ON THE USE OF LYRICAL DURATION FROM MUSICAL SCORE FOR AUTOMATIC LYRICS-TO-AUDIO ALIGNMENT

Georgi Dzhambazov, Xavier Serra

Music Technology Group
Universitat Pompeu Fabra
Barcelona

ABSTRACT

In this work we show that automatic lyrics-to-audio alignment can be improved by explicitly modeling sung vocal durations. The system is a variant of a duration-explicit hidden Markov model (DHMM) phonetic recognizer based on timbral features: mel frequency cepstral coefficients (MFCCs). We modify the standard scheme for text-to-speech alignment to address the differences of phoneme durations, specific for singing. Phoneme durations are inferred from sheet music.

The proposed approach is tested on polyphonic audio from the classical Turkish music tradition, whereby MFCCs are extracted in a way robust to background instrumental sounds. In order to assess the impact of the polyphonic setting, alignment is evaluated as well on an acapella dataset, compiled especially for this study.

We show that the inclusion of duration information improves alignment accuracy by absolute 10 percent on the level of lyrics lines (phrases) and performs on par with state-of-the-art aligners for other languages.

***Index Terms*—** lyrics-to-audio alignment; score-informed alignment; phoneme durations; singing voice tracking; Turkish classical music

1. INTRODUCTION

The automatic synchronization between lyrics and audio is a challenging research problem. It has inherent relation to text-to-speech alignment. For spoken utterances phonemes have relatively similar duration across speakers. Unlike that, in singing durations of phoneme (especially vowels) have higher variation [**?**]. When being sung, vowels are prolonged according to musical note values, which in term have intrinsic relation to musical meter (e.g. duration could align with beats in a musical bar).

Another aspect that distinguishes speech from music is that unlike clean speech, singing voice is accompanied by background instruments. Instrumental accompaniment and non-vocal segments can deteriorate significantly the alignment accuracy.

The goal of this study is to test the hypothesis that extending a state-of-the-art system for automatic lyrics-to-audio alignment with modeling of phoneme durations, can improve its accuracy. More specifically, we aim to show that durations of vocals (inferred from musical score) can guide the recognition process in cases when it looses track in polyphonic audio. Such guidance can be compared to the way modeling prosodic rules helps in automatic speech understanding.

While being aided by sheet music, our modeling approach allows at the same time room for certain temporal flexibility to handle cases of expressive singing, in which vocals are sustained in a way not obeying the reference sheet music. The proposed approach was tested on polyphonic audio from the classical Turkish tradition which is characterized by high degree of expressive singing, thus providing challenging material with versatile temporal deviations.

2. RELATED WORK

To date most of the studies of automatic lyrics-to-audio alignment exploit phonetic acoustic features and state-of-the-art work is based on a phoneme recognizer [**?**, **?**].

An example of such a system [**?**] relies on hidden Markov model (HMM) and was tested on Japanese popular music. To reduce the spectral content of background instruments, the authors perform automatic segregation of the vocal line. Then Viterbi forced alignment [**?**] is run utilizing mel frequency cepstral coefficients (MFCCs) extracted from the vocal-only signal. In both [**?**] and [**?**] the phoneme models are trained on speech and later adapted to singing voice. This is necessary because of the lack of big enough training singing voice corpus. In [**?**] additionally an adaptation to the voice of a particular singer is carried out.

In other works duration of lyrics has been applied as a reinforcing cue: In [**?**] relative estimated durations are inferred directly from textual lyrics. The estimation process is done based on supervised training on singing voice.

A common-occurring drawback of HMMs is that their capability to model exact state durations is restricted. The wait time in a state is implicitly set to a geometric distribution (derived from the self-transition likelihood). Duration is

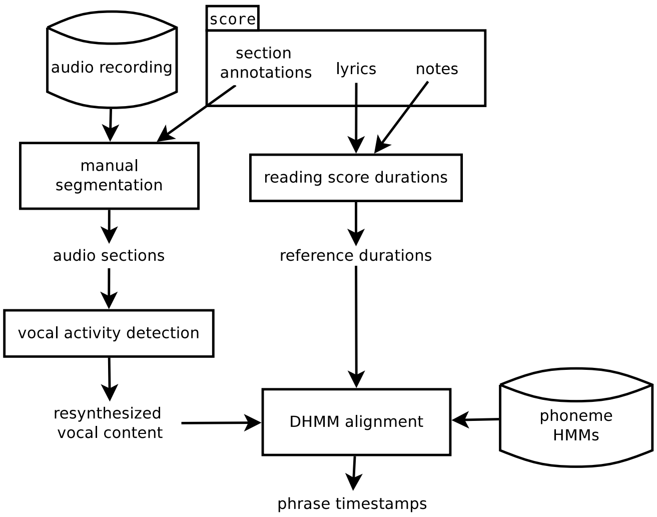

Fig. 1. Overview of the modules of the proposed approach. Leftmost column represents audio preprocessing steps, while the middle column shows how durations are modeled.

usually 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 [**?**]. Each state in turn can be assigned any statistical distribution. DHMMs have been shown to be successful for modeling chord durations in automatic chord recognition [**?**].

3. PROPOSED SYSTEM

Similar to [**?**] in this work we develop a phoneme-recognizer-based forced alignment employing the Viterbi algorithm [**?**] to decode the most optimal state sequence. We have adopted the idea of [**?**] not to explicitly add states for durations in the model, but instead to extend the Viterbi decoding to handle duration of states. For brevity in the rest of the paper our model will be referred to as DHMM.

Figure **??** presents an overview of the proposed system. An audio recording and its corresponding score are input. Relying on HMMs of phonemes the DHMM returns start and end timestamps of aligned lyrical phrases.

First an audio recording is manually divided into sections (e.g. verse, chorus) as indicated in the score, whereby instrumental-only sections are discarded. All further steps are performed on each audio section. If we had used automatic segmentation instead, potential erroneous lyrics and durations could have biased the comparison of a baseline system and DHMM. As we focus on evaluating the effect of DHMM, manual segmentation is preferred.

3.1. Vocal activity detection

Next a predominant singing voice detection (a.k.a. vocal activity detection) method [**?**] is applied on each section to attenuate the spectral content from accompanying instruments, because they have negative effect on the alignment. It performs detection of segments with predominant singing voice and in the same time melody transcription for the detected segments. Based on the extracted melodic contours, the vocal content is resynthesized as separate audio using a harmonic model [**?**]. This resynthesis allowed us to perceptually evaluate the intelligibility of different vocals after vocal detection. More details and examples on the resynthesis step can be found in [**?**].

3.2. Reading score durations

For each lyrics syllable a reference duration is derived from the values of its corresponding musical notes . Then the reference duration is spread among its constituent phonemes, whereby consonants are assigned constant duration and the rest is assigned to the vowel.

Each phoneme is modeled by a 3-state HMM, resulting into a lookup table of reference durations R_i for each state i. We assume that the duration d for a state i may vary according to a normal distribution $P_i(d)$ with mean at the reference duration $d = R_i$ and standard deviation σ. To align a given recording the score-inferred lengths are linearly rescaled to match its musical tempo. In this work the unit of R_i is number of acoustic frames.

3.3. Duration-explicit HMM alignment

For each phoneme a HMM is trained from a corpus of turkish speech utilizing MFCCs. For given lyrics, the words are expanded to phonemes based on grapheme-to-phoneme rules for Turkish [**?**, Table 1] and the the trained HMMs are concatenated into a phoneme network. The network is then aligned to the MFCC features, extracted from the resynthesized audio signal, by means of the duration-explicit decoding. In what follows we describe a variation of Viterbi decoding method, in which maximization is carried over the most likely duration for each state. The decoding is adapted from the procedure described in [**?**]. Let us define:

R_{max} : $\max_i(R_i) + \sigma$

$b_i(O_t)$: observation probability for state i for feature vector O_t (comply with the notation of [**?**])

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

Recursion

For $R_{max} < t \leq T$

$$\begin{aligned}\delta_t(i) &= \max_d\{\delta_{t-d}(i-1). \\ &\qquad P_i(d)^\alpha \left[B_t(i,d)\right]^{1-\alpha}\} \qquad (1)\end{aligned}$$

where

$$B_t(i,d) = \Pi_{s=t-d+1}^{t} b_i(O_s) \qquad (2)$$

is the observation probability of staying d frames in state i until frame t. The domain of d comes from the normal distribution $(\max\{R_i - \sigma, 1\}, R_i + \sigma)$ and is reduced for states with reference duration $R_i < \sigma$.

A duration back-pointer is defined as

$$\begin{aligned}\chi_t(i) &= \arg\max_d\{\delta_{t-d}(i-1). \\ &\qquad P_i(d)^\alpha \left[B_t(i,d)\right]^{1-\alpha}\} \qquad (3)\end{aligned}$$

Note that in forced alignment the source state could be only the previous state $i - 1$.

To be able to control the influence of the duration we have introduced a weighting factor α. Note that setting α to zero is equivalent to using a uniform distribution for $p_i(d)$.

Initialization

For $t \leq R_{max}$

$$\delta_t(i) = \max\{\delta_t(i)^*, \kappa_t(i)\} \qquad (4)$$

where a reduced-duration delta $\delta_t(i)^*$ is defined in the same way as in (**??**) but

$$d \in \begin{cases} emptySet, & t \leq R_i - \sigma \\ (R_i - \sigma, \min\{t-1, R_i + \sigma\}), & else \end{cases} \qquad (5)$$

reduces the duration to t when $t < R_i + \sigma$.

Lastly the probability of staying at initial state i at time t is defined as:

$$\kappa_t(i) = \pi_i P_i(t)^\alpha [\Pi_{s=1}^{t}(O_s)]^{1-\alpha} \qquad (6)$$

for $t \in (1, R_i + \sigma)$.

Finally the decoded state sequence is derived by backtracking starting at the last state N and switching to a source state a number of $d = \chi_t(i)$ frames ahead according to the back-pointer from (**??**).

total #sections	#phrases per section	section duration
75	2 to 5	7-20 seconds

Table 1. Section and phrase statistics for test dataset.

4. EXPERIMENTAL SETUP

Alignment is performed on each manually divided audio section and results are reported per recording (on total for its sections).

To assess the benefit of duration modeling for alignment a comparison to a baseline system with standard Viterbi decoding is conducted. We present results for the most optimal $\alpha = 0.97$. It was found by minimizing the alignment error (see section **??**) on a separate development dataset of 20 minutes Turkish acapella recordings. To assure optimality we aligned on word-level ground truth.

To train the speech model the HMM Toolkit (HTK) [**?**] is employed. The acoustic properties (most importantly the formant frequencies) of spoken phonemes can be induced by the spectral envelope of speech. To this end, we utilize the first 12 MFCCs and their delta to the previous time instant.

A 3-state HMM model for each of 38 Turkish phonemes is trained, plus a silent pause model. For each state a 9-mixture Gaussian distribution is fitted on the feature vector.

4.1. Datasets

The test dataset consists of 12 single-vocal classical Turkish music recordings with accompaniment with total duration of 18:40 minutes. Scores are provided in the machine-readable *symbTr* format [**?**].

Additionally a separate acapella dataset of the same 12 songs sung by semiprofessional singers has been recorded especially for this study. It can be considered a vocal-track-only version of the original polyphonic dataset. Evaluation on the acapella corpus was conducted in order to assess the impact of the vocal extraction step.

Each song section was manually annotated into musical phrases as proposed by [?]. A musical phrase usually corresponds to a lyrical line. If a phrase boundary splits a word we have modified it to include the complete word. Short instrumental motives have not been excluded from the phrases. Furthermore we split or merged some melodic phrases so that phrases within a recording have roughly the same number of musical bars (1 or 2). Table 1 presents statistics about phrases.

4.2. Evaluation metrics

Alignment is evaluated in terms of alignment accuracy (AA) as the percentage of duration of correctly aligned regions from total audio duration (see [**?**, Fig.9] for an example). A value

System variant	**alignment accuracy**	**alignment error**
musical score in-sync	88.14	0.32
HMM polyphonic	67.46	1.04
DHMM polyphonic	77.74	0.63
DHMM acapella	90.04	0.26
HMM+adaptation [**?**]	-	1.4
HMM+ singer adaptation [**?**]	85.2	-

Table 2. Alignment accuracy (in percent) for musical score in-sync; different system variants: baseline HMM and DHMM; state-of-the-art for other languages. Alignment accuracy is reported as total for all recordings. Additionally the total mean phrase alignment error (in seconds) is reported

of 100 means perfect matching of phrase boundaries. Additionally is reported the mean absolute phrase alignment error (AE): measured at the start and end timestamp of a phrase.

We define a metric *musical score in-sync* (MSI) to measure the approximate degree to which a singer performs a recording in synchronization with note values indicated in the musical score. Thus low accuracy of MSI indicates a higher temporal deviation from musical score. We compute MSI per a recording as the AA of score-inferred reference durations R_i (defined in section **??**) compared to ground-truth, as if they were results after alignment.

5. RESULTS

Table 2 presents comparison of the proposed DHMM system performance and a baseline HMM system. It can be observed that modeling of note values with DHMM increases HMM accuracy by 10 absolute percent. One reason for this are cases of long vocals, in which HMM switches to the next phoneme prematurely (due to its inability to stay long in a given state). In contrast, the duration-explicit decoding allows picking the optimal duration (which can be traced in an example in figure **??**).

Figure **??** allows a glance at results per recording, ordered according to MSI[1] . It can be observed that DHMM performs consistently better than the baseline (with some exceptions of where accuracy is close). Unlike the relatively stable accuracy for the acapella case, when background instruments are present, the accuracy variates more among recordings.

Although coming from different genre and language, we compare our alignment results to best hitherto alignment systems for English pop songs [**?**] and for Japanese pop [**?**]. These are abbreviated in table 2 respectively as *HMM + adaptation* and *HMM + singer adaptation*. In these works alignment is evaluated also on the level of a lyrical line/phrase. Except for the duration-explicit decoding scheme, our approach differs from both works essentially in that they conduct speech-to-singing-voice adaptation. Unlike that we did not perform any adaptation of the original speech model. Adaptation data of clean singing voice for a particular singer might not always be available and thus does not allow the system to scale to data from unknown singers. In the end, despite lacking adaptation, our approach yields results comparable to these reference approaches.

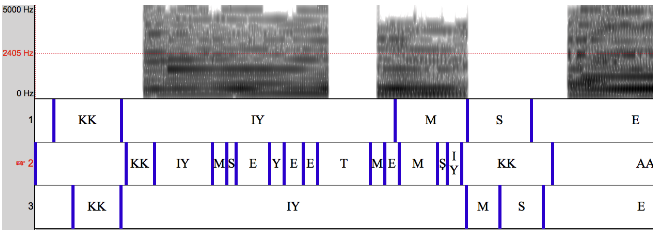

Fig. 3. Example of decoded phonemes. *very top*: resynthesized spectrum; *upper level*: ground truth, *middle level*: HMM; *bottom level*: DHMM; (excerpt from the recording 'Kimseye etmem şikayet' by Bekir Unluater)

Moreover, [**?**] trains a vocal activity detection (VAD) module on data selected from material with same acoustics characteristics as the test data. The VAD module showed to notably increase the average accuracy of 72.1 % for a baseline to accuracy of 85.2 % for their final system. Similarly we observe that for our system evaluation on the acapella dataset yields an accuracy by about the same percent higher than the polyphonic one (see table 2). Investigating our results with low accuracy revealed that false positives of our VAD module is a considerable reason for misalignment. Unlike [**?**] we did not tailor the parameters of the VAD module used in this work (built for Western popular music) to the specificities of our test dataset.

6. 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 duration-explicit modeling for the former setting outperformed the latter for polyphonic Turkish classical recordings.

Importantly our approach reaches accuracy on par with state of the art alignment systems by using an acoustic model trained on speech only. This suggests that steps like adaptation to singing voice and adaptation to a particular singer can be compensated by applying the DHMM. Furthermore, the DHMM performs considerably better on an acapella version of the test dataset, which indicates that improving the vocal

[1] the per-recording results are published here `https://drive.google.com/file/d/0B4bIMgQlCAuqY3hKc25WTm9kTEk/view?usp=sharing`

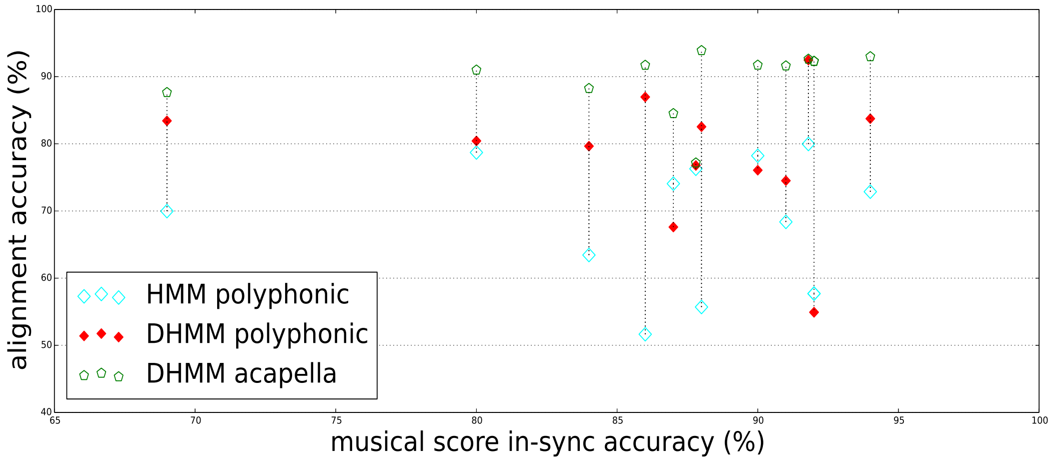

Fig. 2. Comparison between results from DHMM (for both polyphonic and acapella) and baseline HMM. Metric used is alignment accuracy. A connected triple of shapes represents results for one recording. Results are ordered according to *musical score in-sync* (on horizontal axis)

activity detection module will result in even better accuracy.

A limitation of the current alignment system is the prerequisite for manually-done structural segmentation, which we plan to automate in the future.

In general, the proposed approach is applicable not only when musical scores are available, but also for any format, from which duration information can be inferred: for example annotated melodic contour or singer-created indications along the lyrics.

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 grant agreement 267583) and partly by the AGAUR research grant.

REFERENCES

[1] A.B. Smith, C.D. Jones, and E.F. Roberts, “Article title,” *Journal*, vol. 1, no. 1, pp. 1–10, Month Year.

[2] C.D. Jones, E.F. Roberts, and A.B. Smith, “Paper title,” in *Proc. Conference Name*, Location, Year, pp. 1–10.

[3] E.F. Roberts, A.B. Smith, and C.D. Jones, *Book Title*, Publisher, Year.