Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals

Hiromasa Fujihara,† Masataka Goto,‡ Jun Ogata,‡
Kazunori Komatani,† Tetsuya Ogata,† and Hiroshi G. Okuno†

†Dept. of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
{fujihara, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp

‡National Institute of Advanced Industrial
Science and Technology (AIST)
Tsukuba, Ibaraki 305-8568, Japan
{m.goto, jun.ogata}@aist.go.jp

Abstract

This paper describes a system that can automatically synchronize between polyphonic musical audio signals and corresponding lyrics. Although there were methods that can synchronize between monophonic speech signals and corresponding text transcriptions by using Viterbi alignment techniques, they cannot be applied to vocals in CD recordings because accompaniment sounds often overlap with vocals. To align lyrics with such vocals, we therefore developed three methods: a method for segregating vocals from polyphonic sound mixtures, a method for detecting vocal sections, and a method for adapting a speech-recognizer phone model to segregated vocal signals. Experimental results for 10 Japanese popular-music songs showed that our system can synchronize between music and lyrics with satisfactory accuracy for 8 songs.

1. Introduction

A vocal track and its lyrics play an important role in many musical genres, especially in popular music. To represent the theme and story of the song, they are essential elements that create an impression of the song. When a song is heard, for example, most people listen to the vocal melody and follow the lyrics. This is why music videos often display synchronized lyrics as a caption, helping the audiences enjoy the music.

In this paper we propose an automatic synchronization system for polyphonic audio signals of songs and their lyrics. This system can automatically estimate the temporal relationship (alignment) between audio signals and the corresponding lyrics. This approach is different from direct lyrics recognition like speech recognition and takes advantage of the vast amount of lyrics embedded in the web. Our system has various applications, such as automatic generation of music video captions and a music playback interface that can directly access to specific words or passages of interest.

Wang *et al.* [19] have worked on a similar system. They have integrated higher structural information (such as beat tracking and chorus detection) and lower level lyrics alignment. Their lower level lyrics alignment method uses only the duration of each phoneme as a cue. However, this method is not consistently effective because the durations of uttered phonemes differ based on location, even though they are the same phonemes. The method also requires many assumptions about the structure and meter of the song in order to obtain higher structural information. Other related studies have focused on lyrics recognition[8, 16, 7]. They use a speech recognizer for lyrics recognition. These studies presume pure monophonic singing voices without accompaniment, posing additional difficulties for practical use with musical audio signals like CD recordings.

Because current speech recognition techniques are incapable of automatically synchronizing lyrics with music including accompaniments, we developed three methods: a method for segregating vocal (singing) signals from polyphonic audio signals, a method for detecting sections including vocal signals, and a method for adapting a phone model of speech recognizers to segregated vocal signals.

The rest of this paper is organized as follows. In the next section, we describe an overview of our system for automatic synchronization between music and lyrics. From Section 3 to Section 5, we describe our system in detail. In Section 6, we describe our experiments and present the results. In Section 7, we draw conclusions and discuss future directions.

Proceedings of the Eighth IEEE International Symposium on Multimedia (ISM'06)
0-7695-2746-9/06 $20.00 © 2006 IEEE

2. Automatic Synchronization between Music and Lyrics

Given musical audio signals and the corresponding lyrics, our system can locate the start and end times for each phrase of the lyrics by automatic synchronization between music and lyrics. The system deals with real-world musical audio signals such as popular music CD recordings that contain a singer's vocal track and various accompaniment sounds. We assume that the main vocal part is sung by a single predominant singer (except for choruses), but do not assume anything about the number and kind of sound sources in the accompaniment sounds.

To solve this problem, the basic idea is to use the Viterbi alignment (forced alignment) technique that is often used in automatic speech recognition. This technique, however, does not work well when there are accompaniment sounds that are performed together with a vocal and when there are interlude sections during which the vocal is not performed. We therefore first extract and resynthesize the harmonic structure of the vocal melody and obtain the vocal signals by using an *accompaniment sound reduction method* that we proposed previously [4]. Then, we detect the vocal sections (regions) from segregated vocal signals by using a *vocal activities detection method* based on a Hidden Markov Model (HMM). Finally, we align the lyrics with the segregated vocal audio signals by using a *Viterbi alignment technique*. We also propose a method for adapting a phone model to the segregated vocal signals of the target singer.

3. Accompaniment Sound Reduction

To extract a vocal feature representing the phonetic information of a singing voice from polyphonic audio signals, it is necessary to reduce the influence of accompaniment sounds. We therefore resynthesize vocal signals from the harmonic structure of the melody line by the following three parts:

1. Estimating the fundamental frequency (F0) of the melody line (vocal) in CD recordings by using a predominant-F0 estimation method called PreFEst [5].
2. Extracting the harmonic structure corresponding to the F0 of the melody line.
3. Resynthesizing the audio signal (waveform) corresponding to the melody line using a sinusoidal synthesis.

Thus, we obtain a waveform corresponding only to the melody line. Figure 1 shows an overview of the accompaniment sound reduction method. Note that the melody line obtained with this method may contain instrumental (*i.e.*, non-vocal) sounds in interlude sections as well as voices in vocal sections, because the melody line here is defined as the most predominant F0 in each frame [5]. It is therefore necessary to detect the vocal sections by using the method described in section 4.

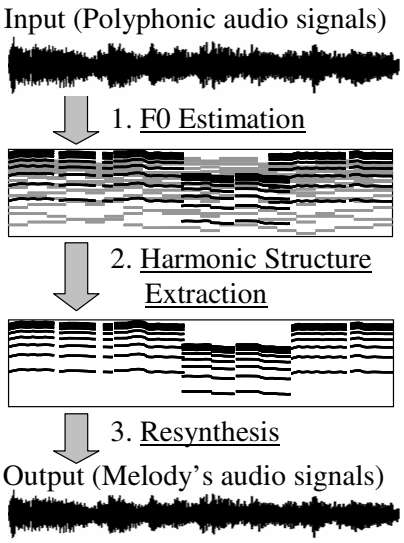

Figure 1. Accompaniment Sound Reduction.

3.1. F0 Estimation

We use Goto's PreFEst [5] to estimate the F0 of the melody line. PreFEst estimates the most predominant F0 in frequency-range-limited sound mixtures. Since the melody line tends to have the most predominant harmonic structure in middle- and high-frequency regions, we can estimate the F0s of the melody line by applying PreFEst with adequate frequency-range limitation.

We describe a summary of PreFEst below. Hereafter, x is the log-scale frequency denoted in units of cents (a musical-interval measurement), and t is discrete time. Although a cent originally represented tone interval (relative pitch), we use it as a unit of absolute pitch using $440 \times 2^{\frac{3}{12}-5}$ Hz as a criterion, according to Goto [5]. The conversion from Hz to cent is expressed as follows:

$$f_{\mathrm{cent}} = 1200 \log_2 \frac{f_{\mathrm{Hz}}}{440 \times 2^{\frac{3}{12}-5}}, \qquad (1)$$

where f_{cent} and f_{Hz} represent frequency in cents and Hz, respectively.

Given the power spectrum $\Psi_p^{(t)}(x)$, we first apply a band-pass filter (BPF) that is designed so that it covers most of the dominant harmonics of typical melody lines. The filtered frequency components can be represented as $BPF(x)\Psi_p^{(t)}(x)$, where $BPF(x)$ is the BPF's frequency response for the melody line. In this paper, we designed the BPF according to Goto's specifications [5]. To make the application of statistical methods possible, we represent each of the bandpass-filtered frequency components as a probability density function (PDF), called an observed PDF,

$p_\Psi^{(t)}(x)$:

$$p_\Psi^{(t)}(x) = \frac{BPF(x)\Psi_p^{(t)}(x)}{\int_{-\infty}^{\infty} BPF(x)\Psi_p^{(t)}(x)dx}. \quad (2)$$

Then, we consider each observed PDF to have been generated from a weighted-mixture model of the tone models of all the possible F0s, which is represented as follows:

$$p(x|\theta^{(t)}) = \int_{\mathrm{Fl}}^{\mathrm{Fh}} w^{(t)}(F)p(x|F)dF \quad (3)$$

$$\theta^{(t)} = \{w^{(t)}(F)|\mathrm{Fl} \leq F \leq \mathrm{Fh}\}, \quad (4)$$

where $p(x|F)$ is the PDF of the tone model for each F0, and Fh and Fl are defined as the lower and upper limits of the possible (allowable) F0 range, and $w^{(t)}(F)$ is the weight of a tone model that satisfies

$$\int_{\mathrm{Fh}_i}^{\mathrm{Fl}_i} w^{(t)}(F)dF = 1. \quad (5)$$

A tone model represents a typical harmonic structure and indicates where the harmonics of the F0 tend to occur. Then, we estimate $w^{(t)}(F)$ using an EM algorithm and regard it as the F0's PDF. Finally, we track a dominant peak trajectory of F0s from $w^{(t)}(F)$ using multiple agent architecture.

3.2. Harmonic Structure Extraction

Based on the estimated F0, we extract the power of fundamental frequency component and harmonic components. For each component, we allow r cent error and extract the peak in the allowed area. The power A_l and frequency F_l of l th overtone ($l = 1, \ldots, L$) can be represented as

$$F_l = \underset{F}{\mathrm{argmax}} |S(F)| \quad (l\overline{F} \cdot (1 - 2^{\frac{\mathrm{r}}{1200}}) \leq F \leq l\overline{F} \cdot (1 + 2^{\frac{\mathrm{r}}{1200}})), \quad (6)$$

$$A_l = |S(F_l)|, \quad (7)$$

where $S(F)$ denotes the spectrum and $\overline{F}$ denotes the F0 estimated by PreFEst. In our experiments, we set r to 20.

3.3. Resynthesis

We resynthesize the audio signals of the melody line from the extracted harmonic structure using a sinusoidal model [12]. We denote the frequency and the amplitude of l th overtone at time t as $F_l^{(t)}$ and $A_l^{(t)}$, respectively. Changes of a phase are approximated using a quadratic function so that a frequency changes linearly, and changes of amplitude are approximated using linear function. Resynthesized audio signals, $s(k)$, are expressed as

$$\theta_l(k) = \frac{\pi(F_l^{(t+1)} - F_l^{(t)})}{K}k^2 + 2\pi F_l^{(t)}k + \theta_{l,0}^{(t)}, \quad (8)$$

$$s_l(k) = \left\{(A_l^{(t+1)} - A_l^{(t)})\frac{k}{K} + A_l^{(t)}\right\}\sin\left(\theta_l(k)\right), \quad (9)$$

$$s(k) = \sum_{l=1}^{L} s_l(k), \quad (10)$$

where k represents a time in units of seconds and defines time t as $k = 0$, K represents the duration between (t) and $(t+1)$ in units of seconds, and $\theta_{l,0}^{(t)}$ means an initial phase. In the first flame, $\theta_{l,0}^{(t)}$ was set to 0. Fron then on, $\theta_{l,0}^{(t)}$ was given by $\frac{\pi(F_l^{(t)} - F_l^{(t-1)})}{2K} + \theta_{l,0}^{(t-1)}$, where $F_l^{(t-1)}$ denotes a frequency of lth overtone in the previous flame and $\theta_{l,0}^{(t-1)}$ denotes an initial phase in the precious flame.

4. Vocal Activities Detection

We remove non-vocal sections using the vocal activities detection method. The melody waveform obtained with the accompaniment sound reduction method contains instrumental sounds in non-vocal sections. The existence of long non-vocal sections negatively influences the execution of the Viterbi alignment between the audio signal and the lyrics, if interlude sections are not removed.

We propose a vocal activities detection method that can control a balance between hit rate and correct rejection rate. Generally, there is a trade-off relationship between hit rate and correct rejection rate, and proper balance between them depends on the application. For example, since our system positions the vocal activities detection method as a preprocessing of Viterbi alignment, we attach importance to hit rate instead of correct rejection rate. In other words, we want to detect all regions that contain vocals. On the other hands, other applications such as singer identification require maintenance of a high correct rejection rate, and detection of the regions that certainly contain vocals.

In previous vocal activities detection methods [2, 18, 13], no studies have ever tried to control a balance between hit rate and correct rejection rate.

4.1. Basic Formulation

We introduce a Hidden Markov Model (HMM) that transitions back and forth between vocal state, s_V, and non-vocal state, s_N, as shown in figure 2. The vocal state means that vocals are present and the non-vocal state means that vocals are absent. Given the feature vectors of input audio signals, the problem is finding the most likely sequence of vocal and non-vocal states, $\hat{S} = \{s_1, \cdots, s_t, \cdots\}$.

$$\hat{S} = \underset{S}{\mathrm{argmax}} \sum_t \left\{\log p(\boldsymbol{x}|s_t) + \log p(s_{t+1}|s_t)\right\}, \quad (11)$$

where $p(\boldsymbol{x}|s)$ represents an output probability of state s, and $p(s_i|s_j)$ represents a state transition probability from state s_j to state s_i.

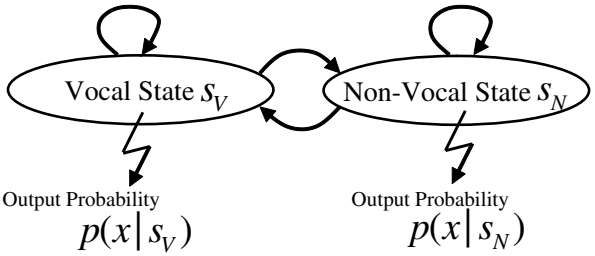

Figure 2. A Hidden Markov Model (HMM) for vocal activities detection.

The output log probability of each state is approximated with the following equations:

$$\log p(\boldsymbol{x}|s_V) = \log \mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x};\theta_V) - \frac{1}{2}\eta, \tag{12}$$

$$\log p(\boldsymbol{x}|s_N) = \log \mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x};\theta_N) + \frac{1}{2}\eta, \tag{13}$$

where $\mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x};\theta)$ denotes the probability density function of the Gaussian mixture model (GMM) with parameter θ, and η represents a threshold parameter that controls trade-off between hit rate and correct rejection rate. The parameters of the vocal GMM, θ_{V}, and the vocal GMM, θ_{N}, are trained on feature vectors extracted from vocal sections and nonvocal sections of the training data set, respectively. We set the number of mixture of the GMM at 64.

4.2. Calculation of Threshold

The balance of vocal activities detection is controlled by changing η in Equations (12) and (13). However, there is bias in the log likelihoods of GMMs for each song and it is difficult to decide universal value of η. Therefore, we divide η into bias correction value, $\eta_{\mathrm{dyn.}}$, and application dependent value, η_{fixed}.

$$\eta = \eta_{\mathrm{dyn.}} + \eta_{\mathrm{fixed}} \tag{14}$$

While the application dependent value, η_{fixed}, is set by hand, the bias correction value, $\eta_{\mathrm{dyn.}}$, is determined by using Otsu's method for threshold selection [15] as follows. We first calculate a difference of log likelihood, $l(\boldsymbol{x})$, of all the feature vectors in input audio signals.

$$l(\boldsymbol{x}) = \log \mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x};\theta_V) - \log \mathcal{N}_{\mathrm{GMM}}(\boldsymbol{x};\theta_N). \tag{15}$$

Then we create a histogram of $l(\boldsymbol{x})$ and select a threshold to maximize the between-class variance.

4.3. Feature Extraction

Feature vectors used in this method consist of the following two features.

- LPC-derived mel cepstral coefficients (LPMCCs):
  We use LPC-derived mel cepstral coefficients (LPMCCs) as spectral feature for vocal/non-vocal discrimination. LPMCCs are mel-cepstral coefficients of the LPC spectrum [1]. We have reported that, in the context of singer identification, LPMCCs represent vocal characteristics better than mel-frequency cepstral coefficients (MFCCs) [3, 10], which are widely used for music modeling [4]. We derive LPMCCs by computing the MFCC from the LPC spectrum because of simplicity of implementation.

- ΔF0s:
  We use the derivative of F0s (ΔF0s) [14], which represent the dynamics of the F0's trajectory, because singing voice tends to have temporal variation of F0s because of vibrato and, therefore, ΔF0s are expected to be good cues for vocal/non-vocal discrimination.

  ΔF0 is calculated as a regression parameter over five frames:

$$\Delta f[t] = \frac{\sum_{k=-2}^{2} k \cdot f[t+k]}{\sum_{k=-2}^{2} k^2}, \tag{16}$$

  where $f[t]$ represents the frequency in units of cents at the time t.

5. Viterbi Alignment

In this section, we describe our method of executing Viterbi alignment between lyrics and segregated signals. We first create a language model from the given lyrics and then extract feature vectors from segregated vocal signals. Finally, we execute the Viterbi alignment between them. We also describe our method of adapting a phone model to the specific singer of the input audio signals.

5.1. Lyrics Processing

Given the lyrics corresponding to input audio signals, we create a language model for forced alignment. In this language model, we deal with only vowel phonemes because the unvoiced consonant phonemes do not have a harmonic structure and cannot be extracted by using the accompaniment sound reduction method. In addition, the voiced consonant phonemes are usually uttered for a short time and it is difficult to estimate F0s stably. We first convert the lyrics to a sequence of the phonemes and then create a language model using the following rules:

- Ignore all the consonant phoneme except the syllabic nasal.

- Convert each boundary of sentences or phrases into multiple short pauses.

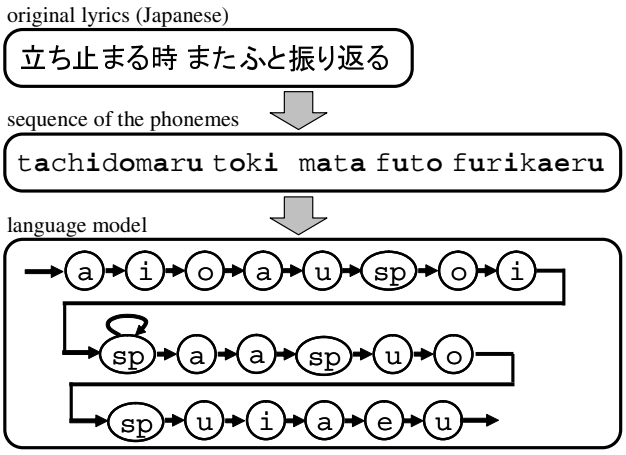

Figure 3. Example of lyrics processing.

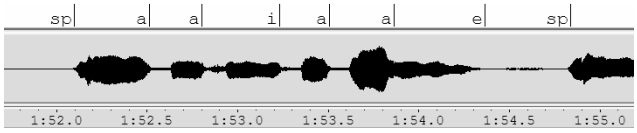

Figure 4. Example of phoneme labels.

- Convert each boundary of words into a single short pause.

Figure 3 shows an example of conversion from lyrics to the language model.

5.2. Adaptation of a Phone Model

We adapt a phone model to the specific singer of input audio signals. Our adaptation method consists of the following three steps:

Step 1. We adapt a phone model for clean speech to a clean singing voice.

Step 2. We adapt the phone model for a clean singing voice to the singing voice segregated using the accompaniment sound reduction method.

Step 3. We adapt the phone model for segregated speech to the specific singer of input audio signals using the unsupervised adaptation method.

Steps 1 and 2 are carried out preliminarily, and step 3 is carried out at runtime.

As an adaptation method, we use MLLR and MAP adaptation. We manually annotated phoneme labels to the adaptation data for supervised adaptation. Figure 4 shows an example of phoneme labels.

5.3. Alignment

We execute the Viterbi alignment (forced alignment) by using the language model created from the given lyrics, the feature vectors extracted from segregated vocal signals, and the adapted phone model for the specific singers. MFCCs [3], Δ MFCCs, and Δ power are used as a feature vector for the Viterbi alignment.

Table 1. Evaluation data set from RWC-MDB-P-2001.

Song #	Singer Name	Gender
012	Kazuo Nishi	Male
027	Shingo Katsuta	Male
032	Masaki Kuehara	Male
037	Hatae Yoshinori	Male
039	Kousuke Morimoto	Male
007	Tomomi Ogata	Female
013	Konbu	Female
020	Eri Ichikawa	Female
065	Makiko Hattori	Female
075	Hiromi Yoshii	Female

6. Experiments

We conducted the experiments to evaluate the performance of our system.

6.1. Experimental Condition

For the evaluation data set, we used the 10 songs listed in Table 1 taken from "RWC Music Database: Popular Music" (RWC-MDB-P-2001) [6]. These songs are mainly sung in Japanese, but small portion of the vocal part is sung in English. In this experiments, we approximate the English phonemes using similar Japanese phonemes. Using these data, we conducted a 5 fold cross validation for each gender, that is, when we evaluated a song by a particular singer we adapted a phoneme model using the remaining songs of the same gender.

As the training data for the vocal activities detection method, we used 19 songs of 11 singers listed in Table 2 which are also taken from "RWC Music Database: Popular Music" (RWC-MDB-P-2001). These 10 singers differ from the singers used for evaluation. We applied the accompaniment sound reduction method to these training data. We set η_{fixed} to 1.5.

Table 3 shows the analysis conditions of the Viterbi alignment. As an initial phone model, we used the gender dependent monophone model of ISRC Software [9]. For conversion from lyrics to a sequence of phonemes, we use the readings created by ChaSen [11], which is a Japanese morphological analysis system. For feature extraction, the Viterbi alignment, and an adaptation of the phone model, we use HCopy, HVite,and HAdapt in the Hidden Markov Toolkit (HTK) [17].

Table 2. Training data for vocal activities detection from RWC-MDB-P-2001.

Singer Name	Gender	Piece Number
Hiroshi Sekiya	M	048, 049, 051
Katsuyuki Ozawa	M	015, 041
Masashi Hashimoto	M	056, 057
Satoshi Kumasaka	M	047
Oriken	M	006
Tomoko Nitta	F	026
Kaburagi Akiko	F	055
Yuzu Iijima	F	060
Reiko Sato	F	063
Tamako Matsuzaka	F	070
Donna Burke	F	081, 089, 091, 093, 097

Table 3. Analysis conditions of Viterbi alignment.

Sampling	16 kHz, 16 bit
Window function	Hamming
Frame length	25 ms
Frame period	10 ms
Feature vector	12th order MFCC 12th order ΔMFCC ΔPower

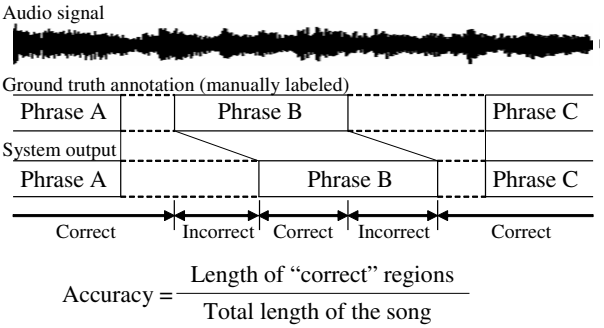

Figure 5. Evaluation measure.

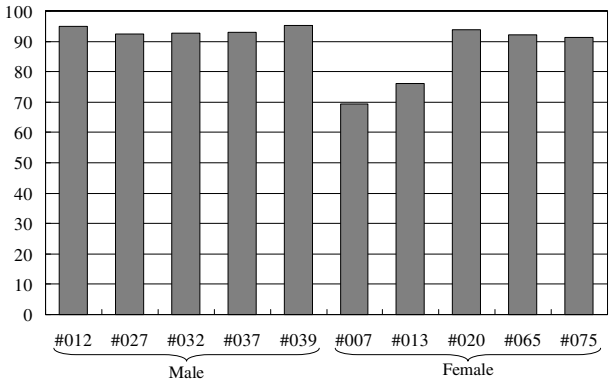

Figure 6. Experimental result: evaluation of whole system.

Evaluation was done by using phrase level alignment. In these experiments, we define phrase as a section that was delimited by a space or a line feed in the original lyrics. As an evaluation measure, we calculate proportion of a length of the sections that are correctly labeled in phrase level to a total length of a song (Figure 5). The system output of a song is judged to be satisfactory if its accuracy was over 90%.

6.2. Evaluation of Whole System

We conducted experiments using the system in which all the methods described in this paper were implemented. Figure 6 shows the results of these experiments.

6.3. Evaluation of Adaptation Method

The purpose of this experiment was to investigate the effectiveness of the adaptation method. We conducted experiments under the following four conditions:

(i) No adaptation: We did not execute the phone model adaptation.

(ii) One-step adaptation: We adapted a phone model for clean speech directly to segregated vocal signals. We did not execute the unsupervised adaptation to input audio signals.

(iii) Two-step adaptation: First, we adapted a phone model for clean speech to clean vocal signals, and then we adapted the phone model to segregated vocal signals. We did not execute the unsupervised adaptation to input audio signals.

(iv) Three-step adaptation (proposed): First, we adapted a phone model for clean speech to clean vocal signals, and then we adapted the phone model to segregated vocal signals. Finally, we adapted the phone model to the specific singer of input audio signals.

In this experiment, the vocal activities detection method was enabled. Figure 7 shows the result of these experiments.

6.4. Evaluation of Vocal Activities Detection

The purpose of this experiment was to investigate the effectiveness of the vocal activities detection method. We also investigated the performance of the vocal activities detection method. We compared the results of disabling the

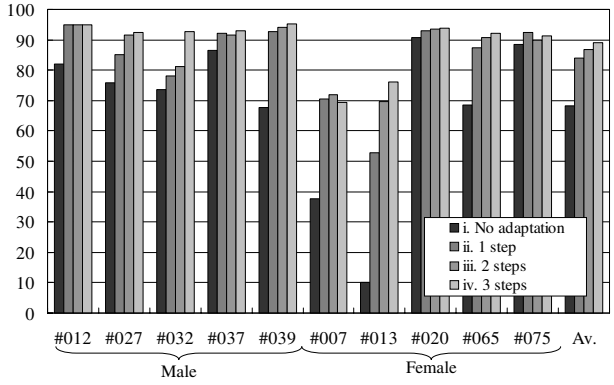

Figure 7. Experimental result: evaluation of adaptation.

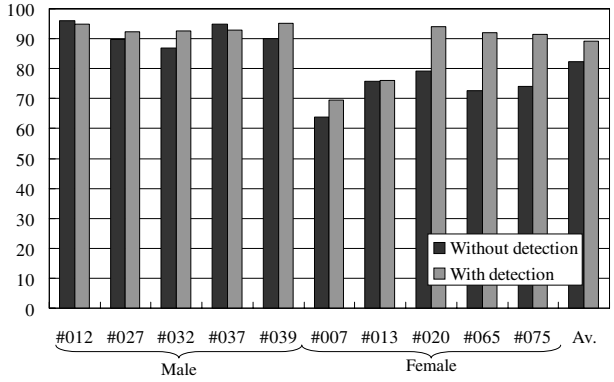

Figure 8. Experimental result: evaluation of vocal activities detection.

vocal activities detection with the results of enabling it. In this experiment, the three-step adaptation was executed to adapt a phone model to the specific singer of input audio signals. Figure 8 shows the results of these experiments and Figure 9 shows hit rate and correct rejection rate of the vocal activities detection method.

6.5. Discussion

As shown in Figure 6, the accuracy was more than 90% for 8 songs. When we compare the results for the males and female singers, the accuracy for the female singers is lower. This is because it is difficult to capture the characteristics of the voices with high F0 [16]. When we analyzed errors in each song, typical errors occurred at sections in which the lyrics is sung in English. This means that it is sometimes difficult to approximate English phonemes using similar Japanese phonemes. To overcome this problem, we plan to use an English phone model in combination with a Japanese one. Other typical errors were caused by a singing

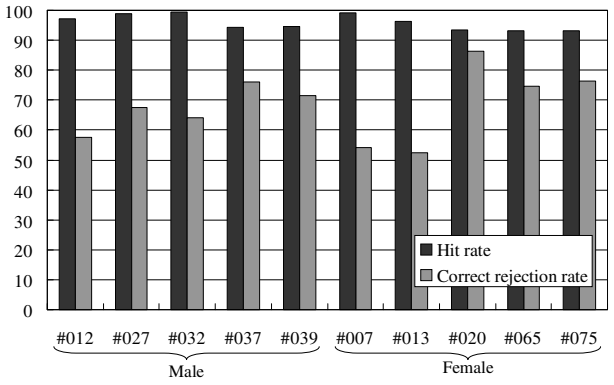

Figure 9. Experimental result: hit rate and correct rejection rate of vocal activities detection.

style of humming that was not written in the lyrics.

As shown in Figure 7, our adaptation method was effective for all 10 songs, and as shown in Figure 8, the vocal activities detection method was also effective when applied to the songs with relatively low accuracy. On the other hand, this method had poor efficacy to the songs #007 and #013, even though the accuracies for these songs are relatively low. This is because correct rejection rates for these songs were not so high as shown in Figure 9. In addition, when this method was applied to songs with higher accuracy, #012 and #037, the accuracy slightly decreased. This is because the sections improperly rejected by the vocal activities detection method are always judged incorrect.

7. Conclusions

We have described a system for automatically synchronizing between musical audio signals and corresponding lyrics. Our system consists of the following three methods: accompaniment sound reduction, vocal activities detection, and Viterbi alignment. We also propose a method for adapting a phoneme model to the segregated vocal signals of the specific singer. Experimental results showed that our system is robust enough to synchronize lyrics with real-world music signals containing sounds of various instruments.

The main contributions of this paper can be summarized as follows:

- We first dealt with the problem of synchronization between music and lyrics forthrightly by segregating vocal signals from sound mixtures and recognizing the phonemes. Due to negative influences caused by accompaniment sounds, no other studies have succeeded in applying the technique of speech recognition to this task.

- We proposed an original vocal activities detection method that can control the trade-off between hit rate

and correct rejection rate by changing the parameter. Although the balance between hit rate and correct rejection rate differs depending on the application, little attention has been given to this viewpoint. We enabled it by dividing the threshold into a bias correction value and an application dependent value, and obtaining the bias correction value automatically by using Otsu's method [15].

- We proposed a method to adapt a phone model for speech to segregated vocal signals. This method is useful for both alignment between music and lyrics and lyrics recognition of polyphonic audio signals.

In the future, we plan to conduct experiments using songs sung in languages other than Japanese. We also plan to incorporate higher-level information such as song structures and achieve more advanced synchronization between music and lyrics.

8. Acknowledgements

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid for Scientific Research (A), No. 15200015, and Information Research Center for Development of Knowledge Society Infrastructure (COE program of MEXT, Japan). We thank Tetsuro Kitahara, Kazuyoshi Yoshii (Kyoto University) and Tomoyasu Nakano (University of Tsukuba) for their valuable discussions.

References

[1] B. S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *the Journal of the Acoustical Society of America*, 55(6):1304–1312, 1974.

[2] A. L. Berenzweig and D. P. W. Ellis. Locating singing voice segments within music signals. In *Proceedings IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001.

[3] S. B. Davis and P. Mermelstein. Comparison of parametric representation for monosyllabic word recognition. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 28(4):357–366, 1980.

[4] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. Singer identification based on accompaniment sound reduction and reliable frame selection. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pages 329–336, 2005.

[5] M. Goto. A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43(4):311–329, 2004.

[6] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC Music Database: Popular, classical, and jazz music databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pages 287–288, October 2002.

[7] T. Hosoya, M. Suzuki, A. Ito, and S. Makino. Lyrics recognition from a singing voice based on finite state automaton form music information retrieval. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pages 532–535, 2005.

[8] C. kai Wang, R.-Y. Lyu, and Y.-C. Chiang. An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech2003)*, pages 1197–1200, 2003.

[9] T. Kawahara, A. Lee, K. Takeda, and K. Shikano. Recent progress of open-source LVCSR engine julius and japanese model repository – software of continuous speech recognition consortium –. In *Proceedings of The Sixth International Conference on Spoken Language Processing (Interspeech2004 ICSLP)*, 2004.

[10] B. Logan. Mel frequency cepstral coefficients for music modelling. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR 2000)*, pages 23–25, 2000.

[11] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara. Japanese morphological analysis system ChaSen. http://chasen.naist.jp/, 2000.

[12] J. A. Moorer. Signal procesing aspects of computer music: A survey. *Proceedings of the IEEE*, 65(8):1108–1137, 1977.

[13] T. L. Nwe and Y. Wang. Automatic detection of vocal segments in popular songs. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, pages 138–145, 2004.

[14] Y. Ohishi, M. Goto, K. Itou, and K. Takeda. Discrimination between singing and speaking voices. In *Proceedings of 9th European Conference on Speech Communication and Technology (Eurospeech 2005)*, pages 1141–1144, 2005.

[15] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transaction on System, Man, and Cybernetics*, SMC-9(1):62–66, 1979.

[16] 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 the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, pages I–237–240, 2005.

[17] The Hidden Markov Model Toolkit (HTK), http://htk.eng.cam.ac.uk/.

[18] W.-H. Tsai and H.-M. Wang. Automatic detection and tracking of target singer in multi-singer music recordings. In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, pages 221–224, 2004.

[19] Y. Wang, M.-Y. Kan, T. L. Nwe, A. Shenoy, and J. Yin. Lyrically: Automatic synchronization of acoustic musical signals and textual lyrics. In *Proceedings of the 12th ACM International Conference on Multimedia*, pages 212–219, 2004.