Structure Analysis of Beijing Opera Arias

Yile Yang

MASTER THESIS UPF / 2016

Master in Sound and Music Computing

Master thesis supervisor:

Xavier Serra

Department of Information and Communication Technologies

Universitat Pompeu Fabra, Barcelona

UNIVERSITAT POMPEU FABRA

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

This thesis has been carried out between Sep. 2014 and Jul. 2016 at the Music Technology Group (MTG) of Universitat Pompeu Fabra (UPF) in Barcelona (Spain), supervised by Dr. Xavier Serra. The work in Chapter 3 has been conducted in collaboration with Rafael Caro Repetto. The work in Chapter 4 has been conducted in collaboration with Georgi Dzhambazov. This work has been supported by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583).

Acknowledgements

This work is realized with the help and contribution of a number of people. First I would like to express my gratitude for my supervisor, Professor Xavier Serra. His knowledge and experience have provided me with valuable guidance. And he is always supportive throughout the whole process. His passion for research, work ethic, and leadership have set an example for me even outside the work of the thesis.

I would also like to thank all the professors in the SMC program. Their efforts in teaching provide the students with abundant amount of knowledge and great resources. They are always willing to help with a great attitude.

I am sincerely grateful to the researchers in the CompMusic project: Sertan Şentürk, Gopala K. Koduri, Sankalp Gulati, Andrés Ferraro, Ajay Srinivasamurthy, Alastair Porter, Shuo Zhang, Rafael Caro Repetto, Georgi Dzhambazov, Rong Gong, Dr. Mohamed Sordo, Kaustuv Kanti Ganguli, Kainan Chen, and Dr. Dmitry Bogdanov. Their knowledge, support, and encouragement have provided me with tremendous help and motivation.

I would like to thank all the colleagues in the SMC master programs. It is a diverse and welcoming group with students from all over the world. I appreciate the friendship that we have established and all the experiences that have made my life better during my stay in Barcelona.

Last but not least, I would like to thank my parents for their unconditional support. None of this could be made possible without them being on my back.

Abstract

Music has been playing an important role in the society. With the wide spread of digital and mobile devices, the consumption of music has been increasing significantly. Therefore, the demand for automatic processing and analyzing music with computational technologies has also been pushed to a new ground. It has opened up opportunities for a variety of applications which were not possible before. Structure, being a fundamental entity of music, is being used in a lot of those applications to help people organize, navigate, and understand music. It is also of great importance for the academic areas such as Computational Musicology and Music Information Research (MIR). However, such application and research are mostly focused on western music, while there is increasing demand for non-western music as well. Beijing opera, being one of the most famous traditional operas in China, has influence all over the world. But few efforts have been made in the study of Beijing opera music with a computational perspective. In this work, we propose computational tools for the structure analysis of Beijing opera arias.

Literature review of state of the art research on related topics are performed. Methodologies, data, and challenges of these topics are analyzed. Then we proceed to the specific tasks chosen for the structure analysis of Beijing opera arias on different levels. The first task is the segmentation of singing, percussion and instrumental sections. Features have been chosen to reflect the nature of the sound in these sections. An Support Vector Machine (SVM) classifier is built using these features. Experiments on a 34 aria dataset have shown the effectiveness of this method. The influence of features have also been analyzed.

The second task is lyrics-to-audio alignment. For this task, two datasets of a capella singing have been created, with human annotations on the phoneme level. The study of the datasets have been performed and have shown interesting characteristics of the lyrics of Beijing opera singing. An approach based on Duration-aware Hidden Markov Model (DHMM) has been proposed to address the challenge of long syllable durations in Beijing opera arias. The evaluation results demonstrate the ability of the proposed method to tackle the task of lyrics-to-audio alignment in the case of Beijing opera arias.

The thesis is concluded with main conclusions and a summary of the contributions of this work.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Motivation and goals

Nowadays there is vast amount of information available from various sources and there is a fast growing demand for people to try to take advantage of machines to help process and analyze the information. Humans, either consciously or unconsciously, tend to identify the underlying structure of the information in order to grasp the meaning of a given message. Structure is of great importance in the understanding of information. And the importance is carried also to the realm of music, if not increased. It is one of the essential topics in the study of Computational Musicology as well as Music Information Research (MIR) [26]. For a variety of tasks in MIR, a step of structural segmentation has to be performed in order to conduct further research. For example, in order to study the melody of singing, the singing voice section has to be segmented first. The same goes to studies in rhythm, lyrics and so on. If musically meaningful segments can be identified automatically, it will be of great help for further research. Structure information is also very useful for music education, which can help people organize, navigate, and understand music better.

However, most research in the field of MIR is centered on western music, despite the world's richness in terms of musical culture [27]. The models built are biased and do not respond to world's multi-cultural reality, which can hardly be applied to music from other origins. Different musical cultures will also provide different perspectives and pose new challenges, and computational tools can be utilized to study them. For example, the music of Beijing opera (also known as 'Jingju'), which is a form of traditional Chinese theatre, has a lot of valuable musical characteristics that are not present in western music, and current technologies developed are not capable of addressing them well. Thus, it is necessary to approach the current research challenges from a cultural specific perspective. Developing

techniques of relevance to Beijing Opera music can provide new insights to the studies and thus contribute to the filed of MIR.

The main goal of this research is to develop computational tools for the structure analysis of the music of Beijing Opera arias. There are a few aspects of music that can be considered as structure information. In the case of Beijing Opera, an example of the music structure can be found in Figure 1.1. It shows the structure for different elements. Under the aria name, the first row shows the section based on melody patterns, which is called 'Shengqiang'. The annotated characters centered in the section is the name of the 'Shengqiang'. The second row represents the sections based on rhythm. The annotations are the types of 'Banshi', which is the term for rhythmic patterns. The third row shows the sections of lyrics lines. The annotations are the lyrics in Chinese characters. Each section corresponds to a lyric line. The last row show the sections of percussion patterns called 'Luogu', which are are a set of special percussion passages used for certain situations.

As seen from the figure, structure analysis can be done on different levels with different focuses. In this work, two specific tasks have been chosen to accomplish. The first is the segmentation of singing, percussion, and instrumental parts, focusing on structure based on singing voice and instruments, which is covered in Chapter 3. The second task is lyrics-to-audio alignment, focusing on structure based on lyrics, which is covered in chapter 4.

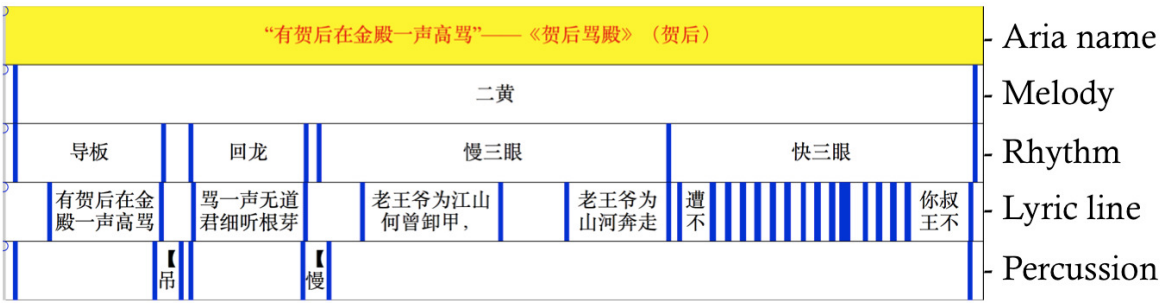

Fig. 1.1 Example of an annotation of a segment of a Beijing opera aria. Horizontal axis is time. Blue lines are section boundaries.

1.2 Beijing opera (Jingju) music

Beijing opera, also known as Jingju in Chinese, is one of the most famous forms of Chinese opera that has history over a hundred years. It originated from Anhui opera from southern China, and was influenced by several operas from other provinces. It quickly gained popularity during the Qing dynasty and became famous throughout China since then. Beijing

Opera has a complicated art form that combines music, vocal performance, mime, dance, and acrobatics.

The characters in Beijing opera have four main role types: Sheng, Dan, Jing, and Chou. Role type can reflect the character's gender, age, social states, and personalities. It can usually be recognized by the character's outfit and singing style.

The vocal production in Jingju is composed of four types: songs with music, verse recitation, prose dialogue, and non-verbal vocalizations [29]. The techniques for Jingju singing are quite different from many other musical cultures. For example, it has special breathing techniques to be used in different situations. It has also techniques to imitate the voice from people of different ages.

The accompaniment for a Beijing Opera performance usually consists of a small ensemble of traditional melodic and percussion instruments, including jinghu, a small high-pitched two-string spike fiddle; erhu, another two-string spike fiddle; yueqin, a four-string guitar like instrument; and percussion instruments like daluo, xiaoluo, naobo, etc. The music for aria falls into one of the two categories: Erhuang and Xipi. These are called 'Shengqiang', which represents the melodic patterns. Similarly, another concept 'Banshi' represents rhythmic patterns, which has several types with different rhythms and tempo that can be used to accompany different emotions and situations.

Lyrics in Beijing opera aria are mostly written in couplets consisting of two lines, whose last syllables usually rhyme with each other. Each lyric line can be further divided into smaller units called 'dou', which has 2, 3 or 4 syllables (characters). Thus most lines can be divided into three 'dou' patterns, with short pauses in between, sometimes too short to be perceivable. Patterns of 2-2-3 and 3-3-4 are most common patterns for lines of 7 and 10 characters. The pronunciation of the lyrics is mainly Mandarin, but mixed with a couple of other dialects. The lyrics carry great significance for Beijing opera arias since they tell the content of the story and can express the emotion directly. The composition of music is influenced by the lyrics structure.

Chapter 2

State of the Art

This chapter reviews the state of the art literature that is related to the topic of this work, including music structure analysis, singing voice detection, lyrics-to-audio alignment, and lyrics searching. The description of each task is introduced. The approaches of each task are surveyed and evaluated. And the challenges of these tasks are identified. We summarize the literature to provide insights and useful information for tackling our tasks.

2.1 Music structure analysis

Structure analysis of music is one of the most fundamental topics in Computational Musicology as well as in Music Information Research. Music is full of structure, including sections, sequences of distinct musical textures, and the repetition of phrases or entire sections. In Western music, the main high-level structural organization of a piece is the musical form, which describes the layout of a composition as divided into sections, segments, or parts, e.g. 'exposition', 'development', 'recapitulation', 'coda' of a sonata. In popular music, compositions are usually divided into segments that alternate or repeat throughout the piece, commonly called 'intro', 'verse', 'chorus', and 'bridge'. In the context of MIR, music structure analysis has been a popular topic and has accumulated a large amount of studies. It aims to find patterns and repetitions underlying the structure of the song [22][21][4][16].

The most commonly used approach for this task is to define a similarity measure between audio frames or segments, then use the features extracted from these frames to compute the similarity values, which form into a similarity matrix [6]. In most cases, the matrix is a self-similarity matrix, which is useful to find repetitions within a song. Figure 1.1 is an example of self-similarity matrix.

The similarity matrix is used to find similar patterns, for example the chorus of a pop song, which can be repeated several times with minor change. However, the arias of Beijing

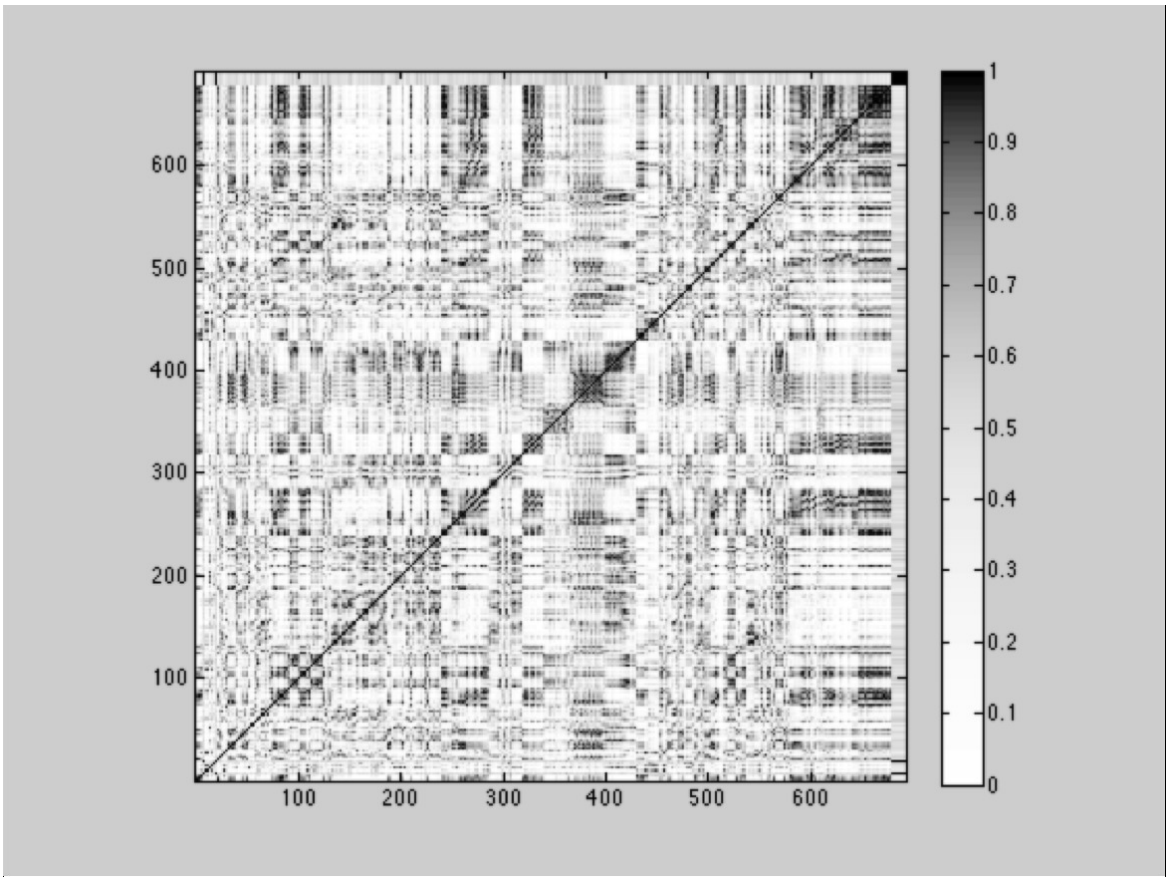

Fig. 2.1 An example of a self-similarity matrix of an excerpt of music based on Chroma feature.

opera apply different compositional approach, where repetitions like the chorus in pop music don't usually happen. The music of Beijing opera applies heterophonic texture to follow the melody of singing. And the singing doesn't not have a representative part like the chorus where a certain melody is repeated and emphasized. Thus this approach has to be adapted for the studying of Beijing opera arias.

Apart from this approach, some researchers have addressed this topic from a cognitive perspective. Focusing on melody segmentation, [25] attempted to formalized the problem of melody segmentation and analyzed segmentation based on different aspects: repetition based, contrast based, template based, and multi-cue. This work provides insights to look at the structure from different perspectives.

2.2 Singing voice detection

As a usually predominant part in music, the study of singing voice has drawn attention of many researchers. The task of singing voice detection is to locate the segment when singing voice is present. It is a necessary step before conducting many other tasks such as lyrics-to-audio alignment. It is also important to isolate singing voice from non-singing voice in order to extract features that are only meaningful when singing voice is present. To tackle this task, one has to find out the features that could differentiate singing voice from other type of musical sound. The distinctive features can be found in both frequency and time domain. Once the features have been selected, a normal approach is to then use machine learning method to classify singing voice and other sounds.

In the work of [19], researchers selected features motivated by human perception of sounds. The features used were regarded to reflect harmonic, vibrato and timbre characters of singing voice, such as cepstral coefficients. Temporal features such as attack-decay envelope have also been used to capture characteristics in the time domain. The results show that the chosen features are effective in identifying singing voice.

Apart from the low level features extracted from the signal, high level features can also be useful. [20] incorporated harmonic content attenuation using higher level musical knowledge like key and structure information. A Multi-Model Hidden Markov Model was built and achieved a classification accuracy of 86.7%, on a dataset of 20 popular songs.

However, a dataset of 20 songs is quite small to validate certain methods, which is also a noticeable issue in the research of MIR, since large datasets are hard to acquire and annotate. To take advantage of abundant unlabeled songs, [13] used co-training algorithm to leverage compatible and partially uncorrelated information across different features to effectively boost the model from unlabeled data. Vibrato, harmonic and MFCC features were tested

using a HMM classifier. The combination of all features yield the best results, achieving an average error rate of 17% in segment level singing voice detection on a dataset of 94 pop songs.

One of the biggest problems in singing voice detection is that the accompanied instrument might have similar sound to singing voice, for instance some string instruments, which would make the system confused. These instruments usually share similar harmonic features to singing voice. To address this issue, new features have been designed. [15] designed two new features to describe temporal characteristics of the signal. One is the Fluctogram, which detects sub-semitone fluctuations of partials by using cross correlation. The other feature is called Vocal Variance which is derived from MFCCs over time. An random forest algorithm was used as a binary classifier on audio window level. The result showed an accuracy of 86.32% on a dataset of 75 songs by 75 different artists, outperforming baseline method, as well as method proposed by [28]. The analysis also showed that the new features improved false positives effectively on string quartet and some wood instruments. Experiments on common benchmark datasets were conducted as well, which showed a performance comparable to methods that are much more sophisticated.

In some cases the singing voice is not predominant and the accompanied instrument is very loud. This could influence the features and cause mistakes. To improve the robustness in this situation, [23] used sinusoidal modeling to attenuate harmonic spectral content of stable-pitch instruments which is likely to be mistaken. Gaussian Mixture Model (GMM) was used as the classifier. This research was done in the context of Indian music and it showed high classification accuracy on seven different north Indian classical vocal performances.

As we can seen, the detection of singing voice has been extensively studied. However, the subjects are mainly popular music, with only a few on other music genres and cultures. To our knowledge, only one previous work has been done on the detection of singing voice in the context of Beijing opera [5], which created a dataset of 1 hour, half of which is singing. The work used 11 features including MFCCs, pitch confidence, inharmonicity, etc. Several classifiers such as ZeroR, NaiveBayes, Decision tress, SVM, and K-NN were tested. Using a 10-fold cross validation, K-NN classifier reported the highest precision of 96.5%.

2.3 Lyrics-to-audio alignment

Given audio signals of singing voices and corresponding textual lyrics as input data, lyrics-to-audio alignment can be defined as a problem of estimating the temporal relationship between them. In this task, start and end times of every block of certain length in lyrics are estimated. Here, the term ”block“ means a fragment of lyrics, the size of which depends on

the application, and can be either phoneme, syllable, word, phrase, line, or paragraph [10]. So far, most studies on lyrics-to-audio alignment are on the phoneme level.

Lyrics-to-audio alignment is heavily influenced by the methodology used in text-to-speech alignment. A widely used approach is to build phoneme models based on acoustic features like MFCCs, with forced alignment and Hidden Markov Models (HMMs). Despite the similar approach, lyrics sung in music lie in a very different context from speech, which normally has no music in the background. Thus it brings certain difficulties for this task. [10] identified 3 of them:

1. Fluctuation of acoustic characteristics. Unlike natural speech, singing voices are usually produced in a special manner to express certain effects and emotions, which would alter the frequency and dynamic characters.

2. Influences of accompanied sounds. It is difficult sometimes even for a human listener to understand the lyrics when accompanied by other instruments. It brings great negative influences on the spectrum of singing voice.

3. Incomplete lyrics. This is a less significant issue but exists when the quality of lyrics is not ideal.

As mentioned above, the HMM speech recognizer with Viterbi forced alignment is a common approach for this task. Apart from that, to address the accompanied instruments in real life music, [18] proposed a pre-processing step, using a voice separation algorithm based on melody transcription and sinusoidal modeling, to attenuate the influence from accompanied instruments. A language model consisting of a sequence of phonemes, pauses and possible instrumental breaks was also created to improve the performance of recognition. The HMMs were adapted from speech data. The test dataset consisted of 17 commercial recordings. An average absolute error of 1.40 seconds in aligning lines of the lyrics was reported.

To address overlapping instrumental accompaniment in CD recordings, [12] followed similar methodologies but utilized more sophisticated approaches. A step of accompany sound reduction were performed to get more clean singing voice. A vocal activity detection method had also been proposed that can control the balance between the hit and correct rejection rates. Moreover, a novel fricative detection was proposed to detect unvoiced consonants, in order to increase the robustness of the system. The adaption from speech model was also used in this work. Comprehensive experiments were conducted to test different part of the system and results showed that the system can accurately synchronize musical audio signals and their lyrics.

High level features can also be used for lyrics-to-audio alignment. [17] integrated chord information into HMM-based lyrics-to-audio alignment and achieved improvement with statistical significance.

Nearly all the research on this task has to build acoustic models on the phoneme level. One thing that is worth noticing is that the data used for building acoustic models. Unlike speech, which has plenty of data annotated on the syllable or phoneme level. There is a lack of such data for singing. Therefore most studies choose to build the model indirectly by adapting from speech data. But efforts have been made to train directly using singing voice as well. [11] used predominant F0 to extract corresponding harmonic structure, and re-synthesized the information to acquire cleaner singing voice. Phone models were trained from scratch on the singing voice. The experiments comparing these models with adapted models from speech had shown that averaged frame level accuracy was drastically improved with the non-adaptation models. It further stated that results indicated that, when enough training data were available, phone models trained from scratch could perform better than the adapted ones trained from speech.

2.4 Lyrics searching and key work spotting

Another related area of research is lyrics searching and keyword spotting. Though most of the studies of searching and spotting are made in speech area, there are still a few efforts that have been made on music. In this context, the data used is mostly a capella singing, partially because current systems are not robust enough to handle accompanied instruments.

[14] presented a keyword spotting system based on keyword-filler HMMs. MFCCs, perceptual linear predictive features, and temporal pattern features were used to train phoneme model directly from singing voice. A best precision of 31% had been reported on a dataset of 19 a capella songs. It also compared models trained from speech adaption and showed that the models trained on singing outperformed the adapted ones. However, the overall performance is not satisfying enough due to the difficult nature of this task.

[7] proposed an approach that addresses the differences of syllable durations, specific for singing. MFCC-based phoneme models were trained and adapted from speech. Dynamic time warping between the phrase and audio was applied to estimate candidate audio segments. Then a novel score-informed HMM, which could model the duration of syllables, was created to give a rank to the candidates. The proposed approach is evaluated on 12 a capella audio recordings of Turkish Makam music. The reported results outperformed a baseline approach unaware of syllable duration.

Chapter 3

Singing, Percussion, Instrumental Segmentation

In this chapter, we report in detail the work that has been done on the segmentation of singing, percussion, and instrumental sections. We first introduce the background knowledge of the music in Beijing opera arias. Then the proposed approach to tackle this task is explained, followed by the discussion on the features that are used. Three experiments have been carried out and the results show that the system is able to yield satisfying performance.

3.1 Music in Beijing opera arias

Beijing opera arias are usually comprised of these three parts: the singing part, the percussion part ('Luogu'), and the rest accompanying instrumental part. The singing part is normally the most predominant, while the passages of percussion part happening from time to time as important structural point, with the instrumental part accompanying in the background. Since these parts have their own significance, it is helpful to segment each part from the whole aria, which could provide information of the structure. It is also a necessary step for further studies on these separated segments. For example, predominant melody could be derived from singing segments, and percussion patterns could be discovered and studied from percussion segments.

The aural dimension of Beijing Opera consists of three main components: singing, instrumental music, and artistic declamation, in which the first two components take up the majority of the content. The singing in Beijing opera aria is accompanied by a small instrumental ensemble, in a heterophonic texture. In the ensemble, the fiddle-like instrument

Jinghu has similar timbre to the female voice, which often sings at a very high pitch. The Jinghu sometimes even plays the exact same pitch/melody as the singing voice.

There is a repertoire of predefined, labelled percussion patterns, called 'Luogu', which is played by the percussion part of the instrumental ensemble only. The functions of these patterns are: accompanying and conducting actor's stage movements, setting the emotional atmosphere, signaling structural points from the play to the aria levels, introducing arias and their sections. Thus identifying section of these percussion patterns is quite meaningful for the structure analysis of Beijing Opera music.

What needs to be clarified is that, in the task of segmentation of singing, percussion, instrumental, 'percussion' refers to 'Luogu' section where only percussion instruments are played, while 'instrumental' refers to the rest of the instrumentation part where no singing voice is present, but percussion instruments may be present as well. Singing part refers to the section as long as singing voice is present.

3.2 Approach

For the segmentation of singing, percussion, and instrumental, this work follows a general machine learning approach. The difference of these segments can be captured by a number of features, which are extracted from short audio frames. These features are then used as input to a classifier, which learns to classify the audio frames into one of the three categories. Finally, a step of grouping and smoothing is performed to connect these short frames into segments.

For the proposed approach, the key factor for the performance is the feature extraction. The better the features could capture the characteristics of the segments, the better the system could perform. Thus it is meaningful to study how the features would affect the performance, and what characteristics of the music they can capture and represent, that would lead to the result.

The features are extracted using Essentia [2] from short audio frames. Then the features are fed to a classifier, which is Support Vector Machine (SVM) in this case, to train a 3-class model. Each input of the classifier is classified into one and only one of singing, percussion, and instrumental categories. After the classifier is trained using the training data, same set of features are extracted from the test data and evaluated on the trained classifier.

After the classification step is done, a post-processing is performed to group these frames into audio segments which corresponds to singing, percussion, and instrumental. The final segmentation can be seen in Figure 3.1.

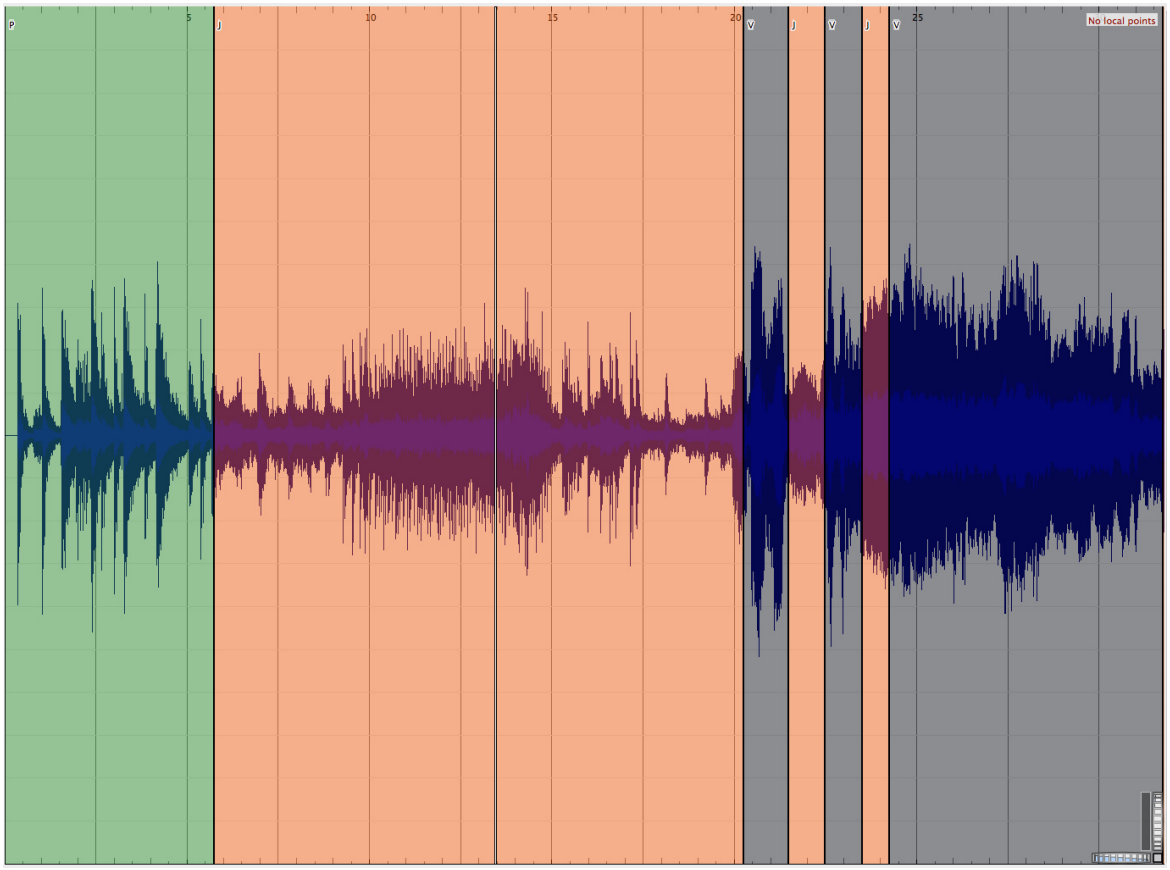

Fig. 3.1 Example of final segmentation result opened in Sonic Visualiser [3]. Green section is percussion. Orange section is instrumental. Grey section is singing.

A similar approach has been proposed in [5], which performed a classification of vocal and non vocal. Some code for feature extraction has been reused from the work. However, the aimed classes are different, and the features used and the settings are also different.

3.3 Features

The features should be able to distinguish different characteristics from singing, percussion, and instrumental. For singing part and the rest, the major difference is the presence of human singing voice, which possesses unique characteristics. For example, the timbre of human singing voice is quite different from other instruments used in Beijing opera arias. This trait can be captured with features such as MFCCs. The percussion instruments used are mostly unpitched, and their sound is inharmonic, which could be captured by the feature inharmonicity. More features can be selected by combining musical knowledge and the analysis of the spectral and temporal information. As a result, a list of features have been chosen that are considered suitable for this task:

- Frequency domain
  - Spectral spread
  - Inharmonicity
  - Spectral complexity
  - Spectral crest
  - Spectral flatness
  - Spectral flux
  - MFCCs
  - delta MFCCs
  - Spectral contrast
- Time domain
  - Zero crossing rate

This results into a feature vector of 44 dimensions. Several lengths of audio frame, which the features are extracted on, are tested. The final length of the audio frame is chosen at 25ms due to the better performance of the features extracted on this length.

3.4 Dataset

The dataset consists of 34 Beijing opera arias by 23 singers from a set of commercial CD recordings, which is a representative collection of Beijing opera arias. The total length of all recordings is 3 hours 40 minutes. Nearly each aria has all three segments, with the singing part takes the longest time, while the instrumental part has a shorter duration and the total duration of percussion part is the shortest. The audio files are in mono wav format with a sample rate at 44100 Hz. The segments are annotated with Praat [1] by experienced Beijing Opera musicologist.

3.5 Experiments

As stated above, the key to the performance of the system are the features. Thus, three sets of experiments have been carried out, with the focus on testing the performance of the features.

The first experiment is to simply use all the features. The result of the first experiment can be used as a reference for the other two experiments. Since perceptually the major difference between the classes are the timbral difference, which can be well represented by MFCC features, the second experiment uses only MFCC features, which contains 12 MFCCs and 12 delta MFCCs, resulting a feature vector of 24 dimensions. The result could reveal how well the MFCC features can achieve alone without the information from other features. For the third experiment, a feature selection method, namely 'InfoGain', is applied. Feature selection can reduce the redundancy of the features as well as reducing overfitting. The feature selection method evaluates the features with a measure of info gain, and produces a rank of features. The first 24 features (dimensions) are then selected from the ranking, and tested with the classifier. All three experiments are tested using the same classifier with the same training and testing data.

3.6 Evaluation and results

The evaluation is done with 3-fold cross-validation. The dataset is randomly divided into three folds, with two folds as the training set and the other one fold as the testing set. For classification, the input of classifier is the set of features extracted from short audio frames. In other words, the evaluation is on the frame level.

The measures used for classification evaluation are precision, recall, and F-measure. These are defined as follows, with tp referring to true positive, fp referring to false positive, and fn referring to false negative:

Table 3.1 Result of experiment 1

	Precision	Recall	F-measure
Singing	0.97	0.968	0.969
Percussion	0.805	0.863	0.833
Instrumental	0.931	0.92	0.925
Weighted Avg.	**0.942**	**0.0.942**	**0.942**

Table 3.2 Confusion matrix of experiment 1

	Singing	Percussion	Instrumental
Singing	**69529**	66	2265
Percussion	268	**8857**	1143
Instrumental	1884	2079	**45677**

$$Precision = \frac{tp}{tp + fp} \tag{3.1}$$

$$Precision = \frac{tp}{tp + fn} \tag{3.2}$$

$$F = 2 \times \frac{precision \times recall}{precision + recall} \tag{3.3}$$

The result of the first experiment, which applies all the features in this list, is shown in Table 3.1 and Table 3.2. It shows that the system is able to achieve high performance for the classification of singing, percussion, and instrumental, with high values in precision, recall, and F-meature. However, the scores of percussion are noticeably lower than the other two classes. A possible reason is that due to the shorter total duration of percussion, it has much fewer training materials for the classifier, thus is not recognized as well as other classes.

It is also interesting to notice that, from the confusion matrix, singing voice is much more likely to be mistaken as instrumental than percussion. The accompanying ensemble of Beijing opera has some high-pitched fiddles, which can be played at very high volume, thus quite dominant. They have similar timbre to the female human singing voice, which also has very high pitch. And the fiddles are played in a heterophonic texture with the singing melody. This could result in similar feature values and make it difficult for the system to decide either it's human voice or instrumental.

Table 3.3 Result of experiment 2

Precision	Recall	F-measure	
Singing	0.85	0.845	0.848
Percussion	0.734	0.656	0.693
Instrumental	0.742	0.764	0.753
Weighted Avg.	**0.8**	**0.8**	**0.8**

Table 3.4 Confusion matrix of experiment 2

	Singing	Percussion	Instrumental
Singing	**60748**	331	10781
Percussion	1111	**6740**	2417
Instrumental	9589	2111	**37940**

The results of experiment 2 is shown in Table 3.3 and Table 3.4. When the features are reduced to only MFCCs, the performance is not as good as the first experiment, which is expected. However, the result is not bad considering the fact that it doesn't have potentially helpful information from other features. It shows the effectiveness of MFCC features in this task, since it could capture the timbral characteristics well.

For the third experiment, the rank of features produced with the feature selection method 'InfoGain' is shown in Table 3.7. Higher score represents that the feature is more significant for the task based on this measure. Using the first 24 features from the ranking, which is the same number of features (dimensions) as the second experiment, the result is shown in Table 3.5 and Table 3.6.

The performance is better than the second experiment, with the same number of features. This proves that the feature selection method is able to find the more useful features from the task and reduce redundancy from the features. The second experiment could be seen as a manual feature selection, which in this case is outperformed by the automatic feature selection method.

The lower score of percussion class and the error mistaking singing voice as instrumental are still present, since the modification of features cannot address the potential underlying reasons that cause the problems.

3.7 Discussion

In this chapter, we proposed an approach for the segmentation of singing, percussion, and instrumental in Beijing opera arias. Three experiments have been carried out with the attempt to evaluate the influence and effectiveness of the features.

Table 3.5 Result of experiment 3

Precision	Recall	F-measure	
Singing	0.943	0.942	0.942
Percussion	0.78	0.766	0.773
Instrumental	0.877	0.881	0.879
Weighted Avg.	**0.905**	**0.905**	**0.905**

Table 3.6 Confusion matrix of experiment 3

	Singing	Percussion	Instrumental
Singing	**67695**	117	4048
Percussion	314	**7867**	2087
Instrumental	3814	2107	**43719**

From the results it can be found that the classification system is able to perform quite well with all the features applied. With only MFCC features, the performance is not as good. With feature selection method, the same number of features can yield better results than MFCCs, which proves the usefulness of the feature selection method.

Also, it is interesting to notice that, the singing voice is much more likely to be mistaken as instrumental in all cases. The reason behind might be that the timbre of instruments in Beijing Opera ensemble has similar feature to the (female) singing voice. To address this issue, new features can be used to better differentiate singing voice and instrumental.

Since very few computational studies have been down on Beijing opera music, it is hard to compare the results with other works. However, there are some possible ways to improve the current system. For example, features that better capture the unique characteristics of human singing, such as the subtle fluctuations, can be used. Also, increasing the data of percussion could possibly help too.

Table 3.7 Rank of features

Score	Ranked features
0.9957	5 pectral_complexity.mean
0.6276	6 spectral_complexity.var
0.5218	41 spectral_contrast.mean-3
0.4364	13 zero_crossing_rate.mean
0.4245	42 spectral_contrast.mean-4
0.4173	12 spectral_flux.var
0.4103	11 spectral_flux.mean
0.4093	40 spectral_contrast.mean-2
0.3959	9 spectral_flatness_db.mean
0.363	43 spectral_contrast.mean-5
0.2864	7 spectral_crest.mean
0.2625	22 mfcc.mean-8
0.2607	10 spectral_flatness_db.var
0.2393	15 mfcc.mean-1
0.2314	14 zero_crossing_rate.var
0.2215	20 mfcc.mean-6
0.1931	18 mfcc.mean-4
0.1896	16 mfcc.mean-2
0.1872	17 mfcc.mean-3
0.1755	28 mfcc.var-2
0.1745	19 mfcc.mean-5
0.1725	32 mfcc.var-6
0.1718	34 mfcc.var-8
0.1594	35 mfcc.var-9
0.1557	39 spectral_contrast.mean-1
0.1551	31 mfcc.var-5
0.154	37 mfcc.var-11
0.1536	36 mfcc.var-10
0.1533	23 mfcc.mean-9
0.1469	8 spectral_crest.var
0.1446	44 spectral_contrast.mean-6
0.1427	33 mfcc.var-7
0.1387	30 mfcc.var-4
0.137	29 mfcc.var-3
0.1367	38 mfcc.var-12
0.1247	27 mfcc.var-1
0.1203	26 mfcc.mean-12
0.1107	25 mfcc.mean-11
0.1045	24 mfcc.mean-10
0.1008	3 inharmonicity.mean
0.0958	21 mfcc.mean-7
0.071	4 inharmonicity.var
0.0232	1 harmonic_spectral_spread.mean
0.023	2 harmonic_spectral_spread.var

Chapter 4

Lyrics-to-Audio Alignment

In this chapter we present detailed information of the work in lyrics-to-audio alignment in Beijing opera arias. We start by introducing the background knowledge of Beijing opera lyrics, followed by the challenges we face in this task. Then the proposed rule-based duration aware system is explained. It is tested on the capella singing dataset created with manual annotation. Evaluation results show that the proposed method is able to handle long syllables in Beijing opera singing, and is able to utilize the duration information to improve the performance.

4.1 Beijing opera lyrics

Lyrics play a crucial role in Beijing Opera arias. Beijing Opera arias are usually presenting a story, and the lyrics set the basic emotion and structure. The singing of the lyrics is the most representative part of Beijing opera arias.

Beijing Opera lyrics are mostly in the form of couplets, where each line in the couplet has the same number of characters. In mandarin Chinese, each character is equivalent to one syllable. Thus syllable and character are interchangeable concepts throughout this thesis. Within each line, there exists smaller units called 'dou', which consists of 2 to 4 characters. An example of Beijing opera lyrics can be found in Figure 4.1.

The unique way of organizing the lyrics makes it very interesting to study the structure of the lyrics along with the audio. Lyrics-to-audio alignment is a suitable task for this analysis. One of the various applications of this task is the navigation with lyrics, which would be beneficial to musicological studies and music education.

有贺后在金殿一声高骂，
骂一声无道君细听根芽。

老王爷为江山何曾卸甲，
老王爷为山河奔走天涯。

遭不幸老王爷晏了御驾，
贼昏王篡了位谋乱邦家。

把一个皇太子逼死殿下，
反倒说为嫂我拦阻有差。

贼好比王莽贼称孤道寡，
贼好比曹阿瞒一点不差。

Fig. 4.1 Example of an excerpt of lyrics in Beijing opera aria. Four couplets are shown in this figure.

4.2 Lyrics-to-audio alignment with Beijing opera arias

The task of lyrics-to-audio alignment aims to estimate the temporal relationship and match the singing voice and the corresponding lyrics. Apart from existing challenges for general lyrics-to-audio alignment, in the case of Beijing opera, there are a couple of more challenges that have been posed. One is the lack of training data. Due to the highly skilled vocal production in Beijing opera singing, it is very different from general Chinese speech. Thus the adaption from speech would not perform very well. So it is better to train the acoustic models directly from singing data. However, most commercially accessible materials have background instrumental music accompanied with the singing voice, which makes it very difficult to train acoustic models with this influence. To tackle this, datasets of clean singing voice (a cappella singing) have been created. The other challenge for this task is the long syllable duration in Beijing opera singing. A single syllable could last up to 10 seconds without pulse. Detailed information of this will be covered in the dataset section. To address this problem, a lyrics-to-audio system that could handle long syllable duration is proposed.

4.3 Approach

For the task of lyrics-to-audio alignment with Beijing opera aria, we propose an approach that could handle the special structure of Beijing opera lyrics based on syllable duration.

An overview of the proposed system is presented in Figure 4.2. For the training phase, which is at the left end of the figure, each phoneme is built as a Gaussian Mixture Model

(GMM) trained on annotated a cappella singing. The first 13 MFCCs and their delta and delta-delta are extracted from 25ms audio frames with the hop size of 10ms. The extracted features are then fit into a phoneme GMM with 40 components, a number of components usually proved as sufficient in speech recognition. A model for silent pause is added at the end of each syllable, which is optional on decoding. This allows to accommodate the frequent for regions of pauses after some syllables.

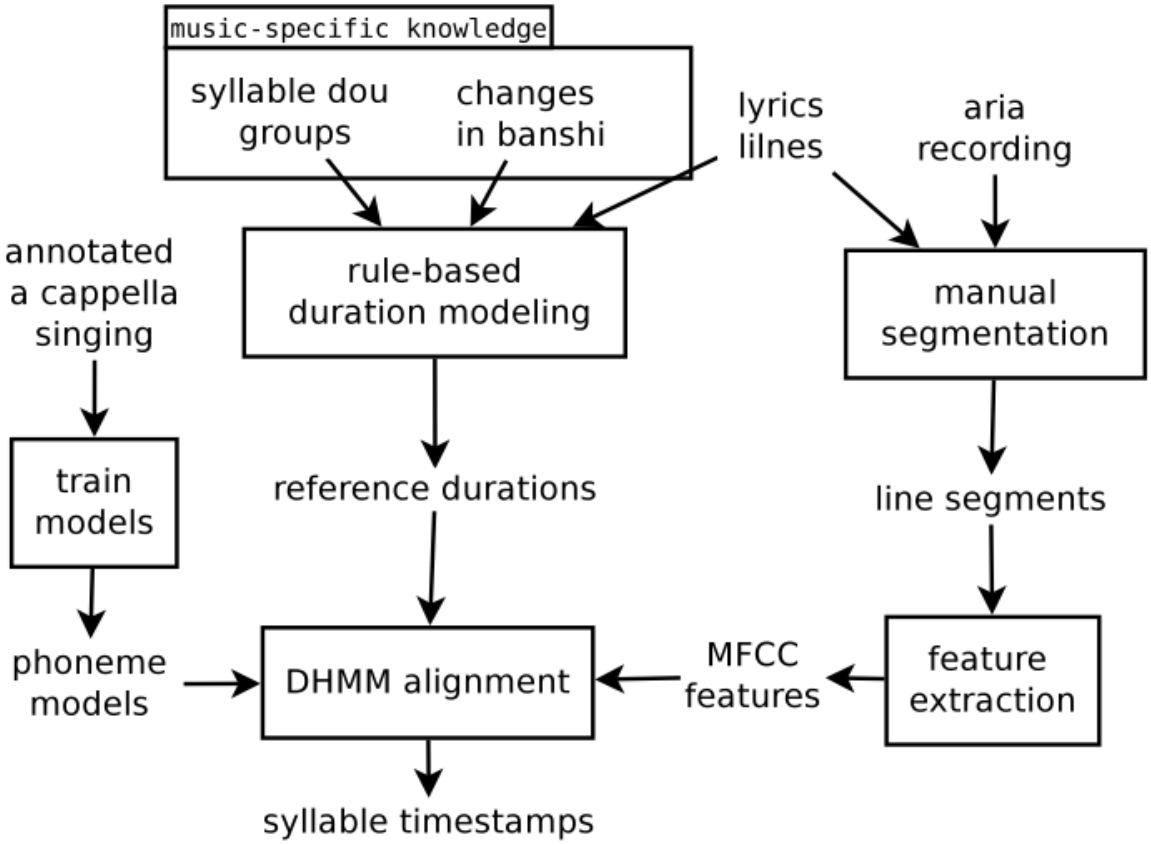

Fig. 4.2 Approach Overview: The middle column shows how reference durations are derived based on music-specific knowledge.

The music-specific knowledge includes syllable dou groups, which are the smaller units within each lyrics line, and the changes of banshi, which refers to the rhythmic patterns. We use the assumption that the last syllable of each dou group is more likely to be prolonged, so as the last syllable before the change of banshi. Thus, this information, i.e. the rule, could be used to inform the system to optimize for certain situations. It is realized by assigning a reference duration to certain syllable.

The alignment step is done with the Duration-aware Hidden Markov Model (DHMM), which is able to handle long syllable duration by setting a large variation value for the duration. Detailed information of this system can be found in [9] and [8].

For the testing step, the audio recording of arias are first manually segmented into lyric lines. Then the same set of features are extracted from these lines. These features are provided to the DHMM to perform the alignment and produce the final time stamps for each syllable.

4.4 Datasets and statistical analysis

Two datasets have been created for this task. The first consists of excerpts from 15 arias of two female singers, chosen from a commercial CD collection of CompMusic corpus [24]. The total length of this dataset is 1 hour and 17 minutes. The audio is in mono wav format with sample rate at 44100 Hz. For each aria, there are two versions present: a recording with voice plus accompaniment, and an accompaniment-only one. Thus the clean singing vocie can be generated by subtracting the instrumental accompaniment from the complete version in the frequency domain. The resulted audio has only high quality clear singing voice left.

Each aria is annotated on the phoneme level by native Chinese speakers and Jingju musicologist. The phoneme set is derived from Chinese pinyin, and represented using the X-sampa standard. Certain phonemes are grouped into phonetic classes as new models based on their perceptual similarity, making a final set of 29 phonemes in total.

The second dataset is made of 14 aria excerpts recorded in studio from 3 singers. The recorded materials are only a capella singing without any accompanying instrument. The total length of this dataset is 20 minutes. The audio is in mono wav format with sample rate at 44100 Hz. The phonemes are annotated by a native Chinese speaker using the same criteria defined in the first dataset.

Statistical analysis has been performed on the datasets, which reveals very interesting characteristics of Beijing Opera structure based on lyrics. As mentioned above, the lyrics of Beijing opera arias are mostly present in the form of couplets, which consists of two lines. The number of syllables in lyrics line may vary. However, based on musical experience, some numbers of syllables are more likely to occur than others. Figure 4.3 and Figure 4.4 shows the statistics of the number of syllables for each lyric line of the two datasets. From the figures we can tell that in dataset 1, the most frequent numbers of syllables for each line are 10, 7, and 8 syllables. For dataset 2, the numbers are similar, with 7 as the most frequent number of syllables.

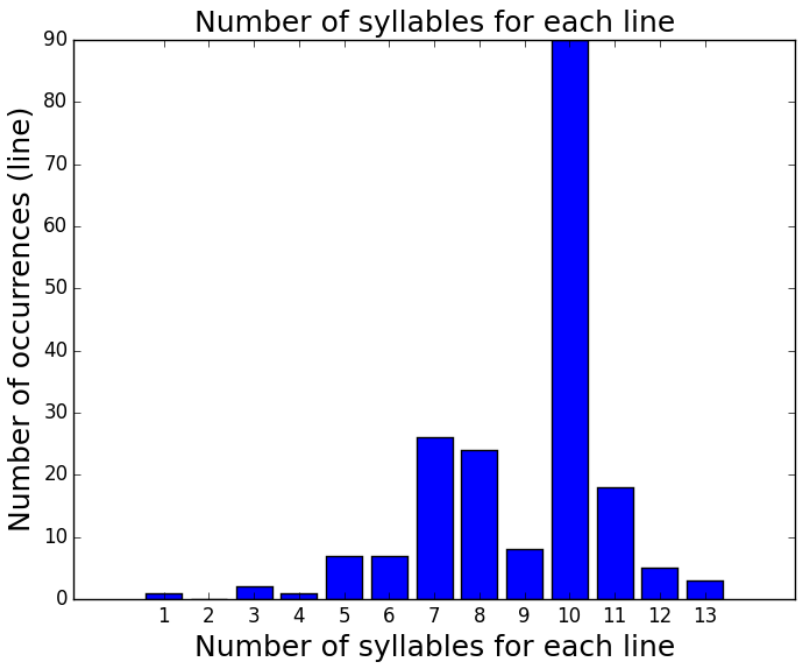

Fig. 4.3 Number of syllables in each lyrics line (Dataset 1).

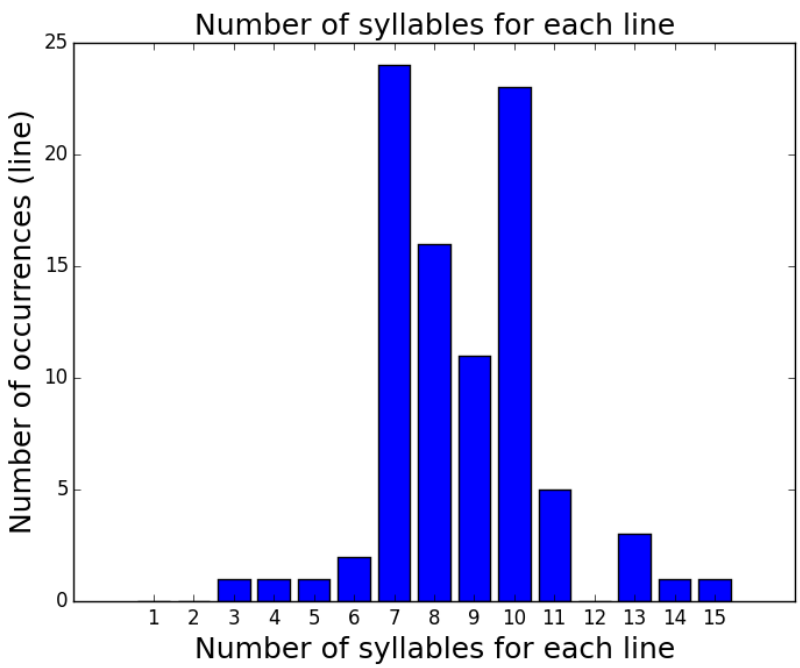

Fig. 4.4 Number of syllables in each lyrics line (Dataset 2).

Table 4.1 Duration statistics (ms)

	Min	Max	Mean	Median
Dataset1	43.31	10876.57	979.43	544.54
Dataset2	36.60	13073.61	873.93	488.17

Another interesting characteristic of Beijing Opera lyrics is the long duration of syllables. The long duration is reflected in two aspects: the average duration is much longer, compared to Mandarin speech, which has been reported as 248.98ms according to [30]. Also, the longest duration is significantly longer than the average duration, which is a common phenomenon in Beijing opera singing, where a certain syllable might be prolonged to a high extent, and sometimes repeated after a short pulse to produce an even longer duration. Figure 4.5 and Figure 4.6 show the phoneme duration distributions, with only vowels considered in this case as an approximation of the syllable duration, since the consonants are normally quite short. From the figures we can see that there are a lot of occurrences of phonemes which have durations over 500ms.

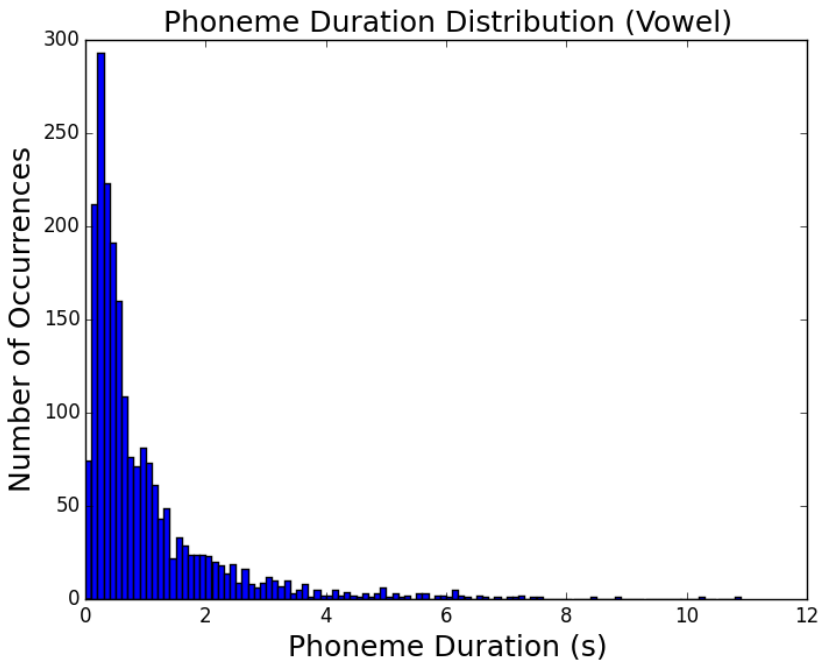

Fig. 4.5 Phoneme duration distribution (Dataset 1).

Table 4.1 shows the min, max, mean, and median values of the phoneme duration of both datasets. It is shown that the longest vowels have the duration over 10 seconds. Besides, the median and mean duration of vowels are also much longer than the mandarin speech.

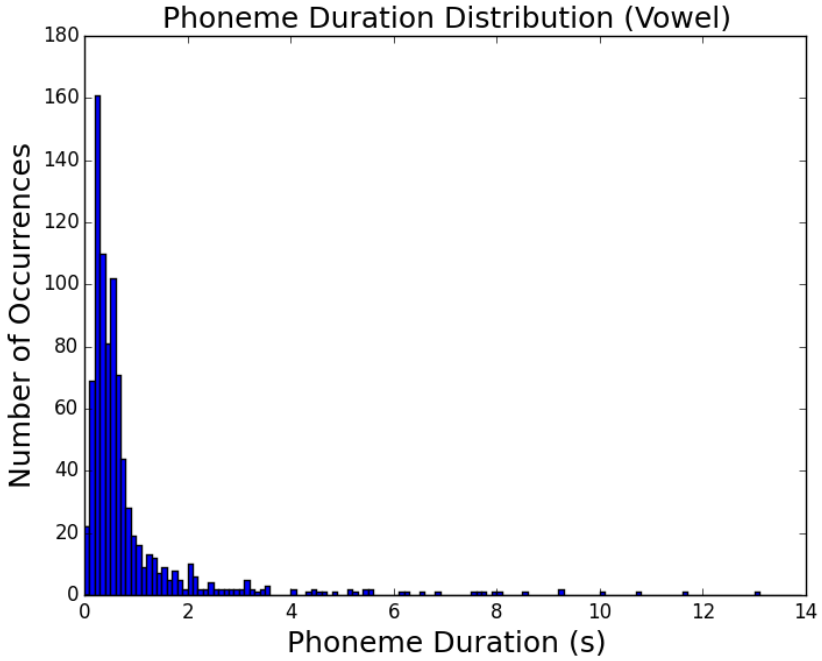

Fig. 4.6 Phoneme duration distribution (Dataset 2).

However, in a lyric line, the syllables at different positions are not equally likely to be prolonged. Syllables at certain positions are more like to be prolonged than others, which is another interesting characteristic of Beijing opera singing. To demonstrate this phenomenon, the syllable durations based on their position have been studied. Figure 4.7 and Figure 4.8 show the normalized syllable duration based on their position in the lines that have ten syllables, which is the most frequent case. The syllable duration is normalized by the length of the whole line.

It is clear that syllables at certain positions are have longer durations than others. In the case of 10-syllable lines, 3rd, 6th, and 10th syllable are more likely to be prolonged. These positions are also the last syllable of ‘dou’s, and the 10-syllable lines mostly consists of three dous with 3, 3, and 4 syllables. So this information could help to improve the lyrics-to-audio alignment system if system can handle possibly longer duration at certain positions.

4.5 Experiments

Due to the previous analysis of Beijing opera lyrics, the system to be evaluated has to address several challenges in the case of Beijing opera: first and foremost, if the phoneme models trained from a cappella singing could actually work and yield reasonable results; second, if

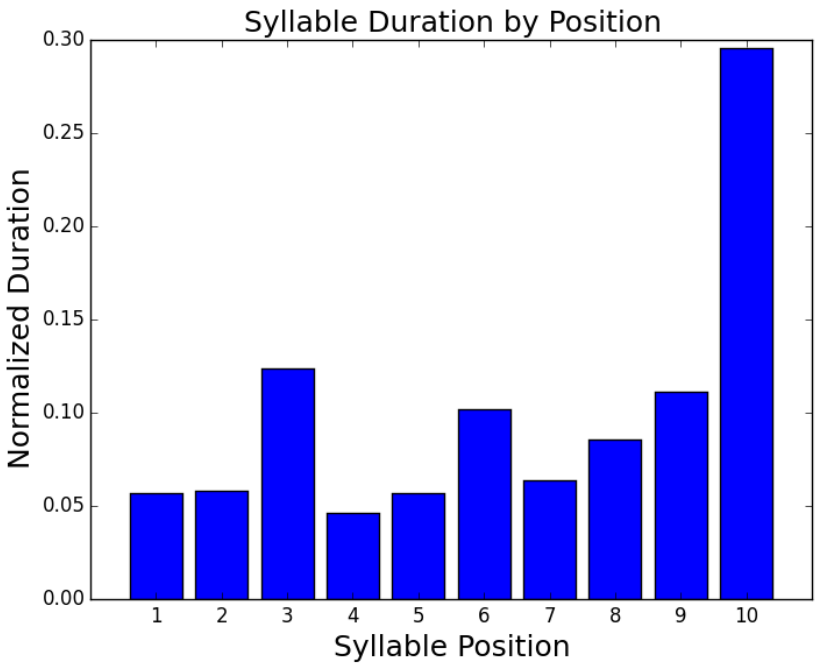

Fig. 4.7 Syllable duration based on position (Dataset 1).

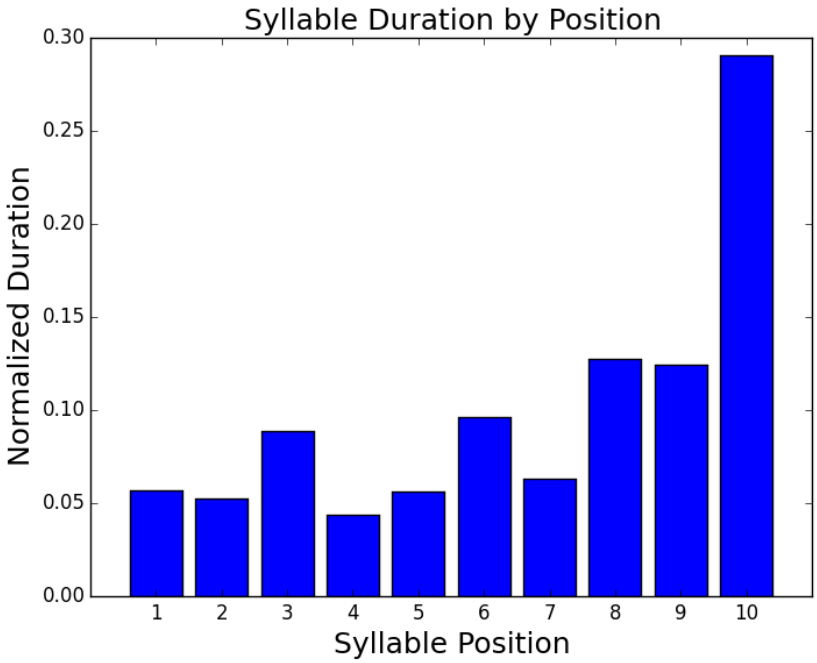

Fig. 4.8 Syllable duration based on position (Dataset 2).

the system could handle the generally long duration of the syllables; and third, if the system could be improved if given the knowledge on the syllable duration based on position.

With this in mind, two experiments have been conducted. One is the oracle experiment, which aims to test if the system could be able to handle long syllable durations if given correct duration information. More specifically, the phoneme annotation is used as an oracle to suggest the system with the duration of the syllable. In this case, duration probability of phoneme is set to 1 during its time interval and 0 otherwise. Another experiment is the baseline comparison, which uses Viterbi decoding from HTK Toolkit [31] as baseline method.

Only the dataset 1 is used throughout the experiments. Besides, a 'canonical' dataset, which is a subset of the dataset 1, is also created to test if the system could utilize the knowledge on the position-informed syllable duration, which is derived from previous analysis. The selected 'canonical' subset consists of lines which satisfies the assumption we made: syllables at certain position (key syllables) are more likely to be prolonged. And the key syllables are those that are at the end of a 'dou', or at the end of a 'banshi' before it changes. Thus, we kept only these lines, for which at most one key syllable is not prolonged, and discarded the rest. A syllable is considered as being prolonged if it is longer than 130% of the average syllable duration for the current line.

4.6 Evaluation and results

Evaluation is done with 3-fold cross validation, with 5 arias in each fold. Phoneme models are trained on 10 arias (2 folds), using the phoneme-level annotations, and evaluated on the last fold of 5 arias. Alignment accuracy is defined as the percentage of duration of correctly aligned syllables from total audio duration. Each lyric line is manually segmented and the lyrics-to-audio alignment is performed on the syllable level for each line. Accuracy is measured for each manually segmented line and accumulated on total for all the recordings.

The final results are shown in Table 4.2. For the oracle experiment, it is shown that the accuracy is able to reach close to 100 percent. This means that the system is generally capable of handling the long syllable duration in Beijing opera singing. For the baseline comparison, the proposed method outperforms the baseline method significantly. It is supposedly because of the ability that the proposed system has to handle high-varying syllable durations, which is something the general lyrics-to-audio alignment method cannot deal with well. For the comparison between the complete dataset and canonical dataset, the result of the proposed method is improved by 6.4%, which shows that the system is able to utilized the prior knowledge on the syllable duration based on position to improve the performance. However,

Table 4.2 Result of lyrics-to-audio alignment

	Baseline	DHMM	Oracle
Complete	56.6%	89.9%	98.5%
Canonical	57.2%	96.3%	99.5%

the baseline method doesn't have significant improvement on the canonical dataset, because the method cannot handle well the long syllable durations in the first place, thus it won't help even if it is given extra duration information.

4.7 Discussion

In this chapter, we presented the work on lyrics-to-audio alignment with Beijing opera arias. There is a lack of data in Beijing opera singing, and the skilled vocal production makes the singing voice very different from speech, thus the adapting data from speech would not yield satisfactory results. Facing this challenge, two datasets have been created with clean singing voice and annotated on the phoneme level.

Statistical analysis has been performed on the datasets, which revealed some musical characteristics of Beijing opera lyrics. One of them is the syllable duration that is much longer compared to Mandarin speech. The other finding is the relationship between the syllable duration and their position in the lyric line.

To address the characteristics of Beijing opera arias, a lyrics-to-audio alignment approach based on Duration-aware HMM has been proposed. With the evaluation of oracle test and comparison to baseline. It is proved to be able to handle the high-varying syllable durations in Beijing opera singing, and is able to utilized musical knowledge based on the syllable position to enhance the performance.

The annotation of Beijing opera singing is not an easy job. First of all, it cannot always be mapped accurately using the X-Sampa standard. Besides, it is sometimes very difficult to identify the boundaries of phonemes, which have no sudden transition. These could affect the quality of the annotation and thus the phoneme model. Also, due to these difficulties in annotation, the current datasets are still quite small, and might not cover well all the phonemes, thus influence the quality of the models. Increasing the size the of the dataset could be a good way to improve the performance. Also, the information from scores could also be incorporated into the system, though the amount of annotation work is also quite large.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Structure analysis is the essential study for Computational Musicology and Music Information Research. Despite of its own significance, it is fundamental for many of other researches as well. From the review of the state of the art, we find that in the field of MIR, most of the structure related research is oriented at western popular music. Thus the methodologies and tools have been designed to fit to the problems and characteristics of that particular type of music. This brings the limitations to the technologies when applied to music from other cultures, which have different nature and characteristics, and may pose questions that have not been answered before. Beijing opera is one of the cases, whose structure is constructed using a different compositional approach.

Two specific tasks have been chose to study the structure on different levels. The first task is the segmentation of singing, percussion, and instrument sections. The proposed approach selected features according to musical knowledge and how it is reflected in the audio signal. The classifier built on these features are tested on a annotated dataset of 34 arias. Three experiments have been conducted. The results show that the proposed approach is able to yield satisfactory classification results. And the experiments on the features show the effectiveness of MFCC features in reflecting the timbral difference. And the feature selection method is proved to be able to reduce the redundancy and help achieve better results with limited number of features.

The second task is lyrics-to-audio alignment. Two datasets have been created with human annotation on the phoneme level. The statistical analysis of these datasets have shown interesting characteristics of the structure of Beijing opera lyrics, namely the long duration of the syllables and the relationship between the duration and the syllable position in the lyric line. The experiments of oracle and baseline comparison have proven that the system is able

to handle the high-varying syllable duration in Beijing opera arias. Further more, experiment on the canonical subset of the dataset has shown that the system is able to utilize the syllalble position information to improve the performance.

5.2 Future work

The analysis of results of the segmentation of singing, percussion, and instrumental has shown that the errors can be made while confusing singing voice with some instruments, due to their similarity in timbre. Thus, better feature can be designed to capture the unique aspects of singing voice and the instruments. The lack of data for percussion sections has resulted in poorer performance in classifying this class. Thus increasing the size of percussion data could contribute to the system as well.

The lyrics-to-audio alignment is one of the very first attempts to address this task in the context of Beijing opera. Currently, the rule based duration model can be improved with the computed statistics of the data. However, this requires the size of the data to reach a certain point so that the statistics are significant enough to be generalized in order to correctly inform the system. Increasing the dataset can also improve the phoneme models, which can then have enough to be able to cover all the phonemes that have appeared, with different singers and singing styles.

5.3 Contributions

This is a summary of the contributions that have been made throughout the course of the thesis[1]:

- Propose and experiment on the approach for the segmentation of singing, percussion and instrumental sections in Beijing opera arias
- Study and analyze the influence of features for the segmentation task
- Create Beijing Opera singing datasets annotated on phoneme level
- Experiment on training phoneme models directly from singing data
- Reveal the characteristics of long syllable duration and lyrics-based structure of Beijing Opera arias

[1]To encourage reproducibility of research, the related data and code are available at http://compmusic.upf.edu/corpora

- The collaborated work on lyrics-to-audio alignment has won the Best Paper Award in the 2016 Folk Music Analysis workshop [9]

References

[1] Boersma, P. et al. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

[2] Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J. R., and Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. pages 493–498.

[3] Cannam, C., Landone, C., and Sandler, M. (2010). Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. pages 1467–1468.

[4] Chai, W. (2005). *Automated analysis of musical structure*. PhD thesis, Massachusetts Institute of Technology.

[5] Chen, K. (2013). *Characterization of Pitch Intonation of Beijing Opera*. PhD thesis, Citeseer.

[6] Cooper, M. and Foote, J. (2003). Summarizing popular music via structural similarity analysis. pages 127–130.

[7] Dzhambazov, G., Şentürk, S., and Serra, X. (2015). Searching lyrical phrases in a-capella turkish makam recordings. pages 687–693.

[8] Dzhambazov, G. and Serra, X. (2015). Modeling of phoneme durations for alignment between polyphonic audio and lyrics.

[9] Dzhambazov, G., Yang, Y., Repetto, R. C., and Serra, X. (2016). Automatic alignment of long syllables in a cappella beijing opera.

[10] Fujihara, H. and Goto, M. (2012). Lyrics-to-audio alignment and its application. *Dagstuhl Follow-Ups*, 3.

[11] Fujihara, H., Goto, M., and Ogata, J. (2008). Hyperlinking lyrics: A method for creating hyperlinks between phrases in song lyrics. pages 281–286.

[12] Fujihara, H., Goto, M., Ogata, J., and 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.

[13] Khine, S. Z. K., Nwe, T. L., and Li, H. (2008). Singing voice detection in pop songs using co-training algorithm. pages 1629–1632.

[14] Kruspe, A. M. and Fraunhofer, I. (2014). Keyword spotting in a-capella singing. 14:271–276.

[15] Lehner, B., Widmer, G., and Sonnleitner, R. (2014). On the reduction of false positives in singing voice detection. pages 7480–7484.

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

[17] Mauch, M., Fujihara, H., and 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.

[18] Mesaros, A. and Virtanen, T. (2008). Automatic alignment of music audio and lyrics.

[19] Nwe, T. L. and Li, H. (2007). Singing voice detection using perceptually-motivated features. pages 309–312.

[20] Nwe, T. L., Shenoy, A., and Wang, Y. (2004). Singing voice detection in popular music. pages 324–327.

[21] Paulus, J. and Klapuri, A. (2006). Music structure analysis by finding repeated parts. pages 59–68.

[22] Paulus, J., Müller, M., and Klapuri, A. (2010). State of the art report: Audio-based music structure analysis. pages 625–636.

[23] Rao, V., Ramakrishnan, S., and Rao, P. (2009). Singing voice detection in polyphonic music using predominant pitch. pages 1131–1134.

[24] Repetto, R. C. and Serra, X. (2014). Creating a corpus of jingju (beijing opera) music and possibilities for melodic analysis. pages 313–318.

[25] Rodríguez López, M. (2016). Automatic melody segmentation.

[26] Serra, J., Müller, M., Grosche, P., and Arcos, J. L. (2014). Unsupervised music structure annotation by time series structure features and segment similarity. *IEEE Transactions on Multimedia*, 16(5):1229–1240.

[27] Serra, X. (2011). A multicultural approach in music information research.

[28] Vembu, S. and Baumann, S. (2005). Separation of vocals from polyphonic audio recordings. pages 337–344.

[29] Wichmann, E. (1991). *Listening to theatre: the aural dimension of Beijing opera*. University of Hawaii Press.

[30] Xu, Y. and Wang, M. (2009). Organizing syllables into groups—evidence from f 0 and duration patterns in mandarin. *Journal of Phonetics*, 37(4):502–520.

[31] Young, S. J. and Young, S. (1993). *The HTK hidden Markov model toolkit: Design and philosophy*. University of Cambridge, Department of Engineering.

Appendix A

Terminology in Beijing opera

- Jingju: the Chinese for Beijing opera, sometimes also called Peking opera
- Shengqiang: the melodic patterns in Beijing opera
- Banshi: the rhythmic patterns in Beijing opera
- Luogu: a pre-defined set of percussion patterns in Beijing opera
- Dou: the smaller units consists of several syllables within a lyric line