Lyrics-to-Audio Alignment and its Application

Hiromasa Fujihara and Masataka Goto

National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
{h.fujihara, m.goto}@aist.go.jp

Abstract

Automatic lyrics-to-audio alignment techniques have been drawing attention in the last years and various studies have been made in this field. The objective of lyrics-to-audio alignment is to estimate a temporal relationship between lyrics and musical audio signals and can be applied to various applications such as Karaoke-style lyrics display. In this contribution, we provide an overview of recent development in this research topic, where we put a particular focus on categorization of various methods and on applications.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, H.5.1 Multimedia Information Systems

Keywords and phrases Lyrics, Alignment, Karaoke, Multifunctional music player, Lyrics-based music retrieval

Digital Object Identifier 10.4230/DFU.Vol3.11041.23

1 Introduction

Music is an important media content in both industrial and cultural aspects, and a singing voice (vocal) is one of the most important elements of music in many music genres, especially in popular music. Thus, research that deals with singing voices is gaining in importance from cultural, industrial and academic perspectives. Lyrics are one of the most important aspects of singing voices. Since the lyrics of a song represent its theme and story, they are essential for creating an impression of the song. When a song is heard, for example, most people would follow the lyrics while listening to the vocal melody. This is why music videos often help people to enjoy music by displaying synchronized lyrics as a Karaoke-style caption.

In this paper we overview several research attempts that deal with automatic synchronization between music and lyrics, also known as lyrics-to-audio alignment. To deal with lyrics in music, one of the ultimate goals is automatic lyric recognition (i.e., the dictation of lyrics in a mixture of singing voices and accompaniments). However, since this goal has not yet been achieved even for ordinary speech in noisy environments with satisfactory accuracy, it is not a realistic way to pursue automatic dictation of the lyrics in the first place. As a matter of fact, though several research attempts have been made to pursue this goal [23, 19, 21, 5, 14], none of them achieved satisfactory performance in natural environments under realistic assumptions so far. From this perspective, it can be said that lyrics-to-audio alignment is a reasonable problem setting because not only does the problem itself have a number of practical applications but knowledge accumulated by tackling this problem can also be a stepping-stone for automatic lyric recognition.

The rest of this paper is organized as follows. We continue with defining the problem of lyrics-to-audio alignment and describing main challenges of this task. Then, Section 3 summarizes numerous study attempts and introduces some representative works. In Section 4, we introduce applications of lyrics-to-audio alignment techniques.


© Hiromasa Fujihara and Masataka Goto;
licensed under Creative Commons License CC-BY-ND
Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 23–36

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

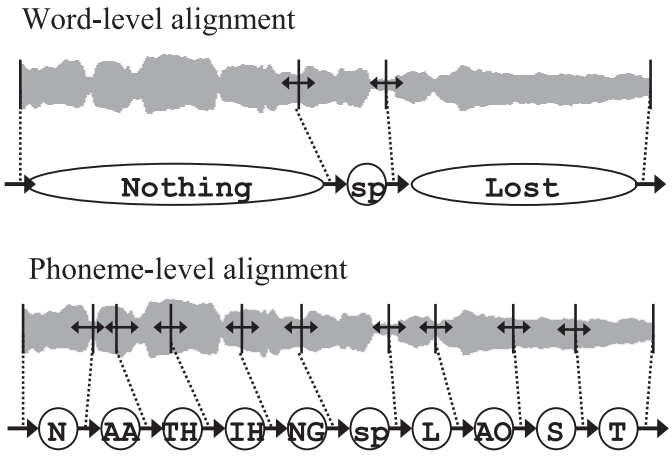

Figure 1 Example of word-level alignment and phoneme-level alignment.

2 Lyrics-to-Audio Alignment

2.1 Problem Definition and Applications

Given audio signals of singing voices and corresponding textual lyrics as input data, lyrics-to-audio alignment can be defined as a problem of estimating the temporal relationship between them. To this end, start and end times of every block of certain length in lyrics are estimated. Here, the term "block" means a fragment of lyrics, the size of which depends on the application as described below, and can be either phoneme, syllable, word, phrase, line, or paragraph (See Figure 1).

Numerous applications of this technique are conceivable, such as a music player with Karaoke-like lyrics display function, automatic generation of subtitles of music videos, and the generation of audio thumbnails. Apart from these consumer-oriented applications, this technique can also be used as a basic building block for other singing voice research, such as singing voice synthesis [16] and analysis of the relationship between musical audio signals and lyrics [17]. In the case of music video subtitles, granularity of synchronization does not have to be very precise and line or phrase level alignment is sufficient. If the precise timing of the lyrics is needed such as in the case of Karaoke-like display, on the other hand, phoneme or word level alignment is imperative.

2.2 Difficulties

The problem of lyrics-to-audio alignment bears a relationship to text-to-speech alignment used in automatic speech recognition research, which is generally conducted by using a forced alignment techniques with mel-frequency cepstral coefficients (MFCCs), phonetic hidden Markov models (HMMs), and the Viterbi algorithm[1] [25]. However, it is difficult to directly apply the forced alignment technique to singing voices because there are several difficulties

[1] Hereafter, we use a term "forced alignment" to refer to this particular technique that can align transcribed text and speech signals by using phonetic HMMs and the Viterbi algorithm.

intrinsic to singing voices:

1. **Fluctuation of acoustic characteristics.** It is known that the singing voice has more complicated frequency and dynamic characteristics than speech [20]. For example, fluctuation of fundamental frequency (F0) [2] and loudness of singing voices are far stronger than those of speech sounds.
2. **Influences of accompaniment sounds.** Singing voice signals are generally accompanied by other instruments, which make it difficult even for a human to understand what is being sung. This is mainly because spectrum of the singing voices are overlapped and distorted by those of accompaniment sounds. Thus, it is necessary to either reduce such negative influences or use features robust to them.
3. **Incomplete lyrics.** Available textual lyrics do not always correspond exactly to what is sung in a song. For example, repetitive paragraphs are sometimes omitted for the sake of simplicity and utterances of interjections (such as "yeah" and "oh") are often excluded in the lyrics. This is particularly problematic when lyrics are taken from the Internet [8].

3 Literature Review

After overviewing studies of this field, this section describes brief explanations of representative works.

3.1 Overview of Previous Studies

A number of studies have been made in the field of lyrics-to-audio alignment [10, 6, 24, 15, 9, 13, 7, 3, 12]. Except for early research [10], most of the studies dealt with singing voices in polyphonic popular music. Since lyrics are inevitably language-dependent, it is not easy to prepare training data for a number of several languages. Thus, evaluations were usually conducted by using songs sung in a single language such as English [6, 7], Chinese [24], and Japanese [3]. With that being said, except for a study that specialized in Cantonese [24], most of them are applicable to any language in principle.

These studies can be categorized according to the following two main viewpoints:

1. **Primary cue for aligning music and lyrics.** To achieve an accurate estimation of a temporal relationship between music and lyrics, it is important to deliberately design features (or representation) that are used to represent music and lyrics and methods to compare these features since such features and methods directly affect the performance of an entire system.
2. **Additional methods for improving performance.** Polyphonic audio signals are so complex that it is not easy to align music and lyrics accurately just by using a single method. Thus, many studies have integrated several additional methods and information to improve performance of alignment such as music understanding techniques and musical knowledge.

Table 1 summarizes the conventional studies from these viewpoints.

3.1.1 Primary Cue for Aligning Music and Lyrics

To characterize algorithms for lyrics-to-audio alignment, it is of central importance to categorize what kind of features they extract from audio and lyrics and how they compare

[2] "F0", which represents how high a sound is, is sometimes referred as "pitch" although, strictly speaking, their definitions are different because the F0 is a physical feature while the pitch is a perceptual feature.

Table 1 Summarization of the conventional studies.

Authors	Primary method	Other additional methods
Loscos *et al.* [10]	The forced alignment with MFCCs	
Iskandar *et al.* [6]	The forced alignment with MFCCs	Song structure Musical knowledge
Wong *et al.* [24]	Comparison of F0 contours	Vocal enhancement Vocal detection Onset detection
Müller *et al.* [15]	Audio-to-MIDI alignment with lyrics-enhanced MIDI files	
Lee *et al.* [9]	Dynamic programming with manually-labeled lyrics segmentation	Structural segmentation
Mesaros *et al.* [13]	The forced alignment	Vocal segregation
Kan *et al.* [7]	Phoneme duration	Beat detection Structure detection Vocal detection
Fujihara *et al.* [3]	The forced alignment	Vocal segregation Vocal detection Fricative detection
Mauch *et al.* [12]	The forced alignment with chord labels	Vocal segregation Vocal detection

them. From this viewpoint, conventional studies can be categorized into the following three categories; those that use acoustic phonetic features, those that use other features, and those that use features taken from external sources.

Studies that fall into the first category [10, 6, 13, 3] adopt the forced alignment. It compares phonetic features (such as MFCCs) extracted from audio signals with a phone model consisting of a sequence of phonemes in the lyrics. Since the forced alignment technique is mainly designed for clean speech signals, the main focus of these studies lies in how to apply it to singing voices with accompaniment sounds. For this purpose, most of the studies incorporate various additional methods described in Section 3.1.2.

The second category contains studies that do not use the forced alignment technique. Wong *et al.* [24] used the tonal characteristics of Cantonese language and compared the tone of each word in the lyrics with the F0 of the singing voice. Kan *et al.* developed a system called LyricAlly [7], which used the duration of each phoneme as a main cue for a fine alignment along with structural information for a coarse alignment.

Instead of directly comparing music and lyrics, studies of the third category deploy external information and use them as a cue for alignment. For example, Müller *et al.* [15] used MIDI files that are manually aligned with lyrics. Then, by executing automatic alignment between music recordings and the MIDI file, they indirectly estimated temporal relationship between music and lyrics. Lee *et al.* [9], assuming that manually-annotated segmentation labels (such as Chorus and Verse) are available, aligned these labels with automatically-estimated structural segmentation by using dynamic programming. While these strategies could result in simpler or more accurate alignments, the range of songs to which the algorithms are applicable is inevitably limited.

In addition to the above mentioned works, Mauch *et al.* [12] used both acoustic phonetic

features and external information at the same time. It can be said that this work belongs to both the first and third categories. More specifically, assuming that the textual chord information provided in the paired chords-lyrics format is available, they integrated lyrics-to-audio alignment and chord-to-audio alignment. Chord alignment, which is more reliable but can be done in only bar or note level, worked as a coarse alignment, followed by a fine alignment achieved by singing voice alignment.

3.1.2 Additional Methods for Improving Performance

In addition to the primary cues described above, most of the studies sought to improve their algorithm by incorporating other music understanding and signal processing methods. For example, some studies used vocal detection methods as a preprocessing step [24, 7, 3]. Regions detected as non-vocal are excluded from the allocation of lyrics. Methods for reducing the influence of accompaniment sounds and enhance singing voices were also used in [24, 13, 3]. This process is usually done before extracting features from audio signals to enable feature extractors to accurately capture the characteristics of singing voices. Fujihara *et al.* [3] and Mesaros *et al.* [13] used singing voice segregation techniques based on harmonic structures, and Wong *et al.* [24] used bass and drum reduction and center signal segregation methods. Other information such as beat [7], song structure [9, 7], onset [24], fricative sound [3], and musical knowledge about rhythms and notes [6] were automatically extracted and incorporated.

3.2 A Lyrics-to-Audio Alignment Method for Cantonese Popular Music

Wong *et al.* developed a lyrics-to-audio alignment method based on tonal characteristics of Cantonese popular music [24]. Cantonese, which is a tone language, distinguishes the meaning of a word by changing the pitch level. Their method took advantage of this fact and tried to align music and lyrics by using pitches extracted from audio signals and those estimated from lyrics, assuming that the contour of the lyrics and that of the musical melody match perfectly. In their method, a vocal signal enhancement algorithm based on center signal estimation and bass and drum reduction methods was used to detect the onsets of the syllables and to estimate the corresponding pitches. They then used a dynamic time warping algorithm to align lyrics and music. Figure 2 shows a block diagram of the method.

To estimate the onsets and the pitch accurately, the authors developed a vocal signal enhancement technique. Based on the assumption that only singing voice and drum signals are located at the center in stereo audio signals, they extracted center parts of the stereo recordings by using a spectral subtraction method. Bass and drum sounds were then removed by subtracting the average spectrum within five-second segments.

Then the onsets of vocal notes, which are expected to correspond to syllables, were detected as the smallest unit of alignment. They first extracted the amplitude envelope of the signal and detected candidates of the onsets by using a difference of the amplitude envelope. Finally they eliminated non-vocal onsets by using a neural network classifier with standard audio features such as spectrum flux, zero-crossing rate, and Mel-frequency cepstral coefficients (MFCCs).

As features for aligning music and lyrics, they used pitch contours. Pitch contours of audio signals were extracted by a standard F0 estimation method, and that of lyrics were estimated from the lyrics based on linguistic rules. These two types of pitch contours are aligned by using dynamic time warping.

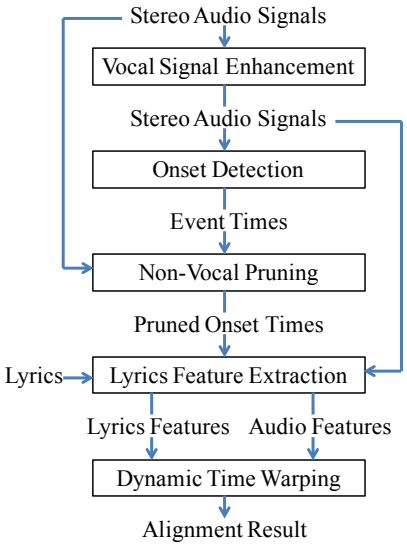

Figure 2 A block diagram of a lyrics-to-audio alignment method in [24].

3.3 LyricAlly

Kan *et al.* developed a lyrics-to-audio alignment system called LyricAlly [7]. It integrates several music understanding techniques such as beat detection, chorus detection, and vocal estimation. They first conducted section-level alignment, which was followed by line-level alignment. Figure 3 shows a block diagram of LyricAlly.

Three kinds of music understanding techniques, namely hierarchical rhythm structure detection, chorus detection, and vocal detection, were executed as a preprocessing step, to constrain and simplify the synchronization process based on musical knowledge. Input lyrics were then analyzed to estimate the duration of the lyrics. This duration estimation process was done based on supervised training.

Assuming a song is consisted of a specific type of song structure (Verse-Chorus-Verse-Chorus) and that each section of lyrics is already marked as a single block, they first conducted section level alignment based on the chorus section detected by using the chorus and vocal detectors. Then, they conducted line-level alignment by using duration information estimated from lyrics.

3.4 A Lyrics-to-Audio Alignment Method based on the forced Alignment

Fujihara *et al.* developed a lyrics-to-audio alignment method based on the forced alignment technique [3]. Because the ordinary forced alignment technique used in automatic speech recognition is negatively influenced by accompaniment sounds performed together with a vocal and also by interlude sections in which the vocal is not performed, they first obtained the waveform of the melody by using a vocal segregation method proposed in [2]. They then detected the vocal region in the separated melody's audio signal, using a vocal detection method based on a Hidden Markov Model (HMM). They also detected the fricative sound and incorporated this information into the next alignment stage. Finally, they aligned the lyrics and the separated vocal audio signals by using the forced alignment technique. Figure

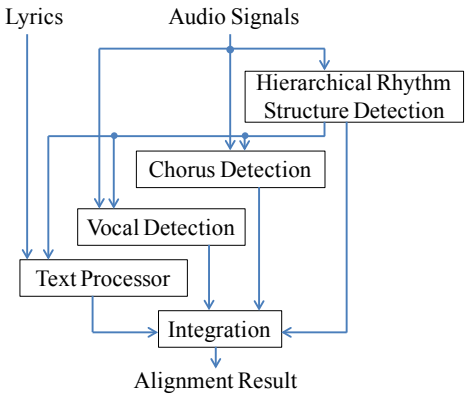

Figure 3 A block diagram of Lyrically [7].

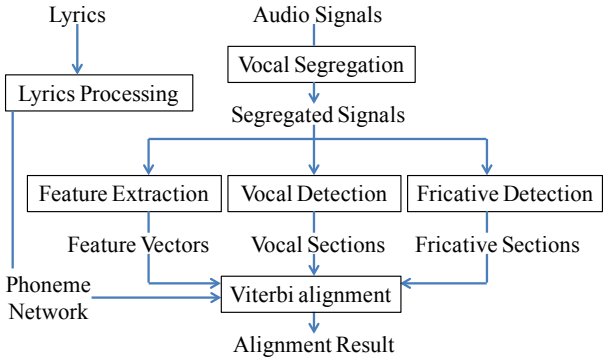

Figure 4 A block diagram of a lyrics-to-audio-alignment method proposed in [3].

4 shows an overview of this method.

Before extracting a feature that represents the phonetic information of a singing voice from polyphonic audio signals, they tried to segregate vocal sound from accompaniment sounds by using a melody detection and resynthesis technique based on a harmonic structure [2]. The technique consists of the following three parts:

1. Estimate the fundamental frequency (F0) of the melody by using a method called PreFEst [4].
2. Extract the harmonic structure corresponding to the melody.
3. Resynthesize the audio signal (waveform) corresponding to the melody by using a sinusoidal synthesis.

This melody resynthesis usually results in vocal signals with bad sound quality for a human perception and it makes it even more difficult for humans to recognize lyrics. However, for a computer, which does not have a sophisticated perceptive system that humans have, this process is important.

The authors developed a vocal detection method to eliminate the influence of non-vocal regions. The method is based on supervised-training of characteristics of singing voices and non-vocal sounds. This method is needed because the melody detection technique assumed that the F0 of the melody is the most predominant in each frame and could not detect

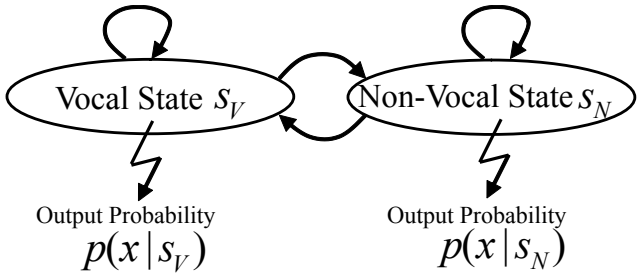

Figure 5 A hidden Markov model (HMM) for vocal activity detection [3].

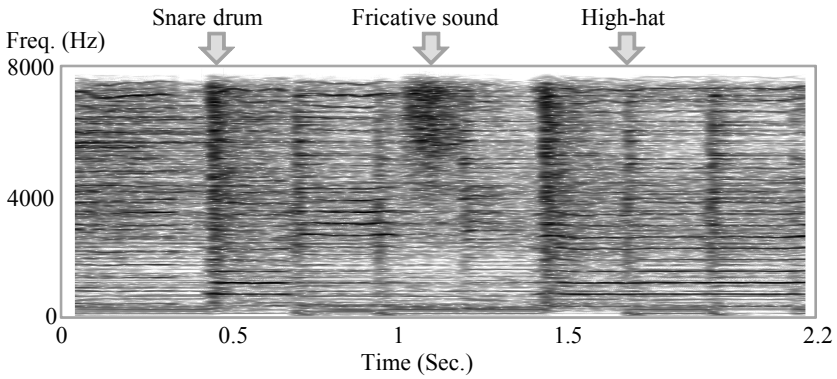

Figure 6 Example spectrogram depicting snare drum, fricative, and high-hat cymbal sounds [3]. The characteristics of fricative sounds are depicted as vertical lines or clouds along the frequency axis, whereas periodic source components tend to have horizontal lines.

regions where vocal does not exist. Thus, such regions have to be eliminated before actually aligning lyrics to segregated signals. An HMM was introduced that transitions back and forth between a vocal state, s_V, and a non-vocal state, s_N, as shown in Figure 5. Vocal state means that vocals are present and non-vocal state means that vocals are absent. Given the feature vectors of input audio signals, x_t, at time t, the problem of vocal detection is finding the most likely sequence of vocal and non-vocal states, $\hat{S} = \{s_1, \cdots, s_t, \cdots\}\,(s_t \in \{s_V, s_N\})$.

$$\hat{S} = \underset{S}{\operatorname{argmax}} \sum_t \{\log p(\boldsymbol{x_t}|s_t) + \log p(s_{t+1}|s_t)\}\,, \tag{1}$$

where $p(\boldsymbol{x}|s)$ represents an output probability of state s, and $p(s_{t+1}|s_t)$ represents a state transition probability for the transition from state s_t to state s_{t+1}. Unlike other previous studies on vocal detection [1, 22, 18], this methods could automatically control the balance between vocal and non-vocal regions.

The forced alignment technique used in automatic speech recognition research synchronizes speech signals and texts by making phoneme networks that consist of all the vowels and consonants. However, since the vocal segregation method, which is based on the harmonic structure of the melody, cannot segregate unvoiced consonants that do not have harmonic structure, it is difficult for the general forced alignment technique to align unvoiced consonants correctly. Therefore, the authors developed a method for detecting unvoiced consonants from the original audio signals. They particularly focused on the unvoiced fricative sounds

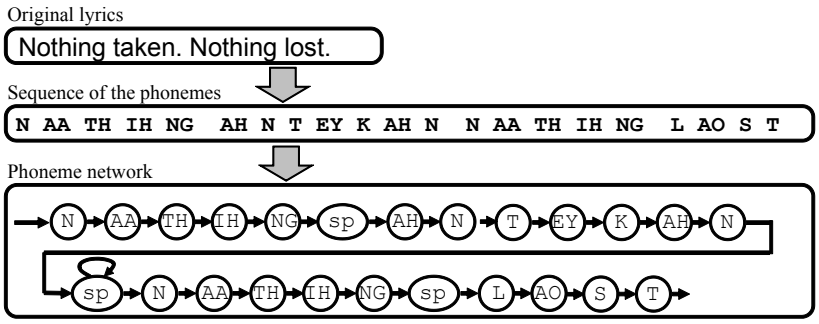

Figure 7 Example for converting from original lyrics to a phoneme network [3].

(a type of unvoiced consonant) because their durations are generally longer than those of the other unvoiced consonants and because they expose salient frequency components in the spectrum. They first suppressed peak components in the spectrum, which is not related to fricative sounds. The fricative sounds were then detected by using the ratio of the power of a band where salient frequency components of fricative sounds exist to that of the other bands. Figure 6 shows an example of a fricative sound to be detected. Then, in the forced alignment stage, fricative consonants were only allowed to appear in the detected candidates of fricative regions.

To actually align lyrics and music, a phoneme network was created from the given lyrics and feature vectors are extracted from separated vocal signals. Figure 7 shows an example of conversion from lyrics to a phoneme network. The phoneme network consists of sequentially connected HMMs of phonemes that appeared in the lyrics. Each HMM represents sound characteristic of a corresponding phoneme and is used to compare likelihood of feature vectors extracted from audio signals. Finally, the forced alignment was executed by calculating the most likely path of a sequence of the feature vectors and the phoneme network. The authors used the proportion of the length of the sections which are correctly labeled as a quality measure and reported that the system achieved 90% accuracy for 8 out of 10 songs.

4 Applications to Music Player and Lyrics-based Music Retrieval

Due to the diffusion of the personal computer and the portable music player, there has been a growing opportunity to listen to songs using devices that have screens. It is natural to consider using that screens to enrich users' experience in music appreciation by displaying lyrics and related information on it. This section introduces three examples of this idea, which display synchronized lyrics estimated by lyrics-to-audio alignment techniques and utilize it to lyrics-based music retrieval or music navigation.

4.1 Lyrics-based Music Retrieval

Müller *et al.* proposed a lyrics search engine based on their lyrics-to-audio alignment [15]. Their system was developed as a plug-in software on the SyncPlayer framework, which is a software framework that integrates various MIR-techniques. Figure 8 (taken from [15]) shows a screenshot of their system. The three windows on the left side display lyrics synchronously and the right window enables lyrics-based search. It should be noted that users can directly

Figure 8 Screenshot of lyric-based search system developed on the SyncPlayer framework [15].

jump to the corresponding matching positions within the audio recordings from the search results.

4.2 LyricSynchronizer

Fujihara *et al.* developed a music playback interface called LyricSynchronizer based on their algorithm for lyrics synchronization [3]. This music playback interface offers the following two functions: displaying synchronized lyrics, and jump-by-clicking-the-lyrics functions. The former function displays the current position of the lyrics as shown in Figure 9. Although this function resembles the lyrics display in Karaoke, manually labeled temporal information is required in it. Using the latter function, users can change the current playback position by clicking a phrase in the lyrics. This function is useful when users want to listen only to sections of interest.

4.3 SongPrompter

Mauch *et al.* developed a software system called SongPrompter [11] by utilizing their lyrics-and-chord-to-audio alignment method [12]. This system acts as a performance guide by showing lyrics, chords, beats and bar marks along with music playback. Unlike the previous two examples, this software is designed for music performer. Figure 10 shows a screenshot of the system. As can be seen in the figure, both lyrics and chords are shown in the horizontal scrolling bar so that players can play music and sing without memorizing lyrics and chords or turning pages.

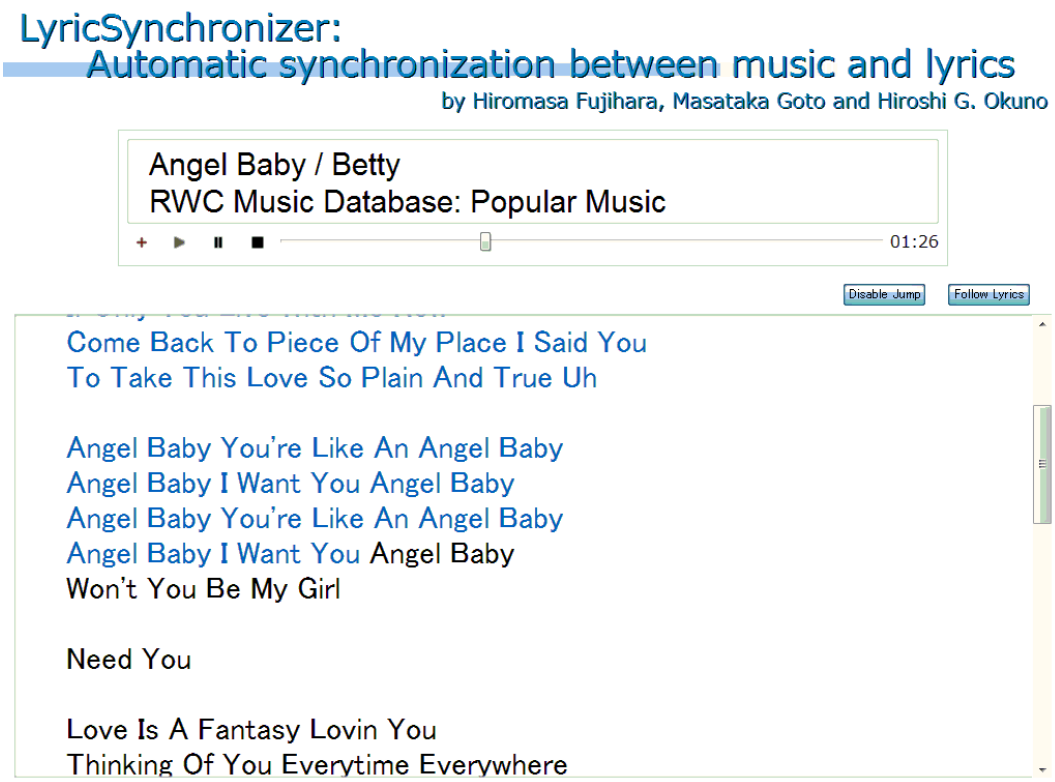

Figure 9 Screenshot of LyricSynchronizer [3].

5 Conclusions

In this paper, we described recent developments in lyrics-to-audio alignment research. We first defined the problem of lyrics-to-audio alignment and then gave an overview of current work. Finally, we introduced several applications of lyrics-to-audio alignment techniques.

Thanks to the advancements of this research fields, it is possible to align lyrics and audio with satisfactory accuracy for songs in which vocals pronounce words clearly and the sounds of vocals are mixed louder. On the other hand, there are still songs of which it is not easy to estimate the correct alignments. As mentioned in Section 3.1, most of lyrics-to-audio alignment techniques have sought to improve their performance by integrating various signal processing and music understanding techniques. This is because singing voices are highly correlated with other elements in music (e.g. melody F0s and chords) and, thus, the understandings of such elements can help aligning lyrics and singing voices.

To advance this field further, we think that the following three approaches can be conceivable:

Integrating of other signal processing and music understanding techniques. We believe that it is a promising direction to integrate fruits of a broader array of research field. For example, recent developments of source separation research can contribute to lyrics-to-audio alignment research. It is also possible to incorporate music classification methods such as genre detection and singer identification to select a model which is most suited for an input song.

A more sophisticated way of integrating information. As a way of integrating various in-

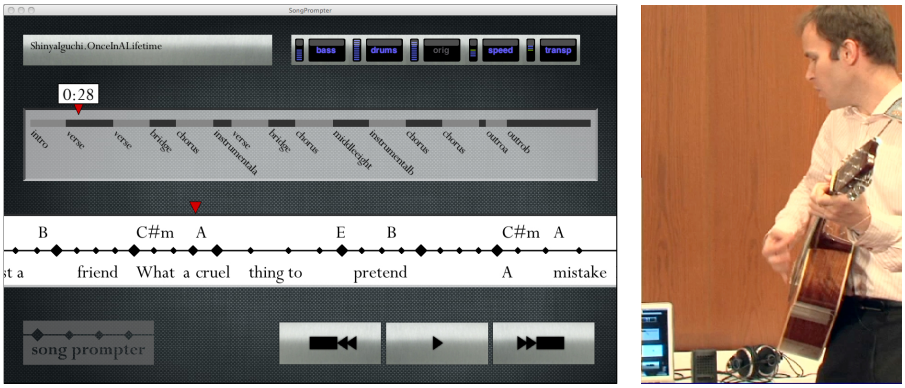

Figure 10 *SongPrompter* interface screenshot and usage example [11].

formation extracted by several music understanding techniques, most of the current studies took a straightforward approach: each music understanding technique was regarded as independent and only the results from different techniques were integrated. However, we believe it is possible to boost the performance by integrating the process of each music understanding technique so that each technique works in a mutually complementary manner.

Practically-oriented approach by utilizing external information. Finally, it is also interesting to incorporate external information available on the Web or other places. This approach, which narrows the range of applicable songs but can lead to interesting applications, was already taken by Müller *et al.* (lyrics-aligned MIDI) [15] and Mauch *et al.* (lyrics with chord annotation) [12] and resulted in the appealing applications as described in the previous section. We think that there could be other sources of information that are easy to obtain and can be beneficial to improve the performance.

References

1 Adam L. Berenzweig and Daniel P. W. Ellis. Locating singing voice segments within music signals. In *Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

2 Hiromasa Fujihara, Masataka Goto, Tetsuro Kitahara, and Hiroshi G. Okuno. A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity based music information retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 18:638–648, 2010.

3 Hiromasa Fujihara, Masataka Goto, Jun Ogata, and Hiroshi G. Okuno. LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics. *IEEE Journal of Selected Topics in Signal Processing*, 5:1252–1261, 2011.

4 Masataka 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.

5 Matthias Gruhne, Konstantin Schmidt, and Christian Dittmar. Phoneme recognition in popular music. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 369–370, 2007.

6 Denny Iskandar, Ye Wang, Min-Yen Kan, and Haizhou Li. Syllabic level automatic synchronization of music signals and text lyrics. In *Proceedings of ACM Multimedia*, pages 659–662, 2006.

7 Min-Yen Kan, Ye Wang, Denny Iskandar, Tin Lay Nwe, and Arun Shenoy. Lyrically: Automatic synchronization of textual lyrics to acoustic music signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):338–349, 2008.

8 Peter Knees, Markus Schedl, and Gerhard Widmer. Multiple lyrics alignment: Automatic retrieval of song lyrics. In *Proceedings of the 6th International Conference on Music Information Retrieval*, pages 564–569, 2005.

9 Kyogu Lee and Markus Cremer. Segmentation-based lyrics-audio alignment using dynamic programming. In *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 395–400, 2008.

10 Alex Loscos, Pedro Cano, and Jordi Bonada. Low-delay singing voice alignment to text. In *Proceedings of the International Computer Music Conference 1999 (ICMC99)*, 1999.

11 Matthias Mauch, Hiromasa Fujihara, and Masataka Goto. Song Prompter: An accompaniment system based on the automatic alignment of lyrics and chords to audio. In *Late-breaking session at the 10th International Conference on Music Information Retrieval*, 2010.

12 Matthias Mauch, Hiromasa Fujihara, and Masataka Goto. Integrating additional chord information into hmm-based lyrics-to-audio alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1), 2012.

13 Annamaria Mesaros and Tuomas Virtanen. Automatic alignment of music audio and lyrics. In *Proceedings of the 11th International Conference on Digital Audio Effects*, pages 1–4, 2008.

14 Annamaria Mesaros and Tuomas Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Advances in Signal Processing*, 2010, 2010.

15 Meinard Müller, Frank Kurth, David Damm, Christian Fremerey, and Michael Clausen. Lyrics-based audio retrieval and multimodal navigation in music collections. In *Proceedings of the 11th European Conference on Digital Libraries (ECDL 2007)*, 2007.

16 Tomoyasu Nakano and Masataka Goto. VocaListener: A singing-to-singing synthesis system based on iterative parameter estimation. In *Proceedings of the 6th Sound and Music Computing Conference*, pages 343–348, 2009.

17 Naoki Nishikawa, Katsutoshi Itoyama, Hiromasa Fujihara, Masataka Goto, Tetsuya Ogata, and Hiroshi G. Okuno. A musical mood trajectory estimation method using lyrics and acoustic features. In *Proceedings of the First International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies (MIRUM 2011)*, pages 51–56, 2011.

18 Tin Lay Nwe and Ye Wang. Automatic detection of vocal segments in popular songs. In *Proceedings of the 5rd International Conference on Music Information Retrieval*, pages 138–145, 2004.

19 Akira Sasou, Masataka Goto, Satoru Hayamizu, and Kazuyo Tanaka. An auto-regressive, non-stationary excited signal parameter estimation method and an evaluation of a singing-voice recognition. In *Proceedings of the 2005 International Conference on Acoustics, Speech, and Signal Processing*, pages I–237–240, 2005.

20 Johan Sundberg. *The Science of Singing Voice*. Northern Illinois University Press, 1987.

21 Motoyuki Suzuki, Toru Hosoya, Akinori Ito, and Shozo Makino. Music information retrieval from a singing voice using lyrics and melody information. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.

22 Wei-Ho Tsai and Hsin-Min Wang. Automatic detection and tracking of target singer in multi-singer music recordings. In *Proceedings of the 2004 International Conference on Acoustics, Speech, and Signal Processing*, pages 221–224, 2004.

23 Chong-Kai Wang, Ren-Yuan Lyu, and Yuang-Chin 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.

24 Chi Hang Wong, Wai Man Szeto, and Kin Hong Wong. Automatic lyrics alignment for Cantonese popular music. *Multimedia System*, 4-5(12):307–323, 2007.

25 Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. *The HTK Book Version 3.4*. Cambridge University Engineering Department, 2006.