Alignment of Lyrics With Accompanied Singing Audio Based on Acoustic-Phonetic Vowel Likelihood Modeling

Yu-Ren Chien, Hsin-Min Wang, *Senior Member, IEEE*, and Shyh-Kang Jeng

***Abstract*—This study addresses the task of aligning lyrics with accompanied singing recordings. With a vowel-only representation of lyric syllables, our approach evaluates likelihood scores of vowel types with glottal pulse shapes and formant frequencies extracted from a small set of singing examples. The proposed vowel likelihood model is used in conjunction with a prior model of frame-wise syllable sequence in determining an optimal evolution of syllabic position. In lyrics alignment experiments, we optimized numerical parameters on two independent development sets and then tested the optimized system on two other datasets. New objective performance measures are introduced in the evaluation to provide further insight into the quality of alignment. Use of glottal pulse shapes and formant frequencies is shown by a controlled experiment to account for a 0.07 difference in average normalized alignment error. Another controlled experiment demonstrates that, with a difference of 0.03, F0-invariant glottal pulse shape gives a lower average normalized alignment error than does F0-invariant spectrum envelope, the latter being assumed by MFCC-based timbre models.**

***Index Terms*—Acoustic phonetics, F0 modification, formant frequency, glottal pulse shape, lyrics alignment, singing voice, vowel likelihood, vowel timbre examples.**

I. Introduction

SONGS typically come with words. The lyrics of a song determine how the song is performed in terms of phonetic articulation, and shape the timbral variations perceived by those who listen to the performance. The timings in which words and syllables in the lyrics are sung are highly variable, both within a song and between different songs—some syllables are short because they are assigned to a short musical note; others are long, associated with a long note or multiple notes; and the tempo adds to the uncertainty in timing. This variability in timing makes it appealing to display lyric syllables synchronously in karaoke applications, or synchronized lyric lines or words as a visual augmentation to song playback. Furthermore, a form of rhythmic information could be derived from timings of lyrics and used as a feature in music information retrieval in place of the rhythm of musical notes. Nevertheless, timing information such as this is typically lacking in commercial distributions of lyrics.

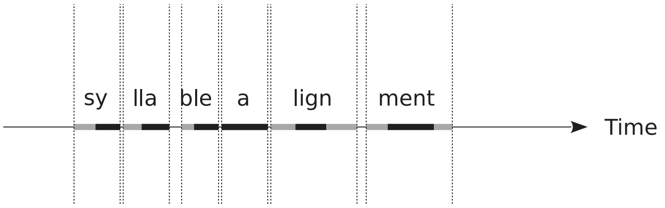


Fig. 1. Intervals plotted on the time axis for the syllables in the phrase “syllable alignment.” Vertical dotted lines mark the boundaries of these syllables. Each syllable is composed of a nucleus in black and possibly a consonant in grey preceding or following the nucleus.

In this paper, we address automatic extraction of lyric timings from accompanied singing audio, i.e., alignment of lyrics text with the audio. The aligned textual units in the lyrics can be syllables, words, phrases, or lines. The desired results of alignment consist of onset and offset time positions of each aligned unit. This is illustrated in Fig. 1 for alignment of lyric syllables. Here we assume that correct lyrics are available, specifying all the words sung in the audio one after another from the beginning to the end. Simply applying a speech recognition tool to the alignment of lyric text with the audio does not give satisfactory results—the audio contains substantial instrumental interference that can confuse a speech model, and the singing signal itself differs much from speech, both in the temporal proportion of consonant sounds, and in the range and variations of fundamental frequency (F0).

The apparent diversity in vocal timbre among singers makes it a natural choice to take advantage of human voice data in singing voice modeling. Iskandar *et al.* [1] used speech data and pop song data to build their phoneme likelihood model, and estimated the probability distribution of syllable durations from pop song data to constrain a Viterbi alignment of lyrics. In the lyrics alignment system of Wong *et al.* [2], singing voice detection was performed by a multi-layer perceptron trained with vocal and non-vocal onset data. To train hidden Markov models (HMM) for vocal segment classification in their lyrics alignment system, Kan *et al.* [3] used vocal and non-vocal segment data. They also used lyric-line duration data to estimate the prior-mean

Manuscript received January 17, 2016; revised April 21, 2016 and June 21, 2016; accepted July 13, 2016. Date of publication July 27, 2016; date of current version August 18, 2016. This work was supported by the Ministry of Science and Technology of Taiwan under Grant NSC 102-2221-E-001-008-MY3. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rodrigo Capobianco Guido.

Y.-R. Chien is with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei 106, Taiwan, and also with the Institute of Information Science, Academia Sinica, Taipei 115, Taiwan (e-mail: yrchien@ntu.edu.tw).

H.-M. Wang is with the Institute of Information Science, Academia Sinica, Taipei 115, Taiwan (e-mail: whm@iis.sinica.edu.tw).

S.-K. Jeng is with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei 106, Taiwan (e-mail: skjeng@ntu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2594282

durations of phonemes in singing, which were then used to calculate a duration estimate for each input lyric line. Mesaros and Virtanen [4] built phoneme likelihood models with speech and unaccompanied singing data. In the lyrics alignment system of Fujihara *et al.* [5], predominant melodic source data was extracted from popular songs to train Gaussian mixture models for vocal activity detection. They also used speech, singing, and separated singing data to build their phoneme likelihood models. In the aforementioned approaches where both speech and singing data is used for phoneme likelihood modeling, speech data is used in training Gaussian mixture models before the models are adapted to a small amount of singing data.

Some approaches in the lyrics alignment literature make use of knowledge in particular fields. Iskandar *et al.* [1] set an alignment time unit to a minimum note duration determined from the tempo of a song. Wong *et al.* [2] aligned tonal contours and non-uniform prosodic rhythms in the lyrics with melodies and onsets in the audio by dynamic time warping. Kan *et al.* [3] performed beat tracking and set their alignment time unit to one beat. To locate chorus sections, they detected repeated sections with chroma features. From a structural audio segmentation of a popular song, Lee and Cremer [6] identified the chorus and verse sections by measuring an audio similarity between instances of the same section type, and aligned them with manually labeled lyric sections by dynamic programming. Fujihara *et al.* [5] attempted to enhance the alignment of unvoiced consonants in the lyrics with the audio by detecting fricative sounds in their specific frequency bands. Mauch *et al.* [7] addressed a closely related task, where timings in the audio signal are estimated for paired chords and lyric words from a song book. Gong *et al.* [8] took the approach of spectral envelope estimation (with the True Envelope algorithm [9], [10]) to timbre modeling, in addressing the task of aligning unaccompanied singing with the score, which is composed of melody and lyrics. Dynamic time warping techniques were applied to lyrics spotting in unaccompanied singing [11] and to melody and lyrics matching for real-time accompaniment [12]. McVicar *et al.* [13] utilized the repetitive structure of popular songs to enhance the performance of automatic lyric transcription. Without using any audio data, Knees *et al.* [14] used the song title and artist name to retrieve via a search engine web pages containing various versions of the lyrics, and conducted "lyrics alignment" among the versions to generate a final consensus version of lyrics output.

To isolate the singing voice from the analyzed audio, use of techniques from audio source separation is made by some approaches to lyrics alignment. Wong *et al.* [2] used the central panning of singing voice in a stereo recording to enhance the vocal signal before alignment. Mesaros and Virtanen [4] aligned lyrics with separated vocal signals, where the separation was based on a reconstruction of accompaniment. They performed the reconstruction by applying non-negative matrix factorization to vocal-free time-frequency regions determined from an estimated vocal melody. The voice signal with which Fujihara *et al.* [5] aligned lyrics was resynthesized from partial frequencies and amplitudes extracted from accompanied singing according to an estimated predominant melody.

For the task of lyrics alignment, there has been no research in the literature that explores the use of formant and glottal pulse models from acoustic phonetics. Models of voice production are closely related to singing: A formant filter model of the vocal tract can represent various vowels sung by a singer [15], and glottal pulse shape models are relevant both to falsetto singing [16] and to personal voice quality. In each of the abovementioned approaches that fit Gaussian-mixture phoneme models to human voice data, one attempts to eliminate the effect that an unknown vocal F0 has on timbre modeling, by extracting Mel-frequency cepstral coefficients (MFCCs) as a representation of the vocal spectrum envelope. The actual quality of this representation depends on the specific distribution of vocal F0 in the analyzed human voice. Since singing presents a much larger vocal pitch range than speech, it could be difficult for the success of MFCCs in automatic speech recognition to generalize to lyrics alignment. Concerning explicit estimation of spectrum envelope, the timbre model sacrifices physical validity in supposing that the spectrum envelope of glottal source is invariant to F0. In contrast, the wide range of vocal F0 in singing can be handled by a basic F0 parameter in a model of voice production. In this paper, we use a voice production model to simulate a vowel timbre at any vocal F0 estimated from the analyzed audio, thereby circumventing pitch-blind representation of spectrum envelope. By implementing an invariance of glottal pulse shape to F0 in timbre modeling, our approach gives a lower average normalized alignment error than otherwise yielded by a form of F0-invariant spectrum envelope, as shown in the experiments. In particular, the glottal pulse shape model adaptively represents various pulse shapes in different timbre examples. This distinguishes our voice production model from the source-filter model of Durrieu *et al.* [17], which assumes a fixed glottal pulse shape with the KLGLOTT88 model [18] and can therefore give rise to an error in the filter estimate whenever the true pulse shape deviates from the assumed shape.

In our approach, a vocal component in the polyphonic audio is isolated according to a vocal F0 sequence estimated from the audio. At each analysis time position, a lyric vowel can be identified for the vocal component by a *timbral fitness score* of the component with respect to the vowel, which is a variant of the timbral fitness measure used in our previous work on vocal melody extraction [19]. Furthermore, in verifying the effectiveness of this timbral fitness measure in lyrics alignment, we fit values of all numerical parameters to 2 labeled development sets, and evaluated the performance on another 2 data sets. The proposed method differs from [19] in the following respects:

1) a text processing procedure for converting the input lyrics into a sequence of vowels;
2) an HMM state encoding scheme for all the syllabic positions in the lyrics;
3) an HMM state transition model that ensures a proper order in which syllabic positions should be visited; and
4) a voicing fitness measure that provides the overall vowel likelihood model with the capability of distinguishing a vowel from a vocal rest.

The rest of this paper is outlined as follows. An overview is presented in Section II for the complete alignment system.

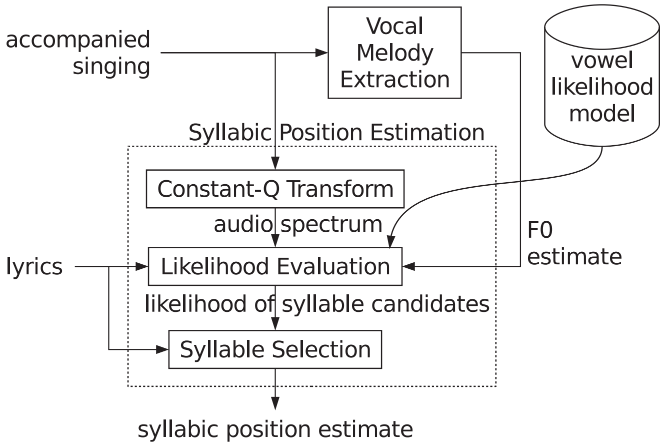


Fig. 2. Block diagram of the complete system.

Section III is dedicated to an acoustic-phonetic model of vowel likelihood, which defines how likelihood scores of vowel candidates are evaluated for each analysis time position. Described in Section IV is an algorithm for lyrics alignment based on the vowel likelihood model. Experiments and results are documented in Section V. Section VI concludes the paper.

II. System Overview

A block diagram of the complete system is shown in Fig. 2. With the lyrics regarded as an alternating sequence of syllables and vocal rests, we construct an alignment of lyrics as a *syllabic position* sequence, which specifies a syllabic position (initial rest, first syllable, rest following first syllable, second syllable, rest following second syllable, etc.) for each and every analysis time position in the accompanied singing signal. This sequence is generated by an estimation procedure that 1) extracts partial amplitudes from the audio spectrum according to an F0 estimate produced by the vocal melody extractor of Chien *et al.* [19], 2) evaluates likelihood scores for syllabic position candidates with the proposed vowel likelihood model, and 3) uses sequential constraints among the lyric syllables to select a syllabic position for each time position. Note that whereas vocal F0 is undefined within vocal rests, the melody extractor gives an F0 estimate for every time position.

III. Acoustic-Phonetic Model of Vowel Likelihood

In this section, we describe a likelihood model of *vowel type*, which is a variant of the F0 likelihood model used in our previous work on vocal melody extraction [19]. In the current model, the observation is represented by a sequence of partial amplitudes extracted from the spectrum of input signal according to a vocal F0 estimate. These partial amplitudes exhibit a specific timbral quality that, when checked against timbre examples of a vowel hypothesis, indicates how likely the hypothesis gives the true vowel type. In addition, overall loudness of the partials is also taken into account in likelihood evaluation so that vocal rests can be distinguished from vowels. In this model, a vowel type can be either a specific vowel or a vocal rest. Let the vowel type at time m be denoted by a discrete random variable v_m. The likelihood function of v_m, given input signal $\mathbf{z}_m$ and vocal F0 estimate w_m, is modeled with a *timbral fitness measure* $F_h(\cdot)$ and a *voicing fitness measure*[1] $F_e(\cdot)$:

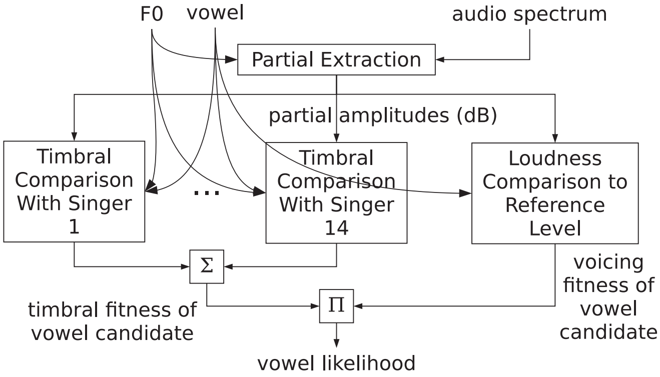

F0
vowel
audio spectrum
Partial Extraction
partial amplitudes (dB)
Timbral Comparison With Singer 1
...
Timbral Comparison With Singer 14
Loudness Comparison to Reference Level
Σ
Π
timbral fitness of vowel candidate
voicing fitness of vowel candidate
vowel likelihood

Fig. 3. Architecture of the vowel likelihood model.

$$\mathcal{L}_{v_m|\mathbf{z}_m,w_m}(v|\mathbf{z},w) \propto F_h(\mathbf{z},w,v)\cdot F_e(\mathbf{z},w,v). \quad (1)$$

Architecture of this model is represented in Fig. 3.

Our choice of dropping consonants from the representation of lyric syllables has been motivated by the nature of singing. To produce sustained pitches in his or her vocal melody, a singer would typically use vowels, which gives vowels a substantially higher proportion of time in a typical sung phrase than when the same phrase is spoken. As a result, consonants usually occur sparsely in singing in terms of temporal proportion, which could be responsible for human confusion of lyric consonants [20]. For a more direct evidence of this sparsity, note that while unvoiced phonation takes up roughly 25% of time in speech, the time ratio of unvoiced phonation in singing is less than 5% [21], given that all unvoiced sounds are consonants. Although capturing consonant features from a singing signal could be beneficial for lyrics alignment, we deem it difficult in practice because of the sparsity of consonants, and propose a vowel-only representation of lyric syllables as a model of the actual vowel dominance in singing signals. In lyrics alignment based on this representation, we expect, and reasonably tolerate, mapping of a lyric vowel to an audio segment that not only contains time positions for the vowel, but also contains (at its boundary) fewer time positions for some consonants actually sung around the vowel.

A. Timbral Fitness Measure

The timbral fitness of vowel hypothesis v with respect to input signal $\mathbf{z}$ and vocal F0 estimate w, is measured by comparing the timbral quality exhibited by observed partials to N_s timbre examples of vowel v that have been F0-modified to w by an offline acoustic-phonetic procedure [19]. The comparison consists in converting a distance between observation and example to a similarity score by a monotonically decreasing function. The specific conversion would also determine sensitivity of the similarity score to timbral differences. For this conversion, we use the (right-hand side of) zero-mean Gaussian function with

[1] These fitness measures are not probability distributions.

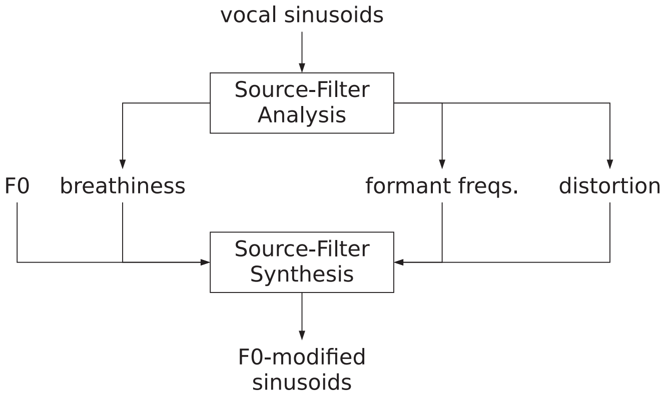


Fig. 4. Block diagram for the F0 modification procedure.

a standard deviation parameter. Lesser values for this standard deviation would represent greater sensitivity of the similarity, and thus a greater weight for the timbral fitness.

In the case where hypothesis v is a vocal rest, we need a timbral fitness score that measures how well the observed partial amplitudes fit our expectation for a vocal rest. At a vocal rest, what the melody extractor [19] gives in addition to an expected symbol for the rest, is a frequency w that does not typically match any vocal or instrumental F0. Such an F0 estimate often results in partial frequencies at non-sinusoidal, low-magnitude positions in the spectrum. In consequence, we assume that any timbre is possible for these partials:

$$F_h(\mathbf{z}, w, v) = 1 \tag{2}$$

whenever v is a vocal rest. Obviously, vocal rests cannot be recognized with the timbral fitness score alone; rather, detection of vocal rests relies on the voicing fitness measure.

The F0 modification procedure synthesizes sinusoids for the new F0 while preserving source-filter timbral discriminants estimated from the timbre-example sinusoids [19]. These timbral discriminants include glottal pulse shape (breathiness), vocal tract transfer function (formant frequencies), and distortion, as shown in Fig. 4. In the synthesis, a waveform is constructed for the glottal flow derivative according to a glottal flow model controlled by a shape parameter as well as by the F0. Meanwhile, frequency response of the vocal tract filter is determined by the formant frequencies. The glottal waveform, the vocal-tract frequency response, and a piecewise-linear spectral distortion are used to calculate amplitudes for the synthesized sinusoids. In the source-filter analysis, the shape parameter, formant frequencies, and distortion are estimated from the timbre example by numerically solving an optimization problem. In this optimization problem, sinusoids are synthesized (without distortion) for the old F0 from the timbral discriminant variables, and amplitude deviations of these sinusoids from the timbre example are minimized with respect to the variables.

We collected N_s timbre examples ($N_s = 14$) for each of the following 6 vowel types: /i/, /ɛ/, /ɑ/, /ɔ/, /u/, and /ə/ [19]. Since apparently many vowels are not included in this set, we are in fact dividing all possible vowels into the 6 categories with a set of rules for categorical mapping. For our experiments on songs with English lyrics, we map each lyric diphthong to its first component vowel, each lyric /æ/ to /ɛ/, and each lyric /Λ/ to /ə/. Here the diphthong mapping is based on the singing practice of gliding toward the end of note, ignoring the typically short ending component vowel. Similarly, for Chinese lyrics, we map each lyric falling diphthong to its first component vowel, each lyric /y/ to /i/, and each lyric /ɨ/ to /ə/. Each timbre example assumes a sinusoidal representation. The 14 timbre examples for each vowel type cover different genders, genres, and voice types [19]. Comparison of 6 different vowel timbre examples from the same singer's performance is presented in Fig. 5.

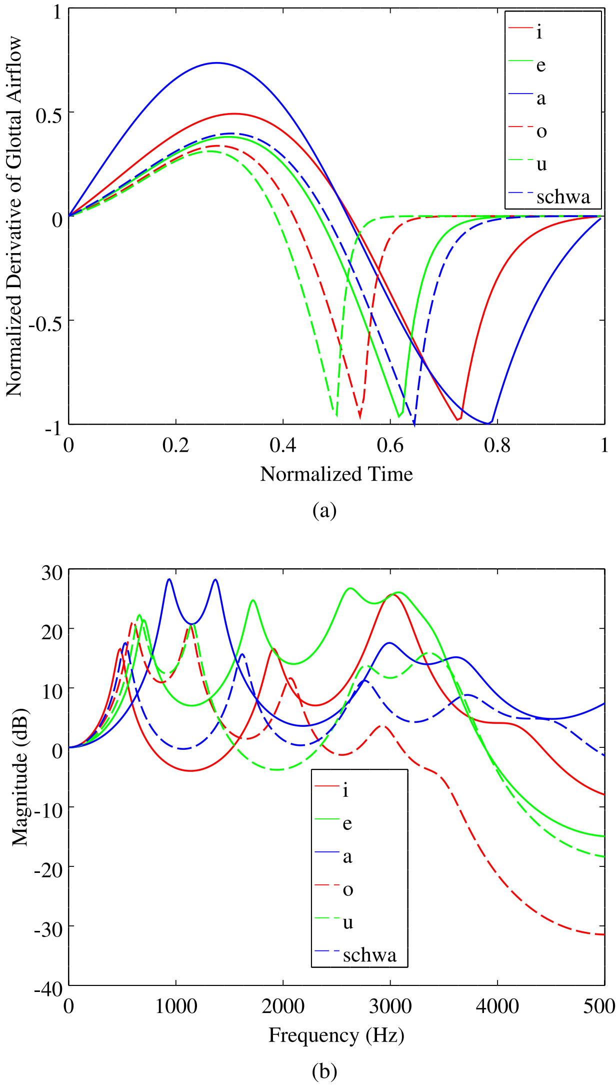


Fig. 5. Source-filter analysis of 6 mezzo-soprano's vowel timbre examples. The word "schwa" refers to the vowel /ə/. (a) Glottal pulse shapes. (b) Vocal tract filters.

B. *Voicing Fitness Measure*

In an attempt to distinguish a vowel from a vocal rest, we estimate the loudness of singing voice from input signal $\mathbf{z}$ and vocal F0 estimate w[2]:

$$\Lambda_e(\mathbf{z}, w) = \sum_{f \in J_w} P_f^z, \tag{3}$$

where J_w denotes the set of partial frequencies and P_f^z denotes the sinusoidal power spectrum of the input signal [19]. A high vocal loudness would indicate a vowel, while a low vocal loudness would indicate a vocal rest. Note that the melody extractor gives a vocal F0 estimate even at a vocal rest, which would in turn give, ideally, a zero loudness estimate, without any sinusoid found at its partial frequencies. In practice, overestimation of the loudness at a vocal rest can occur when the F0 estimate coincides with an instrumental F0, in which case the loudness of the instrument could be spuriously used. When the coincidence does happen, a sinusoid tracking procedure serves to alleviate the overestimation, where many instrumental sinusoid contours are removed for their almost constant frequency [19].

A *voicing fitness score* is calculated to measure the fitness of hypothesis v in terms of presence of singing voice. This score can now be defined by comparing the vocal loudness estimate $\Lambda_e(\mathbf{z}, w)$ to a reference loudness level chosen for the hypothesis: If v is a vowel, regardless of its exact vowel type,

$$F_e(\mathbf{z}, w, v) = \exp\left\{-\frac{(\lambda_e - \lambda_s)^2}{2\sigma_s^2}\right\}; \tag{4}$$

otherwise (if v is a vocal rest),

$$F_e(\mathbf{z}, w, v) = \exp\left\{-\frac{(\lambda_e - \lambda_r)^2}{2\sigma_r^2}\right\}. \tag{5}$$

Here, λ_s (dB) denotes an adaptive loudness level for a vowel, which is calculated by median-filtering (with filter length specified by N_m[3]) the vocal loudness estimate $\Lambda_e(\mathbf{z}, w)$ across all analysis time positions and taking the maximum of the filter output. By median filtering, the procedure rejects spikes in the loudness variations. With the reference level for a vowel defined, that for a vocal rest, denoted by λ_r, is then determined by subtracting a dynamic range ρ_d (dB) from λ_s. In (4) and (5), λ_e (dB) denotes a smoothed, compressed vocal loudness estimate. The initial loudness estimate $\Lambda_e(\mathbf{z}, w)$ is compressed with a lower limit of λ_r and an upper limit of λ_s. After that, the loudness is median-filtered across all analysis time positions (with filter length N_v) to smooth out any isolated, brief loudness dips resulting from F0 estimation errors. The effect loudness deviation has on the likelihood score is scaled by nonnegative parameters σ_s and σ_r. Lesser values for these two parameters would represent greater effect of the loudness deviation, and thus a greater weight for the voicing fitness.

IV. Syllabic Position Estimation

As the first step in aligning lyric syllables with audio, the lyrics text is processed to generate a list of syllabic position candidates. Estimation of syllabic position is performed on a grid of 100 time positions per second. At each time position, likelihood scores are evaluated for all the candidates with the proposed vowel likelihood model. Last, a syllabic position is selected for each time position according to the likelihood scores and the sequential relation among syllabic positions.

A. Lyrics Preprocessing

To convert the sequence of words in each lyric line into a sequence of vowels, each word is looked up in a digital pronunciation dictionary, thereby converting the word sequence into a sequence of phonemes. Next, consonants are removed from the phoneme sequence, and each vowel is mapped to one of the 6 vowel types, /i/, /ε/, /α/, /ɔ/, /u/, and /ə/. When multiple pronunciations are available in the dictionary for a syllable, all vowel variants are kept for the syllable. Our implementation of the lyrics preprocessor can process North American English and Standard Chinese. The English dictionary used is the CMU Pronouncing Dictionary. For Chinese lyrics, a 72 647-word dictionary is used for maximum-matching word segmentation and pronunciation lookup.

To generate a complete set of syllabic position candidates for lyrics alignment, we create a unique candidate for each element in the vowel sequence. For a syllable with vowel variants, a unique candidate is created for each variant, with multiple variant candidates sharing the same syllabic position. Unique candidates are also declared for vocal rests between every two neighboring elements in the vowel sequence, before the first vowel, and after the last vowel. For instance, for a lyrics text file consisting of 50 syllables, at least 101 syllabic position candidates should be created. The resulting set of candidates is denoted by $C = \{1, 2, \ldots, N_c\}$, where each candidate is represented by a positive integer and N_c denotes the number of candidates. Each syllabic position $s \in C$ is associated with a syllable index n_s, which locates s as the n_sth syllable in the lyrics or the vocal rest following the n_sth syllable. A zero syllable index $n_s = 0$ indicates that s is the initial vocal rest. Each syllabic position $s \in C$ is also associated with a vowel type $v^{(s)} \in \{0, 1, \ldots, 6\}$. A zero vowel type $v^{(s)} = 0$ indicates that s is a vocal rest.

B. Likelihood Evaluation

Selection of a syllabic position for an analysis time position is based on likelihood evaluation performed over all syllabic position candidates for the same time position. The proposed vowel likelihood model (1) yields 7 likelihood scores respectively for the 7 vowel types: /i/, /ɛ/, /ɑ/, /ɔ/, /u/, /ə/, and vocal rest. Each syllabic position candidate s is assigned one of the 7 scores according to its vowel type $v^{(s)}$. Notice that with this likelihood evaluation scheme, many candidates can share the same likelihood score because they have the same vowel type, a fact that reveals the importance of selecting syllables jointly over all analysis time positions with sequential relation taken into account.

In the implementation, likelihood scores for the first and last analysis time positions are defined in a way that directly ensures proper syllabic positions for these two time positions. Since the audio should start with the initial vocal rest or the first syllable,

[2] Following the notation in (1), we let $\mathbf{z}$ and w denote specific values on which random variables $\mathbf{z}_m$ and w_m take, respectively.

[3] For values of numerical parameters, see Table II.

all other candidates are assigned a zero score for the first time position. Similarly, all candidates are assigned a zero score for the last time position, except for the final vocal rest and the last syllable.

C. Syllable Selection

1) Sequence Estimation: Let s_m denote the syllabic position of singing voice at time position m. To estimate the sequence $\{s_m\}_{m=1}^{M}$, we let the joint probability distribution of syllabic position evolution and accompanied singing audio,

$$p(s_1, \ldots, s_M, \mathbf{z}_1, \ldots, \mathbf{z}_M) = P(s_1, \ldots, s_M) p(\mathbf{z}_1, \ldots, \mathbf{z}_M | s_1, \ldots, s_M), \tag{6}$$

be approximated by a HMM [22]:

$$\begin{aligned} P(s_1, \ldots, s_M) &= P(s_1) P(s_2|s_1) P(s_3|s_1, s_2) \cdots \\ &\quad P(s_m|s_1, \ldots, s_{m-1}) \cdots \\ &\quad P(s_M|s_1, \ldots, s_{M-1}) \end{aligned} \tag{7}$$

$$\approx P(s_1) \prod_{m=2}^{M} P(s_m|s_{m-1}), \tag{8}$$

$$p(\mathbf{z}_1, \ldots, \mathbf{z}_M | s_1, \ldots, s_M) \approx \prod_{m=1}^{M} p(\mathbf{z}_m|s_m). \tag{9}$$

In (9), likelihood scores of syllabic position candidates are defined as in Section IV-B. In (8), state transition probabilities among candidates encode sequential relation among syllabic positions. With this probabilistic model, the estimation can be achieved by maximizing the posterior probability of state sequence with the Viterbi algorithm. We create a state transition probability matrix from the preprocessed lyrics information, as detailed below.

2) State Transition Probabilities: A uniform initial state distribution can be assumed here because a proper syllabic position at the first time position has been guaranteed by likelihood scores:

$$P(s_1 = s) = \frac{1}{N_c}, \forall s \in C. \tag{10}$$

As shown in Fig. 6, the conditional probability $P(s_m|s_{m-1})$ considers a particular syllabic position given for the previous time position, and defines the probabilities of all possible syllabic positions for the current time position. We derive specific values for these probabilities from the fact that syllabic positions must be visited in the same order as they appear in the lyrics. Since only a continuation of the same syllabic position or a transition to a succeeding syllabic position is allowed, we assign zero probabilities to all other syllabic positions.

First, consider the case where s_{m-1} is a vowel, i.e., $v^{(s_{m-1})} \neq 0$. In this case, the transition probabilities have no influence on the number of transitions between syllables, which has been fixed by the aligned lyrics. Moreover, since the Viterbi algorithm implements a joint optimization of all the elements of the state sequence, an increased probability for continuation does not create a tendency for delayed transition—a delayed transition would be penalized exactly as much as an earlier transition by an increased probability for continuation. Consequently, a uniform distribution can be assumed between continuation and transition, with a probability of 0.5 assigned to the continuation, and the remaining probability of 0.5 distributed among several eligible successors. If $n_{s_{m-1}}$ points to the end of a lyric line, the only eligible successor is the vocal rest immediately following s_{m-1}, which is expected to last relatively long. Otherwise, we make the vocal rest optional, with the rest taking 0.25 and all vowel variants for the next syllable uniformly sharing the remaining 0.25, where uniform distributions are arbitrarily chosen because detection of vocal rests and vowel variants is not the primary goal of this work.

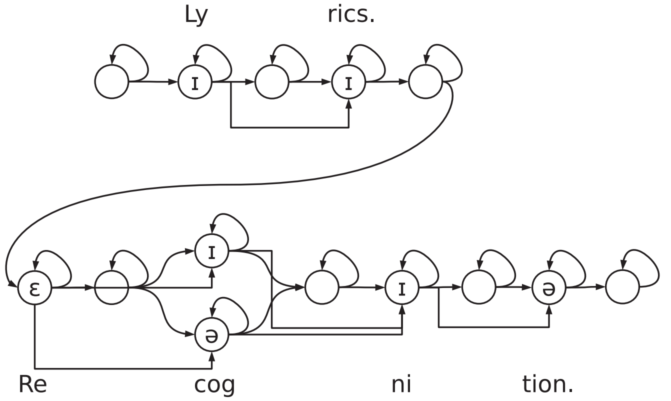


Fig. 6. A state transition diagram depicting possible transitions among syllabic positions in 2 consecutive toy lyric lines, "Lyrics" and "Recognition." Blank nodes represent vocal rests.

Second, consider the case where s_{m-1} is a vocal rest, i.e., $v^{(s_{m-1})} = 0$. Again, if this is at a line break, we use a parameter $P_c < 0.5$ to control the tendency for this rest to last longer than vocal rests within a lyric line, assigning a probability of $1 - P_c$ to the continuation; otherwise, the continuation probability can be set to 0.5 again because of the fixed number of transitions. The remaining probability is again shared uniformly among vowel variants of the $(n_{s_{m-1}} + 1)$th syllable.

V. Experiments

A. Data Sets

In Sections III and IV, the vowel likelihood model and the procedure for syllabic position estimation have been described with a set of numerical parameters, whose values need to be determined empirically. To select appropriate values for these parameters, we considered their effect on alignment performance by carrying out alignment experiments on 2 *development sets*: `adc2004` and `labrosa`. Data set `adc2004` is adapted from the data used for audio melody extraction in the ISMIR 2004 audio description contest (ADC 2004). ADC 2004 in its entirety consists of 20 audio recordings with melody annotations, among which 8 recordings are fully instrumental, and the other 12 are accompanied singing. For lyrics alignment, only recordings with words are included in `adc2004`, including 4 pop song excerpts, 2 song excerpts with synthesized vocal, and 3 opera

TABLE I
SUMMARIES FOR THE DATA SETS

Data Set	#Excerpts	Excerpt Length (s)	Aligned Unit
adc2004	9	∼20	syllable
labrosa	9	∼30	syllable
poly_100	100	9–49	line
slam	130	9–52	phrase

excerpts[4]. Melody annotations are replaced with syllable onset and offset annotations and vowel-sequence transcriptions, the latter serving as a substitute for lyrics lacking in this data set. We intend to produce syllable-level annotations to make full use of this small data set. Data set labrosa is similarly adapted from a data set created for polyphonic melody extraction by LabROSA, Columbia University, consisting of English popular song excerpts. Numbers of excerpts and lengths are listed in Table I for these data sets.

To assess the generalization performance of parameter optimization, we used 2 test sets without any overlap with the development sets: poly_100 and slam. Data set poly_100 was created by Mesaros and Virtanen [4], composed of excerpts extracted from 17 English popular songs (8 female artists and 9 male artists), their lyrics, and lyric-line onset and offset annotations. A total of 4–8 excerpts are extracted from each song, with each excerpt capturing a complete structural section in a song, such as a chorus and a verse. From a collection of 20 Chinese popular songs (10 female artists and 10 male artists), we derived a similar data set slam[5] with 3–9 excerpts extracted from each song. For this data set, we produced onset and offset annotations for a smaller unit of alignment, a lyric phrase. Chinese lyric phrases are separated by spaces or line breaks, and each lyric line consists of one or more lyric phrases. Numbers of excerpts and lengths are listed in Table I for these data sets.

B. Performance Measures

In the experiments documented here, a textual unit for alignment is specified for each alignment task, which can be a lyric syllable, a lyric word, a lyric phrase, or a lyric line. Onset and offset time estimates are extracted for the sequence of aligned units from the evolution of syllabic position given by the tested system. The extraction proceeds by finding onsets and offsets for syllables before identifying specific syllabic onsets and offsets that correspond to boundaries of the aligned units.

To measure the overall performance of a tested system in aligning lyrics of an excerpt with its audio, we calculate the *average absolute alignment error* and the *average normalized alignment error*. The average absolute alignment error is computed by taking the average of distances (in seconds) of all the onset and offset estimates from ground-truth annotations. When each of these distances is normalized by the true duration of the aligned unit, the normalized distance represents an alignment error as a proportion in the true duration. An upper limit of unity can be further applied to the normalized distance to avoid discriminating among any excessive distances. We calculate the average of such normalized, limited distances to give the average normalized alignment error.

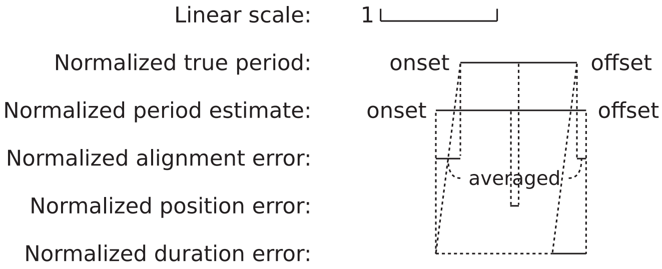

Linear scale:
1
Normalized true period:
onset
offset
Normalized period estimate:
onset
offset
Normalized alignment error:
averaged
Normalized position error:
Normalized duration error:

Fig. 7. Normalized performance measures of lyrics alignment.

Alignment of a textual unit with audio attempts to determine the position and duration of the unit as measured on the time axis of audio. To focus performance measurement on one of these two factors, we calculate the *average normalized position error* and the *average relative duration error*. With the position of a unit defined by the midpoint between onset and offset, the average normalized position error is computed by taking the average of, again, normalized and limited distances of all the position estimates from ground-truth values. The same calculation applied to the duration gives the average relative duration error. An example is depicted in Fig. 7 to illustrate these normalized errors.

C. Results on the Development Sets

The search for optimal parameter values was performed on the development sets in 2 stages [19]. In the first stage, parameters were manually adjusted to eliminate as many errors in syllable alignments as possible, which were indicated by particular syllables aligned with excessive absolute errors. In the second stage, the average absolute alignment error was further minimized by coordinate descent in 3 cycles.

The results of parameter optimization are presented in Table II, where optimal settings are compared to extreme settings with average absolute alignment errors given by syllable alignment experiments conducted on the development sets. When all the optimal settings were used, the resulting average absolute alignment error was 0.448 s. Each row contains 3 parameter settings derived from the optimal setting by assigning 3 different values to the designated parameter: The first value is the smallest sample point, the second value is the optimal value, and the third value is the largest sample point. As shown in Fig. 8, the limited effect that changes in P_c have on the error, suggests that the assumed relative lengthiness of vocal rests at line breaks may not be sufficiently consistent in practice to deserve a special transition probability assignment. What is more, labrosa could present stronger accompaniment than does adc2004, making itself more difficult to analyze both in vocal melody extraction and in lyrics alignment. This could explain the optimal error for labrosa in Fig. 8 being larger than twice the error

[4] Two excluded vocal excerpts (daisy3 and daisy4) do not have words. Another vocal excerpt (opera_male3) was excluded for its excessive difficulty, featuring an exaggerated loudness contrast in singing.

[5] Available by personal contact.

TABLE II
OPTIMAL AND EXTREME PARAMETER SETTINGS AND THE RESULTING AVERAGE ABSOLUTE ALIGNMENT ERRORS EVALUATED ON THE DEVELOPMENT SETS

Param.	Smallest Value	Optimal Value	Largest Value
P_c	(0.05, 0.453)	(0.2, 0.448)	(0.5, 0.487)
ρ_d	(9.0, 2.492)	(27.0, 0.448)	(36.0, 0.965)
σ_s	(0.1, 2.304)	(0.4, 0.448)	(1.0, 1.142)
σ_r	(0.1, 1.384)	(0.4, 0.448)	(1.0, 1.329)
c_h	(0.003, 0.522)	(0.023, 0.448)	(0.043, 0.468)
N_h	(5, 0.660)	(10, 0.448)	(15, 0.497)
N_m	(1, 0.931)	(35, 0.448)	(280, 1.065)
N_v	(10, 0.699)	(22, 0.448)	(30, 0.546)
N_a	(2, 1.656)	(7, 0.448)	(10, 0.475)
θ_p	(0.0, 0.795)	(12.0, 0.448)	(20.0, 1.196)
θ_r	(0.1, 0.833)	(0.9, 0.448)	(2.1, 0.945)
θ_j	(0.5, 0.484)	(1.0, 0.448)	(1.5, 0.484)
θ_g	(0.002, 0.448)	(0.1, 0.448)	(0.5, 0.700)
θ_d	(1, 0.540)	(9, 0.448)	(21, 0.498)
N_e	(6, 0.532)	(14, 0.448)	(16, 0.474)

In each ordered pair, the first entry gives a parameter value and the second entry gives an error in seconds. The reader is referred to Sections III and IV for the units of parameter values. (All error entries in column "Optimal Value" are given by the same optimal setting.)

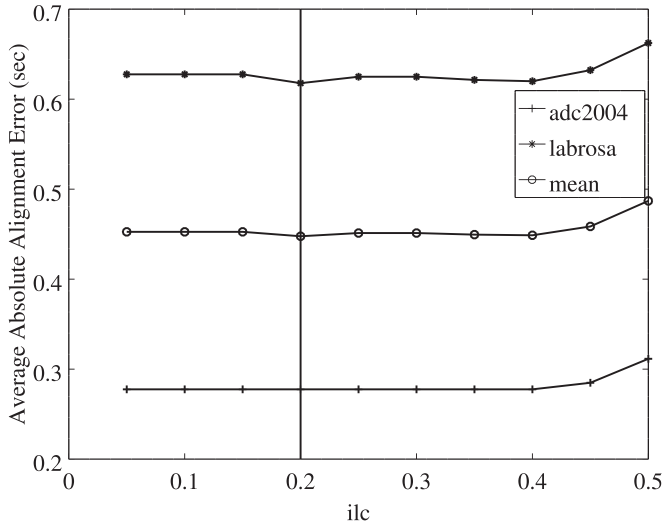


Fig. 8. Average absolute alignment error evaluated on the development sets for sample points of P_c. The vertical dotted line marks the optimal parameter value.

for adc2004. As defined in [19], c_h is a parameter controlling the sensitivity of timbral comparison to magnitude deviation; N_h is the number of partials on which timbral comparison is performed; N_a is the number of frequency bins for magnitude averaging in sinusoid detection; θ_p is a peak prominence threshold for sinusoid detection; θ_r and θ_d are frequency-range and contour-duration thresholds, respectively, for sinusoid pruning; θ_j and θ_g are frequency- and amplitude-change thresholds, respectively, for sinusoid tracking; and N_e is the number of partials on which loudness is evaluated.

D. Results on the Test Sets

Performance evaluation of the proposed approach on the test data sets is presented in Table III. The absolute alignment error is around one second for both sets. Since many Chinese lyric lines in slam contain only one phrase, line-level alignment with poly_100 and phrase-level alignment with slam would not be expected to give dissimilar absolute errors. The normalized alignment error is below 0.3 for both sets. For an onset or offset, a normalized alignment error of 0.3 means an absolute error equal to 30% the true duration of aligned unit, but does not indicate a specific error in estimation of position or duration: If errors of 0.3 are in the same direction for onset and offset, the position error will be 0.3, and the duration error will be zero; otherwise, the position error will be zero, and the duration error will be 0.6. Results in normalized position and relative duration errors indicate that errors associated with these two types of estimation are in fact below or close to 0.3 on average for both data sets.

TABLE III
RESULTS OF PERFORMANCE EVALUATION ON poly_100 AND slam

Data Set	AA	NA	NP	RD
poly_100	0.897 s	0.251	0.229	0.306
slam	1.069 s	0.295	0.260	0.327

AA = Average Absolute Alignment Error; NA = Average Normalized Alignment Error; NP = Average Normalized Position Error; RD = Average Relative Duration Error.

TABLE IV
PERFORMANCE COMPARISON WITH VARIANTS OF THE APPROACH OF MESAROS AND VIRTANEN [4] BY AVERAGE ABSOLUTE ALIGNMENT ERRORS EVALUATED ON poly_100

Approach	Error
This Paper	0.90 s
[4], 8-Class Adaptation	0.94 s
[4], 3-Class Adaptation	0.97 s
[4], 22-Class Adaptation	1.07 s

The approaches are sorted by error in ascending order.

As shown in Table IV, our approach gave an average absolute alignment error slightly lower than those given by variant approaches of Mesaros and Virtanen [4] on poly_100. Their approach to lyrics alignment linearly adapts several (3, 8, or 22) categories of speech phoneme likelihood models to a small amount of singing data, with all the models in a category sharing the same linear transformation. This comparison shows that, if our approach does not significantly outperform their approach, these two approaches could be considered comparable in performance.

E. Example

To gain insight into our test results, consider the excerpt in poly_100 for which the resulting average normalized alignment error was 0.197, which is the 50th lowest error among all the 100 error values and serves as a median item that could represent the entire data set. As shown in Fig. 9, the estimated evolution of lyric position adequately matched the true evolution at all 6 lyric lines except the third line, where the transition to vocal rest occurred prematurely. Inspection of its spectrogram

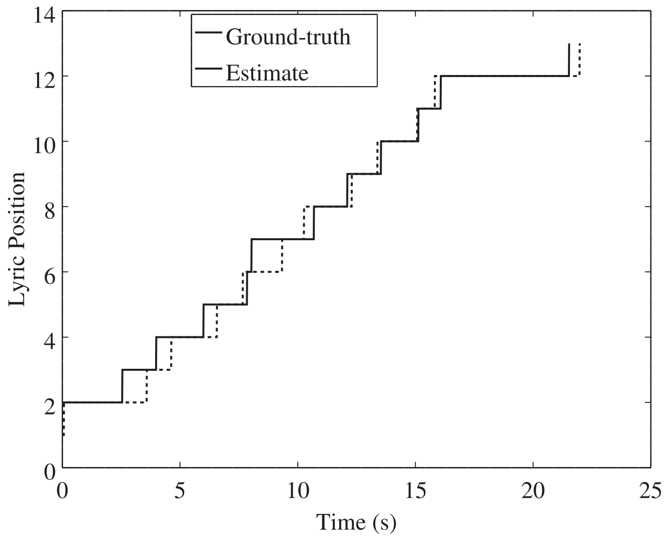


Fig. 9. Lyric position sequence estimated from the excerpt with the 50th lowest average normalized alignment error in `poly_100`. The initial vocal rest is represented by position 1, the first line is represented by position 2, the vocal rest following the first line is represented by position 3, and so on. In other words, odd positions represent vocal rests at line breaks, whereas even positions represent lyric lines.

revealed that the third line is sung with a relatively low loudness, which led our algorithm to treat the segment of signal as a vocal rest. Many instances of low estimated vocal loudness actually indicate vocal rests that separate adjacent lyric syllables or lyric lines; as a result, our approach is inevitably confused by soft singing in some circumstances.

F. Results of Controlled Experiments

1) Timbre Model: In order to assess the effect of timbral fitness measure, we repeated the `poly_100` experiments with vowel likelihood defined by the voicing fitness measure alone:

$$\mathcal{L}_{v_m|\mathbf{z}_m,w_m}(v|\mathbf{z},w) \propto F_e(\mathbf{z},w,v). \tag{11}$$

This timbre-insensitive likelihood model gave the results listed in Table V on row “Timbre.” For lack of timbre checking, the normalized alignment error grows by 0.07. To establish statistical significance for the comparisons in our controlled experiments, we calculated a p-value by the two-sided Wilcoxon signed-rank test on the 100 pairs of average normalized alignment errors in each comparison, which is to be assessed against a significance level of 5%. The p-value for removal of timbre model is 0.001, which confirms the effectiveness of timbral fitness measure in identifying lyric vowels. Moreover, this result reveals that lyrics alignment based solely on estimating the loudness or voicing of singing voice, not performing any phonemic discrimination, could adequately estimate line-level evolution of lyric position. As an example of timbre-blind alignment, an evolution of lyric position is shown in Fig. 10, which is estimated without the timbre model from the same excerpt considered in Section V-E. In this example, the impact of timbre model removal is far above the average, giving a normalized alignment error of 0.772 and a clear deviation from the true evolution in Fig. 10.

TABLE V
RESULTS OF EXPERIMENTS CONDUCTED ON `poly_100` WITH AN ALGORITHMIC FEATURE REMOVED FROM THE PROPOSED APPROACH

Removed	AA	NA	NP	RD
None	0.897 s	0.251	0.229	0.306
Timbre	1.198 s	0.320	0.301	0.385
LF	1.012 s	0.278	0.253	0.328
Gender	0.979 s	0.269	0.253	0.320
Genre	0.993 s	0.266	0.244	0.334
Vx Type	1.032 s	0.277	0.256	0.338
Diversity	1.103 s	0.297	0.281	0.334
Voicing	9.929 s	0.885	0.898	0.979
P_c	0.863/1.137 s	0.244/0.323	0.224/0.310	0.304/0.365
ρ_d	2.192/1.085 s	0.489/0.307	0.418/0.292	0.645/0.330
σ_s	1.745/1.081 s	0.424/0.305	0.358/0.291	0.551/0.333
σ_r	1.199/1.275 s	0.336/0.324	0.321/0.270	0.350/0.435
c_h	0.932/0.918 s	0.257/0.250	0.237/0.226	0.333/0.306
N_h	0.947/0.786 s	0.259/0.227	0.234/0.203	0.328/0.299
N_m	1.020/1.013 s	0.269/0.293	0.231/0.276	0.360/0.325
N_v	0.968/0.920 s	0.275/0.250	0.256/0.228	0.308/0.314
N_a	2.261/0.943 s	0.513/0.251	0.463/0.227	0.621/0.319
θ_p	0.979/2.505 s	0.272/0.482	0.253/0.434	0.322/0.625
θ_r	1.115/1.021 s	0.304/0.286	0.283/0.249	0.336/0.350
θ_j	0.904/0.967 s	0.260/0.263	0.236/0.242	0.311/0.313
θ_g	0.898/0.914 s	0.251/0.256	0.229/0.231	0.307/0.318
θ_d	0.910/1.009 s	0.247/0.278	0.209/0.255	0.342/0.325
N_e	0.885/0.909 s	0.240/0.253	0.215/0.232	0.295/0.312

AA = Average Absolute Alignment Error; NA = Average Normalized Alignment Error; NP = Average Normalized Position Error; RD = Average Relative Duration Error.

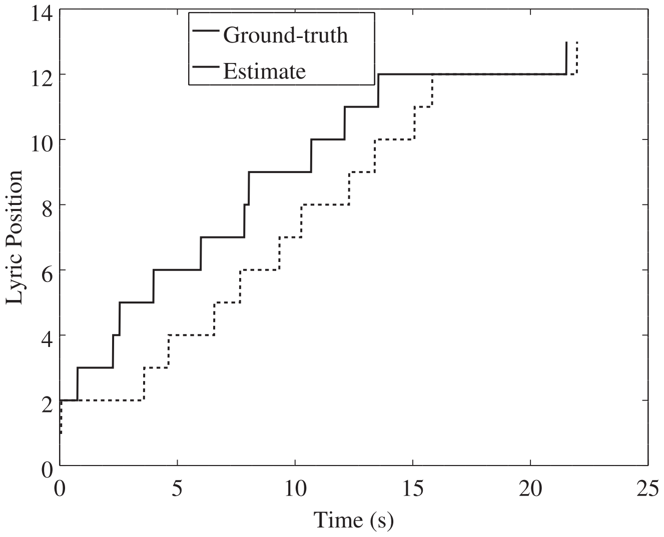


Fig. 10. Lyric position sequence estimated without the timbre model, from the excerpt with the 50th lowest average normalized alignment error in `poly_100`. For the definition of lyric position, see the caption of Fig. 9.

2) Transformed Liljencrants-Fant Model: In the F0 modification of vowel timbre examples, we implement an invariance in glottal pulse shape, which is represented by the transformed Liljencrants-Fant model [23] with shape parameter estimated from the original example. Since the spectrum envelope of the same glottal pulse shape actually varies with F0, this model

TABLE VI
RESULTS (p-VALUES) OF TWO-SIDED WILCOXON SIGNED-RANK TESTS FOR THE `poly_100` CONTROLLED EXPERIMENTS

Timbre	LF	Gender	Genre	Vx Type	Diversity
0.001	0.011	0.169	0.267	0.078	0.000

Voicing	P_c	ρ_d	σ_s	σ_r	c_h	N_h	N_m
0.000	0.000	0.000	0.000	0.001	0.673	0.978	0.072

N_v	N_a	θ_p	θ_r	θ_j	θ_g	θ_d	N_e
0.207	0.000	0.000	0.016	0.285	0.367	0.074	0.823

The significance test is applied to pairs of average normalized alignment errors given by each controlled experiment.

of glottal excitation sets our timbre model apart from MFCC-based models. To examine the effect of this model, we repeated the `poly_100` experiments with the (radiated) glottal excitation modeled in the conventional way by a fixed spectrum envelope [15]:

$$U_R(f^h) = \frac{f^h/100}{1 + (f^h/100)^2}, \tag{12}$$

where f^h denotes frequency in hertz. This glottal spectrum envelope, along with the vocal tract and distortion filters adapted to the timbre example, realizes an F0 modification procedure that preserves spectrum envelope. The repeated experiments gave results listed in Table V on row "LF." The difference of 0.027 in average normalized alignment error, along with a p-value of 0.011, confirms the advantage of transformed Liljencrants-Fant model in modeling vowel timbre.

3) Diversity in Vowel Timbre Examples: In Section III-A, vowel timbre examples are collected with diversity in gender, genre, and voice type. In an effort to evaluate the separate effect of these diversity factors, we repeated the `poly_100` experiments 3 times, each time using a timbre example subset where one of these diversity factors has been eliminated. The 7-singer male-only (7-example) subset gave the results listed in Table V on row "Gender." Labeled "Genre" are the results for the 4-singer opera-only (4-example) subset. The row labeled "Vx Type" is for the 6-example high-voice-only subset, which consists of the tenor, the soprano, and the other 4 high-voice singers. Removal of each diversity factor is found to raise the normalized alignment error by an amount between 0.015 and 0.026, but the corresponding p-value between 0.07 and 0.27, as listed in Table VI, leaves the benefit unconfirmed as to including these individual factors in timbre example collection. For the purpose of evaluating the composite effect of the diversity factors, we repeated the `poly_100` experiments with only one tenor example for each vowel type, giving results listed in Table V on row "Diversity." This shows that complete lack of diversity results in a significant increase of 0.046 (p-value: 0.000) in normalized alignment error. As a consequence, the benefit of including all the diversity factors, as opposed to the single-example vowel representation, has been confirmed. This benefit could be further amplified by including more diversity factors in the collection, such as a more complete genre set and singer age.

4) Voicing Model: To see the effect of voicing fitness measure, we repeated the `poly_100` experiments with vowel likelihood defined by the timbral fitness measure alone:

$$\mathcal{L}_{v_m|\mathbf{z}_m,w_m}(v|\mathbf{z}, w) \propto F_h(\mathbf{z}, w, v). \tag{13}$$

This loudness-insensitive vowel likelihood model gave results listed in Table V on row "Voicing," where the lack of voicing model brought about a dramatic rise in average normalized alignment error. Our parameter optimization efforts made on the development sets determined that, in order to minimize alignment errors, the alignment should predominantly depend on voicing clues provided by the variations in estimated vocal loudness, which is made evident by this result.

5) Optimality of Parameter Values: As mentioned in Section V-C, values of the 15 parameters in this approach have been selected to optimize syllable-level alignment performance measured on the development sets. To demonstrate the necessity of this optimization, we repeated the `poly_100` experiments 15 times, each time with a parameter set to the smallest or largest sample point used in the corresponding one-parameter line search, which typically deviates considerably from the optimal setting and is expected to give larger alignment errors on `poly_100` than does the optimal value. The resulting alignment errors are listed in Table V on rows titled with parameters in question, first for the smallest sample point and then for the largest sample point in each entry. This shows that for each parameter displaced from the optimum in two directions, the average normalized alignment error was consistently raised at least by one of the two displacements. A p-value was calculated only for the displacement that generated the larger error, as shown in Table VI. This confirms that the rise in error is significant for 7 of the 15 parameters displaced. In brief, the evaluated performance of our system can be in part attributed to parameter optimization.

VI. Conclusion and Further Work

An approach to lyrics alignment with audio has been presented that is based on a model of vowel likelihood. The vowel likelihood model is built upon a set of vowel timbre examples for the F0 estimate that are generated by F0-modifying a small set of singing voice samples. The F0 modification is achieved by source-filter analysis and synthesis with state-of-the-art models from the field of acoustic phonetics, which we previously applied to vocal melody extraction.

The proposed method has been evaluated in multiple experiments, not only to test its efficacy, but to look into the importance of various algorithmic features in the method. For two data sets alike, which have different origins and languages, our approach achieved an average normalized alignment error below 0.3 and an average absolute alignment error around one second. A state-of-the-art approach was previously evaluated on one of the two data sets, also giving an absolute error around one second. With a series of controlled experiments, we verified that timbral discrimination in the vowel likelihood model is effective as a mechanism for rendering the alignment phonetically sensitive. Even so, voicing modeling based on estimation of vocal loudness is indispensable for lyrics alignment.

This work demonstrates how acoustic-phonetic models lend themselves to distinguishing different vowel sounds. It would be intriguing to further inquire application of acoustic-phonetic models to other aspects of singing voice in the future. An example would be the personal timbre that characterizes one's singing, which might be represented by such models in singer recognition.

Acknowledgment

The authors would like to thank Dr. Annamaria Mesaros and Lien-Chiao Lin for sharing their ground-truth annotations for the data sets `poly_100` and `slam`. They would also like to thank Dr. Kuan-Yu Chen for his implementation of Chinese word segmentation.

References

[1] D. Iskandar, Y. Wang, M.-Y. Kan, and H. Li, "Syllabic level automatic synchronization of music signals and text lyrics," in *Proc. 14th Annu. ACM Int. Conf. Multimedia*, 2006, pp. 659–662.

[2] C. H. Wong, W. M. Szeto, and K. H. Wong, "Automatic lyrics alignment for Cantonese popular music," *Multimedia Syst.*, vol. 12, pp. 307–323, 2007.

[3] M.-Y. Kan, Y. Wang, D. Iskandar, T. L. Nwe, and A. Shenoy, "LyricAlly: Automatic synchronization of textual lyrics to acoustic music signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 338–349, Feb. 2008.

[4] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP J. Audio, Speech, Music Process.*, vol. 2010, no. 546047, pp. 1–11, 2010.

[5] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno, "LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1252–1261, Oct. 2011.

[6] K. Lee and M. Cremer, "Segmentation-based lyrics-audio alignment using dynamic programming," in *Proc. 9th Int. Conf. Music Inf. Retrieval*, pp. 395–400, 2008.

[7] M. Mauch, H. Fujihara, and M. Goto, "Integrating additional chord information into HMM-based lyrics-to-audio alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 200–210, Jan. 2012.

[8] R. Gong, P. Cuvillier, N. Obin, and A. Cont, "Real-time audio-to-score alignment of singing voice based on melody and lyric information," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, pp. 3312–3316, 2015.

[9] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Proc. 8th Int. Conf. Digital Audio Effects*, 2005, pp. 30–35.

[10] F. Villavicencio, A. Röbel, and X. Rodet, "Improving LPC spectral envelope extraction of voiced speech by True-Envelope estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. I, pp. 869–872, 2006.

[11] C. Dittmar, P. Mercado, H. Grossmann, and E. Cano, "Towards lyrics spotting in the SyncGlobal project," in *Proc. 3rd Int. Workshop Cognitive Inf. Process.*, 2012, 1–6.

[12] P. Papiotis and H. Purwins, "Real-time accompaniment using lyrics-matching QBH," in *Proc. 7th Int. Symp. Comput. Music Modeling Retrieval*, 2010, pp. 279–280.

[13] M. McVicar, D. P. Ellis, and M. Goto, "Leveraging repetition for improved automatic lyric transcription in popular music," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 3117–3121.

[14] P. Knees, M. Schedl, and G. Widmer, "Multiple lyrics alignment: Automatic retrieval of song lyrics," in *Proc. 6th Int. Conf. Music Inf. Retrieval*, 2005, pp. 564–569.

[15] G. Fant, *Acoustic Theory of Speech Production With Calculations Based on X-Ray Studies of Russian Articulations*. The Hague, The Netherlands: Mouton, 1970.

[16] G. L. Salomão and J. Sundberg, "What do male singers mean by modal and falsetto register? An investigation of the glottal voice source," *Logopedics Phoniatrics Vocology*, vol. 34, pp. 73–83, 2009.

[17] J. L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 564–575, Mar. 2010.

[18] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820–857, 1990.

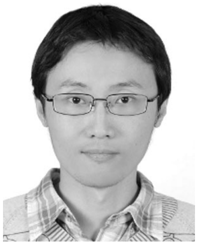
[19] Y.-R. Chien, H.-M. Wang, and S.-K. Jeng, "An acoustic-phonetic model of F0 likelihood for vocal melody extraction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1457–1468, Sep. 2015.

[20] L. Collister and D. Huron, "Comparison of word intelligibility in spoken and sung phrases," *Empirical Musicology Rev.*, vol. 3, no. 3, pp. 109–125, 2008.

[21] P. Cook, "Identification of control parameters in an articulatory vocal tract model with applications to the synthesis of singing," Ph.D. dissertation, Dept. of Music, Stanford Univ., Stanford, CA, USA, 1990.

[22] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[23] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *STL-QPSR*, vol. 2, 119–156, 1995.



Yu-Ren Chien received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering in 2000, 2002, and 2016, respectively, from National Taiwan University, Taipei, Taiwan. He is a Research Assistant at the Institute of Information Science, Academia Sinica, Taipei City, Taiwan. From 2007 to 2008, he was a Senior Engineer at Realtek Semiconductor Corp., Taiwan. In 2013, he was a visiting Ph.D. student at IRCAM, France. His research interests include music signal processing and speech acoustics.

Hsin-Min Wang (S'92–M'95–SM'04) received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1989 and 1995, respectively. In October 1995, he joined the Institute of Information Science, Academia Sinica, Taipei City, Taiwan, where he is currently a Research Fellow and the Deputy Director. He also holds a joint appointment as a Professor in the Department of Computer Science and Information Engineering, National Cheng Kung University. He currently serves an Editorial Board Member of IEEE/ACM Transactions on Audio, Speech and Language Processing and *APSIPA Transactions on Signal and Information Processing*. His major research interests include spoken language processing, natural language processing, multimedia information retrieval, machine learning, and pattern recognition. He received the Chinese Institute of Engineers Technical Paper Award in 1995 and the ACM Multimedia Grand Challenge First Prize in 2012. He was an APSIPA Distinguished Lecturer during 2014–2015. He is a Member of the International Speech Communication Association and ACM.

Shyh-Kang Jeng received the B.S.E.E. and the Ph.D. degrees from National Taiwan University, Taipei, Taiwan, China, in 1979 and 1983, respectively. In 1981, he joined the Faculty of the Department of Electrical Engineering, National Taiwan University, where he is currently a Professor. From 1985 to 1993, he visited the University of Illinois, Urbana-Champaign, IL, USA, as a Visiting Research Associate Professor and a Visiting Research Professor several times. In 1999, he visited the Center for Computer Research in Music and Acoustics, Stanford University, Stanford, CA, USA, for half of a year. His research interest includes computational cognitive neuroscience, software engineering, numerical electromagnetics, ultrawide band wireless system, music signal processing, music information retrieval, intelligent agent applications, and electromagnetic scattering analysis.