AUTOMATIC ALIGNMENT OF LONG SYLLABLES IN A CAPPELLA 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

ABSTRACT

In this study we propose how to modify a standard approach for text-to-speech alignment to apply in the case of alignment of lyrics and singing voice. We model phoneme durations by means of a duration-explicit hidden Markov model (DHMM) phonetic recognizer based on MFCCs. The phoneme durations are empirically set in a probabilistic way, based on prior knowledge about the lyrics structure and metric principles, specific for the Jingju opera music tradition. Phoneme models are GMMs trained directly on a small corpus of annotated singing voice. The alignment is evaluated on a cappella material from Jingju opera, which is characterized by its particularly long syllable durations. Results show that the incorporation of music-specific knowledge results in a very high alignment accuracy, outperforming significantly a baseline HMM-based approach.

1. INTRODUCTION

The task of lyrics synchronization (also known as lyrics-to-audio alignment) has as an aim to find in an automatic way a match between two representations of a musical composition: the singing voice and the corresponding lyrics. Lyrics-to-audio alignment may be used in various applications: for example to automatically match structural sections from lyrics (verse, chorus) to a recording of a particular singer. This facilitates navigation and can thus be beneficial for musicologists or singing students.

The problem of lyrics-to-audio alignment has inherent relation to text-to-speech alignment. Text-to-speech alignment has been a research field for more than 20 years and thus yielded established successful ways for modeling phonemes (Anguera et al., 2014). However, compared to speech, singing voice has some substantially different characteristics including harmonics, pitch range, pronunciation, vibrato, etc. In particular, unlike speech, for singing voice, durations of vocals have on average somewhat higher variation (Kruspe, 2014). This suggests that applying an approach from speech recognition out of the box might not lead to satisfactory results. Traditional music, characterized by frequent local tempo changes, poses an additional challenge: Singers might prolong substantially certain syllables, as a way to emphasize them or as an expressive singing element.

Furthermore, current approaches on modeling lyrics are confined 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 every language or not freely available, as is the case for Mandarin. Recent work has shown that training on singing voice instead might be a viable alternative (Hansen, 2012).

In this paper we propose a lyrics-to-audio alignment method, which relies on some of the specificities of lyrics structure of Jingju opera as an additional cue to an appoach adopted from speech alignment. One of the goals of the study is to show that enhancing computational tasks with music-specific knowledge might improve accuracy.

2. BACKGROUND ON JINGJU OPERA MUSIC PRINCIPLES

Lyrics in Jingju opera (also known as Beijing or Peking opera) come from poetry and are thus commonly structured into couplets: each couplet has two lyrics lines. A line is usually divided into 3 syllable groups: a group is called *dou* and consists of 2 to 4 written characters (Wichmann, 1991, Chapter III) [1] . To emphasize the semantics of a phrase or according to the plot, 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.*

In addition to that, each aria from Jingju opera can be arranged into one or more metrical pattern (called *banshi*): it indicates the mood of singing and is correlated to meter and tempo (Wichmann, 1991). Usually an aria starts with a slow *banshi*, which gradually changes a couple of times to a faster one, to express more intense mood. The language of Jingju is standard Mandarin with some slight dialect.

3. RELATED WORK

Current lyrics-to-audio alignment is mostly based on an approaches, adopted from text-to-speech alignment (Mesaros & Virtanen, 2008; Fujihara et al., 2011): A phonetic recognizer is built from speech corpus, whereby a hidden Markov model (HMM) is trained for each phoneme. The acoustics of phonemes are described by mel frequency cepstral coefficients (MFCCs). In an example of such an approach, polyphonic Japanese and English pop music is aligned (Fujihara et al., 2011). The authors propose to adapt the speech phoneme models to the specific acoustics of singing voice by means of Maximum Likelihood Linear Regression. This is necessary because of the lack of a big enough singing voice corpus for training. Further, an automatic segregation of the vocal line is performed, in order to reduce the spectral content from background instruments.

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

HMMs, being originally applied to model spoken phonemes, have the drawback that, in general, are not capable to represent well vowels with long and highly-variable durations. This is because the waiting time in a state in traditional HMMs cannot be unlimitedly long (Rabiner, 1989). Durations can be modeled instead by a duration-explicit hidden Markov model (DHMM) (also known as hidden semi-Markov model). In DHMMs the underlying process is allowed to be a semi-Markov chain with variable duration of each state (Yu, 2010). DHMMs have been applied to detect keywords from a cappella English pop songs (Kruspe, 2015). The author showed that accuracy of detection increases if the duration of each phoneme is learned from a singing dataset. In addition, DHMMs have been shown to be successful for modeling other problems from the domain of music information retrieval: They have been, for example successful in representing chord durations in automatic chord recognition (Chen et al., 2012).

To our knowledge, very few studies of lyrics-to-audio alignment have been conducted on songs with Chinese language (Wong et al., 2007).

4. APPROACH OVERVIEW

To model phoneme durations, we rely on a DHMM [2]. A general overview of the proposed approach is presented in Figure 1. First an audio recording is manually divided into audio segments corresponding to lyrics lines 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 line is expanded to a sequence of phoneme models, whereby reference syllable durations guide the alignment process.

4.1 Rule-based duration modeling

The idea of the duration modeling is that the actual duration of a phoneme can be seen as being generated by a statistical distribution with highest probability at an expected reference duration. The reference durations can be assigned using any prior knowledge like for example structure of lyrics segments, as has been done by Wang et al. (2004). In this work they are derived as follows:

Firstly, each *key syllable* in a *dou* is assigned longer reference duration according to empirically found ratios, while the rest get equal durations. Additionally, we observed in the dataset that usually the last *key syllable* of the last line in a *banshi* is prolonged additionally. Thus we lengthened additionally the reference syllable duration of these last *key syllables*. Figure 2 depicts an example. According to *dou* groups the 3rd, 6th and last syllable are

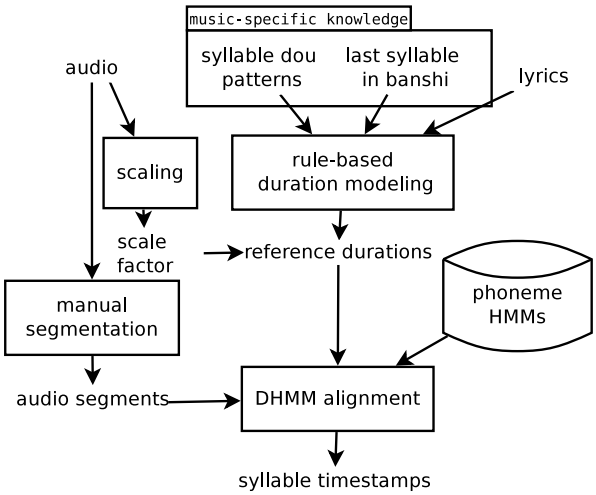

Figure 1: Approach Overview. Leftmost column represents audio preprocessing steps, while the middle column shows how reference durations are derived based on music-specific knowledge

expected to be prolonged. Note that for the example this expectation does not hold for the 3rd syllable.

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). A syllable has a middle part (nucleus) being a simple vowel, a diphthong, or triphthong. An initial part (a consonant) or a final part (a group of consonants) is optional. We assign consonants a fixed reference duration $R_c = 0.3$ seconds, while the rest of the syllable is distributed equally among vowels. The reference durations R_i are linearly scaled to a reference number of frames according to the ratio between the number of phonemes in a lyrics line and the duration of its corresponding audio segment.

4.2 Phoneme models

For each phoneme a GMM is trained on annotated a cappella singing. The first 13 MFCCs and their Δ and $\Delta\Delta$ are extracted from 25ms audio frames with the hop size of 10ms. The extracted features are then fit into a phoneme GMM with 40 components: a number of components usually proved as sufficient in speech recognition. A model for silent pause *sp* is added at the end of each syllable, which is optional on decoding. This allows to accommodate the frequent for Jingju regions of pauses after some syllables.

4.3 DHMM alignment

The syllables for a line are expanded to a sequence of phonemes based on grapheme-to-phoneme rules [3]. Then the trained GMMs are concatenated into a phonemes network, represented by a HMM, where each GMM is a state. The HMM is aligned to the MFCC features, extracted from

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

[3] We built a pinyin-to-X-Sampa mapping available at https://github.com/georgid/AlignmentDuration/blob/noteOnsets/jingju/syl2ph.txt

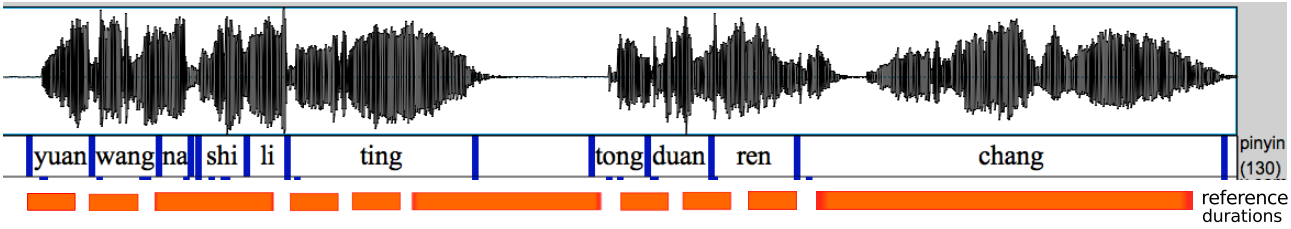

Figure 2: An example of 10-syllable line, being last in a *banshi* (before the *banshi* changes). Actual syllable durations are in pinyin, whereas reference durations are in orange parallelograms (below).

the aria, being aligned. The most likely state sequence is found by means of a forced alignment with Viterbi decoding.

We have adopted the idea of Chen et al. (2012) not to represent durations by an additional counter state in the HMM, but instead to modify the Viterbi decoding stage. Let us define

$\delta_t(i)$: probability for the path with highest probability ending in state *i* at time *t* (comply with the notation of Rabiner (1989, III. B)))

Now maximization is carried over the most likely duration for each state, instead of over different states:

$$\delta_t(i) = \max_d\{\delta_{t-d}(i-1)\, P_i(d)\, [B_t(i,d)]\} \qquad (1)$$

where $B_t(i,d)$ is the observation probability of staying d frames in state i until frame t. The duration d of a phoneme is modeled as a normal distribution $\mathcal{N} \sim (R_i;\sigma)$, with a peak at R_i. Thus, we chose to restrict the domain of d to $(\max\{R_i-\sigma,1\}, R_i+\sigma)$. Note that in forced alignment the source state could be only the previous state $i-1$. More details on the inference with DHMM can be found in our previous work Dzhambazov & Serra (2015). In comparison to our previous work, we opted for dividing the global standard deviation σ into σ_c for consonants and σ_v for vowels. Proper values for σ_c and σ_v assure that a phoneme sung longer or shorter than the expected R_i can be adequately handled. Another modification we did is that *sp* models are assigned an exponential distribution, because the duration of inter-syllable silences cannot be predicted.

5. DATASET

The dataset has been especially compiled for this study and consists of excerpts from 15 arias of two female singers, chosen from a *CompMusic* corpus of Jingju arias (Repetto & Serra, 2014). For a given aria were present two versions: a recording with voice plus accompaniment and an accompaniment-only one. Thus a cappella singing was generated by subtracting the instrumental accompaniment from the complete version [4]. Table 1 presents the average values for lines and syllables.

[4] The resulting monophonic singing is as clean as if it were a cappella, having slightly audible artefacts from percussion on the non-vocal regions

	dataset	'canonical' dataset
duration (minutes)	67	27
#lines per aria	9.2	9.9
#syllables per line	10.7	10.3
line duration (seconds)	18.3	23.4
syllable duration (seconds)	2.4	3.1

Table 1: Line and syllable averages about the dataset

Each aria is annotated on the phoneme level by native Chinese speakers and a Jingju opera musicologist. The phoneme set has 29 phonemes and is derived from Chinese pinyin, and represented using the X-sampa standard [5]. To assure enough training data for each model, certain phonemes are grouped into phonetic classes, based on their perceptual similarity.

Further, we selected a 'canonical' subset of the dataset, consisting of lines, according to the assumptions we made: *key syllables* should be prolonged. Thus, we kept only these audio segments, for which at most one *key syllable* is not prolonged and discarded the rest. We considered a syllable as being prolonged if it is longer than 130% of the average syllable duration for the current line.

6. EXPERIMENTS

Alignment accuracy is evaluated as the percentage of duration of correctly aligned syllables from total audio duration (see Fujihara et al. (2011, figure 9) for an example). Accuracy is measured for each manually segmented line and accumulated on total for all the recordings [6].

6.1 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

[5] Annotations are made available at http://compmusic.upf.edu/node/286

[6] To encourage reproducibility of this research an efficient open-source implementation together with documentation is available at https://github.com/georgid/AlignmentDuration/tree/noteOnsets/jingju. Further, a script for building the models is available at https://github.com/elitrout/lyrics/blob/master/code/htk/buildModelHTKSave.py

	baseline	DHMM	oracle
overall	56.6	89.9	98.5
'canonical'	57.2	96.3	99.5

Table 2: Comparison of accuracy on oracle, baseline and DHMM alignment on total and selected arias. Accuracy is reported as accumulate correct duration over accumulate total duration over all lines from a set of arias.

and 0 otherwise. We found that the accuracy per line of lyrics is close to 100%, which means that the model is generally capable of handling the highly-varying vocal durations of Jingju singing. Most optimal results were obtained with σ_c = 0.7 seconds; σ_v = 2.0 seconds, which are used in experiment 2.

6.2 Experiment 2: comparison with baseline

As a baseline we employ a standard Viterbi decoding, run with the *htk* toolkit Young (1993). 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 10 of the arias using the phoneme-level annotations and evaluated on a 5-aria hold-out subset.We have further evaluated on the 'canonical' selected subset of lyrics lines, introduced in Section 5. Table 2 shows that the proposed duration model outperforms significantly the baseline alignment. The improved accuracy for 'canonical' lyric lines can be attributed to the increased degree, to which prior duration expectations are met.

7. CONCLUSION

In this work we evaluated the behavior of a HMM-based phonetic recognizer for lyrics-to-audio alignment in two settings: with and without utilizing lyrics duration information. Using probabilistic duration-explicit modeling of phonemes for the former setting outperformed the latter on recordings of a cappella Jingju opera. It has incorporated prior expectations of syllable durations, based on knowledge specific for this music genre. In particular, the proposed DHMM aligns remarkably well a selected set of lyrics lines, which comply more precisely with these music-specific principles.

Acknowledgements We are thankful to Wanglei from *Doreso* for providing a dictionary of pinyin syllables. 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 grant agreement 267583) and partly by the AGAUR research grant.

8. REFERENCES

Anguera, X., Luque, J., & Gracia, C. (2014). Audio-to-text alignment for speech recognition with very limited resources. In *INTERSPEECH*, (pp. 1405–1409).

Chen, R., Shen, W., Srinivasamurthy, A., & Chordia, P. (2012). Chord recognition using duration-explicit hidden markov models. In *Proceedings of the 13th International Society for Music Information Retrieval Conference*, (pp. 445–450).

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. *IEEE Journal of Selected Topics in Signal Processing*, *5*(6), 1252–1261.

Hansen, J. K. (2012). Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients. In *Proceedings of the 9th Sound and Music Computing Conference*, (pp. 494–499)., Copenhagen, Denmark.

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.

Kruspe, A. M. (2015). Keyword spotting in singing with duration-modeled hmms. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, (pp. 1291–1295). IEEE.

Mesaros, A. & Virtanen, T. (2008). Automatic alignment of music audio and lyrics. In *Proceedings of the 11th Int. Conference on Digital Audio Effects (DAFx-08)*.

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 *Proceedings of the 15th International Society for Music Information Retrieval Conference*, (pp. 313–318).

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.

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

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

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