AUTOMATIC LYRICS-TO-AUDIO ALIGNMENT IN CLASSICAL TURKISH MUSIC

Georgi Dzhambazov, Sertan Şentürk, Xavier Serra
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
{georgi.dzhambazov,sertan.senturk,xavier.serra}@upf.edu

ABSTRACT

We apply a lyrics-to-audio alignment state-of-the-art approach to polyphonic pieces from classical Turkish repertoire. A phonetic recognizer is employed, whereby each phoneme is assigned a hidden Markov model (HMM). Initially trained on speech, the models are adapted on singing voice to match the acoustic characteristics of the test dataset. Being the first study on lyrics-to-audio alignment applied on Turkish music, it could serve as a baseline for singing material with similar musical characteristics. As part of this work a dataset of recordings from the classical music tradition is compiled. Experiments, conducted separately for male and female singers, show that female singing is aligned more accurately.

1. INTRODUCTION

Lyrics are one of the most important musical components of vocal music. When a performance is heard, most listeners will follow the lyrics of the main vocal melody. From this perspective, the automatic synchronization between lyrics and music poses a user-demanded research question.

By applying a lyrics-to-audio alignment state-of-the-art approach to classical Turkish songs, we aim to outline the research challenges, raised by the musical aspects peculiar to this music tradition. To this end we compile a dedicated evaluation corpus. This work is performed in the context of the CompMusic project [Serra, 2011], which aims to analyze non-western music traditions in a culture-specific manner. In this respect, the corpus is built as well with the intention to be useful for further music retrieval tasks for the Turkish tradition.

2. ELEMENTS OF CLASSICAL MUSIC OF TURKEY

Şarkı - the scope of this study - is a vocal form in the classical repertoire. Typical for it is that vocal and accompanying instruments follow the same melodic contour in their corresponding registers with slight melodic variations. However, the vocal line has usually melodic predominance. This musical interaction is termed heterophony.

Additionally, the *şarkı* form adheres to a well-defined verse-refrain-like structure: a *şarkı* contains *zemin* (verse), *nakarat* (refrain), *meyan* (second verse), *nakarat* (refrain) sections, which are preceded by *aranağme* (an instrumental interlude) [Ederer, 2011].

Concerning language, unlike modern Turkish, Ottoman Turkish is characterized by more loanwords from Arabic and Persian origin. The lyrics language for the *şarkı* compositions in our evaluation dataset spans both modern and Ottoman Turkish. The Turkish phonology comprises 38 distinctive phonetic sounds, 8 of which are vowels. There are no diphthongs, and when vowels come together, they retain their individual sounding. Lengthening of vowels is realized by a non-pronounced character *ğ*. However vowel length has a negligible importance in sung Turkish.

In classical Turkish singing an expressive effect occurs frequently: When performing vibrato, singers tend to alternate between the original vowel and another helper one, simultaneously to alternating the pitch.

3. RELATED WORK

To date most of the studies of singing voice in general and the automatic lyrics-to-audio alignment in particular are focused on western polyphonic popular music. Many approaches exploit phonetic acoustic features.

An example of such a system [Fujihara et al., 2011] relies on a forced alignment scheme and was tested on Japanese popular music. Since the forced alignment technique was originally developed to carry out the alignment between clean speech and text, accompanying instruments and non-vocal sections deteriorate the alignment accuracy. To address this issue, the authors perform automatic segregation of the vocal line and the alignment is run using phonetic features extracted from the vocal-only signal.

A diametrically different approach is to deploy external information sources. Müller et al. [2007] uses MIDI files, which are manually synchronized to lyrics. By performing mapping of timestamps between an audio recording and a MIDI version of the composition, lyrics are implicitly aligned to the audio.

4. METHOD

Combining aspects of these two methods, in this work we develop a system for the automatic synchronization between vocal *şarkı* recordings and their lyrics. Similar to the approach of Fujihara et al. [2011] we train a hidden Markov model (HMM) for each phoneme, present in Turkish language.

Furthermore, we exploit a lyrics representation, for which sections are labeled. Songs are segmented into structural

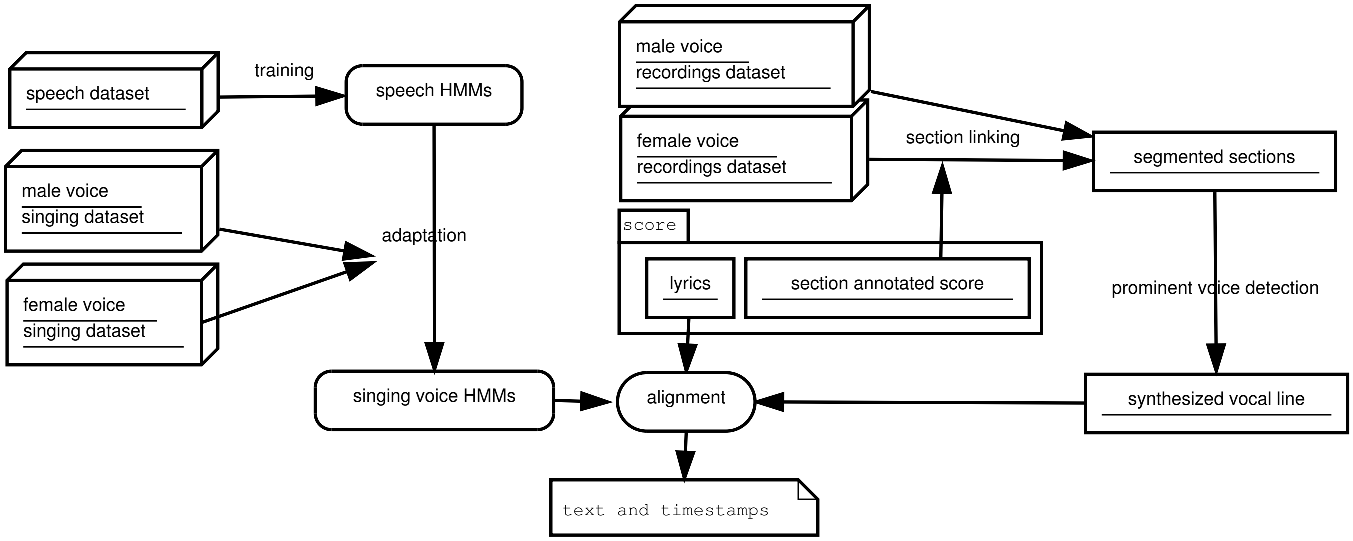

Figure 1: Training steps (on the left) and alignment process (on the right)

parts through an external module, which links sections from the musical score to temporal anchors in the audio [Şentürk et al., 2014]. Using a Viterbi forced alignment the system aligns in a non-linear way the extracted phonetic features to a network of the trained phoneme models.

Figure 1 gives an overview of the layout of the system.

4.1 Training

4.1.1 Training a speech model

In the absence of annotated data of singing phonemes, we train mono-phone models on a big corpus of annotated Turkish speech. Later, we adapt the speech models to match the acoustic characteristics of clean singing voice using a small singing dataset (see Figure 1).

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 mel-frequency cepstral coefficients (MFCCs) and their difference 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 10-mixture Gaussian distribution is fitted on the feature vector.

4.1.2 Adaptation

Mesaros & Virtanen [2008] have proposed to apply an adaptation technique for speaker-dependent speech modeling. The Maximum a posteriori (MAP) transform - applied as well in this work - shifts the mean and variance components of the Gaussians of the speech model towards the acoustic characteristics of the singing voice. An advantage of the MAP transform compared to other adaptation techniques is that it allows the manipulation of each phoneme model independently.

In singing the articulatory characteristics of unvoiced consonants do not vary significantly from these in speech. This is because unvoiced consonants do not bear any melodic line. For this reason we adapted only the vowels and voiced consonants.

4.2 Preprocessing steps

4.2.1 Section Linking

In the *şarkı* form each vocal segment is associated with a section (e.g. *nakarat*). Furthermore, a given performance of a *şarkı* composition typically contains section repetitions or omissions, which are not indicated in the score. Thus, for each audio recording, prior to alignment, we utilize a method for linking score sections to their beginning and ending timestamps [Şentürk et al., 2014] (see Figure 1). Non-vocal *aranağme* sections are discarded.

To each segmented vocal section we assign the corresponding lyrical strophe automatically, because lyrics syllables are manually anchored to musical notes in the score.

4.2.2 Predominant vocal detection

After section linking a melody extraction algorithm is applied [Salamon & Gómez, 2012]. It extracts the contour of the predominant melodic source and generates time series of pitch values. It performs in the same time a dominant source detection: it returns pitch values of zero for regions with no dominant melody.

Re-synthesis Using a harmonic model [Serra & D, 1989], based on the extracted pitch series, we re-synthesize the spectral components corresponding to the first 30 harmonic partials of the singing voice. This results ideally in a vocal line, with no audible instruments. However, some spectral artifacts from accompanying instruments are inevitable, because they follow in parallel the melodic line of the voice. A side effect of the synthesis is that non-voiced consonants are not re-synthesized, which leaves regions of silence (see Figure 2). To handle these special-case-consonants, they are replaced in the pronunciation dictionary by the HMM for silence. MFCCs are extracted from the vocal part, because the harmonic partials keep the information about articulation.

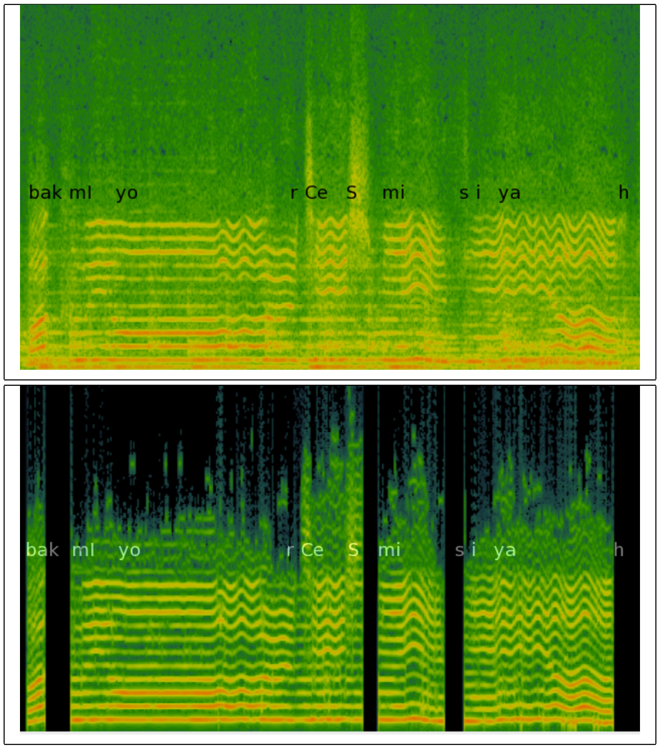

Figure 2: A snippet from the test dataset: spectrum of original audio (above) and after re-synthesis (below)

A challenge for lyrics-to-audio alignment is posed by regions, in which the predominant source is a solo instrument, because its pitch is detected instead of voice and respectively the synthesized audio does not represent voice. In all sections of the *şarkı* form (except *aranağme*), solo instruments can be present at interludes preceding the vocal phrases. To accommodate these instrumental regions, we train a single-state background noise HMM that captures the timbre of background instruments.

4.3 Alignment

The lyrics are expanded to phonemes based on grapheme-to-phoneme rules for Turkish [Özgül Salor et al., 2007, Table 1]. In this way, the HMMs are concatenated into a phoneme network. At the beginning and end of the network for each section, the background noise model (NOISE) is appended. Setting it as optional allows the recognizer to activate it or not, depending on whether sound from background instruments was re-synthesized.

Figure 3 presents an example of such a network.

```
{sil|NOISE} kUSade [sil] taliim {sil|NOISE}
```

Figure 3: Example of recognition grammar. Curly brackets denote one or more occurrence, square brackets denote zero or more occurrence, and vertical bars denote alternatives.

The phoneme network is then aligned to the extracted features by means of the Viterbi forced alignment. The alignment is run for each segmented section separately.

	total #sections	section duration	#phrases per section
male	48	8 to 20 seconds	3 to 7
female	41	8 to 35 seconds	3 to 7

Table 1: Section duration and counts for test dataset, divided into male and female cases. Total count of phrases is around 200 for male and 200 for female

	abs. median	abs. mean	standard deviation
male	0.25	0.95	1.95
female	0.3	1.61	3.15
total	0.28	1.26	2.63

Table 2: Alignment error (in seconds) for the female and male subsets of the test dataset. Evaluation metrics are the median and mean of the absolute value of the misaligned time

5. EXPERIMENTAL SETUP

To train the speech model and adapt it to singing voice, as well as to run the forced alignment, the HMM Toolkit (HTK) [Young & Young, 1993] is employed. The alignment is run on the re-synthesized voice-only counterparts of the original.

5.1 Datasets

The speech dataset, used for training, encompasses clean speech totaling to approximately 500 minutes of speech [Özgül Salor et al., 2007].

The test dataset consists of 10 single-vocal *şarkı* performances (5 distinct male and 5 distinct female). The recordings are selected from a musicBrainz collection of Turkish music [1], whereas scores are provided in the machine-readable *symbTr* format [Karaosmanoğlu, 2012] [2]. The phrase boundaries of each song section were manually annotated, whereby a phrase corresponds roughly to a musical bar and contains 1 or 2 words. Counts of the song sections are presented in Table 1. Phrase-level alignment accuracy of the detection results is reported (in terms of the absolute error in seconds).

5.2 Results

Statistics of the alignment errors are summarized in table 2.

The alignment accuracy for female singing voice is slightly better than for male. A reason for this is that for female voice the extracted melody line has less errors than for male. We observed that, when the extracted pitch is wrong (e.g by an octave), the vowels are not recognized correctly.

Another problem is that alignment performs poorly towards the end of longer sections, which results in outliers of huge magnitude (seen at the distribution of the

[1] http://musicbrainz.org/collection/544f7aec-dba6-440c-943f-103cf344efbb

[2] The dataset will be made available on http://compmusic.upf.edu/datasets

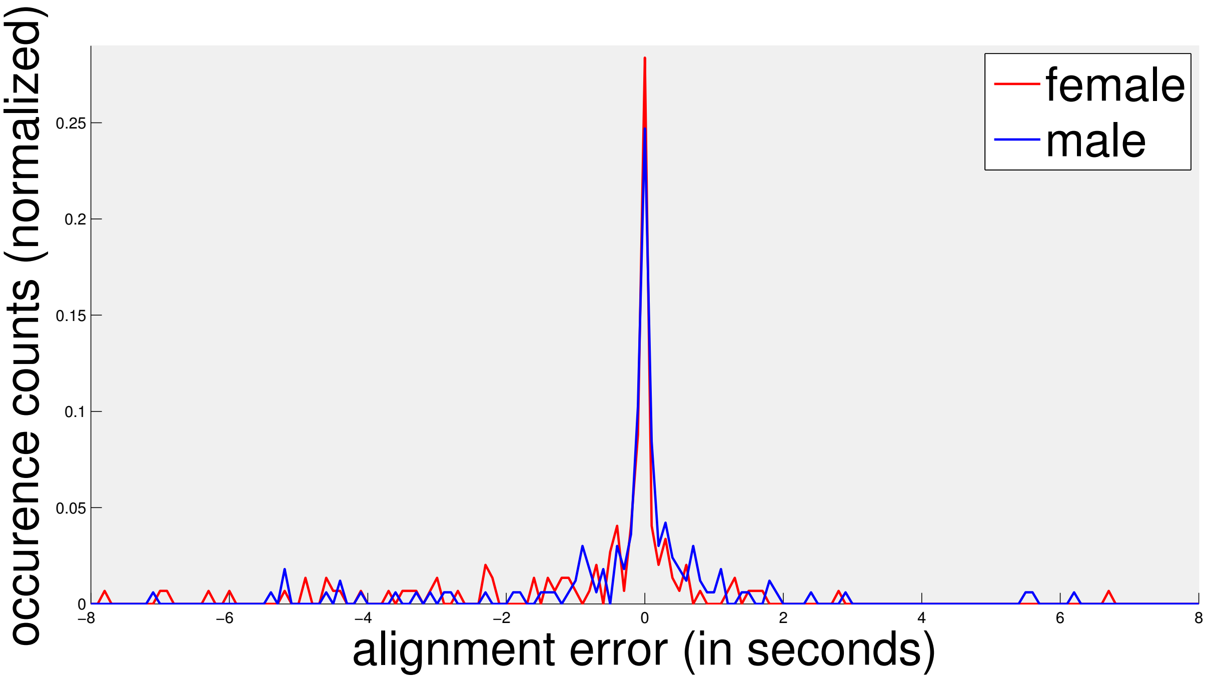

Figure 4: Distribution of the alignment error (in seconds) for male and female. Negative values mean that the beginning timestamp is detected as being later than it actually is

alignment error figure 4). A glance at the distribution of the alignment error of a lyrics-to-audio system for western popular music [Mesaros & Virtanen, 2008] reveals that the frequency and magnitude of outliers are comparable to ours. Further, our mean error lies not far from theirs of 1.4 seconds.

6. CONCLUSION

In this work was presented a method for the automatic alignment between lyrics and audio recordings of vocal compositions in the classical Turkish tradition. Performance was evaluated on a dataset, compiled and annotated by us especially for this task. The method showed better results for female singers, which is partly explained by the greater amount of erroneously recognized male pitch, which distorts the re-synthesized pitch as well.

We expect that the generated phrase-to-audio alignment may be a starting point for subsequent musicological analysis tasks.

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).

7. REFERENCES

Ederer, E. B. (2011). *The Theory and Praxis of Makam in Classical Turkish Music 1910–2010*. University of California, Santa Barbara.

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.

Karaosmanoğlu, M. K. (2012). A turkish makam music symbolic database for music information retrieval: Symbtr. *Proc. Int. Society for Music Information Retrieval (ISMIR)*.

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

Müller, M., Kurth, F., Damm, D., Fremerey, C., & Clausen, M. (2007). Lyrics-based audio retrieval and multimodal navigation in music collections. In *Research and Advanced Technology for Digital Libraries* (pp. 112–123). Springer.

Salamon, J. & Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *Audio, Speech, and Language Processing, IEEE Transactions on*, *20*(6), 1759–1770.

Şentürk, S., Holzapfel, A., & Serra, X. (2014). Linking scores and audio recordings in makam music of turkey. *Journal of New Music Research*, *43*, 34–52.

Serra, X. (2011). A multicultural approach in music information research. In *Int. Soc. for Music Information Retrieval Conf. (ISMIR)*, (pp. 151–156)., Miami, Florida (USA).

Serra, X. & D, X. S. P. (1989). A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition. Technical report.

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

Özgül Salor, Pellom, B. L., Ciloglu, T., & Demirekler, M. (2007). Turkish speech corpora and recognition tools developed by porting sonic: Towards multilingual speech recognition. *Computer Speech and Language*, *21*(4), 580 – 593.