# Automatic alignment of long syllables in acapella Beijing opera

Anonymous Author, Anonymous Author, Anonymous Author, Anonymous Author

December 22, 2015

Anonymous Institute

## 1 Introduction

The task of automatic lyrics-to-audio alignment (LAA) is to find in an automatic way a match between two musical aspects of a composition: the singing voice and the corresponding lyrics. Tracking lyrics may be used to automatically outline the structure of the audio recording (inferred from lyric), which can be beneficial for musicologists or singing students.

With a few exceptions (Wong et al., 2007), to-date computational research on LAA has been focused mainly on western popular music (Fujihara and Goto, 2012). Unlike popular music, in traditional genres, duration of vocals might vary substantially from one syllable to the next. Actors of Beijing opera, in particular, tend to prolong particular vowels to pertain to the poetic rhythm of the story, being sung. More specifically a lyrics line is usually divided into 3 units - *dou*, each consisting of 2 to 5 written characters (Wichmann, 1991, Chapter III). To outline a *dou*, an actor has the option to sustain the vocal of its final syllable [1] (sometimes performing ornamentation/vibrato), resulting in a substantially longer vowel.

In this work we model explicitly phoneme durations by a probabilistic method capturing lyrics principles of Beijing opera.

## 2 Method Overview

First an audio recording is manually divided into lines as indicated in the lyrics script of the aria. Then each line is represented as a sequence of phoneme models, tied into a probabilistic duration-aware model (anonymous, -). Each phoneme model yields an observation probability for singing audio, based on its timbral features. Finally, accuracy (as defined in Fujihara et al. 2011, Fig. 9) is evaluated at beginning and ending timestamps of each Mandarin syllable.

### Durational model

First, all probable *key syllables* with regard to the *dou* grouping patterns (e.g. 3 + 3 + 4 syllables in 10-syllable line) are assigned longer reference durations. Second, the reference duration of each syllable is divided among its constituent phonemes, considering the initial-final division of Mandarin syllables (Duanmu, 2000), which forms a sequence of phoneme reference durations $R_i$

A metrical pattern (*banshi*) in an aria can be changed up to several times (Wichmann, 1991). We observed in the dataset that usually the final *key syllable* of the last line in a *banshi* is prolonged additionally. Thus we defined $R_i$ of final *key syllables* considering *banshi* changes.

We propose a duration-aware hidden Markov model (DHMM). It models the duration of a phoneme as a normal distribution, centered at $R_i$ with a standard deviation $d$. A proper $d$ assures that a phoneme sung longer or shorter than the expected $R_i$ can be adequately handled.

[1] We use the term *syllable* as equivalent to one written character. Final syllables in this work will be referred as *key syllables*

| | oracle | baseline | DHMM |
|---|---|---|---|
| overall | 79 | 56.8 | 66.14 |
| median per lyrics line | 98 | 75.2 | 82.3 |

Table 1: Comparison of total oracle, baseline and DHMM alignment. Accuracy is reported as accumulate correct duration over accumulate total duration over all lines from a set of arias.

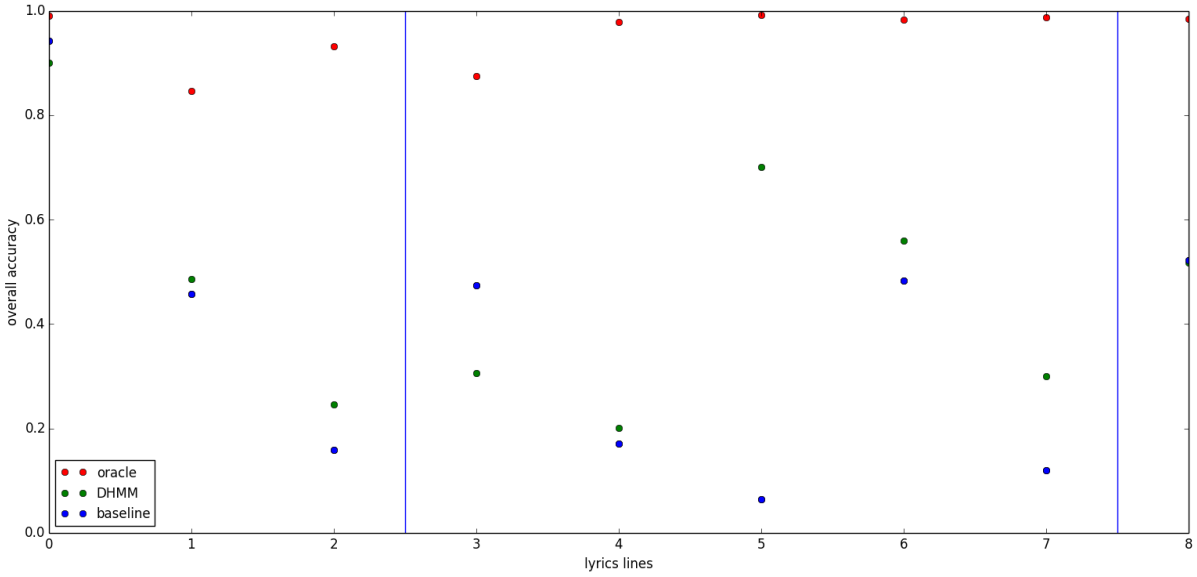


Figure 1: Comparison of oracle, baseline and DHMM results on one aria. Each point represents one lyrics line, vertical lines represent *banshi* changes

# 3 Experimental Setup

Our dataset consists of excerpts from 15 arias with acapella female voice of total duration of 67 minutes. A line has an average duration of 18.3 seconds and 9 syllables [2].

To define a glass ceiling accuracy an alignment was performed considering phoneme annotations as an oracle for acoustic features. We found that accuracy is overall close to 100%[3], which means that the model is generally feasible to hand the highly-varying vocal durations of Beijing singing.As a baseline we considered same models but without duration modeling. For both baseline and DHMM, to assure good generalization of results, evaluation was done by cross validation with 3-equally sized folds.

## Results

Table 1 shows how the proposed duration model outperforms the baseline alignment.

In figure 1 is depicted the model's accuracy for an aria with very long *key syllables*, for which the baseline model performs poor, whereas the DHMM aligns decently [4]. One can see the advantage of the proposed model for example for lyrics line 5. Looking at oracle, one can conclude that reaching closer to it can be achieved in the future by designing features which capture phoneme identities in a more robust way.

[2]The dataset has been especially annotated for this study. Annotations are made available on http://anonymous

[3]We have utilized fixed vowel and consonant standard deviations $d_v$ and $d_c$. Most optimal results were obtained at the values: consonant duration = 0.3, $d_c = 0.7$; $d_v = 3.0$

[4]Please find attached a video recording demonstrating the alignment accuracy

## References

anonymous. suppressed for anonymity. -.

S. Duanmu. *The Phonology of Standard Chinese*. Clarendon Studies in Criminology. Oxford University Press, 2000. ISBN 9780198299875. URL `https://books.google.es/books?id=oDZkAAAAMAAJ`.

H. Fujihara and M. Goto. Lyrics-to-audio alignment and its application. *Multimodal Music Processing*, 3:23–36, 2012.

H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno. Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics. *Selected Topics in Signal Processing, IEEE Journal of*, 5(6):1252–1261, 2011.

E. Wichmann. *Listening to theatre: the aural dimension of Beijing opera*. University of Hawaii Press, 1991.

C. H. Wong, W. M. Szeto, and K. H. Wong. Automatic lyrics alignment for cantonese popular music. *Multimedia Systems*, 12(4-5):307–323, 2007.