SEARCH BY LYRICAL PHRASES IN ACAPELLA TURKISH MAKAM RECORDINGS

First author
Affiliation1
author1@ismir.edu

Second author
Retain these fake authors in submission to preserve the formatting

Third author
Affiliation3
author3@ismir.edu

ABSTRACT

In this work we propose an approach for locating the exact occurrences of a lyrical query in musical audio, a problem known in speech processing research as key-phrase-spotting. A query is constructed by expanding text to MFCC-based phoneme models, which are trained only on speech. To address the differences of syllable durations, specific for singing, durations inferred from musical score are incorporated in the phonetic query. First we apply dynamic time warping (DTW) to estimate candidate segments. In a second step these audio segments are ranked by means of a novel hierarchical hidden Markov model (HHMM), which models a query as a separate structural section. The proposed approach is evaluated on an acapella dataset of 12 recordings from Turkish Makam recordings. By tuning the parameters of DTW, we retrieve a high-recall candidate set. Combined with a high-precision decoding by HHMM, we arrive at promising f-measure, even with standard speech phonetic models. Being on of the first methods for searching by lyrics, and the first on non-eurogenetic music in particular, we expect that it can serve as a baseline for further research on the topic.

1. INTRODUCTION

TODO: rewrite first paragraph and reorder

In this work we investigate the problem of locating the exact occurrences of a lyrical query from performance recording for a particular composition. We address the case when a query represents an entire structural section or phrase from textual lyrics. The composition is known in advance, but no information about the structure of the particular performance is given. This problem is comparable to phrase-spotting when considering speech recordings ([ref]). We assume that the musical score with lyrics is present for the composition of interest.

It has ben shown the durations of singing voice are quite different than in speech [Anna]. Therefore adopting an approach from speech recognition might lack some singing-specific rules (or semantics) including among others note durations. Hitherto approaches do not rely on temporal information. A lot of this information can be inferred from musical scores.

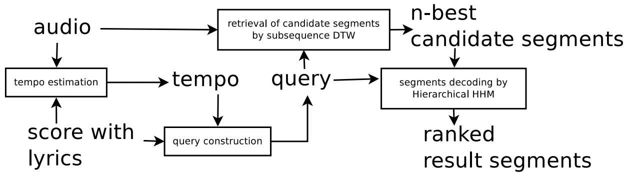
audio
retrieval of candidate segments by subsequence DTW
n-best candidate segments
tempo estimation
tempo
query
segments decoding by Hierarchical HHM
score with lyrics
query construction
ranked result segments

Figure 1. System Overview

Why is it important: Search by lyrics has an inherent connection to the problem of structure discovery. For most types of music a section-long lyrical phrase is a feature that represents the corresponding structural section in a unique way.

2. RELATED WORK

[1]

3. ARCHITECTURE

Figure 1 presents an overview of the proposed approach.

We propose a two-pass retrieval approach: On the first pass a subsequence DTW retrieves a set of candidate audio segments that roughly correspond to a query. On the second pass each candidate segment is separately fed into the HHMM, for which we run a Viterbi decoding to assure that only one (the most optimal) path is detected for an audio segment. Any query-to-audio fullpath match is considered as a hit and all results are ranked according to their respective Viterbi likelihoods.

3.1 Tempo Estimation

TODO

3.2 Query construction

A selected lyrical phrase is expanded it to its constituent syllables and for each 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, resulting into a list of phoneme reference durations R_p for each phoneme p.

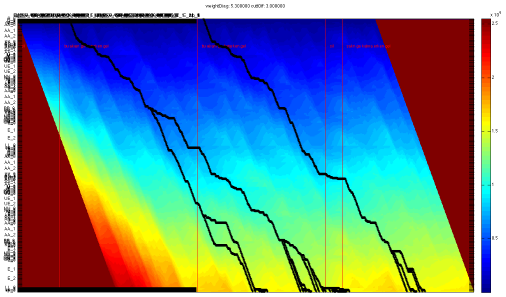

Figure 2. Example candidate audio segment of many overlapped paths (two main)

A query sequence is constructed by substituting the phonemes for their respective models, each being duplicated R_p times. A model has 3 GMMs that represent respectively the beginning, middle and ending acoustic states of a phoneme.

To query a particular performance, the score-inferred R_p are linearly rescaled to match its musical tempo.

4. RETRIEVAL OF CANDIDATE SEGMENTS BY SUBSEQUENCE DTW

Subsequence DTW is a dynamic programming technique that has been applied to phrase-spotting from speech[]. In this example series of timbral features from a recorded spoken utterance appear as a subsequence of features in a target recording, containing the utterance of interest. In our case a query of phoneme models X can be seen as subsequence of the series of MFCC features Y extracted from the whole recording. To this end we define a distance metric for a frame y_m and n^{th} model x_n as a function of the posterior probability

$d(m, n) = -log\, p(y_m|x_n)$

Path computation: For an alignment path p_l we select step sizes $p_l - p_{l-1} = (1, 1), (1, 0), (1, 2)$ according to the notation of [Mueller]. The step size $(0, 1)$ is disallowed because each audio frame has to map to exactly one phonetic model. To counteract the preference for the diagonal step $(1, 1)$ and the skip step $(1, 2)$ local weights w_d and w_s are introduced (as suggested in [Mueller]).

Candidate segment selection: We define an overlapped audio segment to be any segment from the target audio, for which any y_m belongs to more than one path p. Let us call the set of paths for an overlapped audio segment an overlapped path set (see figure 2). Then as candidate segment is considered the audio segment span from the initial timestamp of the leftmost path until the final timestamp of the rightmost path. We assume that a concentration of found paths within a segment indicates at least one true hit.

5. HHMM MODEL

The position in the score in velocity. A unit is different per a composition and is computed so: one minimal unit from score (in number of frames according to a factor of tempo indicated in score). The model moves.

It has two sections S = {Q,F}, where Q is the queried section and F is the filler model

5.1 transition model

velocity variable

asd

position variable

where the modulus operator resets the position to be in a beginning of a new section after it exceeds previous section's duration $D(s_{n-1})$

section variable

$p(s_n|s_{n-1}, p_{n-1,}p_n) = p_s(s_n|s_{n-1}), p_n \leq p_{n-1}1, p_n > p_{n-1}, s_n = s_{n-1}$

A lack of increase in the position is an indicator that a new section should be started.

$p_s(s_n|s_{n-1}) = p_q, s_n = s_{n-1} = Q$ governs the self transitions and can be set according the expected structure of the queried audio. In our case we set p_q=0 and p_f=0.9 to decrease the possibility of having the query section more than once in the candidate audio.

6. EXPERIMENTAL SETUP

A query section consist of 7-10 words.

7. EVALUATION

prec and recall of subsequence-DTW

prec and recall at the end of DTW

8. RESULTS

9. REFERENCES

[1] Dirk Von Zeddelmann, Frank Kurth, and M Muller. Perceptual audio features for unsupervised key-phrase detection. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 257–260. IEEE, 2010.