MODELING OF PHONEME DURATIONS FOR ALIGNMENT BETWEEN POLYPHONIC AUDIO AND LYRICS

Georgi Dzhambazov, Xavier Serra

Music Technology Group, Universitat Pompeu Fabra, Barcelona

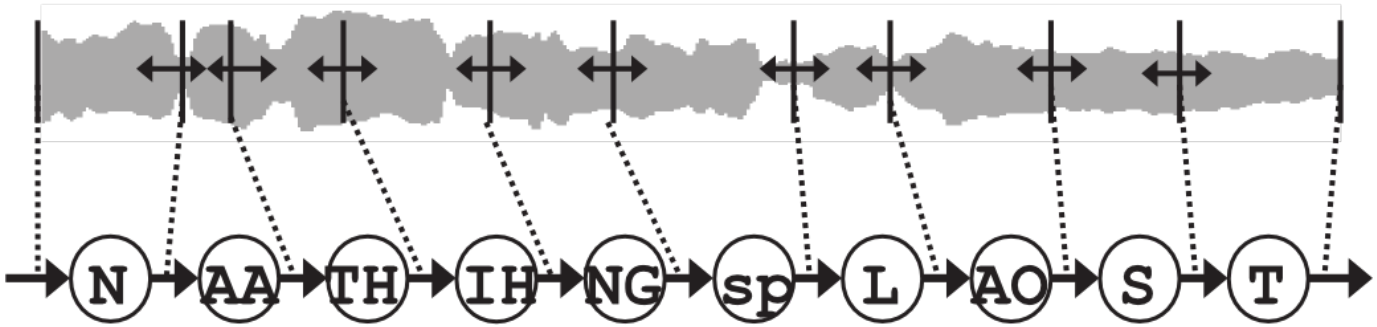
July 31, 2015

Contents

Goal

automatic synchronization between audio and lyrics (a.k.a. lyrics-to-audio alignment)

- input: audio lyrics
- output: boundary timestamps of phonemes
- analogous to score-to-audio alignment



Dataset

Excerpt from classical Turkish Makam

Motivation

- Why automate lyrics-to-audio alignment?
  - automatic karaoke generation
  - structural navigation by lyrics
  - educational purposes: explore interpretations of same lyrics

Lyrics visualization

State of the art

[Mesaros and Virtanen(2008)]
[Fujihara et al.(2011)Fujihara, Goto, Ogata, and Okuno]

	Mesaros et al.	Fujihara et al.
approach	phoneme-HMMs Viterbi forced alignment	
refinements	singing adaptation	singer adaptation
test dataset	English pop	Japanese + English pop

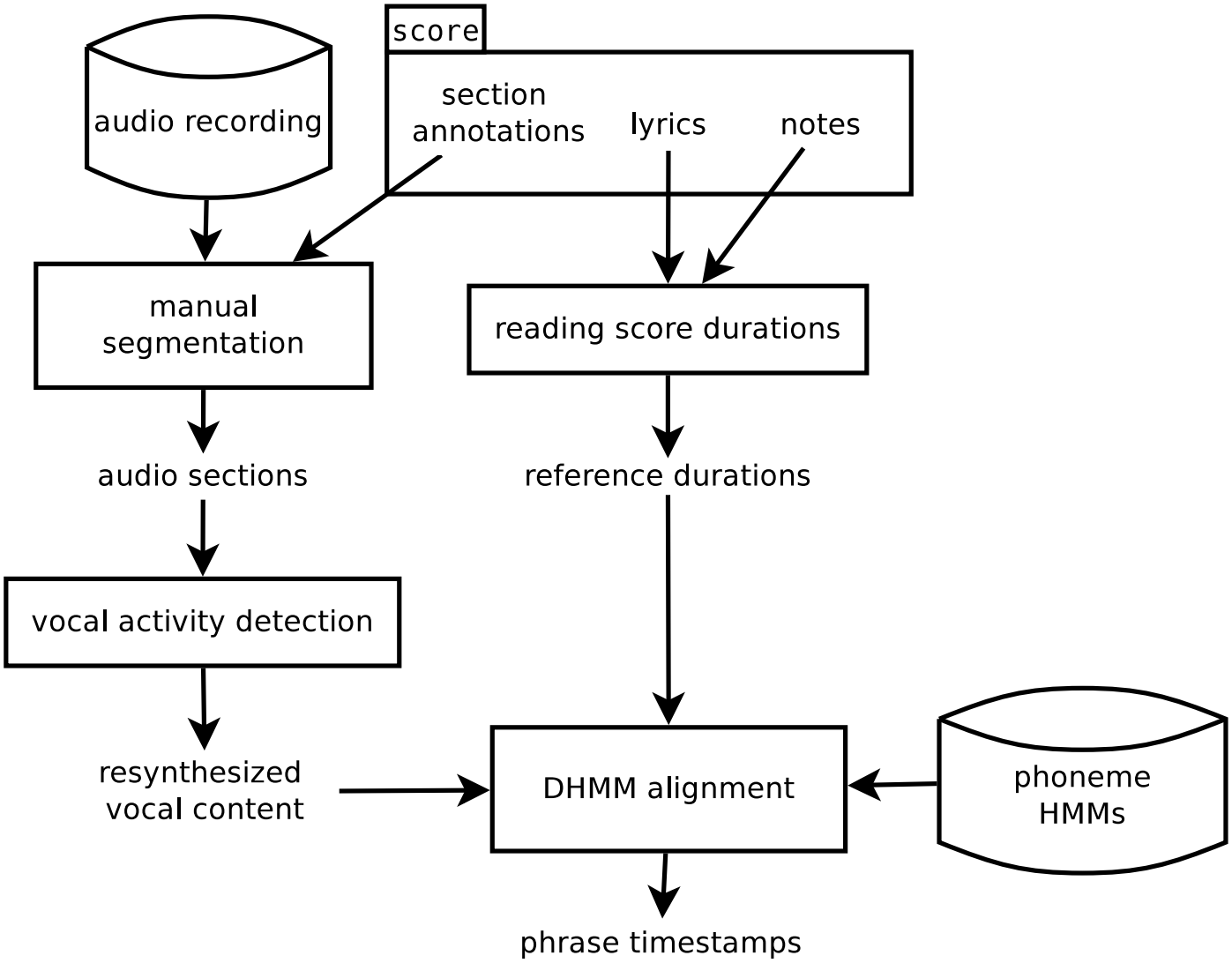
Motivation

- Why research lyrics-to-audio alignment?
  - no work on non-eurogenetic music (singing style and language)
  - state of the art accuracy could be improved
  - build reproducible research

Approach
Main ideas

- extend the phoneme-HMMs Viterbi forced alignment
- explicitly model phoneme durations by use of musical score

Approach Overview



DHMM: Duration-explicit Hidden Markov Model

Dataset

12 recordings (19 min) of classical Turkish Makam

- drawn from CompMusic collection
- segmented into sections
- evaluate phrase boundaries

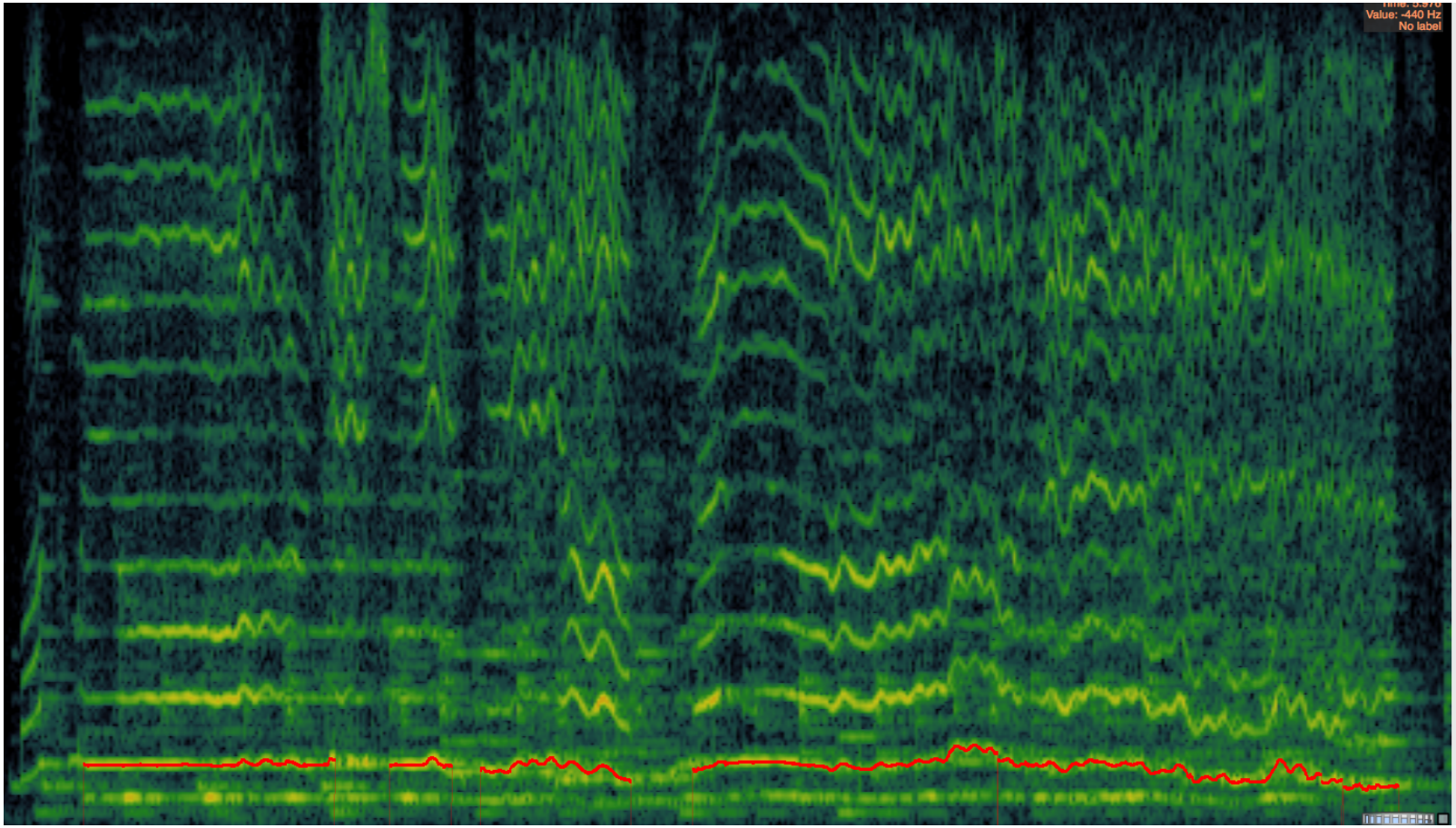
total #phrases	#phrases per section	#words in phrase
220	2 to 5	1-5

Table: Section and phrase statistics for test dataset.

Vocal Activity Detection

original polyphonic audio

Vocal Activity Detection



predominant melody extraction: [Salamon and Gómez(2012)]

- vocal activity detection at the same time

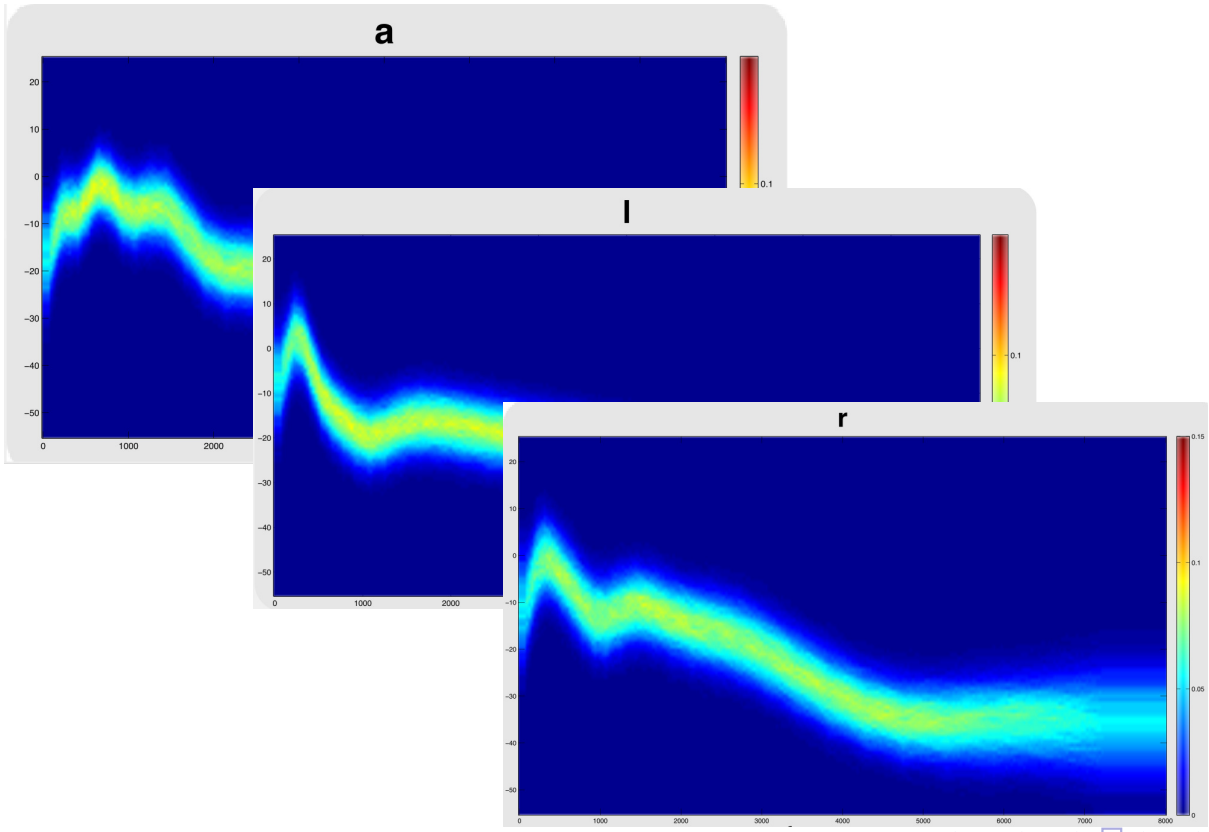
Resynthesis of main vocal

[Serra(1989)]

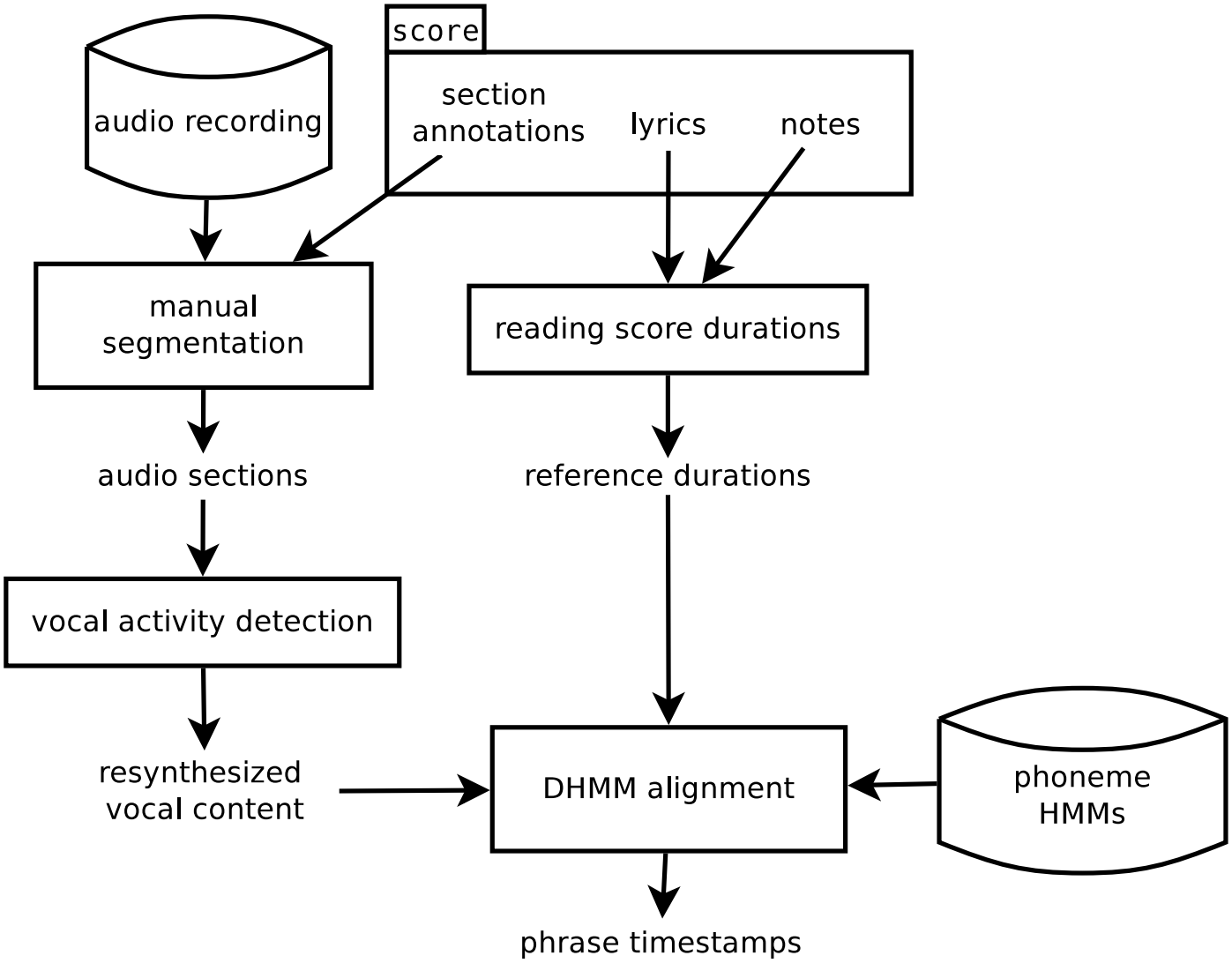
- extract 12 MFCCs from resynthesized vocal

Phoneme models

- 38 Phoneme HMMs trained on clean speech
  - trained on ~6 hours Turkish speech
- $b_i(O_s)$ - from GMM with 9-mixture Normal distribution

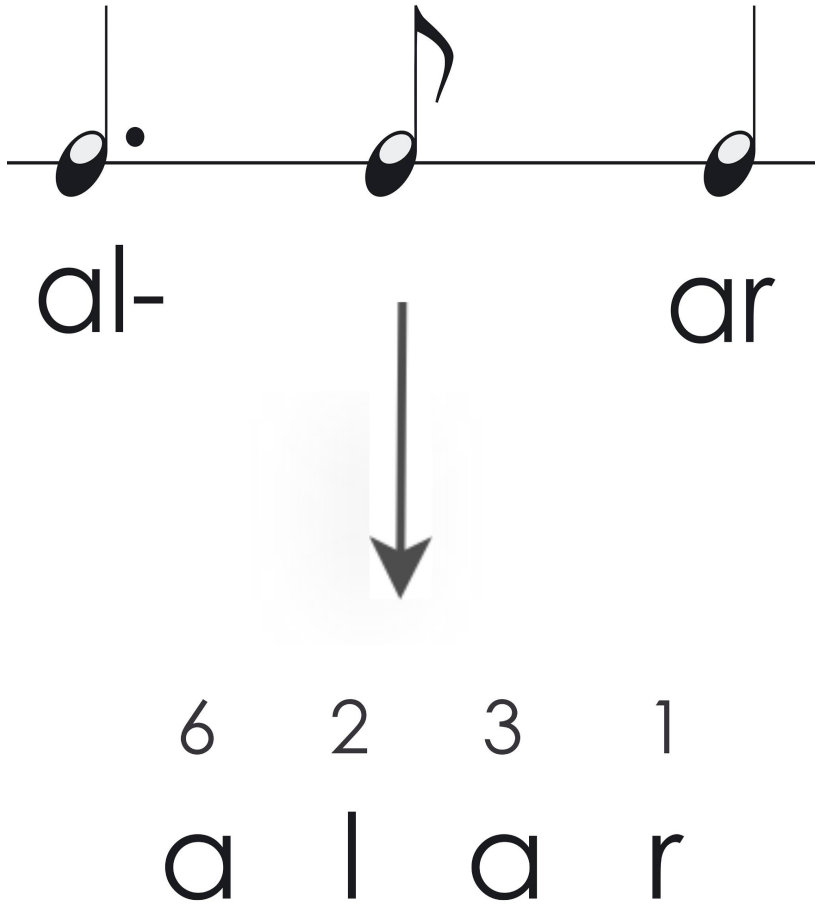


Approach Overview



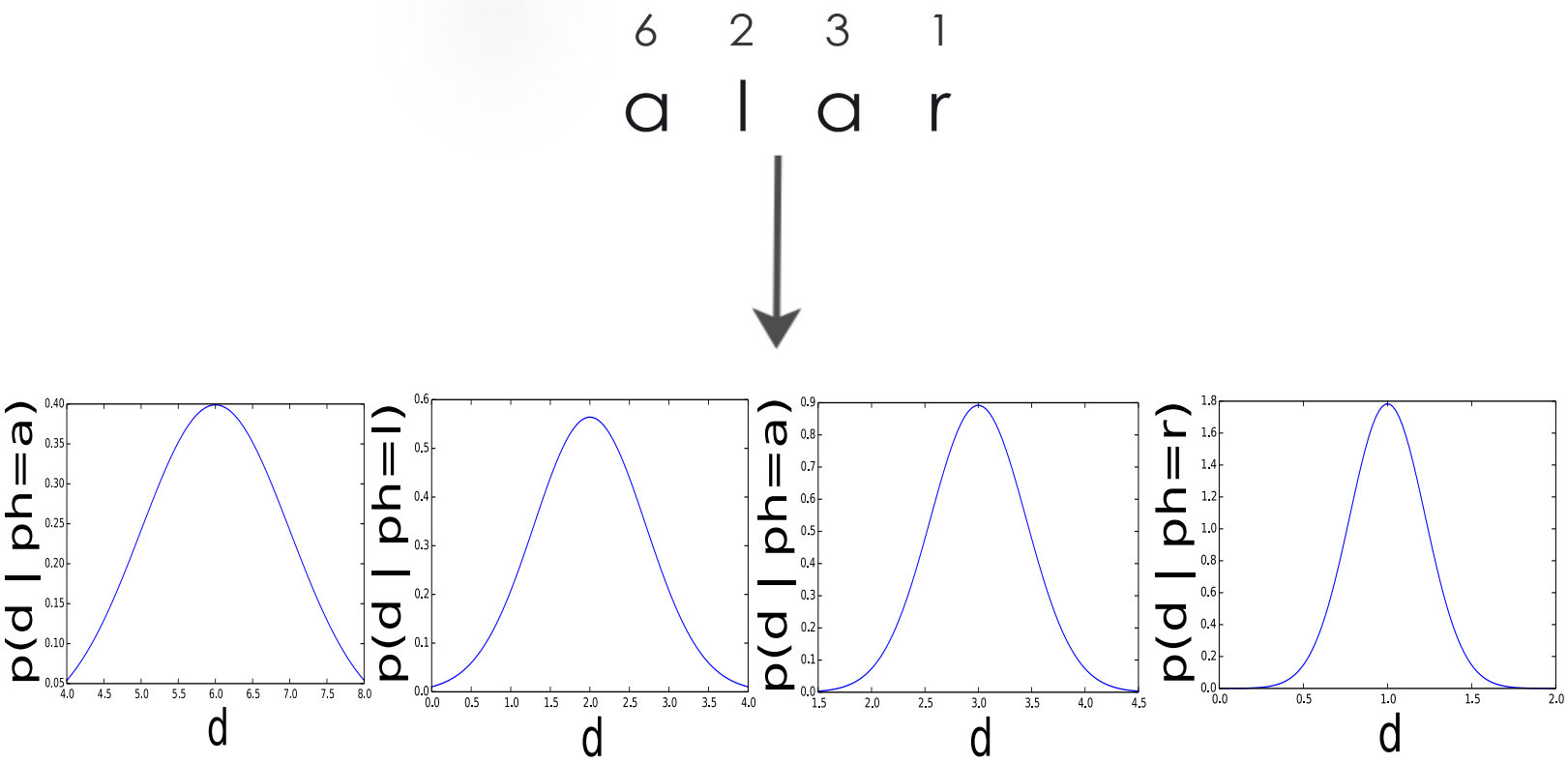
DHMM: Duration-explicit Hidden Markov Model

Read score reference durations



relative reference durations

Model duration distributions $P_i(d)$



durations distributions

Duration-explicit HMM alignment

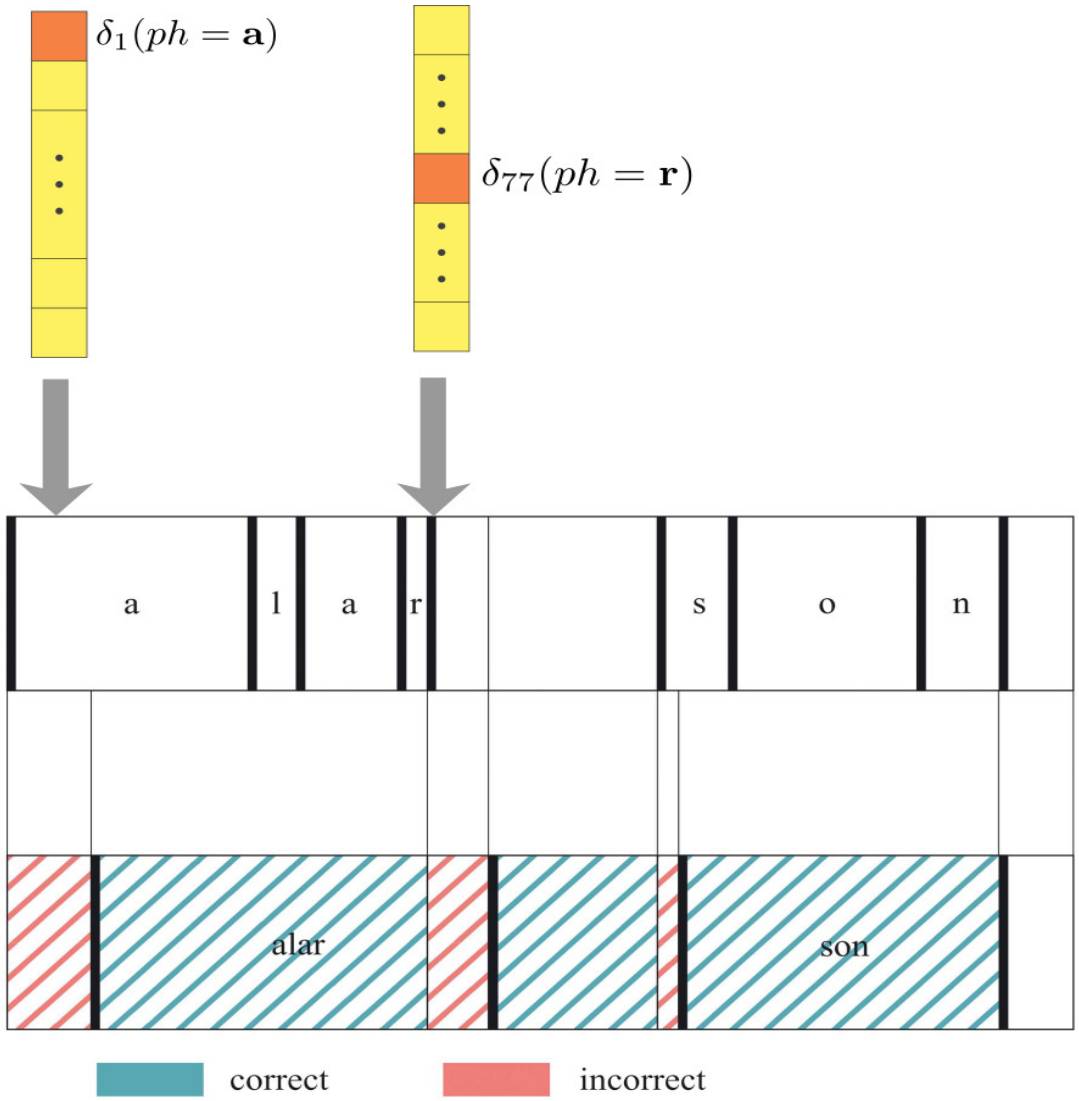
$\delta_t(i)$: probability for optimal path ending in state i at time t

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

and

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

Duration-explicit HMM alignment



Results

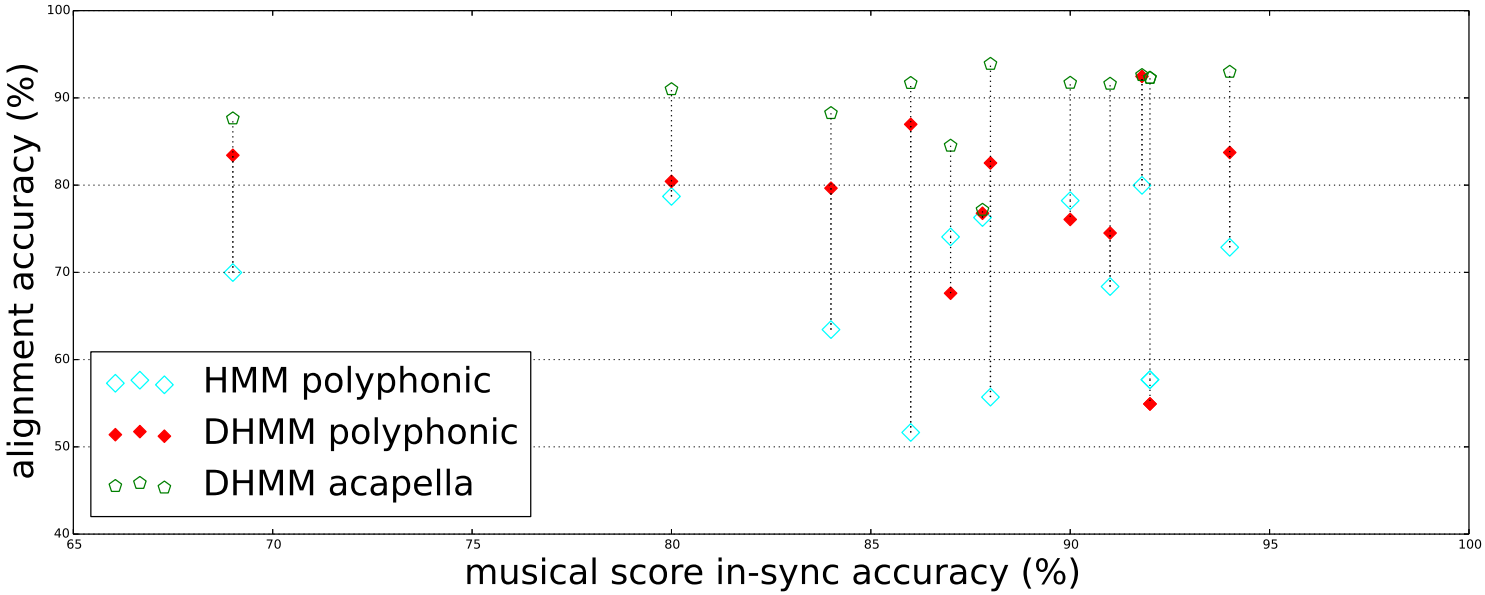
System variant	accuracy	error
HMM polyphonic	67.46	1.04
DHMM polyphonic	77.74	0.63
DHMM acapella	90.04	0.26

Table: Alignment accuracy and error for baseline HMM and DHMM

System variant	accuracy	error
HMM+singing adaptation [2]	-	1.4
HMM+singer adaptation [1]	85.2	-

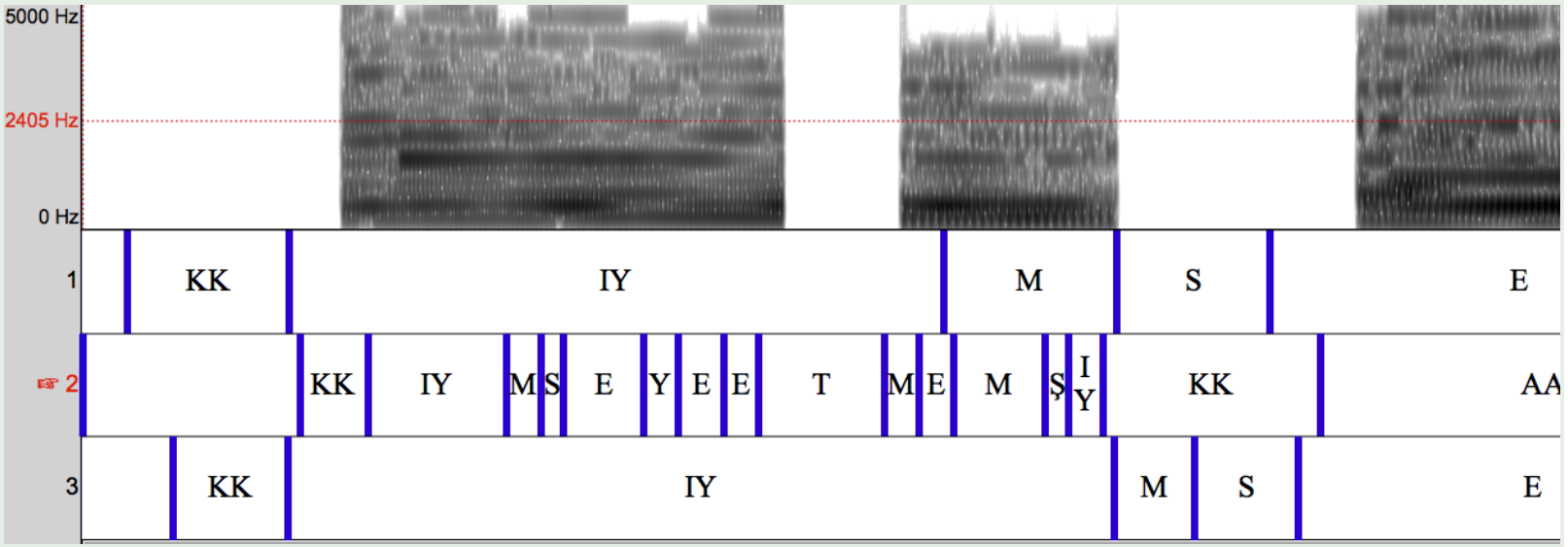
Table: Alignment accuracy and error for related work

Results


alignment accuracy (%)
musical score in-sync accuracy (%)
HMM polyphonic
DHMM polyphonic
DHMM acapella
100
90
80
70
60
50
40
65
70
75
80
85
90
95
100

Results

Example phoneme-level alignment



Conclusion and Future

- extension to a HMM decoding with duration modeling
  - reference durations extracted from score
  - allows tracking duration variability
- vocal segregation by harmonic modeling

Conclusion and Future

- extension to a HMM decoding with duration modeling
  - reference durations extracted from score
  - allows tracking duration variability
- vocal segregation by harmonic modeling

- improve vocal extraction
- rely on automatic note-segmentation instead of score

References


Hiromasa Fujihara, Masataka Goto, Jun Ogata, and Hiroshi G Okuno.
Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics.
Selected Topics in Signal Processing, IEEE Journal of, 5(6):1252–1261, 2011.


Annamaria Mesaros and Tuomas Virtanen.
Automatic alignment of music audio and lyrics.
In Proceedings of the 11th Int. Conference on Digital Audio Effects (DAFx-08, 2008.


Justin Salamon and Emilia Gómez.
Melody extraction from polyphonic music signals using pitch contour characteristics.
Audio, Speech, and Language Processing, IEEE

Questions and answers

georgi.dzhambazov@upf.edu
http://twitter.com/georgi_d_
https://github.com/georgid/AlignmentDuration
http://compmusic.upf.edu/