

Knowledge-based Probabilistic Modeling for Tracking Lyrics in Music Audio Signals

28 June 2017, PhD defense

Georgi Dzhambazov
Dept. of Information and Communication
Technologies
Universitat Pompeu Fabra

Thesis Supervisor
Dr. Xavier Serra
Music Technology Group
Universitat Pompeu Fabra



Thesis Committee:
Dr. Axel Röbel (IRCAM)
Dra. Emilia Gómez (UPF)
Dr. Matthias Mauch (QMUL)



Acknowledgements



Knowledge-based **Probabilistic Modeling** for Tracking Lyrics in Music Audio Signals

- Modeling - different computational models
- Probabilistic - can handle the temporal variability of music

Knowledge-based Probabilistic Modeling for **Tracking Lyrics** in Music Audio Signals

- Tracking the progression of sung lyrics in time
- For automatic alignment between written lyrics and audio

Knowledge-based Probabilistic Modeling for Tracking Lyrics in Music Audio Signals

- Take advantage of knowledge of simultaneous music events

Knowledge-based Probabilistic Modeling for Tracking Lyrics in **Music Audio Signals**

- Recordings of music

Additionally...

- Emphasis on reproducible and open research
- Code – Open source, Data: publicly available



Computational models

for the discovery of the World's Music


UNIVERSITAT
POMPEU FABRA


[HOME](#)

[DESCRIPTION](#)

[TEAM](#)

[PUBLICATIONS](#)

[CORPORA](#)

[SOFTWARE](#)

[EVENTS](#)

[BLOG](#)

[NEWS](#)

[RESOURCES](#)

[GET INVOLVED](#)

🔍

Computational models

for the discovery of the World's Music

English  

KNOWLEDGE-BASED PROBABILISTIC MODELING FOR TRACKING LYRICS IN MUSIC AUDIO SIGNALS

[View](#)
[Edit](#)

This is a companion web page for the PhD thesis of *Georgi Dzhambazov*. All the main resources related with the PhD work (datasets, code, publications, demos etc) are listed on this page.

Short abstract: In this thesis, we devise computational models for tracking sung lyrics in multi-instrumental music recordings. We consider not only the low-level acoustic characteristics, representing the timbre of the sung phonemes, but also higher-level music knowledge, that is complementary to lyrics. We build probabilistic models, based on dynamic Bayesian networks (DBN) that represent the relation of phoneme transitions to two music knowledge facets: the temporal structure of a lyrics line and the structure of the metrical cycle. In one model we exploit the fact the expected syllable durations depend on their position within a lyrics line. Then in another model, we propose how to estimate vocal onsets by tracking simultaneously the position in the metrical cycle, and how these estimated onsets influence the transitions between consecutive phonemes. Using the proposed models sung lyrics are automatically aligned to written lyrics on datasets from Ottoman Turkish makam and Beijing opera, whereby principles, specific for these music traditions are considered. Both models improve a baseline, unaware of music-specific knowledge. This confirms that music-specific knowledge is an important stepping stone for computationally tracking lyrics, especially in the challenging case of singing with instrumental accompaniment.

[Full abstract \(PDF\)](#)

GEORGID

- [People](#)
- [My account](#)
- [Media](#)
- [Create content](#)
- [Administer](#)
- [Log out](#)

LATEST BLOGS

[Technology and Multiculturality](#)
17/04/2016
[Article published in the daily newspaper La Vanguardia on Sunday 17th 2016. English translation of the original text written in catalan.] The violin, typewriter or mobile are examples of technological devices that were born in certain contexts...]

[Two evenings of Chinese traditional music](#) 27/01/2016
Last December (2015), Barcelona's Conservatori Municipal de Música hosted two sessions of Chinese

<http://compmusic.upf.edu/phd-thesis-georgi>

Outline

Outline

- Setting the stage
 - Introduction (Chapter 1)
 - Background (Chapter 2)
 - Baseline lyrics-to-audio alignment (Chapter 3)

Outline

- Setting the stage
 - Introduction (Chapter 1)
 - Background (Chapter 2)
 - Baseline lyrics-to-audio alignment (Chapter 3)
- Main
 - Lyrics-duration aware lyrics-to-audio alignment (Chapter 4)
 - Metrical-accent aware lyrics-to-audio alignment (Chapter 5)

Outline

- Setting the stage
 - Introduction (Chapter 1)
 - Background (Chapter 2)
 - Baseline lyrics-to-audio alignment (Chapter 3)
- Main
 - Lyrics-duration aware lyrics-to-audio alignment (Chapter 4)
 - Metrical-accent aware lyrics-to-audio alignment (Chapter 5)
- Conclusion (Chapter 6) and Applications (Appendix)

Chapter 1

Introduction

Chapter 1

Introduction

The way music is created, shared, distributed and listened to has been recently changing rapidly due to advancements in Information Technology. Music Information Retrieval (MIR) is a research subfield of music technology that aims to advance in automatic music processing. Some of the subjects addressed in MIR research include building computational models for describing music structures and events, as well as their temporal progression.

Any musical instrument is characterized also by an unique timbre. Classes representing the perceived 'timbral colour' of the singing voice can be described by abstract categories, such as *mellow, harsh, dull*. This reflects a quality described as *instrumental* quality of timbre by musicologists ([Durga, 1978](#)). Still, the belonging of a singing excerpt to one particular colour class is rather subjective and varies from one listener to another. This means that there may not be a mutual agreement among listeners on where in time the exact transitions between these classes are.

Few instruments, including singing voice, have their timbre continuously vary in time, prensisng frequent timbral alterations. Unlike other instruments though, the singing voice has a unique characteristic: its ability to articulate actual lyrics. Lyrics are one of the most important musical aspects. They carry a message or a story and attract the attention of the listener. She/he will naturally follow the lyrics while listening to the melody of the main singing voice.

Phonemes - the building block of words - can be considered as a discrete number of timbral classes, wherein each class has a characteristic spectral template. The ability to articulate phonemes is an innate characteristic of human speakers. In fact, singers articulate by means of given vowels even

Research objectives

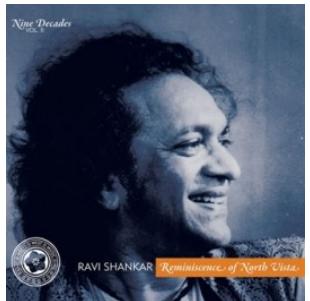
- Creation of computational models for tracking lyrics
- Considering music-specific knowledge
 - for a specific music tradition with its music principles
- Application to music with predominant singing voice
- Evaluation of the models on an end task - lyrics-to-audio alignment
 - contribution of music knowledge compared to baseline
 - perfect alignment not an end goal

Scientific context

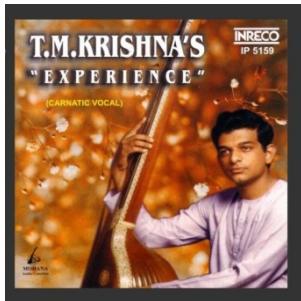
- Singing voice has unique ability - to articulate actual words
- Various works on singing voice (Goto, 2014)
 - little work on sung lyrics
 - almost no work on sung lyrics with music-specific knowledge

Project context – CompMusic

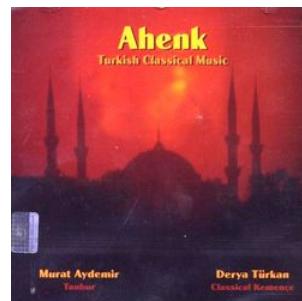
- Music-tradition aware Music Information Retrieval (MIR)



Hindustani



Carnatic



Ottoman
Turkish makam

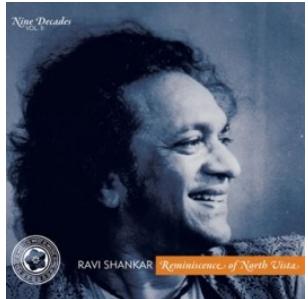


Arab-Andalusian
Beijing Opera

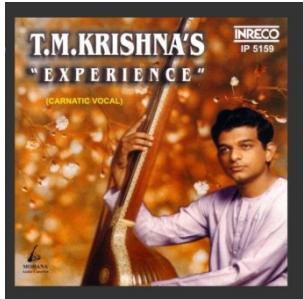


Project context – CompMusic

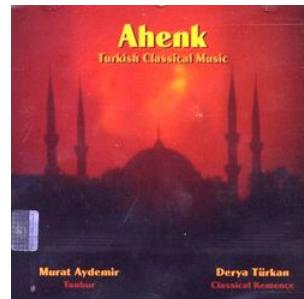
- Music-tradition aware Music Information Retrieval (MIR)
- Vocal-centered traditions



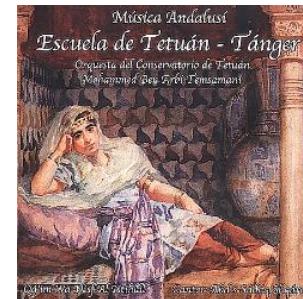
Hindustani



Carnatic



Ottoman
Turkish makam



Arab-Andalusian

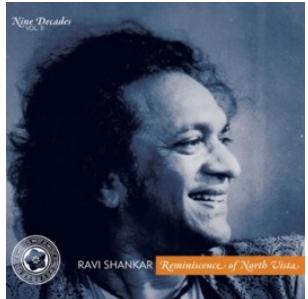


Beijing Opera

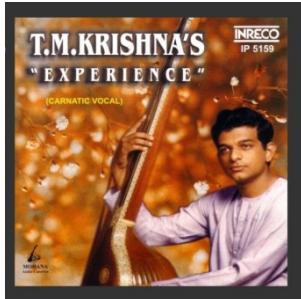
<http://compmusic.upf.edu/>

Project context – CompMusic

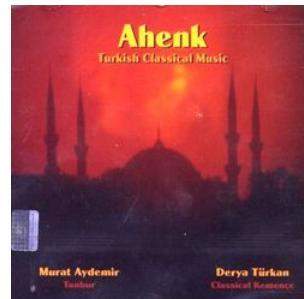
- Music-tradition aware Music Information Retrieval (MIR)
- Vocal-centered traditions
- Well-grounded music principles



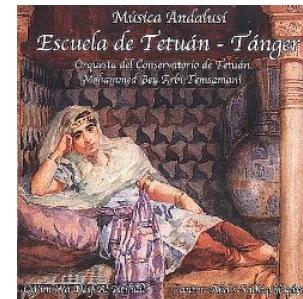
Hindustani



Carnatic



Ottoman
Turkish makam



Arab-Andalusian

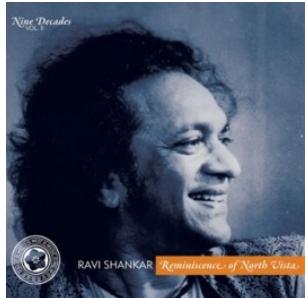


Beijing Opera

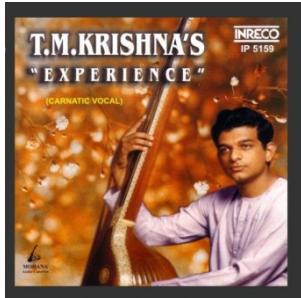
<http://compmusic.upf.edu/>

Project context – CompMusic

- Music-tradition aware Music Information Retrieval (MIR)
- Vocal-centered traditions
- Well-grounded music principles



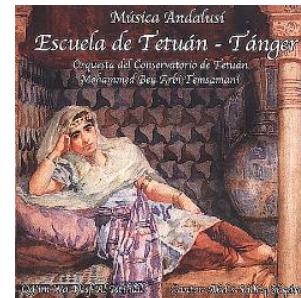
Hindustani



Carnatic



Ottoman
Turkish makam



Arab-Andalusian



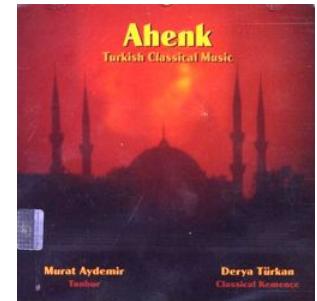
Beijing Opera

<http://compmusic.upf.edu/>

Project context – CompMusic

Ottoman Turkish Makam Music (OTMM)

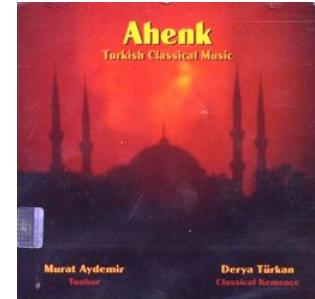
- has machine-readable music scores
- with significant musicological literature
- OTMM seemed fairly familiar to me



Project context – CompMusic

Ottoman Turkish Makam Music (OTMM)

- has machine-readable music scores
- with significant musicological literature
- OTMM seemed fairly familiar to me



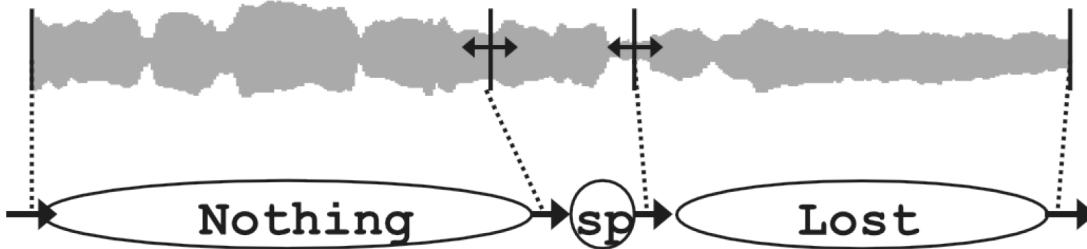
Beijing opera (jingju)

- challenging language
- difference between interpretations and music score



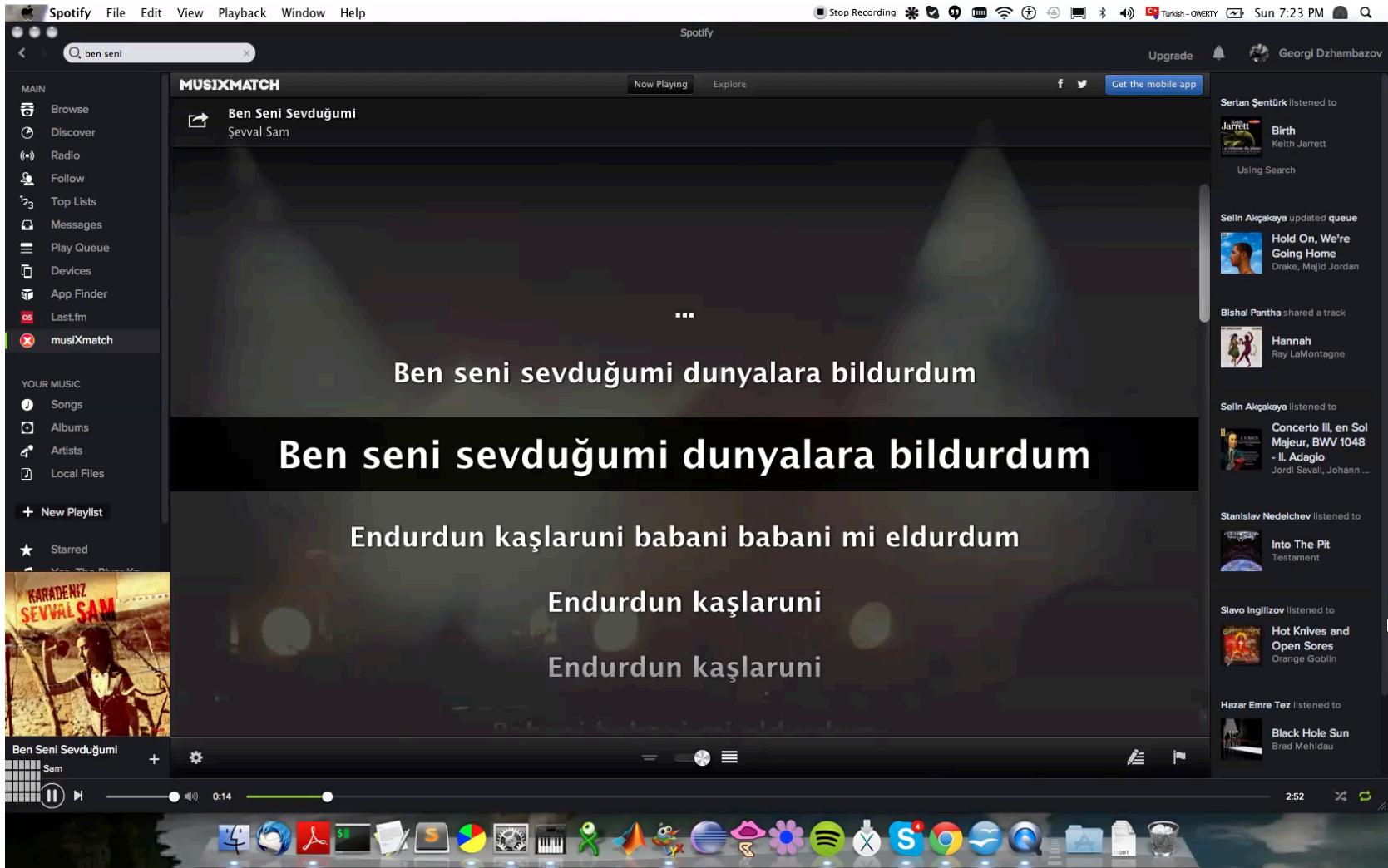
Lyrics-to-audio alignment

- Goal: Automatic synchronisation between an audio recording and its written lyrics: phrases/words/syllables



- State of the art approach: phonetic recognizer adopted from automatic speech recognition
- Applications
 - automatic highlighting of lyrics
 - automatic thumbnail generation
 - lyrics-based navigation

Lyrics-based navigation

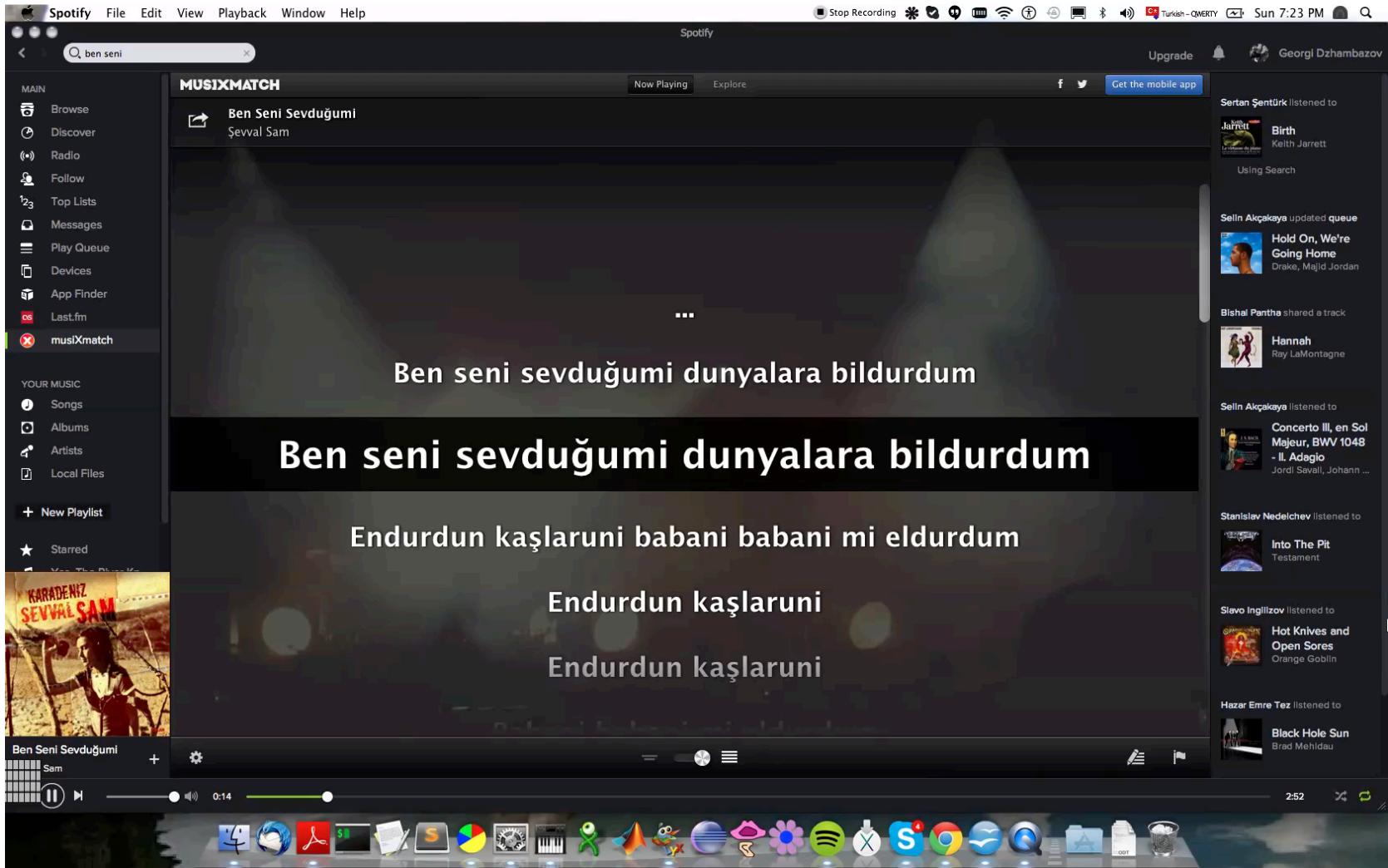


The screenshot shows a Spotify interface with a lyrics-based navigation overlay. The main song playing is "Ben Seni Sevdiğim" by Şevval Sam. The lyrics are displayed in large white text on a semi-transparent black background:

- Ben seni sevdiğim dunyalara bildurdum
- Ben seni sevdiğim dunyalara bildurdum
- Endurdun kaşlarını babanı babanı mı eldurdum
- Endurdun kaşlarını
- Endurdun kaşlarını

The Spotify interface includes a sidebar with navigation links like Browse, Discover, Radio, Follow, Top Lists, Messages, Play Queue, Devices, App Finder, Last.fm, and musixmatch. The music player at the bottom shows the current track, volume, and playback controls.

Lyrics-based navigation



The screenshot shows a Spotify interface with a lyrics-based navigation overlay. The main song playing is "Ben Seni Sevdiğim" by Şevval Sam. The lyrics are displayed in large white text on a semi-transparent black background:

- Ben seni sevdiğim dunyalara bildurdum
- Ben seni sevdiğim dunyalara bildurdum
- Endurdun kaşlarını babanı babanı mı eldurdum
- Endurdun kaşlarını
- Endurdun kaşlarını

The Spotify interface includes a sidebar with navigation links like Browse, Discover, Radio, Follow, Top Lists, Messages, Play Queue, Devices, App Finder, Last.fm, and musixmatch. The music player at the bottom shows the current track, volume, and playback controls.

Why consider music knowledge?

- Lyrics structure has granularity
 - verse-chorus-like structure
 - lyrics lines structure
 - metrical structure

[Verse 2]

Am G
Black bandana, sweet Louisiana,
Dm Am
Robbin' all the banks in the state of Indiana,
Am G
She's a runner, rebel and a stunner,
Dm Am
Condemned everywhere sayin, 'baby whatcha gonna'
Am G Dm Am
Lookin down the barrel of a hot metal .45,
Am G | Dm
Just another way to survive

[Chorus]

F C Dm
California, Rest In Peace,
G F C Dm
Simultaneous release,
G F C Dm
California, show your teeth,
G F C | Dm
She's my priestess, I'm your priest, yeah, yeah

Why consider music knowledge?

- Lyrics structure has granularity
 - verse-chorus-like structure
 - lyrics lines structure
 - metrical structure

[Verse 2]

Am G
 Black bandana, sweet Louisiana,
 Dm Am
 Robbin' all the banks in the state of Indiana,
 Am G
 She's a runner, rebel and a stunner,
 Dm Am
 Condemned everywhere sayin', 'baby whatcha gonna'
 Am G Dm Am
 Lookin down the barrel of a hot metal .45,
 Am G | Dm
 Just another way to survive

[Chorus]

F C Dm
 California, Rest In Peace,
 G F C Dm
 Simultaneous release,
 G F C Dm
 California, show your teeth,
 G F C | Dm
 She's my priestess, I'm your priest, yeah, yeah

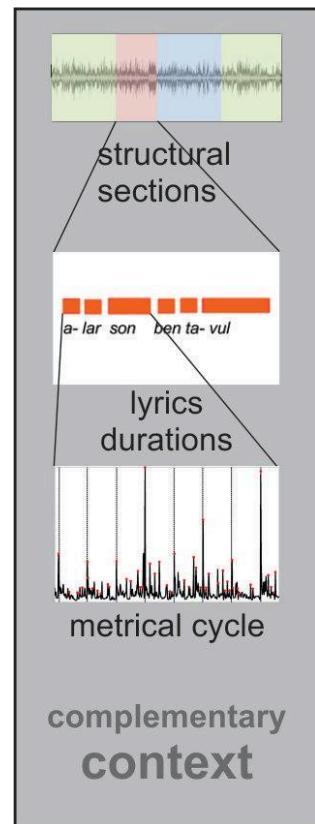
- Lyrics units are related to music facets
 - chords
 - accents in the melody (Nichols, 2009)
 - musical rhythm and meter

Complementary context

- What is complementary temporal music context?
 - any music facet, occurring simultaneously to sung lyrics
- Different granularity levels

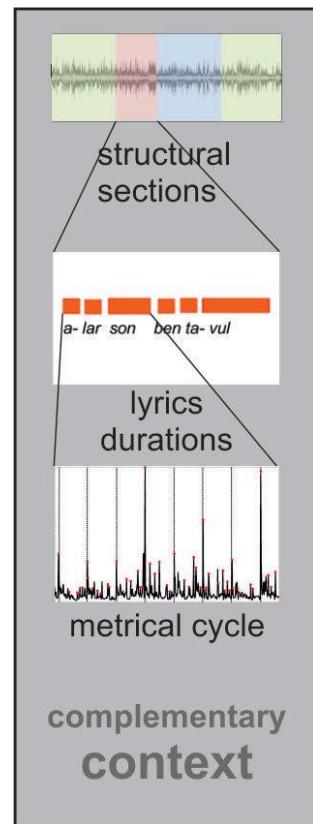
Complementary context

- What is complementary temporal music context?
 - any music facet, occurring simultaneously to sung lyrics
- Different granularity levels



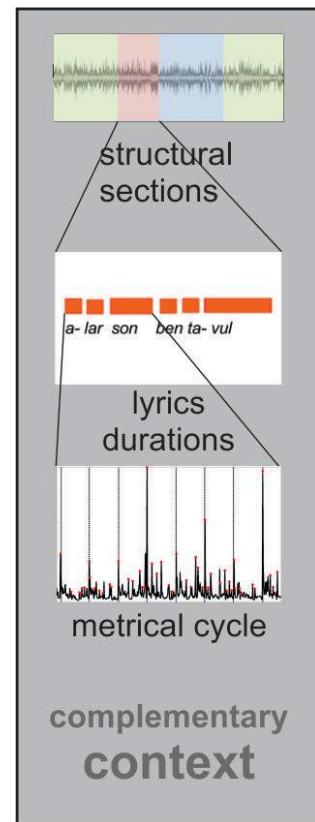
Complementary context

- What is complementary temporal music context?
 - any music facet, occurring simultaneously to sung lyrics
- Different granularity levels
 - Coarse level: structure of composition



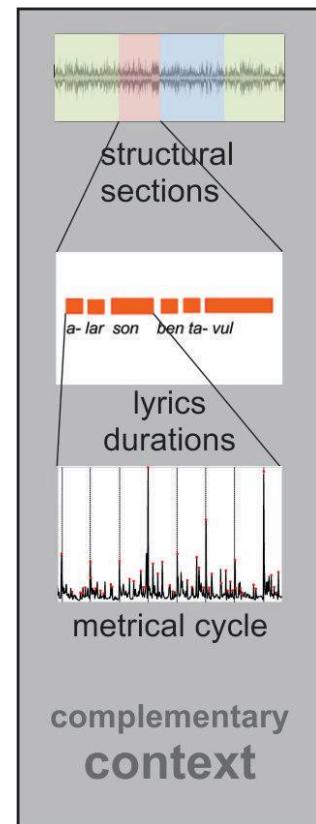
Complementary context

- What is complementary temporal music context?
 - any music facet, occurring simultaneously to sung lyrics
- Different granularity levels
 - Coarse level: structure of composition



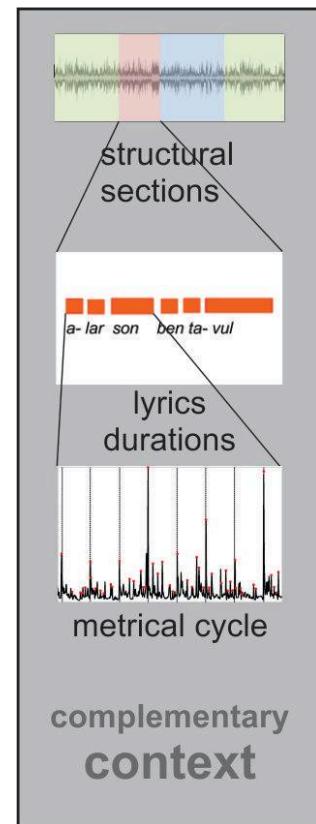
Complementary context

- What is complementary temporal music context?
 - any music facet, occurring simultaneously to sung lyrics
- Different granularity levels
 - Coarse level: structure of composition
 - Middle level: temporal structure of lyrics lines



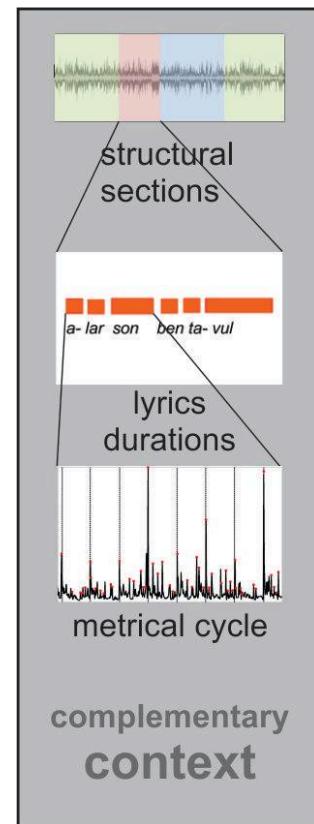
Complementary context

- What is complementary temporal music context?
 - any music facet, occurring simultaneously to sung lyrics
- Different granularity levels
 - Coarse level: structure of composition
 - Middle level: temporal structure of lyrics lines



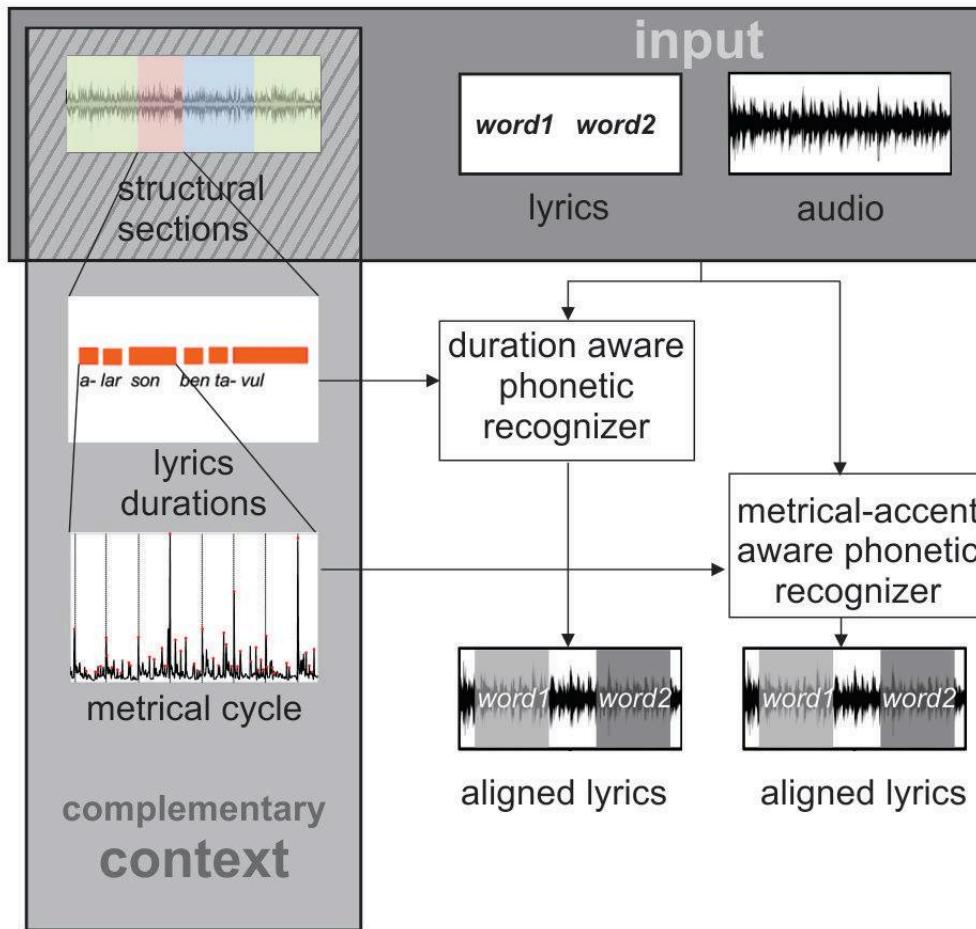
Complementary context

- What is complementary temporal music context?
 - any music facet, occurring simultaneously to sung lyrics
- Different granularity levels
 - Coarse level: structure of composition
 - Middle level: temporal structure of lyrics lines
 - Fine level: structure of metrical cycle



Models overview

- Integrate knowledge from complementary music context into the phonetic recognizer



Outline

- Setting the stage
 - Introduction (Chapter 1)
 - Background (Chapter 2)
 - Baseline lyrics-to-audio alignment (Chapter 3)
- Main
 - Lyrics-duration aware lyrics-to-audio alignment (Chapter 4)
 - Metrical-accent aware lyrics-to-audio alignment (Chapter 5)
- Conclusion (Chapter 6) and Applications (Appendix)

Outline

- Setting the stage
 - Introduction (Chapter 1)
 - Background (Chapter 2)
 - Baseline lyrics-to-audio alignment (Chapter 3)
- Main
 - Lyrics-duration aware lyrics-to-audio alignment (Chapter 4)
 - Metrical-accent aware lyrics-to-audio alignment (Chapter 5)
- Conclusion (Chapter 6) and Applications (Appendix)

Chapter 2

Background

Review of state of the art phonetic
recognizer alignment approaches

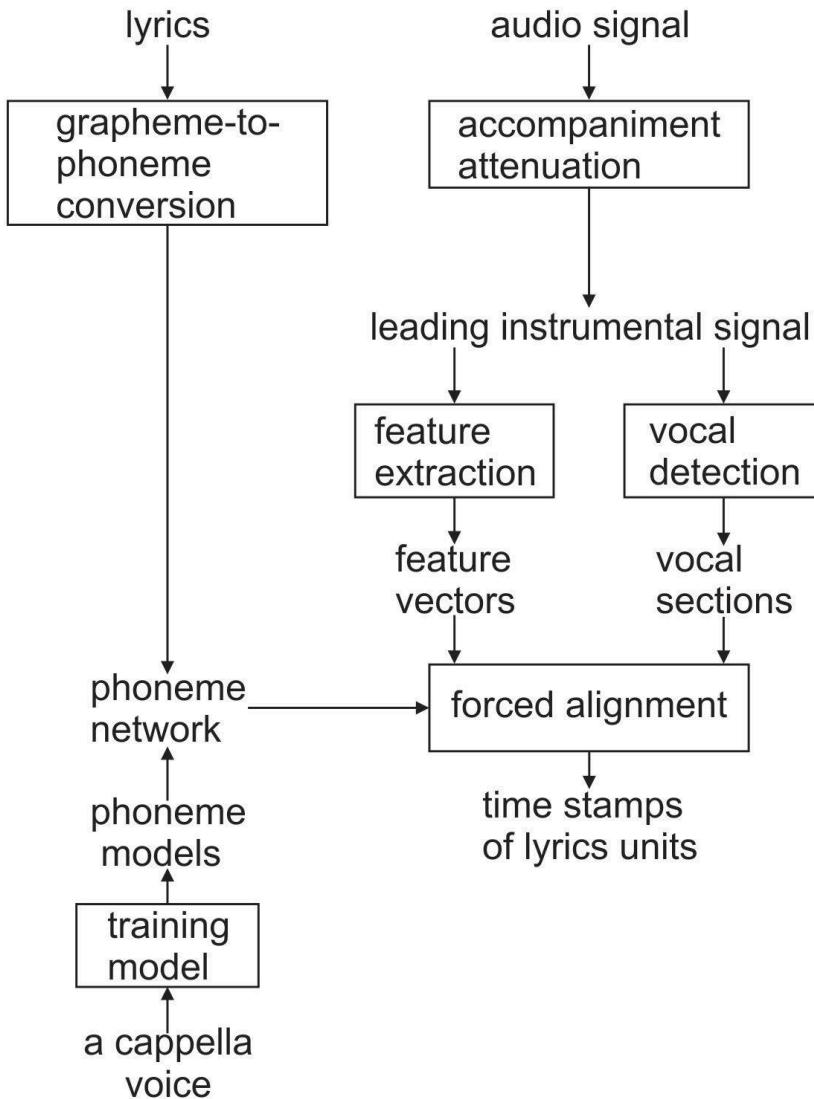
Chapter 2 Background

In Section 2.1.1 we first summarize some of the principles of OTMM, the main music tradition analyzed in this thesis, which influences directly or implicitly the way phonetic timbre progresses in time. We put a focus among all principles on the ones related to the structural form of the compositions: the music scores; and the rhythmic patterns of the music. Language, being one of the important aspects of lyrics, is reviewed in terms of the acoustic characteristics of the phonemes. Analogously, for jingju we review the language and some relevant principles of complementary context (Section 2.1.2). We emphasize the structure of a lyrics line, being the specific context facet we exploit later in Chapter 4.

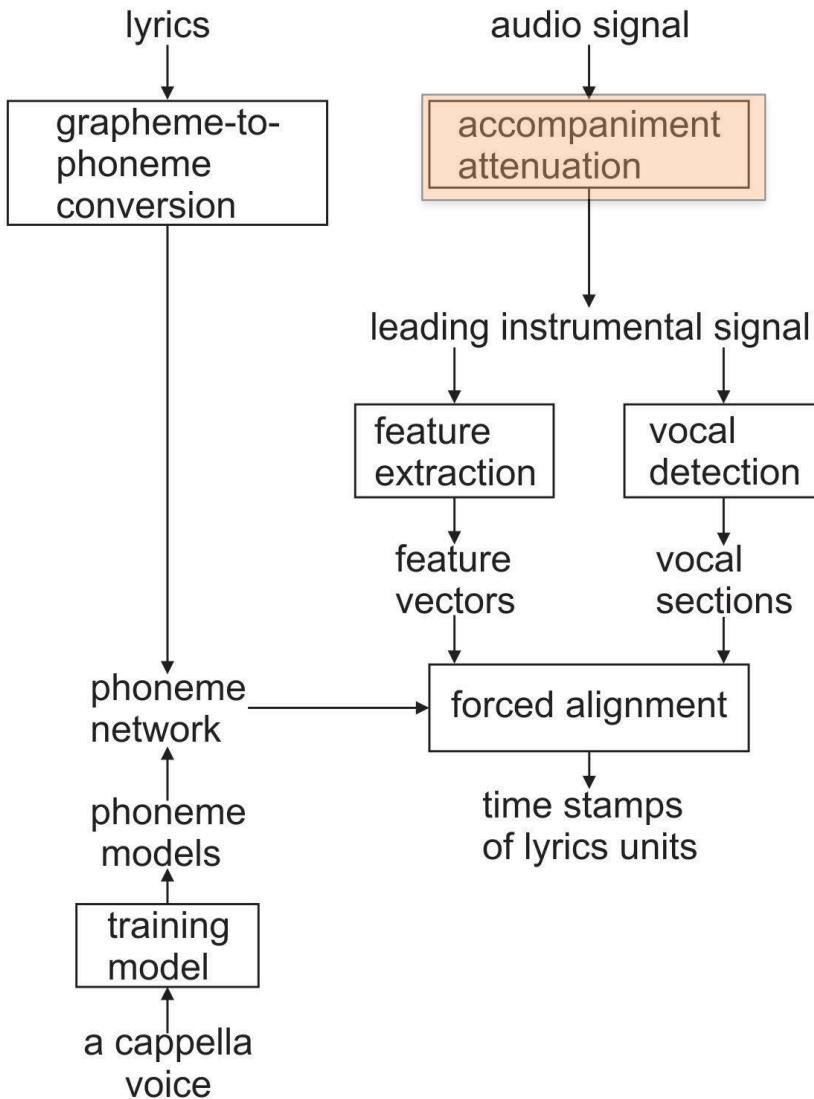
Then in Section 2.2 we summarize the existing approaches to the LAA alignment problem whereby the focus is put on those based on the phonetic recognizer paradigm. Common shortcomings as well as opportunities for extension are identified.

Finally, after introducing briefly the concept of dynamic Bayesian networks (Section 2.3), we review in Section 2.4 particular examples of related work on sung lyrics, in which consideration of concepts of complementary context, complementary to phonetic timbre, proved to be beneficial.

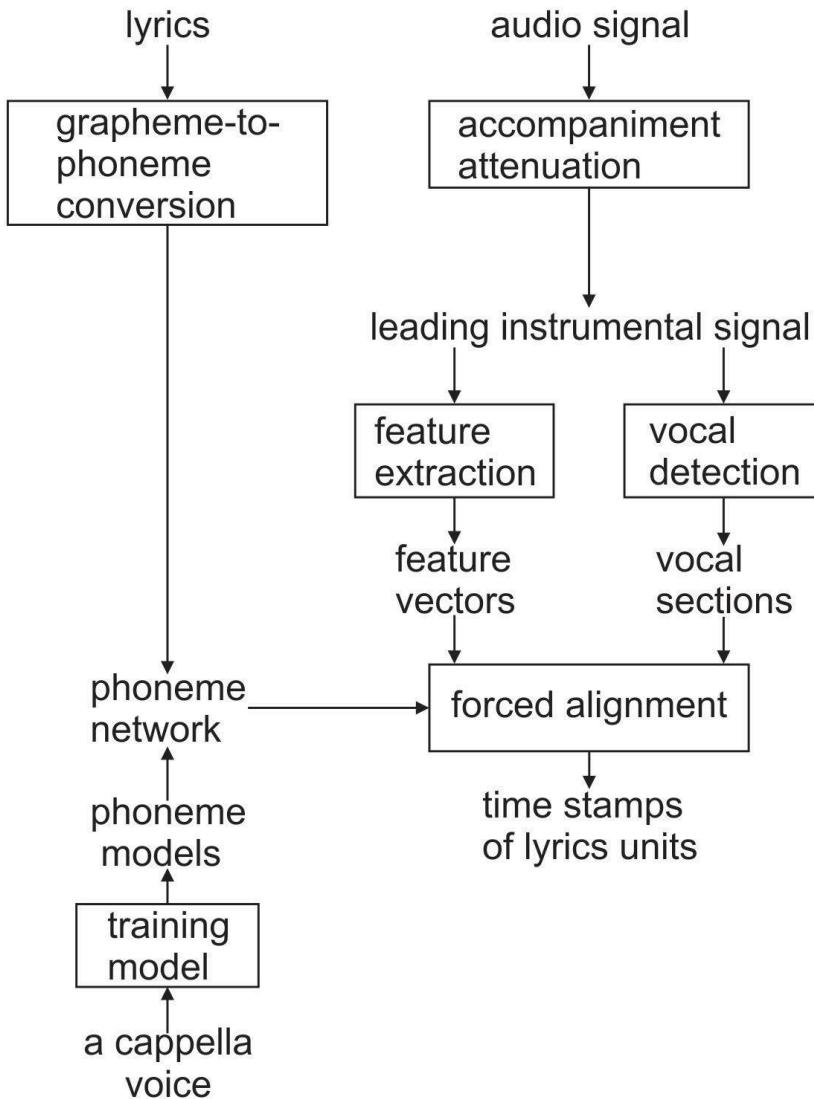
Modules of existing approaches



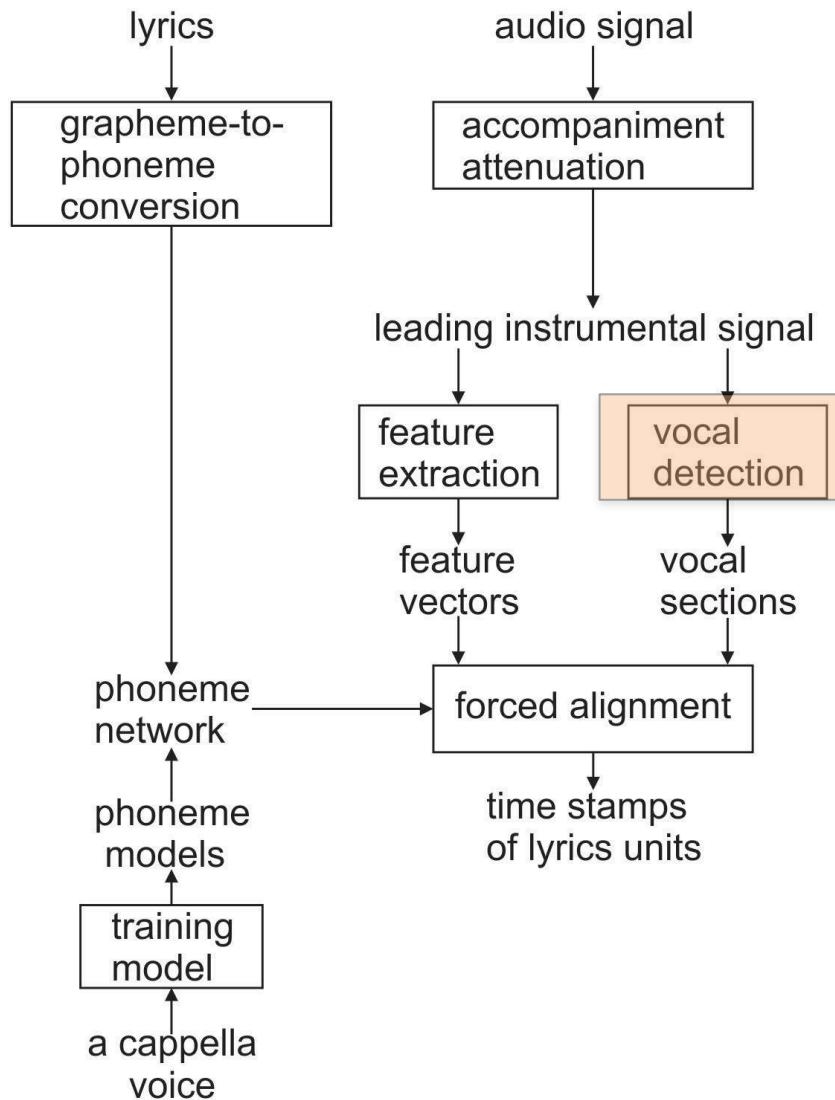
Modules of existing approaches



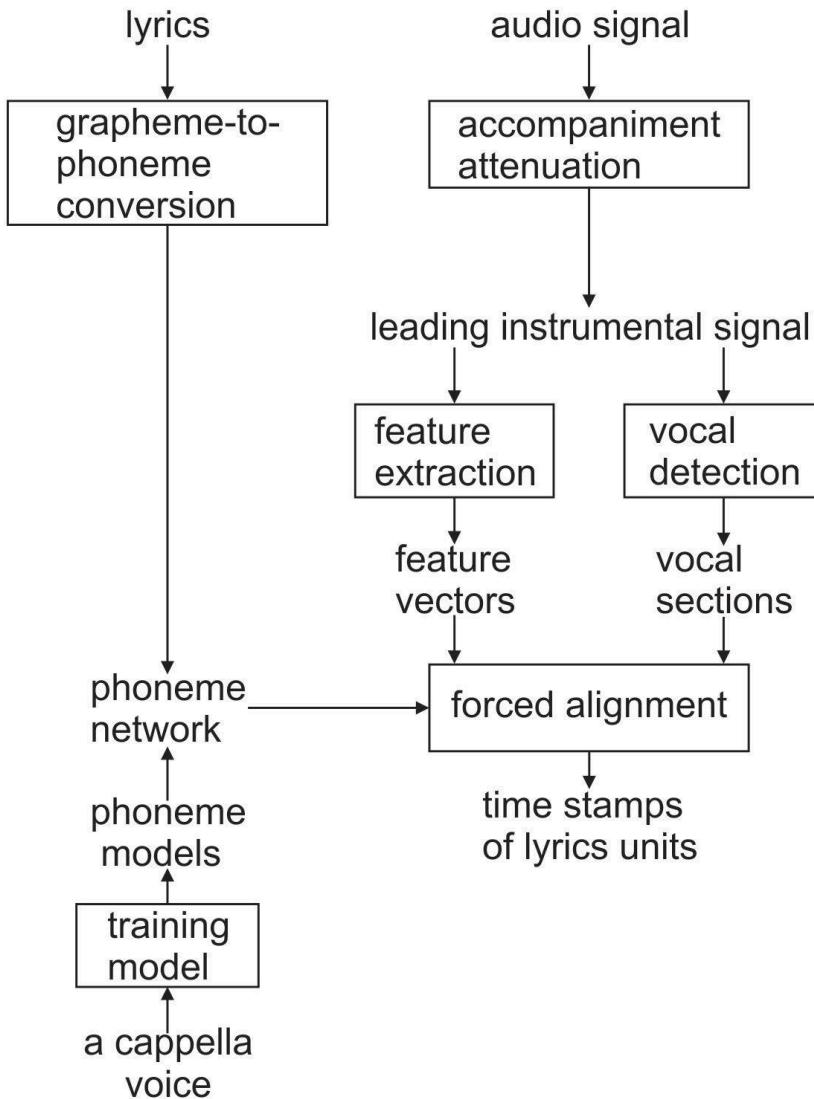
Modules of existing approaches



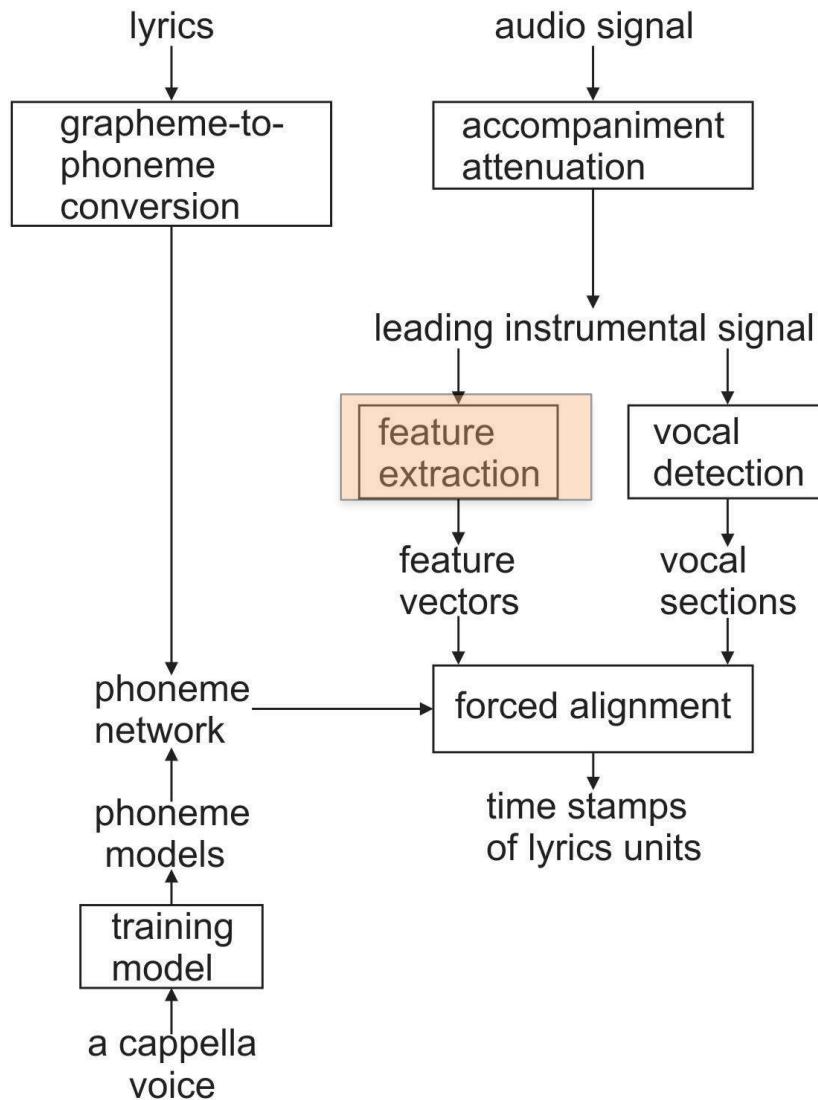
Modules of existing approaches



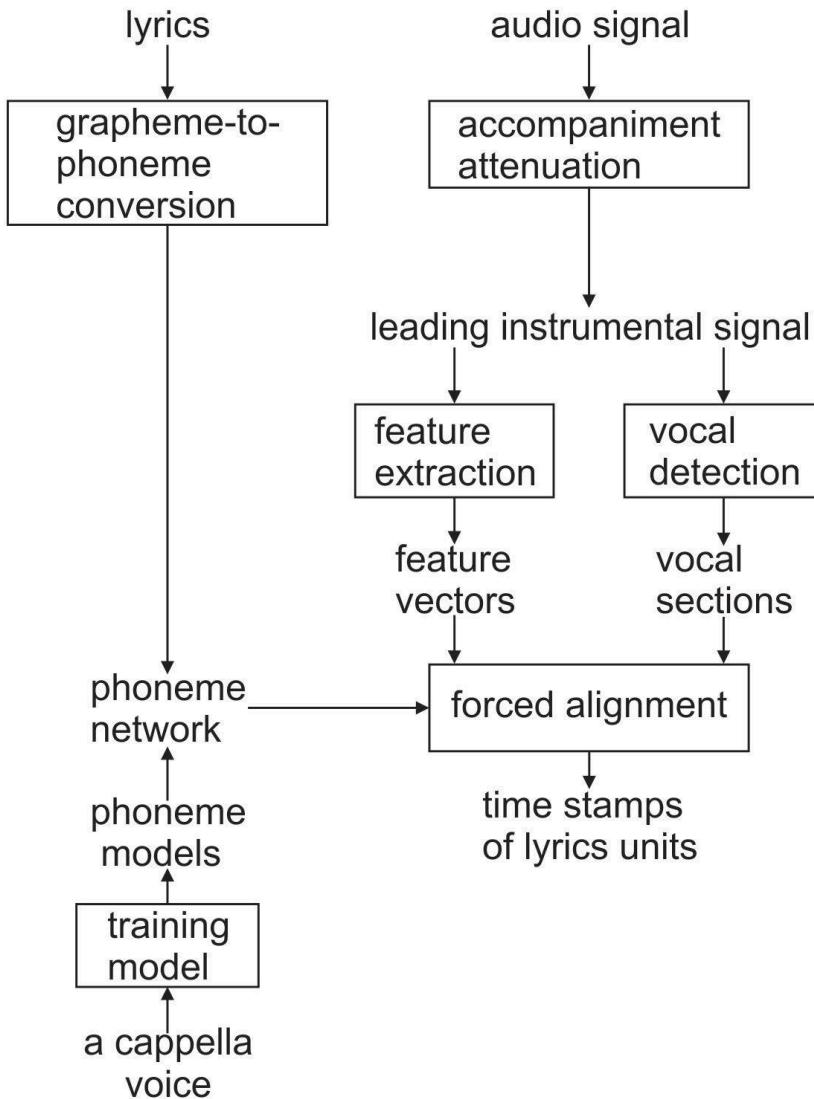
Modules of existing approaches



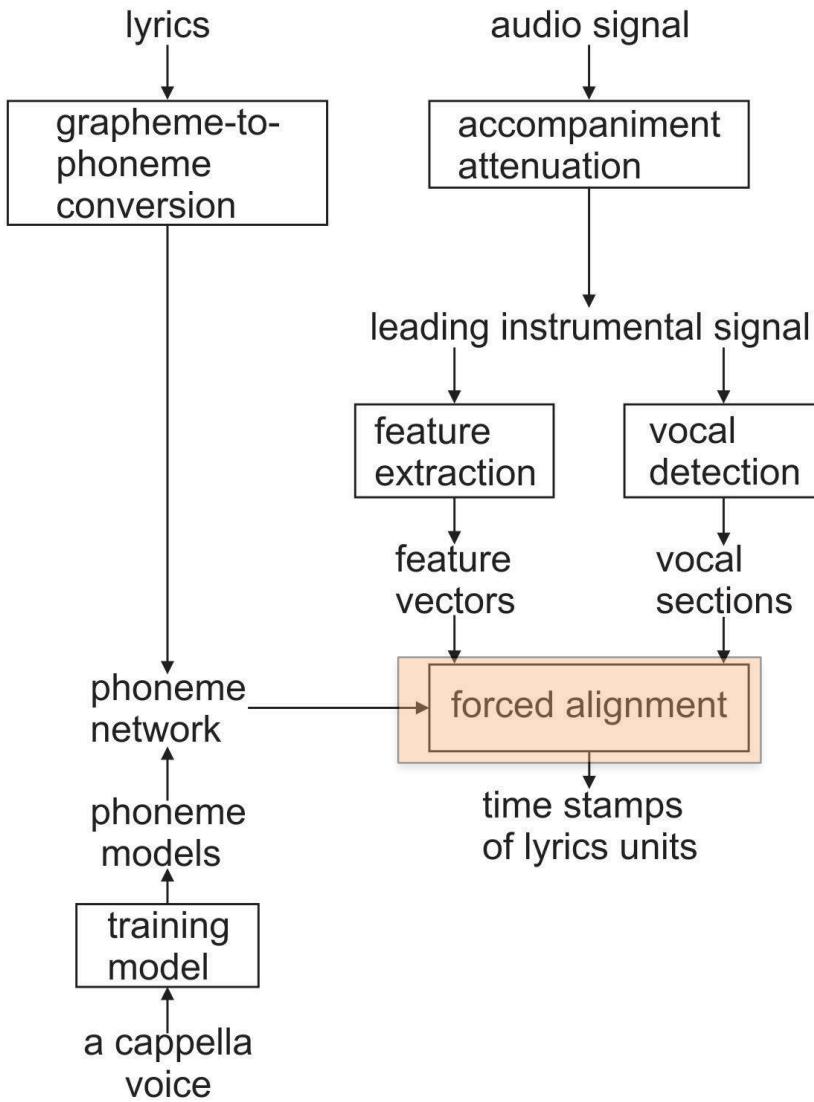
Modules of existing approaches



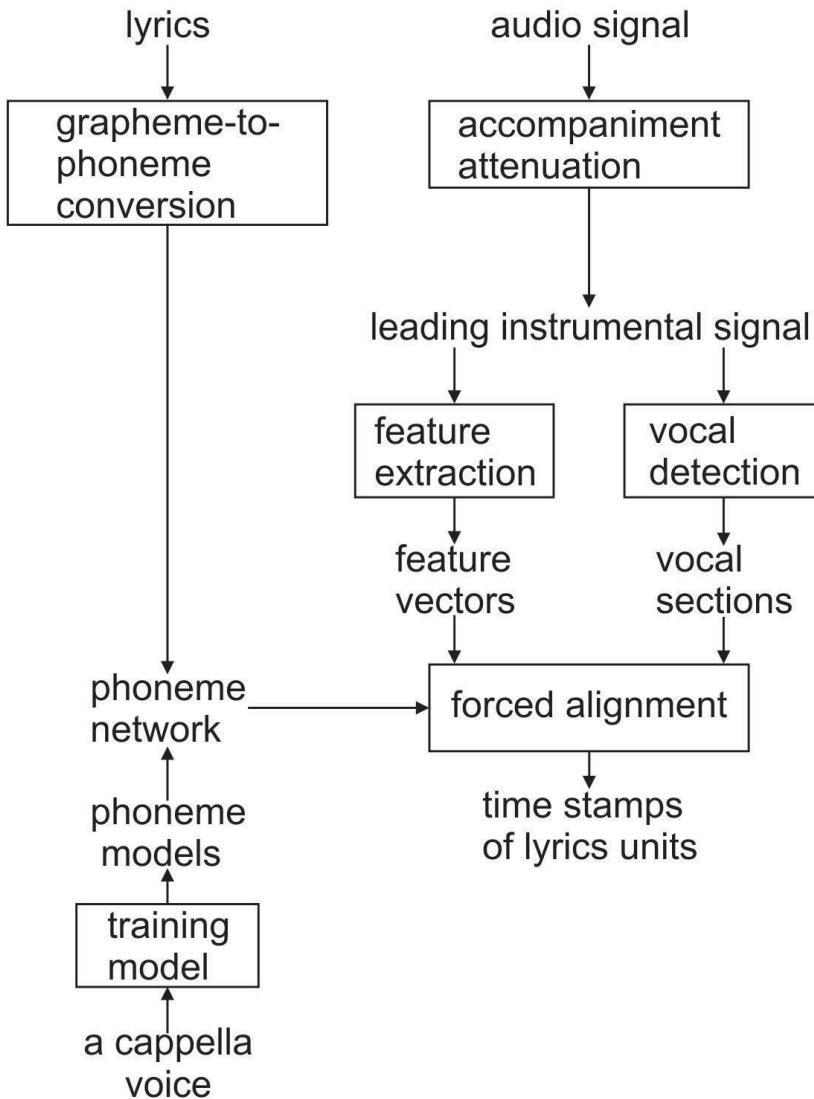
Modules of existing approaches



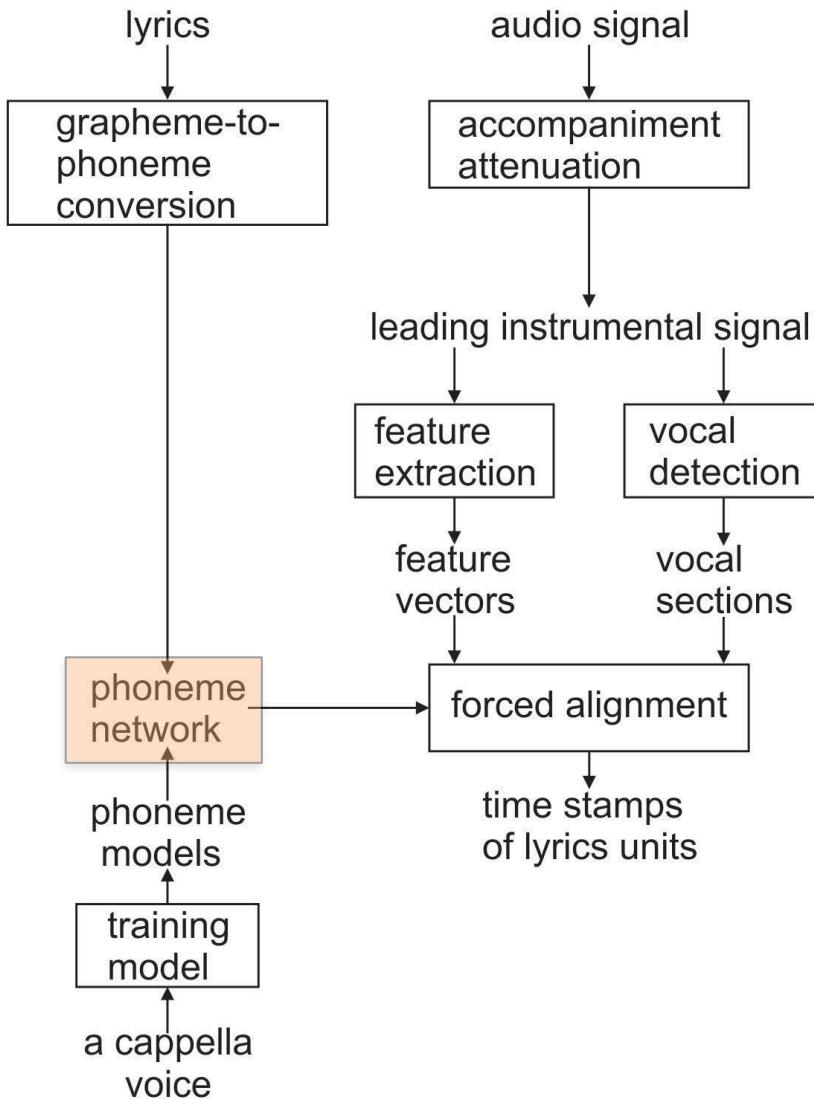
Modules of existing approaches



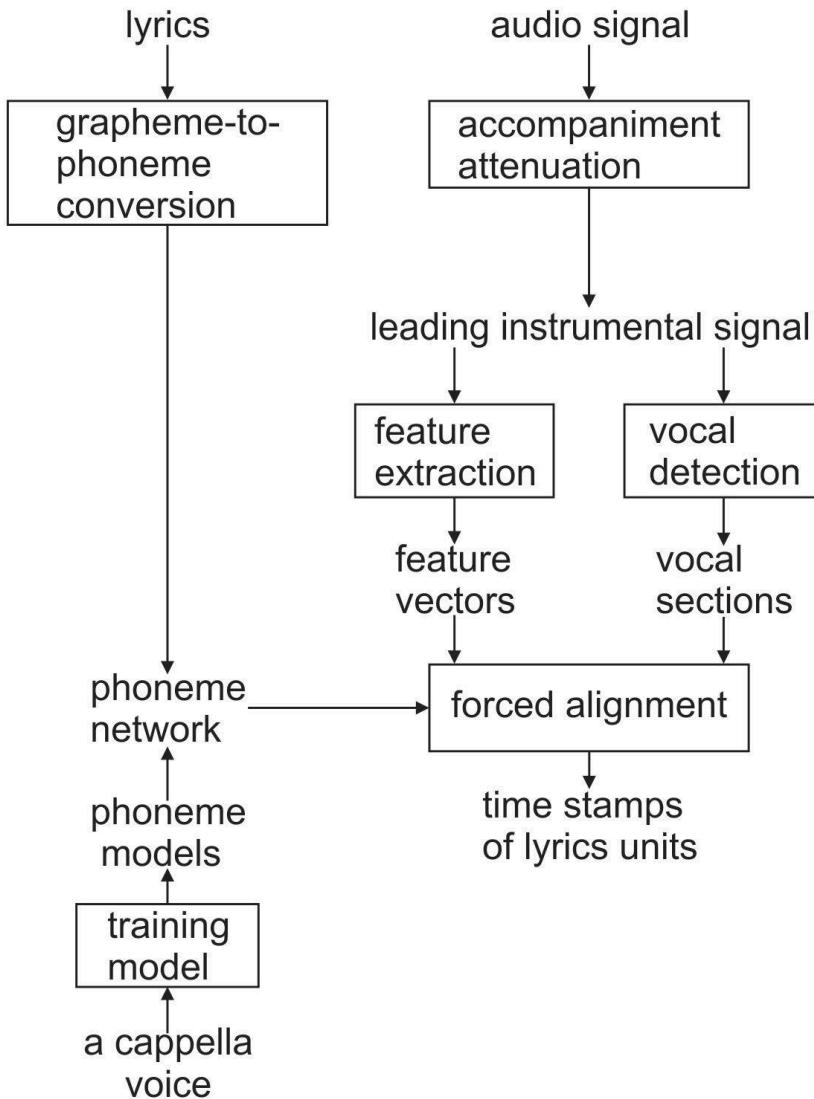
Modules of existing approaches



Modules of existing approaches



Modules of existing approaches



Comparison of existing approaches

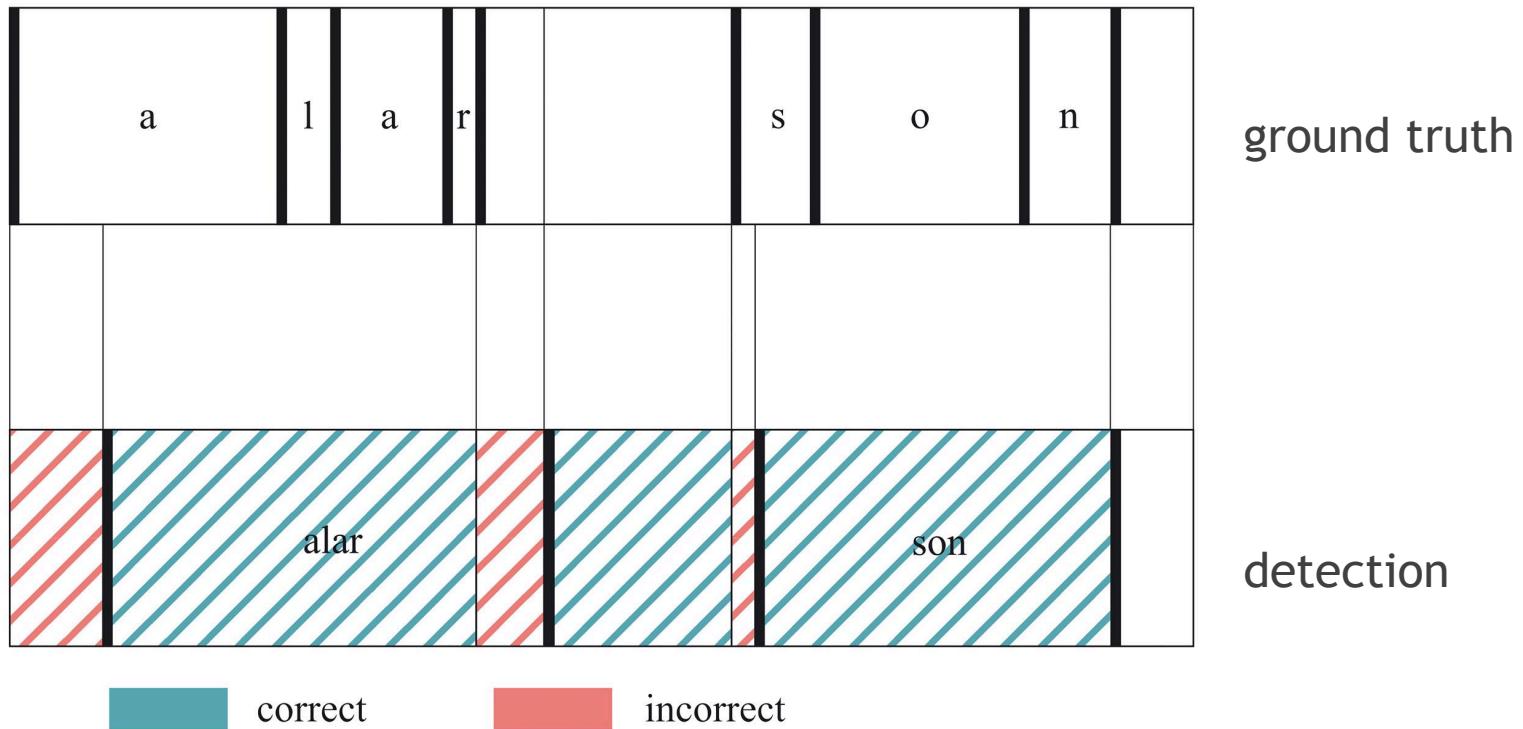
Author	Features	Training approach
(Mesaros, 2008)	MFCC	Speech + adaptation
(Fujihara, 2011)	MFCC	Speech + singer adaptation
(Kruspe, 2016)	MFCC+PLP	Singing

Comparison of existing approaches

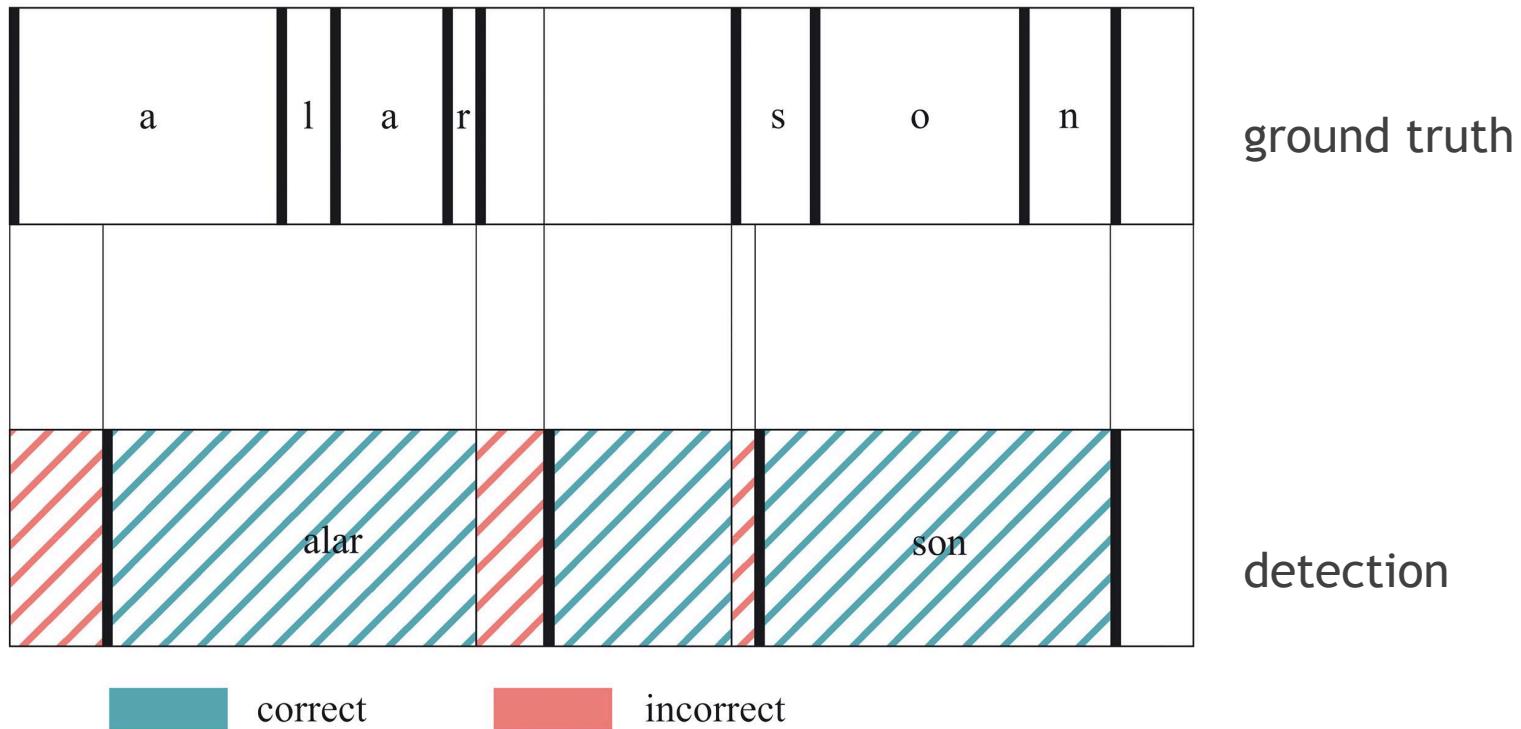
Author	Features	Training approach	>1 language	reproducible
(Mesaros, 2008)	MFCC	Speech + adaptation	✗	✗
(Fujihara, 2011)	MFCC	Speech + singer adaptation	✓	✗
(Kruspe, 2016)	MFCC+PLP	Singing	✗	✓

- Almost all are trained and tested on one language (English)
- None is fully reproducible
- None considers knowledge of simultaneously occurring musical events

Evaluation metrics

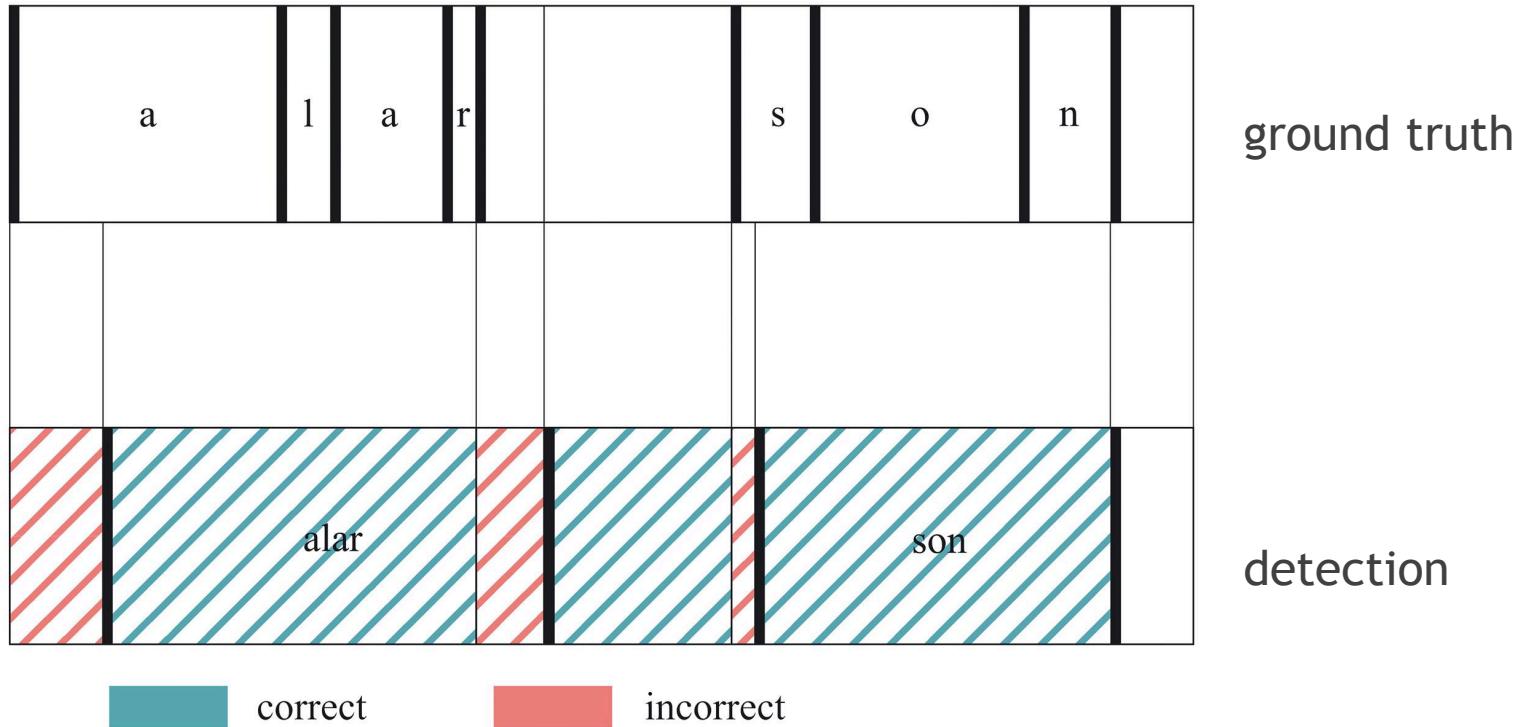


Evaluation metrics



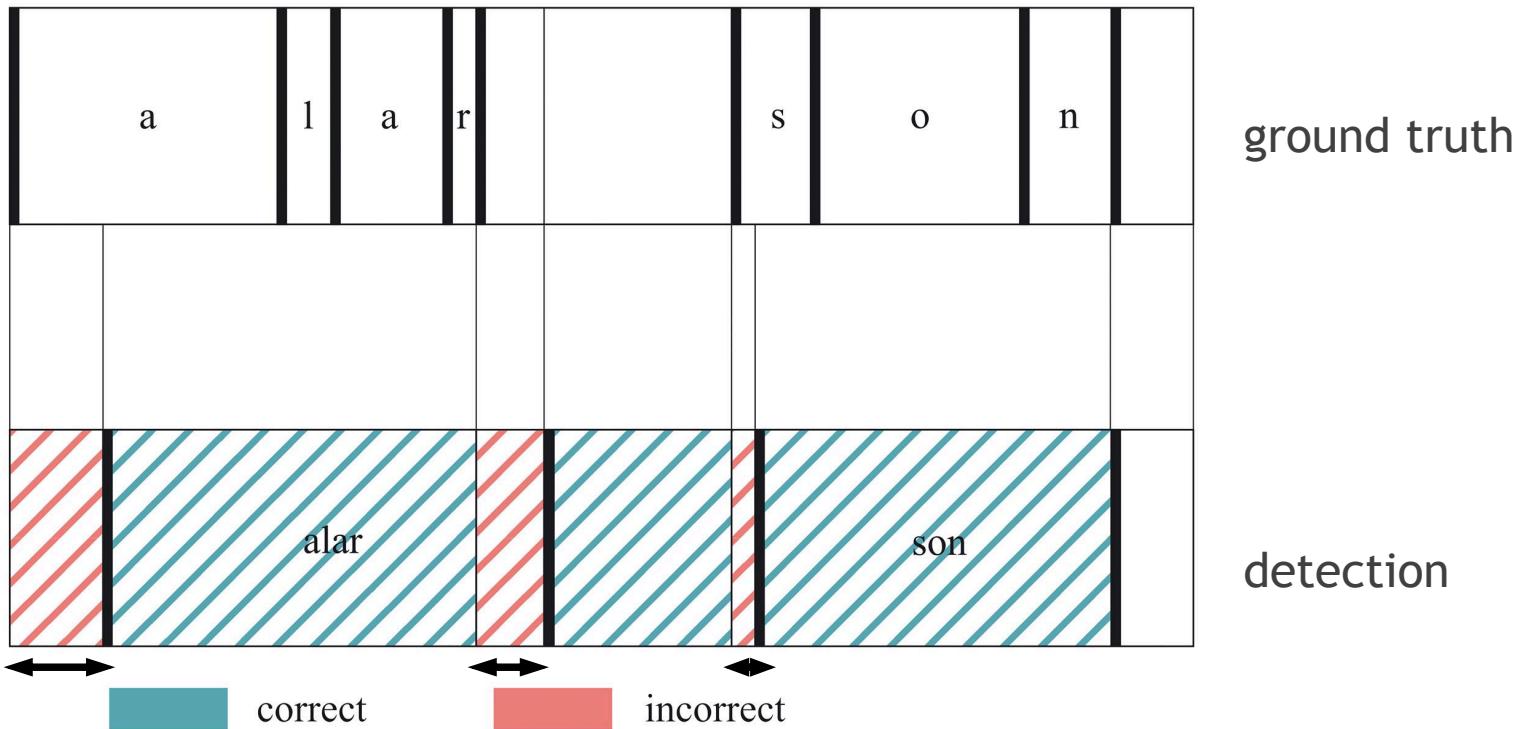
- Alignment accuracy (in %)

Evaluation metrics



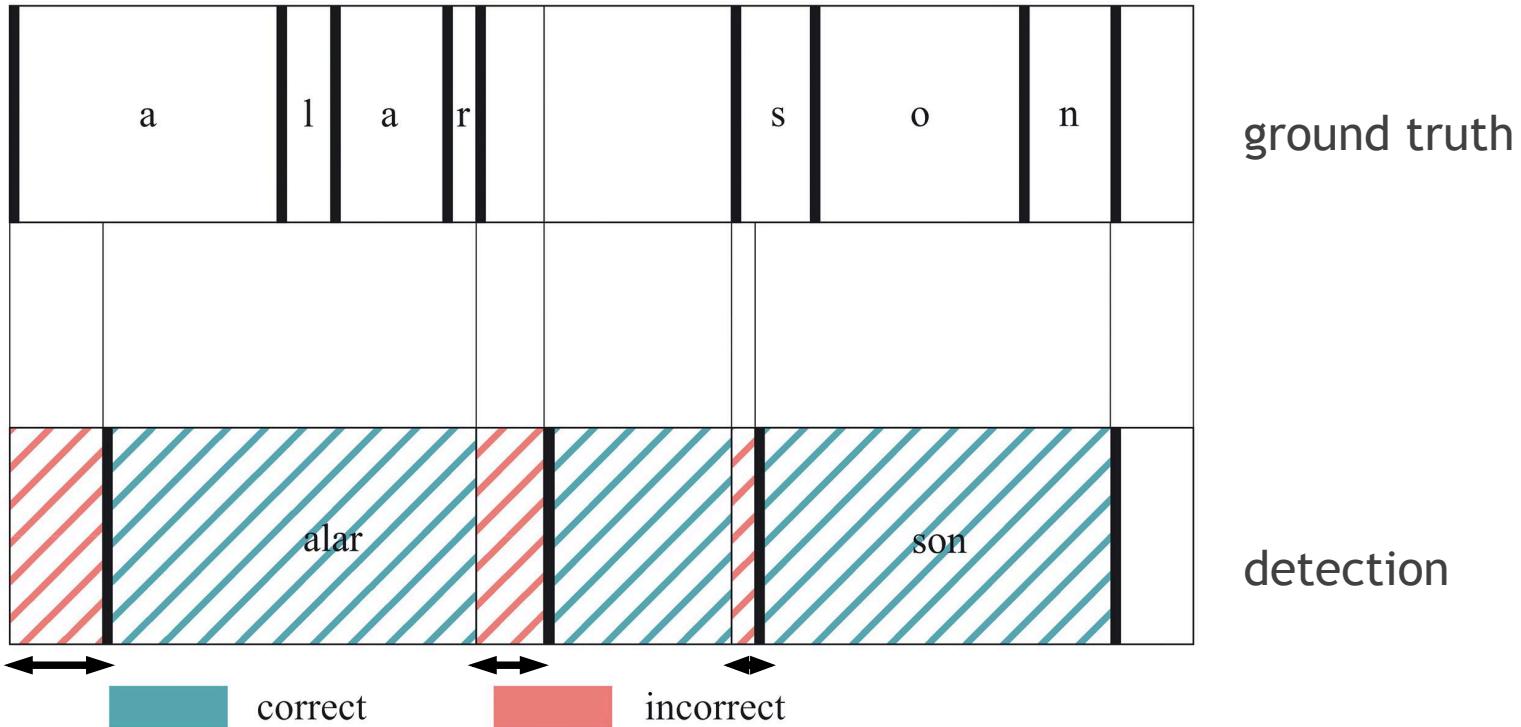
- Alignment accuracy (in %) = $\frac{\text{correct}}{\text{correct} + \text{incorrect}}$

Evaluation metrics



- Alignment accuracy (in %) = $\frac{\text{correct}}{\text{correct} + \text{incorrect}}$
- Average absolute error (in seconds)

Evaluation metrics



- Alignment accuracy (in %) = $\frac{\text{correct}}{\text{correct} + \text{incorrect}}$ (Fujihara, 2011)
- Average absolute error (in seconds)

Outline

- Setting the stage
 - Introduction (Chapter 1)
 - Background (Chapter 2)
 - Baseline lyrics-to-audio alignment (Chapter 3)
- Main
 - Lyrics-duration aware lyrics-to-audio alignment (Chapter 4)
 - Metrical-accent aware lyrics-to-audio alignment (Chapter 5)
- Conclusion (Chapter 6) and Applications (Appendix)

Outline

- Setting the stage
 - Introduction (Chapter 1)
 - Background (Chapter 2)
 - Baseline lyrics-to-audio alignment (Chapter 3)
- Main
 - Lyrics-duration aware lyrics-to-audio alignment (Chapter 4)
 - Metrical-accent aware lyrics-to-audio alignment (Chapter 5)
- Conclusion (Chapter 6) and Applications (Appendix)

Chapter 3

Baseline Lyrics-to-audio Alignment

Chapter 3

Baseline Lyrics-to-audio Alignment Model

3.1 Introduction

In this chapter we depict our lyrics-to-audio alignment (LAA) baseline system. It is a phonetic recognizer, based on phone HMMs. To date most of the studies on LAA are based on the phonetic recognizer approach, as described in Section 2.2. The goal is to describe the key elements of the baseline approach, which are not related to the complementary context of lyrics. In this way we ‘set the scene’ for the methodologies that consider context - the main contribution of this thesis. They will be treated in one of the following two chapters. To this end, we start this chapter through the key elements of phonetic alignment and describe which existing methodologies we plugged in. Some of these are tailored to the specific characteristics of OTMM (see Section 2.1.1). In particular, we explain how we utilized a method for linking structural sections of the composition to their respective audio segments in a recording. Further, we describe the benefit of a predominant melody extraction method. We comment on tuning their parameters. We present in more details the construction of the phoneme network from the lyrics transcription, for which some rules for Turkish language are required.

A major contribution of this chapter is a strategy to represent phonemes in Turkish language by mapping them to phonemes in English. This enables the use of a reliable model for English as a viable replacement for Turkish, for which the available training material is scarce. We also describe the datasets used to evaluate the LAA methods, presented throughout this thesis. Com-

Demo


Gel Güzelim Çamlıca'ya

by Münir Nurettin Selçuk

Album

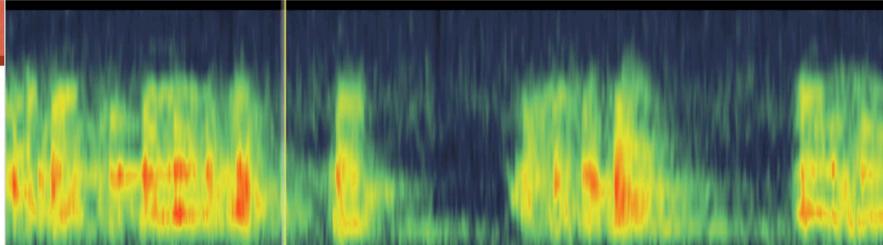
Geçmişten
Günümüze Türk
Müziği - Kalplerden
Dudaklara (Various
Artists)

Compositions

Gel Güzelim
Çamlıca'ya

Performers

Münir Nurettin
Selçuk(Voice)


00:08 00:10 00:12 00:14 00:16

Gel güzelim Çamlıca'ya bu gece
 Gel güzelim Çamlıca'ya bu gece
 Gün doğmadan a canım görüşelim gizlice
 Gün doğmadan a canım görüşelim gizlice
 Bülbüllerin efganını dinleyelim yanyana
 Bülbüllerin efganını dinleyelim yanyana
 Kumru gibi a canım sevişelim can cana
 Kumru gibi a canım sevişelim can cana

00:21
00:10
03:16

Demo


Gel Güzelim Çamlıca'ya

by Münir Nurettin Selçuk

Album

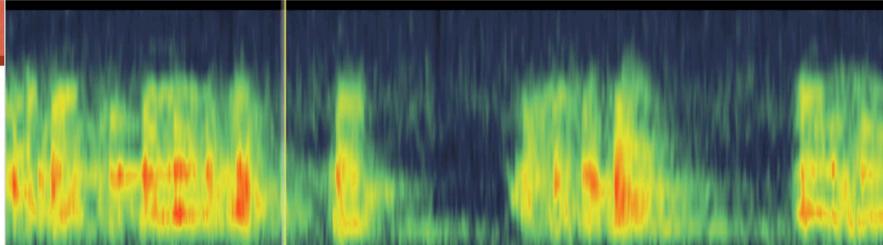
Geçmişten
Günümüze Türk
Müziği - Kalplerden
Dudaklara (Various
Artists)

Compositions

Gel Güzelim
Çamlıca'ya

Performers

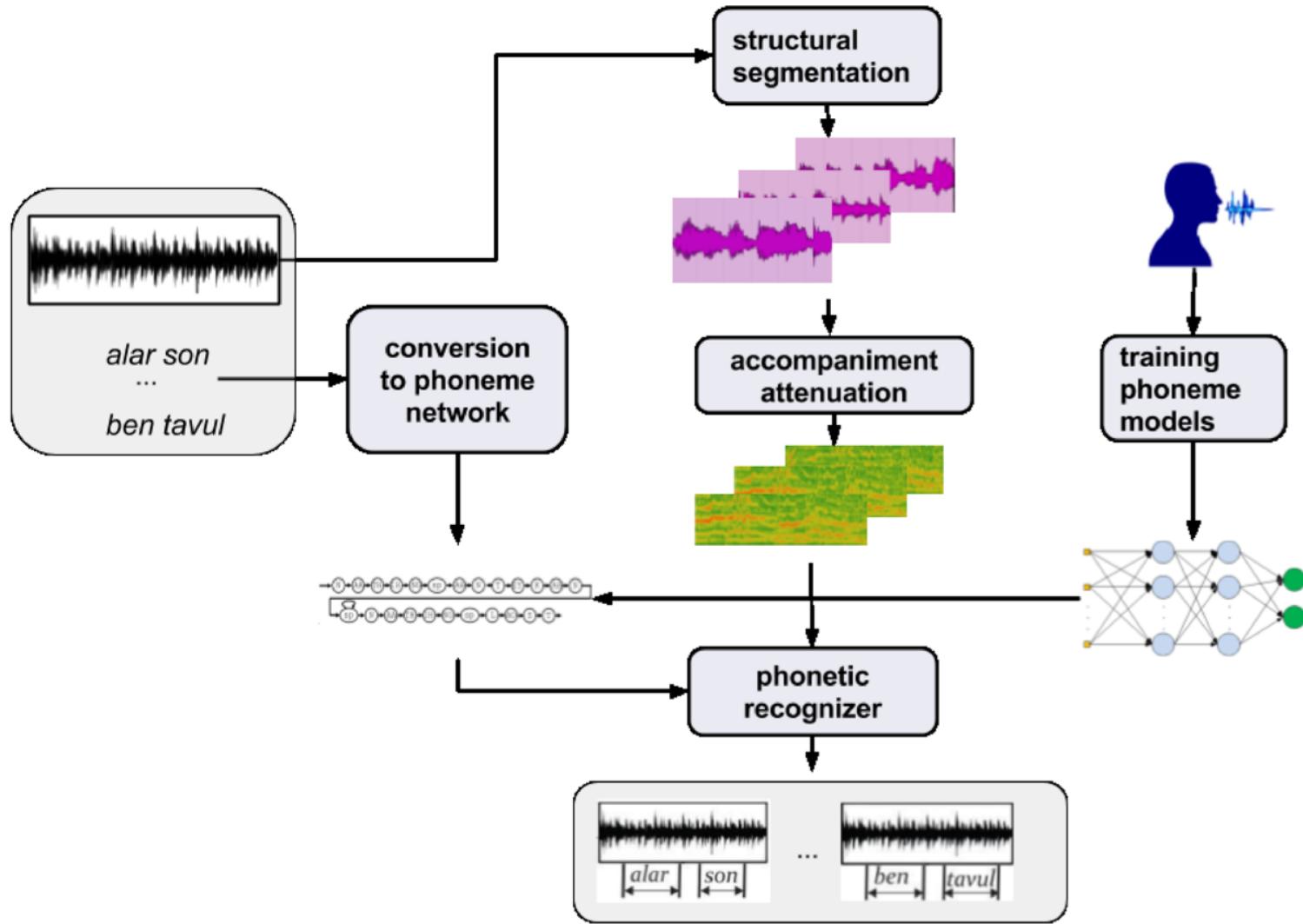
Münir Nurettin
Selçuk(Voice)


00:08 00:10 00:12 00:14 00:16

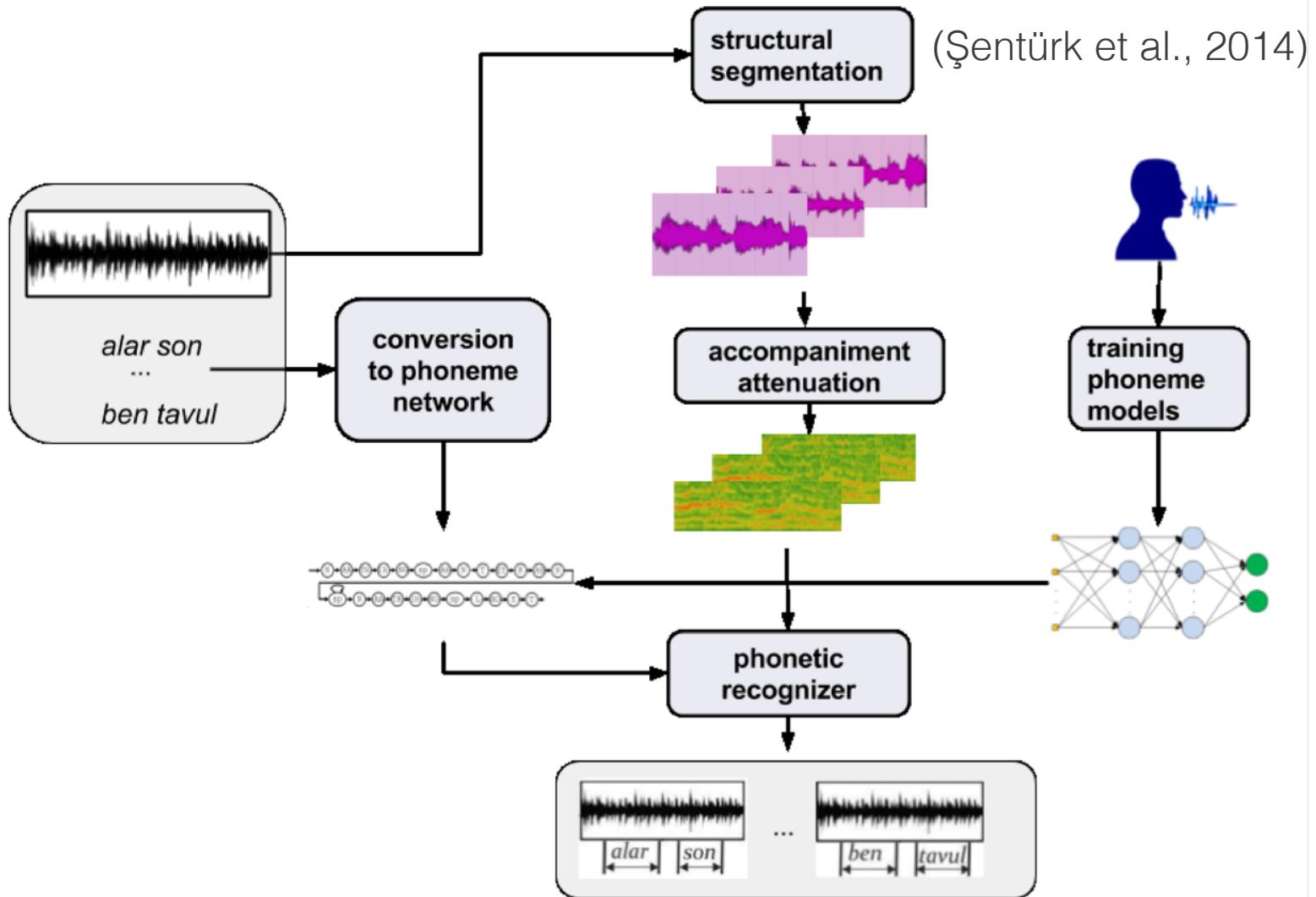
Gel güzelim Çamlıca'ya bu gece
 Gel güzelim Çamlıca'ya bu gece
 Gün doğmadan a canım görüşelim gizlice
 Gün doğmadan a canım görüşelim gizlice
 Bülbüllerin efganını dinleyelim yanyana
 Bülbüllerin efganını dinleyelim yanyana
 Kumru gibi a canım sevişelim can cana
 Kumru gibi a canım sevişelim can cana

00:21
00:10
03:16

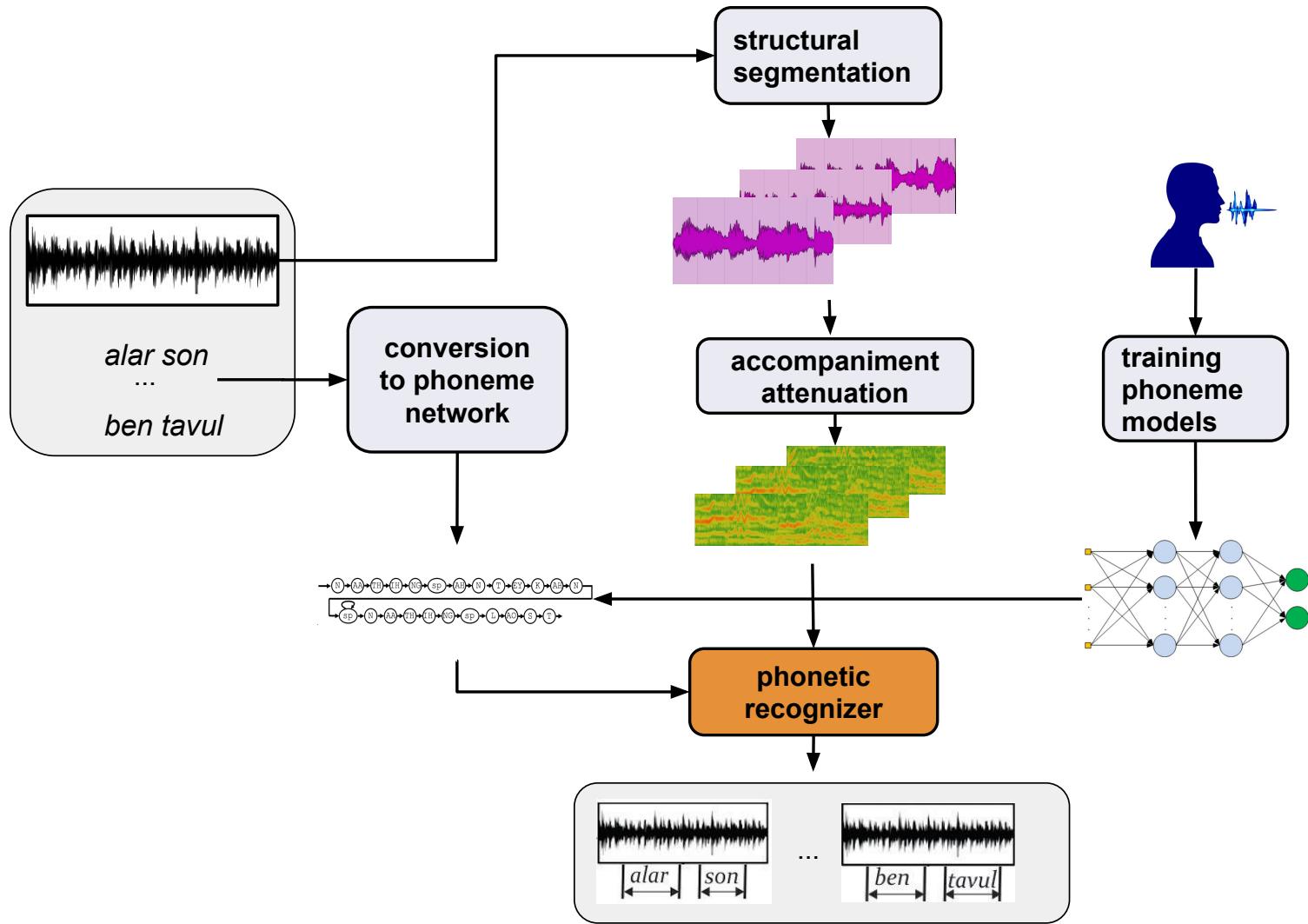
Steps of baseline alignment



Steps of baseline alignment

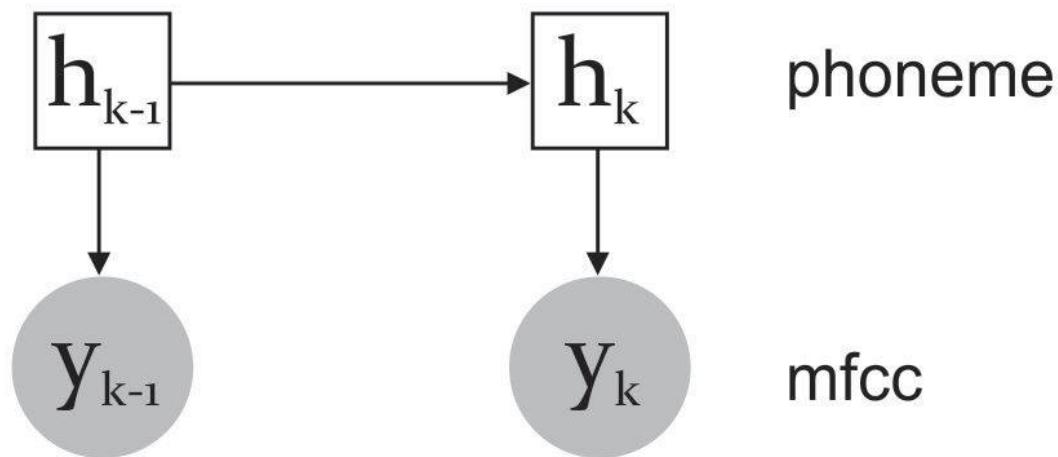


Phonetic recognizer

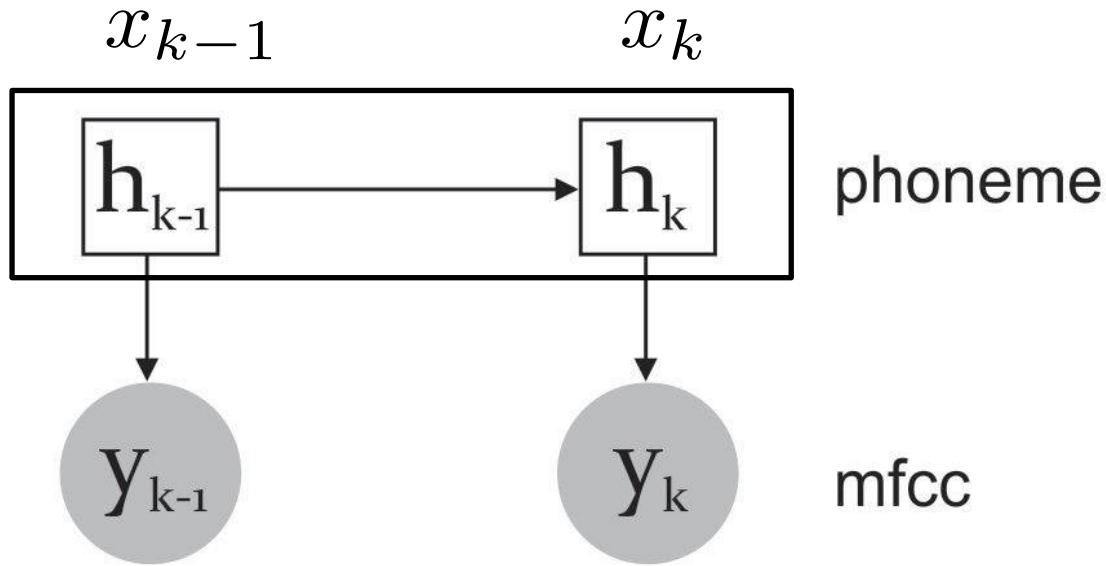


Model components

- based on Hidden Markov models

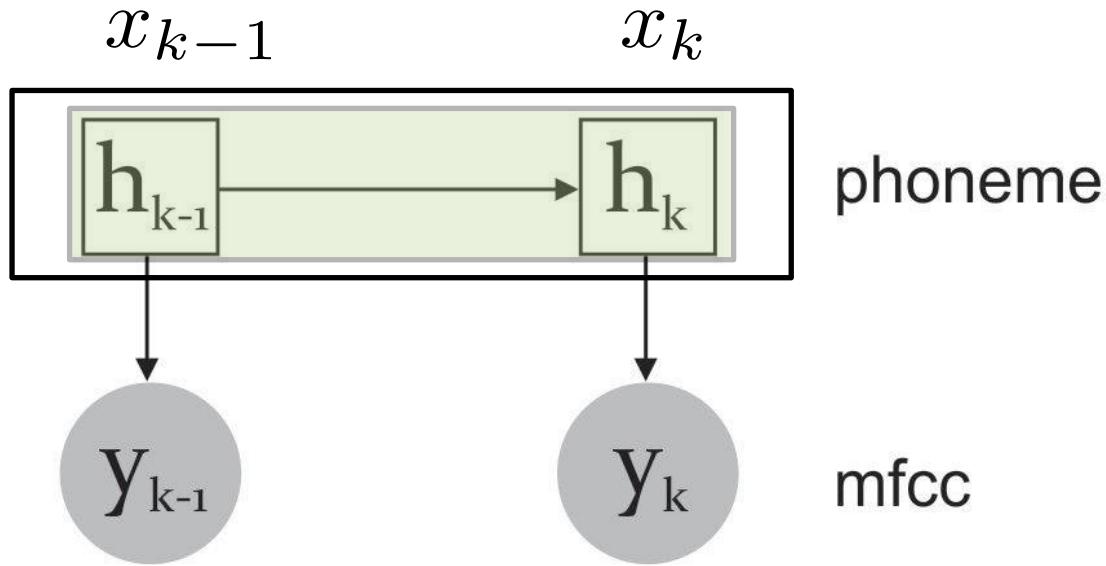


Model components



$$P(x_{1:K}, y_{1:K}) = P(x_0) \prod_{k=1}^K P(x_k | x_{k-1}) P(y_k | x_k)$$

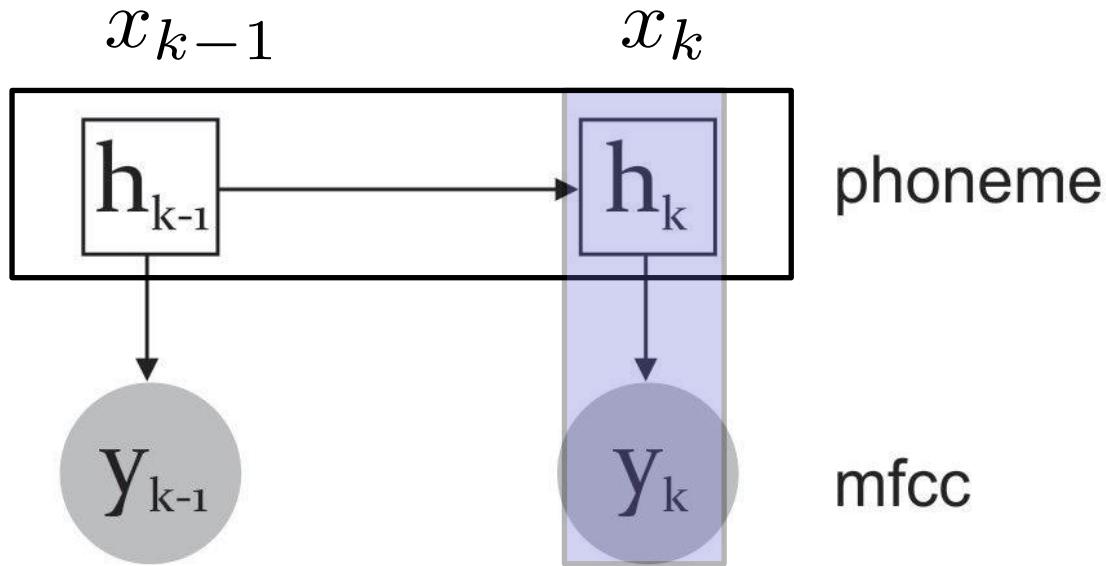
Model components



$$P(x_{1:K}, y_{1:K}) = P(x_0) \prod_{k=1}^K P(x_k | x_{k-1}) P(y_k | x_k)$$

Transition model

Model components



$$P(x_{1:K}, y_{1:K}) = P(x_0) \prod_{k=1}^K P(x_k | x_{k-1}) P(y_k | x_k)$$

Observation model

Viterbi decoding

Viterbi decoding

$$a_{ij} = P(x_k = j \mid x_{k-1} = i)$$

Viterbi decoding

$$a_{ij} = P(x_k = j \mid x_{k-1} = i)$$

$$b_j(O_k) = P(y_k = O_k \mid x_k = j)$$

Viterbi decoding

$$a_{ij} = P(x_k = j \mid x_{k-1} = i)$$

$$b_j(O_k) = P(y_k = O_k \mid x_k = j)$$

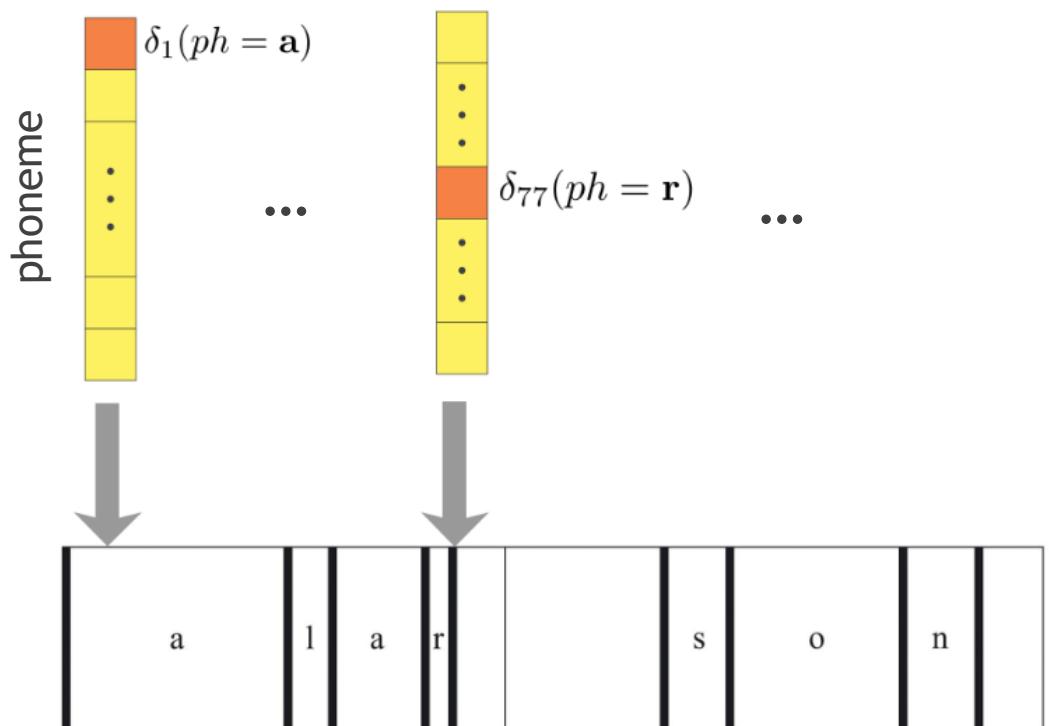
$$\delta_k(j) = \max_{i \in (j, j-1)} \delta_{k-1}(i) a_{ij} b_j(O_k) \quad (\text{Rabiner, 1989})$$

Viterbi decoding

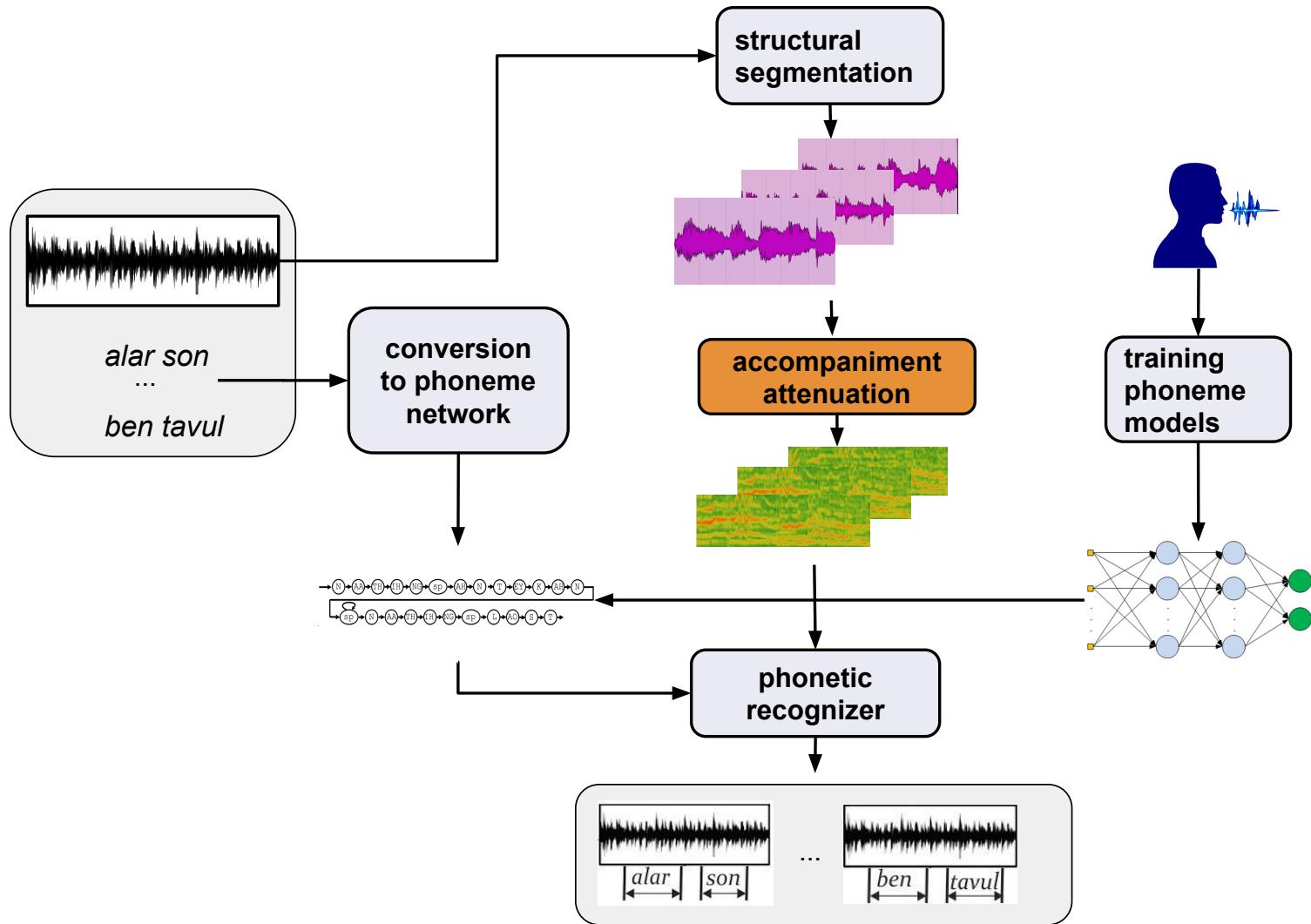
$$a_{ij} = P(x_k = j \mid x_{k-1} = i)$$

$$b_j(O_k) = P(y_k = O_k \mid x_k = j)$$

$$\delta_k(j) = \max_{i \in (j, j-1)} \delta_{k-1}(i) a_{ij} b_j(O_k) \quad (\text{Rabiner, 1989})$$



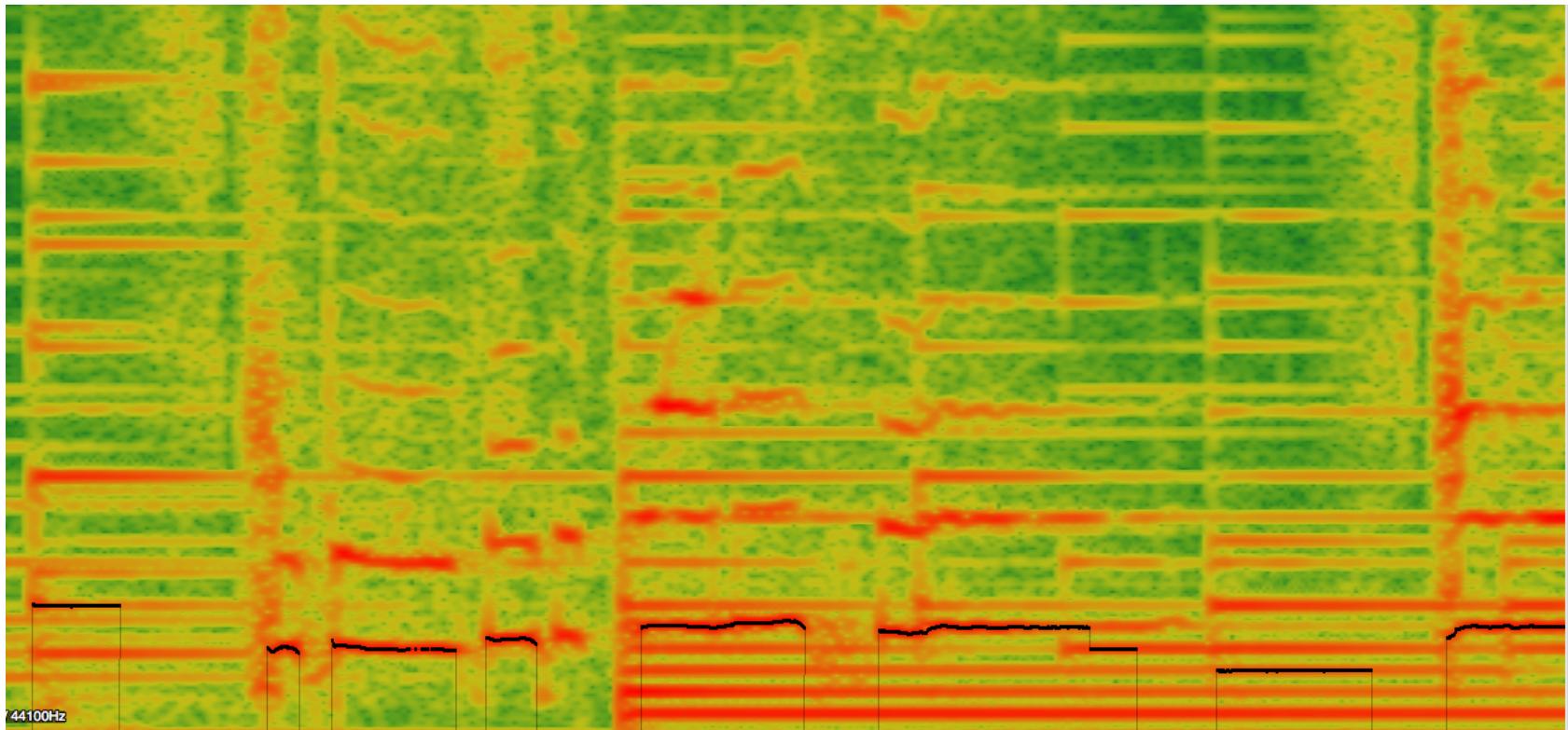
Accompaniment attenuation



Accompaniment attenuation

1. Extract fundamental frequency

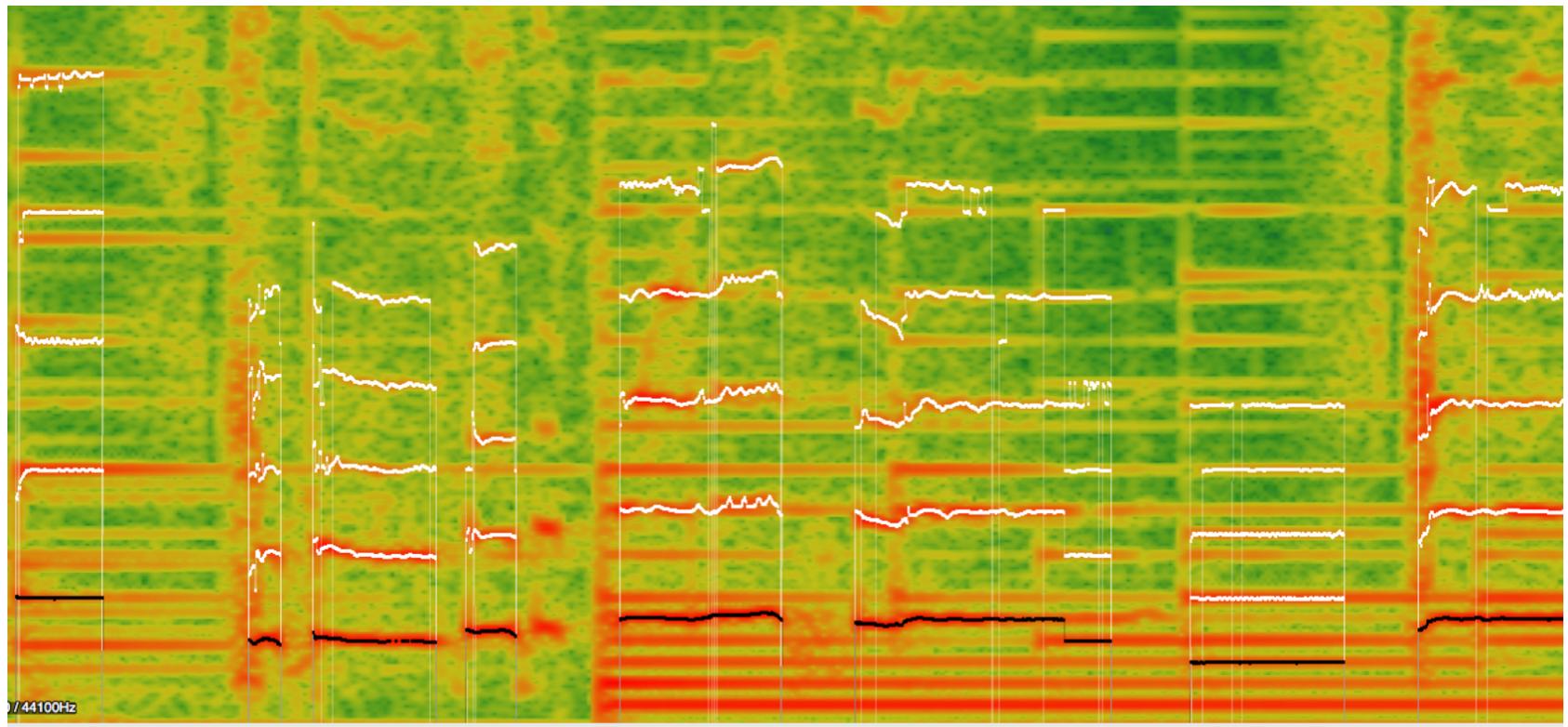
OTMM-tailored predominant melody (Atlı et al., 2014)



Accompaniment attenuation

2. Extract harmonic partials

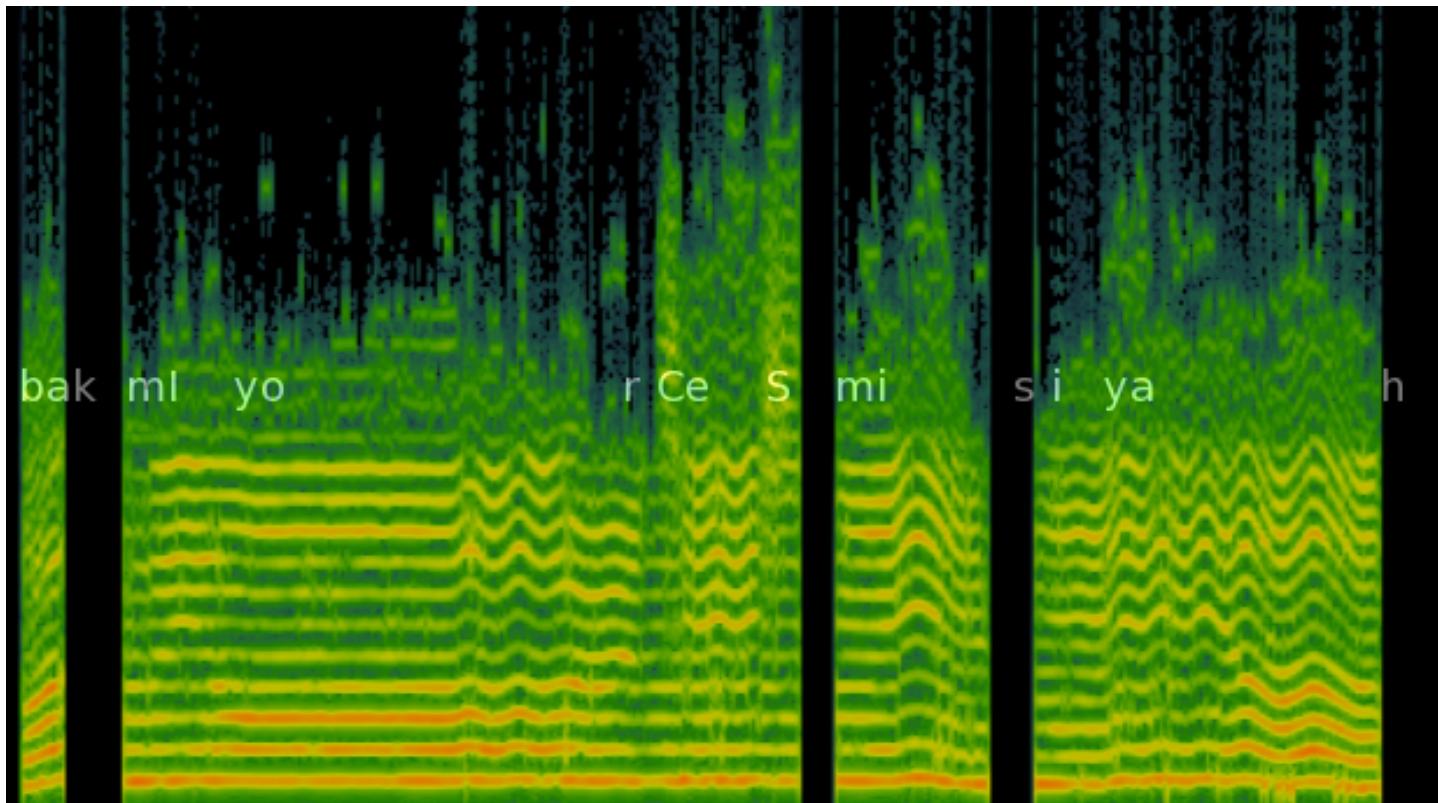
harmonic model (Serra, 1989)



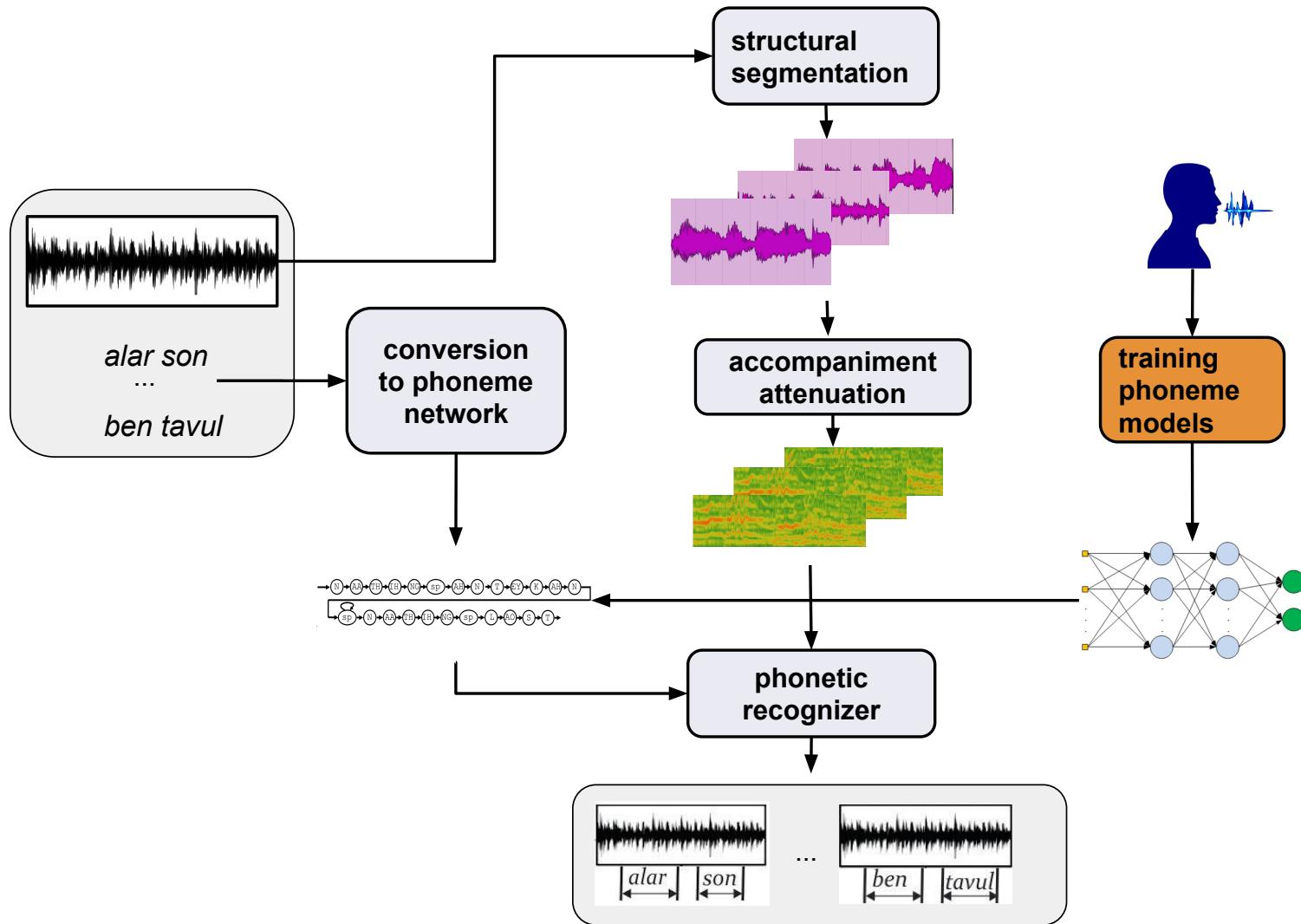
Accompaniment attenuation

3. Resynthesize harmonic partials

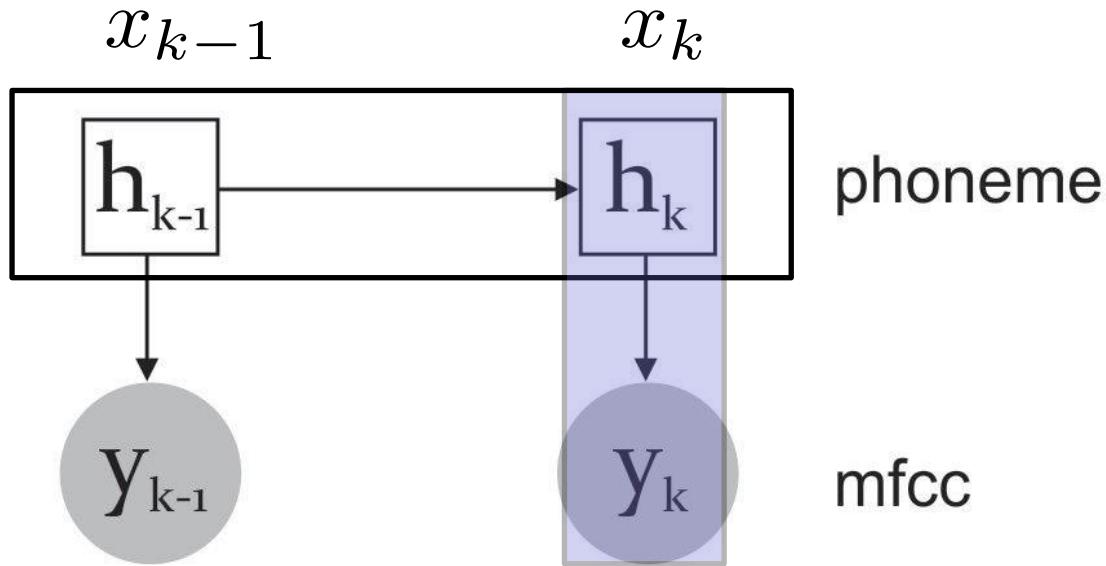
overlap-add synthesis (Serra, 1989)



Training phoneme models



Training phoneme models



$$P(x_{1:K}, y_{1:K}) = P(x_0) \prod_{k=1}^K P(x_k | x_{k-1}) P(y_k | x_k)$$

Observation model

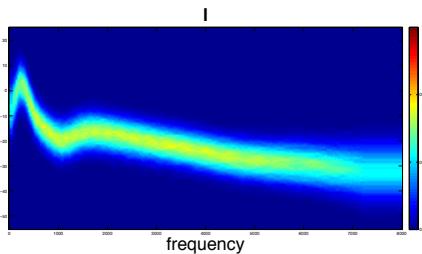
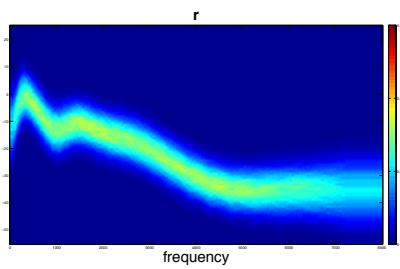
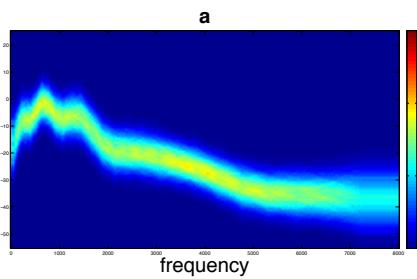
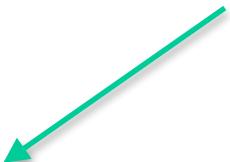
Training phoneme models

- GMMs with MFCCs
 - on Turkish speech
 - 9 mixtures



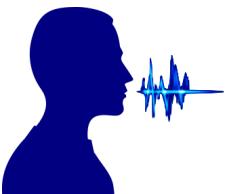
Training phoneme models

- GMMs with MFCCs
 - on Turkish speech
 - 9 mixtures

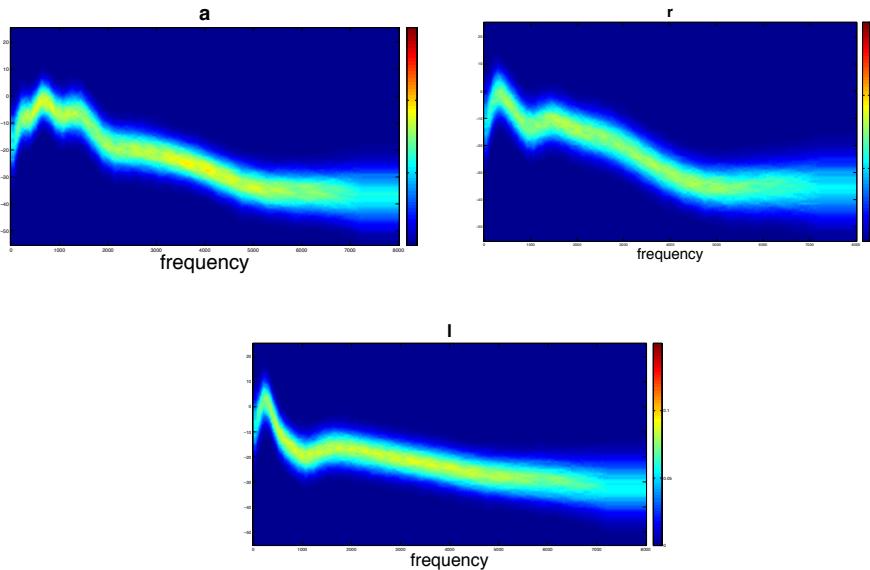
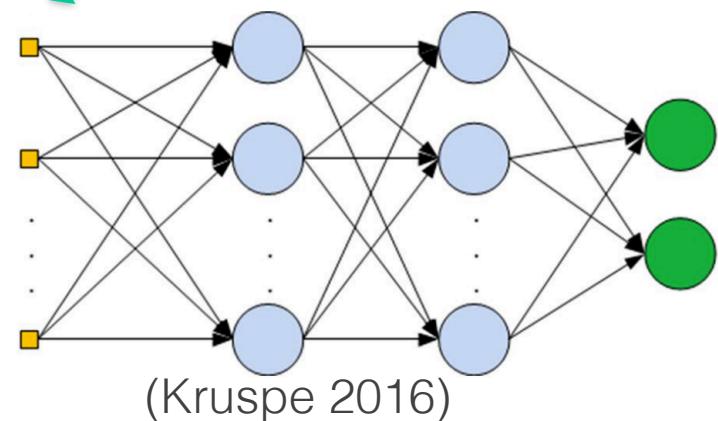


Training phoneme models

- GMMs with MFCCs
 - on Turkish speech
 - 9 mixtures

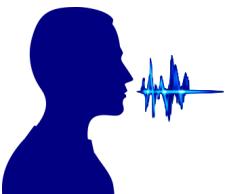


- MLP with MFCCs
 - on English a cappella signing
 - 1024-850-1024 neurons

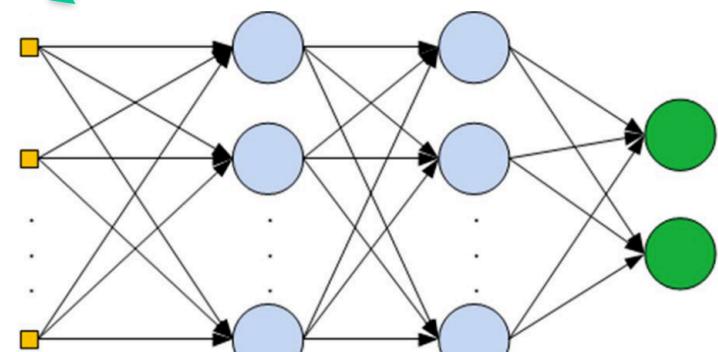


Training phoneme models

- GMMs with MFCCs
 - on Turkish speech
 - 9 mixtures

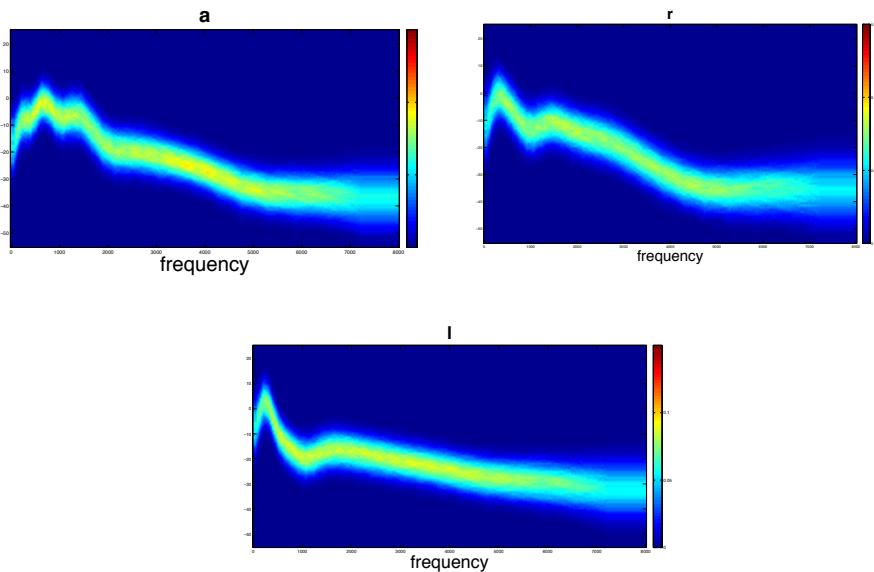


- MLP with MFCCs
 - on English a cappella signing
 - 1024-850-1024 neurons



(Kruspe 2016)

- mapping Turkish phonemes to English ones (MLP-DirectM)



Multi-instrumental lyrics OTMM dataset

- 13 recordings from the *şarkı* form (19 minutes)
- Music scores with section annotations
- Annotations of lyrics phrases

Kimseye Etmem Şikayet
Nihâvent Şarkı

Usul: K. Cucuna
 $\frac{2}{4} \Rightarrow 3$ Dk 29 Sn

Beste: Kemâni Sarkis Efendi (1885 - 12/12/1944)
Güfte: ?



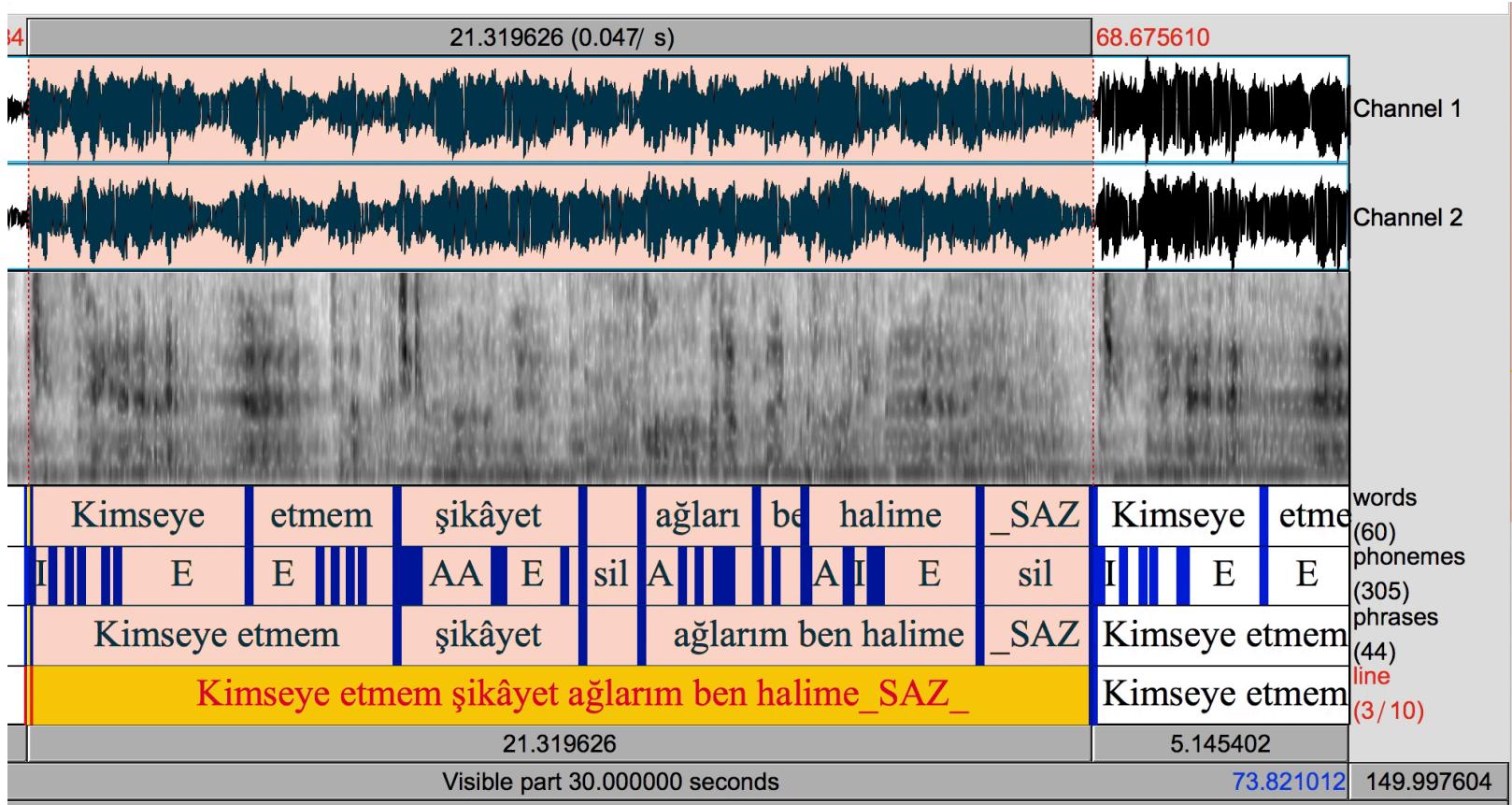
Kimseye Etmem Şikayet
Nihâvent Şarkı

Beste: Kemâni Sarkis Efendi (1885 - 12/12/1944)
Güfte: ?

ARANAĞME... et mem si kâ yet
Kim se ye
1 2
ağ la rim ben ha li me SAZ 1 2
Tit re rim müc rim gi bi bak
tik çä is tik ba li me SAZ [SON] SAZ
Per de i zul met çé kil müş
kor ka rum ik ba li me SAZ
1 2
kor ka rum ik ba li me SAZ

Kimseye etmem şikayet ağlarım ben halime
Titterim mücerim gibi baktıkça istikbalime
Perdeyi zulmet çekilmiş korkarım ikbalime
Titterim mücerim gibi baktıkça istikbalime

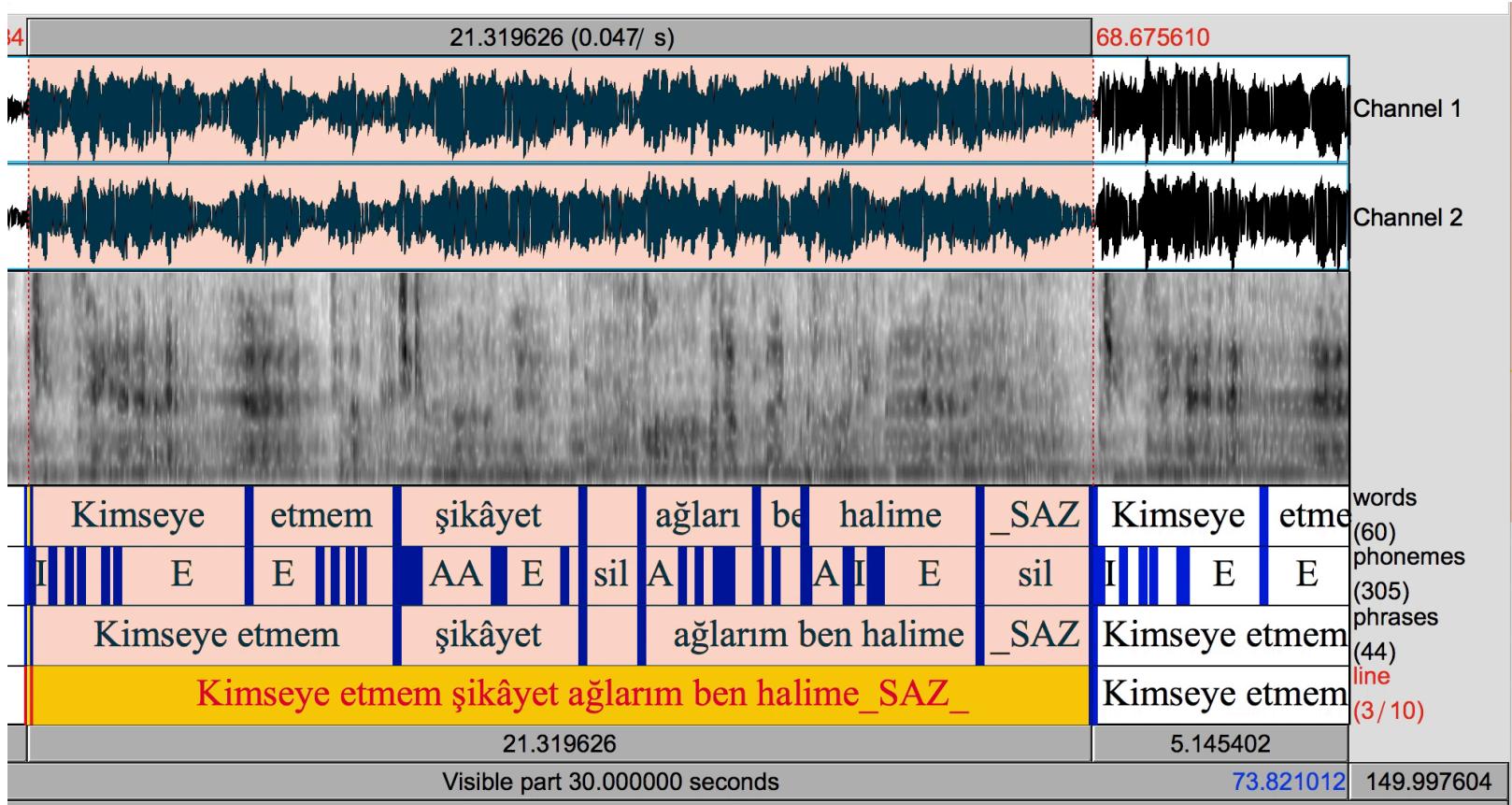
Multi-instrumental lyrics OTMM dataset



<http://compmusic.upf.edu/turkish-sarki>

Sec 3.2.1

Multi-instrumental lyrics OTMM dataset



A cappella lyrics OTMM dataset

<http://compmusic.upf.edu/turkish-makam-acapella-sections-dataset>

Sec 3.2.2

A cappella lyrics OTMM dataset

- A cappella versions recorded in-sync with the original

<http://compmusic.upf.edu/turkish-makam-acapella-sections-dataset>

A cappella lyrics OTMM dataset

- A cappella versions recorded in-sync with the original

<http://compmusic.upf.edu/turkish-makam-acapella-sections-dataset>

Experiments

acoustic model	data	AA	AH	accuracy	error
GMMs	a cappella	-	-	70.2	1.14
MLP-DirectM	a cappella	-	-	79.2	0.57
GMMs	multi-instrumental	N	-	59.1	2.15
GMMs	multi-instrumental	Y	N	63.2	1.98
GMMs	multi-instrumental	Y	Y	67.5	1.26
Mesaros	multi-instrumental	-	-	-	1.4
Fujihara	multi-instrumental	-	-	85.2	-

Dzhambazov, Georgi, Sertan Şentürk, and Xavier Serra (2014). Automatic lyrics-to-audio alignment in classical Turkish music. In 4th International Workshop on Folk Music Analysis (FMA 2014)

Experiments

acoustic model	data	AA	AH	accuracy	error
GMMs	a cappella	-	-	70.2	1.14
MLP-DirectM	a cappella	-	-	79.2	0.57
GMMs	multi-instrumental	N	-	59.1	2.15
GMMs	multi-instrumental	Y	N	63.2	1.98
GMMs	multi-instrumental	Y	Y	67.5	1.26
Mesaros	multi-instrumental	-	-	-	1.4
Fujihara	multi-instrumental	-	-	85.2	-

Dzhambazov, Georgi, Sertan Şentürk, and Xavier Serra (2014). Automatic lyrics-to-audio alignment in classical Turkish music. In 4th International Workshop on Folk Music Analysis (FMA 2014)

Experiments

acoustic model	data	AA	AH	accuracy	error
GMMs	a cappella	-	-	70.2	1.14
MLP-DirectM	a cappella	-	-	79.2	0.57
GMMs	multi-instrumental	N	-	59.1	2.15
GMMs	multi-instrumental	Y	N	63.2	1.98
GMMs	multi-instrumental	Y	Y	67.5	1.26
Mesaros	multi-instrumental	-	-	-	1.4
Fujihara	multi-instrumental	-	-	85.2	-

Dzhambazov, Georgi, Sertan Şentürk, and Xavier Serra (2014). Automatic lyrics-to-audio alignment in classical Turkish music. In 4th International Workshop on Folk Music Analysis (FMA 2014)

Experiments

acoustic model	data	AA	AH	accuracy	error
GMMs	a cappella	-	-	70.2	1.14
MLP-DirectM	a cappella	-	-	79.2	0.57
GMMs	multi-instrumental	N	-	59.1	2.15
GMMs	multi-instrumental	Y	N	63.2	1.98
GMMs	multi-instrumental	Y	Y	67.5	1.26
Mesaros	multi-instrumental	-	-	-	1.4
Fujihara	multi-instrumental	-	-	85.2	-

Dzhambazov, Georgi, Sertan Şentürk, and Xavier Serra (2014). Automatic lyrics-to-audio alignment in classical Turkish music. In 4th International Workshop on Folk Music Analysis (FMA 2014)

Experiments

acoustic model	data	AA	AH	accuracy	error
GMMs	a cappella	-	-	70.2	1.14
MLP-DirectM	a cappella	-	-	79.2	0.57
GMMs	multi-instrumental	N	-	59.1	2.15
GMMs	multi-instrumental	Y	N	63.2	1.98
GMMs	multi-instrumental	Y	Y	67.5	1.26
Mesaros	multi-instrumental	-	-	-	1.4
Fujihara	multi-instrumental	-	-	85.2	-

Dzhambazov, Georgi, Sertan Şentürk, and Xavier Serra (2014). Automatic lyrics-to-audio alignment in classical Turkish music. In 4th International Workshop on Folk Music Analysis (FMA 2014)

Experiments

acoustic model	data	AA	AH	accuracy	error
GMMs	a cappella	-	-	70.2	1.14
MLP-DirectM	a cappella	-	-	79.2	0.57
GMMs	multi-instrumental	N	-	59.1	2.15
GMMs	multi-instrumental	Y	N	63.2	1.98
GMMs	multi-instrumental	Y	Y	67.5	1.26
Mesaros	multi-instrumental	-	-	-	1.4
Fujihara	multi-instrumental	-	-	85.2	-

Dzhambazov, Georgi, Sertan Şentürk, and Xavier Serra (2014). Automatic lyrics-to-audio alignment in classical Turkish music. In 4th International Workshop on Folk Music Analysis (FMA 2014)

Experiments

acoustic model	data	AA	AH	accuracy	error
GMMs	a cappella	-	-	70.2	1.14
MLP-DirectM	a cappella	-	-	79.2	0.57
GMMs	multi-instrumental	N	-	59.1	2.15
GMMs	multi-instrumental	Y	N	63.2	1.98
GMMs	multi-instrumental	Y	Y	67.5	1.26
Mesaros	multi-instrumental	-	-	-	1.4
Fujihara	multi-instrumental	-	-	85.2	-

Dzhambazov, Georgi, Sertan Şentürk, and Xavier Serra (2014). Automatic lyrics-to-audio alignment in classical Turkish music. In 4th International Workshop on Folk Music Analysis (FMA 2014)

Outline

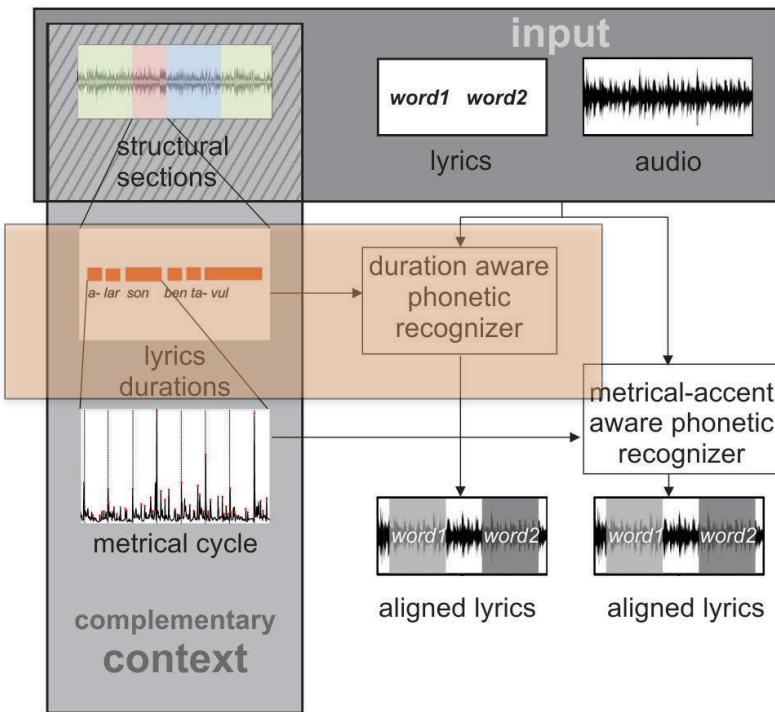
- Setting the stage
 - Introduction (Chapter 1)
 - Background (Chapter 2)
 - Baseline lyrics-to-audio alignment (Chapter 3)
- Main
 - Lyrics-duration aware lyrics-to-audio alignment (Chapter 4)
 - Metrical-accent aware lyrics-to-audio alignment (Chapter 5)
- Conclusion (Chapter 6) and Applications (Appendix)

Outline

- Setting the stage
 - Introduction (Chapter 1)
 - Background (Chapter 2)
 - Baseline lyrics-to-audio alignment (Chapter 3)
- Main
 - Lyrics-duration aware lyrics-to-audio alignment (Chapter 4)
 - Metrical-accent aware lyrics-to-audio alignment (Chapter 5)
- Conclusion (Chapter 6) and Applications (Appendix)

Chapter 4

Lyrics-to-audio Alignment with Middle-level Complementary Context



Context: Syllable durations within a lyrics line

Chapter 4

Lyrics-to-audio Alignment with Middle-level Complementary Context

4.1 Introduction

In this chapter, we propose how to improve the baseline lyrics-to-audio alignment method by considering some context facets, complementary to lyrics. We focus on one particular middle-level facet - the temporal structure of the sung lyrics line. Studies of sheet music have indicated that there is a correlation between the accents of sung syllables and the accents in the melodic motif (Nichols et al., 2009). Singers may often prolong or reduce the duration of some syllables, in order to align them with the accents in the melody.

Music scores provide important contextual information complementary to lyrics, including note values. Nevertheless, the length of sung syllables could deviate considerably from the durations indicated in the music score. Singers in OTMM, in particular, tend to deviate from the music score to a significantly larger extent, in comparison, for example, to classical music. To address this, we propose an extension of the phonetic recognizer that models explicitly some reference syllable durations. The proposed duration aware model is designed to accommodate duration variations. The major technical contribution of this chapter is the derivation an inference method for the model. The reference syllable durations are obtained from the music score. To our knowledge, this study is the first application of modeling of music-score-induced durations for sung syllables.

Chapter 4 outline

- 4.3 Duration aware probabilistic model
- 4.4 Durations derived from music score (on OTMM)
- 4.5 Durations derived from music principles (on jingju)

Waiting time distribution

$$P_i(d) = (1 - a_{ij})^d$$

exponential distribution

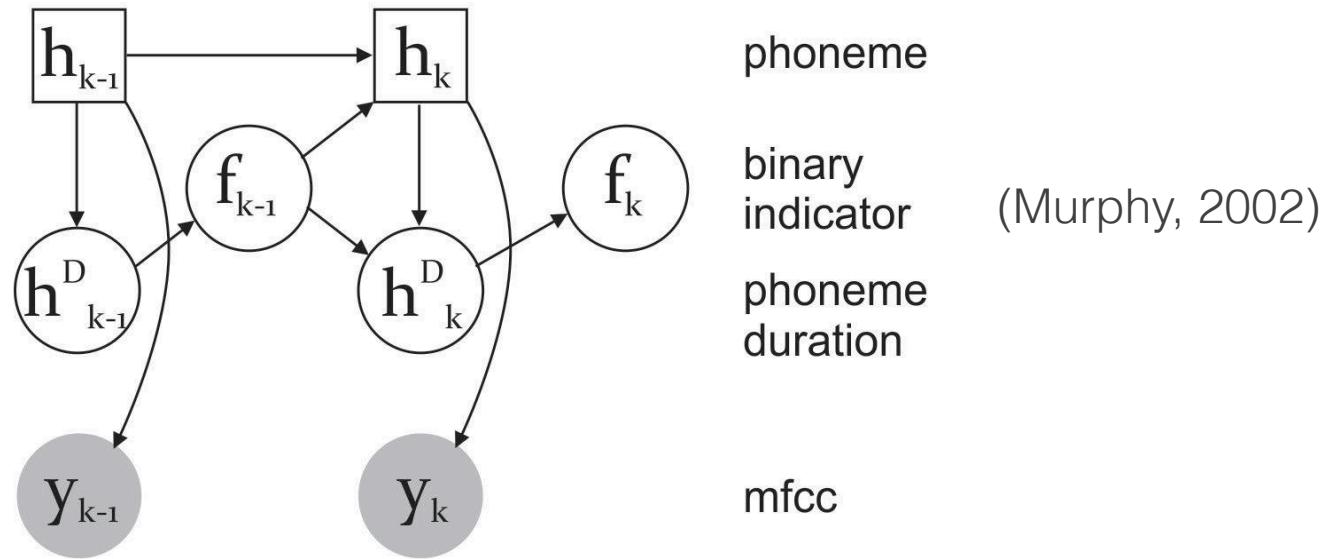
Waiting time distribution

$$P_i(d) = (1 - a_{ij})^d$$

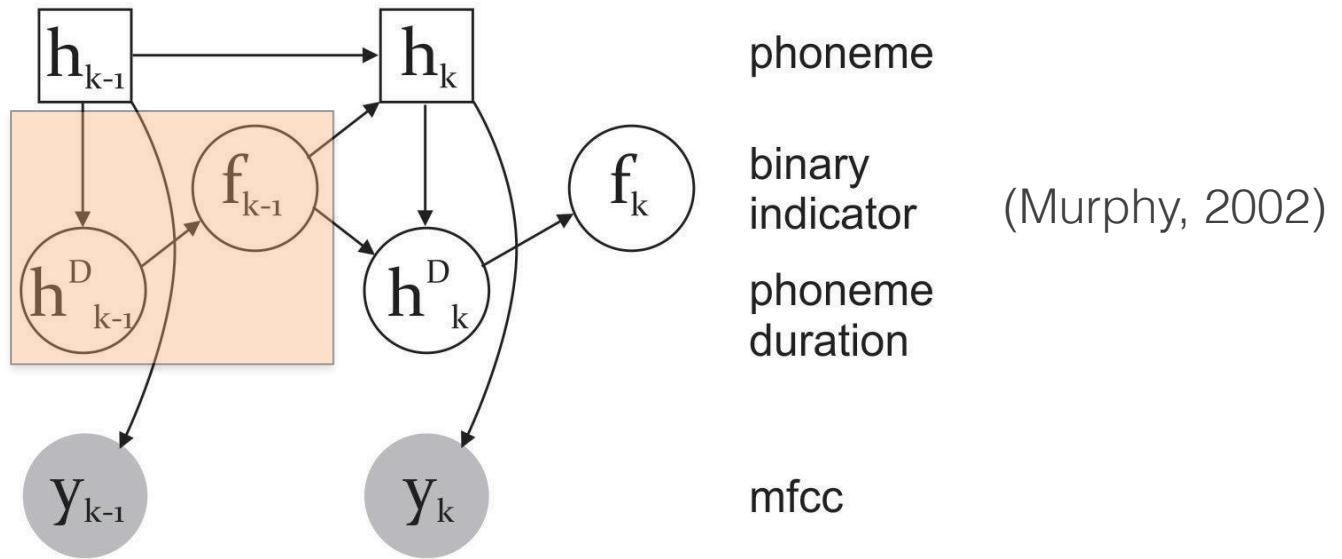
exponential distribution



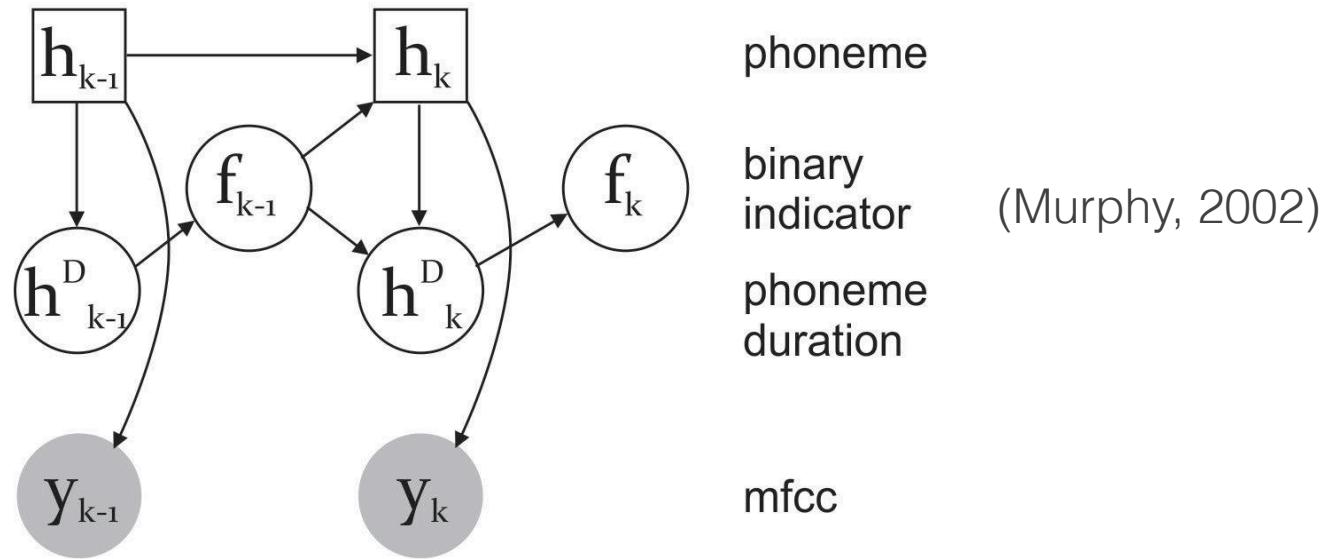
Model components



Model components



Model components



Viterbi decoding

$$\delta_k(j) = \max_{i \in (j, j-1)} \delta_{k-1}(i) [a_{ij}] [b_j(O_k)]$$


Viterbi decoding

$$\delta_k(j) = \max_{i \in (j, j-1)} \delta_{k-1}(i) [a_{ij}] b_j(O_k)$$



$$\delta_k(j) = \max_d \{ \delta_{k-d}(j-1) [a_{ij}] P_j(d) [B_k(j, d)] \} \text{ (Chen et al., 2012)}$$

Viterbi decoding

$$\delta_k(j) = \max_{i \in (j, j-1)} \delta_{k-1}(i) a_{ij} b_j(O_k)$$



$$\delta_k(j) = \max_d \{ \delta_{k-d}(j-1) a_{ij} P_j(d) B_k(j, d) \} \text{ (Chen et al., 2012)}$$

$$B_k(j, d) = \prod_{s=k-d+1}^k b_j(O_s)$$

Viterbi decoding

$$\delta_k(j) = \max_{i \in (j, j-1)} \delta_{k-1}(i) a_{ij} b_j(O_k)$$



$$\delta_k(j) = \max_d \{ \delta_{k-d}(j-1) a_{ij} P_j(d) B_k(j, d) \} \text{ (Chen et al., 2012)}$$

$$B_k(j, d) = \prod_{s=k-d+1}^k b_j(O_s)$$

Viterbi decoding

$$\delta_k(j) = \max_{i \in (j, j-1)} \delta_{k-1}(i) a_{ij} b_j(O_k)$$



$$\delta_k(j) = \max_d \{ \delta_{k-d}(j-1) a_{ij} P_j(d) B_k(j, d) \} \text{ (Chen et al., 2012)}$$

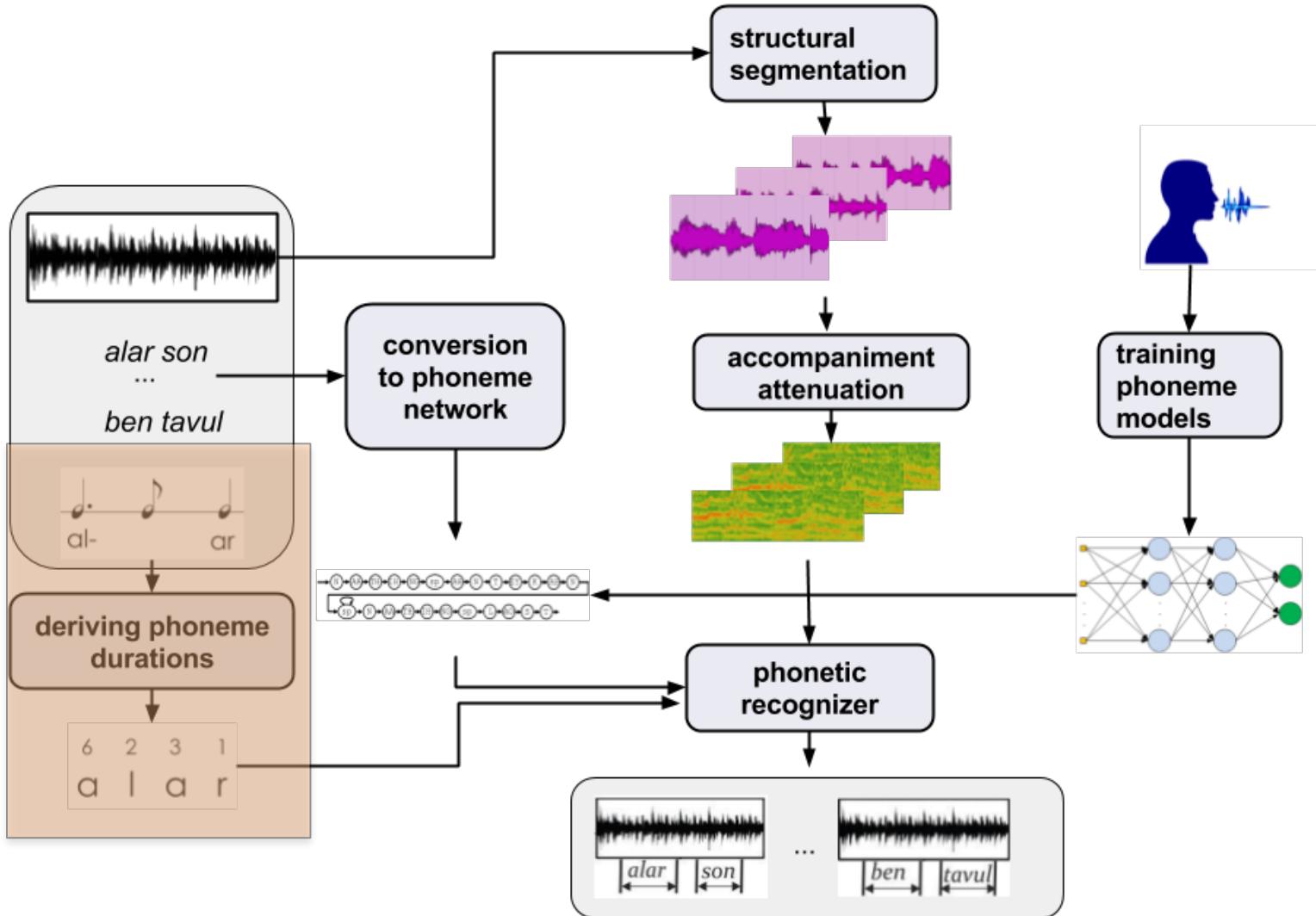
$$B_k(j, d) = \prod_{s=k-d+1}^k b_j(O_s)$$

syllable 1 syllable 2 syllable 3

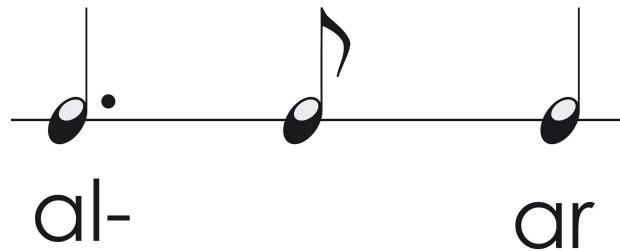


Chapter 4.4

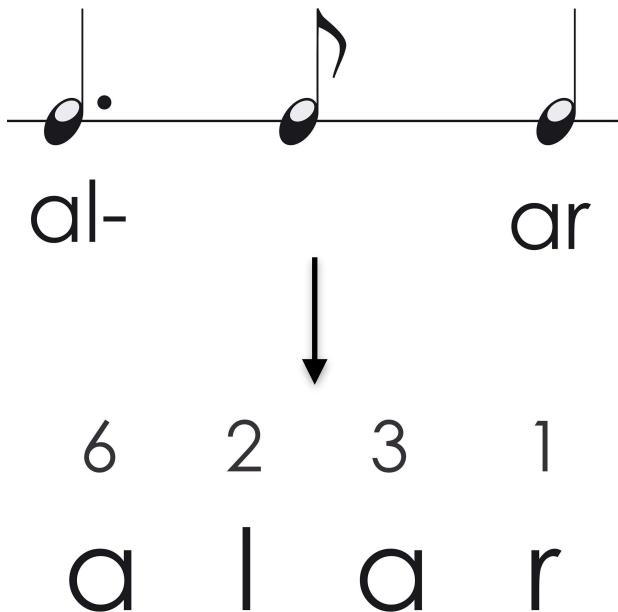
Durations derived from music score



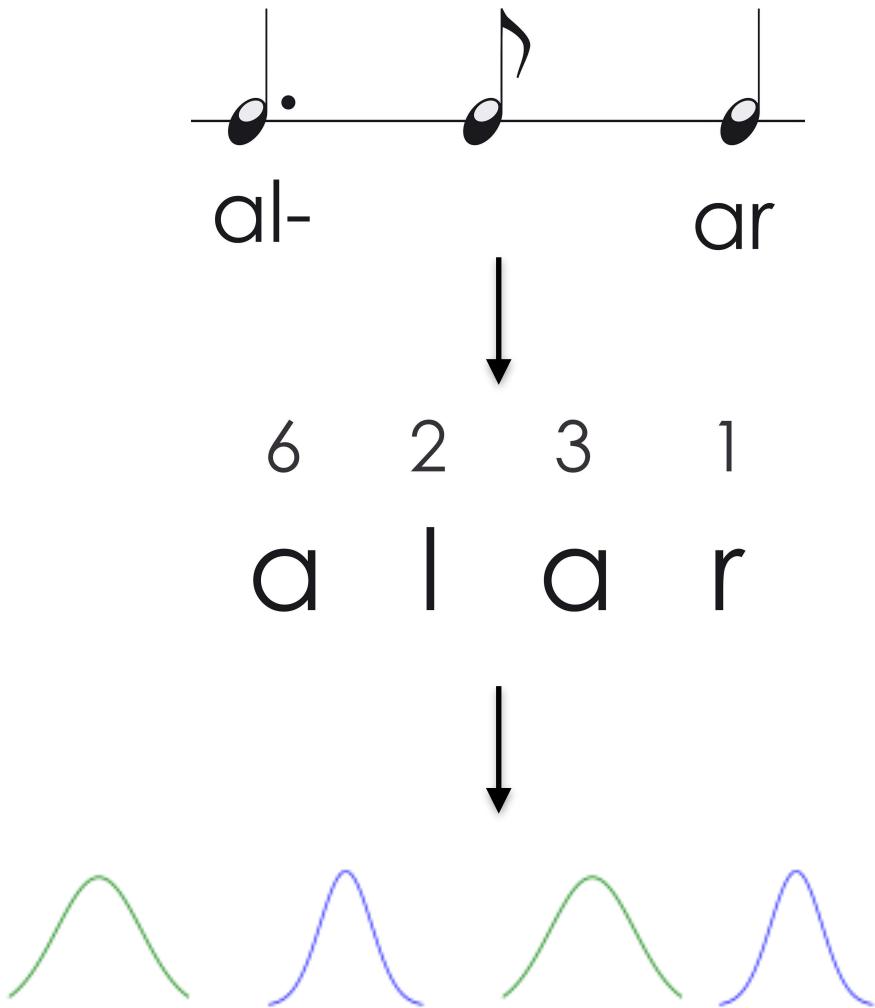
Deriving phoneme durations



Deriving phoneme durations

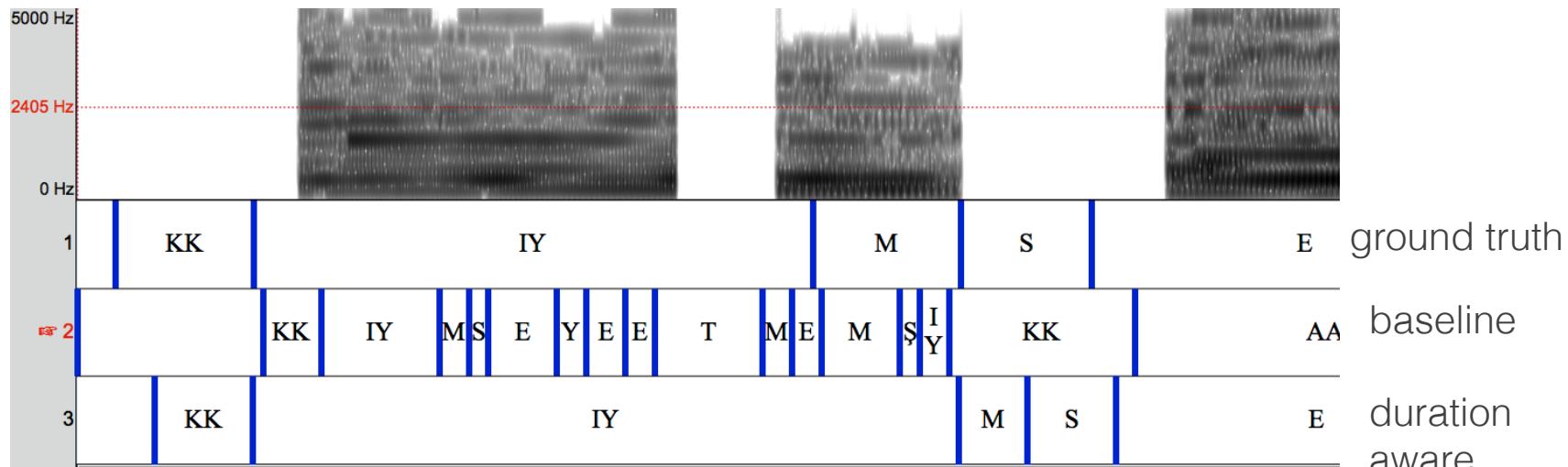


Deriving phoneme durations



Experiments

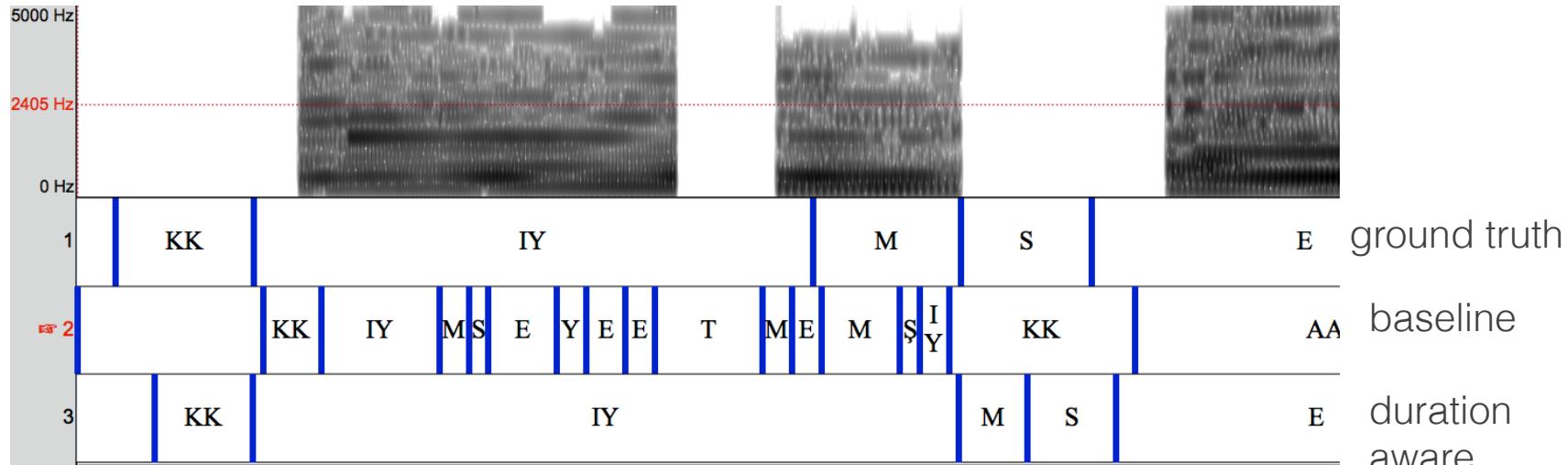
model	data	accuracy	error
baseline GMM	a cappella OTMM	70.2	1.14
duration aware GMM	a cappella OTMM	90.04	0.26
baseline GMM	multi-instrumental OTMM	67.46	1.26
duration aware GMM	multi-instrumental OTMM	77.74	0.63
(Mesaros, 2008)	multi-instrumental	-	1.4
(Fujihara, 2011)	multi-instrumental	85.2	-



Dzhambazov, Georgi and Xavier Serra (2015). Modeling of phoneme durations for alignment between polyphonic audio and lyrics. In Sound and Music Computing Conference 2015 (SMC 2015)

Experiments

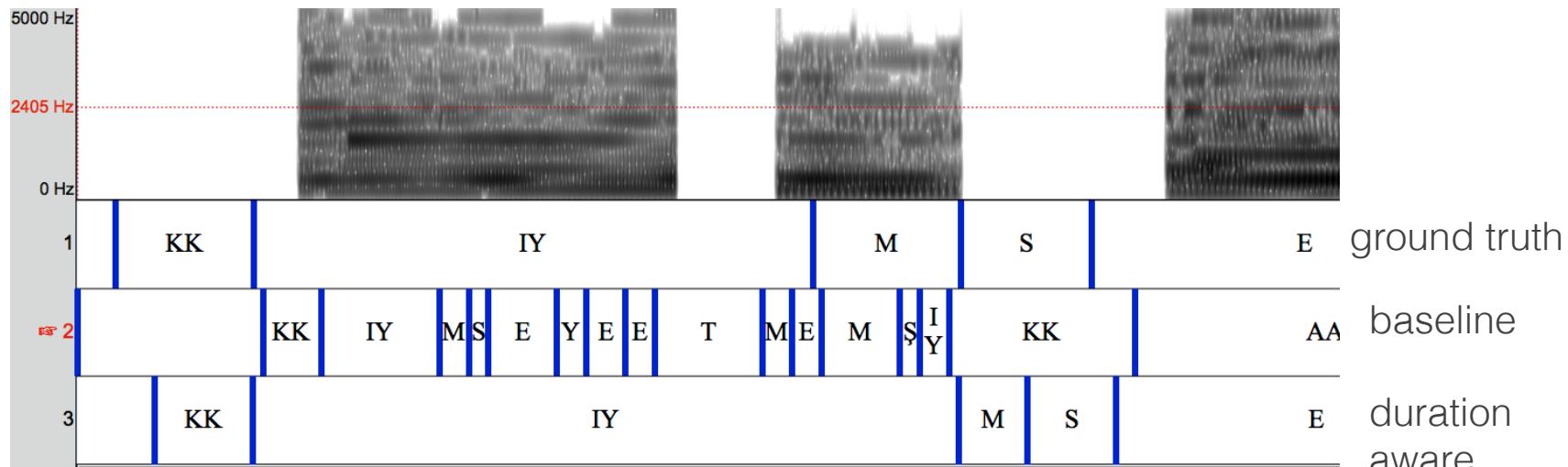
model	data	accuracy	error
baseline GMM	a cappella OTMM	70.2	1.14
duration aware GMM	a cappella OTMM	90.04	0.26
baseline GMM	multi-instrumental OTMM	67.46	1.26
duration aware GMM	multi-instrumental OTMM	77.74	0.63
(Mesaros, 2008)	multi-instrumental	-	1.4
(Fujihara, 2011)	multi-instrumental	85.2	-



Dzhambazov, Georgi and Xavier Serra (2015). Modeling of phoneme durations for alignment between polyphonic audio and lyrics. In Sound and Music Computing Conference 2015 (SMC 2015)

Experiments

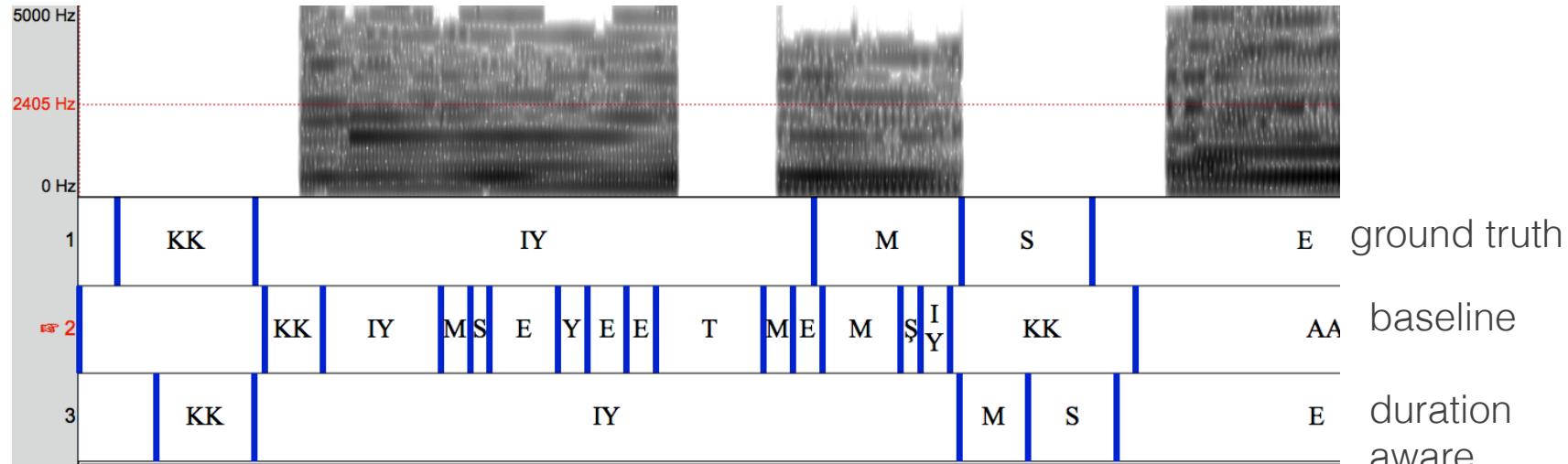
model	data	accuracy	error
baseline GMM	a cappella OTMM	70.2	1.14
duration aware GMM	a cappella OTMM	90.04	0.26
baseline GMM	multi-instrumental OTMM	67.46	1.26
duration aware GMM	multi-instrumental OTMM	77.74	0.63
(Mesaros, 2008)	multi-instrumental	-	1.4
(Fujihara, 2011)	multi-instrumental	85.2	-



Dzhambazov, Georgi and Xavier Serra (2015). Modeling of phoneme durations for alignment between polyphonic audio and lyrics. In Sound and Music Computing Conference 2015 (SMC 2015)

Experiments

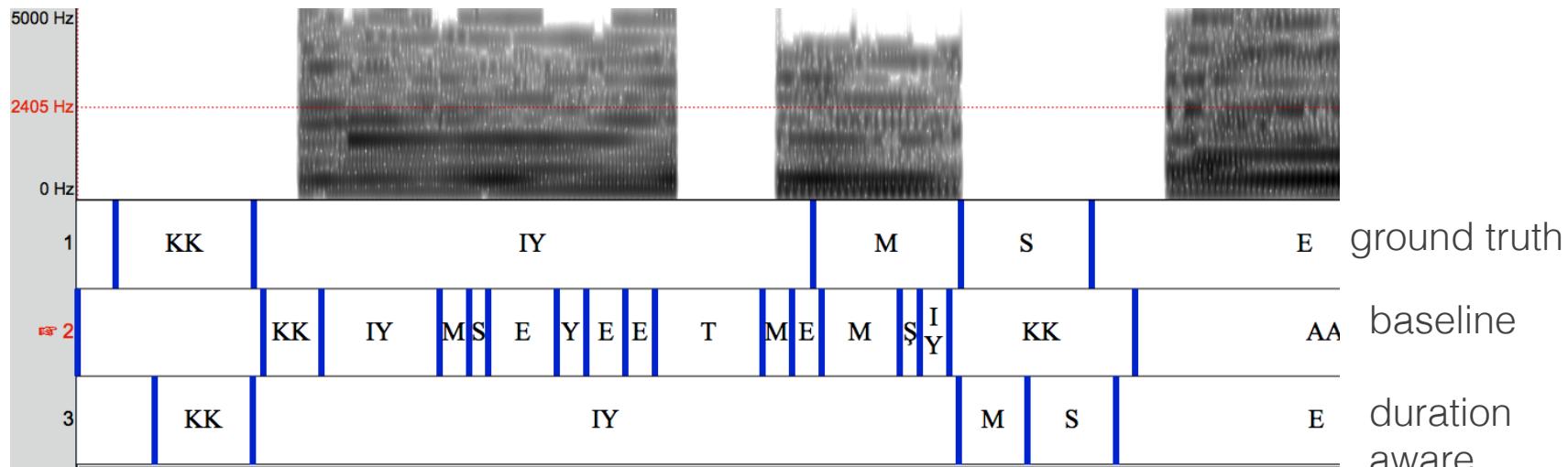
model	data	accuracy	error
baseline GMM	a cappella OTMM	70.2	1.14
duration aware GMM	a cappella OTMM	90.04	0.26
baseline GMM	multi-instrumental OTMM	67.46	1.26
duration aware GMM	multi-instrumental OTMM	77.74	0.63
(Mesaros, 2008)	multi-instrumental	-	1.4
(Fujihara, 2011)	multi-instrumental	85.2	-



Dzhambazov, Georgi and Xavier Serra (2015). Modeling of phoneme durations for alignment between polyphonic audio and lyrics. In Sound and Music Computing Conference 2015 (SMC 2015)

Experiments

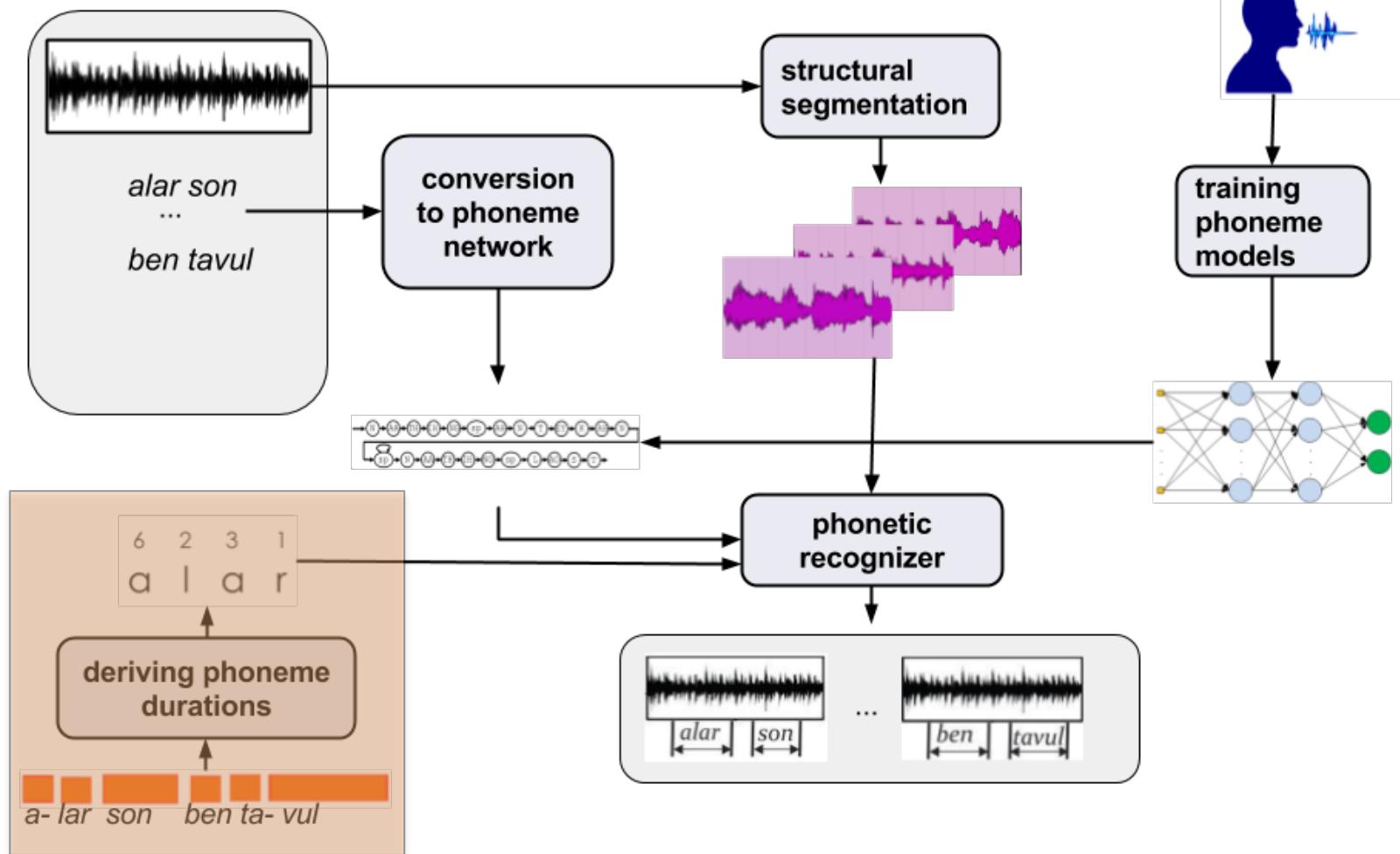
model	data	accuracy	error
baseline GMM	a cappella OTMM	70.2	1.14
duration aware GMM	a cappella OTMM	90.04	0.26
baseline GMM	multi-instrumental OTMM	67.46	1.26
duration aware GMM	multi-instrumental OTMM	77.74	0.63
(Mesaros, 2008)	multi-instrumental	-	1.4
(Fujihara, 2011)	multi-instrumental	85.2	-



Dzhambazov, Georgi and Xavier Serra (2015). Modeling of phoneme durations for alignment between polyphonic audio and lyrics. In Sound and Music Computing Conference 2015 (SMC 2015)

Chapter 4.5

Durations derived from music principles



Introduction to jingju

- Unique singing style
- Language based on Mandarin
- Different role types

dan



laosheng



Jingju-specific principles

- Lyrics: have poetic structure

Jingju-specific principles

- Lyrics: have poetic structure 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。

Jingju-specific principles

- Lyrics: have poetic structure 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。

▪ Usually 3 dou in a lyrics line
到如 今 只 落 得 罪 衣 罪 裙



Jingju-specific principles

- Lyrics: have poetic structure 玉堂春含悲泪忙往前进,
想起了当年事好不伤情!

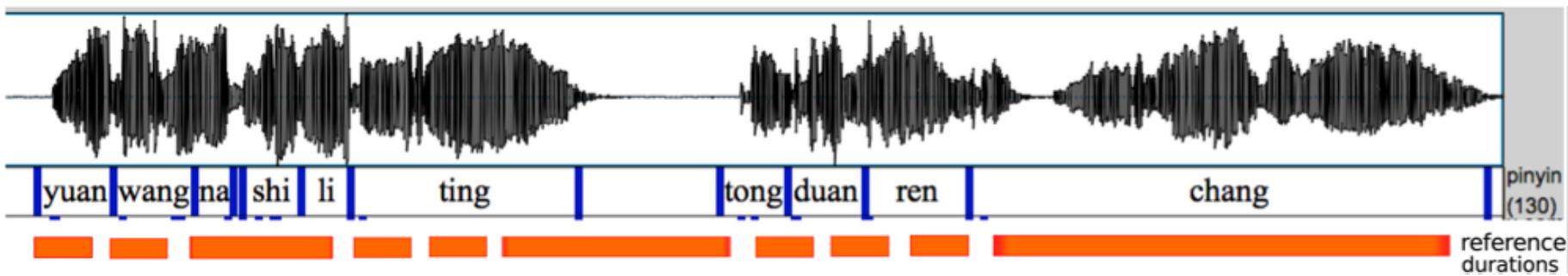
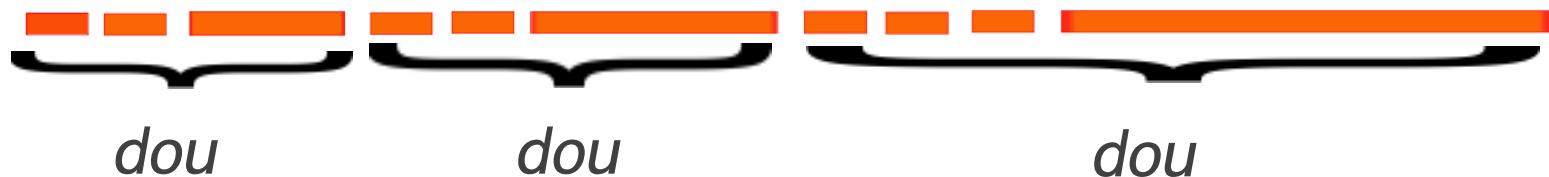
每日里在院中缠头似锦,
到如今只落得罪衣罪裙。

- Usually 3 dou in a lyrics line
到如今只落得罪衣罪裙

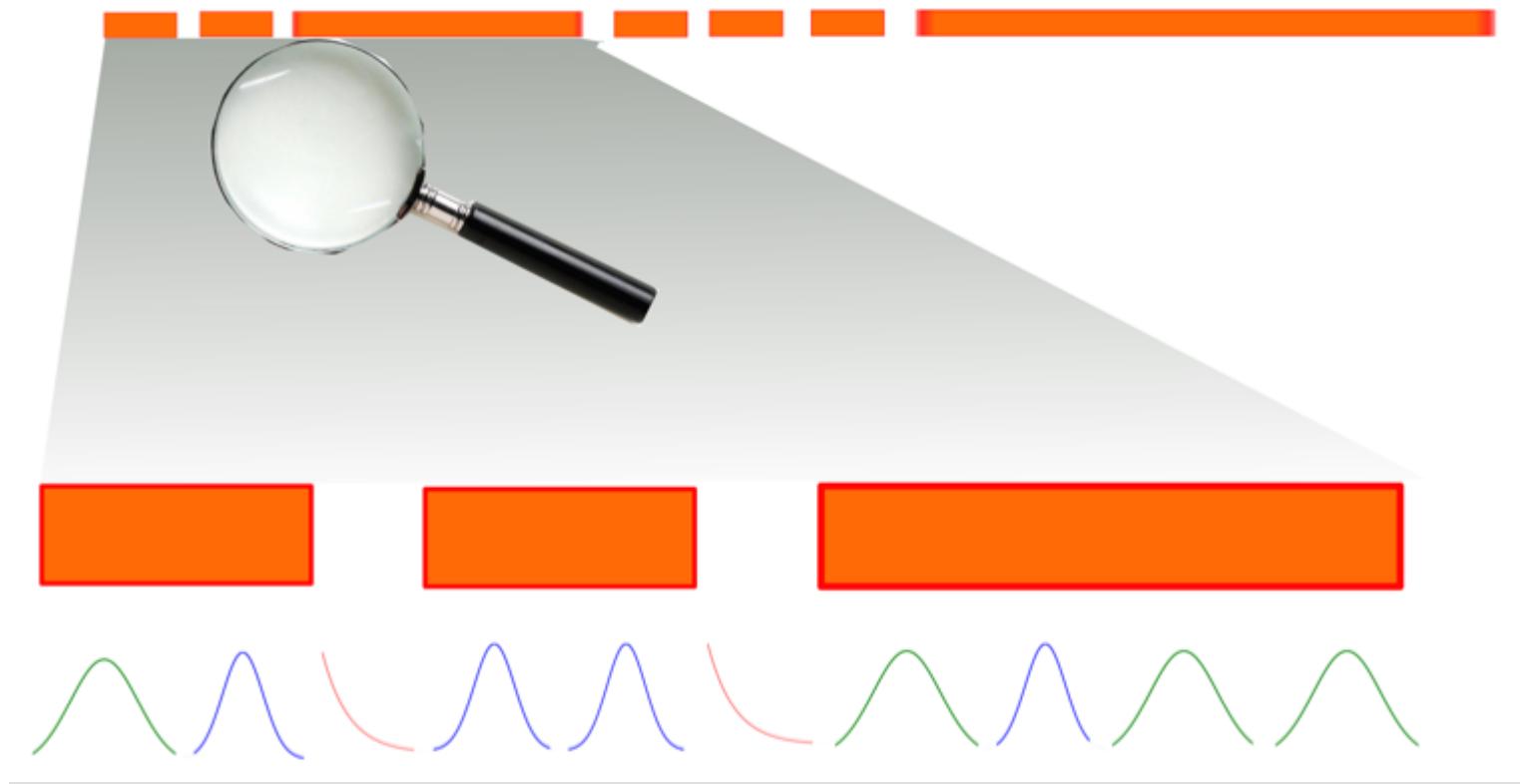


- Long last syllable

Deriving phoneme durations

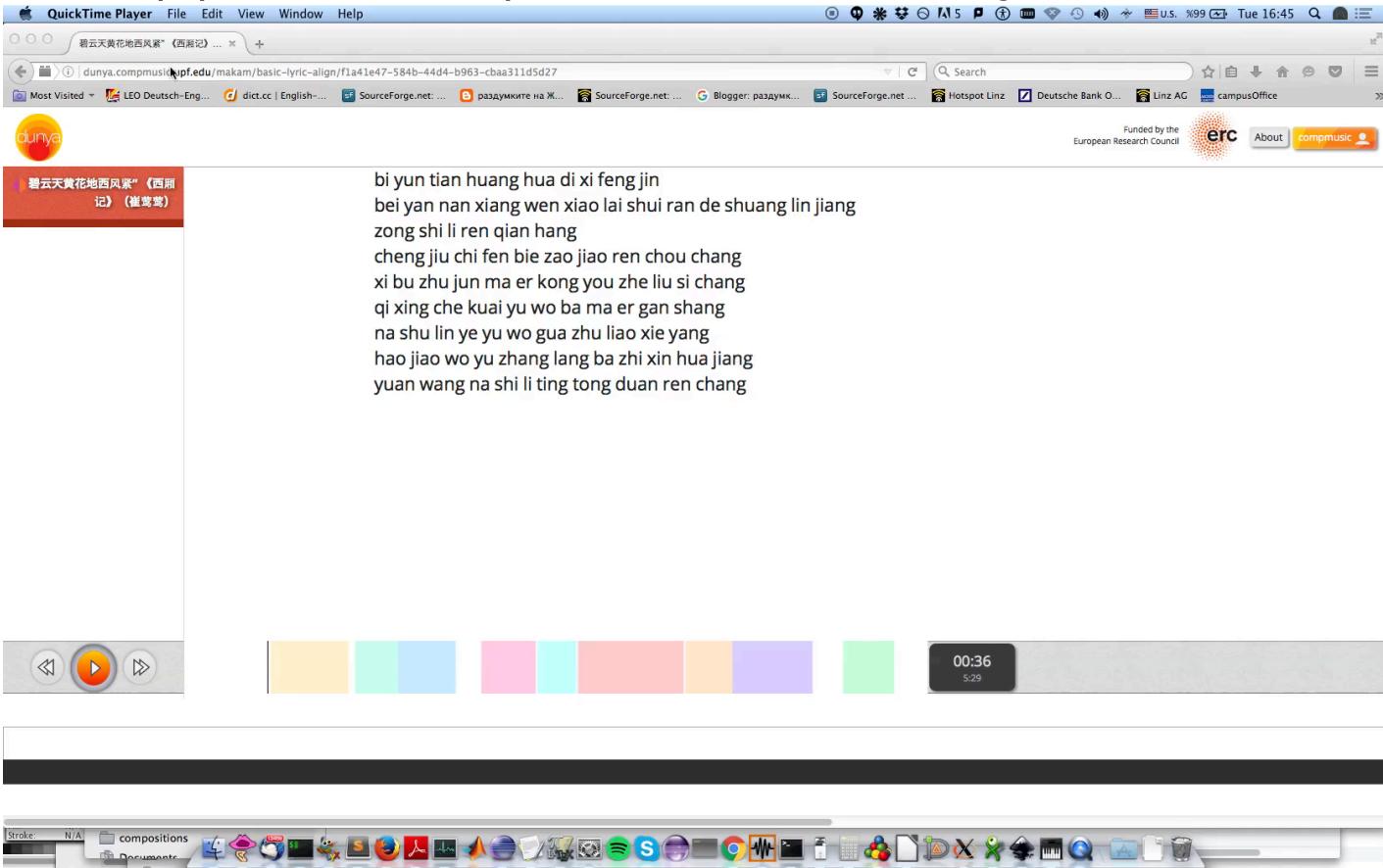


Deriving phoneme durations



A cappella lyrics jingju dataset

- 15 arias of dan role type (80 minutes)
- A cappella is separated manually



The screenshot shows a Mac OS X desktop with a QuickTime Player window open. The window displays a Jingju aria with lyrics in Chinese. The lyrics are:

```

bi yun tian huang hua di xi feng jin
bei yan nan xiang wen xiao lai shui ran de shuang lin jiang
zong shi li ren qian hang
cheng jiu chi fen bie zao jiao ren chou chang
xi bu zhu jun ma er kong you zhe liu si chang
qi xing che kuai yu wo ba ma er gan shang
na shu lin ye yu wo gua zhu liao xie yang
hao jiao wo yu zhang lang ba zhi xin hua jiang
yuan wang na shi li ting tong duan ren chang

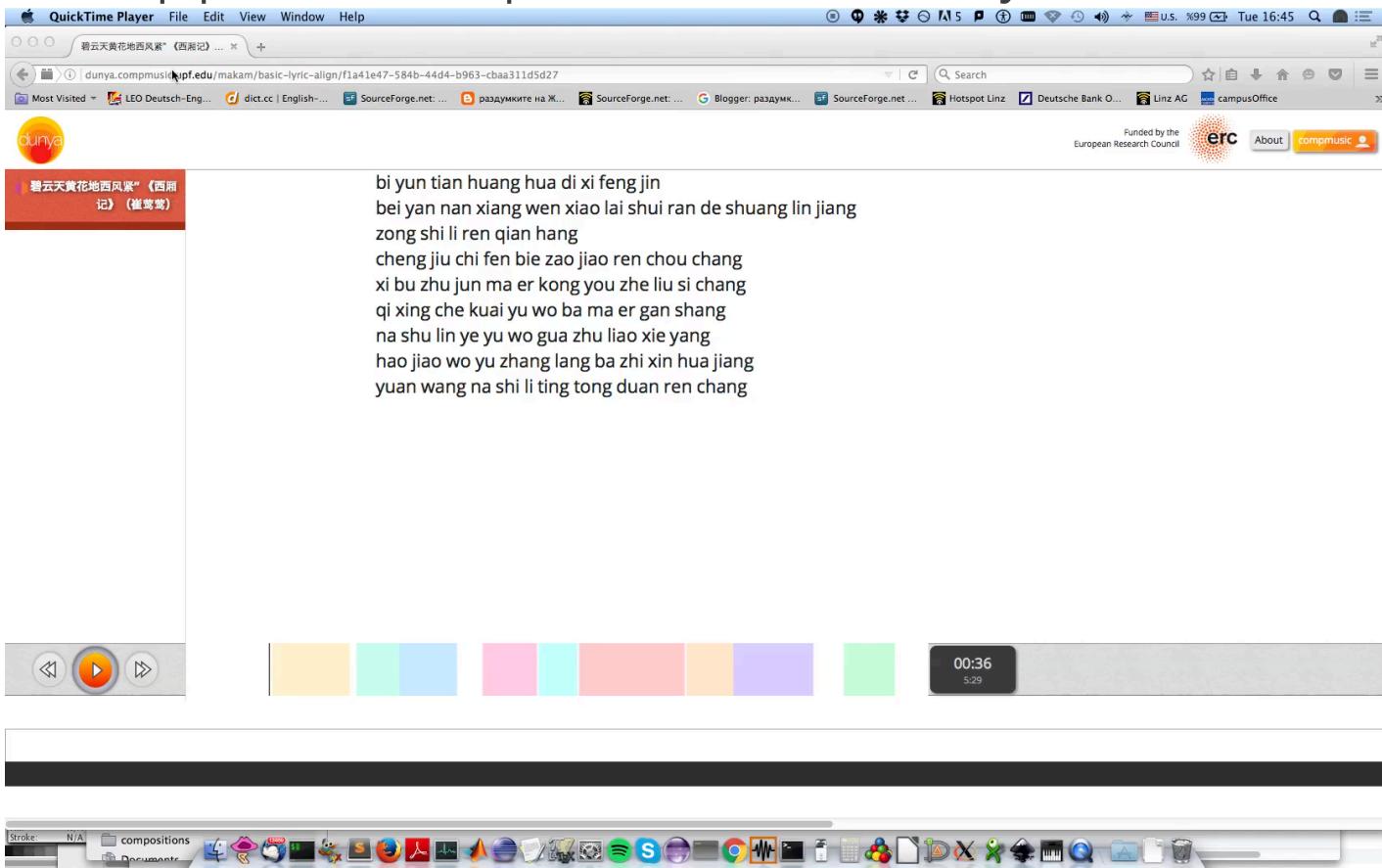
```

Below the lyrics is a color-coded timeline consisting of several colored squares (yellow, light blue, pink, cyan, red, orange, purple, green) followed by a black bar. A progress bar at the bottom indicates the current position at 00:36 of a 5:29 minute duration. The desktop menu bar at the top shows "QuickTime Player" and other system icons. The Dock at the bottom contains various application icons.



A cappella lyrics jingju dataset

- 15 arias of dan role type (80 minutes)
- A cappella is separated manually



The screenshot shows a Mac OS X desktop with a QuickTime Player window open. The window displays a Jingju aria with lyrics in Chinese. The lyrics are:

```

bi yun tian huang hua di xi feng jin
bei yan nan xiang wen xiao lai shui ran de shuang lin jiang
zong shi li ren qian hang
cheng jiu chi fen bie zao jiao ren chou chang
xi bu zhu jun ma er kong you zhe liu si chang
qi xing che kuai yu wo ba ma er gan shang
na shu lin ye yu wo gua zhu liao xie yang
hao jiao wo yu zhang lang ba zhi xin hua jiang
yuan wang na shi li ting tong duan ren chang

```

Below the lyrics is a color-coded timeline consisting of several colored squares (yellow, light blue, pink, cyan, red, orange, purple, green) followed by a black bar. A progress bar at the bottom indicates the current position at 00:36 of a 5:29 minute duration. The desktop menu bar at the top shows "QuickTime Player" and other system icons. The Dock at the bottom contains various application icons.



Experiments

- Experiment 0: duration aware with oracle phonemes

	oracle	baseline	duration aware
average	98.5	56.6	89.9
median per sentence	98.3	75.2	92.3

Dzhambazov, Georgi, Yile Yang, Rafael Caro Repetto, and Xavier Serra (2016). Automatic alignment of long syllables in a cappella Beijing opera. In 6th International Workshop on Folk Music Analysis (FMA 2016)

Experiments

- Experiment 0: duration aware with oracle phonemes

	oracle	baseline	duration aware
average	98.5	56.6	89.9
median per sentence	98.3	75.2	92.3

Dzhambazov, Georgi, Yile Yang, Rafael Caro Repetto, and Xavier Serra (2016). Automatic alignment of long syllables in a cappella Beijing opera. In 6th International Workshop on Folk Music Analysis (FMA 2016)

Experiments

- Experiment 0: duration aware with oracle phonemes

	oracle	baseline	duration aware
average	98.5	56.6	89.9
median per sentence	98.3	75.2	92.3

Dzhambazov, Georgi, Yile Yang, Rafael Caro Repetto, and Xavier Serra (2016). Automatic alignment of long syllables in a cappella Beijing opera. In 6th International Workshop on Folk Music Analysis (FMA 2016)

Experiments

- Experiment 0: duration aware with oracle phonemes
- Allow large standard deviation = 2 sec

	oracle	baseline	duration aware
average	98.5	56.6	89.9
median per sentence	98.3	75.2	92.3

Dzhambazov, Georgi, Yile Yang, Rafael Caro Repetto, and Xavier Serra (2016). Automatic alignment of long syllables in a cappella Beijing opera. In 6th International Workshop on Folk Music Analysis (FMA 2016)

Experiments

- Experiment 0: duration aware with oracle phonemes
- Allow large standard deviation = 2 sec

	oracle	baseline	duration aware
average	98.5	56.6	89.9
median per sentence	98.3	75.2	92.3

Dzhambazov, Georgi, Yile Yang, Rafael Caro Repetto, and Xavier Serra (2016). Automatic alignment of long syllables in a cappella Beijing opera. In 6th International Workshop on Folk Music Analysis (FMA 2016)

Experiments

- Experiment 0: duration aware with oracle phonemes
- Allow large standard deviation = 2 sec

	oracle	baseline	duration aware
average	98.5	56.6	89.9
median per sentence	98.3	75.2	92.3

Dzhambazov, Georgi, Yile Yang, Rafael Caro Repetto, and Xavier Serra (2016). Automatic alignment of long syllables in a cappella Beijing opera. In 6th International Workshop on Folk Music Analysis (FMA 2016)

Outline

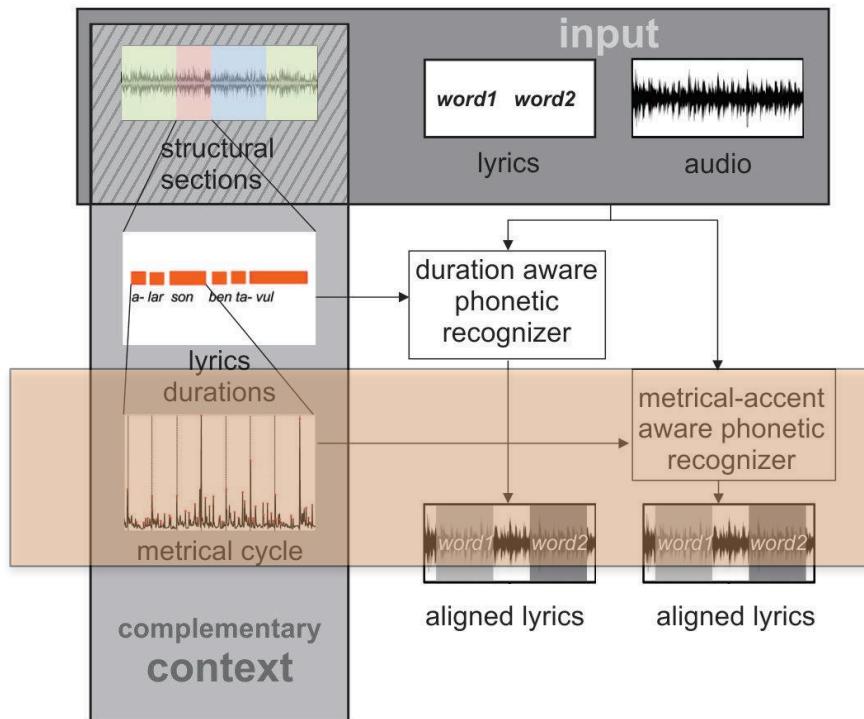
- Setting the stage
 - Introduction (Chapter 1)
 - Background (Chapter 2)
 - Baseline lyrics-to-audio alignment (Chapter 3)
- Main
 - Lyrics-duration aware lyrics-to-audio alignment (Chapter 4)
 - Metrical-accent aware lyrics-to-audio alignment (Chapter 5)
- Conclusion (Chapter 6) and Applications (Appendix)

Outline

- Setting the stage
 - Introduction (Chapter 1)
 - Background (Chapter 2)
 - Baseline lyrics-to-audio alignment (Chapter 3)
- Main
 - Lyrics-duration aware lyrics-to-audio alignment (Chapter 4)
 - Metrical-accent aware lyrics-to-audio alignment (Chapter 5)
- Conclusion (Chapter 6) and Applications (Appendix)

Chapter 5:

Lyrics-to-audio Alignment with Fine-level Complementary Context



Context: Accents within a metrical cycle

Chapter 5

Lyrics-to-audio Alignment with Fine-level Complementary Context

5.1 Introduction

In this chapter, we propose how to improve the baseline lyrics-to-audio alignment method by considering facets of fine-level context, complementary to lyrics. We focus on one particular fine-level facet – the accents in the metrical cycle (i.e., metrical accents). In some voices, transitions between consecutive lyrics are aligned with the metrical accents to a certain degree. However, we found that it is not obvious how to conceptualize the direct relation of metrical accents to syllable transitions. Instead, we investigate the relation of metrical accents to the positions of onsets (attacks) of sung notes in the vocal melody. In this way, the influence of metrical events on syllable transitions is represented implicitly through its influence on note onsets, which are in turn influenced by metrical events. In this sense, metrical accents can be considered a facet of complementary context of lyrics.

With this motivation, we propose in the first part of the chapter a vocal onset detector that considers the simultaneously occurring accents in a metrical cycle. Vocal onset detection can be seen as a subtask of singing voice transcription. That is why we propose how to extend a state-of-the-art probabilistic model for singing voice transcription, in which a priori probability of a note at a specific position in the metrical cycle interacts with the probability of observing a vocal note onset. Designing in a compact manner meter-aware

Chapter 5 outline

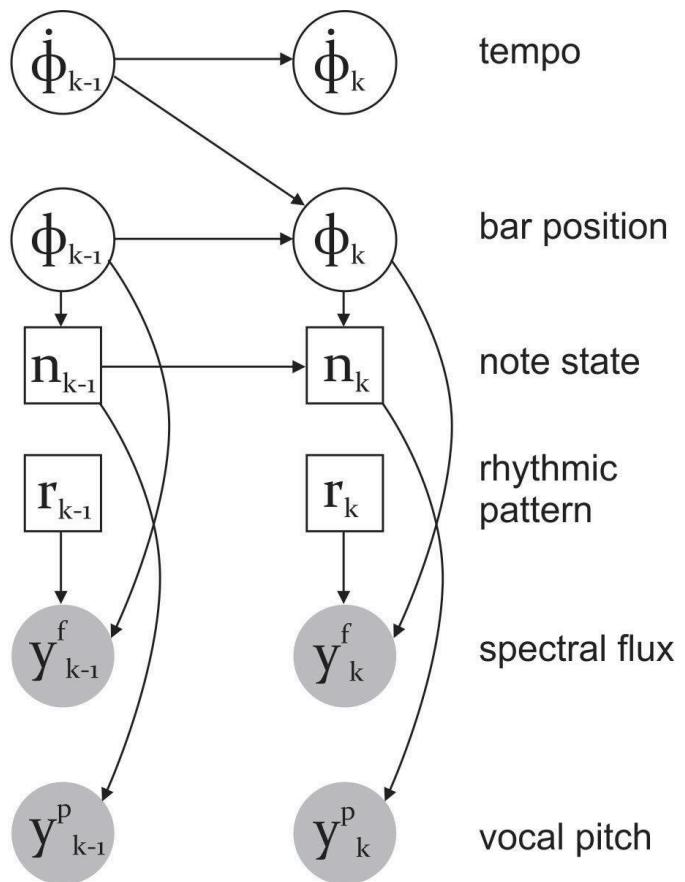
- Metrical-accent aware vocal note onset detection (5.3)
- Vocal-onset aware lyrics-to-audio alignment (5.4)

Chapter 5.3

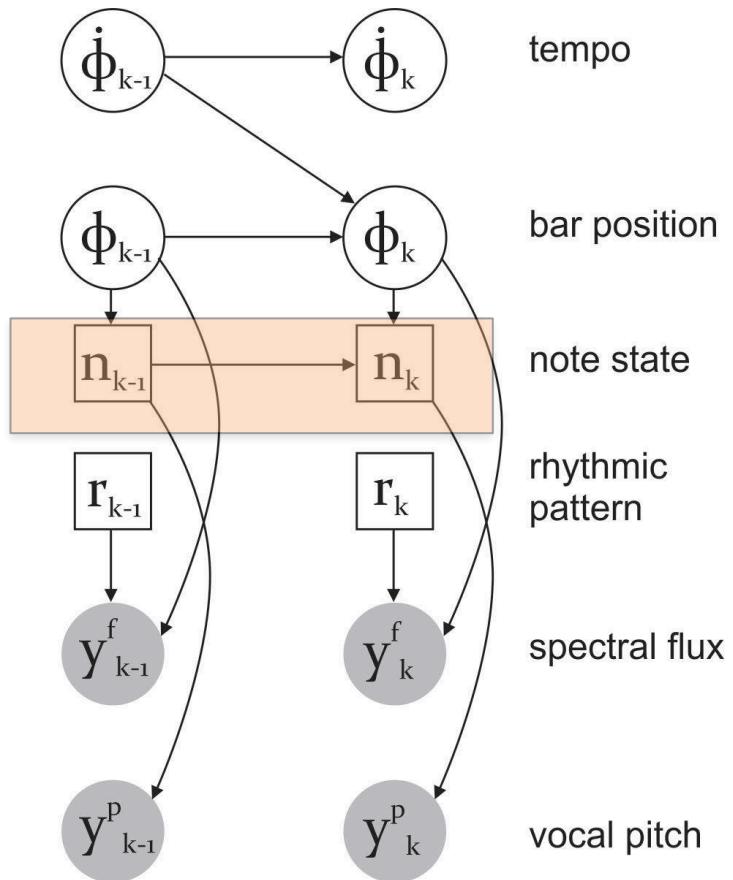
Metrical-accent aware vocal onset detection

- Vocal note events co-occur with accents in a metrical cycle (Holzapfel 2014)
 - OTMM has well pronounced metrical accents

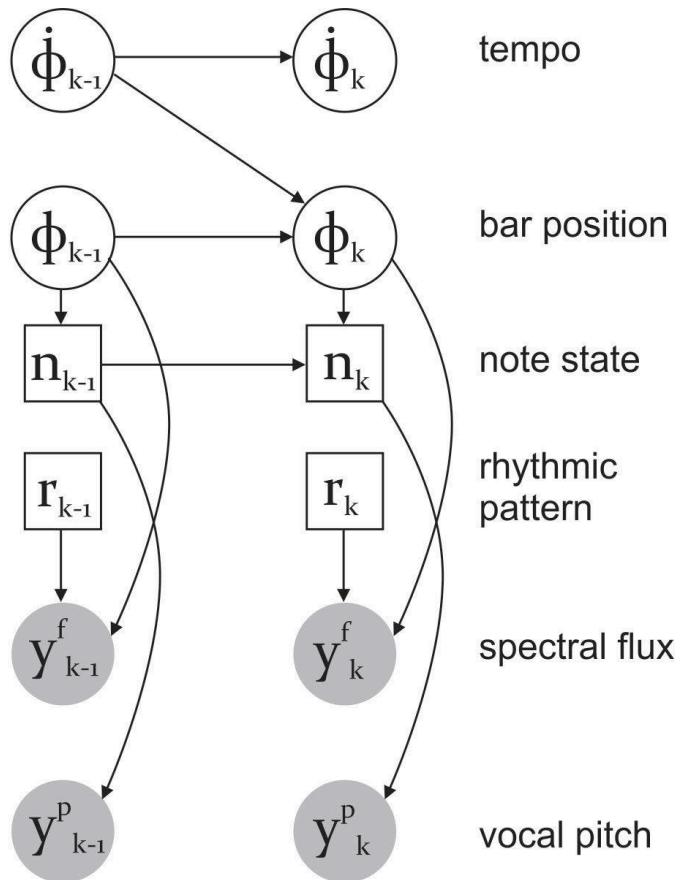
Model components



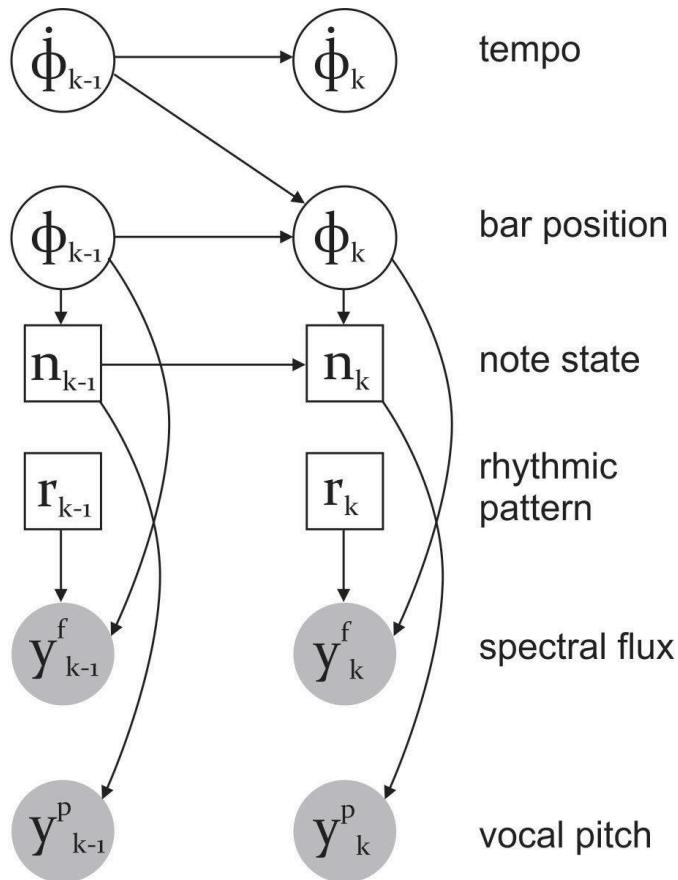
Model components



Model components

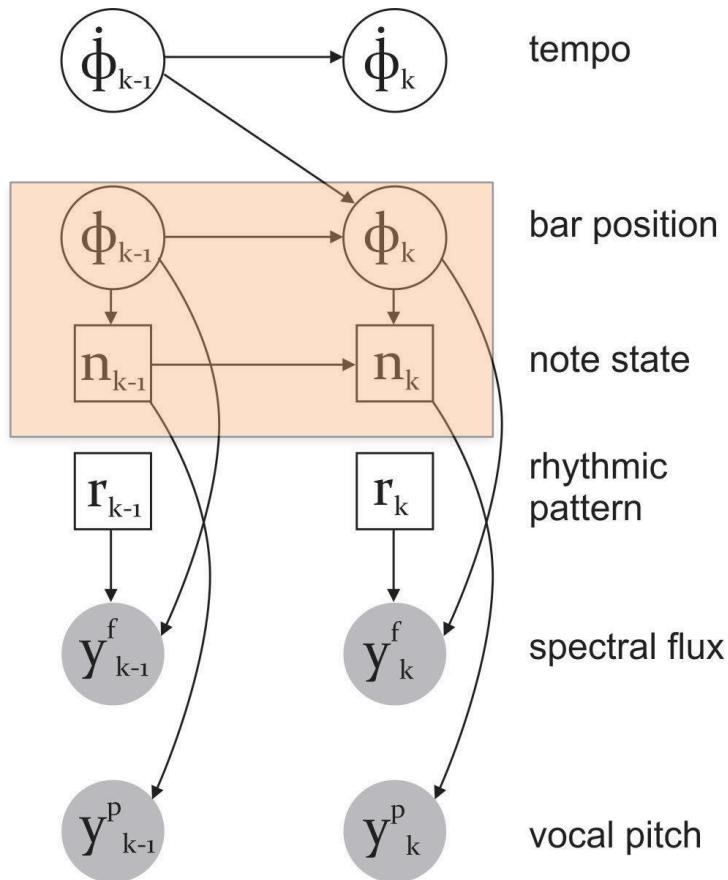


Model components



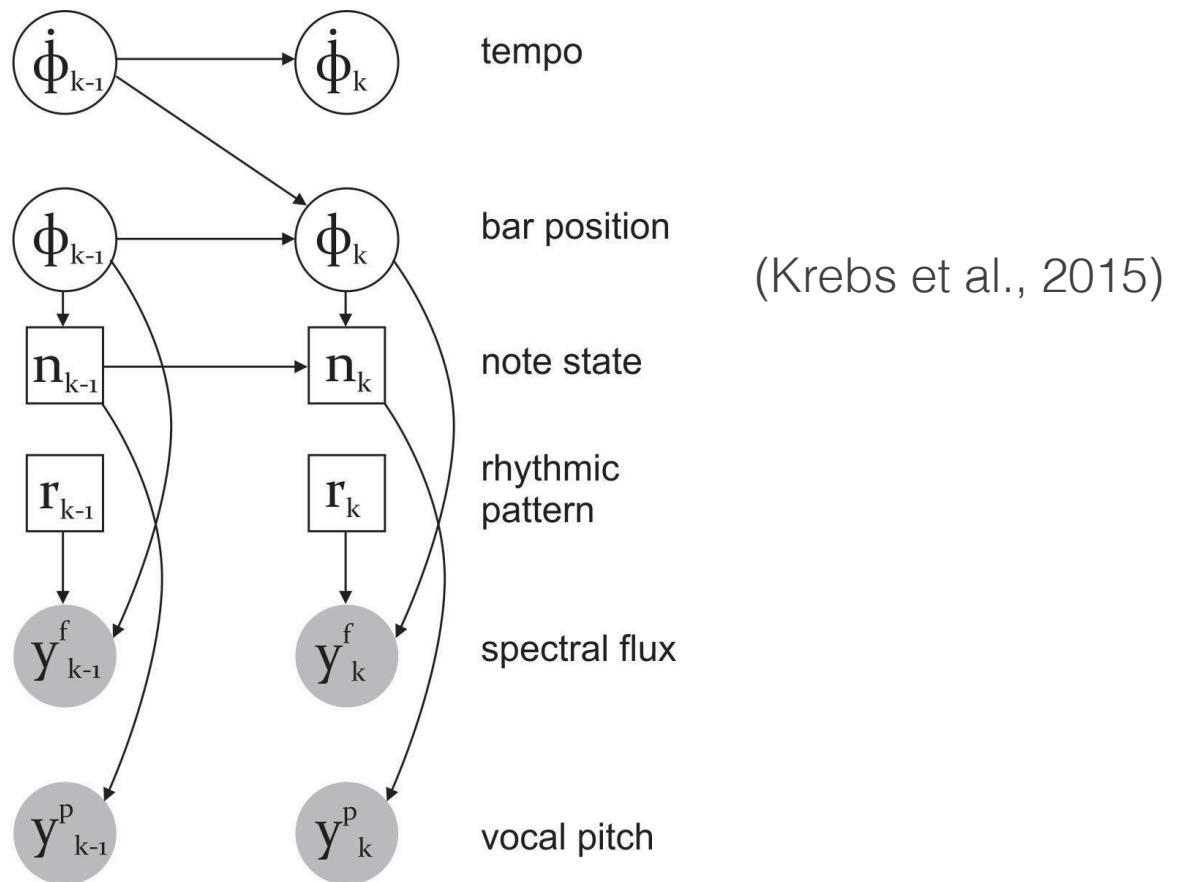
Goal: track simultaneously bar position and vocal note state

Model components

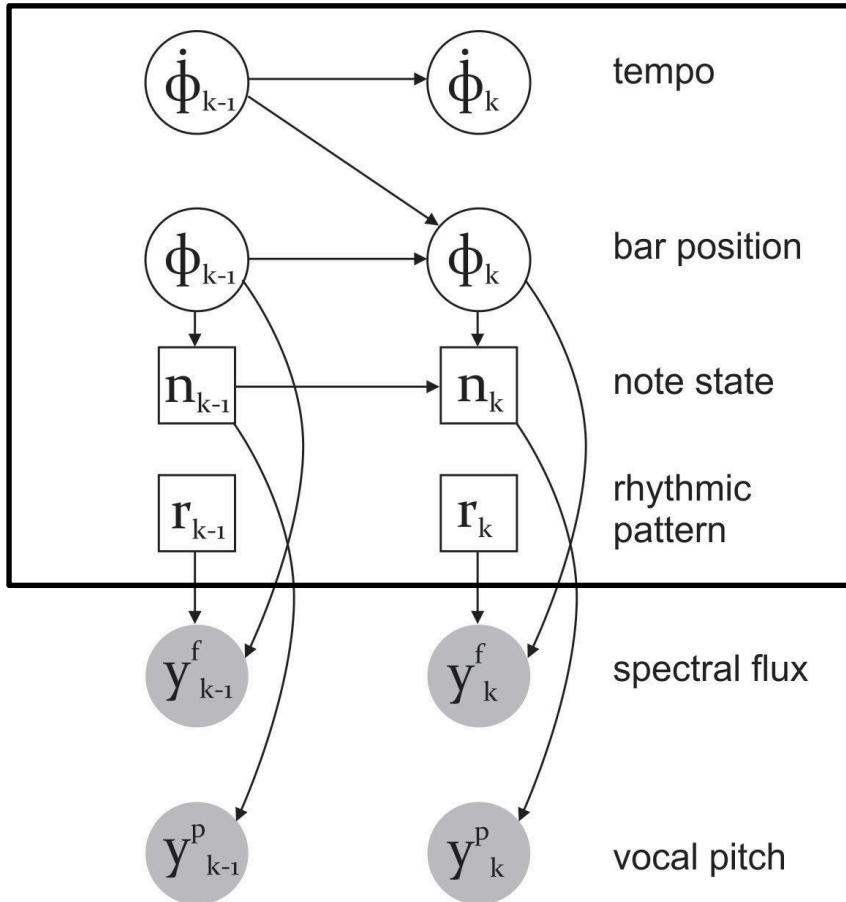


Goal: track simultaneously bar position and vocal note state

Transition model

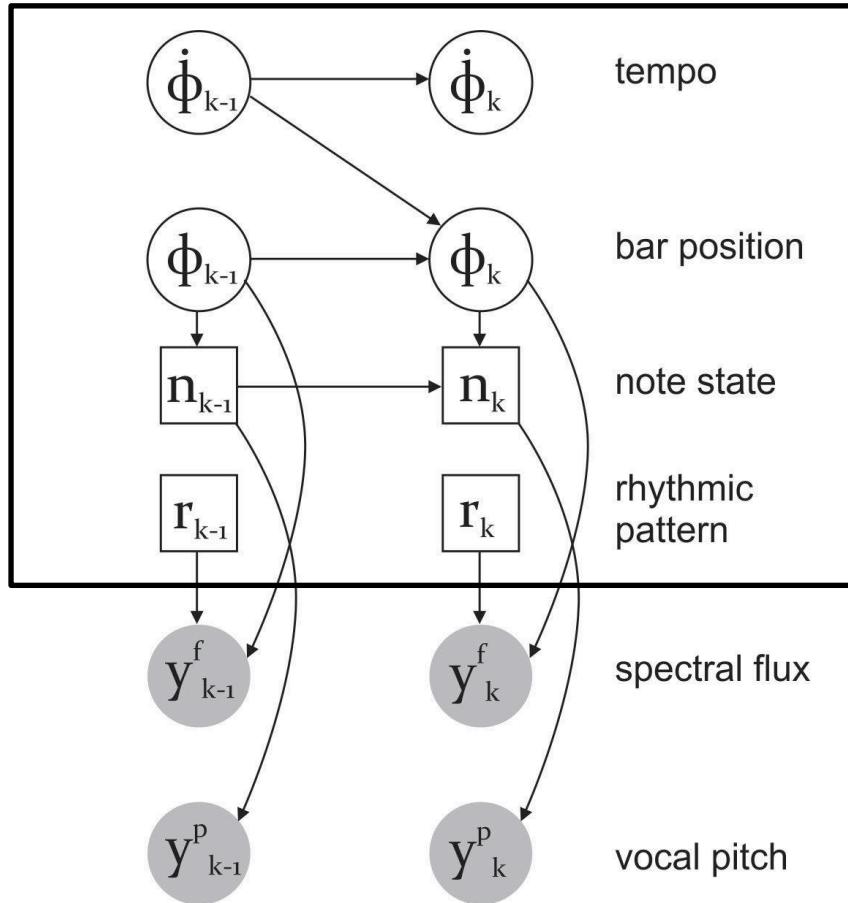


Transition model

 $x_{k-1} \quad x_k$


(Krebs et al., 2015)

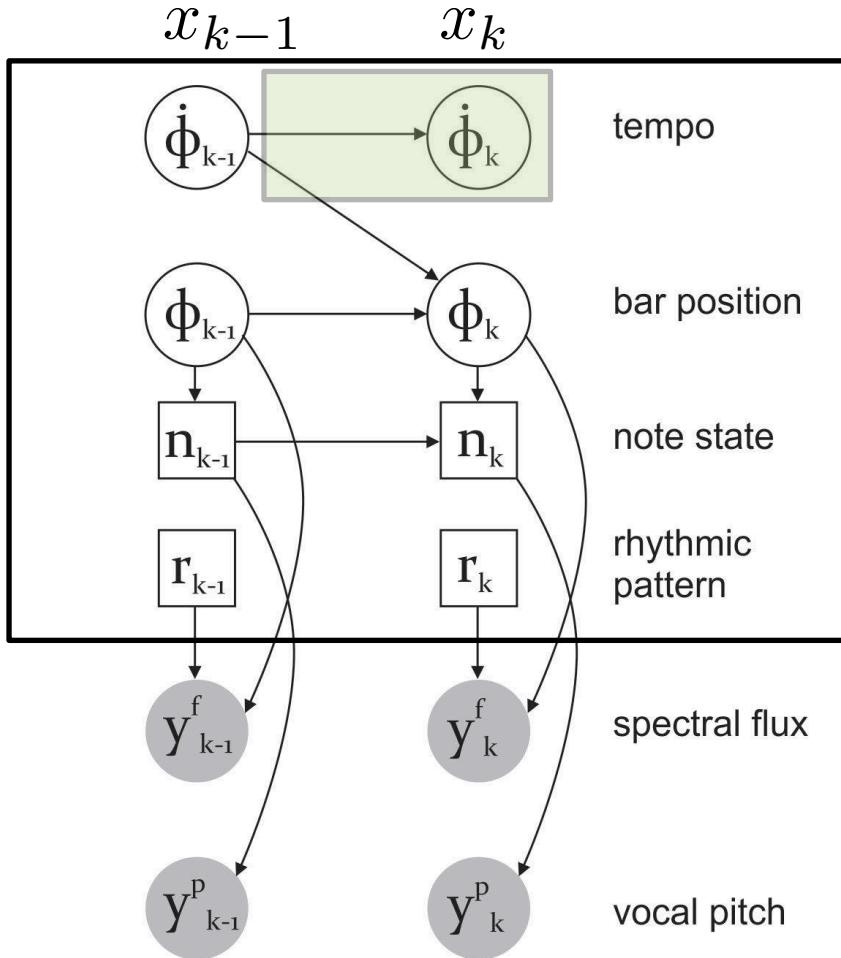
Transition model

 $x_{k-1} \quad x_k$


(Krebs et al., 2015)

$$P(x_k|x_{k-1}) = P(\dot{\phi}_k|\dot{\phi}_{k-1}) \times P(\phi_k|\phi_{k-1}, \dot{\phi}_{k-1}) \times P(n_k|n_{k-1}, \phi_k)$$

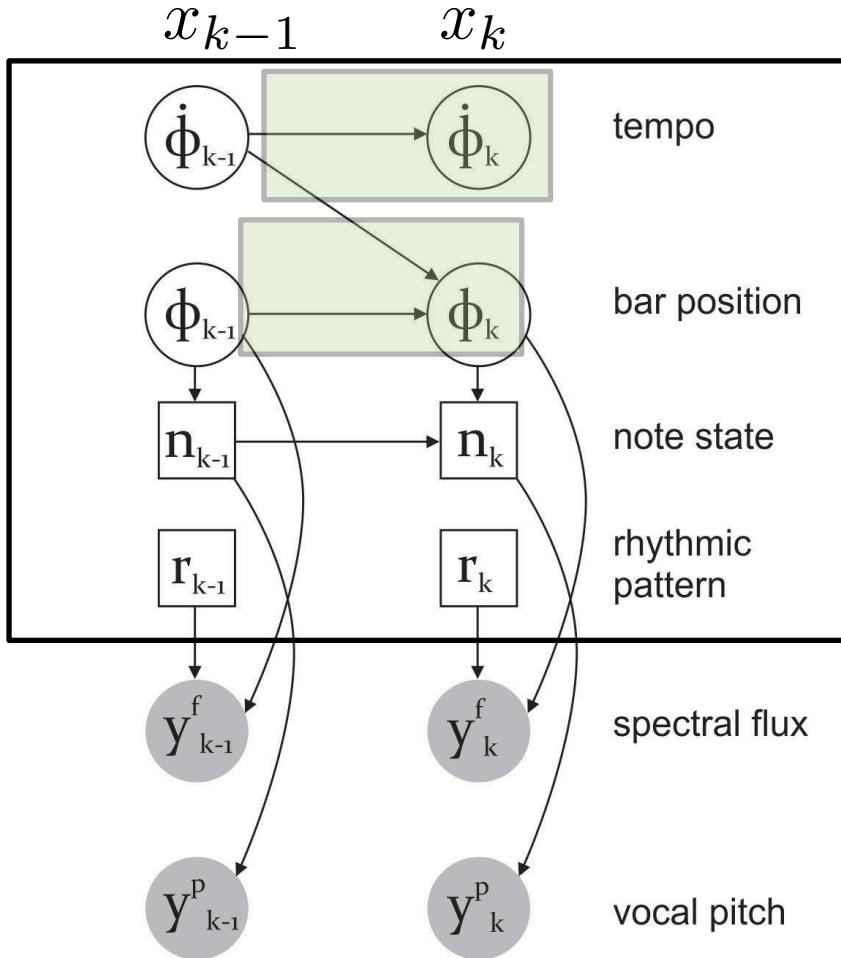
Transition model



(Krebs et al., 2015)

$$P(x_k|x_{k-1}) = P(\dot{\phi}_k|\dot{\phi}_{k-1}) \times P(\phi_k|\phi_{k-1}, \dot{\phi}_{k-1}) \times P(n_k|n_{k-1}, \phi_k)$$

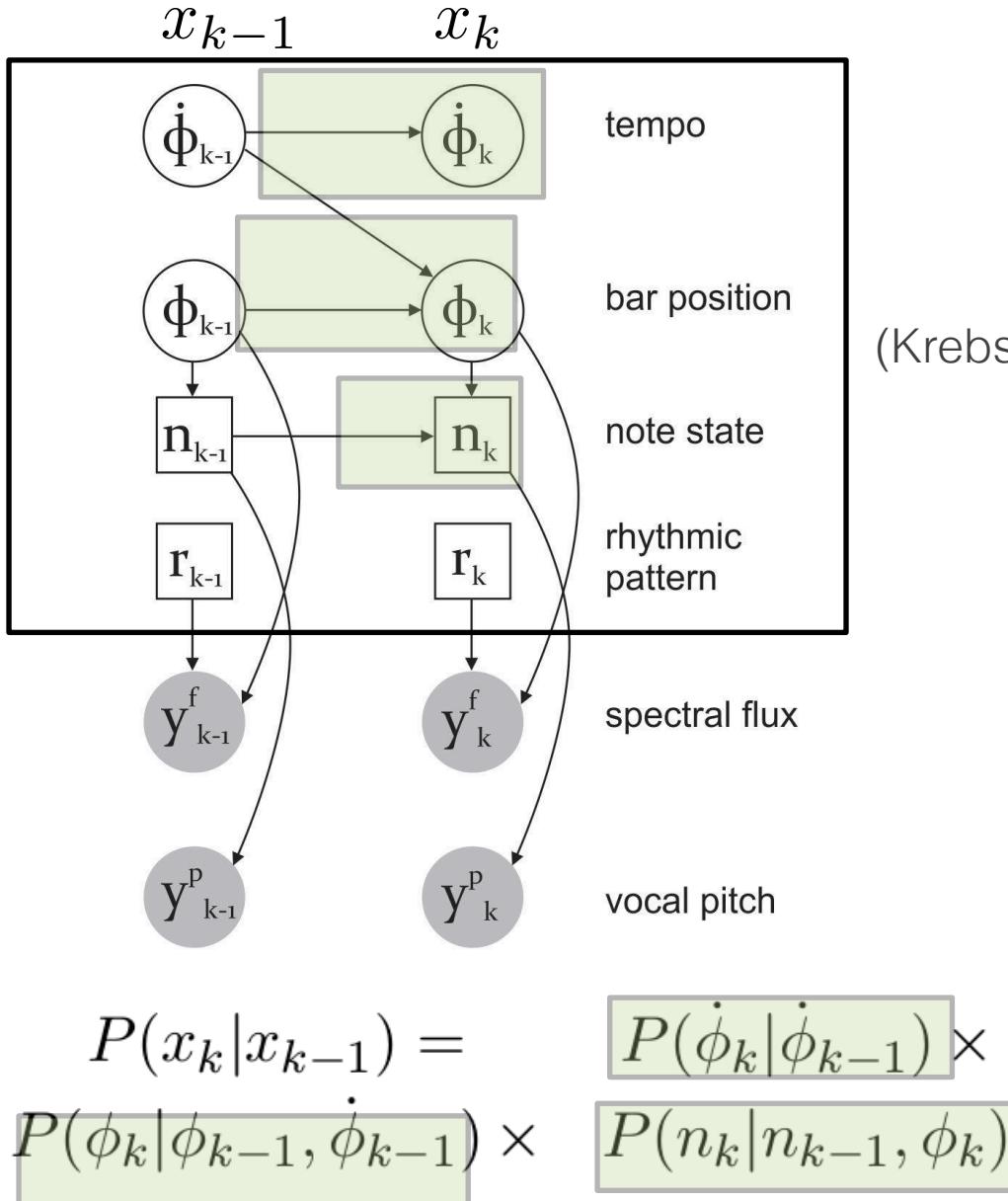
Transition model



(Krebs et al., 2015)

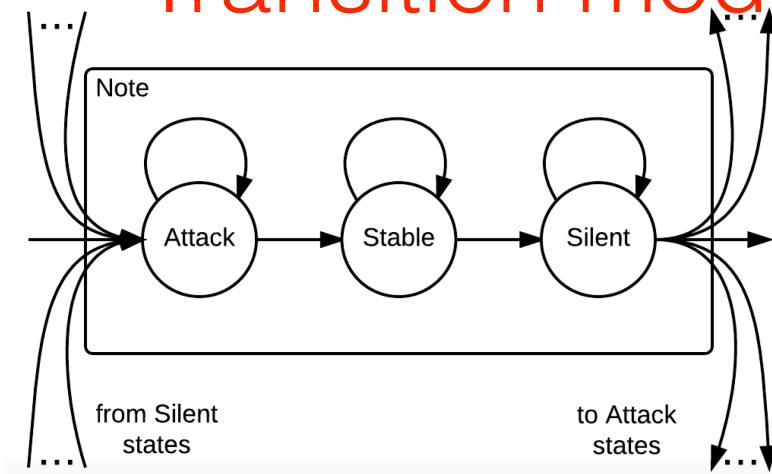
$$P(x_k|x_{k-1}) = P(\dot{\phi}_k|\dot{\phi}_{k-1}, \dot{\phi}_{k-1}) \times P(n_k|n_{k-1}, \phi_k)$$

Transition model

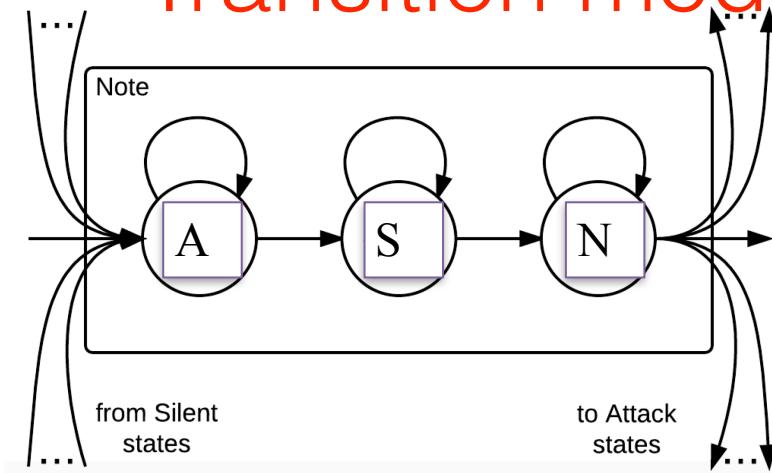


$$P(x_k|x_{k-1}) = P(\dot{\phi}_k|\dot{\phi}_{k-1}) \times P(\phi_k|\phi_{k-1}, \dot{\phi}_{k-1}) \times P(n_k|n_{k-1}, \phi_k)$$

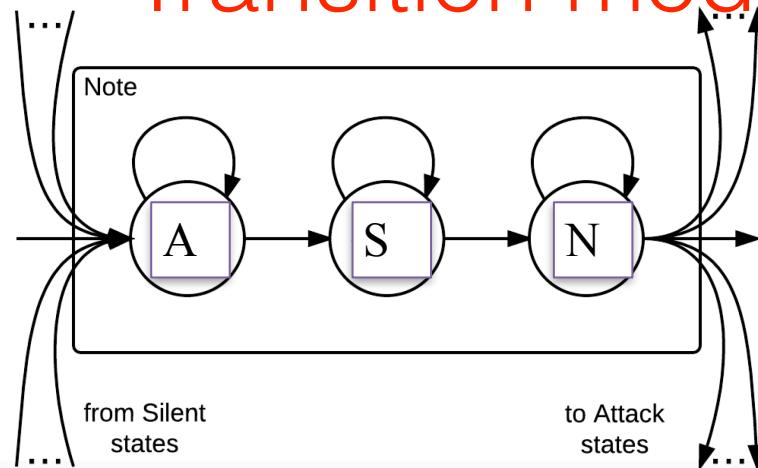
Transition model



Transition model

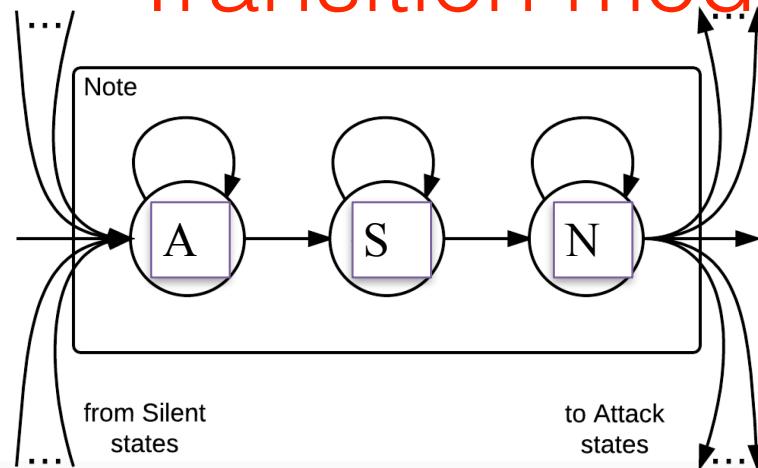


Transition model



$$p(n_k | n_{k-1}) = \begin{cases} P_{N_i A_j}, & n_{k-1} = N_i \quad n_k = A_j \\ 1 - \sum_i P_{N_i A_j}, & n_{k-1} = n_k = N_i \\ 1 - c_A, & n_{k-1} = A_i \quad n_k = S_j \\ c_A, & n_{k-1} = n_k = A_i \\ 1 - c_S & n_{k-1} = S_i \quad n_k = N_j \\ c_S, & n_{k-1} = n_k = S_i \\ 0 & \text{else} \end{cases} \quad (\text{Mauch et al., 2015})$$

Transition model



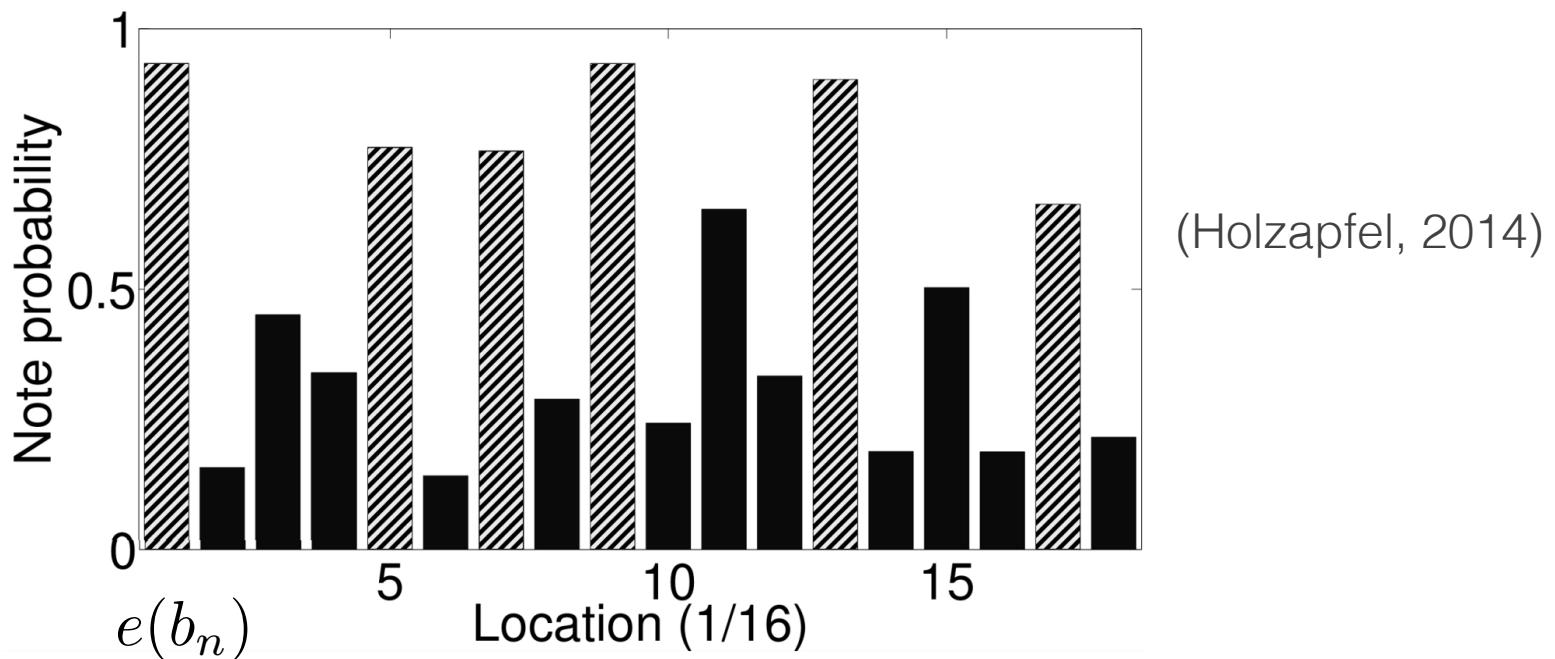
$$p(n_k | n_{k-1}) = \begin{cases} P_{N_i A_j}, & n_{k-1} = N_i \quad n_k = A_j \\ 1 - \sum_i P_{N_i A_j}, & n_{k-1} = n_k = N_i \\ 1 - c_A, & n_{k-1} = A_i \quad n_k = S_j \\ c_A, & n_{k-1} = n_k = A_i \\ 1 - c_S, & n_{k-1} = S_i \quad n_k = N_j \\ c_S, & n_{k-1} = n_k = S_i \\ 0 & \text{else} \end{cases} \quad (\text{Mauch et al., 2015})$$

$$p(n_k | n_{k-1}, \phi_k) = \begin{cases} P_{N_i A_j} \Theta(\phi_k), & n_{k-1} = N_i \quad n_k = A_j \\ 1 - \Theta(\phi_k) \sum_i P_{N_i A_j}, & n_{k-1} = n_k = N_i \\ \dots \end{cases}$$

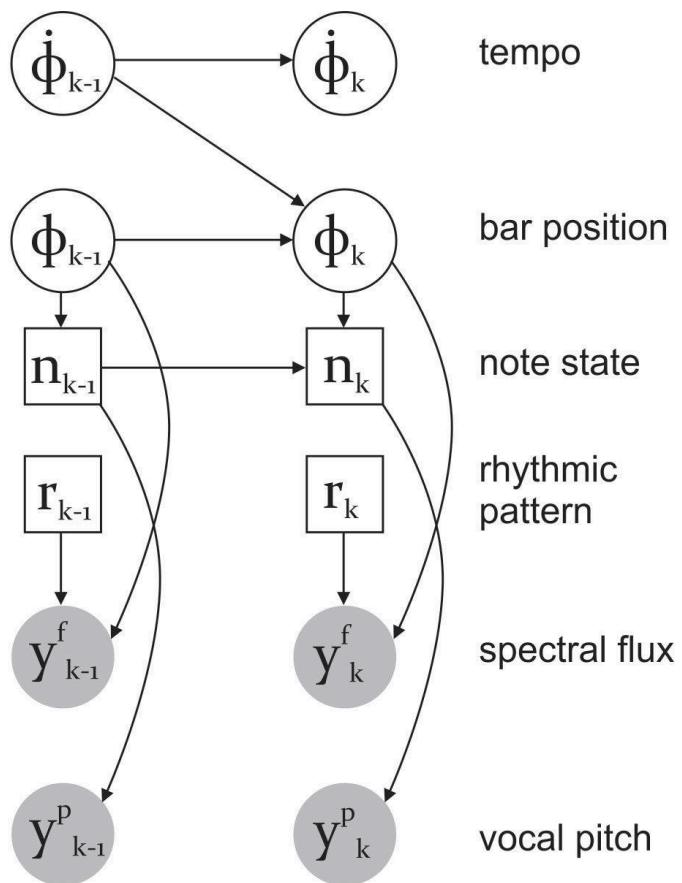
Transition model

$$\Theta(\phi_k) = \begin{cases} q \cdot e(b_n), & \phi_k := b_n \\ 1 & \text{else} \end{cases}$$

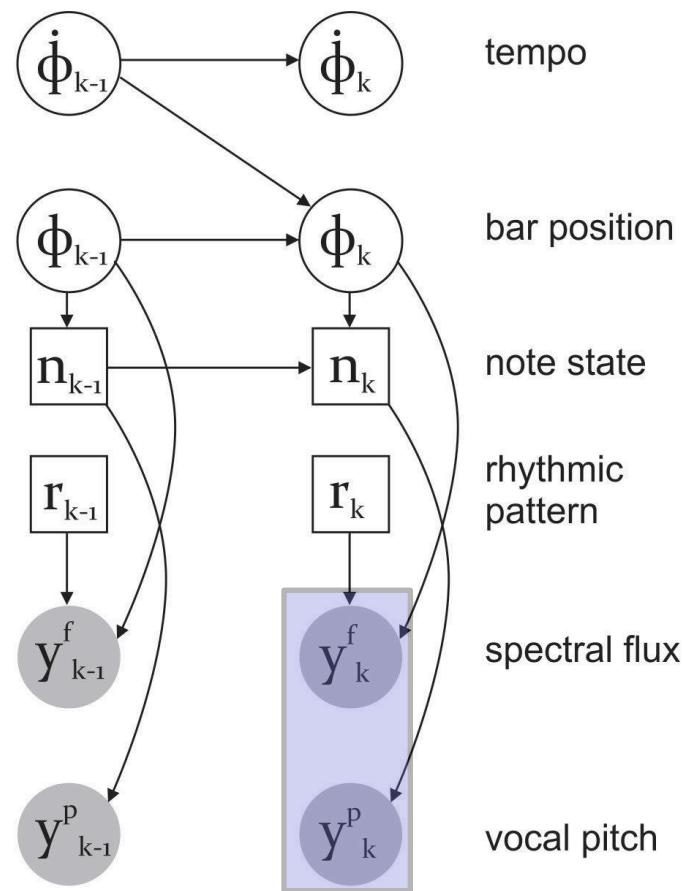
$e(b_n)$: note probability within a metrical cycle



Observation model



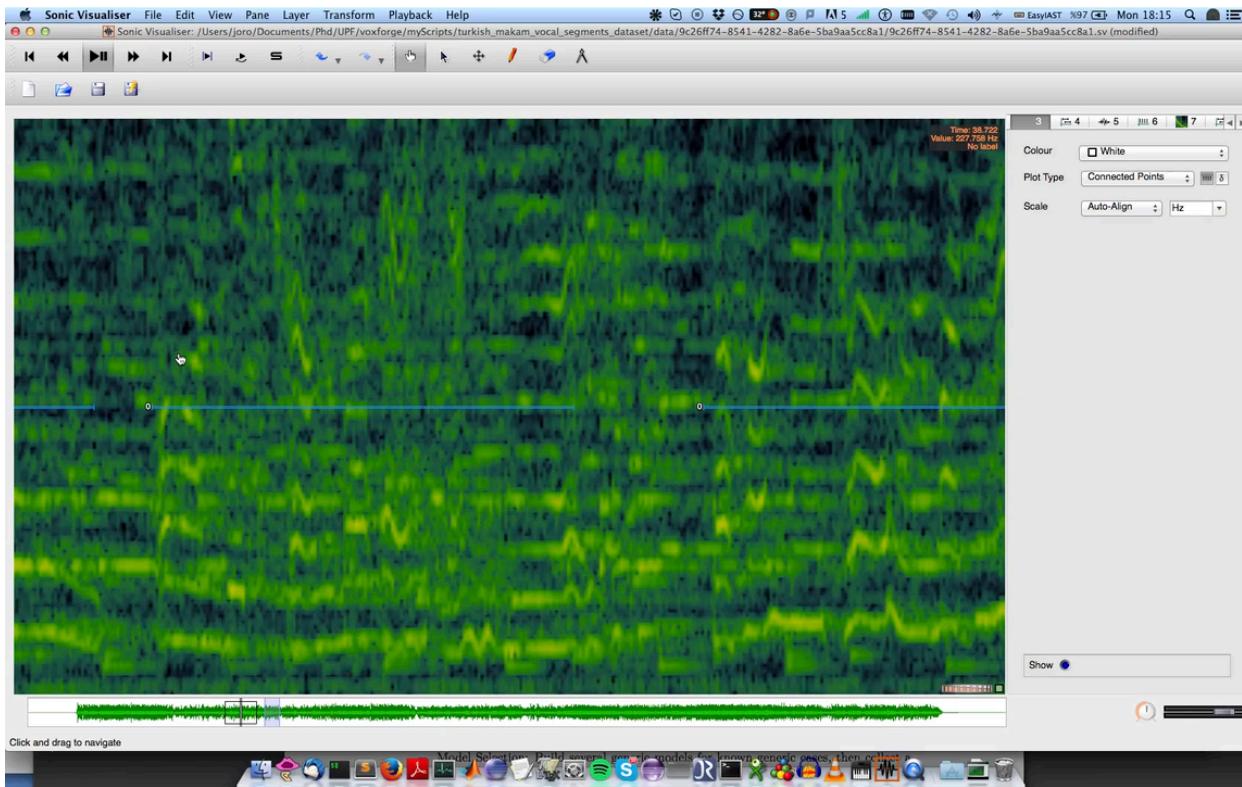
Observation model



$$P(y_k^f, y_k^p | x_k) = P(y_k^f | \phi_k, r_k) P(y_k^p | n_k)$$

Multi-instrumental vocal onsets OTMM dataset

- Annotations on beats + vocal onsets
- 5 1-minute recordings from two meter types

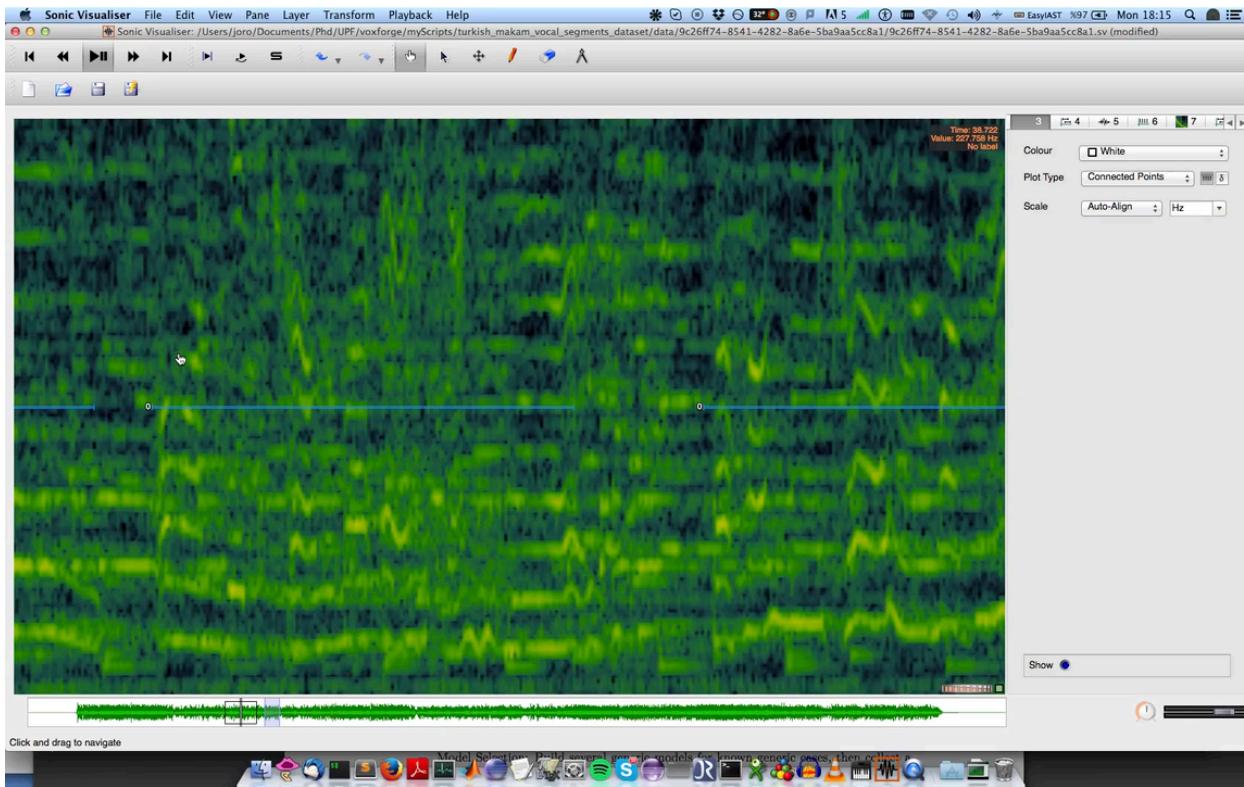


<http://compmusic.upf.edu/otmm-vocal-onsets-dataset>

Sec 3.2.3

Multi-instrumental vocal onsets OTMM dataset

- Annotations on beats + vocal onsets
- 5 1-minute recordings from two meter types



<http://compmusic.upf.edu/otmm-vocal-onsets-dataset>

Sec 3.2.3

Experiments

- Experiment 1 (Ex-1): manually annotated beats

	meter	beat	Fmeas	P	R	Fmeas
düyek	Mauch	-	33.1	31.6	31.6	31.6
	Ex-1	-	40.4	39.5	39.0	39.0
	Ex-2	86.4	37.8	36.1	36.1	36.1
aksak	Mauch	-	42.1	36.9	37.9	37.9
	Ex-1	-	48.4	39.1	43.0	43.0
	Ex-2	72.9	45.0	39.0	40.3	40.3

Dzhambazov, Georgi, Andre Holzapfel, Ajay Srinivasamurthy and Xavier Serra (2017). Metrical-accent aware vocal onset detection. In 18th International Society for Music Information Retrieval Conference (ISMIR 2017)

Experiments

- Experiment 1 (Ex-1): manually annotated beats

meter	beat	Fmeas	P	R	Fmeas
düyek	Mauch	-	33.1	31.6	31.6
	Ex-1	-	40.4	39.5	39.0
	Ex-2	86.4	37.8	36.1	36.1
aksak	Mauch	-	42.1	36.9	37.9
	Ex-1	-	48.4	39.1	43.0
	Ex-2	72.9	45.0	39.0	40.3

Dzhambazov, Georgi, Andre Holzapfel, Ajay Srinivasamurthy and Xavier Serra (2017). Metrical-accent aware vocal onset detection. In 18th International Society for Music Information Retrieval Conference (ISMIR 2017)

Experiments

- Experiment 1 (Ex-1): manually annotated beats

	meter	beat	Fmeas	P	R	Fmeas
düyek	Mauch	-	33.1	31.6	31.6	31.6
	Ex-1	-	40.4	39.5	39.0	39.0
	Ex-2	86.4	37.8	36.1	36.1	36.1
aksak	Mauch	-	42.1	36.9	37.9	37.9
	Ex-1	-	48.4	39.1	43.0	43.0
	Ex-2	72.9	45.0	39.0	40.3	40.3

Dzhambazov, Georgi, Andre Holzapfel, Ajay Srinivasamurthy and Xavier Serra (2017). Metrical-accent aware vocal onset detection. In 18th International Society for Music Information Retrieval Conference (ISMIR 2017)

Experiments

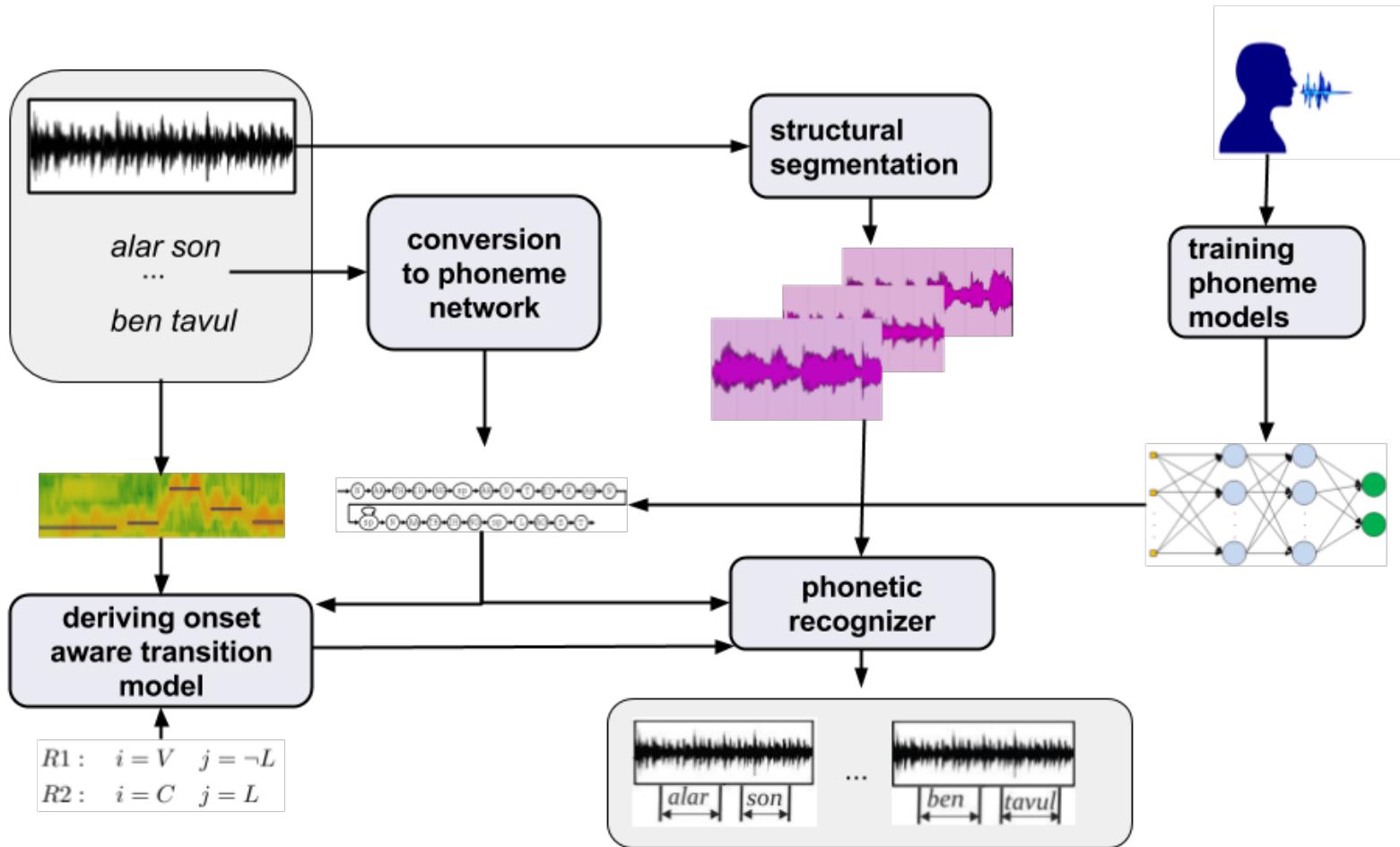
- Experiment 1 (Ex-1): manually annotated beats
- Experiment 2 (Ex-2): simultaneous bar-position and vocal onset

	meter	beat	Fmeas	P	R	Fmeas
düyek	Mauch	-	33.1	31.6	31.6	31.6
	Ex-1	-	40.4	39.5	39.0	39.0
	Ex-2	86.4	37.8	36.1	36.1	36.1
aksak	Mauch	-	42.1	36.9	37.9	37.9
	Ex-1	-	48.4	39.1	43.0	43.0
	Ex-2	72.9	45.0	39.0	40.3	40.3

Dzhambazov, Georgi, Andre Holzapfel, Ajay Srinivasamurthy and Xavier Serra (2017). Metrical-accent aware vocal onset detection. In 18th International Society for Music Information Retrieval Conference (ISMIR 2017)

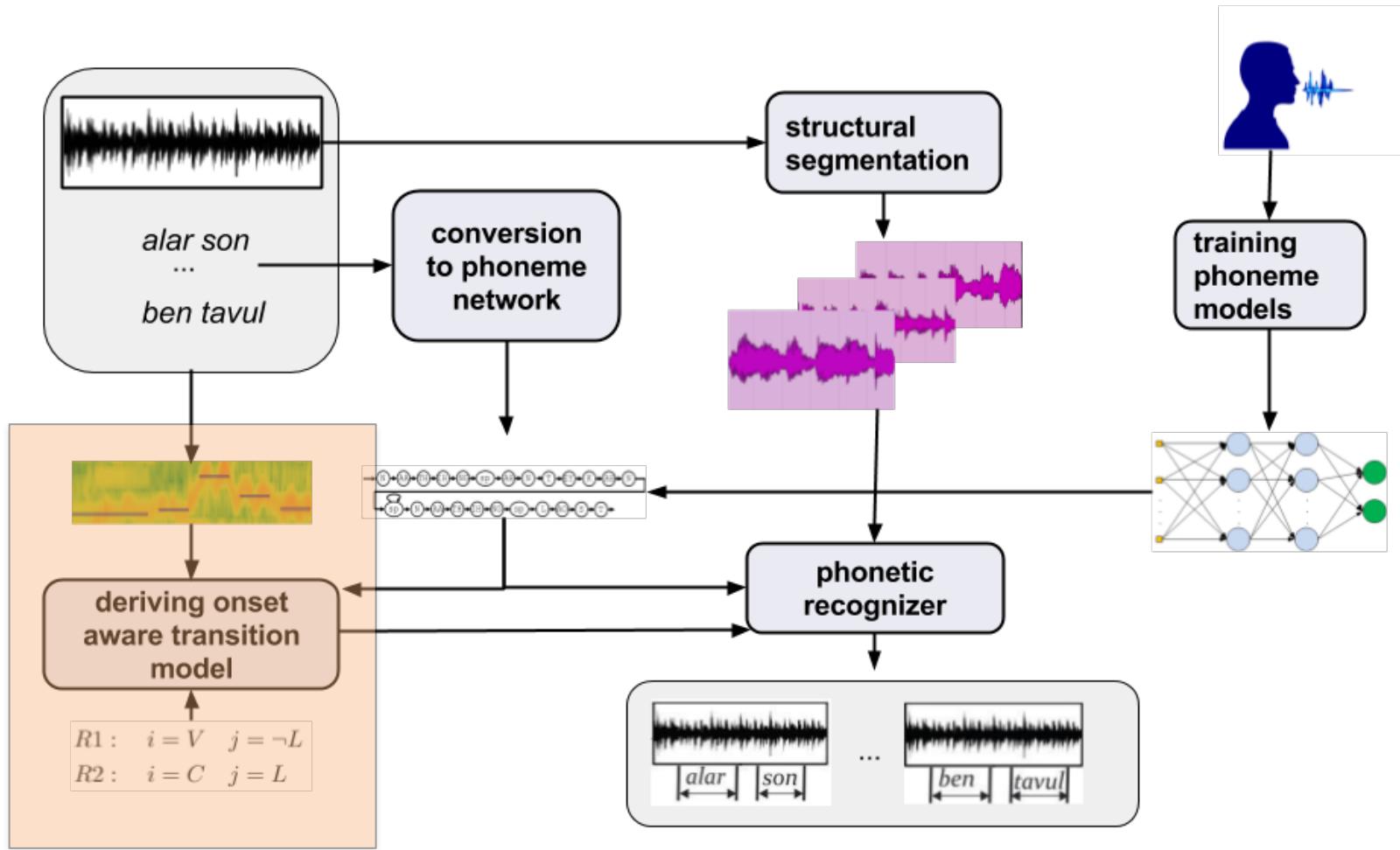
Chapter 5.4

Onset aware lyrics-to-audio alignment



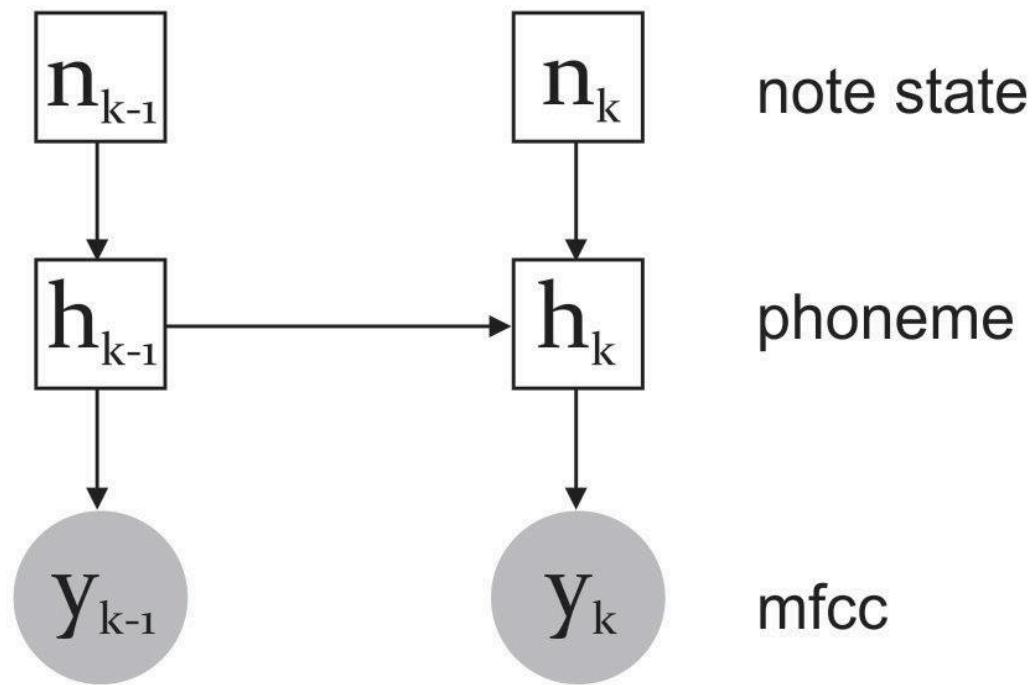
Chapter 5.4

Onset aware lyrics-to-audio alignment



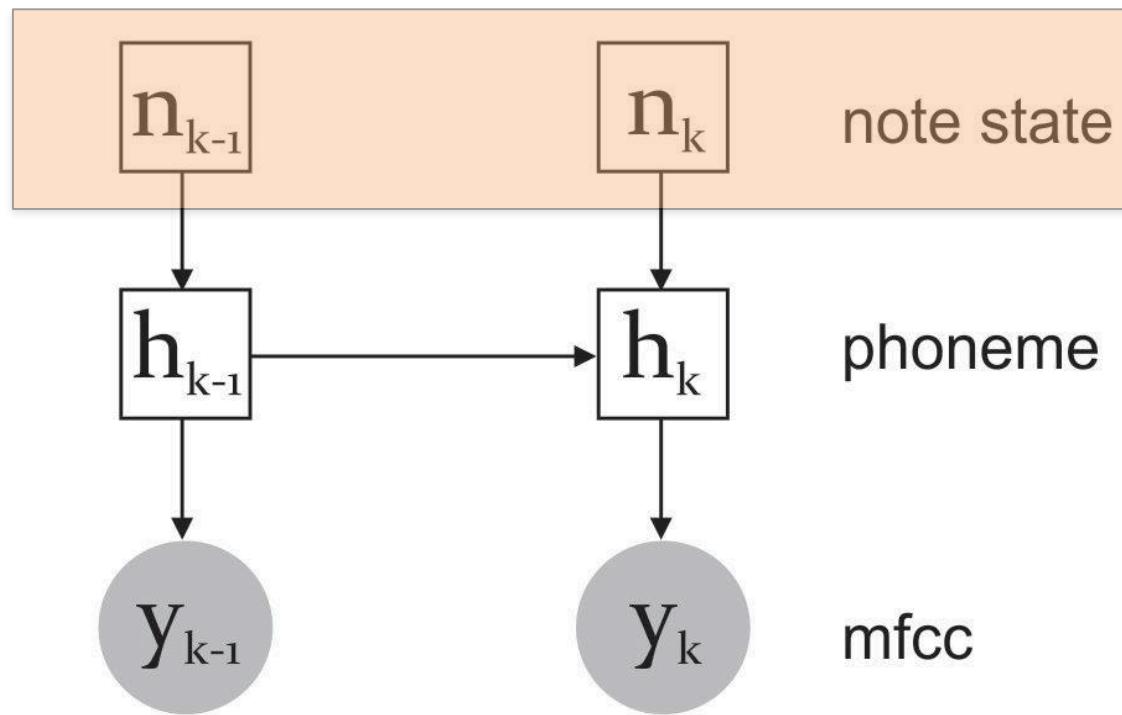
Model components

- Syllable boundaries coincide with segments of vocal notes (e.g. note onsets) (Sundberg 2006)



Model components

- Syllable boundaries coincide with segments of vocal notes (e.g. note onsets) (Sundberg 2006)



Phoneme transition rules

$$\left. \begin{array}{l} R1: i_t = V \quad i_{t+1} = C \setminus L \\ R2: i_t = C \setminus L \quad i_{t+1} = V \text{ or } L^* \\ R3: i_t = V \quad i_{t+1} = C \end{array} \right\} \begin{array}{l} \text{inter-syllable} \\ \text{intra-syllable} \end{array}$$

C: consonant

V: vowel

L: liquid (L,M,N) or semivowel (Y)

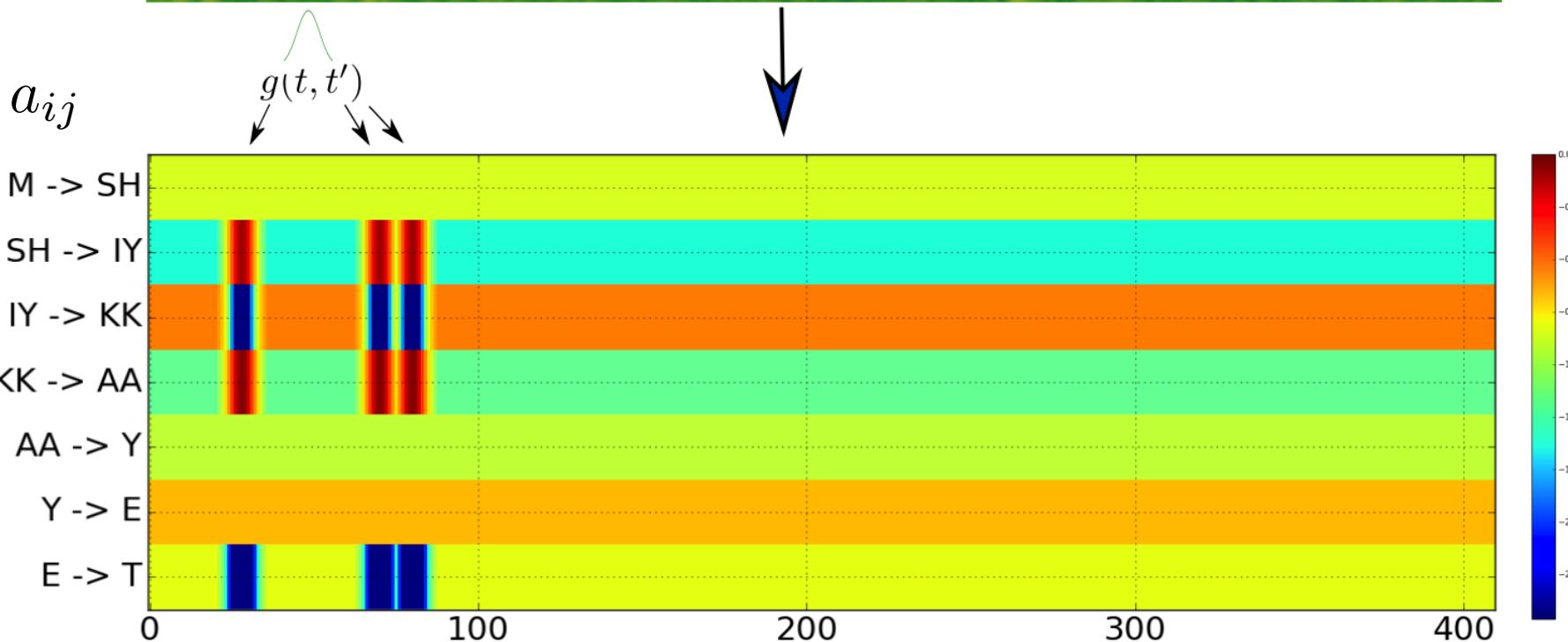
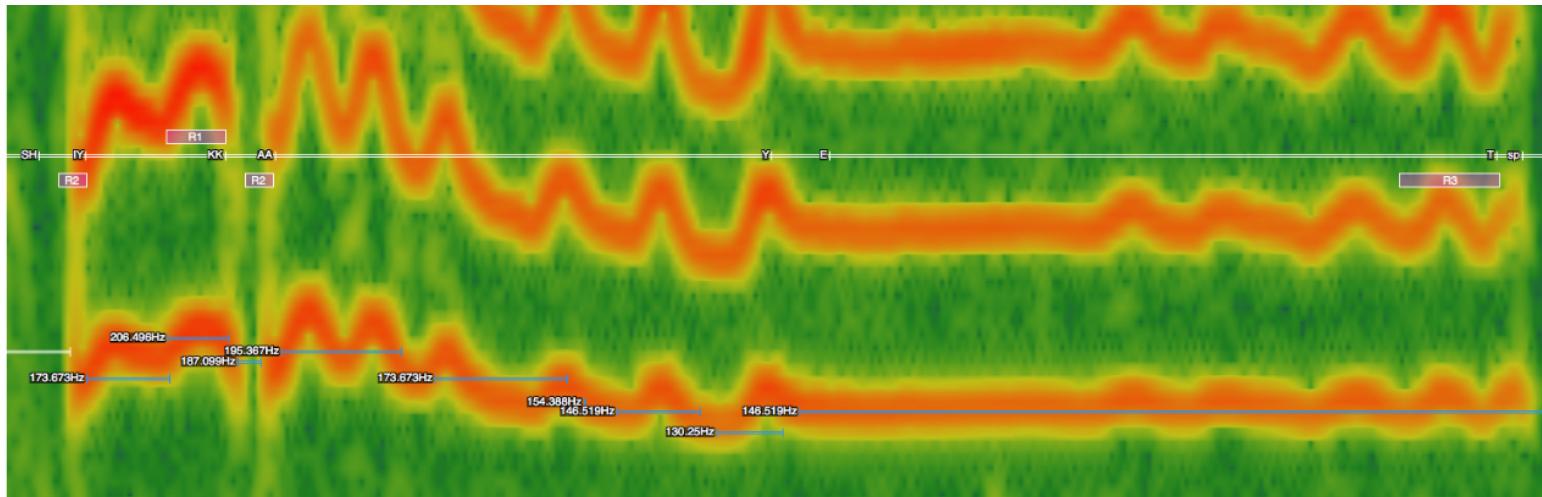
Transition model

$$a_{ij}(k) = \begin{cases} a_{ij} - g(k, k'), & R1 \text{ or } R3 \\ a_{ij} + g(k, k'), & R2 \\ a_{ij}, & \text{else} \end{cases}$$

- Variable-time transitions
 - modified at the presence of adjacent note onsets at time k'
 - unaffected otherwise

$g(k, k')$: normal distribution centered at k'

Transition model



Viterbi decoding

$$\delta_k(j) = \max_{i \in (j, j-1)} \delta_{k-1}(i) \boxed{a_{ij}} b_j(O_k)$$


- Decoding with one hidden variable
- Vocal onset detection as a preprocessing step

Viterbi decoding

$$\delta_k(j) = \max_{i \in (j, j-1)} \delta_{k-1}(i) \boxed{a_{ij}} b_j(O_k)$$

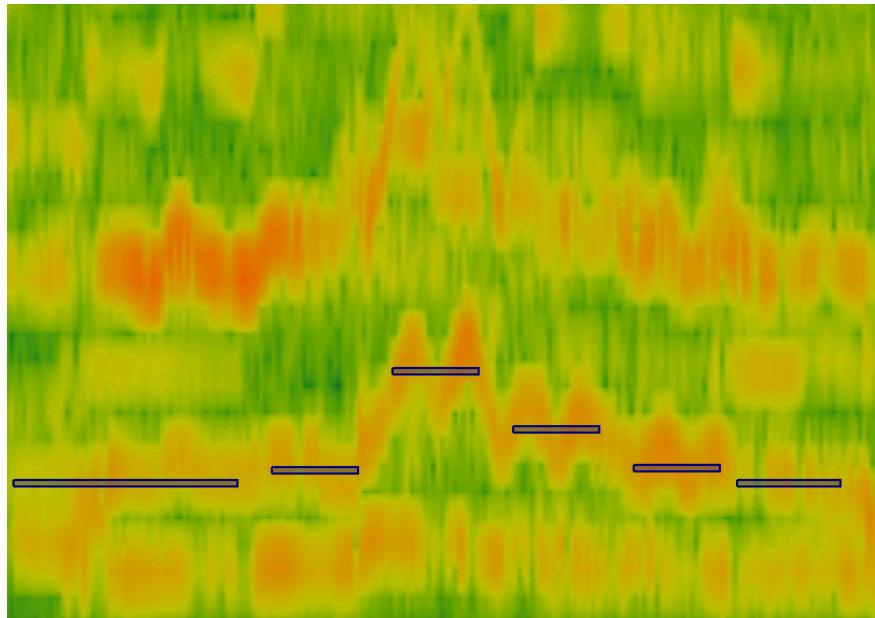
↓

$$\delta_k(j) = \max_{i \in (j, j-1)} \delta_{k-1}(i) \boxed{a_{ij}(k)} b_j(O_k)$$

- Decoding with one hidden variable
- Vocal onset detection as a preprocessing step

A cappella lyrics OTMM dataset

- 6 complete recordings (10 minutes) selected from the Multi-instrumental lyrics OTMM dataset
- added annotations of vocal onsets



Experiments

- Experiment 1: oracle (manually annotated) onsets

HMM	onset-aware HMM	
	automatic onsets	oracle onsets
79.2	81.7	82.5

Dzhambazov, Georgi, Ajay Srinivasamurthy, Sertan Şentürk, and Xavier Serra (2016). On the use of note onsets for improved lyrics-to-audio alignment in Turkish makam music. In 17th International Society for Music Information Retrieval Conference (ISMIR 2016)

Experiments

- Experiment 1: oracle (manually annotated) onsets

HMM	onset-aware HMM	
	automatic onsets	oracle onsets
79.2	81.7	82.5

Dzhambazov, Georgi, Ajay Srinivasamurthy, Sertan Şentürk, and Xavier Serra (2016). On the use of note onsets for improved lyrics-to-audio alignment in Turkish makam music. In 17th International Society for Music Information Retrieval Conference (ISMIR 2016)

Experiments

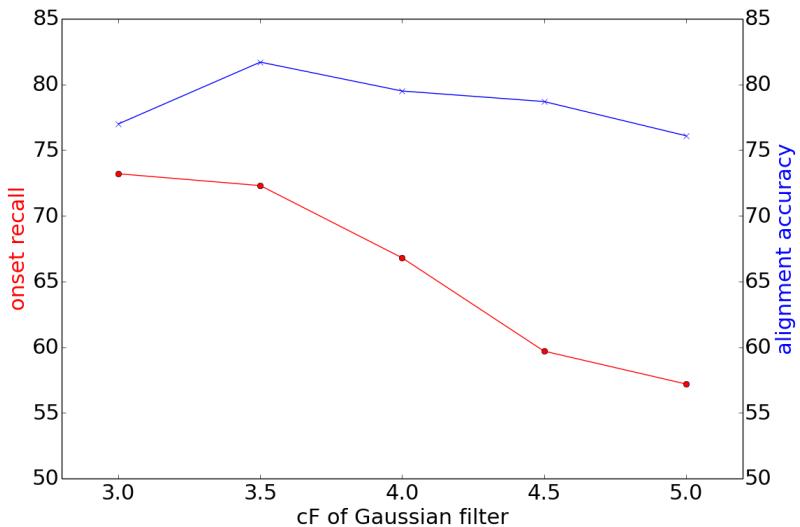
- Experiment 1: oracle (manually annotated) onsets

HMM	onset-aware HMM	
	automatic onsets	oracle onsets
79.2	81.7	82.5

Dzhambazov, Georgi, Ajay Srinivasamurthy, Sertan Şentürk, and Xavier Serra (2016). On the use of note onsets for improved lyrics-to-audio alignment in Turkish makam music. In 17th International Society for Music Information Retrieval Conference (ISMIR 2016)

Experiments

- Experiment 1: oracle (manually annotated) onsets
- Experiment 2: onsets detected with Flamenco singing voice transcription algorithm (Kroher, 2016)

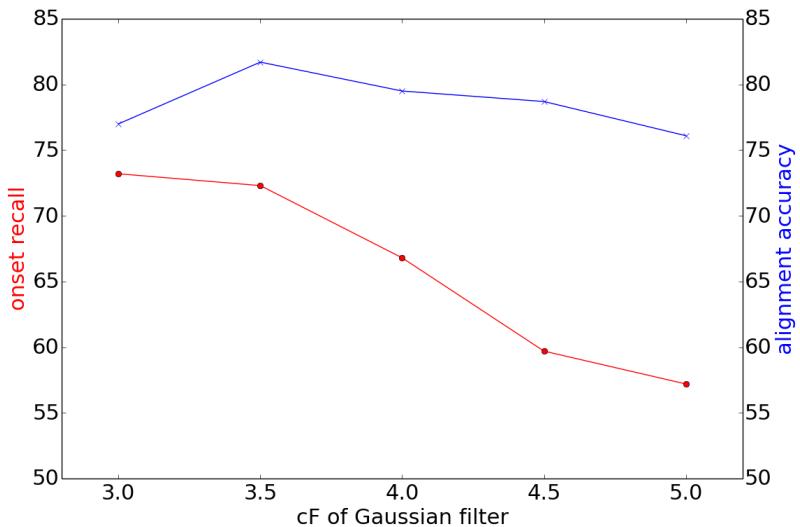


HMM	onset-aware HMM	
	automatic onsets	oracle onsets
79.2	81.7	82.5

Dzhambazov, Georgi, Ajay Srinivasamurthy, Sertan Şentürk, and Xavier Serra (2016). On the use of note onsets for improved lyrics-to-audio alignment in Turkish makam music. In 17th International Society for Music Information Retrieval Conference (ISMIR 2016)

Experiments

- Experiment 1: oracle (manually annotated) onsets
- Experiment 2: onsets detected with Flamenco singing voice transcription algorithm (Kroher, 2016)



HMM	onset-aware HMM	
	automatic onsets	oracle onsets
79.2	81.7	82.5

Dzhambazov, Georgi, Ajay Srinivasamurthy, Sertan Şentürk, and Xavier Serra (2016). On the use of note onsets for improved lyrics-to-audio alignment in Turkish makam music. In 17th International Society for Music Information Retrieval Conference (ISMIR 2016)

Outline

- Setting the stage
 - Introduction (Chapter 1)
 - Background (Chapter 2)
 - Baseline lyrics-to-audio alignment (Chapter 3)
- Main
 - Lyrics-duration aware lyrics-to-audio alignment (Chapter 4)
 - Metrical-accent aware lyrics-to-audio alignment (Chapter 5)
- Conclusion (Chapter 6) and Applications (Appendix)

Outline

- Setting the stage
 - Introduction (Chapter 1)
 - Background (Chapter 2)
 - Baseline lyrics-to-audio alignment (Chapter 3)
- Main
 - Lyrics-duration aware lyrics-to-audio alignment (Chapter 4)
 - Metrical-accent aware lyrics-to-audio alignment (Chapter 5)
- Conclusion (Chapter 6) and Applications (Appendix)

Appendix Applications

Appendix A Applications

Researchers of the CompMusic team have created a web application called Dunya-web¹ to showcase the technologies developed within the CompMusic project. Dunya-web is an application aimed at culture-aware music discovery (Porter et al., 2013). Dunya-web has a makam part, representing algorithms developed for the computational analysis of OTMM (Şentürk et al., 2015). Dunya-web stores all the audio recordings (including the OTMM datasets described in Section 3.2) and music scores, together with the lyrics.

The users can navigate the audio collection by searching or filtering by recordings, compositions, artists, makams, musical forms and/or usuls. Users can play the recordings and examine musical facets synchronous to the audio playback. Musical facets like pitch, the score, the tonic are visualized in a user-intuitive way.

The most successful lyrics-to-audio alignment (LAA) approach for OTMM, developed in this thesis, is the phonetic recognizer, aware of syllable durations. We integrated its python implementation into Dunya-web for a subset of the OTMM corpus available in Dunya-web (see Fig. A.1). This subset includes vocal recordings in the şark form with music scores and lyrics information available.

The ease of use of Dunya-web and its intuitive interface allows expert users (e.g. music aficionados, musicologists and/or music students) to follow the aligned lyrics, while listening to the audio. Simultaneously, the acoustic features (inverse spectral representation of MFCCs) representing the timbral differences of phonemes are displayed.

¹<http://dunya.compmusic.spl.edu/makam>

Dunya-web

dunya

Gel GÜZELİM ÇAMLICA'YA
by Münir Nurettin Selçuk

Album
Geçmişten Günümüze Türk Müziği - Kalplerden Dudaklara (Various Artists)

Compositions
Gel GÜZELİM ÇAMLICA'YA

Performers
Münir Nurettin Selçuk(Voice)

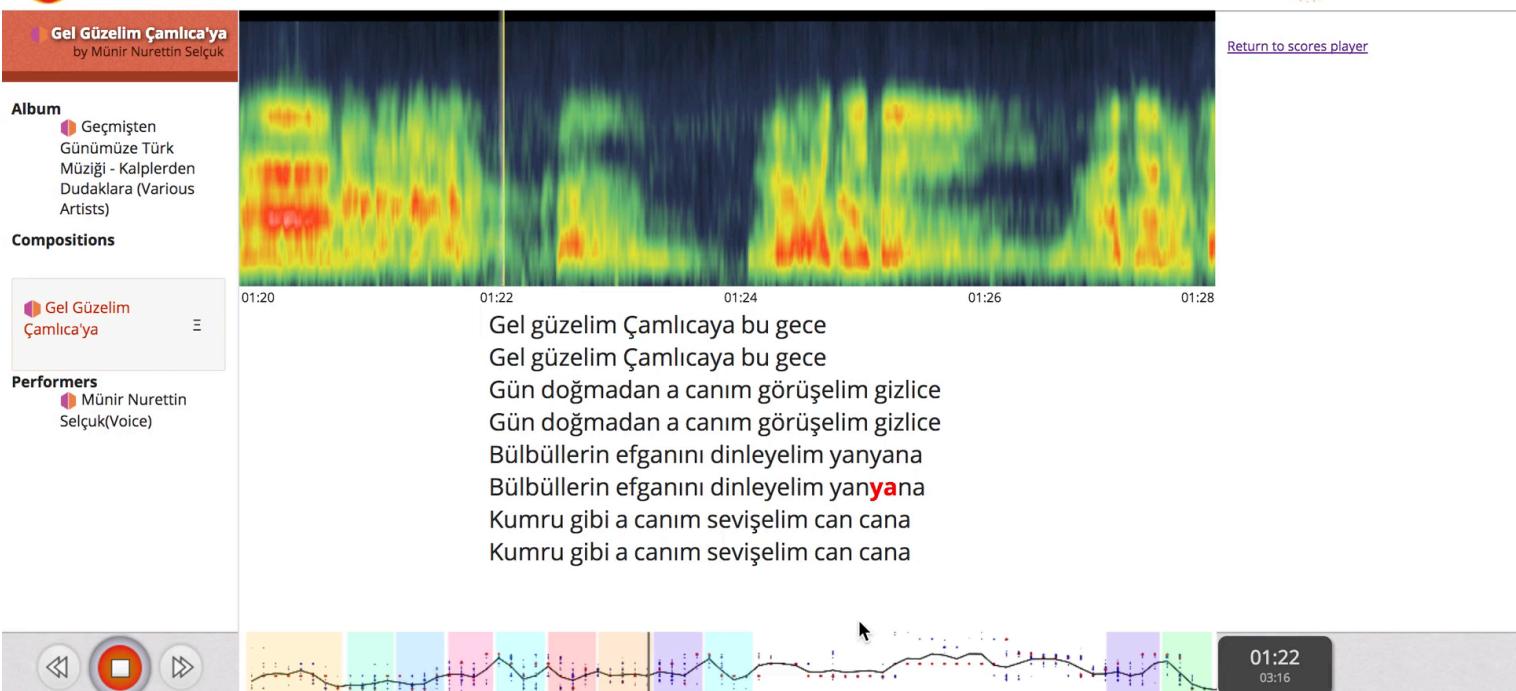
Funded by the European Research Council **erc** [About](#) [compmusic](#)

[Return to scores player](#)

01:20 01:22 01:24 01:26 01:28

Gel güzelim Çamlıca'ya bu gece
 Gel güzelim Çamlıca'ya bu gece
 Gün doğmadan a canım görüşelim gizlice
 Gün doğmadan a canım görüşelim gizlice
 Bülbüllerin efganını dinleyelim yanyana
 Bülbüllerin efganını dinleyelim yanyana
 Kumru gibi a canım sevişelim can cana
 Kumru gibi a canım sevişelim can cana

01:22 03:16



<http://dunya.compmusic.upf.edu/makam/>

Dunya-web

dunya

Gel GÜZELİM ÇAMLICA'YA
by Münir Nurettin Selçuk

Album
Geçmişten Günümüze Türk Müziği - Kalplerden Dudaklara (Various Artists)

Compositions
Gel GÜZELİM ÇAMLICA'YA

Performers
Münir Nurettin Selçuk(Voice)

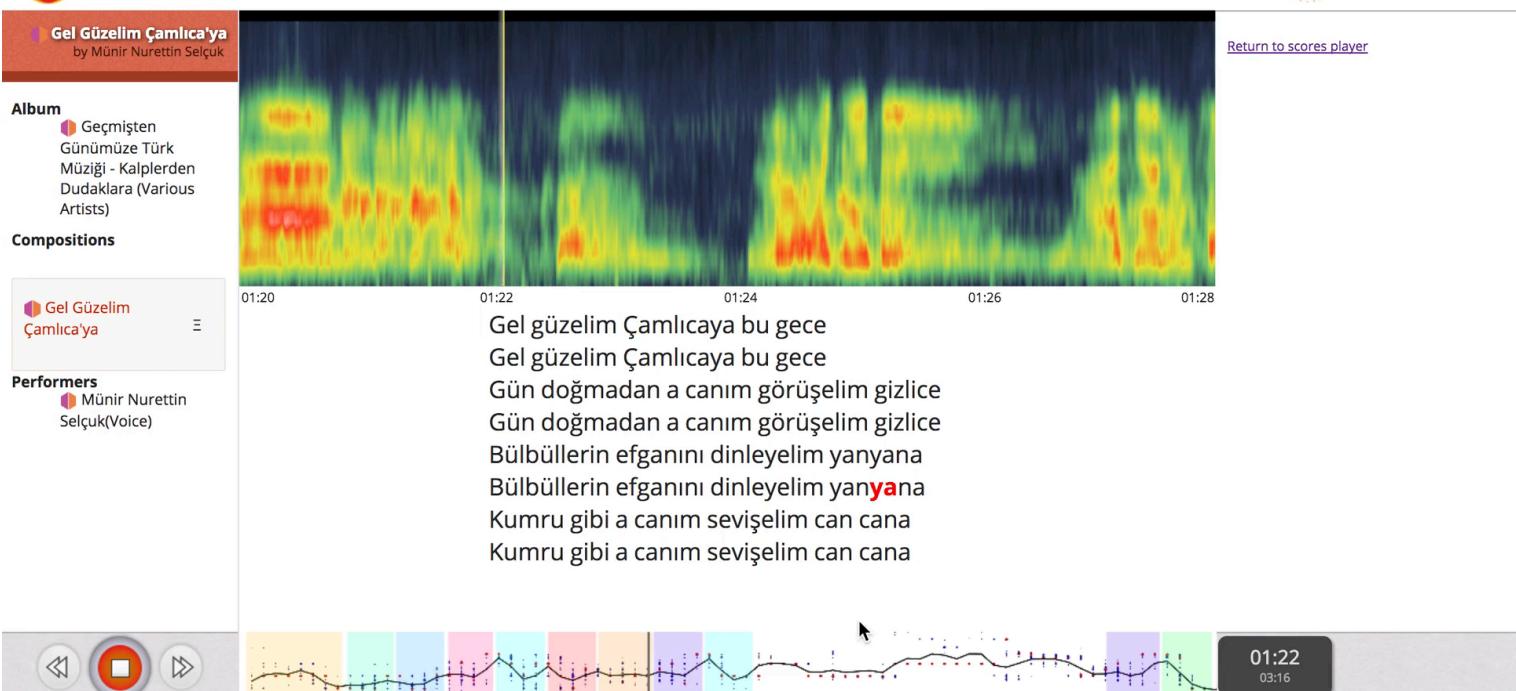
Funded by the European Research Council **erc** [About](#) [compmusic](#)

[Return to scores player](#)

01:20 01:22 01:24 01:26 01:28

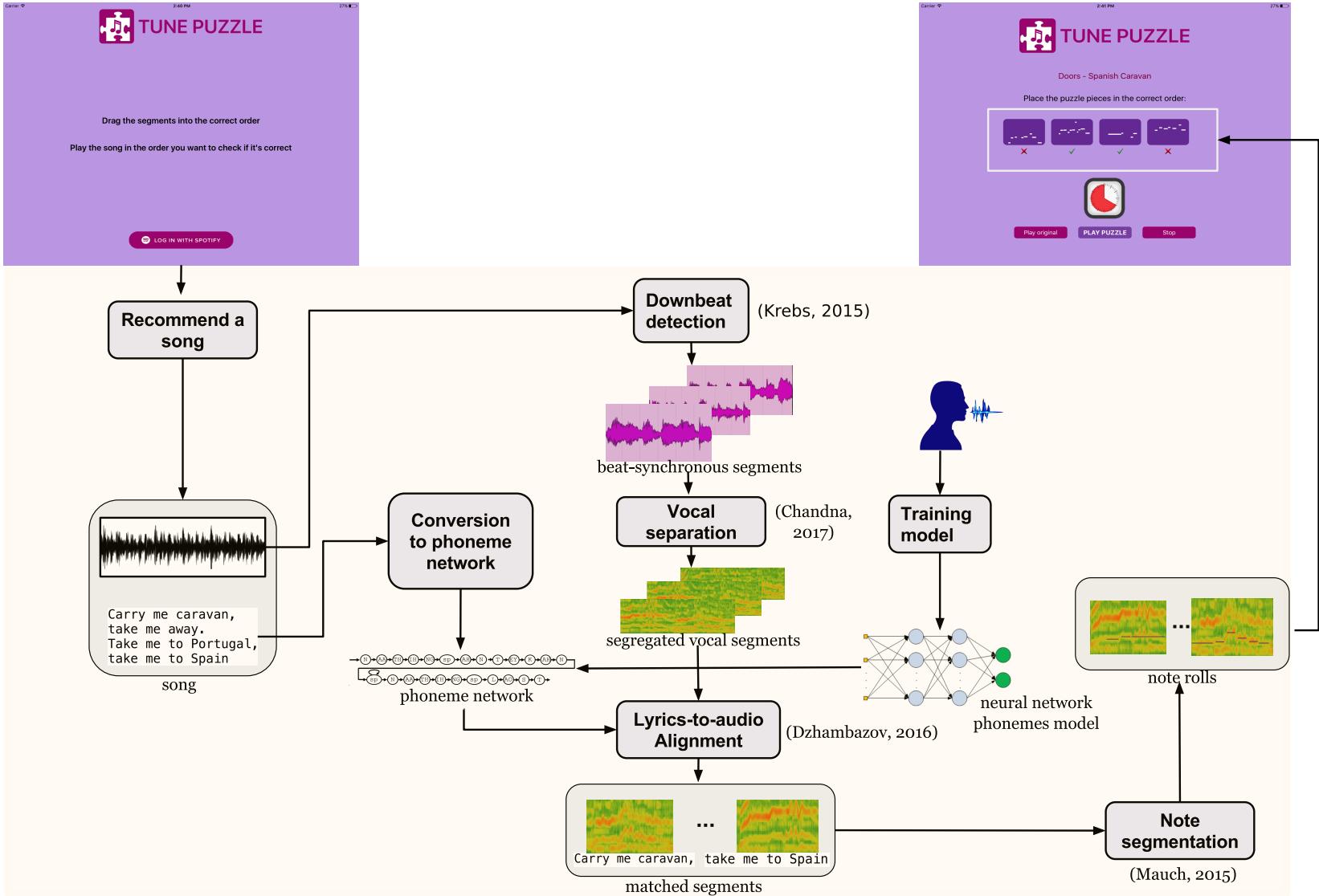
Gel güzelim Çamlıca'ya bu gece
 Gel güzelim Çamlıca'ya bu gece
 Gün doğmadan a canım görüşelim gizlice
 Gün doğmadan a canım görüşelim gizlice
 Bülbüllerin efganını dinleyelim yanyana
 Bülbüllerin efganını dinleyelim yanyana
 Kumru gibi a canım sevişelim can cana
 Kumru gibi a canım sevişelim can cana

01:22 03:16



<http://dunya.compmusic.upf.edu/makam/>

Tune puzzle



Chapter 6

Conclusions

Chapter 6

Conclusions

Broadly, this dissertation aimed to build culture-aware and domain-specific MIR approaches using probabilistic models for tracking lyrics in music audio signals. We proposed specific probabilistic models to represent how the transitions between consecutive sung phonemes are conditioned by different facets of music-domain knowledge. The models we build take into account some of these facets and consider them as *temporal context*, which is *complementary to lyrics*.

In order to evaluate the potential of the proposed models, we built a complete methodology for the automatic alignment of lyrics to audio recording (LAA) and evaluated its performance by the means of the LAA. As a baseline we chose a phonetic recognizer based on hidden Markov models (HMM): a methodology applied in most of existing computational studies on lyrics tracking. We applied the proposed methodologies to a specially compiled for this study datasets that are subsets from the CompMusic research corpora on OTMM and jingji. These music traditions pose a challenge to LAA because of their highly expressive singing style and the resulting thereof high degree of temporal variability and relatively long syllable durations. The reason is that traditional HMMs have waiting time in a state that cannot be too long. The low accuracy of the baseline phonetic recognizer confirmed that.

To this end, we built two separate extensions of the phonetic recognizer: one for middle-level complementary context and a separate one for fine-level context. As middle-level we modeled the influence of the temporal structure of a lyrics phrase on the phoneme transitions of lyrics. As to the fine-level context, we modeled how phoneme transitions interact with the position of the accents in the metrical cycle.

Importance of complementary context

- Duration aware alignment
 - biggest contribution model
 - dependent on music scores
- Metrical-accents improve vocal onsets detection
- Onset aware alignment
 - onset aware phoneme transition rules improve alignment
 - improvement is not substantial
 - note onsets approached by glides in OTMM

Scientific Contributions

- Conceptualize the interaction of phoneme transitions to complementary music knowledge as DBN
- Proposed several implementation simplifications of decoding in DBN
- Duration aware model beneficial in different music traditions

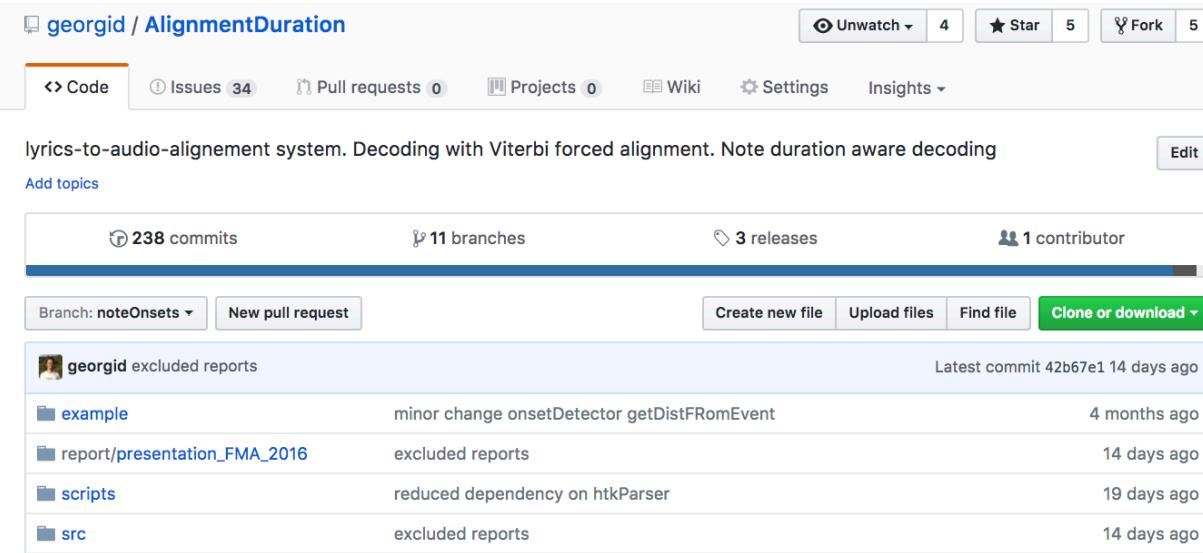
Other contributions

- Compiled several datasets of OTMM and jingju with annotations of different music facets
- Alignment with duration aware model implemented in Dunya-web
 - enriched music listening
 - tracking singing techniques
- MFCC extraction in *Essentia*



Reproducibility

- In python with few dependencies
(Essentia and sms-tools)
- The first open reproducible system for automatically aligning lyrics



The screenshot shows a GitHub repository page for 'AlignmentDuration'. The repository has 238 commits, 11 branches, 3 releases, and 1 contributor. The latest commit was 14 days ago. The repository description is: "lyrics-to-audio-alignement system. Decoding with Viterbi forced alignment. Note duration aware decoding". The repository URL is <https://github.com/georgid/AlignmentDuration>.

Branch	Commit Message	Time Ago
noteOnsets	excluded reports	14 days ago
example	minor change onsetDetector getDistFRomEvent	4 months ago
report/presentation_FMA_2016	excluded reports	14 days ago
scripts	reduced dependency on htkParser	19 days ago
src	excluded reports	14 days ago

github.com/georgid/AlignmentDuration

Future directions

- Combine the duration aware and the metrical-accent aware models into one
- Explore other music context facets (e.g. vibrato)
- Generalize the proposed models to other musics with principles similar to these of OTMM and jingju
- Vocal activity detection to replace structural segmentation

MIREX Lyrics-to-Audio Alignment

[page](#) [discussion](#) [view source](#) [history](#)



2017:Automatic Lyrics-to-Audio Alignment

Contents [hide]

- 1 Description
 - 1.1 Task specific mailing list
- 2 Data
 - 2.1 Audio Formats
- 3 Evaluation
- 4 Submission Format
 - 4.1 Input Data
 - 4.2 Output File Format
 - 4.3 Command line calling format
 - 4.4 README File
 - 4.5 Packaging submissions
- 5 Time and hardware limits
- 6 Submission opening date
- 7 Submission closing date
- 8 Potential Participants

Description

The task of automatic lyrics-to-audio alignment has as an end goal the synchronization between an audio recording of singing lyrics units can be estimated on different granularity: phonemes, words, lyrics lines, phrases. For this task word-level alignment is required.

Task specific mailing list

Data

The evaluation dataset contains 11 songs of popular music with annotations of timestamps of the words and the sentences. The dataset includes an accompaniment and a cappella singing voice only one.

You can read in detail about how the dataset was made here: [Recognition of Phonemes in A-cappella Recordings using Ten](#) has been kindly provided by Jens Kofod Hansen.

Knowledge-based Probabilistic Modeling for Tracking Lyrics in Music Audio Signals

28 June 2017, PhD defense

Georgi Dzhambazov
Dept. of Information and Communication
Technologies
Universitat Pompeu Fabra

Thesis Supervisor
Dr. Xavier Serra
Music Technology Group
Universitat Pompeu Fabra



Thesis Committee:
Dr. Axel Röbel (IRCAM)
Dra. Emilia Gómez (UPF)
Dr. Matthias Mauch (QMUL)

