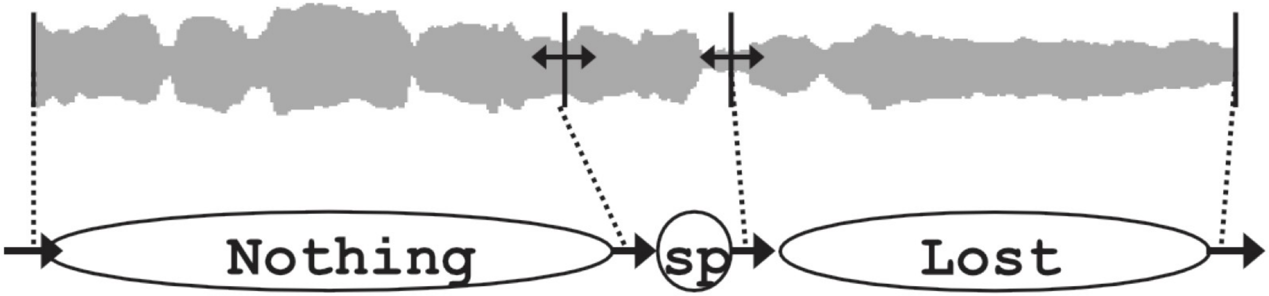
Automatic Alignment of Long Syllables in A Cappella Beijing Opera

Georgi Dzhambazov, Yile Yang, Rafael Caro, Xavier Serra
CompMusic Project – Universitat Pompeu Fabra
georgi.dzhambazov@upf.edu

6th International Workshop on Folk Music Analysis
Dublin Institute of Technology
17 June 2016

Introduction

What is lyrics-to-audio alignment? automatic matching between an audio recording and its lyrics: phrases/words/syllables



State-of-the-art approaches: (Fujihara, 2012):

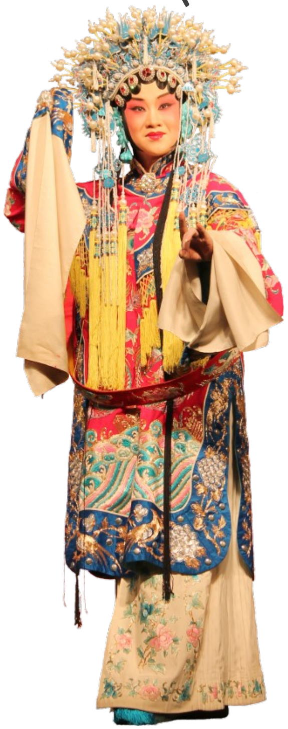
methodology	training	evaluation dataset
phoneme recognizer	speech	English pop, Japanese pop, Cantonese pop

- Most work is on pop music with speech-adopted approach

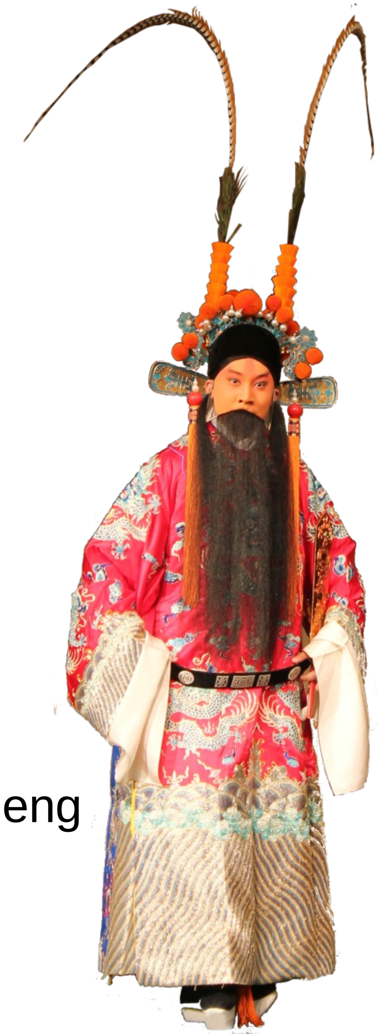
What is Beijing opera (a.k.a. Jingju)

- Unique singing style
- Language: Mandarin + dialects
- Different metrical patterns (*banshi*):
  - *manban* (slow)
  - *kuaiban* (fast)

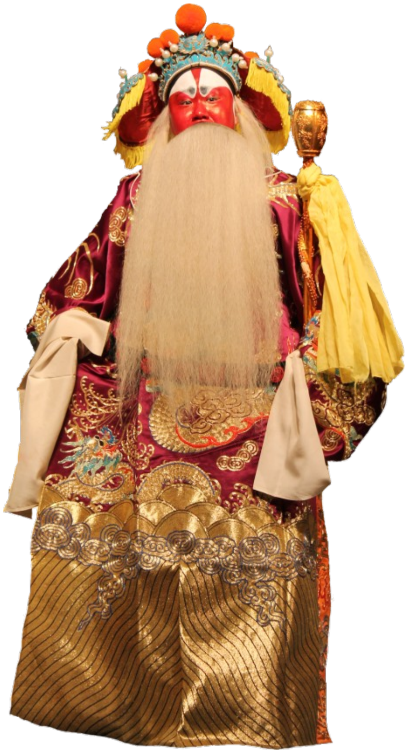
- Different role types



dan

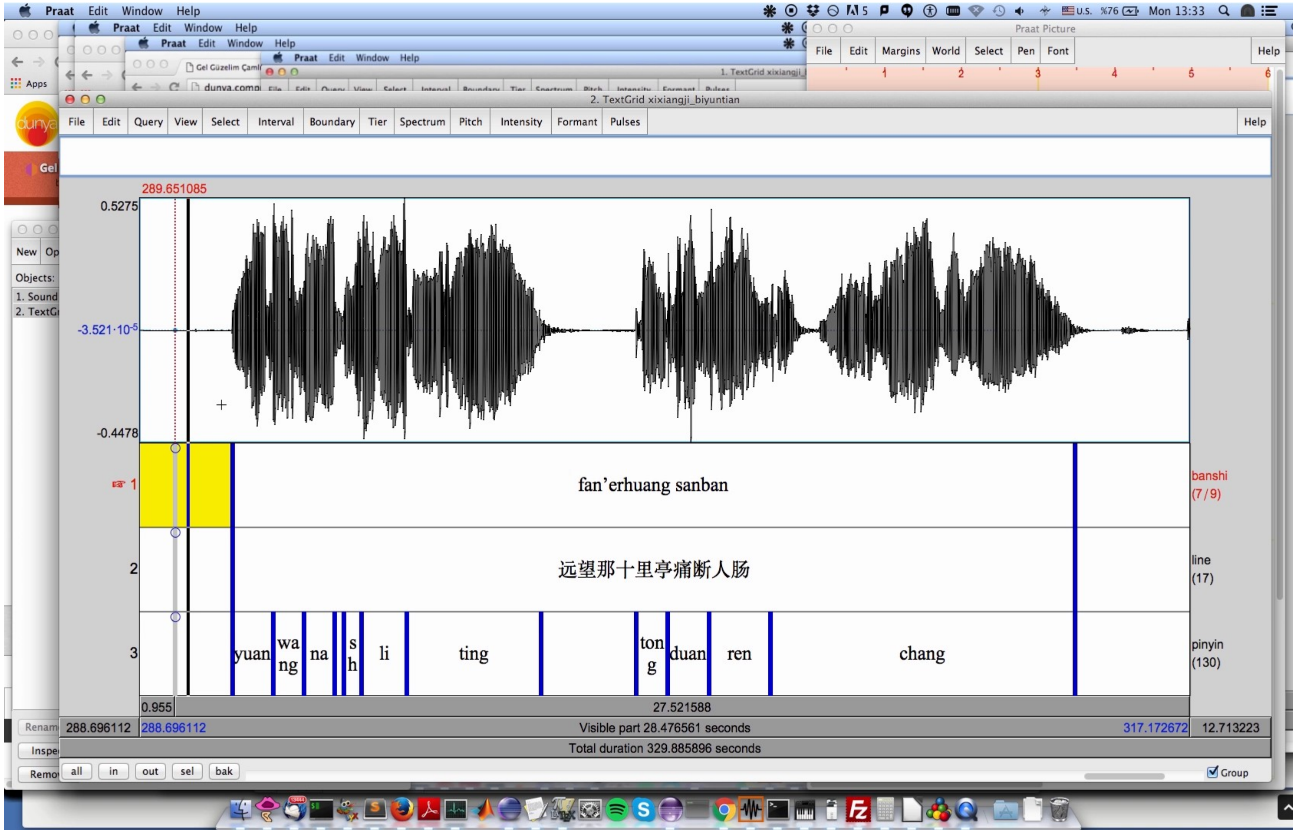


laosheng



jing

Example excerpt from an aria



Average duration: 3.1 sec
Max duration: 8 sec

Example excerpt from an aria: performance of speech-recognition-based alignment

Motivation (why automatic alignment)

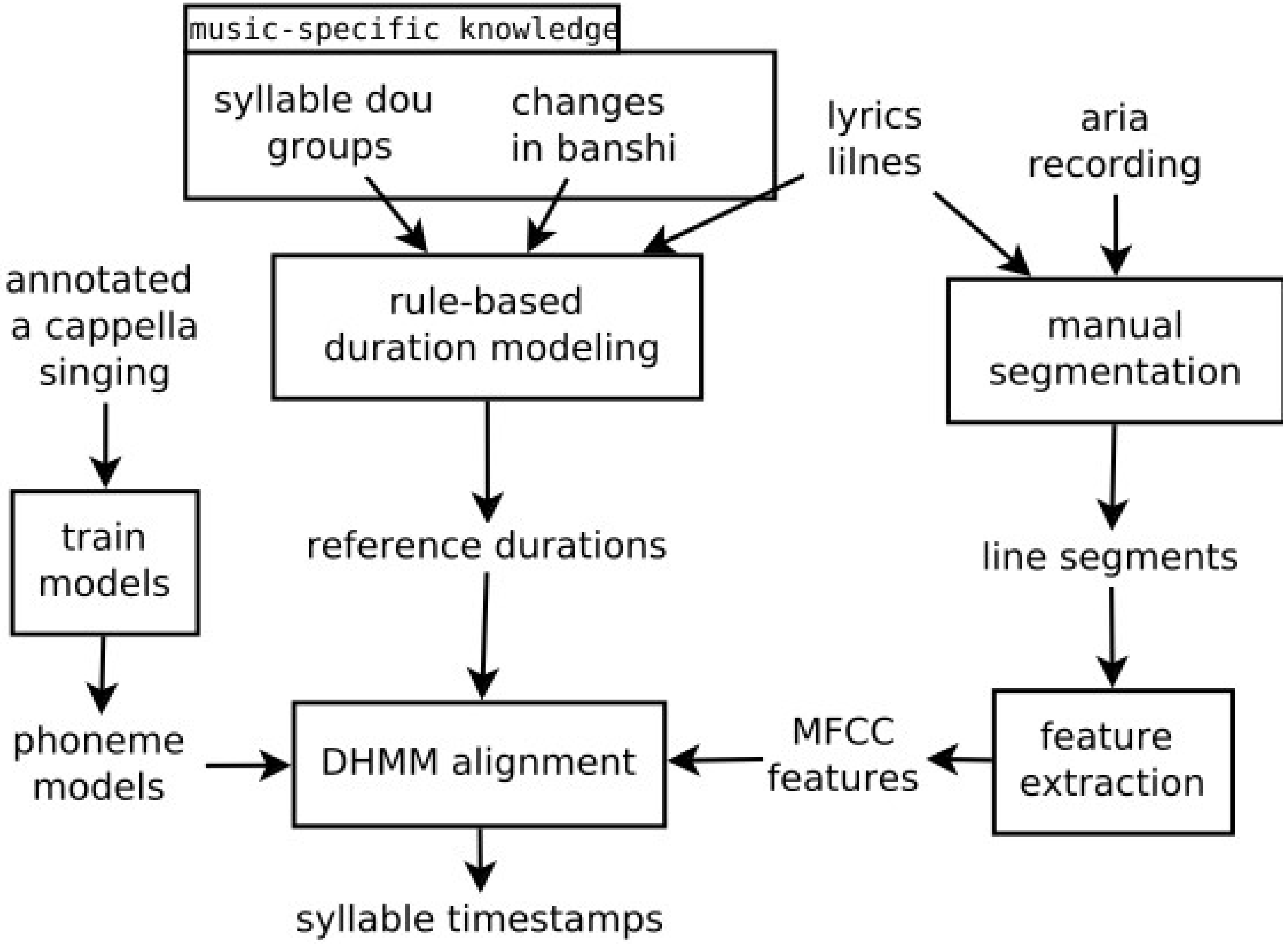
- Alignment: fundamental for following the opera script (Theatre performance)
- Application of state-of-the-art:

  not satisfactory results

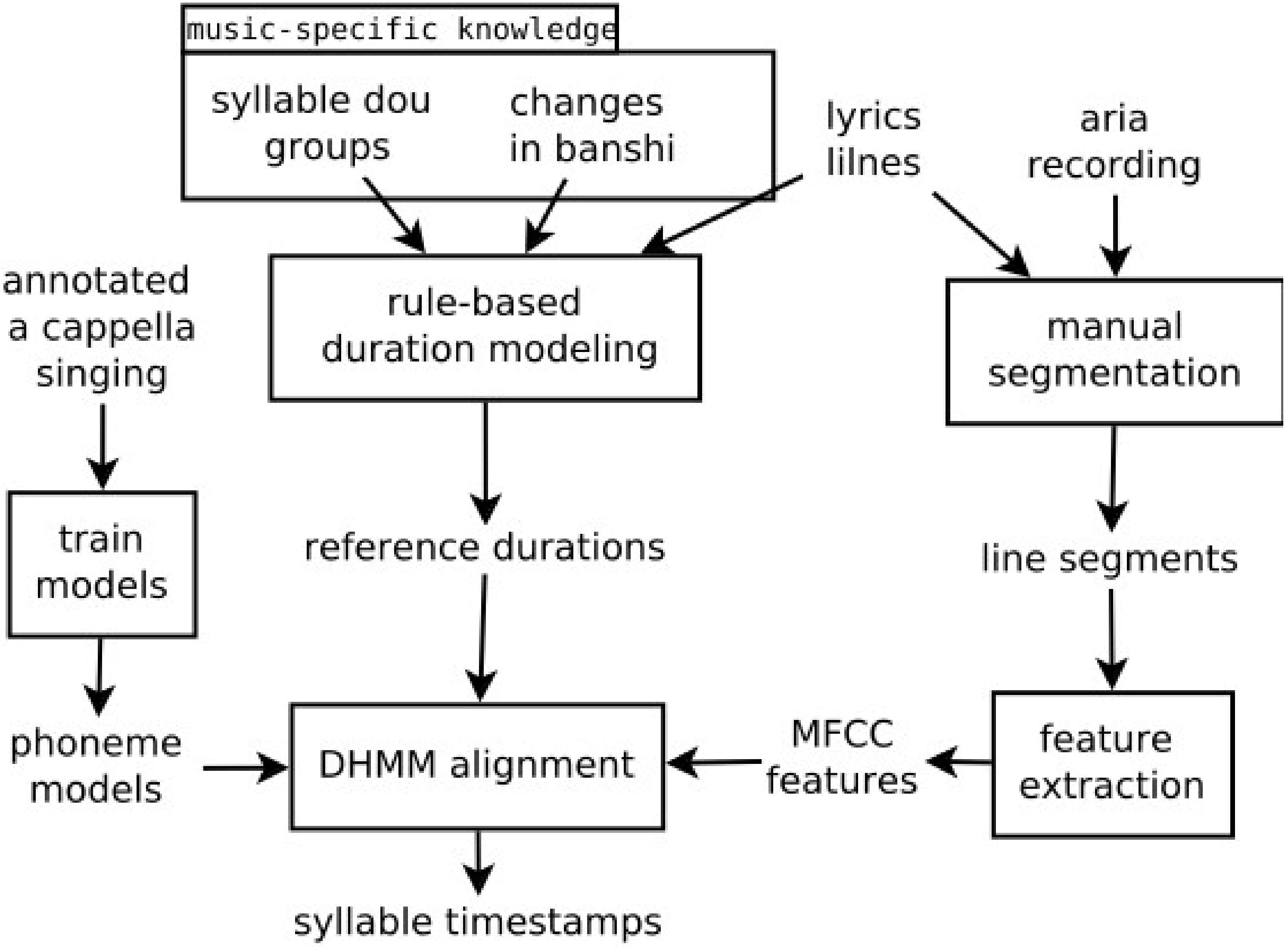
  Unaware of music-specific knowledge
- Can be further used:
  - To facilitate musicological studies (navigation)
  - In Education:

    Aid singing students and teachers (compare performances)

Approach Overview



Approach Overview



music-specific knowledge

syllable dou groups

changes in banshi

lyrics lilnes

aria recording

annotated a cappella singing

rule-based duration modeling

manual segmentation

train models

reference durations

line segments

phoneme models

DHMM alignment

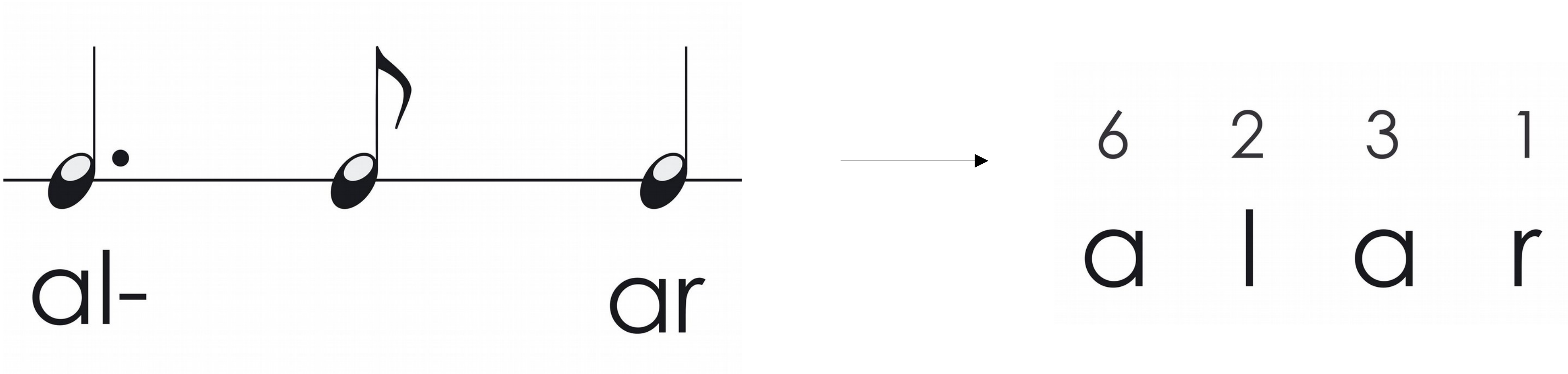
MFCC features

feature extraction

syllable timestamps

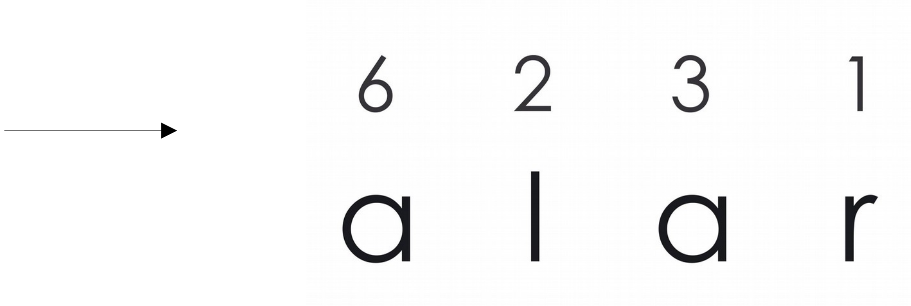
CONTRIBUTION with colors

Syllable reference durations: model from score?



- (Dzhambazov et al. 2015)
- Restriction: need of score

Syllable reference durations: model from training data?

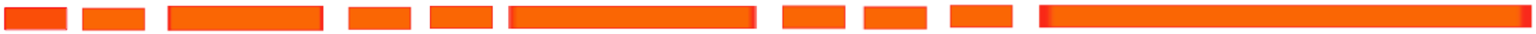


- (Kruspe et al. 2015)
- Restriction: looses context: position of a syllable in a line (same duration)

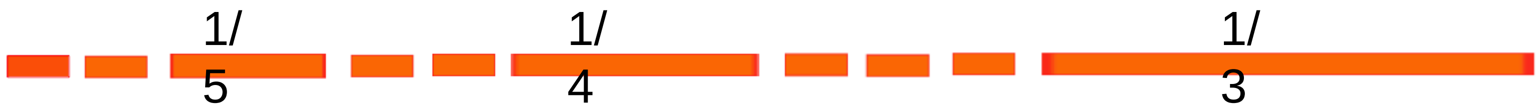
Jingju-specific principles

- Lyrics from poetry: divided into lines
- Each line has 2 or 3 dou

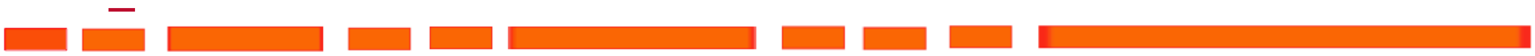
玉堂春含悲泪忙往前进，
想起了当年事好不伤情！
每日里在院中缠头似锦，
到如今只落得罪衣罪裙。



Syllable duration rules



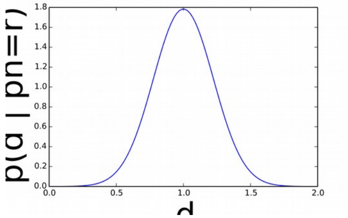
Duration-explicit hidden Markov model



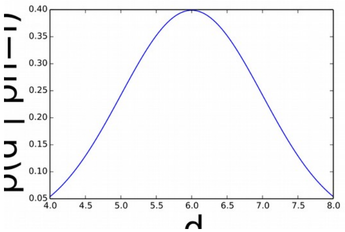
1/ 5

1/ 4

1/ 3



....



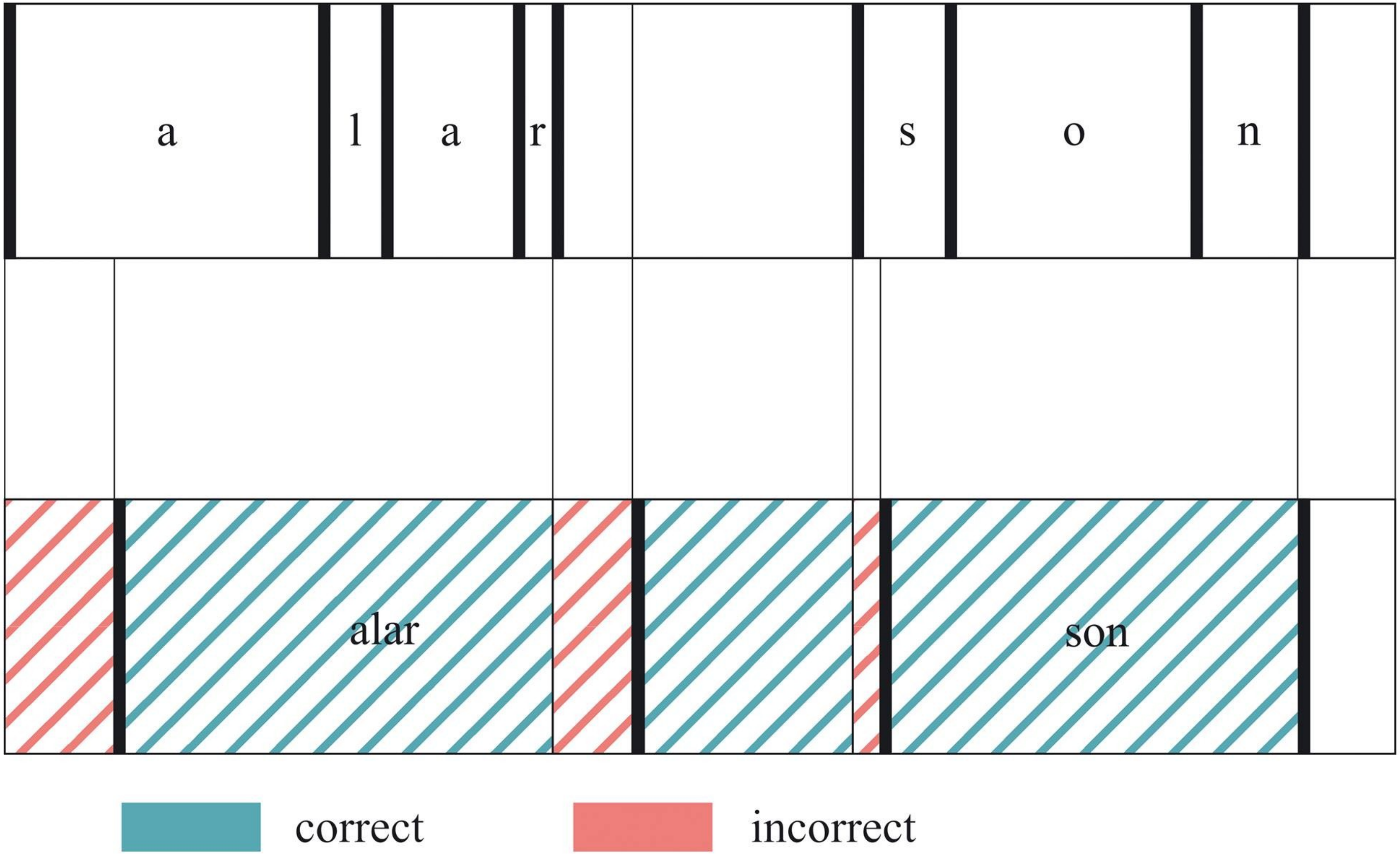
Duration-explicit hidden Markov model

Phoneme models

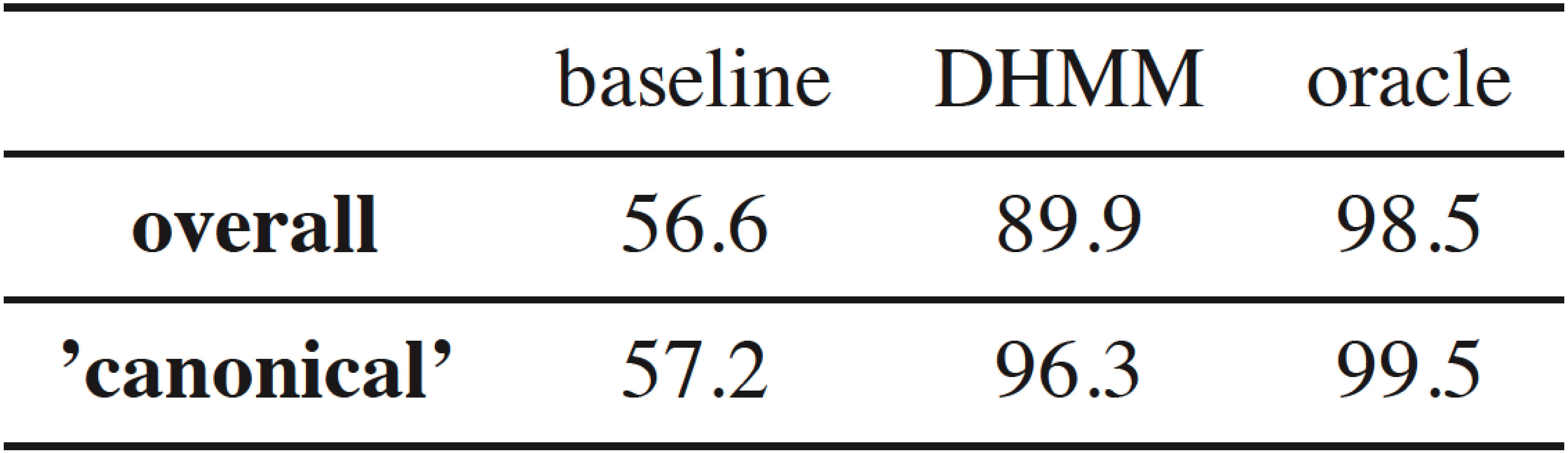
- GMM + MFCCs
- Trained on singing voice

stats

Evaluation



Results



	baseline	DHMM	oracle
overall	56.6	89.9	98.5
’canonical’	57.2	96.3	99.5

Demo

–