Universitat Pompeu Fabra Barcelona

MTG Music Technology Group

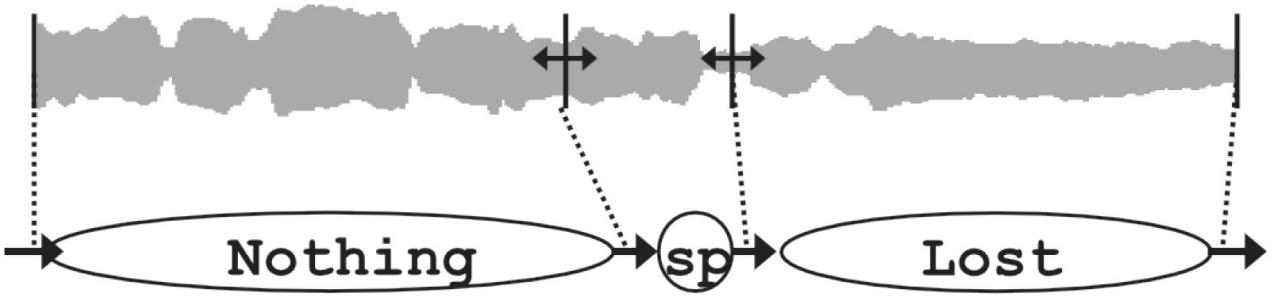
Automatic Alignment of Long Syllables in A Cappella Beijing Opera

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: overview in (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)

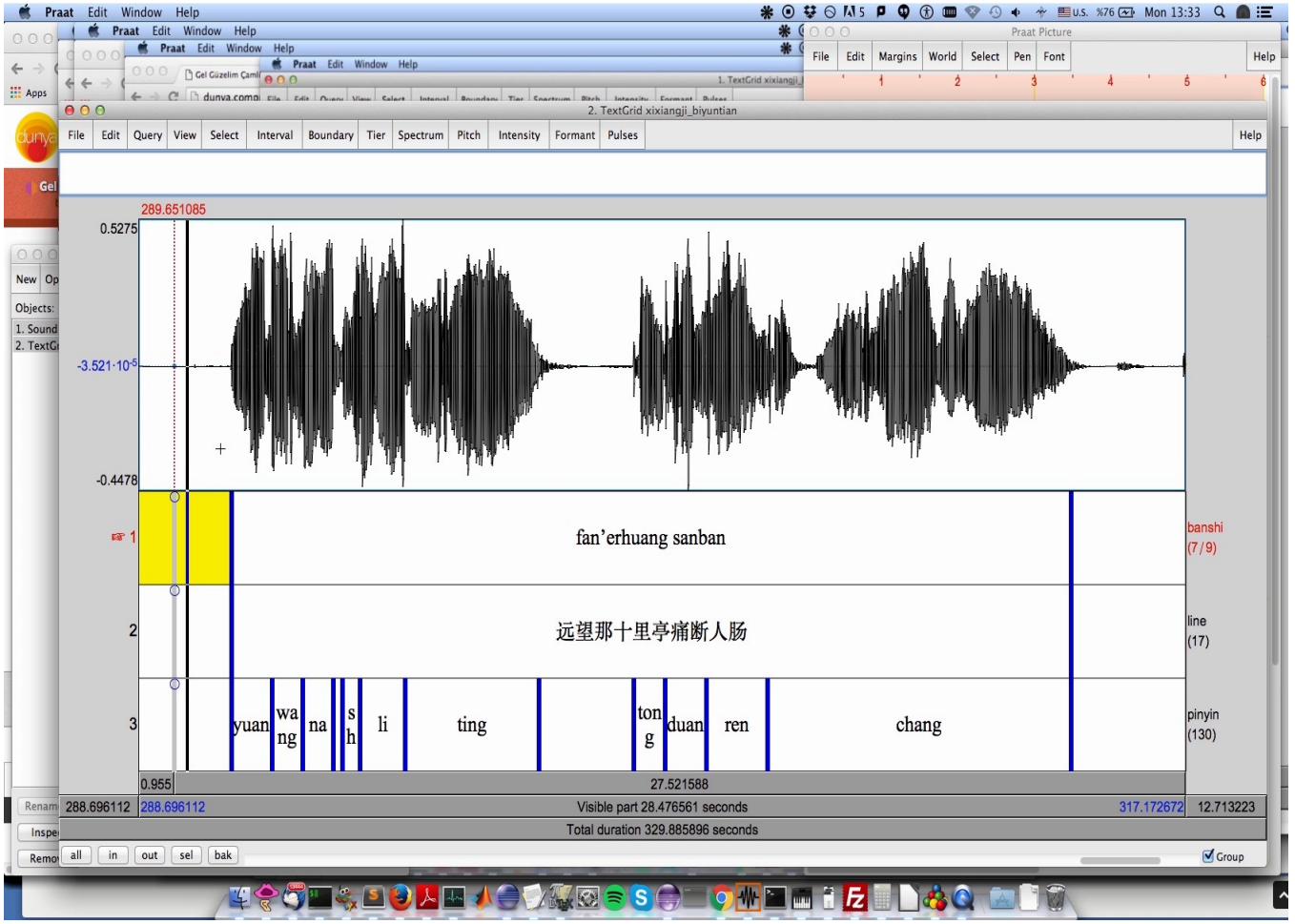
- Different role types

dan

laosheng

jing

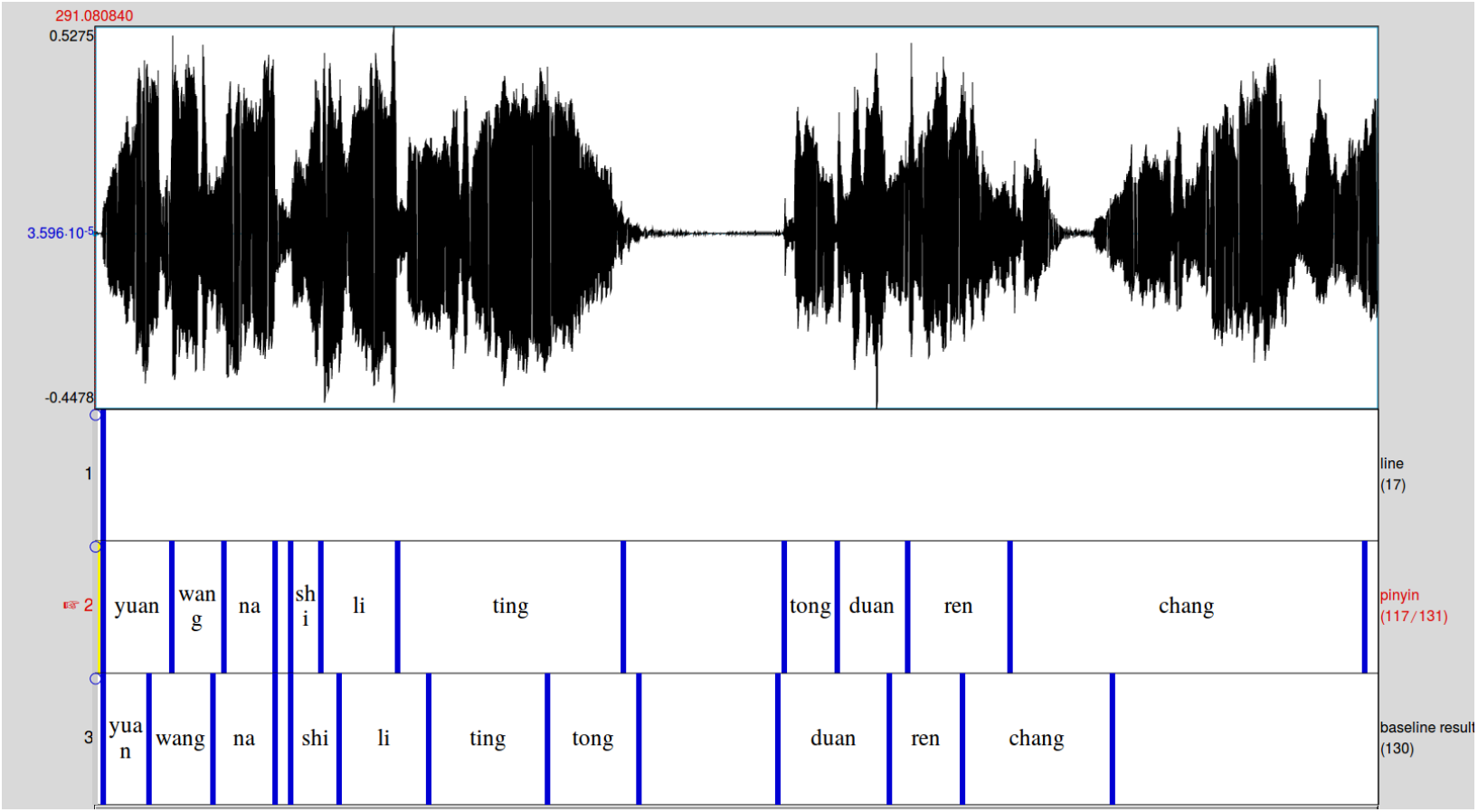
Example excerpt from an aria



Average duration: 3.1 sec

Max duration: 8 sec

Example excerpt from an aria: performance of baseline with phonetic recognizer



- Premature transition to next syllable for long syllables

Motivation

Why alignment?

To facilitate navigation

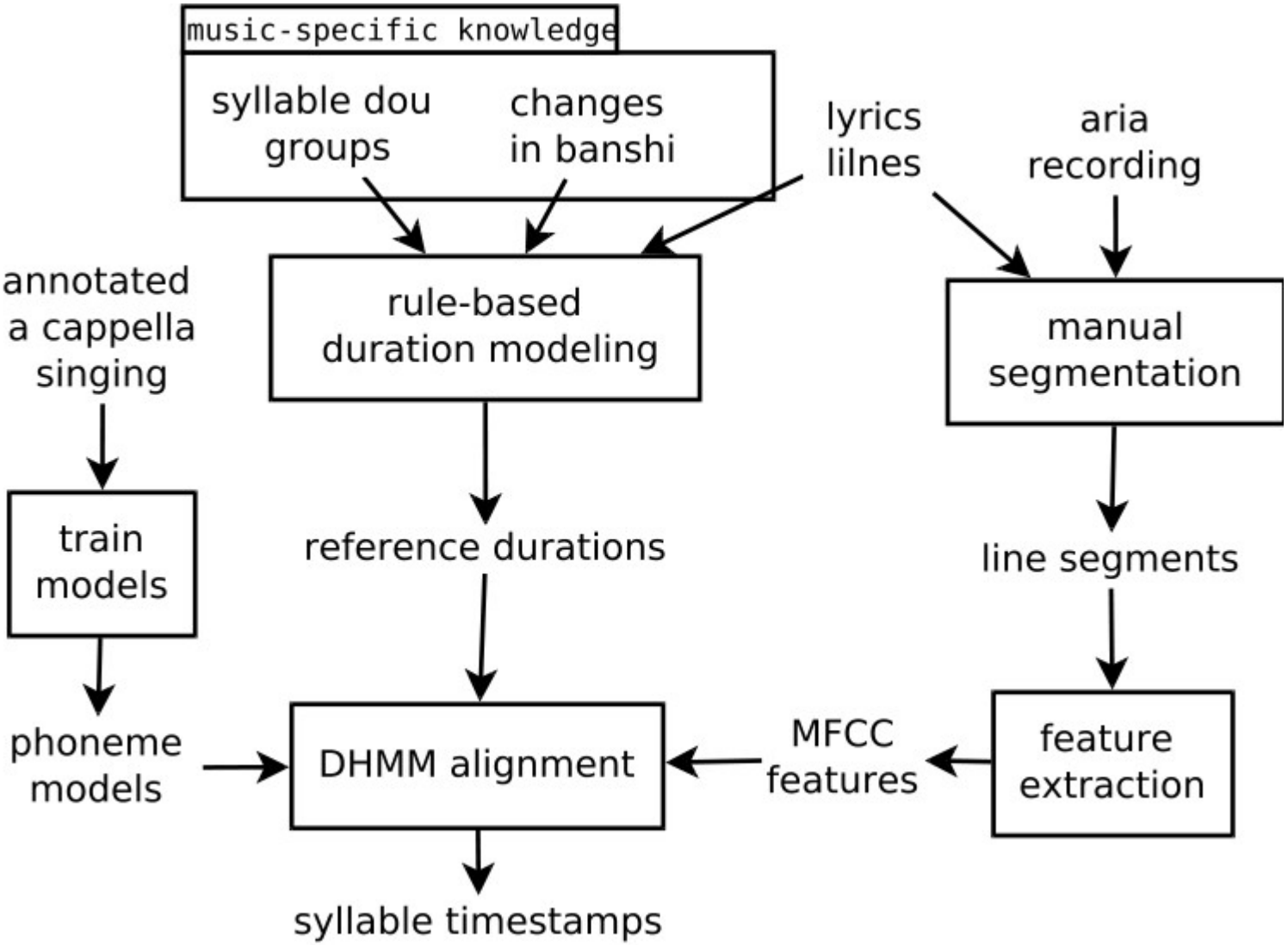
- For musicological studies
- For singing students and teachers

Why design a new alignment method?

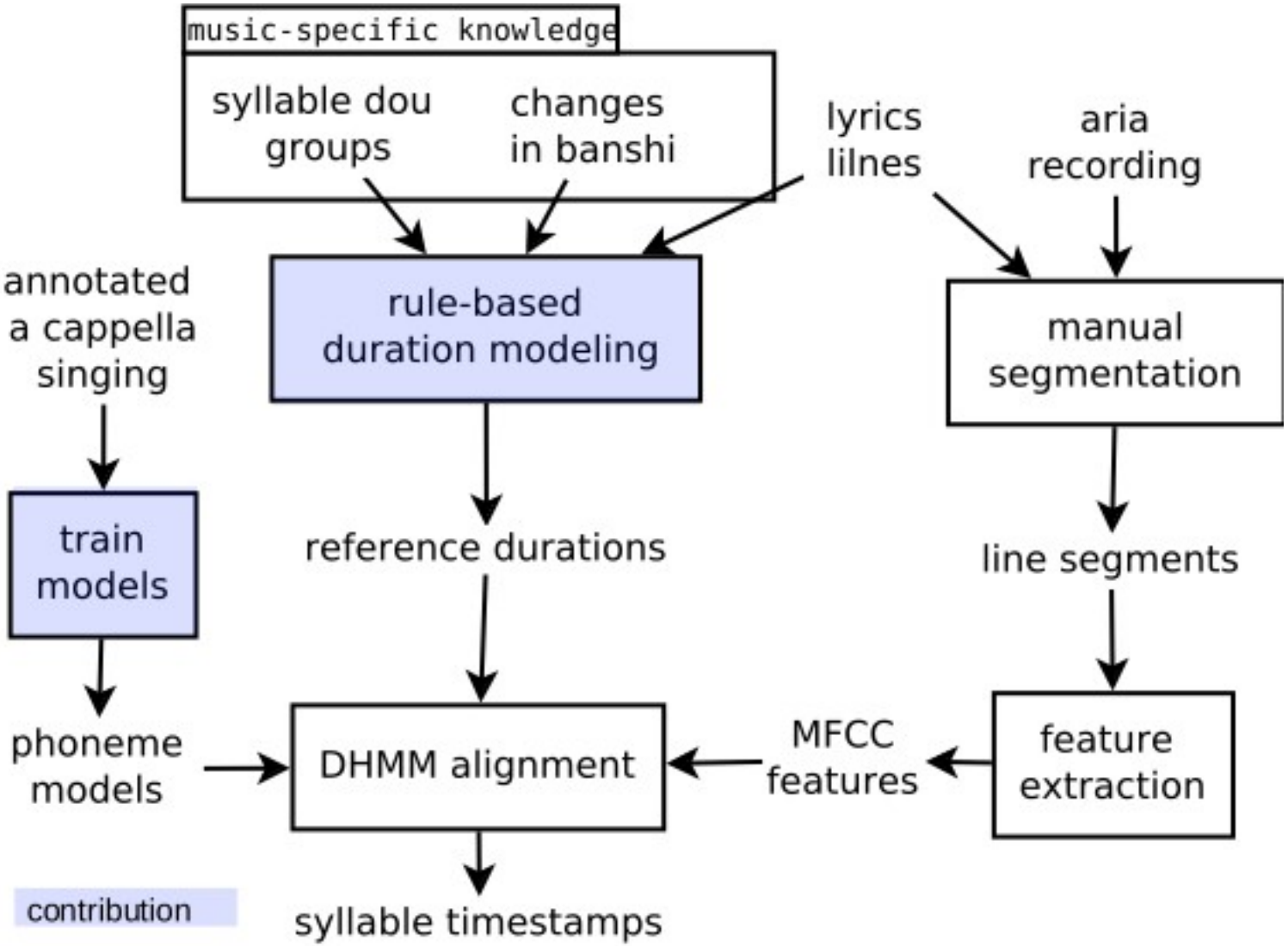
Application of state-of-the-art phonetic recognizer:

- Not satisfactory results
- Unaware of music-specific knowledge

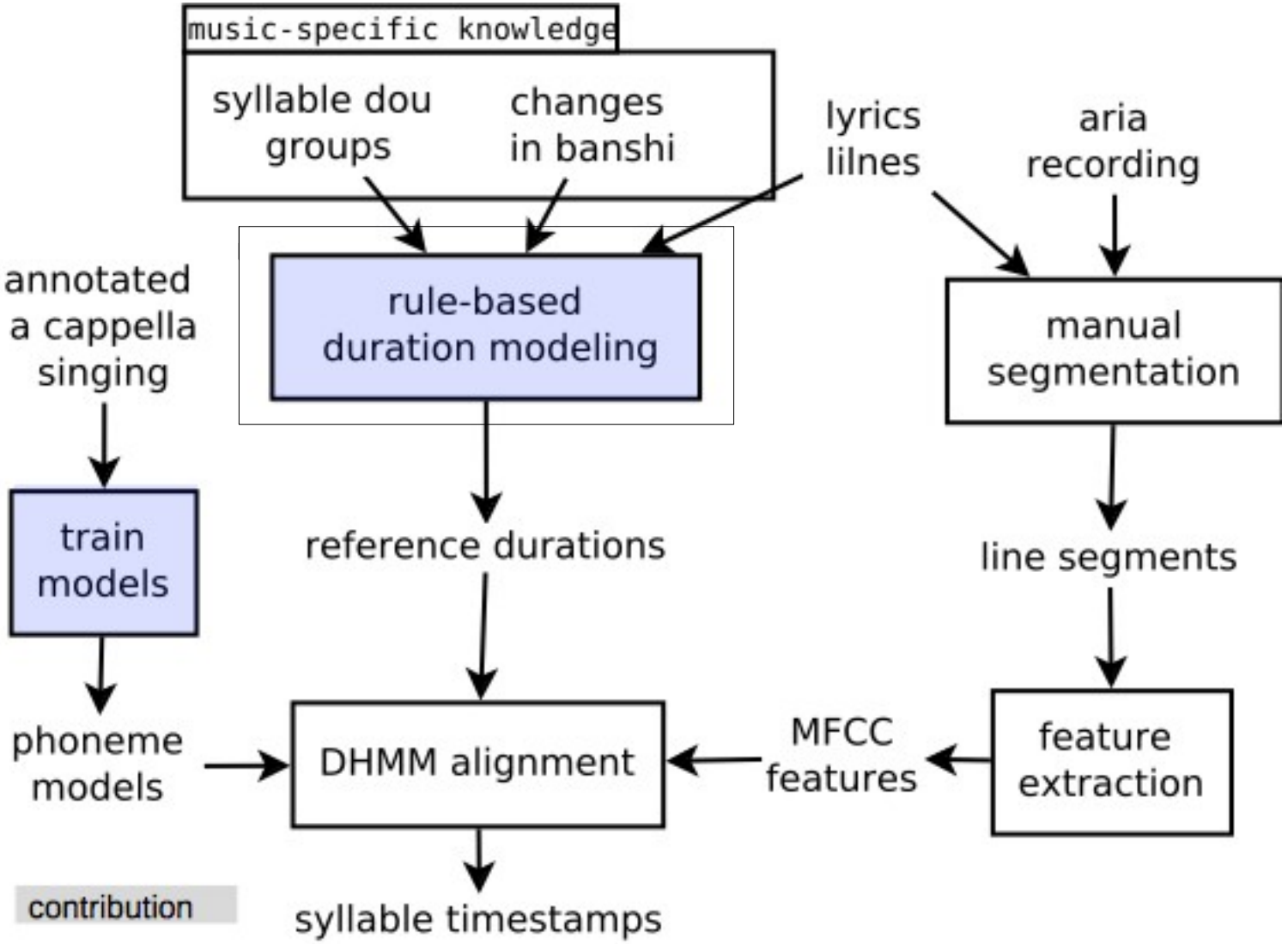
Approach Overview



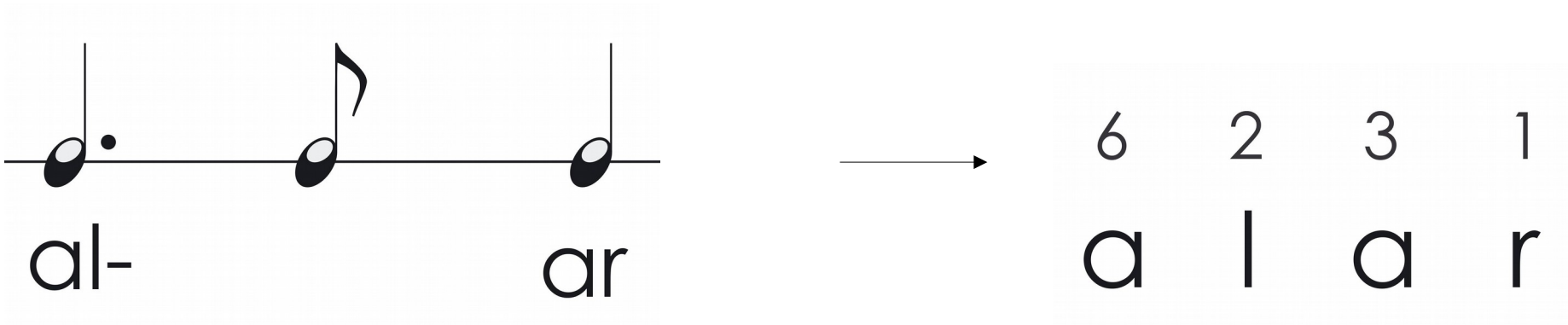
Approach Overview



Approach Overview

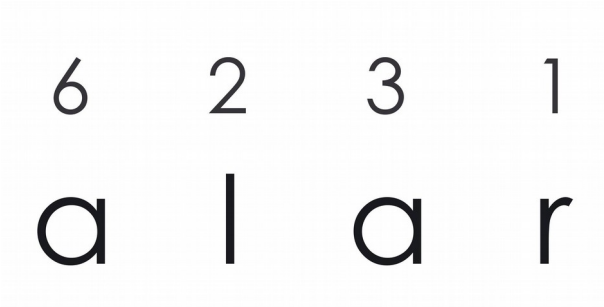


Syllable reference durations: model from score?



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

Syllable reference durations: model from training data?



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

Jingju-specific principles

- Lyrics from poetry: divided into lines
- Each line has 2 or 3 *dou* groups
- Each *dou ends* with a *key syllable*

玉堂春含悲泪忙往前进，

想起了当年事好不伤情！

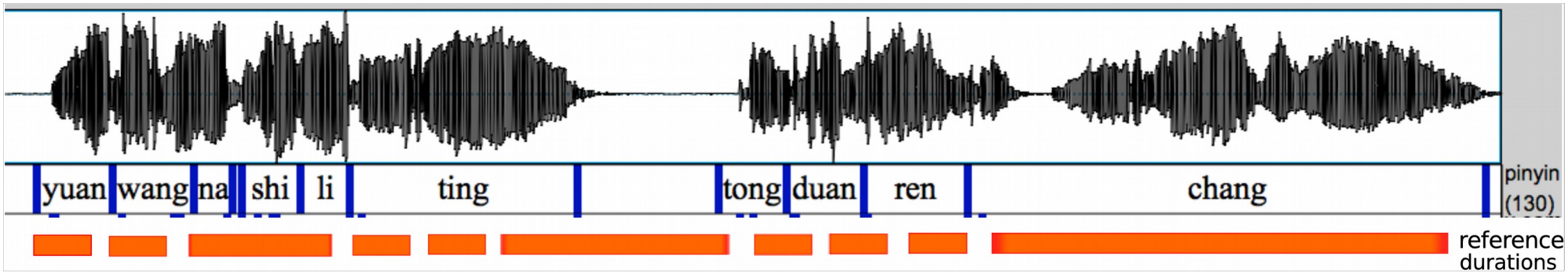
每日里在院中缠头似锦，

到如今只落得罪衣罪裙。

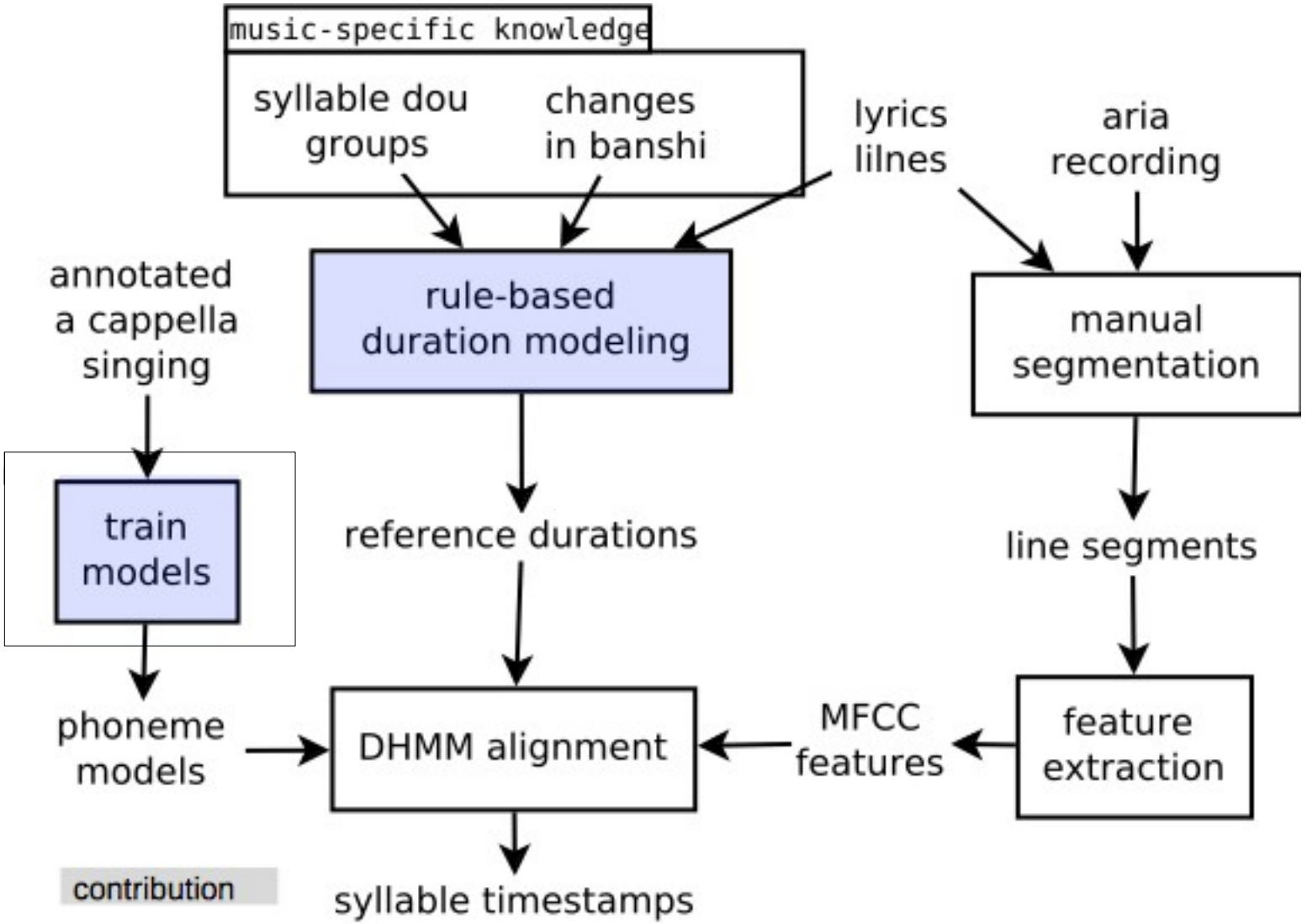
Derivation of phoneme reference durations

- Assign ratios from total line durations to *key syllables*
- Split syllable durations into phoneme reference durations R_i

1/5 1/4 1/3

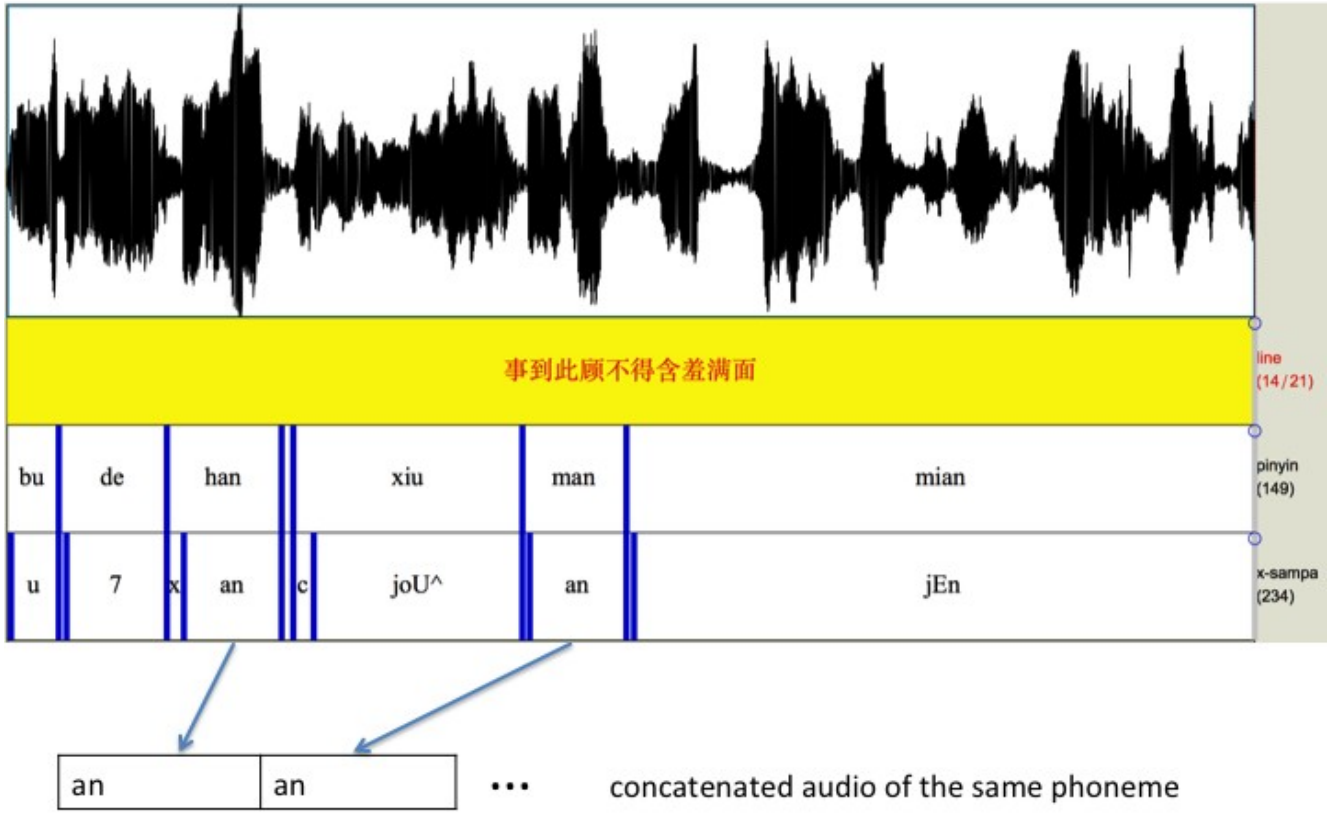


Approach Overview

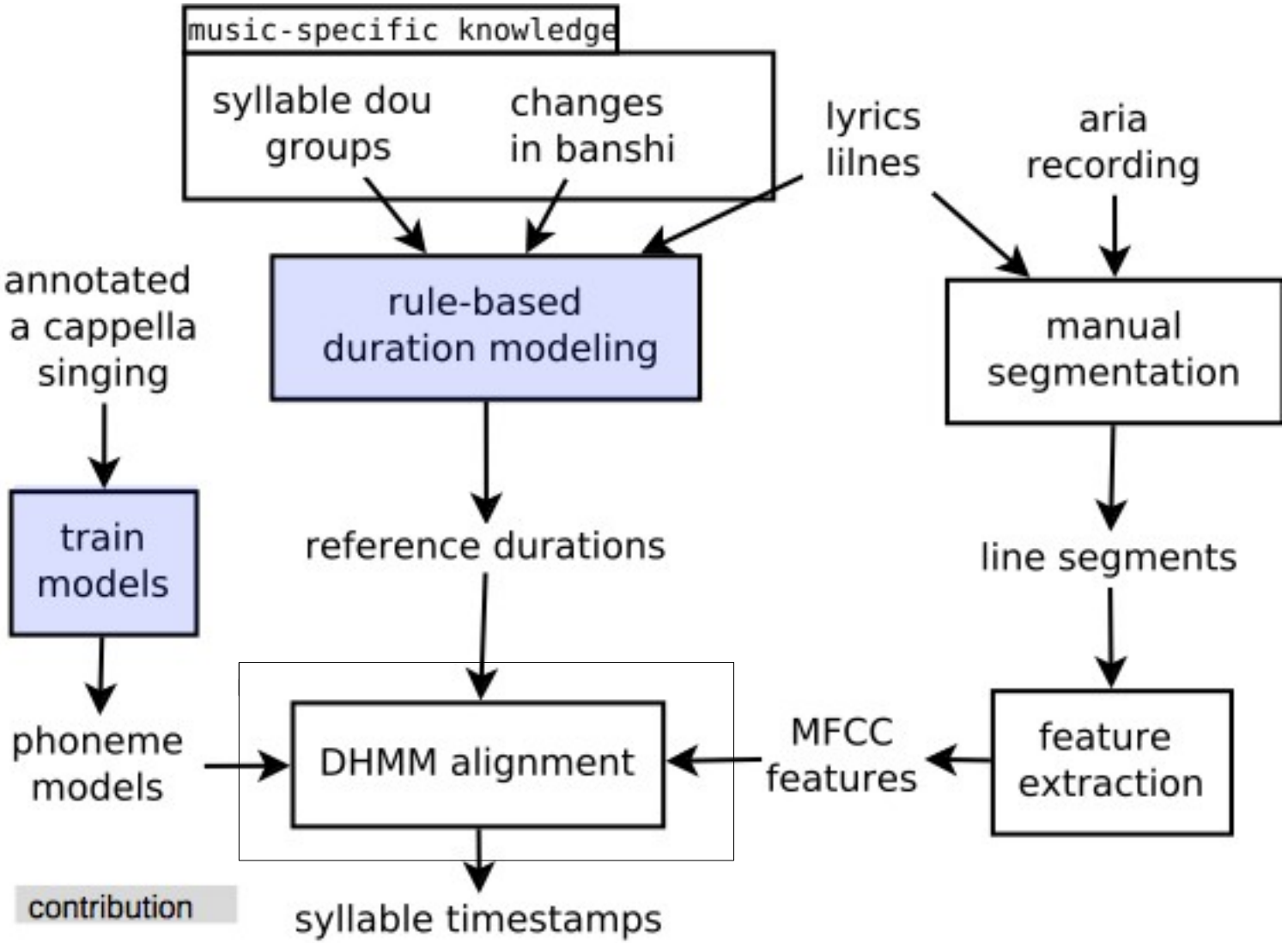


Phoneme models

- Train on singing voice with concatenated excerpts
- 29 phonemes + silence: GMM with MFCCs

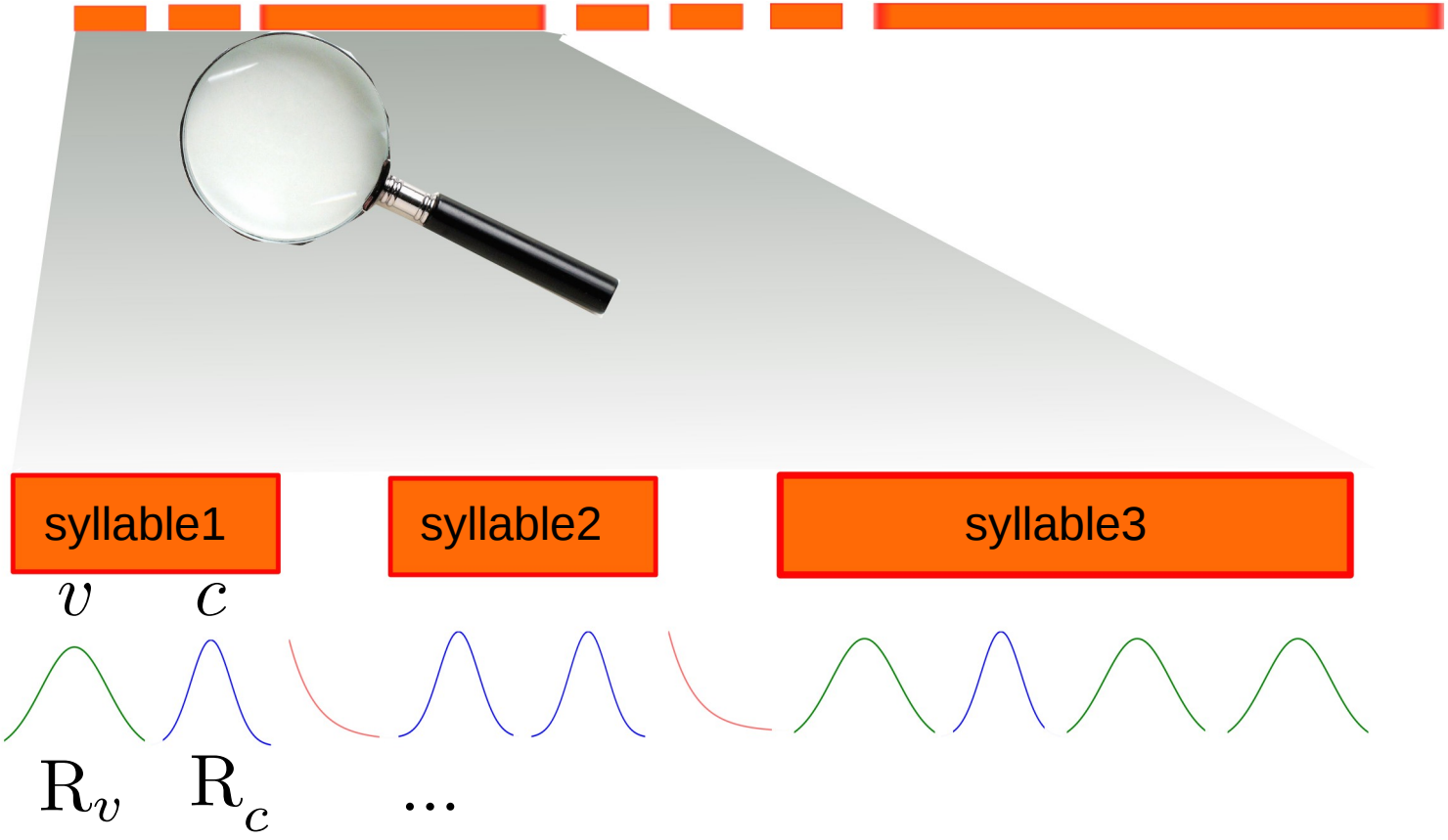


Approach Overview



Duration-explicit hidden Markov model (DHMM)

- Assign normal distributions centered at phoneme reference durations R_i
- Assign exponential distributions at inter-word silences



Duration-explicit hidden Markov model

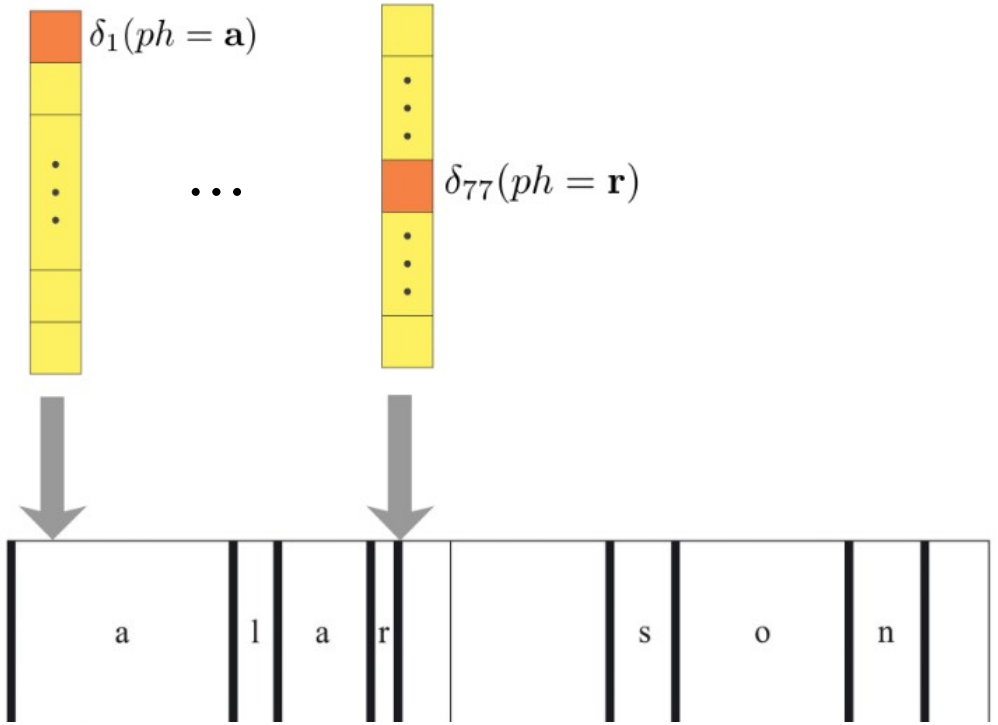
- Consider the sequence of phonemes as a HMM
- Forced Viterbi alignment
  - Maximize duration at each phoneme according to assigned distributions

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

Duration-explicit hidden Markov model

- Consider the sequence of phonemes as a HMM
- Forced Viterbi alignment
  - Maximize duration at each phoneme according to assigned distributions

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

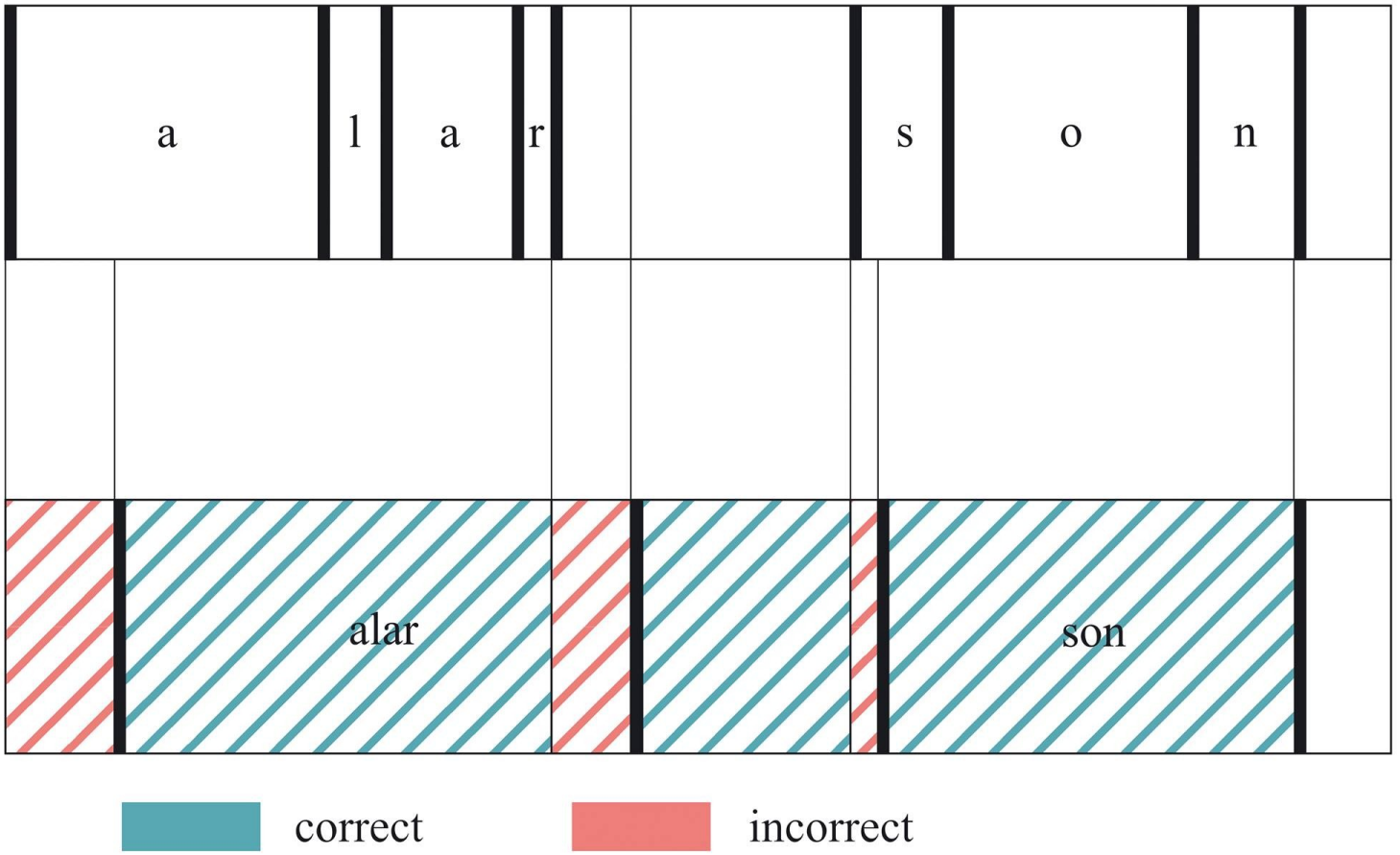


Dataset

- Especially compiled for this study, 2 singers
- 'canonical' dataset: at least two prolonged *key syllables*

	dataset	'canonical' dataset
duration (minutes)	67	27
#lines per aria	9.2	9.9
#syllables per line	10.7	10.3
line duration (seconds)	18.3	23.4
syllable duration (seconds)	2.4	3.1

Evaluation metric



Alignment accuracy = duration of correct regions / total audio duration
Suggested in (Fujihara, 2011)

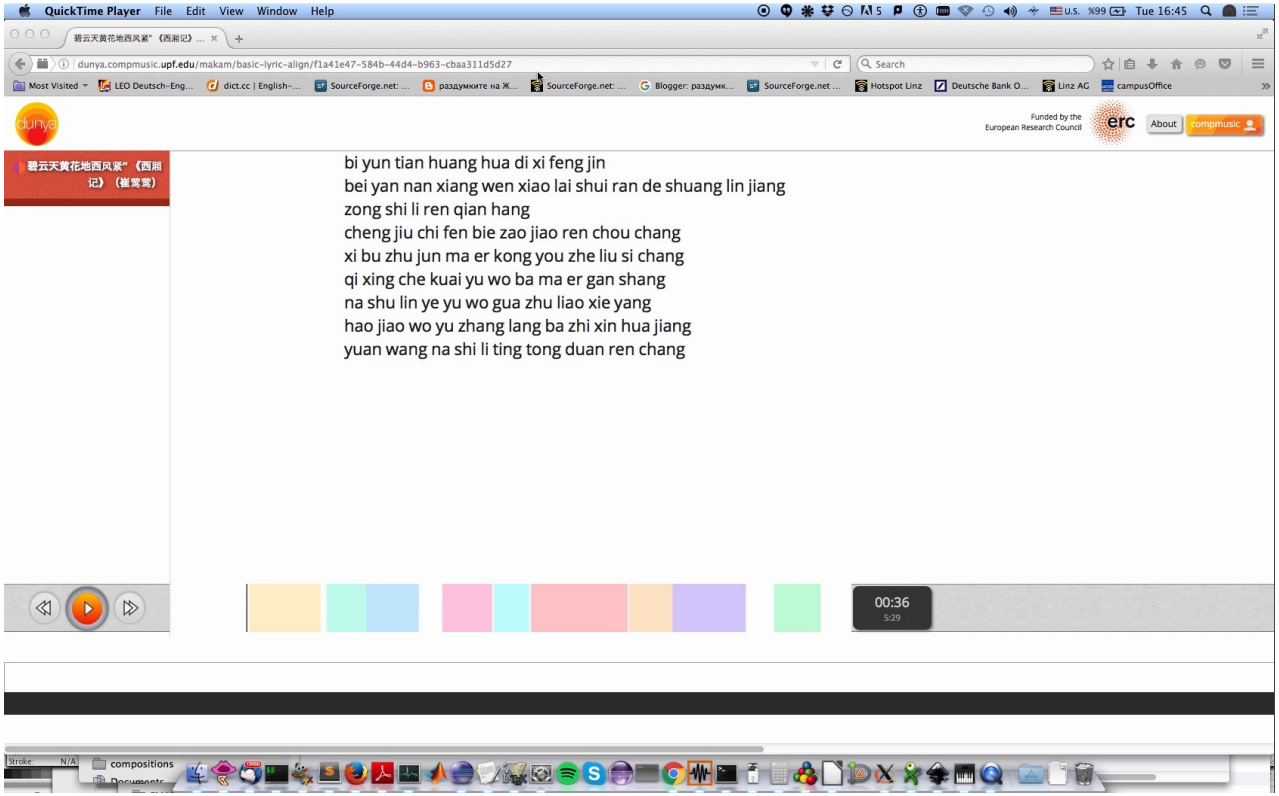
Results

- 3-fold cross validation done
- baseline: same HMM with no duration modeling
- oracle: same DHMM with annotations as if they were acoustic probabilities

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

Demo

Efficient python implementation available under CC license:
https://github.com/georgid/AlignmentDuration/tree/noteOnsets/jingju



Future work

- Extend to work with original polyphonic mix
- Integrate automatic segmentation of lyrics lines

References

(Fujihara, 2012): Lyrics-to-audio alignment and its application. *Multimodal Music Processing*

(Dzhambazov et al. 2015): Modeling of phoneme durations for alignment between polyphonic audio and Lyrics. In *Sound and Music Computing Conference, Maynooth, Ireland*.

(Kruspe et al. 2015): Keyword spotting in a-capella singing. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*

(Fujihara, 2011): Lyric-synchronizer: Automatic synchronization system between musical audio signals and lyrics. *IEEE Journal of Selected Topics in Signal Processing*