

THREE TECHNIQUES FOR IMPROVING AUTOMATIC SYNCHRONIZATION BETWEEN MUSIC AND LYRICS: FRICATIVE DETECTION, FILLER MODEL, AND NOVEL FEATURE VECTORS FOR VOCAL ACTIVITY DETECTION

Hiromasa Fujihara and Masataka Goto

National Institute of Advanced Industrial Science and Technology (AIST)

ABSTRACT

Three techniques are described that improve a previously developed system for automatically synchronizing lyrics with musical audio signals. Although this system achieves state-of-the-art accuracy by extracting vocal vowels from polyphonic sound mixtures and using forced alignment between those vowels and a phoneme network of the lyrics, there was still room for improvement. The first technique detects nonexistence regions in which fricative consonant sounds do not exist, which were not utilized in the previous system, and prohibits the alignment of the fricative phonemes to those regions. The second technique inserts a filler model between phrases of the phoneme network. This model improves the accuracy of the forced alignment by ignoring inter-phrase vowel utterances not included in the lyrics. The third technique introduces novel feature vectors for vocal activity detection that enable a distance calculation between two sets of the harmonic structure without estimating their spectral envelopes. Experimental results showed that all three techniques contribute to improved synchronization.

Index Terms— Music, Lyrics, Fricative sounds, Filler model, Spectral representation

1. INTRODUCTION

Automatic synchronization of lyrics with music compact-disc recordings has many applications, such as an automatic subtitle generation system for music videos and a music playback interface that enables a user to directly access specific words or phrases of interest. However, it has been difficult to achieve such synchronization because singing voices are usually accompanied by many other instruments. Wang *et al.* developed a system called LyricAlly [1] for synchronizing lyrics with music recordings without extracting singing voices from polyphonic sound mixtures. It basically uses the duration of each phoneme as a cue for synchronization, but it is not always effective because the duration of uttered phonemes differs with their location in the lyrics. Wong *et al.* [2] developed an automatic synchronization system for Cantonese popular music. It uses the tonal characteristics of Cantonese language and compares the tone of each word in the lyrics with the fundamental frequency (F0) of the singing voice. Since most languages do not have this tonal feature, this system cannot be generalized to other languages. Loscos *et al.* [3] and Wang *et al.* [4] used a speech recognizer for aligning and recognizing singing voice, respectively, but they presumed pure monophonic singing without accompaniment. Gruhne *et al.* [5] worked on phoneme recognition from polyphonic music. Assuming that boundaries between phonemes were given, they compared several classification techniques. Their experiments were preliminary, and there were difficulties in actually recognizing the lyrics.

We previously developed a system for automatically synchronizing lyrics with the corresponding singing voice (vocal) extracted

from polyphonic sound mixtures [6]. To locate the start and end times of each phrase in the lyrics, our system first segregates the most predominant sounds including vocal vowels from polyphonic sound mixtures on the basis of their harmonic structure (*accompaniment sound reduction*) and discriminates vocal sections from non-vocal sections (*vocal activity detection*). It then adapts speech-recognizer phone models to the segregated vocal (*phone model adaptation*) and uses the forced (Viterbi) alignment to align each vowel of the lyrics with the segregated vocal ignoring the consonants. Our experimental results showed that our system was effective, but that its accuracy could be improved by resolving three issues: 1) Consonants (especially unvoiced ones) were not accurately aligned, 2) Utterances not in the actual lyrics, such as shouting and humming, were judged to be vocal sections, and 3) the vocal activity detection was not always accurate enough when the F0s of the segregated vocal were high.

In this paper, we propose three techniques for resolving these issues and thereby improve the accuracy of our system. The first technique, *fricative detection*, detects *nonexistence* regions in which fricative consonant sounds do not exist and prohibits the alignment of the fricative phonemes to those regions. The novelty of this technique is the detection of *nonexistence* regions rather than *existence* regions. Even if the existence regions (i.e., unvoiced fricative sounds themselves) cannot be accurately detected, it is relatively easy to detect the nonexistence ones (i.e., regions without fricative sounds). The second technique, *introduction of a filler model*, forces the system to ignore any vowels between phrases as these inter-phrase vowel utterances are not actually in the lyrics. Our system sometimes erroneously aligned the lyrics to such utterances [6]. Such errors are reduced by inserting the filler model that matches any vowel sequence at each phrase boundary in the lyrics. Finally, the third technique, *novel feature vectors for vocal activity detection*, uses both the F0 of the vocal's harmonic structure and the power of each harmonic component as feature vectors for the vocal activity detection. Conventional spectral features such as a cepstrum and a linear prediction coefficient (LPC) were used to represent the spectral envelope, which was not always accurately estimated especially for high-pitched sounds. Using our novel features makes it possible to directly calculate the distance between two sets of the harmonic structure without estimating the spectral envelope, so the features are robust to high-pitched sounds.

2. OVERVIEW OF PREVIOUS SYSTEM

Our system for automatically synchronizing music and lyrics [6] is based on the use of the forced alignment (Viterbi alignment), which is often used in automatic speech recognition (ASR). Since the studies of ASR mainly deal with clean speech signals, its application to a singing voice in polyphonic sound mixtures requires the use of the following three steps: (1) segregate candidate vocal (singing voice) sections from the polyphonic mixtures (*accompaniment sound reduction*), (2) identify vocal sections (*vocal activity detection*), and (3) adapt phone models of speech recognizers to the segregated vocal signals.

This work was supported by CrestMuse, CREST, JST.

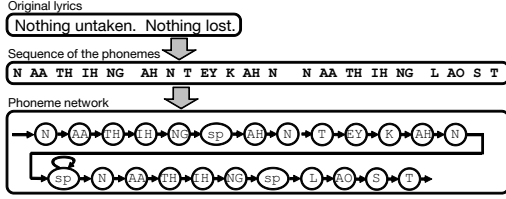


Fig. 1. Example of lyrics processing. Note that we used Japanese songs in our experiments described in Sec. 4, though this figure shows an English example.

2.1. Accompaniment sound reduction

First, the system reduces accompaniment sounds from given audio signals by resynthesizing the vocal signals from the harmonic structure of the melody line (the most predominant F0 in each frame [7]) by performing the following three steps: (1) use *PreFEst* method [7] to estimate the predominant F0 of the melody line (vocal candidate) in input audio signal, (2) extract the harmonic structure corresponding to the estimated F0, and (3) use sinusoidal synthesis of the harmonics to resynthesize audio signal (waveform) corresponding to the melody line.

2.2. Vocal activity detection

The non-vocal sections are then removed since the resynthesized signal corresponding to the melody line often contains instrumental (i.e., non-vocal) sounds in interlude sections. We introduced a hidden Markov model (HMM) that transitioned back and forth between two states, vocal state s_V and non-vocal state s_N . Given feature vectors of the segregated melody line, we calculate the most likely sequence of vocal and non-vocal states. The output probability of each state is approximated with likelihoods of a vocal and a non-vocal Gaussian mixture model (GMM). As feature vectors of these GMMs, we use LPMCCs and $\Delta F0$ s in our previous work [6]. LPMCCs are spectral features that represent spectral envelopes based on the LPC and cepstral analysis. In this paper, however, we propose alternative better feature vectors, which will be described in 3.3.

2.3. Phone model adaptation and forced alignment

Next, the lyrics are aligned with the segregated signals. The lyrics corresponding to the input audio signals are converted into a phoneme network for use in the forced alignment (Figure 1)¹. Each phrase boundary was converted into short pauses (SPs). Note that, although SPs generally represent short silent pauses between words or phrases in automatic speech recognition, we use them for non-vocal sections. Before the forced alignment is executed, each phone model (HMM) is adapted to singing voices in the input audio signals by using the MLLR and MAP adaptation techniques. Finally, the forced alignment is executed using the phoneme network created from the given lyrics, feature vectors (MFCCs, Δ MFCCs, and Δ power) extracted from the segregated vocal signals, and the adapted phone models.

3. NEW TECHNIQUES FOR IMPROVING AUTOMATIC SYNCHRONIZATION BETWEEN MUSIC AND LYRICS

In the following sections, we propose three new techniques to overcome the following three weaknesses our previous system had:

(1) Inaccurate alignment of consonants

Because the accompaniment sound reduction, which is based

¹The phoneme network in this figure includes consonant phonemes used in the proposed system, while the phoneme network in our previous system [6] did not include them.

on the harmonic structure, cannot segregate unvoiced consonants, only the HMMs for vowels were used as a cue for aligning the lyrics and music. Since the HMMs for vowels also covered various consonant sounds, it was difficult to detect the start time of a consonant. This problem can be overcome by incorporating consonants based on fricative detection.

(2) Utterances not in lyrics

Some singers often sing words, such as “Yeah” and “La La La”, not in the actual lyrics during interlude sections and rests between phrases. Such inter-phrase vowel utterances reduced the accuracy of the system because the system inevitably aligned other parts of the lyrics to those utterances. This shortcoming can be eliminated by introducing the filler model.

(3) Inaccurate vocal activity detection

When the vocal activity detection did not work well, the system sometimes did not align lyrics to the vocal regions correctly. We found this happened more frequently with female singers because of the difficulty in estimating the spectrum envelope for high-pitched sounds. This weakness can be overcome by introducing novel feature vectors based on the power of each harmonic component.

3.1. Use of consonants based on fricative detection

The simplest approach to incorporating the consonants is to make the phoneme network for the forced alignment include consonant phonemes as shown in Figure 1. However, since the accompaniment sound reduction based on the harmonic structure cannot segregate unvoiced consonants that do not have the harmonic structure, unvoiced consonants are not aligned correctly in general. We therefore develop a signal processing technique of detecting candidates of fricative sounds (a kind of unvoiced consonants) directly in the input audio signals. Here, we focus on only the fricative sounds because their durations are generally longer than the other unvoiced consonants and because they expose salient frequency components in the spectrum. Note that we do not try to use the detected location of each fricative sound in aligning it but instead use regions in which fricative sounds do not exist.

3.1.1. Nonexistence region detection

It is difficult to accurately detect the existence of each fricative sound because the acoustic characteristics of cymbals and snare drums, for example, sometimes resemble those of fricative sounds. We therefore take the opposite approach and try to detect regions in which there are no fricative sounds, i.e., *nonexistence* regions. Then, in the forced alignment, fricative consonants are prohibited from appearing in the nonexistence regions. Since it is relatively easy to detect the nonexistence regions of fricative sounds, detection errors negligibly degrade the accuracy of the forced alignment. In contrast, in the conventional *existence* detection approach, detection errors can significantly degrade the accuracy.

3.1.2. Fricative sound detection

Figure 2 shows an example spectrogram depicting non-periodic source components such as snare drum, fricative, and high-hat cymbal sounds in popular music. The characteristics of those non-periodic source components are depicted as vertical lines or clouds along the frequency axis in the spectrogram, while other periodic source components tend to have horizontal lines. In the frequency spectrum at certain time, those vertical and horizontal lines correspond to flat and peak (pointed) components, respectively.

To detect flat components from non-periodic sources, we need to ignore peak components in the spectrum. We therefore use the bottom envelope estimation method proposed by Kameoka *et al* [8].

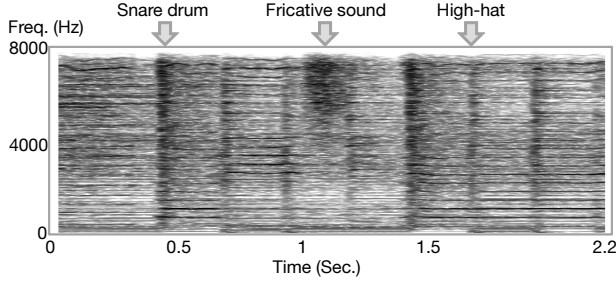


Fig. 2. Example spectrogram depicting snare drum, fricative, and high-hat cymbal sounds.

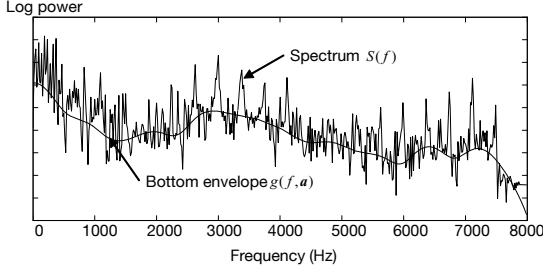


Fig. 3. A bottom envelope $g(f, \mathbf{a})$ in a spectrum $S(f)$.

The bottom envelope is defined as the envelope curve that passes through spectral valleys as shown in Figure 3. The function class of the bottom envelope is defined as

$$g(f, \mathbf{a}) = \sum_{i=1}^I a_i \mathcal{N}(f; 400 \times i, 200^2), \quad (1)$$

where f denotes the frequency in Hz, $\mathcal{N}(x; m, \sigma^2)$ is the Gaussian distribution, and $\mathbf{a} = (a_1, \dots, a_I)$ represents the weights of each Gaussian. The problem here is to estimate \mathbf{a} , which determines the envelope curve. We therefore estimate the $\hat{\mathbf{a}}$ that minimizes the objective function,

$$J = \int \left(\frac{g(f; \mathbf{a})}{S(f)} - \log \frac{g(f; \mathbf{a})}{S(f)} \right) df, \quad (2)$$

where $S(f)$ represents the spectrum at each frame. This objective function is an asymmetric distance measure that penalizes negative errors much more than positive ones. From this objective function, we can derive the following iterative equations to obtain $\hat{\mathbf{a}}$:

$$\hat{a}_i = \frac{\int m_i(f) df}{\int \frac{\mathcal{N}(f; 400 \times i, 200)}{S(f)} df}, \quad (3)$$

$$m_i(f) = \frac{a'_i \mathcal{N}(f; 400 \times i, 200)}{\sum_{\forall i} a'_i \mathcal{N}(f; 400 \times i, 200)}, \quad (4)$$

where a'_i is the estimated value at the previous iteration. In this way, the bottom envelope of the spectrum $S(f)$ is obtained as $g(f, \hat{\mathbf{a}})$.

Among various non-periodic source components, fricative sounds tend to have frequency components concentrated in a particular frequency band of the spectrum. We therefore detect the fricative sounds by using the ratio of the power of that band to that of most other bands. Since the sampling rate in our current implementation is 16kHz, we deal with only the fricative phoneme /SH/ because the other fricative phonemes have main concentrated frequency components above 8kHz, which is the Nyquist frequency of 16kHz sampling. Since the phoneme /SH/ has strong concentrated components from 6kHz to 8kHz, we define the existence degree of phoneme /SH/ as

$$E_{SH} = \frac{\int_{6000}^{8000} g(f, \hat{\mathbf{a}}) df}{\int_{1000}^{8000} g(f, \hat{\mathbf{a}}) df}. \quad (5)$$

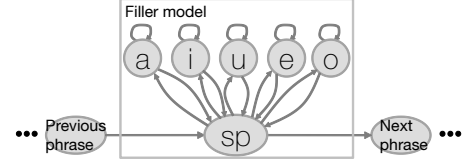


Fig. 4. Filler model inserted at each phrase boundary in the lyrics.

Regions in which E_{SH} is below a threshold (0.4) are identified as *nonexistence* regions in which phoneme /SH/ does not exist. The threshold 0.4 was determined experimentally. Note that we do not use frequency components below 1kHz in the calculation of E_{SH} to avoid any effect from bass drums.

3.2. Filler model

We introduce the filler model to remove errors due to singer's utterances not written in the actual lyrics. As shown in Figure 4, the filler model allows multiple appearances of any vowel between two consecutive phrases. In our previous system, we expected the SPs to represent short non-vocal sections. However, if the singer sung words not in the lyrics in non-vocal sections, the SPs, which were originally trained using non-vocal sections, were not able to represent them. The previous system then incorrectly allocated lyrics from other parts to such non-vocal sections. The vowels from the filler model can cover those inter-phrase utterances.

3.3. Novel feature vectors for vocal activity detection

The vocal activity detection after accompaniment sound reduction can be interpreted as the problem of judging whether the sound source of the given harmonic structure is vocal or non-vocal. In our previous system, we estimated the spectral envelope of the harmonic structure and calculated the distance against spectral envelopes in the training database. However, spectral envelopes estimated from high-pitched sounds by using cepstrum or LPC analysis are strongly affected by spectral valleys between adjacent harmonic components. Thus, there are some songs (especially those sung by female singers) for which the vocal activity detection method did not work well.

This problem boils down to the fact that a spectral envelope estimated from a harmonic structure is not reliable except for the points (peaks) around each harmonic component. This is because a harmonic structure could correspond to different spectral envelopes: the mapping from a harmonic structure to its original spectral envelopes is one-to-many association. When we consider this issue by using the sampling theory, the harmonic components are points sampled from its original spectral envelope at the interval of F0 along the frequency axis. The perfect reconstruction of the spectral envelope from the harmonic components is therefore difficult in general. Because conventional methods, such as MFCC and LPC, estimate only one possible spectral envelope, the distance between two sets of the harmonic structure from the same spectral envelope is sometimes not accurate. To overcome this problem, the distance must be calculated using only the reliable (sampled) points at the harmonic components.

We focus on the fact that we can directly compare the power of harmonic components between two sets of the harmonic structure if their F0s are about the same. Our approach is to use the power of harmonic components directly as feature vectors and compare the given harmonic structure with only those in the database that have similar F0 values. This approach is robust against high-pitched sounds, because the spectral envelope does not need to be estimated.

To ensure that comparisons are done only with feature vectors that have similar F0s, we also use the F0 value as a feature in addition to the power of harmonic components. By modeling the feature vectors using GMM, each Gaussian can cover feature vectors that have similar F0s. When we calculate the likelihood of a GMM, the weights of the Gaussians that have large F0 values are minuscule.

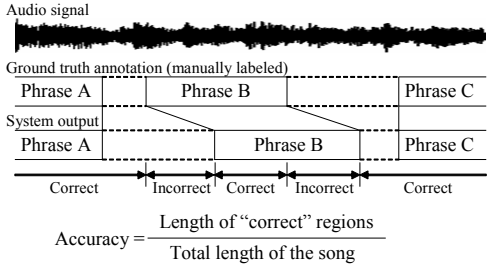


Fig. 5. Evaluation measure.

Thus, we can calculate the distance only with harmonic structures that have similar F0 values.

The absolute value of the power of the harmonic structure is biased depending on the volume of each song. We therefore normalize the power of all harmonic components for each song. The process for normalizing the power is as follows. The normalized power of the h th harmonic component at time t , p_h^t , is expressed as

$$p_h^t = \log p_h^t - \frac{\sum_t \sum_h \log p_h^t}{T \times H}, \quad (6)$$

where p_h^t represents the original power, T is the total number of frames, and H is the number of harmonic components considered.

4. EVALUATION

4.1. Conditions

As an evaluation data set, we used 10 Japanese songs by 10 singers (5 male, 5 female) taken from “RWC Music Database: Popular Music” (RWC-MDB-P-2001) [9]. These songs were generally sung in Japanese, but some English phrases in their lyrics were sung in English. In this experiments, we approximated English phonemes by using similar Japanese phonemes. As the training data for the vocal/non-vocal GMMs for vocal activity detection, we used 19 songs, which were also taken from RWC-MDB-P-2001, by 11 singers who did not sing any song in the evaluation data set. We conducted a five-fold cross-validation.

The evaluation was done using phrase level alignment. We defined a phrase as a section delimited by a space or line feed in the text of the original lyrics. The calculated evaluation measure was the total length of the sections that are correctly labeled at the phrase level to the total length of the song (Figure 5).

We tested our method under five conditions.

- (i) **Baseline:** Previous (unimproved) system used ([6]).
- (ii) **Fricative detection:** Fricative sound detection enabled (3.1).
- (iii) **Filler model:** Filler model enabled (3.2).
- (iv) **Novel feature vector for VAD:** Novel features for vocal activity detection enabled (3.3).
- (v) **Proposed:** All three techniques enabled (3).

4.2. Results and discussion

The results are summarized in Table 1.² With our new techniques (ii), (iii), and (iv) in Table 1), the average accuracy increased by 2.0, 3.3, and 3.7 point, respectively. With all three techniques ((v) in Table 1), the highest accuracy was achieved (85.3%). Of the three techniques, the new feature vectors for vocal activity detection was the most effective. Inspection of the system outputs with the filler model showed that the filler model appeared not only for utterances not in the actual lyrics, but also for non-vocal regions that could not be removed by vocal activity detection. Since our evaluation measure was phrase-based, the effectiveness of fricative detection could

²In our previous experiments [6], we did not use songs with female singers as training data for male singers’ songs, and vice versa. Therefore, some of the accuracies of the baseline were lower than the results in [6]

Table 1. Experimental results (%). Note that M. and F. means male and female, respectively.

Song #*	Gender	(i)	(ii)	(iii)	(iv) VAD	(v)
		Baseline	Fric.	Filler	Feature	Proposed
No. 12	M.	95.7	95.1	96.3	97.8	95.7
No. 27	M.	87.4	87.6	86.3	90.2	91.2
No. 32	M.	66.4	69.6	70.2	81.3	71.7
No. 37	M.	83.7	85.9	89.5	89.5	89.5
No. 39	M.	93.6	93.2	92.4	93.9	93.3
No. 7	F.	62.8	62.5	67.4	79.9	70.0
No. 13	F.	63.6	70.4	67.2	46.0	68.0
No. 20	F.	93.3	93.3	93.1	92.7	94.0
No. 65	F.	73.7	85.4	91.6	91.2	92.0
No. 75	F.	90.6	88.2	90.3	85.9	87.8
Average		81.1	83.1	84.4	84.8	85.3

*A song number of RWC-MDB-P-2001[9].

not be fully evaluated. Inspection of the phoneme-level alignment results showed that phoneme gaps in the middle of phrases were shorter than without fricative detection. We plan to develop a measure for evaluating phoneme-level alignment.

5. CONCLUSION

We have developed three techniques for improving the automatic synchronization of music and lyrics: fricative detection, filler model, novel feature vectors for vocal activity detection. These three techniques are versatile because they do not depend on a specific language or music structure.

The underlying idea of the fricative detection, i.e., the detection of *nonexistence* regions is novel. Experimental evaluation showed that performance is improved by integrating this information, even if it is difficult to detect each fricative sound accurately. Though the filler model is a simple idea, it worked very efficiently. This is because it does not allow a phoneme in the lyrics to be skipped and it appears only when it is needed. The novel feature vectors based on the F0 and the power of harmonic components are robust to high-pitched sounds because a spectral envelope does not need to be estimated. Though we use these vectors just for vocal activity detection, they can also be used as feature vectors for the forced alignment after preparing enough training data. Furthermore, we can use them for studies related to speech signals, such as automatic speech recognition in a noisy environment.

We plan to use other fricative phonemes in addition to the phoneme /SH/. Future plans also include conducting an evaluation using English songs with English phone models, and conducting a larger scale evaluation.

6. REFERENCES

- [1] Y. Wang *et al.*, “Lyrically: Automatic synchronization of acoustic musical signals and textual lyrics,” in *Proc. ACM Multimedia 2004*, pp. 212–219, 2004.
- [2] C. H. Wong *et al.*, “Automatic lyrics alignment for Cantonese popular music,” *Multimedia Systems*, vols. 4-5, no. 12, pp. 307–323, 2007.
- [3] A. Loscos *et al.*, “Low-delay singing voice alignment to text,” *Proc. ICMC 1999*, 1999.
- [4] C. Wang *et al.*, “An automatic singing transcription system with multilingual singing lyrics recognizer and robust melody tracker,” *Proc. Eurospeech 2003*, pp. 1197–1200, 2003.
- [5] M. Gruhne *et al.*, “Phoneme recognition in popular music,” *Proc. ISMIR 2007*, pp. 369–370, 2007.
- [6] H. Fujihara *et al.*, “Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals,” in *Proc. ISM 2006*, pp. 257–264, 2006.
- [7] M. Goto, “A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals,” *Spe. Comm.*, vol. 43, no. 4, pp. 311–329, 2004.
- [8] H. Kameoka *et al.*, “Selective amplifier of periodic and non-periodic components in concurrent audio signals with spectral control envelopes,” in *IPSSJ SIG Technical Report*, 2006-MUS-66-13, pp. 77–84, 2006 (in Japanese).
- [9] M. Goto *et al.*, “RWC Music Database: Popular, classical, and jazz music databases,” in *Proc. ISMIR 2002*, pp. 287–288, 2002.