AUTOMATIC ALIGNMENT OF LONG SYLLABLES IN ACAPELLA BEIJING OPERA

Georgi Dzhambazov, Yile Yang, Rafael Caro Repetto, Xavier Serra
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
{georgi.dzhambazov, yile.yang, rafael.caro, xavier.serra}@upf.edu

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]. Traditional music poses an additional challenge: as a way to express an emphasis, singers might prolong particular syllables to a duration substantially longer compared to other syllables.

A further confinement of current work on modeling lyrics is imposed by the necessity of a large speech corpus, on which phoneme models are typically trained [Fujihara & Goto, 2012]. Such corpora might not be present for any language or not freely available. Recent work has shown that training on singing voice might be a viable alternative [Hansen & Fraunhofer, 2012].

In this study we propose a duration-explicit hidden Markov model (DHMM) for automatic lyrics-to-audio alignment, trained on singing voice [1]. We show that incorporating some prior expectation of syllable durations, based on musical principles, brings improvement over a baseline method. The approach is tested on material from Beijing opera, for which are characteristic particularly long syllable durations.

2. BACKGROUND ON BEIJING OPERA MUSIC PRINCIPLES

Lyric in Beijing opera are commonly divided into couplets: each couplet has two lyrics sentences. A sentence is usually divided into 3 units - *dou*, each consisting of 2 to 5 written characters [2] [Wichmann, 1991, Chapter III]. Actors of Beijing opera tend to prolong particular syllables: to outline a *dou* or to pertain to the poetic rhythm of the story, being sung, an actor has the option to sustain the vocal of the *dou*'s final syllable. In this work we will refer to the final syllable of a *dou* as *key syllable.*

Each aria can have one or more metrical pattern (*banshi*): it indicates the mood of singing and is correlated to tempo[Wichmann, 1991]. Usually an aria starts with a slow *banshi* which gradually changes a couple of time to faster one, to express a more and more intense mood.

[1] For brevity in the rest of the paper the proposed alignment scheme will be referred to as DHMM.

[2] We use the term *syllable* as equivalent to one written character.

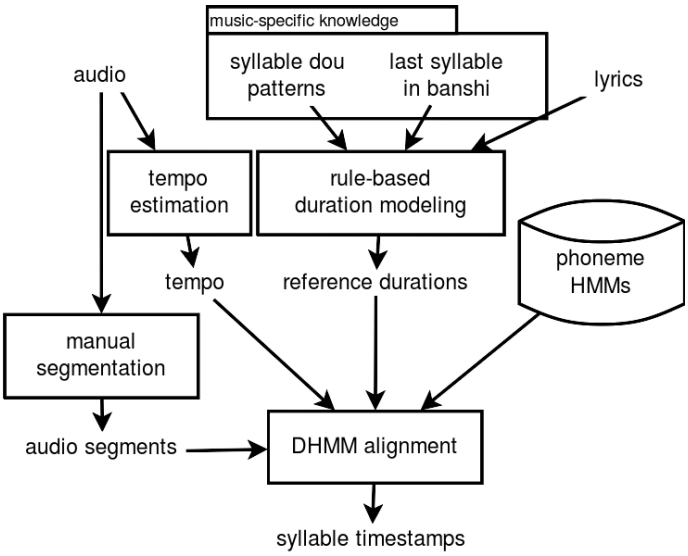

Figure 1: Approach Overview

3. METHOD OVERVIEW

A general overview of the proposed approach is presented in Figure 1. 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 Mel Frequency Cepstral Coefficients (MFCC), whereby reference syllable durations guide the decoding process.

3.1 Rule-based duration modeling

The idea of the DHMM is that the actual duration of a phoneme can be seen as being generated by a statistical distribution with mean centered at an expected reference duration. The reference durations can be assigned using any prior knowledge as, for example, textual structure [Wang et al., 2004]. In this work they are derived as follows:

Firstly, all *key syllables* in a *dou* grouping patterns are assigned longer reference durations, while the rest get equal

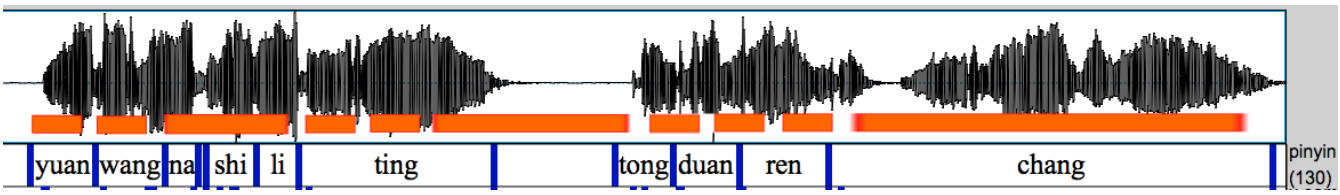

Figure 2: Example the last sentence of a banshi with 10 syllables. Actual syllable durations in pinyin (below), whereas reference durations in orange (above). According to *dou* groups 3rd, 6th and last syllable are prolonged

durations. Additionally, we observed in the dataset that usually the last *key syllable* of the last sentence in a *banshi* is prolonged additionally. Thus we lengthened additionally the reference syllable duration of these last *key syllables* (see Figure for example).

Then, to form a sequence of phoneme reference durations R_i, the reference durations of syllables are divided among their constituent phonemes, according to the initial-middle-final division of syllables in Mandarin [Duanmu, 2000]. The reference durations are linearly scaled to a reference number of frames according to automatically detected tempo.

3.2 DHMM alignment

We have adopted the idea of Chen et al. [2012] not to represent durations by explicitly additional counter states, but instead to modify the Viterbi forced alignment stage: The duration of a phoneme is modeled as a normal distribution, centered at R_i. More details can be found in Dzhambazov & Serra [2015].We opted for a global standard deviation d_c for consonants and d_v for vowels. Proper values for d_c and d_v assure that a phoneme sung longer or shorter than the expected R_i can be adequately handled.

3.3 Phoneme models

The first 13 MFCCs and their Δ and $\Delta\Delta$ are extracted from 25ms audio frames with the hop size of 10ms from the a-cappella singing. The extracted features are then fed to fit a GMM with 40 components for each phoneme

3.3.1 Dataset

The dataset consists of excerpts from 15 arias with acapella female voice of total duration of 67 minutes. The a-cappella singing is acquired by subtracting the instrumental accompany from the original mix. [3] A lyrics sentence has an average duration of 18.3 seconds and 10.7 syllables. Each aria is annotated on the phoneme level by native Chinese speakers and Beijing opera musicologist. The phoneme set has 29 phonemes and is derived from Chinese Pinyin, and represented using X-Sampa standard [4]. To assure enough training data for each phoneme, certain phonemes are combined into classes, based on the perception of the singing voice.

[3] The resulting a-cappella singing are perceived as clean and clear by human ears.

[4] Annotations are made available on http://anonymous

	oracle	baseline	DHMM
overall	94	56.56	68.9
median per sentence	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 sentences from a set of arias.

4. EXPERIMENTS

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.

4.1 Experiment 1: oracle durations

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 sentence of lyrics is close to 100%, which means that the model is generally capable of handling the highly-varying vocal durations of Beijing singing. Most optimal results were obtained with $d_c = 0.7$; $d_v = 3.0$

4.2 Experiment 2: comparison with baseline

As a baseline we employed a standard HMM with Viterbi decoding with the *htk* toolkit Young [1993]. For both HMM 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 10 of the arias using the phoneme-level annotations and evaluated on a 5-aria hold-out subset.Table 1 shows that the proposed duration model outperforms the baseline alignment. 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.

5. REFERENCES

Chen, R., Shen, W., Srinivasamurthy, A., & Chordia, P. (2012). Chord recognition using duration-explicit hidden markov models. In *ISMIR*, (pp. 445–450). Citeseer.

Duanmu, S. (2000). *The Phonology of Standard Chinese*. Clarendon Studies in Criminology. Oxford University Press.

Dzhambazov, G. & Serra, X. (2015). Modeling of phoneme durations for alignment between polyphonic audio and lyrics. In *Sound and Music Computing Conference*, Maynooth, Ireland.

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

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

Hansen, J. K. & Fraunhofer, I. (2012). Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients.

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

Wang, Y., Kan, M.-Y., Nwe, T. L., Shenoy, A., & Yin, J. (2004). Lyrically: automatic synchronization of acoustic musical signals and textual lyrics. In *Proceedings of the 12th annual ACM international conference on Multimedia*, (pp. 212–219). ACM.

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

Young, S. J. (1993). *The HTK hidden Markov model toolkit: Design and philosophy*. Citeseer.