Automatic alignment of long syllables in acapella Beijing opera

Anonymous Author, Anonymous Author, Anonymous Author, Anonymous Author

March 29, 2016

Anonymous Institute

1 Introduction

Compared to speech, singing voice has some substantially different characteristics In particular, unlike speech, for singing voice, durations of vocals have on average somewhat higher variation Kruspe (2014). Singing coming from traditional music sets an additional challenge: prolonging vocals might be utilized as a way to experess an emphasis, which makes particular syllables lasting substantially longer compared to an average duration.

A further confinement of current work is imposed by the speech models: Typically phonemes are trained on a large speech corpus and later adapted to singing voice.It has been recently shown that this way might be suboptimal.

In this work we propose how to model explicitly phoneme durations for lyrics-to-audio alignment by means of a duration-aware probabilistic model. We show that incorporating some prior knowledge based on musical principles for modeling syllable durations brings improvemento over a baseline HMM-based model. The approach is tested on material from Beijing opera, for which particularly long vowel durations are characteristic.

1.1 Background on Jingju music principles

Actors of Beijing opera tend to prolong particular vowels to pertain to the poetic rhythm of the story, being sung. More specifically a lyrics sentence 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. We use the term *syllable* as equivalent to one written character. Final syllables in this work will be referred as *key syllables* (sometimes performing ornamentation/vibrato), resulting in a substantially longer vowel.

Lyric are divided into couplets: each couplet has two lyrics sentences and sentences are related to structure. MORE: what repeats?

A metrical pattern (*banshi*) in an aria can be changed up to several times (Wichmann, 1991). Banshi indicates tempo as well: usually a banshi changes thoughout an aria to gradually increase from slow to faster tempo. MORE

2 Related Work

Current Lyrics-to-audio alignment is mostly based on an speech-adopted approach: phonetic recognizer that models phonemes with hidden Markov models (HMMs) [Fujihara, Mesaros]. Describe Fujihara.

HMMs have been originally developed to model speech phonemes for applications like text-to-speech alignment. HMMs have the drawback that in general they are not capable to represent well vowels with long and/or highly-variable durations, because the waiting time in a state in traditional HMMs cannot be unlimitedly long. Duration can be modeled by duration-explicit hidden Markov models (DHMM) (a.k.a. hidden semi-Markov models). In DHMMs the underlying process is allowed to be a semi- Markov chain with variable duration of each state Yu (2010). DHMMs have been shown to be successful for modeling chord durations in automatic chord recognition [7].

Very few works have trained models directly from singing voice: YILE: Anna Kruspe, Hansen, maybe lyrically

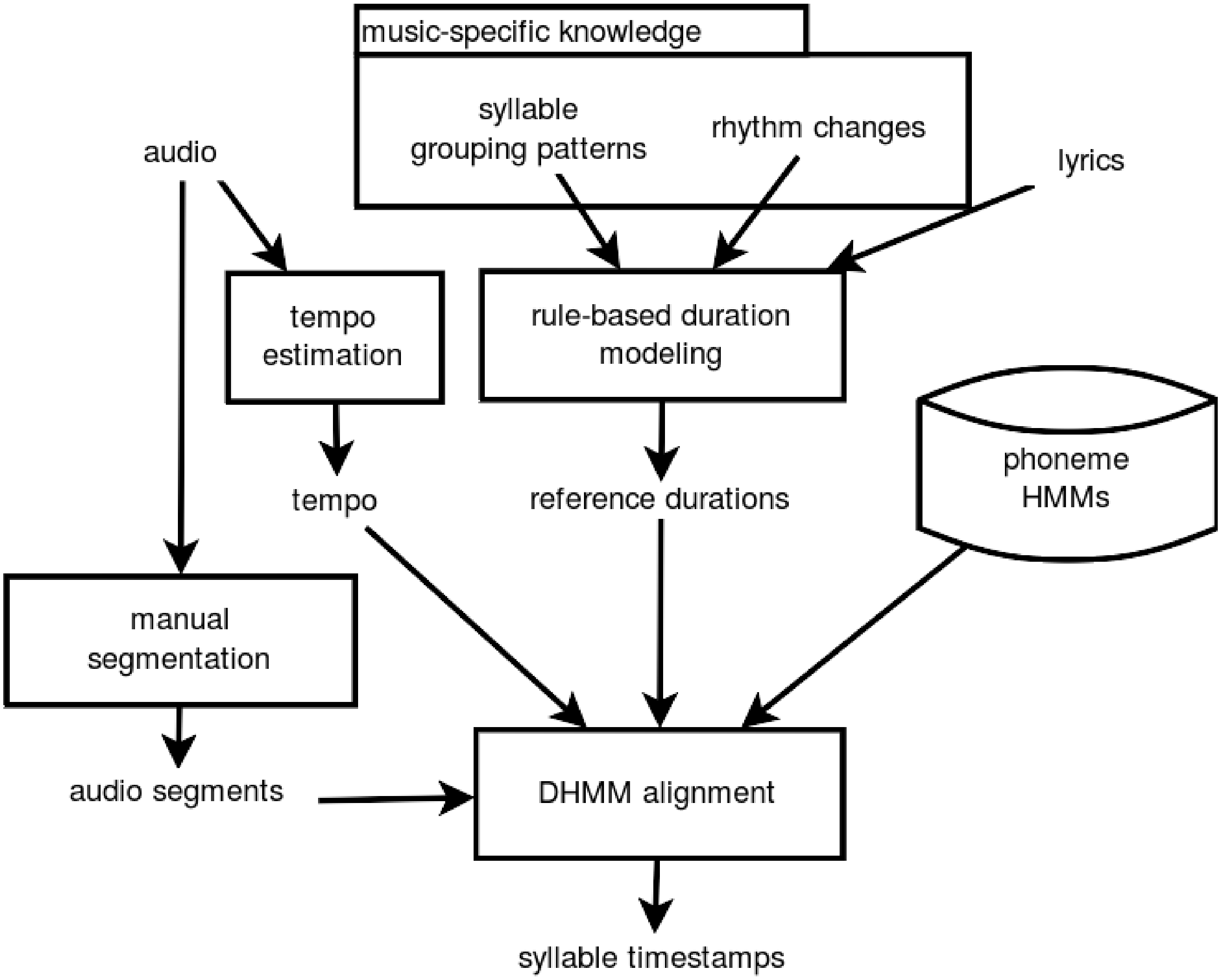

Figure 1: Approach Overview

Not a lot of work on non-western or traditional music: With a few exceptions (Wong et al., 2007), to-date computational research on lyrics-to-audio alignment has been focused mainly on western popular music (Fujihara and Goto, 2012).

3 Method Overview

A general overview of the proposed approach is presented in Figure 1. For brevity in the rest of the paper the proposed alignment scheme will be refered to as DHMM.

First an audio recording is manually divided into sentences as indicated in the lyrics script of the aria, whereby instrumental-only sections are discarded. All further steps are performed on each audio segment. If we had used automatic segmentation instead, potential erroneous lyrics and features could have biased the comparison of a baseline system and DHMM. As we focus on evaluating the effect of DHMM, manual segmentation is preferred. Then each lyrics sentence is expanded to a sequence of phoneme models. Each phoneme model yields an observation probability for singing audio, based on its MFCC features, whereby reference syllable durations guide the decoding process.

3.1 Rule-based duration modeling

In DHMM each state duration in turn can be guided by a custom statistical distribution, with mean centered at a reference state duration. Here we desribe how the reference state durations are derived based on concepts, specific for Beijing opera.

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. Additionally, 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.

Then, to form a sequence of phoneme reference durations R_i, reference durations of syllables are divided among their constituent phonemes. For this purpose we consider the initial-middle-final division of syllables in Mandarin (Duanmu, 2000). TODO: describe how it works...

Tempo is used to scale.

3.2 DHMM alignment

We have adopted the idea of [7] not to explicitly add states that represent durations to the state space, but instead to modify the decoding stage. Thus our model reduces to a standard HMM, whereby preferred decoding is Viterbi with forced alignment. In what follows we desribe how in the Viterbi decoding stage we maximizes over durations.

DHMM 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. Consonant standard deviation d_c is fixed and vowel consonant duration.

3.3 Phoneme models

Recent work shows that recognition of phonemes trained on singing voice can yield better performance compared to the traditional way of training on speech [hansen, kruspe]. Therefore we have trained directly on singing voice....YILE: MORE

4 Dataset

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

Alignment is evaluated in terms of alignment accuracy as the percentage of duration of correctly aligned regions from total audio duration (see Fujihara et al. (2011, figure 9) for an example). In the context of this work a value of 100 means perfect matching of all Mandarin syllable boundaries from evaluated audio.

5 Experiments

5.1 Experiment 1: Oracle duraitons

To define a glass ceiling accuracy, alignment was performed considering phoneme annotations as an oracle for acoustic features. Looking at phoneme annotations, we set the probability of a phoneme to 1 during its time interval and 0 otherwise. We found that the median accuracy per a line of lyrics is close to 100%, which means that the model is generally capable of handling the highly-varying vocal durations of Beijing singing. We have utilized fixed vowel and consonant standard deviations d_v and d_c. Most optimal results were obtained at the values: $d_c = 0.7$; $d_v = 3.0$ As a baseline we considered same models but without duration modeling.

5.2 Experiment 2: comparison with baseline

For both baseline and DHMM, to assure good generalization of results, evaluation is done by cross validation on 3 folds with approximately equal number of syllables: Phoneme models are trained on a subset of the dataset using the phoneme-level annotations and evaluated on a hold-out subset.

	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.

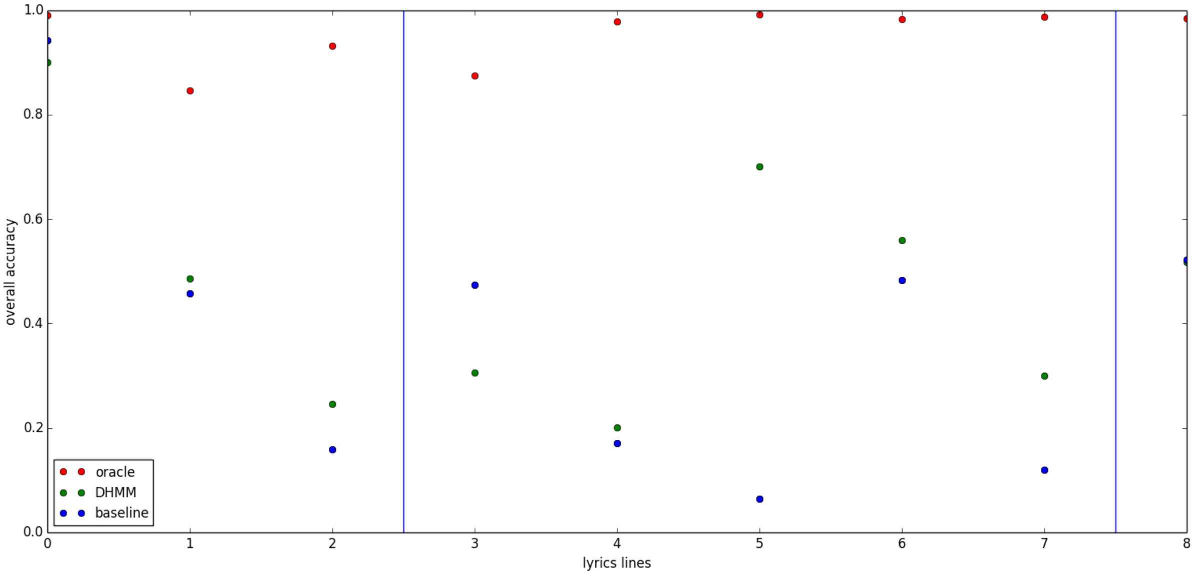

Figure 2: Comparison of oracle, baseline and DHMM results on one aria. Each point represents one lyrics line, vertical lines represent *banshi* changes

5.3 Experiment 3: score reference durations

TODO

5.4 Results

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

In figure 2 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 [1]. 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.

References

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.

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

A. M. Kruspe. Keyword spotting in a-capella singing. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, pages 271–276, Taipei, Taiwan, 2014.

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.

S.-Z. Yu. Hidden semi-markov models. *Artificial Intelligence*, 174(2):215–243, 2010.