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

Georgi Dzhambazov

TESI DOCTORAL UPF / 2017

Director de la tesi:

Dr. Xavier Serra Casals

Music Technology Group

Dept. of Information and Communication Technologies

Universitat
Pompeu Fabra
Barcelona

A web page with links to materials accompanying this manuscript is available at http://compmusic.upf.edu/phd-thesis-georgi

The doctoral defense was held on ...., 2017, at the Universitat Pompeu Fabra and scored as ....

Dr. Axel Röbel

Thesis Committee Member

IRCAM, Paris, France

Dr. Matthias Mauch,

Thesis Committee Member

Queen Mary University of London, UK

Dr. Emilia Gómez

Thesis Committee Member

Universitat Pompeu Fabra, Barcelona, Spain

To the divine voice of nature that gifted the human voice with the amazing capability of singing

Preface

This thesis has been carried out between October 2013 and April 2017 at the Music Technology Group (MTG) of Universitat Pompeu Fabra (UPF) in Barcelona (Spain), supervised by Dr. Xavier Serra Casals. All work has been conducted in collaboration with the CompMusic team at MTG. The singing material for the *a cappella lyrics OTMM dataset* has been recorded in several sessions in the studio of Bahçeşehir Üniversitesi, Istanbul, Turkey in June 2014, as well as in the studio of UPF throughout 2014-2015. The work in Chapter 5.3 has been conducted in collaboration with Dr. Andre Holzapfel (KTH Royal Institute of Technology, Stockholm, Sweden).

This work has been supported by the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) and the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583).

Acknowledgements

This thesis could not have been the same without the input of many mentors, influencers, and friends.

The most valuable piece of support came from the CompMusic team - for me a friendly multicultural atmosphere, where at felt very much at home. At the first place, CompMusic and this work would not have been there without the vision of Xavier Serra. I am extremely grateful to him for believing in me and for his constant readiness to reach a hand and guide me, as well as for the freedom of topic he gave me. My personal thanks go to Sertan Şentürk for his patient in-depth explanations of Makam music concepts as well as the fruitful research ideas; Ajay Srinivasamurthy for his never-ending improvement advices and productive talks on graphical models; Rong Gong (the fellow singing voice explorer); Sankalp Gulati (the great example of multi-productivity), Rafael Caro Repetto (for the excellent music theory hints), Gopala Krishna Koduri (for the startup-related chats), Andrés Ferraro (for his patience until we got together a working software prototype), Mohamed Sordo (my music hack day buddy), Sercan Atlı (for helping with the postproduction of the recorded singing), Yile Yang (for demistifying the Mandarin writing script).

I am greatly thankful to Andre Holzapfel for his guidance and kind reception in Stockholm. I deeply feel that with his multicultural music understanding and long research experience he had a major contribution on my latest pieces of work to stand out.

Being part of the vibrant community of MTG-ers I could benefit to the highest extent from the numerous opportunities for idea exchange. This research has benefited from discussions with Emilia Gómez (she knows the answer to every doubt!), Marius Miron (my Balkan brother from another father), Nadine Kroher (the coolest summer school mate), Dmitry Bogdanov (the green tea provider), Pablo Alonso (the fastest feature coder), Juan José Bosch (for the excellent research hints), Alastair Porter, Swapnil Gupta, Xavier Favory, Oriol Romanì, Giuseppe Bandiera, Néstor Nápoles, Julián Urbano, Jordi Pons, Sergio Oramas, Julio Carabias, Frederic Font, Zacharias Vamvakousis, Rafael Ramírez, Sergio Giraldo, Martì Umbert, Merlijn Blaauw, Panos Papiotis, Stanislaw Gorlow, ...

I am thankful to Cristina, Sonia and Lydia for reducing the administrative burden and last but not least to Alba Rosado for providing immense help with the most critical factor for this research: arranging the AGAUR research grant!

I am grateful to my fellow researchers in the domain of lyrics-related studies Anna Kruspe, Daniel Stoller and Matt McVicar for the productive discussions and their valuable improvement suggestions.

On my hitherto journey in Music Information Retrieval (MIR), the fruit of which is this dissertation, I was fortunate to interact with many researchers that gave me valuable ideas and directions (in chronological order): Sebastian Stober (for opening my eyes for MIR), Rolf Bardeli (for motivating me to pursue a switch to MIR in my orientation time in the area of speech processing), Frank Kurth, Christian Dittmar (for supporting my ideas), Meinard Müller, Maarten Grachten, Gerhard Widmer, Florian Krebs, Markus Schedl, Peter Knees, Sebastian Böck, Andreas Arzt, Tom Collins (for the motivating words in the hardest moments), Robert Kaye, Ryan Groves, Anssi Klapuri, Thomas Lidy, Amélie Anglade, Anja Volk, Jakob Abesser, Colin Raffel, ...

Without my family my productivity would not have been the same. I am grateful to my mother Nikolina for habituating me to be persevering in life and also in this writing journey. I deeply thank my sister Velizara, who, with her brilliant design skills, provided an immense support in depicting the diagrams throughout this thesis. I am happy to also know Georgi Petrov, who was next to me in the last most critical days (nights) before submission.

I deeply believe that my passion for MIR will remain to be strong and life-long mainly because of the amazing people I met and will be meeting...

Abstract

This thesis proposes specific signal processing and machine learning methods for automatically aligning the lyrics of a song to its corresponding audio recording. The research carried out falls in the broader field of music information retrieval (MIR) and in this respect, we aim at improving some existing state-of-the-art methods, by introducing domain-specific knowledge.

The goal of this work is to devise models capable of tracking in the music audio signal the sequential aspect of one particular element of lyrics – the phonemes. Music can be understood as comprising different facets, one of which is lyrics. The models we build take into account the *complementary context* that exists around lyrics, which is any musical facet complementary to lyrics. The facets used in this thesis include the structure of the music composition, temporal structure of a lyrics line, the structure of the metrical cycle. From this perspective, we analyse not only the low-level acoustic characteristics, representing the timbre of the phonemes, but also higher-level characteristics, in which the complementary context manifests. We propose specific probabilistic models to represent how the transitions between consecutive sung phonemes are conditioned by different facets of complementary context.

The complementary context, which we address, unfolds in time according to principles that are particular of a music tradition. To capture these, we created corpora and datasets for two music traditions, which have a rich set of such principles: Ottoman Turkish makam and Beijing opera. The datasets and the corpora comprise different data types: audio recordings, music scores, and metadata. From this perspective, the proposed models can take advantage both of the data and the music-domain knowledge of particular musical styles to improve existing baseline approaches.

As a baseline, we choose a phonetic recognizer based on hidden Markov models (HMM): a widely-used method for tracking phonemes both in singing and speech processing problems. We present refinements in the typical steps of existing phonetic recognizer approaches, tailored towards the characteristics of the studied music traditions. On top of the refined baseline, we devise probabilistic models, based on dynamic Bayesian networks (DBN) that repre-

sent the relation of phoneme transitions to its complementary context. Two separate models are built for two granularities of complementary context: the temporal structure of a lyrics line (higher-level) and the structure of the metrical cycle (finer-level). In one model we exploit the fact the syllable durations depend on their position within a lyrics line. Information about the expected durations is obtained from the score, as well as from music-specific knowledge. Then in another model, we analyse how vocal note onsets, estimated from audio recordings, influence the transitions between consecutive vowels and consonants. We also propose how to detect the time positions of sung note onsets by tracking simultaneously the positions in the metrical cycle (i.e. metrical accents).

In order to evaluate the potential of the proposed models, we use lyrics-to-audio alignment as a concrete task. Each model improves the alignment accuracy, compared to the baseline, which is based solely on the acoustics of the phonetic timbre. This validates our hypothesis that knowledge of complementary context is an important stepping stone for computationally tracking lyrics, especially in the challenging case of singing with instrumental accompaniment.

The outcomes of this study are not only theoretic methods and data, but also specific software tools that have been integrated into Dunya — a suite of tools, built in the context of CompMusic, a project for advancing the computational analysis of the world's music. With this application, we have also shown that the developed methods are useful not only for tracking lyrics, but also for other use cases, such as enriched music listening and appreciation, and for educational purposes.

Resum

La tesi aquí presentada proposa metodologies d'aprenentatge automàtic i processament de senyal per alinear automàticament el text d'una cançó amb el seu corresponent enregistrament d'àudio. La recerca duta a terme s'engloba en l'ampli camp de l'extracció d'informació musical (Music Information Retrieval o MIR). Dins aquest context la tesi pretén millorar algunes de les metodologies d'última generació del camp introduint coneixement específic de l'àmbit.

L'objectiu d'aquest treball és dissenyar models que siguin capaços de detectar en la senyal d'àudio l'aspecte seqüencial d'un element particular dels textos musicals; els fonemes.

Podem entendre la música com la composició de diversos elements entre els quals podem trobar el text. Els models que construïm tenen en compte el context complementari del text. El context són tots aquells aspectes musicals que complementen el text, dels quals hem utilitzat en aquest tesi: la estructura de la composició musical, la estructura de les frases melòdiques i els accents rítmics. Des d'aquesta prespectiva analitzem no només les característiques acústiques de baix nivell, que representen el timbre musical dels fonemes, sinó també les característiques d'alt nivell en les quals es fa patent el context complementari. En aquest treball proposem models probabilístics específics que representen com les transicions entre fonemes consecutius de veu cantanda es veuen afectats per diversos aspectes del context complementari.

El context complementari que tractem aquí es desenvolupa en el temps en funció de les característiques particulars de cada tradició musical. Per tal de modelar aquestes característiques hem creat corpus i conjunts de dades de dues tradicions musicals que presenten una gran riquesa en aquest aspectes; la música de l'opera de Beijing i la música makam turc-otomana. Les dades són de diversos tipus; enregistraments d'àudio, partitures musicals i metadades. Des d'aquesta prespectiva els models proposats poden aprofitar-se tant de les dades en si mateixes com del coneixement específic de la tradició musical per a millorar els resultats de referència actuals.

Com a resultat de referència prenem un reconeixedor de fonemes basat en models ocults de Markov (Hidden Markov Models o HMM), una metodologia abastament emprada per a detectar fonemes tant en la veu cantada com en la parlada. Presentem millores en els processos comuns dels reconeixedors de fonemes actuals, ajustant-los a les característiques de les tradicions musicals estudiades. A més de millorar els resultats de referència també dissenyem models probabilistics basats en xarxes dinàmiques de Bayes (Dynamic Bayesian Networks o DBN) que respresenten la relació entre la transició dels fonemes i el context complementari. Hem creat dos models diferents per dos aspectes del context complementari; la estructura de la frase melòdica (alt nivell) i la estructura mètrica (nivell subtil). En un dels models explotem el fet que la duració de les síl · labes depén de la seva posició en la frase melòdica. Obtenim aquesta informació sobre les frases musical de la partitura i del coneixement específic de la tradició musical. En l'altre model analitzem com els atacs de les notes vocals, estimats directament dels enregistraments d'àudio, influencien les transicions entre vocals i consonants consecutives. A més també proposem com detectar les posicions temporals dels atacs de les notes en les frases melòdiques a base de localitzar simultàniament els accents en un cicle mètric musical.

Per tal d'evaluar el potencial dels mètodes proposats utlitzem la tasca específica d'alineament de text amb àudio. Cada model proposat millora la precisió de l'alineament en comparació als resultats de referència, que es basen exclusivament en les característiques acústiques tímbriques dels fonemes. D'aquesta manera validem la nostra hipòtesi de que el coneixement del context complementari ajuda a la detecció automàtica de text musical, especialment en el cas de veu cantada amb acompanyament instrumental.

Els resultats d'aquest treball no consisteixen només en metodologies teòriques i dades, sinó també en eines programàtiques específiques que han sigut integrades a Dunya, un paquet d'eines creat en el context del projecte de recerca CompMusic, l'objectiu del qual és promoure l'anàlisi computacional de les músiques del món. Gràcies a aquestes eines demostrem també que les metodologies desenvolupades es poden fer servir per a altres aplicacions en el context de la educació musical o la escolta musical enriquida.

(Translated from English by Oriol Romaní Picas)

Resumen

Esta tesis propone metodologías específicas de procesamiento de señales y aprendizaje automático para alinear de manera automática la letra de una canción a su correspondiente grabación de audio. La investigación llevada a cabo recae en el campo más amplio de la recuperación de información musical (MIR), y por lo tanto, pretendemos con ella mejorar algunas de las metodologías más avanzadas de la actualidad, introduciendo conocimiento específico del dominio.

El objetivo de este trabajo es diseñar modelos capaces de rastrear en la señal de audio musical el aspecto secuencial de un elemento particular de la letra, los fonemas. Se puede entender que la música comprende diferentes facetas, una de las cuales es la letra. Los modelos que construimos tienen en cuenta el contexto complementario que existe alrededor de la letra, que es cualquier faceta musical complementaria a las letras. Las facetas utilizadas en esta tesis incluyen la estructura de la composición musical, la estructura temporal de un enunciado de la letra, la estructura métrica. Desde esta perspectiva, analizamos no sólo las características acústicas de bajo nivel, que representan el timbre de los fonemas, sino también las características de alto nivel, en las que se manifiesta el contexto complementario. Proponemos modelos probabilísticos específicos para representar cómo las transiciones entre fonemas cantados consecutivamente están condicionadas por diferentes facetas del contexto complementario.

El contexto complementario, al cual abordamos, se despliega en el tiempo según principios propios de una tradición musical. Para capturar estos principios, hemos creado corpus y conjuntos de datos para dos tradiciones musicales, dichas que tienen un rico conjunto de tales principios: makam turco otomano y ópera de Beijing. Los conjuntos de datos y los corpus comprenden diferentes tipos de datos: grabaciones de audio, partituras y metadatos. Desde esta perspectiva, los modelos propuestos pueden aprovechar tanto los datos como el conocimiento del dominio de la música de determinados estilos musicales para mejorar los enfoques existentes usados como referencia.

Como punto de partida, elegimos un reconocedor fonético basado en modelos ocultos de Markov (HMM): una metodología ampliamente utilizada para el rastreo de fonemas tanto en el canto como en los problemas de procesamiento del habla. Presentamos mejoras en los pasos típicos de los enfoques de reconocimiento fonético existentes, dirigidos hacia las características de las tradiciones musicales estudiadas. Además de los puntos de partida mejorados, usamos modelos probabilísticos basados en redes bayesianas dinámicas (DBN) que representan la relación de las transiciones de fonemas con su contexto complementario. Se construyen dos modelos independientes para dos granularidades de contexto complementario: la estructura temporal de un enunciado de la letra (alto nivel) y la estructura del ciclo métrico (nivel más fino). En un modelo explotamos el hecho de que las duraciones de las sílabas dependen de su posición dentro de un enunciado de la letra. La información sobre las duraciones esperadas se obtiene de la partitura, así como de conocimientos específicos de la música. Luego, en otro modelo, analizamos cómo los onsets de notas vocales, estimados a partir de grabaciones de audio, influyen en las transiciones entre vocales consecutivas y consonantes. También proponemos cómo detectar las posiciones de tiempo de los onsets de nota cantada mediante el rastreo simultáneo de las posiciones en el ciclo métrico (es decir, acentos métricos).

Con el fin de evaluar el potencial de los modelos propuestos, utilizamos la alineación de letra a grabación de audio como una tarea concreta. Cada modelo mejora la precisión de la alineación, en comparación con el modelo de referencia inicial, que se basa únicamente en la acústica del timbre fonético. Esto valida nuestra hipótesis de que el conocimiento del contexto complementario es un factor importante para el seguimiento computacional de las letras, especialmente en el desafiante caso de cantar junto a un acompañamiento instrumental.

Los resultados de este estudio no son sólo metodologías teóricas y datos, sino también herramientas de software específicas que se han integrado en Dunya — un conjunto de herramientas, construido en el contexto de CompMusic, un proyecto para avanzar el análisis computacional de la música del mundo. Con esta aplicación, también hemos demostrado que las metodologías desarrolladas son útiles no sólo para el seguimiento de letras, sino también para otros casos de uso, como una experiencia y apreciación enriquecidas al escuchar música, y fines educativos.

(Translated from English by Néstor Nápoles)

Contents

List of Figures

List of Tables

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 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 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, causing 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 blocks of words — can be considered as a discrete number of timbral classes, wherein each class has a characteristic spectral template. Human speakers possess the ability to articulate phonemes. In fact, singers articulate by means of given vowels even when not singing with actual lyrics. For brevity, in the rest of this thesis we will refer to the aspect of singing voice timbre that makes humans distinguish between the identity of different phonemes as *phonetic timbre*. The transitions between consecutive phonemes can be considered as gradual changes of timbre as opposed to the short-term timbral fluctuations, related to the instability of the human vocal tract. We will refer to these changes as *phonetic timbre changes*. That is to say, the timbre of the singing voice, in addition to carrying the identity and *instrumental* quality, is the reason why we distinguish a particular phoneme in a given time instant. Therefore, despite varying continuously, the singing voice timbre can be considered to belong to one of a discrete set of phonemes at a particular point in time. Unlike the transitions among classes of *instrumental* timbre, the exact time positions, in which singers transition from one phoneme to another, can be distinguished by most listeners unambiguously.

The research carried out in this dissertation focuses on the acoustics of the lyrics of singing voice in polyphonic music and their relation to written lyrics. Sung lyrics can be studied from many different perspectives, whereas this thesis takes an MIR viewpoint, aiming at the analysis of temporal changes of lyrics content with an end goal of automatic synchronization between sung and written lyrics.

1.1 Scientific Context

Singing voice processing is still one of the most challenging subfields of MIR. Challenging remain especially the problems of singing voice detection; transcription of the singing melody and transcription of the lyrics. The timbre of singing voice has multiple functions: One is to articulate actual phonemes; another is to represent the 'instrumental quality', which makes singers stand out from the rest of the accompanying instruments in orchestral performances (Sundberg and Rossing, 1990). Some of the problems related to timbre are summarized by Goto (2014) as 'vocal timbre analysis' and include automatic lyrics processing of voice, singer identification, comparison of timbral similarity.

Looking at MIR in general, there is still a wide gap between what can be automatically extracted from audio recordings and the semantically meaningful high-level musical concepts, which listeners associate with singing (Wiggins, 2009). A possible reason for this semantic gap might be that the approach usually taken is bottom up: low-level features are extracted and then high-level concepts are inferred by aggregating these features. In such approaches often high-level musical knowledge is not reflected in the computational model itself. Most MIR research outcomes have been validated against eurogenetic music and do not generalize to other music cultures of the world[1]. Applying state of the art methods for analysis of non-eurogenetic music yields suboptimal results (Serra, 2011). The lack of explicit modeling of music knowledge in

[1] The term *eurogenetic* is coined in Holzapfel et al. (2014) to avoid the misleading division music into Western and non-Western. It designates the discussed theoretical constructs are motivated by the European common practice period

computational work becomes a more evident disadvantage for material from non-eurogenetic music. This is because these musics are characterized by their own specific music principles. In fact most music to the east of Europe has elaborate rhythmic and melodic framework. Thus extending state of the art approaches by fusing all music-specific concepts, relevant for a given task, would exploit the full potential of the studied music. With this end goal in mind, the project CompMusic[2](Computational Models for the Discovery of the World's Music) was envisioned (Serra et al., 2013). Art music of five different cultures is being studied in the project: Hindustani (North India), Carnatic (South India), Turkish-makam (Turkey), Arab-Andalusian (Maghreb) and Beijing opera (China). The classical music of Turkey, also often referred to as Turkish-makam, is the focus of this study. In this thesis we will refer to it as OTMM[3]. We extended one of the presented models to Beijing opera (also referred to as jingju), too.

In particular for singing voice, in current MIR research little work focuses on methods, which model sung lyrics together with their interdependence on complementary musical aspects like, for example, the progression of a metrical cycle. One possible reason for that could be that such a model is hard to design and develop, because it has to be considerably generic to represent such interdependencies for any music genre in the broad sense. In contrast to that, for each of the music traditions of CompMusic there is a well-defined framework of specific music principles. Therefore it may be more feasible to develop a singing voice model that represents jointly phonetic timbre and these music principles for a particular music tradition. This is mainly because these principles for one music tradition could be summarized into a model in a much more straight-forward way than for multiple genres of music.

The work covered in this thesis has been developed to focus on OTMM. A personal motivation for me is that OTMM has nature very akin to the traditional music of Bulgaria — the music with which I grew up. Being the official music of the Ottoman Empire, it has influenced enormously all Balkan music, and to a rather high extent Bulgarian traditional music. This made me naturally understand and appreciate its musically rich melodic and rhythmic framework throughout the research conducted in this thesis.

1.2 Motivation

1.2.1 Why consider complementary musical facets?

The progression of lyrics in singing is not an isolated aspect: lyrics have an inherent correlation with other music facets. In an abstract sense these music

[2] http://compmusic.upf.edu

[3] For the sake of compliance, this naming is adopted from a related computational study — Şentürk (2016)

facets can be imagined as the 'skeleton' and lyrics as the 'flesh'. Upon songwriting composers often distribute the lyrics syllables, guided by the locations of the 'skeleton' melodic and rhythmic accents. In this respect, studying the temporal aspects of sung lyrics also requires describing their relations to the temporal progression of the underlying music events. These relations unfold in time to form *musical context* in time for the sung lyrics, that is different from and in this sense *complementary* to their timbre. By *complementary musical context* (or simply *complementary context*) we will refer to any music facet, manifesting in events simultaneous to the transitions of lyrical units and having an influence on them[4]. In this thesis we will refer to *unit of lyrics* (or *lyrical units*) as a general concept that stands for different linguistic granularity: lyrics line, a phrase of words, word, syllable, phoneme. For the sake of organization, we suggest dividing the complementary context of lyrics into three hierarchical levels with respect to its time granularity: the overall structure of the composition (coarse-level), the temporal structure of a lyrics line (mid-level) and the structure of a metrical cycle (fine-level).

Each facet of the complementary context manifests itself as the time progression of concrete music events (see Figure 1.1). Firstly, at the highest context level, the overall structure of the composition determines the highest-level of lyrics units: lyrics lines. The transition from current structural section (e.g. verse, chorus) to another one can be considered a musical event, which signals the transition to another lyrics line (or whole lyrics paragraph). Then, on the mid-level of context, the duration of each lyrics syllable is conditioned on events of the sung melodic phrase. Singers may prolong or shorten syllables, in order to align them with accents of the melodic phrase. Finally, onsets of syllables often co-occur with accents within a metrical cycle.

These interdependences are important for OTMM, which has some very specific principles of the main musical facets, explained by a well-grounded theory (Ederer, 2011; Popescu-Judetz, 1996). In addition to that, the sung melodic phrases are rich in expressive ornamentation elements (such as melismas). For all these reasons, OTMM provides an excellent framework to incorporate domain-specific knowledge into a context-aware model of sung lyrics.

Its well-grounded theory also paved the way to computational work on some of these aspects, including among others predominant melody extraction (Atlı et al., 2014); relation of metrical accents and vocal note onsets (Holzapfel, 2015); score-informed structural section discovery (Şentürk et al., 2014). In this context, we can benefit from those studies and use their outcomes as facets of complementary context.

[4] We adopted the term *musical context* from Mauch (2010), where it is introduced for the task of chord estimation to serve a similar function. The authors use it to represent any musical facet, which is complementary to the harmonic content of chords — the main facet being tracked. We decided to use *complementary* instead of *musical* to emphasize the fact it is complementary to phonetic timbre

1.2.2 Why lyrics-to-audio alignment?

In this thesis, we will focus on the concrete problem of LAA. It aims to automatically synchronize the lyrics in their two representations: sung in an audio recording and written as text. An audio recording and its corresponding lyrics are input to an alignment system. It estimates their temporal relationship, providing as output the start timestamp of each phoneme from the phoneme sequence, comprising the lyrics. Among all research questions, related to sung lyrics defined in the context of MIR, we chose to work on LAA for several reasons.

Firstly, the measuring the accuracy of an alignment system provides a quantitative way to access the influence of the complementary context on the transitions between sung lyrical units. From this perspective, we only focused on one aspect of singing voice timbre: the phonetic timbre changes. Secondly, automating the LAA has numerous end-user applications. Building a piece of work with application potential is also a major motivation behind this research. Some applications of alignment include karaoke-like lyrics visualization, automatic thumbnailing and enriched music listening.

Note that some related singing voice language content modeling tasks like singer identification and language identification are not the goals of this thesis, because they can be, in principle, solved solely by signal processing methods, wherein the use of complementary context does not necessarily provide a clear advantage.

1.2.3 Why predominant singing voice?

Characterizing the lyrics content of singing when accompanying instruments are present is challenging. One of the reasons for this is that the audio spectrum is a mixture of many different sources, which for computers are not easily separable from each other.

This complexity is significantly mitigated in music traditions, which are centered around the singing voice, wherein the number of accompanying instruments is often small. That is why, being a largely vocal-centered tradition, OTMM provides a feasible context to validate the modeling developed in this study.

In addition to all the reasons listed above, a strong motivation to pursue this research is that, to our knowledge, this is the first work that designs a computational model of lyrics by considering (relatively) comprehensively the facets of its complementary context.

1.3 Opportunities and Challenges

Computational modeling of the singing voice has been focused to a large extent on transcribing the perceived pitch of the melody, leaving other musical facets, among which sung lyrics, less investigated. In the broad area of computational analysis of the language content of the singing voice, MIR researchers have explored tasks such as singing language identification, LAA, keyword spotting, lyrics transcription, which are well overviewed in (Goto, 2014). In total, however, there have been very few studies per each of these particular lyrics-related tasks.

1.3.1 Challenges of lyrics-to-audio alignment

The topics related to tracking sung lyrics in particular have been approached mostly by adopting the phonetic recognizer paradigm from speech recognition (Fujihara and Goto, 2012). The main idea is that for each phoneme a separate acoustic model is created, which describes the overall timbre of the phoneme (Rabiner and Juang, 1993). However, compared to speech, multi-instrumental music has several substantially different acoustic characteristics. Among them are the presence of accompanying instruments, the longer and more varying durations of vowels (Kruspe, 2014) and sections without singing voice.

1.3.2 Opportunities and challenges of analysing Makam music

In contrast to eurogenetic music, in OTMM the singing voice interacts with its accompanying instruments in a special way: Singers typically perform variations of a simultaneously played instrumental melody. This interaction is commonly referred to as *heterophony* (Cooke, 2013). As a consequence, the harmonics of the singing voice spectrum are interwoven with the harmonics from the spectrum of other instruments. Certain harmonics of the voice can overlap with the harmonics of accompanying instruments, and thus can be distorted by their energy. Therefore a model for lyrics tracking, based on the traditional phonetic timbre features could easily loose track in music with heterophonic voice-instrument interplay. For this reason, we expect that the use the context, complementary to phonetic timbre can provide the ‘stepping stones’ to the tracking of lyrics.

A benefit of the heterophony is that the main vocal melody is approximately doubled by some backing instruments. This has been used among other factors to ease to a certain extent the automatic extraction of the vocal melody contours in the recent work of (Atlı et al., 2014). Several vocal melodic temporal events, such as note onsets, vibrato, glissando are evident looking at the shape of the melodic contours. Therefore, ideally these events could be automatically ’read off’ the melodic contours, if these are reliably extracted.

1.3.3 Opportunities and challenges of analysing a specific music tradition

Modeling lyrics is coupled with the particular language of singing: the pronunciations of the phonemes of any language form an unique set of sounds. Therefore classical approaches on modeling speech are trained and tested on material from the same language. Being a relatively new research field, lyrics modeling follows to a large extent this paradigm. Switching to another target language in this sense would require the complete replacement of the lyrics model with one of the new target language. Building such a model might be a bottleneck, mainly because it depends on the availability of annotated speech/singing corpus (for complete justification see the Background chapter). This thesis, although focused on particular music traditions, aims at building an approach that is not restricted to one specific language. An important motivation for this are the similar characteristics of the traditions within the CompMusic project (in particular being vocal-centered), whereas language is one of the few differing aspects.

Characteristic for the singing in OTMM and jingju is that the sung vowels could be prolonged to a significant extent (Ederer, 2011; Wichmann, 1991), which makes it different from most eurogenetic musics. This lowers the quotient of consonants (a big portion of the language-specific sounds) from the total singing duration and thus mitigates their significance. This allows focusing on modeling of the acoustics of vowels, which makes it easier to adapt the constructed model of lyrics to another language.

When this dissertation was started, the Turkish was the only CompMusic tradition, for which an extensive collection of machine-readable musical scores was available. Music scores provide important contextual information complementary to lyrics, including but not limited to boundaries of structural sections, note durations and metric cycles. Exploiting the information in the musical score to its full extent is a major opportunity, in alignment with the goal of CompMusic to pursue a data-driven study on a music tradition.

1.4 Research Objectives

In alignment with the goals of CompMusic, the goal of this thesis is to build a culture-specific computational approach, which is meaningful for a concrete music repertoire. We have focused on OTMM due to the reasons listed above.

This thesis exploits computational approaches for analysis of music recordings. The approaches applied are taken from the fields of signal processing and machine learning. Signal processing is needed to extract the phonetic timbre of lyrics from the audio signal. The recorded audio is the primary source of information together with the given lyrics. Using complementary context, the proposed alignment models output words together with their aligned timestamps (Fig. 1.1). Two separate phonetic recognizers are created: One represents the influence of the expected syllable durations on phoneme transitions (Chapter 4). Another one represents the influence of the structure of the metrical cycle (Chapter 5). Depending on the nature of the complementary context, different additional data sources or domain knowledge are explored.

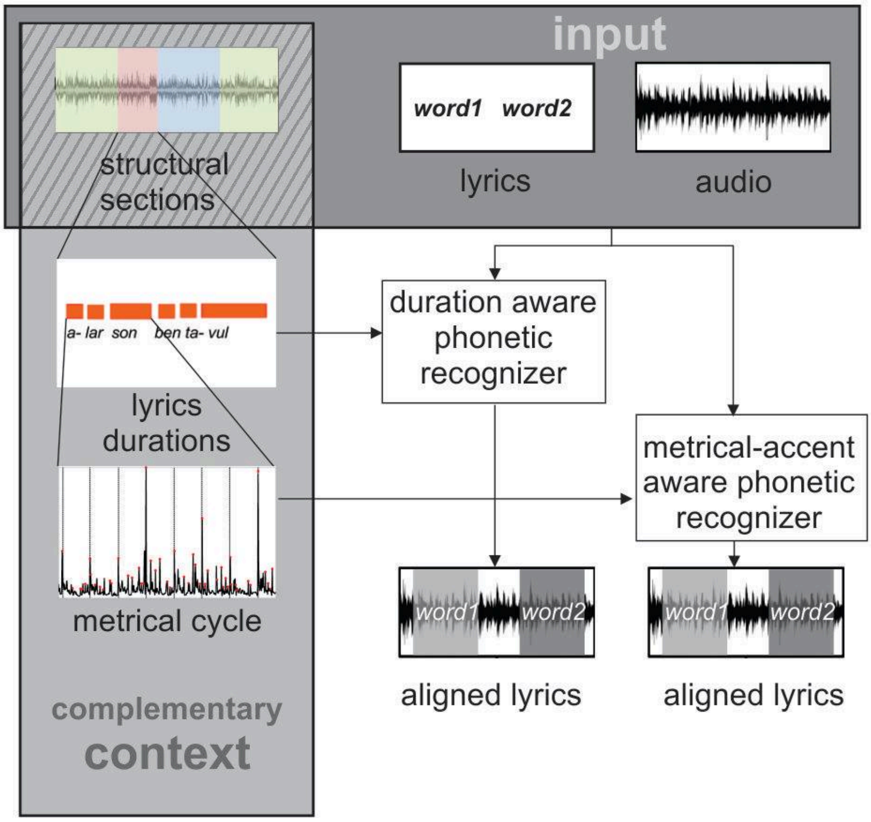


Figure 1.1: Use of different facets of complementary context in the automatic lyrics-to-audio alignment. Structural segmentation of a musical recording into lines of lyrics is considered a 'black box'. The audio signal of the obtained lyrics lines, along with its corresponding lyrics, is input to two separate phonetic recognizers. Both of them perform alignment of the audio signal to lyrics. Timestamps of aligned lyrics units are output.

The baseline method, on which we build upon, extends the HMM-s — a supervised learning method. It is preferred, because its probabilistic generative nature can describe adequately the temporal progression of the singing voice from a phoneme to another one (Rabiner and Juang, 1993).

1.4.1 Broad research objectives

Create a computational approach to describe transitions between sung lyrics that is aware of specific complementary context

The goal is to address those bits of knowledge from the complementary context, which have a clear influence on the phonetic timbre changes. The way music events evolve in time for a given music tradition can be expressed as a set of music principles. As a result of the work of musicologists, such principles specific to a music tradition have been aggregated in terms of concrete patterns and rules (Ederer, 2011; Wichmann, 1991). We aim to create a context-aware machine learning method of tracking sung lyrics, which benefits from the knowledge, compacted in these music patterns. The model has to jointly represent them and their influence on the transitions between consecutive units of lyrics. More precisely, such a joint model will allow the transitions of phoneme timbre be conditioned not only on the acoustic timbral features, but also on the simultaneously occurring complementary context events.

Probabilistic graphical models provide an effective framework to integrate complementary context knowledge in terms of the components of the model. In this thesis, we will extensively use DBN-s — a particular graphical model that can represent not only dependencies between concepts, but also their temporal progression (Murphy, 2002). The phonetic recognizer baseline provides a probabilistic framework (HMM), which allows to be extended to a DBN. We suggest a method that captures the influence on the lyrics transitions of each considered facet of complementary context. To this end, we represent events from complementary context as components in a DBN and their influence on the lyrics as a hierarchical dependence between the components.

The complementary contexts relevant for phonetic transitions, which we explore in this study, are:

- structure of the composition (coarse-level)
- lyrics durations (mid-level)
- structure of the metrical cycle (fine-level)

We do not aim to explicitly model the influence of the structure of the music composition on lyrics. Instead, the segmentation of a recording into its sections is obtained from an external method, which is considered as a 'black box'. Each obtained section contains one or more lyrics lines. Usually a lyrics line corresponds to a melodic phrase — a musically meaningful melodic entity, usually delimited by an instrumental break [5]. The audio signal of each ob-

[5] The term *melodic phrase* is used intentionally instead of a *melodic motif*, which usually stands for a short segment/pattern being a part of a complete melodic phrase. Melodic motif is for example used in this way in Gulati (2016)

tained section, along with its corresponding lyrics line, is input to the proposed phonetic recognizers (see Fig. 1.1). We aim at building a separate phonetic recognizer with mid-level context (Chapter 4) and a separate one with fine-level context (Chapter 5), each of which is a DBN. The mid-level one considers the influence of the temporal structure of a lyrics line on the transitions between consecutive syllables. In particular, we focus on the sequence of some reference durations of sung syllables. As to the fine-level context, we aim at studying how phoneme transitions interact with the position of the accents in the metrical cycle (i.e. the metrical accents). In an initial step we estimate the timestamps of the vocal note onsets (the initial time segments of sung tones), in a manner informed by the metrical accents. Then the goal is to represent how the transition to a consecutive sung syllable is conditioned on the transition to a consecutive note onset.

Since some of these complementary context relations to lyrics have not been previously strictly formalized in a computational study, a major effort of this thesis is conceptualizing them in terms of compact bits of probabilistic knowledge.

Develop a novel approach for lyrics-to-audio alignment

The proposed contextual models are designed with the intention to be generic enough and applicable in different end-tasks in the broader research area of sung lyrics. Having in mind the time limitation of this study, we focused on the particular task of LAA as a way to evaluate the performance of the proposed generic model. However, we expect that due to the ubiquity of the addressed facets of complementary context, the behavior of our model that we asses on alignment will be comparable on neighboring tasks including keyword spotting and lyrics recognition.

As a baseline for LAA we chose phonetic recognizer approach, adopted from speech-to-text alignment, based on HMM-s. They not only have proven to be the most successful strategy for the alignment of lyrics, but they also provide an appropriate temporal probabilistic framework, which we can extend for representing complementary context.

The alignment method, designed in this thesis, is evaluated mainly with singing in Turkish language. Nevertheless, to assure its application to other music genres we aim at devising ways for the easy transfer of the built models of Turkish phonemes to other languages. An ideal solution would be a universal language-independent model of a superset of phonemes representing a set of all languages of interest. Having in mind the reasonable differences between the languages in the CompMusic traditions, this is an elaborate linguistic task, outside the scope of this thesis. The approach commonly taken in existing work is rebuilding a complete model for each language in turn. Training models of phonemes in singing is in fact a laborious task (see Background Chapter) and in general not a flexible strategy. Instead, we set as a reasonable objective to find an adequate scheme for mapping the phoneme models among two different languages. To our knowledge, there has been no work so far in computational modeling of sung lyrics addressing the problem of inter-language phoneme mapping.

Evaluate the contribution of each piece of complementary context knowledge for modeling sung lyrics

Using LAA as a concrete end task allows evaluating the contribution of any particular facet of complementary context in a quantitative way and comparing them.

The novelty of the presented models is that their capability to integrate facets of complementary context into the main alignment step. Some of the context facets explored in this thesis have also been addressed in previous work (Fujihara and Goto, 2012). However their relation to phonetic timbre is not represented explicitly in the main alignment model. Instead, the knowledge from complementary context is isolated: part of a preprocessing or postprocessing step, relative to the main alignment step (see Background Chapter). On top of that, with the exception of Fujihara et al. (2011), almost no work has evaluated the contribution of these separate steps on the final alignment accuracy. To address this research 'vacuum', we compare the alignment accuracy for each different piece of complementary context and the baseline phonetic recognizer, agnostic to any complementary context.

Explore extensions and generalizations of the music-specific models to other traditions in the context of CompMusic

Working in tradition-specific context, there is a danger that the devised models become overfitted to the unique characteristics of the music tradition. To avoid that, the model should not reflect cases, unique for OTMM, but instead induce patterns that are applicable also to other music traditions with similar characteristics.

When a song is performed, the degree of deviation from the musical score is arguably the least, compared to other CompMusic traditions. In jingju, for example, the duration of sung syllables frequently deviates to a bigger extent from the score and could span a very long time interval. To proof the transferability of some of the proposed models outside of OTMM, we evaluate on material from another music tradition. We focused on a particular aspect of complementary context — the temporal structure of sung lyrics lines, for a particular tradition — jingju. Comparing the application of the syllable-duration aware model for two traditions also serves to quantitatively evaluate if a facet of complementary context contributes to a different degree for each of the two traditions (see Chapter 4).

1.4.2 Contributions

In pursuing the above presented goals we build methods, which can be seen as concrete technical and scientific contributions:

1. We extend the existing state of the art phonetic recognizer approach for tracking sung lyrics in a way that involves selected facets of complementary context knowledge. We conceptualize the interaction of phoneme transitions and these facets in a compact way as probabilistic dependencies. These dependences are represented as hidden variables in a DBN.

2. We suggest several implementation strategies for detection with the proposed DBN-s. In some cases the topology becomes relatively complex, because of, for example, the big number of hidden variables. This makes the inference with DBNs computationally demanding and thus model simplifications are required:

   a) integrate the complementary context knowledge in the inference method, instead of being hidden variable

   b) reduce the range of the state-space exploiting all available complementary context knowledge

   c) integrate the complementary context knowledge as a modification of the transition model

3. We develop a clean and modular software framework, which can be easily used to reproduce or extend the outcomes of the research, conducted in this thesis.

1.5 Outline

The thesis is organized into six chapters, wherein the main contributions are contained in Chapters 4 and 5. Chapter 2 covers the research background, summarizing the principles of the musics studied: OTMM and jingju. It also overviews the state of the art in the methodologies used in lyrics-to-audio alignment. A focus is put on describing the pipeline of a phonetic recognizer alignment approach. Finally, the chapter outlines related research on DBN-s — the main probabilistic model, used throughout the thesis. Chapter 3 presents our developed baseline system for lyrics to audio alignment, which is also based on a phonetic recognizer. Refinements in some of the recognizer steps, which makes it tailored to OTMM, are discussed. Chapter 4 describes the first core proposal of the thesis, a lyrics-to-audio alignment system that considers some context facets complementary to lyrics, in particular the sequence of reference durations of sung syllables. Chapter 5 presents a separate model for lyrics-to-audio alignment that considers another facet of complementary information, the accents in the metrical cycle of music. Finally, Chapter 6 concludes the thesis with a review of the key findings and a summary of the contributions.

Chapter 2

Background

In Section 2.1.1 we first summarize some of the principles of OTMM, the main music tradition analysed in this thesis, which influence 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 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.

2.1 Background on the music traditions

As *complementary context* in this thesis we defined the music events that occur simultaneously to and are complementary to the lyrics. OTMM and jingju, the music traditions studied, are characterized by well-defined theory and music principles. In this section we introduce the music traditions in general and exemplify in particular the principles of complementary context for each tradition in turn.

2.1.1 Ottoman Turkish makam music

The term *makam* describes a system of melodic scales used in numerous music traditions in Asia, north Africa and east Europe. Makam music is characterized by solid theory and modal principles. One of these traditions is Turkish classical/art music — the tradition, which proliferated in the Ottoman Empire and Turkey afterwards.

For a comprehensive introduction on the concepts of OTMM from a computational point of view, the interested reader is referred to Bozkurt et al. (2014) and (Şentürk, 2016, Section 2.1).

Principles of complementary context

Examples of complementary context principles can be organized by the levels of granularity, as we suggested in the Introduction Chapter.

Coarse-level: (structure of the composition) Vocal melodic phrases are organised in the course of the performance according to principles of the composition structure. The *şarkı* form (the principal form in *makam* with a lead vocal) adheres to a well-defined verse-refrain-like structure. A şarkı contains three vocal sections: zemin (verse), nakarat (refrain), meyan (second verse). They are preceded/surrounded by aranağme (an instrumental interlude) (Ederer, 2011). Each section is rather short and contains usually one (or 2-3) melodic phrases. In a vocal section through almost all its duration a singing voice is present, except for short instrumental interludes (at the end of a melodic phrase).

Mid-level: (lyrics durations) In this thesis we utilized information about musical note events from the music scores. For most of its existence, OTMM has been predominantly an oral tradition. However, since early 20th century, in parallel to the oral practice, music scores were introduced (Popescu-Judetz, 1996). The scores extend the traditional Western music notation and contain usually also the lyrics organized into sections. Karaosmanoğlu et al. (2014) prepared a machine-readable score collection, in which melodic phrases are annotated into smaller melodic units (motives). A melodic unit in this collection corresponds roughly to a metrical cycle.

Fine-level: (structure of the metrical cycle) The metric structure is explained by *usul*. A certain *usul* roughly defines the metrical cycle, and it can be described by a group of strokes with different velocities, which imply the beats and downbeats in the rhythmic pattern. Some of the common usuls include *düyek* with 8/8 time signature; *aksak* (9/8); curcuna (10/8). In contrast to the eurogenetic music, a metrical cycle can be rather long and have a complex rhythmic pattern with an odd number of beats. The number of pulses (finest metrical accents) in an usul cycle might be up to 120 (Ederer, 2011). The progression of the events in a vocal melody is correlated tightly to the underlying metric pulsation. For example, studies on symbolic music data showed that the likelihoods of vocal note events are influenced by the their position in a metrical cycle (Holzapfel, 2015).

Singing style

OTMM is predominantly a voice-centered tradition. This implies not only that singing voice is the source of predominant melody. It also entails that in performances the vocal melodies are rich in expressive embellishment. Embellishments of the melody is, in fact, a fundamental aesthetic aspect of the music. Singers typically perform variations of a simultaneous instrumental melody in their own register, a musical interaction commonly referred to as heterophony (Cooke, 2013). The melodies are embellished to a high extent, because this way singers can 'stand out' from the instrumental mix and evidence their virtuosity.

The melody contours of singing voice are not flat: skilled singers can control the variation of their voice's pitch to articulate expressive figures such as portamento, vibrato and melisma. Examples of singers very versed at that are Zeki Müren, Melihat Gülses, Kani Karaça. Vocal melodies have often a 'slow start' — the first tone is approached after a long vocal slide (portamento) (Ederer, 2011). Detecting the exact onset timestamp of vocal onsets is hard because of the 'slow start' effect. A further challenge is the ambiguity of note transitions — the transitions to another note are often 'enriched' by melismas.

Language

Unlike modern Turkish, Ottoman Turkish is characterized by more loanwords from Arabic and Persian origin. The lyrics language for the şarkı compositions in our evaluation dataset spans both modern and Ottoman Turkish. Turkish has 38 distinctive phoneme sounds, 8 of which are vowels. There are no diphthongs, and when vowels come together, they retain their individual sounding. Lengthening of vowels is realized by a non-pronounced character ğ. However vowel lengths have a negligible importance in sung Turkish.

2.1.2 Jingju

Jingju[1], also known as Beijing or Peking opera, is one of the around 300 different local genres of Chinese traditional theatre. It formed in Beijing during the 19th century as result of the combination of different local genres from south and west regions of China. Chinese traditional theatre is a comprehensive art form that encompasses disciplines as music, a special style of recitation, acting, dancing and acrobatics. The main dimension of its music component is singing, used for an expressive delivery of a passage of lyrics that can be equivalent to the concept of aria in opera.

Singing in jingju music has attracted the interest of MIR researchers during the recent years, who have studied topics like pitch contour analysis (Caro Repetto et al., 2015) or segmentation into syllables (Gong et al., 2017).

Principles of complementary context

Coarse-level: (structure of the composition) Lyrics in jingju are a central musical facet. Lyrics have poetic structure and are thus commonly organized into couplets. Each couplet has two lyrics lines and can be considered a structural section. The lyrics describe the story of the play and rarely repeat, even though some melodic motives could recur.

Meter is another musical facet that creates the impression of progression in the structure of an aria. Each aria can have one or more metrical pattern (*banshi*): it indicates the mood of the story and is correlated to tempo(Wichmann, 1991). When more than one banshi are present, an aria starts with a slow *banshi* which changes a couple of times to one with faster tempo. In this way the overall tempo of the aria increases gradually up to the fastest tempo to express more intense mood at the culmination point of the aria.

Mid-level: (lyrics durations) In contrast to OTMM, music scores in jingju serve a different purpose. Music in jingju originally was not created by composers, but arranged by performers from a repertory of pre-existing tunes, to fit new lyrics. Once a new play was in this way set to music, it would be transmitted orally from teacher to student by means of imitation. Scores would appear only a posteriori to register specific performances or to provide learning material to aficionados, since professionals rarely rely on scores. As a result, it is common that music scores and audio recordings of the same aria present many differences in many aspects, including note durations. This

[1] jingju literally means theatrical play from the capital (i.e. from Beijing)

is one of the reasons why machine readable music scores of jingju are rarely present.

In jingju a lyrics line (sentence) is usually divided into 3 syllable groupings, called *dou.* Interestingly, in jingju there exist some conventions of the durations of the *dou*-s, which serve as guidance to actors. A *dou* consists of 2 to 4 syllables (Wichmann, 1991, Chapter III). To emphasize the semantics of a phrase, or according to the plot, an actor has the option to sustain the vowel group of the *dou*'s final syllable. One (or in rare cases more) of the the vowels in the belly part can be prolonged substantially (in the order of 20 seconds). There is, however, no indication on which vowel is the prolonged one.

There are also some conventions about the number of *dous:* If a poetry line of the lyrics has 10 syllables, a rule of thumb is that it consists of 2 3-syllable dous, followed by a 4-syllable dou. Respectively, if a poetry line has 7 syllables, it is a rule of thumb that it consists of 2 2-syllable dous, followed by a 3-syllable dou. These rule-like relations present a clear example of some music-specific knowledge that could be probabilistically modeled in a lyrics tracking approach as a complementary source of information.

Language

The language used in jingju is based mainly in the Beijing dialect, but including some characteristics from southern dialects, coming from the different genres that formed this one. Chinese is a predominantly monosyllabic language, meaning that most of the lexical units (words) are conveyed by single syllables. On the other hand, Chinese script is logographic, meaning that each written character represents not a phonetic unit, but a lexical one. As a result, Chinese characters represent lexical units that are pronounced with single syllables. Therefore it makes naturally sense that LAA is evaluated on the syllable level. When referring to jingju we will use the term *syllable* as equivalent to one written character. As a theatrical genre, the delivery of the lyrics is fundamental in jingju performance, therefore mastering correct and clear pronunciation plays a central role in jingju training. With the aim of improved syllable's pronunciation, jingju actors traditionally divide the syllable into three constituent parts head (initial part), belly (middle part) and tail (final part) (Wichmann, 1991). The belly, the middle part, is the main vowel group of the syllable that could be a pure vowel, diphthong or triphthong. The head, not present in all syllables, is the preceding consonant or semi-consonant. Finally, the tail, also not present in all syllables, is the semivowel or final nasalization, following the belly. In this thesis, all Chinese characters are transliterated by the official romanization spelling system *pinyin* [2].

[2] *https://en.wikipedia.org/wiki/Pinyin*

2.2 Background on Lyrics-to-Audio Alignment

Although humans are very versed in making sense of the lyrics, sung in songs, for machines the task of automatically tracking lyrics is very challenging. LAA refers to the automatic synchronization between sung lyrics and their written representation. According to Fujihara and Goto one of the ultimate goals of research in sung lyrics is the automatic transcription of lyrics from a mixture of singing voice and accompaniment. Lyrics transcription (a.k.a. lyrics recognition) is the problem of finding where and what units of lyrics occur in the music signal. LAAs can be seen as a particular subproblem of lyrics recognition, in which the search space is limited: the sequence of lyrical units is known, leaving to find their timestamps (temporal locations). The recognition of ordinary speech in noisy environments itself started only recently to reach satisfactory results. Therefore, it is still not realistic to strive for reasonable results in automatic lyrics recognition. Despite the fact that there has been a few research attempts, none of them has succeeded in achieving satisfactory performance with instrumental-accompanied musical signals (Mesaros and Virtanen, 2010; McVicar et al., 2014). Still, approaching LAA one could gain precious insights, that could be useful as stepping stones to disentangling the riddle of lyrics recognition. LAA relates to lyrics recognition much in the same way speech-to-text alignment relates to speech recognition.

2.2.1 Evaluation metrics

The accuracy of alignment can be evaluated at different level of granularity, which depends on the application. In this sense the accuracy is measured in different level of entities, which we will refer to in what follows as *lyrical units*. A unit could be either phoneme, syllable, word, lyrics line/phrase, or complete lyrics paragraph/section. When generating subtitles for music videos, for instance, line- or phrase-level alignment might suffice. On the other hand, when precise alignment is required, as in the case of automatic generation of highlights for karaoke, syllable- or even phoneme-level alignment is required.

Being a rather under-researched problem, there has not been established a standard evaluation metric. There have been proposed several metrics, whereby each one has been used in only one or two works.

Average absolute error/deviation Initially utilized in Mesaros and Virtanen (2008), the absolute error e measures the time displacement between the actual timestamp t_i and its estimate $\hat{t}_i$ at the beginning and the end of each lyrical unit.

$$e = \frac{1}{N_k} \sum_{word\,i} (\hat{t}_i - t_i) \tag{2.1}$$

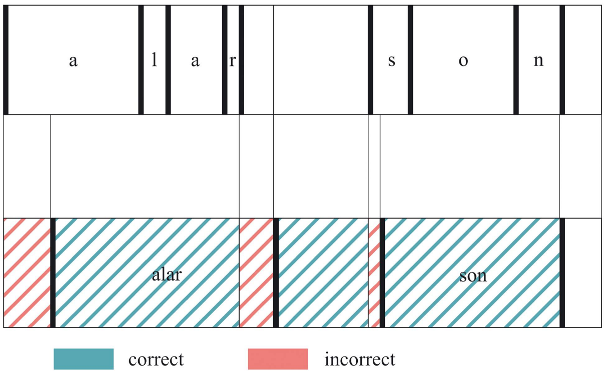


Figure 2.1: Evaluation by percentage of correct segments

The error is then averaged over all N_k words in the dataset. Evaluation was carried on timestamps at boundaries of lyrics lines. The authors themselves note that an error in absolute terms has the drawback that the perception of an error with the same duration can be different depending on the tempo of the song. The granularity of the lyrical units was refined in Mauch et al. (2012), where alignment was evaluated on the word level and further in Chang and Lee (2017) on the syllable level.

Percentage of correct segments The perceptual dependence on tempo is mitigated by measuring for a song k the percentage ρ^k of the total length of the correctly-labeled audio segments to the total length of the song — a metric, suggested by Fujihara et al. (2011, Figure 9):

$$\rho^k = \frac{length\ of\ correct\ segmetns}{total\ length\ of\ the\ song} \times 100 \tag{2.2}$$

Figure 2.1 illustrates the metric by an example.

The granularity on which the authors evaluated was lyric lines. This metric can be seen as a special case of the frame clustering metric for evaluating structural segmentation proposed in the work of Levy and Sandler (2008). This is essentially the same as the percentage of correct segments if we consider a lyrical unit acting as a “section”. Despite being rather unbiased by tempo and rather strict, the percentage of frames does not give a very intuitive estimate from a perceptual point of view, because the correlation to the extent of the absolute error is not obvious.

Percentage of correct estimates according to a tolerance window A metric that takes into consideration that displacements from ground truth below a certain threshold could be tolerated by human listeners, was suggested in Mauch et al. (2012). The authors evaluate the mean percentage of start time estimates $\hat{t_i}$ that fall within τ seconds of the start time t_i of the corresponding ground truth lyrics unit:

$$\rho_\tau^k = \frac{1}{N_k} \sum_{word\, i} 1_{|\hat{t_i} - t_i| < r} \times 100 \tag{2.3}$$

where N_k is the count of words in a given song k. The final metric is computed averaging ρ_τ^k over all songs.

In that particular work evaluation was carried out on the level of words, and τ was set to 1 second. Later in alignment was evaluated for both words and syllables. Further, the authors investigated more elaborately the influence of the its window τ, ranging tolerance values from 0 to 2 seconds.

2.2.2 Phonetic recognizer overview

The only existing overview of LAA approaches can be found in Fujihara and Goto (2012, Literature Review), which is a bit old but still encompasses the most seminal approaches up to date. Here we review only the approaches based on what the authors call a ‘phonetic recognizer’, because it is the alignment strategy, which has resulted in most promising results. The core machine learning model used in phonetic recognizers is HMM-s. They are suitably representing the time-changing nature of lyrics, because they can model time-contiguous, non-overlapping events.

The task of automatically converting spoken speech into text is known as automatic speech recognition (ASR) and has been one of the most well researched acoustic processing problems. One typical way speech recognition is approached is by building a model for each phoneme based on the characteristics of its timbral acoustics (Rabiner and Juang, 1993). The acoustic properties of spoken phonemes can be induced by the spectral envelope of speech.

An end-task, related to ASR is the automatic alignment between speech and its written transcript, also known as text-to-speech alignment. The classical approach of alignment is conducted by using the so called ‘forced alignment’ method: a transcribed piece of text is expanded to a network of phonetic models and matched to an audio recording of a speaker speaking this particular text. Each phonetic model represents the acoustic characteristics of the phoneme and is used to compare the likelihoods of feature vectors, extracted from the audio. A phonetic model is usually a HMM consisting of 1 up to 3 states representing the initial, middle and final acoustic state of a phoneme. The audio is aligned to the phonemes by finding the most likely path for the extracted sequence of feature vectors in the phoneme network (Rabiner and Juang, 1993).

When the vocal audio recordings are monophonic (commonly referred to as *a cappella*, too), LAA can be considered a special case of text-to-speech alignment, which is essentially solved (Anguera et al., 2014). Since the forced alignment technique was originally developed for clean speech, the presence of accompanying instruments and non-vocal sections pose a challenge to migrating it as-is to accompanied singing. Therefore, accompaniment attenuation and singing voice detection are steps that are commonly executed before the actual alignment. A LAA that is based on phonetic recognizer with forced alignment comprises a sequence of typical steps, presented in Figure 2.2.

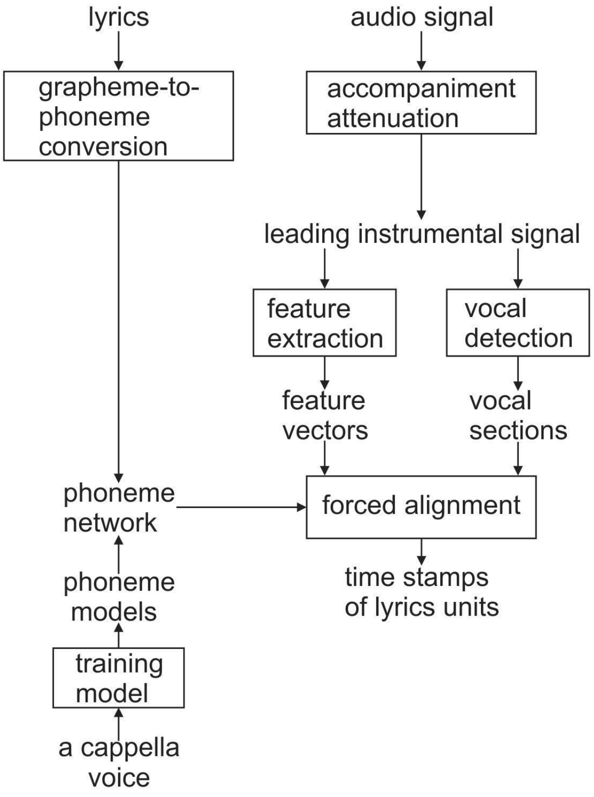


Figure 2.2: Typical steps of lyrics-to-audio phonetic recognizer approach

The goal of accompaniment attenuation (AA) is to isolate from the mixture signal the spectral content, which has its origin in singing voice, while attenuating the rest of the spectrum, with origin in accompanying instruments. Singing Voice Detection also known as Voice Activity Detection (VAD) has the aim to identify time intervals of the music signal, in which singing voice is present. Since AA and VAD can be considered separate problems on their own, in some related work existing prior methods are adopted.

Then, lyrics lines are expanded to a sequence of phonemes based on language-

Author	Features	Training approach	>1 language
Mesaros	MFCC	Speech + adaptation	N
Fujihara	MFCC	Speech + singer adaptation	Y
Kruspe	MFCC+PLP	Singing	N

Table 2.1: Seminal LAA works based on the phonetic recognizer approach. These are respectively: Mesaros, Mesaros and Virtanen (2008); Fujihara, Fujihara et al. (2011); Kruspe, (Kruspe, 2016)

specific grapheme-to-phoneme rules. In this way, the HMM-s are concatenated into a phoneme network. Phonetic models are trained on acoustic characteristics of material from either clean speech or *a cappella* singing.

In what follows we take a reviewing journey through the subsets of existing approaches from Table 2.1, staying some time at each of these steps and scrutinizing how some of the approaches address it. The approach of Kruspe (2016) is evaluated on *a cappella* singing, which excludes the need of both the AA and VAD steps.

2.2.3 Accompaniment attenuation

Compared to *a cappella*, the automatic alignment of lyrics in singing voice accompanied by various instruments is much more challenging. The phonetic models trained on features extracted from unaccompanied voice represent entirely the singing voice properties. In polyphonic mixtures of the voice and accompaniment, however, the vocal properties interfere with the instrumental sounds. Spectral peaks from harmonics of accompanying instruments may occlude the harmonic components of the voice. This means that some timbral characteristics, that are key to detecting the vowel identity, can be distorted. In this setting, phonetic models trained on *a cappella* voice lose their discriminative power. To address this problem researchers have come up with techniques that isolate as much as possible the spectral content, which has its origin in singing voice while attenuating the rest of the spectrum.

In Mesaros and Virtanen (2008) and Fujihara et al. (2011) a method for segregating the predominant melody is utilized: First the spectral components that are multiples of the fundamental frequency of the vocal melody (also known as harmonic partials) are extracted from the sound mixture. Then they are optionally refined and eventually grouped together to form the vocal signal. In the end, the vocal content is resynthesized from these by means of a sinusoidal model. At the core of representing singing voice content in polyphonic mixtures is a model capable of representing the complex interactions between the vocal harmonic partials and other instrumental sources in the mix. Several strategies of harmonic modeling have been proposed (Serra, 1989; Yeh and Röbel, 2009). A key challenge to such models is how to tackle partials from two different sources that have spectral overlap. Yeh and Röbel (2009) describe the expected amplitude of two overlapping partials based on the assumption that the partials overlap at the same frequency.

A drawback of the harmonic modeling presented above is that unvoiced consonant regions are not detected due to their lack of predominant pitch. Fujihara et al. (2011) suggest as a solution a method for fricative (unvoiced consonants) detection. In the alignment stage the time intervals for which the presence of fricative sounds is unlikely are forbidden to be matched to fricatives (actually only 'sh') from the phonemes network. A slight improvement in alignment accuracy was registered, supposedly because phoneme gaps in the middle of lyrics phrases were shorter than they were without fricative detection. However, since alignment accuracy was measured on lyrics phrase level, the effectiveness of the proposed fricative recognition method could not be fully evaluated.

The importance of the accompaniment attenuation method has been confirmed by comparing the alignment performance when disabling it (see Section 3.5). The phrase-level accuracy was improved by 4.8 absolute percent when MFCC were extracted from the vocal segregated signal compared to when extracted directly from the polyphonic mix (rows 3 and 4 of Table 3.6). Apart from that, the quality of the attenuation process can be objectively judged with the metrics used for evaluation of source separation on the segregated vocal (ideally vocal-only) signal. It is however hard to interpret how much the quality of attenuation affects the subsequent processing. To our knowledge no study has taken efforts in carefully examining the correlation between the degree of attenuation and the alignment accuracy, despite it being an important element in dealing with real-world accompanied singing.

2.2.4 Singing voice detection

In early LAA approaches (including the work of Mesaros and Virtanen (2008)) no automatic VAD method was applied. Instead the authors annotate manually structural sections (verse, chorus, bridge) with singing voice present. The sections' durations range from 9 to 40 seconds. The authors assume that in all vocal sections the predominant source is the voice. This permits to apply harmonic modeling, presented in the previous section 2.2.3, without the need of explicitly determining if the source of the main melody is voice. Short instrumental interludes are accommodated by training a model for instrumental accompaniment, which is expected to get activated in such interludes.

2.2.5 Acoustic Features

The timbre of singing voice is related to its harmonic partials — the timbral properties of a sung note depend on the distribution of the energy of its harmonic partials, whereby more energy is concentrated in harmonics around specific frequencies, commonly known as *formant frequencies*.

Formant frequencies

The *formant frequencies* represent resonances of the vocal tract and cavities (Ladefoged, 1996). Formants spectral regions, ordered according to their energy, with first formant (F1) representing the one with highest energy. Findings in phonology have indicated that the two lower order formants (F1-F2) are most important for understanding speech, whereas higher order ones (F3-F5) are related to the identity of the singer (Ladefoged, 1996). The first formant is known to change with the vocal tract shape (mainly by varying the jaw opening), while the second is correlated to the tongue shape. The vowels of speech are determined by specific combinations of F1 and F2, which are relatively stable for each vowel among different speakers.

Mel frequency cepstral coefficients

The MFCC are reliable descriptor of phonetic timbre. It is usually well captured by the first 12 coefficients and their differences to the preceding time instants (Rabiner and Juang, 1993). A commonly adopted variant of MFCC is the default configuration of the HMM toolkit (htk) (Young, 1993).

Ideally the efforts on reducing the influence of accompanying instruments can be mitigated by focusing on designing features that capture phonetic timbre in a way robust to background instruments. There has been some efforts recently to use end-to-end learning: for example encouraging results for singing voice detection were presented in Schlüter (2016). Hopefully, in the future insights from this approach can be adopted to recognizing not only if bits of spectral content originated from singing voice, but also its phonetic class. However, since no such features are yet designed, the working strategy for recognition of phonemes remains to extract features after the accompaniment has been reduced from the original polyphonic mix.

2.2.6 Introduction to Hidden Markov Models

Not only are HMM-s the main machine learning algorithm behind the phonetic recognizer approach, but they can be also considered a reduced case of DBN-s, in which only one hidden variable is present (see Section 2.3). We will give here a very brief overview of HMM-s and interpret them in the context of a phonetic recognizer.

They are probabilistic finite-state automata, where transitions between states $x_k \in 1, ..., S$ (S is the number of states) are ruled by probability functions. Let $x_{1:K} = \{x_1, ..., x_K\}$ be a sequence of hidden states with length K (number of audio frames in an audio excerpt). In speech research, traditionally a 3-state HMM for each phonemes is trained. It has left-to-right topology, which corresponds to how the acoustics of the voice evolve sequentially in time from an initial, through a middle, and ending in a final state. Transition probabilities are assumed to depend only on a finite number of previous states.

$$P(x_k|x_{k-1}, x_{k-2}, ...) = P(x_k|x_{k-1}) \tag{2.4}$$

This assumption is known as the Markov property, i.e. the current state directly depends only on a limited number of previous states (in this example only one). The term $P(x_k|x_{k-1})$ is known as the *transition model* between states, which can be expressed in a stochastic transition matrix (A_{ij}), where $a_{ij} = P(x_k = j \,|\, x_{k-1} = i)$ (Rabiner and Juang, 1993). Transition probabilities can be learned from annotated data or hand-crafted by imposing some musically-meaningful constraints. For example, when the target phoneme transcript is given, at inter-phoneme transitions, the network 'forces' only a single possible transition: to the following phoneme[3].

States (in our case the time phases of the phonemes) are not visible. Instead, one observes acoustic features (in our case the phonetic timbre), which are modeled as a sequence of random variables $y_{1:K}$. The featurey_k is assumed to depend exclusively on the current state, which can be represented in a probability distribution $P(y_k|x_k)$. In short, we refer to this quantity as the *acoustic model* or the *observation model*. It can be learned by maximizing the probability of emitting a given set of sequences of (observed) acoustic features from training data. Although traditionally modeled by GMM-s, the acoustic model could be virtually any machine learning model, which can output a continuous probability distribution.

Inference

Inference in probabilistic models refers to the process, in which we estimate the probability distribution of one or more unknown variables, given that we know the values of other variables. The joint probability distribution of the hidden and observed variables factorizes as:

$$P(x_{1:K}, y_{1:K}) = P(x_0)\Pi_{k=1}^{K} P(x_k|x_{k-1})P(y_k|x_k) \tag{2.5}$$

where $P(x_0)$ is the initial state distribution. The most likely hidden state sequence $x_{1:K}$ can be decoded, among others, by the Viterbi decoding — an efficient dynamic programming algorithm (Rabiner, 1989). Let $\delta_k(j)$ be the probability for the path in the state space with highest probability, among all

[3] This is the reason why it is called 'forced' alignment

paths, which end in state j at time k. The $\delta_k(j)$ is defined recursively in a maximization step:

$$\delta_k(j) = \max_{i \in (j, j-1)} \delta_{k-1}(i)\, a_{ij}\, b_j(O_k) \tag{2.6}$$

Here $b_j(O_k) = P(y_k = O_k \,|\, x_k = j)$ for feature vector O_k (complying with the notation of Rabiner (1989, III. B). Note that in the case of forced alignment we maximize only over two possible transitions — from the current state j and its previous one $j-1$.

A complete discussion on theory and applications of HMM-s can be found in Rabiner and Juang (1993).

2.2.7 Phoneme network

The goal of the grapheme-to-phoneme conversion is to create a phoneme sequence out of the input lyrics. The conversion is carried out using a set of phonemes from a phonetic alphabet, based on a pronunciation dictionary prepared by linguists.

In the phonetic recognizer approach, it is assumed that the observed feature sequence is generated from an HMM. The phoneme network is a super-HMM, concatenated from the individual phoneme HMM-s in the order of the input phoneme sequence. The transition model imposes the 'forced' transition to the consequent phoneme. The only exception are special case phonemes for short silent pauses, which can be optionally skipped. Most LAA approaches adopt this paradigm: Both Mesaros and Virtanen (2008) and Fujihara et al. (2011) utilized the 3-state HMM-s and trained for each state a GMM fitted on a feature vector y_k of MFCC.

Cross-language modeling

As a rule of thumb the phoneme models used in the recognition are trained from the same target language to ensure their integrity. However, often there might not be enough training material for the language of interest, which opens a necessity for finding a cross-language phoneme mapping strategy as an alternative. As a matter of fact cross-language mapping has been an important research direction in speech recognition for long time, but only recently some substantial results were achieved (Sun et al., 2016) (for the particular task of speech synthesis). One of the few LAA research works using phonemes trained on a different language was done by Fujihara et al. (2011). To align English songs the authors mapped English phonemes to their closest in sound entries from a set of Japanese phoneme models. This resulted in suboptimal alignment results though, due to the different language phonetics. In Japanese all vowels are pure (i.e. monophthongs), which is a clear limitation for the more complex acoustic characteristics of English diphthongs.

2.2.8 Training procedure

In the absence of enough singing material with annotated phonemes, the acoustic model $P(y_k|x_k)$ is trained on a big corpus of speech with annotated sentences. Later these phonetic speech models are adapted to match the acoustic characteristics of clean singing voice using a small singing dataset with annotated phonemes. The adaptation techniques are borrowed from research carried on adapting universal speech models to characteristics of a particular speaker.

Training on speech

Compared to speech, the singing voice has more complex frequency and dynamic characteristics: fluctuation of fundamental frequency (F0) and loudness of singing voices are far stronger than those of speech sounds (Sundberg and Rossing, 1990). The fundamental frequency of women in speech is between 165 and 200 Hz, while in singing it can reach 1000 Hz. This is much higher than the normal for speech value range of the first formant (500 Hz). In such cases the first formant moves higher in frequency, so that it corresponds approximately to the fundamental frequency, while the second formant might also move higher. Therefore the first two formants of singing voice are less stable than speech and harder to predict. In addition, some skillful singers are capable of changing drastically their position by moving their vocal cavity, tongue and lips. On top of that, compared to speech, some phenomena including vibrato and singer's formant are present only in singing. To address all these discrepancies an adaptation of the acoustic properties of spoken phoneme models is needed.

Mesaros and Virtanen (2008) proposed to borrow a technique from speech recognition that adapts an universal speech model to the speech for a particular speaker. They used the method Maximum Likelihood Linear Regression (MLLR). In Fujihara et al. (2011) after applying a MLLR, another statistical adaptation technique, the Maximum a posteriori (MAP) transform, was run. MAP shifts the mean and variance components of the Gaussians of the each spoken phoneme model in an acoustic space towards the characteristics of the corresponding sung phoneme. An advantage of the MAP transform compared to other adaptation techniques is that it allows the manipulation of each phoneme model independently.

Training on singing voice

Another fundamental difference between speech and singing voice is that the time a vowel is held in singing is much longer and much more variable than in speech. In a recent study Kruspe (2015b) compared the accuracy of recognition of individual phonemes with model trained on speech and a model trained on the same speech modified with 'sing-like' transformations: In turn pitch shifting, time-stretching and vibrato addition were applied on the same data. The author obtained 18% correctly classified audio frames with the model with all three modifications jointly, improving from the baseline of 12% with the speech model. Furthermore, result showed that a significant accuracy improvement was observed mainly due to time-stretching. The adaptation strategy presented above might compensate to a certain extent for most of the acoustic difference, except arguably for the variation of phoneme durations. One reason might be that when sung vowels are prolonged their transitions to neighboring phonemes have more variability than in speech, too.

A bottleneck for training on actual singing is the lack of singing material with phoneme annotations. Kruspe (2016) proposed a viable strategy for annotation: they trained on a speech corpus monophone one-state HMM-s, wherein each observation model is a GMM. Then the author preselected around 6000 recordings of full songs from the DAMP dataset from Stanford University[4]. DAMP is a huge collection of *a cappella* popular music, sung by amateur singers with lyrics available, but without any annotated word locations. The authors aligned the *a cappella* audio on the phoneme level to its lyrics by means of forced alignment with the fitted speech-trained GMM-s. The aligned phoneme timestamps were fed into a 3-hidden-layer Multi-Layer Perceptron (MLP) with sigmoid activation function, as if they were manual annotations. On material from DAMP the resulting model reached a phoneme recognition of 25% (better than the previous state of the art of Hansen (2012)) of correctly classified frames. This is a remarkable improvement over the 12% with a model trained only on the speech dataset. Results on the word-level alignment were however not reported.

In summary, there has been only a few research works on the problem of LAA. The phonetic recognizer was established as one of the successful alignment strategies. Solutions were proposed for all steps necessary to carry out a complete alignment: including all pre- and post- processing steps. Still, looking back at Table 2.1, some research questions remain open or not fully exploited:

1. Almost all of the presented approaches is trained on material from the language, on which it was tested. This means each time an aligner is

[4] https://ccrma.stanford.edu/damp/

required for a language, different from the one of existing work (for example Turkish), reuse of the existing aligner as a baseline is not straightforward. At least this is not feasible without a modification/adaptation of the acoustic model. This problem is further aggravated by the lack of singing material with reliable phoneme annotations, to use as training material[5].

2. In most approaches the extracted acoustic features (usually MFCC) are agglomerated into classes of phonetic timbre in a bottom-up fashion, without considering the dependence of simultaneously occurring rhythmic musical events or reference durations of syllables. Although phonetic timbre is the core distinguishing facet of sung lyrics, relying solely on the MFCC can be error-prone. They are usually trained on *a cappella* material and extracted (likely with artifacts) from the vocal part of multi-instrumental recordings, which is premise for acoustic mismatch.

3. There is no existing approach on LAA with instrumental accompaniment, which can be reproduced. None of the papers discussed have their implementation available. After personal communication with some of the authors we were able to obtain several pieces of source code, but for several reasons, none of the LAA systems presented in this section is reproducible in its entirety[6].

[5] In fact, to our knowledge, the biggest and only dataset with phoneme annotations is of English pop songs with total duration of less than 30 minutes — the one prepared and used in Hansen (2012)

[6] Personal communication with H. Fujihara in March 2014; with A. Mesaros in October 2013

2.3 Background on dynamic Bayesian networks

A probabilistic graphical model is a probabilistic model that expresses conditional dependence between random variables using a graph. HMM-s can be considered a probabilistic graphical model with a single hidden random variable. A Bayesian network is a probabilistic graphical model that represents a set of random variables and their (conditional) dependencies with a directed acyclic graph.

A DBN is an extension of a Bayesian network that can relate variables over time (Murphy, 2002). In a DBN variables could be either continuous or discrete, which we represent in all diagrams in this dissertation by circles and squares respectively. To build a model of sung lyrics we have at our disposal sequential data from audio features, as well as complementary context events that are interrelated to phonetic timbre. DBN-s hence provide an effective and explicit way to encode the dependence relationships between the phonetic timbre and these context events. Excellent resources on graphical probabilistic models and inference, in general, is Koller and Friedman (2009) and for Bayesian models in time, in particular, is (Barber et al., 2011).

Research by Whiteley et al. (2006) introduced DBN-s to music processing. The authors emphasize the fact that they can natively model higher-level musical facets more intuitively and efficiently than an HMM**.**

2.3.1 Inference in DBNs

To execute inference, one has to obtain the distribution over the required set of hidden variables, by marginalizing over the rest of the variables. This can be achieved by direct marginalization, variable elimination and/or other techniques (D'Ambrosio, 1999). However, in practice this can be complex and without closed form solutions. Therefore, in this dissertation we take a viable workaround, by reducing the proposed DBN-s, without losing their encoded dependence between musical phenomena, to HMM-s and resort to the Viterbi decoding (Section 2.2.6).

2.4 Background on sung lyrics with complementary context

In what follows we review existing studies on sung lyrics, in which knowledge from complementary context contributed to improvement in accuracy. We have organised these studies, according to the levels of granularity of complementary context, to which we comply in this thesis. As some approaches can be considered to benefit from more than one level, we do not aim at strict formal subdivision, but rather at laying out the background in a structured way, which we can start off extending systematically in the following chapters.

2.4.1 Coarse-level context

An experimental evaluation of the relation of structure and sung lyrics remains outside the scope of this study. Instead, we utilize automatic segmentation of complete song recording into its structural segments as a preprocessing step to LAA. However, in a future work, it is desirable to incorporate structural information into the phonetic recognizers, proposed in this thesis.

The use of music structural information has provided guidance for alignments on the higher-level in previous works (Lee and Cremer, 2008; Wang et al., 2004). Lee and Cremer (2008) showed that the results of rough structure segmentation can be used for paragraph-level alignment of lyrics. First, a structural segmentation of the audio recording is performed using acoustic features. Then the chorus section is determined by a clustering method, whereas the vocal ones are determined by a VAD method. The resulting sections are aligned to the hand-labeled lyrics paragraphs by means of dynamic programming.

2.4.2 Mid-level context

Musical chords are a piece of complementary context parallel to lyrics in the granularity of lyrics lines. Mauch et al. (2012) proposed the integration of textual chord information into the baseline phonetic recognizer approach of Fujihara et al. (2011), which we described above. The authors assume that the complete chord annotation is provided together with lyrics in the format of song sheets, which can be obtained from web-sites such as *UltimateGuitar*. The song sheet provides chord annotations anchored to words. To handle the ambiguous mapping of the word-level annotation to the finer-level of syllables, Mauch et al. suggest 'a flexible chord onset' strategy: To allow a chord change in any of the syllable of its corresponding word, for each syllable alternative paths are constructed in the syllable-HMM network . The syllable-HMM-network can be unambiguously expanded to a phoneme-HMM-network.

In this setting, since the phoneme sequence is fixed, a hidden phoneme state h determines several possibilities with equal likelihood for a hidden chord state c, which can be represented as a DBN. The combined transition probability is 'inherited' from the trained phoneme transitions. In addition to the baseline phoneme emission y^m an emission feature y^c for chroma is added. Both are combined into one mutual observation probability on inference. The approach greatly improves the word-level accuracy of the baseline, from 46.4% to 87.5 % in terms of the percentage of correct estimates according to a tolerance window of 1 second.

Chang and Lee (2017) described a method to deal with both syllable- and word-level lyrics-to audio alignments of accompanied music recordings in Korean and English. The approach is to discover repetitive acoustic patterns of vowels in the target audio by referencing vowel patterns appearing in lyrics.

2.4.3 Fine-level context

Few works for tracking lyrics in singing voice have proposed a method that represent features, describing phoneme timbre jointly with other melodic characteristics (Fujihara et al., 2009).

Fujihara et al. (2009) concurrently estimate the phoneme classes and fundamental frequency of singing voice from recordings with instrumental accompaniment. They suggest the use of probabilistic spectral templates of singing voice to represent both phoneme identity and the predominant f0. No temporal progression from one template to the next is modeled though. An important advantage of the approach is that the templates can be trained directly from the polyphonic mix without segregating the predominant voice or affecting the instrumental accompaniment, which is often a necessity in other studies. Accuracy for phoneme estimation is evaluated in terms of the ratio of the number of frames that are correctly estimated to the total number of frames. Frames taken into consideration in this calculation were only the five Japanese vowels a,e,i,o and u. The ratio of 55 % for a baseline with GMM-s and MFCC was increased to 60.1 % with the proposed model, which is arguably the best vowel recognition system in accompanied singing.

In (Korzeniowski, 2011) a hidden state space is proposed that combines the typical 3-state left-to-right HMM-s for phonemes with the note state space introduced in Orio and Déchelle (2001): each note has 3 states corresponding to its temporal phases attack, sustain and release. The goal of the study is to improve automatic score-to-audio alignment by integrating information from the lyrics timbre, available in parallel to the score. However, due to the very large state space, result of the the Cartesian combination of the note and phoneme state space, the authors were not able to implement this strategy. Instead they used the note HMM and incorporated vowel information as additional feature (together with pitch, loudness, etc.) via the observation probabilities of the states.

Summarizing, almost none of the related work that considers complementary context is based on a temporal modeling framework (such is the phonetic recognizer). The only exception is the approach of Mauch et al. (2012), which is however limited to music traditions for which the concept of chords is applicable. Due to the heterophonic interaction of accompaniment instruments with singing voice for traditions like OTMM the harmony does not occur in the form of chords and we cannot benefit from that work.

Typical exploited knowledge about temporal musical facets, complementary to phonetic timbre, is the one of structural sections or the interaction of the vocal with the fundamental frequency. However, this knowledge is modeled outside the main alignment step. For example, the events of transition of a structural section to the next one are used in a preprocessing step (Lee and Cremer, 2008).

Chapter 3

Baseline Lyrics-to-audio Alignment Model

3.1 Introduction

In this chapter we describe our LAA baseline system. It is a phonetic recognizer, based on phoneme HMM-s. 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 methods that consider context — the main contribution of this thesis. They will be the focus of the following two chapters. In this chapter we go through the key steps of a phonetic recognizer and describe which existing methods 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, whereby we comment on tuning its parameters. We present in more detail 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 the 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. Compiling datasets, representative of the music tradition and the key facets of complementary context, is an important effort of this study.

We start the chapter by describing the evaluation datasets, comprising both *a cappella* and multi-instrumental recordings from OTMM (Section 3.2). We then introduce our choices for each of the steps of the standard phonetic recognizer in Section 3.3. We describe the construction of the phoneme network in Section 3.3.2. In Section 3.4, we present a comparison of three strategies to train the acoustic model for Turkish language. Finally, in Section 3.5 we discuss the alignment results by evaluating the baseline model on the presented datasets.

3.2 Datasets

In this thesis we evaluate the proposed lyrics tracking approaches on datasets of selected recordings from OTMM and jingju repertoire. To this end we prepared two datasets: *multi-instrumental lyrics OTMM dataset,* which encompasses original studio recordings with accompaniment of multiple instruments, and an *a cappella lyrics OTMM dataset,* which contains solo signing voice. Additionally, we compiled a *multi-instrumental vocal onsets OTMM dataset* with annotations of vocal note onsets containing performances with well-perceived percussive accents. In all datasets we payed special attention to annotating carefully the timestamps of the music events, in which complementary context manifests.

3.2.1 *Multi-instrumental lyrics OTMM dataset*

The *multi-instrumental lyrics OTMM dataset,* which we compiled, consists of 13 performances with a soloing singer — 5 with male and 8 with female one. The performances are from 11 compositions in the şarkı form and have total duration of 19 minutes. They are drawn from the *CompMusic* corpus of OTMM repertoire (Uyar et al., 2014) and have varying recording quality, including historic recordings. Some these are not necessarily with good studio quality. Music scores are provided in a custom machine-readable format, called *symbTr,* complying with the *humdrum* notation philosophy (Karaosmanoğlu, 2012). These scores contain annotations of the structural sections of the şarkı form. The words in a section are further split adopting the division into musical phrases, proposed by Karaosmanoğlu et al. (2014). What the authors call a musical phrase represents a musically-meaningful melodic motif. A phrase spans roughly the same number of metrical cycles depending on the tempo (1 or 2 cycles). This corresponds to up to 4 words depending on their length. A musical phrase often also contains short instrumental motives before or after the vocal is present. If an original phrase boundary splits a word we have modified it to include the complete word, in order to assure appropriate evaluation on word or phrase level. Table 3.1 presents statistics about the derived phrases of lyrics. The total number of words in the dataset are 732.

The performance recordings contain the annotations of the boundaries of segments corresponding to the score sections, which have been done in the study

total #sections	#phrases per section	#words per phrase
75	2 to 5	1 to 4

Table 3.1: Phrase and section statistics for the *multi-instrumental lyrics OTMM dataset*

of Şentürk et al. (2014). We annotated further the musical phrase boundaries using the *Praat* annotation tool[1]. Whenever needed, we split or merged some lyrics phrases with outlier duration so that phrases within a recording have approximately equal duration[2].

3.2.2 *A cappella lyrics OTMM dataset*

Due to the lack of appropriate *a cappella* material in the şarkı form, we recorded especially for this study an *a cappella* version of the *accompanied vocal OTMM dataset*.

The vocal parts of the *multi-instrumental lyrics OTMM dataset* have been sung by professional singers, especially recorded for this study. A performance has been recorded while listening to the original recording, whereby instrumental sections are left as silence. This assures that the order, in which sections are performed, is kept the same. Therefore, the generated timestamps are valid for the accompanied version, too. Although each recorded singer sings sporadically off-time at some syllables, the recordings are to a very high degree in-sync with the originals. We carefully validated that by listening simultaneously to both the original and the *a cappella* version[3].

Additionally, the singing voice for 6 recordings (with a total duration of around 10 minutes) from the dataset has been manually transcribed with notes, inferred by the music score. A special care is taken to place the onset annotation on the time instant, where a voiced sound starts. In this way, an onset is considered to be always at the beginning of a portamento, when it is present (it is common for some singers)[4]. Similarly, if a syllable starts with an unvoiced consonant, the onset is placed at the beginning of the succeeding vowel (see Figure 5.3).

[1] http://www.fon.hum.uva.nl/praat/

[2] The dataset is available at http://compmusic.upf.edu/turkish-sarki

[3] The audio and the annotations are available under a CC license at http://compmusic.upf.edu/turkish-makam-acapella-sections-dataset

[4] Onset annotations are available at http://compmusic.upf.edu/node/233

#sentences per aria	9.2
#syllables per sentence	10.7
avrg sentence duration (sec)	18.3
avrg syllable duration (sec)	2.4

Table 3.2: Sentence and syllable statistics for the jingju dataset

3.2.3 *Multi-instrumental vocal onsets OTMM dataset*

Unlike the previous two datasets, being designed for LAA, we compiled *the multi-instrumental vocal onsets OTMM dataset* to be used for note onset detection of singing voice. We utilize it for automatic note onset detection, informed by underlying metrical accents. To that end, all recordings have clearly audible percussive strokes, at some of the beats in a metrical cycle. The dataset includes two meter types, referred to as usuls in Turkish makam: the 9/8-usul aksak and the 8/8-usul düyek. It is a subset of the dataset presented in Holzapfel et al. (2014), including only the recordings with singing voice present. The beats and downbeats were annotated by Holzapfel et al. (2014). The vocal note onsets are annotated by a single annotator, whereby only pitched onsets are considered (2100 onsets). To this end, we had the same strategy for annotation onsets as in the *a cappella lyrics OTMM dataset*[5]. Unlike it, however, we used as guidance the annotated beats — being aware of the location of a beat helped to put more precisely the location of an onset. Annotations were done as different layers in *Sonic Visualiser*[6].

3.2.4 *A cappella lyrics jingju dataset*

The dataset has been especially compiled for this study and consists of excerpts from 15 arias, chosen from the *CompMusic* corpus of jingju arias, compiled by Caro Repetto and Serra (2014). It has total duration of 67 minutes and comprises two female singers. For a given aria were present two versions: a recording with voice plus accompaniment and an accompaniment-only one. From these, we generated *a cappella* singing by subtracting manually the instrumental accompaniment from the complete version[7]. Table 3.2 presents the average values per sentence and syllable.

Each aria is annotated on different event granularities: from the *banshi* type, through boundaries of lyrics sentences, down to boundaries of syllables and boundaries of phonemes. Annotations are carefully done by native Chinese

[5] The dataset is available at http://compmusic.upf.edu/node/345

[6] http://www.sonicvisualiser.org/

[7] The resulting monophonic singing is perceived as clean as if it were a cappella, having slightly audible artifacts from percussion on the non-vocal regions

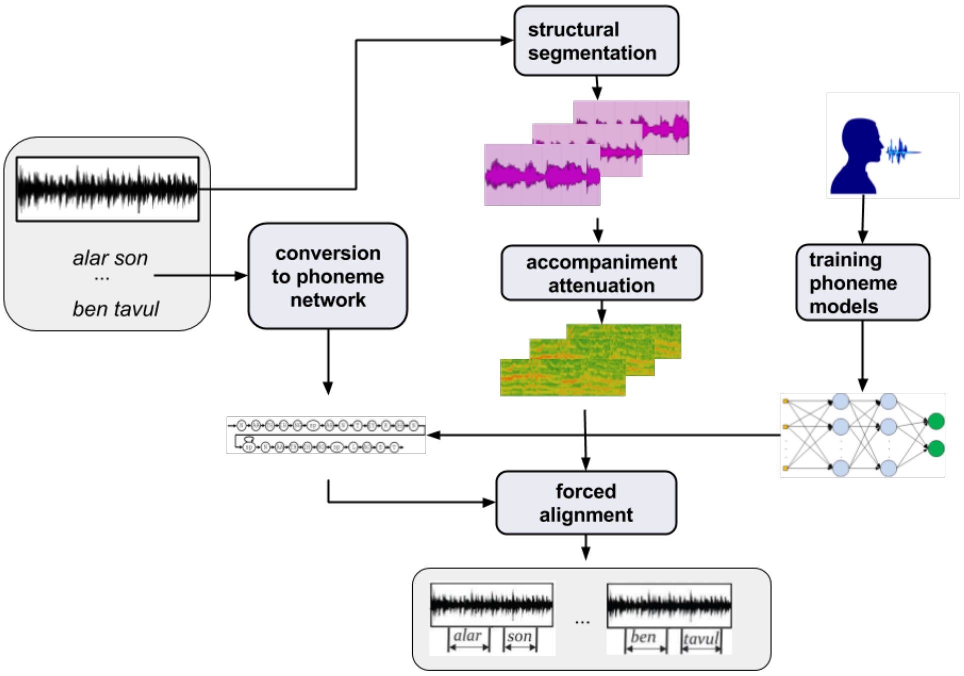


Figure 3.1: Overview of the steps of the baseline lyrics-to-audio alignment system

speakers and a jingju opera musicologist[8]. The phoneme set has 29 phonemes and is derived from Chinese pinyin, and represented using the x-sampa standard[9]. To assure enough training data for each model, certain underrepresented phonemes are grouped into phonetic classes, based on their perceptual similarity.

3.3 Phonetic recognizer

An overview of the steps of the proposed approach can be seen in Figure 3.1. These steps follows some of the the typical steps of existing phonetic recognizer approaches (presented in Fig. 2.2 of the Background Chapter). In what follows we discuss in detail the design choices and the preferred solutions for each step.

[8] These are the *CompMusic* team members Rong Gong, Yile Yang and Rafael Caro Repetto

[9] Annotations are made available at http://compmusic.upf.edu/node/286

3.3.1 Structural segmentation

Being a challenging problem itself, a full-fledged VAD is outside the scope of this study. We instead divided manually each audio recording into sections (e.g. zemin, nakarat, meyan) as indicated in the music score, whereby instrumental-only sections were discarded. In the şarkı form each vocal segment corresponds to a structural section (zemin, nakarat, or meyan). We assign manually to each segmented vocal section its corresponding lyrical line, in order to assure correct lyrics.

All alignment throughout this thesis is performed on an audio recording and text for each vocal section separately. LAA on complete audio recordings was not desirable due to the unpredictability of the sections order in a şarkı form. The sections are often performed in an order differing from that indicated in the score. On top of that, improvisation sections not present in the score are commonly inserted (Popescu-Judetz, 1996).

To verify the feasibility of automating the structural segmentation, we utilized a method for linking score sections to their beginning and ending timestamps in a recording with Makam singing (Şentürk et al., 2014). Due to the high accuracy of this method, almost all sections are mapped correctly with minor section boundary displacements. We showed that integrating section linking as a preprocessing step yields estimated section boundaries that are not detrimental to matching the correct lyrics sections (Dzhambazov et al., 2014).

3.3.2 Accompaniment attenuation

It is difficult to successfully track the phonemes in multi-instrumental music signals by using the models, trained solely on *a cappella* singing. The harmonic partials in unaccompanied singing can be extracted relatively reliably, mainly because they form clear intensity peaks in the spectrogram **(Serra and Smith, 1990)**. A simple intensity-peak-picking strategy is however prone to failure in accompanied singing, because of the interference with instrumental harmonic partials. To handle this case many harmonic partial detection methods were proposed (see Section 2.2.3).

Such a method for the detection of vocal harmonic partials requires a melody contour as an input, being generated by a melodic source. We first extract the vocal contour of the singing voice. Then, based on it, its harmonic partials are derived from the spectrum Y at a given time frame. Then the vocal harmonic partials are resynthesized into an interpolated vocal spectrum Yh. Finally, we extract acoustic features from Yh instead of the original polyphonic spectrum Y.

Singing voice melody extraction

To extract the contour of the predominant singing voice in music with instrumental accompaniment, we utilized the algorithm described in Atlı et al. (2014). It is a method for the extraction of the melody of a predominant instrument. It relies on the basic methodology of Salamon and Gómez (2012), but modifies the way in which the final melody contour is selected from a set of candidate contours, in order to reflect the specificities of OTMM:

1. It chooses a finer bin resolution of only 7.5 cents that approximately corresponds to the smallest noticeable change in Makam melodic scales.
2. Unlike the original methodology, it does not discard time intervals where the peaks of the pitch contours have relatively low magnitude. This accommodates time intervals at the end of the melodies, where Makam singers might sing softer.

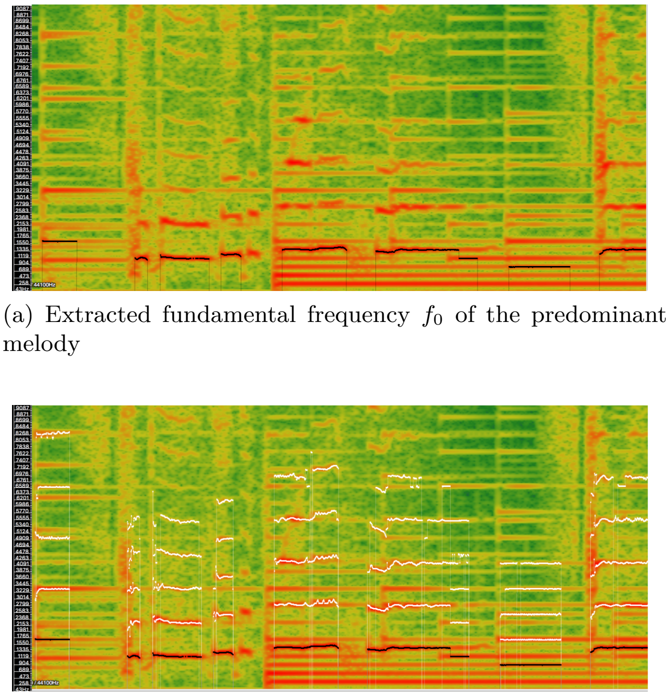
In addition to generating fundamental frequency values (f_0) values, the algorithm performs in the same time a predominant source detection: it returns zero values forf_0 in regions with no predominant melody. The melody contour obtained this way has its origin not only from singing voice but also from accompanying instruments. This happens in short instrumental interludes, where an accompanying instrument carries the main melody.

Harmonic model

We utilized the harmonic model of Serra and Smith (1990) to filter the spectral peaks corresponding to the harmonic partials of the singing voice. The spectral peaks are computed at the expected location of harmonic partials at multiples of the normalized fundamental frequency $\hat{f}_0$. Equation 3.1 represents a spectral bin k as the sum of R harmonic partials from the spectrum of the analysis window W, weighted by their corresponding amplitudes A_r Serra (2016). Parabolic interpolation refines the exact frequency locations. We estimated Yh with a relatively large number of harmonics ($R = 30$), in order to preserve as much as possible the phonetic timbre.

$$Yh[k] = \sum_{r=1}^{R} A_r W[k - r\hat{f}_0] \tag{3.1}$$

It should be noted that it is not an end goal of this study to have the best possible segregation of the singing voice from the polyphonic mix. Segregation methods strive to obtain a representation of the vocal content with the



(a) Extracted fundamental frequency f_0 of the predominant melody

(b) Detected harmonic partials (R=4) with the harmonic model, based on the fundamental frequency f_0 of the predominant melody

Figure 3.2: An example of extracting harmonic partials of predominant voice with the harmonic model

least amount of introduced artifacts. In contrast to that, in our case some artifacts may be acceptable as long as they do not distort significantly the intelligibility of vowels. As a benchmark, we carried out a study, in which we evaluated the quality of voice segregation using the harmonic model (Dzhambazov and Serra, 2016). Results in terms of the common source separation metrics showed that for pop music the harmonic model is inferior to recent separation methods based on convolutional neural networks, like for example the method of (Chandna et al., 2017)[10]. However, a shortcoming of convolutional neural networks is the necessity of a big amount of clean singing voice training data, which was not available for OTMM.

Resynthesis

The interpolated vocal harmonic partials are resynthesized by means of a constant overlap add resynthesis with the *sms-tools* package[11]. Despite being

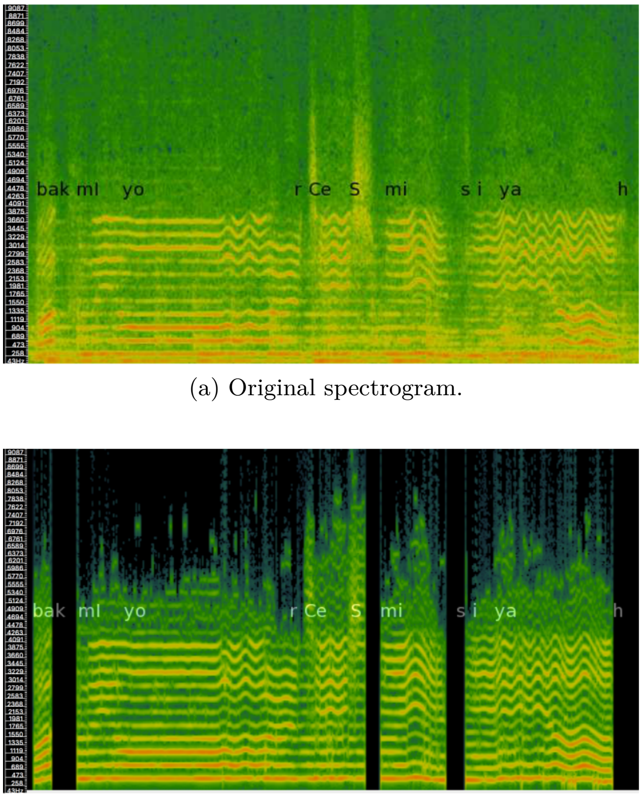
[10] results on the MIREX 2016 task on singing voice separation are available at http://www.music-ir.org/mirex/wiki/2016:Singing_Voice_Separation_Results

[11] http://mtg.upf.edu/technologies/sms

distorted by energy leaks from instruments, the interpolated partials seem to preserve well the overall spectral shape of the singing voice, including the formant frequencies, which encode the phoneme identities. The resynthesis allowed us to listen and verify that vocals are still to a large extent intelligible.

Note that melody resynthesis usually results in singing voice with perceivably worse intelligibility of the phonemes than the original signal. Some unvoiced consonants are dropped, as well as some artifacts are introduced. However, for computers, which are not as versed as human listeners in distinguishing among sources, the accompaniment reduction is an imperative step.

An example for the audio segment with the lyrics phrase *bakmıyor çeşmi siyah* can be seen in Figure 3.3b.



(a) Original spectrogram.

(b) Spectrogram of resynthesized harmonic partials content. Note that some unvoiced consonants are replaced with silences

Figure 3.3: An example of the resynthesized harmonic partials of singing voice for the lyrics phrase *bakmıyor çeşmi siyah.* Content up to 10 kHz is shown.

```
TARGETKIND = MFCC_0_D_A_Z
TARGETRATE = 100000.0
WINDOWSIZE = 250000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
NUMCEPS = 12
HIFREQ = 8000
```

Table 3.3: Parameters of the MFCC extraction (in the *htk* format). The target kind of feature has added 0th coefficient for energy (_0), plus its frame-to-frame difference (_D) and difference of the difference (_A) with zero-mean (_Z). The unit of *htk* is 100 nanoseconds, so the frame size is 25 ms, while hopsize is 10 ms. 26 mel bands and 22 liftering bands are used.

3.3.3 Acoustic Features

The MFCC have several parameters that could be tuned according to the application use case. A standard for their extracting for the characterization of singing voice are the default parameters of the HMM toolkit (*htk*), which is tailored to speech recognition (Young, 1993). The parameters are presented in Table 3.3 and are explained in detail in the *htk book*[12]. We adopted these to assure consistency to previous work, as in fact all the background lyrics-to-audio alignment approaches reviewed in the previous chapter rely on the htk variant of MFCC features. We believe that an important contribution of this work, from a practical point of view, is that we ported the variant of MFCC with the *htk* parameters to the open-source feature extraction library *essentia*[13]. Reducing the dependency on *htk* encourages the easier reproducibility and extensibility of this research[14].

3.3.4 Phoneme network

The phonetic recognizer is an HMM, whose states represent the sequence of phonemes from the phoneme transcription of the lyrics. As we described in Section 2.2.7 the goal of the grapheme-to-phoneme conversion is to create the phoneme transcription out of the word sequence, comprising the input lyrics for a particular vocal section.

A phonetic recognizer HMM can be represented as a DBN with a single hid-

[12] https://www.researchgate.net/publication/236023819_The_HTK_book_for_HTK_version_34

[13] http://essentia.upf.edu/documentation/

[14] A walkthrough on how to reproduce the htk-parameters in *essentia* is available at https://github.com/georgid/mfcc-htk-an-librosa/blob/master/mfcc_parameters_comparison_essentia.ipynb

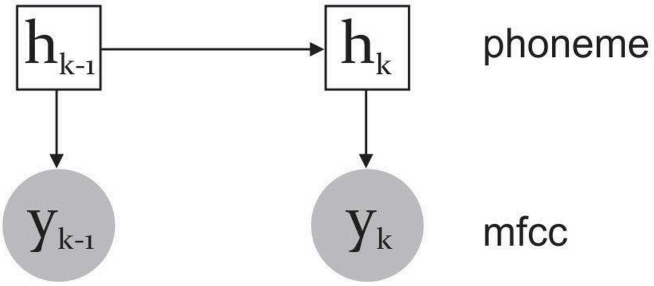


Figure 3.4: DBN for the baseline phonetic recognizer: one hidden variable represents the phoneme state. Circles and squares denote continuous and discrete variables, respectively. Gray nodes and white nodes represent observed and hidden variables, respectively.

den state for the current phoneme (Figure 3.4). In all DBN diagrams in this thesis we use circles and squares to denote continuous and discrete variables, respectively. Also gray nodes and white nodes represent observed and hidden variables, respectively. Although in initial experiments we trained a 3-state-HMM per phoneme, in most of the work presented in this dissertation a single-state-HMM was preferred. Preliminary experiments revealed that the difference in alignment accuracy with 3-states is negligible than that with one state. In this section we present the derivation of the phoneme network for Turkish. While in general the derivation of the phoneme network used for jingju is following the same principles, some Mandarin-particular details are discussed in Section 4.5.1.

Graphene-to-phoneme conversion

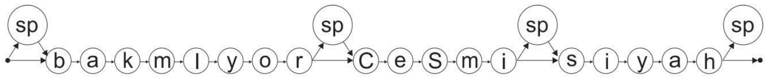
The words are expanded to phonemes based on a phonetic alphabet. Linguists have developed the international phonetic alphabet (IPA) — a language-independent notation system of phoneme sounds[15], because many of them are not language specific. For each language exists one or several options for an alphabet of machine-readable representation of IPA. For Turkish we have adopted the alphabet METUbet, proposed for one of the speech recognition state-of-the art systems for Turkish (Özgül Salor et al., 2007, Table 1). METUbet is very easy to interpret, because of its intuitiveness. All latin written characters are mapped to their corresponding latin phoneme, while the characters ç, ş, ı, ö and ü unique to the Turkish language are mapped to capital letters — respectively C, S, E, OE and UE. The unpronounced ğ is omitted from the transcript, whereas g is represented as GG.

[15] https://en.wikipedia.org/wiki/International_Phonetic_Alphabet

After the grapheme-to-phoneme conversion optional filler silence tokens are inserted in between words. A silence model represents short non-voiced time intervals, when the singing voice is not active to accommodate silent pauses or breaths between words. Using METUbet the lyrics phrase *bakmıyor çeşmi siyah* is expanded to a phoneme sequence seen in Figure 3.5a. Square brackets denote zero or one occurrence of a token, and vertical bars denote alternatives. Its corresponding phoneme network is depicted in Figure 3.5b.

[sp] b a k m I y o r [sp] C e S m i [sp] s i y a h [sp]

(a) Phoneme sequence for the lyrics phrase *bakmıyor çeşmi siyah*.



(b) Phoneme network for the lyrics phrase *bakmıyor çeşmi siyah*. Arrows indicated possible transitions with non-zero probabilities.

Figure 3.5: An example of the phoneme sequence and phoneme network for the phrase *bakmıyor çeşmi siyah* for *a cappella* voice. The phoneme set used is the Turkish METUbet.

Handling accompaniment artifacts

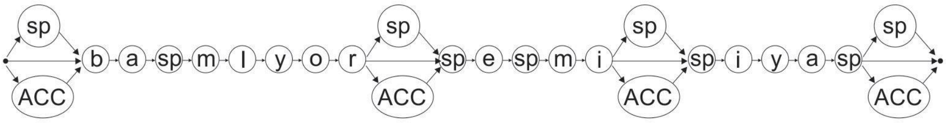
The phoneme network for accompanied singing ideally should be identical to the *a cappella* one presented above. In practice however, some of the phonemes in the accompaniment attenuation process are not accurately resynthesized. To address such cases, we build the network in a flexible way.

Except for silences, another filler model for non-vocal parts is introduced: a model for the instrumental background. We assume that the stochastic characteristics of the background music could be approximated by those of the instrumental-only regions in a music recording. We therefore trained a GMM for accompaniment instruments (ACC) from the time intervals, which are not annotated as words in the test dataset. It has a substantial amount of mixtures (40) to be able to capture the diverse timbral characteristics of background instruments. It is integrated as a single-state-HMM in the phoneme network. Setting the filler models as optional lets the phonetic recognizer activate the ACC model, depending on whether sound from background instruments was re-synthesized by the sinusoidal model, due to short regions, detected falsely as being vocal (see accompaniment attenuation step). In addition, this also accommodates potential instrumental leaks due to automatically detected boundary timestamps of vocal sections, displaced from the actual boundaries of sung lyrics.

A side effect of the resynthesis is that non-voiced consonants are not synthesized, which leaves short time intervals of silence. Looking carefully at Figure 3.3b one can notice that the time intervals for most METUbet unvoiced consonants: *k*, *S*, *s*, and *h* are converted into silences. Fujihara et al. (2011) suggested to tackle this problem by incorporating a separate method for detection of unvoiced consonants in the musical audio. The strategy we used instead is replacing unvoiced consonants by silence in the phoneme sequence. For example, for the phrase *bakmıyor çeşmi siyah* it will look accordingly in Figure 3.6a.

[sp|ACC] b a sp m I y o r [sp|ACC] sp e sp m i [sp|ACC] sp i y a sp [sp|ACC]

(a) Phoneme sequence for the lyrics phrase *bakmıyor çeşmi siyah*.



(b) Phoneme network for the lyrics phrase *bakmıyor çeşmi siyah*.

Figure 3.6: An example of the phoneme sequence and phoneme network for the phrase *bakmıyor çeşmi siyah* when accompanying instruments are present. The phoneme set used is the Turkish METUbet

Figure 3.6b presents its corresponding phoneme network. We evaluated the contribution of this simple resynthesis handling strategy by comparing to the performance of alignment between the resynthesizes audio and the phoneme network of Figure 3.5b that is meant for *a cappella* singing. The results (see Table 3.6) outlined a slight improvement with the accompaniment-aware network. We inspected carefully the flawed alignment cases with the *a cappella* phoneme network. This revealed that sometimes when there is a fricative in the vicinity of an inter-word *sp* (for example the *ş* from *çeşmi* following the *sp* between *bakmıyor* and *çeşmi*) the Viterbi would confuse the model of *sp* with the MFCC for the fricative sound, due to the similarity of the phoneme acoustics of the two. This means that usually a couple of phoneme models (ç and e in this example) are assigned falsely to the regions of the inter-word silence, which is extended in longer time than it should be. Sometimes instead of being delayed, the *sp* model is prematurely 'jumped to' due to the same type of fricative confusion. In contrast, when leaks of accompaniment sounds are present, the added ACC model helps in distinguishing between the fricative and silence/ACC.

3.4 Training the acoustic model

To represent the *acoustic model* (also known as the phonetic model) $P(y_k|x_k)$ of observing the MFCC feature vector y_k at a time instant k, given a phoneme x_k, a classifier of the different phonemes is needed. In essence, for a phonetic recognizer a hidden variable is the current phoneme class x_k (see Figure 3.4). The phoneme classifier has to represent the acoustic specificities of the different phonemes. In this section we present how we trained GMM-s and MLP-s — two different types of classifiers.

3.4.1 Gaussian mixture models

As presented in Section 2.2.7 the GMM-s until recently have been the *de facto* choice of phonetic timbre classifier. They have the ability, given enough mixtures, to approximate arbitrarily shaped densities. It is reasonable to assume that each mixture represents a broad class of a phonetic timbre event. Another reason to be preferred, is the so called *embedded reestimation* training technique. By means of it, it is relatively straightforward to train the model's parameters even from material with no phoneme annotations. Embedded reestimation is an generalization of the Expectation Maximization algorithm over time-series of feature vectors and has an efficient implementation in *htk* (Young, 1993). Utilizing *htk* we fitted a 9-component GMM for each phoneme on feature vectors extracted from a dataset of Turkish speech (Özgül Salor et al., 2007)[16]. The dataset encompasses diverse speech recordings totaling to approximately 500 minutes. Preliminary experiments confirmed that the trained models can successfully recognize withheld speech material from the same dataset.

To address the acoustic differences between speech and singing an adaptation of the trained GMM-s to singing material is needed. However due to lack of sufficient adaptation material we did not perform any adaptation[17]. Instead of that we explored the option of using neural networks for the acoustic model.

3.4.2 Multilayer perceptron neural networks

Recent work on keyword spotting in English *a cappella* singing showed that a MLP trained on singing-like material results in much better accuracy, compared to a GMM, trained on speech Kruspe (2015b).

This motivated us to take the opportunity to consider the deep MLP model the authors trained from amateur singers in their subsequent work — (Kruspe, 2016). We introduced their training procedure in Section 2.2.8 and will refer

[16] Training script is available at https://github.com/georgid/Lyrics2AudioAligner/tree/synthesis/TrainingStep

[17] Some scripts on preliminary adaptation experiments are available at https://github.com/georgid/Lyrics2AudioAligner/tree/synthesis/AdaptationStep

METUbet	IY	AA	UE	E	LL	I	O	M	U	OE	NN	VV
CMU	iy	aa	y	eh	l	ax	ao	m	uw	ow	n	v

METUbet	Z	C	ZH	H	CH	B	D	GG	F	KK	P	S	RR
CMU	z	jh	zh	hh	ch	b	d	g	f	k	p	s	r

Table 3.4: Direct mapping of English CMU phonemes to Turkish METUbet. Upper row vowels and liquids. Lower row all the rest consonants.

to their model as *MLP-English.* The *MLP-English* has 3 hidden layers with sigmoid activation function. The layers have respectively 1024, 850 and 1024 neurons and have as input the first 13 MFCC, extracted with the *htk* extraction parameters, described in 3.3.3 plus their deltas and accelerations. This results in a 39-dimensional feature vector. The phonetic alphabet used is the English-specific encoding of IPA from Carnegie Mellon University (CMU)[18].

Since we did not have as many Turkish singing voice phoneme annotations, sufficient for training a deep MLP, we simply adapted the *MLP-English* to Turkish. We exploited two cross-language phoneme mapping strategies: direct mapping and fuzzy mapping

Direct cross-language mapping

As observation probability for each Turkish phoneme we substituted the probability of an English phoneme from the output layer of the *MLP-English.* The mappings we used are listed in Table 3.4.

To most phonemes in Turkish corresponds an English phoneme that represents a sound with perceivably the same acoustics. The only two Turkish phonemes not existing in English are OE and UE, for which we experimented with different mappings and ended up with respectively *ow* and *y* as most optimal. We will refer to this mapping strategy as *MLP-DirectM.*

Fuzzy cross-language mapping

A more reasonable alternative to enforcing a phoneme to be represented by exactly one phoneme from another language is a weighted sum of the acoustics of a set of similar phonemes. Such types of 'fuzzy' many-to-one mapping strategy has been proposed for speech synthesis of a given speaker from her mother tongue to another language by Sun et al. (2016). Adopting the core idea of their concept, we trained GMM-s with the steps presented in Figure 3.7. First the extracted MFCC features from the *a cappella lyrics OTMM dataset* are input to the *English-MLP.* Then a vector of the posterior probabilities $p(s_n|x_k)$ of the n = 39 English phoneme classes for each time frame k are

[18] http://cmusphinx.sourceforge.net/

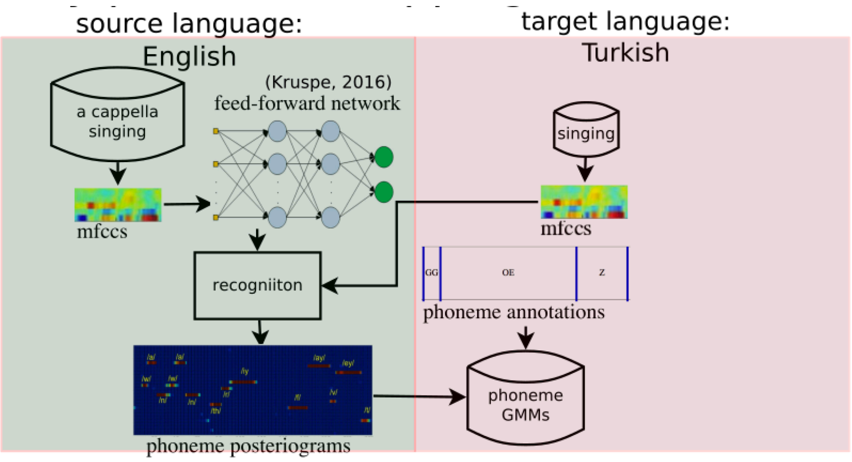


Figure 3.7: Cross-language phoneme mapping strategy from the source language (English) to the target language (Turkish). The *English-MLP* network is trained on a huge *DAMP* singing voice dataset, whereas the GMM-s are trained with phoneme annotations of a subset of the small *a cappella lyrics OTMM dataset*.

generated (see the left hand-side of Figure 3.7). These phonetic posterior probabilities are commonly known as phonetic posteriograms (PPG). Then in a second stage, a new model is trained to capture the mapping relationships between the posteriograms $p(s_n|x_k)$ and the 38 Turkish phoneme classes. The PPGs are fed into the classifier as if they were the acoustic feature vectors. While Sun et al. (2016) built another deep neural network, we preferred a 2-component GMM classifier, because training material with data sizes as small as 30 minutes of phoneme-annotated singing is usually enough for a GMM. Note that one could have trained GMM-s by embedded re-estimation to avoid the need of phoneme boundary annotations. However, we preferred training on annotations to ensure appropriate mappings between the acoustics of the two languages. Then on recognition the English PPGs are generated in the same way as on training. Training was conducted with leave-one-recording-out cross validation. We will refer to this mapping strategy as *MLP-FuzzyM*[19].

We compared the two mapping strategies with the baseline GMM-s trained with Turkish speech. We evaluated the percent of correct frames. To generate binary phoneme activations, we set to 1 the phonemes with maximum posterior probability for each time frame and zero to the rest of the phonemes. Then this sparse activation matrix is intersected with an oracle matrix, inferred from manually annotated phoneme boundaries. The first two models were evaluated on the whole phoneme-annotated subset of the *a cappella lyrics*

[19] Data preparation script available at https://github.com/georgid/englishMLP2turkish

model	% correct frames
GMM	9.8
MLP-DirectM	15.4
MLP-FuzzyM	9.2

Table 3.5: Percentage of correctly identified phoneme frames for the 3 different phoneme models utilized: GMM trained from Turkish speech, *MLP-English* model mapped directly to Turkish phonemes, *MLP-English* model mapped by the proposed fuzzy phoneme mapping strategy.

OTMM dataset, whereas the *MLP-FuzzyM* in the leave-one-out cross validation manner.

Table 3.5 present the percentage of correctly detected frames compared to the phoneme annotations — a metric used for the first time by Kruspe (2015b)[20]. The *MLP-directM* evidences a major improvement over the GMM-s trained on speech. It still scores reasonably worse than the reported 23 % in Kruspe (2016) on excerpts from the same English dataset, with which it was trained. This large margin indicates that the direct mapping strategy may not be the optimal one. Surprisingly, the fuzzy mapping strategy did not yield improvement over the baseline mixture model. We believe that the explanation lies in the very small size of the training singing dataset with phoneme annotations. We attribute the remarkable improvement of the English-to-Turkish directly mapped model to the big learning capacity of a deep feedforward neural network.

3.5 Experiments

Experiments are carried out on the *a cappella lyrics OTMM dataset* (Section 3.2.2) and the *multi-instrumental lyrics OTMM dataset* (Section 3.2.1). To assess the effectiveness of the accompaniment attenuation (AA) step, we aligned the multi-instrumental recordings from the *a cappella lyrics OTMM dataset* with and without AA. In initial experiments we build a *Python* wrapper around *htk* that has efficient Viterbi decoding[21]. To assure the same parameter setting of the baseline and the models that are aware of complementary context, we preferred to implement a custom HMM and Viterbi decoding[22]. The results reported are run with the latter implementation. We make available the source code of all experiments in this dissertation with the

[20] We implemented the percentage of phoneme frames in a script available at https://github.com/georgid/AlignmentEvaluation/blob/master/align_eval/evalPhonemes.py

[21] https://github.com/georgid/Lyrics2AudioAligner/tree/synthesis/AlignmentStep

[22] It is available as part of the repository https://github.com/georgid/AlignmentDuration by setting the parameter *WITH_DURATIONS* to 0. This repository has adopted some classes and Viterbi decoding logic from https://github.com/guyz/HMM

intention to serve as the first fully reproducible system for LAA. In addition, we hope this will encourage future research not only on LAA, but also on related computational topics of lyrics tracking.

When accompanying instruments are present, we employed the modified phoneme network, which can handle possible artifacts from the AA step (see Section 3.3.4).

3.5.1 Evaluation metrics

Throughout this thesis, we evaluate the LAA by the metrics *average absolute error* and *accuracy (percentage of correct segments)*, introduced in Section 2.2.1. We implemented a script for both metrics[23], which we plan to contribute to the collection of evaluation scripts of MIR research *mir_eval*[24]. The alignment error and accuracy are computed at boundaries of the lyrics phrases, which are manually annotated.

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

Table 3.6: Comparison of performance of the baseline phonetic recognizer with different variants of the acoustic model. Evaluation is performed on both *a cappella* and accompanied singing from OTMM. Alignment accuracy and alignment error on the boundaries of lyrics phrases and reported on total for all recordings. Mesaros stands for the approach of Mesaros and Virtanen (2008); Fujihara for the approach of Fujihara et al. (2011). AH stands for handling accompaniment artifacts (see Section 3.3.4).

[23] https://github.com/georgid/AlignmentEvaluation The implementation of the *percentage of correct segments* metric was ported from the script, kindly provided by H. Fujihara used in his work Fujihara et al. (2011)

[24] https://craffel.github.io/mir_eval/

3.5.2 Discussion

Table 3.6 lists results for the different system variants and steps of the recognizer. We compared the performance of the baseline phonetic recognizer on *a cappella* singing with two different variants of the acoustic model: with phoneme GMM-s and *MLP-directM* (the direct mapping to English-phonemes MLP). The GMMs result in rather low accuracy. The most possible explanation is the acoustic mismatch between our phoneme GMMs and the characteristics of singing voice. This is confirmed by the rather low results on *a cappella* singing. Training phoneme acoustics merely on speech is clearly suboptimal. The high score of the *MLP-DirectM* confirms that training on singing voice is a big advantage.

As to multi-instrumental material, adding the accompaniment attenuation improves reasonably accuracy (from 59.1 to 67.5 %). However still below the *a cappella* (70.2) clearly there is still room for improvement. In fact, investigating particular recording excerpts with low accuracy revealed that false positives of the AA module is a considerable reason for misalignment. We realize that the harmonic model (a generic model of singing voice) may not be the best choice for music with heterophonic character.

We observed that a problem is that alignment performs poorly towards the end of longer sections, which results in outliers of huge magnitude.

As a benchmark the best existing alignment systems for English pop songs (Mesaros and Virtanen, 2008) and for Japanese pop (Fujihara et al., 2011) are listed in the table. Comparison to them is not possible because they are developed for different genre and language and evaluated on different datasets. These are short-named in Table 3.6 respectively as Mesaros and Fujihara. Still, in these works alignment is also evaluated with GMM-s on the level of a lyrical line/phrase. Our baseline approach differs from both works essentially in that they conduct speech-to-singing-voice adaptation. In comparison, we did not perform any adaptation of the original speech model. Adaptation data of clean singing voice for a particular singer might not always be available and thus does not allow the system to scale to data from unknown singers.

Moreover, Fujihara et al. (2011) trains a VAD module on data selected from material with same acoustic characteristics as the test data. The module showed to notably increase the average accuracy of 72.1 % for a baseline to accuracy of 85.2 % for their final system.

3.6 Summary

In this chapter we described our LAA baseline system. It is a phonetic recognizer, based on phoneme HMM-s. We described the choices of the key steps of the phonetic recognizer, which are not related to modeling complementary context. Phoneme observation modeled as GMM-s, trained on Turkish speech proved not to be the optimal acoustic model. The alignment accuracy on *a cappella* (70.2 %) is rather low; whereas on multi-instrumental recordings (67.5%) is below the state of the art on LAA on English pop songs (85.2 %). The most possible explanation is the acoustic mismatch between our phoneme model and the characteristics of singing voice. To address this mismatch, we proposed a strategy of mapping a state-of-the-art model for English, trained on English pop songs, to Turkish. We explored two different mapping strategies. The simpler direct mapping increased reasonably the alignment accuracy (79.2 %). To our knowledge, this thesis presents the first work on computational modeling of sung lyrics, addressing the problem of inter-language phoneme mapping.

Despite its superiority, not all experiments (e.g. some presented in Chapter 4) are carried out with the phoneme GMM-s. This is because the mapping strategies were explored once the *English-MLP* became available (towards the end of this thesis[25]). However, we believe that the validity of the experiments in this dissertation is not negatively influenced by that.

[25] August 2016

Chapter 4

Lyrics-to-audio Alignment with Mid-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 mid-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 explicitly models 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 music-score-induced durations as a cue for tracking sung syllables.

To show the transferability of the proposed explicit-duration model outside of OTMM, we also evaluate it on material from jingju. The comparison to jingju has an aim to quantitatively evaluate if the duration knowledge contributes to a different degree for another music tradition. Jingju is a music tradition characterized by sung syllables that span particularly long time intervals. Being a largely oral tradition, it rarely has machine-readable music scores. Instead, to determine how long to sustain a given syllable, actors follow conventions for the structure of the lyrics line. Therefore we apply the previously proposed core modeling, wherein syllable durations are derived from these conventions, instead of score. Among all the approaches presented in this thesis, this is the clearest example of an approach informed by music-specific knowledge.

The chapter is organized as follows: We start off by introducing existing computational approaches of lyrics tracking, which explicitly model durations of syllables (Section 4.2). In Section 4.3 we introduce the duration-aware probabilistic model. Then we describe its application in two different cases: Firstly, in Section 4.4 we study how durations parsed from music scores in *makam* can be utilized as input reference syllable durations. Secondly, the core model is applied to jingju, for which reference durations are obtained from music-specific knowledge in the form of rules (Section 4.5).

4.2 Background on duration-aware lyrics-to-audio alignment

The phonetic recognizer approach is based on phoneme HMM-s. Standard HMMs have the drawback that they do not impose any restrictions on the waiting time in a state, resulting in a geometric distribution. This does not correspond to the naturally occurring durations of phonemes in speech. Introducing restrictions on the state waiting time of the phoneme HMMs improves speech recognition results (Ferguson, 1980).

Unlike speech, for which the variation of the durations of the vowels is relatively small, sung vowels can have significantly bigger variations and long durations. HMM-s are by far not capable to represent vowels with long durations, because the waiting time implied by the geometric distribution cannot be unlimitedly long (Rabiner, 1989). Durations can be modeled instead by a DHMM (also known as hidden semi-Markov models). In DHMMs the underlying process is allowed to be a semi-Markov chain with variable duration of each state (Yu, 2010). The idea is that the actual waiting time in a state can be generated by any statistical distribution. Common choices are the gamma distribution or normal distribution, whereby the distribution's parameters can be set by using some a-priori knowledge about the waiting time. In this respect, DHMM provide a flexible methodology that allows the injection of some music-specific context knowledge, from which the expected waiting time of a phoneme can be derived.

An approach to detect keywords from *a cappella* English pop songs exploiting knowledge about possible phoneme durations is presented in Kruspe (2015a). The authors used a DHMM with a gamma distribution, motivated by findings

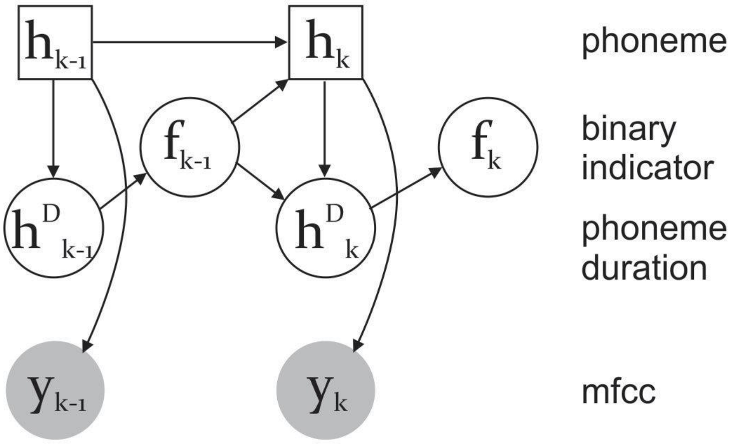


Figure 4.1: A DBN representing the duration-aware phonetic recognizer. Circles and squares denote continuous and discrete variables, respectively. Gray nodes and white nodes represent observed and hidden variables, respectively. A duration counter h^D keeps track of the waiting time in a phoneme state h. When h_D reaches 0, the binary indicator node f is fired, which triggers a change to next phoneme.

that gamma distribution represents well naturally observed phoneme durations in speech. The mean and variance of each phoneme is empirically estimated from a small portion of an *a cappella* dataset. The precision of keyword detection increased when durations were restricted. A limitation is that the learned phoneme parameters do not take into account the temporal structure of the lyrics lines. In other words, the duration of each phoneme is globally estimated, based on some training data. In addition to that, DHMM-s have been successful in modeling other problems from the domain of music information retrieval. In particular, they can represent chord durations in automatic chord recognition (Chen et al., 2012).

4.3 Duration-aware probabilistic model

In this section we describe the syllable-duration-aware probabilistic model, presented first in Dzhambazov and Serra (2015). In Figure 4.1 a DBN represents the duration spent in a phoneme h explicitly as a duration counter variable h^D. When the duration counter expires (reaches 0), the indicator node f_k turns on, the current phoneme h_k can change state, and the next duration counter, h^D_k, is reset. The reason there is no h to f arc is that the duration termination process is deterministic (Murphy, 2002, Figure 2.22). Inference in such a DBN with the Viterbi decoding will have time complexity of $O(TDH^2)$, where D is the maximal duration of the counter, T is the total time of a recording, and H is the number of phonemes in the phoneme network. In the case of forced alignment, it reduces to $O(TDH)$. In comparison to speech, the range of D for sung phonemes (especially in traditions like jingju with reasonably long vowels) can cause a big time complexity.

A limitation of this DBN is that due to the additional hidden counter variable h_D the size of the state space can become memory-demanding. To overcome that, we have adopted the idea of Chen et al. (2012) not to explicitly add the h^D to the model, but instead to extend the inference (Viterbi decoding) by handling the duration of states. Note that this inference does not reduce the time complexity. In what follows we describe a variation of Viterbi decoding method, in which maximization is carried over the most likely duration for each state. The duration counter is controlled by a normal distribution with mean derived from a lookup table of reference durations R_i, where i is the i^{th} phoneme in the phoneme network. The way the lookup table is constructed is related to how the complementary context is exploited and is the topic of sections 4.4 and 4.5

4.3.1 Parameter definitions

The Viterbi decoding is adapted from the procedure described in Chen et al. (2012). We assume that the duration d for a state j may vary according to a normal distribution $P_j(d)$ with mean at the reference duration $d = R_j$ and standard deviation σ. We will use a separate global standard deviation σ_v for all vowels and a global one σ_c for all consonants. For the sake of representation clarity, in the following equations we will use only one standard deviation symbol σ.

4.3.2 Recursion

The recursion step in the Viterbi algorithm from Eq.2.6 is extended by adding a term for the duration distribution $P_j(d)$.

For $R_{max} < t \leq T$

$$\delta_k(j) \quad = \quad \max_d \{\delta_{k-d}(j-1).P_j(d)^{\alpha}\, B_k(j,d)^{1-\alpha}\} \tag{4.1}$$

where

$$B_k(j,d) = \Pi_{s=k-d+1}^{k} b_j(O_s) \tag{4.2}$$

is the observation probability of staying d frames in state j until frame k. For the sake of simplicity, we define the likelihood $P_j(d)$ to be non-zero in the range $d \in (\max\{R_j - \sigma, 1\}, R_j + \sigma)$. The left margin of the range is reduced for states with reference duration $R_j < \sigma$.

A duration back-pointer is defined as

$$\chi_k(j) \quad = \quad \arg\max_d \{\delta_{k-d}(j-1).P_j(d)^{\alpha}\, B_k(j,d)^{1-\alpha}\} \tag{4.3}$$

Note that in forced alignment the source state could be only the previous state from the phoneme sequence $j-1$, therefore the transition probabilities a_{ij} are omitted.

To be able to control the influence of the duration we have introduced a weighting factor α. Note that setting α to zero is equivalent to using a uniform distribution for $P_j(d)$.

4.3.3 Initialization

For $t \leq R_{max}$

$$\delta_k(j) \quad = \quad \max\{\delta_k(j)^*,\, \kappa_k(j)\} \tag{4.4}$$

where a reduced-duration delta $\delta_k(j)^*$ is defined in the same way as in Eq.(4.1) but

$$d \in \begin{cases} \emptyset, & t \leq R_i - \sigma \\ (R_j - \sigma, \min\{t-1, R_j + \sigma\}), & else \end{cases} \tag{4.5}$$

reduces the duration to k when $k < R_j + \sigma$.

Lastly the probability of staying at initial state j at time k is defined as:

$$\kappa_k(j) = \pi_j P_j(k)^{\alpha}[\Pi_{s=1}^{k}(O_s)]^{1-\alpha} \tag{4.6}$$

for $k \in (1, R_j + \sigma)$.

4.3.4 Backtracking

Finally the decoded state sequence is derived by backtracking starting at the last state N and switching to a source state a number of $d = \chi_k(j)$ frames ahead according to the backpointer from Eq.4.3.

4.4 Durations derived from music score

In this section we present an application of the duration-aware model to material from OTMM. Makam singers tend to deviate from the music score to a significantly large extent. The goal is to show that the duration-aware model is capable to accommodate these varying durations. A reference lookup syllable duration table is constructed from note values from the music score.

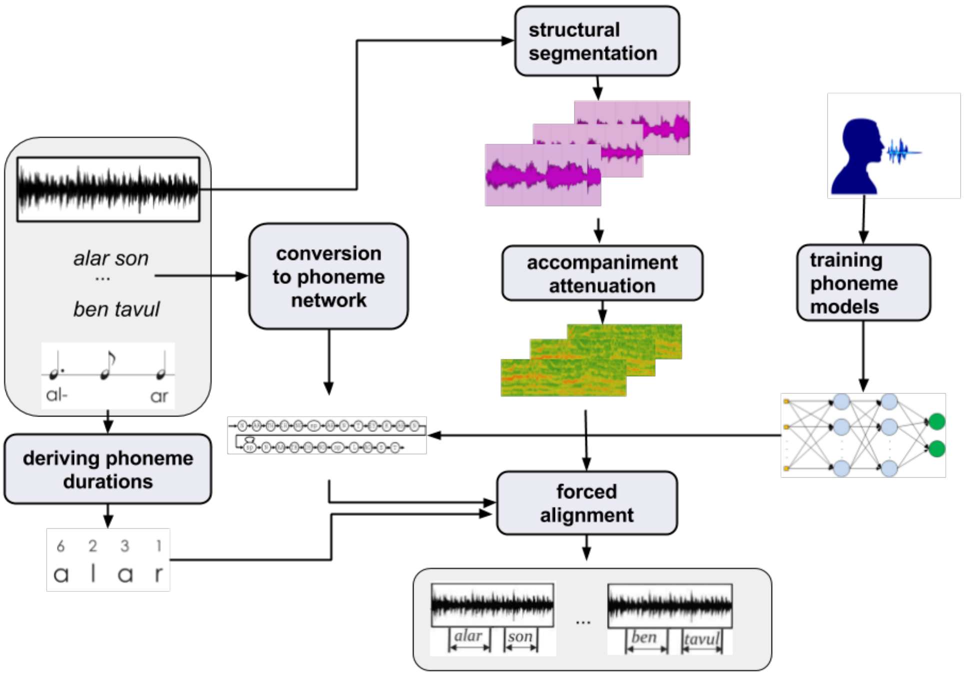


Figure 4.2: Overview of the steps of the lyrics-to-audio alignment system aware of phoneme durations. Durations are derived from the note values in the music score. The phonetic recognizer is a duration-explicit HMM

A general overview of the proposed approach is presented in Figure 4.2. As in all approaches presented in this thesis, first an audio recording is manually divided into segments according to the coarse level complementary context — the sections of the composition. In the case of OTMM the boundaries of vocal section (one of *zemin*, *nakarat*, *meyan*) are indicated in the music score. An audio recording and its corresponding score are input. Relying on HMM-s of phonemes the DHMM returns start and end timestamps of the aligned lyrics phrases.

4.4.1 Deriving phoneme durations

For each lyrics syllable a reference duration is derived by summing the values of its corresponding musical notes (in units of 64th notes). Then the reference duration is spread among its constituent phonemes, whereby consonants are assigned constant duration and the rest is assigned to the vowel. Each consonants in a syllable is assigned a constant reference duration $R_i = 0.3$ seconds. To align the score-inferred R_i to a recording of a performance, its reference musical tempo T_r is inferred, and R_i are linearly scaled according to T_r. After that scaling the unit of R_i becomes the number of acoustic frames.

4.4.2 Experiments

Experiments are carried out on the *a cappella lyrics OTMM dataset* (Section 3.2.2) and the *multi-instrumental lyrics OTMM dataset* (Section 3.2.1). Alignment is performed on each manually divided audio section and results are reported per recording (on total for its sections). To assess the benefit of the DHMM-s, results are compared to the baseline phonetic recognizer, presented in Section 3.5.

We summarize results for the most optimal $\alpha = 0.97$. The most optimal standard deviation for vowels σ_v was found to be 0.7, while we fixed the one for consonants σ_c to 0.1 seconds, based on the fact that consonant durations do not vary significantly. These parameters were optimized by minimizing the alignment error on a separate development dataset of 20 minutes OTMM acapella recordings. To assure precision, we measured alignment of the development dataset on annotations on the word level.

Evaluation metrics

Alignment accuracy is measured as the percentage of duration of correctly aligned regions from total audio duration (see Figure 2.1 for an example).

In addition, we define a metric *musical score in-sync* (MSI) to measure the approximate degree to which a singer performs a recording in synchronization with note values, indicated in the music score. Low accuracy of MSI indicates a higher temporal deviation from the music score. We compute MSI per a recording as the AA of score-inferred reference durations R_i compared to ground-truth, as if they were results after the alignment.

Discussion

Table 4.1 presents comparison of the proposed DHMM system performance and the baseline system. It can be observed that modeling durations with the DHMM increases the accuracy by 10 absolute percent. One reason for this are cases of long vocals, in which the standard HMM switches to the

System variant	**alignment accuracy**	**alignment error**
musical score in-sync	88.14	0.32
baseline a cappella	70.2	1.14
DHMM a cappella	90.04	0.26
baseline polyphonic	67.46	1.26
DHMM polyphonic	77.74	0.63
HMM+adaptation Mesaros and Virtanen (2008)	-	1.4
HMM+ singer adaptation Fujihara et al. (2011)	85.2	-

Table 4.1: Alignment accuracy (in percent) for musical score in-sync; different system variants: baseline HMM and DHMM; state-of-the-art for other languages. Alignment accuracy is reported as total for all recordings. Additionally the total mean phrase alignment error (in seconds) is reported

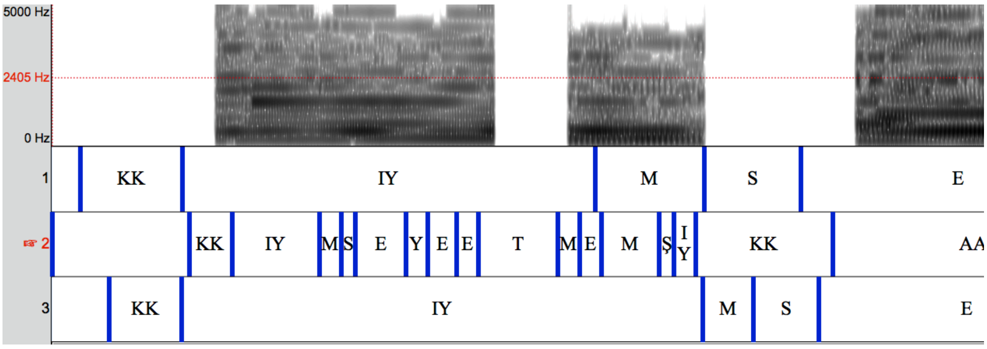


Figure 4.3: An example of decoded phonemes. *very top*: resynthesized spectrum; *upper level*: ground truth, *middle level*: HMM; *bottom level*: DHMM; (excerpt from the recording *Kimseye etmem şikayet* sung by Bekir Unluater).

next phoneme prematurely (due to its low likelihood of staying long in a given state). In contrast, the duration-explicit decoding allows picking the optimal duration (which can be traced in an example in Figure 4.3). [1]

Figure 4.4 allows a glance at results per recording, ordered according to MSI. It can be observed that the DHMM-s performs consistently better than the baseline (with some exceptions, for which accuracy is close). Unlike the rela-

[1] A complete demo of the alignment for this recording can be found at http://dunya.compmusic.upf.edu/makam/lyric-align/567b6a3c-0f08-42f8-b844-e9affdc9d215 after clicking *access lyrics player*. One needs to register in the Dunya-web.

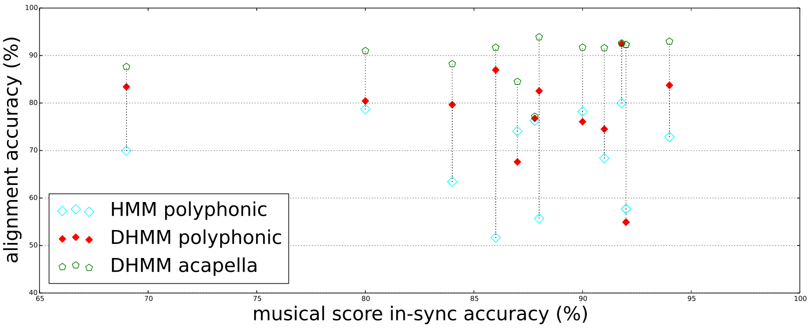


Figure 4.4: Comparison between results from DHMM (for both polyphonic and acapella) and the baseline HMM. The metric used is alignment accuracy. A connected triple of shapes represents results for one recording. Results are ordered according to *musical score in-sync* (on the horizontal axis)

tively stable accuracy for the *a cappella* case, when background instruments are present, the accuracy varies more among recordings.

For the sake of comparison, the alignment results of the best existing LAA systems for English pop songs (Mesaros and Virtanen, 2008) and for Japanese pop Fujihara et al. (2011) are also listed. These are abbreviated in Table 4.1 respectively as *HMM + adaptation* and *HMM + singer adaptation*. In these studies alignment is evaluated also on the level of a lyrical line/phrase. Despite the lack of adaptation, our DHMM-based system, based on an acoustic model trained only on speech, yields results in similar order to these reference approaches.

4.5 Durations derived from music principles

In jingju the durations, indicated in the music scores are not as strictly observed as in OTMM. Instead as a reference usually serve the orally transmitted singing examples of master actors. As we introduced in Section 2.1.2, over time, as part of this oral practice, specific rules have been formed. For example, if a poetry lyrics line has 10 syllables, a rule of thumb is that it consists of 2 3-syllable *dou*-s, followed by a 4-syllable *dou*. Respectively, if a poetry line has 7 syllables, it is a rule of thumb that it consist of 2 2-syllable dous, followed by a 3-syllable *dou* reference phoneme durations. These rules provide an excellent source to derive the phonemes reference durations for a duration-informed LAA. Therefore, we apply the duration-aware probabilistic model (see 4.3), whereby syllable reference durations are derived from these principles, instead of the music score. The experiments in this section are first presented in Dzhambazov et al. (2016).

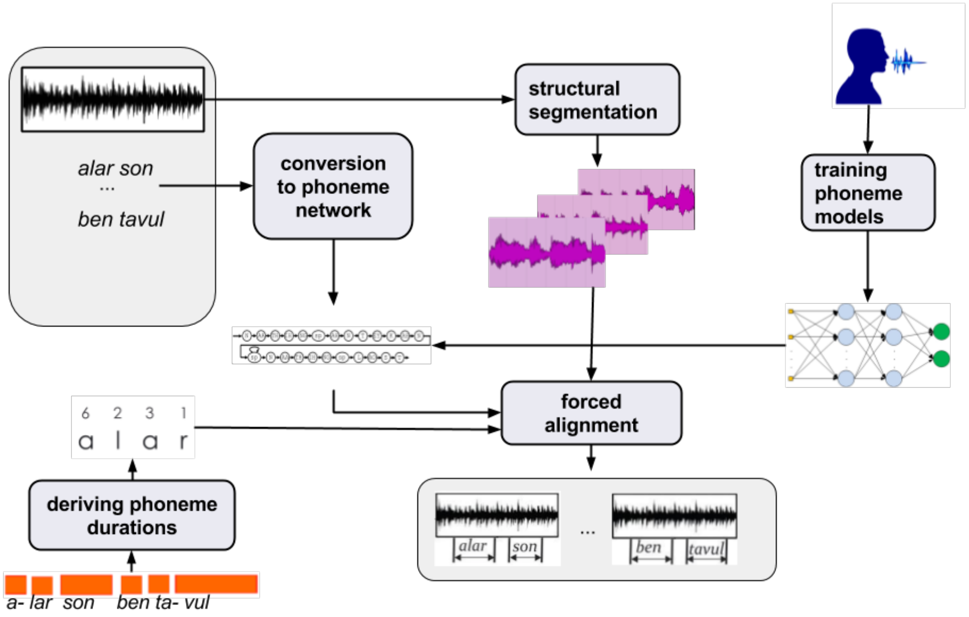


Figure 4.5: Overview of the steps of the lyrics-to-audio alignment system aware of phoneme durations. Durations are derived from music knowledge: the rules of durations of dous (syllable groups). The phonetic recognizer is a duration-explicit HMM

4.5.1 Steps of a phonetic recognizer

A general overview of the proposed approach is presented in Figure 4.5. As in all approaches presented in this thesis, first an audio recording is manually divided into segments according to the coarse level complementary context — the sections of the composition. In the case of jingju as a section serves a lyrics line or a couplet (two lines). Because lyrics in jingju are derived from poetry a lyrics line is in fact a lyrics sentence.

Training phoneme models The lyrics transcription in pinyin, divided into sentences for each aria, is expanded to phonemes based on grapheme-to-phoneme rules for Mandarin. A syllable-to-phoneme mapping table for Mandarin is used. Together with native Chinese speakers in the *CompMusic* team we created a mapping of pinyin syllables to the x-sampa phonetic alphabet [2]. Due to the small amount of training material, and due to their

[2] Part of the mapping (for diphthongs) rules can be found at https://github.com/georgid/AlignmentDuration/blob/noteOnsets/src/for_jingju/syl2ph.txt

relatively small ratio in total recording duration, most unvoiced consonants have been grouped into one class. Due to lack of publicly available Mandarin speech corpus, we trained the phoneme models on the jingju *a cappella* dataset (3.2.4). To assure a reasonable amount of training data, we trained in a 3-fold manner, using 10 of the arias from the dataset. Each fold has 5 arias of around 40 minutes each. A mapping from the *MLP-English* to Mandarin was not endeavored, because it seemed infeasible due to the audibly significant differences in the acoustics of the Mandarin vowels. Diphthongs and triphthongs make the sound of vowels very dependent on the acoustic context.

The first 13 MFCC and their delta and accelerations are extracted from 25 ms audio frames with the hop size of 10 ms from the *a cappella* singing. The extracted features are then fed to fit a GMM with 40 components for each phoneme. Phoneme-level annotations were used to isolate the segments for each phoneme. For jingju we prefer such a big number of mixture components to assure that it fits the varying acoustic conditions of the big number of diphthongs.

4.5.2 Music-knowledge-based durations

In Jingju an actor has the option to sustain the vocal of the *dou*'s final syllable. We will refer to the final syllable of a *dou* as *key syllable*. Therefore, reference phoneme durations are derived according to the *key syllables*, as follows:

Firstly, each *key syllable* in a *dou* is assigned longer reference duration, while the rest gets equal durations. Additionally, we observed in the dataset (see Section 3.2.4) that usually the last *key syllable* of the last sentence in a *banshi* is prolonged additionally. Thus we lengthened additionally the reference syllable duration of these last *key syllables*. Figure 4.6 depicts an example. According to *dou* groups the 3rd, 6th and last syllable are expected to be prolonged. Note that these expectations often do not hold — in this example they do not hold for the 3rd syllable.

To apply the duration-aware probabilistic model, we need to segment further the syllable reference durations down to phonemes reference durations R_i. To this end, the reference durations of syllables are divided among their constituent phonemes, according to the head-belly-tail division of syllables in the Jingju dialect (Wichmann, 1991). We assign consonants a fixed reference duration $R_c = 0.3$ seconds, while the rest of the syllable is distributed equally among vowels. The reference durations R_i are linearly scaled to a reference number of frames according to the ratio between the number of phonemes in a lyrics line and the duration of its corresponding audio segment. In comparison to the model presented for OTMM, we opted for a separate standard deviation d_c for consonants, and d_v for vowels. Proper values for d_c and d_v

assure that a phoneme sung longer or shorter than the expected R_i can be adequately handled.

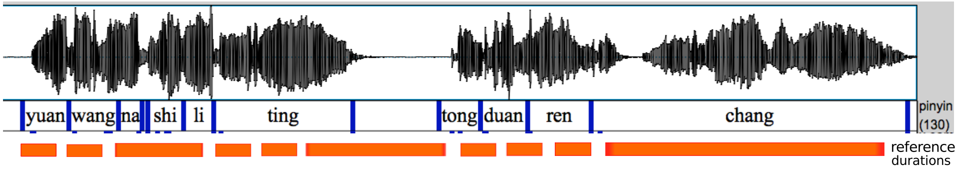


Figure 4.6: An example of 10-syllable sentence, being last in a *banshi* (before the *banshi* changes). Actual syllable durations are in pinyin, whereas reference durations are in orange parallelograms (below).

4.5.3 Experiments

Evaluation is carried on the dataset, presented in Section 3.2.4. Alignment accuracy is measured as the percentage of duration of correctly aligned regions from total audio duration (see Figure 2.1 for an example). In the context of the work, presented in this dissertation a value of 100 means perfect matching of all Mandarin syllable boundaries from evaluated audio. Accuracy is measured for each manually segmented lyrics sentence and accumulated on total for all the recordings.

To define a glass ceiling accuracy, alignment was performed considering phoneme annotations as an oracle for the acoustic features. Considering phoneme annotations, we set the probability of a phoneme to 1 during its time interval and 0 otherwise. We found that the median accuracy per a sentence of lyrics is close to 100%, which means that the model is generally capable of handling the highly-varying vocal durations of jingju singing. Most optimal results were obtained with $\sigma_c = 0.7$; $\sigma_v = 3.0$

As a baseline we employed a standard HMM with Viterbi decoding with the *htk* toolkit (Young, 1993). For both the HMM and the DHMM, because of the small size of the dataset, evaluation is done by cross validation on 3 folds with approximately equal number of syllables. Phoneme models are trained on 10 of the arias and evaluated on a 5-aria hold-out subset. Table 4.2 shows that the proposed duration model outperforms substantially the baseline alignment. Looking at oracle, one can conclude that reaching closer to it can be achieved in the future by improving the phoneme models, to capture phoneme identities in a more deterministic way.

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

Table 4.2: Comparison of total oracle, baseline and DHMM alignment. Accuracy is reported as accumulate correct duration over accumulate total duration over all sentences from a set of arias.

4.6 Summary

In this chapter we proposed how to extend an HMM-based phonetic recognizer for lyrics-to-audio alignment by utilizing lyrics duration information as a cue, complementary to phonetic timbre. An advantage of the presented model is that it allows room for certain temporal flexibility to handle cases of significant deviation of sung vowels from the expected reference durations. We evaluated on material from two music traditions: OTMM and jingju.

For the former reference phoneme durations are inferred from sheet music. The proposed model is tested on polyphonic audio recordings, as well as on an acapella dataset. Results show that the explicit modeling of phoneme durations outperforms a baseline approach, unaware of durations, by absolute 10 percent on the level of lyrics phrases. Information about durations can serves as a an important 'stepping stone' for the alignment process especially in the case of polyphonic audio, for which timbral features may not be deterministic enough.

For jingju we derived the expected syllable durations from music rules, specific for this music tradition.

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 sung voice, transitions between consecutive lyrics units 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 locations 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 transition probabilities between consecutive notes is the first major contribution of this chapter.

In the second part of the chapter, we address the relation of the transitions between consecutive phonemes to the simultaneously occurring vocal onsets.

A well-known fact is that when singing voice advances from the current syllable to another one, simultaneously with the change of timbre a vocal onset is perceived (Sundberg and Rossing, 1990). That is to say, the first voiced sound in a syllable bears the onset of a new note. The second major contribution of this chapter is conceptualizing onset-aware phoneme transition rules, because such relations between vocal onsets and phonemes have not been previously formalized in a computational study. We propose further how to integrate these rules into the transition model of a phonetic recognizer. This contributes to alignment based on knowledge about the vocal onsets. To test the feasibility of the proposed model, we aligned the lyrics utilizing manually annotated onsets. Further, we explore how automatically detected vocal onsets can replace the annotations. Using automatic singing transcription to detect the vocal onsets instead of score-informed methods reduces the need for music scores. Evaluation is carried out on *a cappella* material from OTMM.

We start this chapter off by reviewing existing methods for singing voice transcription and existing methods for tracking metrical accents (in particular tracking beats) (Section 5.2). In Section 5.3 we explore how the accuracy of vocal onset detection can be increased by simultaneously tracking the beats in a metrical cycle. Finally, in Section 5.4 we present a study of how the detected note onsets influence the transitions between consecutive phonemes. The novel phoneme transition rules and their integration into the transitions of an HMM are presented respectively in Sections 5.4.1 and 5.4.2.

5.2 Background

5.2.1 Singing voice transcription

The process of converting an audio recording into some form of music notation is commonly known as automatic music transcription. Current transcription methods use general purpose models, which are unable to capture the rich diversity found in music signals from different instruments Benetos et al. (2013). In particular, singing voice poses a challenge to transcription algorithms because of its soft onsets and expressive elements such as portamento and vibrato. In recent years there has been a substantial amount of work on the extraction of pitch from both *a cappella* singing (Babacan et al., 2013; Molina et al., 2014) and predominant singing voice from polyphonic music (Salamon et al., 2014). This has paved the way to an increased accuracy of singing voice transcription algorithms. One of the reasons for this is that a correctly detected melody contour is a fundamental precondition for singing voice transcription (SVT).

The core subtasks of SVT are detecting note events with a discrete pitch value, an onset time and an offset time from the estimated time-pitch representation. A HMM that describes notes is presented in Ryynänen (2004), wherein a note has 3 states: attack (onset), stable pitch state and silent state. The transition probabilities are learned from data. Recently Mauch et al. (2015) suggested to represent the observation and transition likelihoods by rules compacted from music knowledge, instead of learning them from data. The model covers a range with distinct pitches from a lowest MIDI tone C2 up to B7. Each MIDI pitch is further divided into 3 sub-pitches, resulting in $n = 207$ notes with different pitch, each having the 3 note states. Although being conceptually capable of tracking onsets in singing voice audio with accompaniments, these approaches were tested only on *a cappella* singing. In multi-instrumental recordings, an essential first step is to extract reliably the predominant vocal melody. One of the few works dealing with SVT for polyphonic recordings — Kroher and Gómez (2016); Nishikimi et al. (2016) — are based on the algorithm for predominant melody extraction of Salamon and Gómez (2012). Temporal deviations of the sung onsets from their positions indicated in music score are modeled in a probabilistic way in Nishikimi et al. (2016). In Kroher and Gómez (2016) as a primary step of the note transcription stage, notes are segmented by a set of flamenco-specific onset detection rules, based on pitch contour and volume characteristics.

Vocal onsets are usually soft (a slower attack phrase), in contrast to some instruments with percussive onsets (Sundberg and Rossing, 1990). This makes the precise location of a vocal onset ill-defined. The note onset corresponds to the initial segment of the three temporal segments of a vocal note: attack, sustain and release[1]. As the location of a note onset we will refer to the time instant, in which a pitched segment starts. They could follow (but exclude) a region with higher energy in case a syllable starts with a non-voiced consonant. The other temporal segments — sustain and release (offset) have undoubtedly also impact on the transition of phonemes. However, in this thesis, we consider the impact of vocal note onsets only, for they have arguably the most evident influence on syllable transitions.

The detection of instrumental onsets in polyphonic recordings is a challenging problem itself, which has attracted the attention of researchers for many years[2]. Most algorithms are based on the observation that an onset entails a change of the energy of the signal or of its harmonic content. One successful approach is to distinguish the spectral peaks, which are due to candidate note transient time segments (Röbel, 2009). Soft onsets are treated as a special case: the sensitivity of the generic transient detector is modified whenever the transients appear in a harmonic structure, which is usually a condition for soft onsets. A thorough review of onset detection methods can be found in Benetos et al. (2013, Section 2.2). Vocal onset detection in multi-instrumental music is, in fact, one of the hardest MIR problems. Determining their exact

[1] We stick to the definition of *temporal segments*, adopted from the chapter *Singing Transcription* of Klapuri and Davy (2006)

[2] www.music-ir.org/mirex/wiki/2015:Audio_Onset_Detection

onset timestamp is even harder in OTMM because of expressive singing phenomena: vocal onsets are often approached by portamentos. Therefore any complementary information can be an important 'stepping stone' for increased detection accuracy.

5.2.2 Beat Detection

Recently a Bayesian approach, referred to as the *bar-pointer* model, has been presented (Whiteley et al., 2006). It describes events in music as being driven by their current position in a metrical cycle (i.e. musical bar). The model represents as hidden variables in a DBN the current position in a bar, the tempo, and the type of musical meter, which can be referred to as bar-tempo state space.

The work of Holzapfel et al. (2014) applied this model to recordings from non-Western music, in order to handle jointly beat and downbeat tracking. The authors showed that the original model can be adapted to different rhythmic styles and time signatures, and an evaluation is presented on Indian, Cretan and Turkish music datasets.

Later Krebs et al. (2015) suggested a modification of the bar-tempo state space, in order to reduce the computational burden from its huge size.

5.3 Metrical-accent-aware vocal onset detection

Metrical accents are a facet of complementary context that defines the rhythmic backbone of vocal melodies. Therefore it is worth studying how the transitions between lyrics of units (in particular syllables) interact with these accents. By *metrical accents* we will refer to notes that are emphasized as a result of the context of the musical meter. Naturally, accents occur on the beats, whereby downbeats (the first beat in a meter) will be perceived as being stronger accentuated. Detecting the times of vocal note onsets can benefit from automatically detected metrical events, such as beats. In fact, the accents in a metrical cycle determine to a large extent the temporal backbone of the singing melody. Studies on symbolic music data showed that the timestamps of vocal note onsets are influenced by the their position in a metrical cycle (Huron, 2006; Holzapfel, 2015). Despite that, there have been very few studies on meter-aware analysis of onsets in music audio (Degara et al., 2010).

In this section we make a hypothesis that the knowledge of the current position in a metrical cycle (i.e. metrical accent) can improve the accuracy of vocal note onset detection. To this end we propose a novel probabilistic model to jointly track beats and vocal note onsets.

5.3.1 Model Architecture

The proposed approach extends the beat and meter tracking model, presented in Krebs et al. (2015). We adopt from it the variables for the position in the metircal cycle (bar position) ϕ and the instantaneous tempo $\dot{\phi}$. We also adopt the observation model, which describes how the metrical accents (beats) are related to an observed onset feature vector y_f. All variables and their conditional dependencies are represented as the hidden variables in a DBN (see Figure 5.1). We consider that the *a priori* probability of a note at a specific metrical accent interacts with the probability of observing a vocal note onset. To represent that interaction we add a hidden state for the temporal segment of a vocal note n, which depends on the current position in the metrical cycle. The probability of observing a vocal onset is derived from the emitted pitch y_p of the vocal melody.

In the proposed DBN, an observed sequence of features derived from an audio signal $y_{1:K} = \{y, .., y_K\}$ is generated by a sequence of hidden (unknown) variables $x_{1:K} = \{x_1, ..., x_K\}$, where K is the length of the sequence (number of audio frames in an audio excerpt). The joint probability distribution of hidden and observed variables factorizes as:

$$P(x_{1:K}, y_{1:K}) = P(x_0)\Pi_{k=1}^{K} P(x_k|x_{k-1})P(y_k|x_k) \tag{5.1}$$

where $P(x_0)$ is the initial state distribution; $P(x_k|x_{k-1})$ is the transition model and $P(y_k|x_k)$ is the observation model.

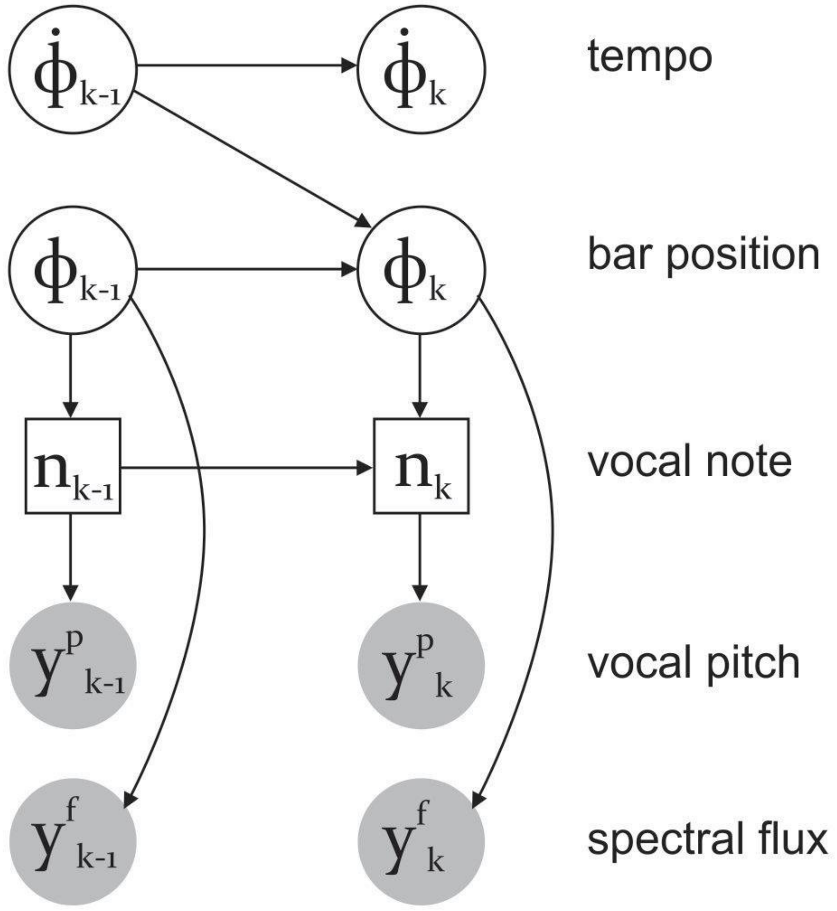


Figure 5.1: A dynamic Bayesian network for the proposed beat and vocal onset detection model. Circles and squares denote continuous and discrete variables, respectively. Gray nodes and white nodes represent observed and hidden variables, respectively.

5.3.2 Hidden variables

At each audio frame k, the hidden variables describe the state of a hypothetical bar pointer $x_k = [\dot{\phi}_k, \phi_k, n_k]$, representing the instantaneous tempo, the bar position and the vocal note respectively.

Tempo state $\dot{\phi}$ and bar position state ϕ

The bar position ϕ points to the current position in the metrical cycle (bar). The instantaneous tempo $\dot{\phi}$ encodes how many bar positions the pointer advances from the current to the next time instant. To assure feasible computational time we relied on the combined bar-tempo efficient state space, presented in Krebs et al. (2015). To keep the size of the bar-tempo state space small, we input the ground truth tempo for each recording, allowing a range for $\dot{\phi}$ within ±10 bpm from it, in order to accommodate gradual tempo changes. This was the minimal margin at which beat tracking accuracy did not degrade substantially. For a study with data with higher stylistic diversity, it would make sense to increase it to at least 20% as it is done in Holzapfel and Grill (2016, Section 5.2). This yields around 100-1000 states for the bar positions within a single beat (in the order of 10000 for the 8-9 beats of the usuls).

Vocal note state n

The vocal note states represent the temporal segments of a sung note. They are a modified version of these suggested in the note transcription model of Mauch et al. (2015). We adopted the first two segments: attack region (A), stable pitch region (S). We replaced the silent segment with non-vocal state (N). Because full-fledged note transcription is outside the scope of this work, instead of 3 steps per semitone, we used for simplicity only a single one, which deteriorated just slightly the note onset detection accuracy. Also, to reflect the pitch range in the datasets, on which we evaluate, we set as minimal MIDI note E3 covering almost 3 octaves up to B5 (35 semitones). This totals to 105 note states.

To be able to represent the DBN as an HMM, the bar-tempo efficient state space is combined with the note state space into a joint state space x. The joint state space is a cartesian product of the two state spaces, resulting in up to $10000 \times 105 \approx 1\,\mathrm{M}$ states.

5.3.3 Transition model

Due to the conditional dependence relations in Figure 5.1 the transitional model factorizes as

$$\begin{aligned} P(x_k|x_{k-1}) = \quad & P(\dot{\phi}_k|\dot{\phi}_{k-1}) \times \\ P(\phi_k|\phi_{k-1},\dot{\phi}_{k-1}) \times \quad & P(n_k|n_{k-1},\phi_k) \end{aligned} \tag{5.2}$$

The tempo transition probability $p(\dot{\phi}_k|\dot{\phi}_{k-1})$ and bar position probability $p(\phi_k|\phi_{k-1},\dot{\phi}_{k-1})$ are the same as in Krebs et al. (2015). Transition from one tempo to another is allowed only at bar positions, at which the beat changes. This is a reasonable assumption for the local tempo deviations in the analyzed datasets, which can be considered to occur relatively beat-wise.

Note transition probability

The probability of advancing to a next note state is based on the transitions of the note-HMM, introduced in Mauch et al. (2015). Let us briefly review it: From a given note segment the only possibility is to progress to its following note segment. To ensure continuity each of the self-transition probabilities is rather high, given by constants c_A, c_S and c_N for A, S and N segments respectively (c_A=0.9; c_S=0.99; $c_N = 0.9999$). Let $P_{N_iA_j}$ be the probability of transition from non-vocal state N_i after note i to attack state A_j of its following note j. The authors assume that it depends on the difference between the pitch values of notes i and j and it can be approximated by a normal distribution centered at change of zero (Mauch et al. (2015), Figure 1.b). This implies that small pitch changes are more likely than larger ones. Now we can formalize their note transition as:

$$p(n_k|n_{k-1}) = \begin{cases} P_{N_iA_j}, & n_{k-1} = N_i \quad n_k = A_j \\ c_N, & n_{k-1} = n_k = N_i \\ 1 - c_A, & n_{k-1=}A_i \quad n_k = S_j \\ c_A, & n_{k-1} = n_k = A_i \\ 1 - c_S & n_{k=1} = S_i \quad n_k = N_j \\ c_S, & n_{k-1} = n_k = S_i \\ 0 & otherwise \end{cases} \tag{5.3}$$

Note that the outbound transitions from all non-vocal states N_i should sum to 1, meaning that

$$c_N = 1 - \sum_i P_{N_iA_j} \tag{5.4}$$

In this study, we modify $P_{N_iA_j}$ to allow variation in time, depending on the current bar position ϕ_k.

$$p(n_k|n_{k-1},\phi_k) = \begin{cases} P_{N_iA_j}\Theta(\phi_k), & n_{k-1} = N_i, n_k = A_j \\ c_N, & n_{k-1} = n_k = N_i \\ \ldots \end{cases} \tag{5.5}$$

where

$\Theta(\phi_k)$: function weighting the contribution of a beat adjacent to current bar position ϕ_k

and

$$c_N = 1 - \Theta(\phi_k) \sum_i P_{N_iA_j} \tag{5.6}$$

The transition probabilities in all the rest of the cases remain the same. We explore two variants of the weighting function $\Theta(\phi_k)$:

1. Time-window redistribution weighting: Singers often advance or delay slightly note onsets off the location of a beat. The work Nishikimi et al. (2016) presented an idea on how to model vocal onsets, time-shifted from a beat, by stochastic distribution. Similarly, we introduce a normal distribution $\mathcal{N}_{0,\sigma}$, centered around 0 to re-distribute the importance of a metrical accent (beat) over a time window around it. Let b_k be the beat, closest in time to a current bar position ϕ_k. Now:

$$\Theta(\phi_k) = [\mathcal{N}_{0,\sigma}(d(\phi_k, b_k))]^w e(b_k) \tag{5.7}$$

where

$e(b)$: probability of a note onset co-occurring with the b^{th} beat (b $\in B$); B is the number of beats in a metrical cycle

w : sensitivity of vocal onset probability to beats

$d(\phi_k, b_k)$: the distance from current bar position ϕ_k to the position of the closest beat b_k

Equation 5.5 means essentially that the original $P_{N_iA_j}$ is scaled according to how close in time to a beat it is.

2. Simple weighting: We also aim at testing a more conservative hypothesis that it is sufficient to approximate the influence of metrical accents only at the locations of beats. To reflect that, we modify the $P_{N_iA_j}$ only at bar positions corresponding to beat positions, for which the weighting function is set to the peak of $N_{0,\sigma}$, and to 1 elsewhere.

$$\Theta(\phi_k) = \begin{cases} [N_{0,\sigma}(0)]^w e(b_k), & d(\phi_k, b_k) = 0 \\ 1 & else \end{cases} \tag{5.8}$$

5.3.4 Observation models

The observation probability $P(y_k|x_k)$ describes the relation between the hidden states and the (observed) audio signal. In this work we make the assumption that the observed vocal pitch and the observed metrical accent are conditionally independent from each other. This assumption may not hold in cases when energy accents of singing voice, which contribute to the total energy of the signal, are correlated to changes in pitch. However, for music with percussive instruments the importance of singing voice accents is diminished to a significant extent by percussive accents. Now we can rewrite Eq.2.5 as

$$\begin{gathered} P(x_{1:K}, y^f_{1:K}, y^p_{1:K}) = \\ P(x_0)\Pi^K_{k=1} P(x_k|x_{k-1}) P(y^f_k|x_k) P(y^p_k|x_k) \end{gathered} \tag{5.9}$$

This means essentially that the observation probability can be represented as the product of the observation probability of a metrical accent $P(y^f_k|x_k)$ and the observation probability of vocal pitch $P(y^p_k|x_k)$.

Accent observation model

For $P(y^f_k|x_k)$ we train GMM-s on the spectral flux-like feature y^f, extracted from the audio signal using the same parameters as in Krebs et al. (2013) and Holzapfel et al. (2014). The feature y^f summarizes the energy changes (accents) that are likely to be related to the onsets of all instruments together. The probability of observing an energy change depends on the position in the bar and the rhythmic pattern, $P(y^f_k|x_k) = P(y^f_k|\phi_k, r_k)$.

Pitch observation model

The pitch probability $P(y^p_k|x_k)$ reduces to $P(y^p_k|n_k)$, because it depends only the current note state. We adopt the idea proposed in Mauch et al. (2015) that a vocal note state emits pitch y^p according to a normal distribution, centered around its average pitch. The standard deviation of stable states and the one of the onset states are kept the same as in the original model, respectively 0.9 and 5 semitones. The melody contour of singing is extracted in a preprocessing step. We utilized an algorithm, extended from Salamon and Gómez (2012) and tailored to Turkish makam. Each audio frame k gets assigned a pitch value and probability of being voiced v_k Atlı et al. (2014).

Based on frames with zero probabilities, one can infer which segments are vocal and which not. Since correct vocal segments is crucial for the sake of this study and the VAD of these melody extraction algorithms are not state of the art, we manually annotated segments with singing voice, and thus assigned $v_k = 0$ for all frames, annotated as non-vocal.

For each state the observation probability $P(y_k^p|n_k)$ of vocal states is normalized to sum to v_k (unlike the original model which sums to a global constant v). This leaves the probability for each non-vocal state be $^{1-v_k}/_n$.

5.3.5 Learning model parameters

Accent observation model

For this study we divided the the *multi-instrumental vocal onsets OTMM dataset* (see section 3.2.3) into training and test subsets. The training dataset spans around 7 minutes of audio from each of the two usuls. Due to the scarcity of material with solo singing voice, several excerpts with choir sections were included in the training data. We trained the accent probability patterns $P(y_k^f|\phi_k, r_k)$ on the training dataset. For each usul we trained one rhythmic pattern by fitting a 2-mixture GMM on the spectral-flux-like feature vector y^f. Analogously to Holzapfel et al. (2014) we pooled the bar positions down to 16 patterns per beat. The feature vector is normalized to zero mean, unit variance and taking moving average. Normalization is done per song.

Probability of note onset

The probability of a vocal note onset co-occurring at a given bar position $e(b)$ is obtained from studies on sheet music. Many notes are aligned with a beat in the music score, meaning a higher probability of a note at beats compared to inter-beat bar positions. A separate distribution $e(b)$ is applied for each different metrical cycle. For the düyek and aksak usuls $e(b)$ has been taken from a recent study Holzapfel (2015, Figure 5. a-c). The authors used a corpus of music scores, on data from the same corpus, from which we derived the dataset. The patterns reveal that notes are expected to be located with much higher likelihoods on those beats with percussive strokes than on the rest.

5.3.6 Inference

We obtain the best state sequence $x_{1:K}$ by decoding with the Viterbi algorithm. A note onset is detected when the state path enters an attack note state after being in non-vocal state.

With manually annotated beats

We explored the option that beats are given as input from a preprocessing step (i.e. when they are manually annotated). In this case, the detection of vocal onsets can be carried out by a reduced model with a single hidden variable: the note state. The observation model is then reduced to the pitch observation probability. The transition model is reduced to bar-position-aware transition probability $a_{ij}(k) = p(n_k = j|n_{k-1} = i, \phi_k)$(see Eq.5.5). To represent this time-dependent self-transition probabilities we we utilize time-varying transition matrix. It falls in the general category of variable-time HMM-s (VTHMMs) Johnson (2005). The standard transition probabilities in the Viterbi maximization step in equation 2.6 are substituted for the bar-position-aware transitions $a_{ij}(k)$

$$\delta_k(j) = \max_{i \in (j, j-1)} \delta_{k-1}(i)\, a_{ij}(k)\, b_j(O_k) \tag{5.10}$$

Full model

In addition to onsets, a beat is detected when the bar position variable hits one of B positions of beats within the metrical cycle.

Note that the size of the state space x poses a memory requirement. A recording of 1 minute has around 10000 frames at a hopsize of 5.8 ms. To use Viterbi thus requires to store in memory pointers to up to 4 G states, which amounts to 40 G RAM (with uint32 python data type).

5.3.7 Experiments

Vocal detection is evaluated on 5 1-minute excerpts from each of the two usuls from the *multi-instrumental vocal onsets OTMM dataset* (see Section 3.2.3), totaling in 10 minutes of audio (on total 780 onsets). The hopsize of computing the spectral flux feature, which resulted in best beat detection accuracy in Holzapfel et al. (2014) is $h_f = 20\,ms$. In comparison, the hopsize of predominant vocal melody detection is usually of smaller order i.e. $h_p = 5.8\,ms$ (corresponding to 256 frames at sampling rate of 44100). Preliminary experiments showed that extracting pitch with values of h_p bigger than this values reasonably deteriorated the vocal onset accuracy. Therefore in this work we used hopsize of 5.8 ms for the extraction of both features. The time difference parameter for the spectral flux computation remains unaffected by this change in hopsize, because it can be set separately.

As a baseline we run the algorithm of Mauch et al. (2015) with the 105 note states, we introduced in Section 5.3.2[3]. The note transition probability is the

[3] We extended the port of the original VAMP plugin implementation to

meter		beat Fmeas	P	R	Fmeas
aksak	Mauch	-	33.1	31.6	31.6
	Ex-1	-	37.5	38.4	37.2
	Ex-2	86.4	37.8	36.1	36.1
düyek	Mauch	-	42.1	36.9	37.9
	Ex-1	-	44.3	41.0	41.4
	Ex-2	72.9	45.0	39.0	40.3

Table 5.1: Evaluation results for Experiment 1 (shown as Ex-1) and Experiment 2 (shown as Ex-2). Mauch stands for the baseline, following the approach of Mauch et al. (2015). P, R and Fmeas denote the precision, recall and F-measure of detected vocal onsets. Results are averaged per usul.

original as presented in Eq.5.3, i.e. not aware of beats. Note that in Mauch et al. (2015) the authors introduce a post-processing step, in which onsets of consecutive sung notes with same pitch are detected considering their intensity difference. We excluded this step in all system variants presented, because it could not be integrated in the proposed observation model in a trivial way. This means that, essentially, in this experiments cases of consecutive same-pitch notes are missed, which decreases somewhat the recall compared to the original algorithm.

Evaluation metrics

Beat detection Since improvement of the beat detector is outside the scope of this dissertation, we report accuracy of detected beats only in terms of their F-measure[4]. This serves solely the sake of comparison to existing work[5]. The F-measure can take a maximum value of 1, while beats tapped on the off-beat relative to annotations will be assigned an F-measure of 0. We used the default tolerance window of 70 ms, also applied in Holzapfel et al. (2014).

Vocal onset detection We measured vocal onset accuracy in terms of precision and recall[6]. Unlike *a cappella* singing, the exact onset times of singing voice accompanied by instruments, might be much more ambiguous. To accommodate this fact, we adopted the tolerance of $t = 50\,ms$, used for vocal

Python https://github.com/ronggong/pypYIN, which we make available at https://github.com/georgid/pypYIN

[4] The evaluation script used is at https://github.com/CPJKU/madmom/blob/master/madmom/evaluation/beats.py

[5] Note that F-measure is agnostic to the phase of the detected beats, which is clearly not optimal

[6] We used the evaluation script available at https://github.com/craffel/mir_eval

onsets in accompanied flamenco singing by Kroher and Gómez (2016), which is much bigger than the $t = 5\,ms$ used by Mauch et al. (2015) for *a cappella*. Note that measuring transcription accuracy remains outside the scope of this thesis.

Experiment 1: With manually annotated beats

As a precursor to evaluating the full-fledged model, we conducted an experiment with manually annotated beats. This is done to test the general feasibility of the proposed note transition model (presented in 5.3.3), unbiased from errors in the beat detection.

We did apply both the simple and the time-redistribution weighting schemes for $\Theta(\phi_k)$, presented respectively in Eq.5.8 and in Eq.5.7. In preliminary experiments we saw that with annotated beats the simple weighting results in much worse onset accuracy than the time-redistributed one. Therefore the experimental results reported are conducted with the latter weighting scheme.

We have tested different pairs of values for w and σ from Eq.5.5.The onset detection accuracy peaked at $w = 1.2$ and $\sigma = 30\,ms$. Table 5.1 presents the accuracies compared to the baseline. Inspection of detections showed that the proposed model added some onsets around beats, which are missed by the baseline.

Experiment 2: Full model

To assure computational speed, we did extend the efficient implementation of the joint bar-tempo state space and the Viterbi algorithm of the *madmom* toolbox[7]. The average F-measure of detected beats for the different metrical cycles can be seen in Table 5.1[8]. For aksak and düyek usuls, the accuracy is somewhat worse than the results of 91 and 85.2 respectively, reported in Holzapfel et al. (2014, Table 1.a-c, R=1). We believe the reason is in the smaller size of our training data. Table 5.1 evidences also a reasonable improvement of the vocal onset detection accuracy for both music traditions. The results reported are only with the simple weighting scheme for the vocal note onset transition model (the time-redistribution weighting was not implemented in this experiment).

Adding the automatic beat tracking improved the baseline, whereas this was not the case with manual beats for simple weighting. This suggests that the concurrent tracking of beats and vocal onsets is a flexible strategy and can accommodate some vocal onsets, slightly time-shifted from a beat. We observe also that the vocal onset accuracy is on average a bit inferior to that

[7] We did a fork of the madmom toolbox https://github.com/CPJKU/madmom/, which we make available at https://github.com/georgid/madmom

[8] per-recoding results can be found in sheet 2 of https://tinyurl.com/kz4mpkz

with manual beat annotations (done with the time-redistribution weighting). All the experiments in this Section are to be presented in Dzhambazov et al. (2017).

5.4 Onset-aware lyrics-to-audio alignment

In the previous section we investigated the relation of metrical accents to the positions of vocal onsets. We proposed a method for automatic vocal onset detection in a way aware of metrical accents. Using as input the detected vocal onsets, in this chapter we propose a strategy to improve LAA by representing the interaction of vocal onsets with syllable transitions. In this way the influence of metrical events on syllable transitions is represented implicitly through its influence on vocal note onsets, which are in turn influenced by metrical events.

As we saw in the previous chapter, automatically determining the time positions of transitions between sung syllables can be greatly assisted by information from the music score. Similarly, by relying on music score, one can infer automatically the timestamps of vocal note onsets. Such timestamps are estimated reasonably well by a recent study on automatic score-to–audio alignment Şentürk (2016, chapter 6). In contrast, with the help of automatic singing voice transcription, vocal note onsets can be derived without the need for music score (Benetos et al., 2013). Since we intend that the proposed methodologies can be applicable for material with no music scores available, we preferred to apply automatic vocal onset detection instead of score-to-audio alignment. Detecting vocal onsets in any setting is arguably one of the hardest MIR problems. Still, for the study of onset-aware phoneme transitions, it is important that onsets timestamps are detected as accurately as possible. To assure better accuracy, experiments in this section are conducted only on *a cappella* material from OTMM, as well as with manually annotated onsets.

An overview of the proposed approach is presented in Figure 5.2. As in all approaches presented in this thesis, first an audio recording is manually divided into segments according to the coarse level complementary context — the sections of the composition. The boundaries of vocal section (one of *zemin*, *nakarat*, *meyan*) are taken from manual annotations. An audio recording and its corresponding lyrics are input. The vocal note onsets (automatically detected or manually annotated) together with phoneme transition rules are fed as input to the transition model. The phonetic recognizer, guided by the phoneme transition rules, returns the start and end timestamps of the aligned phonemes.

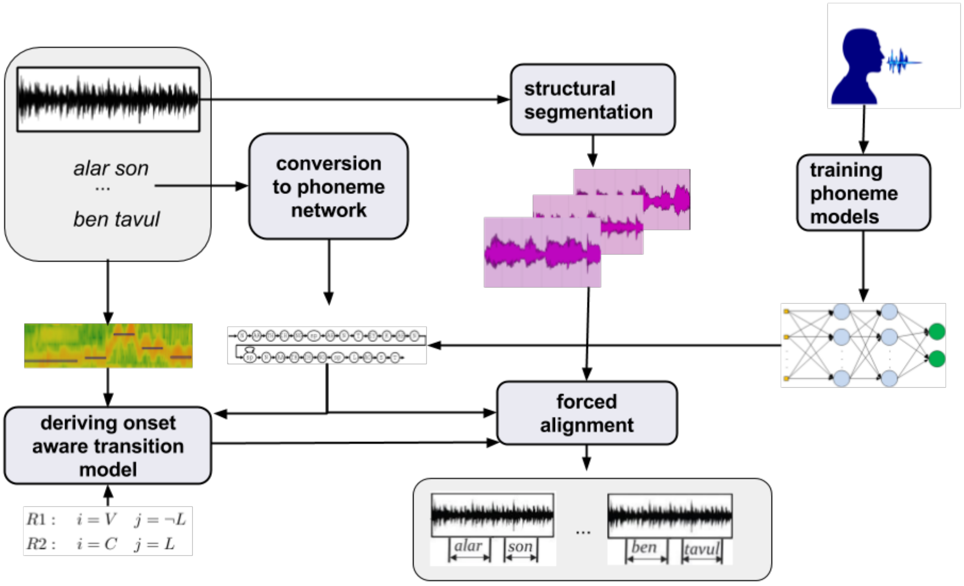


Figure 5.2: Overview of the modules of the proposed approach. The transition model is derived from phoneme transition rules and onset positions from the singing voice transcription. Then it input to the phonetic recognizer, together with the phonemes network and the features, extracted from audio segments.

5.4.1 Phoneme transition rules

The transition to a consecutive lyrics syllable implies a concurrent transition to a new note. The onset of the new note occurs usually at the start of the first voiced sound in the syllable. If we look at this reversely, the occurrence of note attack in a sung melody can signal a phonetic transition. The transition depends on the phoneme types, since, for example, a new note cannot start at unvoiced consonants. Taking advantage of that fact, we formulate rules that guide the transition between consecutive phonemes when a note onset is present. In general, we consider note onsets (attack) events as complementary context of phonetic timbre. Similar phoneme transitions rules have been used successfully to enhance the perceived naturalness of synthesized singing voice (Sundberg, 2006). The onset-aware phoneme transitions rules, we designed, have been presented in Dzhambazov et al. (2016).

We formalize transition rules described in this Section for Turkish, in which each syllable has exactly one vowel. In this sense, the rules could be transferred to another language with single-vowel syllables[9].

Let V denote a vowel, C denote a consonant and L denote a vowel, liquid (LL, M, NN) or the semivowel Y. Rules $R1$ and $R2$ represent inter-syllable transition, e.g. phoneme i is followed by phoneme j from the following syllable:

$$\begin{aligned} R1: &\quad i = V \quad j = \neg L \\ R2: &\quad i = C \quad j = L \end{aligned} \tag{5.11}$$

For example, for rule $R2$ if a syllable ends in a consonant, a note onset imposes with high probability that a transition to the following syllable is done, provided that it starts with a vowel. The same rule applies if it starts with a liquid, according to the observation that pitch change takes place during a liquid preceding the vowel Sundberg (2006, timing of pitch change). Rule R2 is valid also for intra-syllabic phoneme patterns, together with rule R3:

$$R3: \quad i = V \quad j = C \tag{5.12}$$

Essentially, if the current phoneme is vocal and the next one is non-voiced (e.g. $R1$, $R3$) the transition to the next is discouraged. An example of the intra-syllable $R2$ can be seen for the syllable KK-AA in Figure 5.3 where the note onset triggers the change to the vowel AA. Unlike that, an onset for example, to the syllable Y to onset at Y for the syllable Y-E-T.

5.4.2 Transition model

The phoneme transitions are dependent on the current vocal note temporal segment. When a note is in its onset segment, the transition between phonemes could be conditioned differently, compared to when a note is in a another segment. A crucial limitation of the phonetic recognizer HMM is the single latent variable, which can represent only one music facet - phonetic timbre. To represent the influence of events of different music facets, such as vocal note segments, one can use the hidden variables in a DBN (see in Figure 5.4).

[9] Among single-vowel syllabic languages are also Japanese and to some extent Italian

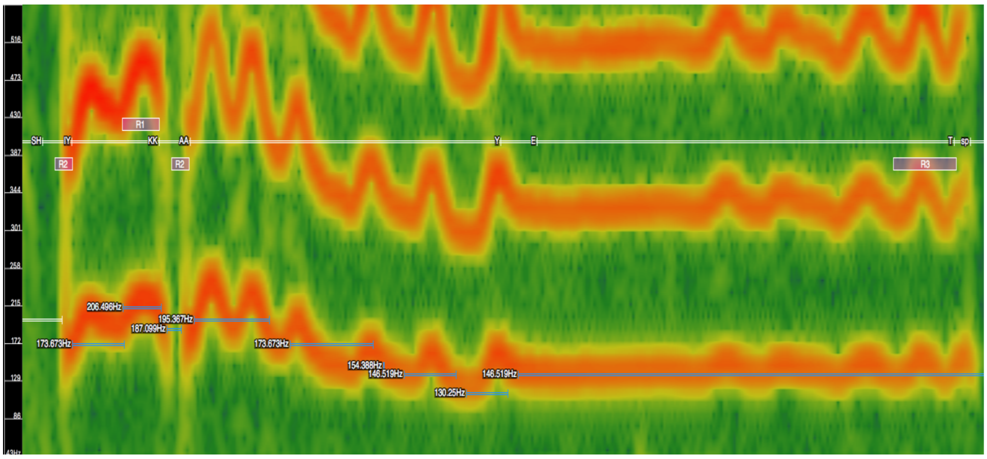


Figure 5.3: Ground truth annotation of syllables (in orange/top), phonemes (in red/middle) and notes (with blue/changing position). Audio excerpt corresponding to the word şikayet with syllables SH-IY, KK-AA and Y-E-T.

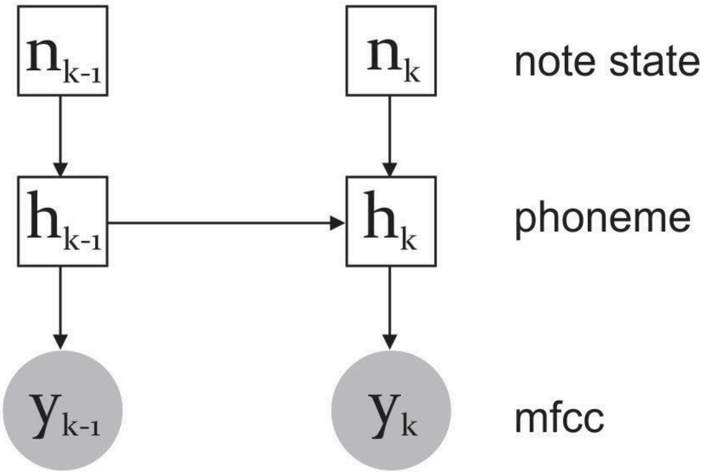


Figure 5.4: A DBN for the simultaneous music note and phoneme states. Circles and squares denote continuous and discrete variables, respectively. Gray nodes and white nodes represent observed and hidden variables, respectively. A phoneme transition is conditioned on the vocal note state. If a note onset is present the likelihood of transition is modified according to what the current h_{k-1} and its following h_k phoneme are.

For particular states, transitions are modified depending on the presence of time-adjacent note onset. Let k' be the timestamp of the onset $\Delta n_{k'} = 1$, which is closest to given time k. Now the transition probability can be rewritten as

$$a_{ij}(k) = \begin{cases} a_{ij} - g(k,k')q, & R1 \,\text{or}\, R3 \\ a_{ij} + g(k,k')q, & R2 \end{cases} \tag{5.13}$$

$R1$ to $R3$ stand the phoneme transition rules, which are applied in the phonemes network by picking the states i and j for two consecutive phonemes. The term q is a constant whereas $g(k,k')$ is a weighting factor sampled from a normal distribution with its peak (mean) at k':

$$g(k,k') = \begin{cases} f(k;k',\sigma^2) \sim \mathcal{N}(k',\sigma^2), & |k-k'| \leq \sigma \\ 0 & else \end{cases} \tag{5.14}$$

Since singing voice onsets are regions in time, they span over multiple consecutive frames. To reflect that fact, $g(k,k')$ serves to smooth in time the influence of the discrete detected Δn_k, where σ has been selected to be 0.075 seconds. In this way an onset influences a region of 0.15 seconds - a value we found empirically to be the optimal. Furthermore, this allows to handle slight timestamp inaccuracies of the estimated note onsets.

5.4.3 Inference

The most likely state sequence is found by means of a forced alignment Viterbi decoding. Similarly to the inference for metrical-accent-aware detection of vocal onsets (see Section 5.3.6), we apply a VTHMM. For the sake of brevity we will refer to the onset-aware alignment model as VTHMM. The standard transition probabilities in the Viterbi maximization step in Eq.2.6 are substituted for the onset-aware transitions $a_{ij}(k)$ from Eq.5.13:

$$\delta_k(j) = \max_{i \in (j,j-1)} \delta_{k-1}(i)\, a_{ij}(k)\, b_j(O_k) \tag{5.15}$$

5.4.4 With automatically detected onsets

To obtain reliable estimate of singing note onsets, we adapt the automatic singing transcription method, developed for polyphonic flamenco recordings Kroher and Gómez (2016). It has been designed to handle singing with high degree of vocal pitch embellishments. We expect that this made it suitable for material from OTMM singing, which has many embellishments, too[10]. We replace the predominant vocal extraction method with the OTMM-tailored pitch detection method of Atlı et al. (2014), which we described in Section 3.3.2.

The algorithm of Kroher and Gómez (2016) considers two cases of onsets: interval onsets and steady pitch onsets. A Gaussian derivative filter detects interval onsets as long-term change of the pitch contour, whereas steady-pitch onsets are inferred from pitch discontinuities. Since phoneme transitions are modified only when onsets are present, we opt for increasing recall at the cost of losing precision. This is achieved by reducing the value of the parameter cF: the minimum output of the Gaussian filter. Since the algorithm cannot be integrated easily in an HMM, note onset segmentation is performed as a preprocessing step to the actual alignment. The extracted note onsets are converted, as in the case of manually annotated onsets, to a binary onset activation at each frame $\Delta n_t = (0, 1)$.

5.4.5 Experiments

LAA is evaluated on the 6-recording subset of the *a cappella lyrics OTMM dataset* (see Section 3.2.3), for which vocal onsets have been annotated. *A cappella* was preferred because of the very low vocal onset detection accuracy on instrumentally-accompanied singing. Experiments are executed with the MLP-DirectM (direct mapping to the *MLP-English* acoustic model from Section 3.4.2).

Evaluation metrics

Alignment accuracy is measured as the percentage of duration of correctly aligned words from the total audio duration (see Figure 2.1 for an example). Unlike previous chapters, in this experiment we preferred to measure accuracy at the finer level of words, since looking at boundaries of phrases, we could potentially miss certain onset locations with improvement over the baseline. To this end, we annotated also word boundaries in the 6-recording subset.

We measured vocal onset accuracy in terms of recall. Similarly to the experiments on vocal onset detection from the previous Section 5.3, we adopted the tolerance of $t = 50\,ms$.

With manually annotated onsets

Unfortunately, as we saw in Section 5.2 note onsets could not be estimated from polyphonic recordings with high accuracy. To assure reasonable accuracy,

[10] We preferred it, because preliminary experiments showed that with default parameters it outperforms the algorithm of Mauch et al. (2015) with default parameters, which we extended in Section 5.3

cF	5	4.5	4.0	3.5	3.0
OR	57.2	59.7	66.8	72.3	73.2
AA	78.1	79.1	81.5	81.7	81.2

Table 5.2: VTHMM performance, depending on the sensitivity parameter cF. Vocal onset recall (OR) and alignment accuracy (AA) are reported as a total for all the recordings.

we utilized firstly manually annotated note onsets. This is done to test the general feasibility of the proposed model, unbiased from errors in the note segmentation algorithm, and to set a glass-ceiling alignment accuracy.

As a baseline we conduct alignment with unaffected phoneme transition probabilities, e.g. setting all $\Delta n_t = 0$, which is equivalent to the baseline, presented in Section 3.5. This resulted in average alignment accuracy of 79.2 %. We have tested with different values of q from Eq.5.13 achieving best accuracy of 82.5% at $q = 0.23$, which is used on in the next experiment, too[11].

With automatically detected onsets

We measured the impact of the note segmentation approach of Kroher and Gómez (2016) (introduced in Section 5.2), varying onset detection recall by changing the minimum output of the Gaussian filter (controlled by the parameter cF). Table 5.2 summarizes the alignment accuracy with the VTHMM depending on recall. On *a cappella* best improvement over the baseline is achieved at recall of 72.3% (at $cF = 3.5$). This is though still much lower than the best recall of 81-84% achieved for flamenco Kroher and Gómez (2016). Setting recall higher than that degraded performance probably because there are too many false alarms, resulting in forcing false transitions.

Figure 5.5 allows a glance at results at the level of detected phonemes: the baseline HMM switches to the following phoneme after some amount of time, relatively similar for all phonemes. One reason for this might be that the waiting time in a state in HMMs with a fixed transition matrix cannot be too long Yu (2010). In contrast, for VTHMM the presence of note onsets at vowels activates rules $R1$ or $R3$, which allows waiting in the same state longer, as there are more onsets (for example AA from the word SH-IY-KK-AA-Y-E-T has five associated onsets). We chose to modify cF because setting it to lower values increases the recall of the *interval onsets* only. Often in our dataset several consecutive notes with different pitch correspond to the same vowel. In fact, due to some characteristic for OTMM descending/ascending melody

[11] Per-recording results can be found at https://tinyurl.com/ksqsqla

progressions, a single syllable may happen to span many notes (up to 12 in our dataset) (Ederer, 2011). However, for cases of vowels held long on same pitch, conceptually VTHMM is not capable of bringing any benefit. This is illustrated in Figure 5.5 by the prematurely detected end boundary of E from the word SH-IY-KK-AA-Y-E-T. Although no separate experiment for each rule was made, inspection of particular cases revalued almost no contribution of *R2*, supposedly due to the difficulty of detecting onsets are syllables starting with unvoiced consonants.

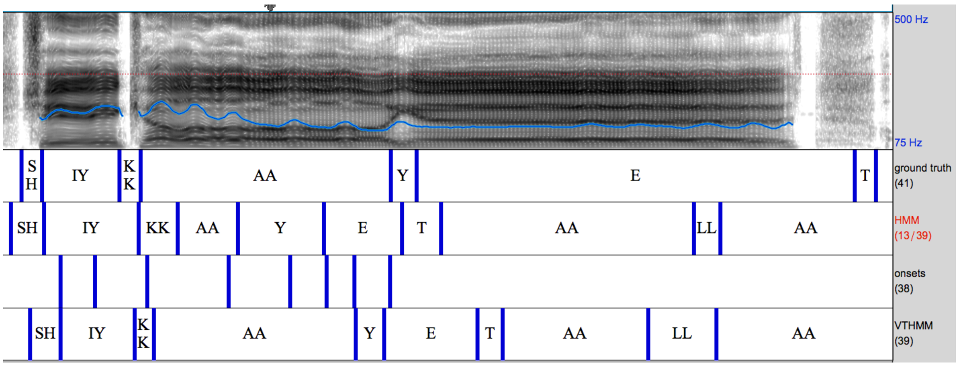


Figure 5.5: An example of boundaries of phonemes for the word *şikayet* (SH-IY-KK-AA-Y-E-T): *on top*: spectrum and pitch; *then from top to bottom*: ground truth boundaries, phonemes detected with the HMM, detected onsets, phonemes detected with VTHMM; (excerpt from the recording *Kimseye etmem şikayet* (in original sung by Bekir Unluater).

5.5 Summary

In this chapter we assessed the contribution of explicitly representing metrical accents (fine-level complementary context) for improving the tracking of sung lyrics. We studied the relation of metrical accents to lyrics in two steps: how metrical accents interact with vocal onsets and how the latter, in turn, interact with phoneme transitions. 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. Therefore, we presented two separate probabilistic models for two separate tasks: metrical-accent-aware vocal onset detection and onset-aware lyrics-to-audio alignment. We carry out an evaluation on material from OTMM.

Metrical-accent-aware vocal onset detection We strived to improve the automatic vocal note onset detection by incorporating information about their position in a metrical cycle (i.e. metrical accents). To this end we proposed a DBN for the simultaneous tracking of metrical position and vocal onsets. The main contribution is that the approach integrates in one coherent model two existing state of the art probabilistic approaches for different tasks: beat tracking and singing voice transcription. We carried out an evaluation on a multi-instrument dataset from OTMM with two different usul (meter) types. Context knowledge about the usul is built within the transition model of the DBN. Results confirmed that the proposed model reasonably improves vocal note onset detection accuracy compared to a baseline model that does not take the metrical position into account. The F-measure rises from 31% to 36 % for the düyek usul, which has better beat detection F-measure and from 38 to 40 % for aksak usul.

Detecting vocal onsets is polyphonic audio is arguably one of the hardest MIR problems. Although not the goal of this thesis, the presented DBN can be used for full-fledged singing voice transcription.

Onset-aware lyrics-to-audio alignment. We extended the phonetic recognizer approach by modeling the singing voice onsets, occurring simultaneously with phoneme transitions. We conceptualized onset-aware phoneme transition rules and proposed how to integrate them into the transition model of the phonetic recognizer. The method was tested on an *a cappella* OTMM dataset. The new model resulted in a slight improvement of from 79.2 % for baseline, unaware of singing voice onsets, to 81.7 %. In particular, the onsets due to rules discouraging premature transition, the states of sustained vowels were allowed to have longer durations. This is, to our knowledge, the first attempt to model explicitly onsets from the vocal melody in the LAA decoding process itself.

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 LAA and evaluated its performance by the accuracy of the LAA. As a baseline we chose a phonetic recognizer based on HMM-s: a methodology applied in most of existing computational studies on lyrics tracking. We applied the proposed methodologies on especially compiled for this study datasets that are subsets from the CompMusic research corpora on OTMM and jingju. 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 conventional HMM-s 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 mid-level complementary context and a separate one for fine-level context. As mid-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.

6.1 Importance of complementary context

We represented events from complementary context as components in a DBN and their influence on the lyrics as a hierarchical dependence between the components. The presented solutions provide an alternative to the prevailing music-knowledge-uninformed approach to modeling lyrics, in which the extracted phonetic timbre features are agglomerated in a bottom-up fashion.

6.1.1 Mid-level context

We first proposed a phonetic recognizer that utilizes lyrics duration information as a cue, complementary to phonetic timbre. It is representing how the position of a syllable in a lyrics phrase influences its duration. An advantage of the model is that it allows room for certain temporal flexibility to handle cases of significant deviation of sung vowels from the expected reference durations. Evaluation showed that syllable durations is the facet of complementary context with biggest contribution to the improvement of the baseline LAA (up to absolute 10 %). For OTMM, despite the accuracy of around 90% for *a cappella* singing, LAA of material with instrumental accompaniment remains somewhat lower and still far from industry-ready order of performance.

For jingju the relative improvement was somewhat bigger than for OTMM. One explanation is the very long durations of sung vowels in jingju, which is a challenge to conventional HMM-s.

6.1.2 Fine-level context

In this thesis we focused on one particular fine-level facet — the accents in the metric cycle. We studied the relation of metrical accents to lyrics in two steps: how metrical accents interact with vocal onsets and how the latter, in turn, interact with phoneme transitions. Therefore, we devised two separate probabilistic models for two separate tasks: vocal-onset-aware lyrics-to-audio alignment and metrical-accent-aware vocal onset detection. We tested the model on recordings from OTMM. Results confirmed that its well-grounded rhythmic framework provides an excellent piece of music knowledge.

For vocal-onset-aware lyrics-to-audio alignment we conceptualized phoneme transition rules that consider the presence of vocal note onsets. We integrated these into the transition model of the phonetic recognizer. Results showed that the improvement of accuracy is not very substantial, even with manually annotated onsets (around 3 absolute %). In fact, for particular cases (for example vowels held long on the same pitch) onsets are not conceptually capable of bringing any benefit. However, we believe that, the derived phoneme transition rules are an important linguistic contribution that can be exploited in other singing styles, because the rules could be easily transferred to languages, other than Turkish.

A limitation of the syllable-duration-aware model is the requirement for external source of syllable reference durations — for example the music scores. In contrast, the onset-aware alignment is not dependent on external sources, since the onsets are automatically extracted. Based on evidence that in OTMM the position of note events in vocal melodies is influenced by the position in a metrical cycle, we designed a model for simultaneously tracking vocal onsets and metrical accents. Vocal onset detection in multi-instrumental music occurred, in fact, to be one of the hardest MIR problems (scoring in the order of 35-40 % f-measure). It is arguably even harder in OTMM because of the expressive singing style: vocal onsets are often approached by portamentos. The complementary metrical accent context proved to be an important 'stepping stone': the accuracy of vocal onset detection was increased reasonably for two different usul types. We believe that the biggest potential of the model lies in its generalisibility — applying it to singing material with different singing style and meter is as easy as tuning its parameters.

The most important advantage of the metric-accent models is that they do not necessarily depend on external sources of information such as music scores.

6.2 Summary of contributions

A summary of the specific contributions from the work presented in the dissertation are listed below.

6.2.1 Scientific contributions

We hope that the outcomes of this work will motivate researchers to use more often music context knowledge in future work. Some particular contributions are:

- We showed that a model of complementary context can be adapted to a different music tradition (the syllable-duration-aware model has been applied to two different traditions). Both the temporal structure and the metrical cycle are facets, characteristic for many music tradition. This means that transferring the model to another music style is a matter of compacting the music knowledge context into an appropriate set of rules/patterns.

- We conceptualized the interaction of phoneme transitions to other musical facets. These interactions were represented as hidden variables and their dependences in DBNs. DBNs are an elegant modeling tool (we presented illustrated the model dependencies in diagrams).

- Inference in DBNs is computationally demanding. Therefore, we proposed several implementation simplifications.

6.2.2 Other contributions

- We compiled several datasets of OTMM and jingju with annotations of different music facets including lyrics, vocal sections, onsets of singing voice, beats.
- The most successful LAA approach developed, the syllable-duration-aware LAA, was integrated into Dunya-web. It can enable musicologists to track not only the aligned lyrics, but also complementary musical facets and music-specific phenomena.
- All the methodologies presented in this thesis are implemented as modular and easy-to-extend software. A special focus has been put on making them reproducible. To our knowledge this is the first open source software for lyrics-to-audio alignment that is based on computational study.

6.3 Future directions

In chapter 5, the relation of metrical accents to lyrics is not modeled directly. Instead, we built two separate DBNs: one for the relation of metrical accents to the positions of onsets (attacks) of sung notes (Section 5.3) and one for the relation of phoneme transitions on vocal onsets (Section 5.4). 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. The two models can be combined in the future by adding to the vocal note state detection DBN (Figure 5.1) a hidden state for the phoneme state that is dependent on the vocal note state. This dependence is presented in the DBN in Figure 5.4.

In fact, for cases of vowels held long on the same pitch, conceptually the presence of the onset is not capable of bringing any benefit. Same pitched long vowels can be handled by the syllable-duration-aware model. In this respect the models aware of different context facets of chapters 4 and 5 complement each other. Therefore, we expect that in the future it will be beneficial that they are combined into one.

We believe that the methods presented in this dissertation generalize to any musics, which share principles akin to these of the evaluated music traditions. Highly variable durations of sung syllables is common in some genres such as for example soul and jazz singing. Still, deviation from the reference note values in the musical score is acceptable only to a limited extent. This setting is similar to that of OTMM. Applying insights and methodologies from this culture-specific study can open up and make the existing computational methods more versatile. We hope that in the future researchers can apply and extend the outcomes of this work to improve and enrich existing MIR methods, thus fulfilling one of the ultimate goals of the CompMusic project (Serra et al., 2013).

A web page with links to materials accompanying this manuscript is available at http://compmusic.upf.edu/phd-thesis-georgi

Appendix A

Applications

Researchers of the CompMusic team have created a web application called Dunya-web[1] 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 datasets described in Section 3.2) and music scores, together with the lyrics.

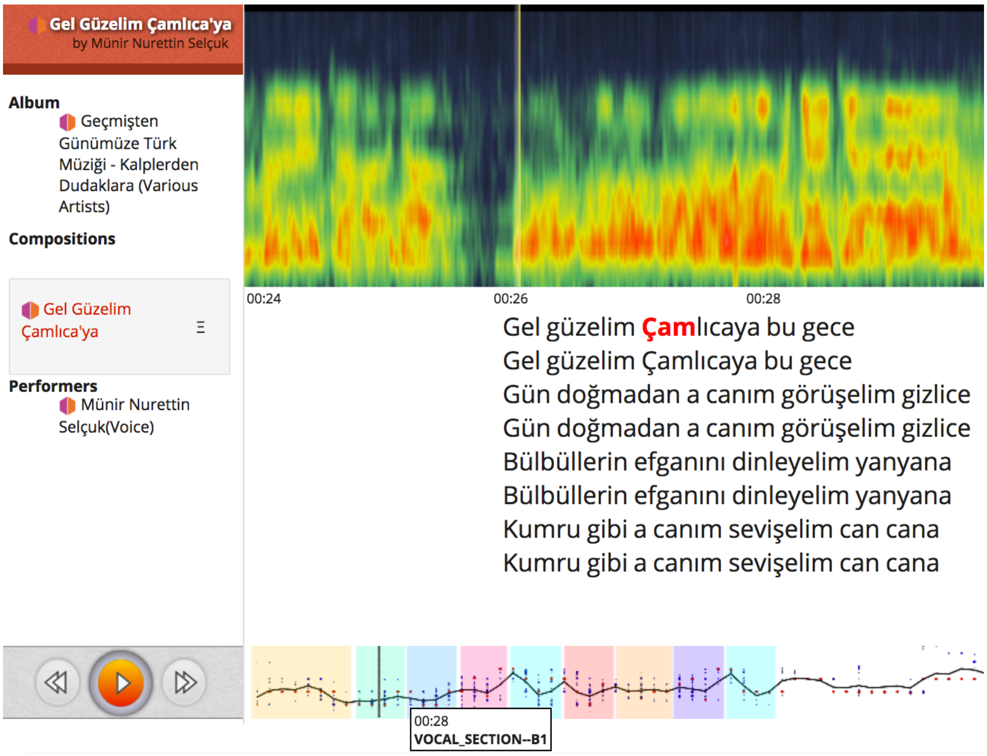
The users can navigate the audio collection by searching or filtering by recordings, compositions, artists, makams, musical forms and/or usuls. Users can play the recordings and examine musical facets synchronous to the audio playback. Musical facets like pitch, the score, the tonic are visualized in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

The most successful LAA approach, 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 şarkı 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, representing the timbral differences of phonemes are displayed. The MFCC feature vectors are hard to interpret visually. Instead, it is a common practice to invert the first 12 coefficients back to mel-bands domain for visualization[2].

[1] http://dunya.compmusic.upf.edu/makam

[2] MFCC are inverted by https://tinyurl.com/y9swfbhf



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)

Gel güzelim **Çam**lıcaya bu gece
Gel güzelim Çamlıcaya 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

Figure A.1: Dunya-web: an interface for the discovery of the music traditions of the world. The part on aligning automatically lyrics in vocal recordings of the OTMM şarkı form is presented.

Abbreviations

List of publications

The following is a list of peer-reviewed conference publications of the author, organized by the relevance to the thesis. An up-to-date list of all publications can be found at:

https://www.researchgate.net/profile/Georgi_Dzhambazov2

Publications in the context of the thesis and the CompMusic project

Georgi Dzhambazov and Xavier Serra. Modeling of phoneme durations for alignment between polyphonic audio and lyrics. In *Sound and Music Computing Conference 2015*, Maynooth, Ireland, 2015. URL `http://mtg.upf.edu/node/3266`.

Georgi Dzhambazov and Xavier Serra. Singing voice separation by harmonic modeling. In *Music Information Retrieval Evaluation eXchange (MIREX)*, 2016. URL `http://mtg.upf.edu/node/3565`.

Georgi Dzhambazov, Sertan Şentürk, and Xavier Serra. Automatic lyrics-to-audio alignment in classical Turkish music. In *Proceedings of 4th International Workshop on Folk Music Analysis (FMA 2014)*, pages 61–64, Istanbul, Turkey, 2014. URL `http://mtg.upf.edu/node/2965`.

Georgi Dzhambazov, Ajay Srinivasamurthy, Sertan Şentürk, and Xavier Serra. On the use of note onsets for improved lyrics-to-audio alignment in Turkish makam music. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, pages 716–722, New York, NY, USA, 2016a. URL `http://mtg.upf.edu/node/3492`.

Georgi Dzhambazov, Yile Yang, Rafael Caro Repetto, and Xavier Serra. Automatic alignment of long syllables in a cappella beijing opera. In *Proceedings of 6th International Workshop on Folk Music Analysis (FMA 2016)*, pages 88–91, Dublin, Ireland, 15/06/2016 2016b. URL `http://mtg.upf.edu/node/3517`.

Georgi Dzhambazov, Andre Holzapfel, Ajay Srinivasamurthy, and Xavier Serra. Metrical-accent aware vocal onset detection in polyphonic audio. *arXiv preprint arXiv:1707.06163*, 2017. URL `http://mtg.upf.edu/node/3805`.

Publications within the CompMusic project, which are outside the context of the thesis

Georgi Dzhambazov, Sertan Şentürk, and Xavier Serra. Searching lyrical phrases in a-capella Turkish makam recordings. In *Proceedings of 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, pages 687–693, 2015. URL http://mtg.upf.edu/node/3321.

Rong Gong, Nicolas Obin, Georgi Dzhambazov, and Xavier Serra. Score-informed syllable segmentation for jingju a cappella singing voice with mel-frequency intensity profiles. In *Proceedings of 7th International Workshop on Folk Music Analysis (FMA 2017)*, Malaga, Spain, 14/06/2017 2017. doi: https://doi.org/10.5281/zenodo.556820. URL http://mtg.upf.edu/node/3732.

Publication outside the CompMusic project, relevant to an extent for the thesis

Georgi Dzhambazov. Towards a drum transcription system aware of bar position. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014. URL https://www.researchgate.net/publication/283599561_Towards_a_drum_transcription_system_aware_of_bar_position.

Georgi Dzhambazov and Rolf Bardeli. Automatic sentence boundary detection for German broadcast news. In *Proceedings of 10. ITG Symposium on Speech Communication*, pages 1–4. VDE, 2012. URL https://www.researchgate.net/publication/260133087_Automatic_Sentence_Boundary_Detection_for_German_Broadcast_News.

Bibliography

Anguera, Xavier, Jordi Luque, and Ciro Gracia (2014). Audio-to-text alignment for speech recognition with very limited resources. In *INTERSPEECH*, pp. 1405–1409. 22

Atlı, Hasan Sercan, Burak Uyar, Sertan Şentürk, Barış Bozkurt, and Xavier Serra (2014). Audio feature extraction for exploring Turkish makam music. In *Proceedings of 3rd International Conference on Audio Technologies for Music and Media (ATMM 2014)*, Ankara, Turkey, pp. 142–153. 4, 6, 41, 77, 88

Babacan, Onur, Thomas Drugman, Nicolas d'Alessandro, Nathalie Henrich, and Thierry Dutoit (2013). A comparative study of pitch extraction algorithms on a large variety of singing sounds. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7815–7819. IEEE. 69

Barber, David, Ali Taylan Cemgil, and Silvia Chiappa (2011). *Bayesian time series models*. Cambridge University Press. 31

Benetos, Emmanouil, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri (2013). Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems 41*(3), 407–434. 69, 70, 83

Bozkurt, Barış, Ruhi Ayangil, and Andre Holzapfel (2014). Computational analysis of Turkish makam music: Review of state-of-the-art and challenges. *Journal of New Music Research 43*(1), 3–23. 15

Caro Repetto, Rafael, Rong Gong, Nadine Kroher, and Xavier Serra (2015). Comparison of the singing style of two jingju schools. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, pp. 507–513. 17

Caro Repetto, Rafael and Xavier Serra (2014). Creating a corpus of jingju (beijing opera) music and possibilities for melodic analysis. In *Proceedings of* *the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, pp. 313–318. 38

Chandna, Pritish, Marius Miron, Jordi Janer, and Emilia Gómez (2017). Monoaural audio source separation using deep convolutional neural networks. In *International Conference on Latent Variable Analysis and Signal Separation*, pp. 258–266. Springer. 42

Chang, Sungkyun and Kyogu Lee (2017). Lyrics-to-audio alignment by unsupervised discovery of repetitive patterns in vowel acoustics. *arXiv preprint arXiv:1701.06078*. 20, 33

Chen, Ruofeng, Weibin Shen, Ajay Srinivasamurthy, and Parag Chordia (2012). Chord recognition using duration-explicit hidden markov models. In *Proceedings of the 13th International Society for Music Information Retrieval Conference*, pp. 445–450. 57, 58

Cooke, Peter (accessed April 5, 2013). Heterophony. Grove Music Online. `http://www.oxfordmusiconline.com/subscriber/article/grove/music/12945`. 6, 16

D'Ambrosio, Bruce (1999). Inference in Bayesian networks. *AI magazine 20*(2), 21–36. 31

Degara, Norberto, Antonio Pena, Matthew EP Davies, and Mark D Plumbley (2010). Note onset detection using rhythmic structure. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5526–5529. IEEE. 72

Durga, SAK (1978). Voice culture-with special reference to south indian music. *Journal of the Indian Musicological Society 9*(1), 5. 1

Dzhambazov, Georgi, Andre Holzapfel, Ajay Srinivasamurthy, and Xavier Serra (2017). Metrical-accent aware vocal onset detection in polyphonic audio. *arXiv preprint arXiv:1707.06163*. 82

Dzhambazov, Georgi, Sertan Şentürk, and Xavier Serra (2014). Automatic lyrics-to-audio alignment in classical Turkish music. In *Proceedings of 4th International Workshop on Folk Music Analysis (FMA 2014)*, Istanbul, Turkey, pp. 61–64. 40

Dzhambazov, Georgi and Xavier Serra (2015). Modeling of phoneme durations for alignment between polyphonic audio and lyrics. In *Proceedings of Sound and Music Computing Conference 2015 (SMC 2015)*, Maynooth, Ireland. 57

Dzhambazov, Georgi and Xavier Serra (2016). Singing voice separation by harmonic modeling. In *Proceedings of Music Information Retrieval Evaluation eXchange (MIREX)*, New York, NY, USA. 42

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 *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, New York, NY, USA, pp. 716–722. 84

Dzhambazov, Georgi, Yile Yang, Rafael Caro Repetto, and Xavier Serra (2016). Automatic alignment of long syllables in a cappella Beijing opera. In *Proceedings of 6th International Workshop on Folk Music Analysis (FMA 2016)*, Dublin, Ireland, pp. 88–91. 63

Ederer, Eric Bernard (2011). *The Theory and Praxis of Makam in Classical Turkish Music 1910–2010.* University of California, Santa Barbara. 4, 7, 9, 15, 16, 90

Ferguson, Jack D (1980). Variable duration models for speech. In *Symposium on the Application of Hidden Markov Models to Text and Speech, 1980*, pp. 143–179. 56

Fujihara, Hiromasa and Masataka Goto (2012). Lyrics-to-audio alignment and its application. *Dagstuhl Follow-Ups 3.* 6, 11, 19, 21

Fujihara, Hiromasa, Masataka Goto, Jun Ogata, and Hiroshi G Okuno (2011). Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics. *IEEE Journal of Selected Topics in Signal Processing 5*(6), 1252–1261. xx, 11, 20, 23, 24, 27, 28, 32, 47, 52, 53, 62, 63

Fujihara, Hiromasa, Masataka Goto, and Hiroshi G Okuno (2009). A novel framework for recognizing phonemes of singing voice in polyphonic music. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '09)*, pp. 17–20. IEEE. 33

Gong, Rong, Nicolas Obin, Georgi Dzhambazov, and Xavier Serra (2017, 14/06/2017). Score-informed syllable segmentation for jingju a cappella singing voice with mel-frequency intensity profiles. In *Proceedings of 7th International Workshop on Folk Music Analysis (FMA 2017)*, Malaga, Spain. 17

Goto, Masataka (2014). Singing information processing. In *12th International Conference on Signal Processing (ICSP)*, pp. 2431–2438. IEEE. 2, 6

Gulati, Sankalp (2016). *Computational Approaches for Melodic Description in Indian Art Music Corpora.* Ph. D. thesis, Universitat Pompeu Fabra, Barcelona. 9

Hansen, Jens Kofod (2012). Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients. In *Proceedings of the 9th Sound and Music Computing Conference*, Copenhagen, Denmark, pp. 494–499. 29, 30

Holzapfel, Andre (2015). Relation between surface rhythm and rhythmic modes in turkish makam music. *Journal of New Music Research 44*(1), 25–38. 4, 16, 72, 78

Holzapfel, Andre and Thomas Grill (2016). Bayesian meter tracking on learned signal representations. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, pp. 262–268. 74

Holzapfel, Andre, Florian Krebs, and Ajay Srinivasamurthy (2014). Tracking the "odd": Meter inference in a culturally diverse music corpus. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, Taipei, Taiwan, pp. 425–430. 2, 38, 71, 77, 78, 79, 80, 81

Huron, David Brian (2006). *Sweet anticipation: Music and the psychology of expectation*. MIT press. 72

Johnson, Michael T (2005). Capacity and complexity of HMM duration modeling techniques. *Signal Processing Letters, IEEE 12*(5), 407–410. 79

Karaosmanoğlu, M Kemal (2012). A Turkish makam music symbolic database for music information retrieval: Symbtr. In *Proceedings of the 13th International Society for Music Information Retrieval Conference*, Porto, Portugal. 36

Karaosmanoğlu, M Kemal, Barış Bozkurt, Andre Holzapfel, and Nilgün Doğrusöz Dişiaçık (2014). A symbolic dataset of Turkish makam music phrases. In *Fourth International Workshop on Folk Music Analysis (FMA2014)*. 15, 36

Klapuri, Anssi and Manuel Davy (2006). *Signal Processing Methods for Music Transcription*. Springer. 70

Koller, Daphne and Nir Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press. 31

Korzeniowski, Filip (2011). Real-time capable singer tracking using pitch and lyrics information. Master's thesis. 33

Krebs, Florian, Sebastian Böck, and Gerhard Widmer (2013). Rhythmic pattern modeling for beat and downbeat tracking in musical audio. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curitiba, Brazil. 77

Krebs, Florian, Sebastian Böck, and Gerhard Widmer (2015, October). An Efficient State-Space Model for Joint Tempo and Meter Tracking. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, Malaga, Spain, pp. 72–78. 71, 72, 74, 75

Kroher, Nadine and Emilia Gómez (2016). Automatic transcription of flamenco singing from polyphonic music recordings. *IEEE Transactions on Audio, Speech and Language Processing 24*(5), 901–913. 70, 81, 87, 88, 89

Kruspe, Anna M (2014). Keyword spotting in a-capella singing. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, pp. 271–276. 6

Kruspe, Anna M (2015a). Keyword spotting in singing with duration-modeled HMMs. In *In Proceedings of 23rd European Signal Processing Conference (EUSIPCO)*, pp. 1291–1295. IEEE. 56

Kruspe, Anna M (2015b). Training phoneme models for singing with "songified" speech data. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*. 29, 48, 51

Kruspe, Anna M (2016). Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing. In *Proceedings of 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, New York, NY, USA. xx, 23, 29, 48, 51

Ladefoged, Peter (1996). *Elements of acoustic phonetics*. University of Chicago Press. 25

Lee, Kyogu and Markus Cremer (2008). Segmentation-based lyrics-audio alignment using dynamic programming. In *Proceedings of 9th International Society for Music Information Retrieval Conference (ISMIR 2008)*, pp. 395–400. 32, 34

Levy, Mark and Mark Sandler (2008). Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech, and Language Processing 16*(2), 318–326. 20

Mauch, Matthias (2010). *Automatic Chord Transcription from Audio Using Computational Models of Musical Context*. Ph. D. thesis, Queen Mary University of London. 4

Mauch, Matthias, Chris Cannam, Rachel Bittner, George Fazekas, Justin Salamon, Jiajie Dai, Juan Bello, and Simon Dixon (2015). Computer-aided melody note transcription using the Tony software: Accuracy and efficiency. In *Proceedings of the First International Conference on Technologies for Music Notation and Representation (TENOR 2015)*, pp. 23–30. xxi, 70, 74, 75, 77, 79, 80, 81, 88

Mauch, Matthias, Hiromasa Fujihara, and Masataka Goto (2012). Integrating additional chord information into HMM-based lyrics-to-audio alignment. *IEEE Transactions on Audio, Speech, and Language Processing 20*(1), 200–210. 20, 21, 32, 34

McVicar, Matt, Daniel PW Ellis, and Masataka Goto (2014). Leveraging repetition for improved automatic lyric transcription in popular music. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3117–3121. IEEE. 19

Mesaros, Annamaria and Tuomas Virtanen (2008). Automatic alignment of music audio and lyrics. In *Proceedings of the 11th International Conference on Digital Audio Effects (DAFx-08).* xx, 19, 23, 24, 27, 28, 52, 53, 62, 63

Mesaros, Annamaria and Tuomas Virtanen (2010). Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing 2010*(1), 546047. 19

Molina, Emilio, Lorenzo J Tardón, Isabel Barbancho, and Ana M Barbancho (2014). The importance of f0 tracking in query-by-singing-humming. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, pp. 277–282. 69

Murphy, Kevin Patrick (2002). *Dynamic bayesian networks: representation, inference and learning.* Ph. D. thesis, University of California. 9, 31, 57

Nichols, Eric, Dan Morris, Sumit Basu, and Christopher Raphael (2009). Relationships between lyrics and melody in popular music. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pp. 471–476. 55

Nishikimi, Ryo, Eita Nakamura, Katsutoshi Itoyama, and Kazuyoshi Yoshii (2016). Musical note estimation for F0 trajectories of singing voices based on a bayesian semi-beat-synchronous HMM. In *Proceedings of the 17th International Society for Music Information Retrieval Conference, (ISMIR 2016)*, pp. 461–467. 70, 76

Orio, Nicola and François Déchelle (2001). Score following using spectral analysis and hidden markov models. In *ICMC: International Computer Music Conference*, pp. 1–1. 33

Özgül Salor, Bryan L Pellom, Tolga Çiloğlu, and Mübeccel Demirekler (2007). Turkish speech corpora and recognition tools developed by porting sonic: Towards multilingual speech recognition. *Computer Speech and Language 21*(4), 580 – 593. 45, 48

Popescu-Judetz, Eugenia (1996). *Meanings in Turkish Musical Culture.* Istanbul: Pan Yayıncılık. 4, 15, 40

Porter, Alastair, Mohamed Sordo, and Xavier Serra (2013). Dunya: A system for browsing audio music collections exploiting cultural context. In *Proceedings of 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curitiba, Brazil, pp. 101–106. 97

Rabiner, Lawrence (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–286. 26, 27, 56

Rabiner, Lawrence and Biing-Hwang Juang (1993). *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc. 6, 8, 21, 25, 26, 27

Röbel, Axel (2009). Onset detection by means of transient peak classification in harmonic bands. In *Music Information Retrieval Evaluation eXchange (MIREX)*. 70

Ryynänen, Matti (2004). Probabilistic modelling of note events in the transcription of monophonic melodies. Master's thesis. 69

Salamon, Justin and Emilia Gómez (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing 20*(6), 1759–1770. 41, 70, 77

Salamon, Justin, Emila Gómez, Dan Ellis, and Gaël Richard (2014, 02/2014). Melody extraction from polyphonic music signals: Approaches, applications and challenges. *IEEE Signal Processing Magazine 31*, 118–134. 69

Schlüter, Jan (2016). Learning to pinpoint singing voice from weakly labeled examples. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, pp. 44–50. 25

Şentürk, Sertan (2016, December). *Computational Analysis of Audio Recordings and Music Scores for the Description and Discovery of Ottoman-Turkish Makam Music*. Ph. D. thesis, Universitat Pompeu Fabra, Barcelona, Spain. 3, 15, 83

Şentürk, Sertan, Andrés Ferraro, Alastair Porter, and Xavier Serra (2015). A tool for the analysis and discovery of Ottoman-Turkish makam music. In *Extended Abstracts for the Late Breaking Demo Session of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, Málaga, Spain. 97

Şentürk, Sertan, Andre Holzapfel, and Xavier Serra (2014). Linking scores and audio recordings in makam music of Turkey. *Journal of New Music Research 43*(1), 34–52. 4, 37, 40

Serra, Xavier (1989). A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition. Technical report. 23

Serra, Xavier (2011, October). A multicultural approach in Music Information Research. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, Miami, USA, pp. 151–156. 2

Serra, Xavier (accessed April 7, 2016). Harmonic model. Harmonic Model. `https://github.com/MTG/sms-tools/blob/master/lectures/06-Harmonic-model/6T1-Harmonic-model.odp`. 41

Serra, Xavier, Michela Magas, Emmanouil Benetos, Magdalena Chudy, Simon Dixon, Arthur Flexer, Emilia Gómez, Fabien Gouyon, Perfecto Herrera, Sergi Jorda, et al. (2013). Roadmap for music information research. 3, 96

Serra, Xavier and Julius Smith (1990). Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal 14*(4), 12–24. 40, 41

Sun, Lifa, Hao Wang, Shiyin Kang, Kun Li, and Helen Meng (2016). Personalized, cross-lingual tts using phonetic posteriorgrams. pp. 322–326. 27, 49, 50

Sundberg, Johan (2006). The KTH synthesis of singing. *Advances in Cognitive Psychology 2*(2-3), 131–143. 84, 85

Sundberg, Johan and Thomas D Rossing (1990). The science of singing voice. *Journal of the Acoustical Society of America 87*(1), 462–463. 2, 28, 69, 70

Uyar, Burak, Hasan Sercan Atlı, Sertan Şentürk, Barış Bozkurt, and Xavier Serra (2014). A corpus for computational research of Turkish makam music. In *1st International Digital Libraries for Musicology Workshop*, London, United Kingdom, pp. 57–63. 36

Wang, Ye, Min-Yen Kan, Tin Lay Nwe, Arun Shenoy, and Jun Yin (2004). Lyrically: automatic synchronization of acoustic musical signals and textual lyrics. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 212–219. ACM. 32

Whiteley, Nick, Ali Taylan Cemgil, and Simon Godsill (2006, October). Bayesian modelling of temporal structure in musical audio. In *Proceedings of the 7th International Society for Music Information Retrieval Conference (ISMIR 2006)*, Victoria, Canada, pp. 29–34. 31, 71

Wichmann, Elizabeth (1991). *Listening to theatre: the aural dimension of Beijing opera*. University of Hawaii Press. 7, 9, 17, 18, 65

Wiggins, Geraint A (2009). Semantic gap?? schematic schmap!! methodological considerations in the scientific study of music. In *Multimedia, 2009. ISM'09. 11th IEEE International Symposium on*, pp. 477–482. IEEE. 2

Yeh, Chunghsin and Axel Röbel (2009). The expected amplitude of overlapping partials of harmonic sounds. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3169–3172. IEEE. 23, 24

Young, Steve J (1993). *The HTK hidden Markov model toolkit: Design and philosophy*. 25, 44, 48, 66

Yu, Shun-Zheng (2010). Hidden semi-Markov models. *Artificial Intelligence 174*(2), 215–243. 56, 89