)$ was calculated as follows:

$$A(R) = \frac{1}{Q}\sum_{i=1}^{Q} T_i(R)$$

$$T_i(R) = \begin{cases} 0 & r(i) > R \\ 1 & r(i) + n_i(r(i)) - 1 \leq R \\ \dfrac{R - r(i) + 1}{n_i(r(i))} & \text{otherwise} \end{cases}$$

where Q is the number of queries, $r(i)$ is the rank of the correct song in i-th query, $n_i(x)$ is the number of songs in the x-th place in i-th query and $T_i(R)$ is the probability that the correct song appears in the top R-th candidates of the i-th 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 had the same part of lyrics consisting of five words.

In our results, about 84% retrieval accuracy was obtained by the proposed method. As the retrieval accuracy itself is not better than that of the query-by-humming-based system (A. Ito, S.-P. Heo, Motoyuki Suzuki and Shozo Makino, 2004), this is a promising result.

3 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. To improve the acoustic model for modeling the singing voice, we tried to adapt the HMM to the singing voice using the 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 user's voice data for the adaptation.

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

In the following experiment, the MLLR (Maximum Likelihood Linear Regression) method (C.J.Leggetter and P.C.Woodland, 1995) was used as an adaptation algorithm. 127 choruses sung by 6 males were used as the adaptation data. These 6 singers were different from those who sang the test data. 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 gave more than 7 points of improvement in the word correct rate. Table 5 shows the retrieval accuracy results. These results prove the effectiveness of the adaptation.

4 IMPROVEMENT OF THE FSA: CONSIDERATION OF JAPANESE PHRASE STRUCTURE

The FSA used in the above experiments accepts any word sequences which are a sub-sequence 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. In Japanese, singing from a particle or a suffix hardly ever occurs. For example, in the following sentence

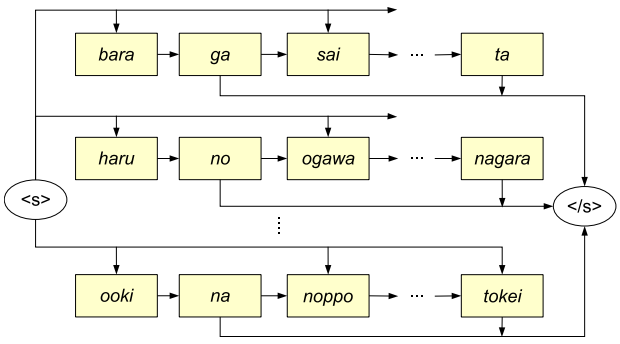

Figure 3: Example of improved grammar

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

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 2.2 so that:

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 Fig. 3.

The lyrics recognition experiment was carried out using the improved FSA. The adapted HMM described in Section 3 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 Table 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.

5 CONCLUSION

In pursuit of a MIR system that uses both melody and lyrics information in the singing voice, we tried to recognize lyrics in the users' singing voice. To exploit the constraints of the input song maximally, we used an FSA that accepts only a part of word sequences in the song database. Then we performed "singing voice adaptation" using MLLR speaker adaptation technology. Finally, we tried to introduce further linguistic constraints into the FSA. From the experimental results, the proposed methods proved to be effective. As a result, about 86% retrieval accuracy was obtained.

REFERENCES

A. Ito, S.-P. Heo, Motoyuki Suzuki and Shozo Makino. Comparison of Features for DP-matching based Query-by-humming System. In *Proc. ISMIR*, 10 2004.

Cambridge University Engineering Department. Hidden Markov Model Toolkit. http://htk.eng.cam.ac.uk/.

C.J.Leggetter and P.C.Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. In *Computer Speech and Language*, 4 1995.

Hironao Ozeki, Takayuki Kamata, Masataka Goto and Satoru Hayamizu. The influence of vocal pitch on lyrics recognition of sung melodies. In *The proceedings of the 2003 autumn meeting of the acoustical society of japan*, 9 2003.

J.S.Roger Jang, H.Lee and J.Chen. Super MBox: An Efficient/Effective Content-based Music Retrieval System. In *In the ninth ACM Multimedia Conference(Demo paper)*, 2001.

J.S.Roger Jang, Jiang-Chun, Ming-Yang Kao. MIRACLE: A Music Information Retrieval System with Clustered Computing Engines. In *International Symposium on Music Information Retrieval*, 2001.

Naoko Kosugi, Yuichi Nishihara, Tetsuo Sakata, Masashi Yamamoto and Kazuhiko Kushima. A Practical Query-By-Humming System for a Large Music Database. In *ACM Multimedia 2000*, 2000.

Rodger J. McNab, Lloyd A. Smith, David Bainbridge and Ian H. Witten. The NewZealand Digital Library MELody inDEX. D-Lib Magazine, May 1997.

Stanford University. Themefinder. http://www.themefinder.org/.

Toru Hosoya, Motoyuki Suzuki, Akinori Ito and Shozo Makino. Song retrieval system using the lyrics recognized vocal. In *The proceedings of the 2004 autumn meeting of the acoustical society of japan*, 9 2004.

University of Bonn. MidiLiB. http://www-mmdb.iai.uni-bonn.de/forschungprojekte/midilib/english/.

University of Karlsruhe. Tuneserver. http://name-this-tune.com/.