SCORE-INFORMED SYLLABLE SEGMENTATION FOR JINGJU A CAPPELLA SINGING VOICE WITH MEL-FREQUENCY INTENSITY PROFILES

Rong Gong
Music Technology Group,
Universitat Pompeu Fabra,
Barcelona, Spain
rong.gong@upf.edu

Nicolas Obin
IRCAM, CNRS,
UPMC-Sorbonne Universités,
Paris, France
nicolas.obin@ircam.fr

Georgi Dzhambazov, Xavier Serra
Music Technology Group,
Universitat Pompeu Fabra,
Barcelona, Spain
georgi.dzhambazov@upf.edu
xavier.serra@upf.edu

ABSTRACT

This paper introduces a new unsupervised and score-informed method for the segmentation of singing voice into syllables. The main idea of the proposed method is to detect the syllable onset on a probability density function by incorporating *a priori* syllable duration derived from the score. Firstly, intensity profiles are used to exploit the characteristics of singing voice depending on the Mel-frequency regions. Then, the syllable onset probability density function is obtained by selecting candidates over the intensity profiles and weighted for the purpose of emphasizing the onset regions. Finally, the syllable duration distribution shaped by the score is incorporated into Viterbi decoding to determine the optimal sequence of onset time positions. The proposed method outperforms conventional methods for the segmentation of syllable on a jingju (also known as Peking or Beijing opera) *a cappella* dataset. An analysis is conducted on precision errors to provide direction for future improvement.

1. INTRODUCTION

1.1 Context and motivations

Indication from both psychoacoustic and psycholinguistical research Massaro (1974); Segui et al. (1990); Greenberg (1996) suggests that the syllable is a basic perceptual unit for speech processing in humans. The syllable was recommended as a basic unit of automatic speech recognition as early as 1975 Mermelstein (1975). The syllabic level offers several potential benefits; for one, contrary to the phoneme system which is specific to a language, the syllable is universally defined in terms of acoustic sonority [1] : a syllable segment is fully determined by a maximum of sonority (the vowel nucleus) surrounded by local minima of sonority. Additionally, the syllable is the basic unit of the prosody analysis of speech or singing voice.

In contrast to speech syllables, the duration of singing voice syllables varies enormously and their vowel nucleus may consists of numerous local sonority maxima due to the various ornaments, typically the vibrato - amplitude and frequency modulation, which poses new challenge for the segmentation task. A musical score contains a wide range of prior information, such as the pitch, the onset time and the duration of the note and the syllable, which can be used to guide the segmentation process.

[1] the relative loudness of a speech sound.

1.2 Related work

Most of existing speech syllable segmentation methods can be divided into two categories: unsupervised Mermelstein (1975); Wang & Narayanan (2007); Obin et al. (2013) and supervised Howitt (2000); J. Makashay et al. (2000). In the Mermelstein method Mermelstein (1975), the syllable onset are detected by recursively searching on the convex hull of the loudness function. Wang & Narayanan (2007) have explored the Mel-frequency spectral representations for syllable segmentation. Most recently, the Syll-O-Matic system Obin et al. (2013) exploited the fusion of Mel-frequency intensity profiles and voicing profiles which gives the best segmentation result for the methods of the first category. Supervised methods Howitt (2000); J. Makashay et al. (2000) adopted from Automatic Speech Recognition need the support of a language model and an acoustic model. The latter is learned from a set of audio recordings and their corresponding transcripts, which takes a considerable amount of time to adapt this method from one language to another.

The syllable segmentation of singing voice is still a research gap which needs to be filled. The related subjects are singing voice phonetic segmentation Lin & Jang (2007), lyrics-to-audio alignment Fujihara & Goto (2012); Dzhambazov et al. (2016), and score-to-audio alignment of singing voice Gong et al. (2015). The approaches adopted in these works are mostly supervised, so the problems of the language specificity and the need for a large amount of training data remain.

Various applications such as score-informed source separation Ewert et al. (2014); Miron et al. (2015), tonic identification Sentürk et al. (2013) and score-to-audio alignment Cont (2010) have been proposed in recent years which exploit the availability of a musical score. Dzhambazov et al. (2016) shows that modeling of duration improves the phrase-level lyrics-to-audio alignment accuracy significantly.

This paper introduces a new unsupervised and score-informed method for the segmentation of singing phrase into syllables. We present the definitions of speech syllable and jingju singing voice syllable, and disclose the issues existing in syllable segmentation in section 2. The

approach is explained in section 3. The evaluation and the error analysis are conducted on a jingju *a cappella* singing voice dataset in section 4.

2. WHAT IS A SYLLABLE?

2.1 Definition

The task of automatically detecting the speech syllable is based on the assumption that a syllable is typically vowel centric and neighboring vowels are always separated by consonants Howitt (2000). A precise characterization of the syllable structure can be made in terms of sonorityAssociation (1999), which hypothesizes that syllables contain peaks of sonority that constitute their nuclei and may be surrounded by less sonorous sounds Goldsmith et al. (2011). According to the Sonority Sequencing Principle Dressler (1992), vowels and consonant sounds span a sonority continuum with vowel nuclei being the most sonorous and obstruents being the least, with glides, liquids, and nasals in the middle.

Mandarin is a tonal language and there are in general 4 lexical tones and 1 neutral tone in it. Every character of spoken Mandarin language is pronounced as monosyllable Lin et al. (1993). The jingju singing is the most precisely articulated rendition of the spoken Mandarin language. Although certain special pronunciations in jingju theatrical language differ from their normal Mandarin pronunciations, due to firstly the adoption of certain regional dialects, and secondly the ease or variety in pronunciation and projection of sound, the mono-syllabic pronouncing structure of the standard Mandarin doesn't change Wichmann (1991).

A syllable of jingju singing is composed of three distinct parts in most of the cases: the "head" (tou), the "belly" (fu) and the "tail" (wei). The head consists of the initial consonant or semi-vowel, and the medial vowel if the syllable includes one, which itself is normally not prolonged in its pronunciation except for the one with a medial vowel. The belly follows the head and consists of the central vowel. It is prolonged throughout the major portion of the melodic-phrase for a syllable. The belly is the most sonorous part of a jingju singing syllable and can be analogous to the nuclei of a speech syllable. The tail is composed of the terminal vowel or consonant Wichmann (1991).

The speech syllable only contains one prominent sonority maximum due to its short duration (average < 250 ms and standard deviation < 50 ms for Mandarin Wang (1994)). In contrast, a singing voice syllable may consists of numerous local sonority maxima, of which the reason is either intentional vocal dynamic control for the needs of conveying a better musical expression or unintentional vocal intensity variation as a by-product of the F0 change Titze & Sundberg (1992).

2.2 Issues in syllable segmentation

The issues of speech syllable segmentation has been summarized in Obin et al. (2013). Jingju singing voice brings up two new issues. Firstly, the syllable duration of jingju singing voice varies enormously. According to the statistics of our dataset, the syllable durations range from 70 ms to 21.7 s and its standard deviation is 1.74 s, which makes it impossible to model the durations with one single distribution as it has been done for speech Obin et al. (2013). Secondly, as mentioned in section 2.1, the syllable's central vowel may consists of numerous local sonority maxima, which introduces noisy information for the syllable segmentation.

A priori syllable duration information is often easy to obtain from the score and this is an advantage which can be exploited. The repertoire of jingju includes around 1400 plays Wichmann (1991), among which are still performed and used in teaching are mostly well transcribed into sheet music. Constructing the syllable duration distribution from the score and using it to guide the segmentation process is a feasible way of solving the two new issues mentioned above.

3. APPROACH

The objective of this study is automatically segmenting singing phrases into syllables by incorporating syllable duration information derived from the score into syllable onset detection. Firstly, Mel-frequency intensity profiles are measured over various frequency regions. An observation probability function of syllable onsets is obtained by selecting candidates over the intensity profiles and weighted for the purpose of augmenting its value in the onset regions. Secondly, the *a priori* duration distribution derived by the score is incorporated into the Viterbi decoding to determine the optimal sequence of syllabic onset time positions (Fig.1). The conventional unsupervised speech syllable segmentation method is based on the detection of syllable onset and landmark Obin et al. (2013). However, we focus the issue only on onset detection because the definition of syllable landmark Howitt (2000) doesn't apply to jingju singing voice due to the numerous local sonority maxima within the central vowel.

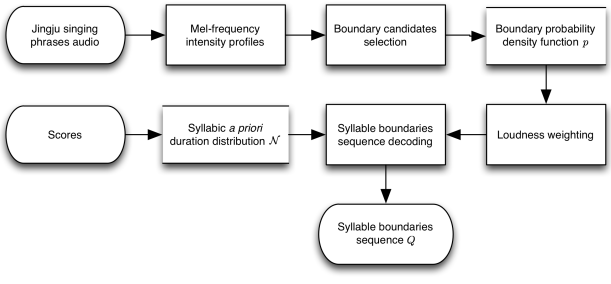

Figure 1: Approach diagram.

3.1 Mel-frequency intensity profiles

A time-frequency representation is used to measure the intensity contained into various frequency regions. For each

frequency region, the specific loudness is measured as:

$$\mathrm{L}_t^{(k)} = \sum_{n=1}^{N^{(k)}} |A(t,n)|^{2\ 0.23} \quad (1)$$

where k denotes the k-th frequency region, $A(t,n)$ the amplitude of the n-th frequency bin at time t in the considered frequency region, and $n = 1$ the start value of the summation index in the k-th frequency region. The specific loudness is related to the sound intensity - the square of the amplitude $A(t,n)$ through a power law with an exponent 0.23 Zwicker & Fastl (2013). In this study, the specific loudness is measured over 40 Mel-frequency bands, with unitary integrated energy in order to enhance the information contained in low-frequency regions relatively to high-frequency regions. The frequency bands are equally spaced on the mel scale Slaney (1998), which approximates the human auditory system's response more closely than the linearly-spaced frequency bands. Then, the specific loudness $\mathrm{L}_t^{(k)}$ is normalized into a probability density function $\mathrm{L}_{t\ \mathrm{norm}}^{(k)}$ so that each intensity profile will be further equally processed (Fig.2-b).

3.2 Onset candidates selection

A syllable has a great probability of starting with a consonant. Stop consonants consist of an interval of complete closure. Because of this, all stops have a period of silence. Affricates consonants have frication portion preceded by stop-like 'silent' portion. Liquids consonants are normally voiced, but have less energy than vowels Johnson (2011). Accordingly, consonants, apart from fricatives and nasals, contain a complete silence or less energy (intensity) than vowels. Additionally, a syllable is usually preceded by some silence or breath frames which also have low intensities in certain frequency regions. These characteristics incite us to conduct the syllabic onset detection on $\mathrm{L}_{t\ \mathrm{norm}}^{(k)}$ by a local maxima-minima detection method Obin et al. (2013), which gives a local minima onset candidate sequence $\mathrm{Onset}^{(k)}$ for each Mel-frequency band k.

The local maxima-minima detection method consists of two steps: in the first step, we conduct a coarse search to find all the maxima and minima positions; in the second step, the positions are selected such that the maxima are required to exceed both neighboring minima by at least a heuristic height threshold (0.01 relative amplitue) and to be separated by at least an heuristic offset threshold (0.025s); otherwise, the maxima together with their neighboring minima are considered as insignificant and suppressed.

This process forms a (K $\times$ T) matrix of onset time-frequency position candidates (Fig.2-c). K and T denote respectively the numbers of the Mel-frequency bands and the time frames. Then, it is summed up into a (1 $\times$ T) probability density function p (Fig.2-d) because the more frequent is observed a time position of a candidate over frequency bands, the more likely is the presence of an onset. However, the exact time position of an onset may differ from one frequency region to the other due to the asynchronism of the information contained in the frequency regions. Thus, a moving average window MA (typically, a 20 ms. window) is employed.

$$p = \mathrm{MA}(\sum_{k=1}^{K} \mathrm{Onset}^{(k)}) \quad (2)$$

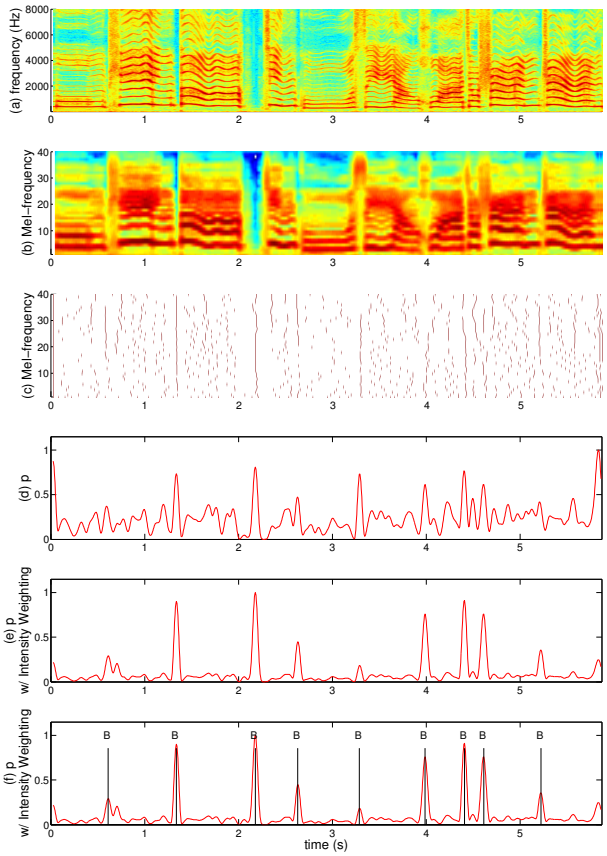

Figure 2: Spectrogram (a), Mel-frequency intensity profiles (b), (K $\times$ T) matrices of onset time/frequency position candidates (c), onset probability density function p (d) and its loudness weighted version (e), determined sequence of syllable onsets (f) for the singing phrase: "Meng ting de jin gu xiang hua jiao sheng zhen."

3.3 Loudness weighting

Certain prominent peak positions can be identified as the syllable onsets on the graph of the probability density function p (Fig. 2-d). However, numerous less prominent peaks can also be found, which do not correspond to the real syllable onsets. This noisy information (less prominent peaks) will eventually degrade the performance of the onset sequence decoding. By observing the graph (Fig.2-d), we clearly see that most of these noisy peaks appear in the vowel regions which usually show a high intensity Dressler (1992). To reduce these noisy peaks, we scale down the high-intensity regions of p by multiplying it by a weighting coefficient.

Inspired by the loudness gating method used in EBU (2016), we employ an absolute gating threshold θ_a, a relative gating threshold θ_r and a sound pressure level storing block SPL_i to detect the high-intensity signal frames.

The intensity of the input singing voice signal is measured frame by frame by the sound pressure level of its RMS amplitude $SPL_{\text{RMS}} = 20\log_{10}(RMS)$. The current frame is detected as high-intensity if its SPL_{RMS} meets both of the following conditions:

$$SPL_{\text{RMS}} \geqslant \theta_a \tag{3}$$

$$SPL_{\text{RMS}} \geqslant \theta_r + \overline{SPL_i} \tag{4}$$

where $\overline{SPL_i}$ is the mean value of the integrated preceding stored SPL_{RMS}, θ_a, θ_r are heuristically selected as -35 dB and -10 dB. Once a frame is detected as high-intensity, its SPL_{RMS} is added to the storing block SPL_i.

A continuous sequence of high-intensity frames is detected as the high-intensity region if it is followed by a continuous sequence of low-intensity frames. The length of the latter should be larger than a threshold θ_l which will be optimized by the grid search method. Finally, the p value in the high-intensity regions is multiplied by a weighting coefficient w_h which will also be optimized later. (Fig.2-e).

3.4 A priori duration distribution

The *a priori* duration distribution $\mathcal{N}(x;\mu_l,\sigma_l^2)$ is modeled by a Gaussian function whose mean μ_l equals to l-th syllable duration of the score and whose standard deviation σ_l is proportional to μ_l: $\sigma_l = \gamma\mu_l$ (Fig.3). The proportionality constant γ will be optimized by the grid search method.

$$\mathcal{N}(x;\mu_l,\sigma_l^2) = \frac{1}{\sqrt{2\pi}\sigma_l}\exp\left(-\frac{(x-\mu_l)^2}{2\sigma_l^2}\right). \tag{5}$$

The relative duration of each note is measured on the quarter note length, so an eighth note has a duration of 0.5. We only keep the relative duration and discard the tempo information of the score. By normalizing the summation of the notes' relative durations to unity, then multiplying it by the duration of the incoming audio recording, we obtain the absolute score duration of the entire singing phrase which is equal to the latter. The note's absolute duration along with its subsequent silence or the summation of the notes' absolute durations (e.g. syllable *gu* in Fig.3) corresponding to l-th syllable is assigned to μ_l. The duration distribution (Eq.5) will be incorporated into Viterbi algorithm as the state transition probability, which holds the highest expectation on its mean value - the syllable duration of the score.

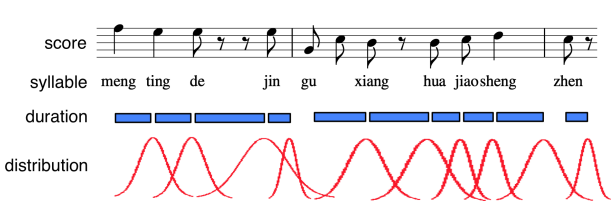

Figure 3: *A priori* relative duration distributions (bottom) of the syllables in the singing phrase: "Meng ting de jin gu xiang hua jiao sheng zhen."

3.5 Decoding of the syllable onsets sequence

A sequence of *a priori* absolute duration $M = \mu_1\mu_2\cdots\mu_L$ is deduced from the score and the length of the incoming audio (section 3.4). To decode the syllable boundaries, we construct an hidden Markov model characterized by the following:

1. The state space is a set of N candidate onset positions $S_1, S_2, \cdots, S_N$ determined by picking the local maxima positions from the probability function p.
2. The state transition probability at decoding time l is defined by *a priori* duration distribution $\mathcal{N}(d_{ij};\mu_l,\sigma_l^2)$, where d_{ij} is the time distance between states S_i and S_j $(j > i)$. The overall decoding time is equal to the total syllable number L written in the score.
3. The observation probability for the state S_j is represented by its corresponding value in the onset detection function p, which is denoted as p_j.

As we assume the onset of the current syllable is also the offset of the previous syllable, the problem is translated into finding the best offset position state sequence $Q = q_1q_2\cdots q_L$, for the given *a priori* duration sequence M, where q_i denotes the offset of the ith decoding syllable or the onset of the $i+1$th decoding syllable. q_0 and q_L are fixed as S_1 and S_N as we expect that the onset of the first syllable is located in the beginning of the incoming audio and the offset of the last syllable is located in the ending of the audio. One can fulfill this assumption by truncating the silences at both ends of the incoming audio. According to the logarithmic form Viterbi algorithm Rabiner (1989), we define

$$\delta_l(i) = \max_{q_1,q_2,\cdots,q_l} \log P[q_1q_2\cdots q_l,\ \mu_1\mu_2\cdots\mu_l]$$

the initially step

$$\delta_1(i) = \log(\mathcal{N}(d_{1i};\mu_1,\sigma_1^2)) + \log(p_i)$$
$$\psi_1(i) = S_1$$

the recursion step

$$\delta_l(j) = \max_{1\leqslant i<j}[\delta_{l-1}(i) + \log(\mathcal{N}(d_{ij};\mu_l,\sigma_l^2))] + \log(p_j)$$
$$\psi_l(j) = \arg\max_{1\leqslant i<j}[\delta_{l-1}(i) + \log(\mathcal{N}(d_{ij};\mu_l,\sigma_l^2))]$$

and termination step

$$\log P^* = \max_{1\leqslant i<N}[\delta_{L-1}(i) + \log(\mathcal{N}(d_{iN};\mu_L,\sigma_L^2))]$$
$$q_L^* = \arg\max_{1\leqslant i<N}[\delta_{L-1}(i) + \log(\mathcal{N}(d_{iN};\mu_L,\sigma_L^2))]$$

Finally, the best offset position state sequence Q is obtained by the backtracking step (Fig.2-f).

4. EVALUATION

4.1 Dataset

The *a cappella* singing dataset [2] used for this study comes from MTG and C4DM Black et al. (2014) and focuses

[2] http://doi.org/10.5281/zenodo.345490

on two most important jingju role-types Repetto & Serra (2014): *dan* (female) and *laosheng* (old man). It contains 39 interpretations of 31 unique arias sung by 11 jingju singers. The syllable onset ground truth is manually annotated in Praat Boersma (2001), which represents 298 phrases and 2672 syllables (including padding written - characters Wichmann (1991)). The average syllable duration is 1.1s and the standard deviation is 1.74s. The syllable duration dataset is manually transcribed from sheet music.

The whole dataset is randomly split into 2 parts with the constraint that each part is selected without role-type bias and contains almost an equal number of onsets. One of them is reserved as the development set for the purpose of parameter optimization. Another part is used as the test set to evaluate the syllable segmentation algorithms.

4.2 Evaluation metrics

The objective of the syllable segmentation for singing phrases is to determine the time positions of syllable boundaries. The evaluation consisted in the comparison of the determined syllable onsets and offsets to the reference one. We use the same metric for the speech syllable segmentation evaluation: recall, precision and F-measure Obin et al. (2013). The definition of a correct segmented syllable is borrowed from the note transcription evaluation Molina et al. (2014): for the syllable onset, we choose a evaluation tolerance $\pm\tau$ ms. For the offset, which is also the onset of the subsequent syllable, $\pm 20\%$ of the reference syllable's duration or $\pm\tau$ ms, whichever is larger, is chosen as the tolerance. If both the onset and the offset of a syllable lie within the tolerance of their reference counterparts, we say it's correctly segmented. As there is no standard tolerance previously defined for the evaluation of singing voice syllable onset detection, and the tolerance for the evaluation of speech syllable onset detection is too strict because the average duration of speech syllable (200 ms) is much shorter than that of singing voice syllable (1.1 s), we decide to report the evaluation results for multiple tolerances, $\tau = [0.05, 0.1, 0.15, 0.2, 0.25, 0.3]$ (second).

4.3 Parameters optimization

The parameters which need to be optimized are: the length threshold θ_l of low-intensity regions, the weighting coefficient w_h for p in high-intensity regions in section 3.3; the proportionality constant γ in section 3.4. The syllable segmentation accuracy can be reported by sweeping these parameters on the development set. Table 1 lists the search bounds and the optimal results.

Table 1: Search bounds, optimal results (OR) of the optimization process for each parameter.

Parameters	Search bounds	OR
θ_l (s)	$[0.01, 0.1]$ with step 0.01	0.02
w_h	$[0.1, 1]$ with step 0.1	0.2
γ	$[0.05, 1]$ with step 0.05	0.35

5. RESULTS AND DISCUSSION

5.1 Syll-O-Matic syllable segmentation

The evaluation includes the speech syllable segmentation method Syll-O-Matic Obin et al. (2013) for a comparison with the unsupervised method. This method performs the same Mel-frequency intensity profiles and onset candidate selection steps introduced in this paper. It detects both the speech syllable onsets and the vowel landmarks. We will not report its landmark detection performance because the definition of syllable landmark - the only and most sonorous peak with the central vowel, doesn't apply to most of jingju singing syllables due to the existence of numerous local sonority maxima within the central vowel.

Our proposed method can be seen as an adaption of the original Syll-O-Matic method to the singing voice, which introduces the loudness weighting to attenuate the noisy peaks in the onset probability density function p, and *a priori* syllable duration distribution to take account into the duration information provided by the score, whereas only a fixed mean (1.1s, the average syllable duration of our dataset) normal distribution has been used in the Viterbi decoding process of the original Syll-O-Matic method.

The Syll-O-Matic method performs bad on our dataset (Fig.4) and causes a low F-measure. There are at least three reasons for this bad performance. First, its Viterbi decoding algorithm doesn't restrict the overall decoding time, so any peak position in p is able to be decoded as a syllable onset if it happens to have a high duration probability. Second, the numerous sonorous peaks in p act as the noisy information, which introduces many insertions. Third, the duration distribution used in Syll-O-Matic is mean-fixed, which doesn't conform to the fact of the variable syllable duration of the jingju singing voice.

5.2 HMM-based lyrics-to-audio alignment

The evaluation also includes a HMM-based lyrics-to-audio alignment method Dzhambazov et al. (2016) for a comparison with the supervised method. The HMM-based system extends Viterbi decoding to handle the duration of states. For each of 40 Mandarin phonemes and diphthongs, a one-state HMM is trained from a 67 minutes corpus of *a cappella* female jingju singing voice. This corpus is different from the one mentioned in section 4.1 in terms of the singer and the repertoire. For each state a 40-mixtures of Gaussian distribution are fitted on the MFCCs feature vector. The HMM-based system outputs the decoded syllable onset positions.

5.3 Proposed method

Even without the loudness weighting step, the proposed method (Score-informed) outperforms all the compared methods. Additionally, the loudness weighting (Score-informed + Loudness weighting) successfully improves the segmentation performance due to the reduction of the noisy information in the high-intensity region of the probability density function. Compared to supervised methods (e.g. HMM-based), the results are encouraging for the use of

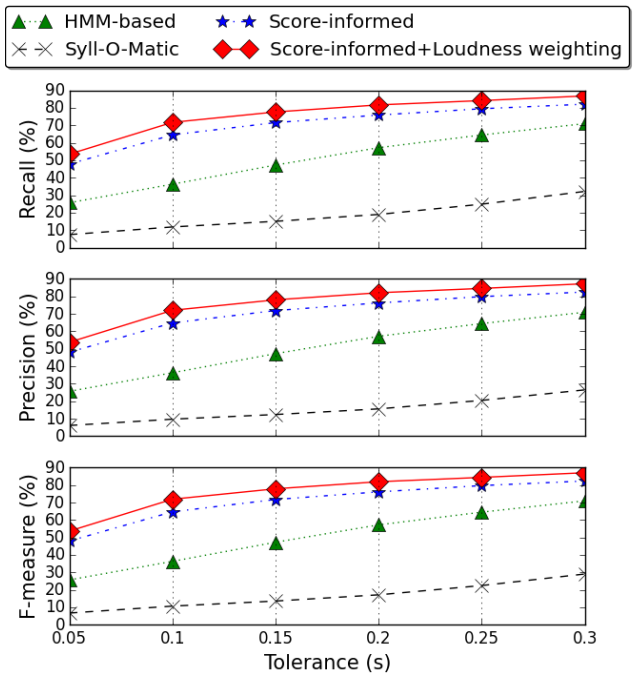

Figure 4: Recall, precision and F-measure results of the syllable segmentation evaluation. The three metrics do not look so different because the number of the segmented syllable and the number of the ground truth syllable are almost the same.

unsupervised score-informed method for singing voice syllable segmentation, which avoids the problems of the language specificity and the need for a large amount of training data.

5.4 Error analysis

We conduct error analysis to make clear the causes of segmentation errors of our proposed method, and also to provide direction for future improvement. First, only the errors occurred in the segmented syllables (precision errors) will be analyzed because the number of the segmented syllable (1329) and the number of the ground truth syllable (1334) are almost equal for the result of the proposed method, which means almost all the syllables in the ground truth are segmented. Second, only the errors out of 0.3s tolerance will be analyzed because the causes of these errors are straightforward to be identified from observing the segmentation plots. 169 syllables are mistakenly segmented out of 1329 evaluated syllables (Table 2).

Table 2: Performance of the proposed method with 0.3s tolerance.

Method	Recall(%)	Precision(%)	F-measure(%)
Score-informed+ Loudness weighting	86.83	87.28	87.05

Four types of error have been identified (Table 3) by observing the plots of detected syllable onsets compared to ground truth onsets:

- Redundant intensity minima: errors caused by redundant intensity minima (redundant peaks) in the onset probability density function p. Silence or large intensity change within the syllable are the main causes of this error type.
- Missed intensity minima: errors caused by missed intensity minima (missed peaks) in p. Long silence followed by the syllable is the main cause of this error type, which usually happens in *laosheng* (old man) singing.
- Ambiguous syllable transitions: errors caused by ambiguous syllable transitions, such as transitions from vowel to vowel or to semi-vowel, from semi-vowel to semi-vowel. This cause has also been reported in the unsupervised speech syllable segmentation research Obin et al. (2013).
- Score and singing incoherent: errors caused by large contrast between syllable duration in score and that in real practice.

Table 3: Error analysis for the result of the proposed method with 0.3s tolerance.

Types of error	Num. errors (frequency %)
Redundant intensity minima	92 (54.3)
Missed intensity minima	34 (20.3)
Ambiguous syllable transitions	32 (18.8)
Score and singing incoherent	11 (6.6)
Sum	169 (100)

The reason for the first three types of error is that our proposed method only uses intensity-related feature and technique (Mel-frequency intensity profiles and loudness weighting) which are not knowledgeable in the phonetic context of the signal frames. By applying phonetic features to shape the peaks of the onset probability density function in the future, for example, comparing the phonetic content before and after the silence, we may reduce these types of error. For the last type of error - Score and singing incoherent, the effort should be put in improving the onset decoding method. Using different duration distribution function, such as gamma distribution, and variable decoding time can be the possible way to tackle this type of error.

6. CONCLUSION

In this paper, we present the definition of jingju singing voice syllable and disclose the new issues arose by this singing form. A new method is then introduced for the segmentation of singing voice into syllables. The main idea of the proposed method is to detect the syllable onset on a syllable onset probability density function by incorporating the syllable duration information of the score into the decoding process. The main contribution of this work is twofold: First, the loudness weighting is applied on the high-intensity regions of the onset probability density function, which reduced the noisy sonorous peaks and augmented the segmentation accuracy. Second, the syllable duration distribution is incorporated into the decod-

ing process of the optimal syllable onset sequence to make use of the *a priori* information of the score. The proposed method outperforms conventional methods for the syllable segmentation of singing voice phrases, and provides a promising paradigm for the segmentation of singing voice into syllables.

7. ACKNOWLEDGEMENTS

This work is partly supported by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grand agreement 267583).

8. REFERENCES

Association, I. P. (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

Black, D. A. A., Li, M., & Tian, M. (2014). Automatic Identification of Emotional Cues in Chinese Opera Singing. In *ICMPC-2014*, Seoul, South Korea.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Cont, A. (2010). A coupled duration-focused architecture for realtime music to score alignment. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *32*(6), 974–987.

Dressler, W. (1992). *Phonologica 1988*. Cambridge University Press.

Dzhambazov, G., Yang, Y., Repetto, R. C., & Serra, X. (2016). Automatic alignment of long syllables in a cappella beijing opera. In *FMA-2016*, Dublin, Ireland.

EBU (2016). 'EBU Mode' metering to supplement Loudness normalisation. Recommendation Tech 3341-2016, Geneva. Version 3.0.

Ewert, S., Pardo, B., Muller, M., & Plumbley, M. (2014). Score-Informed Source Separation for Musical Audio Recordings: An overview. *IEEE Signal Processing Magazine*, *31*(3), 116–124.

Fujihara, H. & Goto, M. (2012). Lyrics-to-audio alignment and its application. *Dagstuhl Follow-Ups*, *3*.

Goldsmith, J. A., Riggle, J., & Yu, A. C. L. (2011). *The Handbook of Phonological Theory*. John Wiley & Sons.

Gong, R., Cuvillier, P., Obin, N., & Cont, A. (2015). Real-Time Audio-to-Score Alignment of Singing Voice Based on Melody and Lyric Information. In *Inter Speech-2015*, Dresden, Germany.

Greenberg, S. (1996). *Understanding Speech Understanding: Towards A Unified Theory Of Speech Perception*.

Howitt, A. W. (2000). *Automatic Syllable Detection for Vowel Landmarks*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

J. Makashay, M., W. Wightman, C., K. Syrdal, A., & Conkie, A. (2000). Perceptual Evaluation of Automatic Segmentation in Text-to-Speech Synthesis. In *ICSLP 2000*, Beijing.

Johnson, K. (2011). *Acoustic and Auditory Phonetics*. John Wiley & Sons.

Lin, C.-H., Lee, L.-s., & Ting, P.-Y. (1993). A new framework for recognition of Mandarin syllables with tones using sub-syllabic units. In *ICASSP-1993*, volume 2.

Lin, C.-Y. & Jang, J.-S. (2007). Automatic Phonetic Segmentation by Score Predictive Model for the Corpora of Mandarin Singing Voices. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7).

Massaro, D. W. (1974). Perceptual units in speech recognition. *Journal of Experimental Psychology*, 199–208.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *The Journal of the Acoustical Society of America*, *58*(4), 880–883.

Miron, M., Carabias-Orti, J. J., & Janer, J. (2015). Improving Score-Informed Source Separation for Classical Music through Note Refinement. In *ISMIR-2015*, Malaga.

Molina, E., Barbancho, A. M., Tardón, L. J., & Barbancho, I. (2014). Evaluation Framework for Automatic Singing Transcription. In *ISMIR-2014*, Taipei, Taiwan.

Obin, N., Lamare, F., & Roebel, A. (2013). Syll-O-Matic: An adaptive time-frequency representation for the automatic segmentation of speech into syllables. In *ICASSP-2013*.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Repetto, R. C. & Serra, X. (2014). Creating a Corpus of Jingju (Beijing Opera) Music and Possibilities for Melodic Analysis. In *ISMIR-2014*, Taipei, Taiwan.

Segui, J., Dupoux, E., & Mehler, J. (1990). *Cognitive Models of Speech Processing*. Cambridge, MA, USA: MIT Press.

Sentürk, S., Gulati, S., & Serra, X. (2013). Score informed tonic identification for makam music of turkey. In *ISMIR-2013*.

Slaney, M. (1998). Auditory toolbox. Technical Report 1998-010, Interval Research Corporation.

Titze, I. R. & Sundberg, J. (1992). Vocal intensity in speakers and singers. *The Journal of the Acoustical Society of America*, *91*(5), 2936–2946.

Wang, D. & Narayanan, S. (2007). Robust Speech Rate Estimation for Spontaneous Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(8), 2190–2201.

Wang, J. (1994). Syllable duration in Mandarin. In *the Fifth International Conference on Speech Science and Technology*.

Wichmann, E. (1991). *Listening to Theatre: The Aural Dimension of Beijing Opera*. University of Hawaii Press.

Zwicker, E. & Fastl, H. (2013). *Psychoacoustics: Facts and models*, volume 22. Springer Science & Business Media.