Evaluation of the components of a lyrics-to-audio alignment method for Turkish classical music

May 2, 2014

Abstract

We propose a method for automatic alignment between lyrics and audio for turkish classical music recordings. 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 one of the first works applied on a non-western music tradition, it could serve as a baseline for singing material with similar musical characteristics. As part of this study a singing voice dataset is compiled. Since state-of-the-art lyrics-to-audio alignment systems usually consist of many steps, it is hard to trace the effect of each step on the performance of the final alignment. A focus of this work is to evaluate how the inclusion of two important such steps affect performance: 1) adaptation of the trained model and 2) reduction of the effect of background instruments through re-synthesis of vocals. Experiments are conducted with different parameters of adaptation and re-synthesis separately for male and female singers. We conclude that the most optimal results are in the case of maximum a posteriori adaptation with re-synthesized male voice.

1 Motivation

In this work we develop a method for the automatic synchronization between vocal *şarkı* recordings and their lyrics. The *şarkı* is a vocal form in the classical Turkish 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.

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 heterophony. For this sake a focus is put on measuring how important for the alignment is the effect of automatic detection of predominant vocal.

A further goal of this work is to evaluate the importance of model adaptation. The adpatation from speech model to a model covering characteristics of singing voice is a crucial component of HMM-based approaches .

2 Related work

Most of the successful hitherto lyrics-to-audio alignment systems rely on phonetic models [[Fujihara et al.(2011)Fujihara, Goto, C Mesaros, Loscos]. All these approaches share common building blocks.

Explain what type of adaptation and vocal detection is carried out in each of the appraoches....

However little work has been carried out to track how these two componenets are correlated to the final alignment's performance.

3 Approach

Figure ?? gives an overview of the layout of our alignement method.

We train a phonetic recognizer, where each phoneme is assigned a hidden Markov model (HMM). Model adaptation to singing voice ensures that after adaptation each state of a phoneme model is more likely to generate the corresponding phoneme, when being sung. A separate model is adapted for female and male voice.

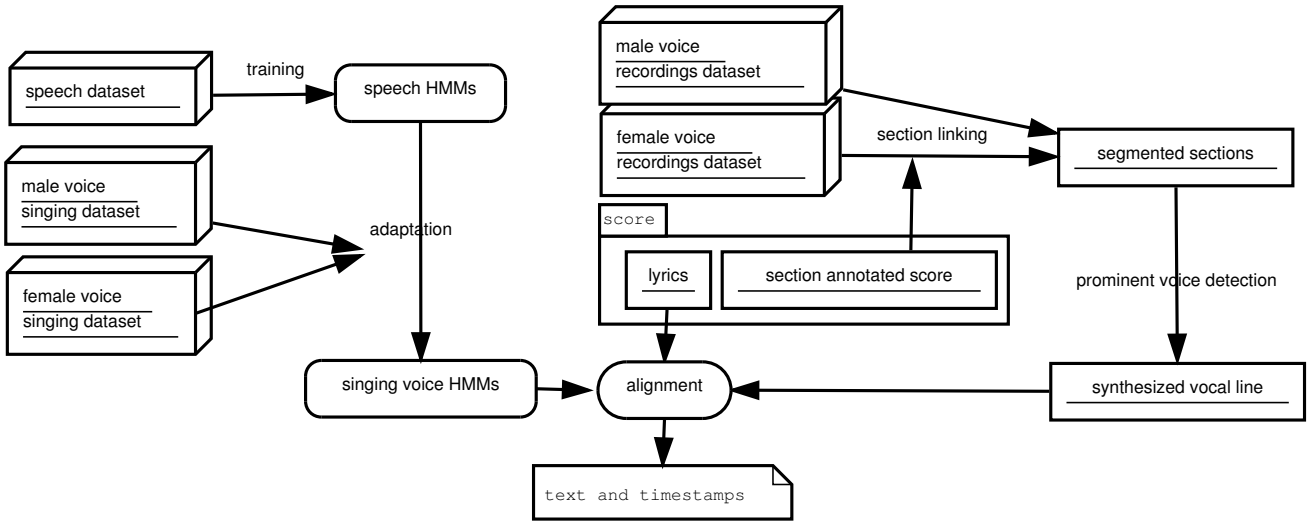

Figure 1: Training steps (on the left) and alignment process (on the right)

Songs are segmented into structural parts through an external module, which links sections from the musical score to temporal anchors in the audio. 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. A test dataset is divided into gender-specific subsets, which are alignment with the correponding adapted mode. This setting allows to isolate the phenomena of formant freqeuncies, due to high-pitched female voice.

4 Training

4.1 Training a speech model

In the absence of annotated data of singing voice phonemes, we train monophone 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 ??).

The acoustic properties (most importantly the formant frequencies) of spoken phonemes can be induced by the spectral envelope of speech. Thus we utilize the first 12 mel-frequency cepstral coe□cients (MFCCs) and their di□erence to the previous time instant.

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

4.2 Adaptation to singing voice

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

Since the consonants last short on singing their spectral dynamic characteristic are not altered significantly. For this reason we adapted only vowels. Due to lack of clean singing voice in Turkish language, a material from the HTS corpus in Japanese language [ref] was utilized. Japanese and Turkish have similar vowel charactestics.

<here spectrum of models before and after adaptation, aka MERLINOgram >

5 Alignment

For each audio recording, prior to alignment, we utilize a method [Şentürk et al.(In Press)Şentürk, Holzapfel, and Serra] for linking score sections to their beginning and ending timestamps. Why needed? Non-vocal *aranağme* sections are discarded.

The words from the lyrics are expanded to phonemes based on grapheme-to-phoneme rules for Turkish [Özgül Salor et al.(2007)Özgül Salor, Pellom, Ciloglu, and Demirekler, 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 instrumental background model is appended. Setting it as optional allows the recognizer to activate it or not, depending on whether an instrumental section is present.

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.

REWRITE THIS SECTION : fragmentary, di□erent things

5.1 Predominant vocal detection

A predominant melody detection algorithm is applied [justin].

It extracts the contour of the predominant melodic source and generates time series of pitch values. The algorithm is as well performing dominant source detection: returns pitch values of zero on regions of non-detected voice.

5.2 Re-Synthesis

Using a harmonic model [Bonada, Serra], based on the pitch series, we re-synthesize the spectral componenets corresponding to the first 30 harmonic partials of the singing voice. This resulted in slightly distorted but instruments-free vocal line. Because it holds the information about articulation, MFCCs are extracted from the segregated vocal line.

<Here figure comparing spectrum of original and synthesized voice>

However a challenge for lyrics-to-audio alignment is posed by regions, in which the predominant source is a soloing instument, because its pitch is detected instead of voice and respectively the synthesized audio does does not represent voice. In the non-aranagme sections of the sarki form, soloing instruments are present at interludes preceding the vocal phrases. To accomodate those, we train a separate model of background instruments.

A side e□ect of the synthesis is that non-voiced consonants are not resynthesized, which leaves regions of silence. To handle these special-case-consonants, they were replaced in the pronunciation dictionary by the HMM for silence.

6 Experimental Setup

To analyze the e□ect of each step in turn on the alignment, we evaluated the alignemnt accuracy in di□erent stages. For each stage the performance is measured with and without the step under consideration, while keeping all other steps present.

...

The mean and standard deviation of the alignment error serve as evaluation metric.

6.1 Datasets

The speech corpus, used for training, encompasses clean speech totaling to approximately 500 minutes of speech [Özgül Salor et al.(2007)Özgül Salor, Pellom, Ciloglu, and Demirekler]. The adaptation dataset consists of vocal-only recordings from di□erent datasets: 5 minutes of female from HTS (japanese), 5 minutes from HTS (japanese),