Automatic Assessment of Singing Voice Pronunciation:

A Case Study with Jingju Music

Rong Gong

TESI DOCTORAL UPF / 2018

Director de la tesi

Dr. Xavier Serra Casals
Music Technology Group
Dept. of Information and Communication Technologies

http://compmusic.upf.edu/phd-thesis-rgong

Music Technology Group (http://mtg.upf.edu/), Department of Information and Communication Technologies (http://www.upf.edu/etic), Universitat Pompeu Fabra (http://www.upf.edu), Barcelona, Spain

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

Dr. Xavier Serra Casals
(Thesis Supervisor)
Universitat Pompeu Fabra (UPF), Barcelona

Dr. Emilia Gómez Gutiérrez
(Thesis Committee Member)
Universitat Pompeu Fabra, Barcelona

Dr. Javier Hernando Pericas
(Thesis Committee Member)
Universitat Politècnica de Catalunya, Barcelona

Dr. Alexander Lerch
(Thesis Committee Member)
Georgia Institute of Technology, Atlanta

致我亲爱的父母. To my beloved parents.

This thesis has been carried out between Oct. 2015 and Sep. 2018 at the Music Technology Group (MTG) of Universitat Pompeu Fabra (UPF) in Barcelona (Spain), supervised by Dr. Xavier Serra Casals. This work has been supported by the Dept. of Information and Communication Technologies (DTIC) PhD fellowship (2015-18), Universitat Pompeu Fabra and the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583).

Acknowledgements

I am grateful to my stuttering – my lifelong companion, who teaches me compassion for the weak, patience, modesty and to never give up.

I am extremely grateful to my advisor Xavier Serra for the guidance, mentoring and support. His academic vision and humility made me admire. I thank him for the opportunity he gave me to be a part of the CompMusic family.

Three years went by in a blink of an eye. Thanks to those colleagues in the CompMusic family who supported me. My personal thanks go to Rafael Caro Repetto (the guru of jingju musicology), Sertan Şentürk (who always be patient and give good advice), Georgi Dzhambazov (the fellow lyrics-to-audio alignment researcher), Oriol Romaní (our dearest developer), Ajay Srinivasamurthy (a real scientist and thinker), Sankalp Gulati (fast and efficient executor), Yile Yang, Gopala Krishna Koduri, Swapnil Gupta, Barış Bozkurt, Błażej Kotowski, Alastair Porter, Hasan Sercan Atlı, Honglin Ma. I also thank the collaborators in NACTA Jin Zhang, Xiaoyu Zhang, Ruoxuan Song, Hao Tian, Jiani Liao and Yuzhu Sun for helping us collect the precious jingju singing recordings.

My colleagues at MTG gave me a lot of thoughts and ideas. I'd like to thank its past and present members: Emilia Gómez, Perfecto Herrera, Jordi Pons (my deep learning master), Eduardo Fonseca (my best workmate), Olga Slizovskaia (deep learning technology explorer), Marius Miron, Pritish Chandna, Xavier Favory,

Sergio Oramas, Dmitry Bogdanov, Vsevolod Eremenko, Emir Demirel, Tae Hun Kim, Pablo Alonso, Frederic Font, Albin Correya, Juan José Bosch, Oscar Mayor, Ángel David Blanco, Sergio Giraldo, Fabio Jose Muneratti, Minz Won, Nadine Kroher, Giuseppe Bandiera. Many thanks to all mentioned here and others I might have missed. Special thanks to Rafael for his help in translating the abstract! Thanks to Cristina Garrido, Sonia Espí, Jana Safrankova, Lydia García, Vanessa Jimenez and Aurelio Ruiz Garcia for their reassuring support in dealing piles of paperwork.

I gratefully acknowledge Nicolas Obin, Sebastian Böck and Jan Schlüter for their inputs on this work and for providing me access to their code and data. I have interacted with and learned from Daniel Povey, Herman Kamper, Shane Settle, Richard Vogl, Chitralekha Gupta, Iris Ren, Zijin Li and Huy Phan who have been eager to give their inputs and share ideas.

A note of thanks to ERC and DTIC-UPF for funding parts of the thesis work.

I am deeply thankful to my parents in China for their selfless support. 我爱你们，愿你们身体健康! And to Charlène, tu rends ma vie plus belle!

The new journey will start again . . .

Rong Gong

17th September 2018

Abstract

Online learning has altered music education remarkable in the last decade. Large and increasing amount of music performing learners participate in online music learning courses due to the easy-accessibility and boundless of time-space constraints. However, online music learning cannot be extended to a large-scale unless there is an automatic system to provide assessment feedback for the student music performances.

Singing can be considered the most basic form of music performing. The critical role of singing played in music education cannot be overemphasized. Automatic singing voice assessment, as an important task in Music Information Retrieval (MIR), aims to extract musically meaningful information and measure the quality of learners' singing voice.

Singing correctness and quality is culture-specific and its assessment requires culture-aware methodologies. jingju (also known as Beijing opera) music is one of the representative music traditions in China and has spread to many places in the world where there are Chinese communities. The Chinese tonal languages and the strict conventions in oral transmission adopted by jingju singing training pose unique challenges that have not been addressed by the current MIR research, which motivates us to select it as the major music tradition for this dissertation. Our goal is to tackle unexplored automatic singing voice assessment problems in jingju music, to make the current eurogeneric assessment approaches more culture-aware, and in return, to develop new assessment approaches which

can be generalized to other music traditions.

This dissertation aims to develop data-driven audio signal processing and machine learning (deep learning) models for automatic singing voice assessment in audio collections of jingju music. We identify challenges and opportunities, and present several research tasks relevant to automatic singing voice assessment of jingju music. Data-driven computational approaches require well-organized data for model training and testing, and we report the process of curating the data collections (audio and editorial metadata) in detail. We then focus on the research topics of automatic syllable and phoneme segmentation, automatic mispronunciation detection and automatic pronunciation similarity measurement in jingju music.

It is extremely demanding in jingju singing training that students have to pronounce each singing syllable correctly and to reproduce the teacher's reference pronunciation quality. Automatic syllable and phoneme segmentation, as a preliminary step for the assessment, aims to divide the singing audio stream into finer granularities – syllable and phoneme. The proposed method adopts deep learning models to calculate syllable and phoneme onset probabilities, and achieves a state of the art segmentation accuracy by incorporating side information – syllable and phoneme durations estimated from musical scores, into the algorithm.

Jingju singing uses a unique pronunciation system which is a mixture of several Chinese language dialects. This pronunciation system contains various special pronounced syllables which are not included in standard Mandarin. A crucial step in jingju singing training is to pronounce these special syllables correctly. We approach the problem of automatic mispronunciation detection for special pronunciation syllables using a deep learning-based classification method by which the student's interpretation of a special pronounced syllable segment is assessed. The proposed method shows a great potential by comparing with the existing forced alignment-based approach, indicates its validity in pronunciation correctness assessment.

The strict oral transmission convention in jingju singing teaching requires that students accurately reproduce the teacher's reference pronunciation at phoneme level. Hence, the proposed assessment

method needs to be able to measure the pronunciation similarity between teacher's and student's corresponding phonemes. Acoustic phoneme embeddings learned by deep learning models can capture the pronunciation nuance and convert variable-length phoneme segment into the fixed-length vector, and consequently to facilitate the pronunciation similarity measurement.

The technologies developed from the work of this dissertation are a part of the comprehensive toolset within the CompMusic project, aimed at enriching the online learning experience for jingju music singing. The data and methodologies should also be contributed to computational musicology research and other MIR or speech tasks related to automatic voice assessment.

Resumen

El aprendizaje en línea ha cambiado notablemente la educación musical en la pasada década. Una cada vez mayor cantidad de estudiantes de interpretación musical participan en cursos de aprendizaje musical en línea por su fácil accesibilidad y no estar limitada por restricciones de tiempo y espacio. Sin embargo, el aprendizaje musical en línea no puede extenderse a gran escala a menos que haya un sistema automático que proporcione una evaluación sobre las interpretaciones musicales del estudiante.

Puede considerarse el canto como la forma más básica de interpretación. No puede dejar de recalcarse el crítico papel que desempeña el canto en la educación musical. La evaluación automática de la voz cantada, como tarea importante en la disciplina de Recuperación de Información Musical (MIR por sus siglas en inglés) tiene como objetivo la extracción de información musicalmente significativa y la medición de la calidad de la voz cantada del estudiante.

La corrección y calidad del canto son específicas a cada cultura y su evaluación requiere metodologías con especificidad cultural. La música del jingju (también conocido como ópera de Beijing) es una de las tradiciones musicales más representativas de China y se ha difundido a muchos lugares del mundo donde existen comunidades chinas. Las lenguas tonales chinas y las estrictas convenciones de transmisión oral adoptadas en la formación del canto del jingju plantean dificultades singulares que no han sido tratadas en la investigación actual de MIR, lo que nos ha motivado para elegirla

como la principal tradición musical para esta tesis. Nuestro objetivo es abordar problemas aún no explorados sobre la evaluación automática de la voz cantada en la música del jingju, hacer que las propuestas eurogenéticas actuales sobre evaluación sean más específicas culturalmente, y al mismo tiempo, desarrollar nuevas propuestas sobre evaluación que puedan ser generalizables para otras tradiciones musicales.

El objetivo de esta tesis consiste en el desarrollo de modelos basados en datos de procesamiento de señal de audio y de aprendizaje automático (aprendizaje profundo) para la evaluación automática de la voz cantada en colecciones de música del jingju. Definimos sus retos y oportunidades, y presentamos varias tareas relevantes para la evaluación automática de la voz cantada en la música del jingju. Los métodos computacionales basados en datos requieren datos bien organizados para el entrenamiento y testeo del modelo, y describimos en detalle el proceso de gestión de las colecciones de datos (audio y metadatos de edición). Después nos centramos en los temas de investigación de segmentación automática de sílaba y fonema, detección automática de pronunciación incorrecta y medición automática de similitud de pronunciación en la música del jingu.

Es de una extrema exigencia en el estudio del canto de jingju que los alumnos pronuncien cada sílaba cantada correctamente y reproducir la calidad de pronunciación que proporciona la referencia del profesor. La segmentación automática de sílaba y fonema, como un paso preliminar para la evaluación, tiene como objetivo dividir la corriente sonora del canto en niveles más específicos, a saber, la sílaba y el fonema. El método propuesto adopta modelos de aprendizaje profundo para calcular las probabilidades de inicio de sílabas y fonemas, y alcanza una precisión de segmentación similar a la más avanzada en el estado de la cuestión actual mediante la incorporación en el algoritmo de información extra, como la duración de las sílabas y los fonemas, estimada a partir de partituras musicales.

El canto del jingju utiliza un sistema de pronunciación único que combina diferentes dialectos de la lengua China. Este sistema de pronunciación contiene varias sílabas con pronunciación especial que no están incluidas en el mandarín estándar. Un paso crucial en

el estudio del canto de jingju es la correcta pronunciación de estas sílabas especiales. El problema de la detección automática de la pronunciación incorrecta de caracteres de pronunciación especial es tratado mediante un método de clasificación basado en aprendizaje profundo por el cual se evalúa la interpretación del estudiante de un segmento silábico de pronunciación especial. El método propuesto muestra un gran potencial comparado con el método actual basado en alineación forzada, indicando su validez para la evaluación de pronunciación correcta.

(*Translated from English by Rafael Caro Repetto*)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We live in a world with explodingly increasing data and information. The information and communication technologies (ICT) help us assimilate, organize, interpret, interact, consume and generate these data and information, enhancing the experience with knowledge of the world. The technologies keep developing to follow the fast-evolving sociocultural context, in which we build tools and devise new methods.

Music is a necessary part of many people's lives. Teaching and learning music is not only a hobby but a professional commitment of many people. The knowledge-learning experience has been updating rapidly in the past few years along with the fast developing ICTs. A large amount of amateur or professional learners are converted into the online education environment, thus benefited from its easy-accessibility, various course content, and most importantly, the automatic practice assessment feedback which is used to keep the learners aware their shortcomings and gets their skill improved. Music performance, as a type of knowledge and skill, although different and more abstract than other subjects. Its online learning requires automatic tools to adapt to such context and widely diverse learners.

Music Information Retrieval (MIR) is a specialized field within the music technology, in which people invent and develop methods and tools to analyze, synthesize, understand and represent the music. It aims to explain the music concepts at various levels and build computational models for the human music perception and

understanding aspects, such as melody, rhythm, timbre, harmony, structure, mood. Automatic music performance assessment in MIR aims to extract perceptual and semantic representations in music performance recordings and devise computational models for the performance assessment.

Music has many elements. The singing voice is one of those elements which plays the central role in songs, and it's the main attraction for listeners. The pronunciation aspect of the singing voice is essential in many music traditions as it conveys the semantic information of the lyrics and the aesthetic pleasure to the listeners. This work takes the MIR point of view dealing with automatic singing voice assessment and focusing on the pronunciation aspect: to segment pronunciation meaningful singing voice events, extract pronunciation meaningful representations and develop computational models to detect the mispronunciation and measure the pronunciation similarity.

The work presented in this dissertation lies in the automatic music performance assessment subject of the MIR field, aiming at domain-specific analysis and assessment approach within a culture-specific context. We now reach the context and motivation of the thesis. The scope and objectives are then defined. Finally, we describe the organization and thesis outline.

1.1 Context and relevance

In the last two decades, many new methods, models and algorithms have been developed in the MIR community, which significantly promoted the advancement of the fields of sound and music computing, and music technology. Initially, the MIR research has been restrained to eurogeneric music. Not until recently, we witness an increasing amount of researchers who devote themselves to the MIR research of non-eurogeneric music. The CompMusic project plays an essential role in boosting this trend.

CompMusic (Computational Models for the Discovery of the World's Music) is focused on the advancement in the field of MIR by approaching new challenges from a culture-specific perspective. CompMusic aims to develop computational models for several non-western music traditions and in the meantime, advance the

overall development of MIR. CompMusic studies five music traditions: Jingju (also known as Beijing opera, China), Hindustani (North India), Carnatic (South India), Turkish-makam (Turkey), and Arab-Andalusian (Maghreb). The current information technologies in MIR field are typically targeting to solve the problems emerging from the western music. However, a wide range of music traditions other than western music can bring many new challenges. The motivation behind CompMusic is to face these new challenges by studying the five non-western music traditions, and to develop MIR technologies to embrace the richness of the world's music.

CompMusic further aims to understand music both perceptually and semantically. The typically MIR methods revolve around the audio-centric idea, which parses the incoming audio into high-level music events or concepts, such as onsets, notes, beats, melodies and chords. Although all music traditions share some common concepts, each one has its unique perceptual or semantic attributes that require different interpretations. Additionally, music is encapsulated in a complex sociocultural and historical context, which affects deeply the way of how we interpret it. Many attributes of the five non-western music traditions studied in CompMusic project cannot be explained by the audio or the western music knowledge themselves. Thus, a deeper understanding can only be achieved by considering additional culture-specific information.

Delving into the problems brought by diverse music traditions will not only help develop technologies for the specific tradition, but also will extend the scope of the existing MIR technologies, making them more adaptable and robust, and eventually open a new path in the MIR research field. Delving into these problems can also break off the limitation of current MIR technologies by posing new issues.

CompMusic focuses on the extraction of the features from music audio recordings related to melody, rhythm, timbre, pronunciation and on the perceptual and semantical analysis of the contextual information of these recordings. The goal is to identify and describe the culture-specific music perspective and to develop perceptual and semantical meaningful computational models with them. The research in CompMusic is data-driven, hence it builds upon research corpora. The types of data collected for the corpora of each music tradition are mainly audio recordings, then accom-

panied with metadata, scores, lyrics, etc. To construct the research corpora is one of the main goals of CompMusic project.

The work presented in this dissertation has been conducted in the context of the CompMusic project, focusing on automatic singing voice pronunciation assessment for jingju music from a data-driven perspective using signal processing and machine learning methodologies. This dissertation assimilates the aimings and context of the CompMusic project as applied to automatic singing voice analysis and assessment. By facing the challenge and building the culture and domain-specific singing voice analysis and assessment models, we also acquire a better understanding the existing MIR tools, and would eventually improve their capabilities. The development of the newer algorithms, models and technologies allow enriching the current knowledge of world's music and provide a novel sociocultural and musical perceptual insight. Such a work in this dissertation is relevant since we push the boundaries of automatic singing voice assessment to address the new challenges of different music traditions in the world.

1.2 Motivation

Singing can be considered the most basic form of music-performing and making since it doesn't require any external musical instrument. The important role of the singing played in the music education and performing cannot be over-emphasized. Since everyone can practice singing without an instrument, all the music aspects – melody, rhythm, timbre, dynamic, pronunciation, expression and so on can be studied by singing and also internalized by singers.

Singing is an act of producing musical sounds by voice using augmented speech tonality, rhythm, pronunciation and various vocal techniques. The music of singing contains events and structures organized in time, in other words, it's an event-based occurrence. Thus segmenting the musical events is an important task in conducting singing analysis and assessment. The automatic segmentation, analysis and modeling of these singing events can help us to elaborate perceptual and semantical meaningful measures for the singing voice assessment.

Musical events are often organized in a hierarchical way which

further forming the musical structure. Estimating event onsets and boundaries related with the singing is indispensable for the further analysis and assessment from a microscopic perspective. All the melodic, rhythmic or lyrical phrases in singing voice are established upon the basic musical or articulative event, such as the musical notes, singing phonemes, syllables and words. Each of the events can be estimated in an isolated way. However, due to the hierarchical structuring of them, a hierarchical estimation approach needs to be exploited as an important MIR task.

Singing voice can be perceptually appreciated from several musical or articulative dimensions (pitch, rhythm, timbre, dynamic, expression, pronunciation). Some of them are musically well-defined and can be assessed by relatively objective measures, such as melody, rhythm, dynamic. Others are more abstract and subjective due to the natural character of these dimensions. To sing in an accurate pitch and rhythm, have a pleasurable dynamic variation and be expressive are some high-quality singing traits commonly shared with many music traditions. However, due to the specificity of the Chinese tonal language and the stringency of the mouth and heart teaching method (口传心授, oral teaching), jingju singing education is extremely demanded in reproducing accurately the teacher's singing pronunciation at syllable and even phoneme level.

Tools developed for automatic singing voice assessment can be useful in a large number of applications such as computer-aided singing teaching, enhanced navigation of music collections and content-based music retrieval. The target users of these tools extend across professional singers who pursue to convey perfect singing details, amateur singing students who seek to have a professional assessment feedback to improve their singing abilities, musicologists who can use these tools for visualizing some singing perceptual concepts and music streaming services who can use these tools to align the lyrics to the audio.

Due to the artistic nature of the music, music performance teaching should be done individually regarding the different skill level of the students, hence it's a time-consuming and resource-intensive work. A music teacher is only able to tutor a limited number of students in a class. However, with large and ever-growing students participating in online music performance courses, the

limited teacher human power cannot meet the requirement of such large amount of audience. Thus, automatic assessment and feedback need to be provided to achieve an effective learning experience in a scalable way.

The automatic assessment of singing voice can be conducted on different singing events granularities (entire song, melodic phrases, lyrical lines, onsets, syllables, phonemes) and on various dimensions (pitch, rhythm, dynamic, timbre, pronunciation, expression). Further, the assessment can be template-based, where the student's singing is compared by measuring the similarity with a reference singing; or non-template-base, where the student's singing is assessed by a predefined model. In the template-based case, there is a need to develop perceptual relevant and content-based similarity measure; and in the non-template-based case, it necessitates to define the assessment model.

As specified earlier, a meaningful singing voice assessment model can be better achieved by taking into account the context of the music tradition - incorporating high-level musical knowledge into the assessment of the singing voice on the culturally meaningful event granularities and musical dimensions. This requires identifying unique challenges for the current MIR technologies and combining information from both raw data sources and high-level musical knowledge to build computational models.

With a unique spoken language system and a strict convention of oral transmission, jingju music singing poses a big challenge to the state of the art in automatic singing voice assessment. Several automatic singing voice assessment tasks in jingju music singing have not or very few studied before. With such unique characteristics, studying the automatic singing voice assessment for jingju music can help to pinpoint the limitations of current approaches, improve their performance, and eventually open up new paths for further research on this topic. As mentioned earlier, the gap between the current state of the art capacities of MIR technologies and the need of the multicultural world is huge. This applies as well to jingju music, in which the current methods come short of using its culture-specific knowledge and restrict the assessment performance. Being well-established music tradition in China and with a large amount of audience around the world, jingju music is an ideal candidate to develop novel automatic singing voice assess-

ment methods.

1.3 Score and objectives

The work presented in this dissertation on automatic singing voice assessment comes to the crossroad of audio signal processing, machine learning, musicology and the application of online music education. Automatic singing voice assessment can be a very broad topic and may extend to many detailed sub-topics. Thus, it is necessary to define the scope of the research in this dissertation and elucidate the research questions and objectives. The objectives of this research are listed below:

- To identify challenges and opportunities in automatic singing voice assessment of jingju music and formulate pertinent automatic singing voice assessment problems. Convert musical definitions and concepts into engineering formulations compliant with computational modelling using signal processing and machine learning approaches.
- To build annotated jingju singing voice audio and symbolic collections focusing on automatic pronunciation assessment for the computational model training and testing.
- To construct culture-aware computational models for automatic jingju singing voice analysis and assessment.
- To develop novel machine learning models for the music event segmentation, pronunciation representation and assessment of jingju singing voice.
- To explore the application of the specific computational models to western music culture with the application of automatic solfège assessment.

The final goal of this dissertation is to devise culture-specific representations for jingju singing voice events, and to use these representations for the automatic assessment modelling. The focus of the research is on jingju music singing voice, while we also explore application to western solfège assessment problem.

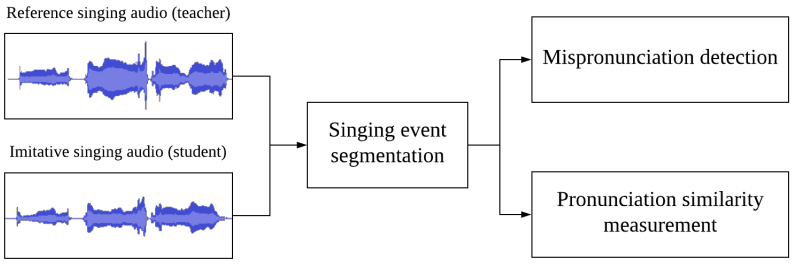


Figure 1.1: Example of automatic singing voice assessment from the reference and imitative singing audio recordings, estimating singing event segments, detecting the mispronunciation and measuring the pronunciation similarity. The methods in this dissertation follows the similar diagram, with the audio recordings as the major information source.

This dissertation investigates data-driven signal processing and machine learning approaches for automatic assessment of singing voice audio recordings. An audio recording is thus the primary source of information on which the computation models are built. Figure 1.1 shows an example of such diagram, presenting three assessment tasks - singing event segmentation, mispronunciation detection and pronunciation similarity measurement all adopt the audio recordings as the major information source. Other musical data sources such as musical scores, lyrics and editorial metadata are secondary, however, used in some tasks. The approaches adopted in this dissertation are mainly audio signal processing and machine learning (deep learning and probabilistic graphical models), investigating supervised learning methods to develop automatic singing voice assessment models.

This dissertation works toward to bring knowledge related to jingju music teaching and jingju music language to the methods. We aim to build knowledge-informed machine learning methods so that the extracted representations of the musical events and computational models are culturally relevant. High-level knowledge is taken as the determinants in the task and computational model designing.

The data-driven methods adopted in this dissertation require good quality and real-world data collections. We carefully collected, annotated and compiled the research datasets in align with

the goal of building assessment models. The algorithms are developed to perform on the real world music teaching scenarios – on the singing audio recordings of the actual classroom teaching.

The investigation in this dissertation is focused on developing novel automatic singing voice assessment methods and technologies, which are based on well-studied musicological knowledge of jingju music culture. Although the data and methods presented can eventually be adopted by musicologists for large-scale corpus analysis, the work does not aim to make any musicological contributions.

The musicological and music teaching knowledge adopted in this work is borrowed partly from the consultation with jingju performing teachers, students and jingju musicologists. The assessment models developed in this dissertation are by no means a replacement of expert music teachers, but only serve within the support of music teachers, musicologists, and as a guidance to jingju singing learners. In addition to developing novel approaches for automatic singing voice assessment, this dissertation aims to answer the following research questions:

1. How do existing automatic music performance assessment methods developed within different musical context extend to jingju music? What limitations can we pinpoint from the current state of the art?

2. We assume the high-level musicological and music teaching knowledge is useful for defining research problems, tasks and helping design computational models. What kind of high-level knowledge are insightful? How and to what extent can such knowledge be used to achieve the goal?

3. It is hypothesized that music performance assessment is conducted on various musical or articulative event granularities and different musical dimensions. Which event granularities and which musical dimensions can bring new and unique challenges to this research topic, and in return, to generalise the existing state of the art methods?

4. What are machine learning methods or deep learning architectures which are able to learn the musical dimension-relevant rep-

resentations on the variable-length musical events? What kind of side information is useful for the singing voice event segmentation and assessment? How can these side information be included in the machine learning framework?

5. It is assumed that the methods devised in this work are culture-specific. Generally, it is desirable to have a generalised method that can be transferred to other music cultures. How can the methods proposed in this work be adapted to a different music culture?

In general, this dissertation identifies the challenges and opportunities in automatic singing voice assessment of jingju music, formulates several assessment problems and tasks, tackles the issues with constructing datasets, and finally focus on the tasks of singing events segmentation, mispronunciation detection and pronunciation, overall quality similarity measurement. The scope of this dissertation within CompMusic is to support singing voice assessment methods and tools to be a part of the inclusive set of content-based analysis approaches.

One of the main advocacies of the CompMusic project is open and reproducible research – openly sharing ideas, objectives, data, code and experimental results. All the data, code and experimental results presented in this dissertation will be openly accessible via open source platform (Github, Zenodo) under open and non-commercial licenses.

1.4 Organization and thesis outline

The dissertation has eight chapters. Each chapter is written on a main topic of the thesis and is aimed to be self-contained unit with introduction, main content and summary. After an introduction of the dissertation in Chapter 1, Chapter 2 presents an overview of the jingju music background and a state of the art review of the related research topics. Chapter 3 is engaged in elucidating several new automatic singing voice assessment problems in jingju music. Chapter 4 discusses the jingju music research corpora and mainly the a cappella singing voice datasets that will be used for several singing voice assessment tasks. Chapter 5, Chapter 7 and Chapter 7 are

the major chapters of this dissertation presenting the works of automatic singing syllable and phoneme segmentation, syllable-level mispronunciation detection and phoneme-level pronunciation similarity measurement. Chapter 8 presents applications, conclusions and future works. The links of external resources such as data, code are listed in Appendix D.

Chapter 2 provides the necessary music and technical background for understanding the work presented in the thesis. We establish a consistent terminology for jingju music concepts. We describe pronunciation related concepts in jingju singing. We present an overview of the state of the art for the automatic singing voice pronunciation assessment problem tackled in the thesis. And finally we describe the technical concepts necessary to understand the algorithms and methods presented in this thesis.

Chapter 3 presents the attempts to open up the research topic of automatic singing voice assessment. We first elucidate the important role of pronunciation played in jingju singing training. Then we introduce several relevant research problems, with a review of the state of the art for jingju music or other music traditions in the context of CompMusic project. We present the background of all the relevant research problems. We formulate the thesis problems of syllable and phoneme segmentation, mispronunciation detection for special pronunciation, and pronunciation similarity measures at phoneme level.

In Chapter 4, we compile and analyse the research corpus and test datasets for the research of this dissertation. We will discuss the corpus building criteria and evaluation methodologies. We describe the corpus and the test datasets, emphasizing the research problems and tasks relevant to this thesis. We describe a set of corpus design criteria and methodologies, then use them to evaluate the jingju a cappella singing voice corpus. We present both corpus-level and test dataset-level musically meaning data analysis and visualization. We mainly emphasize on presenting a scientific approach for corpus building and the evaluation of its coverage and completeness. Apart from the corpus description, the musically meaningful data analysis and visualization is another contribution of this chapter. Finally, the research corpus and test datasets presented in this chapter will be made available for further jingju MIR research.

Chapter 5 aims to address the automatic syllable and phoneme segmentation task within the context of jingju music, presenting several methods and an evaluation of these methods. The problem is formulated in two ways – duration-informed lyrics-to-audio alignment and duration-informed syllable or phoneme onset detection. Several approaches are proposed to address the problem. We present a detailed description of hidden semi-Markov model-based (HSMM) segmentation method and the proposed onset detection-based method for syllable and phoneme segmentation. Finally, we present an evaluation of HSMM-based alignment method and the proposed onset detection-based method and explore various deep learning architectures to improve the onset detection-based method.

Chapter 6 aims to address the automatic mispronunciation detection task within the context of jingju singing, presenting several methods and an evaluation of these methods. The problem is formulated as building discriminative machine learning models to classify binarily the singing syllables into mispronounced or correctly pronounced class. Several neural network architectures are experimented to address this problem. We present a description of the forced alignment-based baseline method and the discriminative model-based method for mispronunciation detection. We present an evaluation of the forced alignment-based method and the discriminative model-based method, and explore two deep learning architectures intending to improve the discriminative detection model.

Chapter 7 aims to address the pronunciation and overall quality similarities measurement task in the context of jingju singing training, presenting several methods and an evaluation of these methods. The problem is formulated as building machine learning models to perform phoneme embedding regarding pronunciation and overall quality aspects. Several neural network architectures are experimented to address this problem. We present a description of the classification model for phoneme embedding, and to explore the siamese network model for the same purpose. Finally, we present an evaluation of the classification model and the siamese model.

Chapter 8 presents some of the applications, conclusions, and the pointers for future work.

To our knowledge, this thesis is the first attempt at singing voice

pronunciation assessment of jingju music. By tacking the problems presented in this thesis, we aim to develop useful tools and algorithms for automatic singing voice assessment of jingju music. In this process, we also hope to obtain a better understanding into the nature of pronunciation in jingju singing, and contribute to improving the state of the art.

Chapter 2

Background

This chapter provides the necessary music and technical background for understanding the work presented in the thesis. The main aims of this chapter are:

1. To establish a consistent terminology for jingju music concepts.
2. To describe pronunciation related concepts in jingju singing.
3. To present an overview of the state of the art for the automatic singing voice pronunciation assessment problem tackled in the thesis.
4. To describe the technical concepts necessary to understand the algorithms and methods presented in this thesis.

2.1 Jingju music

Jingju (also known as Beijing or Peking opera) is the most representative form of Chinese opera which assimilates the essence of various Chinese opera forms such as 徽剧 (Anhui opera), 昆曲 (Kun opera), 秦腔 (Qin qiang) and 高腔 (Gao qiang). It arose in the late 18th century and became fully developed in the mid-19th century. Now it is regarded one of the cultural treasures in China and inscribed in the UNESCO representative list of the intangible cultural heritage of humanity. Jingju is widely practised

over mainland China, Hong Kong, Taiwan, and overseas countries where there is Chinese communities presence. Major jingju performing troupes are located in big mainland China cities such as Beijing, Tianjin and Shanghai. A significant amount of jingju musicological literature can be used to formulate MIR problems. The presence of a large audience and musicological literature are an important motivation to carry a computational research for this music culture.

This section describes the focus of this dissertation – jingju music culture. The emphasis is on singing concepts in this music culture. This section is not a comprehensive introduction to this culture but is aimed to be sufficient to support the following chapters of the dissertation.

We use simplified Chinese characters (computer encoding: GB2312) to introduce jingju terminologies for the first time in this dissertation. We also introduce the pinyin, the romanization system of Mandarin Chinese, for each terminology. Only the pinyin form of the terminology will be used throughout the dissertation.

2.1.1 A synthetic art form

Professor Li in National Academy of Chinese Theatre Arts (NACTA) said "The three basic elements of Chinese opera are 曲 (pinyin: qu, tune), 程式 (pinyin: chengshi, conventions – a strict set of rules) and 虚拟表演 (pinyin: xu ni biao yan, virtual acting). These three elements are ultimately aiming to support 戏 (pinyin: xi), which can be approximately understood as 'entertainment'." Of the three elements, "tune" is the most important one, which represents all the musical dimensions of the jingju music. However, this representation is not only limited to music but constructs the whole skeleton of the jingju performing.

Jingju is a synthetic art form which includes four disciplines – 唱 (pinyin: chang, singing)，念 (pinyin: nian, declamation)，做 (pinyin: zuo, physical acting) and 打 (pinyin: da, acrobatics). Singing is directly related to tune, and the other three disciplines are integrated together by the music and rhythm of jingju performing.

The jingju technical training for performers consists in becoming proficient of the conventions of the four disciplines as mentioned earlier which are established by tradition. The jingju per-

formers use these conventions to construct characters and convey stories. For example, they use singing conventions to express the character's emotional state. The jingju performance is codified through the conventions which are not aimed at hinder the creativity and artistry. The appreciation of the beauty of jingju is to see how the performers are conveying the conventions. In jingju training, a performer will have more creativity if she/he can master more conventions.

2.1.2 Singing and instrumental accompaniment

"In the aural performance of Beijing opera, two types of sounds are actually heard: song and speech vocalized by the stage performers, and instrumental music played by the musicians of the orchestra. The voice of the Beijing opera performer, is the featured component of aural performance." – Elizabeth Wichmann (Wichmann, 1991).

In a jingju play, the sections where singing occurs are 唱段 (pinyin: chang duan, literally translated as singing section). The closest form to chang duan in Western opera is "aria", which signifies "any closed lyrical piece for solo voice (exceptionally for more than one voice) with or without instrumental accompaniment, either independent or forming part of an opera, oratorio, cantata or other large work." The difference between chang duan and aria is that latter is a self-sufficient piece conceptually, whereas chang duan is formulated in a dramatic continuum, although it is usually performed and recorded individually (Repetto, 2018).

Jingju chang duan is started actually before the performer starts to sing. The declaration of the starting point of a chang duan is 叫板 (pinyin: jiaoban, literally translated as "calling the banshi"). Banshi is the rhythmic framework concept that we will introduce it in Section 2.1.7. Jiaoban is included in every chang duan of the commercial recordings and teached in conservatory jingju performing classes. The percussion pattern 住头 (pinyin: zhu tou) is to signal the end of a chang duan (Mu, 2007).

Jingju instrumental ensemble is divided into two sections – 文场 (pinyin: wenchang, literally translated as "civil scene") and 武场 (pinyin: wuchang, literally translated as "martial scene"). Wenchang is the orchestral accompaniment, and wuchang is formed by percussion instruments. There are five basic percussion instru-

ments – 单皮鼓 (pinyin: danpigu, drum), 板 (pinyin: ban, clappers), 铙钹 (pinyin: naobo, cymbals), 大锣 (pinyin: daluo, big gong), 小锣 (pinyin: xiaoluo, small gong). The first two instruments are played normally by the same person, so they have a combined term – bangu. The musician who plays the bangu is called 司鼓 (pinyin: si gu), literally translated as "the man who is in charge of the bangu". The primary functions of wuchang are playing 锣鼓经 (pinyin: luo gu jing, rhythm patterns) and supporting the rhythmic aspect of the actor/actress' performing.

The main instrument of wenchang is 京胡 (pinyin: jinghu). Having loud volume, and very bright and penetrating sound, jinghu is the aural representative of jingju sound. The musician who plays jinghu is called 琴师 (pinyin: qin shi, master instrumentalist). The major role played by qinshi is supporting the jingju melody. Traditionally, qinshi is the musician who has the closest collaboration with the performer. Jingju line sustains the singing line to form an uninterrupted melody stream, which impels the singing (Repetto, 2018).

The other instruments in wenchang are 月琴 (pinyin: yueqin), 三弦 (pinyin: sanxian), 京二胡 (pinyin: jingerhu), 阮 (pinyin: ruan), 中阮 (pinyin: zhongruan) and 大阮 (pinyin: daruan). They all play the same melody as the jinghu line in the same or different octave, and in a heterophonic structure. The performer, sigu and qinshi take turns in coordinating the jingju performing tempo.

2.1.3 Lyrics structure

The primary function of the lyrics in jingju is telling stories. Music structure in jingju is closely related to lyrics structure. The tune sequences in jingju are inherited from the creation principle in poetry of Tang dynasty (618 - 907 AC), that the melody and poetic structure are taken from the preexisting poems or songs, and new lyrics are arranged to fit in that schema (Repetto, 2018). The new lyrics are labelled with the name of the original poem or song, so that the performer knows how to sing the tune. The label of the original poem or song is called 曲牌 (pinyin: qu pai, literally translated as tune label). Different qupai have different forms which represent not only different melodies but also a different number of melodic lines and a different number of characters in each line. This kind

of lyrics structure is called 长短句 (pinyin: chang duan ju, literally translated as long and short lines). A jingju chang duan consists of a sequence of these chang duan ju.

The basic structure of lyrics stanza consists of two symmetrical lines which have the same number of characters. The most common term in jingju circles for describing this two symmetrical lines is 上下句 (pinyin: shang xia ju, literally translated as upper and lower lines). The most common English terminology for shang xia ju is couplet for the stanza, opening line for upper line and closing line for lower line (Repetto, 2018). A standard line has either 7 or 10 characters, grouped in three sections – 2 + 2 + 3 or 3 + 3 + 4. These sections, namely 逗 (pinyin: dou), are the basic semantic and rhythmic units (Wichmann, 1991).

The lyrics structure mentioned above can be modified in actual singing. A typical case is the variation of the number of characters in each line, for example, 衬字 (pinyin: chenzi), the characters do not have semantic meaning, but serve to help the performer prolong the singing of certain nasal or vowel sounds. Another form of increasing the number of characters is 垛字 (pinyin: duozi), which inserts semantic units containing 3 or 4 characters into the line.

2.1.4 Linguistic tones and pronunciation

It is commonly assumed that the linguistic intonation of a tonal language singing needs to agree with its melody to a certain extent to make sure the intelligibility. For jingju which is sung by using mainly Chinese Mandarin language and various dialects, its music features are related to the dialects. In other words, The Chinese dialects used in jingju singing determines its melody characteristics to a certain extent.

In jingju circles, it exists an expression to describe the relation between linguistic tones and melody – 字正腔圆 (pinyin: zi zheng qiang yuan, literally translated as "characters should be straight, tune should be round."). This expression can be understood as that the performer needs to attain both the intelligibility of the lyrics and the smooth sounding of the melody. The most critical problem to be avoided in jingju singing is 倒字 (pinyin: dao zi, literally translated as upside-down character), which means that the lyrics is

misunderstood because the performer mispronounces certain characters.

Most of the jingju scholars agree in that jingju singing uses mainly two Chinese dialects – 北京音 (pinyin: Beijing yin, the dialect of Beijing) and 湖广音 (pinyin: Huguang yin, the dialect of Huguang). Some scholars consider that jingju singing also uses a third Chinese dialect – 中州韵 (pinyin: Zhongzhou yun, literally translated as rhymes from Zhongzhou). All three dialects share the same tone categories 阴平 (yin ping), 阳平 (yang ping), 上 (shang) and 去 (qu), although the pitch contours of the same characters realized in the same categories are different for the three dialects.

The three dialects result in the complexity of the pronunciation in jingju singing. Such complexity influences the linguistic tones as well as the pronunciation of the syllabic initials and finals. The standard Chinese used in Mainland China – 普通话 (pinyin: pu tong hua), very close to Beijing yin, is taken as the reference for jingju pronunciation. All the special pronunciations different from the reference putonghua can be divided into two categories – 尖团字 (pinyin: jian tuan zi, literally translated as pointed and rounded characters) and 上口字 (pinyin: shang kou zi, literally "up to the mouth" characters). Jian tuan zi has two sub-categories of characters – 尖字 (pinyin: jianzi) and 团字 (pinyin: tuanzi), which are separated by the fricative and affricative consonants of a syllable. When studying a new play, jingju performer should learn which characters belonging to tuanzi in putonghua should be pronounced as jianzi. The jian tuan zi qualities are considered extremely important for both listening comprehension and aesthetic effect (Repetto, 2018). Shang kou zi are generally a set of characters of which the pronunciation is different from the standard Mandarin, adopting from southern Chinese dialects – Huguang yin and Zhongzhou yun. By shang kou zi and converting certain tuanzi to jianzi, the language of jingju is made more appealing to speakers of the diverse range of dialects throughout China than is Mandarin alone (Wichmann, 1991). Jian tuan zi and shang kou zi are one of the main study focuses of this dissertation. Thus a more specific extended description will be presented in Section 2.2.3 and Section 2.2.4.

2.1.5 Role-types

行当 (pinyin: hang dang), commonly translated as role-type, is a colloquial term for the "acting profiles" of a jingju performing character. There are four role-types in jingju – 生 (sheng), 旦 (dan), 净 (jing) and 丑 (chou), which respectively have their specific styles of performing, speaking, singing, costume, and make-up. These oral and visual means of expression define the gender, approximate age, social status, profession, personality and singing style (Yung, 1989). Due to the various and complicate conventions that each role-type possesses, every performer has to specialize one role-type and practice these conventions along the performing career.

Table 2.1: Jingju four role-types and their sub role-types. The role-types with * superscript are the main research objects of this dissertation because singing is their major discipline.

Main role-types	sheng	dan	jing	chou
Sub role-types	laosheng*	qingyi*	tongchui	wenchou
	xiaosheng	huadan	jiazi	wuchou
		laodan		
		wudan		

Sheng role-type is specialized in the performance of male characters, whereas dan role-type is specialized in that of female characters. Jing role-type depicts the male characters with an exaggerated temperament. Chou role-type is used for male or female comic characters (Repetto, 2018). The most obvious difference between the male's voice and the female's voice is the timbre. Male role-types sing with chest voice, while female role-types use falsetto. Regarding the singing pitch register, there is a displacement of the pitch range in the female singing to a higher region, where female sings a fourth to an octave higher than male singing. Regarding melodic contours, female singing is usually more melismatic than male singing (Wichmann, 1991).

老生 (pinyin: lao sheng) role-type portrays adult or old male characters, which is also the representative of male singing. All textbooks use the examples of laosheng role-type to explain elements of jingju music system. Two representative sub role-types

in dan are 青衣 (pinyin: qing yi) and 花旦 (pinyin: hua dan). The former is most representative role-type of female singing, and generally used for building female characters from a higher social classes. The latter is used for building female characters with a playful personality.

We list the major jingju role-types in Table 2.1, where laosheng and qingyi are the main research objects of this dissertation since singing is their major discipline.

2.1.6 Shengqiang

There is no agreed definition of jingju 声腔 (pinyin: shengqiang) between scholars. Some of them define shengqiang as tune families of jingju music, meaning a tune which has been evolved into different versions in the performing and transmission process throughout the history. Although these tunes share certain tonal, modal, and dramatic function, they tend to differ from each other in metrical, rhythmic, and melodic details (Yung, 1989). The shengqiang definition of Elizabeth Wichmann deviated from tune family, and characterize a group of related shengqiang as system. Each shengqiang system is identified by its unique patterns of modal rhythm, song structure, melodic contour and construction, and keys and cadences (Wichmann, 1991).

Jingju contains mainly eight shengqiang – 西皮 (xipi), 二黄 (erhuang), 反西皮 (fanxipi), 反二黄 (fanerhuang), 四平调 (sipingdiao), 南梆子 (nanbangzi), 高拨子 (gaobozi) and 吹腔 (chuiqiang). Two shengqiang with the most significant presence in jingju arias are xipi and erhuang. Fanxipi and fanerhuang are respectively first degree shifted versions of xipi and erhuang. Additionally, shengqiang is related to the emotional content of the aria. For example, Wichmann describes the emotional content of erhuang as dark, deep, profound, heavy and meticulous, and xipi as sprightly, bright, clear, energetic, forceful, purposeful (Wichmann, 1991).

2.1.7 Banshi

Ban means the percussion instrument clappers used in jingju wuchang. Banshi can be understood as the jingju rhythmic patterns.

There are four types of banshi in jingju – 一板一眼 (one ban and one eye), 一板三眼 (one ban and three eyes), 有板无眼 (ban but no eye) and 无板无眼 (no ban and no yan), where ban and eye indicate respectively accented and unaccented beats. The first three banshi are metred types, and usually notated in jianpu scores correspondingly with the time signatures 2/4, 4/4 and 1/4. The last is assigned to 散板 (pinyin: sanban), meaning unmetred type. Wichmann describes that each banshi has a characteristic tempo, is associated with certain characteristic melodic tendencies, and is perceived as appropriate for certain dramatic situations (Wichmann, 1991).

Table 2.2: Jingju metred banshi.

Tempo	Banshi	Time signature	Melodic tendencies
slow	manban	4/4	melismatic
↑	zhongsanyan	4/4	↑
	kuaisanyan	4/4	
	yuanban	2/4	
	erliu	2/4	
↓	liushui	1/4	↓
fast	kuaiban	1/4	syllabic

The primary or original banshi is called 原板 (pinyin: yuanban), with time signature 4/4. When it is transformed into 1/4, the corresponding banshi is 快板 (pinyin: kuaiban), and the tempo is also accelerated. When yuanban is transformed into 4/4, the resulting banshi is 慢板 (pinyin: manban), meaning slow banshi. When yuanban is transformed into manban, not only the time duration of syllables are extended, but also the number of ornamentations within each syllable are increased. However, when yuanban is converted to kuaiban, the singing style becomes almost syllabic – one beat for one syllable.

三眼 (pinyin: sanyan, literally translated as three eyes) is another name for manban. Sanyan banshi can be divided into three sub-banshi – 慢三眼 (pinyin: mansanyan, equal to manban), 中三眼 (pinyin: zhongsanyan), 快三眼 (pinyin: kuaisanyan). The tempo of kuaisanyan is faster than zhongsanyan, and they both faster than mansanyan but slower than yuanban. Except for yuanban, the banshi of time signature 2/4 category also contain 二六

(pinyin: erliu, literally translated as two six) because their couplet has six accented beats. In terms of shengqiang xipi, there is another metred banshi – 流水 (pinyin: liushui, literally running water) which uses 1/4 time signature, is faster than yuanban but slower than kuaiban. We list major jingju metred banshi in Table 2.2.

Three main unmetred banshi are 导板 (daoban), 回龙 (huilong) and 哭头 (kutou). Daoban is the first melodic line of a chang duan. Huilong follows daoban, and is used to draw out the metred banshi. Kutou, literally crying head, is used for a grievous outburst and can occur after any section of the couplet (Repetto, 2018). The variety of banshi is needed to convey the emotional content of the lyrics. In general, yuanban is related to neutral and narrative lyrics content; manban reflects introspective, and deep emotions; while kuaiban expresses agitation, nervousness.

The precise, clear pronunciation is critical to jingju listening comprehension, and also form an important aural aesthetic value of jingju. The primary focus of this dissertation is the pronunciation aspect of jingju singing. An in-depth description of jingju pronunciation concepts is given in the following section.

2.2 Pronunciation in jingju singing

Mandarin is a tonal language and there are in general 4 lexical tones and 1 neutral tone in it. Every character of spoken Mandarin language is pronounced as mono-syllable (C.-H. Lin, Lee, & Ting, 1993). When the differences in tones are disregarded, the total number of different pronounced syllables is 408. The jingju singing is the most precisely articulated rendition of the spoken Mandarin language. Basic pronunciation of a jingju syllable is categorized as precisely shaping the throat and mouth to articulate (1) four vowel types – 四呼 (pinyin: sihu) and (2) five consonants types 五音 (pinyin: wuyin). As been briefly discussed in Section 2.1.4, in jingju singing pronunciation, certain sounds are spotted as either jianzi (pointed) or tuanzi (rounded). The jiantuanzi (pointed and rounded sounds) is an extremely important jingju pronunciation aspect, such that the precision and exaggeration of its sound qualities is one of the remarkable attribute of all jingju vocalization. Due to the adoption of certain regional dialects, and the ease or vari-

ety in pronunciation and projection of sound, certain special pronunciations in jingju theatrical language differ from their normal Mandarin pronunciations. However, the mono-syllabic pronouncing structure of the standard Mandarin doesn't change (Wichmann, 1991).

2.2.1 Jingju singing syllable

Definitions for the syllable in speech have been provided from a variety of perspectives; phonetically, Roach (Roach, 2000) describes a syllable as "consisting of a center which has little or no obstruction to airflow and which sounds comparatively loud; before and after that center (...) there will be greater obstruction to airflow and/or less loud sound." This definition allows for a conceivable way for detecting syllables in speech.

A syllable of jingju singing is composed of three distinct parts in most of the cases: 头 (pinyin: tou, head), 腹 (pinyin: fu, belly) and 尾 (pinyin: wei, tail). Some syllables are only composed of an head and a belly or a belly alone. The head is normally not prolonged and consists of the initial consonant or semi-vowel, and the medial vowel if the syllable includes one, which itself is normally not prolonged in its pronunciation except for the one with a medial vowel. The belly follows the head and consists of the central vowel. It is prolonged throughout the major portion of the melodic-phrase for a syllable. The belly is the most sonorous part of a jingju singing syllable and can be analogous to the nuclei of a speech syllable. The tail is composed of the terminal vowel or consonant (Wichmann, 1991). The head, belly, tail structure of a syllable is illustrated in the upper part of Figure 2.1.

Another Mandarin syllable structure describes a syllable consisting of two components – 声母 (initial) and 韵母 (final). An initial is the consonant or semi-vowel, and a final is the combination of optional medial vowel, central vowel and terminal vowel or consonant. The initial and final structure grouping is used widely in Mandarin language teaching textbooks. The initial and final structure of a syllable is illustrated in the bottom part of Figure 2.1.

For example, Mandarin syllable 现 (pinyin: xian) is composed of the initial 'x' and the final 'ian', or the head 'xi', the belly 'a' and the tail 'n'; syllable 坡 (pinyin: po) is composed of the initial

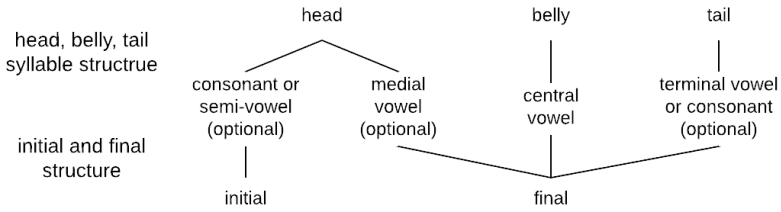


Figure 2.1: An illustration of two Mandarin Chinese syllable structures.

'p' and the final 'o', or the head 'p' and the belly 'o', without a tail. A complete table of Mandarin sounds noted in pinyin, I.P.A and X-SAMPA formats can be consulted in Appendix A.

The speech syllable only contains one prominent sonority maximum due to its short duration (average $<$ 250 ms and standard deviation $<$ 50 ms for Mandarin (J. Wang, 1994)). In contrast, a singing voice syllable may consists of numerous local sonority maxima, of which the reason is either intentional vocal dynamic control for the needs of conveying a better musical expression or unintentional vocal intensity variation as a by-product of the F0 change (Titze & Sundberg, 1992) or vocal ornaments such as vibrato (Horii & Hata, 1988).

2.2.2 Sihu and wuyin – basic jingju pronunciation units

Sihu means four basic shapes for the throat and mouth when pronouncing the semi-vowel and central vowel parts of a syllable, whereas wuyin indicates five portions of the mouth deemed to the articulation of initial consonants. In jingju circles, it is believed that the throat and mouth must be shaped and the correct portion of the mouth needs to be used for producing the desired vowel and consonant sound. The detailed throat and mouth shapes for producing sihu and portions of mouths for articulating wuyin are described in Wichmann's book (Wichmann, 1991). In this section, we list the vowels and consonants which are related to each category of sihu and wuyin.

The four shapes for the throat and mouth are 齐齿呼 (pinyin:

qichi, level-teeth), 合口呼 (pinyin: hekou, closed-mouth), 撮口呼 (pinyin: cuokou, scooped-lips) and 开口呼 (pinyin: kaikou, open-mouth).

- qichi: syllables using the pinyin /i/ as its medial vowel or central vowel, e.g. qi, ji.
- hekou: syllables using the pinyin /u/ as its medial vowel or central vowel, e.g. lu, kuan.
- cuokou: syllables using the pinyin /ü/ (also written 'v') as its medial vowel or central vowel, e.g. yun, jun.
- kaikou: syllables which central vowel is not /i, u, ü/ and without a medial vowel, e.g. da, cheng.

The five portions for the mouth are 喉 (pinyin: hou, larynx), 舌 (pinyin: she, tongue), 齿 (pinyin: chi, jaw and palate), 牙 (pinyin: ya, teeth) and 唇 (pinyin: chun, lips).

- hou: syllables starting with the semi-vowels /i, u/ (written 'y', 'w') and consonants /g, k, h/, e.g. guo, ke, he.
- she: syllables starting with the consonants /d, t, n, l/, e.g. da, ta, ni, liao.
- chi: syllables starting with the consonants /zh, ch, sh, r/, e.g. zhi, chi, shi, ri.
- ya: syllables starting with the consonants /j, q, x, z, c, s/, e.g. zuo, cong, cai, jia, que, xiao.
- chun: syllables starting with the consonants /b, p, m, f/, e.g. bang, fang, miao, fa.

2.2.3 Jiantuanzi – pointed and rounded syllables

The concept of jiantuanzi is associated with certain consonant and vowel combinations, and there is no unified definition for jiantuanzi. Wichmann defineds jiantuanzi in three senses (Wichmann, 1991):

- Strict sense: pointed syllables are those starting with the consonants /z, c, s/ and followed by the vowel /i/ or /u/, e.g. zi,

zu, ci, cu, si, su; rounded syllables are those starting with the consonants /j, q, x, zh, ch, sh, r/ and followed by the vowel /i/ or /ü/, e.g. zhi, chi, shi, ri, ji, jia, qu, quan, xi, xia, xian.

- Broader sense: all syllables that start with /z, c, s/ are pointed, and all syllables that start with /j, q, x, zh, ch, sh, r/ are rounded.
- Broadest sense: all syllables that do not start with /z, c, s/ are rounded.

According to the strict sense of this definition, the rounded syllables are those begin with wuyin types chi and /j, q, x/ of ya, and use two sihu types – qikou and cuokou. Whereas, the pointed syllables are those begin with wuyin type /z, c, s/ of ya, and use sihu types qikou and hekou. Wichmann only uses the aural perspective of Mandarin syllables to define jiantuanzi. However, the polyphonic case of a written-character – whether it should be pronounced as jianzi or tuanzi, is not discussed. As a supplementary to Wichmann's definition, Tonglin Shu (Tonglin Shu, 2011) discussed that certain rounded syllables in Mandarin can be pronounced as pointed syllables in jingju singing or speech due to the influence of several regional Chinese dialects. The rule of this pronunciation alteration is /j, zh/ → /z/, /q, ch/ → /c/ and /x, sh/ → /s/. For example, 晓 (pinyin in Mandarin: xiao) can be pronounced as the pointed sound siao, 期 (qi) can be pronounced as ci, 出 (chu) can be pronounced as cu in certain jingju scenarios. Tonglin shu also pointed that this alteration of pronunciation can be seen as one type of special pronunciation which will be introduced in the next section. To maintain the traditional flavor of jingju singing, it is important for a performer to articulate precisely certain syllables of which their sounds are altered from rounded to pointed.

2.2.4 Special jingju singing pronunciation

The definition of special pronunciation in jingju music is quite simple – All pronunciations of written-character which are different from those in standard Mandarin Chinese. These special pronunciations come from two sources – traditional Chinese sounds and sounds from various regional Chinese dialects. As jingju is evolved

from several regional Chinese opera, such as Anhui opera, Kun opera, Qinqiang and Gaoqiang, certain regional pronunciations of written characters were adopted (Wichmann, 1991).

No overall set of rules can be found by which special pronunciations can be analytically established. In other words, all special pronunciations are set up by tradition. "The performer must simply memorize the sounds and specific written-characters whose pronunciation may be given special pronunciations ... This process of memorization is an ongoing one; it occurs each time a student of professional performer learns an established play from a particular school (流派), in which the words that have special pronunciations and their specific altered pronunciations have been set by tradition." – Elizabeth Wichmann. A non-complete list of special pronunciations is given in Wichmann's book (Wichmann, 1991). However, we will not copy directly this list in the dissertation, but will organize another one regarding our jingju singing data collections that will be presented in Section 4.2.2.

2.3 The pronunciation of jingju singing and Western opera singing: a comparison

We compare some of the pronunciation concepts in jingju and Western opera singing, so that it can be used for better clarification of the unique assessment approaches for jingju singing.

Western opera is a vowel-centric singing genre since vowels are easy to be prolonged and can carry rich resonance. Consonants, especially non-pitch ones which interfere with that goal, are regarded as "unfriendly", hard to manipulate. Thus non-pitched consonants are often sung in a low volume, and pitch consonants are employed far more resonance in the formation than is common in speech (Nair, 1999). Whereas, it is attached much importance to fully pronounce syllable initials (non-pitched consonants) in jingju singing. In fact, in certain cases, to show the physical strength of the mouth, the initials should be pronounced in an "overstressed" way than is in the normal speech. This "overstressed" pronunciation technique of jingju consonants is called 喷口 (pinyin: penkou).

Italian is the most important language for the Western opera singing not only because many long-established opera classics are written in Italian, but also because some characteristics of this language such as the openness of the vowels, richly resonant phonemes and free from any local dialectal influence, makes it very well suitable for singing (Nair, 1999). However, other European languages, especially English, do not possess those innate advantages as Italian do. To facilitate the formation of a rich resonance, opera singers often adopt Italian vowels or make subtle vowel shift (Nair, 1999) when they sing in these languages. Jingju singing is free from the influence of other languages than Chinese since it is only sung in Chinese dialects.

In both Western opera and jingju, to maintain the purity of the phonemes, it is required to sing with a great precision for the reduction of certain coarticulation. For example, the diphthong /ɑi/ in both genres are sung separately as two vowels connected by a fast transition (Nair, 1999). However, for certain coarticulation such as transiting from the semi-vowel /j/ to a central vowel, it is adopted a slow fashion in jingju singing while is maintained fast in Western opera singing.

Finally, it is found that in both genres, singers do formant tuning to adapt the first resonance frequency to the note frequency (Sundberg, Lã, & Gill, 2013).

Weakening the volume of the non-pitched consonants, vowel shift, fast transition between two phonemes in coarticulation and formant tuning are all deemed as the causes to decrease the intelligibility of singing voice. In general, jingju singing is more demanding than Western opera singing regarding a clear pronunciation of both consonant and vowel.

2.4 A review of automatic assessment of musical performance

A review of the state of the art of the automatic assessment of musical performance is presented to provide a starting point for the main research works in this dissertation. The review in this section is generic, and not specific to jingju music.

In most of the previous automatic assessment of music performance studies, the assessment are conducted for the entire music piece (Nakano et al., 2006; Cao et al., 2008; Liu et al., 2011; Tsai & Lee, 2012; Molina et al., 2013; C. Gupta, Li, & Wang, 2017). In the other studies, the assessment are performed at musical event-level (Schramm et al., 2015; Robine & Lagrange, 2006; C. Gupta, Grunberg, et al., 2017). Regarding the latter case, Pre-processing methods should be performed on the entire musical piece to segment it into musical event units. The relevant pre-processing topics that will be reviewed in this section are musical onset detection, text-to-speech alignment and lyrics-to-audio alignment.

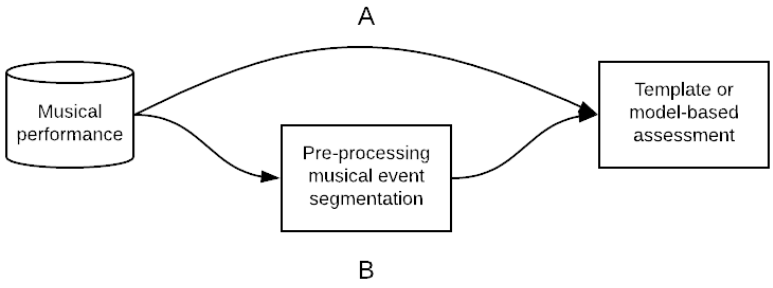


Figure 2.2: The general flowchart of automatic assessment of musical performance. A: assessment for the entire music piece. B: assessment for musical event units.

The automatic assessment methods can be either template-based (Cao et al., 2008; Tsai & Lee, 2012; Molina et al., 2013; Bozkurt et al., 2017; C. Gupta, Li, & Wang, 2017) or model-based (Nakano et al., 2006; Schramm et al., 2015; Robine & Lagrange, 2006; Han & Lee, 2014; Luo et al., 2015; C. Gupta, Grunberg, et al., 2017). The former case means that the reference performance are provided for a comparision with the target assessing performance. While the latter case indicates that the reference performance are not given, and the target performance are assessment using a pre-trained model. Regarding the templated-based assessment, the similarity calculation between the reference and target musical performance segments is usually involved. Thus we review the neural acoustic embedding technique which can faciliate

the similarity calculation. A general flowchart of automatic assessment of musical performance is illustrated in Figure 2.2.

2.4.1 Automatic assessment of musical performance

We introduce the overview of the studies on general automatic assessment of musical performance in this seciton. A significant number of authors adopt the regression or classification model to predict the human rating of the musical performance with acoustic features. In this following part of this section, we firstly present the studies of automatic singing voice assessment. Then we present those of instrumental performance assessment.

Automatic assessment of singing voice

In Table 2.3, we list the goal and methods of each work. A model-based method is presented by (Nakano et al., 2006) for evaluating unknown melodies. The SVM is trained with semitone stability and vibrato features to classify the good and poor singers. The evaluation dataset contains 600 songs sung by 12 singers – 6 good and 6 poor. (Daido et al., 2014) identifies that three features – A-weighted power, F0 fall-down and vibrato extent, are relevant to singing enthusiasm. Then they build a regression model to predict the human rating scores of singing enthusiam by combing these three features.

As we have mentioned above, most of the works are template-based and build regression or classification model for the prediction of human rating. (Cao et al., 2008) calculate features on four categories – intonation, rhythm, timbre brightness and vocal clarity, then adopt SVR regression model to predict the expert rating scores of the singing quality. (Liu et al., 2011) propose a two-step method for solfège assessment. In the first step, they use DTW to align the reference and target performance pitch tracks. In the second step, they calculate intonation and rhythm features using the aligned musical notes. Relative pitch interval and lagged tempo reference are identified respectively as the most correlated features for intonation and rhythm rating. (Tsai & Lee, 2012) model intonation rating using DTW cost between reference and target pitch

tracks, model volume rating using DTW cost between reference and target logarithmic energy curves, model rhythm rating using HMM classification score between pitch strength time sequences, and predict the overall score with linear regression model by combining these three dimension ratings.

(Molina et al., 2013) explore both low-level and high-level features for singing voice intonation and rhythm assessment. The reference used in their work is not symbolic MIDI but singing audio. Low-level features are calculated based on the DTW alignment path, and high-level features are calculated on the transcribed musical notes. They calculate the correlation coefficients between each individual feature and expert rating score, and find that low-level total intonation error and high-level pitch difference are correlated with the intonation rating and low-level RMS of the alignment path is correlated with the rhythm rating. Finally, they use quadratic polynominal regression model with all the features to predict the overall singing quality. (Bozkurt et al., 2017) develop a dataset and a baseline model for the singing assessment of the conservatory entrance exam. Their dataset contains 2599 piano references and 1018 singing performance of 40 different melodies. These singing performances are labeled as pass or fail categories by 3 experts. They use DTW to align the pitch tracks between the singing performance and the piano reference. The baseline is built by using a multilayer perceptron model with 3 featuers – pitch difference histogram, DTW alignment cost and the amount of the length change of the DTW alignment. (C. Gupta, Li, & Wang, 2017) construct the singing assessment model on 6 aspects – intonation accuracy, rhythmic consistency, pronunciaiton, vibrato, volume and pitch range. To avoid the alignment error caused by the intonation mistake, they use MFCC as the representation for the DTW alignment between reference and target. They also experiment cognition modeling for obtaining the perceptual relevant features. Finally, both linear regression and multilayer perceptron models are explored to predict human ratings with various feature combinations.

In the works mentioned above, the model are built to assess the singing performance in the entire musical excerpt granularity. However, several works explore the assessment of detailed musical events, such as note, expression segment and syllable. (Mayor et

al., 2006) proposed a probablistic and rule-based method for note alignment and expression segmentation. They use HMM framework with Viterbi algorithm for the note alignment. The cost probablity is calculated by a set of heurstic rules which are defined on timing, pitch, energy, vibrato and timbre. The similar idea is adopted for expression segmentation. They define several rules for the expressions such as attack, sustain, vibrato, release and transition. The HMM topology of the expression is constrained by the note segmentation. (Schramm et al., 2015) construct a Bayesian classifier to assess the performing correctness of solfège note. They first transcribe the pitch track and notes for the target singing performance, and devise a special DTW algorithm to align the reference score the transcribed notes. For each assessment dimension – note-level pitch, onset and offset, they construct Gamma probability density functions for both correct and incorrect classes. Finally, they identify the fuzzy boundary between two classes using a Bayesian classifier. In (C. Gupta, Grunberg, et al., 2017)'s work, they first generalize the mispronounciation rules for the singing voice of Southeast Asian English dialects. Then they use Automatic Speech Recognition system with an adapted dictionary to detect the mispronunciation.

Table 2.3: Summary table of the previous studies on automatic assessment of singing voice.

Authors	Goal	Methods
(Mayor et al., 2006)	Expression categorization and alignment	Rule-based note and expression alignment using Viterbi algorithm
(Nakano et al., 2006)	Singing skill evaluation for unknown melodies	Building SVM model to classify good and bad performance.
(Cao et al., 2008)	Singing quality evaluation	Building SVM regression model between features and human rating scores.
(Liu et al., 2011)	Singing evaluation	Using correlation coefficient to select the best features.
(Tsai & Lee, 2012)	Karaoke singing evaluation	Building linear regression model for predicting the overall score.
(Molina et al., 2013)	Singing voice assessment	Building nonlinear regression model for predicting the human rating score.
(Daido et al., 2014)	Singing enthusiasm evaluation	Building linear regression model for predicting the human rating of singing enthusiasm.

Table 2.4: Summary table of the previous studies on automatic assessment of singing voice. (continued)

Authors	Goal	Methods
(Schramm et al., 2015)	Solfège assessment	Constructing Gamma probability density functions, building Bayesian classifiers for each note.
(Bozkurt et al., 2017)	Singing voice assessment	Building a Multilayer perceptron model for predicting the human rating.
(C. Gupta, Li, & Wang, 2017)	Singing quality evaluation	Using both linear regression and Multilayer perceptron for predicting the human rating.
(C. Gupta, Grunberg, et al., 2017)	Singing mispronunciation detection	DNN-HMM lyrics-to-audio forced alignment, using an adapted dictionary to detect mispronunciation.

Automatic assessment of instrumental music performance

In Table 2.4, we list the goal and methods of each work. We have identified that in three works, the assessment is conducted at music excerpt level. Several pitch and rhythm score-independent features and score-based features are proposed in (Vidwans et al., 2017)'s work. A SVR regression model is experimented with various feature combinations to predict the expert ratings of Alto Saxophone performance. (Wu & Lerch, 2018) use sparse coding to learn representations in a unsupervised way on the local histogram matrix features. The learned features are used in a SVR model to predict the expert ratings for the percussive music performance. (Pati et al., 2018) use fully-convolutional neural networks and covolutional recurrent neural networks to predict the expert ratings for the saxophone, clarinet and flute music performance. Pitch track and logarithmic Mel band energies are adopted as the input representations of the networks. They also discuss the learned representation for the musicality dimension using network inspection techniques.

In other works, the asessment is done at musical note-level. (Robine & Lagrange, 2006) assess the note quality of saxophone performance. They extract pitch and amplitude related features for stable, crescendo/decrescendo and vibrato notes. Then, the feature values are mapped to the expert ratings. (Knight et al., 2011) develop models to assess the trumpet performance at note-level. They use a SVM classifier to predict the expert ratings with 56 dimensional features of which are mostly spectral features. (Luo et al., 2015) build a bag-of-features classifciation model to detect violin performing mistake. Their note and expression segmentation are achieved using a photo resistor and four rings of surface-mounted light-emitting diodes (SMD LEDs). (Han & Lee, 2014) detect three types of flute performing errors – assembeling error, fluctuated sound and mis-fingering by using handcrafted features and Random Forest classifier.

Table 2.4: Summary table of the previous studies on automatic assessment of instrumental musical performance.

Authors	Goal	Methods
(Robine & Lagrange, 2006)	To assess saxophone notes	Extracting metrics for straight, crescendo/decrescendo and vibrato notes
(Knight et al., 2011)	Trumpet tone quality assessment	Building SVM model to classify trumpet tone quality on 7 scales.
(Han & Lee, 2014)	Detecting common mistakes of flute players	Using handcrafted features, thresholding and Random Forest classifier to detect playing mistakes.
(Luo et al., 2015)	Detection of common violin playing mistakes	Building SVM classifiers for detecting four types of violin playing mistakes.
(Vidwans et al., 2017)	Assessment of student music performance	Building SVR regression model for predicting the human rating score.
(Wu & Lerch, 2018)	Percussive music performance assessment	Using sparse coding to learn the feature, then building SVR model for prediction.
(Pati et al., 2018)	Multi-intrumental student music performance	Using fully-convolutional network or covolutional recurrent network for prediction.

2.4.2 Musical onset detection

Musical onset detection (MOD) is aimed to automatically detect musical onsets such as musical note, singing syllable onsets, in the musical signal. Most of the MOD methods follow this pipeline – (1) calculating audio input representation, (2) onset detection function (ODF) computation, (3) onset selection. In Table 2.5, we list the method used in each MOD work.

Various audio input representations are used for the first step of the pipeline, such as filtered logarithmic magnitude and phase spectrum (Bello et al., 2005; Böck & Widmer, 2013b). The former can be subdivided by the filterbank type – Bark scale bands (Böck, Arzt, et al., 2012), Mel scale bands (Eyben et al., 2010; Schluter & Bock, 2014) or constant-Q bands (Lacoste, 2007; Böck, Krebs, & Schedl, 2012).

Depending on the techniques used, we classify ODF computation methods into three categories:

Unsupervised methods: Methods in this category estimate ODF in an unsupervised way. Earlier methods in this category are based on calculating temporal, spectral, phase, time-frequency or complex domain features, such as energy envelope, high-frequency content, spectral difference, phase deviation and negative log-likelihoods. Bello et al. (Bello et al., 2005) and Dixon (Dixon, 2006) both review these methods thoroughly. The state-of-the-art methods in this category are based on spectral flux feature (Böck, Krebs, & Schedl, 2012). Some variants such as *SuperFlux* (Böck & Widmer, 2013b), local group delay weighting (Böck & Widmer, 2013a) are proposed to suppress the negative effect of vibrato, primarily for pitched non-percussive instruments. The advantage of these methods is that no data is needed for training the ODF, and they are computationally efficient and can often operate in online real-time scenarios.

Non-deep learning-based supervised methods: Some methods in this category are probabilistic model-based, such as using Gaussian autoregressive models to detect the onset change point (Bello et al., 2005). Toh et al. (Toh et al., 2008) propose a method using two Gaussian Mixture Models to classify audio features of onset frames and non-onset frames. Chen (Chen, 2016) detect the onset candidates from two ODFs, extracted features around these candidates,

then used support vector machine technique to classify them.

Deep learning-based supervised methods: The state-of-the-art performance in the MIREX Audio Onset Detection is defined by deep learning-based methods. Lacoste et al. (Lacoste & Eck, 2007; Lacoste, 2007) are the earliest researchers who apply feed-forward or convolutional neural networks (CNNs) to esimtate the ODF. Eyben et al. (Eyben et al., 2010) propose using recurrent neural networks (RNNs) with LSTM units to predict the input frames binarily as onset or non-onset. Schlüter and Böck (Schluter & Bock, 2014) use the similar idea but replace RNNs by CNNs and adopt several novel deep learning techniques, which achieve the best performance in the MIREX Audio Onset Detection task. Huh et al. (Huh et al., 2018) estimate time-to-event (TTE) or time-since-event (TSE) distributions from Mel-spectrograms by a CNN, then use them as a onset density predictor.

The last step of the pipeline – onset selection can be done by peak-picking (Böck, Krebs, & Schedl, 2012) algorithm.

Table 2.5: Summary table of the previous studies on musical onset detection.

Authors	Methods
(Bello et al., 2005)	A tutorial paper, introducing spectral feature-based, probability model-based and negative likelihood onset detection methods.
(Dixon, 2006)	Another review paper, proposing a weighted phase deviation function and a half-wave rectified complex difference for onset detection.
(Lacoste & Eck, 2007)	Feed-forward neural networks for onset detection.
(Lacoste, 2007)	Convolutional neural networks for onset detecion.
(Toh et al., 2008)	Using GMMs model, feature-level and decision-level fusion for singing onset detection.
(Eyben et al., 2010)	Two frame size logarithmic Mel bands input, Bidirectional LSTMs for binary onset classification.
(Böck, Krebs, & Schedl, 2012)	Using logarithmic Constant-Q bands as input, and half wave rectified spectral difference as onset detection function.

Table 2.5: Summary table of the previous studies on musical onset detection. (continued)

Authors	Methods
(Böck, Arzt, et al., 2012)	3 channels logarithmic Bark bands input, using single direction RNN for online onset detection.
(Böck & Widmer, 2013a)	Using local group delay (LGD) weighting of Superflux feature to suppress the false postive onsets caused by vibrato and tremolo.
(Böck & Widmer, 2013b)	Introducing SuperFlux, using Maximum filter to suppress the vibrato ripple in the spectral flux.
(Schluter & Bock, 2014)	3 channels logarithmic Mel bands input, using convolutional neural networks as the detection funcition.
(Chen, 2016)	Multiple onset detection functions, SVM onset classification.
(Huh et al., 2018)	Using time-to-event (TTE) or time-since-event (TSE) prediction for onset detection.

2.4.3 Text-to-speech alignment

Text-to-speech alignment is a process that the orthographic transcription is aligned in temporal axis with the speech audio at word, syllable or phone-level. In Table **??**, we list the method used in each text-to-speech alignment work. Most of the non-commercial alignment tools are built on HTK (Young et al., 2006) or Kaldi (Povey et al., 2011) frameworks, such as Montreal forced aligner (McAuliffe et al., 2017) and Penn Forced Aligner (*Penn Phonetics Lab Forced Aligner*, n.d.). These tools implement an intermediate step of automatic speech recognition (ASR) pipeline, train the HMM acoustic models iteratively using Baum-Welch or Viterbi algorithm and align audio features (e.g. MFCCs) to the HMM monophone or triphone model. Brognaux and Drugman (Brognaux & Drugman, 2016) explore the forced alignment in a small-dataset case using supplementary acoustic features and initializing the HMM silence model by voice activity detection (VAD) algorithm. To predict the confidence measure of the aligned word boundaries and to fine-tune their time positions, Serrière et al. (Serrière et al., 2016) explore an alignment post-processing method using a deep neural network (DNN). Usually, no manually boundary labeled dataset is needed for the HMM acoustic model training which is initialized by flat-start training method (Young et al., 2006). (Pakoci et al., 2016) experiment to train HMM acoustic model by making use of a manually boundary labeled dataset in a small-dataset scenario.

The forced alignment is a language-dependent method, in which the acoustic models should be trained by using a corpus of certain language. Another category of text-to-speech methods is language-independent, which relies on detecting the phoneme boundary change in the temporal-spectral domain (Esposito & Aversano, 2005; Almpanidis et al., 2009). The drawback of these methods is that the segmentation accuracies are usually poorer than the language-dependent counterparts.

Table 2.6: Summary table of the previous studies on text-to-speech alignment.

Authors	Methods
(McAuliffe et al., 2017)	A text-to-speech forced alignment tool built on Kaldi.
(*Penn Phonetics Lab Forced Aligner*, n.d.)	Another text-to-speech alignment tool built on HTK.
(Brognaux & Drugman, 2016)	Experimenting forced alignment with supplementary acoustic features.
(Serrière et al., 2016)	Forced alignment with DNN post-processing.
(Esposito & Aversano, 2005)	Text independent alignment by detecting fast transition of phone onsets.
(Almpanidis et al., 2009)	Detecting phone boundaries using model selection techniques.
(Pakoci et al., 2016)	Forced alignment making use of dataset which has the manually labeled phone boundaries.

2.4.4 Lyrics-to-audio alignment

To goal of lyrics-to-audio alignment is similar to text-to-speech alignment – aligning the lyrics with the singing voice audio at word, syllable or phone-level. In Table 2.7, we list the method used in each lyrics-to-audio alignment work. Most of these works (Mesaros & Virtanen, 2008; Loscos et al., 1999; Fujihara et al., 2011; Mauch et al., 2012; Iskandar et al., 2006; Gong et al., 2015; Kruspe, 2015; G. B. Dzhambazov & Serra, 2015) use the speech forced alignment method accompanied with music-related techniques. Loscos et al. (Loscos et al., 1999) use MFCCs with additional features and also explore specific HMM topologies to take into account of singing aspiration, silence and different pronunciation possibilities. To deal with mixed recordings, Fujihara et al. (Fujihara et al., 2011) use voice/accompaniment separation to extract clean singing voice. They also adopt vocal activity detection, fricative detection techniques to recover the consonant information lost in the separation process.

Additional musical side information extracted from the musical score is used in many works. Mauch et al. (Mauch et al., 2012) use chord information such that each HMM state contains both chord and phoneme labels. Iskandar et al. (Iskandar et al., 2006) constrain the alignment by using musical note length distribution. Gong et al. (Gong et al., 2015), Kruspe (Kruspe, 2015), Dzhambazov and Serra (G. B. Dzhambazov & Serra, 2015) all use syllable/phoneme duration extracted from the musical score as side information, and decode the alignment path by duration-explicit HMM models. Chien et al. (Chien et al., 2016) introduce an approach based on vowel likelihood models. Chang and Lee (Chang & Lee, 2017) use canonical time warping and repetitive vowel patterns to find the alignment for vowel sequence. Some other works achieve the alignment at music structure-level (Müller et al., 2007) or line-level (Y. Wang et al., 2004).

Table 2.7: Summary table of the previous studies on lyrics-to-audio alignment.

Authors	Methods
(Loscos et al., 1999)	Forced alignment with addtional features and special HMM topologies.
(Iskandar et al., 2006)	Forced alignment using musical note length distribution.
(Fujihara et al., 2011)	Treating mixed recordings with voice/accompaniment separation, forced alignment with phone filler topologies.
(Mauch et al., 2012)	Forced alignment using chord side information.
(Gong et al., 2015)	Forced alignment with HSMM model using vowel duration information.
(Kruspe, 2015)	Forced alignment with duration-explicit HMM model.
(G. B. Dzhambazov & Serra, 2015)	Forced alignment with duration-explicit HMM model.
(Chien et al., 2016)	A method based on vowel likelihood model.
(Chang & Lee, 2017)	A method based on canonical time wraping and repetitive vowel patterns.
(Müller et al., 2007)	Structure-level DTW alignment.
(Y. Wang et al., 2004)	Line-level alignment making use of rhythm, chrous and singing voice detection.

2.4.5 Neural acoustic embeddings

Neural acoustic embeddings is a technique to convert variable-length acoustic sequence into fixed-length vector using neural networks. It is a common technique that is adopted in speech field, and applied to various tasks such as Query-by-Sample search and speech recognition. In those tasks that involve measuring the similarity between speech segments, acoustic embeddings generated from neural networks allows a more efficient and accurate computation because the alignment between variable-length speech segments can be avoided. In Table 2.8, we list the method used in each speech neural acoustic embedding work.

To embed variable-length representation of acoustic word segment such as MFCCs into fixed-length vector. (Kamper et al., 2016) experiment two neural network architectures – classification CNN and Siamese CNN. Softmax units are used in the output layer of the classification CNN, which allows it to classify input word segment into word categories in a fully-supervised fashion. The vector output from the last CNN layer is taken as the word embedding. The hinge cosine triplet loss is used by the Siamese CNN of which the network training is done in a semi-supervised way. The penultimate layer of the Siamese CNN is taken as the word embedding such that the dimension is adjustable. RNN is a natural choice for the sequential data modelling. In the work of Settle et al. (Settle & Livescu, 2016), the CNN is replaced by RNN in both classification and Siamese architectures. A weighted random sampling method is also devised to accelerate the Siamese RNN training. The acoustic word embeddings learned from the Siamese RNN is then used in a Query-by-Example search task with a small training dataset (Settle et al., 2017).

Apart from exploring different neural network architectures to obtain an efficient word embedding, we can also take advantage of multiple information sources. Zeghidour et al. (Zeghidour et al., 2016) jointly learn phoneme and speaker embeddings by a single Siamese network which minimize simutanously two objectives. In the work of He et al. (He et al., 2017), acoustic word segment and text word segment are embeded by two different RNNs. The embeddings of the two different sources are projected into a common space and used in the objective function in a mixed way.

Neural acoustic embeddings can also be learned in an unsupervised way. Chung et al. (Chung et al., 2016) adopt sequence-to-sequence model commonly used in natural language processing tasks to learn the word embeddings. They experiment skipgrams and continous bag-of-words training methods and show a superior performance than simply reconstructing the input representation (Chung & Glass, 2018)

Table 2.8: Summary table of the previous studies on neural acoustic embeddings.

Authors	Methods
(Kamper et al., 2016)	Generating acoustic word embeddings using CNN and Siamese network.
(Settle & Livescu, 2016)	Similar to (Kamper et al., 2016), but using RNN rather than CNN.
(Settle et al., 2017)	Using Siamese network word embeddings for Query-by-Example search task.
(Zeghidour et al., 2016)	Using Siamese network to jointly learn phoneme and speaker embeddings.
(He et al., 2017)	Embedding acoustic words and character sequences simutanously by multi-objectives.
(Chung et al., 2016)	Learning acoustic word embeddings in an unsupervised way using sequence-to-sequence model.
(Chung & Glass, 2018)	Similar to (Chung et al., 2016), but using skipgrams and continuous bag-of-words trainings rather than reconstruct the input.

2.4.6 Evaluation metrics

Onset detection metrics

A simple evaluation metric for onset detection – **onset detection accuracy** is adopted in many state of the art works (Eyben et al., 2010; Böck, Krebs, & Schedl, 2012; Böck, Arzt, et al., 2012; Böck & Widmer, 2013a, 2013b; Schluter & Bock, 2014). We use the same metric for our evaluation. To define a correctly detected onset, a tolerance threshold of $\tau = 25$ms is chosen. If the detected onset o_d lies within the tolerance of its ground truth counterpart o_g: $|o_d - o_g| < \tau$, we consider that it's correctly detected. A more strict metric can be defined by requiring that the label of the detected onset and that of the ground truth are identical. The F-measure is a number between 0 and 1 calculated as the harmonic mean of the precision and recall. Precision is the ratio between the number of correctly detected onsets and all detected onsets, and recall is the ratio between the number of correctly detected onsets and the total annotated onsets.

$$\text{Precision} = \frac{\text{number of correctly detected onsets}}{\text{number of all detected onsets}} \tag{2.1}$$

$$\text{Recall} = \frac{\text{number of correctly detected onsets}}{\text{number of total annotated onsets}} \tag{2.2}$$

$$\text{F-measure} = 2 \cdot \frac{\text{Precison} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.3}$$

The metric presented above can also be used to evaluate alignment algorithm. However, we apply another metric for the alignment evaluation – **percentage of correct segments**, which is defined as the ratio between the duration of correctly aligned segments and the total duration of the music piece. This metric has been suggested by Fujihara et al. (Fujihara et al., 2011) in their lyrics alignment work.

$$\text{Percentage of correct segments} = \frac{\text{duration of correctly aligned segments}}{\text{total duration of the music piece}} \tag{2.4}$$

Binary classification evaluation metric

In binary classification task, such as mispronunciation detection, which classifies the singing syllable or phoneme segment into mispronounced or correctly pronounced class, we use classification accuracy as the evaluation metric. The classification accuracy is defined as:

$$\text{Accuracy} = \frac{TP + NP}{\text{number of total population}} \tag{2.5}$$

Where TP is true positive – correctly classified as positive (e.g., mispronunciation in mispronunciation detection task), and NP is true negative – correctly classified as negative (e.g., correct pronunciation in mispronunciation detection task).

Similarity measurement metrics

In this dissertation, the acoustic embedding will be always used as a representation for the similairy (distance) computation. Thus the evaluation of acoustic embedding needs to be done with the help of a simlarity (distance) measure. The ground truth label is set to 1 if two singing segments belong to the same class (phoneme, special pronunciation, etc.), 0 vice versa. We report the average precision (AP) between the pairwise similarities of the segments and the ground truth as the evaluation metric. The AP is used previously to evaluate speech word acoustic embedding (Kamper et al., 2016; Settle & Livescu, 2016). It is also suggested as the metric for imbalanced test set (Davis & Goadrich, 2006), which is the case of the pronunciation aspect evaluation.

AP is defined as the area under the precision-recall curve. In pratice, it is calculated by a finite sum (Su et al., 2015):

$$\text{Average precision} = \sum_{i=1}^{n} p(i)\Delta r(i) \tag{2.6}$$

where $p(i)$ is the precision in index i, and $\Delta r(i)$ is the change in recall from $i-1$ to i.

In Figure 2.3, we illustrate an example of calculating AP for the pairwise similarities of three segments, e.g. segment 1 and 2

belong to the class A, and their ground truth similarity and calculated similarity are respectively 1.0, 0.8.

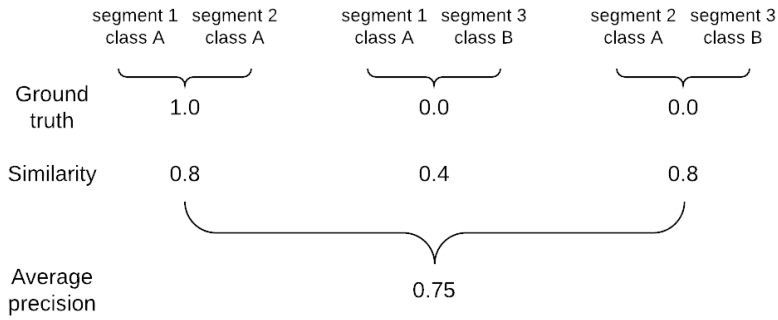


Figure 2.3: An example of calculating average precision for the pairwise similarities of three segments.

2.5 Relevant technical concepts

This thesis uses several well-studied machine learning models and techniques to tackle the automatic singing voice pronunciation assessment problem. There are extensive resources available to study those models and techniques, hence a brief mention of those techniques with references to the resources are provided in this section for the background study.

2.5.1 Deep learning

Deep learning is a subfield of machine learning concerned with algorithms inspired by artificial neural networks. Deep learning refers to large neural networks and deep refers to the number of neural network layers (Dean, n.d.). It is a scalable algorithm which can be fit into huge amounts of data. Almost all value today of deep learning is through supervised learning – learning from labeled data (Ng, n.d.).

Deep learning is a hierarchical feature learning. Yoshua Bengio described deep learning as algorithms seek to exploit the unknown structure in the input distribution in order to discover good

representations, often at multiple levels, with higher-level learned features defined in terms of lower-level features (Bengio, 2012).

Although equally requiring network weights initialization, and using the backpropagation algorithm to update the network parameters, deep learning differs from the traditional artificial neural networks (ANNs) regarding below aspects:

- Using specific layers to deal with different data types, e.g., convolutional layers which learns automatically the representations for image data, recurrent layers for modeling sequential data.
- Using more advanced non-linear activation functions, which facilitates to train very deep architectures.
- Devising new techniques to help model generalization.
- Using new optimizers to facilitate rapid model convergence.

We will briefly introduce each of the core techniques in deep learning. The purpose is to provide adequate references for the background study, and hence the section is not comprehensive.

Convolutional neural networks

Fully-connected neural network (also known as Multilayer perceptron – MLP) is the most basic type of the neural networks. Each layer in MLP is contained by a set of neurons, where each neuron is fully-connected to all neurons in the previous layer (*Stanford University CS231n: Convolutional Neural Networks for Visual Recognition*, n.d.). Convolutional neural networks (CNNs) shares some common characteristics with MLP. For example, they both have learnable weights and biases. Each neuron receives inputs, performs dot product and follows by a non-linear activation function. And the whole network can be expressed by a single differentiable score function, which has loss function on the last layer.

The difference between CNNs and MLP is that firstly the input of CNNs is three-dimensional image or audio spectrogram, whereas the input of MLP can be only a vector. Then, CNNs have more types of layers than MLP. The most important layers in CNNs are convolutional layer (Conv), fully-connected layer (FC) and pooling layer.

Conv layer computes the output of neurons that are connected to local regions in the input, each computing a dot product between their weights and a small region they are connected to in the input volume.

Pooling layer performs a downsampling operation along the spatial dimensions (width, height).

FC layer computes the class scores. As with ordinary ANNs and as the name implies, each neuron in this layer will be connected to all the numbers in the previous volume.

Compared with MLP, CNNs are more efficient in learning spatially local input patterns by using the "filters" and convolution operation. Stacking many Conv layers leads to filters that learn global input patterns. Additionally, the receptive fields of the deeper layers increasingly cover a larger area than their shallow counterparts (Wikipedia, n.d.-a). CNNs benefit as well from the weight sharing, where each filter convolves on the input image and form a feature map. The learned weights are contained in the filter itself, and each output feature map uses the same weights, which greatly reduces the number of parameters. As a consequence, CNNs are more memory-saving and better to learn different levels of representations than MLP. Pooling layer is used in CNNs to reduce the spatial size of the representation, and thus reduce the number of parameters of the network. It is often inserted in-between two Conv layers. The most often used pooling operation is max-pooling, which takes the maximum value of a non-overlapped filter, e.g., 2×2, on the feature map.

Recurrent neural networks

Recurrent neural networks (RNNs) is another type of deep learning architecture which is commonly used to model the symbolic or acoustic sequential data such as text, speech, and music. RNNs is called recurrent because it performs the same operation for the input of each timestamp of the sequence. The calculation of the output of the current timestamp depends on all the previous computations with the help of hidden states. The hidden state is the memory of the network, which is calculated based on the previous state and the input of the current timestamp.

The basic RNN architecture introduced above (also known as vanilla RNNs) suffers from the vanishing or exploding gradient problem (Pascanu, Mikolov, & Bengio, 2013), which make them hard to learn long-term dependencies in a sequence. Certain types of RNNs such as Gated Recurrent unit (GRU), Long-short term memory networks (LSTMs) have been designed to cope with such problems. The most popular extension of the vanilla RNNs is the LSTMs (Hochreiter & Schmidhuber, 1997) of which the calculation of the hidden state is modified to contain various cell states and connections. An extensive walk-through of the complicate LSTMs cells can be found in this blog post (Olah, 2015).

Sometimes we want that the output of the current timestamp not only depends on the previous hidden states but also the future ones. For example, to detect if a frame of the spectrogram is a syllable or instrumental onset, we usually want to check both the spectral context in both left and right directions. Bidirectional RNNs (BiRNNs) (Schuster & Paliwal, 1997) allows us to access to the information of the future timestamps by stacking two RNNs on top of each other.

Non-linear activation functions

Activation functions (or non-linearity) is a set of operations exerted on the output of neurons, which introduce non-linearity to the network in order to adapt to the complexity of the input. The common activation functions could be used in a neural network are sigmoid, softmax, tanh, Rectified Linear Unit (ReLU), Exponential Linear Unit (ELU) (Clevert, Unterthiner, & Hochreiter, 2015), etc. Sigmoid squashes the input into a range between 0 and 1, which is commonly used on the last layer output for the tasks such as multilabel classification and regression (*Stanford University CS231n: Convolutional Neural Networks for Visual Recognition*, n.d.). Softmax is an extension of sigmoid function which is also used on the last layer output, however, for the task of multiclass logistic regression. Tanh function squashes the input into a range between -1 and 1 and gives a zero-centered output, which is commonly used as the default activation function in the vanilla RNNs. ReLU is a very popular activation function which is used extensively in CNNs. It is a linear activation thresholded at zeros. The use of ReLU holds two major benefits – accelerating the convergence of the network train-

ing, computational non-expensive. The main drawback of ReLU is that some neurons in the network could be "died" during the training due to the zero thresholding on the entire left x-axis. Some extensions of ReLU attempt to fix this problem by introducing a negative slope on the left x-axis, such as that ELU uses an exponential function.

Regularization

Large neural networks with many trainable parameters are prone to overfit on small datasets. There are several ways of controlling the capacity of neural networks to prevent overfitting (*Stanford University CS231n: Convolutional Neural Networks for Visual Recognition*, n.d.).

L2 regularization is the most common form of regularization. It is implemented normally by penalizing the squared magnitude of all parameters in the objective. That is, for every weight w in the network, we add the term $1/2\lambda w^2$ to the objective, where λ is the regularization strength.

Dropout is an extremely effective, simple regularization technique introduced by Srivastava et al. (Srivastava, Hinton, Krizhevsky, & Salakhutdinov, 2014). While training, dropout is implemented by only keeping a neuron active with some probability p (a hyperparameter). In practice, it is common to use a single, global L2 strength combining with dropout applied for all layers.

Batch Normalization is a network weights initialization technique developed by Ioffe and Szegedy (Ioffe & Szegedy, 2015), explicitly forcing the activations throughout a network to take on a unit Gaussian distribution at the beginning of the training. In the implementation, applying this technique usually amounts to insert the Batch Normalization layer immediately after fully-connected layers or convolutional layers, and before non-linearities. In practice networks that use Batch Normalization are significantly more robust to bad initialization and accelerate the network training (*Stanford University CS231n: Convolutional Neural Networks for Visual Recognition*, n.d.).

Batch Normalization can also be looked as a way of regularization which is similar to dropout in the sense that it multiplies each hidden unit by a random value at each step of training. In this

case, the random value is the standard deviation of all the hidden units in the minibatch. Because different examples are randomly chosen for inclusion in the minibatch at each step, the standard deviation randomly fluctuates. Batch norm also subtracts a random value (the mean of the minibatch) from each hidden unit at each step. Both of these sources of noise mean that every layer has to learn to be robust to a lot of variation in its input, just like with dropout (Goodfellow, 2016).
Early stopping: it is a simple technique to prevent overfitting by stopping the training iteration when the loss of the validation data doesn't go down certain training epochs (also known as patience).

Loss functions

A loss function or a cost function is a function representing the cost associated with the algorithm output. It is a method that evaluates how well the algorithm models the data. A optimization problem used during the training phase of a deep learning model seeks to minimize a loss function. The choice of the loss function depends on the type of the deep learning task:
Classification: in binary classification, the model prediction p is output from a sigmoid activation function. The loss function is a binary cross-entropy loss:

$$loss_{binary} = -(ylog(p) + (1-y)log(1-p)) \tag{2.7}$$

In multi-class classification, we calculate separate loss for each class label per observation and take the sum:

$$loss_{multiclass} = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{2.8}$$

where c is the class label; p is the predicted probability observation o of class c; y is the binary indicator (0 or 1) if class label c is the correct classification for observation o.
Similarity measures with Siamese network: Siamese network (H. Gupta, 2017) is a special type of neural network architecture. The network learns to differentiate between two or three inputs and learns the similarity between them. Siamese network contains two or three base networks which share the same weights. The objective

of a Siamese network is to differentiate between input samples, thus a contrastive loss is used to achieve this goal.

In a triplet Siamese network, we use weak supervision in the form of pairs of samples labeled as same or different. The output of the base network has linear activation functions. In order to learn the model parameters, we simultaneously feed three samples to the model. One input sample is an "anchor", x_a, the second is another sample with the same label, x_s, and the third is a sample corresponding to a different label, x_d. Then, the network is trained using a "cos-hinge" loss (or triplet contrastive loss) (Settle & Livescu, 2016):

$$l_{coshinge} = \max\{0, m + d_{cos}(x_a, x_s) - d_{cos}(x_a, x_d)\} \tag{2.9}$$

where $d_{cos}(x_1, x_2) = 1 - cos(x_1, x_2)$ is the cosine distance between x_1, x_2.

Other loss functions are used for other types of machine learning tasks. For example, L1 or L2 norm losses are used for the regression task. Since other loss functions will not be applied in the tasks of this dissertation, please consult this reference (*Stanford University CS231n: Convolutional Neural Networks for Visual Recognition*, n.d.) for their details.

Optimizers

The gradient of a deep learning model is calculated by the back-propagation algorithm. The gradient is then used to update the model parameters. There are many approaches – optimizers to perform this update:

Stochastic gradient descent (SGD): This is the simplest form to update the model parameters along the negative gradient direction.

$$x += -lr \cdot d_x \tag{2.10}$$

where x is the vector of parameters, and d_x is the gradient. The learning rate lr is a fixed constant. In deep learning optimization, stochasticity of the gradient descent is represented by randomly choosing using a mini-batch of samples to calculate the gradient.

SGD with momentum: The simplest SGD algorithm has a problem in optimization – the network get stuck in a shallow local minimum. With the momentum, the network can slide through such a minimum (MATLAB, n.d.):

$$v = \mu \cdot v - lr \cdot d_x \tag{2.11}$$
$$x = x + v \tag{2.12}$$

where v is the velocity of the gradient that is initialized at zero. μ is the hyperparameter of momentum. When the gradient is zero, there is still a possibility that the optimization maintains a velocity of $\mu \cdot v$, which helps the network to surpass the local minimum. Another more advanced optimization technique is SGD with Nesterov momentum incorporated with "lookahead" gradient term, which works slightly better than the standard momentum (Bengio, Boulanger-Lewandowski, Pascanu, & Montreal, 2012).

Adam: the standard SGD or SGD with momentum use a fixed learning rate equally for all parameters. In practice, it requires to tune the learning rate to reach a better model convergence. Many adaptive learning rate methods have been devised to automatically meet this requirement. Adam (Kingma & Ba, 2014) is one of the most popular adaptive learning rate methods recommended as the default algorithm to use (*Stanford University CS231n: Convolutional Neural Networks for Visual Recognition*, n.d.). It uses a smooth version of gradient and a "cache" variable to perform the parameter update.

2.5.2 Hidden Markov models and hidden semi-Markov models

A Hidden Markov model (HMM) is a statistical model of which the goal is to recover a data sequence that is not observable. It has a wide range of applications such as lyrics-to-audio alignment, audio-to-score alignment, speech recognition, speech synthesis, handwriting recognition, machine translation and alignment of biosequences. An HMM consists of four basic components:

1. Hidden states: they are not observable and need to be inferred from the observations.

2. Observations: they are observable and depend on the hidden states.
3. State transition probabilities: they are the probabilities of transiting between hidden states.
4. Emission probabilities: they are the probabilities that observations can be emitted from hidden states.

A Markov process has the property that the conditional probability distribution of future states depends only upon the present states, and thus it is memoryless. We give an example of speech recognition application to better illustrate the concepts of HMM. The basic HMM in speech recognition is the monophone model which consists of commonly three sub-phoneme hidden states. The observations are the acoustic representation of the speech signal which needs to be inferred. A common acoustic representation could be Mel-frequency cepstrum coefficients (MFCCs). The state transition probabilities are either the probabilities of transiting between sub-phoneme hidden states or those of transiting between the monophone HMM. The emission probabilities are the probabilities emitting an acoustic representation from sub-phoneme hidden states, which are usually modeled by Gaussian mixture models (GMMs) or neural networks (NNs).

In practice, we usually encounter two types of HMM related problems. The first problem is to train the model parameters, mainly the state transition probabilities and emission probabilities given the hidden states and the observation. The second problem is to recover the hidden state sequence from the observations given the model parameters. The solution of these two problems are quite mature and can be consulted in many references such as HTK book (Young et al., 2006), Rabiner's HMMs tutorial (Rabiner, 1989). Commonly, the first problem can be solved by the Baum-Welch algorithm and the second problem can be solved by Viterbi algorithm.

A hidden semi-Markov model (HSMM) is an extension of the HMM where the time elapses on a hidden state is defined explicitly by an occupancy distribution. In a standard HMM, the occupancy distribution is defined implicitly by a geometric distribution. However, in the HSMM, the probability of being a change in the hidden state depends on the amount of time has been elapsed on the cur-

rent state (Wikipedia, n.d.-b). The two basic problems mentioned above in HMM become more complicated for HSMM such that the Baum-Welch algorithm and the Viterbi algorithm need to be modified. The detailed description of HSMM and the adaptation of these two algorithms for HSMM can be consulted in Guédon's work (Guédon, 2007).

2.5.3 Speech recognition tools

Automatic speech recognition is an important research topic in ICT, which has received a great attention from a large research community over the past few decades and has evolved into a mature research area with state of the art methods (Huang & Deng, 2010). There is a potential to use some of its technologies and tools to analogous task in the automatic assessment of singing voice, such as syllable and phoneme segmentation and mispronunciation detection.

Hidden Markov Model Toolkit (HTK) (*HTK Speech Recognition Toolkit*, n.d.) and Kaldi (Povey et al., 2011) are the two most popular tools for constructing a speech recognition related system. The first version of HTK can be dated to the year 1993. It is a mature toolkit which has a large user and developer communities and a comprehensive documentation (Young et al., 2006). Being a younger project started in 2009, Kaldi is getting more attention because of its extensive functionalities, supportive community and many ready to use recipes for various speech recognition tasks. With the help of the Kaldi recipes, one can configure a speech recognition system in a few lines of code.

Chapter 3

Automatic assessment of singing voice pronunciation of jingju music

Automatic assessment of singing voice of jingju music has not been explored systematically, which means that the challenges, opportunities and relevant research problems have not been formally studied. This chapter presents the attempts to open up this research topic. We first elucidate the important role of pronunciation played in jingju singing training. Then we introduce several relevant research problems, with a review of the state of the art for jingju music or other music traditions in the context of CompMusic project. We present the background of all the relevant research problems. We formulate the thesis problems of syllable and phoneme segmentation, mispronunciation detection for special pronunciation, and pronunciation similarity measures at phoneme level. The main objectives of the chapter are:

1. To present, and discuss the role of pronunciation in jingju singing training.

2. To identify, present, and discuss main challenges to automatic

assessment of singing voice pronunciation in jingju music.

3. To identify, present, and discuss main opportunities in automatic assessment of singing voice pronunciation in jingju music
4. To identify several relevant research problems within the context of jingju music and identify key challenges in addressing them, as a way to indicate future work in singing voice assessment.
5. From the relevant problems, identify a subset of research problems and formulate them in detail, to be addressed in the scope of this dissertation.

3.1 The role of pronunciation in jingju singing training

Assessment of singing performance can be conducted in various musical dimensions such as intonation, rhythm, loudness, tone quality and pronunciation. The automatic assessment method can be devised either for a special dimension or the overall performing quality (C. Gupta, Li, & Wang, 2017). Due to the various and complicate conventions existed in jingju singing performance, and also the strictness of jingju singing training, the automatic system conceived for the assessment of jingju singing needs to have the ability to judge the performance in each dimension. However, due to time and energy constraints, it is not possible to address the relevant research problems of all the musical dimensions in this dissertation.

In this section, we attempt to answer the question: how the jingju singing teachers and students value the importance of each musical dimension? By answering this question, we can identify the most important dimension consistently considered by teachers and students – pronunciation.

3.1.1 Jingju singing training and correction occurrence

Jingju singing is traditionally taught between teacher and student by using the mouth/heart (口传心授, oral teaching) and face-to-face

methods – "Jingju tuition requires demonstration, and teachers tell students the secrets for certain skills that they learned from their masters or that they worked out from their experience. The close relationship of the teacher-student or the master-disciple is based on the mouth/heart teaching method that stresses through oral instruction and intuitive understanding. Imitation is certainly the first step, and it is crucial for our learning proces... not even one component in the 'four skills (singing, speech, dance-acting, combat)' can be learned by the student himself. Much of the nuance of the singing can only be learned from face-to-face teaching." (Li, 2010)

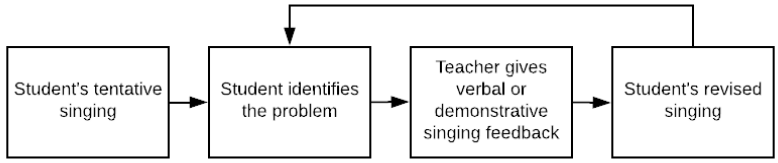


Figure 3.1: The flowchart of a single correction occurrence.

After five months research stay in NACTA (National Academy of Chinese Theatre Arts, leading institute in China dedicated to the training of professionals in performing and studying traditional Chinese opera), we had a firsthand experience of the month/heart teaching method of jingju singing. In class, the teacher teaches several melodic lines selected from an aria. In the first part of the class, the teacher gives a short introduction to the teaching content such as the story and the character setting of the aria, the special pronunciations. Then she/he gives a demonstrative singing of these lines. In the second part of the class, the students imitate the demonstrative singing line by line, and the teacher corrects the imitations. The process of the second part can be generalized as (i) the teacher asks the students to give a tentative singing individually at melodic line-level, or syllable-level. (ii) Then the teacher identifies the singing problems, (iii) gives verbal or demonstrative singing feedback. (iv) Finally, the students do a revised singing with the feedbacks in mind. The step from (ii) to (iv) could be iterated until the student's singing satisfies the teacher's criteria. We name one single such process as a **correction occurrence** (see Figure 3.1).

The verbal feedback is a semantic comment given by the teacher. It is the description that is aimed to help the student to improve her/his singing performance, and it is the most valuable information which can clarify the singing problems.

In paper (Geringer & Madsen, 1998), the musicians have rated the performance of western arias in 5 musical dimensions: phrasing/expression, intonation, rhythm, loudness, tone quality. In our study, we borrow the concept of musical dimensions for music performance assessment and adapt them according to jingju singing background.

Almost all jingju aria contains lyrics, and as we will prove in later chapters – to be able to pronounce the singing lyrics accurately is a key skill in jingju singing, we thus add the pronunciation as an independent dimension to the dimension set mentioned above. Besides, we discard phrasing/expression because it is a "meta-dimension" constructed above other basic dimensions – "A musician accomplishes this by interpreting the music, from memory or sheet music, by altering tone, tempo, dynamics, articulation, inflection, and other characteristics"[1]. Overall, 5 dimensions will be taken into account in this paper – intonation, rhythm, loudness, tone quality and pronunciation. Accordingly, we give their definitions:

- Intonation: accuracy of pitch in singing.
- Rhythm: singing a rhythmic pattern on time, which means that the notes or syllable are not ahead of the time or behind the time.
- Loudness: the dynamic loudness variation between notes/syllables or phrases.
- Tone quality: the color or timbre of the singing voice.
- Pronunciation: the act or result of producing the sounds of speech, including articulation and stress.

In the next section, we explain our methods – classifying teachers' correction occurrence and surveying the students. These methods aim to answer the question: how teachers and students value the

[1] https://en.wikiquote.org/wiki/Musical_phrasing Retrieved 25 July 2018

importance of each jingju singing dimensions – intonation, rhythm, loudness, tone quality and pronunciation. By classifying the correction occurrences, we will find out the dimensions on which teachers lay stress or students tend to have problems. On the other hand, we conduct a simple survey to investigate the importance of each dimension from the students' perspective.

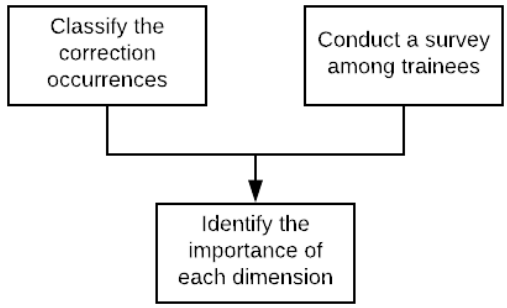


Figure 3.2: The identification process of the importance of musical dimensions.

Correction occurrence analysis

During the research stay in NACTA, we audited and recorded the audio from three singing classes. Three class was taught respectively by three professional teachers, which contain solo and chorus practices. We recorded the class teaching and practicing audio content by using a SONY PCM-D50 stereo portable audio recorder. Only the solo practices are kept for the analysis because they can reveal the singing problems of an individual student, whereas the individual voices are blurred in the chorus practice recordings. The audio excerpts of each correction occurrence are then edited and visualized by signal processing tools – pitch contour, loudness contour and spectrogram, which is helpful in identifying the singing problem, especially when the teacher's verbal feedback is too abstract to extract any effective information.

Table 3.1 depicts the information of aria name, role-type, student number in the class, melodic line number practiced in the class and correction occurrence number collected from the recordings.

Table 3.1: The statistics of the correction occurrence analysis materials.

Aria name	Role-type	#Student	#Melodic line	#Correction occurrence
武家坡 WuJiaPo	laosheng	3	11	20
太真外传 TaiZhen WaiZhuan	qingyi	3	3	21
捉放曹 ZhuoFang Cao	hualian	2	28	21

An example of reading this table is "Three students were involved in the laosheng class WuJiaPo. 11 melodic lines were taught, and 20 correction occurrences were collected from the recordings".

The ratios between the melodic line number and the correction occurrence number are widely different for the three classes (Table 3.1). For example, during the TaiZhenWaiZhuan class, three students practiced three lines and were corrected 21 times, which results in a ratio of 1/7. However, during the ZhuoFangCao class, two students practiced 28 lines and also were corrected 21 times, which has a ratio of 4/3. The correction frequency depends on several factors, such as the students' singing levels, the teacher's teaching method. The low singing level students tend to receive more corrections than those who have high singing levels.

For each occurrence, we analyze the target recordings and the teacher's verbal feedback. Additionally, to achieve the visual analysis, their pitch, loudness contours and spectrogram are also presented.

We firstly classify the correction occurrences into five dimensions – intonation, rhythm, loudness, tone quality and pronunciation. A correction occurrence can be classified into more than one dimension. For example, the correction with the verbal feedback "don't be sloppy, sing it with solidity, make the tone quality sounds round." can be classified into intonation (irregular vibrato), loudness (unstable loudness contour) and tone quality (higher harmonics too clear), by analysing comparatively between the teacher's

demonstrative singing and student's tentative singing. Furthermore, a finer inspection of each correction occurrence is conducted, where we identify the detailed elements.

Five correction occurrences are taken as examples to showcase our analysis. For each one, we list its aria name, melodic line, target syllable, the teacher's verbal feedback, the dimensions classified. Finally, we give a short explanation accompanied by the visualization to justify our classification process.

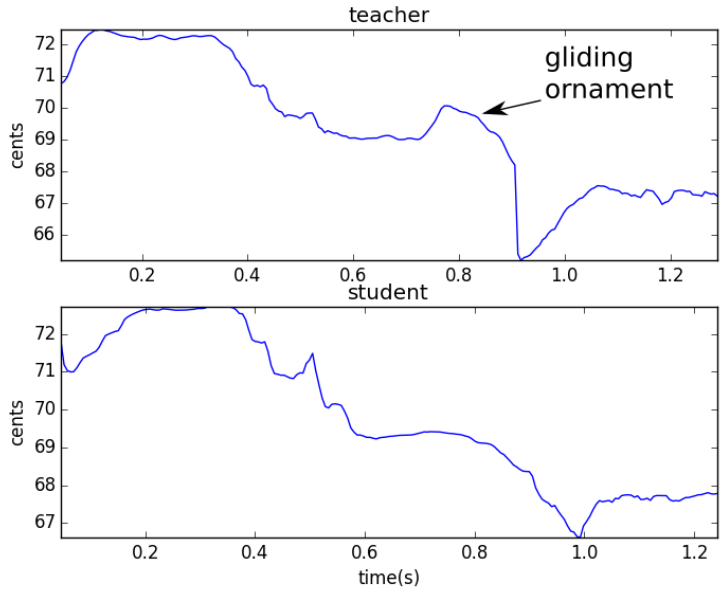


Figure 3.3: The pitch contours of the syllable "yuan" for occurrence 1.

Occurrence 1:

- Aria: TaiZhenWaiZhuan (太真外传)
- Melodic line: yi yuan na chai yu he qing yuan yong ding (一愿那钗与盒情缘永定)
- Target syllable: yuan (缘)
- Teacher's verbal feedback: it didn't jump up. (没跳起来)
- Dimension: intonation
- Explanation: The syllable's second tone in the teacher's demonstrative singing has a pitch gliding (ornament). However, the gliding in the student's version is not apparent (Figure 3.3).

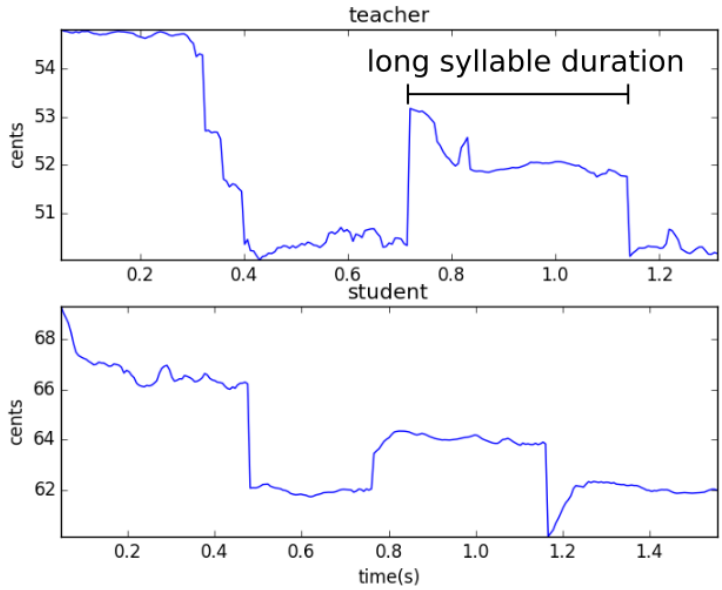


Figure 3.4: The pitch contours of the syllables "dang nian jie bai" for occurrence 2.

Occurrence 2:

- Aria: ZhuoFangCao (捉放曹)
- Melodic line: dang nian jie bai yi lu xiang (当年结拜一炉香)
- Target syllables: dang nian jie bai (当点结拜)
- Teacher's verbal feedback: swing apart these four syllables (1,2,3,4 四个字甩开)
- Dimension: rhythm
- Explanation: In teacher's demonstrative singing, the temporal duration of the third syllable "jie" has been prolonged, in contrast with the other three syllables, which can be observed by the pitch contour (Figure 3.4).

Occurrence 3:

- Aria: TaiZhenWaiZhuan (太真外传)
- Melodic line: yang yu huan zai dian qian shen shen bai ding (杨玉环在殿前深深拜定)
- Target syllable: yang (杨)

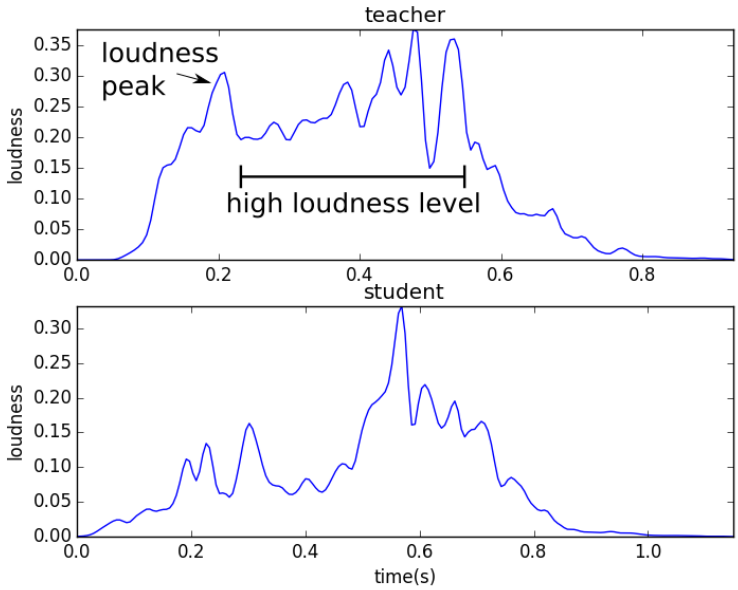


Figure 3.5: The loudness contours of the syllable "yang" for occurrence 3.

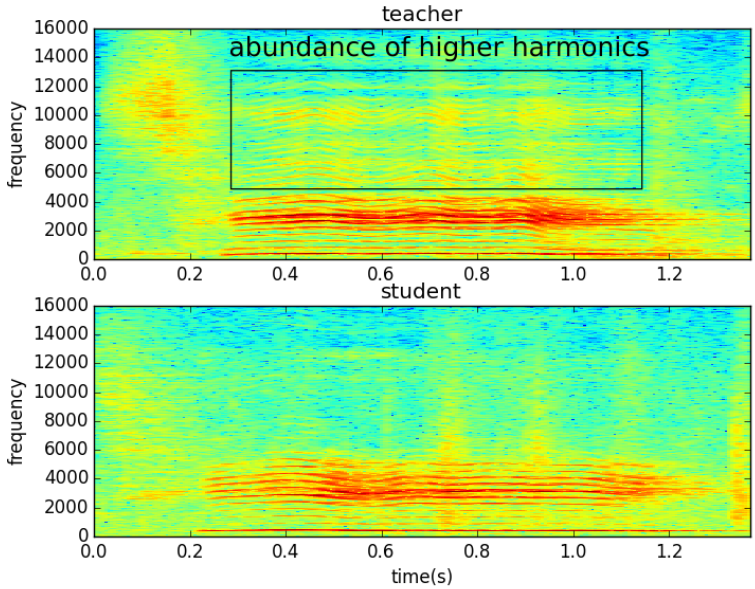


Figure 3.6: The spectrograms of the syllable "yang" for occurrence 3.

- Teacher's verbal feedback: emphasizing the nasal voice (an 鼻音要突出)
- Dimensions: loudness and tone quality
- Explanation: In teacher's demonstrative singing, a prominent loudness peak can be found in the head of the syllable, which maintains a high loudness level in the belly (Figure 3.5). We also can observe that the higher harmonics are abundant from the spectrogram (Figure 3.6).

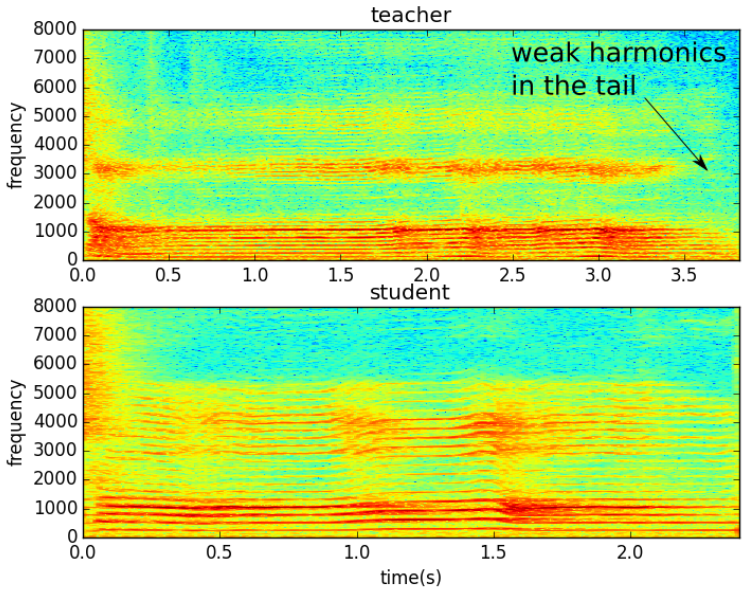


Figure 3.7: The spectrograms of the syllable "shang" for occurrence 4.

Occurrence 4:

- Aria: ZhuoFangCao (捉放曹)
- Melodic line: xian xie zuo le na wa shang shuang (险些做了那瓦上霜)
- Target syllable: shang (上)
- Teacher's verbal feedback: terminate the sound at /ng/ (sound 收音收到 ng)
- Dimension: pronunciation

- Explanation: The teacher's demonstrative singing is one octave lower than the student's singing. The teacher's feedback emphasizes the pronunciation quality of the syllable tail sound – /ng/. His demonstrative singing contains fewer harmonics in the tail than the student's singing (Figure 3.7).

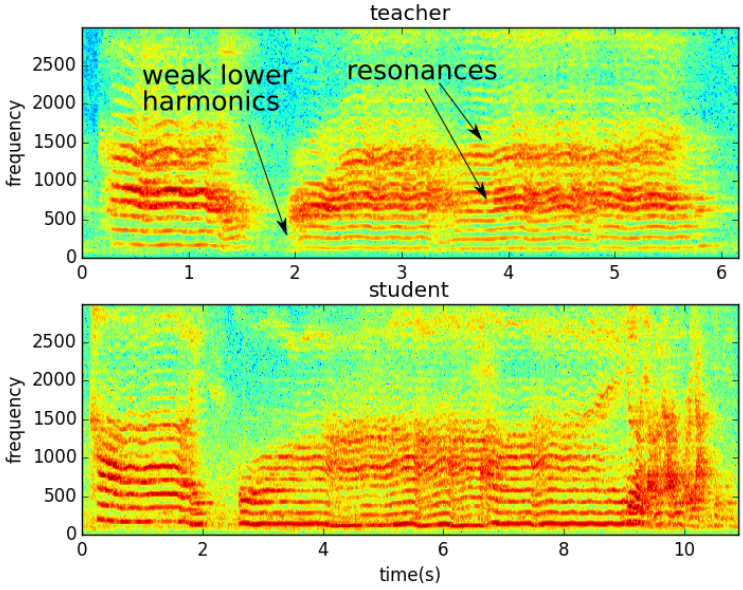


Figure 3.8: The spectrograms of the syllables "kai huai" for occurrence 5.

Occurrence 5:

- Aria: WuJiaPo (武家坡)
- Melodic line: jian liao na zhong da sao xi wen kai huai (见了那众大嫂细问开怀)
- Target syllable: kai huai (开怀)
- Teacher's verbal feedback: adjust the breath, make the sound solid even if you sing in the low register (要用气，低调们也要放实在了)
- Dimension: tone quality
- Explanation: This feedback has twofold of meaning. First is to take enough breath, and have enough air in the chest to sing. Second is to adjust the body's resonance position to make the sound

more solid. This can be observed as in the spectrogram of the teacher's demonstrative singing, which contains less energy for the lower harmonics and the prominent resonances (energy) in around 750 Hz and 1250 Hz (Figure 3.8).

A survey among students

We conduct a simple survey of another nine students to investigate the importance of each dimension from their perspectives. The survey contains two questions:

Please rate the use frequency of the following media when you learn to sing arias – score, audio recording or teacher's classroom teaching. Please rate the importance of the following jingju singing dimensions and elements when you learn to sing arias—intonation, rhythm, loudness, pronunciation, ornament, breath and strength (劲头).

Nine students have participated in this survey; they are different from the ones presented in Section 3.1.1. Among them, five are professional undergraduate students or students already graduated from NACTA, four are amateurs from the jingju associations in two non-art universities in Beijing. We use a five-level Likert scale for each rating term. For example, the "use frequency of the score" in the first question and the "importance of intonation" in the second question can be rated from 1 to 5, where 1 means "never used" or "not important at all" and 5 means "most frequently used" or the "most important". Then, we take the average value of each term for five professional students and four amateurs respectively.

It is worth to mention that three more elements – ornament, breath and strength have been added to the survey. The consideration for this change is that the survey terms need to be adapted to the student's artistic background, and the jingju singing jargons should be easily accessible by them.

3.1.2 Results and discussion

In this section, we report the results of the analysis of teachers' correction occurrences and the students' survey. The correction occurrences are classified into five dimensions by using the method introduced in the Section 3.1.1. Then, we discuss the student' sur-

vey result and compare it with the teachers' correction occurrence classification result.

Correction occurrence analysis

Table 3.2: The statistics of the correction occurrence dimension classification. Inton.: intonation; Loud.: loudness, Pronun.: pronunciation.

	Inton.	Rhythm	Loud.	Pronun.	Tone quality
武家坡 WuJiaPo	8	0	1	6	6
太真外传 TaiZhen WaiZhuan	6	1	9	4	11
捉放曹 ZhuoFangCao	5	2	7	9	3
Sum.	19	3	17	19	20

We observe from the Table 3.2 that among the five dimensions, tone quality, pronunciation and intonation dimensions have the largest and almost equal occurrence number, loudness takes the second place, and rhythm problem was least mentioned. In other words, tone quality, pronunciation and intonation are the dimensions which receive particular attention from teachers and cause problems easily to students.

The correction occurrence analysis results are organized in an Excel spreadsheet, which consists of the teacher's verbal feedback, signal analysis method, and classified dimension.

The survey among students

We gather the survey results by ordering the mean values for each question. For the first question, the usage frequency ordered from high to low of three learning media are:

1. Professional group: classroom teaching, audio recording, score;

2. Amateur group: audio recording, teacher's classroom teaching, score.

The music score has been rated as the lowest use frequency by both professional and amateur groups, which means that the jingju students we investigated do not use the visual clue – music score reading, to learn arias. The teacher's classroom teaching has been rated as the highest use frequency for the professional and the second for the amateurs, which is reasonable because this learning medium is much easier available for the professional. Lastly, the high rating of both teacher's classroom teaching and audio recording shows that the jingju students use mostly the listening and imitation methods to learn arias.

For the second question, the importance order from the most important to most trivial are:

1. Professional group: rhythm, strength, pronunciation, breath, intonation, loudness, ornament;
2. Amateur group: rhythm, pronunciation, strength, ornament, breath, intonation, loudness.

Apart from the terms strength and breath, the others have been analyzed in the correction occurrence perspective. Strength is a stylistic and abstract word to depict the energy used in jingju singing and instrument playing. A jingju singing with strength is conveyed by combining multiple elements, such as loudness (mostly), rhythm, intonation and tone quality. Breath or specific methods of breathing (气口) described in Wichmann's book (Wichmann, 1991) is "these methods allow the exiting breath to control the pitch, timbre or tone color, and energy of the sound produced." In consequence, Strength and breath both are nonspecific terms combining or affecting multiple basic jingju singing elements.

Pronunciation is rated as an essential element by both the professional and amateurs, which is coherent with the result of the correction occurrence analysis. The high importance of rhythm and low importance of intonation and loudness contradict to the result of the correction occurrence analysis. For rhythm aspect, one possible explanation is that the higher importance the students value a

singing dimension, less prone they are going to sing poorly on it. For example, the students consider that rhythm is the most important singing aspect, they pay much attention to it during the practice. Thus they are less prone to have the rhythmic problems. For intonation and loudness, we cannot easily conclude that they are not important in the learning process. The reasons are twofold: on the one hand, the students might think that the intonation accuracy is a basic requirement in jingju singing and its importance is self-evident; on the other hand, because intonation and loudness are jargons used in acoustic, sound and music technology research fields, which might be foreign to these students, so they might avoid them and choose the familiar terms such as strength.

The only jingju singing dimension emphasized in both correction occurrence analysis, and the survey analysis is pronunciation, which shows that its crucial role in jingju singing training. As a consequence, to take advantage of limited time and effort, we will focus on tackling the research problems related to the assessment of singing pronunciation. In the following sections of this chapter, we present challenges, opportunities and research problems which are only related to the pronunciation dimension.

3.2 Challenges and opportunities

Significant challenges are existed to the automatic assessment of singing voice pronunciation in jingju music. We present and discuss challenges and opportunities from the perspectives of jingju singing characteristics and state of the art. These challenges will help us to formulate the research problems to be more comprehensive and akin to jingju music tradition. The opportunities, in turn, help us to pursue new MIR research directions.

3.2.1 Characteristics of jingju singing

We illustrate some signal characteristics of jingju singing voice that will be helpful to identify challenges for automatic assessment.

Figure 3.9: An example of a dan role-type singing phrase. Red vertical lines indicate the syllable onset time positions. Black arrows at the bottom specify the time positions of pause within the syllable. Red horizontal bracket shows a prominent singing vibrato.

Figure 3.9 shows an example of a *dan* role-type singing phrase in which the last syllable lasts approximately 20 seconds. This singing method – 拖腔 (pinyin: tuoqiang, literally translated as the prolonged melody), used more commonly in *dan* than in *laosheng* singing, extends the duration of last syllable of the melodic line or a dou (Section 2.1.3). It is a way of improving artistic expression, and the prolonged syllable can be used to carry various singing skills which include breath, intonational and dynamic control techniques, among others.

In jingju singing, the breath must be under purposeful control at all times (Wichmann, 1991). 偷气 (pinyin: touqi, stealing breath) is one of the primary methods to taking the breath in jingju singing. Performer inhales rapidly without exhaling beforehand. Touqi is performed when a sound is too long to be delivered in one breath and should be undetectable to the audience (Wichmann, 1991). However, this is not the only technique which can lead to pauses within a syllable. Another singing technique (*zu yin*, literally translated as block sound), provokes also pauses without occurring exhalation or inhalation. This kind of pause can be very short in duration and can be easily found in jingju singing syllables.

Vibrato (颤音 chanyin and 波浪音 bolangyin) is extremely important in jingju singing such that a single pitch is rarely prolonged without a vibrato. Compared to the Western opera, jingju singing vibrato is slower and wider regarding vibrato rate and extent (Yang, Tian, & Chew, 2015).

In jingju singing training, correctly pronouncing each written-character (syllable) is essential. The important role of pronunciation in jingju singing training has been discussed in Section 3.1. However, in the actual training scenario, the student is likely to commit the pronunciation errors regarding two types of syllable – jiantuanzi and special pronunciation, where their definitions have been introduced in Section 2.2.3 and Section 2.2.4. Jiantuanzi mispronunciation means that the student mispronounces a pointed sound syllable (jianzi) as the rounded sound (tuanzi). The mispronunciation of the special syllables means that the student mispronounces a special pronounced syllable as the standard pronunciation in Chinese Mandarin. Figure 3.10 and Figure 3.11 shows the Mel spectrograms of the mispronounced syllables. We can observe that the spectral difference between jianzi “siang” and its cor-

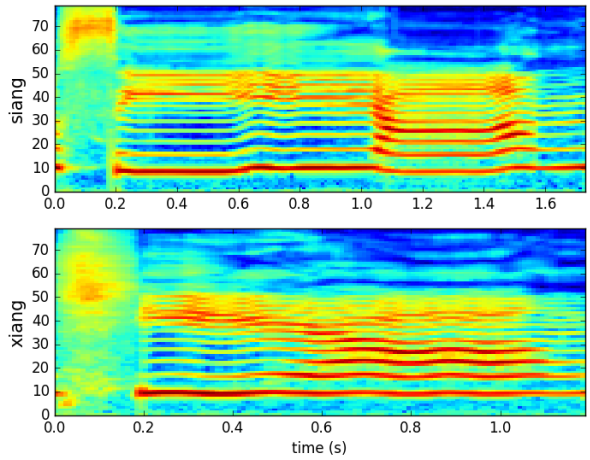


Figure 3.10: The Mel spectrograms of pointed syllable (jianzi) "siang" and its corresponding rounded syllable (tuanzi) "xiang".

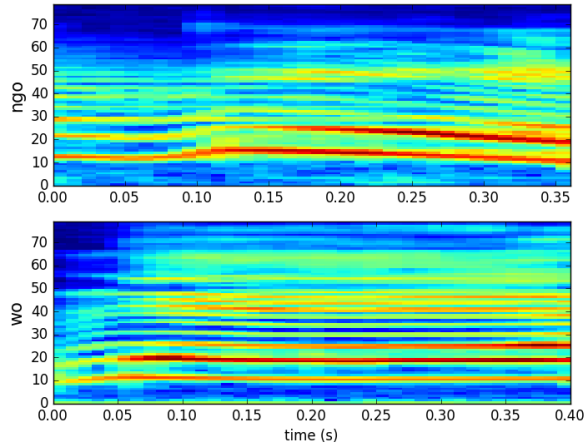


Figure 3.11: The Mel spectrograms of the special pronounced syllable "ngo" and its corresponding normal pronounced syllable "wo".

responding tuanzi "xiang" mainly lies in the non-voiced consonant part, and the difference between special pronounced syllable "ngo" and its corresponding normal pronunciation "wo" also lies in the syllable head part. The mispronunciation in jingju singing training and the formulation of the problem of mispronunciation detection will be continued to discuss in Section 3.3.3 and Section 3.4.3.

The overall quality of the singing syllable or phoneme can be easily illustrated by using spectrogram. Figure 3.12 shows the Mel

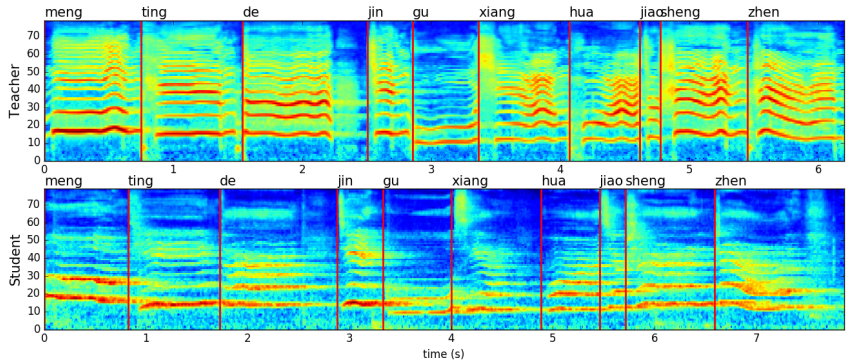


Figure 3.12: The Mel spectrograms of teacher and student singing the same phrase "meng ting de jin gu xiang hua jiao sheng zhen (in pinyin format)". Red vertical lines are the onset time positions of each syllable.

spectrograms of a dan role-type singing phrase taken from the aria 猛听得金鼓响画角声震–《霸王别姬》(meng ting de jin gu xiang hua jiao sheng zhen – Farewell My Concubine). The upper part of the figure is the spectrogram of the teacher's recording, while the lower part is that of a primary school student's recording. Although the student does not commit any mispronunciation, there still exists a significant gap between the overall quality of her singing and that of the teacher singing. The gap is reflected in many aspects if we compare the two spectrograms. For example, the higher harmonics of the student singing is much weaker than those of the teacher; The consonants energies of the student singing are weaker than those of the teacher if we compare the consonants of syllables "xiang" and "sheng"; The intonation of the student singing is flat and lacks variation.

3.2.2 Challenges

The basic music event of jingju singing is syllable. In jingju singing training, the accurate rendition of the syllabic pronunciation is placed in a more important position than that of the melody. In jingju circle, there is a saying 依字行腔 (pinyin: yi zi xing qiang, literally translated as singing according to the syllables), meaning that the singing melody should be consistent with the syllable tone and pronunciation, which also shows the importance of an accurate syllabic pronunciation. In Section 2.2.1, we presented the struc-

tures and the lower-level components of jingju singing syllable. A jingju singing syllable consists of four types of phoneme – an initial consonant or semivowel (optional), a medial vowel (optional), a central vowel and a tail (optional). As a consequence, at a more elaborate level, to pronounce a jingju syllable accurately is to render these elementary phonemes accurately.

According to the jingju singing principles mentioned above, to assess a jingju singing pronunciation at syllable or phoneme level, an automatic assessment system of jingju singing needs to have the ability to segment the singing recording automatically into the syllabic or phonetic unit. As we have mentioned in Chapter 2, a jingju aria is arranged hierarchically in several granularities from the roughest to the finest – banshi, couplet (shangxiaju), melodic line, syllable. Ideally, the segmentation of a jingju singing in a certain granularity needs to be performed on top of its parent one. For example, the segmentation of couplet needs to be done in its parent banshi segment; the segmentation of syllable needs to be done in its parent melodic line segment. Correspondingly, if the target recording for the assessment is an entire aria, which is required to be assessed in syllable or phoneme level, we need systems for different segmentation granularities – automatic banshi, couplet, melodic line, syllable and phoneme segmentation.

One way to approach the segmentation problem of different granularities is the alignment of aria lyrics to audio. Since lyrics can be annotated with boundaries of banshi, couplet and melodic line, once each syllable in the lyrics are time-aligned with the singing audio, the time boundaries of different granularities can be naturally deduced. However, this unified approach might not be optimal regarding the segmentation accuracy. Different banshi has the different singing characteristics. For example, prolonged singing syllables are more likely to be sung in unmetered banshi segments, such as *daoban* and *huilong*, and in slow tempo banshi, such as manban. As it was mentioned in Section 3.2.1, many singing skills such as ornamentation, breath control, are usually used in interpreting a prolonged syllable. Breath control leads to silences within a syllable; ornamentation leads the variation of the spectral pattern. Long syllable, silences within a syllable and spectral pattern variation are the main sources of lyrics-to-audio alignment error. Thus, to avoid the alignment error propagating in different banshi seg-

ments, it is necessary to perform the banshi segmentation.

Different tempi and meters characterise different banshi (Section 2.1.7). Thus banshi segmentation is analogous to meter tracking (Srinivasamurthy, 2016). Unmetered banshi is an important category of jingju banshi of which the singing and instrumental performing do not follow any rhythmic beat. Such unmetered banshi existing in jingju aria present challenge to the segmentation task.

Jingju music tradition does not have the absolute tempo. An expressive performance without a metronome, combined with a lack of annotated tempo can lead to a single composition being performed in different tempi. This lack of an absolute tempo complicates the choice of a relevant timescale for tracking banshi (Srinivasamurthy, 2016).

The jingju music characteristics allow a certain freedom of improvisation in changing the local tempo such as increasing or decreasing the tempo through the melodic line or a few syllables. However, MIR algorithm has difficulty tracking metrical structures that have expressive timing and varying tempo (Holzapfel, Davies, Zapata, Oliveira, & Gouyon, 2012). Thus, the local tempo variation is a potential source of challenge for banshi tracking.

Regarding segmentations in finer granularities than banshi, as we have mentioned above, long syllable, silences within a syllable and spectral pattern variation pose challenge to the relevant segmentation/alignment tasks.

Pronunciation correctness is essential in jing singing training. According to the discussion of mispronunciation in Section 3.2.1, the mispronunciation is revealed usually in some parts of a syllable. If the student mispronounces a jianzi, she/he probably only pronounces badly the non-voiced consonant part of the syllable. For example, the mispronunciation of jianzi “siang” to “xiang” is characterized only by the non-voiced consonant part. If the student mispronounces a special pronounced syllable, she/he might pronounce badly any part of the syllable. Please consult Table B.1 for the mispronunciation patterns regarding the special pronunciation. As a consequence, the model which can discriminate the mispronounced and correctly pronounced syllables should be able to locate the relevant parts in the syllable, which is a potential challenge.

Pronunciation and overall quality of a singing syllable or

phoneme are both abstract and perceptual related concepts. Pronunciation is a subconcept of the timbre which is defined by what is not "a set of auditory attributes of sound events in addition to pitch, loudness, duration, and spatial position". In a signal point of view, timbre is related to the spectral envelope shape and the time variation of spectral content (Pons, Slizovskaia, Gong, Gómez, & Serra, 2017). While overall quality is a more general concept than pronunciation since it is a mixture of different musical dimensions – intonation, loudness, duration and timbre (apart from pronunciation). In consequence, to define a pronunciation similarity measure requires a perceptual related representation of the time-varying spectral content, and to define an overall quality similarity measure requires a representation of all relevant dimensions. To identify the proper representations for similarity measures is a potential challenge.

In summary, the absence of an absolute tempo and local tempo variation are challenging. Long syllable, silences within a syllable and spectral pattern variation pose challenges to existing segmentation approaches. The locality of the mispronunciation in a syllable presents challenges in mispronunciation detection. The fuzziness of pronunciation and overall quality concepts present challenges in finding proper representations for their similarity measurement.

3.2.3 Opportunities

There are several unique features in jingju singing which bring new opportunities to explore new research directions in MIR. The challenges mentioned above also bring new opportunities to explore new approaches for automatic singing voice assessment. The complex metrical structure and syllable-based singing framework requires specific methodologies to perform segmentation and pronunciation description, and will be beneficial to the singing assessment of other music cultures based on similar frameworks.

In this dissertation, we mainly use audio for analysis. However, the corresponding score, lyrics and annotated metadata which also carry pronunciation and duration information can be used for a compound approach for building the singing assessment models.

Another important aspect of jingju singing is that its pronunciation is explicitly shown through the shapes for the throat and mouth

(Section 2.2.2). In jingju singing training, students are required to use a standardized throat and mouth shape to pronounce jingju syllables. It is believed a non-standard throat and mouth shape cannot lead to the correct pronunciation. Thus, a multi-modal approach to jingju singing pronunciation assessment can be done from video recordings of student singing practice, a problem is interesting, but beyond the scope of this dissertation.

The language system of jingju singing is a variant of the standard Mandarin Chinese. Although various Chinese dialects are used in jingju singing and bring certain variations to Mandarin pronunciation, such as special pronunciations – shangkouzi, the syllabic structure of Mandarin language remains unchanged (Section 2.2.1). We can learn methodologies from the mature research area of speech technologies to resolve segmentation and mispronunciation detection problems.

In summary, the unique metric structure, syllable-based singing framework and variants of Mandarin language bring new opportunities for exploring new methods in jingju singing. Additionally, a detailed description of jingju singing pronunciation involves combining various sources of information such as audio, score, lyrics, annotated information related to pronunciation and visual cues.

3.3 Research problems in the assessment of singing voice pronunciation of jingju music

We have identified so far several challenges and opportunities for automatic assessment of singing pronunciation in jingju music. With such context, we will describe relevant research problems, discuss possible methods, and review existing works for each problem. Some associated problems not directly tackled in this dissertation such as banshi segmentation, are also discussed for completeness. Many of the singing assessment problems for jingju singing have not been tackled before, whereas similar problems in speech or other music traditions have been aimed to resolve in speech technology or MIR fields. Most of the tasks for assessment of jingju singing pronunciation need to be reformulated with the

jingju singing background such as onset detector is regarded as useful to develop specific syllable/phoneme segmentation algorithm with the help of other sources of information.

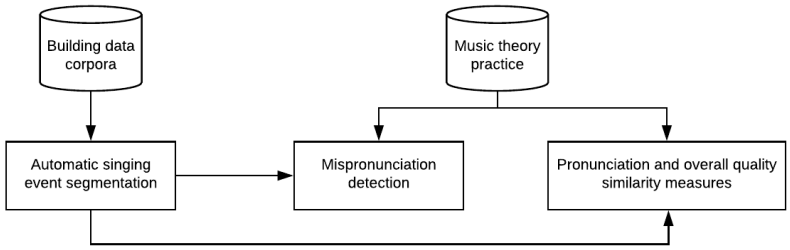


Figure 3.13: Related research topics of automatic assessment of singing voice pronunciation in jingju music.

In this dissertation, the assessment of singing pronunciation will be devised at syllable or phoneme level. There are several subproblems which lead to the final goal – to develop mispronunciation detection models and to define pronunciation similarity measure for jingju singing. In Figure 3.13, we show the information flow between four topics of research problem that will be addressed in this dissertation – building data corpora, automatic singing event segmentation, mispronunciation detection and pronunciation and overall quality similarity measures. There is a significant sequential order while addressing each problem, e.g. to achieve the assessment at syllable or phoneme level, mispronunciation detection and similarity measures benefit from the results of automatic singing event segmentation. The topics of mispronunciation detection and similarity measures use knowledge derived from music theory and practice, making them more culture-aware. Each of the topics will be discussed in detail.

3.3.1 Building data corpora

A crucial part of data-driven research using machine learning approaches requires good quality data. Data corpora of the music tradition under research are crucial for building and testing the automatic assessment models. The data should contain various sources

such as audio, score, lyrics and manual annotation made for automatic assessment research.

The dataset created in the work (Repetto & Serra, 2014) is formed by a collection of commercial recordings, as well as their metadata. Another jingju music corpus gathered in the (Tian & Sandler, 2016) also consists of commercial recordings, and annotated for structural segmentation analysis. These recordings are all mixed with instrumental accompaniment, which means a cappella (clean) singing voice should be separated during the preprocessing step if we want to make use of these recordings for the research of automatic assessment. A collection of 92 jingju music scores gathered for the analysis of jingju musical system is presented in the work (Repetto, Zhang, & Serra, 2017), which is transcribed from published books and stored in machine-readable format. The a cappella singing separated commercial recording dataset and the modified score dataset will be integrated into the data corpora of this dissertation.

The a cappella jingju singing dataset created in the work (D. A. A. Black, Li, & Tian, 2014) consists of 31 unique arias in total around 1-hour recordings. However, due to the small size of this dataset, and that its annotations were made for the task of mood recognition rather than automatic assessment, we have to re-annotate this dataset firstly, and then expand it to a proper scale. One of the main problems tackled in this dissertation is building suitable and scalable data corpora for the singing pronunciation assessment research, a problem that is discussed further in Section 3.4.1.

3.3.2 Automatic singing event segmentation

Automatic singing event segmentation includes a set of problems that aim to infer or time align the boundaries of several musical events in singing recordings related to pronunciation assessment. The common MIR tasks such as musical structure segmentation, lyrics-to-audio alignment and singing event onset detection can be classified as automatic singing event segmentation problems. As we have mentioned in Section 3.4.1, in the context of the assessment of jingju singing pronunciation, the relevant singing events to consider are banshi, couplet, melodic line, syllable and phoneme.

Automatic singing event segmentation is an important preliminary step to achieve automatic assessment, and there are several applications in which the segmentations are useful, such as rhythm-based annotation of audio, beat/melodic line/syllable aligned processing of music, audio summarization. Each of these problems will be described in detail.

Banshi segmentation

Banshi segmentation (or banshi tracking) is not the problem which will be tackled in this dissertation. However, we discuss it for completeness. Banshi segmentation refers to a set of problems that focus on segmenting different banshi sections in a jingju aria. By segmenting the banshi sections, a complete description of jingju metrical structure can be achieved. For such a problem, the subcomponents of a banshi – tempo, accented beat (pinyin: ban, downbeat), unaccented beat (pinyin: yan, beat) can be obtained.

Banshi segmentation can be done either in an uninformed or informed fashion. The former fashion means that inferring the time-varying tempo, beats and downbeats without any prior banshi knowledge of the aria. Informed banshi segmentation is the case to track tempo, beats and downbeats given the information of the banshi sequence of the aria. We can classify the subtasks as tempo tracking, beat tracking and downbeat tracking. Tempo tracking aims to estimate the time-varying tempo over the recording of a jingju aria. The tracked tempo will be useful for the beat tracking tasks. As we have mentioned in Section 3.4.1, the tempo tracking method applied for jingju aria needs to be robust for the local or long-term tempo change. For metered banshi, the rhythmic beats are performed by several percussion instruments – danpigu, ban, naobo, daluo and xiaoluo. Thus, the beat time instance is defined by the onset of each percussion instrument stroke. Although a specific beat tracking algorithm has not been developed for estimating metered jingju banshi, a suitable jingju percussion onset detection method (Tian, Srinivasamurthy, Sandler, & Serra, 2014) and several beat tracking methods (Böck, Böck, & Schedl, 2011; Krebs, Krebs, & Widmer, 2014; Böck, Krebs, & Widmer, 2016) for eurogeneric music can be adapted for this purpose. In jingju performance, each downbeat is usually marked by the ban (wooden

clapper) sound and indicates the first beat a measure (Wichmann, 1991). Thus downbeat tracking can be formulated as the problem of estimating the onset time positions of the ban sound in the beat sequence. Lastly, metered banshi tracking in jingju aria is a task analogous to tala tracking in Indian art music, of which the relevant methods have been studied extensively in Srinivasamurthy's work (Srinivasamurthy, 2016).

Due to the lack of tempo and beat, segmenting metered banshi requires a different framework mentioned above. Banshi segmentation is an important step towards any finer segmentation task of jingju aria. However, since we adopt the melodic line directly as the input of the assessment pipeline, banshi segmentation will not be a problem considered in this dissertation.

Couplet, melodic line, syllable and phoneme segmentation

Couplet or melodic line segmentation refers to estimate the time boundaries of singing couplet or melodic line in a banshi section. The syllable or phoneme segmentation aims to transcribe the audio recording of a melodic line into a time-aligned syllable or phoneme sequence. In jingju singing training scenario, the score, lyrics and relevant annotations such as starting and ending syllables of the couplet or melodic line are usually given beforehand. Thus, these problems can be formulated into a uniformed framework – lyrics-to-audio alignment. Time-aligning lyrics to audio is a fine-grained segmentation task, which can be applied to the syllable or phoneme level singing assessment and analysis.

Jingju is sung in Chinese Mandarin language with regional dialect pronunciations of which each written character is pronounced as a syllable (Section 2.2.1), and several written characters make up a word. Although not many languages in the world adopt the similar writing system, the pronunciation of all languages is built upon basic units – phoneme and syllable (Moran, McCloy, & Wright, 2014). Thus the lyrics-to-audio alignment method can be devised as either language-dependent or language-independent. Both methods can be formulated as supervised learning tasks. The former uses label data to build syllable or phoneme acoustic models; while the latter uses labelled data to build syllable or phoneme boundary models.

As discussed in Section 3.2.2, several jingju singing characteristics such as long syllable, silences within a syllable and spectral pattern variation pose challenge to the lyrics-to-audio alignment task. Apart from that, another challenge is that the mapping from the written characters to syllables is not unique, due to the existence of special pronunciations and multi-pronunciation characters in jingju singing.

The work on lyrics-to-audio alignment for jingju singing has been very limited so far. Dzhambazov et al. (G. Dzhambazov, Yang, Repetto, & Serra, 2016) proposed a modified text-to-speech alignment method for jingju singing. The system is built upon a duration-explicit hidden Markov model, where the phoneme duration is empirically set according to lyrics and metric structures of jingju music.

It is to be noted that prior musical information such as score is usually available for the assessment of jingju singing, and can be exploited to tackle the related challenges. Lastly, since we adopt the melodic line directly as the input of the assessment pipeline, couplet or melodic line segmentation will not be a problem considered in this dissertation. Syllable and phoneme segmentation is one of the problems addressed in this dissertation and is formulated more concretely in Section 3.4.2.

3.3.3 Mispronunciation detection

Mispronunciation detection refers to build the computational model to detect the badly pronounced syllables or phonemes in student's singing voice. The detection could be done in either syllable or phoneme granularities. In this dissertation, we tackle only the problem of building the models to detect the mispronunciation at syllable-level since syllable or written-character is the basic unit of which the teacher corrects the pronunciation in the actual singing training scenario. More specifically, we tackle only the problem of mispronunciation detection for jiantuanzi and special pronounced syllables since these two types of the syllable is the main source of the mispronunciation in jingju singing training. The application of such detection model is not limited to singing voice. Other potential applications are the mispronunciation detection in the second language (L2) learning or broadcasting training.

The challenge of this topic, as mentioned in Section 3.2.2, is to take consideration of the locality of the mispronunciation within a syllable, which is to say, the model should be able to detect the mispronunciation of a syllable according to some parts of the syllable. We can formulate the detection problem as a supervised discrimination task, and a labeled dataset which contains mispronunciation syllable segments (positive samples) and correctly pronounced syllable segments (negative samples) can be used to build the model.

There exist a significant amount of works on the topic of speech mispronunciation detection applied in L2 learning. The most relevant work in this field is the Goodness of Pronunciation (GOP) measure proposed by S. M. Witt and S. J. Young (Witt & Young, 2000), which used forced alignment method with a pronunciation dictionary to generate GOP score for the mispronunciation detection task of English phonemes. In the singing voice application, Gupta et al. (C. Gupta, Grunberg, et al., 2017) first generalized the mispronunciation rules for the singing voice of Southeast Asian English accent. Then they also applied forced alignment with an adapted dictionary for the mispronunciation detection.

Mispronunciation detection is one of the problems addressed in this dissertation. A more comprehensive problem formulation will be presented in Section 3.4.3.

3.3.4 Pronunciation and overall quality similarity measures

Pronunciation and overall quality similarity measures refer to build an objective model to calculate the similarity of corresponding jingju singing segments between teacher and student respecting pronunciation and overall quality aspects. The similarities can be measured in different singing granularities such as banshi section, couplet, melodic line, syllable and phoneme. In this dissertation, we tackle only the problem of building the models of phoneme-level pronunciation and overall quality similarity measures since phoneme is the finest grained pronunciation unit of jingju singing, and the composition basis of any higher singing granularities such as syllable, melodic line, couplet and banshi section. Likewise, phoneme is also the finest grained pronunciation unit of any other

languages. The method developed in building similarity measures at phoneme-level in jingju singing can be easily adapted to singing similarity measurement at phoneme-level in any other languages. The application of such similarity measure is not limited to the assessment of singing voice. Other potential applications are the assessment of pronunciation at phoneme-level in the second language (L2) learning and broadcasting training. In such scenarios, the similarity between the phoneme segments of a language learner and a native speaker needs to be shown to give the learner a clue about how well her/his pronunciation or overall quality is.

The challenge of this topic, as mentioned in Section 3.2.2, is to find proper representations for similarity measures – representation learning. Pronunciation is represented in the signal point of view as the time-varying spectral change. Overall quality is a perceptual concept mixed with different musical dimensions such as intonation, loudness, duration and timbre. The representations need to capture the time-varying and abstractive natures of these two concepts. The representation learning can be formulated as a supervised discriminative or a semi-supervised distance metric learning tasks. Take overall quality aspect as an example, a supervised discriminative learning uses labeled data (e.g. good/bad quality) to build a discriminative model, while a semi-supervised distance metric learning uses data labeled in pairwise or triple-wise similarity to build a model, e.g. the overall quality of samples A and B are similar; that of sample B and C are not similar. As a consequence, the learned representation from either the discriminative model or distance metric learning model is used for the similarity measurement.

We cannot identify any previous work on the topic of pronunciation or overall quality similarity measure at phoneme-level. However, there exist a significant amount of works on the topic of speech phonetic similarity applied in L2 learning. Minematsu et al. (Minematsu, 2004; Minematsu, Kamata, Asakawa, Makino, & Hirose, 2007; Shiozawa, Saito, & Minematsu, 2016) propose a structural representation of speech phoneme sounds. They train HMM for each phoneme class, then compute Bhattacharyya distance between each HMM pairs. The pairwise distance matrix represents the phoneme-level linguistic structure. They also claim that this representation represents purely the linguistic traits of a language

and free from any distortion such as microphone, room, speaker. Thomson (Thomson, 2008) develops a method to measure the English vowel similarity for Mandarin speaker. He builds discriminative models for each English vowel using the recordings of English native speakers, then uses the models to calculate the posterior probability as the similarity measure for vowel segment of Mandarin speakers. Wieling et al. (Wieling, Margaretha, & Nerbonne, 2011) focus on the multidimensional scaling (MDS) representation of vowel segments. They use formant frequencies as the feature to calculate the euclidean distance for each vowel pair, then perform MDS to project each vowel onto a two-dimensional similarity space. Mielke (Mielke, 2012) explores DTW distance for phonetic similarity measure. He uses MFCC as the representation of the phoneme segment, and compute DTW distance between two MFCC vectors. Kyriakopoulos et al. (Kyriakopoulos, Gales, & Knill, 2017) develop a phoneme similarity measure based on Jensen-Shannon divergence. They calculate aggregate PLP feature and fit multivariate Gaussian model for each phoneme. The Jensen-Shannon distance is computed on the multivariate Gaussian models of each phoneme pair.

Pronunciation and overall quality similarity measures are one of the problems addressed in this dissertation. A more comprehensive problem formulation will be presented in Section 3.4.4.

3.4 Formulation of thesis problems

With an overview of the research problems, challenges, review of the state of the art works, a subset of those problems that will be tackled in this dissertation are defined. In this section, we formulate these problems more comprehensively by discussing their assumptions, restrictions, and objectives in an engineering way.

3.4.1 Dataset for research

Building a dataset for MIR research is a scientific problem. Objective criteria are set up for designing, curating and also measuring the goodness of a corpus. One of the goals of CompMusic project is to build such data corpora and make it available for the research

usage. Collection of good quality data and easily accessible audio and metadata is crucial for the research reproducibility.

For developing relevant approaches, we focus on collecting and curating a cappella (clean) jingju singing voice audio in jingju singing training scenario. The jingju a cappella audio dataset includes both professional (teacher) singing audio and amateur (student) imitative singing audio, which accompanied with hierarchical jingju musical events annotations. For all of the tasks addressed in this dissertation, we need singing syllable and phoneme boundary annotations. Specifically, for mispronunciation detection task, we annotate special pronunciation singing syllables.

In general, for the research of automatic assessment of jingju singing voice, we aim to build a data collection which can represent the real world singing training scenarios. The recordings need to include the main role-types disciplined in singing, common teaching repertoire. The datasets built in the context of this dissertation are further presented in Chapter 4.

3.4.2 Syllable and phoneme segmentation

One of the problems addressed in this thesis is syllable and phoneme segmentation of jingju singing recordings. To the best of our knowledge, for jingju music, a system which can achieve a certain segmentation accuracy to be suitable for the needs of automatic assessment at syllable or phoneme level does not exist yet. Additionally, to improve the segmentation accuracy, we also explore incorporating a priori musical information such as the syllable or phoneme duration extracted from music scores or annotations into the segmentation algorithm.

To address the problem, we formulate tasks that can integrate a priori syllable or phoneme duration information – duration-informed syllable and phoneme segmentation. The a priori duration information is extracted either from the musical score or manual annotation, which thus represents the coarse syllable or phoneme duration in the target recording. We then use data-derived audio representation indicative of syllable or phoneme onset events in the recording. Finally, we build hidden Markov models that can incorporate the a priori duration information into the syllable or phoneme boundary selection step. The onset detection-based rather than the

forced alignment-based approach is used since the former is a binary classification task which requires less training data, and onset time stamps annotation is available for model training.

In the scope of this work, the target singing recording is assumed to have been already segmented into pieces that are at melodic line-level. This assumption mainly stems from the fact that in the actual jingju singing training course, the materials are taught and practised line by line. We do not assume any restrictions on banshi type over the melodic line. We restrict our work to two role-types – dan and laosheng in jingju music. The restriction is mainly because dan and laosheng are respectively two major role-types of female and male singing styles, and that singing is the main discipline of these two role-types. The proposed method is likely to extend to the singing of other role-types, provided we have the a priori duration information for them.

The a priori syllable durations of the target melodic line are stored in an array $M^s = \mu^1 \cdots \mu^n \cdots \mu^N$, where μ^n is the duration of the nth syllable. The a priori phoneme durations are stored in a nested array $M_p = M_p^1 \cdots M_p^n \cdots M_p^N$, where M_p^n is the sub-array with respect to the nth syllable and can be further expanded to $M_p^n = \mu_1^n \cdots \mu_k^n \cdots \mu_{K_n}^n$, where K_n is the number of phonemes contained in the nth syllable. The phoneme durations of the nth syllable sum to its syllable duration: $\mu^n = \sum_{k=1}^{K_n} \mu_k^n$. In both syllable and phoneme duration sequences – M^s, M_p, the duration of the silence is not treated separately and is merged with its previous syllable or phoneme. Let the recording of a melodic line can be reduced by short-term Fourier transform (STFT). The goal is to find the best onset state sequence $Q = q_1 q_2 \cdots q_{N-1}$ for a given syllable duration sequence M^s or phoneme duration sequence M_p and impose the corresponding syllable or phoneme label, where q_i denotes the onset of the $i+1$th or the offset of the ith inferred syllable/phoneme.

The approaches, experiments and results for syllable and phoneme segmentation are presented in Chapter 5.

3.4.3 Mispronunciation detection

The problem of mispronunciation detection at syllable-level is the third problem that will be addressed in this thesis. The goal is to

build supervised discriminative deep learning models to classify between mispronounced and correctly pronounced syllabic segments. We explore integrating the attention mechanism into the approach, and the learned model is supposed to concentrate on some certain parts of the syllable. Differ from the widely adopted forced alignment-based approach, the proposed method only requires that the training data has the binary annotation – mispronounced or correctly pronounced, rather than the detailed mispronunciation patterns.

As the preliminary step, the syllables are segmented automatically by using the approach presented in Section 3.4.2. Although this approach will cause some segmentation errors which might be propagated to the mispronunciation detection step, we use it to allow a fair comparison with the baseline forced alignment-based method, since the latter also segment the syllables automatically. We restrict in this dissertation the mispronunciation detection on two types of the syllable – jiantuanzi and special pronunciation since they are the main sources of mispronunciation happened in actual jingju singing training. Jiantuanzi mispronunciation means that the student mispronounces a pointed sound syllable (jianzi) as the rounded sound (tuanzi). The mispronunciation of the special syllables means that the student mispronounces a special pronounced syllable as the standard pronunciation in Chinese Mandarin. The mispronunciation patterns – from correctly pronounced syllable to mispronounced syllable, is shown in Appendix B. Thus, two different models will be explored. The first one classifies the mispronounced special syllables from the correctly pronounced special syllables, and the second one classifies the mispronounced jianzi from the correctly pronounced jianzi.

Let the set of variable-length mispronounced special syllable segments be denoted as $S_{positive}$, and the set of correctly pronounced special syllable segments to be denoted as $S_{negative}$. The discriminative model should be able to classify binarily the samples between these two classes. The similar model can be built for jiantuanzi mispronunciation detection as well.

The approaches, experiments and results for mispronunciation detection are presented in Chapter 6.

3.4.4 Pronunciation and overall quality similarity measures at phoneme level

The problem of pronunciation and overall quality similarity measures is the fourth problem that will be addressed in this thesis. The approach we explore is to learn phonetic pronunciation and overall quality representations using representation learning techniques and compute the similarity between two representations using distance measure. The goal is to test the effectiveness of representation learning techniques in learning pronunciation or overall quality discriminative representations. Differ from the previous similarity measure approaches which use handcrafted features as the representation, and perform DTW related methods to compute the similarity between two variable-length features, we present an approach in this dissertation based on acoustic phoneme embeddings which map the variable-length phoneme segments into fixed-length vectors to facilitate the similarity calculation.

We assume that the singing recordings have been segmented into phoneme units, which is done manually in this dissertation. Although the segmentation can be done automatically by using the approach presented in Section 3.4.2, this assumption can make sure the segmentation accuracy, and thus avoid the error propagated by the automatic segmentation step. We restrict in our work the overall quality to be a binary rubric such that phoneme segments sung by the teacher have a professional overall quality, and those sung by the student have an amateur overall quality. Thus, the overall quality similarity is measured between teacher and student phoneme segment pair which belong to the same phoneme class. For example, the approach can only measure the similarity between phoneme segments A and B, where A is sung by teacher and B is sung by student; A, B belong to the same phoneme class. From the point of view of the actual jingju singing training, only the case mentioned above is valid since the similarity between two phoneme segments sung consistently by teacher or student would not be measured, and measuring the similarity between two segments belonging to different phoneme classes is a problem which can be avoided by mispronunciation detection. Since we can be sure that the student recordings in our dataset can by no means reach the professional level, such a restriction is justified.

Let the set of variable-length phoneme segments be denoted as $\mathcal{A} = A_1, A_2, A_3, ..., A_N$, where each A_j is a subset of phoneme segments belonging to class j. The learned phonetic pronunciation representation for each phoneme segment in A_j should be capable of minimizing the intra-class similarity and maximizing the inter-class similarity. Let the set of variable-length phoneme segments sung by teacher be denoted as $\mathcal{B} = B_1, B_2, B_3, ..., B_N$, and another set of variable-length phoneme segments sung by student be denoted as $\mathcal{C} = C_1, C_2, C_3, ..., C_N$, where each B_j and C_j are the subsets of phoneme segments belonging to class j. The learned phonetic overall quality representation should be capable of minimizing the intra-class similarity between two segments from one single set such as B_j **or** C_j, and maximizing the inter-class similarity between those from different sets such as B_j **and** C_j.

The whole approach can be formulated as a fixed-length representation learning problem with pre-segmented phoneme samples – using the fixed-length representation for the similarity computation. The approaches, experiments and results for syllable and phoneme segmentation are presented in Chapter 7.

Chapter 4

Data corpora for research

Computational data-driven MIR research requires a well-curated data corpus for training and testing the models. The corpus should meet certain criteria so that the models can be built successfully and applicable to real-world scenarios. A research corpus is an evolving data collection which is representative of the research domain under study. A good corpus can be built by a single research institute or by crowdsourcing within a community. Regarding MIR research, a research corpus is a representative subset of one or several music genres, since it is nearly impossible to work with all the relevant music pieces. Computational models developed upon this subset can be assumed generalizable to real-world scenarios.

A test dataset is a subset of the research corpus which is designed for a specific research task. In the research task, the test dataset is used to develop and evaluate computational models. For better reproducibility of experiment results, the test dataset is usually fixed or properly versioned.

Building a research corpus is a research problem itself and has been studied in many fields. There are many repositories for the research of speech such as Linguistic Data Consortium[1], Librispeech[2], and for the research of musicology such as IM-

[1] https://www.ldc.upenn.edu/
[2] http://www.openslr.org/

SLP/Petrucci Music Library[3] and MusicBrainz[4]. There have been efforts to compile large collections for MIR or general sound analysis research such as Million Song Dataset (Bertin-Mahieux, Ellis, Whitman, & Lamere, 2011), FMA dataset (Defferrard, Benzi, Vandergheynst, & Bresson, 2017), AcousticBrainz[5] and Freesound datasets (Fonseca et al., 2017). These music data collections are good resources for developing MIR models on Western Pop music. A systematic way of building a research corpus is essential for the MIR research, and receives attention from the research community. Serra (Serra, 2014) described a set of criteria to build a MIR research corpus – Purpose, Coverage, Completeness, Quality and Reusability. We use these criteria to help develop a corpus for automatic assessment of jingju singing pronunciation.

In this chapter, we compile and analyse the research corpus and test datasets for the research of this dissertation. We will discuss the corpus building criteria and evaluation methodologies. Our main focus in this chapter will be jingju music, while other relevant test datasets are also presented. We aim:

1. To describe the corpus and the test datasets, emphasizing the research problems and tasks relevant to this thesis.

2. To describe a set of corpus design criteria and methodologies, then use them to evaluate the jingju a cappella singing voice corpus.

3. To present both corpus-level and test dataset-level musically meaning data analysis and visualization.

We mainly emphasize on presenting a scientific approach for corpus building and the evaluation of its coverage and completeness. Apart from the corpus description, the musically meaningful data analysis and visualization is another contribution of this chapter. Finally, the research corpus and test datasets presented in this chapter will be made available for further jingju MIR research.

[3] https://imslp.org/
[4] https://musicbrainz.org/
[5] https://acousticbrainz.org/

4.1 CompMusic research corpora

Although different music genres share some basic concepts such as melody and rhythm, some other important aspects can be described only by considering the musical specificity of that tradition. In the context of CompMusic project, Serra (Serra, 2011) highlighted the needs for culture-specific MIR research corpora to develop approaches which benefit from the essential aspects of the music tradition.

In CompMusic project, we work with five music traditions of the world which expose different research problems. A significant effort has been put towards the design the research corpora for the relevant problems of the specific musical traditions. In this chapter, we focus mainly on the a cappella singing voice of jingju music, while jingju commercial audio recording (Repetto & Serra, 2014), lyrics and musical score collections (Repetto et al., 2017) have been presented by other researchers of the CompMusic project. The Carnatic and Hindustani research corpora have been described thoroughly by (Srinivasamurthy, Holzapfel, & Serra, 2014). The Turkish makam music research corpus has been presented in detail by (Uyar, Atli, Şentürk, Bozkurt, & Serra, 2014).

4.1.1 Criteria for the creation of research corpora

Serra (Serra, 2014) listed five criteria for build culture-specific MIR research corpus:

"Purpose: The first step in the design of a corpus is to define the research problem that wants to be addressed and the research approach that will be used. In CompMusic, we want to develop methodologies with which to extract musically meaningful representations from audio music recordings, mainly related to melody, rhythm, timbre and pronunciation. The approaches are based on signal processing and machine learning techniques; thus the corpus has to be aligned with this purpose.

Coverage: A corpus has to include data representative of all the concepts to be studied and given our quantitative approach, there have to be enough samples of each instance for the data to be statistically significant. For our research we need to have audio

recordings, plus appropriate accompanying information, covering the varieties of pronunciation present in the musical culture.

Completeness: In each corpus, every audio recording is complemented by a set of data fields, and the idea of completeness relates to the percentage of fields filled, thus how complete the corpus is. For our corpora, this mainly refers to the completeness of the editorial metadata and the annotations accompanying each audio recording.

Quality: The data has to be of good quality. The audio has to be well recorded and the accompanying information has to be accurate. We have used well-produced recordings whenever possible, and the accompanying information has been obtained from reliable sources and validated by experts.

Reusability: The research results have to be reproducible, and that means that the corpus has to be available for the research community to use. In our case, we have emphasised the use of specific open repositories such as Zenodo.org[6] that are either already suitable or that can be adapted to our needs."

Central to the jingju a cappella singing corpus is the audio recordings with its annotation. We present this corpus in the next section.

4.1.2 Jingju a cappella singing corpus

The jingju a cappella singing corpus mainly contains audio recordings, editorial metadata and musical event annotations. All annotated corpus is the content used by signal processing and machine learning approaches.

Given that aria is the natural unit of jingju music, most audio recordings in this corpus are arias. A unique aria might be sung by different singers with different singing levels – professional or amateur. To facilitate the development of singing voice assessment models, the audio recordings in this corpus are all a cappella version, meaning without instrumental accompaniment. The singer's singing level is most important metadata associated with a recording.

[6] https://zenodo.org/

To build the corpus, we consulted jingju professors and musicologists. The main institutional reference is the National Academy of Chinese Theatre Arts (NACTA)[7], which is the premier institution dedicated to jingju performing training in Beijing, China, and is the only institute of its kind in China that offers both B.A. and M.A. degrees in jingju performing.

We wish to compile recordings sung by both professional and amateur singers from different backgrounds. The professional singers are the professors and students of NACTA. The amateur singers are from various sources – students of jingju associations in non-art universities, amateurs of jingju groups in community activity centers located in Beijing, amateurs of jingju associations located in London and students from several primary schools. We did not keep track the singer information of the amateurs of jingju groups in community activity centers located in Beijing and the students from several primary schools due to that a large number of singers participated in these recording sessions. Otherwise, the singer information of the other recordings is written in the editorial metadata.

The corpus has been collected in three different stages and thus been split into three parts. The audios in the first part are recorded with the joint effort of two music technology research institutes – Center for Digital Music, Queen Mary University of London (C4DM) (D. A. Black, Li, & Tian, 2014) and Music Technology Group, Universitat Pompeu Fabra (MTG-UPF). Additionally, another 15 clean singing recordings separated from the commercial releases have been included in this part. The audios in the second and third parts are recorded by the author of this thesis during his two times research stay in Beijing.

The corpus consists of 2 role-types (dan and laosheng), 121 unique arias with 289 recordings, meaning several arias have been recorded more than once by different singers. The total duration is 13.61 hours. Other information related to the corpus is described in Table 4.1.

[7] https://www.nacta.edu.cn/

Table 4.1: General statistics of the jingju a cappella singing corpus.

Role-type	#Unique aria	#Recording (A: amateur, P: professional)	#Singers	#Total duration (hours)	#Median recording duration (minutes)
dan	73	171 (A: 79, P: 83)	14+	8.26	1.87
laosheng	48	118 (A: 67, P: 51)	13+	5.35	2.01

The editorial metadata associated with each recording has been stored in MusicBrainz, as well as in Zenodo.org. The primary metadata is the name of the aria, the name of the play and the name of the singers. Each entity such as artist, recording, work, in MusicBrainz is assigned a unique MusicBrainz IDentifier (MBID), which helps organize the metadata. The editorial metadata has been entered using simplified Chinese characters and romanization system – pinyin.

A large part of the audio recordings has been annotated. The annotations consist of (i) melodic line onset and offset time boundaries and lyrics in simplified Chinese characters, (ii) syllable onset time stamps and pinyin label, (iii) phoneme onset and offset time boundaries and X-SAMPA label (Appendix A), (iv) labels indicating the melodic lines which contain long syllables and (v) labels indicating special pronunciations. All annotations have been done in Praat speech analysis and annotation tool (Boersma, 2001). Two Mandarin native speakers and one jingju musicologist have dedicated to the annotation. The annotation has been verified and corrected twice by the thesis author to ensure its the boundary accuracy and label correctness.

The whole corpus including audio recordings, editorial metadata and audio annotations are easily accessible from Zenodo.org[8,9,10].

Recording setup

In the first part of the corpus, the information of recording setup for the audios collected by C4DM has not been given in the original release (D. A. Black et al., 2014). However, by listening to each of them, we confirmed that the audios from this part included in the corpus are all good quality. Also in this part, the recordings whose names ending with 'upf' are recorded in a professional studio with jinghu accompaniment. Singers and jinghu accompanists are placed separately in different recording rooms and used two recording channels to avoid crosstalk. Other recordings whose name ending with 'lon' are recorded by using a Sony PCM-50

[8] https://doi.org/10.5281/zenodo.780559
[9] https://doi.org/10.5281/zenodo.842229
[10] https://doi.org/10.5281/zenodo.1244732

portable stereo recorder in a classroom with a certain reverberation. Additionally, another collection of 15 clean singings source-separated from commercial recordings contain audible artefacts of the background accompaniment.

For the second part of the corpus, most of the recording sessions have been conducted in professional recording rooms by using professional equipment. We use two recording equipment sets and two recording rooms:

- Set 1: M-Audio Luna condenser microphone + RME Fireface UCX audio interface + Apple GarageBand for Mac DAW;
- Set 2: Mojave MA-200 condenser microphone + ART voice channel microphone preamp + RME Fireface 800 audio interface + Adobe Audition 2.0 DAW;
- Room 1: The conference room in NACTA's business incubator with reflective walls, carpet-covered floor, conference furniture and medium room reverberation;
- Room 2: The sound recording studio in Institute of Automation, Chinese Academy of Science, with acoustic absorption and isolation.

Commercial audio recordings are used, or jingju players are invited for accompanying the singing. When commercial audio recordings were used as the accompaniment, singers were recorded while listening to the accompaniment sent through their monitoring headphone. Otherwise, when *jinghu* players were used as the accompaniment, to simultaneously record both singing and *jinghu* without crosstalk, we placed them separately in two different recording rooms and used two recording channels. However, they were still able to have visual communication through a window and monitor each other through headphones.

For the third part of the corpus, the recording sessions are done by using a Sony PCM-50 portable stereo recorder. The professional singings are recorded during the primary school jingju courses. The recording sessions of primary school students are done in three classrooms rather than asking the students to come to the studio. We believe that recording in the classrooms can represent the room

acoustic conditions of the actual jingju teaching. The three rooms are (i) a mid-reflected classroom with the hard wall, marble floor, wood tables, chairs and a blackboard; (ii) a high-reflected dancing rehearsal room with mirrors, hard wall and wood floor; (iii) a mid-reflected dancing rehearsal room with carpet floor, mirrors and glass windows. Lastly, the amateurs of jingju groups in community activity centers are recorded in (iv) a high-reflected community entertainment room with marble floor and hard wall.

Coverage

A research corpus needs to be representative of the real world in the concepts that are primary to the music culture (Srinivasamurthy, 2016). The main concepts of jingju music – role-type, shengqiang and banshi, are presented previously in Section 2.1. The concepts of jingju singing – syllable and phoneme are the essential units for the pronunciation assessment. In this work, the coverage analysis is presented for role-type, shengqiang, banshi and phoneme. We do not analyse syllable because there are excessive syllable classes in the languages of jingju singing.

The corpus includes the two jingju role-types whose main discipline is singing – *laosheng* and *dan*. Both professional and amateur singers have been recorded. For *dan* role-type, there are 79 amateur recordings and 83 professional recordings. For *laosheng* role-type, there are 67 amateur recordings and 51 professional recordings (see Table 4.1).

The corpus also includes the two main *shengqiang* - *xipi* and *erhuang*, and a few auxiliary ones, such as *fanxipi*, *fanerhuang*, *sipingdiao*, *nanbangzi*.

In terms of *banshi*, the whole range of metered ones is represented in the dataset - *yuanban*, *manban*, *kuaiban*, *erliu*, *liushui*, *sanyan* and its three variations – *kuaisanyan*, *zhongsanyan* and *mansanyan*. Besides these metered *banshi*, there are a few unmetered ones – *sanban*, *daoban*, *yaoban* and *huilong*, whose occurrence is very punctual in performance. A list of shengqiang and banshi included in the corpus for *dan* and *laosheng* role-types is shown in Table 4.2.

Figure 4.1 represents the number of occurrence for each phoneme for dan and laosheng role-types. We can see that the cor-

Table 4.2: A list of shengqiang and banshi included in the corpus.

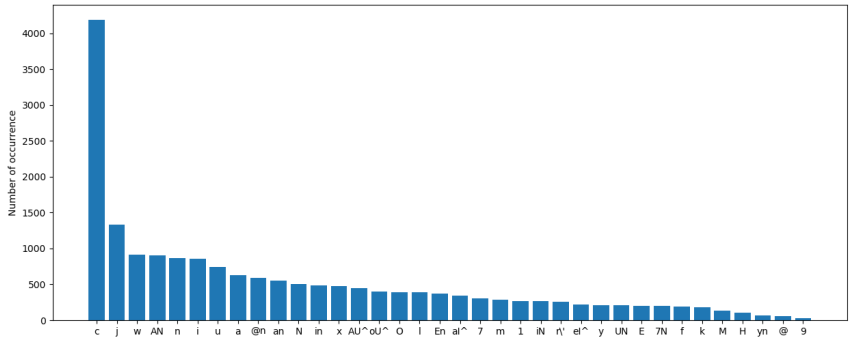
Role-type	shengqiang	banshi
dan	xipi, erhuang, fanxipi, fanerhuang, nanbangzi, sipingdiao, fansipingdiao, gaobozi, handiao	yuanban, manban, kuaiban, liushui, erliu, kuaisanyan, sanyan, pengban, shuban, duoban, daoban, huilong, yaoban, sanban
laosheng	xipi, erhuang, fanxipi, fanerhuang	yuanban, manban, kuaiban, liushui, erliu, kuaisanyan, zhongsanyan, sanyan, daoban, huilong, yaoban, sanban

pus cover all the phoneme classes for both dan and laosheng role-types, although some phonemes such as "@", "yn" have tiny number of occurrence. In all the phoneme classes, "c", a meta-phoneme class which is merged by all the non-voiced consonants, has the largest number of occurrence. An interesting observation is that the semivowel "j" has a large number of occurrence because this semivowel is used for both syllable initial and medial vowel. Another large number of occurrence phoneme "n" is also used for both syllable initial and terminal consonants. Phoneme "AN" is presented much more in dan singing than in laosheng singing, which indicates that it is preferable to use the syllables constituted with this nasal final in dan singing than in laosheng singing.

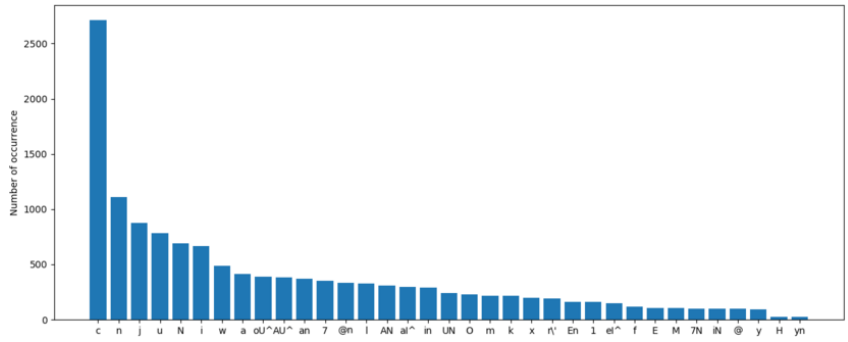
Completeness

In the context of this dissertation, completeness of the corpus refers to the completeness of the associated metadata and annotation for each recording. As the metadata and annotations are important for training and testing singing assessment machine learning models, they should as complete as possible.

The corpus contains the metadata of each recording – the artist's



(a) Dan role-type



(b) Laosheng role-type

Figure 4.1: Number of occurrence for each phoneme of dan and laosheng role-types.

Table 4.3: Metadata and annotations completeness. Table cell format: #annotated recordings/total recordings; percentage.

Role-type	Metadata	Melodic line	Syllable	Phoneme
dan	100%	118/171; 69%	110/171; 64.32%	92/171; 53.8%
laosheng	100%	89/118; 75.42%	89/118; 75.42%	52/118; 44.06%

singing level, the name of the aria, the name of the play and the name of the singing characters. The metadata is 100% annotated for all recording in the corpus.

The annotations are done in a hierarchical way at melodic line, syllable and phoneme levels. Due to time limits, not all record-

Table 4.4: The number of annotated melodic line, syllable and phoneme in the corpus.

Role-type	#Melodic line	#Syllable	#Phoneme
dan	1213	9847	18671
laosheng	893	8239	13378

ings have been annotated. The annotation completeness for each recording in different granularities is shown in Table 4.3. The number of annotated melodic line, syllable and phoneme are shown in Table 4.4.

An important concern in computational research is the reproducibility of the experiments, which requires a corpus openly accessible to the research community. All three parts of the corpus which includes audio, metadata and annotations are stored in Zenodo.org[11,12,13]. The metadata is also organized into releases in MusicBrainz[14].

Dataset analysis

In this section, we present a corpus-level statistic analysis towards the durations of the melodic line, syllable and phoneme. The goal is to draw musically meaningful insights from the analyses.

Table 4.5 shows a basic statistics of duration for melodic line, syllable and phoneme. Some interesting insights can be drawn from the table. Firstly, the mean and standard deviation of melodic line, syllable and phoneme durations of dan role-type are all larger than those of laosheng, which indicates that the length of dan singing regarding melodic line, syllable and phoneme are longer and more varying than that of laosheng. Secondly, the maximum duration of dan melodic line and syllable are more than two times longer than those of laosheng. However, the maximum duration of dan phoneme is shorter than that of laosheng, which indicates that dan role-type tends to prolong the singing syllables and to take more

[11] https://doi.org/10.5281/zenodo.780559

[12] https://doi.org/10.5281/zenodo.842229

[13] https://doi.org/10.5281/zenodo.1244732

[14] https://musicbrainz.org/search?query=jingju&type=release&method=indexed

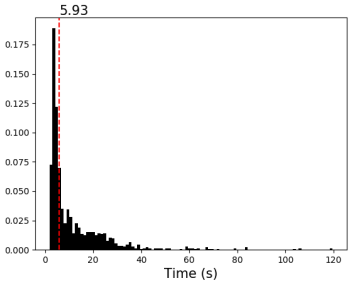
Table 4.5: Mean and standard deviation duration, minimum and maximum duration of melodic line, syllable and phoneme (second).

Role-type	Melodic line		Syllable		Phoneme	
	Mean	Min	Mean	Min	Mean	Min
	Std	Max	Std	Max	Std	Max
dan	11.93	1.81	1.35	0.07	0.42	0.0047
	13.85	119.69	2.66	52.66	0.75	11.08
laosheng	10.02	1.82	1.08	0.07	0.29	0.0025
	8.86	55.92	0.84	20.63	0.62	13.59

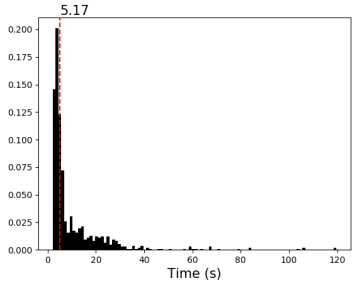
breaths such that a long syllable is split into several phoneme segments by short pauses. Lastly, compared with the mean < 250 ms and standard deviation < 50 ms of the duration of Mandarin speech syllable (J. Wang, 1994), those of jingju singing voice are at least four times longer and more varying. As we have mentioned in Chapter 3, many singing skills such as ornamentation, breath control, are usually used in interpreting a prolonged syllable. Breath control leads to silences within a syllable; ornamentation leads the variation of the spectral pattern. Long syllable, silences within a syllable and spectral pattern variation bring challenges in developing singing assessment methodologies.

Figure 4.2 shows the duration histograms of melodic line, syllable and phoneme for dan and laosheng role-types. The general shapes of the histogram distribution between dan and laosheng are similar. Although there are prominent peaks on all the histograms, the durations are varied, which can be observed by the extended long-tails on each histogram. For example, the median melodic line duration of dan is 5.93s. However, a significant amount of melodic lines are longer than 10s; the median phoneme duration of laosheng is 0.15s, whereas those phonemes whose durations are more prolonged than 0.4s are not the minority.

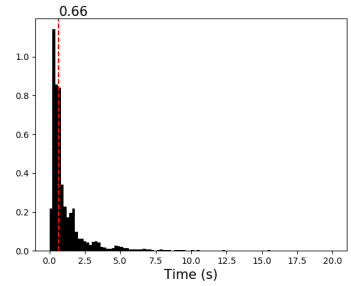
Figure 4.3 and Figure 4.4 show the histograms of durations for the individual phoneme. The phonemes which we are selected to show are "c, l, N, @n, i, a" in X-SAMPA format. 'c' is an ensemble phoneme class of non-voiced consonants; 'l' is a voiced phoneme; 'N' is a terminal consonant; '@n' is a diphthong used as the central vowel in a syllable; 'i' and 'a' are single vowels also used as



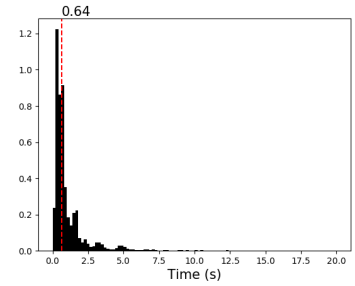
(a) Dan melodic line



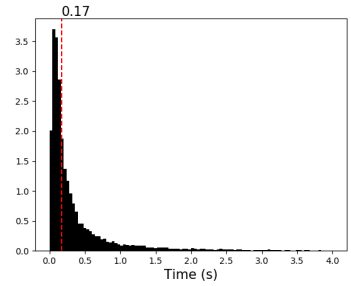
(b) Laosheng melodic line



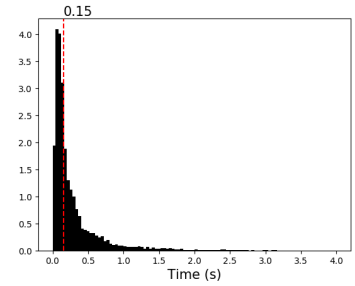
(c) Dan syllable



(d) Laosheng syllable



(e) Dan phoneme



(f) Laosheng phoneme

Figure 4.2: Dan and laosheng melodic line, syllable and phoneme duration histograms normalized to unit density. Vertical red dash lines indicate the median duration.

the central vowel in a syllable. Again, the general shapes of the histogram distribution between dan and laosheng are similar. The median duration of the individual phoneme of dan is longer than that of laosheng except for syllable initial consonants – 'c' and 'l'. The duration of initial consonants does not vary much. However,

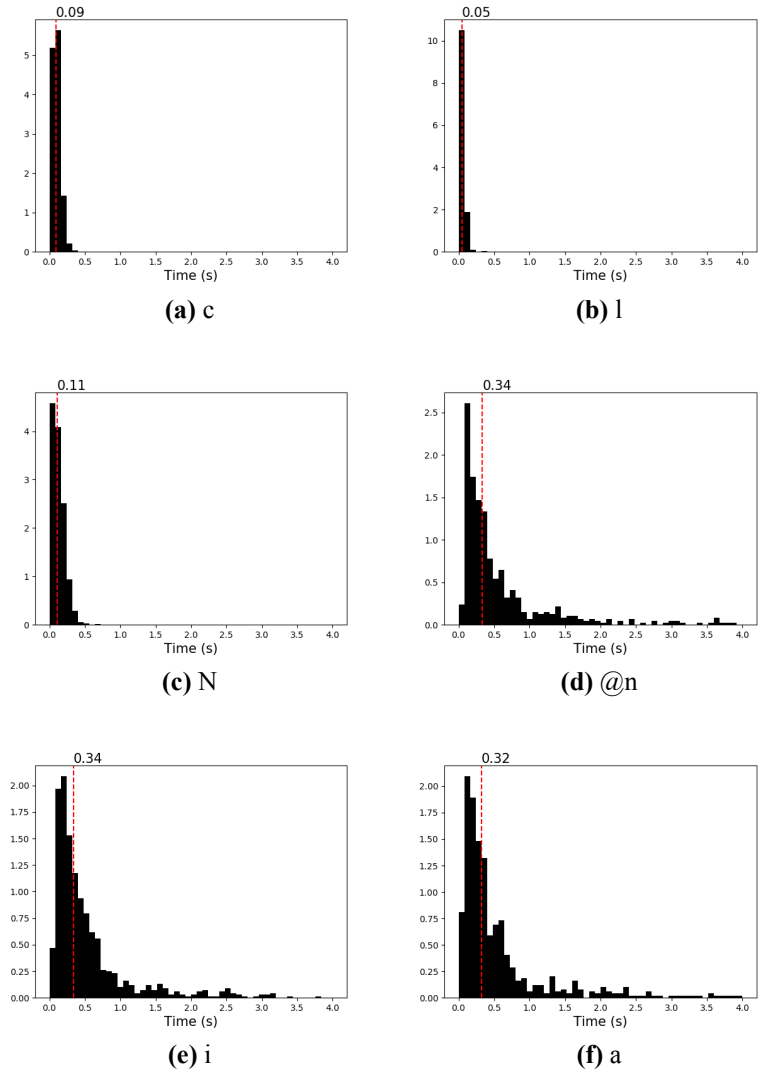
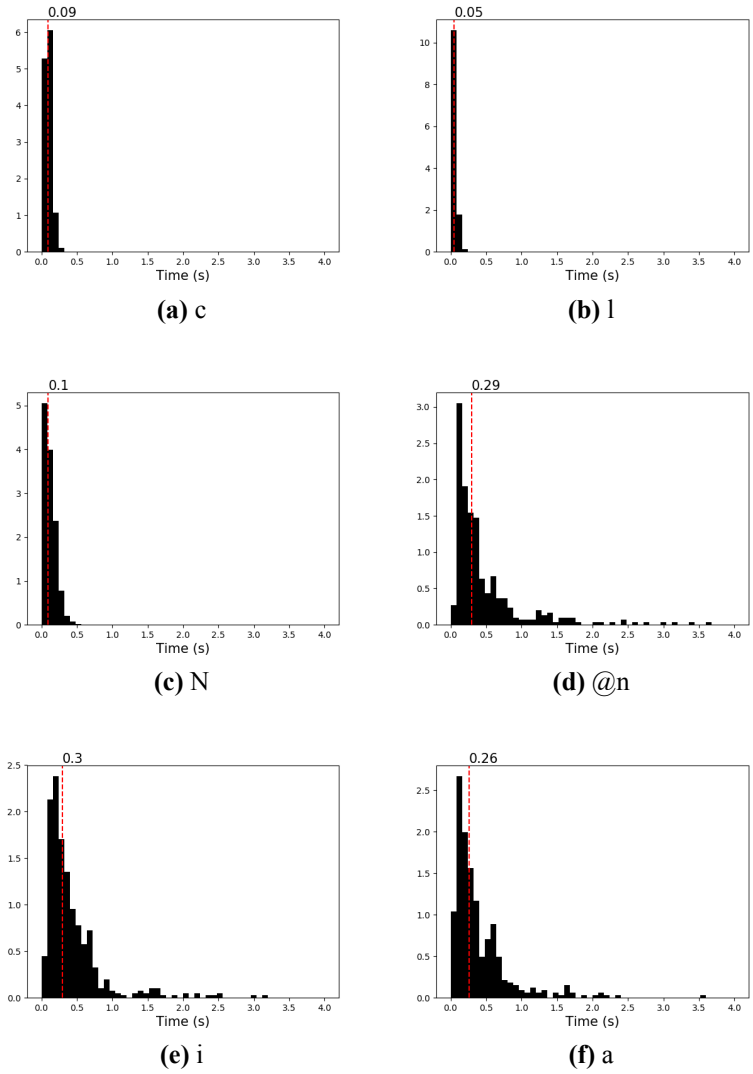


Figure 4.3: Dan histograms normalized to the unit density of durations for phonemes “c, l, N, @n, i, a”. Vertical red dash lines are the median phoneme durations.

the central vowels show a large duration variation since which are the primary part of a syllable sung by a singer in a prolonged way.



(a) c **(b)** l **(c)** N **(d)** @n **(e)** i **(f)** a

Figure 4.4: Laosheng histograms normalized to the unit density of durations for phonemes "c, l, N, @n, i, a". Vertical red dash lines are the median phoneme durations.

4.2 Test datasets

The test datasets are designed for special research tasks. There are several test datasets built in CompMusic for different musical tra-

ditions[15]. We describe in this section only the test datasets for the tasks of automatic assessment of jingju singing pronunciation.

4.2.1 Dataset for automatic syllable and phoneme segmentation

Automatic syllable and phoneme segmentation task require the data having syllable and phoneme time boundary and label annotations. Thus, we select the recordings with associated annotations in the corpus to form the test datasets. Two datasets are prepared – $ASPS_1$ and $ASPS_2$. $ASPS_1$ will be used for setting the baseline syllable and phoneme segmentation model, while $ASPS_2$ will be applied for searching an efficient state of the art syllable segmentation model.

$ASPS_1$ is a subset of the jingju a cappella singing corpus. The recordings in this dataset are selected from all three parts of the corpus. $ASPS_1$ contains two jingju role-types: *dan* and *laosheng*.

Table 4.6: Statistics of the $ASPS_1$ test dataset.

	#Recordings	#Melodic line	#Syllables	#Phonemes
Train	56	214	1965	5017
Test	39	216	1758	4651

The dataset contains 95 recordings split into train and test sets (table 4.6). The recordings in the test set only include student imitative singing. The corresponding teacher's demonstrative recordings can be found in the train set, which guarantees that the coarse syllable/phoneme duration and labels are available as a priori information being used in model testing. Recordings are pre-segmented into melodic line units. The syllable/phoneme ground truth boundaries (onsets/offsets) and phoneme labels are manually annotated. 29 phoneme categories are annotated, which include a silence category and a non-identifiable phoneme category, e.g. throat-clearing sound. The category table can be found in the Github page[16]. The dataset is publicly available[17].

[15] http://compmusic.upf.edu/datasets
[16] https://github.com/ronggong/interspeech2018_submission01
[17] https://doi.org/10.5281/zenodo.1185123

Table 4.7: Statistics of the $ASPS_2$ test dataset.

	#Recordings	#Melodic line	#Syllables
Train	85	883	8368
Test	15	133	1203

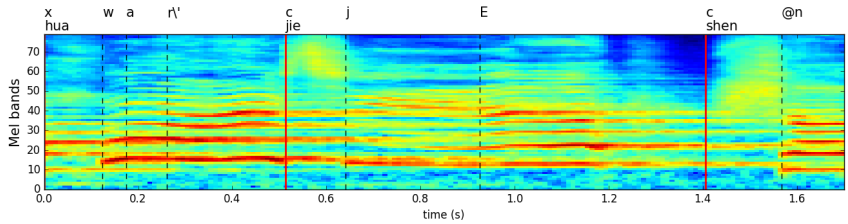
$ASPS_2$ test dataset is also a subset of the jingju a cappella singing corpus. It also includes recordings of dan and laosheng role-types. $ASPS_2$ contains 100 recordings manually annotated for each syllable onset. The syllable segmentation evaluation will be conducted on each melodic line which has been pre-segmented manually. The statistics and train-test sets split are shown in table 4.7. It is worth to mention that the artists, recording rooms and recording equipment used for the test set is completely different from the training set. This train-test split setup avoids the artist/room/equipment filtering effects which might be happening in the evaluation process (Flexer & Schnitzer, 2010). The musical score is also included in this dataset, which provides the syllable duration prior information for the evaluation. This dataset is openly available[18].

As it has been mentioned in Section 3.2.2, pronunciation is a subconcept of the timbre, and in a signal point of view, timbre is related to the spectral envelope shape and the time variation of spectral content. In the following of this section, we show several spectrogram examples of various phoneme categories – syllable initial non-voiced consonant, voiced consonant, medial vowel, central vowel and syllable terminal consonant, and the transition between two phonemes such as from syllable initial non-voice consonant to media vowel and from central vowel to syllable terminal consonant.

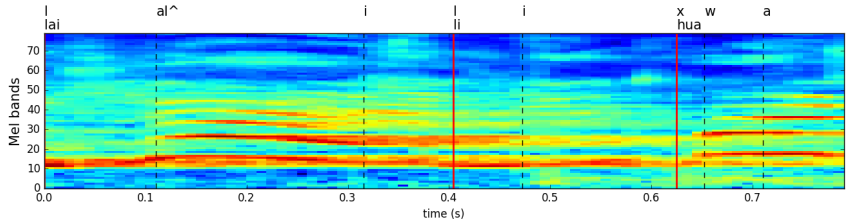
Figure 4.5 show three Mel spectrograms of singing syllable sequence, each of which consists of three syllables. We focus the analysis of the middle syllable of each sequence such that the transition between the first and second syllables and the transition between the second and third syllables can be visualized easily.

The pinyin of middle syllable in Figure 4.5a is “jie”, which consists of three phonemes – non-voiced consonant “c”, medial vowel

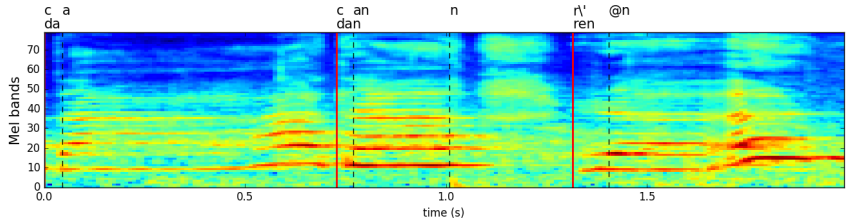
[18] https://doi.org/10.5281/zenodo.1341070



(a) Mel spectrogram of syllables “hua, jie, shen”.



(b) Mel spectrogram of syllables “lai, li, hua”.



(c) Mel spectrogram of syllables “da, dan, ren”.

Figure 4.5: Three examples of syllablic Mel spectrogram. Vertical red solid lines are syllable onsets; vertical black dash lines are phoneme onsets. On top of each subfigure, the first line is the phoneme transcription in X-SAMPA format, the second line is the syllable transcription in pinyin format.

“j” and central vowel “E”. The spectrogram pattern of the non-voiced consonant “c” can be distinguished easily from “j” and “E” by the high-frequency noise-like content since the consonant “c” is an affricate. The spectrogram pattern of the central vowel “E” contains more regular harmonic pattern than the medial vowel “j”. However, the difference between the patterns of “j” and “E” is not that obvious to discriminate.

The pinyin of the middle syllable in Figure 4.5b is “li”, which consists of two phonemes – syllable initial voiced consonant “l” and the central vowel “i”. The difference of spectrogram pattern

between these two phonemes can be hardly distinguished.

The pinyin of the middle syllable in Figure 4.5c is "dan", which consists of three phonemes – syllable initial non-voiced consonant "c", central vowel "an" and syllable terminal consonant "n". The consonant "c" is a short non-voiced stop which doesn't contain harmonics. Thus, it can be distinguished easily from the central vowel "an". The syllable terminal consonant "n" doesn't contain any higher harmonics. Therefore, it can be distinguished easily as well from the central vowel "an".

Given the analysis of the above three spectrogram examples, we can design the methodologies for the syllable and phoneme segmentation task in an intuitive way. Firstly, as most of the phoneme categories can be distinguished between each other except for medial vowel-central vowel and voiced consonant-central vowel, we can develop the segmentation algorithm based on the discrimination between phonemes. Secondly, as there are usually obvious spectrogram pattern transitions between phoneme segments, we can also devise the algorithm based on the detection of these transitions. The segmentation algorithms development and evaluation will be presented in detail in the next Chapter.

4.2.2 Dataset for mispronunciation detection

As we have mentioned in Section 3.4.3, we consider only two types of mispronunciation in jingju singing – the mispronunciation of special pronunciation and that of jianzi. The first type of mispronunciation – special pronunciation, is that some written characters in jingju singing pieces should be pronounced differently than in Mandarin Chinese, however, the student doesn't pronounce them correctly as in teacher's demonstrative singing. The second type of mispronunciation – jianzi, is that certain rounded syllables (团子, pinyin: tuanzi) in jingju singing pieces can be altered to pronounce as the pointed sounds (尖子, pinyin: jianzi), however, the student doesn't pay attention and still pronounce them as rounded syllables.

In the actual jingju teaching scenario, the teacher's demonstrative singing pieces are given, thus we can identify in advance those special pronounced and jianzi written characters in the pieces. After we obtain the student's imitative singing pieces, the detection

process can be carried out only on those special pronounced and jianzi written characters. To this end, we need a model which either can transcribe orthographically each singing syllables considering the special pronunciations and jianzi, or can distinguish between the standard Mandarin pronunciations and the special pronunciations/jianzi. Either way, we need a test dataset where the special pronounced syllables and jianzi are annotated orthographically in pinyin and in phoneme using X-SAMPA format.

Table 4.8: Statistics of the MD test dataset. Syl.: syllable; Phn.: phoneme.

	#Melodic line	#Syl.	#Special pronunciation	#jianzi	#Phn.
Train	662	5797	463	41	15287
Test	345	3106	356	13	7561

The Mispronunciation Detection – MD test dataset is annotated for the above purpose. MD is a subset of the jingju a cappella singing corpus. The recordings in this dataset are selected mainly from the part 1 and 2 of the corpus. Table 4.8 shows the statistics of the MD test dataset which are split into train and test parts, and the test part contains only the recordings of amateur singings. As we can see from the table, the occurrence of the special pronounced syllables is much larger than jianzi. Most importantly, in the test part of the MD dataset, according to the teacher's demonstrative recordings, there are in total 451 syllables of special pronunciation and 50 syllables of jianzi should be pronounced correctly by the students. However, in the actual recordings of the test part of the MD dataset, there are 102 syllables of special pronunciation and 37 syllables of jianzi which have been mispronounced, and 349 syllables of special pronunciation and 13 syllables of jianzi which have been pronounced correctly. These mispronounced syllables are labeled manually by comparing the annotation of the test recording and that of the corresponding teacher's demonstrative recording. For example, if in the teacher's recording, there is a special pronunciation /ngo/ for the syllable "wo", however, in the amateur's recording, the corresponding syllable is still pronounced as /wo/, this syllable is labeled as a mispronunciation.

Figure 4.6 shows the occurrence of each special pronounced syllables in MD dataset. The most frequently occurred syllables are /ngo, shen, qin, tin, bin, yin, chen, go, min, lin, xin, ho/. The pronunciation /ngo/ is altered from /wo/ in standard Mandarin by changing the semivowel /w/ to the nasal consonant /ng/. The syllables /go/ and /ho/ are altered from /ge/ and /he/ in standard Mandarin by changing the vowel /e/ to /o/. The other syllables mentioned above are the alteration from velar nasal to alveolar nasal, for example, changing from /eng/ and /ing/ to /en/ and /in/. The full table of the alteration from Mandarin pronunciation to the special pronunciation appeared in the MD test dataset is presented in Table B.1.

Figure 4.7 shows the occurrence of each jianzi in MD dataset. As we have mentioned in Section 2.2.3, the rule of this pronunciation alteration is /j, zh/ $\rightarrow$ /z/, /q, ch/ $\rightarrow$ /c/, /x, sh/ $\rightarrow$ /s/. For example, the pronunciation /siang/ is altered from /xiang/, and /zeng/ is altered from /zheng/. The full table of rounded syllables to the special pronunciation appeared in the MD test dataset is presented in Table B.2.

4.2.3 Dataset for pronunciation and overall quality similarity measures

The test dataset for the task of pronunciation and overall quality similarity measures (POQSM) needs to have phoneme-level onset and offset time boundary and label annotation. These recordings are also a subset of the jingju a cappella singing corpus. The phoneme segments of the dataset are randomly split into the train, validation and test sets, except that we deliberately use the recordings of the amateurs of jingju groups in community activity centers (see Section 4.1.2 for the information of the recording artists of the corpus) as the amateur part of the test set. The purpose of using this special amateur part of the test set is to avoid artist and room filtering effects by checking if the trained assessment model overfits on certain singers or the acoustic room conditions of the train and validation sets.

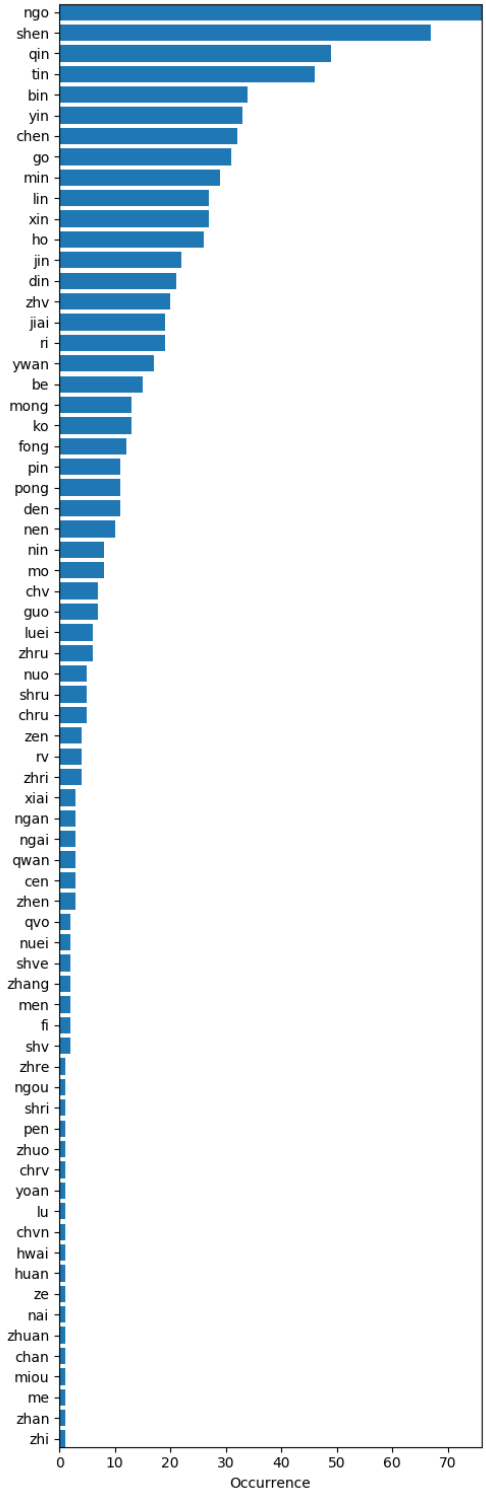


Figure 4.6: The occurrence of each special pronounced syllable.

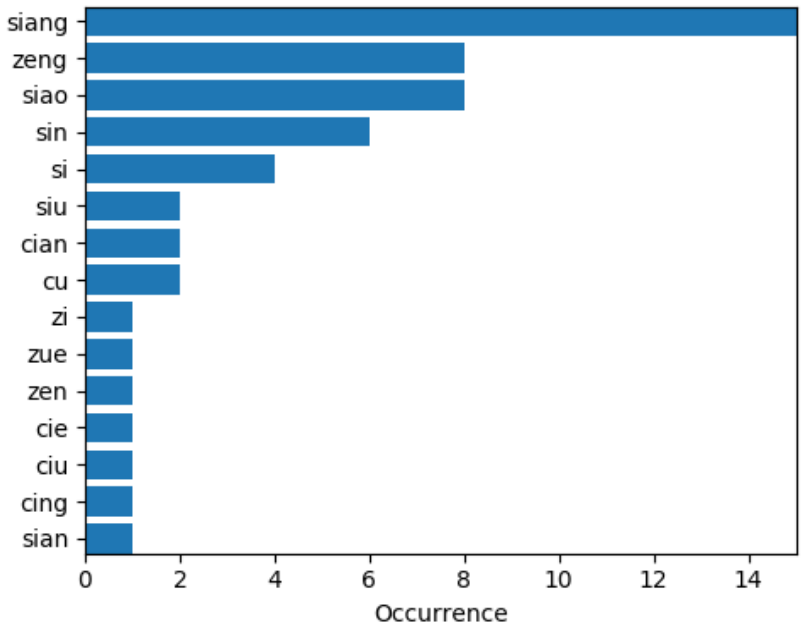


Figure 4.7: The occurrence of each jianzi syllable.

Table 4.9: POQSM test dataset split, numbers of the professional and amateur singing phonemes and the source of the professional and amateur singers.

	#Professional phonemes	#Amateur phonemes	Professional singers	Amateur singers
Train	6888	5673	Adult conservatory students or graduates	Mainly primary school students
Validation	1733	1429		
Test	2167	2021		Adult amateur singers

We consider the fact that, after dataset splitting, the train, validation and test sets would contain both professional and amateur phoneme segments. Additionally, the amateur part of the train and validation sets mainly include the phoneme segments of the primary school students, while that of the test set contains exclusively the segments of the adult singers recorded in a different room (see the room iv in section 4.1.2). This special split of the test set would verify the artist or room filtering effect of the assessment model (Flexer & Schnitzer, 2010). Please check table 4.9 for the phoneme numbers and the singers in each split. For the detailed information on the phoneme numbers per phoneme class and recording file names used for each split, please consult this link[19]. The test dataset can be download in this link[19].

An example of spectrogram visualization between teacher and student singing melodic line has been presented already in Section 3.2.1. In such an example, although the student does not commit any mispronunciation, there still exists a significant pronunciation quality and an overall quality gap between her singing and the teacher singing. Building a pronunciation and overall quality similarity model can help detect the relevant singing problems automatically apart from mispronunciation.

[19] https://doi.org/10.5281/zenodo.1287251

Chapter 5

Automatic syllable and phoneme segmentation

Automatic syllable and phoneme segmentation of singing voice is an important MIR task. It provides preliminary syllable or phoneme time boundary and label information to achieve a fine-grained singing voice assessment.

Syllable and phoneme segmentation aim to time-align a piece of singing voice audio recording with syllable or phoneme sequence. It tags the recording with the time-aligned syllable or phoneme boundary timestamps and labels. Within the context of jingju music, syllable and phoneme segmentation aim to time-align a recording with a syllable sequence in pinyin format or a phoneme sequence in X-SAMPA format.

This chapter aims to address the automatic syllable and phoneme segmentation task within the context of jingju music, presenting several methods and an evaluation of these methods. The main aims of this chapter are:

1. To address automatic syllable and phoneme segmentation task for jingju music. The problem is formulated in two ways – duration-informed lyrics-to-audio alignment and duration-informed syllable or phoneme onset detection. Several approaches are proposed to address the problem.

2. To present a detailed description of hidden semi-Markov model-

based (HSMM) segmentation method and the proposed onset detection-based method for syllable and phoneme segmentation.

3. To present an evaluation of HSMM-based alignment method and the proposed onset detection-based method and explore various deep learning architectures to improve the onset detection-based method.

5.1 Task description

We describe the automatic syllable and phoneme segmentation task addressed in this dissertation. We will also describe how the set of approaches described in this chapter can be adapted to this task, making the task of syllable and phoneme segmentation flexible to the available audio recordings and the related annotations. The task description presented in this section is continued building on the problem formulation presented in Section 3.4.2.

Given the singing audio recording pre-segmented into pieces of melodic line level, and the prior coarse syllable or phoneme duration information extracted from the musical score or the annotation of teacher's recording, the most relevant syllable and phoneme segmentation tasks for jingju music are duration-informed lyrics-to-audio alignment or duration-informed syllable or phoneme onset detection. In the context of jingju music, lyrics-to-audio alignment aims to time-align the a priori phoneme sequence in X-SAMPA format with the melodic line singing audio piece. The coarse phoneme duration information can be incorporated into the alignment system by using an HSMM-based model, which sets up the baseline segmentation system stemmed from various HMM-based text-to-speech alignment and lyrics-to-audio alignment methods presented in Section 2.4.3 and Section 2.4.4. Syllable and phoneme onset detection aim to find the onset timestamps for the syllables and phonemes in a melodic line singing audio piece. The a priori syllable and phoneme duration information can be used as a post-processing step in the detection algorithm to help select the correct onsets. In the context of this dissertation, because the a priori duration information is always accompanied with syllable or phoneme label, the post-processing onset selection method using a priori du-

ration information is equal to time-aligning the syllable or phoneme sequence with the melodic line singing audio piece.

The two main tasks of this chapter are setting up the HSMM-based baseline segmentation method and proposing the onset detection-based segmentation method. As the third task of this chapter, we explore various deep learning architectures for the syllable onset detection and try to identify and explain the most efficient architecture. The performance of all the three tasks will be evaluated on $ASPS_1$ and $ASPS_2$ test datasets. The results and the pros and cons of two segmentation methods and various deep learning architectures will be discussed in detail.

5.2 Prerequisite processing

In this section, we present two prerequisite processings that will be used in the segmentation approaches – logarithmic Mel input representation and a priori coarse duration model. The former converts the singing voice audio waveform to a perceptual representation - Mel spectrogram, which is then used as the input representation of both HSMM-based and onset detection-based segmentation methods. The latter utilizes the coarse syllable or phoneme durations extracted from the annotation of teacher's recording to build the duration model, as the teacher's recording and its annotation is always prior information for an assessment system. The phoneme duration model is then integrated into the HSMM-based segmentation method as the state occupancy distribution, and the syllable and phoneme duration models are both used in the onset detection-based segmentation method to help select the correct syllable and phoneme onsets.

5.2.1 Logarithmic Mel input representation

We use Madmom (Böck, Korzeniowski, Schlüter, Krebs, & Widmer, 2016) Python package to calculate the log-mel spectrogram of the singing voice audio. The frame size and hop size of the spectrogram are respectively 46.4ms (2048 samples) and 10ms (441 samples). The low and high frequency bounds of the log-mel calculation are 27.5Hz and 16kHz. We use log-mel input features with a

overlapped context window of 15 frames and 80 bins as the input to the networks. The classification acoustic model used in HSMM-based segmentation task takes a categorical phoneme label for every context window. While the onset detection model takes a binary onset/non-onset decision sequentially for every context window. This audio pre-processing configuration is almost the same as in Schlüter and Böck's work (Schluter & Bock, 2014) except that 3 input channels with respectively frame sizes 23ms, 46ms and 93ms have been used in their work, whereas only 1 channel with frame size 46.4ms input is used in this research.

5.2.2 Coarse duration and *a priori* duration model

The syllable durations of the teacher's singing phrase are stored in an array $M^s = \mu^1 \cdots \mu^n \cdots \mu^N$, where μ^n is the duration of the nth syllable. The phoneme durations are stored in a nested array $M_p = M_p^1 \cdots M_p^n \cdots M_p^N$, where M_p^n is the sub-array with respect to the nth syllable and can be further expanded to $M_p^n = \mu_1^n \cdots \mu_k^n \cdots \mu_{K_n}^n$, where K_n is the number of phonemes contained in the nth syllable. The phoneme durations of the nth syllable sum to its syllable duration: $\mu^n = \sum_{k=1}^{K_n} \mu_k^n$ (figure 5.1). In both syllable and phoneme duration sequences – M^s, M_p, the duration of the silence is not treated separately and is merged with its previous syllable or phoneme.

The *a priori* duration model is shaped with a Gaussian function $\mathcal{N}(d; \mu_n, \sigma_n^2)$. It provides the prior likelihood of an onset to occur according to the syllable/phoneme duration of the teacher's singing. The mean μ_n of the Gaussian represents the expected duration of nth teacher's syllable/phoneme. Its standard deviation σ_n is proportional to μ_n: $\sigma_n = \gamma\mu_n$ and γ is heuristically set to 0.35 for the onset detection-based method. Figure 5.1 provides an intuitive example of how the *a priori* duration model works. The a priori phoneme duration model will be used as the state occupancy distribution in the HSMM-based segmentation method, and the a priori syllable and phoneme duration models will be incorporated into a duration-informed HMM as the state transition probabilities to inform that where syllable/phoneme onsets is likely to occur in

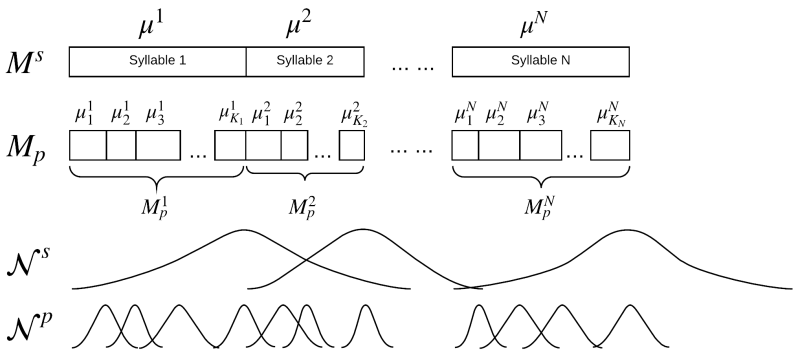


Figure 5.1: Illustration of the syllable M^s and phoneme M_p coarse duration sequences and their *a priori* duration models – $\mathcal{N}^s$, $\mathcal{N}^p$. The blank rectangulars in M_p represent the phonemes.

student's singing phrase.

5.3 HSMM-based segmentation method

As a baseline, we develop an lyrics-to-audio alignment system which also makes use of the prior phoneme duration information. This lyrics-to-audio alignment system is a 1-state monophone DNN/HSMM model. We use monophone model because our small dataset doesn't have enough phoneme instances for exploring the context-dependent triphones model, also Brognaux and Drugman (Brognaux & Drugman, 2016) and Pakoci et al. (Pakoci et al., 2016) argued that context-dependent model can not bring significant alignment improvement. It is convenient to apply 1-state model because each phoneme can be represented by a semi-Markovian state carrying a state occupancy time distribution. The audio preprocessing step is presented in Section 5.2.1.

5.3.1 Discriminative acoustic model

We use a CNN with softmax outputs as the discriminative acoustic model. According to the work of Renals et al. (Renals, Morgan, Bourlard, Cohen, & Franco, 1994), a neural network with softmax

outputs trained for framewise phoneme classification outputs the posterior probability $p(q|x)$ (q: state, x: observation), which can be approximated as the acoustic model at the frame-level if we assume equal phoneme class priors. In Pons et el.'s work (Pons, Slizovskaia, et al., 2017), a one-layer CNN with multi-filter shapes has been designed. It has been experimentally proved that this architecture can successfully learn timbral characteristics and outperformed some deeper CNN architectures in the phoneme classification task for a small jingju singing dataset. The convoluational layer of the architecture has 128 filters of sizes 50×1 and 70×1, 64 filters of sizes 50×5 and 70×5, and 32 filters of sizes 50×10 and 70×10. These filters are large in the frequency axis and narrow in temporal axis, which are designed to capture timbral relevant time-frequency spectrogram patterns. A max-pool layer of $2\times N'$ follows before the 32-way softmax output layer with 30% dropout, where N' is the temporal dimension of the feature map. Max-pooling of $2\times N'$ was chosen to achieve time-invariant representations while keeping the frequency resolution. The detailed model architecture is shown in Table 5.1.

Table 5.1: One-layer CNN architecture of the acoustic model. N' is the temporal dimension of the feature map.

Layer1: Conv 128x 50×1, 64x 50×5, 32x 50×5
128x 70×1, 64x 70×5, 32x 70×10
Layer2: Max-pooling $2\times N'$
Layer3: Dropout 0.3
Output layer: 29-way softmax

Model training

We use this one-layer CNN acoustic model for the baseline method. The log-mel context window representation presented in Section 5.2.1 is used as the model input. The target labels of the training set are prepared according to the ground truth annotations. We set the label of a spectrogram context window to its categorical phoneme class. The model predicts the phoneme class posterior probability for each log-mel spectrogram context window.

The model parameters are learned with mini-batch training (batch size 256), adam (Kingma & Ba, 2014) update rule and early stopping – if validation loss is not decreasing after 15 epochs. ELUs activation functions and weight decay regularization are used in the first convolutional layer.

5.3.2 Coarse duration and state occupancy distribution

The HSMM-based segmentation method receives the phoneme durations of teacher's singing phrase as the prior input. The phoneme durations are stored in a collapsed version of the M^p array (section 5.2.2): $M_c^p = \mu_1^{s1}\mu_2^{s1}\cdots\mu_{N_{s1}}^{s1}\cdots\mu_1^{sN}\mu_2^{sN}\cdots\mu_{N_{sN}}^{sN}$. The silences are treated separately and have their independent durations.

The state occupancy is the time duration of the phoneme state of the student's singing. It is expected in the best case to be the same duration as that of the teacher's singing. However, in the actual scenario, the phoneme duration of the student's singing always deviates from that of the teacher's singing in varying degrees. We build the state occupancy distribution as a Gaussian, which has the same form $\mathcal{N}(d;\mu_n,\sigma_n^2)$ as in section 5.2.2, where μ_n indicates in this context the nth phoneme duration of the teacher's singing. We set γ empirically to 0.2 as we found this value works well in our preliminary experiment.

We construct an HSMM for phoneme segment inference. The topology is a left-to-right semi-Markov chain, where the states represent sequentially the phonemes of the teacher's singing phrase. As we are dealing with the forced alignment, we constraint that the inference can only be started by the leftmost state and terminated to the rightmost state. The self-transition probabilities are set to 0 because the state occupancy depends on the predefined distribution. Other transitions – from current states to subsequent states are set to 1. We use a one-layer CNN with multi-filter shapes as the acoustic model (Pons, Slizovskaia, et al., 2017) and the Gaussian $\mathcal{N}(d;\mu_n,\sigma_n^2)$ introduced in section 5.2.2 as the state occupancy distribution. The inference goal is to find best state sequence, and we use Guédon's HSMM Viterbi algorithm (Guédon, 2007) for this purpose. The baseline details and code can be found in the Github

page[16]. Finally, the segments are labeled by the alignment path, and the phoneme onsets are taken on the state transition time positions.

5.3.3 Experimental setup

We use $ASPS_1$ test dataset presented in Section 4.2.1 and two metrics to evaluate the algorithm performance – onset detection accuracy and percentage of correct segments, where we also consider the phoneme label correctness in calculating onset detection accuracy. These two metrics have been presented in Section 2.4.6. We trained the CNN acoustic model 5 times with different random seeds, and report the mean and the standard deviation score on the testing part of the dataset.

5.3.4 Results and discussions

We only show the F1-measure of the results of the HSMM-based method in Table 5.2. The full results including precision and recall can be found in the Github page[16]. The performance of the HSMM-based method is mediocre in the sense that none of the onset detection accuracy and percentage of correct segments reaches an ideal level. The low onset detection accuracy – 44.5% for phoneme detection, 41% for syllable detection, means that the HSMM-based method cannot maintain more than half of the detected onsets within the 50ms tolerance window, which is crucial for the onset detection or segmentation of the consonants since they usually have a short duration. The low percentage of correct segments – 53.4% for phoneme and 65.8% for syllable, means that many phoneme boundaries including vowel boundaries are not detected correctly. As a consequence, the segmentation error will propagate to the automatic assessment step and reduce the assessment accuracy.

Table 5.2: Evaluation results table. Table cell: mean score±standard deviation score.

Onset F1-measure %		Segmentation %	
phoneme	syllable	phoneme	syllable
44.5±0.9	41.0±1.0	53.4±0.9	65.8±0.7

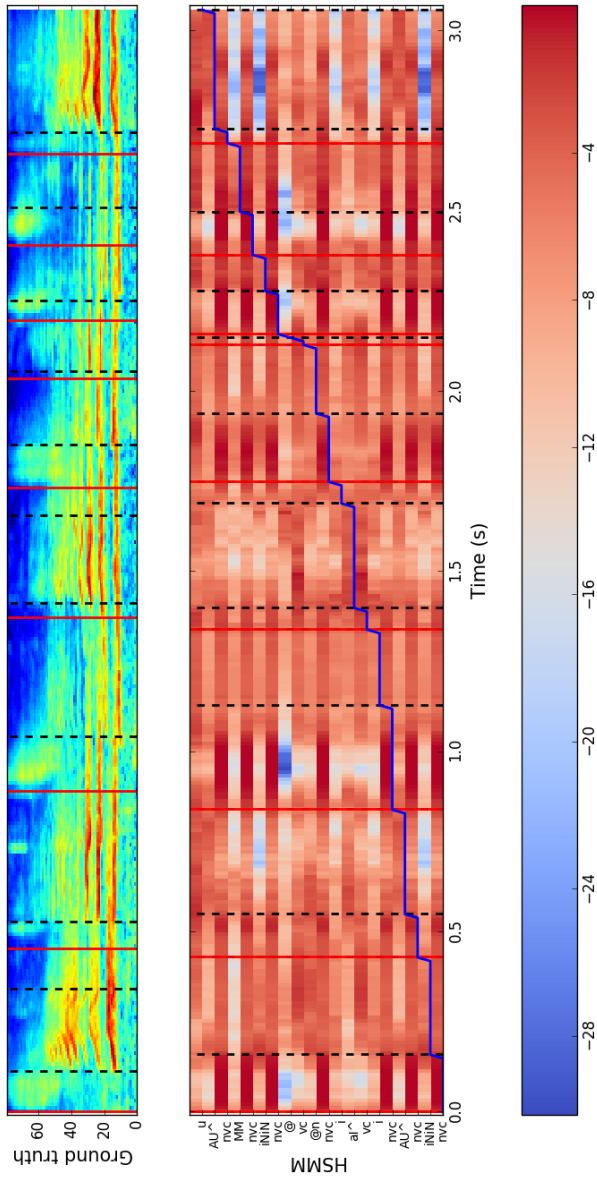


Figure 5.2: An illustration of the result for a singing phrase in the testing part of the dataset. The red solid and black dash vertical lines are respectively the syllable and phoneme onset positions. 1st row: ground truth, 2nd row: HSMM-based segmentation method. The staircase-shaped curve in the 2nd row is the alignment path.

Figure 5.2 shows a result example for a singing phrase in the testing part of the dataset. Notice that there are some extra or missing onsets in the detection. This is due to the inconsistency between the coarse duration input and the ground truth, for example, students might sing extra or miss some phonemes in the actual singing. We can also observe the deviations between the detected and ground truth phoneme onsets. Some of the deviations are quite large, for example, the first detected syllable onset after 2 seconds in the 2nd row, which is an indication that the HSMM-based segmentation method cannot meet the need of having a precise segmentation, and it has to be improved or replaced by a better method.

The unsatisfactory performance of the HSMM-based segmentation method might be due to the lack of a large training dataset. The DNN acoustic model usually requires a certain amount of training dataset such that it can effectively learn the temporal-spectral patterns of each phoneme class.

5.4 Onset detection-based segmentation method

The unsatisfactory performance of the HSMM-based segmentation method motivates us to search for a more accurate segmentation method. As we mentioned in Section 5.3.4, the lack of enough training dataset might be the cause of the unsatisfactory performance. In this section, we devise a coarse duration-informed syllable and phoneme segmentation method based on syllable and phoneme onset detection. As the onset detection is generally a binary detection problem – to classify the spectrogram of each frame into onset or non-onset class, it can greatly reduce the amount of the required training dataset. The coarse syllable and phoneme durations extracted from the annotation of teacher's recording can be used in the algorithm to boost the segmentation performance.

In the proposed onset detection-based segmentation method, the syllable and phoneme onset detection functions (ODFs) are jointly learned by a hard parameter sharing multi-task CNN model. The syllable/phoneme boundaries and labels are then inferred by an HMM using the *a priori* duration model as the transition probabil-

ities and the ODFs as the emission probabilities.

5.4.1 CNN onset detection function

We build a CNN for classifying each log-mel context and output the syllable and phoneme ODFs. We extend the CNN architecture presented in Schlüter's work (Schluter & Bock, 2014) by using two predicting objectives – syllable and phoneme (figure 5.3). The two objectives share the same parameters, and both are using the sigmoid activation function. Binary cross-entropy is used as the loss function. The loss weighting coefficients for the two objectives are set to equal since no significant effect has been found in the preliminary experiment.

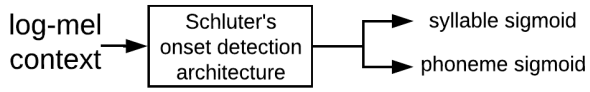


Figure 5.3: Diagram of the multi-task CNN model.

Model training

The target labels of the training set are prepared according to the ground truth annotations. We set the label of a certain context to 1 if an onset has been annotated for its corresponding frame, otherwise 0. To compensate the human annotation inaccuracy and to augment the positive sample size, we also set the labels of the two neighbor contexts to 1. However, the importance of the neighbor contexts should not be equal to their center context, thus we compensate this by setting the sample weights of the neighbor contexts to 0.25. A similar sample weighting strategy has been presented in Schluter's paper (Schluter & Bock, 2014). Finally, for each log-mel context, we have its syllable and phoneme labels. They will be used as the training targets in the CNN model to predict the onset presence.

The model parameters are learned with mini-batch training (batch size 256), adam (Kingma & Ba, 2014) update rule and early stopping – if validation loss is not decreasing after 15 epochs. The

ODFs output from the CNN model is used as the emission probabilities for the syllable/phoneme boundary inference.

5.4.2 Phoneme boundaries and labels inference

The inference algorithm receives the syllable and phoneme durations and labels of teacher's singing phrase as the prior input and infers the syllable and phoneme boundaries and labels for the student's singing phrase.

We present an HMM configuration which makes use of the coarse duration and label input (section 5.2.2) and can be applied to inferring firstly (i) the syllable boundaries and labels on the ODF for the whole singing phrase, then (ii) the phoneme boundaries and labels on the ODF segment constrained by the inferred syllable boundaries. To use the same inference formulation, we unify the notations N, K_n (both introduced in section 5.2.2) to N, and M^s, M_p^n to M. The unification of the notations has a practical meaning because we use the same algorithm for both syllable and phoneme inference. The HMM is characterized by the following:

1. The hidden state space is a set of T candidate onset positions $S_1, S_2, \cdots, S_T$ discretized by the hop size, where S_T is the offset position of the last syllable or the last phoneme within a syllable.
2. The state transition probability at the time instant t associated with state changes is defined by *a priori* duration distribution $\mathcal{N}(d_{ij}; \mu_t, \sigma_t^2)$, where d_{ij} is the time distance between states S_i and S_j ($j > i$). The length of the inferred state sequence is equal to N.
3. The emission probability for the state S_j is represented by its value in the ODF, which is denoted as p_j.

The goal is to find the best onset state sequence $Q = q_1 q_2 \cdots q_{N-1}$ for a given duration sequence M and impose the corresponding segment label, where q_i denotes the onset of the $i+1$th inferred syllable/phoneme. The onset of the current segment is assigned as the offset of the previous segment. q_0 and q_N are fixed as S_1 and S_T as we expect that the onset of the first syllable(or

phoneme) is located at the beginning of the singing phrase(or syllable) and the offset of the last syllable(or phoneme) is located at the end of the phrase(or syllable). One can fulfill this assumption by truncating the silences at both ends of the incoming audio. The best onset sequence can be inferred by the logarithmic form of Viterbi algorithm (Rabiner, 1989):

Algorithm 1 Logarithmic form of Viterbi algorithm using the *a priori* duration model

$\delta_n(i) \leftarrow \max_{q_1,q_2,\cdots,q_n} \log P[q_1 q_2 \cdots q_n,\ \mu_1 \mu_2 \cdots \mu_n]$

procedure LogFormViterbi(M, p)

Initialization:

$\delta_1(i) \leftarrow \log(\mathcal{N}(d_{1i}; \mu_1, \sigma_1^2)) + \log(p_i)$

$\psi_1(i) \leftarrow S_1$

Recursion:

$tmp\_var(i,j) \leftarrow \delta_{n-1}(i) + \log(\mathcal{N}(d_{ij}; \mu_n, \sigma_n^2))$

$\delta_n(j) \leftarrow \max_{1 \leqslant i < j} tmp\_var(i,j) + \log(p_j)$

$\psi_n(j) \leftarrow \arg\max_{1 \leqslant i < j} tmp\_var(i,j)$

Termination:

$q_N \leftarrow \arg\max_{1 \leqslant i < T} \delta_{N-1}(i) + \log(\mathcal{N}(d_{iT}; \mu_N, \sigma_N^2))$

Finally, the state sequence Q is obtained by the backtracking step. The implementation of the algorithm can be found in the Github link[16].

5.4.3 Experimental setup

We use ASPS_1 test dataset presented in Section 4.2.1 and two metrics to evaluate the algorithm performance – onset detection accuracy and percentage of correct segments, where we also consider the phoneme label correctness in calculating onset detection accuracy. These two metrics have been presented in Section 2.4.6. We trained the onset detection neural network model 5 times with different random seeds, and report the mean and the standard deviation score on the testing part of the dataset.

5.4.4 Results and discussions

We only show the F1-measure of the results of both HSMM-based and onset detection-based methods in Table 5.3. The full results including precision and recall can be found in the Github page[16].

Table 5.3: Evaluation results of HSMM-based and onset detection-based methods. Table cell: mean score±standard deviation score.

Methods	Onset F1-measure %		Segmentation %	
	phoneme	syllable	phoneme	syllable
HSMM-based	44.5±0.9	41.0±1.0	53.4±0.9	65.8±0.7
Onset detection-based	75.2±0.6	75.8±0.4	60.7±0.4	84.6±0.3

On both metrics – onset detection and segmentation, the proposed method outperforms the baseline. The proposed method uses the ODF which provides the time "anchors" for the onset detection. Besides, the ODF calculation is a binary classification task. Thus the training data for both positive and negative class is more than abundant. Whereas, the phonetic classification is a harder task because many singing interpretations of different phonemes have the similar temporal-spectral patterns. Our relatively small training dataset might be not sufficient to train a proper discriminative acoustic model with 29 phoneme categories. We believe that these reasons lead to a better onset detection and segmentation performance of the proposed method.

Fig 5.4 shows an result example for a singing phrase in the testing part of the dataset. Notice that there are some extra or missing onsets in the detection. This is due to the inconsistency between the coarse duration input and the ground truth, for example, students might sing extra or miss some phonemes in the actual singing. Also notice that in the 3rd row, the two detected phoneme onsets within the last syllable are not in the peak positions of the ODF. This is due to that the onsets is inferred by taking into account both ODF and the *a priori* duration model, and the latter partially constraints the detected onsets.

The biggest advantage of the proposed method is the language-independency, which means that the pre-trained CNN model can

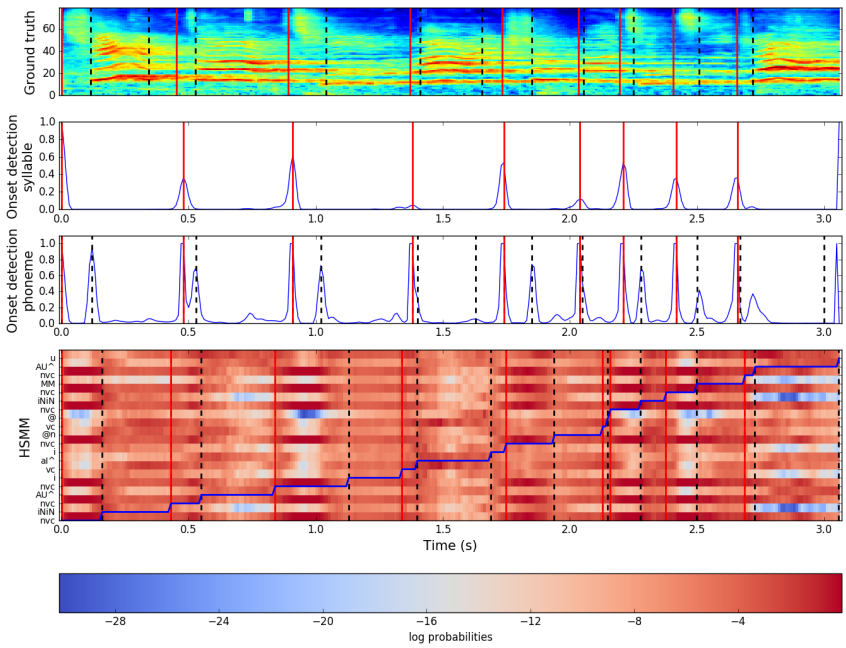


Figure 5.4: An illustration of the result for a singing phrase in the testing part of the dataset. The red solid and black dash vertical lines are respectively the syllable and phoneme onset positions. 1st row: ground truth, 2nd and 3rd rows: onset detection-based method, 4th row: HSMM-based segmentation method. The blue curves in the 2nd and 3rd row are respectively the syllable and phoneme ODFs. The staircase-shaped curve in the 2nd row is the alignment path.

be eventually applied to the singing voice of various languages because they could share the similar temporal-spectral patterns of phoneme transitions. Besides, the Viterbi decoding of the proposed method (time complexity $O(TS^2)$, T: time, S: states) is much faster than the HSMM counterpart (time complexity $O(TS^2 + T^2S)$). To showcase the proposed algorithm, an interactive jupyter notebook demo is provided for running in Google Colab[1].

[1] https://goo.gl/BzajRy

5.5 Improving the deep learning-based onset detection model

In the last section, we devised an onset detection-based syllable and phoneme segmentation method which firstly estimates the onset detection function (ODF) by using a deep learning classification model, then selects the correct onsets on the ODF by using a duration-informed HMM model. It is obvious that the accuracy of the onset selection step depends on largely on the quality of the ODF. Thus, it is necessary to explore an effective and efficient deep learning architecture for estimating the ODF.

In this section, we experiment with seven deep learning architectures for estimating syllable onset detection functions to find the most effective and efficient one for the onset detection task. The seven deep learning models are compared and evaluated on a jingju a cappella singing test dataset presented in Section 4.2.1.

5.5.1 Deep learning onset detection functions

We introduce the neural network setups and training strategies for the experiment which aims to find the most efficient network architecture trained separately on a jingju singing test dataset for syllable onset detection.

Searching for the most efficient neural network architecture

Following the terminology used in Pons et al.'s work(Pons, Gong, & Serra, 2017), we regard a neural network architecture as two parts – front-end and back-end. According to their work, the front-end is the part of the architecture which processes the input features and maps it into a learned representation. The back-end predicts the output given the learned representation. In this research, we don't restrict the functionality of back-end to prediction. However, we use it as terminology to differentiate from the front-end. We present the front-ends in table 5.4 and back-ends in table 5.5. **Conv** means convolutional layer. **10x 3 $\times$ 7** means 10 filters of which each convolves on 3 frequency bins and 7 temporal frames. All the Conv layers use ReLU activations. The first Conv layer in the front-end B has 6 different filter shapes. Each Conv layer in back-end

C and D follows by a batch normalization layer to accelerate the training(Ioffe & Szegedy, 2015). **BiLSTMs** means bidirectional RNN layers with LSTM units. In back-end B, both forward and backward layers in BiLSTMs have 30 units with Tanh activations. The activation function type of **Dense** layer – ReLU or Sigmoid used in back-end A depends on the architecture.

Table 5.4: Architecture front-ends

Front-end A	Front-end B
Conv 10x 3×7	Conv 24x 1×7, 12x 3×7, 6x 5×7 24x 1×12, 12x 3×12, 6x 5×12
Max-pooling 3×1	Max-pooling 5×1
Conv 20x 3×3	Conv 20x 3×3
Max-pooling 3×1	Max-pooling 3×1
Dropout 0.5	Dropout 0.5

Table 5.5: Architecture back-ends

Back-end A	Back-end B
Dense 256 units	Flatten
Flatten	BiLSTMs 30 units
Dropout 0.5	Dropout 0.5
Back-end C	**Back-end D**
Conv 40x 3×3	Conv 60x 3×3
Conv 40x 3×3	Conv 60x 3×3
Conv 40x 3×3	Conv 60x 3×3
Conv 80x 3×3	Flatten
Conv 80x 3×3	Dropout 0.5
Conv 80x 3×3	
Conv 135x 3×3	
Flatten	
Dropout 0.5	

We present seven architectures which are the combination pipelines of the front-ends and back-ends. All back-ends are connected with a sigmoid unit to output the ODF for the input log-mel contexts.

Baseline: Front-end A + back-end A with sigmoid activations. This architecture is the same as the one described in Schlüter and Böck's work (Schluter & Bock, 2014).
ReLU dense: Front-end A + back-end A with ReLU activations. In Schlüter and Böck's work (Schluter & Bock, 2014), using ReLU activations in the back-end A caused a drop in performance when evaluating on Böck dataset. However, ReLU activation function has been shown to enable better training of deeper networks because it has several advantages compared with Sigmoid, such as reducing the likelihood of vanishing gradient (Glorot, Bordes, & Bengio, 2011). We want to (re-)test the performance of ReLU activation on both Böck and jingju dataset.
No dense: Front-end A + Flatten layer. We use this architecture to test the effect of removing the dense layer in the baseline.
Temporal: Front-end B + back-end A with sigmoid activations. This one is similar to the "Temporal architecture" presented in Pons et al.'s work (Pons, Gong, & Serra, 2017), and uses various filter shapes in the first convolutional layer. In this work, we use 6 different filter shapes which are wide in temporal axis and narrow in frequency axis. Such kind of filter shape design aims to capture the onset spectral-temporal patterns on the spectrogram. It has been shown experimentally that on a smaller jingju dataset, this architecture outperformed the baseline by effectively learning the temporal onset patterns.
BiLSTMs: Front-end A with time-distributed Conv layers + back-end B. This one is similar to the CRNNs architectures presented in Vogl et al.'s work (Vogl, Dorfer, Widmer, & Knees, 2017). We use the sequence of the log-mel contexts as the architecture input and we experiment 3 different sequence lengths – 100, 200 and 400 frames. At the training phase, two consecutive input sequences are overlapped but their starting points are distanced by 10 frames. At the testing phase, the consecutive input sequences are not overlapped. We use this architecture to test the effect of replacing the dense layer in the baseline by RNN layer.
9-layers CNN: Front-end A + back-end C. We use this architecture to test the performance of deep CNN without using dense layer.
5-layers CNN: Front-end A + back-end D. As our datasets are relatively small, the above 9-layers CNN could be overfitting. Thus, we test also this shallow architecture with 5 CNN layers.

Table 5.6: Total numbers of trainable parameters (TNoTP) of each architecture.

Baseline	ReLU dense	No dense	Temporal
289,273	289,273	3,161	283,687
BiLSTMs	9-layers CNN	5-layers CNN	
278,341	288,286	81,541	

The Total numbers of trainable parameters (TNoTP) of each architecture is shown in table 5.6. To keep a fair comparison, we maintain a similar TNoTP between the baseline, ReLU dense, Temporal, BiLSTMs and 9-layers CNN architectures. We reduce the parameter numbers in No dense and 5-layers CNN architectures to explore the model efficiency. Notice that 9-layers and 5-layers CNNs are not fully-convolutional architectures (Long, Shelhamer, & Darrell, 2015) since we don't perform average pooling to the last Conv layer.

Model training

We use the same target label preparing strategy been described in Schlüter and böck's work (Schluter & Bock, 2014). The target labels of the training set are prepared according to the ground truth annotations. We set the label of a certain context to 1 if an onset has been annotated for its corresponding frame, otherwise 0. To compensate the human annotation inaccuracy and to augment the positive sample size, we also set the labels of the two neighbor contexts to 1. However, the importance of the neighbor contexts should not be equal to their center context. Thus the sample weights of the neighbor contexts are compensated by being set to 0.25. The labels are used as the training targets in the deep learning models to predict the onset presence.

Binary cross-entropy is used as the loss function. The model parameters are learned with mini-batch training (batch size 256), Adam (Kingma & Ba, 2014) update rule. 10% training data is separated in a stratified way for early stopping – if validation loss is

not decreasing after 15 epochs. In all experiments, we use `Keras`[2] with `Tensorflow`[3] backend to train the models.

5.5.2 Onset selection

The ODF output from the model is smoothed by convoluting with a 5 frames Hamming window. Onsets are then selected on the smoothed ODF. Two onset selection methods are evaluated. The first is a peak picking method which has been used in many MOD works (Böck, Krebs, & Schedl, 2012; Schluter & Bock, 2014; Vogl et al., 2017). We use the `OnsetPeakPickingProcessor` module implemented in Madmom (Böck, Korzeniowski, et al., 2016) package. Please refer to our code for its detailed parameter setting. Another onset selection method is based on the score-informed HMM presented in Section 5.4.2, which has been used to take advantage of the prior syllable duration information of the musical score.

5.5.3 Experimental setup

We use $ASPS_2$ test dataset presented in Section 4.2.1 and the metric – onset detection accuracy presented in Section 2.4.6 to evaluate the performance of each algorithms. We report the evaluation results for both peak-picking and score-informed HMM onset selection methods on jingju dataset. The pick-peaking results are reported by grid searching the best threshold on the test set, and the score-informed HMM results are evaluated directly on the test set since no optimization is needed.

We report only F1-measure in this paper. For jingju dataset, to average out the network random initialization effect, each model is trained 5 times with different random seeds, then the average and standard deviation results are reported. To measure the statistical significance of the performance improvement or deterioration, we calculate the Welch's t-test on the 5 training times results for jingju dataset. We report two tails p-value and reject the null hypothesis if the p-value is smaller than 0.05.

[2] `https://github.com/keras-team/keras`
[3] `https://github.com/tensorflow/tensorflow`

5.5.4 Results and discussions

In this section, we report and analyze the results for the most efficient architecture searching experiments. In tables 5.7, the p-value is calculated by comparing each model results with **Baseline**.

Table 5.7: Jingju dataset peak-picking (upper) and score-informed HMM (bottom) results of different architectures.

	F1-measure	p-value
Baseline	76.17±0.77	–
ReLU dense	76.04±1.02	0.840
No dense	73.88±0.44	0.002
Temporal	76.01±0.61	0.749
BiLSTMs 100	78.24±0.83	0.006
BiLSTMs 200	77.82±0.68	0.013
BiLSTMs 400	76.93±0.68	0.178
9-layers CNN	73.83±0.92	0.005
5-layers CNN	76.68±1.04	0.457
	F1-measure	p-value
Baseline	83.23±0.57	–
ReLU dense	82.49±0.28	0.057
No dense	82.19±0.44	0.021
Temporal	83.23±0.57	1
BiLSTMs 100	82.99±0.31	0.479
BiLSTMs 200	83.29±0.37	0.882
BiLSTMs 400	82.47±0.54	0.087
9-layers CNN	80.90±0.67	0.001
5-layers CNN	83.01±0.76	0.649

Observing table 5.7 – the results of jingju dataset, **BiLSTMs 100** and **200** outperform **Baseline** with peak-picking onset selection method but not with score-informed HMM method. **9-layers CNN** overfits and significantly performs worse than **Baseline**, which means this architecture is too "deep" and overfitted for this test dataset (check the Github page[4] for its loss curve). **Temporal** architecture has the p-value of 1 when evaluating by score-informed

[4] https://github.com/ronggong/musical-onset-efficient

HMM method, and we confirm that it is a coincidence after having checked its 5 training times F1-measures. **No dense** architecture performs significantly worse than **Baseline**. However, considering its tiny TNoTP – 3,161, this performance is quite acceptable. The similar case has been reported in Lacoste and Eck's work(Lacoste & Eck, 2007), where their 1 unit 1 hidden layer architecture achieved a remarkable result (only 4% F1-measure worse than their best architecture). This means that if the state-of-the-art performance is not required, one can use a quite small and efficient architecture. The score-informed HMM onset selection method outperform the peak-picking by a large margin. Also notice that the score-informed HMM method is able to compensate both good and bad performance of peak-picking, which can be seen by comparing upper and bottom results regarding **No dense**, **BiLSTMs 100** and **200** models.

Finally, we choose **5-layers CNN** as the most efficient architecture because it performs consistently equivalent to **Baseline** but only contains 28.3% TNoTP. Although **Temporal** architecture performs equally well, it is not selected because its equal TNoTP to **Baseline** and the complex configuration of its front-end B. **BiLSTMs** outperforms **Baseline** on jingju dataset, however, due to its overfitting on Böck dataset and slow training, we don't consider it as an efficient architecture.

Experiment code and pre-trained models used in the experiments are available in Github[4]. A Jupyter notebook running in Google Colab is prepared for showcasing the performance of different network architectures[5].

5.6 Conclusions

We formulate the syllable and phoneme segmentation problem within the context of jingju singing from two different perspectives – lyrics-to-audio alignment and onset detection. After setting up the baseline HSMM-based segmentation (alignment) method, we proposed the duration-informed onset detection-based method for tackling the segmentation problem. Finally, we explored various

[5] https://goo.gl/Y5KAFC

deep learning architectures for improving the syllable onset detection performance.

A detailed evaluation of HSMM-based segmentation method, onset detection-based method and various deep learning onset detection models was discussed for two jingju a cappella singing test datasets. Jingju singing, with distinct musical characteristics, is an ideal case to study the performance of novel methods for syllable and phoneme segmentation.

The duration-informed onset detection-based method explicitly considered coarse syllable and phoneme duration information for the segmentation. However, the algorithm is language-independent, and thus can easily adapt to the singing voice of various languages and even to instrumental playing. Since onset detection-based method is a binary onset/non-onset classification model, it requires a small amount of syllable or phoneme onset annotated training data.

The duration-informed onset detection-based method shows significant promise in syllable and phoneme segmentation task. It provides a significant improvement in both onset detection and segmentation performance compared with the baseline HSMM-based method for jingju a cappella singing. An exploration of various deep learning syllable onset detection models showed that the architecture of the deep learning model cannot affect significantly the onset detection performance, however, one can design an efficient architecture to reach the state of the art performance.

One main limitation of the onset detection-based method presented was the assumption that a similar duration of each syllable or phoneme and the same syllable or phoneme sequence should be sung in both teacher's and student's singing pieces. While this is a fair and realistic assumption for jingju professional training since the students can usually imitate the teacher's singing very well, amateur singers might imitate very badly because of a large deviation of the syllable or phoneme duration and missing or extra syllable or phoneme. A coarse syllable and phoneme duration correction might be necessary there before applying the segmentation algorithm. The syllable or phoneme recognition might be the method used to tackle the problem of missing or extra syllable in student's singing.

The onset detection-based segmentation method utilized

duration-informed HMM to select the correct onsets on the onset detection function (ODF). A good quality ODF is essential to reach a desirable syllable or phoneme onset detection/segmentation accuracy. Various deep learning architectures were experimented to search for a most effective and efficient one which can lead to a superior syllable onset detection performance. Although the experiment did not show an improvement in onset detection accuracy, we witnessed that a CNN architecture without dense connection reached the state of the art performance, while it has much less trainable parameters than the baseline architecture, which indicates that the segmentation model efficiency can be improved by using this deep learning architecture.

The presented onset detection-based segmentation method can be further improved to incorporate other linguistic information such as the phoneme class of each time frame. This is in addition to the ideas explored already – HSMM-based method utilized phoneme class and duration information, while onset detection-based method utilized syllable/phoneme onset and duration information. Such a model which makes use of all of the three information – onset, phoneme class and duration need to be further explored.

The automatic syllable and phoneme segmentation methods discussed in the chapter are aligned with the goal to lead towards an automatic pronunciation assessment system in a fine granularity for jingju a cappella singing. Syllable and phoneme segmentation is the first step towards this goal. The methods of mispronunciation detection and pronunciation similarity measures are built based on the results of the segmentation.

Chapter 6

Mispronunciation detection

Mispronunciation detection is a popular speech assessment task, and the developed system is used for Computer-aided language learning (CALL). As we have discussed in Section 3.1, an accurate pronunciation of each singing syllable is an essential aspect in jingju performing, which is stressed by both teacher and student in the actual jingju singing training scenario. A system which can detect the mispronounced singing syllable automatically is a crucial component of the automatic system for assessing the jingju singing pronunciation.

Mispronunciation detection aims to detect automatically the mispronounced syllables or phonemes in student's singing. It tags each syllable or phoneme in the singing recording as either mispronunciation or correct pronunciation. Within the context of the jingju singing, we constrain the detection to (1) syllable-level and (2) two types of mispronunciation – jiantuanzi and special pronunciation. We will explain in detail these two constraints in the next section.

This chapter aims to address the automatic mispronunciation detection task within the context of jingju singing, presenting several methods and an evaluation of these methods. The main aims of this chapter are:

1. To address automatic mispronunciation task for jingju singing.

The problem is formulated as building discriminative machine learning models to classify binarily the singing syllables into mispronounced or correctly pronounced class. Several neural network architectures are experimented to address this problem.

2. To present a description of the forced alignment-based baseline method and the discriminative model-based method for mispronunciation detection.

3. To present an evaluation of the forced alignment-based method and the discriminative model-based method, and explore two deep learning architectures intending to improve the discriminative detection model.

The implementation code used in the experiments of this chapter is openly available[1].

6.1 Task description

We describe the automatic mispronunciation detection task in this section. We will also describe how the approaches presented in this chapter can be adapted to this task, making them flexible to the available audio recordings and annotations. The task description presented in this section is continued building on the problem formulation presented in Section 3.4.3.

Given the singing audio recording pre-segmented into pieces of melodic line-level, the pronunciation dictionaries (lexicon), the most relevant detection task is to detect the mispronunciations of the special pronounced syllables or jiantuanzi syllables in the melodic line. We constrain the detection at syllable-level because it is the basic pronunciation unit in jingju singing teaching which has semantic meaning (Section 2.2.1). We also constrain the detection task to two types of mispronunciation – jiantuanzi and special pronunciation, since they are two main sources of mispronunciation in jingju singing training (Section 2.2.3 and Section 2.2.4). In the context of jingju singing, forced alignment aims to time-align a priori syllable sequence in pinyin format with the melodic line

[1] https://github.com/ronggong/mispronunciation-detection

singing audio recording. The forced alignment method uses a dictionary with multiple pronunciations for a particular syllable entry, and then the pronunciation which matches best with the singing acoustics will be decoded in the aligned syllable sequence. The mispronunciation detection result can be obtained by comparing the decoded syllable sequence with the teacher's syllable sequence. A mispronunciation discriminative model aims to classify a syllable segment into either mispronounced or correctly pronounced class. The onset detection based syllable segmentation method presented in Section 5.4 will be used as the preliminary step to obtaining the syllable segment from the melodic line. In the context of this thesis, as the syllable sequence of the teacher's demonstrative singing is always available, the information of the syllable type in a melodic line is known in advance, which is to say, we know which syllables in a melodic line are special pronunciation or jiantuanzi. Such information is necessary for the algorithm evaluation step.

The two main tasks in this chapter are setting up the forced alignment-based baseline detection method and proposing the discriminative model-based method. As the third task of this chapter, we explore two deep learning architectures intending to improve the discriminative models. The performance of all the three tasks will be evaluated on MD dataset (Section 4.2.2). The results and the pros and cons of two misrponunciation detection methods will be discussed in detail.

6.2 Forced alignment-based method

Forced alignment is a technique which time-align the syllable or phoneme orthographic transcription with the speech or singing voice audio. It is a preliminary step in a speech recognition system for training the acoustic model. The baseline method for syllable and phoneme segmentation presented in Section 5.3 also used the forced alignment technique. In this section, we will build a forced alignment system based on Kaldi toolkit (Section 2.5.3). This system will make use a special pronunciation dictionary to decode the syllable sequence of the jingju singing recording. The decoded syllable sequence is intended to reflect the actual pronunciation by inspecting the acoustics of the recording. Then the evaluation of the

mispronunciation detection performance is done by comparing the decoded syllable sequence with the teacher's demonstrative syllable sequence.

6.2.1 Preparing lexicons for the forced alignment

Kaldi is a toolkit for constructing speech recognition system based on Finite State Transducers (FSTs) (Mohri, Pereira, & Riley, 2002). The advantage of using Kaldi to build a forced alignment system is that many code recipes are provided for some speech datasets, and only minimal effort is required to modify a certain recipe to our singing voice dataset.

The principle idea of performing forced alignment for the mispronunciation detection is that the system could make use of a pronunciation dictionary (lexicon) with multiple pronunciations for each syllable entry to decode the syllable sequence. The decoded sequence can reflect the actual pronunciation of the singing recording. The critical steps are preparing the pronunciation lexicons, which are the dictionaries of the syllables and their corresponding phoneme transcriptions. In the forced alignment system training step, we provide the dictionary with the exact pronunciation because the phonetic level annotation of the training set is known in advance. An example lexicon for the system training is:

HAO0	x	AUˆ	u				
WANG0	w	AN					
MIN0	m	in					
NGO0	w	O					
JIN0	c	in					
ZAO0	c	AUˆ	sil_phone	AUˆ	u		
HAO1	x	AUˆ	sil_phone	AUˆ			
WANG1	w	AN	sil_phone	AN	sil_phone	AN	N
MIN1	m	in	N				
NGO1	N	O					

Where the first column of each line is the syllable and the following characters are the phonetic pronunciation of this syllable in X-SAMPA format (Appendix B, sil_phone indicates the silence).

Each syllable is postpended with numbers, which indicates different pronunciations of this syllable. The Baum-Welch algorithm of the system learns the monophone acoustic model by giving the lexicon and the syllabic level transcription of each singing melodic line. We use the MFCCs with the energy as the feature representation of the audio.

In testing phase, the exact pronunciation of the testing singing melodic line is unknown. To make Kaldi choose the pronunciation of a syllable which matches the best with the singing acoustics, we merge the syllable pronunciation entries and remove the postpended numbers. For example, we merge the syllable "wo" and its corresponding special pronounced syllable "ngo" to an identical syllable entry "wo", and merge the syllable "xiang" and its corresponding jianzi syllable "siang" to an identical syllable entry "xiang". Then the above lexicon becomes:

WO	x	AUˆ	u				
WANG	w	AN					
MIN	m	in					
WO	w	O					
JIN	c	in					
ZAO	c	AUˆ	sil_phone	AUˆ	u		
HAO	x	AUˆ	sil_phone	AUˆ			
WANG	w	AN	sil_phone	AN	sil_phone	AN	N
MIN	m	in	N				
NGO	N	O					

The alignment decoding graph in Kaldi will contain the alternative pronunciations for a single syllable and decode the phoneme sequence which matches the best with the acoustics. The syllable sequence of the testing melodic line can be then inferred from the decoded phoneme sequence.

6.2.2 Experimental setup

We use the MD test dataset presented in Section 4.2.2 and classification accuracy metric presented in Section 2.4.6 to evaluate the performance of the forced alignment system. We only evaluate the syllables that teacher pronounces as the special pronunciation

or jianzi. If the student pronounces a syllable wrongly, the true negative is that the decoded syllable is not equal to the teacher's syllable transcription. While the student pronounces a syllable correctly, the true positive is that the decoded syllable is equal to the teacher's syllable transcription. The detection accuracy is reported separately for special pronunciation syllable type and jiantuanzi type.

6.2.3 Results and discussions

The results are shown in Table 6.1. 69.08% of the special pronunciation syllables and around half of the jianzi syllables in the testing set are correctly detected.

Table 6.1: The evaluation result table of the forced alignment mispronunciation detection method. #Correctly detected: number of correctly detected syllables; #Total: number of total syllables; Accuracy: binary classification accuracy; Special: special pronunciation task; jianzi: jiantuanzi task.

#Correctly detected special	#Total special	Accuracy special
324	469	69.08%
#Correctly detected jianzi	#Total jianzi	Accuracy jianzi
26	50	52%

To our surprise, the detection for the special pronunciation syllable type reaches an average level detection accuracy – 69.08%, as the true positive criterion is quite strict, which requires that the decoded syllable be equal to the teacher's syllable transcription. While the detection accuracy for jianzi syllable type is undesirable. The possible reason could be that the forced alignment system is not able to decode the non-voiced consonants correctly. As we have discussed in Section 4.2.2, the difference between the mispronounced jianzi syllable and correctly pronounced jianzi syllable mainly lies on the different pronunciations of the non-voiced consonant. In the next section, we will explore the discriminative model-based detection method which intends to make the decision based on a particular part of the syllable, for example, the non-voiced consonant part.

6.3 Discriminative model-based method

The unsatisfactory performance of the forced alignment-based model motivates to explore an alternative mispronunciation detection method. In this section, we devise a mispronunciation detection method based on the syllable segmentation and the discriminative model. We use the same syllable segmentation method presented in Section 5.4 to segment automatically the jingju singing melodic line into syllable segments. As the testing set contains only the student's recordings, we use the coarse syllabic durations extracted from the corresponding teacher's recordings to build the a priori duration model. Although the segmentation algorithm will inevitably cause the segmentation errors which can be propagated to the mispronunciation detection step, we still adopt the automatic segmentation rather than using the ground truth annotation of the syllable boundary in order to perform a fair evaluation with the baseline algorithm.

As the input representation for the discriminative model, we use the same logarithmic Mel (log-mel) representation presented in Section 5.2.1, except that no overlapped context window will be used. Thus the input to the model is variable-length syllable segments which are represented by two dimensional log-mel spectrogram.

We construct two discriminative models respectively for the mispronunciation detection of the special pronunciation syllable and the jiantuanzi. We present various deep learning techniques in the next section for building the model.

6.3.1 Discriminative deep learning models

As mentioned in Section 2.5.1, recurrent neural networks (RNNs) are the natural choice to model acoustic sequential data. Thus our initial model is a bidirectional RNNs with Long short-term memory (LSTM) units. We also explore three deep learning techniques – using additional convolutional layers to learn the local connectivity of the input representation, using feed-forward attention mechanism to allow the model to make the decision by weighting the

most important syllable part, and using dropout to overcome the overfitting.

Bidirectional LSTMs

Due to the small size of the training data (Section 6.3.2), we experiment with a one-layer bidirectional LSTM (BiLSTM) recurrent model, which has 8 LSTM units in each direction. The output layer has one sigmoid unit for the binary classification.

Additional convolutional layers

Convolutional layer uses the receptive field to capture the local connectivity of the input representation and can extract music meaningful features by designing the kernel shape (Pons, Slizovskaia, et al., 2017). We stack a 2-layers CNN between the input and the RNN layer.

Table 6.2: 6-layers CNN, "8x 1×3 ReLU" means 8 kernels of which each convolves on 1 frequency bins and 3 temporal frames, using ReLU activation function.

Conv 8x 1×3 ReLU
Max-pooling 1×3
Conv 16x 1×3 ReLU
Max-pooling 1×3

Table 6.2 shows the CNN architecture. It does convolution and max-pooling only in frequency axis because we only want to capture the frequential local connectivity and maintain the temporal resolution.

Feed-forward attention mechanism

In the initial BiLSTM network, the output sigmoid layer takes the last time stamp hidden state of the RNN as the input. Attention mechanism provides a way to capture the global sequence information rather than only to classify based on the last hidden state. The original attention has been proposed in the context of sequence-to-sequence model for the machine translation purpose (Bahdanau,

Cho, & Bengio, 2014). Then this mechanism or its variants are applied for image caption generation, video clip description, machine reading comprehension and speech recognition (Cho, Courville, & Bengio, 2015; Xu et al., 2015; Hermann et al., 2015). In this work, we use the feed-forward attention proposed by C. Raffel and D. P. W Ellis (Raffel & Ellis, 2015) because it is suitable for the classification task. This mechanism can be seen as producing a fixed-length embedding of the input sequence by computing an adaptive weighted average of the entire state sequence.

Dropout

To prevent our model from overfitting on the small size training set, we experiment 0.5 rate dropout for both input and output of the RNN.

Models training

The target labels of the training set are prepared according to the ground truth annotation. We set the label of the mispronunced syllable to 1, and the correctly pronounced syllable to 0. The model parameters are learned with mini-batch training (batch size 1 due to the variable-length of each training sample), adam update rule (Kingma & Ba, 2014), and early stopping – if validation loss is not decreasing after 15 epochs.

6.3.2 Experimental setup

The experimental setup is similar to the one mentioned in Section 6.2.2. We also use the MD test dataset and classification accuracy metric the performance of the discriminative model. The task aims to discriminate between the mispronounced syllable and the correctly pronounced syllable. Thus we subsample from the MD dataset the special pronunciation syllables, jianzi syllables as the positive samples and their standard pronunciation syllables as the negative samples. Table 6.3 shows the numbers of the special pronunciation and jiantuanzi syllables in the entire training set. The average syllable duration is 86.09 frames (2.15 s) and the standard deviation duration is 119.89 frames (3.0 s).

Table 6.3: Numbers of the special pronunciation (special) and jiantuanzi syllables in the training set.

#special positive	#special negative	#jiantuanzi positive	#jiantuanzi negative
463	1083	41	242

We use 5-folds cross-validation to report the classification accuracy for the model architecture selection. In each fold, 75% samples of the entire training set is split as the train set, and another 25% samples is reserved for the validation set. The mean validation loss (MVL) is reported separately for models of special pronunciation and jiantuanzi tasks. In the testing phase, the best architectures which have the minimum MVL is chosen to train on the entire training set once, and then the trained models are evaluated on the test set. We also report the results for the automatic syllable segmentation evaluation. The evaluation metrics for the segmentation – onset detection F1-measure and segmentation accuracy, are described in Section 2.4.6.

6.3.3 Results and discussions

We show in Table 6.4 the F1-measure onset detection and segmentation accuracy results which indicate the automatic syllable segmentation performance. We can observe a high segmentation accuracy 95.19% and an average onset detection F1-measure 78.74%, which means that a certain amount of the detected onsets do not lie within the 50 ms tolerance window constrained by the ground truth onsets. As the non-voiced consonants usually have a short duration, those onset detection errors might cause an inaccurate segmentation of the non-voiced consonants and can be propagated into the mispronunciation detection step.

Table 6.5 shows the number of parameter of each model architecture and MVL results of the model architecture selection step for each special pronunciation and jiantuanzi models. All of the additional deep learning techniques – CNN, attention and dropout, help improve the model performance of the vanilla BiLSTM. For the detection task of the special pronunciation syllables, the result of

Table 6.4: Evaluation results of the preliminary automatic syllable segmentation step. Onset detection F1-measure and segmentation accuracy are reported.

Onset F1-measure	Segmentation accuracy
78.74%	95.19%

the dropout technique reaches the minimum MVL – 0.6152, which means that this technique to avoid overfitting is crucial for such a small training set. While for the task of the jiantuanzi syllables, the combination of all the techniques reaches the minimum MVL – 0.3457.

Table 6.5: The number of parameters of each model architecture and the mean validation loss (MVL) results of the special pronunciation (special) and jiantuanzi models. CNN: additional convolutional layers, Att.: feed-forward attention mechanism, Comb.: combine BiLSTM, CNN, attention and dropout architectures.

	BiLSTM	CNN	Att.	Dropout	Comb.
MVL special	0.7488	0.6600	0.6560	0.6152	0.6574
MVL jiantuanzi	0.5046	0.3523	0.3892	0.3754	0.3457
#params	5713	9217	5730	5713	9234

We use these two architectures to train respectively the final special pronunciation and jiantuanzi models on the entire training set. Then we evaluate the trained models on the test dataset.

Table 6.6 shows the mispronunciation detection results for both special pronunciation syllables and jiantuanzi syllables. We can observe that the discriminative model degrades the detection performance for special pronunciation syllables compared with the baseline forced alignment results – from 69.08% to 64.68%, which might due to that the discriminative model training only accessed a subset of the MD dataset, while the baseline model training utilised the entire MD dataset. On the other hand, the detection accuracy for the jiantuanzi task is improved significantly. To illustrate the effect of the attention mechanism, we visualize the logarithmic Mel

Table 6.6: The evaluation result table of the discriminative model-based mispronunciation detection method. #Correctly detected: number of correctly detected syllables; #Total: number of total syllables; Accuracy: binary classification accuracy; Special: special pronunciation task; jianzi: jiantuanzi task.

#Correctly detected special	#Total special	Accuracy special
304	469	64.68%
#Correctly detected jianzi	#Total jianzi	Accuracy jianzi
34	50	68%

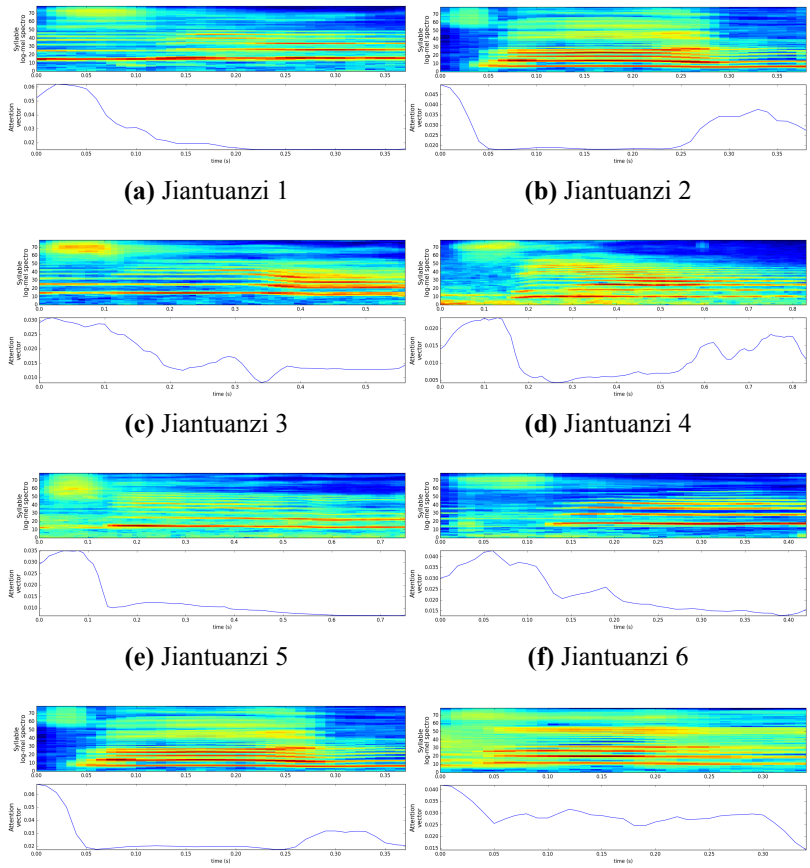
spectrogram and the attention vector output from the model decoding process in Figure 6.1.

We can notice from Figure 6.1 that the attention vectors have a relatively high value towards the non-voiced consonant part of the syllable (the noise-like spectrogram at the syllable beginning), which means that the attention mechanism allows the model to make the decision mainly on the non-voiced consonant part of the syllable, which is the segment to discriminate a mispronounced and a correctly pronounced jiantuanzi syllable.

6.4 Improving the discriminative mispronunciation detection models

In the last section, we devised a discriminative model-based mispronunciation detection method. The discriminative model is based on deep learning architecture which classifies the input syllable segment binarily into mispronounced or correctly pronounced class. The classification accuracy largely depends on the deep learning architecture, the size and quality of training dataset. To collect more training data would involve the participation of multiparty, e.g., artists, recording engineers, which is more difficult in coordination and more time-consuming than experimenting new deep learning architectures.

In this section, we experiment with two new deep learning architectures and intend to improve the mispronunciation accuracy. These two architectures have been proposed recently for the se-



(a) Jiantuanzi 1 **(b)** Jiantuanzi 2

(c) Jiantuanzi 3 **(d)** Jiantuanzi 4

(e) Jiantuanzi 5 **(f)** Jiantuanzi 6

(g) Jiantuanzi 7 **(h)** Jiantuanzi 8

Figure 6.1: The visualization of the logarithmic Mel spectrogram and the attention vector output from the model decoding process for jiantuanzi syllables.

quential data modelling.

6.4.1 Temporal convolutional networks

Temporal convolutional networks (TCNs) is a deep learning architecture proposed by Bai et al. (Bai, Kolter, & Koltun, 2018). TCNs adopt three novel deep learning techniques – causal convolutions, dilated convolutions and residual connections to perform sequential modelling rather than using RNN-related architectures. The author evaluated TCNs along with several traditional sequential modelling

architectures, such as LSTM, GRU and RNN, on various sequential modelling tasks, such as polyphonic music modelling, word- and character-level language modelling. The experiment shows that TCNs outperformed those architectures in most of the tasks. The main component of the TCNs is a 1D full-convolutional network (FCN). The causal convolutions restrict that there can be no leakage of information from the future into the past. The dilated convolutions can achieve a large history size while maintaining a relatively small layer number (not too deep) and filter size (not too large). Residual connections can stabilize a large and deep network by only learning the modifications to the identity mapping. The memory retention analysis shows that TCNs exhibit substantially longer memory than the LSTMs and GRUs of the same size.

TCNs hold several hyperparameters which need to be tuned when applying the model to the mispronunciation detection task. The most important factor for choosing hyperparameters is to make sure the TCN has a sufficiently large receptive field to cover the amount of the sequential context. The relevant hyperparameters are the number of stacks n of the FCN, dilation factor d and filter size k. The receptive field size which is equal to $d \times k \times n$ needs to be adapted to the average syllable length of our training dataset – 86.09 frames. We experiment with three hyperparameter configurations:

Table 6.7: Configurations of three TCNs. n: number of stacks, k: filter size, d: dilation factor.

	Receptive field size (frames)	n	k	d for each stack
TCNs1	8192	4	8	[2, 4, 16, 256]
TCNs2	128	4	8	[1, 2, 4, 16]
TCNs3	96	4	8	[1, 2, 8, 32]

6.4.2 Self-attention mechanism

The feed-forward attention mechanism presented in Section 6.3.1 aims to learn a fixed-length embedding vector by weighted summing the RNN hidden states of all timestamps. The attention vector

is learned by a multilayer perceptron (MLP) network, which allows the model to emphasize certain hidden states. Lin et al. (Z. Lin et al., 2017) proposed a structured self-attention mechanism which learns a 2D embedding matrix rather than a vector. Applying for the natural language sentence embedding task, they claim that each row of the matrix can attend on a different part of the sentence.

They argue that the feed-forward attention mechanism using an embedding vector usually focus on a specific component (time region) of the input sequence. To allow the model to attend to multiple components, they proposed to use multiple attention vectors, which forms an embedding matrix. Consequently, to compute the embedding matrix is to learn a weighting matrix of which each row is the weighting vector of the hidden state sequence. In the implementation, we use a 2-layer MLP without bias to learn this weighting matrix.

The theoretical framework of the self-attention mechanism is compelling, and also has practical meaning when applying to the mispronunciation detection. For example, when a student mispronounces a syllable, she/he might commit errors on multiple parts of the syllable. E.g., the mispronunciation of "sian" could be "xiang", where the student made the errors on both non-voiced consonant "s" $\rightarrow$ "x" and the terminal consonant "n" $\rightarrow$ "ng".

In this work, we experiment with the self-attention mechanism with the BiLSTM architecture presented in Section 6.3.1. To prevent overfitting, we restrict the number of parameters in the architecture and use 16 hidden units for each layer of the MLP.

6.4.3 Experimental setup

The experimental setup is the same as it has been mentioned in Section 6.3.2.

6.4.4 Results and discussions

Table 6.8 shows the MVL results of three TCNs hyperparameter configurations. For special pronunciation task, the TCNs3 with the smallest receptive field size performs the best. While for jiantuanzi task, the TCNs1 with the largest receptive field size performs the best. The possible reason is that the model needs a large receptive

field to retain the long history in order to detect the mispronunciation of a jianzi syllable, of which the mispronunciation usually happens at the non-voiced consonant part which is also the beginning of the syllable. While for special pronunciation task, the mispronunciation usually happens at the syllable belly or tail position, which doesn't require the model to have a large receptive field to retain the long history. However, the best results among the three configurations still lag behind those of the previous experimented BiLSTM models (Section 6.5), that we need a further study to understand the poor results of the TCNs. An assumption could be that TCNs requires more training data to work properly. However, our current training dataset is too small.

Table 6.8: Mean validation loss (MVL) of three TCN hyperparameter configurations for special pronunciation and jiantuanzi detection tasks.

	TCNs1	TCNs2	TCNs3
MVL special	1.2831	1.1702	1.0787
MVL jiantuanzi	0.8877	0.9986	1.2243
#params	14609	7505	8097

Table 6.9 shows the MVL results of self-attention and feed-forward attention mechanism. We can observe an improvement by adopting self-attention, which means that using the self-attention mechanism individually with BiLSTM leads to a better performance than using feed-forward attention. However, while combining self-attention with other deep learning techniques mentioned in Section 6.3.1, the performance for the special pronunciation task does not surpass the best result reported in Table 6.5, and the performance for the jiantuanzi task is worse than using self-attention individually.

Because of the inferior results of TCNs and self-attention mechanism, we would not include them in the final models for the mispronunciation detection tasks. Consider that the relatively small training data size might be the bottleneck of improving the deep learning-based detection models, it is more reasonable to collect

Table 6.9: Mean validation loss (MVL) of self-attention and feed-forward attention architectures for special pronunciation and jiantuanzi detection tasks. Self-att.: self-attention, Feed-forward: feed-forward attention, Comb.: combine BiLSTM, CNN, self-attention and dropout.

	Self-att.	Feed-forward	Comb.
MVL special	0.6458	0.6560	0.6313
MVL jiantuanzi	0.3512	0.3892	0.3943
#params	6257	5730	9761

more training data firstly, then study the performance of different deep learning architectures.

6.5 Conclusions

This chapter presented a detailed formulation of the task of mispronunciation detection in jingju singing voice. The approaches utilized the automatic syllable segmentation algorithm presented in the last chapter and a deep learning-based discriminative model to perform the detection on two types of jingju singing syllables. Evaluation on an amateur jingju singing dataset showed the possibility of this approach and its limitations. The goal of developing such model was to present a methodology for mispronunciation detection in automatic singing voice assessment system of jingju music. The work presented in this chapter was preliminary and not comprehensive, with a great possibility for further study and improvement. However, the basic idea of using a deep learning-based discriminative model to achieve the mispronunciation detection is valid.

We mainly addressed the problem of the detection of two types of mispronounced syllables in jingju singing recordings – special pronunciation and jiantuanzi. The presented method firstly used the onset detection-based automatic syllable segmentation algorithm to obtain the segment of each syllable, then classified each syllable segment to mispronounced or correctly pronounced class by using a deep learning-based discriminative model. Compared to a baseline

forced alignment method, we showed that the proposed method is more advantageous in detecting the mispronunciation of the jiantuanzi syllable type. By illustrating the attention vector, we found that the attention mechanism is useful in putting more weights in the non-voiced consonant part of, and thus to help the model to make a better detection of the jiantuanzi mispronunciation. Additionally, intending to improve the detection accuracy of the discriminative model, we adopted two newly developed deep learning techniques for sequential modelling to our mispronunciation detection task. However, the results showed that their performance was not ideal, and inferior to our initial discriminative model.

For future work, we aim to improve the discriminative model performance by collecting more training dataset. Deep learning techniques are known to be data-consuming. However, our current dataset is too small to train a proper deep-learning based discriminative model and to outperform the forced alignment-based model which usually requires much less training data. The next steps would be performing an extensive hyperparameter tuning for the deep learning models since the performance of such models can be optimized by considering the coordinative effect between the hyperparameters and the size of the training data.

Chapter 7

Pronunciation and overall quality similarity measures

Pronunciation and overall quality similarities measurement is a subtask in singing voice assessment. which is useful in the online singing training scenario to assess the pronunciation quality and the overall quality of the student's singing. In the last chapter, we have discussed the possibility of using computational models to detect the mispronunciation syllables in jingju singing. However, as we have mentioned in Section 3.2.1, in some cases, although the student does not commit any mispronunciation, there still exists a clear gap on pronunciation and overall quality between the singing of teacher and student. The rigour of the jingju singing training and the learning by imitation training method require the student, especially the professional one, to imitate the timbre quality of the jingju master. Thus, a system which can measure the pronunciation and overall quality similarities between the singing of teacher and student is a useful component of the automatic system for jingju singing assessment.

In the context of this dissertation, pronunciation and overall quality similarities measurement aims to measure the pronunciation and overall quality similarities between the teacher and student's corresponding phoneme segments. This chapter aims to address the

pronunciation and overall quality similarities measurement task in the context of jingju singing training, presenting several methods and an evaluation of these methods. The main goal of this chapter are:

1. To address the similarity measure problem for jingju singing. The problem is formulated as building machine learning models to perform phoneme embedding regarding pronunciation and overall quality aspects. Several neural network architectures are experimented to address this problem.

2. To present a description of the classification model for phoneme embedding, and to explore the siamese network model for the same purpose.

3. To present an evaluation of the classification model and the siamese model.

7.1 Task description

Task description of the pronunciation and overall quality similarities measurement in this section is continued building on the problem formulation presented in Section 3.4.4.

Given the singing audio recording pre-segmented into phoneme-level, the most relevant task is to develop the computational models which can measure the pronunciation and overall quality similarities between phoneme segments. We constrain the granularity at phoneme-level because it is the smallest pronunciation unit in jingju singing teacher, and it is also the basic component to constitute the high-level singing unit, such as syllable and phrase. In the context of this thesis, we mainly consider using phoneme embedding to distill the pronunciation and overall quality information of the phoneme segment, then apply distance measures to define the similarity between two segments. The advantages of using phoneme embedding rather than the traditional sequential alignment method for the similarity measures have been discussed in Section 3.4.4. We adopt deep learning-based methods for generating phoneme embeddings from variable-length phoneme segments. The deep learning-based

classification model aims to classify the phoneme segment into phoneme and overall quality categories. The output of the second last layer of the classification model will be used as the embedding. As an exploration, we also experiment the siamese network architecture for phoneme embedding learning task since this architecture was designed for measuring similarity between multiple inputs. Then, the similarity between two phoneme segments can be obtained by calculating the distance measure of their phoneme embeddings. Differ from the last chapter, we will evaluate the model performance directly on the manually pre-segmented phoneme segments rather than involving any automatic phoneme segmentation step into the pipeline. Thus, we leave the joint evaluation of phoneme segmentation and similarity measure for future work.

The two main tasks in this chapter are setting up the classification phoneme embedding model, proposing several improvements, and explore the siamese phoneme embedding model. The performance of these two tasks will be evaluated on POQSM test dataset. The results of the two models will be discussed in detail.

7.2 Baseline phoneme embedding networks

We introduce a phoneme embedding neural network as the baseline model, which is able to convert variable-length phoneme segments into fixed-length vectors. We use the logarithmic Mel (log-mel) spectrogram of the phoneme segment as the input. The frame size and hop size of the spectrogram are respectively 46.4ms (2048 samples) and 10ms (441 samples). The low and high-frequency bounds of the Mel bank filters are 27.5Hz and 16kHz. This input representation is similar as we have been mentioned in Section 6.3.

7.2.1 Fully-supervised classification network

We call this network the fully-supervised classification network, because we use fully-supervised training method and provide to the network the phoneme class label for the pronunciation classification, or the professional/amateur binary label for the overall quality

classification. Figure 7.1 shows a diagram of this network. The main part of the network is a single or multi recurrent layers. The optimal layer number will be decided in the Section 7.2.4. The last layer of the network use softmax units for the categorical classification. We take the output vector from the last layer as the embedding - either a 27 dimensional vector for the pronunciation embedding (figure 7.1 left part) or a 2 dimensional vector for the overall quality embedding (figure 7.1 right part). We use categorical cross-entropy loss during the network training. The embeddings learned by this network are expected to capture either the pronunciation or overall quality characteristics of the phoneme segment. We also experimented sharing the weights between the left and right branches of the architecture, so that we could use one network to learn both pronunciation and overall quality embeddings, which is the idea of multi-task learning (Ruder, 2017). However, our experiment shows that it doesn't work better than individual task learning.

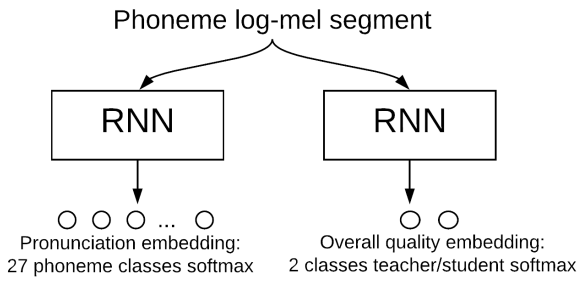


Figure 7.1: Fully-supervised classification phoneme embedding network for learning pronunciation (left part) or overall quality (right part) embeddings. RNN: recurrent network network.

7.2.2 Model training

The weights of the network are learned with mini-batch training (batch size 64), adam (Kingma & Ba, 2014) update rule and early stopping of 15 epochs. To accelerate the network training, we bucket the training phoneme segments which have a similar frame length into the same mini-batch. The segments are then zero-padded so that they have the same length as the longest segment of the mini-batch.

7.2.3 Experimental setup

We use pronunciation and overall quality similarity measures (PO-QSM) test dataset presented in Section 4.2.3 for the evaluation purpose. The train, validation and test split of this dataset can be consulted in Table 4.9.

For the pronunciation aspect, we measure the cosine similarity for every phoneme embedding pair in the test set. The ground truth label of an embedding pair is 1 if they belong to the same phoneme class, or 0 vice versa. For overall quality aspect, we only measure the cosine similarity of the phoneme embedding pairs of the same phoneme class. The ground truth label is 1 if two embeddings belong to the same overall quality class – professional or amateur, 0 vice versa. We report the average precision (AP) between the cosine similarities and the ground truth as the evaluation metric. The AP is used previously to evaluate speech word acoustic embedding (Kamper et al., 2016; Settle & Livescu, 2016). It is also suggested as the metric for imbalanced test set (Davis & Goadrich, 2006), which is the case of the pronunciation aspect evaluation.

We experiment 9 RNN architectures of bidirectional LSTM (BiLSTM) recurrent layer and fully-connected layer combinations and report their AP on the validation set. Each recurrent layer is bidirectional with 32 LSTM units in each direction (BiLSTM). Each fully-connected layer has 64 ReLU activation units and followed by a dropout layer with 0.5 dropout rate. We train each model 5 times with different random seeds, and take the mean value of the average precisions. Two optimal architectures are decided separately for pronunciation embedding and overall quality embedding. Finally, we evaluate the performance of the optimal architectures on the test set.

7.2.4 Results and discussion of the baseline

Table 7.1 shows the AP results on the validation set for 9 different architectures. We observe that fully-connected layer doesn't help increase the AP. The pronunciation and overall quality aspects reach their highest AP respectively by using 2 BiLSTM layers and 1 BiLSTM layer.

Table 7.1: Mean value of average precision on the validation set over 5 runs, using classification network embedding. R: # recurrent layers, F: # fully-connected layers.

R	F	Pronunciation AP	Overall quality AP
1	0	0.690	**0.934**
1	1	0.694	0.926
2	0	**0.695**	0.915
2	1	0.694	0.928
2	2	0.689	0.927
3	0	0.691	0.924
3	1	0.695	0.920
3	2	0.684	0.924
3	3	0.673	0.920

Table 7.2: Average precision on the test set over 5 runs, using optimal network architectures.

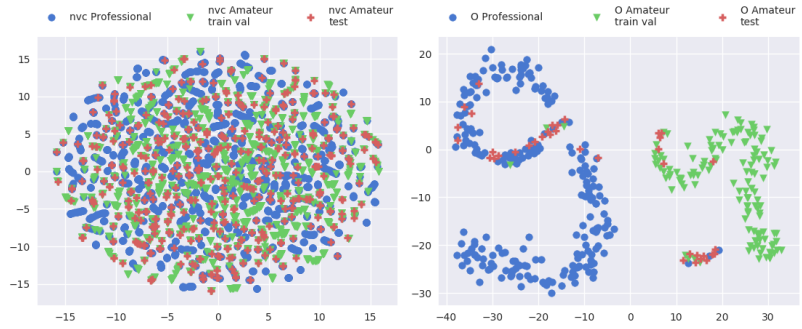
Pronunciation AP	Overall quality AP
0.645	0.632

Table 7.2 shows the evaluation results for the baseline classification network with the optimal architectures on the test set. We observe that the classification network test AP is much worse than the validation AP – a 0.302 difference. We have two assumptions to explain this observation:

Assumption i: the test set amateur phoneme segments are very different from those of amateur train and validation sets, and similar to the professional segments.

Assumption ii: the model is heavily overfitted on the train and validation sets.

We have the assumption i because the amateur part of the test set is special, which is recorded by the adult singers. However, the amateur segments in the train and validation sets are recorded mostly by primary school students. To show that the learned overall quality embeddings are not able to discriminate between professional and test set amateur phoneme segments for some phoneme classes, we use t-SNE technique (Van Der Maaten & Hinton, 2008)



(a) Non-voiced consonant phoneme class

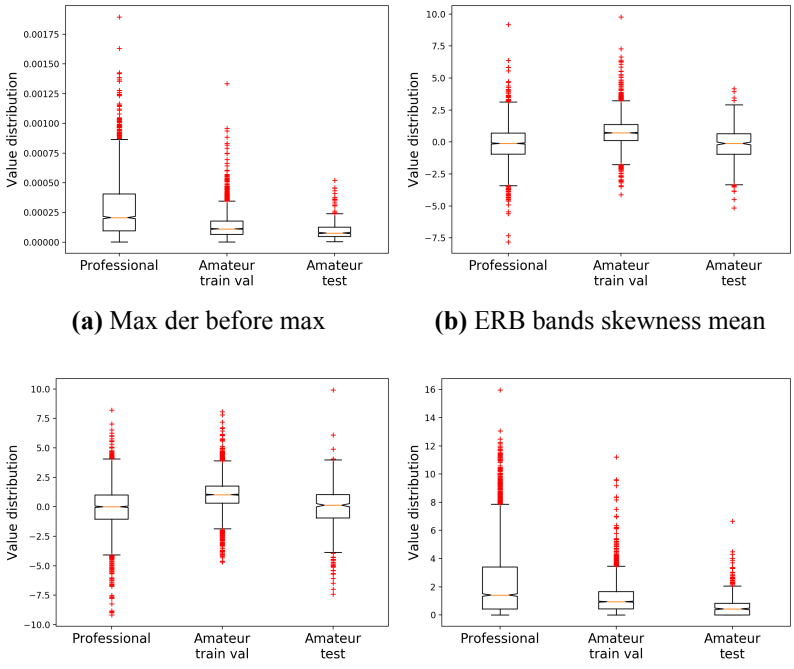
(b) Phoneme O class

Figure 7.2: t-SNE projection of classification network phoneme embeddings for overall quality aspect. nvc: non-voiced consonant; Blue dots: phoneme embeddings of the professional singers; Green triangles: training and validation sets phoneme embeddings of the amateur singers; Red plus: test set phoneme embeddings of the amateur singers.

to project the embeddings of three different groups – professional embeddings, amateur train and validation embeddings and amateur test embeddings into a 2-dimensional space.

Figure 7.2 shows two examples of t-SNE projection for two phoneme classes – non-voiced consonant and phoneme O, of which the test set APs are 0.626 and 0.677. For the non-voiced consonant class, we can't observe three separated groups of phoneme embeddings on figure 7.2a. For the phoneme O class of figure 7.2b, we can observe a clear separation between the professional phoneme segments (blue dots) and the amateur train and validation sets segments (green triangles). However, many amateur test set segments (red plus) are mixed up within the professional cluster.

The mixing up of the test set amateur segments with the professional ones doesn't necessarily mean that these amateur test set phoneme segments have reached the professional singing quality, but perhaps we haven't learned a suitable phoneme embedding to distinguish them. To check if the amateur test segments can be discriminated from the professional segments by using acoustic features, we conduct a feature analysis. We first extract 151 features for the segments of the three groups using Essentia FreesoundEx-



(a) Max der before max

(b) ERB bands skewness mean

(c) Bark bands skewness mean

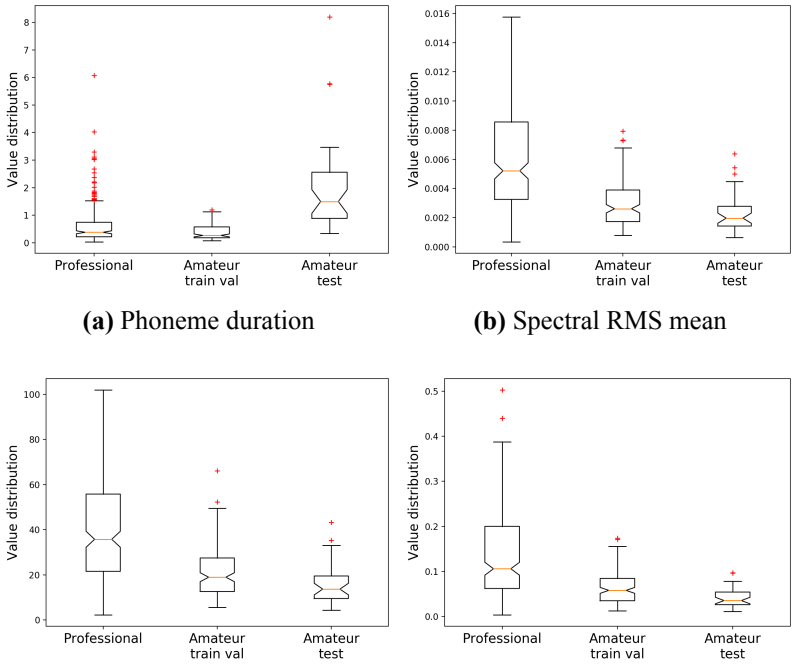
(d) Spectral complexity std

Figure 7.3: Value distributions of the most discriminative features for non-voiced consonant phoneme segments. For the definition of each feature, please consult online[1].

tractor[1]. The feature name list can be checked online[19]. Then we compute ANOVA F-value for each feature, and sort the F-values to select the best individual features which are capable to separate between the three groups (Stoller & Dixon, n.d.). We use `f_classif` function in scikit-learn python package to compute the ANOVA F-value. Figure 7.3 and figure 7.4 shows the value distributions of individual feature for phoneme classes – non-voiced consonant and O.

Figure 7.3 shows that, for the non-voiced phoneme class, no individual feature can separate the amateur test segments from the professional segments. Figure 7.4 indicates that, for the phoneme O class, all these four features can effectively separate the amateur

[1] `http://essentia.upf.edu/documentation/freesound_extractor.html`



(a) Phoneme duration

(b) Spectral RMS mean

(c) Strong decay

(d) Spectral flux mean

Figure 7.4: Value distributions of the most discriminative features for "O" phoneme segments. For the definition of each feature, please consult online[1].

test segments from the professional segments. However, its low test set AP (0.667) indicates that the learned phoneme embeddings are far from discriminative. So it is possible that for such phoneme classes, the learned embeddings are overfitted on the train and validation sets, and not discriminative for the test sets. In the next section, we are going to explore four experiments to overcome the overfitting and to improve the phoneme embedding discriminability.

7.2.5 Techniques to improve the baseline model

We explore four experiments to improve the phoneme embedding:

1. adding **attention** mechanism in the network architecture.
2. using **32 embedding** dimensionality.

3. stacking convolutional layers **CNN** before the RNN layers.
4. adding RNN input and output **dropout**.

Attention: The attention mechanism used in this experiment is the same as it has been presented in Section 6.3.1.

32 embedding: We consider that the embedding dimension, especially for the overall quality aspect (2-dimensional embedding) is too few to capture sufficient information. To address this problem, we insert an intermediate fully-connected layer with 32 linear activation units between the RNN and the output softmax layers. Then we take the 32-dimensional output of this intermediate layer as the embedding.

CNN: The architecture of the convolutional layers is similar as it has been presented in Section 6.3.1. We stack a 6-layers CNN between the input and the RNN layer. Table 7.3 shows the CNN architecture.

Table 7.3: 6-layers CNN, "8x 1×3 ReLU" means 8 kernels of which each convolves on 1 frequency bins and 3 temporal frames, using ReLU activation function.

3 Conv 8x 1×3 ReLU
Max-pooling 1×3
3 Conv 16x 1×3 ReLU
Max-pooling 1×3

Dropout: To overcome the overfitting, we experiment 0.5 rate dropout for both input and output of the RNN.

7.2.6 Results and discussion of the improved model

Figure 7.5 shows the results of the four experiments. For the pronunciation aspect, the attention, CNN or dropout improves the AP. 32 embedding performs worse than the original 29 dimensions, which indicates that increasing the embedding dimensionality cannot always bring the improvement. By combing attention, CNN and dropout, we obtain the best AP 0.753 on the test set (best combination). For the overall quality aspect, the 32 embedding and

CNN improves the AP. However, combine these two architectures cannot bring improvement than using only 32 embedding. Another observation is that dropout failed in improving the overall quality embedding. Thus it probably doesn't help improve the generalization ability of this embedding.

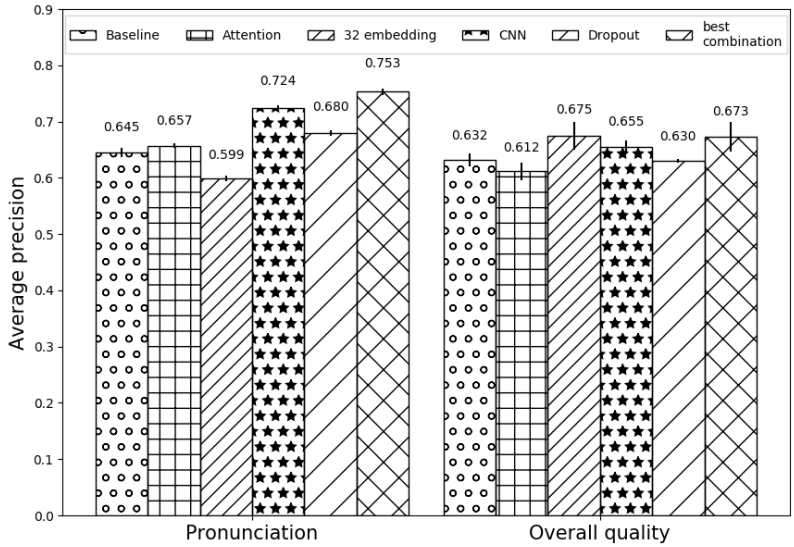
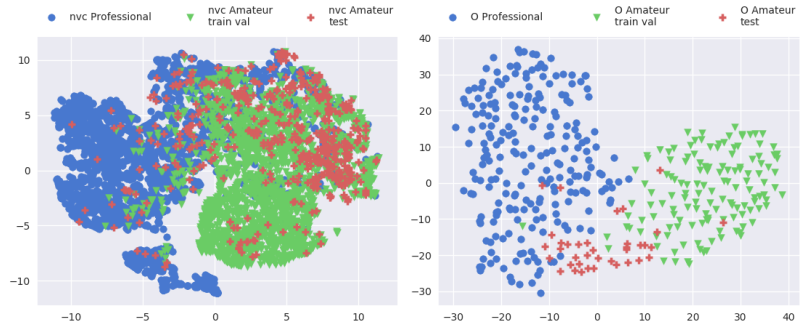


Figure 7.5: Average precision on the test set over 5 runs, using optimal classification networks (baseline), and four architectures to improve the baseline. The best combination of pronunciation aspect is to combine attention, CNN and dropout; that of overall quality aspect is to combine 32 embedding and CNN.

Figure 7.6 shows the t-SNE projection of the 32-dimensional overall quality embeddings. For non-voiced consonant phoneme class, we can observe clearly two separated professional and amateur clusters, and many test set amateur segments are distributed in the train and validation sets amateur cluster.

We can also notice that, for phoneme O class, most of the test set amateur segments are no longer mixed up within the professional cluster, although some segments lie on the border between the professional and the amateur clusters. It worth to notice that the amateur segments in the train and validation sets are entirely composed by the singing samples of the primary school students, while the professional segments are entirely composed by the recordings



(a) Non-voiced consonant phoneme class

(b) Phoneme O class

Figure 7.6: t-SNE projection of classification network 32 dimensional phoneme embeddings for overall quality. nvc: non-voiced consonant; Blue dots: phoneme embeddings of the professional singers; Green triangles: training and validation sets phoneme embeddings of the amateur singers; Red plus: test set phoneme embeddings of the amateur singers.

of the adult singers. The segregation between the amateur test segments and the professional segments means that the model is not overfitted by the age of the singers. Additionally, compared to figure 7.2, 32 dimensional embedding presents a remarkable improvement in separating professional and amateur groups for these two phoneme classes.

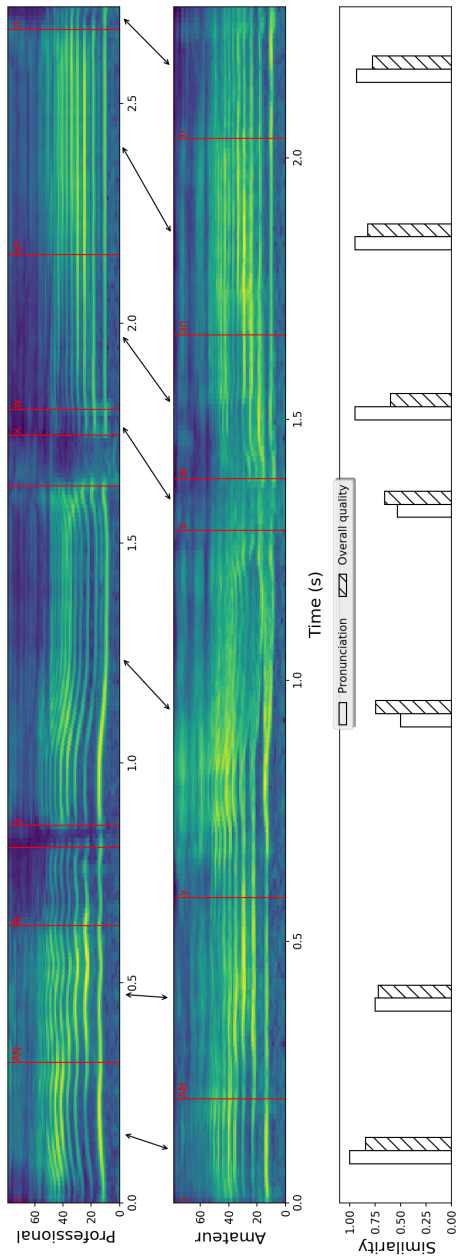


Figure 7.7: Illustration of the similarities of embeddings used in grading the singing phonemes on a singing excerpt of three syllables – yang yu huan. First row: professional singing log-mel spectrogram; Second row: amateur singing log-mel spectrogram; Third row: pronunciation and overall quality similarities by comparing the corresponding professional and amateur phonemes. Vertical red lines: phoneme onsets. Red label following the vertical line is the phoneme name in XSAMPA format.

Figure 7.7 shows the similarity measurement results by using the phoneme embeddings to grade the amateur singing at phoneme-level. We measure the cosine similarity between the professional and amateur corresponding phoneme embeddings, and ignore the extra or missing phonemes, e.g., the third phoneme "N" in the professional singing. The pronunciation and overall quality similarity of each phoneme segment are indicated in the third row of the figure. The similarity measures have a potential application for the automatic assessment of jingju singing voice in online education scenario, where the visualization feedback of the pronunciation and overall quality similarities could be an effective guidance for the students to improve their singing.

7.3 Siamese phoneme embedding networks

Siamese network is a network architecture which receives multiple inputs, and shares the same weights. It uses a contrastive loss to learn the similarity between multiple inputs. This network architecture is more complicated than the classification network. However, it outperforms the classification network in learning speech word acoustic embeddings (Settle & Livescu, 2016). Additionally, the siamese network and the contrastive loss have been initially proposed to learn the similarity between multiple inputs, such as the similarity between images or sounds, which is coherent with the task we are dealing with – to model the pronunciation and overall quality similarities between phonemes. Thus in this section, we explore the performance of the siamese network on singing voice phoneme embedding.

7.3.1 Semi-supervised Siamese network

Figure 7.8 illustrates an example of the siamese network experimented in this work for learning the overall quality embedding.

The network receives three inputs – **anchor**, **same** and **different**. For instance, we can feed a *teacher phoneme class A* sample into the **anchor** input, another *teacher phoneme class A* sample into the **same** input, and a *student phoneme class A* sample into the

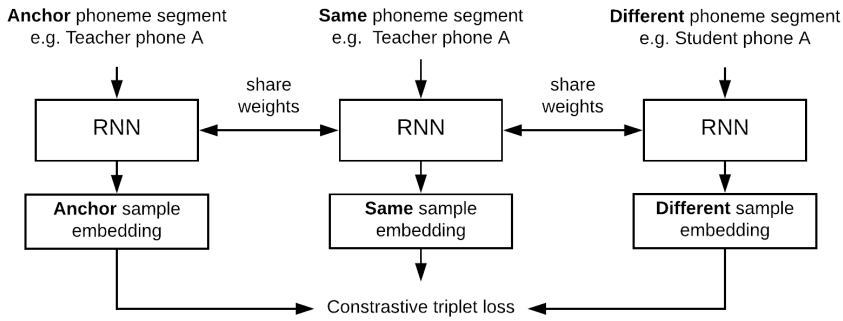


Figure 7.8: Semi-supervised siamese phoneme embedding network example for learning overall quality aspect.

different input. We disclose the teacher or student labels in this example only for clarifying the network training process. However, in the actual case, we don't need to know the exact labels of these samples, but instead the fact that the **anchor** and **same** samples belonging to the same class and the **anchor** and **different** samples belonging to different classes, which is also why we name it the semi-supervised network.

The network outputs are the embeddings for the three input samples – x_a, x_s and x_d. Then we use contrastive triplet loss (also known as cos-hinge triplet loss) to minimize the cosine distance between the **anchor** x_a and **same** x_s embeddings, and maximize the cosine distance between the **anchor** x_a and **different** x_d embeddings. The formula of the contrastive triplet loss is:

$$l = \max\{0, m - d_{cos}(x_a, x_s) - d_{cos}(x_a, x_d)\} \tag{7.1}$$

where $d_{cos}(x_1, x_2)$ is the cosine distance between x_1 and x_2, and m is margin parameter that will be optimized by using the validation set.

To learn the pronunciation embeddings, we only need to feed the network different training samples. For example, a valid training sample combination could be a *phoneme of class A* for the **anchor** input, another *phoneme of class A* for the **same** input, and a *phoneme of class B* for the **different** input.

7.3.2 Model training

We use N **anchor** samples, where N is the sample number of the training set, then randomly choose another N examples which each match the word of a corresponding **anchor** sample, then choose another $5N$ random samples for **different** class to their corresponding **anchor** samples. This training data sampling strategy leads to N training sample buckets, each includes 5 triplet combinations, where each **anchor** and **same** sample pair are repeated 5 times to match with the 5 **different** samples. Then we calculate the contrastive triplet loss for each 5 samples combination, and choose the one with the maximum loss to update the network weights. By doing this, we choose the most similar **different** sample for each **anchor** sample. This sampling strategy is recommended by S. Settle (Settle & Livescu, 2016) through a personal communication. It has been provided by him that this strategy improved the performance of training speech word acoustic embedding.

7.3.3 Experimental setup

We train two phoneme embedding models respectively for pronunciation and overall quality similarities. The optimal architectures are used directly for the evaluation of the siamese network – a 2 layers BiLSTM architecture for pronunciation similarity and a single layer BiLSTM architecture for overall quality similarity. The evaluation procedure and metrics are the same as they have been mentioned in Section 7.2.3. To find the best-performed margin parameter m for the network, we grid search 5 different values of m. Additionally, to test if the network learns useful information, we give the results of the model with randomly initialized weights.

7.3.4 Results and discussion

Table 7.4 shows a much inferior performance on the validation set compared with the classification network embeddings (Table 7.1), and the best-performed margin parameter $m = 0.15$.

Then we show in the Figure 7.1 the results of the siamese network model on the test dataset, along with the baseline classification model and the siamese network model with random weights.

Table 7.4: Mean value of average precision on the validation set over 5 runs, using siamese network with the optimal architectures. m: margin parameter.

m	Pronunciation AP	Overall quality AP
0	0.275	0.507
0.15	**0.354**	**0.511**
0.3	0.332	0.508
0.45	0.323	0.510
0.6	0.279	0.510

We can observe that (1) the classification embedding outperforms the siamese embedding in a large margin; (2) the siamese network with random weights performs equally than the trained siamese network for the overall quality aspect.

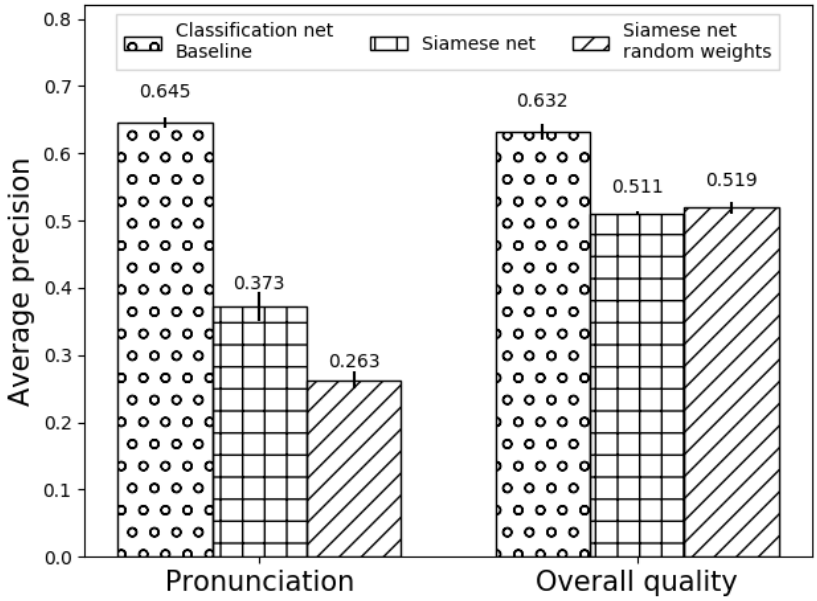


Figure 7.9: Average precision on the test set over 5 runs, using optimal network architectures and margin parameter $m = 0.15$.

The observation (1) is contradicted by the results in paper (Kamper et al., 2016; Settle & Livescu, 2016), where they found that siamese network consistently works better than the classifica-

tion network in learning speech word embedding. A possible reason could be singing voice, especially jingju, is quite different from the speech in terms of the pitch range, spectral variation, syllable or phoneme duration, etc. Another possible reason could be that the training set used for training siamese word embedding (100k word segment pairs) is much larger than our jingju phoneme segment training set (12.5k phoneme segment pairs). We have tried to increase the training set by creating more phoneme segment pairs, however, this makes the training too long to iterate one train-validation loop, which also indicated that siamese network is much hard to train to obtain the equal performance than the classification network. The observation (2) shows that the trained siamese network doesn't learn any useful information for overall embedding. This is contra-intuitive such that we thought that there should have mistakes in the network training, e.g., mistakenly selected **same** or **different** samples. However, for the pronunciation aspect, the trained siamese embedding works better than the random weights one, and they use the same experiment pipeline except for the training data preparation step. We examined the data preparation step, and confirmed that in both pronunciation and overall quality aspects, we fed to the network the correct samples and ground truth labels. Thus, the application of learning a siamese network-based phoneme embedding model needs further study.

7.4 Conclusions

This chapter presented a detailed formulation of the task of pronunciation and overall quality similarities measure in jingju singing voice. The approaches utilized the deep learning-based classification and siamese network models to generate phoneme embeddings, and then calculated the similarity measure for jingju singing phonemes. The evaluation of the specific testing dataset showed the possibility of this approach and its limitations. The work presented in this chapter is evaluated on the manually pre-segmented phoneme segments. Thus a future study needs to be carried out for the joint phoneme segmentation and similarity measure.

We mainly addressed the problem of similarity measurement by using fixed-length phoneme embeddings. The presented method

firstly used a recurrent neural network with the classification objectives to obtain the phoneme embeddings, then calculated the cosine distance between two embeddings as the similarity measure. Additionally, we experimented with several deep learning techniques aiming to improve the model generalization. As the results, the combination of all the techniques was effective in improving the model performance regarding the pronunciation aspect. While expanding the embedding dimension improved the model performance regarding the overall quality aspect prominently. As an exploration, we also tested the siamese phoneme embedding network since it has been designed for the similarity measurement of multiple inputs. However, the performance was much inferior to the classification model, which requires further study.

For future work, to have an overall assessment of the system pipeline, we will evaluate the similarity measure performance by jointly performing automatic phoneme segmentation and similarity measurement models. The next steps would be investigating deeply in the training data preparation, training speed optimization aspects for the siamese network.

Chapter 8

Applications, Summary and Conclusions

This chapter aims to present a concrete application of the singing voice pronunciation assessment and results presented in the previous chapters. The application section is followed by a summary of the work presented in the thesis, along with some key results and conclusions. The last section opens up some open problems and directions for future work.

8.1 Applications

There are quite a few applications for the research work presented in the thesis. Some of these applications have been identified in Chapter 1. This section aims to present concrete examples of such applications and further propose other applications that might be built from the thesis work. This section also describes in detail one application that has resulted from the work in MusicCritic[1] project. Possible future applications are also discussed briefly.

The primary objective and application of the methodologies related to automatic singing voice assessment – syllable and phoneme segmentation, mispronunciation detection and pronunciation and overall similarity measures, is to use them for online or classroom

[1] https://musiccritic.upf.edu/

singing voice training. Additionally, there are many ways to use the methodologies developed in this thesis for various applications.

Jingju singing exercises can be organized into melodic lines in an online jingju singing teaching course. The teacher's singing recordings are provided in the exercise as the demonstrative singing example. The students who take this exercise are required to imitate the teacher's singing. To automatically assess the imitative singing of the students, the assessment tools can segment the student's recording into syllable and phoneme units, detect the mispronunciation for special pronunciation and jiantuanzi syllables, and measure the pronunciation and overall quality similarities between teacher's and student's singing phonemes. The assessment results can serve as initial guidance for the students to improve their singing skills in the absence of a singing instructor.

The methods developed in this thesis also have the potential to be applied in the classroom jingju singing training scenario. As it has been mentioned in Section 3.1, jingju singing is taught between teacher and student by using the oral teaching and face-to-face methods. The students need to understand the teacher's verbal or singing feedbacks firstly, assimilate them, then do a lot of singing practice to improve their singing skills. However, the teacher's verbal feedback is usually mixed with many personal comments (please check the teacher's feedbacks in correction occurrence in Section 3.1.1), and thus cannot describe the student's singing problems in an objective and precise way. The syllable and phoneme segmentation method developed in our thesis can automatically segment and label the teacher's singing melodic line into syllable or phoneme units. With the help of some singing voice or speech visualization technologies, such as pitch, loudness tracking, formant frequency tracking, the students could better assimilate the teacher's verbal feedback by benefiting from the visual cues of their singing voice segmented into syllable and phoneme units.

Automatic singing voice assessment technologies find their application in helping navigate through jingju singing voice recordings and in content-based music retrieval. Applications such as search by pronunciation traits can be conceived, such as query by mispronunciation rate, query by pronunciation similarity. Additionally, the automatic syllable and phoneme segmentation method applied on the jingju singing recordings allows a clear visualiza-

tion of the syllable/phoneme boundaries and labels, which could be applied to a semantic exploration of the singing corpus.

Musicologists working with jingju pronunciation would benefit from the corpora and tools developed in this thesis. The jingju a cappella singing voice datasets are representative and well-curated with useful metadata, annotations, and can be used to derive musicological findings. Automatic syllable and phoneme segmentation tool can lead to a precise segmentation of the singing syllable/phoneme units and hence to analyze large corpora of recordings, which would be otherwise time-consuming if done manually. The mispronunciation detection tool is useful to annotate automatically the pronunciation correctness of the singing recordings at syllable level, which could be interesting to the musicologists who study the pronunciation trait of the jingju singing.

To conclude, one specific application built with MusicCritic project is described below – sofège assessment. This application is the collaborative effort of the MusicCritic team. A brief introduction to the application is provided, and then we put stress on how the pronunciation assessment methods developed in this thesis applied and integrated into this application.

8.1.1 MusicCritic solfège assessment

MusicCritic is a music performance technology with which to evaluate musical exercises sung or played by students, giving meaningful feedback. It is a service that uses the Basic LTI standard[2] and can be easily integrated into online applications or education platforms, such as Coursera[3] and Kadenze[4]. It contains four sub-technologies – solfège, melodic imitation, chord playing and improvisation assessment, developed collaboratively by the researchers and developers in MTG. The solfège assessment tool stems from the automatic singing voice segmentation and assessment methods conceived in this thesis.

MusicCritic solfège assessment tool can receive the student's solfège singing recording, then return the pitch, rhythm and pro-

[2] http://www.imsglobal.org/activity/learning-tools-interoperability

[3] https://www.coursera.org/

[4] https://www.kadenze.com/

nunciation feedback visually and automatically. It also generates the pitch and pronunciation assessment scores for the student's recording. With the help of the MusicCritic LTI standard integration, the solfège assessment tool can be easily integrated into online applications or education platforms that support this standard.

The research results from this thesis on singing voice assessment are partly integrated into MusicCritic solfège assessment tool. A Kaldi-based syllable recognition system extended from the mispronunciation detection baseline presented in Section 6.2 is used to recognize the solfège syllable and detect its boundaries. The pitch, rhythm and pronunciation accuracy visualization is done based on the recognition results.

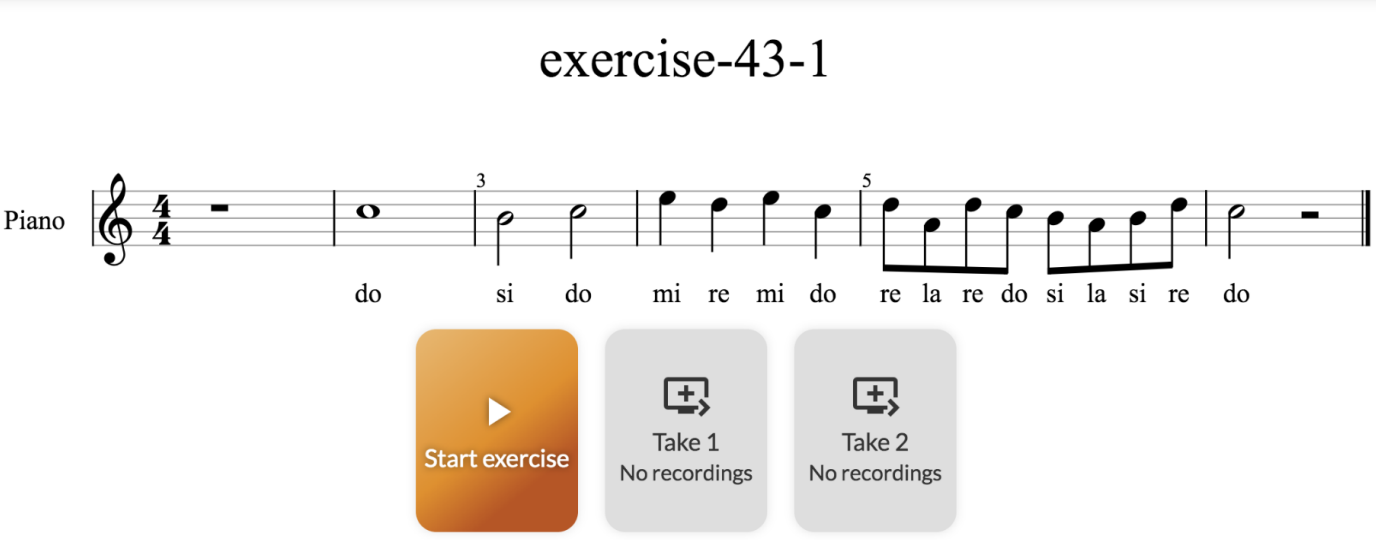


Figure 8.1: A screenshot of the recording page of the solège assessment tool. The student can listen the demonstrative singing by clicking "Start exercise", then record their singing twice by clicking "Take 1" and "Take 2" buttons.

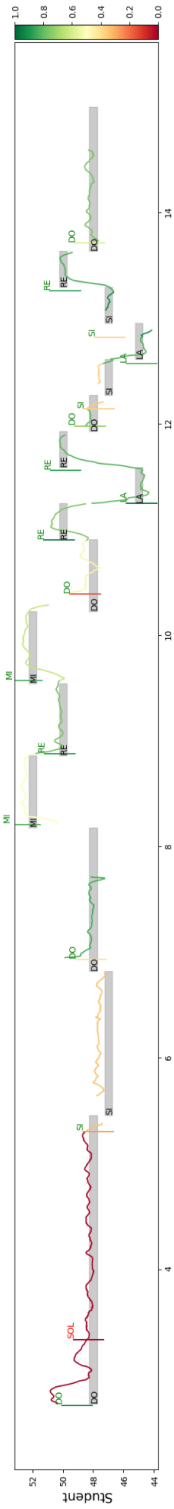


Figure 8.2: The visualzation of a student's recording. The rectangles in grey are the ground truth solfège notes of which the name is labeled at the beginning of each note. The curve indicates the pitch. The vertical lines before each detected notes indicate the note onset positions. The character on top of the vertical line indicates the detected note name. The singing quality is indicated by the color system – green: good performance, red: bad performance.

Figure 8.1 shows a recording interface of the solfège assessment tool, where the student can listen to the demonstrative singing recording, then record and submit their singing voice to the assessment tool. Figure 8.2 shows the assessment result visualization automatically generated by the tool on which the singing quality of three aspects – pitch, note onset and pronunciation, are visualized by the color system.

8.2 Contributions

A summary of the specific contributions from the work presented in the thesis is listed below.

8.2.1 Contributions to creating research corpora and datasets

Building research corpora for MIR is one of the primary tasks of CompMusic project. Significant efforts have been put into building research corpora and datasets. The relevant datasets are listed below. The link to access all these datasets are provided in Appendix D.

- Automatic syllable and phoneme segmentation (ASPS) dataset (Section 4.2.1): This dataset has two sub-datasets ($ASPS_1$ and $ASPS_2$) which contain in total 197 jingju singing recordings. Each recording in $ASPS_1$ is annotated with syllable and phoneme boundaries and labels. While the recordings in $ASPS_2$ are annotated at syllable-level. The annotation is done by the author with the support of Rafeal Caro Repetto and Yile Yang.

- Mispronunciation detection (MD) dataset (Section 4.2.2): Special pronunciation, jiantuanzi syllables and the mispronunciation labels for these two types of the syllable are annotated for 1007 jingju singing melodic lines in this dataset.

- Pronunciation and overall quality similarity measures (POQSM) dataset (Section 4.2.3): This dataset contains 19911 jingju singing phoneme segments recorded by both professional and

amateur singers. These segments are split deliberately to build pronunciation and overall quality similarity models.

8.2.2 Technical and scientific contributions

- Identification of the critical role of pronunciation in jingju singing training (Section 3.1).
- Identification of challenges, opportunities and applications of automatic singing voice pronunciation assessment of jingju music (Section 3.2).
- Identification of four problems which are the most relevant to automatic singing voice pronunciation assessment of jingju music, along with a review of the state-of-the-art methods (Section 3.3).
- Formulation the problems of automatic syllable and phoneme segmentation, mispronunciation detection, and pronunciation and overall quality similarity measures in jingju music (Section 3.4).
- An evaluation of the jingju a cappella singing corpus based on the methodology by Serra (Serra, 2014) (Section 4.1).
- A demonstration of the utility of corpus for musicological analysis – melodic line, syllable and phoneme duration analysis (Section 4.1.2).
- Duration-informed syllable and phoneme segmentation for jingju singing. Developing approaches that combine the feature learning power of the deep learning models and the inference of the syllable or phoneme onset positions by using the coarse duration information explicitly. New onset selection HMM model is proposed, which shows improvement in both syllable or phoneme onset detection and segmentation (Chapter 5).
- Demonstration of the validity of applying deep learning-based classification models in mispronunciation detection of jingju singing syllables (Chapter 6).

- Phoneme embedding-based approaches for measuring pronunciation and overall quality similarities at phoneme-level (Chapter 7).

8.3 Summary and conclusions

In this section, we present a summary, conclusions and the key results from the thesis, organized based on the chapters of the thesis. Broadly, the thesis aimed to build culture-specific data-driven MIR approaches using deep learning and probabilistic models for automatic pronunciation assessment in jingju music, focusing mainly on the tasks of syllable and phoneme segmentation, mispronunciation detection and pronunciation and overall quality similarity measures. Such approaches would lead to tools and technologies that can improve our experience of jingju singing training, within the context of jingju music culture. The applications lie in computer-aided jingju singing training, music navigation, content-based music retrieval and musicology, and as pre-processing steps for MIR tasks extracting semantic information such as singing syllable and phoneme.

This thesis focused on automatic singing voice pronunciation assessment tasks within the scope of CompMusic project, which is limited to developing data collections and computational models for singing voice pronunciation assessment in jingju music.

An introduction to singing in jingju art music was presented in Chapter 2 with the concentration on jingju singing pronunciation concepts. The introduction provided a background to music concepts encountered in this thesis. A comparison of the pronunciation characteristics between jingju singing and Western opera singing showed the contrasting differences between two singing genres. A review of state of the art in automatic singing voice pronunciation assessment-related tasks provided a basis for understanding relevant methodologies in jingju singing.

Chapter 3 identified the critical role of pronunciation in jingju singing training, and thus justified the pronunciation assessment as the main focus of this thesis. Additionally, this chapter identified some of the challenges and opportunities of jingju singing pronunciation assessment. Important and relevant research problems in

the assessment of singing voice pronunciation of jingju music were identified and described, which will be useful to a researcher who is looking to solve relevant problems in this area of research.

Four core research problems – building data corpus, syllable and phoneme segmentation, mispronunciation detection, and pronunciation and overall quality similarity measures, are formulated.

The problem of creating research corpora and datasets for data-driven MIR research, addressed in Chapter 4 shows that significant efforts to build relevant datasets for automatic singing voice assessment research. The corpora and datasets are built and evaluated according to a set of corpus design criteria and methodologies. Furthermore, both corpus-level and dataset-level data analysis and visualization were conducted to draw some musicological inferences. To promote the idea of open data and reproducible research, the presented corpus, test datasets including audio recordings, metadata and annotations are openly available through Zenodo.org.

Automatic syllable and phoneme segmentation were one of the main problems addressed in this thesis. Chapter 5 presented a comprehensive methodology of syllable/phoneme onset detection and segmentation for jingju singing voice. The experiment on the baseline HSMM-based method showed an unideal performance on syllable/phoneme onset detection or segmentation tasks, indicating the need for a better method that can use the a priori coarse syllable/phoneme duration information. The duration-informed syllable/phoneme detection-based segmentation method that allows incorporating the coarse durations into the syllable and phoneme decoding process.

An evaluation of the duration-informed onset detection-based segmentation method clearly showed the improvement in both onset detection and segmentation tasks, compared with HSMM-based method. Additionally, the HMM onset selection model used in the proposed method allows a faster inference than the HSMM-based method. The segmentation performance depends on the goodness of the onset detection function. An exploration of various onset detection functions generated by different deep learning architectures showed that the onset detection function outputted from a basic convolutional architecture can achieve the state of the art segmentation accuracy, and is more efficient than other more complicated architectures. Lastly, the proposed segmentation model is capable

of generalizing to other singing voice in languages different from Mandarin Chinese since its language-independency.

The mispronunciation detection problem for jingju singing was addressed in Chapter 6. The task scope was constrained to the syllable-level special pronunciation and jiantuanzi mispronunciation detection because they are the primary source of the mispronounced syllables in jingju singing training. A forced-alignment baseline method and the proposed discriminative model-based method were experimented to tackle this problem.

The baseline method used a pronunciation dictionary with multiple pronunciations for each syllable entry. The mispronunciation detection worked as that the model decoding algorithm select the pronunciation which is matched best to the acoustics of the singing. The experiment on the baseline method showed a mediocre performance on special pronunciation syllable mispronunciation detection, and an unsatisfactory performance on jiantuanzi syllable mispronunciation detection.

The proposed method used firstly the syllable segmentation method presented in Chapter 5 to segment the singing recording into the syllable units, then used deep learning-based discriminative models to classify binarily the syllable units into mispronunciation or correct pronunciation class. Several deep learning techniques and two new architectures were experimented to augment the classification and generalization abilities. The results showed that the discriminative model-based method improved the mispronunciation detection accuracy on jiantuanzi syllables. The additional visualization of the attention vector showed that the attention mechanism worked well in making the classification decision based on certain essential time regions of a syllable.

A framework for measuring pronunciation and overall quality similarities for jingju singing phoneme, along with an exploratory experiment was the subject matter of Chapter 7. Utilizing the phoneme embedding allows us to convert the variable-length syllable segments into fixed-length embedding vectors. The similarity measure can then be obtained by calculating the distance metric between two phoneme embeddings.

An RNN-based classification model was proposed to generate the phoneme embedding. The model predicted the input phoneme segment into (1) different phonetic categories and (2) professional

or amateur singing categories. The penultimate layer output was taken as the phoneme embedding as it was expected to embed the (1) phonetic and (2) overall quality information of the phoneme segment. Various deep learning techniques were experimented to improve the quality of the similarity measure and the generalization ability of the model. As an exploratory experiment, a siamese network which is designed originally for measuring the similarity between multiple inputs was tested. However, the performance was much worse than the classification model, and further experiments is needed to suggest improvements.

8.4 Future directions

There are several directions for future work based on the thesis. One of the goal of the thesis was to present relevant research problems in pronunciation assessment of jingju singing voice. Some of these problems presented in Section 3.3 are a good start to extend the work presented in the thesis. Several tasks for jingju singing voice pronunciation assessment were proposed, while only a part of them was addressed in this thesis. The problems such as banshi (metrical structure) segmentation, melodic line segmentation, and automatic intonation or rhythm assessment have received little attention from the research community so far.

Automatic singing voice assessment of jingju music is a very board topic, and the assessment can be done in several musical dimensions, such as intonation, rhythm, loudness and pronunciation. The musical dimension which has been addressed extensively is pronunciation. The automatic assessment methods of the other musical dimensions related to jingju singing are worth to be explored in furture. Besides, the work in the thesis used mainly audio recordings along with syllable/phoneme duration and pronunciation annotations to develop computational models for the assessment. However, using additional information such as score, editorial metadata may lead to better automatic assessment models.

The curated a cappella jingju singing voice corpus and test data provide an opportunity to be used for a variety of research problems in future, such as jingju Query-by-singing, jingju singing transcription and jingju singing synthesis. The research corpus evolves. Im-

proving the research corpus and building additional datasets for automatic jingju singing assessment are important tasks for the future. The use of the corpus and the datasets for musicological research was hinted in the thesis. However, a rigorous study of the suitability of the corpus and datasets for musicology, and comprehensive musicological research using the corpus is one direction to pursue in future.

Syllable and phoneme segmentation task was addressed in detail in the thesis. However, several open questions still need more exploration. The presented onset detection-based segmentation method can be extended by incorporating more side information other than duration, such as linguistic information. While the performance of the baseline lyrics-to-audio alignment method can be improved by including syllable/phoneme duration or onset information. The current onset detection-based segmentation method can not deal with the situation that missing or extra syllables are sung in the recording. To develop a recognition-based method for the segmentation is a path to explore in future.

The mispronunciation detection task was addressed in the thesis with a preliminary result presented on a small dataset. The discriminative model-based detection method used deep learning architectures which require a large of training data to outperform the forced alignment-based method. To expand the amount of the training set by collecting more singing recordings, and reevaluate these two methods is a work to be done in future.

The tasks of phoneme segmentation and pronunciation similarity measure were addressed as independent tasks in the thesis. An overall assessment of the similarity measurement pipeline requires to combine these two tasks. Additionally, the experiment results of using the siamese network in similarity measure are very preliminary. Extensive experiments to study this architecture including accelerating the model training, studying different data preparation methods need to be done in future.

Integration of these algorithms into practical applications requires additional effort. The evaluation of the validity of the integrated applications needs to be conducted in the real jingju singing training scenario. In the future, an integration of all the described singing voice pronunciation assessment approaches into one application is necessary and it helps to improve the algorithms through

user feedback.

Appendix A

The sounds in Mandarin Chinese

pinyin spelling	I.P.A symbols	X-SAMPA symbols
b	p	p
d	t	t
z	ts	ts
j	tɕ	t_s\
zh	tʂ	ts‘
g	k	k
p	p^h	p_h
t	t^h	t_h
c	ts^h	ts_h
q	$tɕ^h$	t_s_h
ch	$tʂ^h$	ts‘_h
k	k^h	k_h
f	f	f
s	s	s
x	ɕ	s\
sh	ʂ	s‘
r	ɻ	r\‘

Table A.1: Initial consonants

pinyin spelling	I.P.A symbols	X-SAMPA symbols
h	x	x
m	m	m
n	n	n
l	l	l

Table A.1: Initial consonants (continued)

pinyin spelling	I.P.A symbols	X-SAMPA symbols
n	n	n
ng	ŋ	N

Table A.2: Terminal consonants (nasal finals)

pinyin spelling	I.P.A symbols	X-SAMPA symbols
a	a	a_”*
ia	ja	ja_”*
ua	wa	wa_”*
an	an	an
ian	jɛn	jEn
uan	wan	wan
üan	yɛn	yEn
en	ən	@n
in	in	in
uen (un)	wən	w@n
ün	yn	yn
ang	ɑŋ	AN
iang	jɑŋ	jAN
uang	wɑŋ	wAN
eng	əŋ	EN˙

Table A.3: Final vowels

pinyin spelling	I.P.A symbols	X-SAMPA symbols
ong	ʊŋ	UN
ing	iŋ	iN
iong	jʊŋ	jUN
ueng	wəŋ	wEN
e	ɤ	7
o	ɔ	O
uo	wɔ	wO
ai	ai	aI_^*
uai	wai	waI_^*
ao	ɑʊ	AU_^*
iao	jɑʊ	jAU_^*
ou	oʊ	oU_^*
iou (iu)	joʊ	joU_^*
i	i	i
ü	y	y
zhi	tʂɻ̩†	ts‘1
chi	ʈʂʰɻ̩†	ts‘_h1
shi	ʂɻ̩†	s‘1
ri	ɻɻ̩†	r\‘1
zi	tsɹ̩†	tsM
ci	tsʰɹ̩†	ts_hM
si	sɹ̩†	sM
u	u	w
ê	ɛ	E
ie	jɛ	jE
üe	yɛ	yE
ei	ei	eI_^*
uei (ui)	wei	weI_^*

Table A.3: Final vowels (continued)

*The X-SAMPA symbols a_”, I_^ and U_^ are annotated respectively as a”, I^and U^in the phonetic annotation of the dataset used in this dissertation for the simplicity.

•The X-SAMPA symbol En is annotated as 7N in the phonetic annotation of the datasets used in this dissertation.

†These are not final vowels, since they consist of both an initial

consonant and a vowel. However, the vowels ɿ and ʅ occur in Mandarin Chinese only when preceded by these consonants.

Appendix B

Special pronunciations and jianzi

Mandarin pronunciation	Special pronunciation
bei	be
bai	be
zei	ze
mai	me
feng	fong
meng	mong
peng	pong
peng	pen
sheng	shen
bing	bin
ting	tin
qing	qin
ping	pin
jing	jin
ling	lin
ming	min

Table B.1: Mandarin pronunciations and their special pronunciations in pinyin format in MD test dataset Section 4.2.2.

Mandarin pronunciation	Special pronunciation
ning	nin
ding	din
ying	yin
xing	xin
zeng	zen
ceng	cen
cheng	chen
zheng	zhen
neng	nen
meng	men
chang	chan
zhang	zhan
deng	den
zhuang	zhuan
hai	huan
yuan	ywan
yuan	yoan
quan	qwan
chun	chün
zheng	zhang
zhan	zhang
ji	jin
ai	ngai
an	ngan
e	ngo
wo	ngo
wu	ngo
luo	nuo
zhao	zhuo
ge	guo
na	nuo

Table B.1: Mandarin pronunciations and their special pronunciations in pinyin format in MD test dataset Section 4.2.2. (continued)

Mandarin pronunciation	Special pronunciation
ke	ko
que	qüo
ge	go
he	ho
me	mo
nei	nuei
jie	jiai
hai	xiai
xie	xiai
lei	luei
zhu	zhi
zhi	zhri
shi	shri
ru	rü
chu	chü
chu	chrü
zhu	zhü
zhu	zhru
shu	shru
chu	chru
mao	miou
lü	lu
wo	ngou
zhe	zhre
shuo	shüe
lai	nai
fei	fi
ri	ri \i\

Table B.1: Mandarin pronunciations and their special pronunciations in pinyin format in MD test dataset Section 4.2.2. (continued)

Mandarin pronunciation	Special pronunciation
ji	zi
jue	zue
qian	cian
qiu	ciu
qing	cing
qie	cie
xi	si
xiao	siao
xian	sian
xiang	siang
xiu	siu
xin	sin
zheng	zen
zheng	zeng
chu	cu

Table B.2: Mandarin pronunciations and their jianzi in pinyin format in MD test dataset Section 4.2.2.

Appendix C

List of Publications

First author full articles in peer-reviewed conferences

- Gong, R., Cuvillier, P., Obin, N., & Cont, A. (2015, September). Real-time audio-to-score alignment of singing voice based on melody and lyric information. *In Interspeech 2015*; Dresden, Germany.
https://hal.archives-ouvertes.fr/hal-01164550

- Gong, R., Yang, Y., & Serra, X. (2016). Pitch contour segmentation for computer-aided jinju singing training. *13th Sound & Music Computing Conference*; 2016 Aug 31-Sep 3; Hamburg, Germany.
http://mtg.upf.edu/node/3537

- Gong, R., Obin, N., Dzhambazov, G. B., & Serra, X. (2017). Score-informed syllable segmentation for jingju a cappella singing voice with mel-frequency intensity profiles. *In Proceedings of the 7th International Workshop on Folk Music Analysis*; 2017 Jun 14-16; Málaga, Spain.
https://doi.org/10.5281/zenodo.556820

- Gong, R., Pons, J., & Serra, X. (2017). Audio to score matching by combining phonetic and duration information. *In ISMIR*; 2017 Sep; Suzhou, China.
https://arxiv.org/abs/1707.03547

- Gong, R., Repetto, R. C., & Serra, X. (2017, October). Creating an A Cappella Singing Audio Dataset for Automatic Jingju Singing Evaluation Research. *In Proceedings of the 4th International Workshop on Digital Libraries for Musicology*; 2017 Sep; Shanghai, China.
https://doi.org/10.1145/3144749.3144757

- Gong, R., & Serra, X. (2017). Identification of potential Music Information Retrieval technologies for computer-aided jingju singing training. *In Chinese traditional music technology session - China conference on sound and music technology*; 2017 Sep; Suzhou, China.
https://arxiv.org/abs/1711.07551

- Gong, R., & Serra, X. (2018). Singing voice phoneme segmentation by hierarchically inferring syllable and phoneme onset positions. *In Interspeech 2018*; Hyderabad, India.
https://arxiv.org/abs/1806.01665

Second author full articles in peer-reviewed conferences

- Caro Repetto, R., Gong, R., Kroher, N., & Serra, X. (2015). Comparision of the singing style of two jingju schools. *16th International Society for Music Information Retrieval Conference*; 2015 Oct 26-30; Málaga, Spain.
http://mtg.upf.edu/node/3317

- Fonseca, E., Gong, R., Bogdanov, D., Slizovskaia, O., Gómez Gutiérrez, E., & Serra, X. (2017). Acoustic scene classification by ensembling gradient boosting machine and convolutional neural networks. *Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*; 2017 Nov 16; Munich, Germany.
http://hdl.handle.net/10230/33454

- Pons, J., Gong, R., & Serra, X. (2017). Score-informed syllable segmentation for a cappella singing voice with convolutional neural networks. *In ISMIR 2017*; 2017 Sep; Suzhou, China.
https://arxiv.org/abs/1707.03544

- Fonseca, E., Gong, R., & Serra, X. (2018). A Simple Fusion of Deep and Shallow Learning for Acoustic Scene Classification. *15th Sound & Music Computing Conference*; 2018 July; Limassol, Cyprus.
  https://arxiv.org/abs/1806.07506

Third author full articles in peer-reviewed conferences

- Pons, J., Slizovskaia, O., Gong, R., Gómez, E., & Serra, X. (2017, August). Timbre analysis of music audio signals with convolutional neural networks. *In EUSIPCO 2017*; Kos, Greece.
  https://arxiv.org/abs/1703.06697

Appendix D

Resources

This appendix is a compendium of links to resources and additional material related to the work presented in the thesis. An up-to-date set of links is also listed and maintained on the companion webpage http://compmusic.upf.edu/phd-thesis-rgong.

Some of the results not reported in the dissertation are presented on the companion webpage. The companion webpage will also be updated with any additional resources and material that will be built in the future.

Corpora and datasets

Access to the corpora and datasets will be through the Zenodo.org and MusicBrainz.org

Research corpus

Jingju a cappella singing voice dataset part 1

https://doi.org/10.5281/zenodo.780559

Jingju a cappella singing voice dataset part 2

https://doi.org/10.5281/zenodo.842229

Jingju a cappella singing voice dataset part 2

https://doi.org/10.5281/zenodo.1244732

Jingju a cappella singing voice dataset metadata on MusicBrainz
https://musicbrainz.org/search?query=MTG-UPF&type=release&method=indexed

Test dataset

Automatic syllable and phoneme segmentation test dataset part 1 – $ASPS_1$
https://doi.org/10.5281/zenodo.1185123

Automatic syllable and phoneme segmentation test dataset part 2 – $ASPS_2$
https://doi.org/10.5281/zenodo.1341070

Pronunciation and overall quality similarity measures test dataset – POQSM
https://doi.org/10.5281/zenodo.1287251

Code

The links to code related to the thesis are listed. Up-to-date links to code (including future releases) will be available on: https://github.com/ronggong

Automatic syllable and phoneme segmentation baseline code
https://github.com/ronggong/interspeech2018_submission01

Automatic syllable and phoneme segmentation onset detection function improvement code
https://github.com/ronggong/musical-onset-efficient

Mispronunciation detection code
https://github.com/ronggong/mispronunciation-detection

Pronunciation and overall quality similarity measures code
https://github.com/ronggong/DLfM2018

Bibliography

The numbers in brackets at the end of each bibliographic entry indicate the pages in which it is cited.

Almpanidis, G., Kotti, M., & Kotropoulos, C. (2009, feb). Robust Detection of Phone Boundaries Using Model Selection Criteria With Few Observations. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(2), 287–298. doi: 10.1109/TASL.2008.2009162 [43, 44]

Bahdanau, D., Cho, K., & Bengio, Y. (2014, sep). Neural Machine Translation by Jointly Learning to Align and Translate. Retrieved from http://arxiv.org/abs/1409.0473 [156, 157]

Bai, S., Kolter, J. Z., & Koltun, V. (2018, mar). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *eprint arXiv:1803.01271*. Retrieved from http://arxiv.org/abs/1803.01271 [161]

Bello, J., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. (2005, sep). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 1035–1047. doi: 10.1109/TSA.2005.851998 [39, 41]

Bengio, Y. (2012). Deep Learning of Representations for Unsupervised and Transfer Learning. In *Jmlr: Workshop and conference proceedings* (Vol. 27, pp. 17–37). Retrieved from http://dl.acm.org/citation.cfm?id=3045796.3045800 [53]

Bengio, Y., Boulanger-Lewandowski, N., Pascanu, R., & Montreal, U. (2012). *Advances in Optimizing Recurrent Networks* (Tech. Rep.). Retrieved from https://arxiv.org/pdf/1212.0901v2.pdf [59]

Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., & Lamere, P. (2011). The Million Song Dataset. In *Ismir*. Retrieved from http://ismir2011.ismir.net/papers/OS6-1.pdf [100]

Black, D. A., Li, M., & Tian, M. (2014). Automatic identification of emotional cues in Chinese opera singing. In *13th int. conf. on music perception and cognition (icmpc-2014)*. [103, 105]

Black, D. A. A., Li, M., & Tian, M. (2014). Automatic identification of emotional cues in {Chinese} opera singing. In *Icmpc*. Seoul, South Korea. [87]

Böck, S., Arzt, A., Krebs, F., & Schedl, M. (2012). Online Real-time Onset Detection with Recurrent Neural Networks. In *the 15th international conference on digital audio effects (dafx-12)*. Retrieved from https://www.dafx12.york.ac.uk/papers/dafx12_submission_4.pdf [39, 42, 50]

Böck, S., Böck, S., & Schedl, M. (2011). Enhanced Beat Tracking with Context-Aware Neural Networks. In *The 14th international conference on digital audio effects (dafx-11)*. [88]

Böck, S., Korzeniowski, F., Schlüter, J., Krebs, F., & Widmer, G. (2016, may). Madmom: a new Python Audio and Music Signal Processing Library. Retrieved from http://arxiv.org/abs/1605.07008 [127, 144]

Böck, S., Krebs, F., & Schedl, M. (2012). Evaluating the Online Capabilities of Onset Detection Methods. In *Ismir*. Retrieved from http://ismir2012.ismir.net/event/papers/049_ISMIR_2012.pdf [39, 40, 41, 50, 144]

Böck, S., Krebs, F., & Widmer, G. (2016). Joint Beat and Downbeat Tracking with Recurrent Neural Networks. In *Ismir*. Retrieved from http://www.cp.jku.at/research/papers/Boeck_etal_ISMIR_2016.pdf [88]

Böck, S., & Widmer, G. (2013a). Local Group Delay Based Vibrato and Tremolo Suppression for Onset Detection. In *Ismir*. Retrieved from http://www.cp.jku.at/research/papers/Boeck_Widmer_Ismir_2013.pdf [39, 42, 50]

Böck, S., & Widmer, G. (2013b). Maximum Filter Vibrato Suppression For Onset Detection. In *the 16th international conference on digital audio effects (dafx-13)*. Retrieved from http://dafx13.nuim.ie/papers/09.dafx2013_submission_12.pdf [39, 42, 50]

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345. Retrieved from http://hdl.handle.net/11245/1.200596 [105]

Bozkurt, B., Baysal, O., & Yüret, D. (2017). A Dataset and Baseline System for Singing Voice Assessment. In *Cmmr*. Retrieved from http://cmmr2017.inesctec.pt/wp-content/uploads/2017/09/43_CMMR_2017_paper_31.pdf [31, 33, 36]

Brognaux, S., & Drugman, T. (2016). {HMM}-based speech segmentation: Improvements of fully automatic approaches. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *24*(1), 5–15. doi: 10.1109/TASLP.2015.2456421 [43, 44, 129]

Cao, C., Li, M., Liu, J., & Yan, Y. (2008). An Objective Singing Evaluation Approach by Relating Acoustic Measurements to Perceptual Ratings. In *Interspeech*. doi: https://doi.org/10.1016/S0892-1997(98)80038-6 [31, 32, 35]

Chang, S., & Lee, K. (2017). Lyrics-to-Audio Alignment by Unsupervised Discovery of Repetitive Patterns in Vowel Acoustics. *IEEE Access*, *5*, 16635–16648. doi: 10.1109/ACCESS.2017.2738558 [45, 46]

Chen, C. (2016). An Onset Detection Algorithm by ODF Fusion. In *Mirex 2016 audio onset detection*. Retrieved from http://www.music-ir.org/mirex/abstracts/2016/CC3.pdf [39, 42]

Chien, Y. R., Wang, H. M., & Jeng, S. K. (2016, nov). Alignment of Lyrics With Accompanied Singing Audio Based on Acoustic-Phonetic Vowel Likelihood Modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(11), 1998–2008. doi: 10.1109/TASLP.2016.2594282 [45, 46]

Cho, K., Courville, A., & Bengio, Y. (2015, jul). Describing Multimedia Content using Attention-based Encoder–Decoder Networks.

doi: 10.1109/TMM.2015.2477044 [157]

Chung, Y.-A., & Glass, J. (2018). Speech2Vec: A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech. In *Interspeech*. Retrieved from https://arxiv.org/pdf/1803.08976.pdf [48, 49]

Chung, Y.-A., Wu, C.-C., Shen, C.-H., Lee, H.-Y., & Lee, L.-S. (2016). Audio Word2Vec: Unsupervised Learning of Audio Segment Representations using Sequence-to-sequence Autoencoder. In *Interspeech*. Retrieved from https://arxiv.org/pdf/1603.00982.pdf [48, 49]

Clevert, D.-A., Unterthiner, T., & Hochreiter, S. (2015, nov). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). Retrieved from http://arxiv.org/abs/1511.07289 [55]

Daido, R., Ito, M., Makino, S., & Ito, A. (2014, mar). Automatic evaluation of singing enthusiasm for karaoke. *Computer Speech & Language*, *28*(2), 501–517. doi: 10.1016/j.csl.2012.07.007 [32, 35]

Davis, J., & Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. In *the 23rd international conference on machine learning*. Retrieved from https://www.biostat.wisc.edu/~page/rocpr.pdf doi: 10.1145/1143844.1143874 [51, 171]

Dean, J. (n.d.). *Large-Scale Deep Learning for Intelligent Computer Systems* (Tech. Rep.). Retrieved from http://static.googleusercontent.com/media/research.google.com/en//people/jeff/BayLearn2015.pdf [52]

Defferrard, M., Benzi, K., Vandergheynst, P., & Bresson, X. (2017, dec). FMA: A Dataset For Music Analysis. In *Ismir*. Retrieved from http://arxiv.org/abs/1612.01840 [100]

Dixon, S. (2006). Onset Detection Revisited. In *the 9th int. conference on digital audio effects (dafx'06)*. Montreal, Canada. Retrieved from http://www.dafx.ca/proceedings/papers/p_133.pdf [39, 41]

Dzhambazov, G., Yang, Y., Repetto, R. C., & Serra, X. (2016). Automatic Alignment of Long Syllables In A cappella Beijing Opera. In *Fma-2016*. Dublin, Ireland. Retrieved from http://hdl.handle.net/10230/32889 [90]

Dzhambazov, G. B., & Serra, X. (2015). Modeling of phoneme durations for alignment between polyphonic audio and lyrics. In *12th sound and music computing conference*. Maynooth, Ireland. Retrieved from http://hdl.handle.net/10230/27614 [45, 46]

Esposito, A., & Aversano, G. (2005). Text independent methods for speech segmentation. In *Nonlinear speech modeling and applications* (pp. 261–290). Springer. [43, 44]

Eyben, F., Böck, S., Schuller, B., & Graves, A. (2010). Universal Onset Detection with Bidirectional Long Short-term Memory Neural Networks. In *Ismir*. Retrieved from http://ismir2010.ismir.net/proceedings/ismir2010-101.pdf [39, 40, 41, 50]

Flexer, A., & Schnitzer, D. (2010). Effects of Album and Artist Filters in Audio Similarity Computed for Very Large Music Databases. *Computer Music Journal*, *34*, 20–28. doi: 10.2307/40963029 [116, 124]

Fonseca, E., Pons Puig, J., Favory, X., Font Corbera, F., Bogdanov, D., Ferraro, A., Oramas, S., Porter, A., & Serra, X. (2017). Freesound datasets: a platform for the creation of open audio datasets. In China (Ed.), *Ismir*. Suzhou: International Society for Music Information Retrieval (ISMIR). Retrieved from https://repositori.upf.edu/handle/10230/33299 [100]

Fujihara, H., Goto, M., Ogata, J., & Okuno, H. G. (2011). LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics. *IEEE Journal of Selected Topics in Signal Processing*, *5*(6), 1252–1261. doi: 10.1109/JSTSP.2011.2159577 [45, 46, 50]

Geringer, J. M., & Madsen, C. K. (1998). Musicians' Ratings of Good versus Bad Vocal and String Performances. *Journal of Research in Music Education*, *46*(4), 522–534. doi: 10.2307/3345348 [66]

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. In *the fourteenth international conference on artificial intelligence and statistics*. Retrieved from http://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf [142]

Gong, R., Cuvillier, P., Obin, N., & Cont, A. (2015). Real-

time audio-to-score alignment of singing voice based on melody and lyric information. In *Proceedings of interspeech 2015*. Dresden, Germany. Retrieved from https://hal.archives-ouvertes.fr/hal-01164550 [45, 46]

Goodfellow, I. (2016). *Is there a theory for why batch normalization has a regularizing effect? - Quora*. Retrieved 2018-08-13, from https://www.quora.com/Is-there-a-theory-for-why-batch-normalization-has-a-regularizing-effect [57]

Guédon, Y. (2007). Exploring the state sequence space for hidden Markov and semi-Markov chains. *Computational Statistics & Data Analysis*, *51*(5), 2379–2409. doi: https://doi.org/10.1016/j.csda.2006.03.015 [61, 131]

Gupta, C., Grunberg, D., Rao, P., & Wang, Y. (2017). Towards Automatic Mispronunciation Detection in Singing. In *Ismir*. Retrieved from https://ismir2017.smcnus.org/wp-content/uploads/2017/10/56_Paper.pdf [31, 34, 36, 91]

Gupta, C., Li, H., & Wang, Y. (2017). Perceptual Evaluation of Singing Quality. In *Apsipa annual summit and conference*. Retrieved from https://www.smcnus.org/wp-content/uploads/2013/09/WP-P2.5.pdf [31, 33, 36, 64]

Gupta, H. (2017). *One Shot Learning with Siamese Networks in PyTorch – Hacker Noon*. Retrieved 2018-08-13, from https://hackernoon.com/one-shot-learning-with-siamese-networks-in-pytorch-8ddaab10340e [57]

Han, Y., & Lee, K. (2014). Hierarchical Approach to Detect Common Mistakes of Beginner Flute Players. In *Ismir*. [31, 37, 38]

He, W., Wang, W., & Livescu, K. (2017). Multi-view Recurrent Neural Acoustic Word Embeddings. In *6th international conference on learning representations*. Retrieved from https://arxiv.org/pdf/1611.04496.pdf [47, 49]

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015, jun). Teaching Machines to Read and Comprehend. Retrieved from http://arxiv.org/abs/1506.03340 [157]

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term mem-

ory. *Neural computation*. doi: 10.1162/neco.1997.9.8.1735 [55]

Holzapfel, A., Davies, M. E. P., Zapata, J. R., Oliveira, J. L., & Gouyon, F. (2012, nov). Selective Sampling for Beat Tracking Evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(9), 2539–2548. doi: 10.1109/TASL.2012.2205244 [83]

Horii, Y., & Hata, K. (1988). A note on phase relationships between frequency and amplitude modulations in vocal vibrato. *Folia Phoniatrica*, *40*(6), 303–311. doi: https://doi.org/10.1159/000265924 [26]

HTK Speech Recognition Toolkit. (n.d.). Retrieved 2018-08-13, from http://htk.eng.cam.ac.uk/ [61]

Huang, X., & Deng, L. (2010). An Overview of Modern Speech Recognition. In *Handbook of natural language processing, second edition, chapter 15 (isbn: 1420085921)* (pp. 339–366). Chapman & Hall/CRC. Retrieved from https://www.microsoft.com/en-us/research/publication/an-overview-of-modern-speech-recognition/ [61]

Huh, J., Martinsson, E., Kim, A., & Ha, J.-W. (2018). Modeling Musical Onset Probabilities via Neural Distribution Learning. In *The 2018 joint workshop on machine learning for music*. Stockholm, Sweden. [40, 42]

Ioffe, S., & Szegedy, C. (2015, feb). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Retrieved from http://arxiv.org/abs/1502.03167 [56, 141]

Iskandar, D., Wang, Y., Kan, M.-Y., & Li, H. (2006). Syllabic level automatic synchronization of music signals and text lyrics. In *Proceedings of the 14th acm international conference on multimedia* (pp. 659–662). Santa Barbara, CA, USA. doi: 10.1145/1180639.1180777 [45, 46]

Kamper, H., Wang, W., & Livescu, K. (2016). Deep Convolutional Acoustic Word Embeddings using Word-pair Side Information. In *Icassp*. doi: 10.1109/ICASSP.2016.7472619 [47, 49, 51, 171, 183]

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. [59, 131, 135, 143, 157, 170]

Knight, T., Upham, F., & Fujinaga, I. (2011). The Potential for Automatic Assessment of Trumpet Tone Quality. In *Ismir*. Retrieved from http://ismir2011.ismir.net/papers/PS4-17.pdf [37, 38]

Krebs, F., Krebs, F., & Widmer, G. (2014). A Multi-model Approach to Beat Tracking Considering Heterogeneous Music Styles. In *Ismir*. Retrieved from http://www.terasoft.com.tw/conf/ismir2014/proceedings/T108_367_Paper.pdf [88]

Kruspe, A. M. (2015). Keyword spotting in singing with duration-modeled HMMs. In *23rd european signal processing conference (eusipco)* (pp. 1291–1295). Nice, France. doi: 10.1109/EUSIPCO.2015.7362592 [45, 46]

Kyriakopoulos, K., Gales, M. J. F., & Knill, K. M. (2017). Automatic Characterisation of the Pronunciation of Non-native English Speakers using Phone Distance Features. In *7th isca workshop on speech and language technology in education*. Retrieved from http://www.slate2017.org/papers/SLaTE_2017_paper_41.pdf [93]

Lacoste, A. (2007). Turbo convolutron 2000. In *Mirex*. Retrieved from http://www.music-ir.org/mirex/abstracts/2007/OD_lacoste.pdf [39, 40, 41]

Lacoste, A., & Eck, D. (2007). A Supervised Classification Algorithm for Note Onset Detection. *EURASIP Journal on Advances in Signal Processing*, *43745*. doi: 10.1155/2007/43745 [40, 41, 146]

Li, R. (2010). *The soul of Beijing opera : theatrical creativity and continuity in the changing world*. Hong Kong University Press. doi: https://doi.org/10.1017/S0305741012000215 [65]

Lin, C.-H., Lee, L.-s., & Ting, P.-Y. (1993). A new framework for recognition of Mandarin syllables with tones using sub-syllabic units. In *Icassp* (Vol. 2). doi: 10.1109/ICASSP.1993.319276 [24]

Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017, mar). A Structured Self-attentive Sentence Embedding. Retrieved from http://arxiv.org/abs/1703.03130 [163]

Liu, Y. X., Jin, Z. Y., Jia, J., & Cai, L. H. (2011, oct). An Automatic

Singing Evaluation System. *Applied Mechanics and Materials*, *128-129*, 504–509. doi: 10.4028/www.scientific.net/AMM.128-129.504 [31, 32, 35]

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. In *The ieee conference on computer vision and pattern recognition (cvpr)* (pp. 3431–3440). Boston, USA. doi: 10.1109/CVPR.2015.7298965 [143]

Loscos, A., Cano, P., & Bonada, J. (1999). Low-Delay Singing Voice Alignment to Text. In *International computer music conference*. Beijing, China. [45, 46]

Luo, Y.-J., Su, L., Yang, Y.-H., & Chi, T.-S. (2015). Detection of Common Mistakes in Novice Violin Playing. In *Ismir*. Retrieved from http://ismir2015.uma.es/articles/197_Paper.pdf [31, 37, 38]

MATLAB. (n.d.). *Gradient descent with momentum backpropagation - MATLAB traingdm*. Retrieved 2018-08-13, from https://www.mathworks.com/help/nnet/ref/traingdm.html [59]

Mauch, M., Fujihara, H., & Goto, M. (2012). Integrating additional chord information into HMM-based lyrics-to-audio alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 200–210. doi: 10.1109/TASL.2011.2159595 [45, 46]

Mayor, O., Bonada, J., & Loscos, A. (2006). The Singing Tutor: Expression Categorization and Segmentation of the Singing Voice. In *Aes 121st convention*. Retrieved from http://mtg.upf.edu/files/publications/a13c70-AES121-omayor-jonada-aloscos.pdf [33, 34, 35]

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proceedings of interspeech 2017* (pp. 498–502). Stockholm, Sweden. doi: 10.21437/Interspeech.2017-1386 [43, 44]

Mesaros, A., & Virtanen, T. (2008). Automatic alignment of music audio and lyrics. In *Proceedings of the 11th int. conference on digital audio effects (dafx-08)*. Espoo, Finland. Retrieved from http://www.cs.tut.fi/sgn/arg/

music/tuomasv/autalign_cr.pdf [45]

Mielke, J. (2012, jan). A phonetically based metric of sound similarity. *Lingua*, *122*(2), 145–163. doi: 10.1016/J.LINGUA.2011.04.006 [93]

Minematsu, N. (2004). Yet Another Acoustic Representation of Speech Sounds. In *Icassp*. doi: 10.1109/ICASSP.2004.1326053 [92]

Minematsu, N., Kamata, K., Asakawa, S., Makino, T., & Hirose, K. (2007). Structural Representation of Pronunciation and its Application for Classifying Japanese Learners of English. In *Speech and language technology in education*. Retrieved from https://www.gavo.t.u-tokyo.ac.jp/{~}mine/paper/PDF/2007/SLaTE{_}2{_}t2007-10.pdf [92]

Mohri, M., Pereira, F., & Riley, M. (2002, jan). Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, *16*(1), 69–88. doi: 10.1006/CSLA.2001.0184 [152]

Molina, E., Barbancho, I., Gómez, E., Barbancho, A. M., & Tardón, L. J. (2013). Fundamental Frequency Alignment vs. Note-based Melodic Similarity for Singing Voice Assessment. In *Icassp*. doi: 10.1109/ICASSP.2013.6637747 [31, 33, 35]

Moran, S., McCloy, D., & Wright, R. (2014). *PHOIBLE Online*. Retrieved 2018-07-25, from http://phoible.org [89]

Mu, W. (2007). 京剧打击乐教程 *(Course on jingju percussion performance)*. Beijing, China: Renmin yinyue chubanshe. [17]

Müller, M., Kurth, F., Damm, D., Fremerey, C., & Clausen, M. (2007). Lyrics-based audio retrieval and multimodal navigation in music collections. In *International conference on theory and practice of digital libraries* (pp. 112–123). Budapest, Hungary. [45, 46]

Nair, G. (1999). *Voice tradition and technology : a state-of-the-art studio*. Singular Pub. Group. [29, 30]

Nakano, T., Goto, M., & Hiraga, Y. (2006). An Automatic Singing Skill Evaluation Method for Unknown Melodies Using Pitch Interval Accuracy and Vibrato Features. In *Interspeech*. Retrieved from https://staff.aist.go.jp/m.goto/PAPER/INTERSPEECH2006nakano.pdf [31, 32,

35]

Ng, A. (n.d.). *What data scientists should know about deep learning* (Tech. Rep.). Retrieved from https://www.slideshare.net/ExtractConf [52]

Olah, C. (2015). *Understanding LSTM Networks – colah's blog.* Retrieved 2018-08-13, from http://colah.github.io/posts/2015-08-Understanding-LSTMs/ [55]

Pakoci, E., Popović, B., Jakovljević, N., Pekar, D., & Yassa, F. (2016). A phonetic segmentation procedure based on hidden markov models. In *International conference on speech and computer* (pp. 67–74). Budapest, Hungary. [43, 44, 129]

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *In international conference on machine learning.* Retrieved from http://proceedings.mlr.press/v28/pascanu13.pdf [55]

Pati, K., Gururani, S., & Lerch, A. (2018, mar). Assessment of Student Music Performances Using Deep Neural Networks. *Applied Sciences*, *8*(4), 507. doi: 10.3390/app8040507 [37, 38]

Penn Phonetics Lab Forced Aligner. (n.d.). Retrieved from https://web.sas.upenn.edu/phonetics-lab/facilities/ [43, 44]

Pons, J., Gong, R., & Serra, X. (2017). Score-informed Syllable Segmentation for A Cappella Singing Voice with Convolutional Neural Networks. In *Ismir.* Retrieved from https://ismir2017.smcnus.org/wp-content/uploads/2017/10/46_Paper.pdf [140, 142]

Pons, J., Slizovskaia, O., Gong, R., Gómez, E., & Serra, X. (2017). Timbre analysis of music audio signals with convolutional neural networks. In *25th european signal processing conference (eusipco)* (pp. 2744–2748). Kos, Greece. doi: 10.23919/EUSIPCO.2017.8081710 [84, 130, 131, 156]

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *Ieee 2011 workshop on automatic speech recognition and understanding.* IEEE Signal Processing Society. [43, 61]

Rabiner, L. R. (1989). A tutorial on hidden Markov models and

selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. doi: 10.1109/5.18626 [60, 137]

Raffel, C., & Ellis, D. P. W. (2015, dec). Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems. Retrieved from http://arxiv.org/abs/1512.08756 [157]

Renals, S., Morgan, N., Bourlard, H., Cohen, M., & Franco, H. (1994). Connectionist probability estimators in HMM speech recognition. *IEEE Transactions on Speech and Audio Processing*, *2*(1), 161–174. doi: 10.1109/89.260359 [129]

Repetto, R. C. (2018). *The musical dimension of Chinese traditional theatre. An analysis from computer aided musicology* (Unpublished doctoral dissertation). Universitat Pompeu Fabra. [17, 18, 19, 20, 21, 24]

Repetto, R. C., & Serra, X. (2014). Creating a corpus of jingju (Beijing opera) music and possibilities for melodic analysis. In *Ismir*. Taipei, Taiwan. Retrieved from http://mtg.upf.edu/node/3017 [87, 101]

Repetto, R. C., Zhang, S., & Serra, X. (2017). Quantitative analysis of the relationship between linguistic tones and melody in jingju using music scores. In *Proceedings of the 4th international workshop on digital libraries for musicology - dlfm '17* (pp. 41–44). Shanghai, China: ACM Press. doi: 10.1145/3144749.3144758 [87, 101]

Roach, P. (2000). *English Phonetics and Phonology: A Practical Course*. Cambridge University Press. [25]

Robine, M., & Lagrange, M. (2006). Evaluation of the Technical Level of Saxophone Performers by Considering the Evolution of Spectral Parameters of the Sound. In *Ismir*. Retrieved from http://ismir2006.ismir.net/PAPERS/ISMIR06129_Paper.pdf [31, 37, 38]

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*. [170]

Schluter, J., & Bock, S. (2014, may). Improved musical onset detection with Convolutional Neural Networks. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6979–6983). IEEE. doi: 10.1109/ICASSP.2014.6854953 [39, 40, 42, 50, 128, 135, 142, 143, 144]

Schramm, R., De, H., Nunes, S., & Jung, C. R. (2015). Automatic Solfège Assessment. In *Ismir*. Málaga. Retrieved from http://ismir2015.uma.es/articles/75_Paper.pdf [31, 34, 36]

Schuster, M., & Paliwal, K. K. (1997). Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, *45*(11). doi: 10.1109/78.650093 [55]

Serra, X. (2011). A Multicultural Approach in Music Information Research. In *Ismir*. Miami, USA. doi: https://doi.org/10.1177/025576149702900111 [101]

Serra, X. (2014). Creating research corpora for the computational study of music: the case of the Compmusic project. In *Audio engineering society conference: 53rd international conference: Semantic audio*. Retrieved from http://mtg.upf.edu/node/2899 [100, 101, 194]

Serrière, G., Cerisara, C., Fohr, D., & Mella, O. (2016). Weakly-supervised text-to-speech alignment confidence measure. In *International conference on computational linguistics (coling)*. Osaka, Japan. Retrieved from http://www.aclweb.org/anthology/C16-1192 [43, 44]

Settle, S., Levin, K., Kamper, H., & Livescu, K. (2017). Query-by-Example Search with Discriminative Neural Acoustic Word Embeddings. In *Interspeech*. Retrieved from https://arxiv.org/pdf/1706.03818.pdf [47, 49]

Settle, S., & Livescu, K. (2016). Discriminative Acoustic Word Embeddings: Recurrent Neural Network-based Approaches. In *Ieee spoken language technology workshop (slt)*. Retrieved from https://arxiv.org/pdf/1611.02550.pdf [47, 49, 51, 58, 171, 180, 182, 183]

Shiozawa, F., Saito, D., & Minematsu, N. (2016, dec). Improved prediction of the accent gap between speakers of English for individual-based clustering of World Englishes. In *2016 ieee spoken language technology workshop (slt)* (pp. 129–135). IEEE. doi: 10.1109/SLT.2016.7846255 [92]

Srinivasamurthy, A. (2016). *A Data-driven bayesian approach to automatic rhythm analysis of indian art music* (Doctoral dissertation, Universitat Pompeu Fabra). Retrieved from http://hdl.handle.net/10803/398986 [83, 89, 107]

Srinivasamurthy, A., Holzapfel, A., & Serra, X. (2014, jan). In

Search of Automatic Rhythm Analysis Methods for Turkish and Indian Art Music. *Journal of New Music Research*, *43*(1), 94–114. doi: 10.1080/09298215.2013.879902 [101]

Srivastava, N., Hinton, G., Krizhevsky, A., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958. Retrieved from http://jmlr.org/papers/volume15/srivastava14a.old/srivastava14a.pdf [56]

Stanford University CS231n: Convolutional Neural Networks for Visual Recognition. (n.d.). Retrieved 2018-08-13, from http://cs231n.stanford.edu/ [53, 55, 56, 58, 59]

Stoller, D., & Dixon, S. (n.d.). Analysis and Classification of Phonation Modes in Singing. In *Ismir-2016.* Retrieved from https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/233_Paper.pdf [174]

Su, W., Yuan, Y., & Zhu, M. (2015). A Relationship between the Average Precision and the Area Under the ROC Curve. In *the 2015 international conference on theory of information retrieval - ictir '15* (pp. 349–352). New York, New York, USA: ACM Press. doi: 10.1145/2808194.2809481 [51]

Sundberg, J., Lã, F. M., & Gill, B. P. (2013, may). Formant Tuning Strategies in Professional Male Opera Singers. *Journal of Voice*, *27*(3), 278–288. doi: 10.1016/J.JVOICE.2012.12.002 [30]

Thomson, R. I. (2008). *Modeling L1/L2 interactions in the perception and production of English vowels by Mandarin L1 speakers: A training study* (Unpublished doctoral dissertation). University of Alberta, Edmonton, Alberta, Canada. [93]

Tian, M., & Sandler, M. B. (2016, oct). Towards Music Structural Segmentation across Genres. *ACM Transactions on Intelligent Systems and Technology*, *8*(2), 1–19. doi: 10.1145/2950066 [87]

Tian, M., Srinivasamurthy, A., Sandler, M., & Serra, X. (2014, may). A study of instrument-wise onset detection in Beijing Opera percussion ensembles. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 2159–2163). IEEE. doi: 10.1109/ICASSP

.2014.6853981 [88]

Titze, I. R., & Sundberg, J. (1992, may). Vocal intensity in speakers and singers. *The Journal of the Acoustical Society of America*, *91*(5), 2936–2946. doi: https://doi.org/10.1121/1.402929 [26]

Toh, C. C., Zhang, B., & Wang, Y. (2008). Multiple-feature Fusion Based Onset Detection for Solo Singing Voice. In *Ismir*. Retrieved from http://ismir2008.ismir.net/papers/ISMIR2008_127.pdf [39, 41]

Tonglin Shu. (2011). 余叔岩“十八张半”唱腔浅析 *(Analysis of Shuyan Yu's arias)*. Tianjin gu ji chu ban she. [28]

Tsai, W.-H., & Lee, H.-C. (2012, may). Automatic Evaluation of Karaoke Singing Based on Pitch, Volume, and Rhythm Features. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(4), 1233–1243. doi: 10.1109/TASL.2011.2174224 [31, 32, 35]

Uyar, B., Atli, H. S., Şentürk, S., Bozkurt, B., & Serra, X. (2014). A Corpus for Computational Research of Turkish Makam Music. In *Proceedings of the 1st international workshop on digital libraries for musicology - dlfm '14* (pp. 1–7). New York, New York, USA: ACM Press. doi: 10.1145/2660168.2660174 [101]

Van Der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605. [172]

Vidwans, A., Gururani, S., Wu, C.-W., Subramanian, V., Swaminathan, R. V., & Lerch, A. (2017). Objective descriptors for the assessment of student music performances. In *Aes conference on semantic audio*. Retrieved from http://www.aes.org/e-lib/browse.cfm?elib=18758 [37, 38]

Vogl, R., Dorfer, M., Widmer, G., & Knees, P. (2017). Drum Transcription via Joint Beat and Drum Modeling Using Convolutional Recurrent Neural Networks. In *18th international society for music information retrieval conference*. Suzhou, China. Retrieved from https://ismir2017.smcnus.org/wp-content/uploads/2017/10/123_Paper.pdf [142, 144]

Wang, J. (1994). Syllable duration in Mandarin. In *the fifth international conference on speech science and technology*. [26,

111]

Wang, Y., Kan, M.-Y., Nwe, T. L., Shenoy, A., & Yin, J. (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). New York, NY, USA. doi: 10.1109/TASL.2007.911559 [45, 46]

Wichmann, E. (1991). *Listening to theatre : the aural dimension of Beijing Opera*. University of Hawaii Press. [17, 19, 20, 21, 22, 23, 25, 26, 27, 29, 76, 79, 89]

Wieling, M., Margaretha, E., & Nerbonne, J. (2011). Inducing phonetic distances from dialect variation. *Computational Linguistics in the Netherlands Journal*, *1*, 109–118. doi: 10.1016/j.wocn.2011.12.004 [93]

Wikipedia. (n.d.-a). *Convolutional neural network.* Retrieved from https://en.wikipedia.org/wiki/Convolutional_neural_network [54]

Wikipedia. (n.d.-b). *Hidden semi-Markov model.* Retrieved from https://en.wikipedia.org/wiki/Hidden_semi-Markov_model [61]

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*. doi: 10.1016/S0167-6393(99)00044-8 [91]

Wu, C.-W., & Lerch, A. (2018). Learned Features for the Assessment of Percussive Music Performances. In *International conference on semantic computing (icsc)*. doi: 10.1109/ICSC.2018.00022 [37, 38]

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015, feb). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Retrieved from http://arxiv.org/abs/1502.03044 [157]

Yang, L., Tian, M., & Chew, E. (2015). Vibrato Characteristics and Frequency Histogram Envelopes in Beijing Opera Singing. In *the fifth international workshop on folk music analysis*. Retrieved from http://qmro.qmul.ac.uk/xmlui/handle/123456789/16062 [79]

Young, S. J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (2006). *The HTK Book Version 3.4*. Cambridge University Press. [43, 60, 61]

Yung, B. (1989). *Cantonese opera : performance as creative process*. Cambridge University Press. [21, 22]

Zeghidour, N., Synnaeve, G., Usunier, N., & Dupoux, E. (2016). Joint Learning of Speaker and Phonetic Similarities with Siamese Networks. In *Interspeech*. doi: 10.21437/Interspeech.2016-811 [47, 49]