INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Lyrics recognition from singing voice focused on correspondence between voice and notes

Motoyuki Suzuki, Sho Tomita and Tomoki Morita

Faculty of Information Science and Technology, Osaka Institute of Technology, Japan

moto@m.ieice.org

Abstract

Lyrics recognition from singing voice is one of the most important techniques for query-by-singing music information retrieval systems. Lyrics information realizes a higher retrieval performance than retrieval using only melody information.

However, recognizing a song lyrics from singing voice is very difficult. In order to improve recognition, a new method focused on correspondence between voice and notes has been proposed. Note boundary scores are calculated for each frame, and these values are included in feature vectors by expanding their dimensions. The marker HMM is defined to correspond to feature vectors located at note boundaries, and the marker HMM is inserted among all morae in a pronunciation dictionary. As a result, the recognizer restricts an individual mora to correspond to only one note.

We also modified the marker HMM in order to account for short pauses in a particular position. A short pause corresponding to a musical rest or breath may occur after any morae, even if inside a word. The short pause HMM is concatenated to the marker HMM, and a skip transition arc of the short pause HMM is also introduced.

From experimental results, the proposed model provided higher word accuracy than the baseline model. It improved word accuracy from 85.71% to 93.18%, which means that 52.3% of the word error rate decreased. Insertion errors, especially, were drastically suppressed.

Index Terms: lyrics recognition, note boundary information, onset detection function score

1. Introduction

Lyrics recognition from singing voice is one of the most important techniques for query-by-singing music information retrieval systems. A lyrics text is extracted from an input singing voice by using a speech recognizer and is used as a retrieval key. If lyrics recognition is carried out without recognition error, then this information realizes higher retrieval performance than retrieval using only melody information.

However, recognizing a song lyrics from singing voice is very difficult, even though speech recognition can be used for practical use. Our previous experiment showed that word accuracy was only 12.1% for singing voice. On the other hand, speech data, which consist of the same text as lyrics, could be recognized with 88.5% accuracy. In this experiment, LVCSR (Large Vocabulary Continuous Speech Recognition system) for speech data was used without any adaptation to singing voice.

In general, singing voices have several features that reduce recognition accuracy:

1. Different utterance style. Spectral features of singing voice are different from that of read speech. It causes a mismatch between acoustic models and feature vectors.
2. Vowels may have a very long duration. Each mora in lyrics corresponds to each musical note. If a note has a long duration (whole note, half note, and so on) then a mora assigned to the note is sung with a long duration. In general, consonants are not stretched, but vowels are. It leads to increasing insertion errors.
3. Short pauses (silence) may be inserted in any position. In read speech, a short pause is not inserted into the inner word, but such a pause may occur in sung lyrics because a musical rest may be inserted in any position. Moreover, other types of short pauses corresponding to breath may also be inserted. Such short pauses cannot be recognized as a short pause because language models do not suppose these situations. As a result, these are recognized as plosive phonemes (ex. /k/, /b/) because these phonemes start with a short pause.

The first problem can be solved by adapting the acoustic model for singing voice[1, 2, 3] by using a speaker adaptation method such as MLLR[4]. This method drastically improves recognition accuracy for singing voice. From our previous experiments, the adapted acoustic model provided 85.7% recognition accuracy.

In order to solve the second and third problems, a special language model was proposed[5, 6]. Thinking of a song input for music information retrieval, it seems reasonable to assume that the input lyrics are part of a song in the lyrics database. In other words, the output text from a lyrics recognizer should be a part of the text in the lyrics database. To introduce this assumption, a finite state automaton (FSA) is used as a language model instead of an n-gram.

This is a very strong constraint for a speech recognizer, which does not accept any lyrics that have small difference. If you sing a song with a different word, it cannot be recognized correctly. FSA grammar is "hard" grammar compared with *n*-gram grammar. This is one of the problems of this method.

In this paper, we propose a new singing voice recognition method. The primary cause of the second and third problems is that lyrics are sung in correspondence with musical notes. We focus on the correspondence between notes and lyrics and introduce time information of note boundaries into a recognition algorithm.

2. New lyrics recognition method focused on note boundaries

2.1. Suppression of insertion error

In most of songs, especially Japanese songs, one note corresponds to one mora in sung lyrics. If time information for each note boundary is given, then a singing voice recognizer can restrict that one period corresponding to one note should correspond to one mora. We introduce such a restriction.

http://dx.doi.org/10.21437/Interspeech.2019-1318

A note boundary time, especially the start time of each note, is estimated by several methods called "onset detection method"[7]. However, estimation accuracy is not perfect. On the other hand, one note sometimes corresponds to two or more morae, and vice versa. In order to deal with such inaccuracies and ambiguities, "onset score" is introduced instead of determining the "onset frame."

The onset score[8] is calculated for each frame. A higher score (maximum is 1) indicates that the frame seems to be the onset frame. Onset scores for each frame are added into each feature vector by expanding dimension, and a pronunciation dictionary restricts note boundary frames to only appear at mora boundaries. A total score for a hypothesis is calculated by the weighted sum of acoustic score (calculated from acoustic model), linguistic score (calculated from language model), and onset likelihood (calculated from onset score). The hypothesis with the highest total score is selected as the final recognition result.

Details of the proposed algorithm is as follows:

1. Expand dimension of feature vector
   The "onset score" is calculated for each frame by using the onset detection method[7], and these values are normalized so that their range becomes $0 \sim 1$. Afterward, the acoustic feature vector is expanded in one dimension, and an onset score is stored in it.

2. Expand dimension of mean and variance vectors in all phoneme HMMs
   In this algorithm, HMMs are used as an acoustic model. Each state of HMMs has a diagonal Gaussian mixture distribution. Both mean and variance vectors in all Gaussian distributions are expanded in one dimension corresponding to the onset score dimension. The value of the expanded dimension is set to 0 (which means "inner note") in all mean vectors, and a pre-defined (appropriate) value is used in all variance vectors.

3. Define the "marker HMM"
   The special HMM, called "marker HMM," is introduced in order to correspond to a note boundary frame. The marker HMM has only one state without self-loop. It has only one Gaussian distribution, which is trained by all training data. After this, both mean and variance vectors are expanded in one dimension. The value of an expanded dimension in the mean vector is set to 1, and that of the variance vector is the pre-defined value that is the same as other phoneme HMMs.

4. Add the marker HMM in a pronunciation dictionary
   The marker HMM is inserted at the end of all morae in the pronunciation dictionary. For example, the Japanese word "onsei" (which means "voice") has four morae (/o/, /N/, /se/, and /i/). The entry in the pronunciation dictionary is "o # N # s e # i #." In this example, /#/ indicates the marker HMM.

The pronunciation dictionary constrains the number of morae that is between the marker HMMs to one. The marker HMM provides a higher likelihood for the feature vector with a higher onset score. As a result, the hypothesis in which each mora corresponds to each note is provided higher likelihood. A total score is calculated from acoustic and linguistic likelihood in addition to onset score. For the final result, we can use the highest score to determine the most appropriate hypothesis.

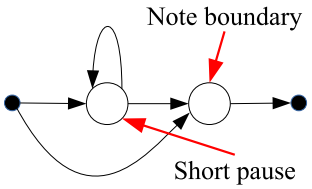

Figure 1: *Marker HMM*

2.2. Considering the short pause problem

Short pauses corresponding to a musical rest or breath may be inserted at the end of any note. In order to account for it, the marker HMM is modified.

The marker HMM corresponds to note boundary frames, and a short pause may be inserted in front of it. Therefore, "short pause(sp) HMM" is attached in front of the marker HMM. The sp HMM is normally trained in traditional speech recognition system. Typically, it has one state with a self-loop. The new marker HMM is constructed by concatenating with the sp HMM and the marker HMM.

However, short pause is not always inserted in front of a note boundary. The sp HMM should be skipped when a short pause is not inserted. Therefore, a skip transition arc, which skips the first state, is added in the new marker HMM.

In conclusion, the structure of the marker HMM is shown in Fig. 1. It has two states, the first being the same as the sp HMM, and the second is the same as the marker HMM described in the Sec. 2.1. The skip transition of the first state is added.

3. Experiments

In order to investigate the effectiveness of the proposed method, singing voice recognition experiments were carried out.

3.1. Database

Singing voice data, consisting of 198 units sung by 19 males and 8 females, were used. The data were pieces of 48 different Japanese children's songs. Table 1 shows a breakdown of data by correspondence between notes and morae. There were only one-by-one correspondences (all notes in data correspond to each note) for 103 units of data. Some one-by-two (one note corresponds to two morae) correspondences existed in 61 units of data, and some two-by-one correspondences existed in 21 units of data. Of the data units that included both types of correspondence, there were 13, and there was no data that included any correspondence between one note and three or more morae, and vice versa.

3.2. Configuration of experiment

Table 2 shows the experimental setup. An acoustic model was trained by using all singing voice data sung by 26 singers, and other data sung by another singer was used as test data. Each singer was selected as test data, and a total of 27 experiments were carried out. The average accuracy of these experiments was then calculated. In these experiments, monophone HMM was used as the acoustic model.

A word trigram was used as the language model. The trigram was trained by using only the lyrical text of 48 Japanese children's songs in the database. The number of vocabulary words was 314, and there were no out-of-vocabulary words.

Table 1: *Breakdown of the data by correspondence*

		2 notes : 1 mora	
		not included	included
1 note : 2 morae	not included	103	21
	included	61	13

Table 2: *Experimental Setup*

Database		Tokushima Univ. singing database
	Song	48 Japanese children's song
	Singer	19 males, 8 females
	Number of units	198
Speech recognizer		julius[10]
	Acoustic model	monophone trained by singing data
	Language model	trigram trained by lyrics
	Vocabulary size	314 (No out-of-vocabulary)
Onset score		Rectified complex deviation[8]

Several onset detection methods have been proposed, and several types of onset scores are employed in these methods. For these experiments, the rectified complex deviation[8] calculated by OnsetsDS Vamp plugin[9] was used as the onset score.

3.3. Lyrics recognition results

Table 3 shows the lyrics recognition accuracy and error rates for each of the methods. "Acc." indicates word accuracy, and "Sub.," "Del.," and "Ins." indicate substitution, deletion, and insertion error rate, respectively. In this table, "baseline" refers to the recognition method without note boundary information. In this model, both acoustic and language models were adapted to singing voice recognition, but correspondence between notes and morae were not considered. The use of "+ boundary score" indicates that the model considered the note boundary score and the marker HMM to be comprised of only one state. It does not connect with sp HMM. The sp HMM is connected with the marker HMM in the "+ sp HMM" model.

From these results, we found that both models are effective for lyrics recognition, considering that the note boundary score improved 4.17 points, and the addition of sp HMM to the marker HMM improved scores by an additional 3.30 points. Totally, recognition accuracy was improved 7.47 points, or, in other words, the proposed method provided a 52.3% decrease in the word error rate.

We also checked the details for errors. When considering that the boundary score led to a decrease in insertion error (9.04% $\rightarrow$ 5.26%), we can conclude that many insertion errors into long vowel periods were suppressed. Figure 2 shows an example of time alignment output as determined by each model.

Table 3: *Lyrics recognition accuracy*

Model	Acc.	Sub.	Del.	Ins.
baseline	85.71%	4.43%	0.83%	9.04%
+ boundary score	89.88%	3.69%	1.17%	5.26%
+ sp HMM	93.18%	2.82%	0.87%	3.13%

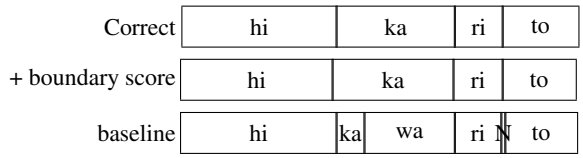

Figure 2: *Time alignment of recognized result*

In this figure, the horizontal axis indicates time, and the width of each box indicates time duration for each mora. As you can see from "Correct" boxes, this example consists of four morae, and the first two morae correspond to long notes.

The first mora, /hi/, was correctly recognized by the baseline model, but the second mora, /ka/, was split into two morae, /ka/ and /wa/. It is a typical insertion error. On the other hand, the "+ boundary score" model could recognize it correctly. The mora, /N/, was also inserted between /ri/ and /to/ with a very short duration in the "baseline" model. The reason for this error is not a long duration of the mora, /ri/. The insertion error of /wa/ led to the insertion error of /N/ because the word *n*-gram gave a higher score to /hikawariNto/ than /hikawarito/.

The sp HMM also decreased the insertion error rate (5.26% $\rightarrow$ 3.13%). In both "baseline" and "+ boundary score" models, some morae were inserted at the short pause position, but the "+ sp HMM" model could recognize such a short pause as is, and it reduced the insertion of error words.

Figure 3 shows an example of the recognition results. The figure displays sung data of the song "Umi," for which the musical score is shown in Fig. 4. In Fig. 3, the text indicates the recognition result given by the "baseline" model, and vertical red lines indicate the boundaries of recognized morae.

Figure 4 shows that the mora, /na/, corresponds to a quarter note. Therefore, a duration of this mora should be the same as the duration of the mora, /i/. However, waveform in Fig. 3 shows that the duration of /na/ was similar to the duration of /hi/ or /ro/, which correspond to an eighth note. It means that a singer did not sing with the anticipated duration but with a shorter duration, and inserted a breath or musical rest after the mora, /na/. Such an insertion of a short pause frequently

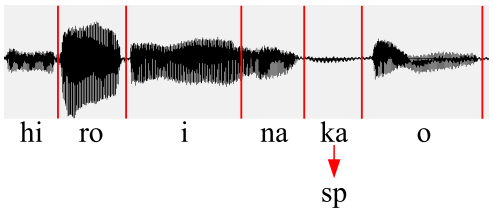

Figure 3: *Example of recognition results*

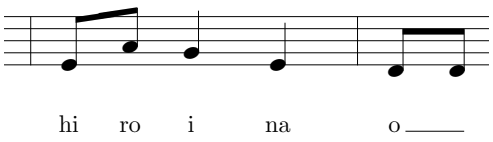

Figure 4: *A Part of the musical score of the song "Umi"*

occurred, especially in *a cappella* singing style. Because the baseline model cannot deal with such short pauses, the mora, /ka/, was inserted at the short pause position. This insertion error could be recovered by adding sp HMM to the marker HMM.

4. Conclusions

In order to improve the performance of singing voice recognition, a new recognition method was proposed. Note boundary scores are calculated for each frame, and these values are included in feature vectors by expanding their dimensions. The marker HMM is defined to correspond to feature vectors located at note boundaries, and the marker HMM is inserted among all morae in the pronunciation dictionary. As a result, the recognizer restricts an individual mora to correspond to only one note.

We also modified the marker HMM in order to account for short pauses in a particular position. The sp HMM is concatenated to the marker HMM, and the skip transition arc of the sp HMM is also introduced. This modification made it possible to account for short pauses corresponding to a musical rest or breath located after any morae.

From these experimental results, the proposed model provided higher word accuracy than the baseline model. It improved word accuracy from 85.71% to 93.18%, which means that 52.3% of the word error rate decreased. Insertion errors, especially, were drastically suppressed.

5. Acknowledgements

A part of this work was supported by JSPS KAKENHI Grant Number JP18K11321.

6. References

[1] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, 2010, article ID 546047, doi:10.1155/2010/546047.

[2] D. Kawai, K. Yamamoto, and S. Nakagawa, "Consideration of lyrics recognition in monophonic singing using DNN-HMM," in *IPSJ SIG Technical Report*, vol. 2015-MUS-107, no. 58, 2015, pp. 1–6, (in Japanese).

[3] H. Ozeki, T. Kamata, M. Goto, and S. Hayamizu, "The influence of vocal pitch on lyrics recognition of sung melodies," in *Proc. 2003 Autumn Meeting of The Acoustical Society of Japan*, Sep. 2003, pp. 637–638, (in Japanese).

[4] 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.

[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] M. Suzuki, T. Hosoya, A. Ito, and S. Makino, "Music information retrieval from a singing voice using lyrics and melody information," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007, article ID 38727, 8 pages, doi:10.1155/2007/38727.

[7] D. Stowell and M. Plumbley, "Adaptive whitening for improved real-time audio onset detection," in *Proc. International Computer Music Conference*, 2007, pp. 312–319.

[8] S. Dixon, "Onset detection revisited," in *Proc. 9th International Conference on Digital Audio Effects*, 2006, pp. 133–137.

[9] C. Cannam and D. Stowell, "OnsetsDS," 2007, https://code.soundsoftware.ac.uk/projects/vamp-onsetsds-plugin.

[10] A. Lee, T. Kawahara, and K. Shikano, "Julius — an open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH*, 2001, pp. 1691–1694.