Computational Modelling and Analysis of Vibrato and Portamento in Expressive Music Performance

Luwei Yang

A thesis submitted to the University of London
for the degree of Doctor of Philosophy

School of Electronic Engineering and Computer Science
Queen Mary University of London

2016

Abstract

Vibrato and portamento constitute two expressive devices involving continuous pitch modulation and is widely employed in string, voice, wind music instrument performance. Automatic extraction and analysis of such expressive features form some of the most important aspects of music performance research and represents an under-explored area in music information retrieval. This thesis aims to provide computational and scalable solutions for the automatic extraction and analysis of performed vibratos and portamenti. Applications of the technologies include music learning, musicological analysis, music information retrieval (summarisation, similarity assessment), and music expression synthesis.

To automatically detect vibratos and estimate their parameters, we propose a novel method based on the Filter Diagonalisation Method (FDM). The FDM remains robust over short time frames, allowing frame sizes to be set at values small enough to accurately identify local vibrato characteristics and pinpoint vibrato boundaries. For the determining of vibrato presence, we test two alternate decision mechanisms—the Decision Tree and Bayes' Rule. The FDM systems are compared to state-of-the-art techniques and obtains the best results. The FDM's vibrato rate accuracies are above 92.5%, and the vibrato extent accuracies are about 85%.

We use the Hidden Markov Model (HMM) with Gaussian Mixture Model (GMM) to detect portamento existence. Upon extracting the portamenti, we propose a Logistic Model for describing portamento parameters. The Logistic Model has the lowest root mean squared error and the highest adjusted R-squared value comparing to regression models employing Polynomial and Gaussian functions, and the Fourier Series.

The vibrato and portamento detection and analysis methods are implemented in AVA, an interactive tool for automated detection, analysis, and visualisation of vibrato and portamento. Using the system, we perform cross-cultural analyses of vibrato and portamento differences between erhu and violin performance styles, and between typical male or female roles in Beijing opera singing.

Statement of Originality

I, Luwei Yang, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: Luwei Yang
Date: 19 September 2016

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors Prof. Elaine Chew and Dr. Khalid Z. Rajab for their invaluable support during my whole PhD study and research. Their patience, immense knowledge and enthusiasm guide me to finish this thesis. I also would like to thank my independent assessor Dr. Simon Dixon who checked and looked my work at different stages.

I would like to thank MuPaE (Music Performance and Expression) group members, Madeleine Le Bouteiller, Adrian Gierakowski, Tomack Gilmore, Dorien Herremans, Katerina Kosta, Sarah Sauvé, Jordan Smith, Janis Sokolovskis, Carlos Vaquero, Bogdan Vera, and Simin Yang for their stimulating discussions and wide backgrounds.

I would like to thank everyone in C4DM for their kindness and willingness for help. To name a few, Mathieu Barthet, Dawn Black, Chris Cannam, Tian Cheng, Gyorgy Fazekas, Matthias Mauch, Andrew McPherson, Alessia Milo, Giulio Moro, Mark Plumbley, Mark Sandler, Siddharth Sigtia, Yading Song, Dan Stowell, Mi Tian, and Siying Wang.

A special thank to my father for his help in erhu recording selections and teaching me for erhu playing; my mother for her continuous support since I was born; and my girlfriend Yinqiu Zhang for taking care of me during the period in London.

Many thanks to Laurel S. Pardue and Jian Yang for their violin recordings, which are used in my PhD research.

Contents

List of Figures

List of Tables

List of abbreviations

AM	Amplitude Modulation
ANOVA	Analysis of Variance
BR	Bayes' Rule
df	Degrees of Freedom
DFT	Discrete Fourier Transform
DT	Decision Tree
ESPRIT	Estimation of Signal Parameters via Rotational Invariance Technique
f_0	Fundamental Frequency
FDM	Filter Diagonalisation Method
FFT	Fast Fourier Transform
FM	Frequency Modulation
GMM	Gaussian Mixture Model
GUI	Graphical User Interface
HMM	Hidden Markov Model
KDE	Kernel Density Estimation
KS	Kolmogorov-Smirnov test
MIDI	Musical Instrument Digital Interface
MIR	Music Information Retrieval
MUSIC	Multiple Signal Classification
PDF	Probability Density Function
STFT	Short-Time Fourier Transform

Chapter 1

Introduction

This chapter introduces the motivations for this thesis, providing the aims, and summarises each chapter and the main contributions. A list of publications associated with this work is also given.

1.1 Motivations

Music is a complex phenomenon comprising of entities and attributes such as melody, rhythm, timbre and silence. The music score is a symbolic format for representing music information where each note is assigned categorical pitch and duration properties. Musicians interpret and transform the score into vivid, lively, and expressive performances. Thus, music performance not only reflects the composer's intention but also blends this with the performers' interpretation and understanding. Kendall & Carterette (1990) wrote that music performance is a multi-side communication system where composers code the musical idea in basic notations, performers encode the notation into acoustic signals, and listeners decode the acoustic signals to ideas.

There are numerous ways for performers to create unique performances, although one performer may have a consistent preference for one way to interpret a piece of music. This allow performer to be distinguished one from the other. Palmer et al. (2001) ran a number of experiments to prove that listeners are able to distinguish between different performances through memorising instance-specific acoustic features, similar to speaker identification. Expression is an important aspect of music performance. It represents the value added to a composition and is part of the reason that music is interesting to listen to and sounds **alive** (Canazza et al., 2004).

A particular style of performing may not be designated only to a specific performer. One group of performers, one instrument, one music genre, or even one

culture can be identified by a specific expressive performance. Thus, research on expressive music performance can have direct impact on ethnomusicological studies around the world, the tracing of musical influences in musicological phylogenetic studies, and expressive performance pedagogy, analysis and synthesis. Despite these far-ranging application, expressive music performance has not received as much research attention as speech and linguistics.

There exist a large number of expressive devices, such as tempo variation, dynamic shaping, pitch variation (e.g. vibrato and portamento), and timbral modulation, to name a few. It would be an impossible task to model all expressive devices in a doctoral thesis. We have chosen to focus on vibrato and portamento, two of the most important and frequently used expressive devices in music performance created on string, woodwind, and brass instruments, and with voice.

Degrees of fluctuation are extremely difficult to measure by ear since their perception is influenced by complex physio-psychological factors (Gable, 1992). Regardless, it is still possible to model these expressive devices in computational and mathematical ways. We are interested in exploring the questions:

- how can we model vibrato and portamento?
- how can we automatically detect vibratos and portamenti from expressive music audio?
- how can we systemetise the large-scale study of expressive performance?
- how can the models be used to explore expressive performance in an analysis-by-synthesis approach?
- how can we create interactive computer music interfaces that can be used in computer-aided tutoring?

To answer these questions, we have developed new vibrato and portamento detection methods. We create a new method to quantitatively and computationally describe portamento. An interface toolbox integrating vibrato and portamento detection and analysis modules has also been created.

1.2 Aims

The main aim of this thesis is to develop better analytical models and computational methods for vibrato and portamento detection. The second aim is to integrate vibrato and portamento detection and analysis modules to create a system that can greatly speed and enhance vibrato and portamento analysis, systematic expressive performance research, music expression synthesis, and

computer-aided tutoring. Another aim is to demonstrate the method through case studies in the analysis of vibrato and portamento differences across performers, instruments, and music genres drawn from different cultures.

1.3 Thesis Structure

The outline of the thesis is as follows:

Chapter 2 starts with a description of music expressivity and reviews existing musicological literature on vibrato and portamento, including performance trends from recent centuries. The chapter describes changes in playing styles for these two expressive devices and summarises the types of each expressive device. We then present related work on expressive music performance modelling, and on computational vibrato and portamento modelling. This chapter also presents state-of-the-art vibrato detection methods with a focus on three frame-wise methods. Relevant single pitch detection methods are also surveyed.

Chapter 3 describes the vibrato modelling and detection. It begins with an introduction to the anatomy of vibrato, showing the parameters that can quantitatively characterise vibrato properties. Then, the chapter presents a novel approach to frame-wise vibrato detection and estimation in music signals using the Filter Diagonalisation Method (FDM). In contrast to conventional methods based on the Fast Fourier Transform, the FDM's output remains robust over short time frames, allowing frame sizes to be set at values small enough to accurately identify local vibrato characteristics and pinpoint vibrato boundaries. The FDM decomposes the local fundamental frequency into sinusoids and returns their frequencies and amplitudes, which the system uses to determine vibrato presence and parameter values. The results show the FDM-based techniques to consistently perform best in both frame-level and note-level evaluations. Furthermore, the FDM method with Bayes' Rule leads to better F-measure results—0.84 (frame-level), 0.41 (note-level)—than the FDM method with Decision Tree—0.80 (frame-level), 0.31 (note-level). The FDM methods' accuracy for determining vibrato rates are above 92.5%, and for vibrato extents are about 85%.

Chapter 4 investigates the feasibility of using the Logistic Model for modelling portamento. We compare models of portamento using the Logistic function, a Polynomial, a Gaussian, and the Fourier Series applied to pitch contours, where each model is constrained to six coefficients. The Logistic

Model is shown to have the lowest root mean squared error and the highest adjusted R-squared value; an ANOVA shows the difference to be significant. Furthermore, the Logistic Model produces musically meaningful outputs: transition slope, duration, and interval; and, time and pitch of the portamento inflection point. A case study comparing portamenti employed in erhu and violin playing the same musical phrase shows transition intervals to be piece-specific—as it is constrained more by the notes in the score than other factors—but transition slopes, durations, and inflection points to be performer-specific. A Hidden Markov Model-based method is employed to explore portamento detection. The results show that the HMM+GMM has better performance than HMM+Gaussian. However, the returns of increasing Gaussian mixture numbers quickly diminish, and so the performance does not significantly improve as the value increases. The addition of delta pitch and energy to the input does not improve portamento detection performance significantly, compared to the use of delta pitch alone.

Chapter 5 presents the Matlab toolbox created for the AVA system, an interactive visual and quantitative analysis system for vibrato and portamento, which integrates vibrato modelling and detection in Chapter 3 and portamento modelling and detection in Chapter 4. The system detects vibratos and extracts their parameters from audio input using a FDM-based method, then detects portamenti using a Hidden Markov Model and presents the parameters of the best fit Logistic Model for each portamento. The graphical user interface (GUI) allows the user to interact with the system, to visualise and hear the detected vibratos and portamenti and their analysis results, and to identify missing vibratos or portamenti and remove spurious detection results. The GUI provides an intuitive way to see vibratos and portamenti in music audio and their characteristics, and has potential for use as a performance analysis and pedagogical tool.

Chapter 6 presents case studies for investigating vibrato and portamento performance styles as found in recorded music audio. The first experiment investigates vibrato characteristic differences between erhu and violin playing of the same piece of music. A manual selection and annotation process has been employed to report the results. The second case study uses the AVA system to analyse vibrato and portamento in string instrument (erhu v.s. violin) playing and Beijing opera (Laosheng v.s. Zhengdan roles) singing. The AVA system is employed to detect vibratos and portamenti, and to refine the detection results; the parameters are output for further analysis.

Chapter 7 summarises the thesis and describes further research directions.

1.4 Contributions

The principal contributions of this thesis are as follows:

- Chapter 3: Presented a novel frame-wise vibrato detection and estimation method using the Filter Diagonalisation Method. FDM-based techniques are shown to consistently yield the best results in both frame-level and note-level evaluations, when compared to state-of-the-art methods.
- Chapter 4: Proposed a new mathematical and computational model, using the Logistic Model to quantify the portamento parameters. This, together with the Hidden Markov Model with Gaussian Mixture Model detection technique represents pioneering work on portamento detection.
- Chapter 5: Implemented and tested the AVA system, an interactive system for visual and quantitative analysis of vibrato and portamento.
- Chapter 6: Analysed the vibrato and portamento performing styles on erhu & violin and in Beijing opera singing datasets, reporting differences in the analytical results.

1.5 Associated Publications

Portions of the work detailed in this thesis have been presented in international scholarly conferences and international journals:

- Chapter 3: The FDM-based vibrato detection and estimation method (Section 3.2) was accepted for publication in the **Journal of Mathematics and Music** (Yang et al., 2017).

- Chapter 4: The Logistic Modelling for portamento (Section 4.1) was presented at the Fifth Biennial International Conference on Mathematics and Computation in Music (MCM2015) (Yang et al., 2015).

- Chapter 5: The integration of the vibrato and portamento modules and the implemented AVA Matlab interface was accepted at the 42nd International Computer Music Conference (ICMC2016) (Yang et al., 2016a) and as a part of a paper published at the 17th International Society for Music Information Retrieval Conference (ISMIR2016) (Yang et al., 2016b).

- Chapter 6: The analyses of vibrato performance style comparing erhu and violin (Section 6.1) have been published at the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR2013) (Yang et al., 2013). The related player-wise analyses have been published at the 4th International Workshop on Folk Music Analysis (FMA2014) (Yang et al., 2014).

  The vibrato characteristics of two roles in Beijing opera singing (Section 6.2.1) have been reported at the 5th International Workshop on Folk Music Analysis (FMA2015) (Yang et al., 2015).

  The analyses of vibrato and portamento for erhu & violin and the Beijing opera singing using the AVA system were published at the 17th International Society for Music Information Retrieval Conference (ISMIR2016) (Yang et al., 2016b).

Chapter 2

Background

In this chapter, we introduce the background and related work for this thesis. First, we discuss the concept of expressivity in music. A musicological review of vibrato and portamento in music performance follows. Then, we survey work on the modelling of expressive music performance followed by that on computational modelling of vibrato and portamento. Next, we focus on state-of-the-art vibrato detection methods, and conclude by showing evaluation results for state-of-the-art single-pitch detection methods.

2.1 Music Expressivity

As noted by Daniel Leech-Wilkinson (2009), "performing musically, or stylishly, involves modifying those aspects of the sound that our instrument allows us to modify, and doing it in a way that brings to the performance a sense that the score is more than just a sequence of pitches and durations." Music performance is not merely the realisation of the categorical pitch and duration information on the score. It is instead a vivid and lively object encoding the players' expressivity. A performance combines both the composer's ideas as notated in the score and the players' interpretation and understanding of the score.

As described in Kendall & Carterette (1990), music performance constitutes a multi-faceted communication system. Composers code their music ideas in common music notation, performers decode the notation into music ideas and encode the ideas into acoustic signals, and listeners decode the acoustic signals to ideas.

People are able to develop, and readily become sensitive to developing, preferences for particular performances of a music piece. Experiments reported by Palmer et al. (2001) prove that, similar to speaker identification, listeners are able to distinguish one performance from another through the memorising of

instance-specific acoustic features.

The music properties or elements added by performers are called *music expressive devices*, which communicate music expressivity. From a cognitive psychology point of view, music expressivity can transfer emotions (Juslin & Sloboda, 2001). Generally speaking, faster tempi and larger dynamic variations present high arousal (happy and angry) emotions, whereas slower tempi and smaller dynamic ranges convey low arousal (leisurely and sad) emotions. Musical form and structure, sometimes indicated by a composer in the music score (Lerdahl & Jackendoff, 1983), can be implied by a performer through the manipulating of expressive devices (Randel & Apel, 1986).

We make a distinction between performance style and expressive interpretation of music. The strategies a performer employs to expressively perform a composition leads to an expressive interpretation of the music. Performance style can refer to the systematic acoustic changes common to performers of a time period; it can also refer to that typical of a particular performer. Beyond performance styles and individual interpretation, Timmers (2007) describes general performance strategies to communicate structure and emotion across diverse performing styles.

Many researchers have offered definitions of music expressivity. According to Snyder (2000),

> The patterns of rhythm, melodies, and so on that we are able to remember from music consist of sequences of musical categories. Each occurrence of a category, however, is shaded in a particular way by its nuance, which constitutes the expressive aspects of the music.

This *nuance*, or so called *expressivity*, is defined as "continuous variations in the pitch or rhythm of a musical event". Besides pitch and rhythm, other elements such as dynamics and timbre can be manipulated in expressive performance.

Palmer & Hutchins (2006) later gave the following definition of music expressivity,

> Performers add variation to music; they manipulate the sound properties, including frequency (pitch), time, amplitude, and timbre (harmonic spectrum) above and beyond the pitch and duration categories that are called "musical expression".

Again, it is clear that there are multiple dimensions to music expressivity, including *time*, *pitch*, *amplitude*, and *timbre*.

There has been a recent surge in expressive performance research. According to (Fabian et al., 2014), expressiveness or expressivity is

1. the effect of auditory parameters of music performance (loudness, intensity, phrasing, tempo, frequency spectrum, etc.) – covering acoustic, psychoacoustic, and/or musical factors;
2. the variation of auditory parameters away from a prototypical performance, but within stylistic constraints (e.g. too much variation is unacceptable, and does not fall within the gamut of expressiveness);
3. used in the intransitive sense of the verb (no emotion or mood or feeling is necessarily being expressed; rather the music performance sounds "expressive" to different degrees).

Here, expressivity is more clearly restricted to the auditory factors of music performance. The extent of the manipulations and variations of these auditory parameters should lie within certain constraints. Excessive expressive variations may, in turn, impact the original aesthetics. A clear distinction is made between expressivity and emotion. Music performance evokes emotions in listeners; however, no specific emotions need to be connected to specific expressive devices. Finally, music expressivity depends on historical and cultural contexts. In this thesis, changes in vibrato and portamento use will be described chronologically in Sections 2.1.1 and 2.1.2.

Tempo and timing is perhaps the most studied aspect of expressivity in music. Tempo variation is essential to modern music performance. The composer may indicate the tempo of the whole or part of a piece, for instance, using text descriptors such as *Largo* (a slow tempo), *Adagio* (slow but not as much as Largo), *Allegro* (moderately fast), and *Presto* (very fast). Furthermore, the composer may use *Rubato* to mark the freeing of tempo and timing. To a large extent, performers possess the freedom to alter the tempo continuously to communicate their unique interpretations. Accurate tempo control is very important as the serial organisation of notes without time constraints would not be understandable, and inaccurate temporal control in music performance would decrease the aesthetic qualities of a performance (Vorberg & Hambuch, 1978).

The altering of pitch is another manifestation of music expressivity. Except for instruments without the capability of changing pitch such as the piano, many instruments—including string, voice, woodwind and brass—allow performers to manipulate the pitch in the process of playing a note. One of the widely adopted pitch variation devices is vibrato. Vibrato is the systematic and periodic modulation of pitch, usually accompanied by amplitude and timbral fluctuations (Sundberg, 1994). Another pitch variation device is portamento, the audible and continuous sliding between two notes of different pitches (Brown,

1988).

Loudness variation is one of the most effective ways for musicians to convey expressivity. Sometimes, loudness is referred to as "dynamics", "amplitude", or "intensity". Similar to tempo, instructions for loudness levels and variations can be found in music notations—for instance, *Crescendo* (becoming louder), *Decrescendo* (becoming softer), and *Forte* (very loud), to name a few.

Articulation is linked to loudness variation and duration and timbral control. On string instruments, articulation is controlled through the bowing. For example, different violinists may play the passages of successive detached notes in a moderate-to-fast tempo quite differently; the bow could bounce or remain firmly on the string. Brown (1988) presents an example from the first movement, *Allegro con brio*, of Beethoven's String Quartet Op. 18 No. 6:

> The passage, for instance, can quite effectively be played either with a *spiccato* or *sautillé* bowing in the middle (which is the bowing most modern violinists would use), a broad bowing in the upper half, a *martelé* bowing near the point, or a slurred *martelé* (staccato) in the upper third of the bow. Each of these methods of performance produces a markedly different effect.

In most cases, there is no indication on the score on which bowing is preferred. Thus, the interpretation depends on the violinist's choice. The above dimensions of expressivity are not independent one from another. Instead, musicians combine them to create expressive performances.

Music expressivity does not merely exist in Western classical music; it transcends music cultures and styles. For instance, the expressive devices in popular music may differ from those found in classical performance, and may include audible signs, glissandi, or variations in the spatialisation of sounds. Dibben (2014) explored the possibilities and limits of performance expressivity in popular music recordings. She reported that the popular music not only maintains the expressive devices in classical music, but also has some mechanistic and robotic expressive devices not included in classical music expression. Focusing on the legacy of Louis Armstrong, Bauer (2014) shows that jazz musicians use the manipulation of vocal gestures and vocalisations on trumpet to create expressive gestures. Ashley (2014) reports that expressivity in funk music relies more on the variation of rhythm and timbre than tempo and dynamics.

van der Meer (2014) examines the ways in which expressive vocal performance techniques are used to trigger emotion in Hindustani music. Subtle vocal pitch inflections are employed to encode emotion in order to bring the characteristics of *raga* [1] to life; however, it is hard to specify these emotions. Lippus

[1] In Indian music, *Raga* comprises of a set of melodic rules to which the musician must

& Ross (2014) focuses on the Estonian songs and investigates to what extent variations in the duration of song syllables are determined by their phonetic length. The correspondence between linguistic duration and musical rhythm in Estonian was found to be loosely defined. It inspired further research on the tone-tune relationship in melodies performed in a tone language (e.g. a variety of Chinese and various African languages). In the African Bedzan Pygmies' music, Marandola (2014) highlights the challenge to study the expressivity when no reference can be made to a typical, or even prototypical version of a piece. Nonetheless, two different levels of expressivity, the individual and collective level, can be identified.

In the following section, we will focus on vibrato and portamento, two key expressive devices, from a musicological view, in chronological order. Vibrato and portamento trends have changed over the years; according to (Philip, 1990):

> In the early years of the century, there was a clearer and more detailed differentiation between levels of expression in a piece of music—between accented and unaccented notes, between long and short notes, between portamento and non-portamento, between vibrato and non-vibrato, and between faster and slower passages.
>
> ...The trend in later years, and continuing into the late 20th century, was towards greater evenness and regularity of expression—evenness of rhythmic emphasis and of tempo, regularity of vibrato, avoidance of disruptive portamento, and a style of rubato based on gradual flexibility rather than rhythmic distortion.

In this thesis, we focus on vocal and string instruments as they represent the media through which vibrato and portamento are frequently applied. Sundberg (1977) describes the vocal organ is an instrument consisting of a power supply (the lungs), an oscillator (the vocal folds), and a resonator (the larynx, pharynx and mouth). The nature of the vocal organ accords the instrument a high degree of freedom to create expressivity. Sundberg (1998) shows that singers are able to modulate various parameters, such as tempo, vibrato, ascending glides (portamento) to target pitch, shaving off or sharpening loudness and pitch contours, to create expressivity and convey emotions. Generally, string instruments are considered to most closely mimic the human voice. This may be due to the fact that the nature of the instruments allows performers the freedom to apply parameter modulations similar to that are used in singing. Furthermore, we also explore definitions of vibrato and portamento, and the use of these expressive devices across music cultures, in Chinese as well as Western

adhere. It is also considered an entity in the sense of a living being that must be brought to life before the audience (van der Meer, 2014).

classical music.

2.1.1 Vibrato

In general, vibrato is the periodic low-frequency—around 5-8Hz (Fletcher, 2001)—modulation applied by a performer to a steady musical sound. Vibrato is usually produced by conscious physical manipulations such as the regular oscillation of the left hand of a string player. However, in some cases, at least for experienced singers, the vibrato arises unconsciously and naturally through the oscillation of abdominal and laryngeal muscles. The typical vibrato rate of 5 to 8Hz is close to the resting alpha rhythm of the human brain; thus, the generation and the perception of vibrato may be closely related to innate human physiological and psychological characteristics (Fletcher, 2001). More details of the psychological aspects can be found in Seashore (1938). Skilled musicians are able to vary the rate and extent of vibrato as the notes of the melody develop. This section reviews vibrato use in Western and Chinese music.

Vibrato in Western Music

When singing, if the voice is sounding near its maximum loudness or breath-pressure level, some relief of the larynx muscular tension is required to maintain the sound production for any length of time. This repeated slackening of the muscles and their return to a high-tension level causes pitch and intensity fluctuations known as vibrato (Gable, 1992).

The preferences for vocal vibrato has not always remained the same. Before the 20th century, vibratos were applied as ornamentation to notes less frequently. It is deemed that the vibrato was introduced in the context of a combination of *messa di voce*, whereby the voice crescendos and then vibrato is introduced. At that time, vocal vibratos had smaller extents. In addition, sometimes the vibrato did not exist for an entire note; instead, it had a delayed start with the beginning of the note characterised by clean pitch intonation (Gable, 1992).

From the 20th century onwards, vibrato is employed as part of normal tone production. The use of vibrato may have been precipitated by large opera houses that require continual loud singing, which in turn led to singers developing the vibrato muscular movement habit. As a result, vibrato has been adopted as a natural means of sound production, with increasing vibrato extent. Earlier musicians would consider today's sound production as a cover-up for faulty or careless intonation. Consequently, Seashore (1938) advocated for a reduction in vibrato extent. However, his suggestion did not have a large impact on musical practice. Howes et al. (2004) showed that singers are able to manipulate vibrato characteristics to convey emotions and expressivity in the same way that they

modulate dynamics and tempi.

In violin playing, vibrato is produced by oscillating finger, wrist, or arm movements that shifts the pitch above and below the target pitch (Stein, 2016). The speed, width and oscillation of the finger control the sonic properties of the vibrato. The pitch variations are typically accompanied by time, amplitude, and timbral modulations.

Vibrato is the most widely used expressive device in violin playing, but trends in vibrato use differs from that in vocal music. During the nineteenth century, vibrato was not encouraged in violin playing. According to Leopold Auer (1845–1930), a pupil of Joseph Joachim (1831–1907):

> Like the *portamento* the *vibrato* is primarily a means used to heighten effect, to embellish and beautify a singing passage or tone. Unfortunately, both singers and players of string instruments frequently abuse this effect just as they do the portamento, and by so doing they have called into being a plague of the most inartistic nature, one to which ninety out of every hundred vocal and instrumental soloists fall victim. (Auer, 1921)

Similarly, Joachim–Moser *Violinschule* writes: "A violinist whose taste is refined and healthy will always recognize the steady tone as the ruling one, and will use the vibrato only where the expression seems to demand it." (Joachim & Moser, 1905)

Nonetheless, the intensifying of vibrato use was brought to a peak by Wicniawsky (1835–1880). By this time, "beauty and nobility of sound were increasingly equated with a continuous vibrato tone, produce by the left hand, whereas before they had been equated with a steady and pure tone produced by the bow" (Brown, 1988). After the death of Joachim in 1907, this new idea of vibrato was widely accepted by the public. As noted by Hodgson (1916), "I see no harm in its presence at all times if the player has such a perfect control that he can reduce it at will to such a slight movement as to be inconspicuous and emotionless."

However, written instructions on how to execute vibrato are infrequent. An exception is found in Baillot (1834): "place one finger on the string, keep the other three fingers raised and rock the left hand as a unit with a more or less moderate movement, so that this rocking or shaking of the left hand is conveyed to the finger on the string." One interesting thing of note is that he required that the pure note should be heard at the beginning and the end which is quite similar to the practice of earlier vocal vibratos.

There is no clear and consistent sign for vibrato use in the music score. Louis Spohr (1784–1859) classified the vibrato speeds into four different cate-

gories: fast, slow, speeding-up, and slowing-down. These four types of vibratos are indicated by different wavy lines (see Figure 2.1). In this period, the condition that a pure tone should be heard at the beginning and end of the note was relaxed. Charles de Bériot (1802-1870) also used similar wavy lines to indicate vibratos but gave no detailed differentiation between vibrato speeds. Joachim used the wavy lines and the word *vibrato* interchangeably. César Franck (1822–1890) preferred the word *vibrato* and Edward Elgar (1857–1934) wrote 'ff "vibrato" molto espress'.

Figure 2.1: Spohr's four vibrato signs. top left: fast, top right: slow, bottom left: speeding-up, bottom right: slowing-down. Reproduced from (Brown, 1988).

Vibrato in Chinese Music

Vibrato is also extensively employed in Chinese music. The erhu, a Chinese bowed string instrument, is one of the main instruments in the Chinese orchestra and is widely considered to be the *Chinese Violin* (see Figure 2.2). The instrument has two strings and one bow. The bow is operated by the player against the string. Strings' vibrations are transferred to the snake skin-covered resonating box via a small bridge. The pitch change is implemented by pressing the strings with fingers. Note that there is no fingerboard for the two strings which means that there is nothing supporting the player's finger pressing the string except the string itself.

Vibratos in erhu playing aims to mimic the human voice to make the sound expressive, lively, and colourful (Yang, 2012). Normally, one can find three vibrato playing techniques for erhu (Yang, 2008) as shown in Figure 2.3. The *sliding vibrato* is a traditional technique formed by sliding the finger up and down the string. This process, usually led by the wrist and palm, creates a periodic change in the string's length. This technique is widely used in the Chinese compositions to imitate Chinese opera singing. The *pressing vibrato* technique uses the finger to press the string to change the string's tension. Usually, the finger is led by the hand muscles in the palm. The *rolling vibrato* is a new technique that emerged only in the early part of the twentieth century, and has its origins in violin vibrato technique. In this kind of vibrato, the finger rolls on the string, instead of sliding or pressing on the string, controlled by periodic

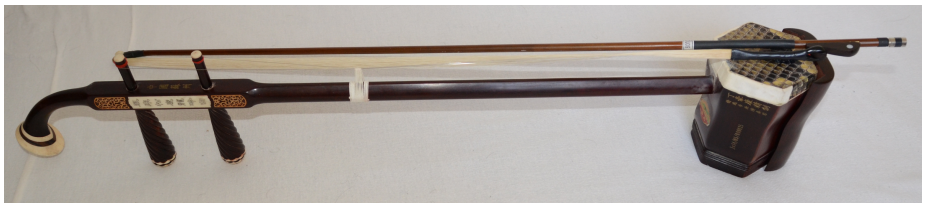


Figure 2.2: Erhu.

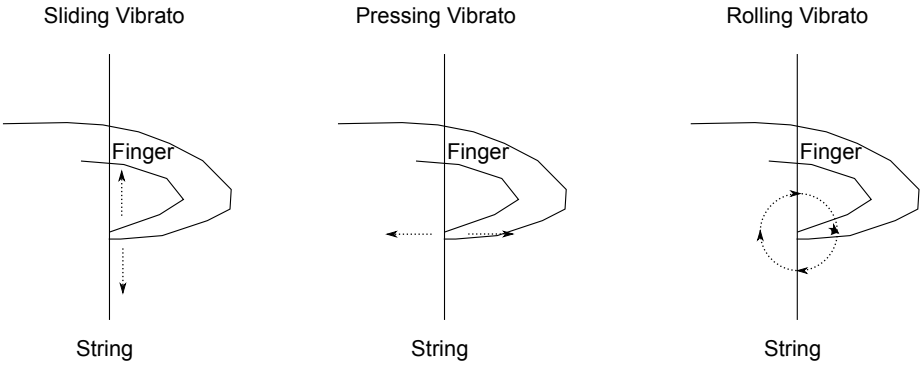


Figure 2.3: Erhu vibrato types. Left to right: sliding vibrato, pressing vibrato and rolling vibrato.

wrist movements. Due to the lack of a fingerboard, the rolling vibrato in erhu is different from that in violin in that it is always mixed with some degree of pressing vibrato (Wang, 2012).

Similar with Western music notation, there is no explicit indication for vibrato use in erhu compositions. When, where, and how vibrato is executed depends on the erhu player. Vibrato use forms an important criterion by which to evaluate performers' expression. Vibrato should not be applied indiscriminately to all notes. Vibrato use is connected to the desired affect and the characteristics and styles of the music piece (Yang, 2010). For instance a delayed vibrato used in conjunction with crescendo has an extremely poignant effect if used in the right place at the right time.

2.1.2 Portamento

Portamento is a widely used expressive device, especially in vocal and string music. This section reviews portamento use in Western and Chinese music.

Portamento in Western Music

The portamento was first described in the literature by Western musicians and music educators. Before the invention of the gramophone, Giovanni Battista

Mancini (1714–1800) described the function of the portamento as "...the blending of the voice from one tone to another, with perfect proportion and union, in ascending as well as descending" (Potter, 2006). More from García (1856): "To slur (portamento) is to conduct the voice from one note to another through all intermediate sounds." Figure 2.4 shows a portamento exercise from Melba (1926), in which a portamenti are indicated by a slur between two notes separated by variable distances.

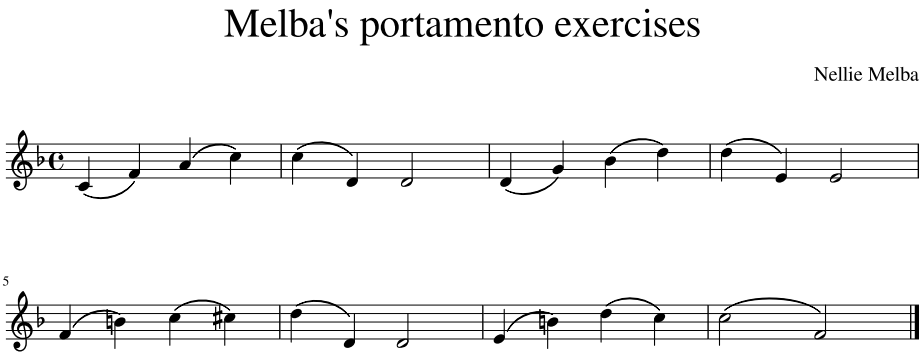


Figure 2.4: Melba's portamento exercise, where portamenti are indicated by slurs. Reproduced from (Melba, 1926).

Portamenti, which continuously slide though all intermediate pitches between two different pitches, should not be confused with *legatos* [2]. Legato is characterised by the passing "... from one sound to another in a neat, sudden, and smooth manner,...; yet not allowing it to drag or slur over any intermediate sound... As an example of this we may instance the organ and other wind instruments, which connect sounds together without either portamento or break" (García, 1856).

Portamenti are slides between two distinct notes of variable length. On the shorter end of the spectrum is the *swoop* (Leech-Wilkinson, 2006), a fast slide lasting between $50ms$ and $300ms$. The *glissando* sits on the longer end of the spectrum. Proper use of portamento is one criterion through which to distinguish good singers. According to the music critic and singing teacher, Klein (1991),

> One attribute of the art of phrasing that immediately distinguishes the accomplished vocalist is the correct employment of the portamento, both in the upward and downward movement of the voice.

[2] We use the terms *portamento*, *glissando*, *glide* or *slide* interchangeably; they each refer to the same thing, which is the continuous note transition from one note to another. Sometimes, the glissando can be distinguished from the portamento by sounding the intermediate pitches neatly, discretely and suddenly.

Portamento has been a significant expressive device employed in performance for over two hundred years. It is suggested that portamento draws on innate emotional responses to human sound and on our earliest memories of secure, loving communication, in order to bring to performance a sense of comfort, sincerity, and deep emotion (Leech-Wilkinson, 2006).

However, it has recently been refused by musicians for several decades (Leech-Wilkinson, 2006; Potter, 2006). The Second World War marks the start of a decline in the use of portamento in sung performances. Intensely expressive performances, which included the generous use of portamenti, were considered old-fashioned, exaggerated and unrealistic. This radical change in taste was prompted by many factors. One reason may be that the political context and cultural change led to a cynicism that eliminated the naivety reflected by expressive devices such as portamento (Leech-Wilkinson, 2006). With the decline in portamento use came an increase in vibrato use in classical and romantic music performance, which became increasingly common in sound production (see Section 2.1.1).

Although there has been a decline in portamento use, it was never abandoned by musicians; furthermore there are signs showing its return in the early 2000s. Leech-Wilkinson (2006) suggests some reasons for why this may be the case. First, portamento use may help to attract audience's attention, leading to closer associations between singer and audience. Another reason is that the portamento is used for maintaining a traditional sense of continuity prevalent in earlier operatic singing. Third, portamento is considered as the link between singing and speaking, showing singers' unique characteristics and expressivity.

Nevertheless, we cannot deny the fact that portamento is an effectively expressive device in music performance, regardless of the innocence and naivety with which it is associated in earlier performance practices or its use to mark irony and laughter after the Second World War.

Compared to singing, the use of portamenti in violin playing is more complex in its types, a result of the constraints on sound production. Louis Spohr, a German violinist and composer, wrote on portamento in his violin methods book: *Violinschule* (Spohr, 1852):

> The violin possesses, among other advantages, the power of closely imitating the human voice, in the peculiar sliding from one tone to another, as well in soft as in passionate passages. [...] The sliding must be made so quick... as not to make a vacancy or break appear in the slide, between the lowest and highest notes.

As a free-pitch instrument, the violin allows the musician to closely imitate the human voice especially in the effect of sliding from one note to another.

Violin portamento can be categorised into three types (Milsom, 2003). In *Type-1* portamenti, the same finger slides between two notes. This is the simplest and most commonly used portamento type in both ascending and descending directions. Figure 2.5 shows an example of such a portamento created using the same first finger to slide from B4 to F♯5. It should be noted that the bending of a note, i.e. slight lowering or raising of the pitch followed by a return to the target pitch, is not usually considered a portamento. Pitch bends are prevalent in Jewish music (Stein, 2016), and in Beijing opera (Yang et al., 2015) and Flamenco singing (Gómez & Bonada, 2013). If these pitch bends are classed as portamento playing, they would fall under Type-1.

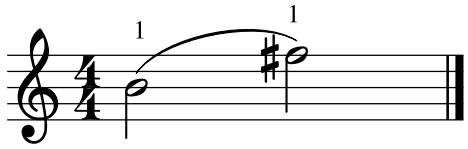


Figure 2.5: Type-1 portamento, where sliding is executed by the same finger. Reproduced from (Lee, 2006).

The *Type-2* portamento is produced by sliding between two notes using more than one distinct fingers. There are three sub-categories of Type-2 portamenti: the B-portamento, the L-portamento, and the combination of B- and L-portamenti.

The *B-portamento* uses the finger pressing the string for the precedent note to slide to an intermediate position, another finger then presses a different string for the subsequent pitch. This is sometimes called a *French slide*. Figure 2.6 shows an example with a B-portamento between the pitches B4 and F♯5; in this B-portamento, the first finger[3] covers the range B4 to D5; as soon as the third finger arrives at the correct position, it drops down to play F♯5 and the first finger is lifted off the string. This process can take place in either an upward or backward direction.

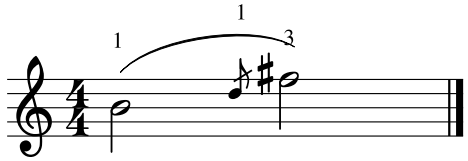


Figure 2.6: Type-2(B) portamento, where the first finger slides from the preceding note. Reproduced from (Lee, 2006).

The *L-portamento* is executed using the finger that will hit the target note,

[3]In violin playing, the first finger is the index finger instead of the thumb as in piano playing.

and is typically applied only in the ascending direction. The executing finger slides from an intermediate note to the target pitch. This is also called a *Russian slide*. An example is shown in Figure 2.7. In this example, the third finger slides from D5 to F♯5.

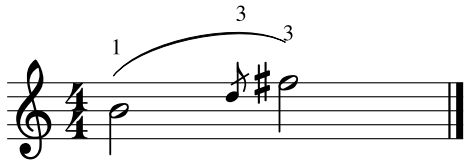


Figure 2.7: Type-2(L) portamento, where the third finger performs the slide to the subsequent note. Reproduced from (Lee, 2006).

The *BL-portamento* is the combination of a B- and L-portamento. Consequently, there are two intermediate notes for this slide: one is the byproduct of the B-portamento, and the other is the byproduct of the L-portamento. The example in Figure 2.8 shows the first slide from B4 to E5 using the first finger, followed by a second slide from F♯5 to A5 using the second finger.

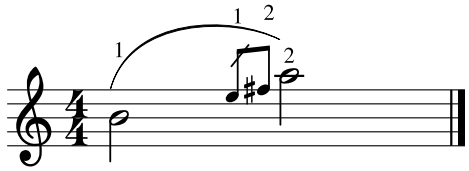


Figure 2.8: Type-2(BL) portamento, a combination of a B- and an L-portamenti. Reproduced from (Lee, 2006).

The *Type-3* portamento results from the swapping of the finger on the same note. According to Milsom (2003), this portamento type is used to alter the tonal quality, and the ways to do this include:

> Placing a finger on the primary note, placing a lower finger on a lower "intermediate" note, and sliding back up to the primary note. Placing a finger on the primary note, sliding down with that finger, and placing a higher finger back on the primary note. Placing a finger on the primary note, placing a higher finger on a higher "intermediate" note, and sliding back down to the primary note. (Stein, 2016)

The sliding range for this type of portamento is small, no larger than four semitones. Thus, it is used more as an ornament than as a device for connecting notes.

Similar to singing, portamento use in string playing became less common after the Second World War. As observed by Rosand (2014), "in the 1960s a change in violin playing began—a trend towards more literal and accurate interpretations of the written note and a purification of performance mannerisms."

Portamento in Chinese Music

Despite the reduction in portamento use in Western music, portamenti continue to be widely and actively employed in other musical cultures, such as in Chinese music.

The use of portamento is essential to the performance of erhu compositions, especially to communicate emotions (Yang & Zhen, 2015). The erhu has only two strings and no finger board. The prevalence of portamento use in erhu playing is due in part to the physical constraints of the instrument, necessitating shifts of the hand position to transition between notes, and in part to the general cultural style of erhu music.

The types of portamenti employed in erhu playing are similar to those in violin playing. The one-finger portamento in erhu playing is identical to the Type-1 portamento in violin (Zhao, 1999). The multiple-finger portamento category is the same as the Type-2 portamento. There also exists three sub-categories of Type-2 portamenti, including Type-2(B), executing the portamento using the precedent finger and using another finger to press the target pitch; Type-2(L), which is the converse of Type-2(B), where the portamento is executed by the finger that will hit the target pitch; and Type-2(BL), the combination of L and B subtypes.

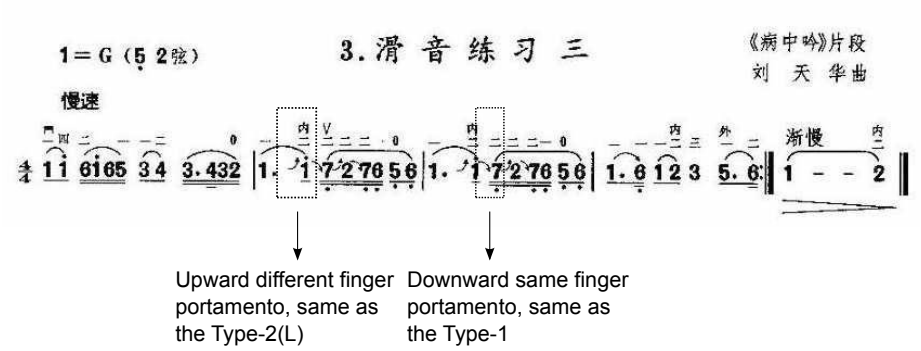


Figure 2.9: Erhu portamenti from a part of the erhu composition *Bingzhongyin* (Liu, 1930).

Figure 2.9 shows the erhu portamenti from a part of the erhu composition[4]

[4]Note that in traditional Chinese music, a larger number of compositions are written in numeric notation where numbers 1–7 represent the musical notes in the diatonic major scale

Bingzhongyin[5] (Liu, 1930). The portamento direction is indicated by a downward curving arrow, ⌒↘, or an upward curving arrow, ⌒↗, to the left of the note. An upward multi-finger portamento (Type-2) is indicated in Figure 2.9. The fingering[6] above the preceding note 1 indicates the first finger, and the second note $\dot{1}$ implies the second finger, suggesting that a Type-2 portamento should be used. An upward mark ⌒↗ near the $\dot{1}$ suggests that a Type-2(L) portamento should be used. As in violin playing, the fingering determines if the portamento is Type-1 or Type-2. Another example is the downward portamento using the same finger (the second finger) from $\dot{1}$ to $\dot{7}$. Note that the performer is also free to use portamento when there is no indication in the score.

Other portamento types exist in erhu playing that are not part of violin playing. One special portamento type, the *intermediate portamento*[7] (Zhao, 1999), in erhu playing involves three fingers, notated as an arrow with a dot in the middle, ⌒•↘. This portamento type requires 1–3 fingers or 2–4 fingers[8], where the second or the third finger, respectively, plays the intermediate role. At the start, all three fingers press on the string. The downward intermediate portamento begins from the high pitch finger led by the wrist. When the high pitch finger touches the intermediate finger, the high pitch finger is released from the string and passes the portamento on to the intermediate finger. The intermediate finger slides on the string until it reaches the lower pitch finger, which is on the target pitch. The upward direction starts from the lower pitch finger using the same principles.

There is another popular portamento type in erhu playing, the *round portamento*[9] (Zhao, 1999; Ling, 2007). This portamento is executed by a finger sliding to a higher (or lower) pitch and back to the original pitch again. A further variant of the round portamento is called *S portamento*, which is a concatenation of the upward and downward round portamenti, forming an S shape. The round and S portamenti are used more as embellishments rather than for connecting two notes.

In addition, there are categories based on the sliding range, *large portamento* and *small portamento*. The large portamento refers to any slide spanning a distance larger than a major third (four semitones) or a minor third (three semitones); the small portamento refers to a slide traversing a distance smaller than or equal to a major or minor third.

---

and dots above or below indicate higher or lower octave(s), respectively.

[5]In Chinese, 《病中吟》.

[6]In erhu composition, Chinese numbers are used to indicate the fingering. Here is the translation, 一: first, 二: second, 三: third, 四: fourth.

[7]In Chinese, 垫指滑音; in pinyin, dianzhi huayin.

[8]As in violin playing, the first finger in erhu playing is the index finger.

[9]In Chinese, 回转滑音; in pinyin, huizhuan huayin.

2.2 Expressive Music Performance Modelling

There has been increasing interest in the modelling and analysis of expressive music performance using mathematical and computational models, and engineering technologies. In this section, we give a brief summary of recent research and directions in expressive music performance modelling.

Widmer (2003) proposed an ensemble learning method, the PLCG (**P**artition + **L**earn + **C**luster + **G**eneralise), that combines multiple models into one final rule set via clustering, generalisation and heuristic rule selection, to investigate simple and robust performance rules from music recordings, showing that it is possible to find novel and musically meaningful discoveries this way. Widmer & Goebl (2004) reviewed four computational models of expressive music performance. The *KTH Model* (Friberg et al., 2000), which is a rule-based performance model; the *Todd Model* (Todd, 1992), which comprises of structure-level models of timing and dynamics; the *Mazzola Model* (Mazzola & Göller, 2002), which proposes mathematical models of musical structure and expression; and a *Machine Learning Model* (Widmer, 2002; Widmer & Tobudic, 2003a), which combines note-level rules with structure-level expressive patterns. An empirical evaluation of these four models has been reported. These models capture common performance principles , i.e. commonalities between performances and performers. There is room for further research focusing on "expressive intentions" and the way in which they are expressed, or new control spaces and devices for music performance.

As timing and dynamics are two of the most important expressive devices in music performance, these aspects of expressivity have attracted much music research. Repp (1995) examined the relationship between the global tempo and expressive timing microstructure in musical performances, and found that an increase in tempo was associated with a decrease in expressive timing variation. The timing and dynamics analyses of Chopin's *Etude in E major* were reported by Repp (1998, 1999). Widmer & Tobudic (2003b) learned the timing and dynamics at different levels (e.g. note-level and phrase-level) from real recordings to predict expressive timing and dynamics. An interactive interface "Air Worm" allowing users to manipulate the tempo and loudness is proposed by Dixon et al. (2002, 2005). It provides a musical interface for non-expert music-lovers to manipulate parameters of music expressivity.

Pitch modulations are also widely used to generate music expressivity. Devaney et al. (2011) demonstrated the feasibility of using the discrete cosine transform to characterise the singing voice fundamental frequency. It was shown to be useful for describing similarities in the evolution of fundamental frequencies in different notes. Devaney et al. (2012) also created an automatic music

performance analysis and comparison toolbox in Matlab. This toolbox provides functions to extract characteristics related to pitch for performance comparison. Using novel time-frequency analysis techniques, Jure et al. (2012) proposed an alternate pitch intonation and tuning analysis scheme using improved melodic content representation from polyphonic audio. It enables more precise representation of pitch fluctuations. Özaslan et al. (2012) analysed the use of expressive devices in pitch, i.e. vibrato and kaydırma [10], in Turkish makam music. Vibratos in Turkish makam music modulate at rates between 2 to 7Hz, which differs significantly from those in Western music (4-12Hz) stated by Desain & Honing (1996).

Instead of focusing on one or two specific features, some researchers employ multiple features to model expressive performance. Poli et al. (1998) analysed expression by different performers in their recordings of a piece of music, proving that performers' expressions can be identified from note acoustic parameters. Sundberg (1998) showed that singers were able to incorporate meaningful modulations (such as tempo and vibrato extents) on different parameters to convey emotional ambiances. Moelants (2004) analysed tremolo, trill, and vibrato use in bowed string instruments (e.g. violin, viola, cello and double bass). They found the rates of all three to be similar; low pitched notes were performed slower, and more advanced players performed faster. Sung & Fabian (2011) analysed Bach's *Suite No. 6 for Solo Cello* recorded between 1961 and 1998, focusing on the tempo, rhythmic flexibility, vibrato, portamento and articulations. In contrast to the common belief in a "general globalisation style," a wide variety of different performance choices were found to prevail since the last decade of the 20th century as a result of the interaction between mainstream and historically informed performance practices. Liebman et al. (2012) proposed a novel phylogenetic analysis approach to music performance analysis. Ten different categorical features (e.g. bowing, vibrato, duration, tempo, etc) are extracted from recording audio. The phylogenetic tree shows the relationship between different performances; some relationships observed contrasted with previous assumptions. Li et al. (2015) used a score-informed method to classify different expressive terms in note-level features, including dynamics, durations and vibratos. The contrast of feature values between expressive and non-expressive performances have been found to be important in modelling musical expression.

Some researchers focused on detecting and extracting of expressive devices [11]. Maestre & Gómez (2005) proposed an automatic feature (dynamics and fundamental frequency) extraction scheme relating to musical expressivity from

[10] A Turkish makam music term, the literal translation is *sliding*. The purpose of this behaviour is to give the feeling of non-edge connections all through the piece rather than sliding between notes. The closest equivalent Western music term is *portamento*.

[11] A survey of vibrato detection research is given in Section 2.3

monophonic musical audio. They extracted features at different scales, from features relating to an analysis frame to global features for entire performances. This description scheme was shown to be reliable for representing expressivity. To capture the expressivity in a performance, Friberg et al. (2007) presented a system for extracting tempo, sound level, articulation, onset velocity, spectrum, and vibrato rate and extent parameters from monophonic music audio recordings. Barbancho et al. (2009) proposed a transcription system for violin music with detection functions for expressive devices such as vibrato, pizzicato, tremolo, and spiccato. Parameters for these expressive devices are determined by time and frequency domain characteristics. Kóküer et al. (2014) presented an automated detection scheme for ornaments in Irish traditional flute playing; audio signal envelopes and fundamental frequencies were employed to detect the ornaments.

2.2.1 Computational Vibrato Modelling

Research on vibrato in Western music dates back to the beginning of the 1930s, when Seashore (1932) analysed vibrato in the singing voice and other instruments. Desain & Honing (1995) explored the algorithmic descriptions of vibrato accompanied with portamento, and Desain et al. (1999) investigated the rhythmic aspect of vibrato. Increase of vibrato rates towards the end of a note and relationships between tempo and vibrato rates were found. Prame (1994, 1997) reported vibrato rates and extents of Western singing voice. He found that, averaged across 10 singers, vibrato rates had a mean of 6.0 Hz and extents a mean of 71 cents. It was noted that singers tended to increase the vibrato rate towards the end of the note. Bretos & Sundberg (2003) examined vibrato in long crescendo sung notes, and confirmed Prame's finding that vibrato rates increased towards the ends of the notes. The means with which singers changed the vibrato rate as they tried to match a target stimulus was explored in (Dromey et al., 2003). Bloothooft & Pabon (2004) found that the vibrato rate and extent became more unstable as singers aged. Amir et al. (2006) examined the assessment of vibrato quality of singing students instead of accomplished professional singers. They found that the features extracted using the FFT and autocorrelation of the pitch contour performed well for predicting vibrato existence. Geringer & Allen (2004) investigated vibratos by music students, but for violin and cello, showing no significant difference in vibrato rates between instruments or performing experience. MacLeod (2008) found musicians used faster rates and wider extents during high pitches; and wider extents in *forte* passages. Mitchell & Kenny (2010) presented research on how singing students' vibratos improved over time by examining their vibrato rates

and extents. They found that the standard deviation of the vibrato rates decreased and the vibrato extent increased significantly with practice. Driedger et al. (2016) proposed an approach to analysing vibrato from the spectrum directly instead of the fundamental frequency, which was shown to be more robust than fundamental frequency-based strategies.

The perception of vibrato has also been subject to research. The relationship between vibrato characteristics and perception in Western singing was examined by Howes et al. (2004). d'Alessandro & Castellengo (1994) showed that pitch perceived for short vibratos were different from that for long vibrato tones, and proposed a numerical model consisting of a weighted time average of the f_0 pattern for short vibrato pitch perception. Diaz & Rothman (2003) showed that vibratos considered to be good (as rated by subjects) were the most periodic ones, and also that vibrato extent was the dominant factor for determining the quality of the vibrato. Verfaille et al. (2005) proposed a perceptual evaluation of vibrato models that also considered spectral envelope modulation. Fritz et al. (2010) investigated the perception of violin notes while varying the magnitude of the vibrato and the damping modes.

Vibrato synthesis forms another focus of research. Xue & Sandler (2008) introduced the use of harmonic sinusoids in the analysis and synthesis of vibrato. The proposed method is capable of retrieving vibrato properties and modifying them to create new vibratos. Gu & Lin (2008) employed an artificial neural network to generate vibrato parameters for singing voice synthesis. More vibrato synthesis systems can be found in Mellody & Wakefield (2000); Meron & Hirose (2000); Järveläinen (2002); Gough (2005); Roebel et al. (2011); Zhu et al. (2014).

2.2.2 Computational Portamento Modelling

Comparing to vibrato, portamento has not received as much attention in music analysis. A type of continuous note transition, it is prevalent in music for string, voice, and other instruments. Portamento is sometimes referred to as "glissando", "glide", or "slide". We shall use the terms "portamento" and "continuous note transition" interchangeably. The other form of note transition is the discrete note transition, which is the default mode in piano and keyboard playing. In this mode, the player is unable to or does not wish to alter the pitch in the process of moving from one note to another.

The definition of portamento from a mathematical point of view is ambiguous, and there is a lack of technical and scientific literature on portamento. Upon examining violin portamenti from eight master violinists, Lee (2006) found that the violinists tend to use portamenti as a highly personalised device to show off their musicianship. Liu (2013) investigated violin glide differences between ca-

dential and non-cadential sequences, comparing their proportional duration and intonation. Maher (2008) focused on vibrato synthesis over portamento transitions, and found that the vibrato rate should be in phase with the note onset so that the note duration is an integer multiple of the vibrato period. Krishnaswamy (2003) explored pitch perception, including vibrato and portamento, in South Indian classical music.

2.3 Vibrato Detection

In this section, we focus on prior work related to vibrato detection. Rossignol et al. (1999) first proposed five methods—based on spectrum modelling, spectral envelope distortion, AR prediction, and analytic signal and minima–maxima detection—to detect, estimate, extract, and modify vibrato. More recently, in the literature, there exist two classes of vibrato detection methods: note-wise and frame-wise methods. *Note-wise methods* require a note segmentation preprocessing step, usually carried out via manual annotation, before determining if the note contains vibrato. This makes real-time detection impossible. *Frame-wise methods* can be applied in real-time, by dividing the audio stream, or the extracted f0 information, into a number of uniform frames. Potential vibrato notes in the frame are then detected.

2.3.1 Note-wise Methods

Rossignol et al. (1999) presented the first note-wise vibrato detection method: for each note, the peaks and troughs of f0 are identified, then their periods are obtained and compared to the typical vibrato period range. Pang & Yoon (2005) proposed a probabilistic approach: they modelled the probability of vibrato existence as the product of two probabilities, one related to the vibrato rate, and the other to the normalised extent. Another note-wise method is presented by Özaslan & Arcos (2011), who compared the peaks and troughs of f0 to an ideal vibrato model, a uniform distribution of peaks and troughs. Weninger et al. (2012) outlined several feature extraction techniques based on f_0, root mean square energy and auditory spectrum of each note, then applied a classifier on these features for singing vibrato recognition in polyphonic contexts.

2.3.2 Frame-wise Methods

We now turn our attention to frame-wise methods. Figure 2.10 shows a flow chart describing a general frame-wise method. The time-varying signal is first subdivided into overlapping frames. The fundamental frequency, f_0, or together

with amplitude[12], $\mathcal{A}$, of the audio signal is extracted from each frame to form a time series. The time series extracted serves as input to the feature extraction module where salient features of the vibratos are determined. Finally, vibrato existence is determined via a decision-making mechanism. We next describe the three state-of-the-art frame-wise vibrato detection methods. Table 2.1 summarises the key components of the above three frame-wise methods.

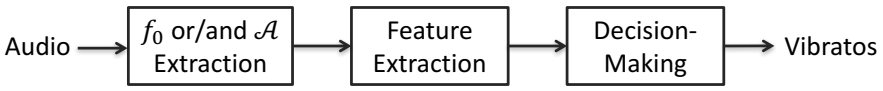


Figure 2.10: Basic framework of frame-wise vibrato detection methods.

Method	Input	Feature Extraction	Decision-Making
Herrera–Bonada	f_0	STFT(f_0)+Parabolic Interpolation	DT(F)
Ventura–Sousa–Ferreira	f_0	STFT(f_0)+RecSine Peak Estimation	DT(F)
Coler–Roebel	f_0 and $\mathcal{A}$	Cross-correlation of STFT(f_0_mod) and STFT($\mathcal{A}$_mod)	DT(corr)

Table 2.1: Comparison of existing vibrato detection methods. f_0: fundamental frequency. $\mathcal{A}$: amplitude of audio signal. DT(F): Decision Tree using sinusoid frequency. DT(corr): Decision Tree using cross-correlation. f_0_mod: modulation of fundamental frequency. $\mathcal{A}$_mod: modulation of amplitude.

Herrera-Bonada Figure 2.11 shows the basic flowchart of the vibrato detection method described in (Herrera & Bonada, 1998). The fundamental frequency time series, f_0, were extracted using Spectral Modelling Synthesis (SMS) analysis (Serra, 1989) with a sampling rate of 345Hz. As a vibrato shape is that of quasi-sinusoid (Sundberg, 1994), Herrera & Bonada (1998) applied a short-time Fourier transform (STFT). The peak-picking process from the resulted spectrogram was improved by using the parabolic interpolation. The vibrato existence was decided whether the peak-frequency is around 5Hz or 6Hz which is deemed to be the vibrato rate frequency.

Ventura-Sousa-Ferreira Figure 2.12 shows the basic flowchart of another vibrato detection method described in (Ventura et al., 2012). Similar to Herrera & Bonada (1998), Ventura et al. (2012) proposed a method which utilised STFT to detect vibrato based on f_0. However, the fundamental frequency time series were detected using the SearchTonal method (Ferreira, 1995; Ferreira et al., 2008; Sousa & Ferreira, 2010). A non-iterative frequency estimation method,

[12]Usually obtained from the root mean square of the audio intensity.

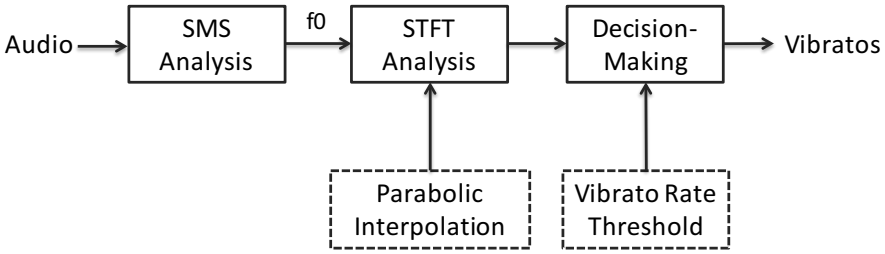


Figure 2.11: Flowchart of Herrera-Bonada method. STFT: short-time Fourier transform. f_0: fundamental frequency.

which is a combined rectangular-sine window frequency interpolation method, has been used to improve the peak-frequency. Decision-making process is to check whether the resulted peak-frequency lay on the vibrato rate frequency range. More comprehensive evaluation has not been reported by the author. For instance, the long audio test.

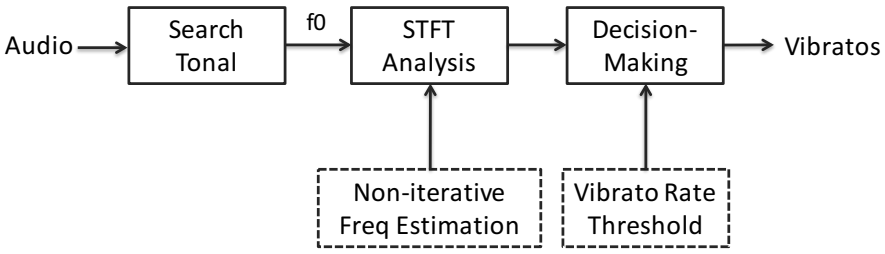


Figure 2.12: Flowchart of Ventura-Sousa-Ferreira method. STFT: short-time Fourier transform. f_0: fundamental frequency.

Coler-Roebel von Coler & Roebel (2011) presented a cross-correlation based frame-wise method for vibrato detection shown in Figure 2.13. Their method assumes that frequency modulations in physical instruments cause amplitude modulations. They first extracted the fundamental frequency using SuperVP method and audio amplitude employing root-mean-square method. Then the modulation of the fundamental frequency time series and amplitude time series were extracted. A first-order differentiation to f_0 modulation time

series was necessary. Then the STFT was applied to both differentiated f_0 modulation time series and amplitude modulation time series. The two spectral outputs were fed into a cross correlation in order to obtain the cross correlation coefficient. The binary vibrato result (on/off) for each frame was decided based on a threshold comparing to the cross correlation results. The resulting curve showed positive peaks in parts with vibrato. Note that, the threshold is depended on different instrument groups. Considering the assumption on the frequency modulation and amplitude modulation simultaneously, one of the shortcomings of this method is that it cannot apply to pure frequency modulation vibrato.

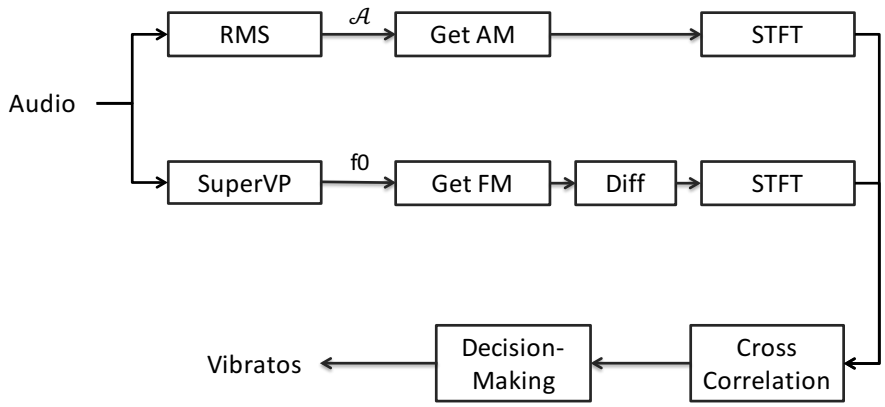


Figure 2.13: Flowchart of Coler-Roebel method. RMS: root-mean-square. Diff: first-order differentiation, STFT: short-time Fourier transform. f_0: fundamental frequency. $\mathcal{A}$: amplitude of audio signal. AM: amplitude modulation. FM: frequency modulation.

2.4 Pitch Detection Methods

As vibratos and portamenti mainly pertain to the pitch curve, we briefly review literature on single pitch detection for audio signals. Since we mainly use the pYIN (Mauch & Dixon, 2014) method in our thesis, we shall expand on its description. For a complete pitch detection methods review, please see (Benetos, 2012). In the review to follow, pitch detection is sometimes referred to as fundamental frequency detection.

2.4.1 What is Pitch?

In (Klapuri & Davy, 2007), pitch is described as "a perceptual attribute which allows the ordering of sounds on a frequency-related scale extending from low to high." In other words, pitch can be described by the fundamental frequency of a music note, a property that helps order the sounds on a frequency-related scale.

2.4.2 Time Domain Methods

In the time domain, autocorrelation is the most widely used method. Autocorrelation-based pitch detection methods are described in (Rabiner, 1977; Boersma, 1993). The autocorrelation function is given by:

$$ACF(\tau) = \frac{1}{N} \sum_{n=0}^{N-v-1} x(n)x(n+\tau), \tag{2.1}$$

where $x(n)$, the input signal, is usually the audio waveform signal, N is the length of the signal, and τ is the time lag for the autocorrelation. Note that the fundamental frequency time series is obtained using sliding overlapping windows. The fundamental frequency for each window is given by the inverse of the fundamental period of the waveform, which is the first major peak in the autocorrelation curve. Note that the major peak does not always appear at the fundamental period, instead there could be multiple fundamental periods. Some variants of the autocorrelation method have been proposed. Ross et al. (1974) created the *average magnitude difference function.* de Cheveigné (1998) proposed the *squared-difference function* using the Euclidean distance in the autocorrelation function. The squared-difference function is defined as:

$$SDF(\tau) = \frac{1}{N} \sum_{n=0}^{N-v-1} \left(x(n) - x(n+\tau)\right)^2 . \tag{2.2}$$

Later, a widely used single pitch detection method, YIN, was proposed by de Cheveigné & Kawahara (2002), which uses a cumulative normalised form of the squared-difference function:

$$CNSDF(\tau) = \begin{cases} 1 & ,\tau = 0 \\ SDF(\tau) / \left(\frac{1}{\tau} \sum_{j=0}^{\tau} SDF(j) \right) & ,otherwise \end{cases} \tag{2.3}$$

The main improvement provided by this method is that it avoids any spurious peaks near the zero lag, which in turn reduces the harmonic errors. YIN has

been show to be robust and reliable for single pitch fundamental frequency estimation (de Cheveigné, 2006; Klapuri, 2004; Yeh, 2008; Pertusa, 2010; Klapuri & Davy, 2007).

Building on YIN, the probabilistic YIN method, pYIN, was proposed by Mauch & Dixon (2014). In the original YIN method, the fundamental period is given by the smallest period, τ, for which $CNSDF(\tau)$ has a local minimum and $CNSDF(\tau) < s$ for a fixed threshold s. Instead of being limited to a fixed threshold, pYIN uses a probabilistic threshold, i.e. a Beta probabilistic threshold distribution S for 100 thresholds ranging from 0.01 to unity in steps of 0.01. Then the probability that a period τ is the fundamental period τ_0 is calculated as,

$$P(\tau = \tau_0 | S, x_n) = \sum_{i=1}^{N} a(s_i, \tau) P(s_i) [Y(x_n, s_i) = \tau], \tag{2.4}$$

where $P(s_i)$ is the probability of threshold s_i, $[\cdot]$ is the Iverson bracket evaluating to unity for a true expression and to zero otherwise, $Y(x_n, s_i)$ is the period estimated by YIN, and

$$a(s_i, \tau) = \begin{cases} 1 & : CNSDF(\tau) < s_i \\ p_a = 0.01 & : otherwise. \end{cases} \tag{2.5}$$

As a result, each time frame is associated with a number of fundamental frequency and corresponding probability pairs. The best likelihood path over time is then obtained using the Viterbi algorithm.

2.4.3 Spectral Domain Methods

In the spectral domain, Lahat et al. (1987) proposed a pitch detection method that also uses the autocorrelation function to obtain the fundamental frequency. Instead of applying the autocorrelation function to the audio waveform signal, they applied the autocorrelation function to the spectrum of the music signal. Another pitch detection method in the spectral domain presented by Childers et al. (1977) uses cepstral analysis. The cepstrum is the inverse Fourier transform of the logarithm of the signal magnitude spectrum. Schroeder (1968) proposed the product spectrum method to obtain the fundamental frequency. The fundamental frequency was measured by the higher harmonic components and the largest common divider of the harmonics, or by measuring the periods of individual harmonics and finding the smallest common multiple. Brown (1991) used the constant-Q spectrum of a music signal. The resulting log-spectrum produces a constant distance between harmonics for all pitches. The pitch is

detected by obtaining the cross-correlation between the log-spectrum and an ideal spectral pattern. A Hidden Markov Model-based single pitch detection approach was proposed by Doval & Rodet (1993), who modelled the spectrum as a set of sinusoids to find the best likelihood path amongst these sinusoids. Noll (1967) used the cepstral analysis to identify the pitch; peaks in the ceptrum provided the reciprocal of the fundamental frequency.

Spectrotemporal Methods

Spectral domain methods tend to introduce errors in pitch detection which are the integer multiples of the fundamental frequency (harmonic errors); time domain methods more easily result in pitch detection errors that are sub-multiples of the fundamental frequency (subharmonic errors) (Klapuri, 2003). Spectrotemporal methods combine both time domain and spectral domain methods to ameliorate these problems. In addition, some methods (Meddis & O'Mard, 1997; Slaney & Lyon, 1990) usually apply filterbanks according to human auditory models. Autocorrelation is then applied to each channel, and the final results are summed across all channels.

2.5 Conclusions

In this chapter, we have summarised the background and motivated the basis for this thesis. We reviewed musicological literature on music expressivity, with particular focus on vibrato and portamento. We briefly reviewed expressive music performance modelling, and computational vibrato and portamento modelling. After considering a number of single pitch detection methods, in the following sections, we choose to use pYIN for its computational efficiency and simple implementation (Mauch & Dixon, 2014). The fundamental frequency time series obtained from pYIN serves as input for vibrato and portamento detection and analysis. The state-of-the-art vibrato detection methods summarised here will be compared against our FDM-based vibrato detection method.

Chapter 3

Vibrato Modelling and Detection

In this chapter, we introduce the vibrato anatomy including: vibrato rate, extent, sinusoid similarity and envelope. Then we propose a novel frame-wise vibrato detection and analysis method based on the Filter Diagonalisation Method. For more details on the state-of-the-art vibrato detection methods, please refer to Section 2.3.

3.1 Vibrato Anatomy

In this section, we introduce the parameters that are used for the purpose of characterising vibratos. They are vibrato rate, extent, sinusoid similarity, and envelope. Much of the analysis will be performed on the extracted fundamental frequencies. For the case of a vibrato, where the fundamental frequency will oscillate between higher and lower pitches, we refer to the locally highest pitch as a peak and the locally lowest pitch as a trough.

3.1.1 Rate

The vibrato rate parameter is used to describe the tempo of a vibrato. Typically, vibrato rate may be estimated from the peaks and troughs of the vibrato fundamental frequency, as shown in Figure 3.1. Since vibrato follows a periodic shape (Sundberg, 1994), the interval between one peak and one trough is assumed to be a half cycle of the vibrato period. Then, the reciprocal of the interval gives the vibrato rate for the corresponding half cycle. The average vibrato rate for all half cycles results in the vibrato rate for the corresponding

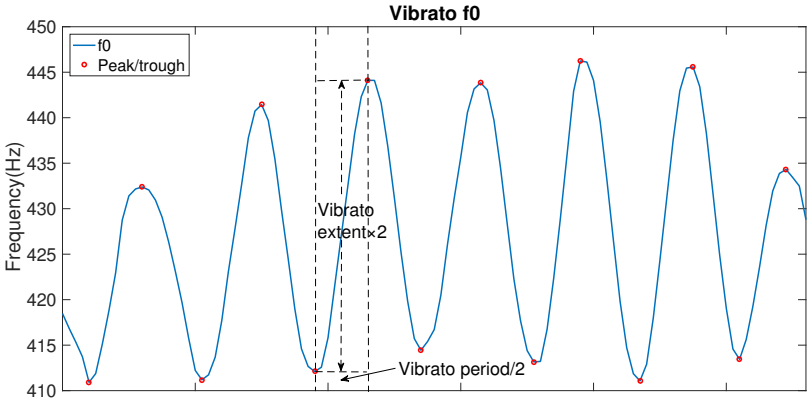


Figure 3.1: Demonstration of vibrato rate and extent.

note.

$$VR = \frac{1}{N-1}\sum_{n=2}^{N}\frac{1}{2(t_n - t_{n-1})}, \tag{3.1}$$

where t_n is the nth peak or trough time, and N is the total number of peaks and troughs for a vibrato.

3.1.2 Extent

The vibrato extent parameter is used to describe the variation in pitch of a vibrato, i.e. the difference between the highest and mean fundamental frequencies required to represent the vibrato. Similar to vibrato rate, the vibrato extent can be calculated from the peaks and troughs of the vibrato fundamental frequency. The vibrato extent for one half cycle is half of the difference between the peak and the trough of the corresponding half cycle, as shown in Figure 3.1. The average vibrato extent for all half cycles gives the vibrato extent for the corresponding note.

$$VE = \frac{1}{N-1}\sum_{n=2}^{N}\frac{|p_n - p_{n-1}|}{2}, \tag{3.2}$$

where p_n is the nth peak or trough pitch (it could be in Hz or semitone scale), and N is the total number of peaks and troughs for a vibrato.

3.1.3 Sinusoid Similarity

The underlying structure (shape) of a vibrato is another important aspect of vibrato research. Usually, the vibrato shape is that of a quasi-sinusoid (Sundberg, 1994). To find a parameter that capable of use for comparative analysis, we make use of a vibrato sinusoid similarity parameter, as published in (Yang

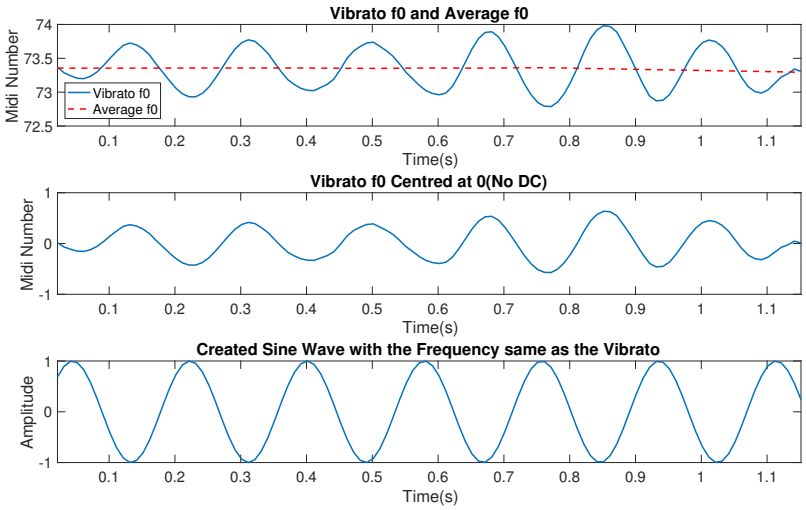


Figure 3.2: Vibrato and sine wave signals for calculating the vibrato sinusoid similarity. Top: the original vibrato fundamental frequency and its average vibato fundamental frequency. Middle: zero-centered fundamental frequency. Bottom: sine wave with the same frequency as the vibrato.

et al., 2013). The vibrato sinusoid similarity is the cross-correlation of a vibrato shape and the relevant sinusoid shape, which describes how similar of the vibrato shape is to that of a sinusoid. Different vibrato notes exhibit different vibrato rates and extents, and even different phases, and so it is impossible to create a unique and general sinusoid as a standard reference for this cross-correlation. Instead, we let every vibrato have its own reference sinusoid by creating a sinusoid having the same frequency as the vibrato. The following steps are used to obtain a vibrato sinusoid similarity:

1. Convert the fundamental frequency of the vibrato from linear scale to MIDI (musical instrument digital interface) scale.

2. Apply the local regression using weighted linear least squares, and a first degree polynomial model with 80 points span to smooth the fundamental frequency, in order to get the vibrato's average fundamental frequency.

3. Subtract the vibrato's average fundamental frequency from the MIDI scale fundamental frequency to block the DC component and centre the vibrato fundamental frequency at 0. The upper and middle parts of Figure 3.2 show a vibrato fundamental frequency waveform and its zero-centred fundamental frequency waveform, respectively. The data has been transformed to MIDI scale. The x-axis shows the performance time.

4. Compute the FFT of the 0-centred fundamental frequency.

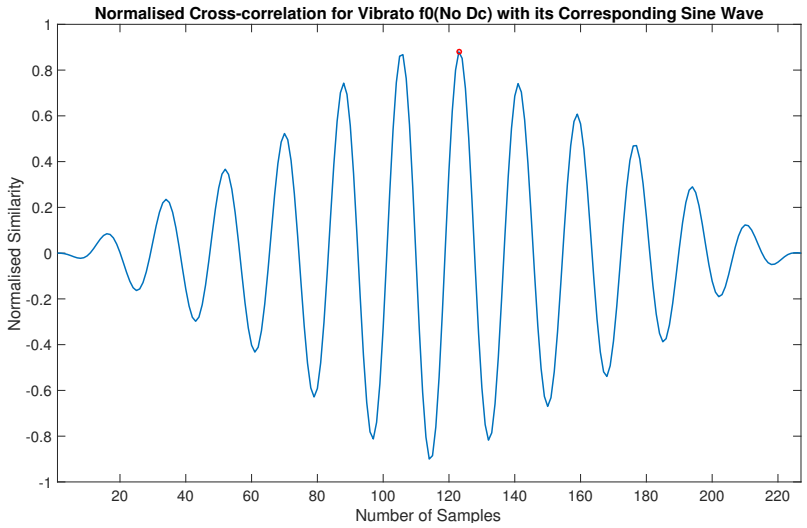


Figure 3.3: Vibrato sinusoid similarity. The normalised cross-correlation between the zero-centred fundamental frequency waveform and the sine wave.

5. Pick the peak from the spectrum to get the vibrato frequency.

6. Use this vibrato frequency to create a sine wave, and set the amplitude of the sine wave to 1. The amplitude does not affect the final result when the normalised cross-correlation is applied.

7. Calculate the normalised cross-correlation between the zero-centred fundamental frequency waveform and the sine wave. The correlation index (vibrato sinusoid similarity) lies between 0 and 1. The larger the value, the more similar the vibrato waveform is to the sine wave. The resulting sinusoid similarity in Figure 3.2 is presented by Figure 3.3.

8. Set the vibrato sinusoid similarity as the maximum of the normalised cross-correlation results.

3.1.4 Envelope

Average vibrato parameters for one note have been explored extensively. However, the vibrato parameters can change as a function of time, even within a single note. How the vibrato changes is an aspect of the vibrato's characteristics. Prame showed that the vibrato rate in opera singing increased towards the end of the note (Prame, 1994). Bretos and Sundberg confirmed this result for long sustained crescendo notes in opera singing (Bretos & Sundberg, 2003). In the experiment of Section 6.1, how the vibrato extent changes within one note was examined. As the analysis later in the Section 6.1 showed, a significant

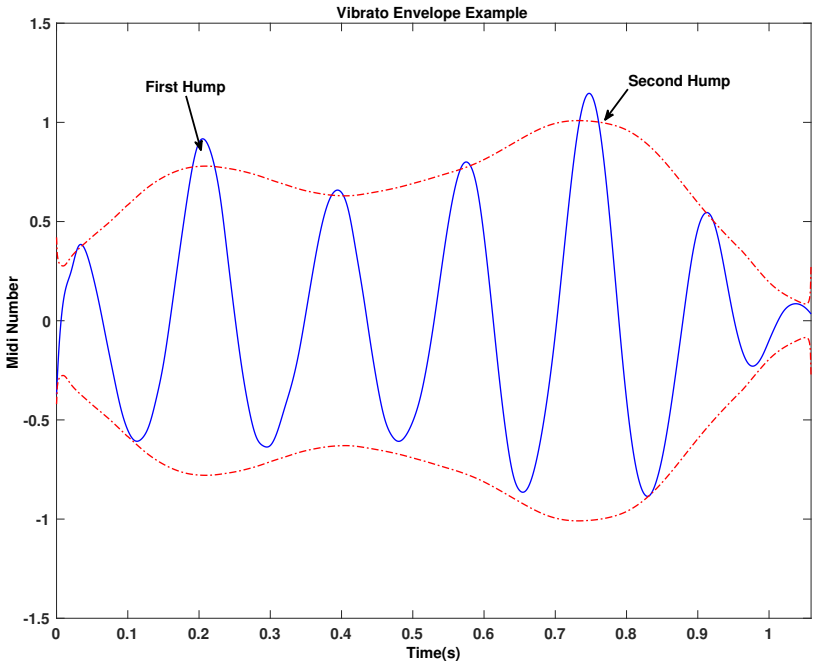


Figure 3.4: vibrato envelope of one vibrato. Solid line: fundamental frequency of the vibrato. Dashed line: vibrato envelope.

difference between erhu and violin vibrato is noticeable in the average vibrato extent. As a consequence, it is interesting to examine closely the make-up of a note, which could give insight into how the vibrato extent changes within the note.

The vibrato envelope was extracted by applying the Hilbert transform to the vibrato fundamental frequency contour. The result is an analytic signal of the fundamental frequency, which is a complex signal. Then the amplitude envelope of the original signal (vibrato fundamental frequency contour) can be obtained by taking the absolute value of this analytic signal.

$$x_a(t) = A(t)e^{j\varphi(\tau)} \tag{3.3}$$

$$A(t) = |x_a(t)| = \sqrt{x^2(t) + \hat{x}^2(t)} \tag{3.4}$$

Finally, the envelope was smoothed by applying a moving-average filter with a 0.2 second span to filter out the noise. Figure 3.4 shows the fundamental frequency of one vibrato and its envelope. There are two humps in the vibrato extent envelope. For this vibrato, the vibrato extent is relatively small at the start of the note, and it then reaches its first hump at around 0.2 s. The extent

decreases, then increases again. It reaches its second hump at around 0.75 s. A decreasing trend completes the vibrato. Thus, this vibrato has two envelope humps. In the Section 6.1.3, for each note, the number of humps in the envelope was recorded to reflect the vibrato extent variation.

3.2 The FDM-based Vibrato Detection and Analysis Method

In this section, we present a novel solution to the problem of vibrato detection and estimation. We show that the Filter Diagonalisation Method (FDM) presents a highly competitive alternative technique for frame-wise vibrato detection. Vibratos constitute an important expressive device in music performance whereby the musician modulates the fundamental frequency of a pitch in a periodic fashion at a rate typically between 4–8 Hz. Precise characterisation and measurements of vibrato features via computational means can reveal differences between performance styles and performers' skills, and has direct impact on ethnomusicological studies of the use of vibrato in world musics, the tracing of musical influences in musicological phylogenetic studies, and expressive performance pedagogy, analysis, and synthesis.

Automatic detection and estimation of vibrato, the focus of this section, would greatly speed vibrato analysis, systematic expressive performance research, music expression synthesis, and automatic music transcription. As an illustration, Figure 3.5 demonstrates the vibrato detection process. The top graph shows the spectrogram, while the middle one plots the f_0 (the estimated of fundamental frequency time series). The bottom graph shows the vibratos detected by the FDM with a Decision Tree, and the FDM with Bayes' Rule. The FDM method and the decision mechanisms will be described in following sections.

Prior efforts in automatic vibrato detection have focused primarily on applying the Fourier transform to the f_0 of the audio, to determine whether the spectral peak resides in the expected vibrato frequency range (Herrera & Bonada, 1998; Ventura et al., 2012). Due to the uncertainty principle for Fourier transform, choosing the best window size for computing the spectrogram presents a challenge when applying the Fourier transform to the f_0. The Fourier transform decomposes the f_0 into a number of sinusoids. In frame-wise vibrato detection, spectral peaks (sinusoids with largest amplitudes) will be blurred if the frame size is too large, containing both vibrato and non-vibrato segments; the precise location of the boundary would also be hard to identify. If the window is too small, the resolution in the frequency domain will be too low to show if the

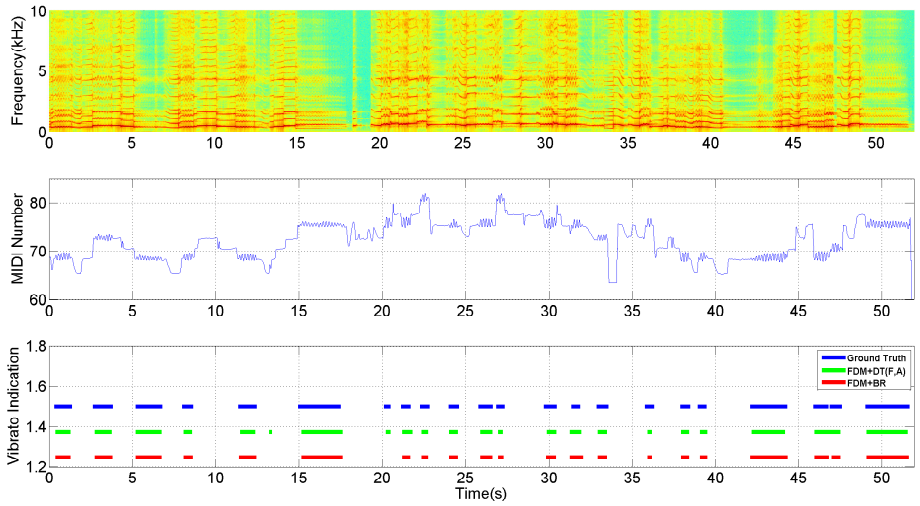


Figure 3.5: Vibrato detection demonstration of a real vibrato passage. Upper plot: spectrogram, middle plot: f_0, lower plot: vibrato detection results. FDM+DT(F,A) and FDM+BR are the proposed methods.

spectral peak resides in the vibrato rate range. This section offers a new solution to this problem via a computational technique that can be applied to short term signals with high frequency resolution.

The FDM is a harmonic inversion method. It is realised by constructing a filtered local frequency-domain signal matrix and then diagonalising it (details will be described in followed sections). The FDM is especially well suited to extraction of spectral features that occur over a very small time span. The greater precision afforded by the FDM allows for high resolution extraction of vibrato boundaries and characteristics, above and beyond current techniques. The method is also amenable to real-time implementation due to the frame-wise processing and small window size. Like other frame-wise methods , the FDM bypasses the segmentation process of note-wise vibrato detection methods(Rossignol et al., 1999; Pang & Yoon, 2005; Özaslan & Arcos, 2011; Weninger et al., 2012), and thus presents a first step towards a fully automatic vibrato detection system. We will show that the performance of the FDM-based system exceeds that of state-of-the-art vibrato detection methods and obtains high accuracy values for vibrato parameter estimation. To our knowledge, our study represents the first application of the FDM to the music domain.

The FDM offers two advantages over Fourier transform-based methods. It can obtain the frequencies and amplitudes of a given number of sinusoids in a selected frequency range for minute time frames. Its parametric fitting technique is capable of extracting the sinusoids directly from the original signal—in this

article, the original signal refers to the fundamental frequency, f_0, time series—without deriving them from spectral information, bypassing the error-prone peak-picking (Keiler & Marchand, 2002) intermediate step in the frequency domain. We will show that these two advantages significantly improve vibrato detection beyond the state-of-the-art in vibrato detection for short time frames.

3.2.1 The Filter Diagonalisation Method

The Filter Diagonalisation Method was developed as a tool to efficiently extract high resolution spectral information from short time signals, and has been used for a range of applications ranging from nuclear magnetic resonance to quantum dynamical systems (Neuhauser, 1990; Wall & Neuhauser, 1995; Mandelshtam & Taylor, 1997; Mandelshtam, 2001; Martini et al., 2013). As the FDM is new (to the best of our knowledge) to musical analysis, we will briefly describe its formulation and application to harmonic inversion.

While the FDM can, in general, be used to determine all fundamental frequencies and harmonics of a waveform audio signal $x(t)$ over an arbitrary frequency band, we are concerned in particular with the characterisation of vibrato. As a vibrato is an oscillating pitch, it can be characterised by properties of the oscillations over very short time periods. One may simply apply the Fourier transform to the time-varying fundamental frequency but, as will be discussed, the STFT is ill-suited to this task even if some peak-picking methods have been employed. We will, however, show that the FDM algorithm outputs a good representation of the vibrato signal.

First, we define the fundamental frequency time series, which describes the variation of the fundamental frequency with time, as $f_0(t) = f_0(n\tau)$, where $n = 0 : N$ and τ is the sampling period for the fundamental frequency. The time series of the fundamental frequency can be extracted from a musical waveform audio signal, $x(t) = x(n'\tau')$, where $n' = 0 : N'$ and τ' is the sampling period for waveform audio signal. A frame $x(nN_s\tau'+1 : (n+1)N_s\tau')$ is defined over a short time segment of N_s samples, and then one of a number of fundamental frequency extraction methods (Boersma, 1993; de Cheveigné & Kawahara, 2002; Mauch & Dixon, 2014; Childers et al., 1977) can be applied to determine the fundamental frequency of that particular frame. By iterating the frames across the entire time segment signal, a time-dependent fundamental frequency function results,

$$f_0(t) = f_0(n\tau) = f_0(nN_s\tau') = \mathcal{T}\{x(nN_s\tau' + 1 : (n+1)N_s\tau')\}, \tag{3.5}$$

where $\mathcal{T}$ stands for the fundamental frequency extraction transform, $N = N'/N_s$. If the final Δ samples of a frame are used as the first Δ samples of the next frame, then the total number of frames is given by $N = (N'-N_s)/(N_s-\Delta)$.

Further analysis will be applied to the signal $f_0(t)$ in order to characterise the spectra resulting from vibrato oscillations.

3.2.2 Outline of FDM

Determination of the vibrational spectrum of a dynamical system is typically performed using one of two classes of techniques: through calculation of the Fourier transform of a signal; or by diagonalisation or inversion of a matrix representing a short-time segment of a signal. It is well understood that the straightforward application of the Fourier transform, while effective at extracting large numbers of frequencies at arbitrary spectral ranges, is restricted by the uncertainty principle. For discretely sampled data this implies the need for a long time signal $T = 1/\Delta f$, which makes computation prohibitively expensive. Furthermore in many dynamical systems, including music, the harmonic profile may be changing rapidly enough that there is insufficient time to capture the data necessary for the inversion, resulting in low resolution results.

The second class of techniques is typically more useful in these applications as they rely on determination of relevant information (harmonic frequencies, decay rates, amplitudes, and phases) simultaneously through manipulation of a short-time segment of signal. Essentially, the purpose of these algorithms (including Prony's method, MUSIC, ESPRIT, etc.) is to fit the relevant parameters to represent the signal as a sum of exponentially decaying sinusoids,

$$f_0(t) = f_0(n\tau) = \sum_{k=1}^{K} d_k e^{-in\tau\omega_k}, \text{ for } n = 0, 1, \ldots, N, \tag{3.6}$$

where K is the number of sinusoids required to represent the signal to some tolerance. ω_k and d_k are fitting parameters which are defined as the complex frequency and complex weight of the k-th sinusoid, respectively. In general, the real part of the complex frequency represents the sinusoidal frequency while the imaginary part represents the decay rate (damping factor). The complex weighting parameter d_k represents the relative amplitude (real part) and phase (imaginary part) of each sinusoidal component. The aim is to solve for a total of $2K$ unknowns, representing all ω_k and d_k.

While these techniques use a variety of methods to achieve this approximation, a common feature necessary for computational efficiency is the need to convert a nonlinear fitting problem to a linear algebraic one. Typically this may lead to large or ill-conditioned problems when there are "too-many" frequencies (Mandelshtam & Taylor, 1997). The method of Wall and Neuhauser was introduced for high resolution spectral analysis of a time signal defined over a short time segment, and was shown to be exceptionally efficient (Wall &

Neuhauser, 1995) compared to linear prediction algorithms (MUSIC, ESPRIT, etc.), not least because all parameters ω_k and d_k are given through a single diagonalisation, rather than requiring multiple procedures (Hu et al., 1998).

While the technique was applied to continuous segments, it was later extended to discrete signals in Mandelshtam & Taylor (1997). The key novelty in this procedure was the association of the time signal, $f_0(t)$, with an autocorrelation function, such that

$$f_0(t) = f_0(n\tau) = \left(\Phi_0, e^{-in\tau\hat{\Omega}}\Phi_0\right), \tag{3.7}$$

where $(\cdot,\cdot)$ denotes the complex symmetric inner product without complex conjugation, i.e. $(a,b) = (b,a)$; and Φ_0 is a $K \times 1$ size vector representing the unknown and arbitrary initial state which does not need to be known explicitly. At this stage the exact form of the inner product is not important; rather, it is the complex symmetric properties that are most pertinent.

Suppose there are orthonormalised eigenvectors $\{Y_k\}$ that can diagonalise the complex symmetric evolution operator $\hat{U} \equiv e^{-i\tau\hat{\Omega}}$, then we have

$$\hat{U} = \sum_{k=1}^{K} u_k Y_k Y_k^T, \tag{3.8}$$

with $u_k \equiv e^{-i\tau\omega_k}$ are eigenvalues.

Inserting Eq. (3.8) into Eq. (3.7) [1], we are left with the same form as Eq. (3.6),

$$\begin{aligned} f_0(n\tau) &= (\Phi_0, \sum_{k=1}^{K} u_k^n Y_k Y_k^T \Phi_0) = \sum_{k=1}^{K} u_k^n (\Phi_0, Y_k)(Y_k, \Phi_0) \\ &= \sum_{k=1}^{K} d_k u_k^n = \sum_{k=1}^{K} d_k e^{-in\tau\omega_k}, \end{aligned} \tag{3.9}$$

with

$$d_k \equiv (\Phi_0, Y_k)(Y_k, \Phi_0) = (Y_k, \Phi_0)^2. \tag{3.10}$$

Thus, extracting spectral information, ω_k and d_k, from the signal, $f_0(t)$, is therefore equivalent to diagonalising the evolution operator, $\hat{U}$. The Filter Diagonalisation Method is then used to extract the eigenvalues, or harmonics, of $\hat{U}$, given as

$$u_k = e^{-i\tau\omega_k}. \tag{3.11}$$

The method is particularly well suited for spectra modelled as sums of dis-

[1] Note the eigenvalue and eigenvector properties, $\hat{U}^n = \sum_{k=1}^{K} u_k^n Y_k Y_k^T$.

crete frequencies. It is summarised below, with more detailed treatment found in (Neuhauser, 1990; Wall & Neuhauser, 1995; Mandelshtam & Taylor, 1997; Hu et al., 1998; Mandelshtam, 2001).

3.2.3 The Filter Diagonalisation Algorithm

In common with many other techniques, including the STFT, the purpose of the FDM is to decompose the original time series, $f_0(t)$, into a sum of sinusoids, as described in Eq. (3.6)[2]. As the musical signal can be quite complex, with a number of harmonics, we restrict our search to extracting the frequency and amplitude of the sinusoid with the largest amplitude. This gives sufficient accuracy at reduced computational cost, but we note that the technique may be generalised to output further harmonics.

The initial state Φ_0 and the evolution operator $\hat{U}$ are not explicitly known, and so we cannot directly solve Eq. (3.8). Instead, one can solve a generalised eigenvalue problem,

$$\mathbf{U}^{(1)}\mathbf{B}_k = u_k\mathbf{U}^{(0)}\mathbf{B}_k, \tag{3.12}$$

with

$$U^{(1)}_{jj'} = (\Psi_j, \hat{U}\Psi_{j'}), U^{(0)}_{jj'} = (\Psi_j, \Psi_{j'}), \tag{3.13}$$

where $\{\Psi_j\}, j = 1, 2, \ldots, K$, is a complete basis set. And amplitudes

$$d_k = \left(\sum_{j=1}^{K} B_{jk}(\Psi_j, \Phi_0)\right)^2, \tag{3.14}$$

with $Y_k = \sum_{j=1}^{K} B_{jk}\Psi_j$.

The primary problem here is the consturction of the evaluation operator matrix, $\mathbf{U}^{(p)}, p = 0, 1$. The use of the input signal $f_0(n\tau)$, in conjunction with an appropriately chosen basis can be used to determine its elements (Chen, 2002).

The selection of a primitive Krylov basis,

$$\Phi_n = \hat{U}^n\Phi_0, n = 0, 1, \ldots, M, \tag{3.15}$$

where $M = K - 1$, reduces the problem to that of a linear prediction algorithm such as ESPRIT. Applying the primitive Krylov Basis to Eq. (3.13), and using the symmetry property, the element of the evolution operator matrix can be

[2]Here, the direct component (DC) of $f_0(t)$ for a frame is removed by subtracting the mean.

evaluated [3],

$$U^{(p)}_{nn'} = (\Phi_n, \hat{U}^p \Phi_{n'}) = (\hat{U}^n \Phi_0, \hat{U}^{n'+p} \Phi_0) = f_0(n + n' + p). \tag{3.16}$$

Applying the primitive Krylov Basis to Eq. (3.14),

$$d_k = \left(\sum_{n=0}^{M} B_{nk} f_0(n) \right)^2. \tag{3.17}$$

However the use of a rectangular window Fourier basis significantly improves computational efficiency as the resultant matrix is almost automatically diagonalised, see e.g. (Hu et al., 1998). Essentially, the idea is to split up the frequency range of interest into a discretised frequency grid over which the measured signal $f_0(n)$ can be analysed. Selecting the Fourier basis set,

$$\Psi(\phi) = \sum_{n=0}^{M} e^{in\phi} \Phi_n \equiv \sum_{n=0}^{M} e^{in(\phi - \hat{\Omega}\tau)} \Phi_0, \tag{3.18}$$

the set of ϕ values defines the grid of frequency components which represents an estimation of the location of the spectral peaks of interest. An important result of this basis selection is that the matrix elements of the operator $\hat{U}^p = exp(-ip\tau\hat{\Omega}), p = 0, 1, 2, \ldots$, of any two functions $\Phi(\phi)$ and $\Phi(\phi')$ can be evaluated purely in terms of the elements of the signal $f_0(n)$. By applying the Fourier basis, the matrix operator can be calculated to be,

$$U^{(p)}(\phi', \phi) = (\Psi(\phi'), \hat{U}^p \Psi(\phi)) = \sum_{n'=0}^{M} \sum_{n=0}^{M} e^{in\phi} e^{in'\phi'} f_0(n + n' + p), \tag{3.19}$$

which may be recognised as a 2D discrete Fourier transform.

To evaluate the operator $\mathbf{U}^{(p)}$ it is first necessary to define the grid of frequency components. It is prudent to define a uniformly spaced grid, such that,

$$\phi_j = -2\pi(j\Delta f + f_{min})\tau; \Delta f = \frac{f_{max} - f_{min}}{K}, \tag{3.20}$$

where $j = 1, \ldots, K$ and $K + 1 = (f_{max} - f_{min})N_s\tau/2$ represents a suitable selection for the number of frequency points on the evaluation grid, as it will give the maximum spectral resolution which could give a unique fit to the specified signal length. On substitution of this equation into Eq. (3.19) we have,

$$U^{(p)}_{jj'} = \sum_{n'=0}^{M} \sum_{n=0}^{M} f_0(n + n' + p) e^{-2i\pi f_{min}\tau(n+n')} e^{-2i\pi\Delta f\tau jn} e^{-2i\pi\Delta f\tau j'n'}. \tag{3.21}$$

[3] For neat representation purpose we use $f_0(n)$ which is short for $f_0(n\tau)$.

By setting $M = K - 1$, one may solve the former equation for the evolutionary operator very efficiently by taking a 2D Fast Fourier Transform of the function $f_0(n + n' + p)e^{-2i\pi f_{min}\tau(n+n')}$. This is the primary achievement of the FDM algorithm.

Once the evoluation operator $\mathbf{U}^{(p)}$ has been determined, the generalised eigenvalue problem of Eq. (3.12) can be solved, from which the k eigenvalues (each giving fundamental/harmonic resonant frequencies and damping coefficients) are determined from u using Eq. (3.11), and the complex amplitudes (giving amplitude and phase information) are determined using,

$$d_k = \left(\sum_{j=1}^{K} B_{jk} \sum_{n=0}^{M} f_0(n) e^{in\phi_j} \right)^2 . \tag{3.22}$$

To summarise, the real benefit of the FDM algorithm lies in very efficient and accurate determination of harmonic information using short-time series. A computationally efficient 2D FFT is performed over a small frequency window $[f_{min}, f_{max}]$ in which there are up to K harmonics. A generalised eigenvalue equation then gives all relevant information. This technique, which is highly suited to vibrato detection, can reduce the linear algebraic computational effort and roundoff errors. As the vibrato rate is usually between 4–8Hz, the frequency window can be set around this range to reduce the computational cost of the generalised eigenvalue decomposition. We set the frequency window as 2–20Hz.

The basic steps for the FDM in the feature extraction module are summarised by the pseudocode in Algorithm 1. We consider only the frequency and amplitude, denoted by $F_H = f_{d_{max}}$ and $A_H = 2||d_{max}||$, respectively, for the sinusoid having the largest amplitude.

Here is an example that demonstrates the advantage afforded by the FDM over the FFT. Figure 3.6 shows the spectrogram obtained using the FDM and FFT for the f_0 which begins with a vibrato note, followed by a portamento (slide) to a lower non-vibrato note, then a higher non-vibrato note. Each set of three plots consist of: (top) the f_0; (middle) the spectrogram constructed from the FDM output; and, (bottom) the spectrogram output of the FFT using a hamming window.

Observe that there are clear peaks around 5–8Hz in the FDM spectrogram, representing the vibrato. Only when the window size increased to 0.5s does the FFT provide acceptable frequency resolution for identifying the presence of the vibrato. On the other hand, if the window size is large, it makes it difficult to pinpoint the vibrato boundaries.

In this simple example, raw outputs of the respective algorithms are presented for comparison; in practice, state-of-the-art FFT methods employ peak-

Algorithm 1: The FDM algorithm

Input: f_0
Output: $f_{d_{max}}$, d_{max}
$f_{\min} = 2; f_{\max} = 20;$
Filter
$\omega_{\min} = 2\pi f_{\min}, \omega_{\max} = 2\pi f_{\max};$
$K = (f_{max} - f_{min})N_s\tau/2 - 1;$
$\Delta f = \frac{f_{max} - f_{min}}{K};$
$\phi_j = -2\pi(j\Delta f + f_{min})\tau;$
Diagonalisation
$N_{iteration} = 4;$
for $n = 1 : N_{iteration}$ **do**
  for $p = 0 : 2$ **do**
    obtain $\mathbf{U}^p$ through 2D FFT of $f_0(n + n' + p)e^{-2i\pi f_{min}\tau(n+n')}$;
  end
  Solve $\mathbf{U^{(1)}B_k} = u_k\mathbf{U^{(0)}B_k}$;
  Get $u_k = e^{-i\tau\omega_k}$ and $\mathbf{B_k}$;
  if $\left|\left|(\mathbf{U}^{(2)} - u_k^2\mathbf{U^{(0)}})\mathbf{B_k}\right|\right| < \varepsilon$ **then**
    accept u_k;
  end
  $z_k = u_k;$
end
Calculate $d_k;$
Return d_{max} and corresponding $f_{d_{max}}$;

picking to refine the results. Section 3.3 provides further evaluations.

3.2.4 Deciding Vibrato Presence

Following the application of the FDM, FFT, or other method, a further decision making step is required to determine vibrato existence. In this section, we propose two alternative methods: the Decision Tree and Bayes' Rule. Both methods use frequency (F_H) and amplitude (A_H) information.

Decision Tree

A decision tree is constructed to support the vibrato detection process like in (Herrera & Bonada, 1998; Ventura et al., 2012). In contrast to the previous methods, which use only frequency information, we use both frequency and amplitude information provided by the FDM. The method requires the frequency range thresholds $F_{thd} = [f_{\min}, f_{\max}]$ Hz and the amplitude range thresholds $A_{thd} = [a_{\min}, a_{\max}]$ to be pre-determined. Figure 3.7 shows the decision tree for deciding vibrato existence.

The rationale for using both frequency and amplitude information is as fol-

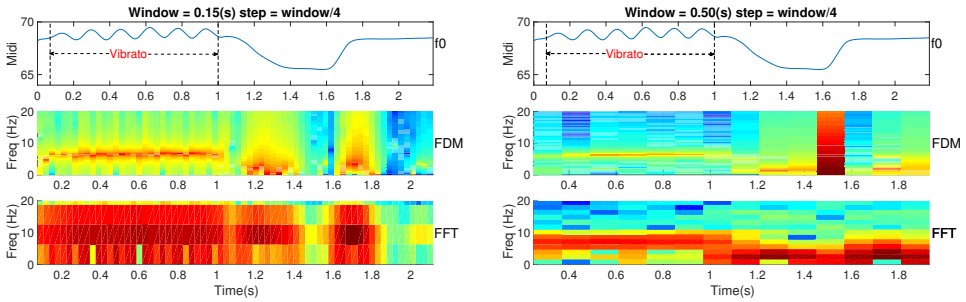


Figure 3.6: FDM and FFT spectrogram results for window sizes 0.15s and 0.50s.

lows: unintended and subconsciously applied movements by a performer can lead to small modulations and fluctuations in f_0. These can have similar frequencies as vibratos, and so only frequency analysis can result in erroneous classification. Figure 3.8 shows f_0 from an erhu audio passage. It consists of a non-vibrato segment followed by a vibrato segment. Symbols at zero indicate that no vibrato was detected. Any mark above the zero indicates positive detection. The red asterisks indicate detection of vibrato using only frequency information, while green triangles mark vibrato detection using both frequency and amplitude information. Both methods correctly detected the vibrato beginning at around $0.78s$. Note the small fluctuations in the non-vibrato part; with vibrato detection using only frequency, these small fluctuations lead to false positive identification due to their frequency in the vibrato range.

The frequency range thresholds can be obtained from the reported vibrato rate in the literature: $f_{\min} = 4\text{Hz}$, $f_{\max} = 12\text{Hz}$ for Western classical music (De-

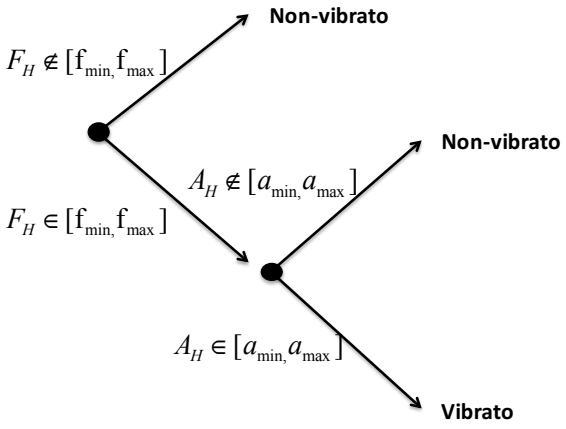


Figure 3.7: Decision Tree for deciding vibrato existence.

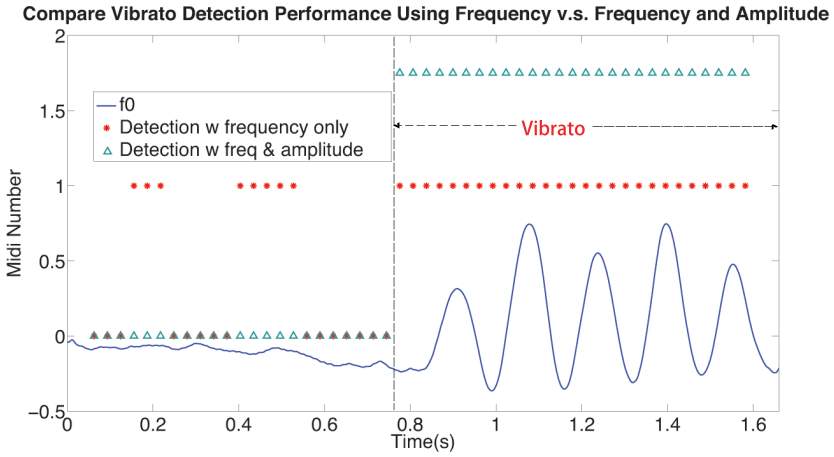


Figure 3.8: Vibrato detection using frequency v.s. frequency and amplitude.

sain & Honing, 1996); $f_{\min} = 4$Hz, $f_{\max} = 9$Hz for singing voice (Prame, 1994); and, $f_{\min} = 5$Hz, $f_{\max} = 8$Hz for erhu music (Yang et al., 2013). The amplitude range thresholds can be determined empirically: for instance, for voice and erhu we used $a_{\min} = 0.15$ semitone, $a_{\max} = +\infty$; and, violin $a_{\min} = 0.07$ semitone, $a_{\max} = +\infty$.

Bayes' Rule

The second technique applies Bayes' Rule, which assigns a probability of vibrato existence, rather than a binary answer, to each frame. Again, we consider the frequency and amplitude, F_H and A_H, respectively, of the sinusoid with the largest amplitude. Let V indicate vibrato existence, $\neg V$ implies no vibrato. Suppose that $\mathrm{P}(F_H) \neq 0$, the probability of vibrato existence given F_H,

$$\mathrm{P}(V|F_H) = \frac{\mathrm{P}(V \cap F_H)}{\mathrm{P}(F_H)}, \tag{3.23}$$

and $\mathrm{P}(A_H) \neq 0$, the probability of vibrato existence given A_H,

$$\mathrm{P}(V|A_H) = \frac{\mathrm{P}(V \cap A_H)}{\mathrm{P}(A_H)}. \tag{3.24}$$

According to Bayes' theory, we can re-write Eqs. (3.23) and (3.24) as

$$\mathrm{P}(V|F_H) = \frac{\mathrm{P}(F_H|V)\mathrm{P}(V)}{\mathrm{P}(F_H)}, \text{ and} \tag{3.25}$$

$$\mathrm{P}(V|A_H) = \frac{\mathrm{P}(A_H|V)\mathrm{P}(V)}{\mathrm{P}(A_H)}, \tag{3.26}$$

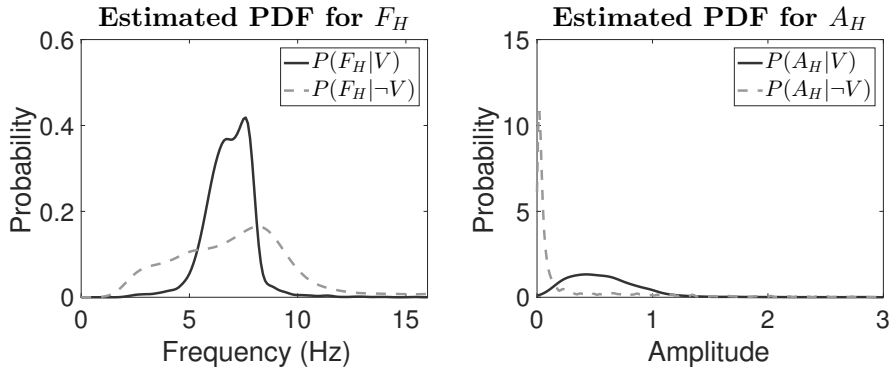


Figure 3.9: PDFs of $\mathbf{P(F_H|V)}$, $\mathbf{P(F_H|\neg V)}$, $\mathbf{P(A_H|V)}$ and $\mathbf{P(A_H|\neg V)}$ estimated using an erhu sample.

where $\mathrm{P}(F_H|V)$ and $\mathrm{P}(A_H|V)$ are the probabilities of observing F_H and A_H, respectively, given vibrato existence. $\mathrm{P}(F_H|V)$ and $\mathrm{P}(A_H|V)$ can be obtained from the estimated probability density function (PDF) for F_H and A_H, respectively. $\mathrm{P}(V)$ is the prior probability of vibrato.

Eqs. (3.25) and (3.26) lead to

$$\mathrm{P}(V|F_H) = \frac{\mathrm{P}(F_H|V)\mathrm{P}(V)}{\mathrm{P}(F_H|V)\mathrm{P}(V) + \mathrm{P}(F_H|\neg V)\mathrm{P}(\neg V)}, \text{ and} \tag{3.27}$$

$$\mathrm{P}(V|A_H) = \frac{\mathrm{P}(A_H|V)\mathrm{P}(V)}{\mathrm{P}(A_H|V)\mathrm{P}(V) + \mathrm{P}(A_H|\neg V)\mathrm{P}(\neg V)}, \text{ respectively.} \tag{3.28}$$

$\mathrm{P}(F_H|\neg V)$ and $\mathrm{P}(A_H|\neg V)$ can be obtained from the estimated PDF for F_H and A_H from non-vibrato frames. One such example, where the PDFs are estimated using Gaussian kernels, is given in Figure 3.9. The graph suggests that high values of $\mathrm{P}(F_H|V)$ lie between 5Hz and 9Hz, which is the typical vibrato frequency range. $\mathrm{P}(A_H|V)$ is larger than $\mathrm{P}(A_H|\neg V)$ for amplitudes between around 0.2 and 1. And, $\mathrm{P}(\neg V)$ is obtained using $\mathrm{P}(\neg V) = 1 - \mathrm{P}(V)$.

The five probabilities that need to be estimated from data are as follows: $\mathrm{P}(V)$, $\mathrm{P}(F_H|V)$, $\mathrm{P}(F_H|\neg V)$, $\mathrm{P}(A_H|V)$, and $\mathrm{P}(A_H|\neg V)$. For simplicity, we set the prior probability of vibrato[4] $\mathrm{P}(V) = 0.5$. Thus, the prior probability of non-vibrato is $\mathrm{P}(\neg V) = 0.5$. We then multiply Eqs. (3.27) and (3.28) to get the probability of vibrato existence:

$$\mathrm{P}(V) = \mathrm{P}(V|F_H) \times \mathrm{P}(V|A_H). \tag{3.29}$$

A threshold needs to be set or tuned to determine vibrato presence. We assume no prior information, i.e. that the probability of vibrato given any frequency is

[4]This quantity can be tailored to specific performers, instruments, genres, and cultures.

0.5 and that given any amplitude is 0.5. Thus, empirically, for the experiments presented here, the threshold set at 0.25 by assigning 0.5 each to $P(V|F_H)$ and $P(V|A_H)$, respectively.

3.3 Evaluation

This section provides details on the two evaluation datasets.

3.3.1 Datasets

Coler-Roebel and CMMSD Datasets

The first evaluation dataset consists of a combination of the existing Coler and Roebel dataset (von Coler & Roebel, 2011) and Coler and Lerch's Classical Monophonic Music Segmentation (CMMSD) dataset (von Coler & Lerch, 2014). Both datasets consist of monophonic samples. The Coler-Roebel dataset contains samples from 28 solo instrument passages of lengths ranging from 2s to 12s. The full details can be found in Appendix A. The samples are classified into four instrument groups: violin, voice, woodwind and brass. Vibrato annotations were completed by two persons, each using the Audacity[5] software. The CMMSD dataset consists of 36 solo instrument (string, woodwind, and brass) excerpts. The vibrato annotations were created by the first author using Tony (Mauch et al., 2015) [6].

Moon Reflected in Second Springs Dataset

In contrast to the short excerpts of the previous evaluation dataset, we created another dataset featuring long passages of music, which readily allows different parts of the same passage to be used as training and held out data, for example, for the FDM+BR method. This new dataset contains entire recordings of four performances of the traditional Chinese piece, *Moon Reflected in Second*

[5] http://audacityteam.org.

[6] In an effort to create a more robust dataset, we had another annotator generate a separate set of annotations. Because the second annotator was less experienced, the quality of the annotations was noticeably poorer (more inconsistent in the use of criteria for determining boundaries), and combining the two sets of annotations would have diluted the quality of the dataset. The perception of the vibrato onsets and offsets also varied from one person to another, and taking the average would not have produced a musically meaningful number. Thus, we chose a high-quality one-annotator dataset over a lower-quality two-annotator dataset, and decided to stick with only the original set of annotations. This highlights the difficulty in obtaining robust datasets. It remains to be validated that the single annotator is representative of a larger set of knowledgeable listeners; until that can be confirmed, we cannot reject the possibility that the model may be capturing an individual's perception. This is the case for the following dataset as well.

No	Ins.	Performer	Durations(s)	# Vibratos
1	Erhu	Jiangqin Huang[a]	445.83	170
2		Guotong Wang[b]	387.53	168
3	Violin	Jiang Yang[c]	254.54	124
4		Laurel S. Pardue[c]	325.50	120

Table 3.1: *Moon Reflected in Second Springs* Dataset for vibrato evaluation. *a*: (Huang, 2006), *b*: (Wang, 2009) and *c*: Recorded by the performer.

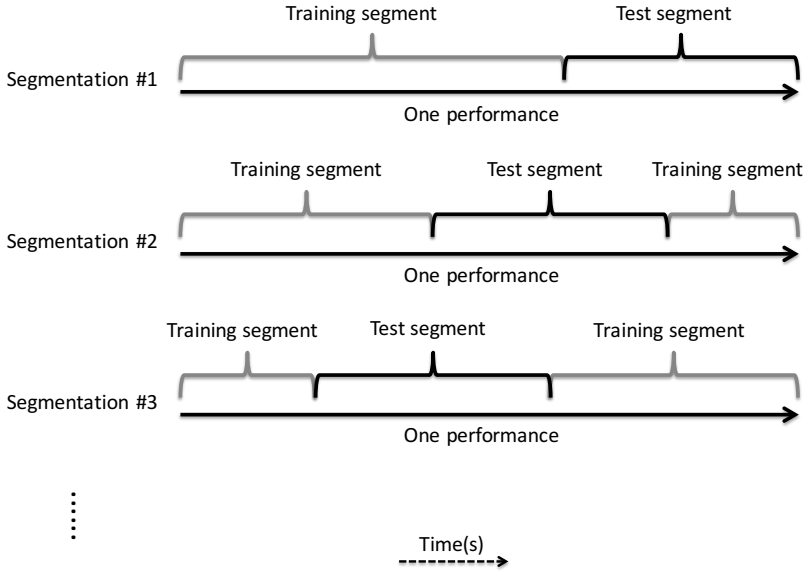


Figure 3.10: Demonstration of training/test data selection. The training segment spans 70% of the recording, with the remaining 30% held out as test data. The circular view of the recording is demonstrated in Segmentations #2 and #3.

Springs[7] (Hua, 1958). Two performances were recorded on the Chinese erhu and the other two on the Western violin. See Table 3.1 for more details.

Vibrato presence was annotated by the author using Tony. We divide each performance into contiguous training and test segments that are proportionally 70% (about 16.5 minutes of the recording or 31.7k consecutive 0.125-second frames) and 30% of the total length, respectively. We maintain a circular view of the recording so that the 70% training segment may begin in the tail end of the recording and loop back to the front. Figure 3.10 demonstrates the segmentation process. The process is iterated 10 times in order to obtain more stable results.

[7]There are a number of English translations for the title of this piece. The original Chinese name is 《二泉映月》 and the pinyin transliteration is Erquanyinyue.

3.3.2 Vibrato Detection Comparison

In this section, we compare the Herrera-Bonada (HB) (Herrera & Bonada, 1998), Ventura-Sousa-Ferreira (VSF) (Ventura et al., 2012), and Coler-Roebel (CR) (von Coler & Roebel, 2011) methods against our proposed FDM-based methods (with the two alternate decision mechanisms). The core components of the individual methods are outlined in Table 3.2. Please see Section 2.3.2 for more details. We re-implemented the HB and VSF methods, the CR method's Matlab code was provided by the authors.

Method	Input	Feature Extraction	Decision-Making
Herrera–Bonada	f_0	STFT(f_0)+Parabolic Interpolation	DT(F)
Ventura–Sousa-Ferreira	f_0	STFT(f_0)+RecSine Peak Estimation	DT(F)
Coler–Roebel	f_0 and $\mathcal{A}$	Cross-correlation of STFT(f_0_mod) and STFT($\mathcal{A}$_mod)	DT(corr)
FDM+DT (proposed)	f_0	FDM(f_0)	DT(F,A)
FDM+BR (proposed)	f_0	FDM(f_0)	BR

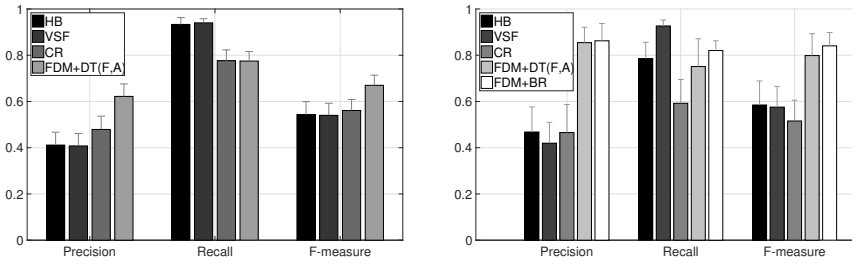
Table 3.2: Experiment setup for comparison of candidate frame-wise vibrato detection methods. f_0: fundamental frequency. $\mathcal{A}$: amplitude of audio signal. DT(F): Decision Tree using sinusoid frequency. DT(F,A): Decision Tree using sinusoid frequency and amplitude. DT(corr): Decision Tree using cross-correlation. f_0_mod: modulation of fundamental frequency. $\mathcal{A}$_mod: modulation of amplitude.

We use as input to the FDM-based techniques, and for the HB and VSF algorithms, the f_0 obtained using pYIN (Mauch & Dixon, 2014), a probabilistic version of the original YIN method (de Cheveigné & Kawahara, 2002). We left the CR f_0 and $\mathcal{A}$ (amplitude) extraction modules untouched as the method requires cross-correlation of the STFT of both the frequency and amplitude modulation time series. The two variants of the FDM-based method used the Decision Tree (denoted as FDM+DT(F,A)) and Bayes' Rule (FDM+BR) decision mechanisms, respectively. For fine time resolution, we set the window size to $w = 0.125s$ and step size $s = w/4$. When two or more methods required the same kinds of thresholds—for example, the vibrato frequency range threshold, F_{thd}, employed by HB, VSF, and FDM+DT(F,A)—the threshold was made uniform across the methods. The CR method had different thresholds for different instruments, as published in (von Coler & Roebel, 2011); we used the thresholds as published for the CR method.

Frame-level Results

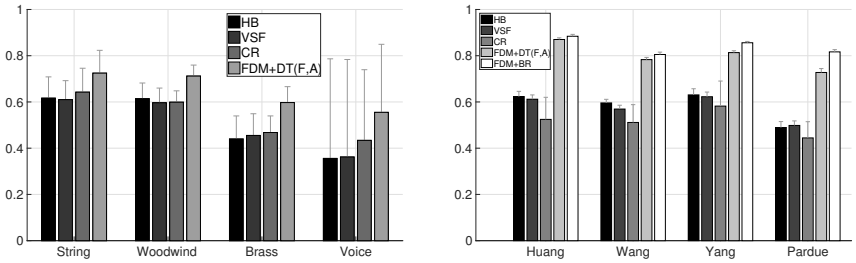
For frame-by-frame evaluation, a ground truth vector is created using the same sampling rate as the detection vector. The performance was then evaluated using the F-measure

$$F = \frac{2PR}{P+R}, \tag{3.30}$$



(a) Coler-Roebel and CMMSD datasets. (b) *Moon Reflected in Second Springs* dataset.

Figure 3.11: Frame-level evaluation. Results shown are averaged over all excerpts and iterations. Error bar shows the 95% confidence interval around the corresponding mean values.



(a) Instrument-wise F-measure for Coler-Roebel and CMMSD datasets. (b) Performer-wise F-measure for *Moon Reflected in Second Springs* dataset.

Figure 3.12: Frame-level F-measure evaluation for (a) each instrument group—results shown are averaged over the Coler-Roebel and CMMSD datasets; and, (b) each performer in the *Moon Reflected in Second Springs* dataset—results shown are averaged over all iterations. Error bar shows the 95% confidence interval around the corresponding mean value.

where precision, P, is defined as the number of true positive vibrato frames divided by the total positive vibrato frames, and recall, R, is defined as the number of true positive vibrato frames divided by the total number of vibrato frames. The F-measure was calculated for each excerpt.

Figure 3.11 shows the precision, recall and F-measure results for each vibrato detection method. The subplot (a) shows the evaluation on the Coler-Roebel and CMMSD datasets between our proposed FDM+DT(F,A) system and the HB, VSF, and CR methods. The FDM+BR system was omitted due to insufficient data for training the priors. Subplot (b) presents the evaluation performed using *Moon Reflected in Second Springs*.

The FDM-based methods perform significantly better with respect to the F-measure for both datasets, reflecting the better balance they strike between

precision and recall. The statistical Bayes' Rule gives better results than the Decision Tree. The HB and VSF have higher recall values and lower precision values. This implies that these two FFT-based methods correctly identified most vibrato frames, but at the cost of a substantial number of false positives; in fact, almost all frames were classified as vibratos, likely due to the low frequency resolution of the FFT at the short window size. The alternate mechanism CR method obtains a slightly higher precision for the Coler-Roebel and CMMSD datasets but a lower recall for both datasets, resulting in F-measure values similar to HB and VSF. This may be due to the CR method using the cross-correlation of the STFT of the frequency and amplitude modulations.

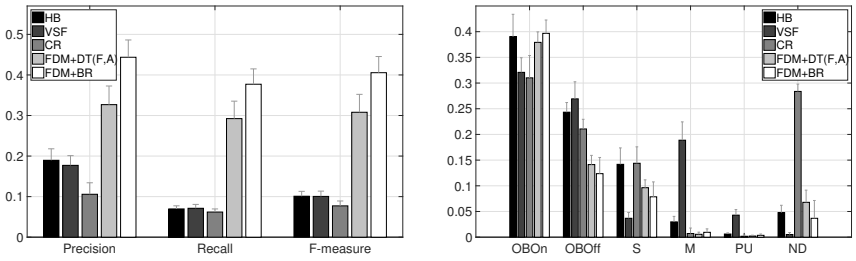
Figure 3.12 presents a further analysis. Subplot (a) shows the instrument-wise F-measure. There is an agreement amongst all methods that vibratos in string and woodwind instruments are easier to be detected than those in brass instruments and voice. The data shows that vibratos in brass instruments have small frequency but high amplitude modulations; and those in voice, at least for the datasets we studied, are less well controlled and more irregular. The error bars for the voice samples are wider because there are fewer voice samples in the Coler-Roebel and CMMSD datasets. Subplot (b) represents the performer-wise F-measure, showing little difference in vibrato detection between erhu and violin recordings. The vibrato detection can be improved by using performer-specific decision-making mechanisms.

Note-level Results

Vibrato is a continuous phenomenon operating at the level of the musical note—see examples in Figure 3.5. To evaluate the accuracy of vibrato boundaries (onset and offset), we employ the note boundary evaluation metric described in (Molina et al., 2014), originally applied to singing voice melody transcription. This is a more stringent evaluation than the frame-level evaluation described in the previous section. To the best of the authors' knowledge, this is the first note-level evaluation of vibrato detection.

We assume that a vibrato spans at least five consecutive frames, i.e. that it has duration > 0.25s. A detected vibrato onset is considered to be correct if it is within $\pm$100ms of the ground truth onset. Note that this threshold is higher than that for music transcription, which is typically 50ms. The detected vibrato offset is considered correct if it is within $\pm 100ms$ of the ground truth offset or no more than $\pm$20% of the ground truth vibrato duration from the ground truth offset.

We compute the F-measure for each excerpt, where the F-measure is defined as given in Eq. (3.30); instead of using vibrato frames, the precision, P, is defined



(a) Precision, Recall and F-measure. (b) Error rate.

Figure 3.13: Note-level evaluation for the *Moon Reflected in Second Springs* dataset. The results are the average values across all excerpts and iterations. Error bar shows the 95% confidence interval around the corresponding mean value.

as the number of true positive vibratos divided by the total positive vibratos, and recall, R, is defined as the number of true positive vibratos divided by the total number of vibratos. A true positive (correctly identified) vibrato is defined as a detected vibrato for which both onset and offset information are correct.

To analyse the errors, we use the six error types as defined in (Molina et al., 2014). Only-Bad-Onset (OBOn) error refers to the case where an onset error occurs but the offset is correct. Only-Bad-Offset (OBOff) error refers to the case where the onset is correct but the offset is not. A split (S) error is said to occur when the answer splits the ground truth vibrato into a number of consecutive detected vibratos. A merge (M) error refers to the case where a number of consecutive ground truth vibratos is merged as one detected vibrato. Spurious (PU) error refers to the case where a transcribed note does not overlap with any ground truth note. A non-detected (ND) error occurs when the ground truth vibrato does not overlap with any detected vibrato. The error rates for these error types are obtained by dividing each count by the number of ground truth vibratos; the PU error rate is divided by the number of detected vibratos.

Because the Coler-Roebel and CMMSD datasets contain short excerpts having one or two or several vibratos, one false positive or false negative will have significant impact on the results for each excerpt. Thus, we choose the *Moon Reflected in Second Springs* dataset, which has sufficient numbers of vibratos for each passage, for the note-level evaluations.

The note-level evaluation results are shown in Figure 3.13. As expected, the note-level precision, recall and F-measure results are lower than those for the frame-level. FDM-based methods produce higher F-measure than that of HB, VSF and CR. More specifically, the FDM+DT(F,A) has an F-measure value of 0.31 and the FDM+BR improves the F-measure to 0.41. Compared to the

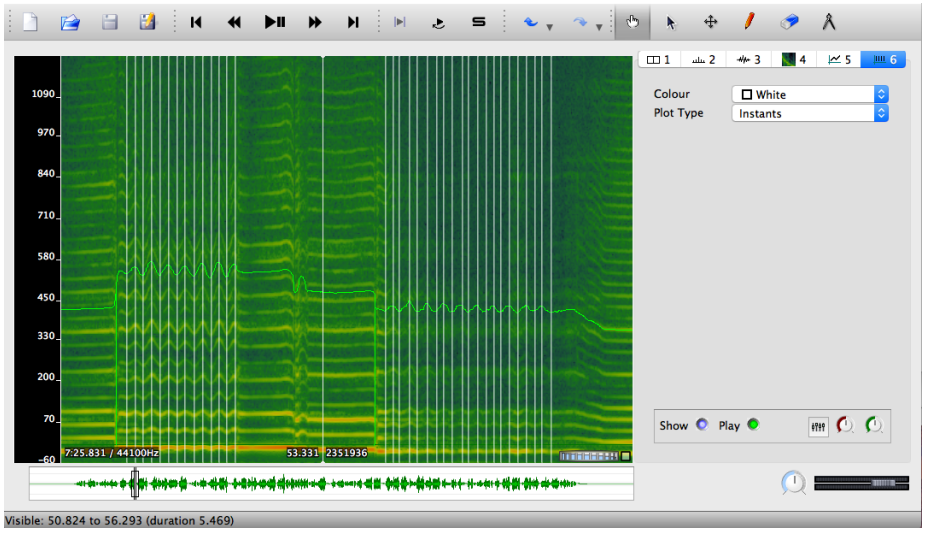


Figure 3.14: Annotation of vibrato peaks and troughs for vibrato parameter ground truth calculation using Sonic Visualiser.

corresponding frame-level results shown in subplot (b) of Figure 3.11, the Bayes' Rule improves much in note-level than frame-level.

3.3.3 Vibrato Estimation Evaluation

The output of the FDM also provides parameters for the vibratos detected. In this section, we evaluate the accuracy of these vibrato parameters, namely, the vibrato rate and extent. F_H, the output frequency having the largest amplitude, is used directly as the vibrato rate for that frame. The vibrato extent is A_H (as described in Section 3.2.3). Each vibrato's rate and extent are aggregated from its consecutive vibrato frames. We maintain the assumption of a vibrato spanning at least five consecutive frames.

We manually annotated the vibrato rates and extents for the *Moon Reflected in Second Springs* Dataset using Sonic Visualiser (Cannam et al., 2010). The peaks and troughs of each vibrato were marked based on the spectrogram and f_0 information as shown by Figure 3.14. The ground truth rate and extent for each vibrato were calculated from each half cycle. Assuming the interval between one peak and one trough is the duration of a half cycle, and the vibrato rate is the inverse of the cycle length, the vibrato extent is the half difference between the peak and trough measured in semitones. The vibrato rate and extent for each note is the mean value over all half cycles.

Vibrato parameter estimation accuracy is complicated by the fact that sometimes a ground truth vibrato is detected as two vibratos, and sometimes more

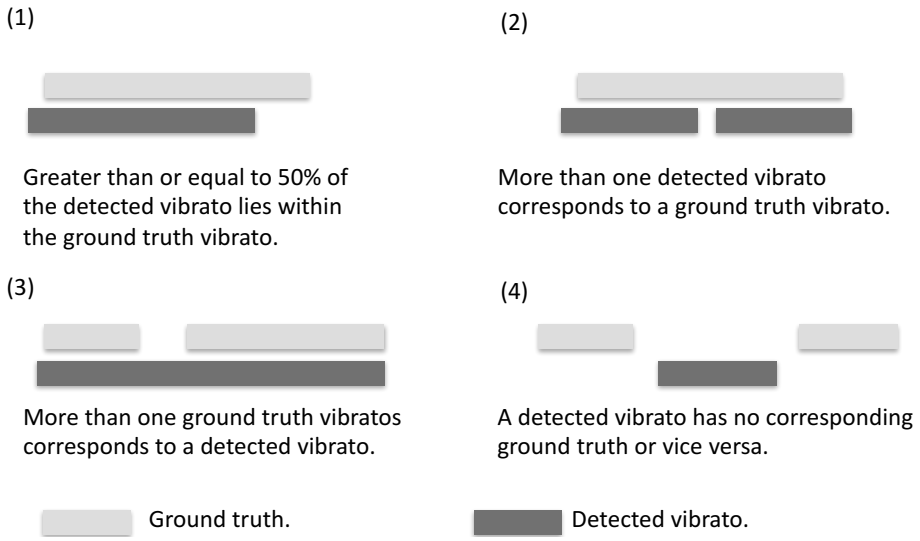


Figure 3.15: Illustration of determining corresponding ground truth and detected vibratos.

than one ground truth vibrato can be detected as one vibrato. To systematically determine corresponding ground truth and detected vibratos to assess parameter estimation accuracy, we apply the following rules, as illustrated in Figure 3.15:

1. for any ground truth vibrato, the corresponding detected vibrato is one for which at least half its interval lies within that of the ground truth vibrato;
2. if there is more than one corresponding detected vibrato, the *average* of the parameters of the detected vibratos will be used for assessing accuracy;
3. if more than one ground truth vibrato corresponds to a detected vibrato, the detected vibrato's parameters will be used for comparison with those of each ground truth vibrato; and,
4. if a detected vibrato has no corresponding ground truth vibrato or vice versa, no comparisons will be done.

The accuracy percentage is defined as

$$A_p = \begin{cases} 1 - \frac{|\hat{p}-p|}{p} & : \hat{p} \leq 2p \\ 0 & : \hat{p} > 2p \end{cases} \tag{3.31}$$

where, $\hat{p}$ is the estimated vibrato parameter (rate or extent) and p is the corresponding ground truth value.

No	Ins.	Performer	HB	VSF	FDM+DT(F,A)	FDM+BR
1	Erhu	Jiangqin Huang	92.68%	86.08%	93.84%	93.61%
2		Guotong Wang	87.99%	76.86%	90.64%	90.79%
3	Violin	Jiang Yang	90.73%	85.30%	93.27%	93.03%
4		Laurel S. Pardue	92.04%	85.10%	92.97%	92.94%
Average			90.86%	83.33%	92.68%	92.59%

Table 3.3: Vibrato rate accuracy for *Moon Reflected in Second Springs* Dataset.

No	Ins.	Performer	FDM+DT(F,A)	FDM+BR
1	Erhu	Jiangqin Huang	88.02%	89.53%
2		Guotong Wang	73.36%	79.46%
3	Violin	Jiang Yang	87.59%	90.90%
4		Laurel S. Pardue	87.23%	90.49%
Average			84.05%	87.59%

Table 3.4: Vibrato extent accuracy for *Moon Reflected in Second Springs* Dataset.

Table 3.3 and Table 3.4 show the accuracy for estimation of vibrato rate and extent, respectively. The accuracy values reported are the average over all iterations. For vibrato rate accuracy, we compare FDM+DT(F,A) and FDM+BR with the HB and VSF methods. The CR method is excluded here as it does not output vibrato rate nor extent. All methods achieve relatively high vibrato rate accuracies. FDM+DT(F,A) and FDM+BR obtained the highest vibrato rate accuracies, 92.68% and 92.59%, respectively. HB has a value of 90.86% and VSF a lower value at 83.33%. HB and VSF use decision trees to determine vibrato existence from vibrato rates; thus, even though their vibrato detection performance may be lower, for the vibratos that were correctly detected, the vibrato rates have been reasonably accurately assessed.

Regarding vibrato extent, we only report the results from our two methods, FDM+DT(F,A) and FDM+BR, because the other three methods do not have a direct vibrato extent output. Extending these methods to give vibrato extent is out of the scope of this article. FDM+BR has better vibrato extent accuracy than FDM+DT(F,A), 87.59% vs. 84.05%. For both proposed methods, the vibrato rate accuracy values are better than the vibrato extent accuracy values. This suggests that FDM-based methods are better at determining vibrato rates than vibrato extents. This may due to the fact that they consider only the sinusoid with the largest amplitude.

3.4 Conclusions

In this chapter we have described the anatomy of vibrato characteristics, including rate, extent, sinusoid similarity and envelope, all of which are necessary in

characterising a vibrato note. Then a novel frame-wise vibrato detection and estimation method that uses the Filter Diagonalisation Method is presented. The FDM is capable of extracting sinusoid frequency and amplitude information for a very short time signal, making it possible to determine vibrato frequency and pinpoint vibrato boundaries over a short time span. A natural byproduct of the FDM algorithm is the vibrato parameters themselves (rate and extent); thus, no additional computation is necessary to obtain the vibrato parameters.

We have also created a new monophonic dataset consisting of erhu and violin performances of an entire piece of music, *Moon Reflected in Second Springs*, for vibrato detection and vibrato parameter estimation. The long sequences allow for training and test data to both be excerpted not only from different performances by the same player, but from a single performance. The performances on Chinese versus Western instruments also allows for cross-cultural style comparisons.

The proposed FDM-based methods outperform existing state-of-the-art methods when evaluated on monophonic datasets comprising of string, wind, brass, and voice excerpts. The FDM method with Decision Tree had a significantly higher F-measure value than (Herrera & Bonada, 1998; Ventura et al., 2012; von Coler & Roebel, 2011) for vibrato detection when tested on the Coler-Roebel + CMMSD dataset; furthermore, the FDM-based methods had more balanced precision and recall values. For all methods tested, vibratos produced on string and woodwind instruments are more easily identified than those on brass instruments and by voice.

We also evaluated the FDM-based technique against other competing methods using frame-level and note-level vibrato detection metrics. The FDM-based methods performed best in both cases, with the Bayes' Rule decision mechanism achieving better results—F-measure 0.84 (frame-level) and 0.41 (note-level)—than Decision Tree—0.80 (frame-level) and 0.31 (note-level)—when tested on the *Moon Reflected in Second Springs* dataset. Bayes' Rule has the advantage (over Decision Trees) of greater flexibility and ability to adapt. Future work includes applying other machine learning methods to vibrato existence classification.

We further evaluated the vibrato parameter estimation capabilities of the FDM method using the *Moon Reflected in Second Springs* dataset. The accuracy of vibrato rate estimation is above 92.5%, and that of the vibrato extent estimation is on the order of 85% for both decision methods with FDM.

Finally, the FDM can be applied not only to vibrato detection and estimation but also to other music research domains requiring the extraction of sinusoids from a short time signal; for instance, it may be worth exploring ways to adapt the FDM to f_0 extraction. We have not completely exploited the full capabilities

of the FDM outputs; the imaginary component of the FDM's sinusoid frequency output could be used to improve vibrato onset detection.

Chapter 4

Portamento Modelling and Detection

There is little technical and scientific literature on portamento modelling and detection. Research on portamento modelling and detection has a further challenge of that a portamento, like pitch, is a perceived entity, and its effect must be heard to some degree in order for it to be recognised.

In this chapter, we propose the use of the Logistic Model in the modelling of portamento. We observe the prevalence of the S shape in the portamento pitch shifts, especially in string playing and vocal music. The S shape indicates that the execution of a portamento consists of an accelerating process followed by a decelerating process. In comparisons with other methods—the Polynomial Model, Gaussian Model, and Fourier Series Model—we show that the Logistic Model performs best. Parameters that convey musically meaningful information also make the Logistic Model stand out from other methods.

Considering the nature of the portamento time series, we propose a portamento detection method that uses the Hidden Markov Model and two observation probability distributions, a Gaussian distribution and Gaussian Mixture Distribution. In the evaluations, the GMM will be shown to have better performance, but returns on increasing Gaussian mixture numbers quickly diminish. We conduct experiments to detect portamenti using pitch alone, and pitch and energy, with the surprising result that combining energy with pitch does not improve portamento detection performance compared to pitch alone.

A portamento is a type of note transition. Note transitions can be classified into two types. The first is a discrete note transition, which is the default mode in piano playing. In discrete note transitions, the player is unable to or does not wish to alter the pitch in the process of moving from one note to another. The

other is the continuous note transition, which is prevalent in string, voice, and other instruments. In continuous note transitions, the player adjusts the pitch continuously. This type of note transition is usually referred to as portamento. Portamento is sometimes referred to as "glissando", "glide", or "slide". Here, we use "portamento" and "continuous note transition" interchangeably. A large range of expressivity exists in continuous note transitions.

In this chapter, we first propose a Logistic Model to mathematically and computationally describe the portamento. This section seeks to achieve the following aims:

1. to model portamenti quantitatively using a mathematical model;
2. to provide a tool for investigating and comparing note transition, and;
3. to provide a note transition model that can be used for synthesising natural-sounding music.

Then secondly, we explore the portamento detection using Hidden Markov Model method. Two observation probability distributions, i.e. Gaussian distribution and Gaussian Mixture Distribution, have been used. This section aims

1. to explore the feasibility of portamento detection;
2. to provide a portamento detection method.

4.1 Logistic Modelling for Portamento

To the authors' knowledge, there is yet no mathematical model designed or tested for note transitions. The aim of this study is to shed light on the mathematical and computational modelling of note transitions. We observe that the S shape is prevalent in many note transitions, especially in string playing and vocal music. Practically speaking, the execution of a portamento consists of an accelerating process followed by a decelerating process. In a portamento, the player's finger will start to accelerate to a target speed, then decelerate to arrive at the target note position. This usually results an S shape in the spectrogram and pitch contour, as shown in Figure 4.1.

Inspired by the model for population growth (Verhulst, 1838; Pearl, 1927), we propose to use the *Logistic Model* to fit the S shape of the portamento. We will show that the Logistic Model fits the shape of note transitions very well, and that it has the distinct advantage that its coefficients have direct musical meanings and interpretations.

The primary goal of this section is to introduce the Logistic Model for note transitions. At the same time, we offer some alternative modelling methods for

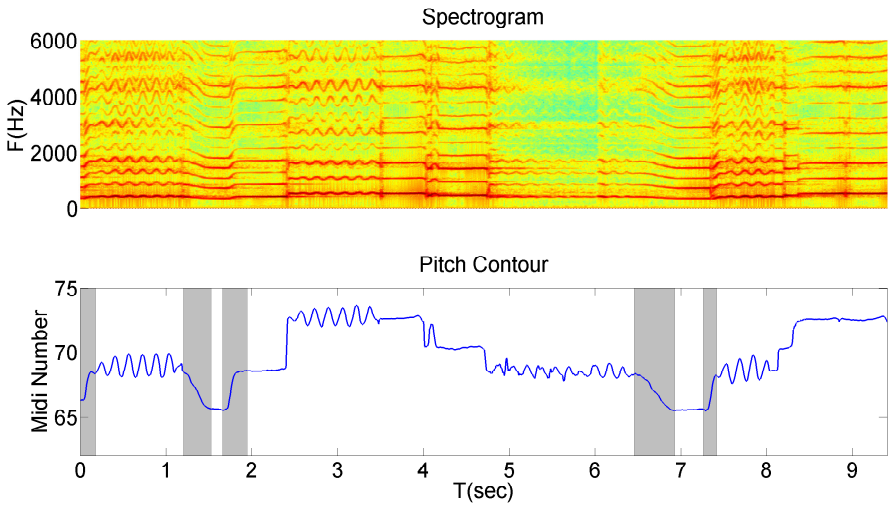


Figure 4.1: Spectrogram and the corresponding pitch contour from a passage of erhu. The portamenti are highlighted by grey area in the lower plot.

comparison, namely, the Polynomial Model, Gaussian Model, and Fourier Series Model. We show that, in general, the Logistic Model has better explanatory value than the other methods. Moreover, other methods are not able to provide direct outputs with meaningful musical interpretations. All models mentioned are used to fit the pitch curve of note transitions.

4.1.1 Logistic Model

The Logistic Model was originally proposed to solve problems in population dynamics (Verhulst, 1838; Pearl, 1927). It has been applied successfully to the physical growth of organisms and to forestry growth (Payandeh, 1983). Moreover, the Logistic Model has been extended to other fields: Marchetti & Nakicenovic (1979) applied the Logistic Model to energy usage and source substitution; Herman & Montroll (1972) presented the industrial revolution as modelled by the Logistic Model.

In string playing and singing, the players' portamento pitch curve (the log of the fundamental frequency) tends to exhibit an exponential start and an exponential end. In other words, the start and the end of portamenti have similar exponential-style increasing and decreasing shapes. The Logistic Model is especially well-suited to model such features.

To the best of the authors' knowledge, the Logistic Model has yet to be applied to note transitions or other relevant music areas. Inspired by the Richards' function (Richards, 1959; Tsoularis & Wallace, 2002), the Logistic function used

here is defined as

$$p(t) = L + \frac{(U - L)}{(1 + Ae^{-G(t-M)})^{1/B}}, \tag{4.1}$$

where L and U are the lower and upper horizontal asymptotes, respectively. Musically speaking, L and U are the antecedent and subsequent pitches of the transition. A, B, G, and M are constants. G can further be interpreted as the growth rate, indicating the degree of slope of the transition.

An important characteristic to model is the inflection point of the transition, where the slope is maximised. The time of the inflection point is given by

$$t_R = -\frac{1}{G} \ln \left(\frac{B}{A} \right) + M. \tag{4.2}$$

This value is obtained by setting the second derivative of Eq. (4.1) to zero. The details of the derivation can be found in Appendix B. Since Eq. (4.1) is monotonically increasing, the second order derivative has only one zero point. In other words, the zero point of the second derivative is the maximum of the first derivative, where the slope changes. The inflection point in pitch is calculated by substituting t_R into Eq. (4.1).

4.1.2 Alternative Models

Polynomial Model

The Polynomial Model is given by

$$p(t) = a_n t^n + a_{n-1} t^{n-1} + \ldots + a_2 t^2 + a_1 t + a_0, \tag{4.3}$$

where n is the degree of the polynomial. The model then requires $n + 1$ coefficients. Although the Polynomial Model is widely used in many applications for curve fitting, this model performs poorly, especially outside the intermediate range of the transition. It cannot model data having asymptotic lines. There is also a trade off between performance and polynomial degree. The larger the number of coefficients, the better the performance; however, the complexity and computational cost would also increase.

Gaussian Model

The Gaussian Model is given by

$$p(t) = \sum_{n=1}^{N} a_n \exp \left[- \left(\frac{t - b_n}{c_n} \right)^2 \right], \tag{4.4}$$

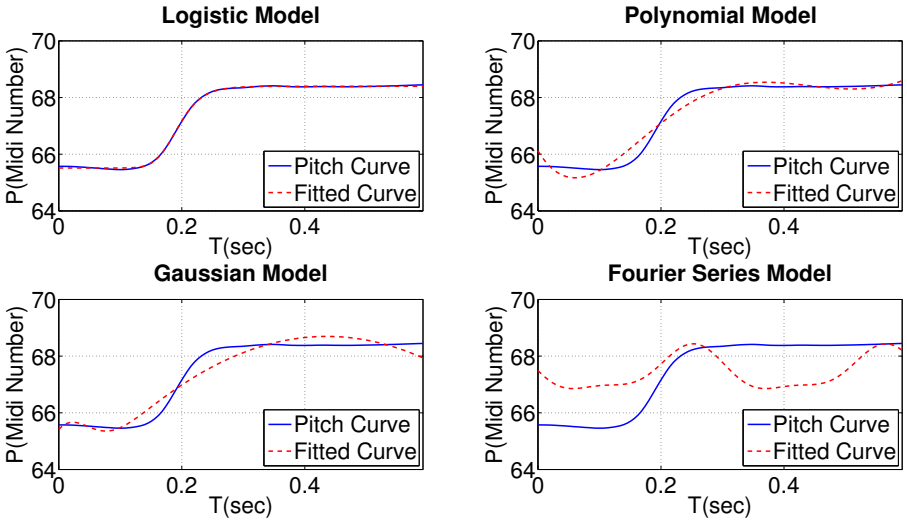


Figure 4.2: Modelling of a note transition using the Logistic Model, Polynomial Model, Gaussian Model, and Fourier Series Model. The coefficients are all constrained to six.

where a_n is the height of the model, b_n the location of the peak, and c_n controls the width of the Gaussian shape. The constant N denotes the number of Gaussian peaks, giving $3 \times N$ coefficients.

Fourier Series Model

The Fourier Series Model is given by

$$p(t) = a_0 + \sum_{n=1}^{N} a_n \cos(n\omega t) + b_n \sin(n\omega t). \tag{4.5}$$

Here, a_0 is the constant term, and ω is the fundamental frequency. The parameters a_n and b_n are amplitudes of the cosine and sine terms, respectively. The constant N is the number of the sinusoids used to fit the data, which results in $2 + 2 \times N$ coefficients.

Figure 4.2 shows the above modelling methods fitting a continuous note transition. For the purpose of unifying the comparison, each method is modelled by six coefficients. Note that the Logistic Model fits the transition better than the other three methods, and while the Fourier series in particular would show an improved match with more coefficients, it shows a poor match with the six used. A further statistical evaluation will follow.

4.1.3 Evaluation

The purpose of this section is to provide a computational evaluation on the feasibility of the Logistic Model-based fit of the portamento, in comparison to the three alternative modelling methods. The Logistic Model in Eq. (4.1) has six coefficients. For comparison, the other three modelling methods were also constrained to the same number of coefficients. As a result, we choose a 5-degree Polynomial Model (i.e., $n = 5$ in Eq. (4.3)), a 2-degree Gaussian Model was selected ($N = 2$ in Eq. (4.4)), and a 2-degree Fourier Series Model ($N = 2$ in Eq. (4.5)).

Dataset

First, pitch contours were extracted from audio files. As there is no prior note transition detection method, there also is no existing database for evaluation. Portamenti can take place over an extremely short period of time, and it can be challenging to annotate the transition duration accurately. This is one of the reasons we choose to annotate a transition from the midpoint (of the note's duration) of the antecedent note to the midpoint of the consequent note. We create a note transition database using the following rules[1].

1. The portamento starts from the midpoint of the antecedent note and ends at the midpoint of the subsequent note[2].

2. If the portamento starts from an intermediate note[3], then the start point is the beginning of the intermediate note.

3. If the subsequent note is not the target of the portamento[4], then the end point is the end of the subsequent note.

4. If either of the two notes contains a vibrato, the vibrato is flattened to the average fundamental frequency of the note.

Following the annotation rules above, we manually annotated portamenti for erhu and violin performances of a phrase in a well known Chinese piece *Moon Reflected in Second Springs* using Sonic Visualiser (Cannam et al., 2010). The violin score of this phrase is shown in Figure 4.3. This phrase forms the backbone of the entire piece, and is the phrase that is least changed when adapting the score from erhu to violin.

[1] To define these rules, we mainly considered the Type-1, Type-2(B) and Type-2(L) portamento types. As the Type-2(BL) and Type-3 portamento are very rare in performance, we ignore them for simplicity.

[2] This is the Type-1 portamento in Section 2.1.2.

[3] This is the Type-2(L) portamento in Section 2.1.2.

[4] This is the Type-2(B) portamento in Section 2.1.2.

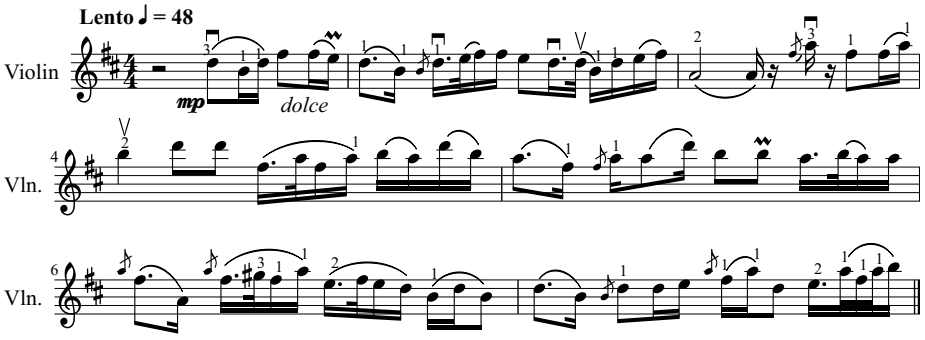


Figure 4.3: A phrase of *Moon Reflected in Second Springs* (Hua, 1958). Numbers above notes indicate fingering.

Instrument	Player	Duration(s)	No. of Transitions
Erhu	Jiangqin Huang	55.65	31
	Guotong Wang	42.04	36
Violin	Jiang Yang	37.78	20
	Laurel Pardue	35.73	24
Total	N/A	171.21	111

Table 4.1: Note transition dataset (corresponding to phrase shown in Fig. 4.3).

The two erhu performances used are from recordings by Huang (2006) and Wang (2009), and the two violin performances used are from solo recordings provided by Jian Yang and Laurel Pardue. Details about the excerpted portamenti can be seen in Table 4.1. The numbers in Table 4.1 show that erhu players tend to use more portamenti than violin players. This may be due to the fact that the erhu has only two strings while the violin has four. Thus, erhu players have to initiate more slides to reach the target pitches while violin players are able to change strings to reach the target pitch without sliding. Except for the cases where portamenti are indicated in the score, the physical form of the instrument may be an important factor influencing the number of portamenti the player employs.

This dataset is used in both this and the next sections. For the purposes of the study in this chapter, we focus only on the continuous note transitions. It is worth pointing out that discrete note transitions can also be modelled by a Logistic Model by giving the slope an extremely high value.

Model Fitting

We use the Curve Fitting Toolbox in Matlab (MATLAB, 2013) to perform the note transition modelling. In this package, the non-linear least squares method was used.

Setting the correct search ranges and initial solutions can have a high impact on curve fitting performance. Unlike the Logistic Model, the Polynomial, Gaussian, and Fourier Series models do not have coefficients having direct musical meanings relating to note transitions. As a result, the search ranges of these methods were set to $(-\infty, +\infty)$, and the initial points were decided (randomly) by Matlab. For the Logistic Model, we found that its performance improves when we set the initial value of L to be the lowest pitch in the note transition, and the initial value of U to be the highest pitch in the note transition. The search ranges and initial points for the Logistic Model coefficients are given in Table 4.2, where p_{min} and p_{max} are the lowest and highest pitches in the note transition, respectively.

Coefficient	$\boldsymbol{A}$	$\boldsymbol{B}$	$\boldsymbol{G}$	$\boldsymbol{L}$	$\boldsymbol{M}$	$\boldsymbol{U}$
Search Range	$(0, +\infty)$	$(0, +\infty)$	$(-\infty, +\infty)$	$[1, 128]$	$[0, +\infty)$	$[1, 128]$
Initial Point	0.8763	1	0	p_{min}	0.1	p_{max}

Table 4.2: Search ranges and initial points for coefficients of Logistic Model.

For each portamento (a finite pitch time series), the Root Mean Squared Error (RMSE) and Adjusted R-Squared values were calculated for each model. The RMSE represents the sample standard deviation of the differences between predicted values and observed values. The smaller the value the better the modelling performance. The adjusted R-squared is a statistical measure of how well the model fits the real data points. The higher the value the better the moelling performance. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance.

The performance of the four modelling methods is presented in Figure 4.4, which shows the average Root Mean Squared Error (RMSE) and Adjusted R-Squared values. Note that the Logistic Model has the lowest RMSE and the highest Adjusted R-Squared value, showing that the Logistic Model performed better in the note transition modelling than any other methods. The Polynomial Model has the second best performance. The Gaussian Model obtains the third place. While the Fourier Series Model gives the poorest modelling performance. The superiority of the Logistic Model is confirmed by an ANOVA (Analysis of Variance) (Ott & Longnecker, 2010) analysis. We performed the ANOVA analysis between the Logistic Model and other modelling methods to confirm that the mean values given in Figure 4.4 are significant. From Table 4.3, all p-values are lower than the significant level of 0.01.

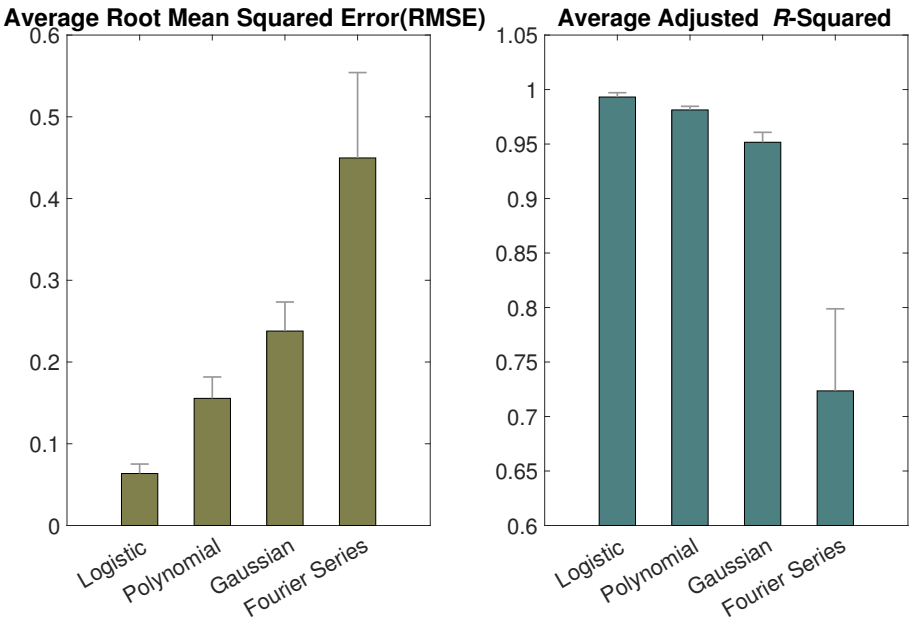


Figure 4.4: Modelling performance of Logistic Model, Polynomial Model, Gaussian Model and Fourier Series Model. Error bar shows the 95% confidence interval around the corresponding mean value.

4.1.4 A Case Study of Erhu and Violin Music

We present here the results of a case study investigating the behaviour of portamenti as performed by erhu versus violin players based on the dataset in Section 4.1.3. The Logistic Model is employed here to show the feasibility of such expressive performance analyses.

Parameters of Interest

Using the Logistic Model as defined in Eq. (4.1), we examine the following characteristics of the note transitions:

1. The slope of the transition, which is the coefficient G in Eq. (4.1).
2. The transition duration. Once the Logistic Model is set up, the first derivative of the Logistic curve that is larger than a threshold value can be employed to identify the transition duration. Empirically, this threshold is 0.861 semitones per second.
3. The transition interval. The interval is obtained by calculating the absolute semitone difference between the lower and upper asymptotes.

Root Mean Squared Error

p-value	Polynomial	Gaussian	Fourier
Logistic	1.11×10^{-9}	2.13×10^{-17}	5.81×10^{-12}

Adjusted R-Squared

p-value	Polynomial	Gaussian	Fourier
Logistic	2.03×10^{-6}	3.94×10^{-15}	1.64×10^{-11}

Table 4.3: ANOVA Analysis (p-value) of Root Mean Squared Error and Adjusted R-Squared between Logistic Model and other three model methods where each pair comparison has $df = 221$.

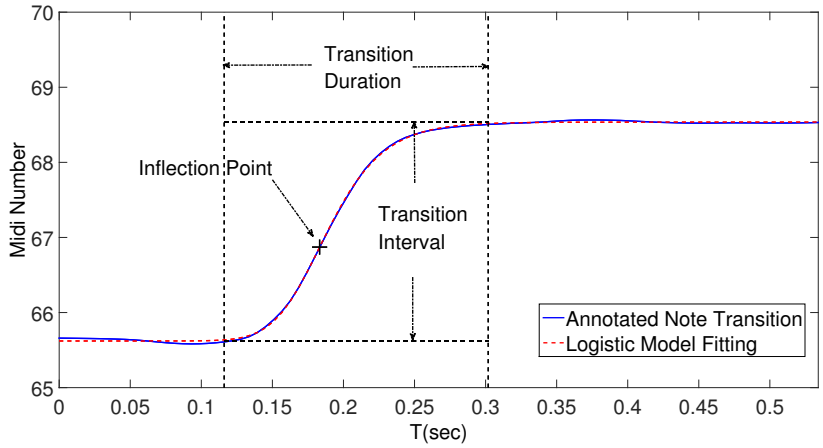


Figure 4.5: Illustration of transition duration, transition interval, and inflection time and pitch from an erhu excerpt.

4. The normalised inflection time. The actual time of the inflection point is given by Eq. (4.2). As transition durations are different one from another, this time is normalised to lie between 0 and 1, where 0 marks the beginning and 1 the end of the transition duration.

5. The normalised inflection pitch. This is similar to the normalised inflection time; this parameter is also normalised to lie between 0 and 1, where 0 indicates the lower asymptote and 1 the higher asymptote in the transition interval.

An example of a note transition is given in Figure 4.5, where a slope of 42.25 can be observed. The transition duration and interval are 0.19 seconds and 2.91 semitones, respectively. The inflection point appears to lie in the first half of the transition duration and interval; this is confirmed by the normalised inflection time of 0.32 and pitch of 0.43.

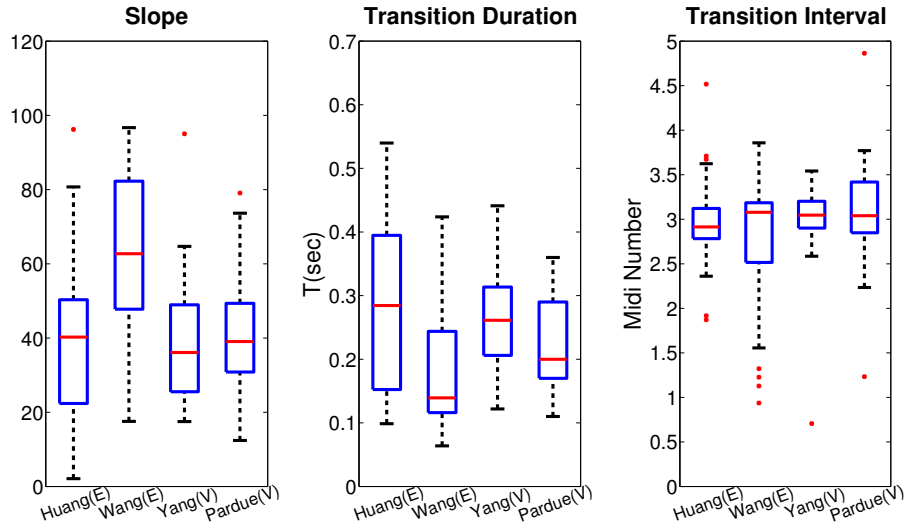


Figure 4.6: Boxplots of slope, transition duration, and transition interval for all four players of *Moon Reflected in Second Springs*. E: Erhu, V: Violin.

Results

The slope, transition duration, and interval statistics are shown in Figure 4.6. The middle bar in the box indicates the median value. The lower and upper edges mark the 25th and 75th percentiles, $Q1$ and $Q3$, respectively. The dotted lines extend from $(Q1 - 1.5 \times (Q3 - Q1))$ to $(Q3 + 1.5 \times (Q3 - Q1))$, while dots beyond these boundaries mark the outliers.

Consider the transition interval. All four players' transition intervals are on the order of three semitones wide, with insignificant differences. A reason could be that the pitches are constrained by the musical score, which limits the range of the transition interval. Wang and Pardue exhibit wider variabilities in their transition intervals, as indicated by the taller boxes, but it is likely that the transition intervals may vary more widely across musical pieces than between players. This hypothesis warrants further experiment and exploration.

Wang has the largest average slope value. Since the four players have similar transition intervals, it is expected that Wang, due to the high slope, has the lowest average transition duration, as confirmed by Figure 4.6. As expected, the slope and the transition duration are negatively correlated, where a larger slope indicates a lower transition duration or verse vera.

Figure 4.7 shows boxplots of the normalised time and pitch of the inflection point. Both erhu players tended to time their inflection points in the first half of the transition duration, while the violin players chose to put their inflection points around the middle of the transition duration. In contrast, the inflection pitch of the erhu players tended to lie around the middle of the transition

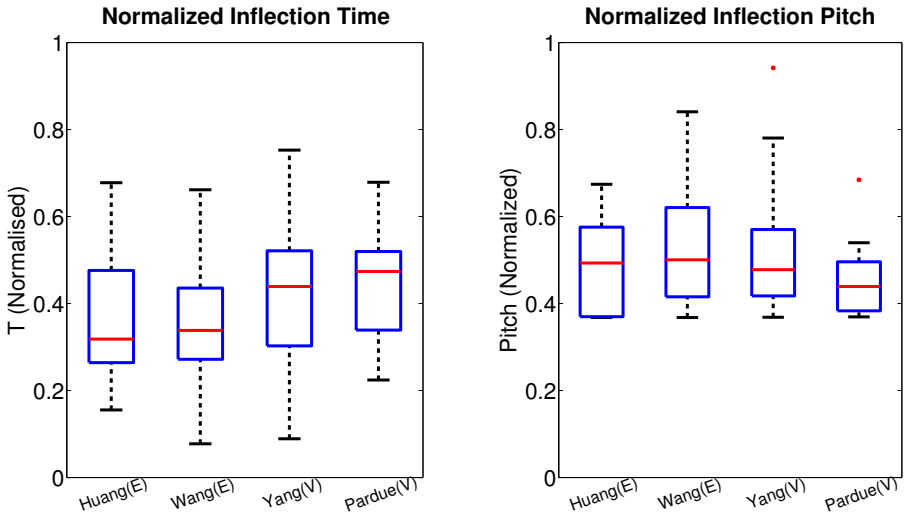


Figure 4.7: Boxplots of normalised inflection time and normalised inflection pitch for all four players of *Moon Reflected in Second Springs*. E: Erhu, V: Violin.

interval, while that of violin players are located lower in the interval.

4.2 Portamento Detection using the Hidden Markov Model

Portamento literature has been focused mainly on the modelling, and the musicological meaning of the portamento. To the best of the author's knowledge, there has not yet been developed an existing portamento detection algorithm. As has been stated in Section 2.1.2, portamento is an important expressive device. The automatic portamento detection method is beneficial to systematic portamento performance analysis, music expression synthesis, and automatic music transcription. In this section, we will describe the portamento detection method using the Hidden Markov Model method.

4.2.1 The Hidden Markov Model

The Hidden Markov Model (HMM) has been widely used in the analysis of speech and music in recent decades. HMM applications on speech recognition can be found in (Rabiner, 1989; Bishop, 2007). For music analysis, HMM has been applied to many music information retrieval areas, including music transcription (Raphael, 2002; Ryynänen & Klapuri, 2008) and score following (Pardo & Birmingham, 2005; Joder et al., 2010).

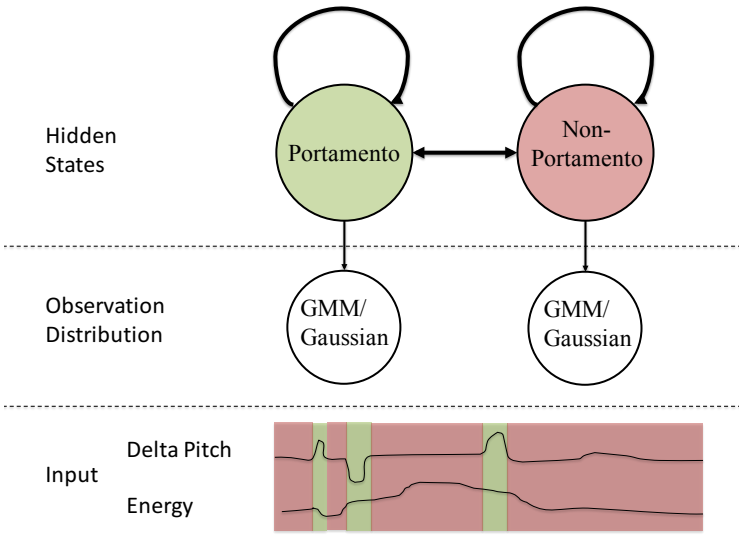


Figure 4.8: The Hidden Markov Model for portamento detection. The lower plot presents the delta pitch and energy time series of a music passage, the green part indicates the portamento and the red part indicates the non-portamento.

Observation of portamento reveals that modulation is present primarily in the pitch, and so we create a frame-wise portamento detection method using the Hidden Markov Model. To detect portamenti, we employ a fully-connected two-state HMM using the delta pitch (Δf_0) and/or energy ($\mathcal{A}$) time series as input as shown in Figure 4.8. The two states are portamento and non-portamento. Each state has an observational probability distribution describing the characteristics of delta pitch and/or energy for the portamento or non-portamento states. The best (most likely) path is decoded using the Viterbi algorithm.

The pitch time series were obtained from the pYIN method (Mauch & Dixon, 2014). The energy time series were calculated using the Root-Mean-Square equation:

$$\mathcal{A}(t_n) = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2)}, \tag{4.6}$$

where x_n is the nth audio sample's amplitude within a window. Consistent with the parameters set for the pYIN method, we choose a window size of 2048 samples and step size of 256 samples.

To explore the effect of the state observation probability distribution on performance, two different types of observation probability distribution were explored: Gaussian distribution, and Gaussian Mixture Model (GMM).

Observation Distribution using Gaussian Distribution

For the Gaussian observation distribution, we assume HMM each state has a single Gaussian distribution for each dimension, e.g. either one of the delta pitch or the energy has a single Gaussian distribution respectively. The multivariate Gaussian probability distribution is defined as:

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{4.7}$$

where d is the dimension of the $\boldsymbol{x}$. $\boldsymbol{\mu}$ is the d-dimensional mean vector and $\boldsymbol{\Sigma}$ is the $d \times d$ positive definite covariance matrix.

Observation Distribution using Gaussian Mixture Model

For the Gaussian Mixture Model observation distribution, we assume each state has more than one Gaussian distribution, which forms an observation distribution for either the delta pitch or energy. It has a better fitting performance when the distribution is not a single Gaussian distribution. The Gaussian mixture model can be considered a superposition of M Gaussian distributions having the form

$$p(\boldsymbol{x}) = c_m \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_m|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_m)\right), \tag{4.8}$$

where $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ are mean vector and covariance matrix for the mth Gaussian component. c_m is the mixing coefficient and $\sum_{m=1}^{M} c_m = 1$.

Figure 4.9 demonstrates the idea of a single Gaussian distribution and the GMM distribution with inputs delta pitch and energy. We can see that there are two peaks for the GMM with mixture number 2, which is better to accommodate the data characteristics than the single Gaussian distribution if the data has more than one peak.

4.2.2 Datasets

To evaluate the portamento detection method, we selected the violin, erhu and Beijing opera singing performances.

Moon Reflected in Second Springs Dataset

The first dataset used is *Moon Reflected in Second Springs* dataset, which was previously used in evaluation of automatic vibrato detection and analysis in Section 3.3. The dataset contains entire recordings of four performances of the traditional Chinese piece, *Moon Reflected in Second Springs*, which contains

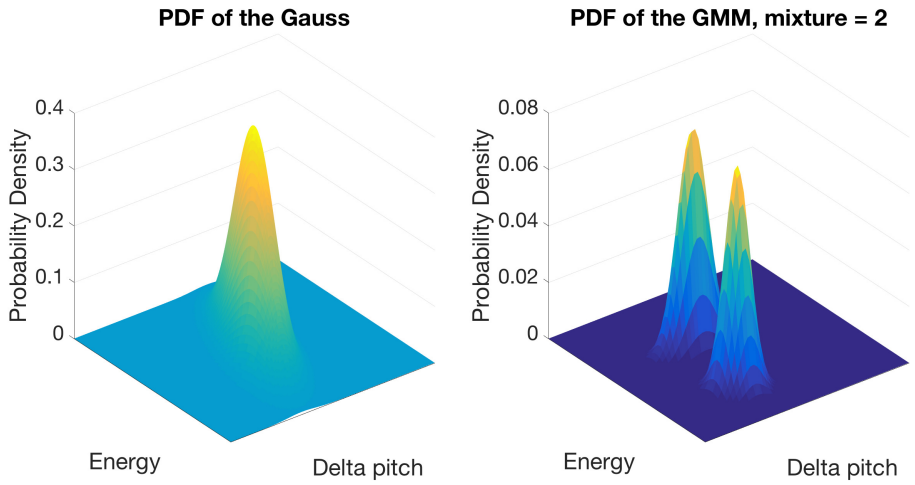


Figure 4.9: The Gaussian probability distribution and GMM probability distribution with mixture as 2.

No	Ins.	Performer	Durations(s)	# Portamenti
1	Erhu	Jiangqin Huang[a]	445.83	186
2		Guotong Wang[b]	387.53	169
3	Violin	Jiang Yang[c]	254.54	91
4		Laurel S. Pardue[c]	325.50	81

Table 4.4: *Moon Reflected in Second Springs* Dataset for portamento evaluation. *a*: (Huang, 2006), *b*: (Wang, 2009) and *c*: Recorded by the performer.

many portamenti. Two performances were recorded on the Chinese erhu and the other two on the Western violin. Table 4.4 lists the details of the dataset, which comprises 23.6 minutes of music, and the annotation of 527 portamenti. The portamenti were annotated by the author using the AVA interface, which will be described in Chapter 5. We can see that the erhu players are more likely to use portamento than the violinists. This may be due to the physical form of the instrument. The erhu has only two strings while the violin has four. Thus, erhu players have to initiate more slides to reach the target pitches while violin players are able to change strings to reach the target pitch without sliding.

Beijing Opera Singing Dataset

Observing the high presence of the portamento in Beijing opera singing, we use the selected recordings from the Beijing opera singing dataset create by Black et al. (2014). 16 performances from six different Chinese opera singers have been selected. The eight performances are from the Beijing opera role Laosheng(老生), while the other eight are from the Zhengdan(正旦) role. The portamento

annotation was created by the author using the AVA interface. The portamento number of each performance is given in Table 6.5 in Section 6.2.1. Note that these annotated portamenti will be used for a portamento performance analysis in Section 6.2.1.

4.2.3 Evaluation

To make the evaluation fair and effective we choose the k-fold cross-validation to test the portamento detection method. For each passage, we partition it into k equal sized subsamples. A single subsample is selected for use as the validation dataset, in order to test the algorithm; the other $k-1$ subsamples are used as training datasets. This process is repeated k times with each of the subsamples used once as the validation dataset. For each passage result, it would be the mean value for all of k results. In our experiment, we choose $k = 5$. The open source HMM Matlab toolbox[5] is employed to train our HMM model and to perform Viterbi decoding.

Train the HMM

The aims of the HMM training is to find out the initial state probabilities, state transition probabilities and the observation distribution parameters (i.e. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for Gaussian distribution and c_m, $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ of all components for GMM). For initialisation, the initial state probabilities and state transition probabilities are set randomly between 0 and 1. However, we found that the initialisation of parameters for Gaussian distribution and GMM are quite important to train the observation distributions. Instead of setting random initial parameters, we used the portamento part of the training set to train the observation distribution parameters of the portamento state, and the non-portamento part of the training set to train the observation distribution parameters of the non-portamento state.

Metrics

For frame-by-frame evaluation, a ground truth vector is created using the same sampling rate as the detection vector. The performance was then evaluated using the F-measure

$$F = \frac{2PR}{P+R}, \tag{4.9}$$

where precision, P, is defined as the number of true positive portamento frames divided by the total positive portamento frames, and recall, R, is defined as the number of true positive portamento frames divided by the total number of

[5] `github.com/qiuqiangkong/matlab-hmm`

portamento frames. Another metric is the overall accuracy which is defined by Dixon (2000):

$$A'_p = \frac{TP}{TP + FP + FN}, \tag{4.10}$$

where TP is the number of true positive portamento frames, FP is the number of false positive frames and FN is the number of false negative frames. Note that it is different from the accuracy, Eq. (3.31), which was used for vibrato parameter evaluation in Section 3.3.3. The F-measure and accuracy were calculated for each excerpt.

Results

Figure 4.10 presents the precision, recall and F-measure values using different observation distributions, i.e. Gaussian distribution and Gaussian Mixture Models with different mixture numbers. As expected, the HMM+GMM method has better performance than HMM+Gaussian. These improvements likely result from the improved capability of GMM to fit the observation distribution of the input data, i.e. delta pitch and/or energy. To be more specific, GMM increases the precision and recall values for *Moon Reflected in Second Springs*. Regarding the *Beijing Opera Singing* dataset, even though precision has deteriorated, the largely increased recall values pull the F-measure up. One of the reasons for the deteriorated precision is that the portamenti in vocal are lack of regularity.

Notably, it is observed that performance is not always improved significantly when the GMM mixture number increases. For the *Moon Reflected in Second Springs* sample, a mixture number $M = 3$ leads to a stable F-measure. While for *Beijing Opera Singing* dataset, $M = 2$ would be enough to obtain a reasonable result.

Portamento is a pitch-related expressive device, but a portamento also involves some degree of variation in energy. We thus explore whether adding the energy feature could improve portamento detection performance. To compare whether the addition of the energy improves the performance, we aggregate each passage's precision, recall or F-measure values using the same input data (i.e. f_0 or f_0 & $\mathcal{A}$) to obtain a mean value. The results are presented by Figure 4.11. Unexpectedly, the addition of the energy does not influence the precision, recall and F-measure values significantly. For the *Moon Reflected in Second Springs* dataset, the ANOVA results were $F = 4.44, P = .0367, df = 159$ for the precision, $F = 1.10, P = .295, df = 159$ for the recall, and $F = 2.68, P = .104, df = 159$ for the F-measure. For the *Beijing Opera Singing* dataset, the ANOVA results were $F = 4.79 \times 10^{-3}, P = .945, df = 639$ for the precision, $F = 1.44 \times 10^{-2}, P = .904, df = 639$ for the recall, and

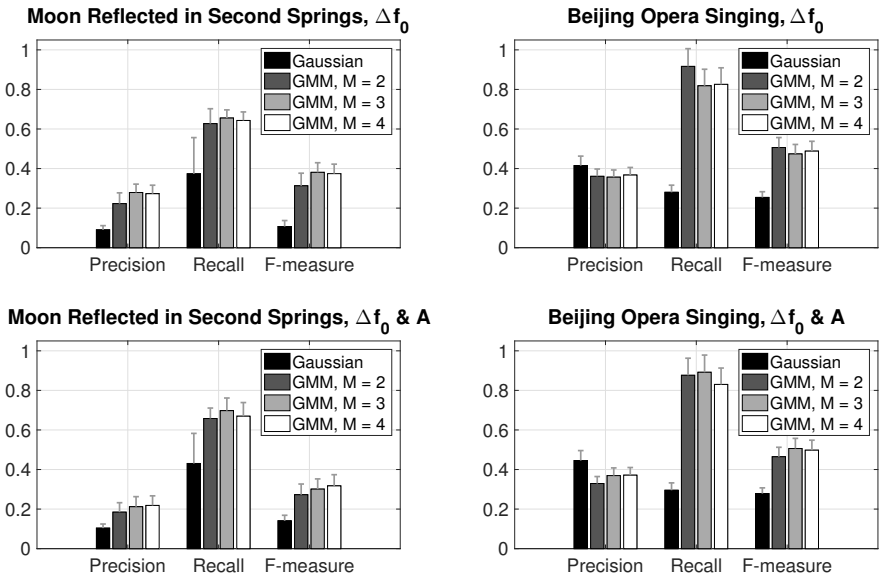


Figure 4.10: Comparison of the precision, recall and F-measure using different observation distribution for two datasets, *Moon Reflected in Second Springs* and *Beijing Opera Singing*. The two subplots in the first row used delta pitch as input and the two subplots in the second row used delta pitch and energy as input. Each precision, recall and F-measure value is the mean value of all passages for corresponding dataset. The error bars show the 95% confidence intervals around the corresponding mean values.

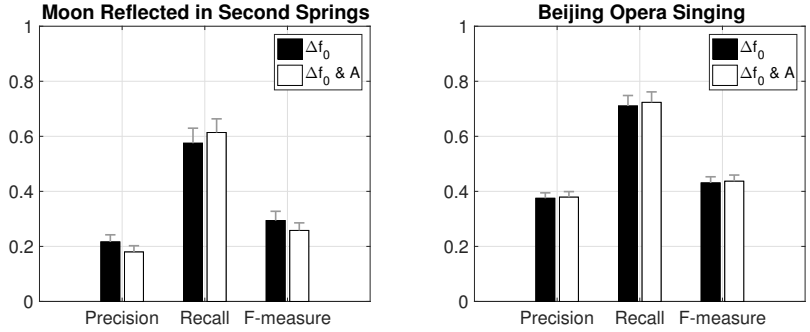


Figure 4.11: Comparison of the precision, recall and F-measure using Δf_0 v.s Δf_0 & $\mathcal{A}$ for two datasets, *Moon Reflected in Second Springs* and *Beijing Opera Singing*. Each precision, recall and F-measure value is the mean value of all passages for the corresponding dataset. The error bars show the 95% confidence intervals around the corresponding mean values.

$F = 9.01 \times 10^{-3}, P = .924, df = 639$ for the F-measure.

The accuracy comparison using different observation distributions is pre-

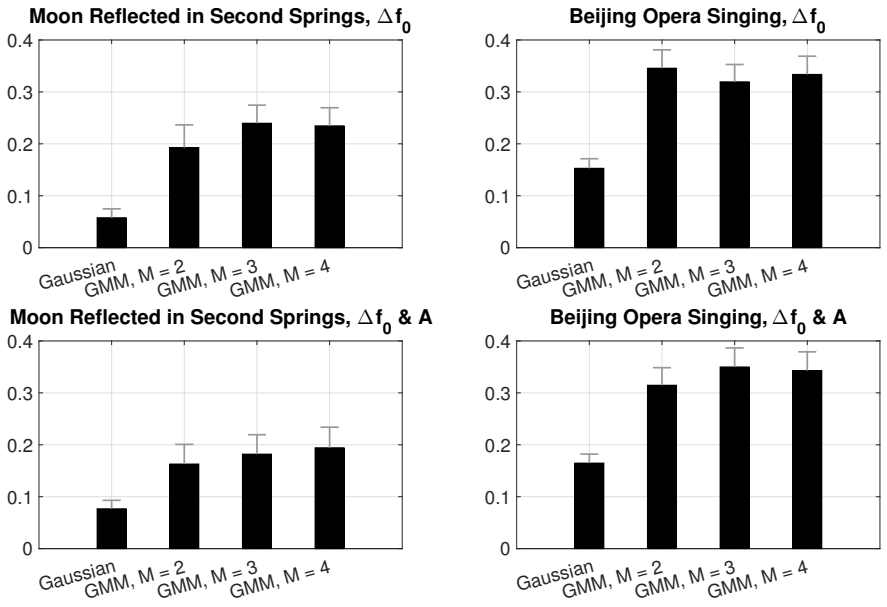


Figure 4.12: Comparison of accuracy resulting from different observation distributions for two datasets, *Moon Reflected in Second Springs* and *Beijing Opera Singing*. The two subplots in the first row used delta pitch as input and the two subplots in the second row used delta pitch and energy as input. Each accuracy value is the mean value of all passages for the corresponding dataset. The error bars show the 95% confidence intervals around the corresponding mean values.

sented in Figure 4.12. The accuracies also confirm that the HMM+GMM obtains the best performance than the combination HMM+Gaussian. Within the HMM+GMM combination, the accuracies are not always better with the increasing of the mixture number.

In terms of the accuracy, the portamento detection performance is not affected by adding the energy together with the delta pitch significantly. The results are shown by Figure 4.13. For the *Moon Reflected in Second springs* dataset, the ANOVA test results were $F = 3.31, P = 7.07 \times 10^{-2}, df = 159$. For the *Beijing Opera Singing* dataset, the ANOVA results were $F = 1.28 \times 10^{-2}, P = .9098, df = 639$.

4.3 Conclusions

In this chapter, we proposed a computational model of note transitions employing the Logistic Model. This model is able to fit discrete and continuous (portamento) note transitions. The Logistic Model is shown to have better performance than other methods, namely the Polynomial Model, Gaussian Model,

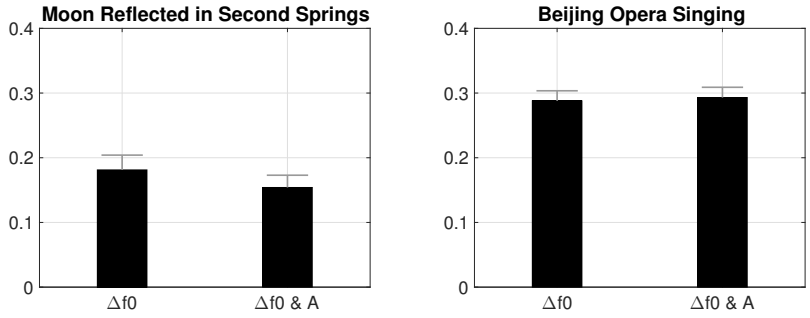


Figure 4.13: Comparison of the accuracy using Δf_0 versus Δf_0 & $\mathcal{A}$ for two datasets, *Moon Reflected in Second Springs* and *Beijing Opera Singing*. Each accuracy value is the mean value of all passages for the corresponding dataset. The error bars show the 95% confidence intervals around the corresponding mean values.

and Fourier Series Model. Moreover, parameters that convey musically meaningful information make the Logistic Model stand out from other methods. A case study on erhu and violin data was presented to demonstrate the feasibility of the Logistic Model in expressive music analyses.

For the four players analysed, the transition interval was found to be largely constrained by the score, the transition duration varied by player, with the duration being inversely related to the slope. The two erhu players tended to place the inflection time in the first half of the duration, while the two violin players tended to put it around the middle; the two violin players tended to place the inflection pitch in the lower half of the interval, while erhu players tended to put it around the middle.

This chapter represents the first effort towards portamento detection. The Hidden Markov Models with observation distribution modelling as Gaussian distribution and Gaussian Mixture Distribution have been employed. These two combinations were tested on a string and violin dataset and a Beijing opera singing (vocal) dataset. The results show that the HMM+GMM has better performance than HMM+Gaussian. However, the returns of increasing Gaussian mixture numbers quickly diminish, and so the performance does not significantly improve as this value increases. The delta pitch and energy have been used for detecting portamento. The combination of delta pitch and energy does not improve the portamento detection performance significantly, as compared to the use of the delta pitch alone. Further research should include a for extensive investigation into the use of energy or loudness features. Future work includes incorporating more features—for example, spectral flux, flatness and centroid—to the HMM-based portamento detection method, and employing the slope parameter in the Logistic Model in portamento detection.

Chapter 5

AVA: An Interactive Visual and Quantitative Analysis Tool of Vibrato and Portamento

The primary goal of this chapter is to introduce the AVA system[1] for interactive vibrato and portamento detection and analysis integrating previous vibrato (Chapter 3) and portamento (Chapter 4) detection and analysis modules. AVA seeks to fill the gap in knowledge discovery tools for expressive feature analysis of continuous pitch instruments. The AVA system is built on recent advances in automatic vibrato and portamento detection and analysis. As even the best algorithm sometimes produces erroneous vibrato or portamento detections, the AVA interface allows the user to interactively edit the detection solutions so as to achieve the best possible analysis results.

Vibrato and portamento use are important determinants of performance

[1]The beta version of AVA is available at `luweiyang.com/research/ava-project`.

styles across genres and instruments (Nwe & Li, 2007; Özaslan et al., 2012; Yang et al., 2013; Lee, 2006; Yang et al., 2015). Vibrato is the systematic, regular, and controlled modulation of frequency, amplitude, or the spectrum (Verfaille et al., 2005). Portamento is the smooth and monotonic increase or decrease in pitch from one note to the next (Yang et al., 2015). Both constitute important expressive devices that are manipulated in performances on instruments that allow for continuous variation in pitch, such as string and wind instruments, and voice. The labor intensive task of annotating vibrato and portamento boundaries for further analysis is a major bottleneck in the systematic study of the practice of vibrato and portamento use.

While vibrato analysis and detection methods have been in existence for several decades (see Section 2.3), there is currently no widely available software tool for interactive analysis of vibrato features to assist in performance and musicological research. Portamenti have received far less attention than vibratos due to the inherent ambiguity in what constitutes a portamento—beyond a note transition, a portamento is a perceptual feature that can only exist if it is recognisable by the human ear.

Applications of AVA include music pedagogy and musicological analysis. AVA can be used to provide visual and quantitative feedback in instrumental learning, allowing students to inspect their expressive features and adapt accordingly. AVA can also be used to quantify musicians' vibrato and portamento playing styles for analyses on the ways in which they use these expressive features. It can be used to conduct large-scale comparative studies, for example, of instrumental playing across cultures. AVA's analysis results can also serve as input to expression synthesis engines, or to transform expressive features in recorded music.

5.1 Integration of Vibrato and Portamento Detection and Analysis

Figure 5.1 shows AVA's system architecture. The system takes monophonic audio as input. The pitch curve, which is given by the fundamental frequency, is extracted from the input using the pYIN method (Mauch & Dixon, 2014). The first part of the system focuses on vibrato detection and analysis. The pitch curve derived from the audio input is sent to the vibrato detection module, which detects vibrato existence using a Filter Diagonalisation Method (FDM). The vibratos extracted are then forwarded to the module for vibrato analysis, which outputs the vibrato statistics.

The next part of the system deals with portamento detection and analysis.

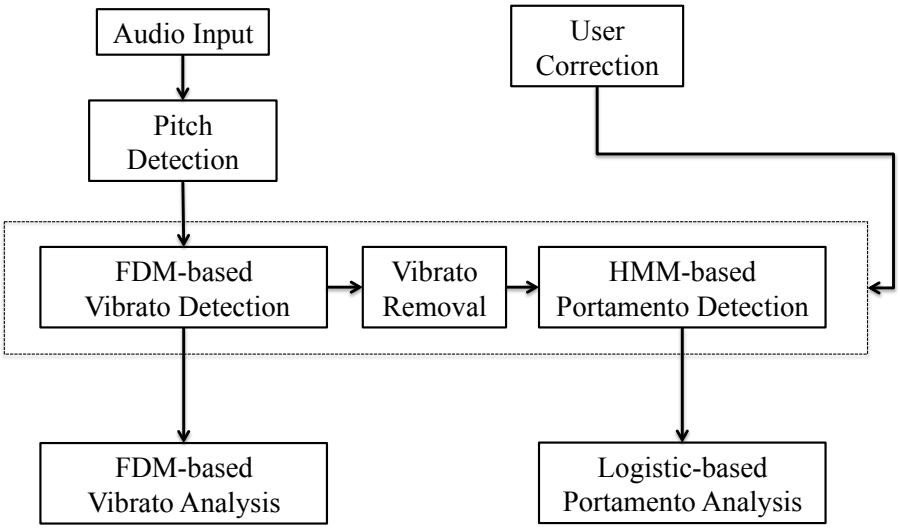


Figure 5.1: The AVA system architecture.

The oscillating shapes of the vibratos degrade portamento detection. To ensure the best possible performance for portamento detection, the detected vibratos are flatten using the built-in MATLAB function 'smooth'. The portamento detection module, which is based on HMM, uses this vibrato-free pitch curve to identify potential portamenti. A Logistic Model is fitted to each detected portamento for quantitative analysis.

For both the vibrato and portamento modules, if there are errors in detection, the interface allows the user to make up missing vibratos or portamenti and delete spurious results.

5.2 The Matlab Toolbox

This section describes the AVA Matlab interface that implemented the vibrato and portamento detection and analysis modules in Chapter 3 and Chapter 4 respectively.

AVA's Graphical User Interface (GUI) consists of three panels accessed through the tabs: Read Audio, Vibrato Analysis, and Portamento Analysis. The Read Audio panel allows a user to input or record an audio excerpt and obtain the corresponding (fundamental frequency) pitch curve. The Vibrato Analysis and Portamento Analysis panels provide visualisations of vibrato and portamento detection and analysis results, respectively.

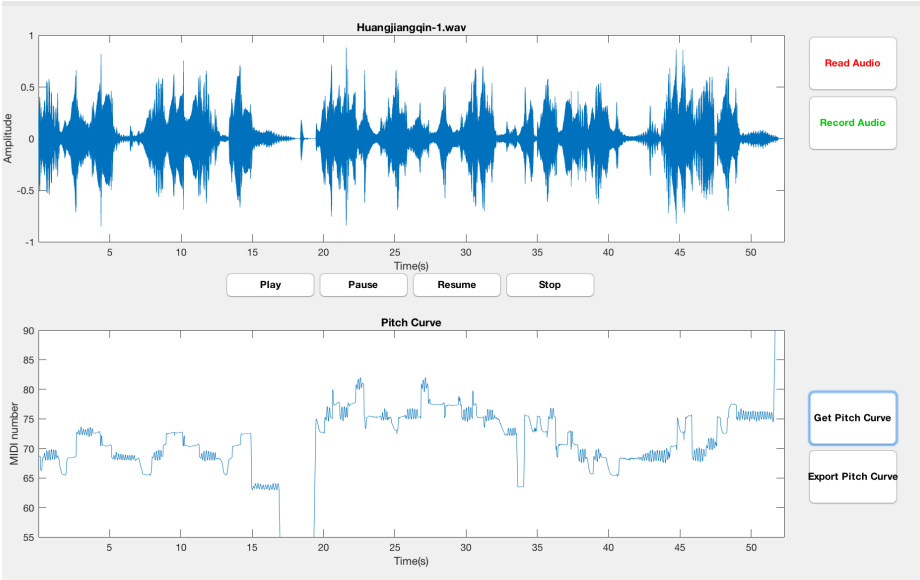


Figure 5.2: AVA screenshot: the read audio.

Figures 5.2, 5.3 and 5.4 shows screenshots of the AVA interface. Figure 5.2 presents the Read Audio panel for an erhu excerpt. Figures 5.3 and 5.4 shows the Vibrato Analysis and the Portamento Analysis panels analysing the same excerpt.

Our design principle was to have each panel provide one core functionality while minimising unnecessary functions having little added value. As vibratos and portamenti relate directly to the pitch curve, each tab shows the entire pitch curve of the excerpt and a selected vibrato or portamento in that pitch curve.

To allow for user input, the Vibrato Analysis and Portamento Analysis panels each have "Add" and "Delete" buttons for creating or deleting highlight windows against the pitch curve. Playback functions allow the user to hear each detected feature so as to inspect and improve detection results. To enable further statistical analysis, AVA can export to a text file all vibrato and portamento annotations and their corresponding parameters.

5.2.1 Read Audio Panel

We first describe the Read Audio Panel, shown in Figure 5.2. There is a "Read Audio" button in the top right allowing the user to input audio signal. Or the user is able to record an excerpt by clicking the button "Record Audio". The audio waveform of the entire excerpt is shown in the upper plot. A playback function is provided for the user. The corresponding fundamental frequency, f_0,

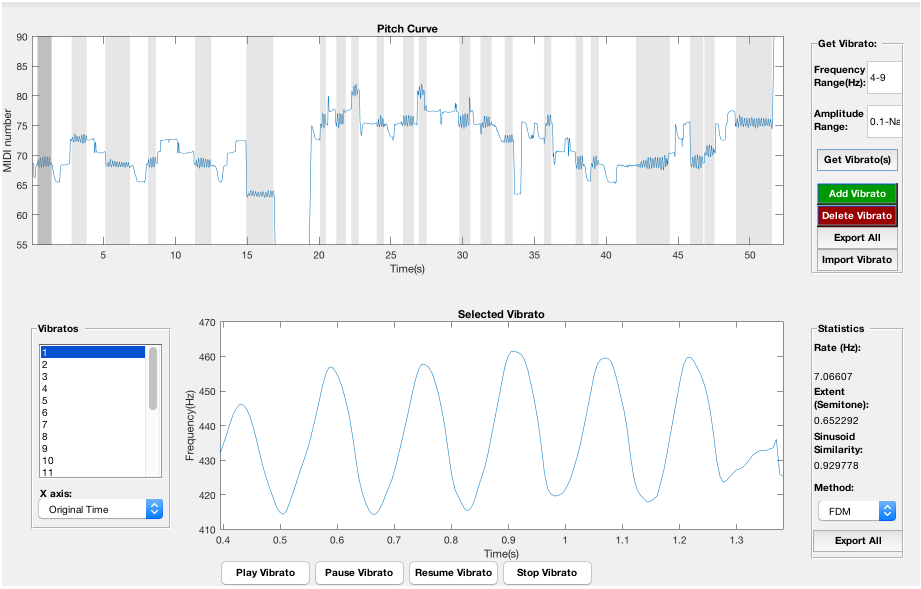


Figure 5.3: AVA screenshot: the vibrato analysis.

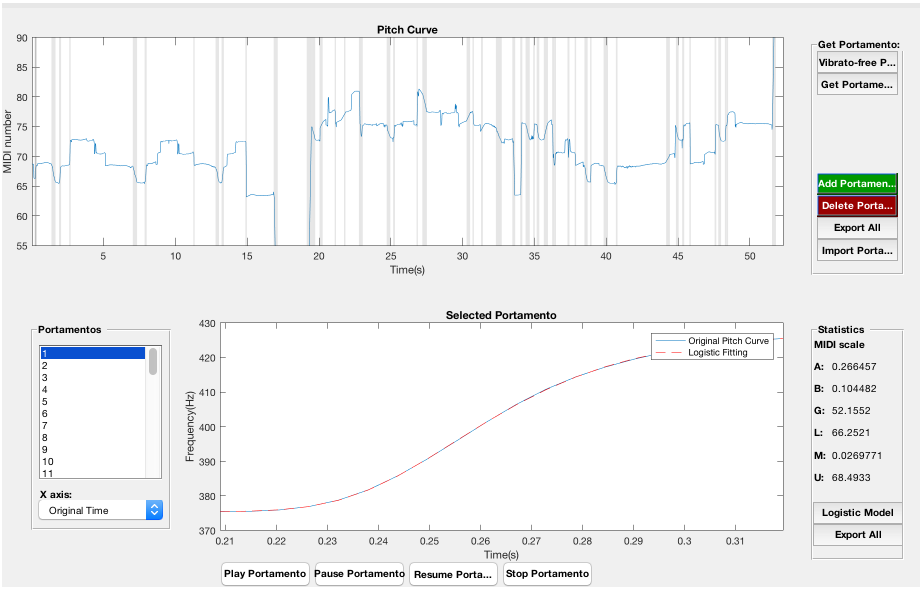


Figure 5.4: AVA screenshot: the portamento analysis.

is extracted by the button "Get Pitch Curve" using the pYIN method (Mauch & Dixon, 2014). The resulted f_0 is shown in the bottom plot. The button "Export Pitch Curve" is able to export the fundamental frequency for further pitch curve statistical analysis.

5.2.2 Vibrato Analysis Panel

The Vibrato Analysis Panel is described shown in Figure 5.3. The pitch curve of the entire excerpt is presented in the upper part, with the shaded areas marking the detected vibratos. Vibrato existence is determined using the method described in Section 3.2. The computations are triggered using the "Get Vibrato(s)" button in the top right, and the detected vibratos are highlighted by grey boxes on the pitch curve. Users can correct vibrato detection errors using the "Add Vibrato" and "Delete Vibrato" buttons.

The interface allows the user to change the default settings for the vibrato frequency and amplitude ranges; these adaptable limits serve as parameters for the Decision Tree vibrato existence detection process. In this case, with the vibrato frequency range threshold $[4, 9]$ Hz and amplitude range threshold $[0.1, \infty]$ semitones.

On the lower left is a box listing the indices of the detected vibratos. The user can click on each highlighted vibrato on the pitch curve, use the left- or right-arrow keys to navigate from the selected vibrato, or click on one of the indices to select a vibrato. The pitch curve of the vibrato thus selected is presented in the lower plot with corresponding parameters shown to the right of that plot.

In Figure 5.3, the selected vibrato has frequency 7.07 Hz, extent 0.65 semitones, and sinusoid similarity value 0.93. These parameters are obtained using the FDM-based vibrato analysis technique except the sinusoid similarity. Alternatively, using the drop down menu currently marked "FDM", the user can toggle between the FDM-based technique and a more basic Max-min method that computes the vibrato parameters from the peaks and troughs of the vibrato pitch contour.

Another drop down menu, labeled "X axis" under the vibrato indices at the bottom left, lets the user to choose between the original time axis and a normalised time axis for visualising each detected vibrato. A playback function assists the user in vibrato selection and inspection. All detected vibrato annotations and parameters can be exported to a text file at the click of a button to facilitate further statistical analysis.

5.2.3 Portamento Analysis Panel

Next, we present the functions available on the Portamento Analysis Panel, shown in Figure 5.4.

In the whole-sample pitch curve of Figure 5.4, the detected vibratos of Figure 5.3 have been flattened to improve portamento detection. Clicking on the "Get Portamenti" button initiates the process of detecting portamenti. The "Logistic Model" button triggers the process of fitting Logistic Models to all the detected portamenti.

Like the Vibrato Analysis panel, the Portamento Analysis panel also provides "Add Portamento" and "Delete Portamento" buttons for the user to correct detection errors. The process for selecting and navigating between detected portamenti is like that for the Vibrato Analysis panel.

When a detected portamento is selected, the best-fit Logistic model is shown as a red line against the original portamento pitch curve. The panel to the right shows the corresponding Logistic Model parameters. In the case of the portamento highlighted in Figure 5.4, the growth rate is 52.16 and the lower and upper asymptotes are 66.25 and 68.49 (in MIDI number), respectively, which could be interpreted as the antecedent and subsequent pitches. From this, we infer that the transition interval is 2.24 semitones.

As with the Vibrato Analysis panel, a playback function assists the user in portamento selection and inspection. Again, all detected portamento annotations and parameters can be exported to a text file at the click of a button to facilitate further statistical analysis.

5.3 Conclusions

In this chapter, we have presented an interactive visual and quantitative analysis tool of vibrato and portamento, AVA. The system was implemented in MATLAB, and the GUI provides interactive and intuitive visualisations of detected vibratos and portamenti and their properties.

For vibrato detection and analysis, the system implements a Decision Tree for vibrato detection based on FDM output and an FDM-based vibrato analysis method. The system currently uses a Decision Tree method for determining vibrato existence; a more sophisticated Bayesian approach taking advantage of learned vibrato rate and extent distributions is described in (**?**). While the Bayesian approach has been shown to give better results, it requires training data; the prior distributions based on training data can be adapted to specific instruments and genres.

For portamento detection and analysis, the system uses an HMM-based por-

tamento detection method with Logistic Models for portamento analysis. Even though the combination of HMM and GMM has been employed, the portamento detection method sometimes mis-classifies normal note transitions as portamenti, often for notes having low intensity (dynamic) values. While there were significant time savings over manual annotation, especially for vibrato boundaries, corrections of the automatically detected portamento boundaries proved to be the most time consuming part of the exercise. It is worth trying more features in the future.

There is some room to further improve AVA. The vibrato and portamento detection and analysis are based upon the pitch curve. Thus, the pitch detection method exerts vital influence on AVA. Although the one of state-of-the-art single pitch detection methods, pYIN, has been incorporated, AVA is still restricted on monophonic audio. One way to improve the AVA is to incorporate and adopt state-of-the-art melody detection method, e.g. MELODIA (Salamon & Gómez, 2012), or polyphonic pitch detection methods, e.g. (Klapuri, 2003; Benetos & Dixon, 2010; Yeh, 2008). Another way to improve AVA is to complete it as a computer-aided music tutoring system that assists in training in performance and expressivity. Timmers & Sadakata (2014) has shown that performers are open-minded to use technology to improve their expressive aspects of performance. Besides the vibrato and portamento analysis, AVA could consist of intonation analysis which assists performers in tune, and rhythmic analysis which helps comprehension of rhythm. The visual feedback and quantitative analysis of performances allows performers to inspect expressive features and adapt accordingly. One may re-implement AVA as a real-time system using live coding. The instant feedback would be more convenient and effective for performance learning.

Chapter 6

Vibrato and Portamento Analysis: Case Studies

This chapter presents two case studies on vibrato and portamento analysis using actual music performances.

The first case study provides the main motivation for the entire thesis. It examines differences in vibrato characteristics between erhu and violin performances of the same piece of music. This experiment compares the same notes in each performance and employs manual annotation of vibrato onset and offset. Significant differences were found between erhu and violin vibrato playing styles. Erhu performers employed larger vibrato extents than violin performers. They also used wider vibrato extent ranges than their violin-playing counterparts. This case study demonstrates that the vibrato performance styles can vary widely even when considering performances of the same piece of music on different instruments.

The nature of the time and labour intensive experiment prompts the question of whether we can design and create a vibrato and portamento analysis system that can be scaled to large datasets and made accessible to users without strong technological and programming backgrounds.

The second case study aims to demonstrate the feasibility and usability of the AVA system for vibrato and portamento analysis on real performances. For this case study, we analysed two datasets. The first is drawn from Beijing opera singing, where AVA reports larger vibrato extents in the singing of the Laosheng role comparing to that of the Zhengdan role, and the singing of the Zhengdan role has faster portamento slides than that for the Laosheng role. The second is a smaller dataset of erhu and violin performances, where AVA is able to confirm the discoveries of the first case study on vibrato and determines the portamento

differences between erhu and violin playing.

6.1 Cross-cultural Analysis of Vibrato Performance Style on Erhu and Violin

To investigate the difference in vibrato style between Chinese and Western music performance, we compare performances on the erhu (see Figure 2.2) and violin as a case study. In this section, we present computational analyses of vibrato in performances of the same piece of music created by erhu and violin players.

In this experiment, we consider the performances of a single piece of music by a number of erhu and violin players. We compare recordings by six different erhu players, and six distinct violin performers, all of the same piece of music. 20 notes are selected from each performance for close examination. We will compare the exact 20 notes on vibrato performance from each performance.

We utilise a number of vibrato characteristics in our study. Vibrato rate and vibrato extent are the most important parameters as vibrato rate determines the speed of the vibrato, while vibrato extent gives its depth. Vibrato structure is represented by vibrato sinusoid similarity. As for the form of the vibrato, the envelope hump number(extracted from vibrato f_0 envelope) is used to assess the variation in vibrato extent.

We expect that the vibrato characteristics will help reveal the differences in musical genre and instrumental styles. It is our aim to answer the following questions:

1. What are the vibrato characteristics of erhu players?
2. What are the vibrato characteristics of violin players performing Chinese music?
3. Is there any difference between the ways erhu players and violinists play vibratos when performing the same piece of music?

It has been posited that vibrato is influenced by several factors including musical context, tone length and musical expression (Prame, 1997). To investigate vibrato differences that are introduced by erhu and violin and minimise the effect of other factors, comparisons are made between performances of the same piece of music, and the same notes within the piece.

Instrument	Index	Performer	Nationality
Erhu	1	Guotong Wang(王国潼)	China
	2	Jiangqin Huang(黄江琴)	China
	3	Wei Zhou(周维)	China
	4	Jiemin Yan(严洁敏)	China
	5	Huifen Min(闵惠芬)	China
	6	Changyao Zhu(朱昌耀)	China
violin	7	Laurel Pardue	U.S.A
	8	Lina Yu(俞丽拿)	China
	9	Baodi Tang(汤宝娣)	China
	10	Takako Nishizaki(西崎崇子)	Japan
	11	Yanling Zhang(张延龄)	China
	12	Yangkai Ou(欧阳锴)	China

Table 6.1: Selected Performances for *Moon Reflected in Second Springs*.

6.1.1 Dataset

Performance selection

We choose a well known Chinese piece called *Moon Reflected in Second Springs*. This piece of music describes the sufferings of a blind composer, and is idiomatic of Chinese traditional music.

The 12 performances that form the focus of this analysis are shown in Table 6.1. Performances 1 to 6 are commercial CD recordings by professional erhu players. These six erhu performers are famous in the erhu music community, and they have each received systematic education in music conservatories. Performances 7 to 12 are recordings by six violin players. Since this piece of music was originally composed for erhu, professional violin performances of this piece on commercial CD recordings are relatively scarce. Only 8, 9 and 10 are commercial CD recordings. 7 is a recording of an unaccompanied performance of the piece by Laurel Pardue. 11 and 12 were found online; the performers are Chinese violin pedagogues. With respect to nationality, the 6 erhu players are all from China. For violin, the 7th player is from the U.S.A, and the 10th player is from Japan. The other violinists are from China.

Notes Selection

To select the notes, the following criteria were employed:

1. The note should not be played on an open string as performers cannot apply vibrato to such notes.

2. The note should be of relatively long duration. Since vibrato rate is typically around 4-8 Hz, if the note duration is too short it is difficult for

Performance Number		Notes Number																		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Erhu	1	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	3	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	*	✓	✓	✓	✓	✓
	4	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	5	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	6	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	*	✓	✓	✓	✓
Violin	7	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	*	✓	*
	8	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	*	✓	✓	✓	✓
	9	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	10	✓	✓	*	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	11	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	12	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	*	*	✓	*

Table 6.2: Notes selection for each performance of *Moon Reflected in Second Springs* for manual vibrato comparison. ✓indicates the note has vibrato. * indicates the note has no vibrato. ✗indicates there is no such note.

the performer to apply vibrato, and for listeners to perceive it as vibrato. In this case study, a note of long duration was one lasting more than 0.5 seconds.

3. The note should be of high amplitude. If the note has low amplitude or exhibits a diminuendo, it will pose difficulties in the measurement of the vibrato parameters, especially in pitch detection. For low amplitude notes, any noise will significantly impact signal acquisition and extraction, providing less reliable data for pitch detection.

After applying the above rules, 20 notes in the second performance were selected. To make the results as unbiased as possible, the same notes were selected from all performances. When a composition is transcribed for other instruments, it is not uncommon to see some degree of changes and recomposition. Here, all six erhu performers used almost exactly the same composition. However, the transcription for violin, while preserving most of the original composition, had introduced some changes to the erhu version. This difference is evident in Table 6.2 which shows the same selected 20 notes for each performance. It tells whether this note is included in the corresponding performance or not, and whether this note has vibrato. The 7th violin performer applied vibrato to almost all the same notes as the erhu performers. In contrast, the 8th, 9th, 11th and 12th performances all used the same transcription. This variation did not include the first phrase, which contained the first 5 selected notes, and they were thus not found. The 10th performance, which was by a Japanese, used another version. This version did not include two notes numbered 7 and 8. Consequently, there were 204 vibrato notes out of total 218 notes.

6.1.2 Methodology

Vibrato Fundamental Frequency Extraction

The vibrato parameters were extracted from the note fundamental frequency, and the fundamental frequency obtained using the Praat (Boersma, 2001) software. Praat performs robustly for fundamental frequency extraction of monophonic audio. The six erhu performances were monophonic and without accompaniment; the vibrato fundamental frequencies were directly extracted using Praat. For polyphonic audio, Praat cannot provide the same reliability in extracting the fundamental frequencies as for monophonic audio. Except for the 7th performance, all other violin performances had accompaniment.

For polyphonic textures, we applied the method described by Desain & Honing (1996). With knowledge of the expected pitch, the spectrum was filtered around the pitches of the melody. As the violin melody may be close to the accompaniment, a higher harmonic was chosen for filtering instead of the violin,Äôs fundamental frequency. The method is demonstrated in Figure 6.1, where one of the higher harmonics is selected for filtering. The filter pass band and stop band are readily identified in a higher frequency range. This filtered area could then be used by Praat to provide robust fundamental frequency extraction.

Vibrato Parameters Extraction

In this experiment, we will report vibrato rate, extent, sinusoid similarity and the envelope hump number. The details of the vibrato parameter extractions are described in Section 3.1.

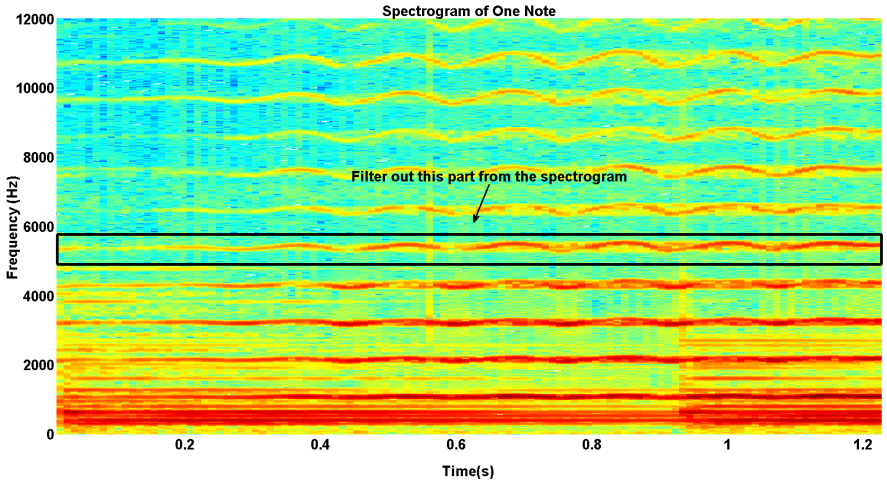


Figure 6.1: Higher harmonic filtered out from spectrogram of one note.

6.1.3 Results and Discussions

In this section, we report the vibrato rates, extents and sinusoid similarities, vibrato envelope hump numbers for all erhu and violin players.

Vibrato Rate

Figure 6.2 shows the erhu and violin performers' mean, min and max vibrato rates on the left, middle and right, respectively. Here, the mean vibrato rate is defined as the mean vibrato rate of all of an individual performer's extracted vibratos. Min/max vibrato rate refers to the min/max value that occurred for an individual. The bar in the middle of the box indicates the median vibrato rate. The upper and lower edges of the box indicates the 75th and 25th percentiles. Dotted lines extend to the extreme values. Plus signs mark the outliers.

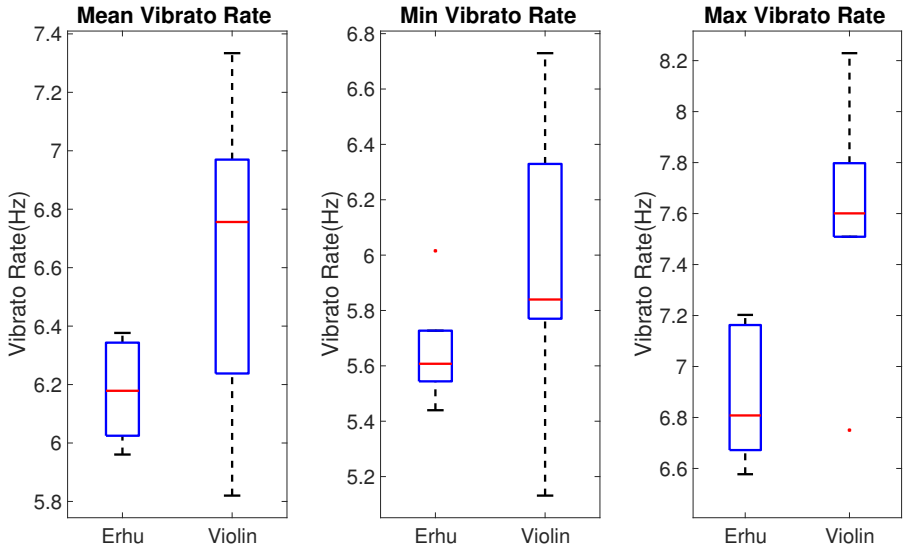


Figure 6.2: Boxplots of mean, minimum and maximum vibrato rates for erhu and violin instruments.

Regardless of whether one considers the mean, min or max vibrato rates, the violin is always larger in value than the erhu. Violin performers tend to apply faster vibrato rates than erhu performers. The violin vibratos also had a wider range than those for erhu for both the mean and min rates, which means that the vibrato rate varied sharply among our violin performers. Note that although the median of the violin vibrato rate is greater than that for erhu, the lower extreme of the violin vibrato rate is lower than that of erhu. The violin performers we considered seem to demonstrate more variability in vibrato rate.

Table 6.3 provides more details of the results for each performer. Erhu players have vibrato rate values ranging between 5.44 and 7.20Hz, with a mean

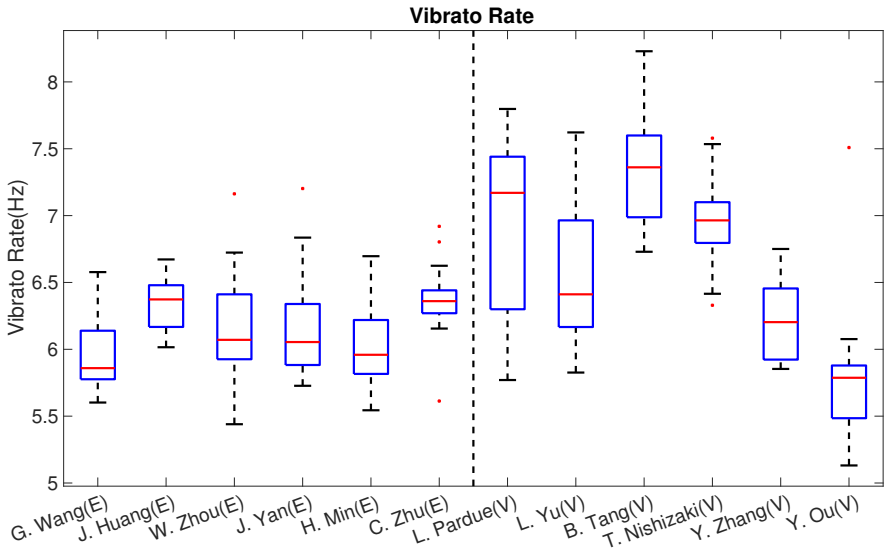


Figure 6.3: Boxplots of vibrato rates for erhu and violin players.

value of 6.18Hz. Violin performers have a vibrato rate ranging between 5.13 to 8.23Hz, with a mean value of 6.65Hz.

Previous studies have shown that, in the Western classical music, the vibrato range is 4 to 12Hz (Desain & Honing, 1996). Seashore noted that vibrato rates between 5 and 6Hz were more common (Seashore, 1938). Specifically, Prame observed that the vibrato rate range amongst 10 singers was between 4.6 and 8.7Hz, with an average value of 6Hz (Prame, 1994). For violin, Seashore found that violinists had 6.5Hz vibrato rate in average with a small range between individuals. Besides, he also discovered violinists and vocalists had same vibrato rate. Desain et al confirmed Seashore's result by seeing the violin vibrato rate was generally around 6.5Hz (Desain et al., 1999). The vibrato rate of violin performers is similar to the values reported in Western music studies, even though the violinists were playing Chinese traditional music.

If vibrato characteristics relate to musical styles and instruments, one might expect erhu and violin vibrato rates to be significantly different. Özaslan et al. (2012) observed that the vibrato rate in Turkey's Mamkan music was between 2 to 7Hz, which is significantly different from that for Western music.

An ANOVA (Ott & Longnecker, 2010) analysis reveals an unexpected result, that the erhu recordings showed no significant difference from the violin recordings in terms of mean vibrato rate ($F = 3.93, P = 7.57 \times 10^{-2}, df = 11$), min vibrato rate ($F = 1.43, P = .259, df = 11$), and max vibrato rate ($F = .950, P = 1.02 \times 10^{-2}, df = 11$). The results show that we can reject the hypothesis that the mean vibrato rates are the same with $P < 0.10$, and we can

reject the hypothesis that the max vibrato rates are the same with $P < 0.05$. There is little evidence to reject the hypothesis that the min vibrato rates are the same.

A further player-wise vibrato rates boxplots are represented by Figure 6.3. Violin players have much variations across players. The first four violin players, L. Pardue, L. Yu, B. Tang and T. Nishizaki, have higher vibrato rates than erhu players. However, violin players Y. Zhang and Y. Ou have similar vibrato rates with erhu players. Note that Y. Ou has smaller vibrato rate values.

Vibrato Extent

Figure 6.4 shows the erhu and violin performers' mean, min and max vibrato extents on the left, middle and right, respectively. Here, the mean vibrato extent is defined as the mean vibrato extent of all of an individual's vibratos. Min/max vibrato extent refers to the min/max value of the extent of the vibratos played by an individual.

Unlike vibrato rate, the mean, min and max vibrato extent values for erhu were much larger than for violin as shown in Figure 6.4. Erhu performers tend to play vibratos with larger extents than violin performers. Moreover, erhu has wider vibrato extent ranges than violin for all these parameters. The vibrato extent varies more widely among erhu performers, indicating that erhu performers have more variability in their vibrato extents.

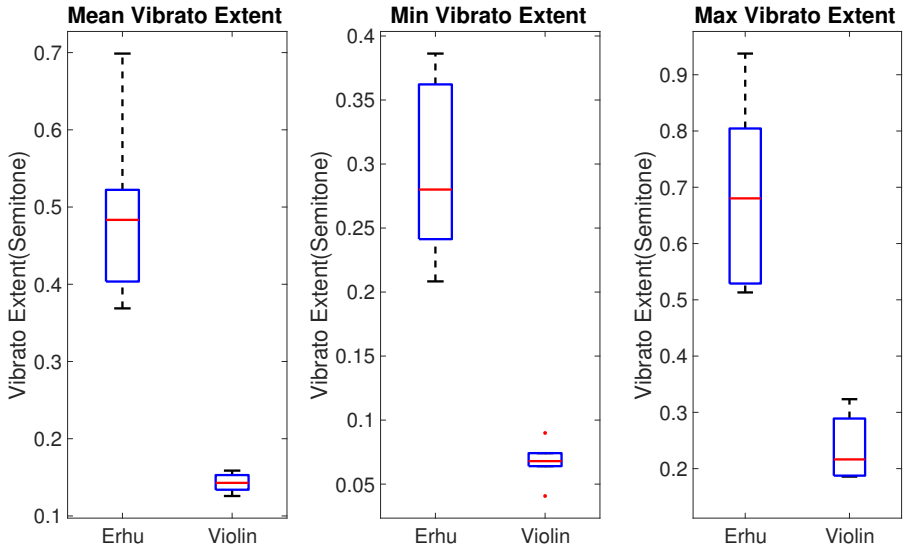


Figure 6.4: Boxplots of mean, minimum and maximum vibrato extents for erhu and violin instruments.

Table 6.3 shows that the six violin performers have vibrato extents ranging

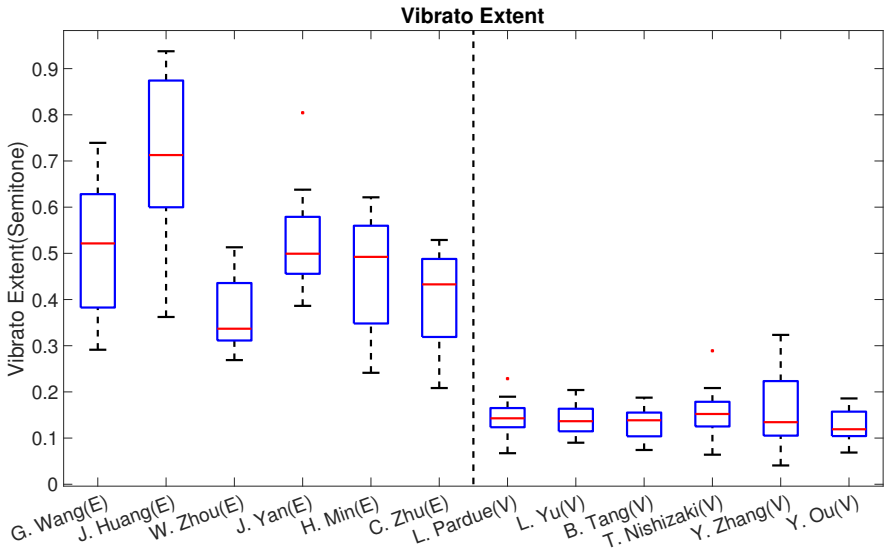


Figure 6.5: Boxplots of vibrato extents for erhu and violin players.

from 0.04 to 0.32 semitones, with a mean of 0.14 semitones. In the literature, Seashore found the violinist's vibrato extent was half as wide as singer's with a value 0.25 semitone. Prame stated that violin vibrato extents were no more than half that of the vibrato of singers; the ten singers he studied had vibrato extents ranging from 34 to 123 cents, with a mean of 71 cents (Prame, 1997). Thus, the statement suggests that the violin extent would be no more than 17 to 62.5 cents, with a mean of 30.5 cents. In another study, the vibrato extent for Western string players was shown to be 0.2 to 0.35 semitones (Timmers & Desain, 2000). The vibrato extent of violin in the present study is very close to these reported in the literature. Although the violin performers were playing Chinese traditional music, their vibrato extent did not exceed that reported in the Western music studies.

Erhu performers, on the other hand, have a larger vibrato extent, from 0.21 to 0.94 semitones, with a mean of 0.49 semitones. Not only are the lower limit, upper limit and the mean value of the erhu vibrato extents larger than that of the violin, the range of the vibrato extent of erhu (0.73 semitones) is also wider than that for violin (0.28 semitones). This implies that erhu performers exercised greater variability in changing the vibrato extent than violin performers.

This observation may stem from differences in the left hand movements when playing the two instruments. For the violin, the lower left arm of the player angles up to the finger board and the vibrato movements are lateral along the horizontal finger board. For the erhu, the lower left arm of the player is more or less horizontal, and the vibrato movements are up and down along the vertical

Performance Number		Vibrato Rate(Hz)				Vibrato Extent(Semitone)			
		Mean	**Min**	**Max**	**SD**	**Mean**	**Min**	**Max**	**SD**
Erhu	1	5.96	5.60	6.58	0.70	0.51	0.29	0.74	0.11
	2	6.34	6.02	6.67	0.64	0.70	0.36	**0.94**	0.12
	3	6.18	**5.44**	7.16	1.01	0.37	0.27	0.51	0.08
	4	6.17	5.73	**7.20**	0.90	0.52	0.39	0.80	0.10
	5	6.02	5.54	6.70	0.58	0.45	0.24	0.62	0.07
	6	6.38	5.61	6.92	0.72	0.40	**0.21**	0.53	0.06
	Average	6.18	5.66	6.87	0.76	0.49	0.29	0.69	0.09
Violin	7	6.97	5.77	7.80	0.92	0.14	0.07	0.23	0.03
	8	6.55	5.83	7.62	1.05	0.14	0.09	0.20	0.04
	9	7.33	6.73	**8.23**	0.85	0.13	0.07	0.19	0.04
	10	6.97	6.33	7.58	0.92	0.15	0.06	0.29	0.04
	11	6.24	5.85	6.75	0.83	0.16	**0.04**	**0.32**	0.05
	12	5.82	**5.13**	7.51	0.97	0.13	0.07	0.19	0.03
	Average	6.65	5.94	7.58	0.92	0.14	0.07	0.24	0.04

Table 6.3: Statistics of vibrato rate and extent for 12 Performances.

strings. The vertical alignment of the erhu strings and the corresponding hand motions may allow for larger vibrato movements.

The fingerboard exists in the violin, but not the erhu. When a violin player presses the string, the string touches the fingerboard. However, when an erhu performer presses the string, nothing else is touched. This absence of a fingerboard may give erhu performers more flexibility to create wide vibratos. This hypothesis requires further research. The ANOVA analysis shows high confidence that the mean vibrato extent ($F = 53.2, P = 2.62 \times 10^{-5}, df = 11$), min vibrato extent ($F = 60.6, P = 1.49 \times 10^{-5}, df = 11$), and max vibrato extent ($F = 39.9, P = 8.70 \times 10^{-5}, df = 11$) are significantly different between erhu and violin players.

Figure 6.5 shows the individual vibrato extent value for each erhu and violin performer. Erhu players have much variation across performers. Compared to vibrato rate, violinists have a more consistence in vibrato extent. They have similar median, 75th and 25th percentiles except Y. Zhang who has a bigger standard deviation. The standard deviation of the erhu vibratos are markedly larger than that for violin.

Vibrato Rate Range and Vibrato Extent Range

The left and middle parts of Figure 6.6 show the vibrato rate range and vibrato extent range for erhu and violin. The vibrato rate range for individual players results from the difference between their min vibrato rates and max vibrato rates. The vibrato extent range was obtained in a similar way. Table 6.4 provides further details. Violin performers have slightly wider vibrato rate ranges than erhu performers. However, this difference is not significant ($F = 2.54, P = .142, df = 11$), meaning that erhu performers could have a similar vibrato rate

Performance Number		Rate Range(Hz)	Extent Range(Semitone)	Sinusoid Similarity
Erhu	1	0.98	0.45	0.85
	2	0.66	0.58	0.85
	3	1.72	0.24	0.77
	4	1.48	0.42	0.86
	5	1.15	0.38	0.89
	6	1.31	0.32	0.88
Average		1.21	0.40	0.85
Violin	7	2.03	0.16	0.68
	8	1.80	0.11	0.74
	9	1.50	0.11	0.66
	10	1.25	0.22	0.80
	11	0.90	0.28	0.77
	12	2.38	0.12	0.77
Average		1.64	0.17	0.74

Table 6.4: Vibrato Rate Range and Vibrato Extent Range for Each Performer.

range as violin performers. For the vibrato extent range, erhu performers have much larger values than violin performers. This difference is significant ($F = 17.6, P = 1.80 \times 10^{-3}, df = 11$), meaning that erhu performers vary their vibrato extents more widely than violin performers. This suggests that erhu performers may use vibrato extent more expressively in their playing.

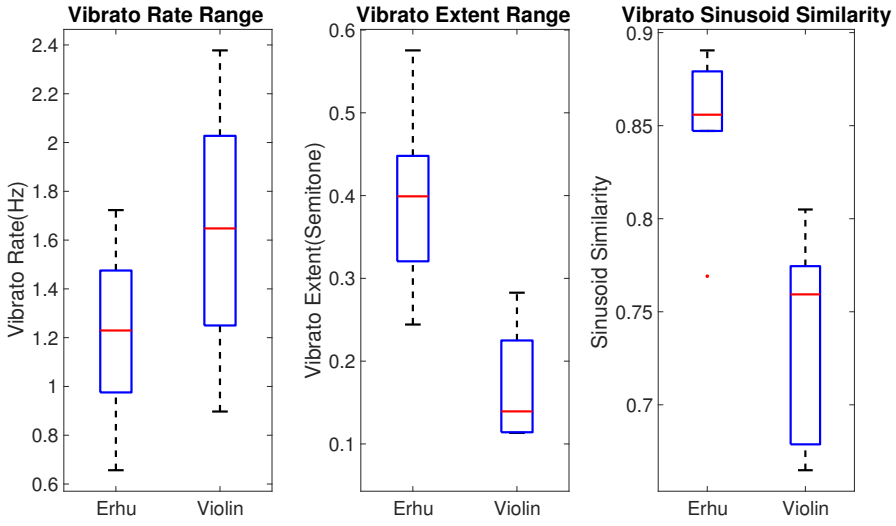


Figure 6.6: Boxplots of vibrato rate range (in Hz), vibrato extent range (in semitones) and mean vibrato sinusoid similarity for erhu and violin instruments.

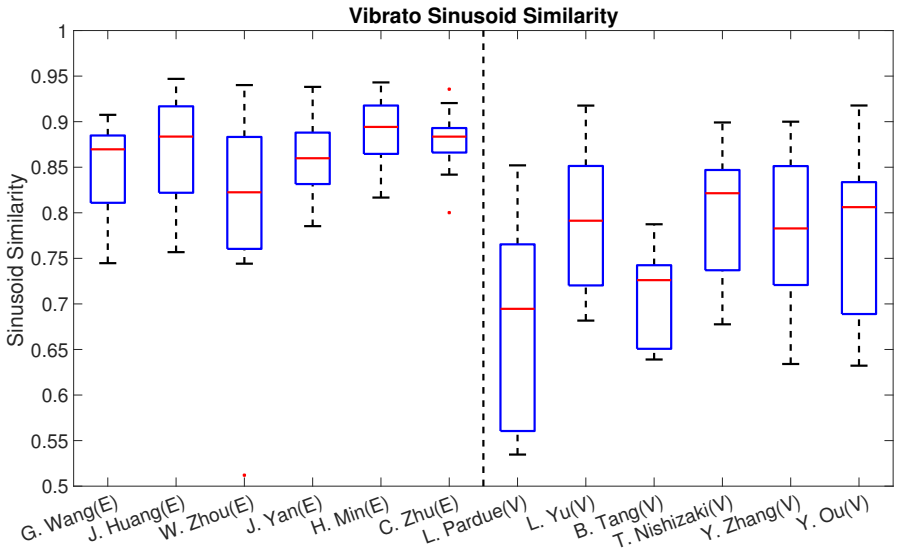


Figure 6.7: Boxplots of vibrato sinusoid similarities for erhu and violin players.

Vibrato Sinusoid Similarity

The mean vibrato sinusoid similarity is shown on the right in Figure 6.6. The vibrato shape of erhu performers is much more similar to a sinusoid than that for violinists. This difference is significant ($F = 14.3, P = 3.60 \times 10^{-3}, df = 11$), as validated by the ANOVA analysis.

In order to try to find any relationships between sinusoid similarity and other parameters, Pearson correlation analysis has been applied. However, no significant relationships were found. The vibrato sinusoid similarity is thus uncorrelated with the other parameters. Whether the vibrato shape is related to the vibrato rate and extent needs further study.

From the player-wise aspect, Figure 6.7 presents the vibrato sinusoid similarity across erhu and violin performer. L. Pardue shows the widest sinusoid similarity range and lowest values.

Vibrato Envelope Hump Number

Pearson correlation shows that the number of the vibrato envelope humps is strongly and positively correlated with note duration (in beats). This phenomenon was observed in both erhu ($P = 7.92 \times 10^{-4}, r = .688, df = 18$), as well as in violin ($P = 6.94 \times 10^{-4}, r = .694, df = 18$) performances, indicating that erhu and violin performers introduced more vibrato extent variation cycles as the number of beats increases. We noted that the r-value for erhu and that for violin are very close, meaning that erhu and violin players employed the

same degree of vibrato envelope variation with the number of beats.

Figure 6.8 shows the number of vibrato envelope humps and note duration (in beats) indexed by vibrato number on the x-axis. It is interesting to observe that erhu and the violin vibrato hump numbers varied in the same ways. The hump numbers peak for vibratos 2, 7, 9, 13, 15 and 18, while the numbers drop for vibratos 3, 8, 12, 14 and 16. With the exception of vibratos 15 and 18, high values of note duration (number of beats) correspond to peaks in the number of envelope humps. Except for vibrato 16, low hump envelope numbers occur at vibratos with similarly low note durations. To some extent, erhu and violin players adapted vibrato extent variations to the number of beats in the same fashion. One exception is vibrato 15: it has the same number of beats as vibratos 14, 16 and 17; however, both erhu and violin players alike introduced more extent variations for this vibrato than the other vibratos.

Although erhu and violin players exhibit significantly different vibrato extents, they vary the vibrato extents in a similar way. Long notes produce typically more variations in vibrato extent. Due to human physiology, it is difficult to maintain the same vibrato rate and vibrato extent over a long period. Another explanation could be the musical context. Performers used the vibrato to indicate the beat positions. This observation warrants further exploration.

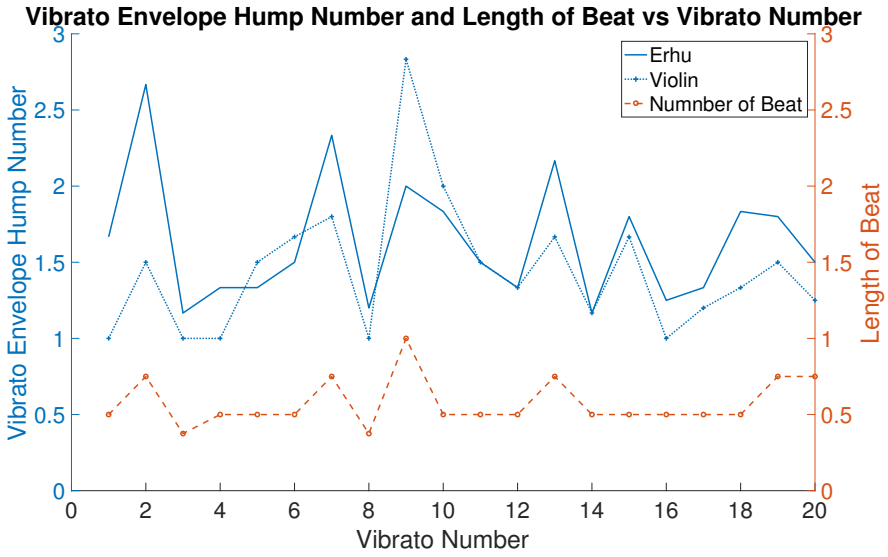


Figure 6.8: Vibrato envelope hump number and number of beats across vibratos.

6.1.4 Remarks

In this study, we examined the differences in vibrato playing styles on two different instruments, the erhu and the violin, for performances of the same piece of music. Vibrato was characterised in terms of vibrato rate, vibrato extent, vibrato sinusoid similarity, and vibrato envelope hump number.

Although violin performers showed slightly higher vibrato rates and ranges than erhu players, the difference is not significant. Erhu players had significantly larger vibrato extents than violinists. They also employed wider vibrato extent ranges than their violin-playing counterparts. The results reveal that violin players exhibit more variability in vibrato rates than erhu players, whilst erhu performers showed more variability in vibrato extents than violinists. The vibrato shapes of the erhu samples were more similar to that of a sinusoid than those in the violin samples. The analyses suggest that the vibrato shape is an independent (having no relationship to vibrato rates and extents) and intrinsic feature of the vibrato. Erhu and violin playing shared the same vibrato envelope hump number variation pattern; both varied the vibrato extent according to the beats in the note.

The piece chosen for the analyses may be traditional Chinese music, but the violin players' vibrato rates and extents were consistent with those reported in the literature for Western music. This suggests that either vibrato performance styles are more dependent on the musical instrument than on the musical genre. Moreover, performances by the U.S. and Japanese violinists each showed the same vibrato characteristics as those by the Chinese violinists. Thus, the cultural background of the violin players may not exert a large influence on the vibrato style. Overall, the results suggest that the physical form of the instrument and how it is played may be the most dominant factor affecting the differences in vibrato style in erhu and violin playing.

6.2 Vibrato and Portamento Performance Analysis using AVA

In this section, we will show the feasibility and usability of the AVA system for vibrato and portamento analyses on real performances. We analyse two datasets, one based on Beijing opera singing and the other derived from a small dataset of erhu and violin performances.

6.2.1 Beijing Opera Singing

Beijing opera, also known as Peking opera, is the predominant opera genre in China. The compound art form comprises of singing, instrumental playing, and acting. There is now growing interest in the study of Beijing opera from a computational perspective in recent years. Repetto & Serra (2014) created a dataset of sung Beijing opera melodies for computational analysis. Tian et al. (2014) investigated onset detection for Beijing opera percussion instruments using non-negative matrix factorization. Srinivasamurthy et al. (2014) utilized a Hidden Markov Model to transcribe and recognize percussion patterns. Gong et al. (2016) proposed a contour segmentation method for Beijing opera singing with the aim for a computer-aided training. Sundberg et al. (2012) studied acoustically the singing of two Beijing opera roles, Laosheng and Dahualian, and found that the singing sound pressure level and pitch are higher than that of speech; furthermore, the vibrato rate was reported to be around 3.5Hz, which is lower than that generally found in Western classical singing. However, research into the expressivity of the featured singing and instrumentation in Beijing opera is still lacking in the literature.

Oriental opera singing possesses characteristics distinct from Western opera. This study aims to investigate the expressive characteristics of Beijing opera singing focusing on pitch, vibrato and portamento. Like the Sundberg et al. (2012)'s study, we also focus on two major Beijing opera roles, those of Laosheng (老生) and Zhengdan (正旦) (also known as Qingyi (青衣)). Our study involves a larger dataset, 16 performances instead of 7. Sundberg et al. (2012)'s study focused on sound pressure level, pitch, and long-term average spectra, with vibrato being a side discussion, we will consider fundamental frequency distributions and details of vibrato parameters such as rate, extent, and sinusoid similarity and portamento parameters such as slope, duration, interval, normalised inflection time and normalised inflection pitch. The vibrato and portamento parameters are annotated and extracted using the AVA system.

We choose to focus on pitch because it is one of the most important features in computational music analysis; this is also true for analysis of world music.

No.	Role	Duration(s)	# Vibratos	# Portamenti
1	Laosheng	102	49	94
2		184	62	224
3		91	56	159
4		95	49	166
5		127	26	115
6		147	51	106
7		168	47	144
8		61	19	89
9	Zhengdan	159	47	173
10		80	24	49
11		119	42	176
12		50	24	71
13		155	57	212
14		41	12	48
15		180	59	219
16		70	34	87

Table 6.5: Beijing opera singing dataset.

For example, Koduri et al. (2012) examined pitch histograms in Indian Carnatic music. Similar research is lacking for Beijing opera singing. Moreover, Chinese traditional music relies on a system with unique characteristics different from that of other music cultures (Tian et al., 2013).

Vibrato and portamento are two of the most important ornamental performing techniques in Beijing opera (Wichmann, 1989). Investigation into the nature of vibrato and portamento use in Beijing opera singing will thus assist in the understanding of the overall plots and the motifs of the story and the roles.

Dataset

For the current study, we use selected recordings from the Beijing opera singing dataset created by Black et al. (2014). The selected dataset consists of 16 monophonic recordings by six different Chinese opera singers performing well-known phrases from the Beijing opera roles Laosheng(老生) and Zhengdan(正旦). Table 6.5 shows the numbers of selected vibratos and portamenti using the AVA system.

Results and Discussions

Figure 6.9 shows the smoothed frequency histogram envelopes for the Laosheng and Zhengdan data, respectively. The sung fundamental frequencies were extracted using the pYIN method. The extracted frequencies were summed into one-cent bins, and the results were smoothed to obtain the histogram envelope.

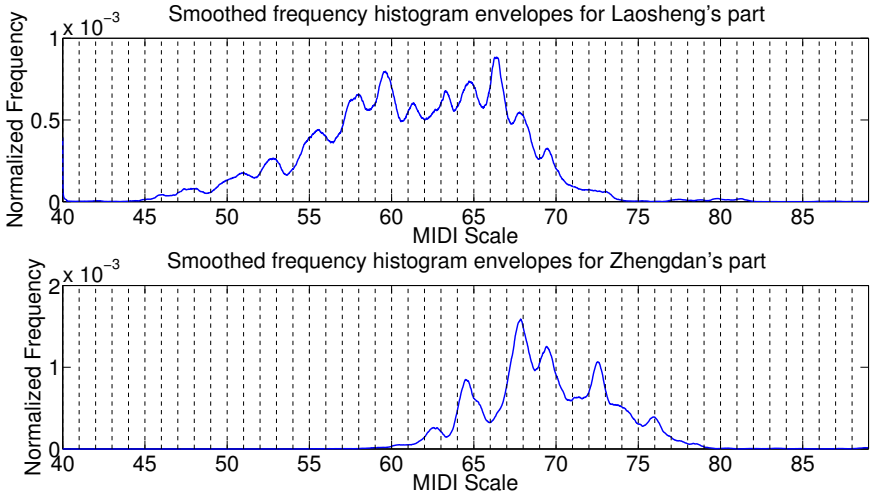


Figure 6.9: Smoothed fundamental frequency histogram envelopes for Laosheng's and Zhengdan's parts.

Because Zhengdan is a female role, we expect the part's pitches to be higher than those of the Laosheng role. The results show that this is indeed the case; however, Laosheng's phrases utilize a wider pitch range than that of Zhengdan, and the highest pitches are in fact higher than most of Zhengdan's pitches. This maybe helpful in role type classification. It is interesting to note that the peaks in the frequency plot show that the Chinese opera melodies also use a semitone scale like that in Western music, although the most prevalent pitches use the traditional Chinese pentatonic scale.

Each recording was uploaded to the AVA system for vibrato and portamento detection and analysis. Detection errors were readily corrected using the editing capabilities of AVA. Figure 6.10 and 6.11 present the resulting histogram envelops of the vibrato and portamento parameter values, each normalised to sum to 1, for the Zhengdan (red) and Laosheng (blue) roles. Translucent lines show the parameter's distributions for individual recordings, and bold lines show the aggregate histogram for each role.

The histograms show the similarities and differences in the underlying probability density functions. Visual inspection shows that the singing of the Zhengdan and Laosheng roles to be most contrastive in the vibrato extents, with peaks at around 0.5 and 0.8 semitones, respectively. A Kolmogorov-Smirnov (KS) test[1] shows that the histogram envelopes of vibrato extent from Laosheng and Zhengdan to be significant different ($p = 2.86 \times 10^{-4}$) at 1% significant level. To be more specific, Zhengdan has the tendency to use smaller vibrato

[1] http://uk.mathworks.com/help/stats/kstest2.html

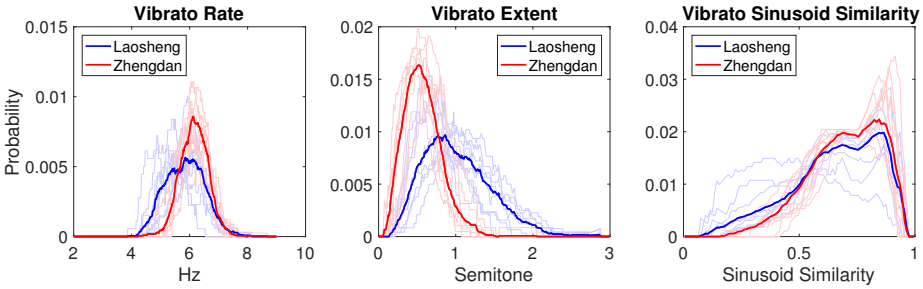


Figure 6.10: Histogram envelopes of vibrato parameters for Beijing opera roles: Laosheng (blue) and Zhengdan (red). Translucent lines show histograms for individual singers; bold line shows aggregated histograms for each role.

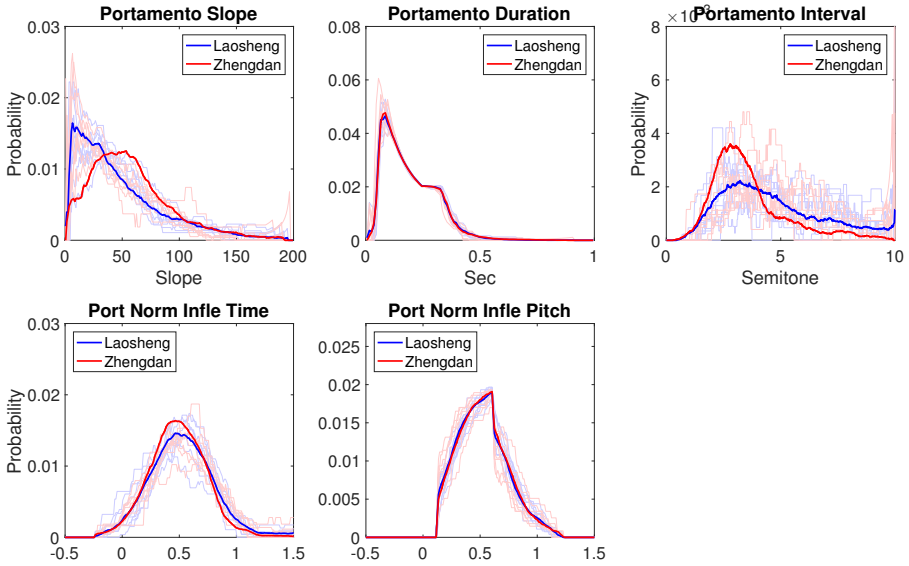


Figure 6.11: Histogram envelopes of portamento parameters for Beijing opera roles: Laosheng (blue) and Zhengdan (red). Translucent lines show histograms for individual singers; bold line shows aggregated histograms for each role.

extents than Laosheng. Regarding the vibrato rate, the difference is not significant ($p = 5.36 \times 10^{-2}$). And for vibrato sinusoid similarity, the difference is also not significant ($p = 2.05 \times 10^{-2}$). Comparing the manually annotated vibrato results from (Yang et al., 2015), the results from the AVA indeed reveal the tendency.

For the portamento, significant differences are found between the singing of the Laosheng and Zhengdan roles for the portamento slope ($p = 1.80 \times 10^{-3}$) and interval ($p = 2.30 \times 10^{-34}$) after testing using the KS test. It reveals that Zhengdan role has a bigger slope value which implies a faster sliding. Differences in duration ($p = .345$), normalised inflection time ($p = .114$) and normalised

inflection pitch ($p = 1.00$) are not significant.

6.2.2 Revisit Vibrato and Portamento Performance Styles on Erhu and Violin using AVA

Here, we revisit vibrato and portamento performance styles on erhu and violin to demonstrate the usability of the AVA system on the analysis of vibrato and portamento performance styles on erhu and violin. Considering the pitch detection methods (i.e. pYIN and YIN) used in AVA, they only work well for monophonic audio. There are only four monophonic performances during the dataset analysed in Section 6.1. Thus, in this section, we will have a small erhu and violin dataset.

Dataset

The study also centres on a well known Chinese piece *Moon Reflected in Second Springs* (《二泉映月》) (Hua, 1958). The study uses four recordings, two for erhu and two more for violin which are same as the dataset listed in Table 3.1 in Section 3.3.1. Table 6.6 lists the details of the test set, which comprises of a total of 23.6 minutes of music; with the help of AVA, 556 vibratos[2] and 527 portamenti were found, verified, and analysed.

No	Ins.	Performer	Durations(s)	# Vibratos	# Portamenti
1	Erhu	Jiangqin Huang[a]	445.83	164	186
2		Guotong Wang[b]	387.53	157	169
3	Violin	Jiang Yang[c]	254.54	131	91
4		Laurel S. Pardue[c]	325.50	104	81

Table 6.6: *Moon Reflected in Second Springs* dataset. *a*: (Huang, 2006), *b*: (Wang, 2009) and *c*: Recorded by the performer.

Results and Discussions

The histograms of the vibrato parameters are summarised in Figure 6.12. Again, we use the KS test to assess the difference in the histograms between violin and erhu. As with the case for the Beijing opera roles, the most significant difference between the instruments is found in the vibrato extent ($p = 2.70 \times 10^{-3}$), with the vibrato extent for the erhu about twice that for violin (half semitone vs. quarter semitone). There is no significant difference found between erhu and violin for vibrato rate ($p = .352$) and sinusoid similarity ($p = .261$). The findings supported our first experiment's conclusions in Section 6.1 that the most

[2]The number of vibratos here has a little difference from the number of vibratos in Table 3.1 used for vibrato detection evaluation. The difference may stem from the tool for annotation. The AVA is used here, but in Section 3.3.1, the Tony is employed.

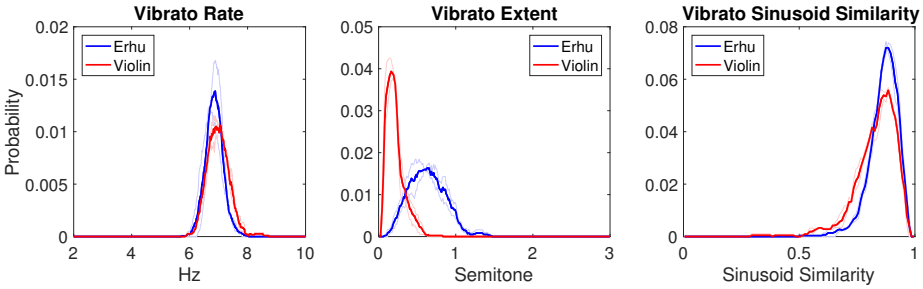


Figure 6.12: Histogram envelopes of vibrato parameters for two instruments: erhu (blue) and violin (red). Translucent lines show histograms for individual players; bold line shows aggregated histograms for each instrument.

significant difference between erhu and violin is vibrato extent. The insignificant differences in vibrato rate (violin players tend to have higher vibrato rate) and sinusoid similarity (violin players tend to have a lower sinusoid similarity) are also observed here. This in turn supports the AVA's feasibility and usability on the performance analysis.

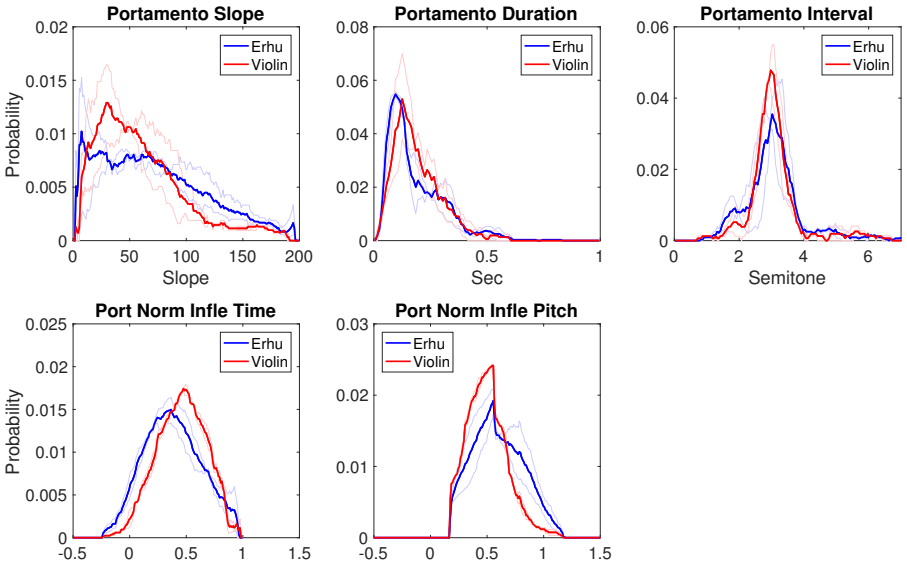


Figure 6.13: Histogram envelopes of portamento parameters for two instruments: erhu (blue) and violin (red). Translucent lines show histograms for individual players; bold line shows aggregated histograms for each instrument.

The histograms of the portamento parameters are presented by Figure 6.13. Regarding portamento, the portamento interval histogram has a distinct peak at around three semitones for both violin and erhu, showing that notes separated by this gap is more frequently joined by portamenti. The difference between

the histograms is highly insignificant ($p = .363$). The most significant difference between violin and erhu portamenti histograms is observed for the slope ($p = 1.51 \times 10^{-4}$). The duration ($p = .344$), normalised inflection time ($p = .256$) and normalised inflection pitch ($p = .382$) don't show significant results.

6.3 Conclusions

In this chapter, we have shown two case studies on vibrato and portamento analysis based on music recordings of actual performances.

The first case study, which provides the main motivations for this thesis, examined the differences between vibrato playing styles as performed on the erhu vs. the violin for the same piece of music. The main and significant difference was found in the vibrato extent. The erhu performers employed significantly larger extents than the violin players. They also used wider vibrato extent ranges than their violin-playing counterparts. The difference between erhu and violin vibrato rates were not significant; the violin recordings had slightly higher vibrato rates and ranges than erhu recordings. The violin players tended to have more variability in vibrato rates, while erhu performers showed more variability in vibrato extents. The analyses suggest that vibrato shape is an independent and intrinsic feature of the vibrato.

The second case study analyses vibrato and portamento performing styles on a Beijing opera singing dataset and a erhu and violin dataset using the AVA system. In Beijing opera singing, the Laosheng and Zhengdan roles had significantly different vibrato extents: performers of the Laosheng role tended to use larger vibrato extents than those for the Zhengdan role. Another significant difference lies in the portamento slope. Singing of the Zhengdan role had faster slides than that for the Laosheng role. A feasibility and usability test applying the AVA system to the Chinese piece, *Moon Reflected in Second Springs*, performed on violin and erhu was also presented. The results further confirmed the first case study's findings about vibrato performance style. It also presents the portamento performance difference between erhu and violin showing that violin players employed steeper portamento slopes than erhu players.

Chapter 7

Conclusions

We demonstrated computational and mathematical models for accurate representation of vibratos and portamenti. To conclude the thesis, we review the primary contributions of each chapter. Then, we present the research challenges and suggest some future directions to further develop this work. Third, we summarise the primary contributions of this thesis. Finally, we present some closing remarks.

7.1 Summary

In Chapter 3, we presented a novel frame-wise vibrato detection and estimation method using the Filter Diagonalisation Method. A thorough evaluation shows that our proposed method is superior to alternative state-of-the-art methods using frame-level and note-level metrics. The FDM is able to extract sinusoidal frequency and amplitude information from a very short time signal, making it possible to determine vibrato frequency and pinpoint vibrato boundaries over a short time span. The byproduct of the FDM algorithm, i.e. the vibrato rate and extent, helps to reduce the computational overhead required to obtain the vibrato parameters. We have created a new monophonic dataset consisting of erhu and violin performances of an entire piece of music, *Moon Reflected in Second Springs*, for vibrato detection and vibrato parameter estimation. The long sequences allow for training and test data to both be excerpted not only from performances by the same player, but from the same performance. We further evaluated the vibrato parameter estimation capabilities of the FDM method. The accuracy of vibrato rate estimation is above 92.5%, and that of the vibrato extent estimation is on the order of 85% for both decision methods with FDM. Besides the common parameters to describe vibrato, i.e. rate and extent, we created a new parameter to describe the shape of a vibrato. This

parameters is sinusoid similarity to describe how similar of the vibrato pitch contour comparing to a reference sinusoid.

In Chapter 4, we proposed a computational and mathematical model of portamento employing the Logistic Model. This model was shown to exhibit better performance as compared to the Polynomial Model, Gaussian Model, and Fourier Series Model. This model has parameters that could convey the musically meaningful information. A case study was used to show the feasibility of the Logistic Model in expressive music analyses. Besides the continuous note transition (portamento), this model is able to be expanded to model discrete note transition. In this chapter, we also make the first effort towards portamento detection. A Hidden Markov Model with different observation likelihood distributions (Gaussian Model and Gaussian Mixture Model) has been employed. The results show that the use of HMM+GMM has better performance than HMM+Gaussian. However, the returns of increasing Gaussian mixture numbers quickly diminish, and so the performance does not significantly improve as this value increases. The delta pitch and energy have been used for detecting portamento. The combination of delta pitch and energy does not improve the portamento detection performance significantly, as compared to the use of the delta pitch alone. In fact, in some cases it shows worse results.

In Chapter 5, with the aim of making available a software tool for interactive analysis of vibrato and portamento in performance and musicological research, we integrated the vibrato and portamento detection and estimation modules to create AVA, an interactive visual and quantitative analysis tool for vibrato and portamento. This toolbox provides the user with the functionality to correct the errors introduced from the automatic detection process. A playback function allows the user to hear each detected vibrato/portamento in order to inspect and improve detection results.

In Chapter 6, we first presented an case study investigating vibrato characteristics between erhu and violin for the same piece of music. Violin performers had slightly higher vibrato rates and ranges than erhu players, but the difference is not significant. Erhu performers had significantly larger vibrato extents than violin performers. The vibrato shape of the erhu samples was more similar to that of a sinusoid than the violin samples. Then we used AVA on erhu & violin datasets, as well as the Beijing opera singing dataset for vibrato and portamento analysis in order to show the feasibility and usability of the AVA system. The analysis on erhu and violin dataset confirms the discoveries of the first case study. In Beijing opera singing, the Laosheng and Zhengdan have significant differences in vibrato extent, where Laosheng tended to use larger vibrato extents than Zhengdan. Another significant difference is in portamento slope; this implies that Zhengdan has faster slides than Laosheng.

7.2 Challenges and Future work

Regarding current status of this work, we discuss existing challenges and possible future directions of this research.

Vibrato detection: We proposed the frame-wise FDM-based vibrato detection method. The FDM is able to extract sinusoid frequency and amplitude information for a very short time signal, making it possible to determine vibrato frequency and pinpoint vibrato boundaries over a short time span. Two decision-making methods have been used to determine vibrato existence. One is the Decision Tree, which uses the explicitly defined vibrato rate and extent to decide vibrato existence for each time frame. The other is the Bayes' Rule, which calculates the probability of vibrato for each frame. The present method outperformed state-of-the-art methods, especially at the note-level, but there is still room for improvement. One possible direction may be to employ supervised machine learning methods such as the support vector machine, a non-probabilistic binary linear classifier, which could handle data points outside the training set; the binary classification also makes it suitable for vibrato detection. Other potential supervised learning methods include the neural network and deep learning, which currently provides the best solutions for many problems in image recognition, speech recognition and natural language processing.

Filter Diagonalisation Method: In using the FDM method, we only extracted the sinusoid with the largest amplitude from the outputs for vibrato detection. The FDM has potential applications in many signal processing domains requiring the extraction of sinusoid parameters from a short time signal. For example, it may be worth exploring ways to adapt the FDM for pitch detection. A preliminary test has shown that it is important to select appropriate values for the number of sinusoids and the frequency window. Large frequency windows may result in too many sinusoids to fit, which degrades the FDM performance, while small frequency windows may not cover the desired sinusoid.

Portamento detection: We explored the Hidden Markov Model based method to detect portamento. Portamento detection deserves more research attention as it is the first step towards the automatic portamento analysis. With the assumption that portamento should be heard by people, we have tried to use energy together with delta pitch to detect portamento. However, the performance was not improved significantly. A thorough investigation into energy or loudness-based detection mechanisms could be a possible direction. Considering that the portamento constitutes a sweep

of the spectrum, it is also worth testing the use of spectral features such as the spectral flux, flatness and centroid. Another direction could employ the the slope parameter in the Logistic Model for portamento detection.

Vibrato and portamento modelling: We used the rate, extent and sinusoid similarity to model each vibrato. These parameters fail to describe the time-varying aspects of the vibrato. For portamento modelling, the Logistic Model was shown to be superior to other alternative methods. Further research could consider its performance on specific portamento types (see Chapter 2). In this thesis, we modelled vibrato and portamento separately. However, vibrato and portamento use may be related (Desain & Honing, 1995)—a reason why AVA requires vibrato removal before portamento detection—and future work could investigate the dependencies between vibrato and portamento, such as scenarios in which a vibrato is followed by a portamento, or vice versa..

AVA system: There are many possible improvements that could be made for this software toolbox. One possible direction is to integrate more music analysis modules, and to expand it into a computer-aided music tutoring system, for example in intonation analysis or rhythmic analysis. Recently, computer music education has attracted much attention (Salgian & Vickerman, 2016; Hancock, 2016; Johnson et al., 2016). AVA is able to provide visual feedback and quantitative analysis of students' performances, allowing students to inspect their expressive features and adapt accordingly. At present, AVA is restricted analyses only monophonic audio; future extensions may incorporate melody detection (for example, using MELODIA (Salamon & Gómez, 2012)) or polyphonic pitch detection modules. Another way to enhance AVA is to make it real-time.

Expression synthesis: Another new direction of this work is expression synthesis. Saitou et al. (2005) presented a pitch-based expression synthesis for singing-voice. They examined the vibrato, portamento (overshoot and preparation in their article) and subtle fluctuations in pitch curve to create a natural singing voice. The psychoacoustical experiments showed the feasibility of the expressions synthesis on pitch curve. By analysing the expressions in different performers, genres and cultures, AVA could be used as the first step for expression synthesis. Further work are investigations of expression synthesis, and examinations of expression transfers. To name a few further questions, "is it possible to add expressions into a MIDI score in a natural way?", "can erhu's expressions be transferred to violin?" or "is it possible to transfer Chinese music expressions into Western instruments?"

7.3 Summary of Key Contributions

Novel methods and materials

- A novel frame-wise vibrato detection method using the Filter Diagonalisation Method has been proposed. It is superior than current state-of-the-art methods (Chapter 3).
- Created a new dataset for vibrato and portamento detection, and vibrato parameter estimation.
- A new portamento modelling method using the Logistic Model is proposed (Chapter 4).
- Explored the Hidden Markov Model for portamento detection (Chapter 4).
- Created an interactive visual and quantitative analysis system for vibrato and portamento (Chapter 5).

New discoveries

- Vibratos in string and woodwind instruments are easier to be detected than those in brass instruments and voice. (Chapter 3).
- Erhu players have larger vibrato extents than violin players (Chapter 6).
- Even when playing Chinese music, violin players' vibrato rates and extents were consistent with those reported in the literature for Western music (Chapter 6).
- In Beijing opera singing, Laosheng has larger vibrato extents than Zhengdan; while Zhengdan has larger portamento slopes than Laosheng (Chapter 6).

7.4 Closing Remarks

Music expressivity modelling, or expressive music performance modelling, is an emerging interdisciplinary area that cuts across musicology, signal processing, and machine learning. The modelling of expressivity introduced by performers not only reveals how humans decode and manipulate music compositions, but also delves into how people interpret and communicate ideas and emotions. Focussing our attention on vibrato and portamento, we have showed that it is possible and feasible to computationally model these two expressive devices.

Regarding data collection, as vibrato has been integrated into music sound production since the 1950s, vibrato data collection has proved to be much easier than that of portamento data. The decline of portamento further made it more difficult to collect data from commercial recordings, especially of Western classical music.

We found vibrato characteristics to be highly different across violin and erhu performances of the same piece of music. These results are surprising especially considering the fact that modern erhu players are more likely to use the *rolling vibrato* which traces its origins to violin vibrato technique. The differences may result from the physical form of the instrument; for example, the violin's fingerboard may constrain violinists' hand movements. We observed that violin players' vibrato characteristics are consistent with those reported in the literature. This suggests that vibrato performance styles may be more dependent on the instrument than the musical genre. Similar to vibrato, we found differences in the ways string players and singers deploy portamenti.

We also instigated a step change in automatic expressivity analysis by AVA, an interactive visual and quantitative tool for vibrato and portamento analysis. AVA's potential applications include music performance analysis, music education, and music expression synthesis.

In this work, we only explored two expressive devices, which is but a small corner in the wider expressivity space. We hope this work will inspire further research on music expression that can help build more accurate, robust, and versatile expressivity modelling systems.

Appendix A

Coler-Roebel Dataset

No	Ins.	Name	Dur.(s)	No	Ins.	Name	Dur.(s)
1		Vioin-1	4.87	15		Alto01	6.38
2		Vioin-2	4.77	16		AltoBalladinAmin01	12.50
3		Vioin-3	9.79	17		alto-sax_short	10.15
4	Violin	Vioin-4	8.63	18	Woodwind	ClarinetOneShot07	12.81
5		Vioin-5	7.61	19		ClarinetSolo12	3.23
6		Vioin-6	7.52	20		Flute_63-03	5.05
7		Vioin-7	4.03	21		Flute_64-09	3.70
8		female02	7.73	22		TromboneSolo14	12.73
9		FemaleVocal16	4.86	23		Trumpet_1	4.67
10		FemaleVocal21	2.77	24		Trumpet10-01	3.38
11	Voice	Fior	11.36	25	Brass	Trumpet10-06	3.46
12		Maria	7.23	26		Trumpet55-01	3.93
13		ProgressiveVocals	4.80	27		TrumpetMelody01	20.47
14		yesterday_short1	11.56	28		TrumpetSolo12	7.39

Table A.1: Coler and Roebel's dataset from (von Coler & Roebel, 2011).

Appendix B

Derivations for Inflection Point of the Logistic Function

During the Section 4.1.1, the time of the inflection point of the Logistic function Eq. (4.1) is defined by finding the extremum of the first derivative. Considering the Logistic function is monotonically increasing or decreasing, it can only have one extremum value on the first order derivative which is implied by the zero point on the second order derivative. The followings are the derivations of the time of the inflection point.

According to Logistic function Eq. (4.1), the first derivative is

$$P'(t) = \frac{\frac{(U-L)GA}{B} e^{-G(t-M)} (1 + Ae^{-G(t-M)})^{\frac{1-B}{B}}}{(1 + Ae^{-G(t-M)})^{\frac{2}{B}}}. \tag{B.1}$$

If we let $\mathcal{X} = 1 + Ae^{-G(t-M)}$, the second derivative is

$$P''(t) = \frac{\frac{(U-L)GA}{B} \left[\frac{(B-1)GA}{B} \mathcal{X}^{\frac{1-2B}{B}} e^{-2G(t-M)} - Ge^{-G(t-M)} \mathcal{X}^{\frac{1-B}{B}} \right]}{\mathcal{X}^{\frac{4}{B}}} - \frac{\frac{2(U-L)G^2A^2}{B^2} e^{-2G(t-M)} \mathcal{X}^{\frac{1-2B}{B}}}{\mathcal{X}^{\frac{4}{B}}}. \tag{B.2}$$

Let $P''(t) = 0$, the denominator is eliminated, then we have

$$\frac{(B-1)A}{B} \mathcal{X}^{\frac{1-2B}{B}} e^{-2G(t-M)} - e^{-G(t-M)} \mathcal{X}^{\frac{1-B}{B}} = -\frac{2A}{B} e^{-2G(t-M)} \mathcal{X}^{\frac{1-2B}{B}}. \tag{B.3}$$

Dividing $\mathcal{X}^{\frac{1-B}{B}}$ from both sides,

$$\frac{(B-1)A}{B} \mathcal{X}^{-1} e^{-2G(t-M)} - e^{-G(t-M)} = -\frac{2A}{B} e^{-2G(t-M)} \mathcal{X}^{-1}. \tag{B.4}$$

Multiplying $\mathcal{X}$ on both sides, then we have

$$\frac{(B-1)A}{B}e^{-2G(t-M)} - e^{-G(t-M)}\mathcal{X} = -\frac{2A}{B}e^{-2G(t-M)}, \tag{B.5}$$

and multiplying $e^{G(t-M)}$ on both sides, then

$$\frac{(B-1)A}{B}e^{-G(t-M)} - \mathcal{X} = -\frac{2A}{B}e^{-G(t-M)}, \tag{B.6}$$

thus,

$$\frac{(B+1)A}{B}e^{-G(t-M)} = \mathcal{X}. \tag{B.7}$$

Substituting $\mathcal{X}$ by $1 + Ae^{-G(t-M)}$,

$$\frac{(B+1)A}{B}e^{-G(t-M)} = 1 + Ae^{-G(t-M)}. \tag{B.8}$$

Finally, the time of the inflection point is

$$t_R = -\frac{1}{G}\ln\left(\frac{B}{A}\right) + M. \tag{B.9}$$

Bibliography

Amir, N., Michaeli, O., & Amir, O. (2006). Acoustic and perceptual assessment of vibrato quality of singing students. *Biomedical Signal Processing and Control*, *1*(2), 144–150.

Ashley, R. (2014). *Expressiveness in music performance: Empirical approaches across styles and cultures*, chapter Expressiveness in Fund, (pp. 154–169). Oxford University Press.

Auer, L. (1921). *Violin playing as I teach it*. Frederick A. stokes Company.

Baillot, P. (1834). *L'art du violon*.

Barbancho, I., de la Bandera, C., Barbancho, A. M., & Tardon, L. J. (2009). Transcription and expressiveness detection system for violin music. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*.

Bauer, W. R. (2014). *Expressiveness in music performance: Empirical approaches across styles and cultures*, chapter Expressiveness in Jazz Performance: Prosody and Rhythm, (pp. 133–153). Oxford University Press.

Benetos, E. (2012). *Automatic transcription of polyphonic music exploiting temporal evolution*. PhD thesis, Queen Mary University of London.

Benetos, E. & Dixon, S. (2010). Multiple-f0 estimation of piano sounds exploiting spectral structure and temporal evolution. In *SAPA@ INTERSPEECH*, (pp. 13–18).

Bishop, C. (2007). *Pattern Recognition & Machine Learning*. Springer.

Black, D. A. A., Ma, L., & Tian, M. (2014). Automatic identification of emotional cues in Chinese opera singing. In *Proc. of the 13th International Conference on Music Perception and Cognition and the 5th Conference for the Asian-Pacific Society for Cognitive Sciences of Music (ICMPC 13-APSC0M 5)*.

Bloothooft, G. & Pabon, P. (2004). Qualities of a voice emeritus. *LOT Occasional Series*, *2*, 17–26.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proc. of the institute of phonetic sciences*, volume 17, (pp. 97–110).

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot international*, *5*(9/10), 341–345.

Bretos, J. & Sundberg, J. (2003). Measurements of vibrato Parameters in long sustained crescendo notes as sung by ten sopranos. *Journal of Voice*, *17*(3), 343–352.

Brown, C. (1988). Bowing styles, vibrato and portamento in nineteenth-century violin playing. *Journal of the Royal Musical Association*, *113*(1), 97–128.

Brown, J. C. (1991). Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, *89*(1), 425–434.

Canazza, S., De Poli, G., Drioli, C., Roda, A., & Vidolin, A. (2004). Modeling and control of expressiveness in music performance. *Proc. of the IEEE*, *92*(4), 686–701.

Cannam, C., Landone, C., & Sandler, M. (2010). Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proc. of the 18th ACM International Conference on Multimedia*, (pp. 1467–1468).

Chen, J. (2002). *Nonlinear methods for high resolution spectral analysis and their applications in nuclear magnetic resonance experiments.* PhD thesis, UNIVERSITY OF CALIFORNIA, IRVINE.

Childers, D. G., Skinner, D. P., & Kemerait, R. C. (1977). The cepstrum: A guide to processing. *Proceedings of the IEEE*, *65*(10), 1428–1443.

de Cheveigné, A. (1998). Cancellation model of pitch perception. *The Journal of the Acoustical Society of America*, *103*(3), 1261–1271.

de Cheveigné, A. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, chapter Multiple F0 estimation, (pp. 45–79). IEEE Press/Wiley.

de Cheveigné, A. & Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917–1930.

Desain, P. & Honing, H. (1995). Towards algorithmic descriptions of continuous modulations of musical parameters. In *Proc. of the International Computer Music Conference*.

Desain, P. & Honing, H. (1996). Modeling continuous aspects of music performance: Vibrato and portamento. In *Proc.of the International Music Perception and Cognition Conference*.

Desain, P., Honing, H., Aarts, R., & Timmers, R. (1999). Rhythmic aspects of vibrato. *Rhythm Perception and Production*, 203–216.

Devaney, J., Mandel, M., & Ichi (2012). A study of intonation in three-part singing using the automatic music performance analysis and comparison toolkit (ampact). In *Proc. of the International Society for Music Information Retrieval Conference*.

Devaney, J. C., Mandel, M. I., & Fujinaga, I. (2011). Characterizing singing voice fundamental frequency trajectories. In *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (pp. 73–76).

Diaz, J. A. & Rothman, H. B. (2003). Acoustical comparison between samples of good and poor vibrato in singers. *Journal of Voice*, *17*(2), 179–184.

Dibben, N. (2014). *Expressiveness in music performance: Empirical approaches across styles and cultures*, chapter Understanding Performance Expression in Poluar Music Recordings, (pp. 117–132). Oxford University Press.

Dixon, S. (2000). On the computer recognition of solo piano music. In *Proc. of Australasian computer music conference*.

Dixon, S., Goebl, W., & Widmer, G. (2002). The performance worm: Real time visualisation of expression based on langner's tempo-loudness animation. In *Proc. of the International Computer Music Conference*.

Dixon, S., Goebl, W., & Widmer, G. (2005). The" air worm": An interface for real-time manipulation of expressive music performance. In *Proc. of the International Computer Music Conference*, (pp. 614–617).

Doval, B. & Rodet, X. (1993). Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, (pp. 221–224).

Driedger, J., Balke, S., Ewert, S., & Müller, M. (2016). Template-based vibrato analysis of music signals. In *Proc. of the 17th International Society for Music Information Retrieval Conference*.

Dromey, C., Carter, N., & Hopkin, A. (2003). Vibrato rate adjustment. *Journal of Voice*, *17*(2), 168–178.

d'Alessandro, C. & Castellengo, M. (1994). The pitch of short-duration vibrato tones. *The Journal of the Acoustical Society of America*, *95*(3), 1617–1630.

Fabian, D., Timmers, R., & Schubert, E. (2014). *Expressiveness in music performance: Empirical approaches across styles and cultures*, chapter Introduction, (pp. xxi–xxx). Oxford.

Ferreira, A., Abreu, F., & Sinha, D. (2008). Stereo acc real-time audio communication. In *Proc. of the Audio Engineering Society Convention 125*.

Ferreira, A. J. S. (1995). Tonality detection in perceptual coding of audio. In *Proc. of the Audio Engineering Society Convention 98*.

Fletcher, N. H. (2001). Vibrato in music. *Acoustics Australia*, *29*(3), 97–102.

Friberg, A., Colombo, V., Frydén, L., & Sundberg, J. (2000). Generating musical performances with director musices. *Computer Music Journal*, *24*(3), 23–29.

Friberg, A., Schoonderwaldt, E., & Juslin, P. N. (2007). Cuex: An algorithm for automatic extraction of expressive tone parameters in music performance from acoustic signals. *Acta acustica united with acustica*, *93*(3), 411–420.

Fritz, C., Woodhouse, J., Cheng, F. P.-H., Cross, I., Blackwell, A. F., & Moore, B. C. J. (2010). Perceptual studies of violin body damping and vibrato. *The Journal of the Acoustical Society of America*, *127*(1), 513–524.

Gable, F. K. (1992). Some observations concerning baroque and modern vibrato. *Performance Practice Review*, *5*(1).

García, M. (1856). *Garcia's New Treatise of the Art of Singing* (Revised version ed.). BOSTON: OLIVER DITSON COMPANY.

Geringer, J. M. & Allen, M. L. (2004). An analysis of vibrato among high school and university violin and cello students. *Journal of Research in Music Education*, *52*(2), 167–178.

Gómez, E. & Bonada, J. (2013). Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing. *Computer Music Journal*, *37*(2), 73–90.

Gong, R., Yang, Y., & Serra, X. (2016). Pitch contour segmentation for computer-aided jingju singing training. In *Proc. of the 13th Sound and Music Computing Conference*.

Gough, C. E. (2005). Measurement, modelling and synthesis of violin vibrato sounds. *Acta Acustica united with Acustica*, *91*(2), 229–240.

Gu, H.-Y. & Lin, Z.-F. (2008). Mandarin singing voice synthesis using ann vibrato parameter models. In *Proc. of the IEEE International Conference on Machine Learning and Cybernetics*, volume 6, (pp. 3288–3293).

Hancock, O. (2016). a band is born: a digital learning game for max/msp. In *Proc. of the 42nd International Computer Music Conference*.

Herman, R. & Montroll, E. W. (1972). A manner of characterizing the development of countries. In *Proc. of the National Academy of Sciences*, volume 69, (pp. 3019–3023).

Herrera, P. & Bonada, J. (1998). Vibrato extraction and parameterization in the spectral modeling synthesis framework. In *Proc. of the Digital Audio Effects Workshop*, volume 99.

Hodgson, P. (1916). 8 ways to vary your vibrato. *The Strad*, *http://www.thestrad.com/8-ways-vary-vibrato/*. Accessed in April 2016.

Howes, P., Callaghan, J., Davis, P., Kenny, D., & Thorpe, W. (2004). The relationship between measured vibrato characteristics and perception in western operatic singing. *Journal of Voice*, *18*(2), 216–230.

Hu, H., Van, Q. N., Mandelshtam, V. A., & Shaka, A. J. (1998). Reference deconvolution, phase correction, and line listing of nmr spectra by the 1d filter diagonalization method. *Journal of Magnetic resonance*, *134*(1), 76–87.

Hua, Y. (1958). *Erquanyingyue*《二泉映月》 (Violin ed.). Zhiruo Ding and Zhanhao He. Musical Score.

Huang, J. (2006). The Moon Reflected on the Second Spring, on The Ditty of the South of the Jiangsu. CD. ISBN: 9787885180706.

Järveläinen, H. (2002). Perception-based control of vibrato parameters in string instrument synthesis. In *Proc. of the International Computer Music Conference*.

Joachim, J. & Moser, A. (1905). Violinschule. *Trans. Alfred Moffat. Berlin: Simrock*, *3*.

Joder, C., Essid, S., & Richard, G. (2010). An improved hierarchical approach for music-to-symbolic score alignment. In *Proc. of the International Society for Music Information Retrieval Conference*.

Johnson, D., Dufour, I., Damian, D., & Tzanetakis, G. (2016). Detecting pianist hand posture mistakes for virtual piano tutoring. In *Proc. of the 42nd International Computer Music Conference*.

Jure, L., Lopez, E., Rocamora, M., Cancela, P., Sponton, H., & Irigaray, I. (2012). Pitch content visualization tools for music performance analysis. In *Proc. of the 13th International Society for Music Information Retrieval Conference*.

Juslin, P. N. & Sloboda, J. A. (2001). *Music and emotion: Theory and research.* Oxford University Press.

Keiler, F. & Marchand, S. (2002). Survey on extraction of sinusoids in stationary sounds. In *Proc. of the Digital Audio Effects Conference*, (pp. 51–58).

Kendall, R. A. & Carterette, E. C. (1990). The communication of musical expression. *Music Perception: An Interdisciplinary Journal*, *8*(2), 129–163.

Klapuri, A. (2004). *Signal processing methods for the automatic transcription of music*. PhD thesis, Tampere University of Technology.

Klapuri, A. & Davy, M. (2007). *Signal Processing Methods for Music Transcription*. Springer Science & Business Media.

Klapuri, A. P. (2003). Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing*, *11*(6), 804–816.

Klein, H. (1991). *Herman Klein and the Gramophone*, (pp. 347–8). Oregon, Portland.

Koduri, G. K., Serrà, J., & Serra, X. (2012). Characterization of intonation in carnatic music by parametrizing pitch histograms. In *Proc. of the International Society for Music Information Retrieval Conference*.

Köküer, M., Jančovič, P., Ali-MacLachlan, I., & Athwal, C. (2014). Automated detection of single-and multi-note ornaments in irish traditional flute playing. In *Proc. of the 15th International Conference on Music Information Retrieval Conference*.

Krishnaswamy, A. (2003). Pitch measurements versus perception of south indian classical music. In *Proc. of the Stockholm Music Acoustics Conference*.

Lahat, M., Niederjohn, R. J., & Krubsack, D. A. (1987). A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech. *IEEE transactions on acoustics, speech, and signal processing*, *35*(6), 741–750.

Lee, H. (2006). Violin portamento: An analysis of its use by master violinists in selected nineteenth-century concerti. In *Proc. of the 9th International Conference on Music Perception and Cognition*.

Leech-Wilkinson, D. (2006). Portamento and musical meaning. *Journal of Musicological Research*, *25*(3–4), 233–261.

Leech-Wilkinson, D. (2009). *The Changing Sound of Music: Approaches to Studying Recorded Musical Performances*, chapter 8. Expressive gestures. London: CHARM.

Lerdahl, F. & Jackendoff, R. (1983). *A generative theory of tonal music*. MIT Press.

Li, P.-C., Su, L., Yang, Y.-H., & Su, A. W. Y. (2015). Analysis of expressive musical terms in violin using score-informed and expression-based audio features. In *Proc. of the 16th International Society for Music Information Retrieval Conference*.

Liebman, E., Ornoy, E., & Chor, B. (2012). A phylogenetic approach to music performance analysis. *Journal of New Music Research*, *41*(2), 95–222.

Ling, G. (2007). Erhu huayin dui yinse de yingxiang 二胡滑音对音色的影响 ((the portamento influence on timbre in erhu). *Guangxi yishu xueyuan xuebao* 广西艺术学院学报《艺术探索》, *21*(6).

Lippus, P. & Ross, J. (2014). *Expressiveness in music performance: Empirical approaches across styles and cultures*, chapter Temporal Variation in Singing as Interplay between Speech and Music in Estonian Songs, (pp. 185–200). Oxford University Press.

Liu, J. (2013). Properties of violin glides in the performance of cadential and noncadential sequences in solo works by bach. In *Proc. of Meetings on Acoustics*, volume 19. Acoustical Society of America.

Liu, T. (1930). *Bingzhongyin*《病中吟》. Chinese erhu composition.

MacLeod, R. B. (2008). Influences of dynamic level and pitch register on the vibrato rate and widths of violin and violia players. *Journal of Research in Music Education*, *56*(1), 43–54.

Maestre, E. & Gómez, E. (2005). Automatic characterization of dynamics and articulation of expressive monophonic recordings. In *Proc. of the 118th Audio Engineering Society Convention*.

Maher, R. C. (2008). Control of synthesized vibrato during portamento musical pitch transitions. *Journal of the Audio Engineering Society*, *56*(1/2), 18–27.

Mandelshtam, V. A. (2001). Fdm: the filter diagonalization method for data processing in nmr experiments. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *38*(2), 159–196.

Mandelshtam, V. A. & Taylor, H. S. (1997). Harmonic inversion of time signals and its applications. *The Journal of chemical physics*, *107*(17), 6756–6769.

Marandola, F. (2014). *Expressiveness in music performance: Empirical approaches across styles and cultures*, chapter Expressiveness in the Performance of Bedzan Pygmies' Vocal Polyphonies: When the same is Never the Same, (pp. 201–217). Oxford University Press.

Marchetti, C. & Nakicenovic, N. (1979). The dynamics of energy systems and the logistic substitution model. Technical report, PRE-24360.

Martini, B. R., Mandelshtam, V. A., Morris, G. A., Colbourne, A. A., & Nilsson, M. (2013). Filter diagonalization method for processing pfg nmr data. *Journal of Magnetic Resonance*, *234*, 125–134.

MATLAB (2013). Version: R2013b.

Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., Bello, J., & Dixon, S. (2015). Computer-aided melody note transcription using the Tony software: Accuracy and efficiency. In *Proc. of the 1st International Conference on Technologies for Music Notation and Representation*, (pp. 23–30).

Mauch, M. & Dixon, S. (2014). PYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (pp. 659–663).

Mazzola, G. & Göller, S. (2002). Performance and interpretation. *Journal of New Music Research*, *31*(3), 221–232.

Meddis, R. & O'Mard, L. (1997). A unitary model of pitch perception. *The Journal of the Acoustical Society of America*, *102*(3), 1811–1820.

Melba, N. (1926). *Melba method*. Chappell Co. LTD.

Mellody, M. & Wakefield, G. H. (2000). The time-frequency characteristics of violin vibrato: Modal distribution analysis and synthesis. *The Journal of the Acoustical Society of America*, *107*(1), 598–611.

Meron, Y. & Hirose, K. (2000). Synthesis of vibrato singing. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2.

Milsom, D. (2003). *Theory and practice in late nineteenth-century violin performance: an examination of style in performance, 1850–1900*. Ashgate Publishing.

Mitchell, H. F. & Kenny, D. T. (2010). Change in vibrato rate and extent during tertiary training in classical singing students. *Journal of Voice*, *24*(4), 427–434.

Moelants, D. (2004). The timing of tremolo, trills and vibrato by string instrument players. In *Proc. of the 8th International Conference on Music Perception and Cognition*.

Molina, E., Barbancho, A. M., Tardón, L. J., & Barbancho, I. (2014). Evaluation framework for automatic singing transcription. In *Proc. of the 15th International Society for Music Information Retrieval Conference*, (pp. 567–572).

Neuhauser, D. (1990). Bound state eigenfunctions from wave packets: Time energy resolution. *The Journal of Chemical Physics*, *93*(4), 2611–2616.

Noll, A. M. (1967). Cepstrum pitch determination. *The journal of the acoustical society of America*, *41*(2), 293–309.

Nwe, T. L. & Li, H. (2007). Exploring vibrato-motivated acoustic features for singer identification. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(2), 519–530.

Ott, R. L. & Longnecker, M. (2010). *An introduction to statistical methods and data analysis* (6 ed.). Brooks/Cole.

Özaslan, T. H. & Arcos, J. L. (2011). Automatic vibrato detection in classical guitar. In *TR-IIIA-2011-05*.

Özaslan, T. H., Serra, X., & Arcos, J. L. (2012). Characterization of embellishments in ney performances of makam music in turkey. In *Proc. of the International Society for Music Information Retrieval Conference*.

Palmer, C. & Hutchins, S. (2006). *What is musical prosody?*, chapter 46, (pp. 245). ELSEVIER.

Palmer, C., Jungers, M. K., & Jusczyk, P. W. (2001). Episodic memory for musical prosody. *Journal of Memory and Language*, *45*(4), 526–545.

Pang, H.-S. & Yoon, D.-H. (2005). Automatic detection of vibrato in monophonic music. *Pattern Recognition*, *38*(7), 1135–1138.

Pardo, B. & Birmingham, W. (2005). Modeling form for on-line following of musical performances. In *Proc. Of The National Conference on Artificial Intelligence*.

Payandeh, B. (1983). Some applications of nonlinear regression models in forestry research. *The Forestry Chronicle*, *59*(5), 244–248.

Pearl, R. (1927). The growth of populations. *The Quarterly Review of Biology*, *2*(4), 532–548.

Pertusa, A. (2010). *Computationally efficient methods for polyphonic music transcription*. PhD thesis, Universidad de Alicante.

Philip, R. (1990). *Performance Practice: Music after 1600*, chapter The 20th Century: 1900–1940, (pp. 478). Macmillan.

Poli, G. D., Rodà, A., & Vidolin, A. (1998). Note-by-note analysis of the influence of expressive intentions and musical structure in violin performance*. *Journal of New Music Research*, *27*(33), 293–321.

Potter, J. (2006). Beggar at the door: the rise and fall of portamento in singing. *Music and Letters*, *87*(4), 523–550.

Prame, E. (1994). Measurements of the vibrato rate of ten singers. *The Journal of the Acoustical Society of America*, *96*(4), 1979–1984.

Prame, E. (1997). Vibrato extent and intonation in professional western lyric singing. *The Journal of the Acoustical Society of America*, *102*(1), 616–621.

Rabiner, L. (1977). On the use of autocorrelation analysis for pitch detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *25*(1), 24–33.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proc. of the IEEE*, volume 77, (pp. 257–286).

Randel, D. M. & Apel, W. (1986). *The new Harvard dictionary of music*. Belknap Press.

Raphael, C. (2002). Automatic transcription of piano music. In *Proc. of the International Society for Music Information Retrieval Conference*.

Repetto, R. C. & Serra, X. (2014). Creating a corpus of jingju (beijing opera) music and possibilities for melodic analysis. In *Proc. of the International Society for Music Information Retrieval Conference*.

Repp, B. H. (1995). Quantitative effects of global tempo on expressive timing in music performance: Some perceptual evidence. *Music Perception: An Interdisciplinary Journal*, *13*(1), 39–57.

Repp, B. H. (1998). A microcosm of musical expression. i. quantitative analysis of pianists' timing in the initial measures of chopin's etude in e major. *The Journal of the Acoustical Society of America*, *104*(2), 1085–1100.

Repp, B. H. (1999). A microcosm of musical expression: Ii. quantitative analysis of pianists' dynamics in the initial measures of chopin's etude in e major. *The Journal of the Acoustical Society of America*, *105*(3), 1972–1988.

Richards, F. J. (1959). A flexible growth function for empirical use. *Journal of experimental Botany*, *10*(2), 290–301.

Roebel, A., Maller, S., & Contreras, J. (2011). Transforming vibrato extend in monophonic sounds. In *Proc. of the 14th International Conference on Digital Audio Effects*.

Rosand, A. (2014). Aaron Rosand on portamento. *The Strad*, *http://www.thestrad.com/cpt-latests/aaron-rosand-on-portamento/*. Accessed in June 2016.

Ross, M. J., Shaffer, H. L., Cohen, A., Freudberg, R., & Manley, H. J. (1974). Average magnitude difference function pitch extractor. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *22*(5), 353–362.

Rossignol, S., Depalle, P., Soumagne, J., Rodet, X., & Collette, J.-L. (1999). Vibrato: detection, estimation, extraction, modification. In *Proc. of the Digital Audio Effects Workshop*.

Rossignol, S., Rodet, X., Soumagne, J., Collette, J.-L., & Depalle, P. (1999). Automatic characterisation of musical signals: Feature extraction and temporal segmentation. *Journal of New Music Research*, *28*(4), 281–915.

Ryynänen, M. P. & Klapuri, A. P. (2008). Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, *32*(3), 72–86.

Saitou, T., Unoki, M., & Akagi, M. (2005). Development of an f0 control model based on f0 dynamic characteristics for singing-voice synthesis. *Speech Communication*, *46*(3), 405–417.

Salamon, J. & Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(6), 1759–1770.

Salgian, A. & Vickerman, D. (2016). Computer-based tutoring for conducting students. In *Proc. of the 42nd International Computer Music Conference*.

Schroeder, M. R. (1968). Period histogram and product spectrum: New methods for fundamentalfrequency measurement. *The Journal of the Acoustical Society of America*, *43*(4), 829–834.

Seashore, C. E. (1932). *University of Iowa Studies in the Psychology of Music (Vol. 1: The Vibrato)*. Iowa City, Iowa: The University Press.

Seashore, C. E. (1938). *Psychology of Music*. New York: Dover Publications.

Serra, X. (1989). *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*. PhD thesis, Stanford University.

Slaney, M. & Lyon, R. F. (1990). A perceptual pitch detector. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (pp. 357–360).

Snyder, B. (2000). *Music and memory: An introduction*. MIT press.

Sousa, R. & Ferreira, A. (2010). Non-iterative frequency estimation in the dft magnitude domain. In *Proc. of the IEEE 4th International Symposium on Communications, Control and Signal Processing*, (pp. 1–4).

Spohr, L. (1852). *Grand Violin School*, volume 98. Boston: Oliver Ditson.

Srinivasamurthy, A., Repetto, R. C., Sundar, H., & Serra, X. (2014). Transcription and recognition of syllable based percussion patterns: The case of Beijing opera. In *Proc. of the International Society for Music Information Retrieval Conference*.

Stein, M. R. (2016). Sliding into jewishness: A pentimento of portamento. Honors theses, Wesleyan University.

Sundberg, J. (1977). The acoustics of the singing voice. *American Scientific*, *236*(3), 82–86.

Sundberg, J. (1994). Acoustic and psychoacoustic aspects of vocal vibrato. *Quarterly Progress and Status Report*, *35*(2-3), 45–68.

Sundberg, J. (1998). Expressivity in singing. A review of some recent investigations. *Logopedics Phoniatrics Vocology*, *23*(3), 121–127.

Sundberg, J., Gu, L., Huang, Q., & Huang, P. (2012). Acoustical study of classical Peking opera singing. *Journal of Voice*, *26*(2), 137–143.

Sung, A. & Fabian, D. (2011). Variety in performance: A comparative analysis of recorded performances of bach's sixth suite for solo cello from 1961 to 1998. *Empirical Musicology Review*, *6*(1).

Tian, M., Fazekas, G., Black, D., & Sandler, M. (2013). Towards the representation of Chinese traditional music: A state of the art review of music metadata standards. In *Proc. of the International Conference on Dublin Core and Metadata Applications*.

Tian, M., Srinivasamurthy, A., Sandler, M., & Serra, X. (2014). A study of instrument-wise onset detection in Beijing opera percussion ensembles. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Timmers, R. (2007). Vocal expression in recorded performances of Schubert songs. *Musicae Scientiae*, *11*(2), 237–268.

Timmers, R. & Desain, P. (2000). Vibrato: the questions and answers from musicians and science. In *Proc. of the International Conference on Music Perception and Cognition*.

Timmers, R. & Sadakata, M. (2014). *Expressiveness in music performance: Empirical approaches across styles and cultures*, chapter Training Expressive Performance by Means of Visual Feedback: Existing and Potential Applications of Performance Measurement Trechniques, (pp. 304–327). Oxford University Press.

Todd, N. P. M. (1992). The dynamics of dynamics: A model of musical expression. *The Journal of the Acoustical Society of America*, *91*(6), 3540–3550.

Tsoularis, A. & Wallace, J. (2002). Analysis of logistic growth models. *Mathematical biosciences*, *179*(1), 21–55.

van der Meer, W. (2014). *Expressiveness in music performance: Empirical approaches across styles and cultures*, chapter Audience Response and Expressive Pitch Inflections in a Live Recording of Legendary Singer Kesar Bai Kerkar, (pp. 170–184). Oxford University Press.

Ventura, J., Sousa, R., & Ferreira, A. (2012). Accurate analysis and visual feedback of vibrato in singing. In *Proc. of the IEEE 5th International Symposium on Communications, Control and Signal Processing*, (pp. 1–6).

Verfaille, V., Guastavino, C., & Depalle, P. (2005). Perceptual evaluation of vibrato models. In *Proc. of the Conference on Interdisciplinary Musicology*.

Verhulst, P.-F. (1838). Notice sur la loi que la population suit dans son accroissement. correspondance mathématique et physique publiée par a. *Quetelet*, *10*, 113–121.

von Coler, H. & Lerch, A. (2014). Cmmsd: A data set for note-level segmentation of monophonic music. In *Proc. of the Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*.

von Coler, H. & Roebel, A. (2011). Vibrato detection using cross correlation between temporal energy and fundamental frequency. In *Proc. of the Audio Engineering Society Convention 131*. Audio Engineering Society.

Vorberg, D. & Hambuch, R. (1978). On the temporal control of rhythmic performance. *Attention and performance VII*, 535–555.

Wall, M. R. & Neuhauser, D. (1995). Extraction, through filter-diagonalization, of general quantum eigenvalues or classical normal mode frequencies from a small number of residues or a short-time segment of a signal. i. theory and application to a quantum-dynamics model. *The Journal of chemical physics*, *102*(20), 8011–8022.

Wang, G. (2009). Track 4, Disk 2, An Anthology of Chinese Traditional and Folk Music – A Collection of Music Played on the Erhu. CD. ISBN: 9787799919928.

Wang, Y. (2012). Wenhua jiaorong zhiyu erhu rouxian 文化交融之于二胡揉弦 (culture conmunication on erhu vibrato). *Yueqi* 乐器, *7*.

Weninger, F., Amir, N., Amir, O., Ronen, I., Eyben, F., & Schuller, B. (2012). Robust feature extraction for automatic recognition of vibrato singing in recorded polyphonic music. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 85–88).

Wichmann, E. (1989). *Listening to theatre: the aural dimension of Beijing opera*. University of Hawaii Press.

Widmer, G. (2002). Machine discoveries: A few simple, robust local expression principles. *Journal of New Music Research*, *31*(1), 37–50.

Widmer, G. (2003). Discovering simple rules in complex data: A meta-learning algorithm and some surprising musical discoveries. *Artificial Intelligence*, *146*(2), 129–148.

Widmer, G. & Goebl, W. (2004). Computational models of expressive music performance: The state of the art. *Journal of New Music Research*, *33*(3), 203–216.

Widmer, G. & Tobudic, A. (2003a). Playing mozart by analogy: Learning multi-level timing and dynamics strategies. *Journal of New Music Research*, *32*(3), 259–268.

Widmer, G. & Tobudic, A. (2003b). Playing mozart by analogy: Learning multi-level timing and dynamics strategies. *Journal of New Music Research*, *32*(3), 259–268.

Xue, W. & Sandler, M. (2008). Analysis and synthesis of audio vibrato using harmonic sinusoids. In *Proc. of the Audio Engineering Society Convention 124*.

Yang, C. & Zhen, X. (2015). Erhu biaoyanzhong fengge tezheng de queli: hun huayin de zhongyaoxing 二胡表演中风格特征的确立:论滑音的重要性 (build the style chracteristics in erhu playing: the importance of portamento). *Dangdai Yinyue* 当代音乐, *31*.

Yang, G. (2010). Erhu rouxian de linian yu shijian 二胡揉弦的理念与实践 (erhu vibrato's theory and practice). *Jiefangjun yishu xueyuan xuebao* 解放军艺术学院学报, *1*.

Yang, L. (2008). Erhu yanzou de rouxian jiqiao 二胡演奏的揉弦技巧 (the vibrato technique in erhu). *Juying yuebao* 剧影月报, *3*.

Yang, L. (2012). Qiantan erhu rouxian xiaoguo yu shengyue chanyin xianxiang de gongxing 浅谈二胡揉弦效果与声乐颤音现象的共性 (on the commons of vibratos in erhu and vocal phenomenon). *Huanghe zhisheng* 黄河之声, *14*.

Yang, L., Chew, E., & Rajab, K. Z. (2013). Vibrato performance style: A case study comparing erhu and violin. In *Proc. of the 10th International Conference on Computer Music Multidisciplinary Research*, (pp. 904–919).

Yang, L., Chew, E., & Rajab, K. Z. (2014). Cross-cultural comparisons of expressivity in recorded erhu and violin: Performer vibrato styles. In *Proc. of the 4th International Workshop on Folk Music Analysis*.

Yang, L., Chew, E., & Rajab, K. Z. (2015). Logistic modeling of note transitions. In *Mathematics and Computation in Music* (pp. 161–172). Springer.

Yang, L., Rajab, K. Z., & Chew, E. (2016a). AVA: A graphical user interface for automatic vibrato and portamento detection and analysis. In *Proc. of the 42nd International Computer Music Conference*, (pp. 547–550).

Yang, L., Rajab, K. Z., & Chew, E. (2016b). AVA: An interactive system for visual and quantitative analyses of vibrato and portamento performance styles.

In *Proc. of the 17th International Society for Music Information Retrieval Conference*.

Yang, L., Rajab, K. Z., & Chew, E. (2017). Filter diagonalisation method for music signal analysis: Frame-wise vibrato detection and estimation. *Journal of Mathematics and Music*.

Yang, L., Tian, M., & Chew, E. (2015). Vibrato characteristics and frequency histogram envelopes in beijing opera singing. In *Proc. of the 5th International Workshop on Folk Music Analysis*, (pp. 139–140).

Yeh, C. (2008). *Multiple fundamental frequency estimation of polyphonic recordings*. PhD thesis, UNIVERSITÉ PARIS VI - PIERRE ET MARIE CURIE.

Zhao, H. (1999). Erhu yanzouzhong huayin de yunyong 二胡演奏中滑音技法的运用 (the application of portamento in erhu playing). *Zhongyang yinyue xueyuan xuebao* 中央音乐学院学报, *2*, 53–57.

Zhu, F., Fan, Z., & Wu, X. (2014). A nonlinear digital feedback oscillator based vibrato control model for singing synthesis. In *Proc. of the 2014 IEEE China Summit & International Conference on Signal and Information Processing*, (pp. 115–119).