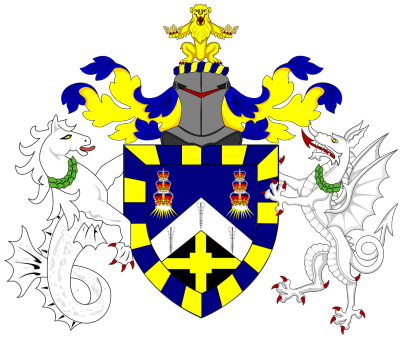
Segmenting and Analysing Pitch Contours across Singing Styles



Yukun Li

A thesis submitted for the degree of

Doctor of Philosophy

School of Electronic Engineering and Computer Science

Queen Mary University of London

Aug 2024

Statement of originality

I, Yukun Li, 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 herein.

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: Yukun Li

Date: 31/Aug/2024

Abstract

While previous studies have explored pitch contours in singing, a systematic computational analysis of their characteristics across diverse singing styles has yet to be conducted. This research develops a computational framework for the automated characterisation and segmentation of pitch contours, with the goal of describing and comparing vocal styles, and evaluates the framework in three case studies.

The first study introduces a novel methodology for automatically delineating three distinct pitch contour elements: steady, modulating, and transitory. Initial pitch tracks were extracted using the PYIN algorithm, and a 'pitch contour unit' was proposed to tokenise pitch contours. This unit formed the basis for a hidden Markov model (HMM) that detected sequences of pitch contour elements. The proposed method outperformed established benchmarks in segmenting Jingju (Peking opera) pitch contours. Additionally, it demonstrated adaptability in identifying sustained notes in Georgian vocal music and detecting portamento and vibrato in Jingju.

The second study analysed pitch contours at the note level in selected Alpine yodel and Russian folk music songs. Results indicated consistency in note annotations made by different cultural experts. This analysis revealed distinct approaches employed by singers in each style to shape pitch contours for connecting and holding notes.

The third study examined pitch contours in Chinese Chaozhou folk music, where vocal style analysis often utilises syllable-level segments corresponding to Chinese characters. The study used the discrete cosine transform (DCT) to characterise pitch contours at the syllable level, examining the effects of lexical tone in speech on singing pitch contours. The analysis employed statistical models to identify the effect of lexical tones and other factors, such as training background, frequency of singing in the dialect

and melodic interval, on sung pitch contours in Chaozhou vocal folk music data.

O time, thou must untangle this, not I. It is too hard a knot for me to untie!

Twelfth Night

Shakespeare

Acknowledgements

First and foremost, I'd like to thank myself—because someone has to! Ten years ago, I started to dream of researching music through computational methods, but I had no idea if such research existed or where to begin. Despite this uncertainty, I bravely pursued my passion, and life gradually revealed the path—through books, researchers, the ISMIR conference, PhD students in C4DM, and eventually meeting my supervisor, Prof. Dixon. This journey transformed my dream into a clear path, leading me to this PhD.

I am deeply grateful to my supervisor, Prof. Dixon, whose guidance has been invaluable. He has taught me to think critically and scientifically, and to uncover the essence behind surface-level engineering problems. His dedication to scientific research has set an exemplary standard of what it means to be a great researcher. I also thank my additional supervisor, Dr. Polina Proutskova, for contributing her expertise in singing to support my research and for organising academic activities that broadened my perspective. Their mentorship has been instrumental in shaping my academic path.

I would also like to extend my heartfelt thanks to the people at QMUL and C4DM, as well as to my flatmates and friends at the QMUL St Benet's Chaplaincy. Your support and the time we spent together, filled with joy, love, and peace, have been truly amazing and have enriched my experience in countless ways.

Special thanks go to my family for their unwavering support, and to the China Scholarship Council and Queen Mary University of London joint Ph.D. Scholarship for covering the financial costs that made this research possible. I am also immensely grateful to my psychoanalyst, Dr. Xue Ying, for her mental support during the challenging

times of this journey.

Finally, I want to acknowledge everyone who has contributed to my academic and personal growth throughout this process. Your encouragement and belief in me have been a source of strength and inspiration.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis focuses on characterising vocal style through fundamental frequency (f0) contour analysis. In this chapter, the motivations and aims of this research are outlined in Sections 1.1 and 1.2. Then, the overall structure of the thesis is presented in Section 1.3, along with the key contributions of this work. Finally, Section 1.4 concludes the chapter with a list of publications of the author related to the thesis.

1.1 Overview

Singing is a common component of music, with vocal music exhibiting diverse characteristics across various styles and cultures. The systematic characterisation of vocal music styles is essential for gaining deeper insights into the music. Pitch, a fundamental element in vocal music, significantly contributes to the conveyance of vocal style. Previous studies have employed f0 to characterise vocal styles within specific cultural contexts, as demonstrated by Ganguli et al. (2017), Yang, Tian, Chew et al. (2015*a*), and Devaney (2011). However, a systematic method for characterising vocal styles across different cultures remains lacking.

It is important to note the distinction between f0, which is an objective physical measurement of vocal fold vibration, and pitch, which is the subjective perception of frequency by listeners. While this thesis aims to analyse f0 estimated from audio to characterise vocal styles, the terms “f0” and “pitch” are used interchangeably through-

out, following common practice in music information retrieval research. This approach is supported by Ma et al. (2022), which argued that despite the differences between f0 and pitch, f0 can be considered a valid proxy for pitch due to its close correlation with perceived pitch. Furthermore, f0 is one of the most accessible and cost-effective measures for studying pitch variations. Since this research focuses on analysing pitch contour shapes rather than intonation, any minor discrepancies between f0 measurements and perceived pitch do not impact the validity of the analysis.

The f0 of singing voice is continuous, with rich variations in f0 contour that can be categorised into various expression types. While certain expression types, such as vibrato (Wen & Sandler 2008) and portamento (Yang, Chew & Rajab 2015), have been modelled and analysed, there are few computational systems that generally define and model multiple expression types or f0 contour elements simultaneously. Mayor et al. (2006) defined and modelled several types of expressions, such as normal, scoop up/down, fall-down, portamento up/down, and other expressive labels, and Gong et al. (2016) abstracted three basic f0 contour elements: steady, transitory, and vibrato. However, no objective evaluation of the automatic segmentation of their methods was reported.

This motivates the exploration of f0 contour segmentation methods and their evaluation in Chapter 3. Section 2.4 reviews common f0 contour elements across different cultures and styles by examining ornaments documented in Western art, Western pop, Indian art, and Chinese traditional opera. Then Chapter 3 focuses on developing a method to detect these basic f0 contour elements and evaluate the segmentation objectively, for the purpose of detecting expression segments specific to musical cultures or styles. For example, Jingju portamento and Georgian steady regions are detected by fine-tuning the method.

Moreover, comparative analyses of vocal style in terms of f0, such as those conducted by Sundberg et al. (2012) and Caro Repetto et al. (2015), primarily focus on specific expressive features like vibrato. Building on this foundation, Chapter 4 extends the analysis by performing a note-level investigation to explore how notes are sustained and transitioned through f0 contour shaping in vocal music across various cultures and

styles.

Syllable-level f0 contour analysis is also necessary for some genres of music such as Chinese traditional music and Indian art music, in which f0 is shaped by the Chinese character or Indian svara. Previous studies, such as Caro Repetto et al. (2017*a*), have used documented musical scores to analyse syllable-level f0 changes. This approach discarded the f0 contour from real singing. While Zhang (2024) made recordings of singing, she lacked robust techniques to characterise f0 contours effectively. Thus, Chapter 5 aims to develop syllable-level f0 contour analysis methods to enhance systematic singing style characterisation, particularly in examining the effects of lexical tone of Chinese characters.

1.2 Aim and Research Questions

The primary aim of this research is to develop a systematic approach for characterising singing style through f0 contour analysis, viewed from three key perspectives:

- **Defining and detecting basic pitch contour elements:** This perspective involves defining and detecting basic f0 contour elements that can be grouped to form complete f0 contours across various musical cultures and vocal styles. Pitch contours refer to the temporal evolution of fundamental frequency in musical phrases, as illustrated in Figure 3.3. These contours can manifest as basic patterns (ascending, descending, or fluctuating) or complex trajectories. Ornaments are specialised melodic embellishments characterised by distinctive f0 contour shapes, such as portamento (continuous f0 slides between musical notes) and vibrato (periodic oscillation of f0 around a central frequency), as well as short auxiliary notes that precede or follow a main structural note.

- **Note-level pitch contour analysis:** This perspective focuses on detecting and analysing the f0 contours that occur at transitions between notes and the held regions within notes, based on a segmentation of the singing into notes.

- **Syllable-level pitch contour analysis:** This stage aims to characterise f0 contour shapes based on syllable segmentation, with particular emphasis on in-

vestigating tone effects in singing. Tone effects refer to how the lexical tones of Chinese characters are realized in the f0 contours during singing, making the f0 contour of a sung syllable tend to preserve characteristics of the original spoken tone pattern. For example, a character with a rising tone may exhibit an overall rising trajectory in its sung f0 contour, even within the constraints of the musical melody.

The research questions guiding this study are as follows:

- How can the singing pitch contour be segmented into a set of basic f0 contour elements?
- Can a hidden Markov model be effectively employed to detect these defined basic f0 contour elements?
- Is the f0 contour element detection method applicable to downstream tasks, such as detecting steady regions and ornaments (e.g., vibrato, portamento, glissando, mordent) across different musical cultures or styles?
- How can singing styles be compared at the note level when the dataset does not contain the same song performed in different styles?
- What is the relationship between phoneme and note boundaries? Can the accuracy of automatic note segmentation be improved by incorporating phoneme segments?
- Can syllable-level visualisation and characterisation of f0 contour effectively demonstrate the influence of tonal effects on singing f0?

1.3 Thesis Structure

Chapter 1: Introduction

This chapter outlines the motivations behind this research and establishes the research aims and questions. It also includes a list of relevant publications of this author and highlights the main contributions of the thesis.

Chapter 2: Background and Previous Work

This chapter lays the groundwork for understanding vocal style in terms of f0 analysis and contextualises related research. It begins by presenting an overview of the physical mechanics behind vocal production, detailing the roles of the lungs, vocal folds, and vocal tract. The chapter then lists previous work exploring how the auditory system perceives and interprets these sound waves, transforming them into musical experiences. It covers related work about various aspects of vocal music, including musical form, content, and performance. A special focus is given to f0 and melodic ornaments in performance, which are important to reflect vocal style. The chapter also provides descriptions of vocal ornaments across Western art, Western pop, Indian art, and Chinese traditional music, highlighting both universal and culture-specific practices. Furthermore, it outlines previous research approaches to detection methods and computational modelling of melodic ornaments. The chapter ends with a critical review of existing vocal style studies, underscoring the limitations of current computational techniques in vocal style analysis.

Chapter 3: Pitch Contour Segmentation and Characterisation Methods

Chapter 3 introduces the concept of the 'pitch contour unit' (PCU), which is used to segment and characterise pitch contours across musical cultures. The chapter begins by defining PCUs as discrete segments of the f0 signal delineated by consecutive local peaks and troughs in f0. It then details the dataset used for training and evaluation of the model, which includes annotated recordings from Jingju and Georgian music. The methodology section outlines the training and inference processes for a Hidden Markov Model (HMM) used to detect primary elements of pitch contours: steady, modulating, and transitory. The evaluation of the detection of these pitch contour elements is followed by specific evaluations of portamento, steady regions, and vibrato detection. Each evaluation provides detailed results and comparisons with existing methods, highlighting the effectiveness and revealing the weaknesses of the proposed approach.

Chapter 4: Note-Level Pitch Contour Analysis

This chapter presents a comparative analysis of pitch contours in examples of Alpine and Russian singing at the note level. The dataset comprises singing recordings, f0 data, and two versions of note segments for each culture, transcribed by two cultural experts from each tradition. These annotated note segments are utilised to evaluate the proposed automatic note segmentation method and to highlight the limitations of the automatic method. The comparative analysis of different transcription versions reveals a consistency in note annotations, alongside distinct preferences, among experts from the same cultural background. This chapter systematically examines various features of both the held and transitional regions of musical notes, using visual and statistical methods to highlight the differences and similarities between these two distinct vocal traditions. The chapter contributes to establishing a computational framework for note-level pitch contour analysis across diverse musical traditions.

Chapter 5: Syllable-Level Pitch Contour Analysis

This chapter investigates the correlation between lexical tones and syllable-level pitch contours in Chaozhou folk singing. The dataset consists of recordings from 34 singers performing the same song. The discrete cosine transform (DCT) is employed to quantify the linear tendency and curvature of the pitch contour for each sung syllable. Linear mixed models are applied to assess the significance of the effects of lexical tones and other factors, such as training background, experience in singing in the Chaozhou dialect, tone sandhi, vowel type, and melodic interval, on the sung pitch contours. The results confirm that lexical tones have a significant effect on the linear tendency of sung syllable pitch contours, while other factors also influence the pitch contour to varying degrees.

Chapter 6: Conclusions and Future Perspectives

This chapter summarises the key achievements of this thesis and discusses future directions for further research and potential applications of this work.

1.4 Associated Publications

This thesis encompasses research on vocal pitch contour and vocal style analysis conducted by the author from October 2018 to August 2024 at Queen Mary University of London, under the supervision of Simon Dixon. Portions of this work have been presented at international peer-reviewed conferences.

Peer-Reviewed Conference Paper

(i) Li, Y., Demirel, E., Proutskova, P., and Dixon, S. (2021). Phoneme-informed Note Segmentation of Monophonic Vocal Music. In *Proceedings of the 2nd Workshop on NLP for Music and Spoken Audio (NLP4MusA)*, pages 17–21.

Other Publications

(ii) Proutskova, P., McBride, J., Ozaki, Y., Chiba, G., Li, Y., Yu, Z., Yue, W., Crowdus, M., Zuckerberg, G., Velichkina, O., *et al.* (2023). The VocalNotes Dataset. In *Late-Breaking/Demo Session at the 24th International Society for Music Information Retrieval Conference (ISMIR 2023)*, Milan.

(iii) Proutskova, P., Chiba, G., Crowdus, M., Nikolaenko, I., Ozaki, Y., Shuster, L., Velichkina, O., Yue, W., Zuckerberg, G. A., Li, Y., *et al.* VocalNotes: Investigating the Perception of Note Pitch and Boundaries through Varying Transcriptions of Vocal Performances from Five Musical Cultures. In *Analytical Approaches to World Musics (AAWM)*.

(iv) Proutskova, P., Velichkina, O., McBride, J., Chiba, G., Crowdus, M., Nikolaenko, Y., Ozaki, Y., Shuster, L., Yu, Z., Yue, W., Zuckerberg, G., Killick, A., Li, Y., Phillips, E., and Savage, P. E. (2024). VocalNotes Methodology: Framework, Challenges and Lessons. (Under review)

Conclusions

This chapter has established the motivation for characterising vocal style through pitch contour analysis and articulated the primary aim of this research: to develop a systematic approach for analysing singing style across different cultures. The chapter has also outlined the structure of the thesis, identifying the key contributions within each chapter. Chapter 3 introduces the concept of pitch contour unit and details the methods for their detection. Chapter 4 presents a comparative analysis of pitch contours at the note level across different musical cultures, while Chapter 5 extends the analysis to the syllable level, focusing on the effects of lexical tones in Chaozhou folk singing. Finally, the chapter concludes by listing the author's publications related to this research. The upcoming literature review synthesises the necessary background and contextual information, highlighting the limitations of previous studies to illustrate the motivation behind this research.

Chapter 2

Background

The singing voice stands as a unique auditory phenomenon, distinct from both environmental sounds and other forms of musical expression. Unlike environmental sounds, the singing voice is a human creation, and unlike instrumental music, it emanates from the most natural of instruments—the human vocal system. Furthermore, while sharing similarities with speech, the singing voice transcends mere communication to become an experience of musical artistry.

This chapter provides a foundational overview aimed at enhancing understanding of various aspects of the singing voice, while acknowledging that not all complexities can be covered in a single chapter. It begins by delving into the *Physical Essentials of the Singing Voice*, exploring the mechanics behind vocal production. The journey continues through the *Auditory Journey: From Singing Voice to Subjective Musical Experience*, examining how the voice is perceived and conceived. The chapter then shifts its focus to *Vocal Music and Vocal Style*, followed by an exploration of the *Musical Context and Melodic Ornaments of Vocal Music in Different Musical Cultures.* It further narrows down to the focus of this thesis, *Computational Modelling of Melodic Ornaments for Vocal Style Understanding*, laying the groundwork for computational approaches to vocal style analysis, summarising key research while critically examining their limitations. The chapter concludes with a *Review of Vocal Style Analysis in Pitch Contour Studies*, which also serves to both summarise previous studies in the field and discuss their shortcomings.

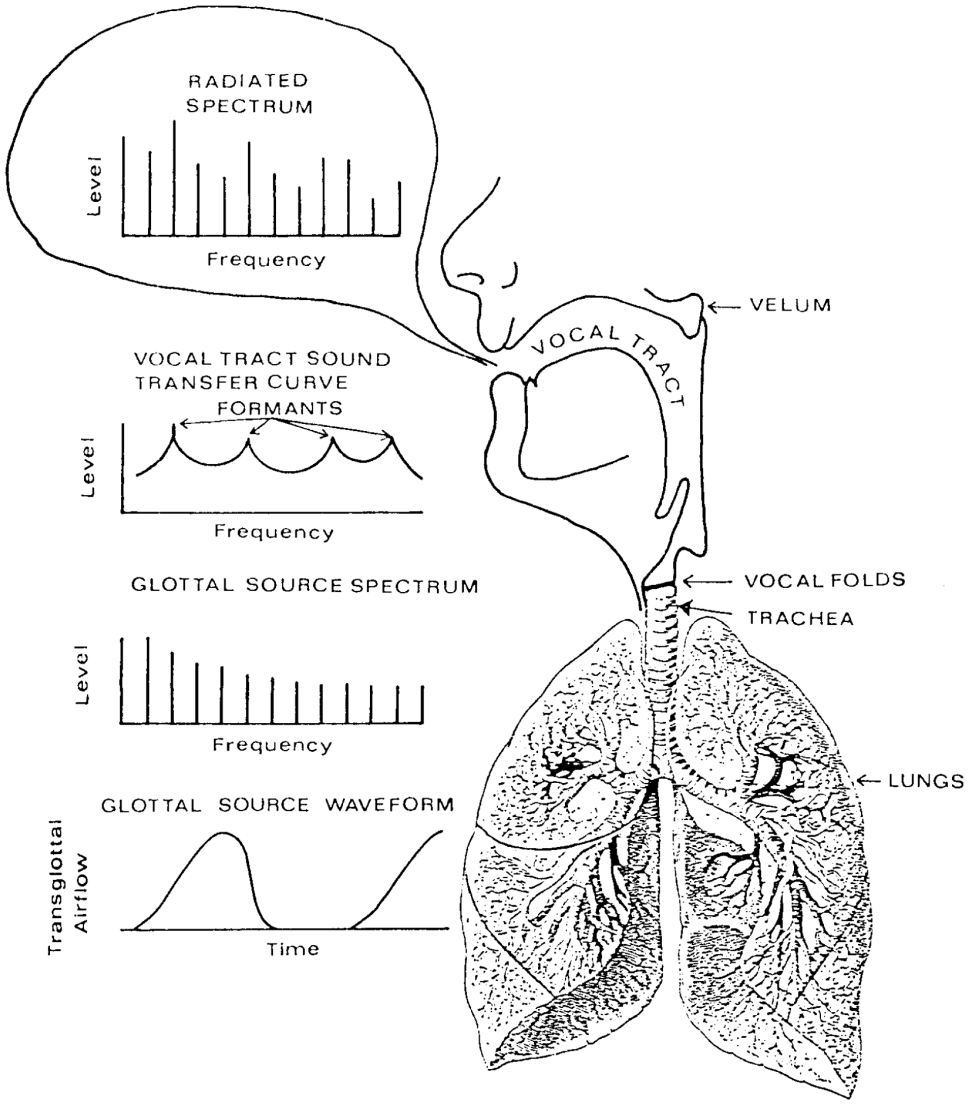


Figure 2.1: A schematic representation of the human vocal organs and physical properties of sound, cited from Sundberg (1995*a*).

2.1 The Physical Essential of the Singing Voice

The singing voice, not only encompasses elements such as perception and music theory, but also rooted in the study of the human voice organs and the physical properties of sound. This section will primarily focus on the latter aspects.

2.1.1 Production of the Singing Voice

To understand how singing is produced, we must delve into the anatomy and functioning of voice organs. Modern physiology and studies like Sundberg (1995*a*) have illuminated

the complexity of this system, providing an illustration which is shown in Figure 2.1. Here is a brief summary.

The vocal apparatus consists of three parts: the lungs, the vocal folds, and the vocal tract. The lungs act as a power source, generating an airstream. The vocal folds, located at the bottom of the larynx, are brought to vibration by the airstream, thus creating pulsations of air - the sound waves or the voice source, with their tension, closure, and length controlled by muscles within the larynx. The vocal tract, comprising the larynx, pharynx, and nasal cavity, resonates and modulates the voice source, with its shape determined by the positioning of the lips, jaw and tongue.

The process of singing involves a complex interplay of these elements. The closed glottis causes excess pressure from the airstream, forcing the vocal folds apart, and a subsequent Bernoulli force closes the glottis. The repeating cycles form the vibration of the vocal folds, with the frequency depending on various factors such as tension, thickness, length of vocal folds and air pressure. The amplitude of the vibration is controlled by air pressure and the degree of closure of the vocal folds. The vocal tract acts as an acoustic filter, selectively attenuating different frequency components of the voice source (glottal waveform). The resulting resonances, known as formants, shape the spectral characteristics of the final acoustic output. These resonance frequencies shift by varying the shape of the vocal tract. During both speech and singing, the vocal tract continuously alters its shape, resulting in varying resonance frequencies throughout phonation.

2.1.2 Physical Properties of the Singing Voice

The physical properties of the singing voice are characterised by two key elements: the voice source and the resonating effect of the vocal tract. The voice source consists of a fundamental frequency (f0), the lowest partial of the voice source spectrum perceived as pitch, and its higher harmonics. The amplitude of these harmonics, typically decreases as their frequency increases. Although the voice source spectrum remains fairly consistent among different singers, it undergoes significant transformation when passing through the vocal tract. The vocal tract forms about four or five major resonances

known as formants, which significantly reshape the sound spectrum. These formants not only create the distinct sound of each singer's voice, altering the spectrum substantially to produce unique vocal qualities, but also play a pivotal role in determining the specific vowel sounds.

Overall, the singing voice is a complex interplay of sound waves produced and shaped by the human vocal organs. These sound waves possess distinct physical properties, manipulated by the movement and coordination of the lungs, vocal folds, and vocal tract. As these waves travel through the air, they carry the unique characters of the singer's voice to the listeners.

2.2 The Auditory Journey: From Sound of Singing Voice to Subjective Musical Experience

The process by which the singing voice is perceived and cognitively interpreted as vocal music is a nuanced one that involves intricate aural perception and cognitive interpretation. It is through this process that sound waves are appreciated as music, encapsulating elements such as melody, lyrics and harmony.

The perception of the singing voice begins with the detection of sound waves by our auditory system, encompassing two basic listening conditions. As listeners, our outer ear captures external sound waves, and our inner ear converts them into electrical signals that the brain interprets. When singers listen to their own voices, the perception includes not only the outer and inner ear but also the transmission of sound through bone and muscle conduction.

In the auditory perception process following sound reception, the physical properties of sound undergo a subjective transformation into auditory sensations. The f0 is commonly associated with our sense of pitch, whereas the sound's power spectrum ——produced by vocal fold vibration and filtered through the vocal tract ——shape our perception of loudness. The timbre of a sound is crafted by the time-varying vocal tract's influence on the sound spectrum. These perceptual translations are governed by a complex interplay between the physiology of the vocal mechanism, the acoustic

properties of the sound waves, and the resultant auditory experiences. This interplay is characterised by a detailed mapping among these elements, which is affected by both linear and non-linear phenomena, as delineated in Table 2.1.

Recent research from Edmonds & Howard (2023) found that the listening condition influences pitch perception. The listening condition affects how we perceive pitch because it involves two different transmission pathways: air conduction for external sounds and bone conduction for our own voice. These pathways deliver different frequency balances, adding a layer of complexity to vocal self-perception and creating a perceptual difference between how we hear our own voice compared to external voices.

In addition, an important consideration in pitch perception of singing is the intrinsic pitch of vowels. Research by Stoll (1984) demonstrated that even when the f0 remains constant, changes in vowel quality can result in perceived pitch differences. This phenomenon occurs due to the interaction between the spectral envelope and the auditory system's pitch perception mechanisms. For example, the vowel /a/ has a spectral envelope with prominent energy peaks at lower frequencies, causing a perceived downward pitch shift, while /i/ has prominent energy peaks at higher frequencies, resulting in an upward pitch shift. These pitch shifts can be significant, with differences up to 1.4% in pitch perception between vowels such as /a/ and /y/ when the fundamental frequency is 125 Hz. While these differences exceed the typical pitch discrimination threshold of 0.25-0.3%, their impact varies by application: they significantly affect measurements of absolute pitch accuracy but have minimal influence on pitch contour analysis where the relative shape of the melody is more important than absolute pitch values.

Upon perceiving a singing voice, humans engage in a cognitive process that transforms this auditory input into vocal music. Unlike speech, this transformation incorporates both verbal and nonverbal sound elements, systematically organising them into structured, meaningful sequences. The complexity of this process lies in converting continuous auditory features into coherent units, crafting the rich and nuanced experience of music. This cognitive aspect is crucial in shaping these sounds into a form that transcends mere auditory sensations, rendering them as structured and significant musical expressions.

Voice Organ Effects	Sound Waves Properties	Auditory Perception
Vocal Folds Vibration Amplitude and Vocal Tract Resonance	Amplitude	Loudness
Vocal Folds Vibration Frequency	f0	Pitch
Vocal Tract Resonance	Spectrum	Timbre

Table 2.1: Mapping among vocal organ's movements, generated sound waves, and auditory perception

The journey from a singing voice to vocal music is a sophisticated interplay of perception and cognition, deeply rooted in individual experience and interpretation. Initially, the human auditory system captures sound waves produced by the singer. These sounds are then processed through subjective perceptual and cognitive faculties. Key musical elements such as melody, harmony, rhythm, and dynamics are discerned and recognised through this process. The inherent nature of auditory perception and cognition, shaped by the structure of our auditory system, lays the foundation for this interpretation. However, the way in which music is ultimately understood and appreciated is greatly influenced by personal habits, experiences, and knowledge structures (Heng & Wang 2022).

In essence, transforming singing into vocal music is not just a matter of hearing sounds; it's about constructing a meaningful and subjective musical experience. This transformation is guided by the listener's ability to decode and contextualise musical elements, thereby converting auditory signals into an enriched musical narrative.

2.3 Vocal Music and Vocal Style

Vocal music is a rich and multifaceted art form that encompasses three fundamental aspects: musical form, content, and vocal performance. Musical form provides the structural framework, content offers the foundational material, and vocal performance focuses on the skilful interpretation and expressive delivery of music.

2.3.1 Musical Form: Composition, Performance, and Understanding

Musical form, which refers to the structure of a musical composition or performance, is shaped by the arrangement and organisation of various elements that contribute to creating a cohesive and expressive work. Mode decides a specific type of scale with a unique arrangement of whole and half steps. Straehley & Loebach (2014) provides insights into the historical use of the term "mode" in discussions of musical structure and affect, highlighting its significance over thousands of years. The concept of tonality in music theory revolves around the organisation of pitches and chords around a

central note, known as 'tonic'. Research by Dibben (1994) provides evidence for the internal representation of tonal music in terms of a hierarchy of events. Metre sets temporal organisation of beats, where certain beats are perceived as more salient than others across various time scales (Grahn 2012). Additionally, compositional structures like the sonata, fugue, and rondo, each with their distinct rules and patterns, play a crucial role in guiding the development and organisation of musical ideas, significantly influencing the build-up and resolution of musical tension. In the composition process, these forms serve as a blueprint, guiding how the piece is constructed. During performance, they inform interpretation and delivery, providing cues for phrasing, dynamics, and expression. In understanding music, these forms act as a roadmap, shaping the musical journey and enhancing listener engagement and comprehension by providing a coherent framework.

2.3.2 Content in Vocal Music

Vocal music presents a unique fusion of musical elements such as melody, rhythm, and texture, with the distinct addition of lyrics, thus creating two primary categories of study: musical content and linguistic content. In the realm of musical content, significant research has been conducted. Notable examples include Panteli et al. (2018), which investigated into melodic contour and Mzhavanadze & Scherbaum (2020), which analysed harmonic intervals in Georgian homophonic vocal music. These studies explore the intricacies of the musical aspects of vocal music. Parallel to musical analysis, researchers have also examined verbal content, with studies such as Fell et al. (2023) and Anisah (2023) conducting investigations into lyrical analysis. While acknowledging this research direction, it falls beyond the scope of this thesis. Additionally, the interplay between musical and linguistic content has been explored by several researchers, with studies like Caro Repetto et al. (2017*a*) and Zhang et al. (2017), Zhang & Cross (2021*a*) revealing complex interdependencies between these two facets. Specifically, these studies demonstrate how the lexical tones of Chinese characters influence both melodic composition and the pitch contours of individual syllables within vocal performances.

2.3.3 The Dual Aspects of Vocal Performance: Vocal Technique and Vocal Expression

Building upon the musical form and content of vocal music as its foundational elements, this thesis primarily concentrates on studies of vocal performance, viewing it through the lens of two interrelated but distinct aspects: vocal technique and vocal expression. While vocal technique focuses on the mechanical aspects of singing, providing the necessary tools for precise control over pitch, rhythm, timbre, and dynamics, vocal expression is concerned with the artistic interpretation and conveyance of emotion in music. This thesis investigates the realm of vocal expression, exploring the style of vocal music. In this exploration, vocal technique is not the primary subject but rather serves as a critical backdrop, offering insights into the mechanics that support and enhance the understanding of vocal expression.

Vocal techniques refer to the physical and mechanical aspects of singing, including breath control, phonation mode, resonance, articulation, and pitch accuracy and modulation. These physiological methods form the foundation for controlled and effective sound production. Research has examined various vocal techniques and their musical effects, exploring how phonation mode (Sundberg 1995*b*) and vocal tract shape (Mainka et al. 2015) influence timbre. Phonation mode affects timbre through variations in vocal fold vibration patterns. For example, breathy phonation produces a softer, airier timbre while pressed phonation creates a harder, more strident sound. The vocal tract shape modifies the acoustic resonances (formants), by adjusting the positions of the throat, mouth, and tongue. These configurations alter the timbral quality of the voice by selectively reshaping its frequency components.

Vocal expression represents the artistic and emotional dimension of singing. It involves interpreting music and lyrics to convey emotions, feelings, and meaning to the audience. This includes dynamics, phrasing, and subtle nuances that animate the music. Studies have examined various facets of vocal expression, such as timing (Yang, Huang & Everett 2017), timbre (Rossing & Sundberg 1984), loudness (Yang, Huang & Everett 2017), pitch contours (Mayor et al. 2006), and pronunciation (Gong 2018).

The two aspects are inseparable in a complete vocal performance, with technique

serving as the vessel for expression. A singer's technical prowess allows for the freedom and flexibility to explore various expressive possibilities, while the expressive choices, in turn, may influence the application of technique. This dynamic interplay creates a rich and complex vocal music experience.

2.3.4 Understanding Vocal Style: A Focus on Pitch and Melodic Ornaments

In this thesis, vocal style is defined as a specific manner of vocal performance that typically occurs within a particular musical context. The term "musical context" here encompasses both the musical form and the content of vocal music.

Vocal style involves multiple elements like pitch, rhythm, timbre, and dynamics. Among these, pitch is universally significant across musical traditions (Brown & Jordania 2013) and is quantifiable. Vocal music presents a complex pitch evolution with time, characterised by continuous variations and expressive gestures. Within the realm of pitch, melodic ornaments and intonation serve as two distinct but critical aspects. While intonation concerns how closely the singer's pitch aligns with the intended melody, melodic ornaments add layers of stylistic and emotional complexity. Given their intricate nature, melodic ornaments offer deeper insights into vocal style, making them the focal point of this thesis.

Vocal style is multifaceted, reflecting factors like musical schools, regional genres, cultural traditions, and individual singers' expression, background, and experience. It can be characteristic of broad entities such as musical schools, regional genres, and even entire musical cultures, which have their own special repertoire. These large entities often employ specialised vocal techniques, which must align with the tradition for a singer to be regarded as proficient. Besides, vocal style can be affected by the individual singer's expression, cultural background, and musical experience.

The organisation and interpretation of melodic ornaments are profoundly influenced by cultural context. This leads to unique musical forms and hierarchical structures of content within different traditions. Each culture has its own way of segmenting musical content into various hierarchical levels, such as repertoire, song, section, and individual

notes or syllables. This thesis will investigate the nuances of pitch contours of melodic ornaments at the foundational level of notes or syllables, where melodic nuances are prominently explored. Detailed analyses across Western, Indian, and Chinese musical traditions will be provided in Section 2.4.2.

2.4 Musical Context and Melodic Ornaments of Vocal Music in Different Musical Cultures

Melodic ornaments, realised through intricate variations in pitch, are a universal aspect of musical expression across diverse cultures. These ornaments are not merely decorative but are essential in defining the vocal style. For instance, in Chinese traditional opera, preliminary observations suggest a distinctive stylistic feature where vibrato rate appears to accelerate towards phrase endings, though this phenomenon awaits systematic investigation. This section investigates the intricate world of melodic ornaments across Western, Indian, and Chinese musical traditions. It places a particular emphasis on Western pop music, Western art music, Indian Art Music, and Chinese traditional opera. The exploration aims to uncover both the shared elements and the distinct characteristics that define the vocal pitch contours within each cultural context. The section is divided into subsections focusing on the musical context of melodic ornaments and types of melodic ornaments in different cultures and making comparisons across cultures.

2.4.1 Musical Context in Different Musical Cultures

The musical context in which melodic ornaments are employed varies significantly across different cultural traditions. Understanding this context is crucial for appreciating the role and significance of melodic ornaments in each tradition.

Musical Scale and Musical Notes in Western Pop Music and Western Art Music

In Western pop music and Western art music, the basic unit is the musical note, which represents a specific perceived pitch and duration (Brown 2017). A musical scale, a fundamental concept in music theory, is a set of musical notes ordered by pitch. In Western music, this typically consists of seven pitch classes. These ordered notes serve as the foundational structure for musical composition, forming the basis for melodies and harmonies. Additionally, melodic ornaments function at the note level (Thompson et al. 2023), adding expressiveness by either transitioning between notes or sustaining a note for an extended duration.

Raga and Swaras in Indian Art Music

In Indian classical music, the basic unit is the swara or svara, which is roughly equivalent to musical notes in Western music. Swaras are used to construct the musical scale, typically comprising seven swaras: Sa, Ri (Carnatic) or Re (Hindustani), Ga, Ma, Pa, Dha, and Ni. Unlike Western notes, swaras are enriched and characterised by specific melodic ornaments. The raga system in Indian classical music further defines the melodic framework, with each raga having a unique combination of swaras and associated ornaments. A distinctive feature of Indian vocal art music is that it is normally sung using swara syllables without the need for meaningful lyrics (Rao et al. 2023).

ShengQiang and Characters in Chinese Traditional Opera

The basic unit in Chinese traditional opera is the Chinese character. Chinese is a tonal language where the meaning of a word can change depending on the tone used. Chinese characters are spoken with a single syllable and tone. The musical audio stream is perceptually segmented according to linguistic syllabic boundaries rather than melodic notes (Shen 1982).

Traditional Chinese defines four types of tones (“平” (Ping), “上” (Shang), “去” (Qu), “入” (Ru)). In contrast, the modern system, introduced by Chao (1930), uses

a five-level tone mark, representing different degrees of pitch. For example, Mandarin tones include four types of tone, “55”, “35”, “214”, “51”. The numbers provided in the context of Mandarin tones represent the relative pitch levels of each tone in the tonal system. These numbers are based on a scale where 1 represents the lowest pitch level and 5 represents the highest. For example:

- “55” represents a high, steady tone (first tone in Mandarin).
- “35” starts at a mid-level and rises to a high level (second tone).
- “214” begins at a low level, dips to an even lower level, then rises to a high level (third tone).
- “51” starts high and sharply falls to a low level (fourth tone).

Chinese dialects exhibit rich variations in tone. For example, in the Chaozhou dialect, there are eight types of tones (Zhang & Cross 2021*b*), “33,” “55,” “53/21,” “35,” “213,” “11,” “21,” and “54.” In Chinese, the sequencing of characters in speech can trigger tone sandhi, a phenomenon where tones undergo modification. If two adjacent characters possess similar tones (e.g., both 55), tone sandhi may alter the first character’s tone to create a smoother phonetic transition. Chaozhou dialect is particularly rich in tone sandhi and the lexical tones are varied to “23,” “213,” “24/35,” “21,” “42/53,” “12,” “33/54,” “21” (Zhang & Cross 2021*a*).

ShengQiang, akin to Raga in Indian music, serves as the musical form and melodic framework in Chinese traditional music. Each ShengQiang encapsulates a distinctive linguistic dialect and its associated ornaments, following the principle encapsulated by the phrase 依字行腔 (Pinyin: yī zì xíng qiāng, literally translating to “singing according to the syllables”) (Gong 2018). This implies that the melody’s pitch contour should align with the tonal quality of each syllable, ensuring a harmonious integration of linguistic tone and musical expression. This application of melodic ornaments, while conventionally established in different ShengQiangs and passed down through oral transmission, is not rigid. It follows a common guideline but also grants singers the flexibility to personalise ornaments subtly, allowing for individual expression and interpretation (Guo 2021).

Comparison of Western, Indian, and Chinese Traditions

The structure and cultural context of melodic ornaments across Western, Indian, and Chinese musical traditions reveal distinct characteristics and intriguing similarities. In Western music, the basic unit is the musical note, representing specific perceived pitches defined by a musical scale, emphasising the precise tonal quality without a higher-level structure for melodic ornaments. Contrastingly, Indian classical music employs the swara, a pitch defined by the raga system, characterised by associated melodic ornaments and sung using specific syllables without meaningful lyrics, removing a linguistic dimension. Chinese traditional opera takes a unique approach with the basic unit being the Chinese character; the pitch contour of the sung syllable aligns with the character's tone and is shaped by melodic ornaments, creating a profound connection between language and music. These distinctions highlight the diverse ways in which different cultures approach the fundamental building blocks of musical perception and analysis, each reflecting unique cultural contexts, language functions and musical concepts.

Interestingly, Indian arts music and Chinese traditional opera share more similarities with each other than with Western music. Both traditions perform swaras and characters as sung syllables enriched with melodic ornaments, whereas in Western art music and Western pop music, notes are tied to specific scaled pitches (Brown 2017). Furthermore, in both Indian and Chinese music, melodic ornaments are defined by Raga or ShengQiang. This contrasts with Western music, where no such higher-level structure prescribes the use of melodic ornaments.

This comparison underscores the rich diversity and underlying commonalities in the approach to melodic structure across different musical cultures. It highlights how the integration of pitch, linguistic elements, and cultural context shapes the unique musical identity of each tradition, while also revealing shared principles that transcend cultural boundaries.

2.4.2 Types of Melodic Ornaments in Different Musical Cultures

Different musical cultures have developed unique sets of ornaments that reflect their distinct musical traditions and aesthetics. The following sections explore the types of

melodic ornaments found in Western pop music, Western art music, Indian Art Music, and Chinese traditional opera, highlighting the similarities and differences in how these cultures approach the art of ornamentation.

Melodic Ornaments in Western Music

It is well-recognised that vibrato is a common and important technique in Western classical singing, as reported by Seashore (1931). Vibrato, an Italian term initially meaning "vibration in pitch", is a musical effect employed in both vocal and instrumental music. However, this thesis will only discuss vocal vibrato as a vocal technique in Western music.

Though a clear definition of vibrato doesn't exist, as noted by Sundberg (1995*a*), Seashore (1931) described it as a "periodic oscillation in pitch" with a relatively stable rate and extent. Lee et al. (2011) emphasised that the major characteristic of vibrato tones is their periodic regularity, which can assume any arbitrary shape, not necessarily sinusoidal. Hence, rate, extent, regularity, and waveform are the four most critical characteristics of vibrato, elaborated in detail by Sundberg (1995*a*). The vibrato rate refers to "the number of undulations per second", while the extent, often indicated with plus and minus signs, shows "how far the phonation frequency diverges upwards and downwards from its average during a vibrato cycle". Regularity measures the similarity between each frequency fluctuation within its cycle. The vibrato waveform denotes the pitch contour shape, usually (but not necessarily) resembling a sine wave.

The function of vibrato in singing, particularly in Western music, remains an intriguing and complex subject. Despite extensive research, a definitive understanding of why singers use vibrato has been elusive, sparking debate for over a century. Various music psychology studies have investigated this issue, exploring vibrato's artistic and emotional significance. For instance, some research have examined vibrato as an expressive device (Howes et al. 2004, Seashore 1937), while others have focused on its physiological aspects (Fletcher et al. 2001, Seashore 1931, Sundberg 1995*a*). What is certain, however, is the importance of vibrato in characterising vocal styles. It has been identified as a key feature in the analysis of singing styles, in music style classification

and vocal style analysis by several studies (Caro Repetto et al. 2015, Panteli et al. 2017, Sundberg et al. 2012).

Trill is similar to vibrato but refers to a rapid alternation between two adjacent notes. Trills can be used to add an ornamental flourish to a piece of music. They require precise control of the vocal folds and strong breath support. Trills are a technique often found in opera and classical music.

Portamento is a significant and common melodic ornament in the singing voice, especially in Western music. The term encompasses a range of similar concepts, and understanding its precise definition requires careful consideration of various expert opinions.

Yang (2017) provides a comprehensive overview, defining portamento as a continuous slide through all intermediate pitches between two different notes. This definition draws on the insights of two prominent figures in the Western classical music field. Giovanni Battista Mancini, a professional soprano castrato and voice educator, described 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). García, another expert, defined it as "Slur (portamento) is to conduct the voice from one note to another through all intermediate sounds" (Garcia 1856).

In this thesis, the term "portamento" is used specifically when referring to a slide between two notes. In contrast, "pitch slide" or "pitch glide" are used when describing modifications to a single note where the slide occurs between the note and silence. These can be further categorised based on the specific position of a note and the direction of the slide, such as "scoop", which describes a pitch slide at the beginning of a note from a lower pitch, and "release," which refers to a pitch fall at the end of a note. To distinguish from "portamento", "glissando" refers to discrete, stepped glides across brief notes.

Two types of fine fluctuations in pitch slide are "overshoot" and "preparation", as discussed by Saitou et al. (2005). "Overshoot" refers to a transitional f0 that exceeds the target note just before settling on it (de Krom & Bloothooft 1995, Mori et al. 2004). This can be likened to an under-damped system, where the voice briefly overshoots the

target pitch. In contrast, "preparation" involves a pitch adjustment in the opposite direction of the following pitch slide.

These sliding pitch ornaments play a crucial role in expressing emotions. For example, Leech-Wilkinson (2006) emphasised the historical significance of portamento, suggesting that it draws on obligatory emotional responses to human sound, bringing a sense of comfort, sincerity, and profound emotion to the performance. Additionally, these sliding pitch ornaments are key in characterising the vocal style, as other studies have considered (Devaney 2011, Mayor et al. 2006, Yang 2017).

Grace notes are brief notes played right before a longer main note. They add decoration to the melody and harmony, and can be included or left out without changing the core structure of the music (Windsor et al. 2000). A mordent is a decorative element that instructs the performer to quickly alternate the main note with the note immediately above or below it during the note.

A more extended melodic ornament in duration is the "run", which is a quick sequence of notes that are sung in one breath, usually more elaborate than a simple scale. They can be used as a form of ornamentation or to show off a singer's vocal agility. They're often used in genres like pop, R&B, and gospel music. A well-executed run can add excitement and emotional expressiveness to a performance.

In summary, Western art music and Western pop music employ a variety of melodic ornaments like vibrato, trill, portamento, grace note and run to add expressiveness and complexity to musical pieces.

Melodic Ornaments in Indian Art Music

In Hindustani Classical music, ornamental pitch variations are referred to as *alankars*, encompassing ornaments such as *meend* (glide), *andolan* (oscillation), and *kan* (touch note). The *kan* lasts less than 300ms and is used to introduce "a slight pause on one or more intermediate notes or even a small stretch of low rate of pitch change" between two meends Datta et al. (2017). These are integral to vocal performances, each contributing unique stylistic nuances.

Meend, as described by Datta et al. (2017), is a continuous sliding pitch from one

melodic note to another, with a duration greater than 300 ms. It can be categorised into several types based on the shape of its pitch contour, ranging from straightforward to complex. The basic ones are smooth and unidirectional, either ascending or descending. More complex types combine both directions, and the third type, known as "the undulating meend", exhibits an up-down or wave-like movement. Meends can be further combined with kan, introducing touch notes between two meends. Datta explained that this combination may be caused by "a slight pause on one or more intermediate notes or even a small stretch of low rate of pitch change" (Datta et al. 2017).

Andolan, another significant alankar, is marked by a gentle, nuanced oscillation around a specific note. This oscillation reaches the boundaries of an adjacent note and touches the microtones or shrutis that lie between, creating a controlled and subtle swing that explores the pitch intervals between the notes. Within the context of Andolan, the specific note undergoing this oscillation is referred to as an andolit swar. It is essential to recognize that the application of these andolit swars is determined by the particular raga being performed, and they are not be used indiscriminately across different ragas (ITC Sangeet Research Academy 2008).

In summary, alankar serves as "a transitory segment which joins two steady segments smoothly," as observed by Ganguli & Rao (2015). Guided by the specific raga in which they are performed, alankars play a vital role in conveying not only style, but also personal characteristics and emotions, thereby adding depth and individuality to a performance.

Carnatic music, a prominent form of South Indian classical music, encompasses a diverse array of ornamental techniques collectively known as gamaka. Musicologists have identified two primary classification schemes for gamaka, consisting of either 10 or 15 types (Sambamoorthy 1958). The first scheme categorises gamakas based on the organization of note groups, often employing Western terms such as grace notes. In contrast, the second scheme emphasizes the pitch contour of the melodic ornamentation. Given that the focus of this thesis aligns more closely with the description of gamaka in the second scheme, it has been selected for further exploration. Among the 15

types of gamaka identified in Carnatic music, a distinction can be made between those related to loudness, instrumental execution and those performed through continuous pitch variations in singing. While all 15 types contribute to the overall aesthetic appeal of the music, it is specifically the gamakas involving continuous pitch variations that play a crucial role in characterising the vocal style of a raga (Rao et al. 2023).

Here are definitions of three types of continuous melodic gamakas in singing: 1. Kampita: A shake that delicately manipulates a note with such a restrained extent and precision that there is not even the slightest suggestion or hint of the adjacent notes. 2. Andolita: A sustained note that eventually glides to a higher note, executed with a free-swinging approach. 3. Ullasita: A glide in either upward or downward direction, transitioning smoothly between notes without emphasising the individuality of intermediate notes, creating a seamless connection.

In summary, Indian classical music uses alankars and gamakas to add stylistic nuances and emotional depth to performances.

Melodic Ornaments in Chinese Traditional Opera

In Chinese traditional opera, the categorisation of melodic ornaments is complex and lacks a standardised system. This complexity arises from the vast number of genres within Chinese traditional music, each with subtle differences in vocal expressions. Influenced by Chinese culture and language, the definitions of melodic ornaments are often subjective, with metaphorical names. It is not uncommon for a term to denote different ornaments in various genres or for the same ornament to have different names across genres or among different individuals.

In modern times, some musicologists have attempted to summarise and categorise these ornaments. Some have adhered to traditional definitions, while others have described the ornaments from a Western music perspective. However, no one has claimed that the system they have built is authentic and complete.

In this thesis, we refer to some literature within our knowledge and synthesise them to create a categorisation from the perspective of pitch contour. Below is a summary of several categories, covering all the common melodic ornaments documented in Chinese

traditional opera literature:

1. Pitch Fluctuations:

   (a) "Chanyin" ("颤音") involves oscillation around one note and the rate is slower than vibrato in Western classical singing (Shu 2018). It is an decorative approach of making a stable tone in the end of a phrase richer by variation (Wang 2011).

   (b) "Souyin" ("擞音") (Shu 2018), is normally after a main note, which serves a principal scale degree, and is generated by rapidly alternating once or twice with the adjacent note and characterised by the shaking or oscillating pitch, akin to a trill or mordent in Western music, but uniquely often transitioning from slow to fast (Miao et al. 1985).

2. Pitch Slide:

   (a) "Luoyin" ("落音"), also known as "Dunyin" ("顿音"), or "Huoyin" ("霍音"), means pitch drop in Qunqu, involving singing at a higher pitch followed by a subtle drop (Miao et al. 1985).

   (b) "Huoyin" ("豁音") in Qunqu, in contrast to "Luoyin", involves singing at a lower pitch followed by a rise. The rise part of the tone has an interval typically with a major second or minor third (one pitch step in a pentatonic scale) and form a very short note (Miao et al. 1985).

   (c) "Huaqiang" ("滑腔"), meaning pitch slide, connects two notes in a descending direction, similar to descending portamento in Western music (Shanghai Art Research Institute & Shanghai Branch of the Chinese Dramatists Association 1981).

3. Pitch Short Break:

   (a) "Duanyin" ("断音" or "duàn yīn") in Qunqu, involves singing the first note for an extremely short duration, introducing a very short rest, and then turning to other notes. This short-burst singing method adds a unique rhythm to the performance (Miao et al. 1985).

Dong (2004) explained the function of melodic ornaments in Chinese traditional opera. Primarily, melodic ornaments are realized to shape the pitch contour of the sung Chinese character to meet the "依字行腔" ("YiZiXingQiang") principle (singing according to the tone of syllables), see Section 2.4.1. For example, "Huoyin" ("霍音") is commonly applied to characters with Qu tone (falling), "Huoyin" ("豁音") is primarily used in characters with Shang tone (rising), and "Duanyin" ("断音") is primarily used in characters with Ru tone (Miao et al. 1985). Besides, melodic ornaments reflect the vocal style defined by the ShengQiang or school and the singer's expression. Overall, these ornaments contribute to the rich and intricate soundscape of Chinese traditional opera, each adding its unique flavour to the performance.

In summary, Chinese traditional music employs a complex set of melodic ornaments, influenced by linguistic and cultural factors, to create a rich and intricate soundscape.

Synthesis of Melodic Ornaments Across Cultures

The exploration of melodic ornaments in Western art, Western pop, Indian art, and Chinese traditional opera reveals common underlying structures that can be categorised into several pitch contour patterns. These patterns highlight the universality of musical expression across diverse cultural contexts:

1. **Pitch Oscillations Around One Note:** These ornaments involve periodic fluctuations in pitch around one note.
   (a) **Vibrato (Western):** A periodic oscillation in pitch with a relatively stable rate and extent (Sundberg 1995*a*).
   (b) **Andolan (Indian Hindustani):** A gentle, nuanced oscillation around a specific note, extending to the periphery of an adjacent note and engaging the microtones or *shrutis* in between (ITC Sangeet Research Academy 2008).
   (c) **Kampita (Indian Carnatic):** A delicate shake that manipulates a note without hinting at adjacent notes (Rao et al. 2023).
   (d) **Chanyin (Chinese):** Similar to vibrato but richer in variation, with various ways the amplitude and rate can vary (Wang 2011).

2. **Rapid Alternation Between Notes:** These ornaments involve a rapid back-and-forth movement between two distinct pitches.

   (a) **Trill (Western):** Rapid alternation between two adjacent notes.

   (b) **Souyin (Chinese):** Shaking or oscillating pitch, characterised by the alternation between two notes, akin to a trill in Western music, but uniquely often transitioning from slow to fast (Miao et al. 1985).

3. **Sliding Pitch Ornaments Involving a Single Note:** These ornaments add subtle inflections to individual notes, enhancing expressiveness.

   (a) **Scoop (Western):** "Scoop" describes the pitch slide at the beginning of a note from a lower pitch.

   (b) **Release (Western):** "release" refers to the pitch fall at the end of a note.

   (c) **Huoyin (Chinese):** "Huoyin" ("豁音") involves singing at a lower pitch followed by a rise (Miao et al. 1985).

   (d) **Luoyin (Chinese):** "Luoyin" ("落音") describes a pitch drop involving singing at a higher pitch followed by a subtle drop (Miao et al. 1985).

4. **Simple Sliding Pitch Ornaments Between Two Notes:** These ornaments involve a continuous movement between notes, creating a smooth and connected sound, but each has its unique characteristics.

   (a) **Portamento (Western):** A continuous slide through all intermediate pitches between two distinct notes (Yang 2017).

   (b) **Ullasita (Indian Carnatic):** Similar to portamento in definition, but exhibiting a distinct pitch curve.

   (c) **Huaqiang (Chinese):** A pitch slide connecting two notes in a descending direction, analogous to descending portamento in Western music (Shanghai Art Research Institute & Shanghai Branch of the Chinese Dramatists Association 1981).

5. **Complex Sliding Pitch Ornaments:** These ornaments are characterised by more intricate and nuanced movements, often involving variations in direction, shape, or additional notes.

   (a) **Glissando (Western):** Refers to discrete, stepped glides across notes.

   (b) **Overshoot and Preparation (Western):** Encompasses bending of pitch curve, consisting of one upward and one downward slide connected with each other.

   (c) **Run(Western):** A quick sequence of notes sung in one breath, usually more elaborate than a simple scale. Unlike oscillations or rapid alternations, a run moves fluidly through a series of notes, creating a flowing and connected sound.

   (d) **Andolita (Indian Carnatic):** A sustained note that eventually glides to a higher note, executed with a free-swinging approach, allowing for more expressive and personalised interpretation. The direction is customarily upward with no specific restriction on the shape of pitch contour.

   (e) **Meend (Indian Hindustani):** Continuous sliding pitch from one melodic note to another, with variations such as smooth and unidirectional glides, attached touch notes, or complex undulating movements, covering all the possible shapes of pitch slides as defined above (Datta et al. 2017).

6. **Short Note in Melody**

   (a) **Grace note (Western):** Grace notes are brief notes played right before a longer main note. They add decoration to the melody and harmony, and can be included or left out without changing the core structure of the music (Windsor et al. 2000).

   (b) **Kan (Indian Hindustani):** Known as a touch note, this ornament lasts less than 300ms and is used to introduce “a slight pause on one or more intermediate notes or even a small stretch of low rate of pitch change” between two meends, thereby emphasising a brief connection and adding a subtle complexity to the melody (Datta et al. 2017).

(c) **Huoyin (Chinese):** "Huoyin" ("豁音") in Qunqu forms a very short note after a pitch rise (Miao et al. 1985).

(d) **Duanyin (Chinese):** Duanyin ("断音"), in Qunqu, contrasts with Kan by involving an extremely short duration of the first note, followed by a very short rest, and then a transition to other notes. While Kan focuses on connection, Duanyin introduces a distinct rhythm through a brief interruption (Miao et al. 1985).

Exploring musical context and melodic ornaments across Western art music, Western pop music, Indian art music, and Chinese traditional opera reveals a fascinating interplay of universality and diversity. While the basic units of melody vary (notes, swaras, characters), the application of ornaments draws upon a shared vocabulary of pitch contour patterns. This comparison highlights the expressive power of melodic ornamentation and suggests a degree of universality in how music manipulates pitch to create beauty and meaning.

2.5 Computational Modelling of Melodic Ornaments for Vocal Style Understanding

As elaborated in Section 2.4, the understanding of vocal styles through melodic ornaments is deeply rooted in the musical context, which comprises two main elements: musical form and basic musical units. To computationally model these ornaments, this thesis outlines a multi-step approach: 1) recognition of the musical form, 2) estimation of the pitch trace, 3) note-level and syllable-level transcription, 4) detection of pitch contour elements, 5) melodic ornament labelling, and 6) characterisation of pitch contour segments through computational models.

2.5.1 Challenges and Opportunities in Human and Computational Approaches

Understanding vocal music is a multifaceted endeavour that involves transcription, representation, and analysis. Both manual and computational approaches have their own

sets of advantages and limitations. This section aims to dissect these methods in terms of the three aforementioned aspects, with the goal of highlighting how a hybrid approach can offer a more comprehensive understanding of vocal styles.

Manual approaches to music transcription, representation or visualisation, and analysis are predominantly utilised by musicologists who prioritise traditional methods. These approaches are deeply rooted in the expertise of vocal and instrumental styles and rely significantly on musical notation for transcription. Although this traditional method offers a comprehensive and nuanced understanding of musical compositions, it is not without its limitations:

- **Transcription**: The manual transcription requires musical training and careful work to achieve accurate results. It is inherently subjective, with potential biases and inconsistencies due to the transcriber's personal experience and interpretation of music.
- **Representation**: While traditional musical notation provides a historic and detailed method for representing music, it may not capture the full intricacies of every performance, such as microtonal variations or the subtle dynamics within live performances.
- **Analysis**: The qualitative nature of manual analysis offers deep insights but it is subjective. It might also present challenges in scalability and objective comparison, particularly in large-scale or cross-cultural studies, potentially limiting its applicability in broader research contexts.

The advent of computational methods, supported by advances in Music Information Retrieval (MIR) technologies, presents several advantages over traditional manual approaches. These computational techniques are revolutionising the way music is transcribed, visualised, and analysed:

- **Transcription**: Computational methods offer a swift and efficient process for transcribing music. They can handle large volumes of data with a high degree of objectivity and consistency, mitigating subjective biases inherent in manual transcription.

- **Visualisation**: With the use of sophisticated software, computational approaches provide advanced representation interfaces. These can display a broader array of musical elements in detail, offering insights beyond what is possible through traditional notation alone.

- **Analysis**: The quantitative analysis enabled by computational methods supports more systematic and scalable research. It facilitates large-scale and cross-cultural studies, promoting a broader and more inclusive understanding of music across different regions and styles.

However, despite these advantages, it is crucial to recognise that computational methods may not fully replicate the nuanced understanding and interpretive depth provided by skilled human experts, particularly in areas where cultural context and emotional expression are key. As such, musicological studies may benefit from a hybrid approach, combining the strengths of both manual and computational methodologies to achieve a more complete and multi-faceted understanding of music.

To amalgamate the strengths of both manual and computational methods, several hybrid approaches have been developed:

- **Transcription**: Software like Tony allows for the manual correction of automated transcriptions, combining speed with human nuance (Mauch et al. 2015).

- **Visualisation and Representation**: Efforts like Dunya (Porter et al. 2013) and Global Notation System (Killick 2020) offer interfaces that merge computational detail with musical symbols that are intuitive to musicologists.

- **Analysis**: Hybrid methods can incorporate human annotations or corrections into computational models, offering a balanced approach for in-depth studies.

This thesis will explore transcription and representation in greater depth within the subsequent segments of this section, and will examine analysis in detail in Section 2.6.

2.5.2 Musical Form Recognition

While automatic methods for musical form recognition, such as raga recognition (Sharma & Salgaonkar 2023) and scale detection (Kawase 2017), have been developed, they often fall short of the reliability and authenticity offered by expert annotations. Consequently, this thesis will primarily rely on expertly annotated musical forms. Nonetheless, the utility of computational techniques, such as pitch histograms for indicating scales, is acknowledged as a valuable resource for researchers who may not be well-versed in specific musical traditions.

2.5.3 Pitch Estimation

The field of computational pitch estimation for monophonic sounds has seen significant advancements over the past half-century, with a multitude of methods being developed to improve accuracy. A comprehensive review of these techniques was provided by Kim et al. (2018). Early approaches to pitch estimation commonly employed specific mathematical functions to generate candidate pitch values. These were often accompanied by pre-processing and post-processing steps to refine the resulting pitch curve. Among the functions used in these early methods are the cepstrum (Noll 1967), the autocorrelation function (ACF) (Dubnowski et al. 1976), the average magnitude difference function (AMDF) (Ross et al. 1974), and the normalised cross-correlation function (NCCF) as introduced in RAPT (Talkin 1995) and PRAAT (Boersma 1993). Another noteworthy method is the cumulative mean normalised difference function, which was proposed for YIN (De Cheveigné & Kawahara 2002).

In more recent years, advanced techniques have emerged that leverage modern computational capabilities. For instance, SWIPE (Camacho & Harris 2008) employs template matching with the spectrum of a sawtooth waveform. Another example is PYIN (Mauch & Dixon 2014), a variant of YIN that incorporates a Hidden Markov Model (HMM) to decode the most probable sequence of pitch values. With the advent of deep learning, CREPE, which utilises a deep convolutional neural network, has established itself as the state-of-the-art open-source pitch extractor (Kim et al. 2018).

While physiological techniques like electroglottography (EGG) offer highly accu-

rate f0 measurements through direct monitoring of vocal fold activity (Howard 1993), their requirement for specialised hardware during recording makes them unsuitable for analysing existing audio recordings. This thesis therefore focuses on computational pitch estimation from audio signals alone.

2.5.4 Note-Level and Syllable-Level Transcription

This thesis concentrates on three fundamental units: notes, swaras, and Chinese characters, as detailed in subsection 2.4.1. Transcription in this context entails the delineation of the temporal boundaries for each basic unit and the subsequent labelling of these segments. This subsection is dedicated to discussing the advantages and drawbacks of existing methods for transcribing both notes and syllables, which include swaras and Chinese characters.

Note Transcription Methods

Automatic note transcription refers to converting an acoustic waveform into musical notes. While monophonic instrument transcription is often considered to be a solved problem in music information retrieval (Benetos et al. 2013), this is not the case for singing, where pitch is rarely stable (Dai & Dixon 2019). Even when singers aim to maintain a steady pitch, the f0 shows small fluctuations rather than remaining perfectly constant. Numerous note segmentation methods have been proposed. Early singing transcription systems (Clarisse et al. 2002, De Mulder et al. 2004, Haus & Pollastri 2001, McNab et al. 1995) implemented simple rule-based methods based on pitch or amplitude variations and the presence of vocal activity. Taking advantage of hidden Markov models (HMMs), more robust systems were then proposed (Mauch et al. 2015, Ryynänen & Klapuri 2004, Viitaniemi et al. 2003) that rely on similar musical features. However, these methods perform poorly on soft onsets and offsets, pitch oscillations within notes (such as vibrato and other expressive modulations) and glides between temporally adjacent pitches.

"Soft" onsets and offsets occur when two adjacent notes are smoothly connected without obvious loudness variations. In most cases, however, there is a phonetic change

between notes. Various spectral features have been used to detect timbre changes, either by selecting as boundaries peaks above a threshold in the measure of timbre change (Gómez & Bonada 2013, Yang, Maezawa, Smith & Chew 2017), or by modelling vowels and their transitions using an HMM (Heo & Lee 2017, Hsuan-Huei Shih et al. 2002). In Chapter 4, a PYIN (Mauch & Dixon 2014) variant (Li et al. 2021), taking phonemes extracted by a state-of-the-art automatic lyrics transcription system (Demirel et al. 2020) as an input, made a positive contribution on this task.

Additionally, pitch fluctuations within notes or pitch glides between notes, whether intentional or not, cause some notes to be separated mistakenly into multiple notes. To address the within-note fluctuation problem, Molina et al. (2015) used hysteresis of pitch and dynamic averaging to avoid the effects of small or short pitch deviations. Yang, Maezawa, Smith & Chew (2017) proposed a pitch dynamic model to address problems with pitch variation.

In recent years, deep neural networks (DNNs) have significantly advanced the field of vocal note transcription. These DNN-based methods often set new benchmarks, eclipsing previous state-of-the-art performances. For instance, Nishikimi et al. (2019) employed an attention-based encoder-decoder network with long short-term memory (LSTM) modules. Fu & Su (2019) enhanced their models by incorporating onset- and offset-related features. Wang et al. (2022) innovatively applied object detection techniques, fine-tuning a pre-trained model with their sight-singing dataset (SSVD), to markedly improve singing voice onset/offset detection. Yong et al. (2023) designed a neural network architecture that leverages a convolutional recurrent neural network (CRNN) backbone and phonetic posteriorgram (PPG) to achieve state-of-the-art performance on two datasets, ISMIR2014 (Molina, Barbancho, Tardón & Barbancho 2014) and SSVD version 2.0 (Wang et al. 2022).

However, these advancements are not without challenges. One major issue is the limited scale of existing annotated datasets for training and testing, primarily due to the labor-intensive nature of manual annotation. Wang & Jang (2021) attempted to address this by creating the MIR-ST500 dataset, comprising over 160,000 notes from 500 pop songs. Yet, the dataset's reliability is questionable as non-experts performed

the annotations, and its genre is restricted mainly to Chinese pop songs. Gu et al. (2023) proposed a self-supervised learning approach adapted from the speech domain, which reduces the need for annotated data but suffers from poor note offset detection.

Moreover, the DNN-based methods often overlook the nuanced challenges intrinsic to vocal note transcription. They tend to focus on outperforming previous methods on well-known datasets, without addressing the subjectivity and context-dependency of note transcription. Factors such as cultural background, perceptual sensitivity, and transcription purposes can influence how different individuals transcribe the same piece of music. Current DNN models lack the flexibility to adapt to these varying contexts, as they are trained on annotations from a limited number of individuals and datasets.

In summary, despite the rapid advancements in automatic note transcription methods, they still fall short of human expertise, as reported by Ozaki et al. (2021). Furthermore, computational metrics used in Music Information Retrieval (MIR) only partially align with human expert assessments (Holzapfel et al. 2022). Consequently, for most musicologists, computer-aided manual note transcription remains the most reliable approach. Tony software (Mauch et al. 2015) is a popular tool for this task, offering an interface that displays both the audio waveform and the pitch trace estimated by PYIN (Mauch & Dixon 2014). Users can perform note transcription with the aid of this visualisation and can validate their annotations by playing the audio, pitch track, and notes both simultaneously and separately.

Syllable Transcription Methods

Swara and Chinese characters share a common property: they can both be considered as syllables in terms of pronunciation. While there are specialised transcription methods for swara, as cited in Singh et al. (2023), this section will focus on general syllable transcription methods. This is because the thesis primarily relies on manual annotations by experts for both swara and Chinese character transcription.

Transcribing syllables in singing presents unique challenges, as sung syllables differ from spoken syllables in both pitch and rhythm. Gong & Serra (2018) tackled this issue by proposing a two-step, language-independent method that utilises a convolutional

neural network (CNN) and a duration-informed hidden Markov model (HMM). Their model was trained on a Jingju dataset, which they annotated themselves at both the phoneme and syllable levels.

Recently, the broader task of lyrics transcription has gained attention. This involves outputting phonetic labels at multiple levels, from phonemes to words, and can also provide syllable segments. Demirel et al. (2021) developed a phoneme-level lyrics transcription system for English-language singing. However, their model was trained and evaluated exclusively on English singing datasets, making it less suitable for cross-cultural music analysis. This limitation is particularly relevant to my research, which focuses on analysing vocal music across different cultural and linguistic traditions. The model's dependence on English phoneme sets and language-specific features makes it inadequate for analysing music from non-Indo-European languages, such as Chinese, where different phonological systems and tonal features play crucial roles in vocal expression.

The most recent advancement comes from Zhuo et al. (2023), who integrated two major AI breakthroughs: Whisper (Radford et al. 2023), a robust automatic speech recognition (ASR) model, and GPT-4 (OpenAI 2023), a powerful text-based natural language processing (NLP) model. Whisper serves as the "ear," transcribing the singing into text, while GPT-4 acts as the "brain," selecting and correcting the output based on context. This model's flexibility and generality make it well-suited for transcribing Indian swaras and Chinese characters in vocal music. However, the word error rates reported in their study indicate that AI performance on this task is still far from human-level accuracy.

For most musicologists, manual annotation remains the most reliable method for syllable transcription. Praat (Boersma 1993), a well-known phonetic software, is commonly used for this purpose. The software interface allows users to observe the audio's spectral patterns and waveform, play the sound, and annotate each segment with boundaries and phonetic labels.

2.5.5 Detection of Pitch Contour Elements

As previously discussed in Section 2.4.2, melodic ornaments across different musical cultures often share underlying pitch contour patterns. Automatic detection of these patterns is foundational for identifying melodic ornaments. This task is challenging due to the intricate shapes of pitch contours. However, this thesis simplifies these patterns into three essential elements for easier machine detection: steady, modulating and transitory. The following sections elaborate on these categories and review existing methods for their detection.

Steady Elements

Steady elements, also referred to as "sustained" or "stable" regions, in pitch trajectories are segments where pitch values remain within a small range around a mean value for at least 50ms. These regions contribute to conveying the tonality and melody of a musical piece. Various methods have been developed to automatically detect these steady elements in vocal pitch.

Koduri et al. (2012) employed the local slope of the pitch trajectory to identify steady regions. Datta et al. (2017) used a more nuanced approach, calculating the deviation between the current frame's pitch and the mean pitch of the preceding steady element, setting a minimum duration of 60ms for a element to be considered steady. Molina, Tardón, Barbancho & Barbancho (2014) utilised pitch chroma contour and its moving average for stable note change detection. Ganguli & Rao (2018) took a global approach, used pitch histograms to identify scale intervals and approximate steady regions ($\pm$35 cents, 250 ms). However, this method has limitations for vocals with significant pitch drift. Mauch et al. (2015) used a Hidden Markov Model (HMM) and PYIN (Mauch & Dixon 2014), a pitch extractor, to identify stable regions where pitch values deviate minimally from a centre pitch. However, this method relies on twelve-tone equal temperament, which limits its accuracy when analysing music using other tuning systems. Rosenzweig et al. (2019) developed two methods for Georgian vocal music, which do not adhere to the any tuning system, using morphological operations and binary time-frequency masks.

In summary, while these methods have been effective for specific datasets and research goals, they often lack generality. Most are rule-based and rely on one or two thresholds, making them less versatile. Additionally, many are constrained by musicological assumptions like musical scale and octave equivalence.

Modulating Elements

In the realm of pitch analysis, the second category of interest is regions of modulation or undulation. In these regions, the frequency of the pitch signal varies according to a secondary signal, such as a sine-wave, to create a vibrato effect (Wen & Sandler 2008). Previous studies listed below considered that modulating regions are synonymous with vibrato regions and define vibrato as a pitch oscillation around a central pitch, with a specific rate range (e.g., fmin = 4Hz, fmax = 9Hz for singing voice reported by Prame (1994)) and a minimum duration threshold.

Vibrato extraction methods can be broadly categorised into two classes: spectrum-based and f0-based methods. Spectrum-based methods (Driedger et al. 2016, Regnier & Peeters 2009, Rossignol et al. 1999) directly analyse the audio spectrum and are advantageous when dealing with polyphonic music, as they are less prone to errors in f0 estimation. On the other hand, f0-based methods excel in monophonic settings where f0 can be accurately estimated (Driedger et al. 2016). Given that this research focuses solely on monophonic audio, the subsequent discussion will centre on f0-based methods.

These f0-based methods themselves are divided into note-wise and frame-wise approaches. Note-wise methods (Ozaslan & Arcos 2011, Pang & Yoon 2005, Rossignol et al. 1999, Weninger et al. 2012) start by segmenting the audio track into individual notes and then detect vibrato within each note. This approach is beneficial as it avoids merge errors, where two distinct vibratos could be mistakenly identified as one. However, the segmentation process can be either time-consuming if done manually or inaccurate if automated. Assuming that an ideal vibrato closely resembles a sinusoidal shape, frame-wise methods decompose the f0 into sinusoids by estimating the frequency and amplitude of sinusoid components frame by frame. These methods either employ

Short-Time Fourier Transform (STFT) (Herrera & Bonada 1998, Nakano et al. 2006, Ventura et al. 2012, Von Coler & Roebel 2011) or parametric fitting techniques (Pang & Yoon 2005, Yang, Rajab & Chew 2017). STFT-based methods face the limitation of the Fourier Transform's uncertainty principle, which imposes a trade-off between temporal and frequency resolution. Parametric fitting methods, however, can achieve high frequency resolution by decomposing the f0 signal into a predefined set of sinusoids, thus avoiding error-prone peak picking.

Among these, one of the most advanced methods was proposed by Yang, Rajab & Chew (2017), who used the Filter Diagonalisation Method (FDM) for frequency and amplitude estimation and employed either a Decision Tree (DT) or Bayes' Rule (BR) for vibrato decisions. Despite its high frequency resolution, this method has significant limitations. It lacks flexibility due to its reliance on pre-defined, empirically set thresholds for vibrato detection. Moreover, by making frame-by-frame decisions, it overlooks the inherent regularity of vibrato, which is a crucial characteristic as a time-series pattern. To address these issues, there is a need for methods that can capture the time-series nature of vibrato effectively. However, the development of such methods appears to have stagnated in recent years, potentially due to the research community's emphasis on employing Deep Neural Networks (DNNs) for higher-level tasks.

Transitory Elements

Transitory elements in pitch contours differ fundamentally from steady and modulating elements in that they lack clear regularity. In the scope of our research, transitory elements are best defined as pitch contours that are neither modulating nor steady. These regions often serve expressive functions in singing and are the subject of several studies aimed at automatic detection and analysis within specific musical traditions.

The concept of a transitory pitch contour was first introduced by Indian musicologists. Ganguli & Rao (2015) and Datta et al. (2017) both approached the identification of transitory regions by first removing all detected steady segments from the pitch contour. While Ganguli focused on raga recognition, Datta specifically investigated "meends," categorizing them based on their shape and defining them as a subset of

transitory regions.

Methods for identifying transitory regions in other musical traditions are more complex, as they must also distinguish between modulating and transitory pitch contours. In Western music, transitory regions can correspond to various melodic ornaments like portamento, pitch slide, pitch release, and glissando. Yang et al. (2016) employed a HMM-based method to detect these ornaments after removing vibratos using a method described in Yang (2017). Gong et al. (2016) explored transitory regions in Jingju music, aiming to assess the similarity of pitch contours between teachers and students. Unlike previous methods that detected different pitch contour types separately, Gong's approach employed the standard deviation of the cumulative differences of local extrema (StdCdLe) as the criterion to segment the pitch contour and label the pitch contour as steady, vibrato, and transitory regions using a K-Nearest Neighbor (kNN) classifier.

The limitations of these methods are noteworthy. The first two methods, designed specifically for Hindustani music, do not account for vibrato and are not easily adaptable to other musical traditions. Yang et al. (2016) requires a pre-processing step to remove vibratos, which is error-prone and could compromise the detection of portamento. Moreover, it does not consider steady regions, limiting its applicability for detecting complex transitory regions with touch notes. Although Gong et al. (2016) addresses some of these limitations, its segmentation performance is less than satisfactory, with an accuracy rate below 40%.

2.5.6 Melodic Ornament Labelling

Based on three basic pitch contour elements: steady, modulating, and transitory, this thesis provides a systematic categorisation of melodic ornaments as detailed in Table 2.2.

Various systems have been developed to label ornaments based on specific characteristics observed in the corresponding pitch segments. Although steady elements are straightforward and offer limited scope for variation, Mayor et al. (2006) identified a specific expression within steady segments in Western pop singing performances.

Pitch contour elements	Ornaments
Steady	• Simple sustained notes longer than 100ms • "Fall-down" • Touch note within complex meend (Indian Hindustani) or glissando (Western)
Modulating	• Vibrato (Western) • Andolan (Indian Hindustani) • Kampita (Indian Carnatic) • Chanyin (Chinese) • Trill (Western) • Souyin (Chinese)
Transitory	**Simple transitory:** • Scoop, Release (Western) • Huoyin, Luoyin (Chinese) • Portamento (Western) • Ullasita (Indian Carnatic) • Huaqiang (Chinese) **Complex transitory:** • Glissando (Western) • Overshoot and Preparation • Run (Western) • Andolita (Indian Carnatic) • Meend (Indian Hindustani)

Table 2.2: Basic pitch contour elements and their corresponding melodic ornaments

Termed as "fall-down," this ornament signifies a gradual lowering of pitch during the sustain phase. The label is derived from empirical definitions crafted by the authors. The label is derived from empirical definitions crafted by the authors. Additionally, brief steady elements lasting less than 50ms within complex ornaments such as meend or glissando are labelled as touch notes (Datta et al. 2017).

Modulating elements can manifest various melodic ornaments, but most existing methods focus solely on detecting vibrato. Other ornaments like trills, andolan (oscillations), and "Souyin" ("撽音") are often considered variations of vibrato, without any specialised methods to distinguish them.

Transitory elements are more complex and can be labelled with a variety of melodic ornaments. Yang, Rajab & Chew (2017) can directly detect pitch slides but does not differentiate them based on their position within a note or their shape. In contrast, Mayor et al. (2006) developed a system for Western singing performances that labels position-related sub-level note segments—such as attack, release, and transition—with specific melodic ornaments like scoop up, scoop down, portamento up, and portamento down.

The most intricate melodic ornaments may consist of multiple pitch slides with varying directions and shapes, or even a combination of transitory, steady, and modulation regions. These complex ornaments are especially common in Indian art music and Chinese traditional opera. Despite their prevalence, automated labelling methods for these ornaments remain in their early stages. This is primarily due to a lack of consistent naming conventions in musicology. However, researchers are starting to bridge this gap with computational methods. For example, a method has been developed to automatically categorise 'meends' in Indian art music based on pitch contours (Datta et al. 2017). This type of approach offers a systematic way to categorise and label complex ornaments, supporting further development of automated analysis techniques. In conclusion, the diversity and complexity of melodic ornaments underscores the need for more comprehensive and nuanced labelling methods.

2.5.7 Characterising the Pitch Contour Segment through Models

The objective of characterising pitch contour segments is to quantify specific features that reflect their shape. This is typically achieved by constructing a mathematical model that serves as a simplified representation of the pitch contour, making it more manageable for analysis (De Cheveigne 2005). In essence, the model is defined by the features we aim to measure, and these features are represented through parameters that can be mathematically derived. Such models are often expressed as mathematical functions that capture the relationships between variables in the system. The parameters of these models are determined either through predefined rules or by data fitting techniques that minimise a particular loss function.

While Deep Neural Networks (DNNs) have gained popularity for modelling complex systems, their large number of parameters often results in low interpretability despite high performance. In contrast, the models used for pitch contour segments only need a few parameters to adequately capture the overall shape of the contour, which is enough to distinguish the vocal style. Although various types of melodic ornaments have been identified in the 2.4.2, prior research has mainly focused on a limited set of commonly occurring shapes.

Characterising the Pitch Contour of Vibrato

In this thesis, vibrato is considered a specific type of modulating pitch contour, characterised by four key features: rate, extent, regularity, and waveform (for details, see Section 2.4.2). The waveform is commonly assumed to be sinusoidal, based on observations that real-world vibratos often exhibit quasi-sinusoidal shapes (Sundberg 1995*a*). Consequently, numerous studies have modelled the fundamental frequency (f0) of vibrato as a sinusoid. While most of these studies, such as Dai & Dixon (2016), focus primarily on rate and extent, only a few, like the works of Wen & Sandler (2008) and Yang (2017), also consider regularity.

Wen & Sandler (2008) proposed a method to decompose the original f0 into a smooth component (the carrier) and a vibrating component (the modulator). This involved observing complete f0 modulation cycles and calculating an average frequency

for each cycle to construct a vibrato-free frequency track. The modulator was then obtained by subtracting this smooth component from the original signal. From the modulator, they measured the vibrato's regularity, rate, and extent. Regularity was quantified using the maximum value of the autocorrelation coefficient, excluding the value at time zero. Rate was estimated by calculating an overall modulation rate that maximized regularity. Two features related to extent were measured: maximal pitch departure from the pitch centroid and average pitch departure, calculated as the root-mean-square of the modulator.

In contrast, Yang et al. (2013) estimated rate and extent directly from the original f0, reserving the decomposed modulator solely for calculating regularity. Yang assumed that the interval between one peak and one trough represents a half cycle of the vibrato, and the overall rate and extent were calculated as the average across these half cycles. Additionally, Yang measured the envelope of the vibrato f0 contour to capture the evolution of extent. This was achieved by taking the absolute value of the analytic signal obtained from the Hilbert transform of the vibrato f0 contour, followed by moving average post-processing. Finally, regularity was assessed by calculating the normalised cross-correlation between the modulator and a relevant sine wave, thereby quantifying how closely the shape of the vibrato resembles a sinusoid.

Characterising the Pitch Contour of Portamento

To the best of our knowledge, only one study has developed a model to characterise the pitch contour of portamento. Yang (2017, Section 4.1) employed a logistic model to capture the f0 contour of portamento with S shape, which suggests that a portamento involves both an acceleration phase and a deceleration phase during its execution. The logistic model is represented by Equation 2.1:

$$p(t) = L + \frac{(U - L)}{\left(1 + Ae^{-G(t-M)}\right)^{1/B}} \tag{2.1}$$

Here, L and U denote the initial and final pitches of the transition, while A, B, G, and M are constant parameters. G can be interpreted as the rate of growth, indicating the steepness of the transition's slope. These parameters were estimated numerically

using Matlab's Curve Fitting Toolbox, employing the non-linear least squares method for optimisation.

Five key features were identified to describe the shape of the portamento f0 contour:

1. The slope of the transition, represented by the coefficient G in Equation 2.1.

2. The transition duration, estimated by measuring the duration of the continuous region where the first derivative of the logistic curve exceeds a threshold of 0.861 semitones per second, based on empirical data.

3. The transition interval, calculated as the absolute semitone difference between the initial and final pitches.

4. The normalised inflection time, which is the time at which the slope reaches its peak. This is calculated using Equation 2.2 and standardised to fall within a range of 0 to 1.
$$t_R = -\frac{1}{G} \ln\left(\frac{B}{A}\right) + M \tag{2.2}$$

5. The normalised inflection pitch, standardised to fall within a range of 0 to 1, where 0 corresponds to the lower asymptote and 1 to the upper asymptote within the transition interval.

In addition to the logistic model, alternative models like Polynomial, Gaussian, and Fourier Series were also tested. Their curve-fitting performance was found to be inferior to the logistic model in terms of portamento with S shape, as evaluated by Root Mean Squared Error (RMSE) and Adjusted R-Squared values. However, the logistic model would not be the best choice for portamento with other shapes.

Characterising the Pitch Contour of the Pitch Slide

Different with portamento, pitch slide, in this thesis's definition, does not have an antecedent or subsequent sustained pitch for a duration at least 0.1s. In these situations, the logistic model which is characterised with an S shape would fail to fit the signal. However, several studies utilised different models to characterise the pitch slide.

Dai & Dixon (2016), for the purpose of synthesising stimulus and characterising the sung pitch contour in imitation, used first-order and second-order polynomial functions for pitch ramps and initial or final pitch slides. We summarise the detailed explanations from chapter three of Dai (2019).

Below are the three distinct types, each with its own mathematical model for characterisation:

1. **Initial Pitch Slide:** This type of pitch slide starts with an initial quadratic pitch glide and transitions into a constant pitch. The mathematical model is given by:

$$p(t) = \begin{cases} at^2 + bt + c, & 0 \leqslant t \leqslant d \\ p_m, & d < t \leqslant 1 \end{cases} \tag{2.3}$$

2. **Final Pitch Slide:** Here, a constant pitch is followed by a final quadratic pitch glide. The equation for this model is:

$$p(t) = \begin{cases} at^2 + bt + c, & 1 - d \leqslant t \leqslant 1 \\ p_m, & t < 1 - d \end{cases} \tag{2.4}$$

3. **Pitch Ramp:** This is a linear pitch ramp, modeled by the following equation:

$$p(t) = p_m + p_D \times (t - 0.5), \quad 0 \leqslant t \leqslant 1. \tag{2.5}$$

In these equations, the duration is normalized to 1 second. The models use three key variables to characterize the pitch slide:

- p_m represents the main or central pitch.
- d denotes the duration of the transient part of the stimulus.
- p_D is the extent of pitch deviation from p_m.

Additional parameters a, b, and c are decided based on specific conditions. For example, these parameters are determined such that the curve passes through the

points $(0, p_m + p_D)$ and (d, p_m), and has its vertex at (d, p_m). The values are given by $a = p_D/d^2$, $b = -2 \times p_D/d$, and $c = p_m + p_D$.

For parameters p_m, d, and p_D, the estimation is performed numerically through curve fitting. For both initial and final pitch slides, a grid search is conducted to find the breakpoint of the piecewise function, as represented by Equations 2.3 and 2.4. The optimal parameters are those that minimize the mean square error. Once the breakpoint is determined, the two segments of the piecewise function can be estimated using regression methods. For pitch ramps, the parameter p_m is calculated as the median pitch over the middle 80% of the duration, and a linear regression is used to model the slope as given by Equation 2.5.

Devaney (2011) employed the type-II Discrete Cosine Transform (DCT) as an alternative to polynomial models for characterising pitch slides. The DCT method has the advantage of providing multiple, independent coefficients, with the 0th, 1st, and 2nd coefficients specifically corresponding to the mean, slope, and curvature of the pitch slide, respectively.

$$\begin{aligned} y(k) &= \omega(k) \sum_{n=0}^{N-1} x(n) \cos \frac{k(2n+1)\pi}{2N} \\ \text{where } \omega(k) &= \begin{cases} \frac{1}{\sqrt{N}}, & k = 0 \\ \sqrt{\frac{2}{N}}, & 1 \leq k \leq 2 \end{cases} \end{aligned} \tag{2.6}$$

In Equation 2.6, the variable x denotes the input signal, N specifies its length, and n serves as the index for each sample in the signal. Coefficients can be calculated according the equation 2.6. Specifically, the 0th DCT coefficient represents the mean value of the signal, normalised by the square root of N, the total number of samples. A positive DCT coefficient indicates a negative slope, while a negative DCT coefficient indicates a positive slope. The 2nd DCT coefficient provides insights into the curvature of the pitch slide, detailing both its magnitude and direction（concave is negative value and convex is positive value). Beyond these, the higher DCT coefficients represent more complex components.

Limitations of the DCT Method:

1. **Coefficient Interpretation:** Although the DCT coefficients do reflect the pitch slide's slope and curvature, they are not directly comparable to mathematical definitions of these terms. This makes it challenging to compare DCT-based measurements with those derived from other mathematical models.

2. **Detail Loss:** The DCT method is best suited for capturing the broad contours of simple shapes like parabolic curves. It may not capture fine details, such as pitch slide modulations.

3. **Phase Sensitivity:** DCT is sensitive to the phase of the input sinusoidal signal. Vibrato in the starting or ending points of the signal can significantly affect the DCT coefficients. This necessitates pre-processing steps, such as moving averages or precise boundary location, to remove the noises.

Characterising the Overall Pitch Contour in a Note and a Syllable

Dai (2019, Section 6.2.4) outlines a method for modelling the broad pitch contour of a note as three distinct components: the initial transient (comprising the first 15% of the note's duration), the note's middle section, and the final transient (occupying the last 15% of the note's duration). Based on linear regression approximations of these components, the overall pitch contour of a note is categorised into one of four types: Concave, Convex, Upward, or Downward. These categories are determined by the slopes of the initial and final transients, which can be either positive or negative. Beyond categorisation, the model also allows for the measurement of the slope and variance for each of the three components: initial transient, note middle, and final transient. However, this approach has limitations, as it overlooks finer details such as overshoot and preparation.

For syllable segments, which may contain multiple notes, the overall pitch contour can be indicative of specific musical or linguistic features. For instance, in Chinese, it may correspond to the tone of a character, while in Indian art music, it could indicate the Raga. The DCT serves as a useful tool for capturing the broad contours of these syllables, particularly in terms of slope and curvature. While this approach is

well-suited for modelling the tones of Chinese characters, it may not be sufficient for capturing the complexity of $f0$ contours in Indian art music. In summary, for a more nuanced characterisation of $f0$ trajectories at the note or syllable level, segmentation into individual steady, modulating, and pitch slide regions is essential.

2.6 Review of Vocal Style Analysis in Pitch Contour Studies

This section reviews prior research in the field of computational vocal style analysis, with a special emphasis on studies related to pitch contour. Statistics serves as a cornerstone in this domain, encompassing various stages such as data annotation, description, and analysis. The application of statistical methods offers multiple advantages. Firstly, it enables the condensation and meaningful presentation of large datasets. Secondly, it facilitates the identification of patterns and trends, thereby aiding in hypothesis formulation and predictive modelling. Lastly, statistical analysis provides a robust framework for quantifying uncertainties and assessing the reliability of predictions and hypotheses. Beyond this, the theoretical interpretation of statistical results and the empirical validation through practical applications further enhance our understanding of the study's reliability. The depth and approach to statistical methods vary across different studies in vocal style analysis; specific examples illustrating these variations will be provided in the subsequent sections.

2.6.1 Data Annotation in Vocal Style Analysis

It is generally accepted that annotations in a dataset may contain noise, which can often be averaged out. However, this approach is not without risks. There are two examples below. Dai (2019) focused on the regularity and characteristics of note pitch trajectories. The note trajectory is divided into three components: the initial transient, the note middle, and the final transient. The slope of these transient parts is estimated using linear regression and is defined as the first 15% and the last 15% of the average note pitch trajectory. However, this 15% duration is a rough estimation based on

observations on the average note trajectory. The study acknowledges that this could be inaccurate for most note trajectories.

Furthermore, the note segmentation may influence the results. For example, the final slope, which is dominated by vibrato, could be classified as either positive or negative depending on the phase of the vibrato at the offset of the note. A close examination of a random sample revealed only a small fraction of cases where the sign of the slope could be ambiguous and the few ambiguous cases would not change the results significantly. Despite this, the study does not consider the amount of slope measurement errors, the effect of transient time, and the compensation of them by averaging. This leaves open the question of whether the measured slope genuinely reflects real-world f0 characteristics.

The second example comes from Section 3.2 of thesis (Devaney 2011), which aims to estimate the slope and curvature of f0 during the transition between two notes. The study straightforwardly cuts the last 250 ms of each note to represent the transition part. However, this fixed-duration cut may include vibrato from the note's middle section, introducing noise into the measurement. Given that the DCT is sensitive to the phase of the sinusoidal signal, the vibrato phase can significantly influence the slope measure.

To mitigate this, a moving average smoothing is applied to the original f0 signal with a window size of 200 ms, and the last 150 ms of the f0 trace is considered as the transient part. Despite these adjustments, the study reports a large amount of variability in the measurements and questions the reliability of the slope estimation method. As a result, the author abandoned this method for measuring slope in subsequent experiments in other sections.

In summary, the process of data collection and organization is intricate and multi-faceted. It involves several steps and considerations, each of which has its own set of challenges and limitations. Careful planning and execution are therefore essential to ensure the reliability and validity of the research findings.

2.6.2 Data Analysis

Data analysis in statistics is a multi-step process that begins with exploratory analysis to understand the data's characteristics and identify dependencies between variables. This is followed by statistical testing to confirm the significance of these dependencies.

Exploratory Analysis in Vocal Style Research

Exploratory analysis serves as the foundational phase in research, where the primary aim is to understand the key characteristics of the data. This is usually accomplished by examining data distributions and summary statistics. The overarching goal is to identify patterns or regularities in the data, which can manifest either as specific distributions of a variable or as functional dependencies between variables. It is crucial to note that these distributions and dependencies can often be mathematically represented as functions. Additionally, such regularities may be conditional, appearing under specific circumstances or conditions, which are relative to the content of music, such as note pitch or melodic interval and to the musical form, such as raga.

Various studies that have employed exploratory analysis in the realm of vocal style research. These studies aim to discern clear patterns or regularities specific to different vocal styles. The examples are organised by the type of variable under investigation:

1. **Categorical Variables:**

    - **Nominal Variable Example:**

      In a study detailed in section 6.2.4 of Dai (2019), the authors aimed to model note trajectories in singing and investigate how these trajectories vary across different vocal parts—Soprano, Alto, Tenor, and Bass (SATB). The variable of interest is the type of note pitch trajectory, categorized as a nominal variable. Utilizing a dataset of 400 recordings from five different groups of singers, focusing on two specific songs, the study yielded a total of 49,200 annotated single notes.

      The shapes of these note trajectories were categorised into four types, Concave, Convex, Upward, and Downward, based on the sign of the slope during

the attack and release phases. The distribution of these shapes was analysed conditionally, based on the specific vocal part, and was represented as frequency percentages.

The study revealed that the most common trajectory shapes across all vocal parts were Convex and Downward, both characterised by a negative note release. This led to the hypothesis that there is a consistent tendency for notes to end with a negative slope, irrespective of the vocal part. This finding suggests a commonality in how singers approach the ending of notes, possibly related to the relaxation of vocal muscles.

- **Ordinal Variable Example:**

  In a study by Caro Repetto et al. (2017*b*), the focus was on the relationship between linguistic tones and melodic contours in Jingju opera. The variable of interest is the shape of the pitch contours of a syllable, categorised as an ordinal variable. The dataset consisted of 7,283 syllabic contours from 92 Jingju scores.

  The study was conditional on two dialects —Beijing (BJ) and Huguang (HG)— and further refined by four tonal categories common to both dialects. The distribution of these pitch contours was presented as frequencies for each tonal category within the dialects.

  The study identified specific preferences for each tone and hypothesised a slight preference for the HG dialect. This research offers valuable insights into the complex relationship between linguistic tones and melody in Jingju opera, particularly how to infer dialects from syllabic pitch contours.

2. **Continuous Variables:**

   - **Vibrato Rate Example:**

     In a study by Caro Repetto et al. (2015), the aim was to compare the variability of vibrato rates between two vocal styles developed by the Cheng and Mei schools. The variable of interest is the vibrato rate, which is a continuous variable. The dataset included four recordings for each vocal

style, performed by six singers.

The study used mean and standard deviation (SD) to assess the distribution of vibrato rates. Although the observed regularity aligned with the hypothesis that the Cheng style would exhibit more variability, the results were not statistically significant, possibly due to the limitation of the small sample size.

- **Expressive Characteristics Example:**

  In another study by Yang, Tian, Chew et al. (2015*b*), the focus was on the expressive characteristics of Beijing opera singing, specifically examining vibrato rates, extents, and sinusoid similarity, which are continuous variables. The dataset comprised 16 monophonic performances, resulting in a total of 344 vibrato examples for the Laosheng role and 273 for the Zhengdan role. The distribution of these variables was visualised using Box plots. The study found that the Laosheng role exhibited a broader range of vibrato features compared to the Zhengdan role. However, the study did not formulate hypotheses based on existing musicology literature.

- **Duration of Meends Example:**

  In a separate study by Datta et al. (2017), the objective was to explore the duration of different categories of simple meends, a musical ornamentation technique in Indian classical music. The dataset used for this research consisted of 3,328 meends (longer than 300ms) that were automatically extracted from 116 songs performed by 41 eminent singers.

  The distribution of this continuous variable was visualized using histogram envelopes, which displayed the frequency of occurrences across different duration categories. Although no specific hypothesis was formulated, the study found that most meends had a duration of less than 600 ms, with the majority falling within the 300—500 ms range. Additionally, less than 1% of the total number of meends had a duration of less than 200 ms. These findings offer valuable insights into the temporal characteristics of meends, revealing variations in duration across different categories and performances.

3. **Discrete Variables:**

   Discrete variables are rarely the focus in vocal style analysis as vocal pitch is generally continuous. When they do appear, it is usually in the context of musical content analysis, such as the number of notes in a scale.

In summary, these exploratory analyses serve as initial investigations into various aspects of vocal styles. Some studies aim to validate existing theories or hypotheses in musicology, while others generate new hypotheses based on observed regularities. However, it is important to note that these exploratory findings are not definitive and should be further validated through rigorous statistical methods, which will be discussed in the subsequent section.

Statistical Testing Methods in Vocal Style Research

Statistical methods are essential in the realm of vocal style analysis. They offer rigorous techniques for hypothesis testing and validation of observed patterns. Several statistical techniques are commonly employed in vocal style analysis, including Analysis of Variance (ANOVA), Kolmogorov-Smirnov (KS) tests, linear regression analysis, and linear mixed models. Each of these methods comes with its own set of assumptions, application scenarios, and advantages, which will be discussed in detail along with previous vocal style analysis studies as examples.

Analysis of Variance, commonly known as ANOVA, serves as a powerful statistical tool for comparing means across multiple groups. Unlike the t-test, which is limited to comparing two groups, ANOVA can handle comparisons among more than two groups and is relatively robust against certain violations of its assumptions.

ANOVA operates under three main assumptions:

1. Observations within each group are normally distributed.
2. Variances within each group are approximately equal.
3. Observations are independent of each other.

The core objective of ANOVA is to test the null hypothesis, which posits that there are no significant differences between the group means. The F-statistic is employed

to compare the variance between groups against the variance within groups. A high F-statistic value can lead to the rejection of the null hypothesis, thereby indicating significant differences among the groups. Subsequent post hoc tests, like Tukey's HSD, can be used to pinpoint which groups significantly differ from each other.

One common variant is the one-way ANOVA, which focuses on the relationship between a single categorical independent variable and a continuous dependent variable. For instance, Dai (2019) used one-way ANOVA to identify significant differences in the mean pitch error between male and female vocal parts, ($F(1, 198) = 734.99, p < .001$). The study concluded that male singers tend to start notes at a higher pitch and adjust downwards, whereas female singers generally begin at a lower pitch, overshoot the target, and then adjust downwards.

Factorial ANOVA, which includes two-way, three-way, and higher-level ANOVAs, extends the scope of one-way ANOVA by allowing for the analysis of effects of multiple categorical independent variables (factors) simultaneously. This enables the investigation of interactions between factors, represented by interaction terms like $A \times B$. In the same thesis (Dai 2019), a two-way factorial ANOVA was conducted to explore interaction effects among various factors, such as note number in trial, singing condition, listening condition, and vocal part, revealing significant interactions for most combinations of factors.

The Kolmogorov-Smirnov (KS) test is a non-parametric test used to compare two distributions. It is commonly applied to continuous variables and is especially useful when the data do not meet the assumptions of other statistical tests. The KS test quantifies the distance between two distributions. The null hypothesis in the KS test posits that the samples are drawn from the same distribution. A low p-value (usually $p < 0.05$) indicates that you should reject the null hypothesis in favor of the alternative hypothesis, which states that the distributions are different.

In (Yang 2017), the exploratory analysis from (Yang, Tian, Chew et al. 2015*b*) was further tested. The KS test was used to compare the distributions of vibrato extents between the Zhengdan and Laosheng roles. The test showed a significant difference between the two distributions, with a p-value of 2.86×10^{-4} at the 1% significance level.

This 1% significance level is more stringent than the commonly used 5% level, providing stronger evidence to reject the null hypothesis and thereby reducing the likelihood of committing a Type I error.

Linear Regression Analysis: Linear regression analysis is a statistical technique used to model and analyse the relationships between a dependent variable and one or more independent variables. The primary advantage of linear regression is its flexibility to incorporate various types of predictors, including continuous and categorical variables.

The general form of the linear regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon \tag{2.7}$$

where y is the dependent variable, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients of the independent variables $x_1, x_2, \ldots, x_p$, and ϵ is the residual or error term, which captures the difference between the observed value of the dependent variable and the value predicted by the model.

The use of coefficients β, in the analysis indicates the size and direction of the effect that the predictor has on the variable being predicted, allowing for a nuanced understanding of the relationships between variables. The error term accounts for the random variation in the data that cannot be explained by the model.

The assumptions for applying linear regression include:

- Linearity: The relationship between the dependent and independent variables is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: The variances of the residuals are equal across all levels of the independent variables.
- Normality: The residuals are normally distributed.

In one study by Devaney (2011), linear regression was used to examine the effects

of various factors on the slope of the last part of the pitch contour of the first note in a melodic interval. The study utilised 306 melodic intervals and found a significant effect of certain intervals and of the professionalism of the singer on the slope. The linear regression yielded a small R^2 value ($R^2 = 0.04, p < 0.0001$), indicating a significant effect on the slope of A-Bb/Bb-A intervals versus other intervals. Additionally, a significant effect was observed for group identity, with the professional group averaging a 10 cents/second larger slope than the non-professional group (95% confidence interval = [3,18]).

However, the study also noted the potential for violating the assumption of independence, due to the possible correlation between sung notes close in time. This leads us to the next model, the linear mixed model, designed to address this issue. Linear Mixed Models (LMMs), also known as linear mixed-effects regression (LMER), provide a robust statistical framework that extends the capabilities of traditional linear regression models. One of the key advantages of LMMs is their ability to handle data that violate the assumption of independent observations, a limitation inherent in standard linear regression models. This makes LMMs particularly useful for analyzing correlated data, such as repeated measures on the same subjects, observations within clusters, or data points that are spatially or temporally close.

The strength of LMMs lies in their incorporation of both fixed and random effects into a single model. Fixed effects function similarly to standard regression coefficients, capturing the primary relationships between the predictors and the response variable. In contrast, random effects account for unexplained variability within clusters or among subjects. This dual structure allows LMMs to provide a more nuanced understanding of complex data, accommodating different baseline response values for each level of a random factor.

The mathematical representation of an LMM is $y = X\beta + Z\gamma + \epsilon$, where y is the response variable, X and Z are design matrices for fixed and random effects, respectively, β represents fixed effects, γ represents random effects, and ϵ is the error term. The assumptions for LMMs include those of linear regression—linearity and independence —along with additional assumptions concerning the distribution of random effects and

the error term, such as homoscedasticity and normality.

Dai (2019) employed an LMM to investigate various factors affecting pitch differences between the f0 contour and the score note pitch. The fixed effects used in the LMM are singing condition, listening condition, vocal part and note number in trial, and a random effect, the individual singer. Two examples of traditional Western church choral music were sung by 16 amateur female singers eight sopranos and eight altos, in different conditions for several times to generate 384 recordings and 18176 annotated notes. The results showed that the effects of all the tested factors are significant and some even with p-value smaller than 0.001.

2.6.3 Theoretical Interpretation

The theoretical interpretation of vocal styles serves as a crucial follow-up to the statistical analyses discussed earlier. While the data analysis section focused on identifying patterns and regularities in vocal styles, this section investigates how researchers have theorised these findings. The interpretations range from physiological and psychological factors to cultural and musical contexts. For instance, the study by Dai (2019) found a consistent tendency for notes to end with a negative slope across different vocal parts. The authors theorised that this could be due to the relaxation of vocal muscles at the end of a note, a physiological explanation that aligns with the statistical findings.

Additionally, it was observed that singers often exhibit a rising inflection towards the end of a note, just before the pitch falls at the very end. This pattern is thought to be a form of psychological preparation for hitting a higher pitch in the subsequent note. This specific pitch contour at the end of a note could be a characteristic feature influencing vocal style. Similarly, in chapter 4 of Devaney (2011), the research measured the curvature of the last part of the first note in melodic intervals. The intervals could either be between two chord tones or non-chord tones. The study found that the curvature values were smaller for intervals ending in a chord tone compared to those ending in non-chord tones. The authors suggest that this could be because singers are preparing for the stability of the subsequent note by introducing increased stability in the current note.

Panteli et al. (2017) employed an unsupervised K-means clustering method to explore similarities in singing styles. Their results indicated that clusters often grouped recordings from neighbouring countries or those with similar languages and cultures. The study also noted that the speed of syllabic singing plays a significant role in distinguishing between different singing styles. Separately, Shen (1982) provided an empirical explanation for the influence of language on singing styles. According to Shen, Western music, which aligns with Indo-European languages, often involves the use of ornaments to combine groups of notes with varied duration and loudness. This is because meaning in these languages is often formed through multiple syllables with stress emphasising meaning. In contrast, Chinese music, reflecting the characteristics of Sino-Tibetan languages, emphasises the shaping of pitch contours within individual notes or syllables. In these languages, single syllables carry independent meaning, and the tone itself conveys meaning.

Sundberg et al. (2012) compared Peking and Western opera, attributing vocal style differences to timbral variations in their orchestral accompaniments. Specifically, the absence of a singer's formant cluster (Sundberg 1995*a*) in Peking opera singers was linked to these timbral differences. Dai (2019) also suggested that the use of vibrato might be less marked in unaccompanied ensemble singing, where the goal is for voices to blend rather than stand out.

In summary, the theoretical interpretations in vocal style research often serve to explain the patterns and regularities identified through statistical analyses. These interpretations, grounded in various domains like physiology, psychology, and culture, not only provide a deeper understanding of the data but also offer avenues for future interdisciplinary research.

Chapter 3

Pitch Contour Segmentation and Characterisation Methods

This chapter introduces the concept of 'pitch contour unit' (PCU), which represents a discrete segment of the f0 signal delineated by consecutive local peaks and troughs, aimed at providing a method for segmenting and characterising pitch contours across diverse musical cultures. Traditional methods often struggle with the continuous nature of pitch contours in vocal music, which varies significantly from one style to another. While previous studies (Gong et al. 2016) have utilised similar concepts, this thesis innovates by formalising PCU as a novel unit for pitch segmentation and analysis. The segmentation level of PCUs, positioned between frame-level analysis and individual notes, effectively bridges the gap between the excessive granularity of frame-level analysis, which does not align with human music cognition, and the subjective variability inherent in note definition. This segmentation strategy provides a resolution that captures unidirectional movements within the pitch contour, making PCUs particularly suited for analysing the subtle nuances and ornaments of diverse pitch contour patterns. By dividing complex pitch contours into manageable PCUs and employing a Hidden Markov Model (HMM) for their analysis, this methodology offers a novel way to universally characterise the primary elements of pitch contours: steady, modulating, and transitory elements, which have been elaborated in Section 2.5.5.

3.1 Dataset

This section introduces four datasets. The first dataset, consisting of pitch contour element segments, is divided into a training set and a test set for training and evaluating the method's capability in detecting the three element types, transitory, steady and modulating. Furthermore, three additional datasets are employed solely for testing purposes, to assess the method's general effectiveness in detecting a specific type of melodic feature individually in datasets annotated by different annotators and datasets in different musical genres.

Pitch Contour Segments Dataset

A subset of the dataset from Gong et al. (2016) was employed for pitch contour segmentation tasks, consisting of acapella singing recordings. This dataset primarily focuses on two Jingju role-types: Dan (female) and Laosheng (elderly man), featuring 41 interpretations of 33 arias by 13 Jingju singers. Manual annotation of the pitch contour segmentation was performed, identifying the three elements: steady, unidirectional transitory, and vibrato, totalling 14,467 segments, which were considered as ground truth by Gong et al. (2016) to evaluate the pitch contour segmentation method they proposed.

However, labels of the three elements are not published with the segments in Gong et al. (2016). Recognising the importance of such labels for evaluating algorithm performance in detecting pitch contour elements, this thesis includes manual correction and labelling of these segments. Due to time constraints, this thesis selected 12 recordings from the total dataset of 41 recordings (containing 14,467 segments). This subset, comprising 8 tracks for the training set (of which 1/10 was allocated for validation purposes, detailed in Section 3.2.3) and 4 for the testing set, was chosen to balance the need for thorough manual verification against time constraints. The manual analysis involved correcting and labelling pitch contour elements in these selected recordings to create a validated ground truth dataset for algorithm evaluation. While this sample size is relatively small, it is sufficient for this study because the model employed uses a simple Hidden Markov Model with fewer than 10 parameters—specifically, the transition

probabilities between three states and the observation likelihood distribution functions of these states. The selection was also carefully made to ensure a diverse representation of singers, roles, and emotions, thereby achieving a balanced dataset. Tables 3.1 and 3.2 provide detailed metadata for the training and test sets, respectively. The role type and emotion labels are made by Black et al. (2014).

File Name	Role Type	Emotion
bcn_001	Dan	Positive
bcn_007	Dan	Negative
fem_01_neg_1	Dan	Negative
fem01_pos_1	Dan	Positive
male_01_neg_1	Laosheng	Negative
male_01_pos_2	Laosheng	Positive
male_02_neg1	Laosheng	Negative
male_13_pos1	Laosheng	Positive

Table 3.1: Training set recording metadata containing 3,096 manually annotated pitch contour segments from 8 recordings

File Name	Role Type	Emotion
fem_07_pos_1	Dan	Positive
fem_11_pos_1	Dan	Positive
londonRecording-Laosheng-01	Laosheng	Negative
bcn_003	Laosheng	Negative

Table 3.2: Test set recording metadata containing 1,135 manually annotated pitch contour segments from 4 recordings

The annotation methodology involves grouping Pitch Contour Units (PCUs) into higher-level segments to form three types of pitch contour elements based on their shared boundaries and characteristics. Adjacent PCUs sharing the same boundary are grouped together, and those with similar characteristics are categorized into either steady or modulating regions. In steady regions, PCUs typically have small durations and low amplitude variations, while in modulating regions, neighbouring PCUs exhibit similar durations and intervals, indicating a vibrato-like modulation. PCUs that do not fit into these categories are labelled as transitory. To save time, entire segments are made for steady or modulating elements, rather than individually segment each PCU within these regions. In contrast, PCUs were segmented and labelled for transitory

elements. Noise regions, such as attacks, echo, or unvoiced consonants, are discarded.

However, it is not always that each PCU clearly belongs to a specific element. For example, some transitory PCUs, have a steady region in the PCU either in the middle or at ends. This characteristic requires to do a finer segmentation to distinguish the steady part from the transitory within the individual PCU. Therefore, the curvature is an important feature in the process of segmentation.

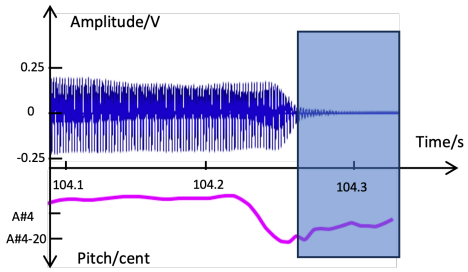
To accurately annotate the segments, the following steps are followed:

1. Listening through the entire track.
2. Labelling each segment based on subjective perception of the author.
3. If more than one element are observed in a segment, breaking it into multiple regions.

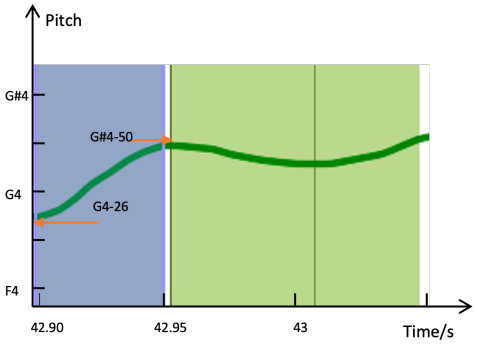
The annotation challenges include:

1. Visual Effect Risks: The visualised pitch track can influence decision-making during labelling, which should be based on hearing. Figure 3.1a illustrates an f0 signal may indicate an outlier (highlighted by the blue box) in a segment, which might not be audible due to the low loudness. Sole reliance on visual signals is risky, as the perceived interval could be influenced by the vertical scale. For example, the pitch slide in the blue region (Figure 3.1b), if the vertical scale is compressed, the pitch slide would flatten and look like a steady region. Therefore, visual f0 signals should not be the definitive reference for annotation, and caution must be exercised against over-reliance on visual details.
2. Hearing Perception Risks: Dependence on auditory perception means that annotations cannot be considered absolute ground truth. Hearing perception is inherently subjective, and annotations may vary with different annotators. Moreover, the consistency of annotations is influenced by factors such as musical context, the purpose of the annotation, and the duration of listening.

In conclusion, despite the challenges posed by auditory perception, since music is fundamentally based on hearing rather than visualisation, the most viable, albeit



(a) Example of an f0 contour (purple) with an outlier (blue box) in the vocal track fem_01_neg_1. The outlier may not be perceptually significant due to low amplitude.



(b) The perceived interval influenced by the vertical scale.

Figure 3.1: Visual effects of f0 signals on decision-making during labelling.

imperfect, approach is to rely on auditory rather than visual perception.

The adjustments made to the original segments made by Gong et al. (2016) include:

- Exclusion of silent parts, where the pitch is detected but it is too soft to be heard, from the annotations.
- Splitting of transitory regions into PCU.
- Exclusion of noises.

Portamento Dataset

For the evaluation of portamento, this study utilises the dataset annotated by Yang et al. Yang et al. (2016). This dataset is composed of Beijing opera recordings, sourced from the collection by Black et al. (2014). These recordings coincide with those used in the pitch contour segment dataset, discussed in Section 3.1, and include a subset of audio tracks common to both collections. Portamento annotations for these opera pieces were conducted utilising the AVA interface, as documented by Yang et al. (2016). To maintain the integrity of the testing environment and prevent data leakage, audio tracks (number 3, 5, 8, 11 and 13) that were previously used in the training set (see Section 3.1) have been omitted from this dataset. The statistics on portamento annotations are detailed in Table 3.3, where the 'neg' and 'pos' in the filename indicate the emotion and the last column is the number of portamento annotations.

Filename	Role	Duration (s)	Portamenti
fem_01_neg_3	Zhengdan	51	71
fem_01_pos_3	Zhengdan	41	48
fem_01_pos_5	Zhengdan	181	219
fem_01_pos_7	Zhengdan	71	87
fem_10_pos_1	Zhengdan	160	173
fem_10_pos_3	Zhengdan	81	49
male_01_neg_4	Laosheng	148	106
male_01_pos_1	Laosheng	171	144
male_12_neg_1	Laosheng	104	94
male_12_pos_1	Laosheng	185	224
male_13_pos_3	Laosheng	95	166

Table 3.3: Summary of portamento dataset

Within the dataset, there are 39 instances where the majority of the annotated portamento falls within unvoiced regions. Figure 3.2 depicts an example of such a case. The green line represents the interpolated pitch curve, while the purple shaded area indicates the originally annotated portamento. This issue in portamento annotation is attributable to the limitations of the AVA system (Yang et al. 2016), which renders a continuous, smooth pitch representation that may not align with the true pitched and dynamic characteristics of the audio (see Figure 3.3). To rectify these inaccuracies, non-pitched portions were excised from the annotation. Furthermore, any residual pitched segments shorter than 50 ms were eliminated.

Steady Dataset

For the evaluation of steady regions in vocal recordings, this study employs a dataset annotated using the methodology described by Rosenzweig et al. (2019). This dataset comprises a selection of five audio tracks from the Erkomaishvili dataset (Rosenzweig et al. 2020), a repository of Georgian chants. The selected recordings have been analysed to identify and annotate the stable regions of their pitch traces, using the interactive tool developed by Müller et al. (2017). The detailed statistics on the duration and the number of stable regions for each audio track are presented in Table 3.4.

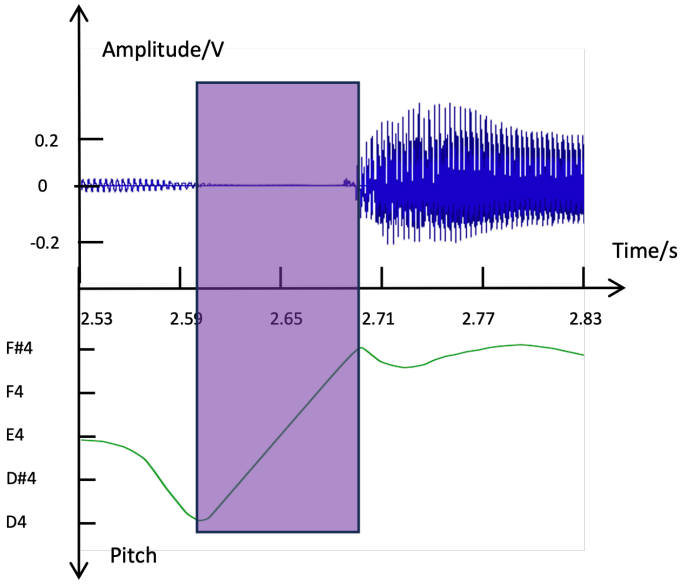


Figure 3.2: Example of annotated portamento in unvoiced region of vocal track fem_01_neg_1, with f0 shown in green and unvoiced region indicated by purple block.

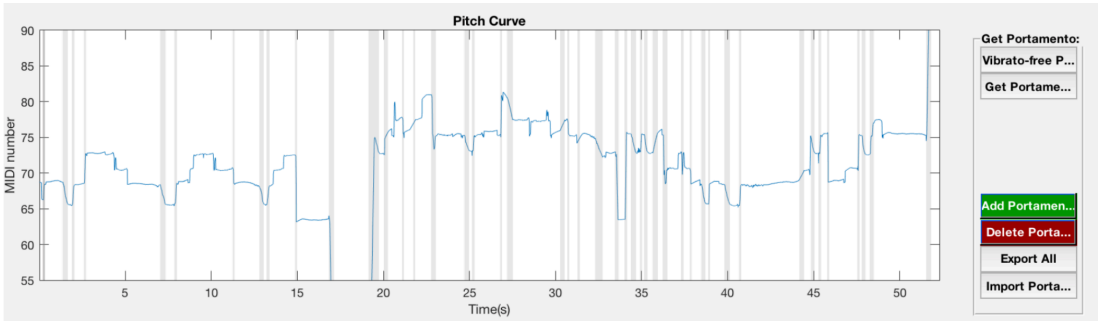


Figure 3.3: A typical pitch curve as displayed by the AVA interface

Vibrato Dataset

For the evaluation of vibrato, this study utilises the dataset annotated by Yang, Tian, Chew et al. (2015*a*). In this dataset, the recordings are identical to those in the portamento dataset. The vibratos are annotated by the first two authors, Yang and Tian, using Tony Software (Mauch et al. 2015). The statistics on vibrato annotations is detailed in Table 3.5.

Audio No.	Duration (s)	Number of Stable Regions
1	129.30	71
2	181.76	176
3	204.28	237
4	55.48	27
5	47.49	24

Table 3.4: Steady region dataset summary

Filename	Role	Duration (s)	Vibrato Regions
fem_01_neg_3	Zhengdan	51	21
fem_01_pos_3	Zhengdan	41	14
fem_01_pos_5	Zhengdan	181	47
fem_01_pos_7	Zhengdan	71	25
fem_10_pos_1	Zhengdan	160	48
fem_10_pos_3	Zhengdan	81	23
male_01_neg_4	Laosheng	148	52
male_01_pos_1	Laosheng	171	36
male_12_neg_1	Laosheng	104	53
male_12_pos_1	Laosheng	185	61
male_13_pos_3	Laosheng	95	40

Table 3.5: Summary of vibrato dataset

3.2 Methods

3.2.1 Pitch Extraction and Pitch Curve Modification

Accurate pitch tracking is crucial for our study of pitch contour analysis. We employ the PYIN algorithm (Mauch & Dixon 2014), which is widely used for pitch extraction in monophonic signals due to its ability to provide accurate and high-resolution f0 contours (5.8 ms hop size and 10 cent pitch resolution). This level of detail is essential for capturing the nuanced pitch variations characteristic of singing performances.

In order to refine the continuous pitch trace segment for further analysis, two key steps are employed to each continuous pitch trace segment separately: interpolation and smoothing. Interpolation is applied before smoothing to reduce alterations to the original pitch data during the smoothing process. Piecewise cubic spline interpolation is selected because the pitch contour of singing is complex and nonlinear, making linear interpolation unsuitable. Moreover, piecewise cubic spline is a widely used method for curve interpolation. Next, in line with Yang et al. (2016), a method of 10-point

moving average smoothing, a common choice suggested, is applied to the interpolated f0 to remove minor local extremes in the f0 curve, which are considered noise in the following detection method.

3.2.2 Pitch Variation Features Extraction

After acquiring the pitch, the subsequent phase involves the extraction of pitch variation features.

PCU Characteristics: Each PCU, visualised in Figure 3.4, is characterised by its duration and extent:

- **Duration**: Duration is defined as the time difference between the PCU's start and end points, representing the interval between consecutive peaks and troughs (Figure 3.2), as 3.4 shows.
- **Extent**: Extent is half of the pitch interval between the PCU's start and end points, with signed values to indicate upward or downward direction.

Figure 3.4 illustrates the application of the PCU concept to a pitch contour. Red circles indicate local peaks and troughs. The pitch interval is marked by the vertical distance between a peak and the following trough, while the duration captures the horizontal extent of the PCU.

3.2.3 HMM-based Pitch Contour Element Detection

The Hidden Markov Model (HMM) is a statistical tool ideal for analysing time series data, such as pitch contours in music. It is based on the concept of Markov processes with unobserved or *hidden* states, to infer the hidden states from the observed sequence.

HMM Structure and Parameters

- *Hidden State:* Denoted by X_t, it represents the hidden state in the model at time t. For instance, $X_t = j$ indicates that the system is in the j^{th} state at time t.

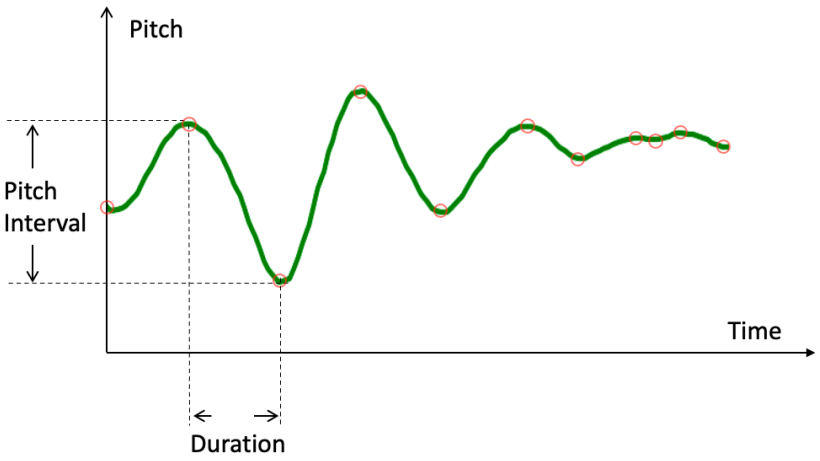


Figure 3.4: Illustration of PCU characteristics in pitch contour. Red circles indicate local peaks and troughs. Horizontal axis is time in seconds and vertical axis is pitch.

- *Observed Sequence:* Represented as $O = \{o_1, o_2, ..., o_T\}$, where each o_t corresponds to a feature vector observed at time t.

- *Emission Probability or Observation Probability:* Denoted as $b_j(o_t)$ (Equation 3.1), indicating the likelihood of observing a specific feature o_t given a specific hidden state j.

$$b_j(o_t) = P(o_t|X = j) \tag{3.1}$$

- *Transition Probability:* Represented by a_{ij} (Equation 3.2), these probabilities form a matrix indicating the likelihood of transitioning from state i to j.

$$a_{ij} = P(X_{t+1} = j|X_t = i) \tag{3.2}$$

- *Initial Probability:* Given by π_i (Equation 3.3), this sets the initial conditions of the Markov process.

$$\pi_i = P(X_1 = i) \tag{3.3}$$

An illustrative diagram of the HMM structure is provided in Figure 3.5).

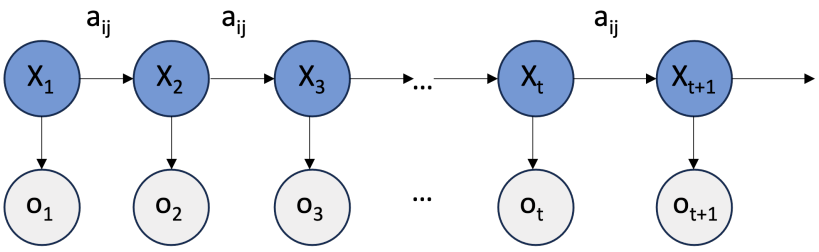


Figure 3.5: Basic structure of a HMM for pitch elements

Sequence Inference in HMM

The goal of an HMM is to infer the most likely sequence of hidden states from the observed sequence. To realise the inference, the Viterbi algorithm is commonly employed. This dynamic programming algorithm calculates the most probable path through the hidden states that results in the observed sequence.

HMM Structure for Pitch Contour Element Detection

The HMM is structured to reflect the nuanced behaviours of pitch elements as follows:

- **Initial States:** The model allows for any of the three states—steady, modulating, and transitory—to be the initial state of a pitch sequence, reflecting the natural variability in the onset of vocal expressions.
- **State Transition Structure:** The transitions between states are not arbitrary but follow a probabilistic structure that encapsulates the natural progression of pitch elements. This structure is visually depicted in Figure 3.6, which details the likelihood of transitioning from one state to another within the pitch contour context.
- **Observation Sequences:** The sequences of observation has two dimensions, one is duration and the other is extent.

Model Training

The training data comprise eight recordings selected from Gong et al. (2016), with pitch elements manually annotated on PCU level based on the original pitch curve and audio.

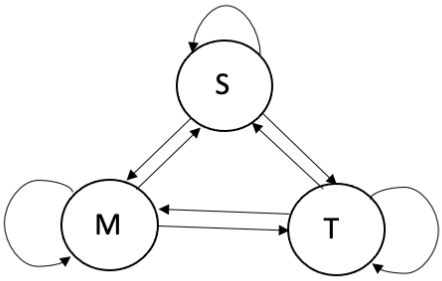


Figure 3.6: State transitions of the HMM for pitch elements

For more details, see Section 3.1. Given the availability of annotated data, this thesis adopts a supervised learning approach for training the HMM. This approach uses the labelled hidden states of each PCU and observed sequence of features to estimate the transition probability matrix and observation probability distributions of each state.

Initial Probability

Since the starting pitch is unknown, this method introduces an assumption of equal probability among states for initiating a pitch track. Specifically, each state—Transitory (T), Steady (S), and modulating (M)—is assigned an identical initial probability. Consequently, the initial probability for each state is uniformly set to $\frac{1}{3}$, reflecting the equal likelihood of any state commencing a pitch sequence.

Transition Probability Matrix Estimation

The transition probabilities within an HMM are statistically derived from the labelled state transitions within pitch contour segments. Each transition links two successive states, denoted as X_t to X_{t+1}, indicating the progression from X_t to X_{t+1}. Given three distinct states—Transitory (T), Steady (S), and Modulating (M)—there are a total of $3 \times 3 = 9$ possible transitions.

Utilising Maximum Likelihood Estimation (MLE), the estimated transition probability $\hat{a}_{ij}$ is:

$$\hat{a}_{ij} = \frac{N_{ij}}{\sum_{j=1}^{3} N_{ij}}, \quad \text{for } i = 1, 2, 3; j = 1, 2, 3 \tag{3.4}$$

where N_{ij} represents the count of transitions from i to j and 3 is the total number of

states.

The estimated transition probability matrix for the three states in the HMM is presented in Table 3.6.

State	Transitory	Steady	Modulating
Transitory	0.59	0.35	0.06
Steady	0.15	0.84	0.01
modulating	0.09	0.04	0.87

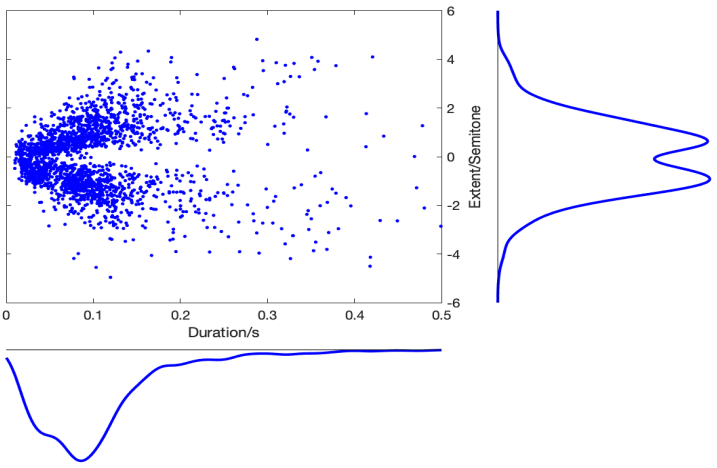
Table 3.6: Estimated transition probability matrix for HMM states

Observation Likelihood Distribution Estimation Method

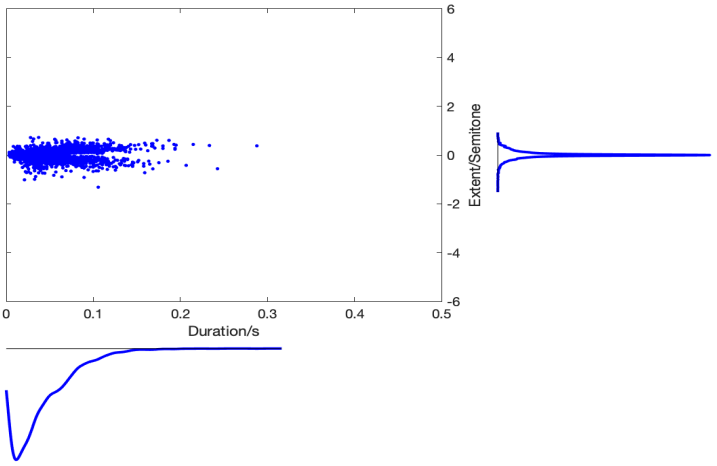
The observation probability for each state in a Hidden Markov Model (HMM) would be calculated from the estimated observation likelihood distribution. It is important to note that the likelihood is denoted by $L(\theta) = P(o|\theta)$ where parameters θ are unknown while the data o is known and is utilised for parameter estimation, whereas the probability $P(o|\theta)$ relies on known parameters to calculate the probability of observed data from the model which is defined by parameters θ. The operation of estimation of observation likelihood distribution involves first establishing the distribution of observed features, and then fitting this to a chosen theoretical distribution while simultaneously optimising its parameters to determine the estimated observation likelihood distributions.

Step 1: Feature distribution analysis: The features of each PCU, namely duration and extent, are extracted and analysed for their distribution across different labelled states. The scatter plots of duration and extent are depicted in Figure 3.7. The figures illustrate the distribution of duration and extent of each PCU for different pitch contour state, providing insight into the temporal and dynamic aspects of pitch variation.

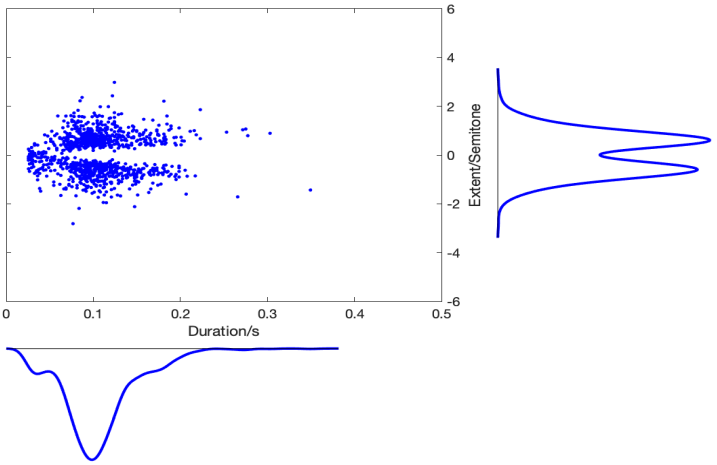
- **Transitory State:** The scatter plot shows a concentration of data points around 0 to 0.2 seconds, centred near the zero semitone mark, indicating that most events have shorter durations and are centred around a specific extent. As duration increases, the spread of points widens, indicating greater variability. The bottom histogram reveals a high frequency of short-duration events, tapering off with



(a) Transitory state



(b) Steady state



(c) Modulating state

Figure 3.7: Scatter plot of duration and extent distribution in three states

longer durations. The right histogram shows a central trough around zero semitones, with two modes around 1 and -1 semitone extent.

- **Steady State:** The duration histogram leans towards an exponential distribution, with most values concentrated at shorter duration. The extent histogram, appearing more symmetric around its mode near zero, suggests a normal distribution with low variance, alluding to minimal pitch variation. The scatter plot's tight clustering around lower values corroborates a pitch that is stable and steady, with negligible fluctuations.

- **Modulating State:** The histogram for duration indicates a mode around 0.1 seconds, which is similar to that observed in transitory state. However, it exhibit a more rapid decline for longer durations, indicating a lower likelihood of PCU duration longer than 0.2 seconds in modulation. The histogram for extent has similar modes with that in transitory state while the frequency declines more rapidly on both sides. The scatter plot is densely packed, forming three distinct clusters. One cluster is centred around smaller duration and extents, reflecting rapid and slight pitch fluctuations. The other two, symmetrically positioned around higher duration and larger intervals, suggest patterns of deliberate and controlled pitch modulating. These clusters suggest the dual nature of pitch variation in the modulating state potentially: both involuntary micro-variations and purposeful modulations.

Additionally, these observed patterns reflect the interplay between the physical mechanics of vocal production and the acoustic manifestation of pitch, with shorter pitch movements generally corresponding to smaller extents due to the physiological constraints of the vocal apparatus.

Step 2: Observation Likelihood Distribution Fitting with KDE:

Given the interdependence of duration and extent, their joint modelling in a two-dimensional space is crucial. The Gaussian Mixture Model (GMM) is less suitable for this task as it assumes Gaussian distributions, which do not fit the observed scatter in transitory and steady states. Consequently, Kernel Density Estimation (KDE) is

employed, a non-parametric approach that does not assume specific parametric forms for the distribution, thus more accurately representing the intricate and diverse pitch patterns observed in these states. KDE is characterised by a smoothing function and a bandwidth value, which controls the smoothness of the estimated density curve. Formally, the KDE is expressed as:

$$\hat{f}_h(o) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{o - o_i}{h}\right), \tag{3.5}$$

where K denotes the kernel function, and h represents the bandwidth, o_i denotes a specific observation from the data, while n signifies the count of all observations. Each o_i contributes to the density estimation at a point o, with the summation across n ensuring normalisation of the density estimate.

The selection of hyperparameters and preprocessing is below:

- **Kernel Choice:** Given that the choice of kernel has a relatively minor impact on the fit, it is opted for the widely used Gaussian kernel function, one of the most common choices in kernel density estimation.

- **Bandwidth Selection:** The choice of bandwidth in KDE is critical. A small bandwidth may lead to overfitting, resulting in a "noisy" or "spiky" estimation, as it closely follows individual data points. Conversely, a large bandwidth can cause underfitting, overly smoothing the data and losing significant distribution features. For selecting the appropriate bandwidth for each feature, the well-established bandwidth optimisation method proposed by Botev et al. (2010) is employed, which is widely recognised for its accuracy and fastness. To prevent overfitting, bandwidth values ranging from 1 to 10 times the acquired bandwidth will be experimented with to identify the optimal value during the HMM optimisation process, which is presented in Step 3.

- **Scaling Methods:** Even though different bandwidths can be chosen for each feature in Kernel Density Estimation (KDE), scaling is taken into consideration since it is a standard pre-processing step that is beneficial for mitigating the disproportionate influence of scale differences across features. **Min-Max Scaling**

is able to normalise features to a fixed range, typically [0, 1] as defined by equation 3.6:

$$X_{\text{scaled}} = \frac{O - \min(O)}{\max(O) - \min(O)} \tag{3.6}$$

where O represents the set of observed values of the feature, and $\min(O)$ and $\max(O)$ are the minimum and maximum values, respectively.

Step 3: Further Optimisation of Hyperparameters Based on Validation: Although the optimisation of hyperparameters has been discussed previously, further optimisation based on validation data is still necessary. Using k-fold cross-validation, the optimisation encompasses the following steps:

1. **Data Partitioning:** All the pitch contours in the training set is divided into k mutually exclusive subsets of approximately equal size. The choice of k typically depends on the size of the dataset, with $k = 5$ or 10 being common choices. Since the dataset utilised is large enough, k is set as 10. Considering the variance of the number of PCU in each subset, this study performs partitioning 100 times and select the partition that exhibit the smallest variance.

2. **Model Training and Validation:** For each fold, the HMM is trained on $k - 1$ subsets and then validated on the remaining subset. This process is repeated k times, with each subset serving as the validation set exactly once.

3. **Performance Aggregation:** The performance at the frame level of the HMM for each parameter set is aggregated across all k folds. The performance metrics are illustrated in Section 3.3.1. This aggregation includes both the average and the variance of the performance metric. To succinctly represent the performance of each parameter set with a single value, the average and variance of the performance metric are combined using the equation 3.7:

$$\text{Score} = \bar{P} - \lambda \times s^2 \tag{3.7}$$

where $\bar{P}$ is the sample mean of the performance metric, representing the average performance across k-folds. s^2 denotes the sample variance, reflecting the perfor-

State	Duration	Extent
Transitory	0.0271	0.2572
Steady	0.0032	0.0091
modulating	0.0037	0.0436

Table 3.7: Optimised bandwidth of features of pitch contour states

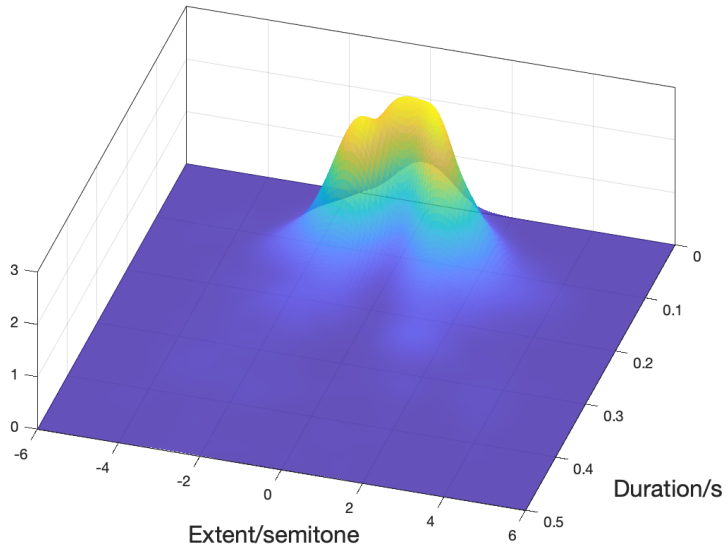


Figure 3.8: Observation probability density function for transitory state

mance variability. The parameter λ, set as $\frac{1}{10}$ empirically, to control the trade-off between mean performance and its variability.

A common practice is to repeat the above k-fold cross-validation process 3 times to achieve more robustness against the randomness in data splitting. The bandwidth that yields the best aggregate performance score is selected as the optimal bandwidth to train the HMM.

Estimated Observation Probability Density Functions of Each State Using KDE

With the optimised bandwidth (see Table 3.7) for Kernel Density Estimation (KDE) fitting, the observation probability density functions (PDFs) of each state are estimated using the training set. The estimated PDFs are utilised to calculate the observation probability of unknown data for each state in the HMM. The PDFs for the transitory, steady, and modulating states are illustrated in Figures 3.8, 3.9, and 3.10 respectively.

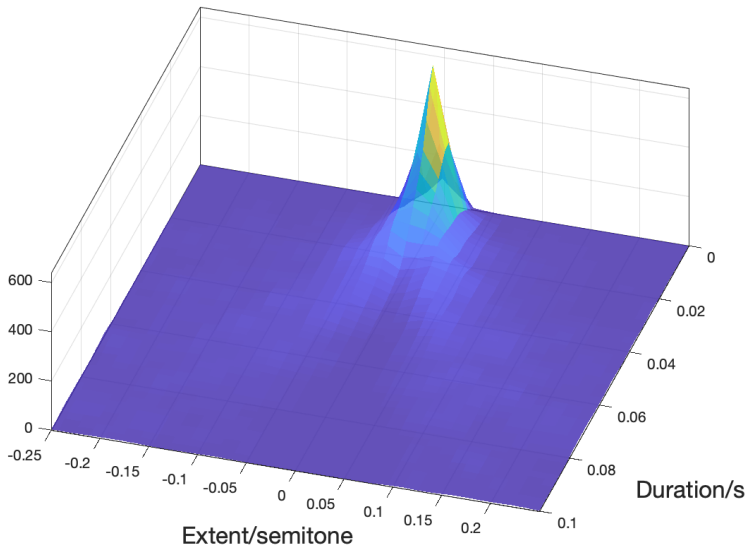


Figure 3.9: Observation probability density function for steady state

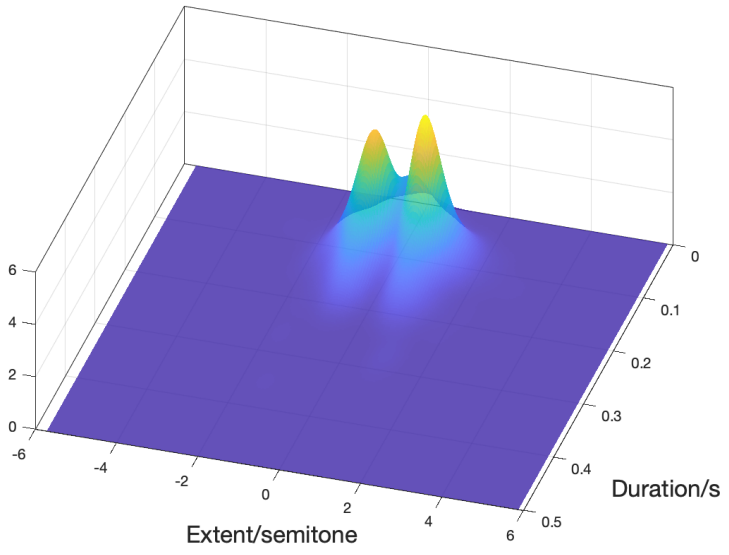


Figure 3.10: Observation Probability Density Function for modulating State

Post-processing

The post-processing step is designed to treat any modulated region comprised of fewer than three PCUs as transitory, as two PCUs are insufficient to manifest a modulating behaviour—at least three points are required to establish a clear oscillatory pattern. This approach aligns with the understanding that a substantial modulated region, indicative of vibrato, typically spans across at least three PCUs. Consequently, regions not meeting this criterion are considered as transitory.

3.2.4 Finetuning the HMM-Based Method for Portamento and Steady Region Detection

Other than basic pitch contour elements, this method has the flexibility to be used to detect ornaments and steady regions in specific vocal style by finetuning the parameters of the HMM empirically.

Portamento Detection in Jingju Singing

The finetuning process leverages the pitch contour element detection method introduced in this chapter, adapting it from general transitory state detection to specifically target portamento detection. Portamento, recognised as a specialised subset of transitory elements, adheres to more strict criteria than transitory regarding the pitch interval. It is also inherently unidirectional, in this thesis's definition, manifesting either as ascending or descending motions. These nuances are exemplified in Figure 3.11, where the green curve represents the pitch trace of a song from the utilised dataset. The red regions illustrate portamenti as annotated by Yang et al. (2016), contrasting with the grey areas at the top that indicate the horizontal range of the transitory region.

This conceptual distinction is critical for the finetuning procedure, which incorporates an exponential decay factor in the observation probability of transitory states. The decay factor in Equation 3.8 is calibrated to diminish the probability of mistaking slight transitory events for portamenti. If a PCU's pitch extent falls below predetermined thresholds, the decay factor reduces the observation probability accordingly, thereby refining the detection of true portamenti.

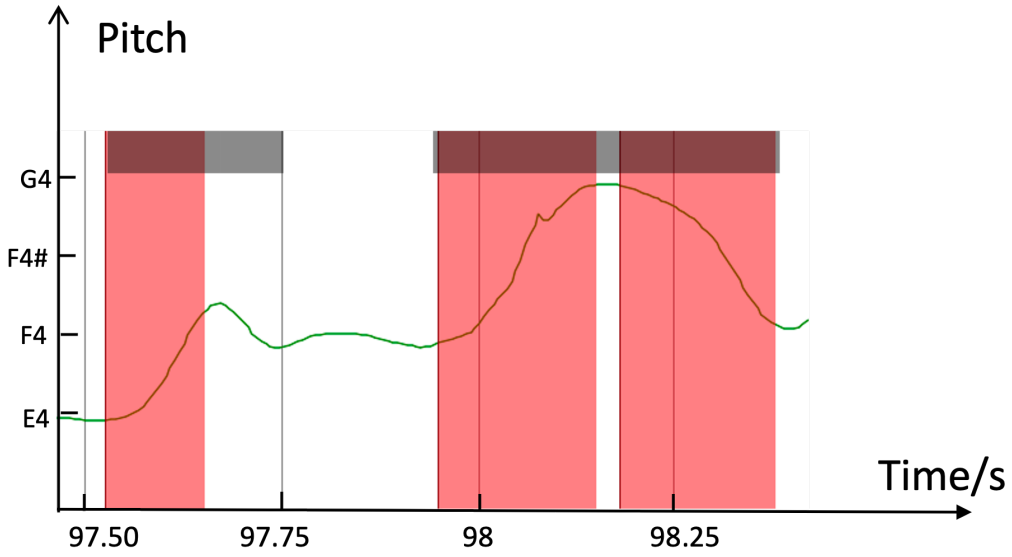


Figure 3.11: A pitch contour with portamento (red) and transitory region (grey)

$$b_j(o_t) = P(o_t \mid X = j) \cdot \exp(\text{decay\_level} \cdot \left(\frac{T_o - |o_t|}{T_o}\right)) \tag{3.8}$$

Here, decay_level is set to a negative value. The more negative the decay level, the more significant the reduction in observation probability, with T_o representing the threshold and o_t denoting the PCU's feature value, whether it be duration or pitch extent. The proximity of o_t to T_o inversely affects the decay. $X = j$ corresponds to transitory state. The decay levels are set empirically at -5 for duration and -10 for extent through qualitative assessment of the decay function's behaviour. This intuitive approach was chosen primarily to test the feasibility of the decay function concept, without pursuing formal optimisation, which could potentially cause over-fitting.

Steady Region Detection in Georgian Chant

For the steady region, it is essential to distinguish it from the steady elements of pitch contour as defined. The primary distinction is that the steady region functions as an indicator of the stable pitch within a melody. It is characterised by a sufficient duration of 0.15 seconds, exceeding the 0.1 second threshold typically used in monophonic singing, and is set to ensure clear perception in polyphonic data. Additionally, its pitch level is consistent with the melody's scale.

In alignment with this empirical knowledge, three modules are applied in the finetuning process. To decrease the observation likelihood of transitory states, a distinctive decay method is applied to steady-state regions, differing from the finetuning approach used for portamento (see Section 3.2.4). This method is formalised by the decay function,

$$f_{\text{decay}}(o) = \exp\left(-\frac{(|o|-M)^2}{2(M-m)^2}\right), 0 < x < M \tag{3.9}$$

where o denotes the value of extent of a PCU, M represents the half of the maximum pitch interval between adjacent degrees of the musical scale, and m signifies the half of the minimum pitch interval between adjacent degrees, determining the point of maximum decay rate. The standard deviation is set as $\sigma = M - m$. The function decreases as o surpasses M, with the decay rate reaching its apex when o equals m. This tailored decay sets two decay ranges: the first, from m to M, where a transitory region may link pitches across scales as a portamento; and the second, from 0 to m, where, disregarding microtonal variations, a pitch is considered part of a steady region without a connecting transitory. The second module augment the self-transition probability of steady states by improving it from 0.84 to 0.99, which mitigates the fragmentation of steady regions by transitory states. The third module eliminate steady-state regions detected with durations shorter than 0.15 seconds. These three models are called 'Observation Probability Decay' (OPD), 'Steady Self-Transition Increase' (SSTI), and 'Postprocessing Removing Short Region' (PRSR) respectively.

3.3 Evaluation Results

To evaluate the proposed approach to pitch contour element segmentation and labelling, this section employed four datasets introduced in Section 3.1. The evaluation focused on several key components. First, the detection of modulating, transitory, and steady elements is assessed individually and the pitch contour element segmentation results are compared against a prior published method. Then, portamento, vibrato and steady region detection are evaluated separately, employing ablation studies with standard metrics followed by comparisons with state-of-the-art systems. For all tables presented

in this section, metrics with a downward arrow (↓) indicate that lower values are better, while metrics with an upward arrow (↑) indicate that higher values are better. If no arrows are present, it means all metrics have an upward arrow (↑).

3.3.1 Evaluation of Pitch Contour Element Detection on Pitch Contour Dataset

The evaluation of pitch contour element detection algorithms in this study is conducted using two metrics: frame-level accuracy and confusion matrix analysis.

Frame-Level Accuracy: The frame-level accuracy metric is a quantitative measure of an algorithm's performance, comparing the predicted state sequence against the ground truth for each individual frame. As delineated in Equation 3.10, this metric calculates the proportion of frames that are correctly predicted concerning their respective state of pitch contour elements. Here, N represent the total number of frames, and S_i and $\hat{S}_i$ denote the ground truth state and the predicted state for the i-th frame, respectively.

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(S_i = \hat{S}_i) \tag{3.10}$$

This evaluation focuses on the overall proportion of correctly detected frames, providing a holistic perspective of the algorithm performance across all elements. Table 3.8 presents the mean and variance of frame-level detection accuracy of each pitch contour for this method.

Method	Mean Accuracy (↑)	Variance (↓)
Proposed	0.66	0.10

Table 3.8: Mean and variance of frame-Level detection accuracy for the proposed method

Confusion Matrix Analysis: Tables 3.9 and 3.10 present the algorithm's performance through recall and precision metrics for each state. The recall values (Table 3.9) show that the algorithm is most effective at identifying transitory states (77.7% recall), moderately successful with steady states (67.3% recall), but less reliable in detecting modulating states (44.1% recall). The precision values (Table 3.10) indicate similar

Actual state	Transitory	Steady	Modulating	Total frames	Recall
Transitory	54009	6491	9032	69532	77.7%
Steady	10710	29565	3624	43899	67.3%
Modulating	15558	7183	17930	40671	44.1%

Table 3.9: Per-state recall showing how often each actual state was correctly identified. Rows represent the actual states, while columns show how these states were predicted by the algorithm. For example, of the 69,532 actual transitory frames, 54,009 were correctly identified (77.7% recall), while 6,491 were misclassified as steady and 9,032 as modulating.

Predicted as	Transitory	Steady	Modulating	Total predictions	Precision
Transitory	54009	10710	15558	80277	67.3%
Steady	6491	29565	7183	43239	68.4%
Modulating	9032	3624	17930	30586	58.6%

Table 3.10: Per-state precision showing the reliability of each predicted state. Columns represent the predicted states, while rows show the actual states of these predictions. For example, of the 80,277 frames predicted as transitory, 54,009 were correct (67.3% precision), while 10,710 were actually steady and 15,558 were actually modulating.

patterns in prediction reliability: predictions of steady states are the most trustworthy (68.4% precision), followed closely by transitory states (67.3% precision), while modulating state predictions are less reliable (58.6% precision). The lower performance in modulating state detection, shown by both metrics, suggests there is a confusion between modulating and transitory states, with 15,558 modulating frames misclassified as transitory, indicating a notable tendency to confuse modulating states with transitory ones. This confusion specifically highlights a key limitation of the current method: the algorithm does not model the differences in pitch level between consecutive PCUs, making it particularly challenging to distinguish between modulating and transitory states. This limitation represents a clear direction for future methodological improvements. In addition, Figure 3.12 displays the confusion matrix in a colorbar format after normalising the counts, which enhances the interpretability of the matrix, making it easier to discern the magnitudes of correct and incorrect classifications.

Comparison to the State-of-the-Art at Segment Level: Existing literature reveals a singular study by Gong et al. (2016), which develops a method for pitch contour segmentation based on elements, transitory, steady, and vibrato. Their method's

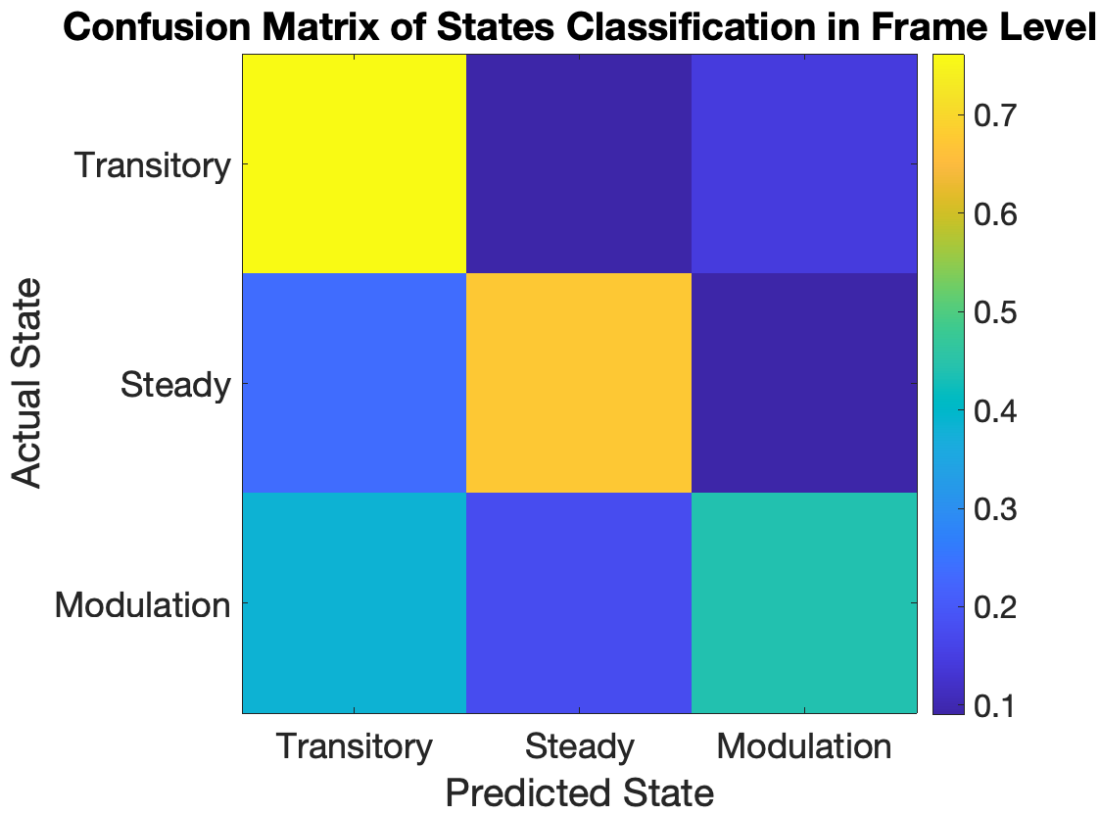


Figure 3.12: Colorbar confusion matrix of states classification: transitory, steady, and modulating. The scale is from 0 to 1, and the values are normalized from the counts in Table 3.9.

efficacy was evaluated using the Jingju dataset comprising 41 recordings. Three quarters of the data were allocated for training, facilitating parameter optimisation, while the remaining part served as the test set to assess segmentation accuracy. Section 3.1 details the selection of 12 recordings from the same dataset, with modifications to segment annotations conducted by the author of this thesis. For fairness in comparative analysis, metrics reported by Gong et al., based on their annotations, are utilised, whereas the proposed method's evaluation leverages the revised annotations.

In Table 3.11, the evaluation metrics are COnPOff and COnP, which are defined in the MIREX protocols (Downie et al. 2004). These measures assess the precision in detecting the start and end boundaries of state regions. COnPOff is the stricter of the two, accounting for the accuracy of the segment's onset time within a margin of 50 milliseconds, and offset time, which considers a 50 millisecond or 20% duration threshold relative to the ground truth, whichever is greater. In the absence of the state

label in Gong et al. (2016) and a requirement for pitch accuracy, the evaluation criteria have been simplified to two metrics: COnOff, which assesses the correctness of both the segment's onset and offset, and COn, which evaluates the accuracy of the segment's onset alone.

Method	**COnOff**			**COn**		
	F-measure	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**
Gong et al.	0.388	0.480	0.326	0.642	0.793	0.539
Proposed	**0.527**	**0.557**	**0.523**	**0.720**	**0.795**	**0.740**

Table 3.11: Comparative results of pitch contour segmentation using COnOff and COn metrics.

Precision, recall and F-measure are defined as Equations 3.11, 3.12, and 3.13:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.11}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.12}$$

$$F\text{-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.13}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Although the comparison in table 3.11 suggests the proposed method achieving higher performance metrics than the previous approach, these results must be interpreted with some caution as they are derived from different, though related, test sets —both sourced from the same corpus of Jingju recordings. It is noted that replicating the method of Gong et al. would indeed be the ideal approach to ensure an equitable comparison. However, given that the algorithm is outdated and difficult to replicate, relying on their reported results offers a practical alternative. This comparison approach enables an assessment of the proposed method's performance against established benchmarks, which, despite not being based on identical test sets, still offers valuable insights into relative efficacy.

3.3.2 Evaluation on Portamento Detection

Evaluation Metrics for Portamento Detection: The evaluation of the portamento detection method is multi-faceted, incorporating both frame-level and segment-level metrics. At the frame-level, precision, recall, and F-measure are employed. In addition, two types of accuracy are considered. The first is the conventional accuracy A_p, which is defined in equation 3.14:

$$A_p = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.14}$$

The second type, accuracy (A'_p), focuses solely on the correct detections and is given by equation 3.15, which was proposed by Dixon (2000), aiming to exclude the influence of true negatives (TN) to provide a more focused assessment of the model's performance in identifying portamento instances.

$$A'_p = \frac{TP}{TP + FP + FN} \tag{3.15}$$

At the segment-level, except F-measure of COnOff and COn as used in this section, additional metrics that assess segmentation errors are untilised, including "Merged" errors, "Split" errors, "Spurious" state regions, and "Non-detected" errors which are defined by Molina, Barbancho, Tardón & Barbancho (2014). A "Merged" error means the multiple ground truth state regions are merged into one region in the detection, while a "Split" error is the opposite. A "Spurious" state region error occurs when a detected state region does not overlap in time with any ground truth state region, while a "Non-detected" error is the opposite. The measures of these metrics are the proportion of all the ground truth portamenti which meet this error (except "Spurious" error which is the proportion of all the detected portamenti).

Ablation Experiment The ablation experiments are designed to compare the original pitch contour element detection method (see Section 3.2.3) with the finetuned approach for portamento detection proposed in Section 3.2.4, particularly focusing on the efficacy of the decay module introduced during finetuning. In the finetuning process, the decay in observation probability is triggered under two conditions: when observed

duration is less than 0.1 seconds or its extent is less than 0.5 semitones.

This dual-threshold approach is grounded in both empirical data and the theoretical framework of Beijing Opera's musical scale. The 0.5 semitone threshold for extent, corresponding to the minimum pitch interval in the scale commonly used in Beijing Opera, is substantiated by the investigation of Beijing Opera modes by Li & Li (2006). Additionally, the study in Section 6.2 in Yang (2017) provides statistical support for these thresholds, as shown in the histogram of pitch distribution, where the distance between any pairs of peaks is larger than one semitone, and the histogram envolopes of portamento duration indicate a rarity of durations below 0.1 seconds. The deciding of duration threshold also aligns with Mauch et al. (2015) for discarding short notes. The decay levels are set empirically at -5 for duration and -10 for extent (see Equation 3.8). This distinction emphasises the greater influence of extent over duration in the accurate identification of portamento in the context of this data and singing style.

The evaluation of the portamento detection methods demonstrates a marked improvement when using the fine-tuned approach. Table 3.12 and 3.13 summarise the results, highlighting the enhanced accuracy of the fine-tuned method compared to the original in frame-level and segment-level. Notably, there is a trade-off indicated by a slight increase in the Non-detected rate of the fine-tuned method, suggesting a more conservative detection strategy.

Method	A_p	A'_p	**Precision**	**Recall**	**F-Measure**
Original	0.68	0.38	0.41	**0.86**	0.55
Fine-tuned	**0.78**	**0.44**	**0.51**	0.76	**0.60**

Table 3.12: Comparison of original and fine-tuned methods on frame-level evaluation metrics.

Method	**COnOff(↑)**	**COn(↑)**	**Split(↓)**	**Merged(↓)**	**Spurious(↓)**	**Non-detected(↓)**
Original	0.37	0.41	0.08	**0.01**	0.66	**0.11**
Fine-tuned	**0.50**	**0.55**	**0.03**	**0.01**	**0.48**	0.22

Table 3.13: Segment-level evaluation metrics for original and fine-tuned methods. An upward arrow (↑) indicates that a higher value is better, while a downward arrow (↓) indicates that a lower value is better.

Comparison to the State-of-the-Art at Frame Level: Only one study, which is in the chapter four of Yang (2017), developed a method to do portamento detection.

This method is developed based on the same portamento dataset this thesis used. The difference is that they choose k-fold cross-validation to test the portamento detection method, in which approach k-1 parts of the data are used to train the model, while the method proposed by this thesis is tested on the whole dataset. From Table 3.14, the proposed method outperforms Yang et al's method in all metrics.

Method	A'_p	Precision	Recall	F-Measure
Yang et al.	0.35	0.39	0.72	0.44
Proposed	**0.44**	**0.51**	**0.76**	**0.60**

Table 3.14: Comparison of proposed method and Yang's method on frame-level evaluation metrics.

3.3.3 Evaluation on Steady Region Detection

Evaluation Metrics for Steady Region Detection: The evaluation of the steady region detection method adopts the same metrics on evaluation on portamento detection, applied at both the frame level and the segment level. The approach incorporates accuracy, precision, recall, and F-measure at the frame level. At the segment level, the metrics "COnOff", "COn", "Split", "Merged", "Spurious", and "Non-detected" are utilised.

Ablation Experiment: The ablation experiments are designed to compare the original pitch contour element detection method (see Section 3.2.3) with the finetuned approach for steady region detection proposed in Section 3.2.4. The hyperparameters used to finetune the model are set based on the empirical knowledge of the vocal style. The parameters M and m in the decay function detailed in Equation 3.9 is set at 1.065 and 0.73 semitones, respectively, based on the 213 cents and 146 cents for maximal and minimal pitch intervals within a scale, as investigated by Rosenzweig et al. (2020) for the dataset of Erkomaishvili's recordings. Additionally, the steady-state self-transition probability is heightened to 0.99. Lastly, the minimal duration threshold for detecting a steady region in polyphonic singing is set at 0.15 seconds, which exceeds the 0.1-second threshold typically used in monophonic singing. While these parameters are specifically set for Georgian music, the model maintains flexibility to accommodate

different musical cultures through parameter adjustment according to their respective theoretical foundations.

The evaluation results as presented in Tables 3.15 and 3.16 indicate that each module within the fine-tuning approach contributes to the overall enhancement of steady region detection performance, both at the frame level and segment level. The three modules are Observation Probability Decay (OPD), Steady Self-Transition Increase (SSTI), and Postprocessing Removing Short Region (PRSR).

Frame-level analysis reflects the basic classification accuracy of steady versus non-steady states for individual time points. At this level, all module combinations achieve similar F-Measure scores around 0.89, with the full combination (OPD, SSTI, and PRSR) and the OPD-PRSR pair reaching 0.892, and PRSR alone achieving 0.890. These marginal differences in frame-level performance suggest that this metric may not fully capture the musical relevance of the detected regions.

More importantly, segment-level analysis evaluates the musical coherence of the detected steady regions by considering their temporal continuity and boundaries. At this level, the PRSR module distinctly excels, evidenced by achieving the highest ConOff of 0.371 and Con of 0.522. In contrast, the OPD and SSTI modules negatively impact ConOff and Con, although they reduce the errors related to Split, Merged, and Non-detected events.

Overall, the PRSR module emerges as the most impactful. The OPD module's role is ostensibly to complement the PRSR module by diminishing split errors. It is suggested to reconsider the SSTI module's inclusion due to its negligible or even detrimental impact on detection performance.

Comparison to the State-of-the-Art: Table 3.17 compares the performance of the two proposed methods with two steady region detection methods by Rosenzweig et al. (2019) on the same dataset. Specifically, the analysis focuses on the OPD+PRSR and PRSR techniques to benchmark against the results reported in Rosenzweig et al. (2019). This comparison underscores the precision, recall, and F-measure values for each method, providing an overview of their effectiveness in steady region detection. The highest F-measure for each dataset is highlighted in bold. Regarding the F-measure,

Modules	A_p	A'_p	**Precision**	**Recall**	**F-Measure**
OPD+SSTI+PRSR	0.837	0.805	0.840	0.953	**0.892**
OPD+SSTI	0.815	0.787	0.811	0.965	0.880
OPD+PRSR	0.837	**0.806**	0.840	0.954	**0.892**
SSTI+PRSR	0.835	0.799	0.855	0.926	0.888
OPD	0.817	0.790	0.813	**0.966**	0.882
SSTI	0.819	0.788	0.823	0.949	0.881
PRSR	**0.838**	0.802	**0.856**	0.929	0.890
None	0.821	0.790	0.826	0.950	0.882

Table 3.15: Steady region detection performance at frame-level of different combinations of finetune modules

Modules	**COnOff(↑)**	**COn(↑)**	**Split(↓)**	**Merged(↓)**	**Spurious(↓)**	**Non-detected(↓)**
OPD+SSTI+PRSR	0.339	0.500	0.086	0.086	0.115	0.083
OPD+SSTI	0.266	0.399	0.119	0.086	0.425	**0.046**
OPD+PRSR	0.344	0.503	**0.075**	0.090	0.115	0.081
SSTI+PRSR	0.360	0.513	0.104	**0.035**	0.114	0.095
OPD	0.274	0.406	0.110	0.090	0.409	0.047
SSTI	0.272	0.397	0.155	**0.035**	0.414	**0.046**
PRSR	**0.371**	**0.522**	0.100	0.037	**0.107**	0.093
None	0.284	0.412	0.141	0.037	0.403	0.047

Table 3.16: Steady region detection performance at segment-level of different combinations of finetune modules

the methods introduced by Rosenzweig et al. (2019) slightly outperform those proposed in this thesis on recordings 001, 087, and 110.

ID	γ_{Morph}			γ_{Mask}			OPD+PRSR			PRSR		
	P	R	F	P	R	F	P	R	F	P	R	F
001	0.82	0.94	**0.88**	0.82	0.94	**0.88**	0.77	0.96	0.85	0.80	0.93	0.86
002	0.94	0.85	0.89	0.93	0.87	**0.90**	0.89	0.90	**0.90**	0.89	0.87	0.88
010	0.87	0.92	**0.89**	0.84	0.95	**0.89**	0.83	0.95	**0.89**	0.84	0.93	0.88
087	0.88	0.98	**0.93**	0.87	0.98	0.92	0.86	0.99	0.92	0.88	0.97	0.92
110	0.90	0.96	**0.93**	0.88	0.97	0.92	0.85	0.97	0.90	0.86	0.95	0.90

Table 3.17: Comparison of methods γ_{Morph}, γ_{Mask} proposed by Rosenzweig et al. (2019), and OPD+PRSR, and PRSR across Precision (P), Recall (R), and F-measure (F), with the highest F-measure values in each row highlighted in bold. The first column is the ID of the recording.

3.3.4 Evaluation on Vibrato Detection

This section compares the performance of the proposed method with the FDM-based vibrato detection method by Yang, Rajab & Chew (2017) on the same dataset introduced in Section 3.1. Table 3.18 compares the performance of two methods at the

frame level. The HMM demonstrates a higher accuracy in the frame level evaluation, as reflected in the A'_p, Precision, Recall, and F-Measure values. Table 3.19 compares the performance of two methods on segment level. The FDM method exhibits superior performance in terms of the "Spurious" and "Non-detected" metrics, indicating a lower spurious rate of detected vibrato and a lower non-detected rate of ground truth vibrato, respectively. These results suggest that the FDM method is particularly more effective in identifying vibrato segments than HMM. On the other hand, the HMM outperforms FDM particularly in the metrics of COnPOff, COnP, denoting a higher accuracy in detecting the correct onset and offset of vibrato, a better precision in characterising vibrato. Interestingly, although FDM has a better spurious rate and non-detected rate, this did not translate into a higher F-Measure at the frame level, which may be due to its lower precision of onset and offset of vibrato compared to HMM.

Method	A_p	A'_p	**Precision**	**Recall**	**F-Measure**
FDM	**0.80**	0.39	0.62	0.52	0.56
HMM	0.78	**0.44**	**0.70**	**0.55**	**0.60**

Table 3.18: Comparison of methods FDM and HMM on frame level across two types of accuracy, precision, recall, and F-measure, with the highest values in each column highlighted in bold.

Method	**COnOff(↑)**	**COn(↑)**	**Split(↓)**	**Merged(↓)**	**Spurious(↓)**	**Non-detected(↓)**
FDM	0.05	0.10	0.04	0.04	**0.34**	**0.20**
HMM	**0.17**	**0.25**	**0.00**	**0.02**	0.48	0.44

Table 3.19: Comparison of methods FDM and HMM on vibrato level across COnOff in F-Measure, COn in F-Measure, Split (Split rate of ground truth vibrato), Merged (Merged rate of ground truth vibrato), Spurious (Spurious rate of detected vibrato), Non-detected (Non-detected rate of ground truth vibrato), with the best values in each column highlighted in bold.

3.4 Conclusion

This chapter proposes a novel pitch contour segmentation method that addresses the limitations of previous research by enabling cross-cultural vocal music analysis. The concept of the Pitch Contour Unit (PCU) was introduced to segment and characterise pitch contours. By formalising the PCU, this thesis offers a novel approach that bridges

the gap between note-based methods and frame-based methods. Utilising the duration and extent of PCU sequences as input, a Hidden Markov Model (HMM) is employed to detect the primary elements of pitch contours: steady, modulating, and transitory elements.

The results from the confusion matrix analysis and frame-level accuracy metrics demonstrate the effectiveness of the proposed method. The fine-tuning of HMM parameters specifically for portamento and steady region detection in a Jingju dataset and a Georgian dataset further highlights the versatility and robustness of this approach. Comparisons with state-of-the-art methods at the frame level show that our proposed methods achieve comparable but generally slightly lower F-measures across the test recordings.

Future research could extend this work by exploring the application of PCUs and HMM-based analysis to other vocal music styles. Additionally, further refinement of the model could improve the robustness and accuracy of the proposed methods to reduce the confusion between transitory and vibrato elements.

Utilising the method proposed in this chapter, the next chapter will delve into note-level pitch contour analysis across two datasets related to Russian and Alpine vocal traditions. The analysis of various ornaments will be based on the pitch contour elements detected through this method.

Chapter 4

Note-Level Pitch Contour Analysis

This chapter presents a comparative analysis of note-level pitch contours in Alpine and Russian singing. Section 4.1 introduces the datasets used, consisting of two versions of note-level segmentations from recordings transcribed by two experts. Section 4.2 details both an automatic note segmentation method and a manual note segmentation approach applied in building the dataset. Section 4.3 evaluates the automatic note segmentation and examines the consistency between the two versions of manually annotated note segments in each culture, considering the importance of reliable segmentation as the ground truth for this chapter. Two characteristics of note annotation are used for the comparison: note types and note boundary displacements. In the analysis of note boundary displacements, two key concepts are defined for musical notes: the "held region" and the "transitional region." In the following sections, to compare singing styles between two cultures, the concepts of held region and transitional region are used again. Section 4.4 characterises held regions and extracts features for comparison. Section 4.5 characterises transitional regions and extracts features for comparison.

The primary aim of this exploratory study is to demonstrate the use of a computational framework for note-level pitch contour analysis across different cultures, focusing on entirely different songs rather than different versions of the same song. Nevertheless, this study does not aim to test any musicological hypothesis regarding the singing style

of the two cultures. It is crucial to bear in mind that these explorations of vocal styles focus on a specific dataset from Proutskova et al. (2023) for each culture and do not fully represent all aspects of Alpine and Russian vocal traditions.

4.1 Dataset Overview

The VocalNotes dataset, as detailed by Proutskova et al. (2023), encompasses audio recordings alongside annotations of vocal performances from five diverse musical traditions: Japanese Min'yo, Chinese Hebei Bangzi opera, Russian traditional singing, Alpine yodel, and Jewish Romaniote chant. Each tradition is represented by approximately ten minutes of audio coupled with comprehensive metadata regarding the origin of the song excerpts. Other annotations, meticulously performed by two or three experts per tradition, comprise f0 data, along with independent onset, offset, and pitch information for each note.

This chapter demonstrates the proposed computational methods by analysing Russian traditional singing and Alpine yodel. These two traditions were selected because yodel offers a clean and simple melody, and there is expertise available in Russian traditional singing, making them more suitable for the detailed pitch contour analysis employed in this study. In contrast, the other traditions incorporate more complex musical elements, which complicates the analysis. Tables 4.1 and 4.2 provide the data statistics of the selected recordings. Variations in note counts between two annotators from the same culture reflect the inherent subjectivity in note segmentation

Audio Filename	Song title	Performer	Duration (s)	Location	LS	YW
Franz Lustenberger Entlebuch.wav	De Schratte zue	Franz Lustenberger	26.204	Entlebuch	26	28
Ehrler - Juuz.wav	Schwyzerjuuz	Paul Ehrler	44.431	Schwyz	91	98
Sophie Brunner.wav	Solojodel	Sophie Brunner	68.023	Appenzell	65	69
Dr Braemiser.wav	Dr Braemiser	Beny Betschart	64.639	Muotatal	98	98
Juz Paul Fetz.wav	Juz	Paul Fetz	27.133	Vorarlberg	54	56
Juez Sepp Schneider.wav	Juez	Sepp Schneider	18.895	Vorarlberg	31	20
hechobeA.wav	Hech Obe	Ruedi Rymann	20.997	Obwalden	30	31
hechobeB.wav	Hech Obe	Ruedi Rymann	14.472	Obwalden	21	23

Table 4.1: Metadata table for Alpine songs. Note: LS and YW are initials for the names of the annotators. The two columns indicate the annotated note count of each annotator.

Audio Filename	Song title	Performer(s)	Duration (s)	Location	OV	PP
Da_po_zoriushke_1.mp3	Da po zoriushke	Basova Tatiana Timofeevna	40.565	Kursk, Russia	72	68
Da_po_zoriushke_2.mp3	Da po zoriushke	Lamanova Maria Antonovna	39.741	Kursk, Russia	92	90
Da_po_zoriushke_3.mp3	Da po zoriushke	Khodosova Daria Semenovna	40.281	Kursk, Russia	97	91
Da_po_zoriushke_4.mp3	Da po zoriushke	Motorykina Ekaterina Illarionovna	39.776	Kursk, Russia	93	90
Kak_letala_jara.mp3	Kak letala jara	Britikova Anna Afanasievna	32.067	Pskov, Russia	98	72
Milyj_moj_zhalkij.mp3	Milyj moj zhalkij	Pilant Natalia Osipovna	31.643	Pskov, Russia	26	48
Neumyvataja.mp3	Neumyvataja	N.Kalosha	27.777	Briansk, Russia	56	54
Oj_kumushki.mp3	Oj kumushki	Sergeeva Olga Fedoseeva	25.420	Pskov, Russia	47	53
Uzh_ja_dumala_2.mp3	Uzh ja dumala	Eliseeva Pelageja Sidorovna	41.169	Tver, Russia	60	72
Zhil_byl_Lazar.wav	Zhil byl Lazar	Koroleva Maria Vasilievna	18.907	Kursk, Russia	61	59

Table 4.2: Metadata table for Russian songs. Note: OV and PP are initials for the names of the annotators. The two columns indicate the annotated note count of each annotator.

4.2 Note Segmentation Methods

The note segmentation process assigns meaning to sounds by emphasising certain perceptions or interpretations while disregarding others. The results are largely influenced by the transcription's purpose. For example, analytical transcriptions may be very detailed, while transcriptions intended for performers might only include essential information for reading the score, assuming familiarity with the style. Additionally, segmentation can focus on various aspects such as note boundaries, pitch, mode, linguistic elements (like vowels), rhythm, dynamics, and vocal style. Often, focusing on one aspect can detract from others. Therefore, this study will mention the transcription's purpose and the characteristic focused on when describing the note segmentation method.

The pitch of the recording is instrumental in note segmentation. Using a computational approach, an automated pitch curve estimation is followed by manual correction. The pitch curve visually represents the singer's pitch trajectory over time, where pitch values are estimated for each audio frame based on PYIN's probabilistic framework, which computes pitch probabilities and uses a Hidden Markov Model to optimise the pitch value sequence. Tony software (Mauch et al. 2015), utilising the PYIN algorithm (Mauch & Dixon 2014), provides automated pitch curve estimation. This algorithm returns candidate pitches with probabilities, selecting the highest probability across the entire track using an HMM. A digital interface is available for manual correction of errors, such as octave mistakes, by selecting different candidates. In rare cases, such as unclear pitch sounds, PYIN may struggle to provide meaningful candidates. Taking the estimated and corrected pitch curve, note segmentation can be approached both automatically and manually.

4.2.1 Automatic Note Segmentation Approach

The details of this automatic note segmentation approach are documented in Li et al. (2021). This proposed automatic note segmentation method considers acoustic features while not limiting itself to domain knowledge of a specific musical tradition in terms of

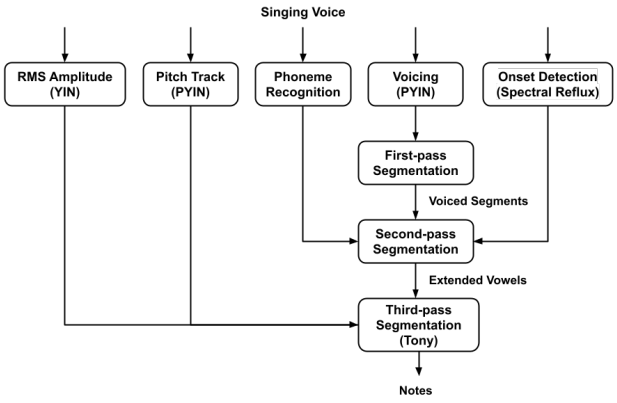


Figure 4.1: The proposed three-step note segmentation method.

language, rhythm, vocal style, and mode. Based on the annotation approach of Molina, Barbancho, Tardón & Barbancho (2014), this method assumes that note boundaries can be categorised into four types: (1) the beginnings and ends of voiced segments; (2) phonetic changes; (3) pitch changes[1]; and (4) amplitude changes. The four types of note boundaries are detected, and the vocal track is segmented using a three-step cascading approach which produces successively finer segmentations at each step (Figure 4.1).

In Step 1, voiced segments (segments of continuous pitch activity) are determined based on the PYIN pitch track. In Step 2, the voiced segments are further segmented based on phonetic change, to create what are termed *extended vowel* regions, which are defined in the following paragraph. In Step 3, extended vowel segments are further divided based on pitch and amplitude changes using the algorithm from PYIN. The main novelty of this approach is the incorporation of phonetic information into an existing framework for note segmentation through the introduction of the second step, which addresses "soft" onsets and offsets. These occur when two adjacent notes are smoothly connected without obvious pitch and loudness variations. In most cases, however, there is a phonetic change between notes.

In order to detect phonetic change, the phonemes are automatically transcribed and temporally aligned using the state-of-the-art speech transcription system by Xu et al. (2021). The Spectral Reflux onset detection function proposed by Sapp (2006) is then used to fine-tune the note boundaries. To detect note boundaries more reliably,

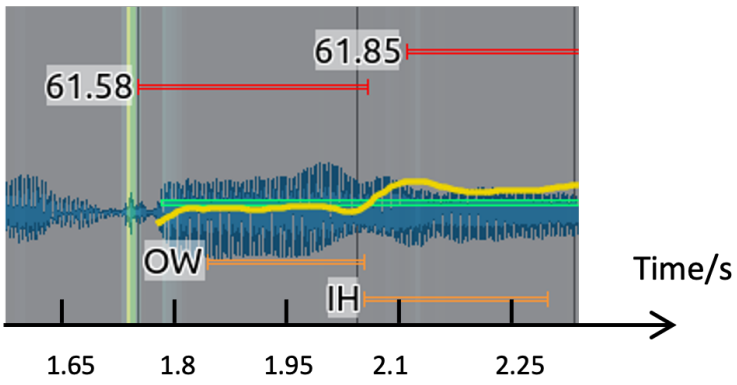
[1]PYIN (Mauch et al. 2015) is followed in setting the threshold of pitch change required for a note boundary to $\frac{2}{3}$ of a semitone.

the phonetic output is fine-tuned with a simple additional signal processing step. First, the phonemes are categorised into vowels and consonants, determining the inter-vowel regions. The inter-vowel regions are then expanded by 50 ms on each side to account for the system's boundary accuracy tolerance. Finally, the maximum of spectral reflux in the expanded inter-vowel region determines the exact note boundary.

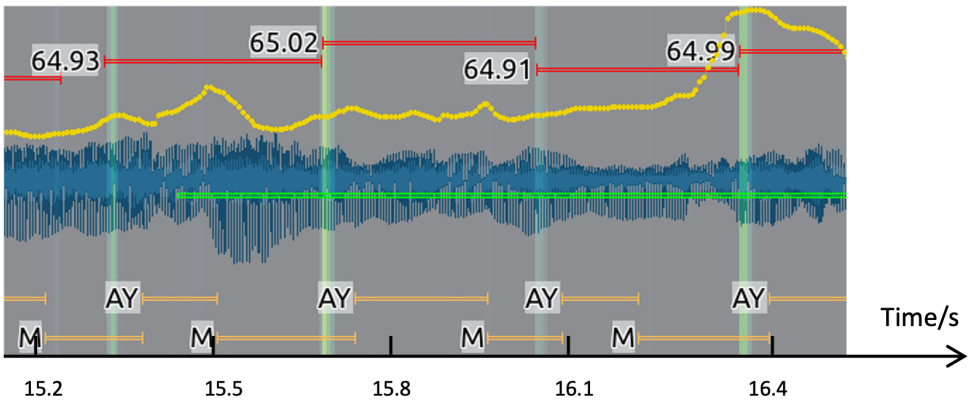
Although Demirel et al. (2020) developed a system specifically for phoneme-level lyrics transcription, which improved note segmentation in Li et al. (2021), it was trained only on English and is no longer accessible. The speech transcription system (Xu et al. 2021) instead leverages the wav2vec 2.0 model (Baevski et al. 2020), pretrained on 53 languages using self-supervised learning. This model, fine-tuned for phoneme recognition across multiple languages, maps phonemes from training to target languages during inference using articulatory features. A beam-search decoder with an integrated language model generates the phoneme sequences, enabling effective transcription of unseen languages without task-specific modifications.

Figure 4.2 illustrates the need for this step, showing examples where Tony makes the systematic error of under-segmentation of successive notes having continuous steady pitch tracks during note transitions. These instances occur generally when consecutive notes are sung either without any consonants or silent gaps (breathing, articulation, etc.), or with short voiced consonants between successive vowels. When there are two adjacent vowels with no gap in between (Figure 4.2a), the note boundary is determined by the timing of the vowel transition. For instances where there is a gap between consecutive vowels (Figure 4.2b), the note boundary is determined as the location of the local maximum of the spectral flux between the vowels in question.

Steps 1 and 2 detect inter-vowel note boundaries, but there are also note boundaries within vowels that are communicated via pitch and amplitude changes. In such cases, phoneme-based segmentation is expected to fail at determining the note boundaries. In order to retrieve the timings of such boundaries, the HMM-based segmentation method of Tony (Mauch et al. 2015) is applied within the extended vowel regions resulting from Steps 1 and 2. This HMM takes as input the pitch and amplitude estimates from PYIN and outputs discrete notes, based on Viterbi HMM-decoding. The HMM models pitches



(a) Adjacent vowels with similar pitches erroneously merged into a single note.



(b) Successive notes sung on similar pitches with voiced phonemes between vowels, resulting in multiple merge errors.

Figure 4.2: Examples of soft onset errors made by the Tony software in vocal tracks 'afemale2' and 'afemale4' from the dataset proposed by Molina, Barbancho, Tardón & Barbancho (2014). The waveform is shown in blue, the ground truth segmentation is in red, labelled with median pitch in semitones (MIDI). The pitch track from PYIN is yellow, the note region extracted by Tony is bright green, detected phoneme boundaries are orange, and spectral flux is represented by the brightness of vertical lines.

from B1 to C♯6 at three steps per semitone, and each pitch has three states representing its attack, stable part, and silence, respectively. The observation probabilities model the fact that the beginnings of notes and note transitions tend to vary more in pitch than the main, stable parts of notes, and the transition model favours continuity in pitch transitions.

4.2.2 Manual Approach

This section describes the existing manual annotation methodology developed by Proutskova et al. (2024), which provides the ground truth data for computational analysis in this chapter. In their approach, two transcribers from each culture are tasked with manually segmenting the music using the interface of Tony software. To mitigate differences in transcription purposes, within each team, transcribers may agree on a detailed transcription objective tailored to their specific repertoire. Once the initial common objective is agreed upon, transcribers work independently without discussing their transcriptions or the challenges encountered. To ensure consistency, teams must agree on a correct pitch curve before independent segmentation.

The transcription process benefits significantly from technical affordances and visualisation tools. Tony software provides features for creating, splitting and merging note segments, and adjusting note boundaries. Users can listen to the original recording, pitch curve, and note segments either simultaneously or individually. They can also observe the displayed waveform and the spectrogram, with the temporal resolution being easily adjustable to aid in determining note boundaries. The software includes looping mechanisms for repeated listening to specific passages, which is particularly useful when uncertain.

However, the segmentation is greatly influenced by technical support. For example, visualising the pitch curve can affect the interpretation of sounds. Another consideration is the number of times transcribers are allowed to listen to a fragment or context, as repeated listening can lead to new cognitive constructs. Constraints may be imposed on the length of the context in which a transcribed element should be heard. Tony allows for slowing down the recording, a technique commonly used in some eth-

nomusicological communities but one that can alter the perception of note boundaries. Therefore, despite having a predetermined segmentation objective, the segmentations produced by different transcribers are likely to be varied and flexible. This variability is influenced by personal perception and cognition, individual characteristics, the effects of the tools used, and the difficult or ambiguous decisions encountered during the process. Accordingly, the next section analyses the differences between annotations of different versions.

4.3 Analysis of Note Segmentation Characteristics

Note segments form the cornerstone of this note-level analysis and significantly impact the analysis results. Given that multiple versions of note segmentation are available and they can differ significantly, it is necessary to analyse the segmentation characteristics of different versions.

4.3.1 Evaluation of Automatic Note Segmentation

To evaluate the note segmentation made by the method proposed in Section 4.2.1, the manual annotations introduced in Section 4.2.2 are used as ground truth. The first and third tracks listed in Table 4.1 are excluded from the evaluation due to very few phonemes being recognised. Since two versions of manual annotations are available, two separate evaluations are conducted for each culture, with each evaluation using one version of the manual annotations as the ground truth to assess the other manual annotation and the automatic segmentation.

Five evaluation metrics are employed. “COnOff”, defined in the MIREX protocols (Downie et al. 2004), accounts for the accuracy of the note’s onset time within a margin of 50 milliseconds, and the offset time within either 50 milliseconds or 20% of the note’s duration relative to the ground truth, whichever is greater. As this study focuses solely on note segmentation rather than note transcription, it is not required to evaluate pitch accuracy. Additional metrics are used to assess segmentation errors, including “Merged” errors, “Split” errors, “Spurious” notes, and “Non-detected” errors, as defined

Metric	YW	Step1+2	Step1+3	Step1+2+3
COnPOff (F-measure) ↑	**0.68**	0.09	0.45	0.16
Split ↓	**0.04**	0.20	0.21	0.35
Merge ↓	**0.04**	0.61	0.09	0.07
Spurious ↓	**0.02**	0.28	0.04	0.03
Non-detected ↓	0.01	**0.00**	0.01	0.08

Table 4.3: Comparison of segmentation performance for different parts of the automatic note segmentation system and the manual annotations of YW on the Alpine dataset, using LS's manual annotations as ground truth. All measurements represent the mean values across all recordings. The arrows indicate whether a higher (↑) or lower (↓) value is better for each metric.

by Molina, Barbancho, Tardón & Barbancho (2014). A "Merged" error occurs when multiple ground truth notes are merged into a single note in the detection, while a "Split" error is the opposite. A "Spurious" note error is identified when a detected note does not overlap in time with any ground truth note, and a "Non-detected" error occurs when a ground truth note is not detected. These metrics are measured as the proportion of all ground truth notes that exhibit the respective error, except for the "Spurious" error, which is measured as the proportion of all detected notes.

The results are shown in Tables 4.3 to 4.6. For both the Alpine and Russian datasets, the manual annotations achieve the best performance, with the note segmentation from Step1+3 being the second best. The Step1+2+3 method results in the lowest split error among the three automatic segmentation across all four tables, but this comes at the cost of a high merged error. Although Li et al. (2021) demonstrated that phoneme segments improve note segmentation on the dataset in Molina, Barbancho, Tardón & Barbancho (2014), the poor performance of Step1+2+3 is not surprising. The phoneme segments transcribed by Xu et al. (2021) are inaccurate because the model was trained on speech rather than singing. In summary, the consistency between two versions of manual note segmentation is higher than that between manual and automatic segmentation. Therefore, automatic note segments will not be used as the ground truth for analysing singing style in this chapter.

Metric	LS	Step1+2	Step1+3	Step1+2+3
COnPOff (F-measure) ↑	**0.68**	0.09	0.51	0.20
Split ↓	**0.02**	0.21	0.21	0.32
Merge ↓	**0.08**	0.64	0.10	0.10
Spurious ↓	**0.04**	0.30	0.05	**0.04**
Non-detected ↓	0.01	**0.00**	0.02	0.09

Table 4.4: Comparison of segmentation performance for different parts of the automatic note segmentation system and the manual annotations of LS on the Alpine dataset, using YW's manual annotations as ground truth. All measurements represent the mean values across all recordings. The arrows indicate whether a higher (↑) or lower (↓) value is better for each metric.

Metric	PP	Step1+2	Step1+3	Step1+2+3
COnPOff (F-measure) ↑	**0.61**	0.10	0.38	0.16
Split ↓	**0.04**	0.49	0.07	0.16
Merge ↓	0.13	0.39	0.24	**0.07**
Spurious ↓	0.06	0.11	0.13	**0.05**
Non-detected ↓	**0.00**	**0.00**	0.03	0.10

Table 4.5: Comparison of segmentation performance for different parts of the automatic note segmentation system and the manual annotations of PP on the Russian dataset, using OV's manual annotations as ground truth. All measurements represent the mean values across all recordings. The arrows indicate whether a higher (↑) or lower (↓) value is better for each metric.

Metric	OV	Step1+2	Step1+3	Step1+2+3
COnPOff (F-measure) ↑	**0.62**	0.08	0.29	0.10
Split ↓	**0.06**	0.52	0.07	0.15
Merge ↓	0.08	0.36	0.23	**0.07**
Spurious ↓	0.04	0.10	0.12	**0.03**
Non-detected ↓	0.05	**0.01**	0.04	0.10

Table 4.6: Comparison of segmentation performance for different parts of the automatic note segmentation system and the manual annotations of OV on the Russian dataset, using PP's manual annotations as ground truth. All measurements represent the mean values across all recordings. The arrows indicate whether a higher (↑) or lower (↓) value is better for each metric.

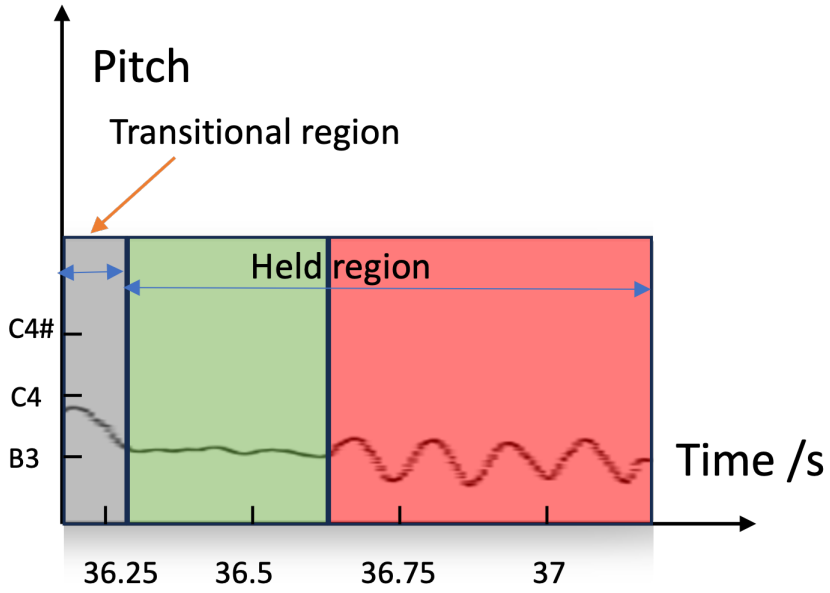


Figure 4.3: Illustration of held and transitional regions. The black curve is the pitch contour, the coloured blocks are pitch contour elements (red: modulating; green: steady; grey: transitory) and the double arrows indicate the held region and transitional region.

4.3.2 Comparison of Manual Note Segmentation

To delve into the distinct tendencies of the two manual segmentation versions of each culture, this section elucidates the distribution of types of transcribed notes alongside the analysis of note boundary demarcation from different transcribers. Two pivotal concepts, namely held and transitional regions, are introduced to facilitate categorising note types and quantifying note boundary locations. Held regions are identified as segments where the pitch stability of a note is maintained by the singer. These regions consist of either a single steady element or modulating element described in Chapter 3, or a mix of both of them within a note's span. Each element was set to be longer than 50 ms empirically. Conversely, transitional regions - bridging two distinct notes or marking the commencement or conclusion of a note - are composed of individual or a series of transitory elements along with brief steady and modulating elements (shorter than 50 ms). For visual reference, Figure 4.3 presents an example. All pitch contour element segments are automatically estimated via the algorithm proposed in Chapter 3 and subsequently undergo manual verification and adjustments by the author.

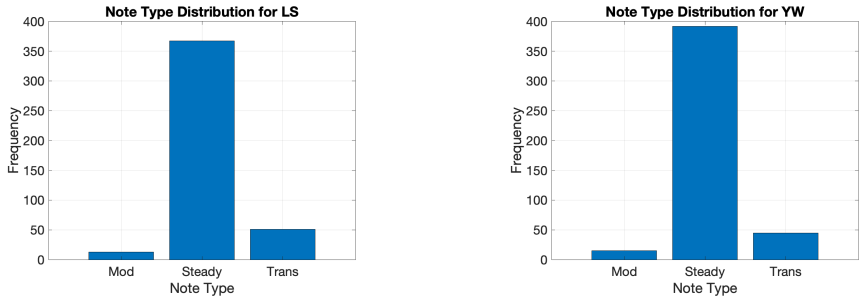


Figure 4.4: Distribution of note types in Alpine data as annotated by transcribers

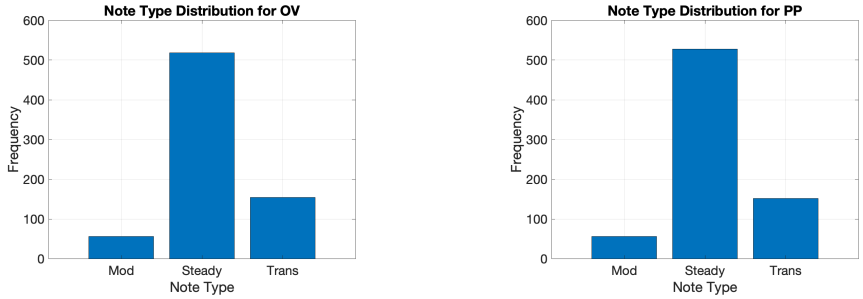


Figure 4.5: Distribution of note types in Russian data as annotated by transcribers

Transcribed Note Type Distribution

Note types are categorised into three distinct classes: steady-dominant notes, modulation-dominant notes, and transitory-dominant notes. The analysis of the distribution of these note types provides insight into the tendencies of a transcriber, to reveal specific patterns and preferences in how notes are transcribed.

- *Steady-dominant note*: Characterised by a held region, where the longest individual pitch contour element is steady.
- *Modulation-dominant note*: Characterised by a held region, where the longest individual pitch contour element is modulating.
- *Transitory-dominant note*: A note without any held region or with one or more held region but the longest individual pitch contour element is transitory.

Figure 4.4 and 4.5 present the distribution of different types of note annotations (steady-dominant, modulation-dominant, transitory-dominant) as annotated by two transcribers within the Alpine and Russian dataset.

The distributions of note types show similarities between annotations from two annotators within the same dataset, with only minor differences observed. Annotator YW tended to annotate slightly more steady-dominant notes and fewer transitory-dominant notes compared to Annotator LS. Similarly, Annotator PP favoured a higher number of steady-dominant notes and fewer transitory-dominant notes than Annotator OV. To determine if there is a significant difference between two categorical distributions, this study performs a chi-squared test on the Alpine and Russian datasets separately. The chi-squared test results for the Alpine dataset are $\chi^2(2, N = 883) = 0.842, p = 0.656$, and for the Russian dataset, $\chi^2(2, N = 1464) = 0.065, p = 0.968$. These p-values suggest that any observed differences in their annotations are likely due to chance, rather than indicating a systematic difference in annotation styles.

Note Boundary Displacement Distribution

Note boundary analysis focuses predominantly on how boundaries mark the onset and offset of steady-dominant and modulation-dominant notes. Crucially, the investigation probes into whether note boundaries are placed within transitional regions around the note and assesses the extent of inclusion of such regions. Figure 4.6 provides more details.

Several metrics have been established to facilitate this analysis:

- *Note boundary displacement: Dis* quantifies the displacement between the onset or offset (t_{on} or t_{off}) and the transition points (t_{tr}). Onset displacement (Dis_{on}) and offset displacement (Dis_{off}) are given by Equation 4.1 and 4.2.

$$Dis_{on} = t_{tr} - t_{on} \tag{4.1}$$

$$Dis_{off} = t_{off} - t_{tr} \tag{4.2}$$

- *Note boundary displacement proportion of transitional region:* This is denoted by Dis_{pro}, measuring the extent of displacement relative to the duration of the

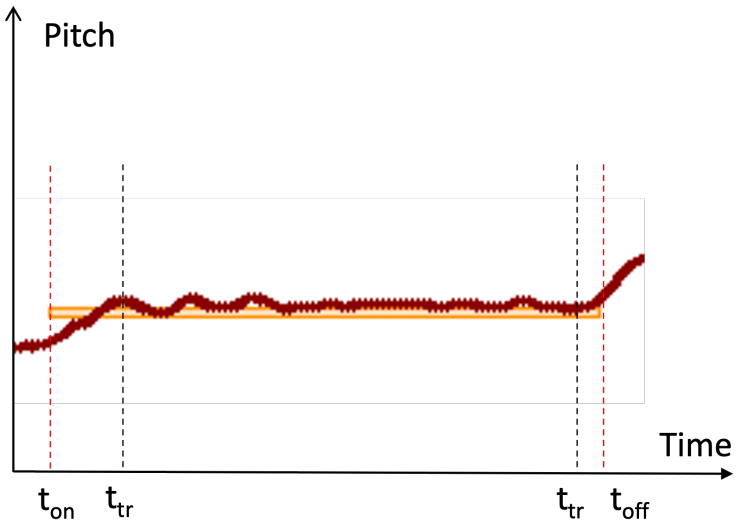


Figure 4.6: Illustration of note boundary analysis. The red curve is a pitch contour, the orange bar is a note segment, the black vertical lines indicate the transition points which connect the transitional regions and the held region, and the red vertical lines indicate the annotated onset and offset of the note.

associated transitional regions (d_{trans}). This proportion is expected to be greater than 0 and can exceed 100% if the note boundary extends beyond the transitional region boundary. Dis_{pro} is defined by:

$$Dis_{pro} = \frac{Dis}{d_{trans}}, \quad Dis \geq 0 \tag{4.3}$$

The distribution of onset and offset displacements, their proportion of transitional regions and the relationship between the proportion and transitional region durations are illustrated here. For Alpine data, figures 4.7 and 4.8 illustrate the displacements for note onset and offset across different segmentation versions. The analysis indicates that the distributions of onset and offset displacements annotated by transcribers LS and YW are right-skewed. Notably, LS's distribution is more balanced compared to YW's, suggesting that although their annotations consistently cover the entire held region within a note, LS is more stringent than YW.

To compare the distributions of annotations between two transcribers, A two-sample Kolmogorov-Smirnov test is conducted, as it is suitable for continuous distributions and does not assume the data follows any specific distribution. The test statistic

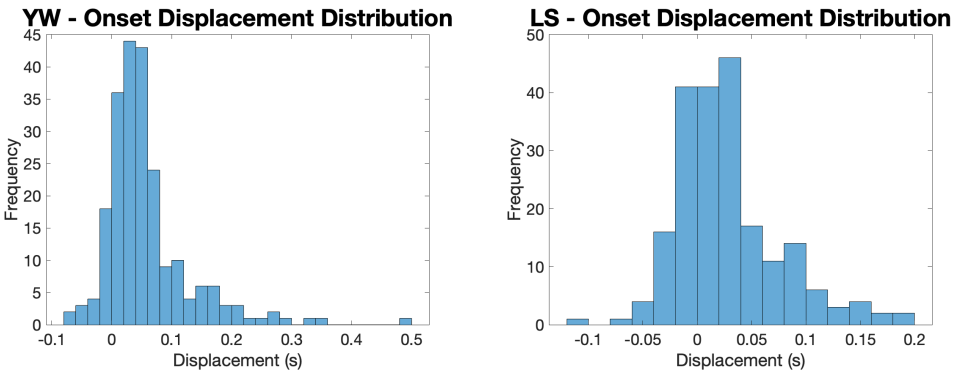


Figure 4.7: Comparative analysis of onset displacement for Alpine data

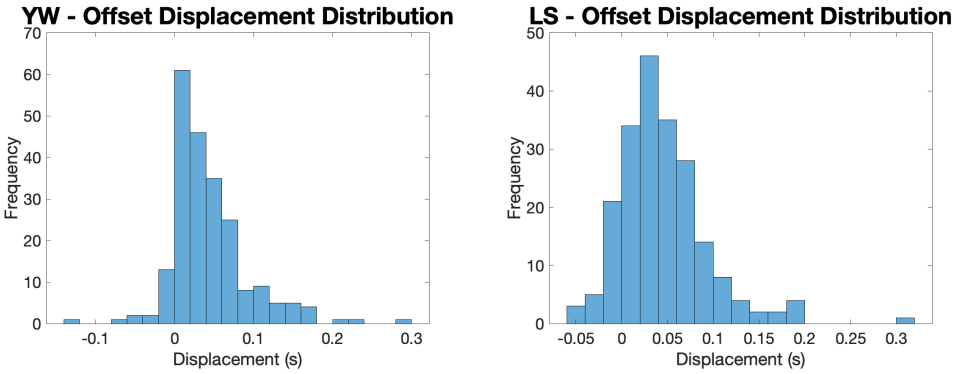


Figure 4.8: Comparative analysis of offset displacement for Alpine data

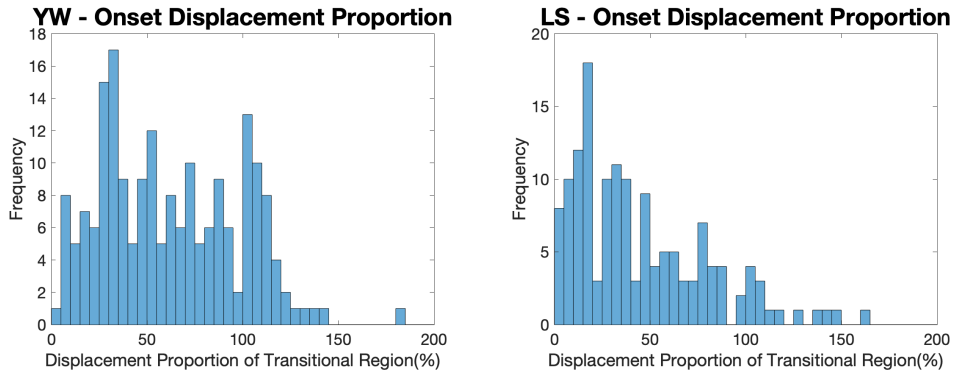


Figure 4.9: Comparative analysis of onset displacement proportion for Alpine data

are: $D(380, 407) = 0.252, p < .001$ for onset and $D(380, 407) = 0.095, p > .05$ for offset, indicating a significant difference for onset, while, for offset displacements, no significant difference between the two transcribers.

Figures 4.9 and 4.10 depict the proportions of onset and offset displacements within the transitional region. LS's annotations for onset demonstrate a mode around 20% followed by an exponential decay, whereas YW's distribution exhibits a bimodal pattern with modes at 30% and 100%. For offset annotations, YW shows a mode at 100% displacement, highlighting a tendency to position offsets at the end of the transitional region. In contrast, LS's data shows a mode around 25% with a secondary mode at 100%, reflecting a more varied approach. The two-sample Kolmogorov-Smirnov test statistics are: $D(380, 407) = 0.241, p < .001$ for onset and $D(380, 407) = 0.121, p > .05$ for offset. These results indicate that the two transcribers have systematically different approaches to marking note onsets: LS tends to place onsets late in the transition region, while YW varies between early (to include portamento at note beginnings) and late placements. However, their approaches to marking note offsets are more similar, showing no statistically significant difference.

Figure 4.11 and Figure 4.12 display scatter plots that elucidate the relationship between the duration of transitional regions and onset or offset displacements. The distributions of each variable are presented marginally. The visualisations reveal observable trends, specifically, an increase in displacements corresponding to longer durations of transitional regions, as indicated by the positive slopes in the linear fitting lines. However, the linear models exhibit limited efficacy in capturing the data's variability, as

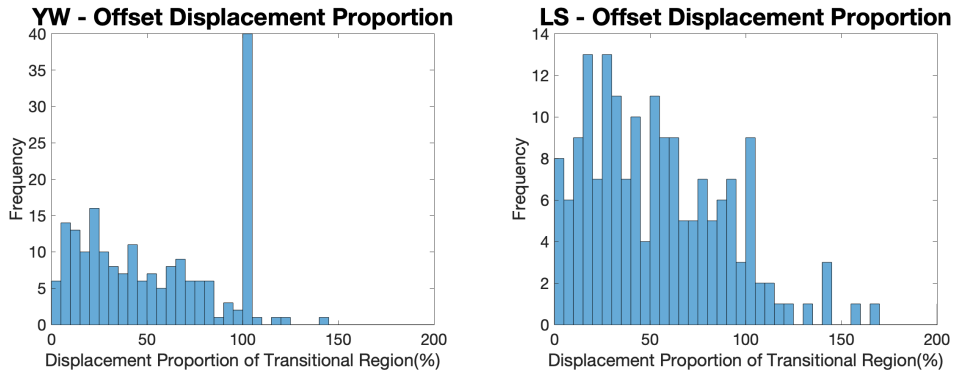


Figure 4.10: Comparative analysis of offset displacement proportion for Alpine data

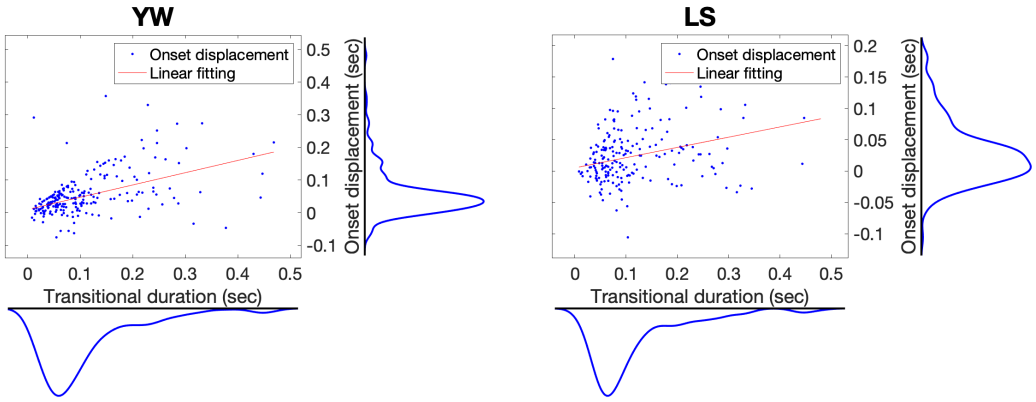


Figure 4.11: Comparative analysis of scatter plots for onset displacement for Alpine data

reflected by the low R-squared values: 0.11 and 0.10 for LS, and 0.26 and 0.09 for YW, respectively. These values suggest that linear models may not adequately describe the complex relationships present within the data. To measure the relationship between the variables more robustly, given the potential skewness or non-normality in the marginal distributions, Spearman's Rank Correlation was chosen. This non-parametric method, advantageous for its indifference to data distribution assumptions, reveals moderate to strong positive correlations: Spearman coefficients of 0.29 (onset) and 0.35 (offset) for the LS, and 0.53 (onset) and 0.42 (offset) for the YW, substantiate the observed trends. YW demonstrates higher coefficients with 0.53 for onset and 0.42 for offset, suggesting a more consistent approach to marking note boundaries relative to the transitional durations. In contrast, LS records lower coefficients of 0.29 for onset and 0.35 for offset, indicating a potentially more flexible annotation style.

Figures 4.13 and 4.14 present the distributions of onset and offset displacements

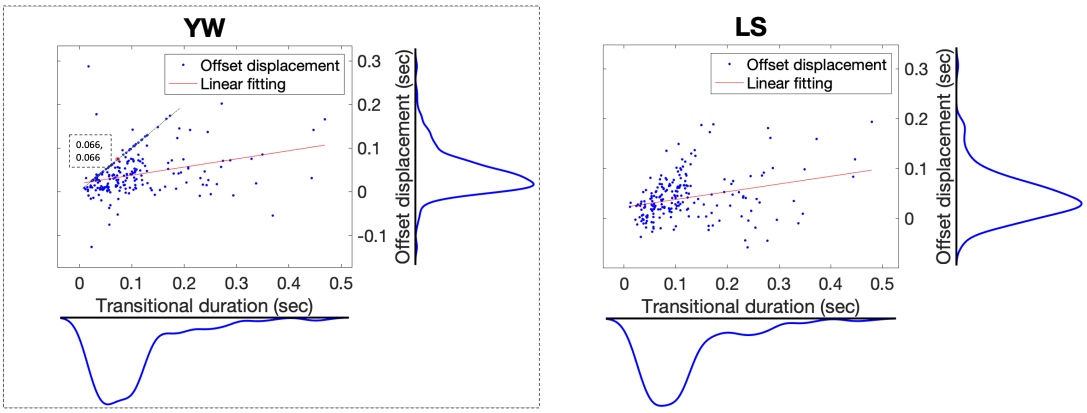


Figure 4.12: Comparative analysis of scatter plots for offset displacement for Alpine data. In YW's annotations, a notable diagonal constraint appears in the upper-left region where points follow a line with approximately equal x and y values (as shown by the point (0.066, 0.066) marked with a red dot). This pattern suggests that note offset displacements are constrained by their corresponding transitional durations for most time in YW's offset annotation, indicating that offsets are typically placed within, not beyond, the transitional region's ending point.

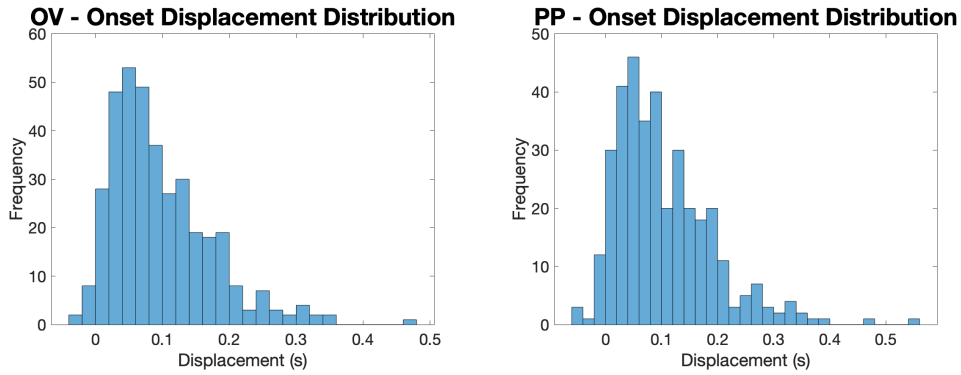


Figure 4.13: Comparative analysis of onset displacement for Russian data

in the Russian dataset, as annotated by two different transcribers, OV and PP. Interestingly, despite the fact that they conducted their note annotations completely independently, the patterns displayed by both transcribers are remarkably similar. The two-sample Kolmogorov-Smirnov test statistics are: $D(574, 584) = 0.054, p > .05$ for onset and $D(574, 584) = 0.081, p > .05$ for offset. These values indicate that there is no statistically significant difference in the onset and offset displacements between the two transcribers' annotations. This similarity in annotation styles, despite the lack of an intentional agreement, underscores the potential for implicit shared understanding of note segmentation in vocal data between the two annotators.

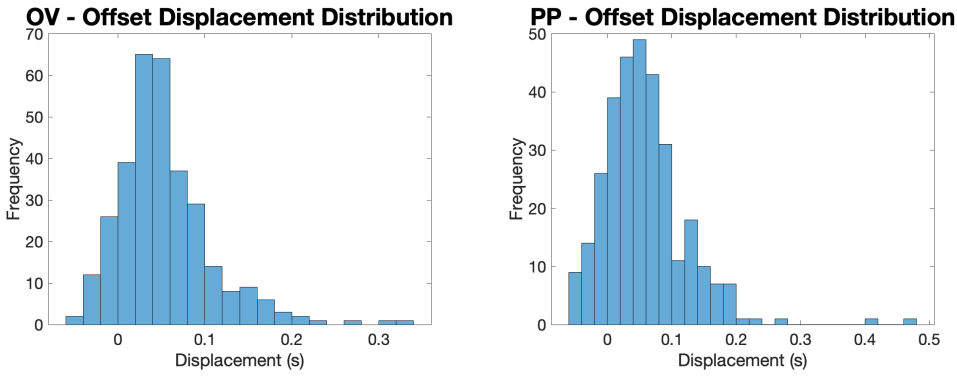


Figure 4.14: Comparative analysis of offset displacement for Russian data

The distributions of both onset and offset displacements are right-skewed. This skewness indicates a general tendency for both transcribers to mark the onset and offset beyond the held region. However, the offset displacements show a more symmetric distribution than the onset displacements. This suggests a more stringent approach when marking the offsets, as indicated by the lower spread of values compared to onsets.

Figures 4.15 and 4.16 present the proportions of onset and offset displacements within the transitional region. The onset displacement proportions for both transcribers predominantly concentrate at 100%, indicating that they often place note onsets at the starting point of the transitional region. The offset displacement proportions, however, exhibit multi-modal patterns, indicating a more variable annotation style for offsets. Particularly, OV shows a stronger tendency to place offsets at the endpoint of the transitional region compared to PP (see the peak at 100% position). The two-sample Kolmogorov-Smirnov test statistics are: $D(574, 584) = 0.103, p = 0.047$ for onset and $D(574, 584) = 0.113, p > .05$ for offset. These results shows that the differences in onset and offset marking styles between OV and PP are significant for proportions of onset displacements, and not significant for proportions of offset displacements.

Figure 4.17 and Figure 4.18 show scatter plots that examine the relationship between the durations of transitional regions and the associated onset and offset displacements, with the distributions of each variable depicted marginally alongside the main plot. The visualisations of onset displacement identify clear trends where displacements increase as the transitional durations extend, while the offset displacement does not. To

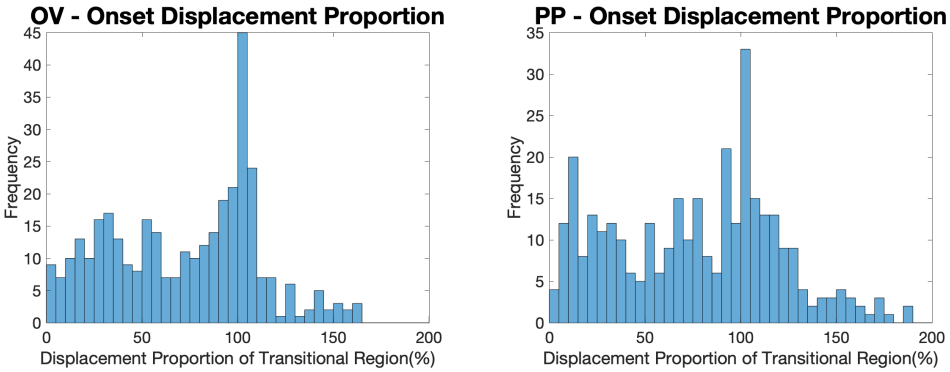


Figure 4.15: Comparative analysis of onset displacement proportion for Russian data

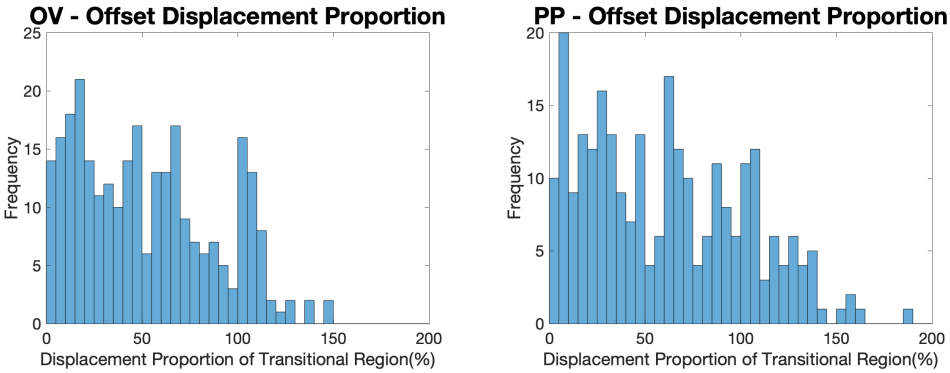


Figure 4.16: Comparative analysis of offset displacement proportion for Russian data

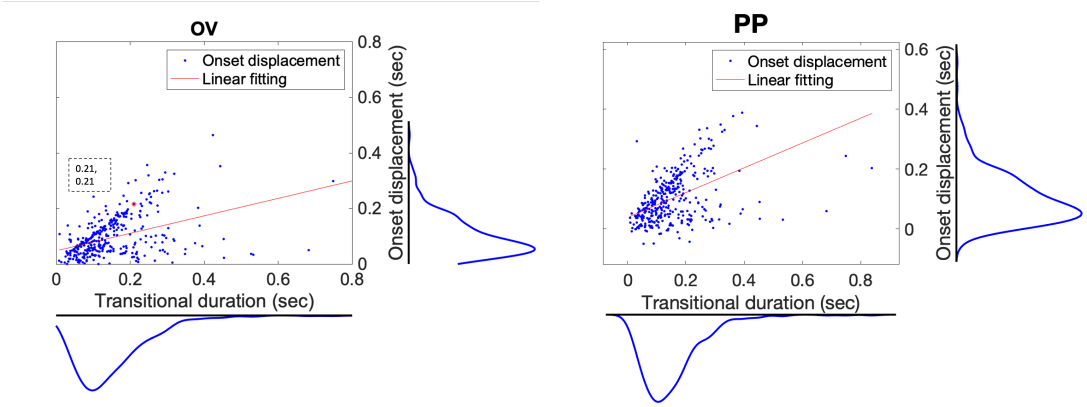


Figure 4.17: Comparative analysis of scatter plots for onset displacement for Russian data. In OV's annotations, a notable diagonal constraint appears in the upper-left region where points follow a line with approximately equal x and y values (as shown by the point (0.21, 0.21) marked with a red dot). This pattern suggests that note onset displacements are constrained by their corresponding transitional durations for most time in OV's onset annotation, indicating that onsets are typically placed within, not beyond, the transitional region's starting point.

demonstrate the linear relationship, linear fitting and Spearman's Rank Correlation are employed. The linear models show limited capacity to encapsulate the full variability of the data, with R-squared values indicating a relatively poor model fit: 0.11 for onset and 0.01 for offset in OV, alongside 0.22 for onset and nearly -0.003 for offset in PP. Spearman coefficients of 0.42 (onset) and 0.20 (offset) for OV, and 0.46 (onset) and 0.17 (offset) for PP confirm the observed trends. These results highlight that transcribers in the Russian dataset adopt a relatively consistent approach in annotating note onsets relative to transitional durations, while showing more flexibility in offset annotation.

4.3.3 Conclusion

In conclusion, although there is no significant difference in note types between the different versions of note annotations, there may still be significant differences in note boundary markings. For instance, discrepancies are observed in the onset markings between two transcribers in the Alpine data, as well as in the proportions of onset displacements within the transitional region in the Russian data. Additionally, the transitional region can influence note boundary markings differently for different individuals, such as the two transcribers in the Alpine dataset.

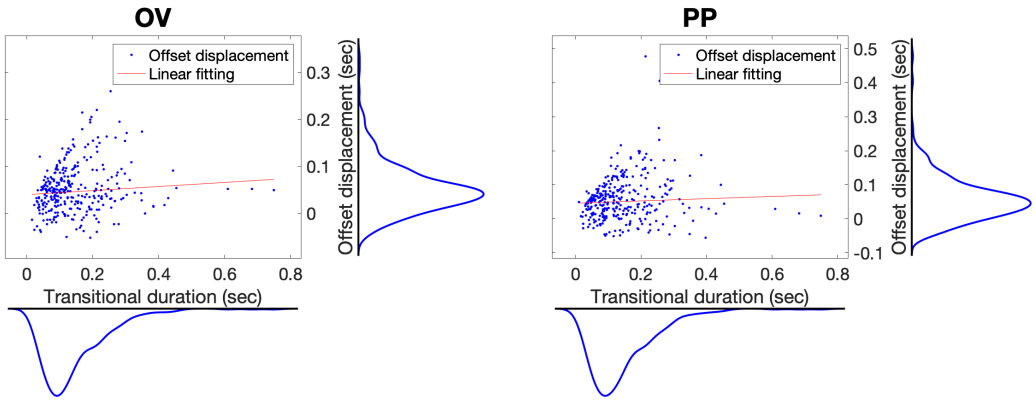


Figure 4.18: Comparative analysis of scatter plots for offset displacement for Russian data

This suggests that note boundaries are set subjectively, and comparative analysis of singing styles based on note boundaries could be influenced by the transcriber. To mitigate this influence, the held regions and transitional regions of musical notes can be used to set more robust boundaries when comparing two singing styles. The next two sections compare Alpine and Russian singing styles in terms of the held region and the transitional region, respectively.

4.4 Held Region Analysis

Held regions include both steady and modulating elements, each characterised by distinct features. The steady element is analysed through several measures: the slope, estimated via linear fitting to diminish the effects of the endpoints of f0 (as depicted in Figure 4.19); the duration; the instability, measured by the variance of the f0 values; and the median of the f0. The modulating element is characterised by regularity, mean and evolution of vibrato rate and extent, instability and slope of carrier, duration and overall pitch. The method proposed by Wen & Sandler (2008) (see details in Section 2.5.7) is utilised to demodulate the original modulating element signal into modulator and carrier (see Figure 4.20 as an example). The regularity of the modulator, mean vibrato rate and extent of the modulating element, along with the instability and slope of the carrier, are estimated. Regularity is quantified using the maximum value of the autocorrelation coefficient of the modulator, excluding the value at time zero, as pro-

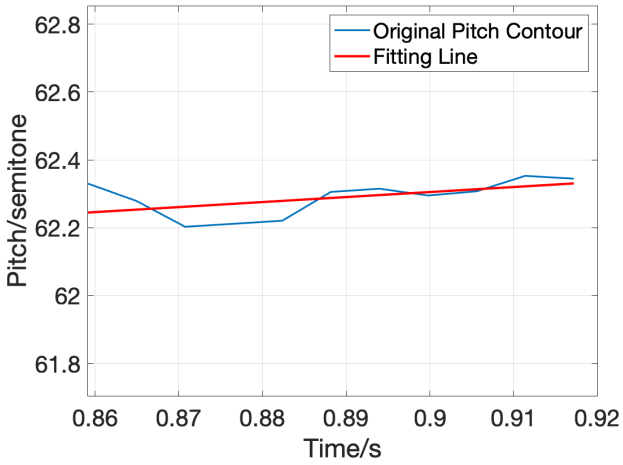


Figure 4.19: Linear fitting for pitch contour of a steady element

posed by Wen & Sandler (2008). This method is selected because it does not limit the modulator to ideally conforms to any specific function, such as a sinusoid. Following Yang et al. (2013), which assumes that the interval between one peak and one trough of the pitch curve represents a half cycle of the modulating element, the rate and extent of each half cycle are calculated. The overall vibrato rate and extent are then calculated as the average across these half cycles.

To investigate how vibrato rate and extent change over time within a modulating element, this study proposes using the Discrete Cosine Transform (DCT) on the sequences of the rate and extent of half cycles. The evolution is measured by the 1st to 7th order coefficients of the DCT. The details of the DCT have been introduced in Section 2.5.7, where the DCT is applied to pitch slides, which are sequences of pitch data.

4.4.1 Analysis Results of Steady Elements in Held Region

This section compares the distribution characteristics of slope, instability, and duration of the steady elements in held regions between Alpine and Russian data. To get a clear visual comparison between two groups of data, distributions are obtained by applying kernel density estimation or exponential fitting on the histogram. Additionally, this section explores the interrelationships among these parameters by regarding the pitch change, which is represented by the vertical distance from one end of the linear regres-

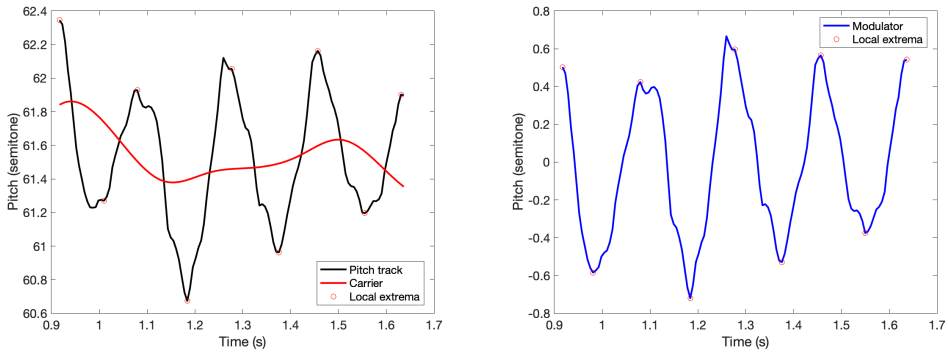


Figure 4.20: An example of demodulation of a modulating element. The graph on the left displays the original pitch track and the carrier signal, with local extrema points marked. The graph on the right illustrates the pitch contour of the modulator with local extrema. In this representation, semitone 69 is set as the reference point A4, equivalent to 440 Hz.

sion line to the other, and instability as dependent variables and duration and median pitch as independent variables, to provide deeper insights into singing behaviour.

Figure 4.21 displays the characteristics of steady elements within the Alpine and Russian vocal data in this analysis. The leftmost plot, which contrasts the slope of steady elements, indicates that both Alpine and Russian distributions are symmetrically centred around zero, forming a bell-shaped curve. This pattern implies a dominant singing style that upholds a relatively unchanging pitch. The Alpine data exhibits a denser distribution with a more pronounced peak, suggesting a higher prevalence of flat steady elements compared to the expansive curve observed in the Russian data. The two-sample Kolmogorov-Smirnov test statistic is : $D(812, 1177) = 0.279, p < .001$, indicating significant differences between Alpine and Russian.

The second plot, which focuses on the instability of steady elements, reveals an exponential distribution for both styles, with a majority of values clustering towards lower instability. This pattern infers that both Alpine and Russian singers generally maintain a steady pitch, yet the Alpine samples display marginally less variability. The two-sample Kolmogorov-Smirnov test statistic is : $D(812, 1177) = 0.225, p < .001$, indicating significant differences.

The third plot, which shows the duration of steady elements, underscores a right-skewed distribution for both styles, with short durations of less than 0.5 seconds being

more prevalent. The Russian distribution peaks sharply around 0.25 seconds, while the Alpine distribution has a broader range, implying that Alpine singers in this data might utilise a wider array of steady durations. The observed difference is notable, with a two-sample Kolmogorov-Smirnov test statistic of $D(812, 1177) = 0.313, p < .001$.

Collectively, these distributions suggest that while both the Alpine and Russian singing styles exhibit a tendency towards a steady pitch and short steady elements, the Alpine data might display a more stable pitch of steady elements.

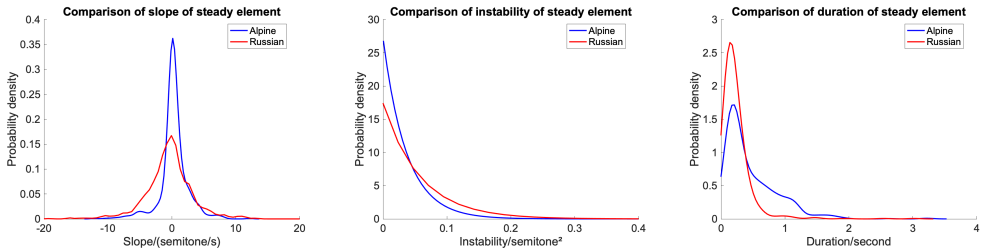


Figure 4.21: Comparative analysis of features of steady elements

Figures 4.22, 4.23, 4.24, and 4.25 present scatter plots with marginal distributions of variables, providing comparative analyses of the relationship between discussed characteristics of steady elements in Alpine and Russian vocal data. The red lines in the scatter plots represent linear fits to the data, yet these yield R-squared values below 0.1. Such low values suggest that the linear model fails to offer a reliable or meaningful explanation of the relationship between the variables. Furthermore, visual inspection of the scatter plots reveals non-linear relationships between variables. Given these characteristics, this analysis considers the Spearman correlation coefficients instead. Figures 4.22 and 4.23 display scatter plots of slope versus pitch and instability versus pitch, respectively. For the Alpine data, the Spearman Correlation Coefficients are -0.11 (p-value < 0.01) for slope and 0.01 for instability, while the Russian data exhibit Spearman Correlation Coefficients of 0.01 for slope and -0.06 (p-value < 0.05) for instability. Although statistically significant correlations were observed for Alpine slope (= -0.11, p < 0.01) and Russian instability (= -0.06, p < 0.05), the small magnitude of these coefficients suggests very weak relationships between pitch level and both slope and instability characteristics in both Alpine and Russian steady elements. Furthermore, Figures 4.24 and 4.25 present the relationship between steady characteristics and

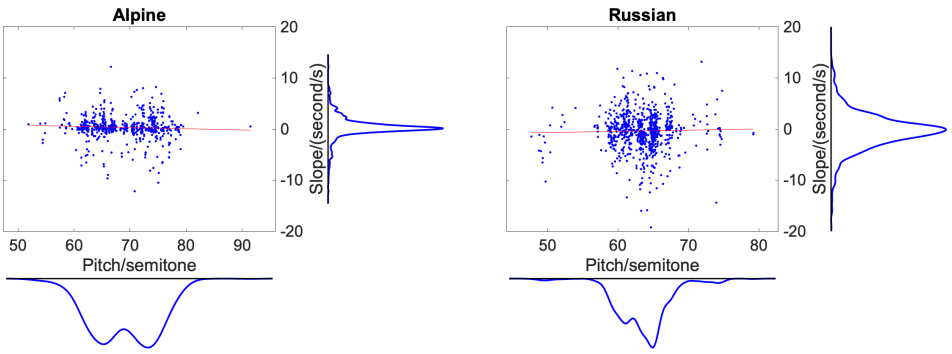


Figure 4.22: Comparative analysis of relationship between slope and pitch of steady elements

the duration of steady elements. For the Alpine data, the Spearman Correlation Coefficient is 0.04 for pitch change versus duration, and 0.16 (p-value < 0.001) for instability versus duration, indicating a subtle trend. Conversely, the Russian data show Spearman Correlations of -0.05 for pitch change and 0.34 (p-value < 0.001) for instability, suggesting a slight relationship. These results indicate that while pitch change shows no significant correlation with duration, instability exhibits a weak positive correlation with the duration of steady elements in both vocal styles, with the correlation being notably stronger in Russian (0.34) than in Alpine (0.16) data, suggesting that Alpine steady elements maintain more consistent pitch stability as duration gets longer compared to Russian steady elements. The correlation between duration and instability raises an important question: whether this increased variability in longer notes stems from motor control limitations or represents deliberate expressive choices by singers. This distinction requires further investigation.

4.4.2 Analysis Results of Modulating Elements in Held Regions

Figure 4.26 presents a comparative analysis of the mean vibrato rate and vibrato extent between Alpine and Russian vocal data. The graph on the left depicts the probability density of the vibrato rate, measured in Hz. The Alpine style is characterised by two peaks around 6 and 8.5 Hz, indicative of rapid vibrato rates. On the other hand, the Russian style displays a distinct peak around 7.5 and 10 Hz, suggesting a faster vibrato rate. The observed difference is notable, with a two-sample Kolmogorov-Smirnov test

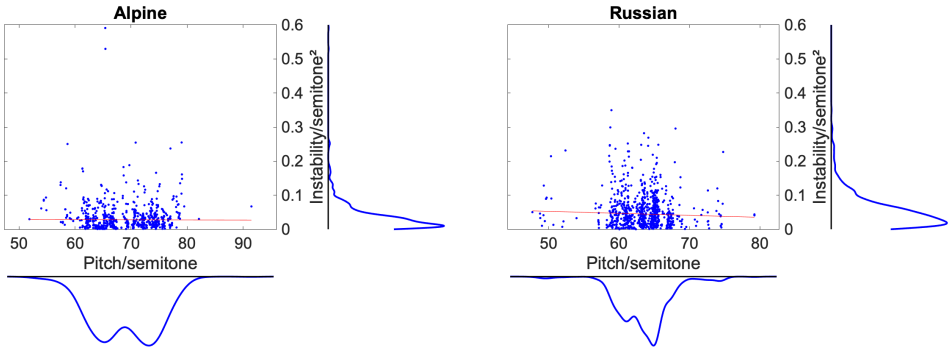


Figure 4.23: Comparative analysis of relationship between instability and pitch of steady elements

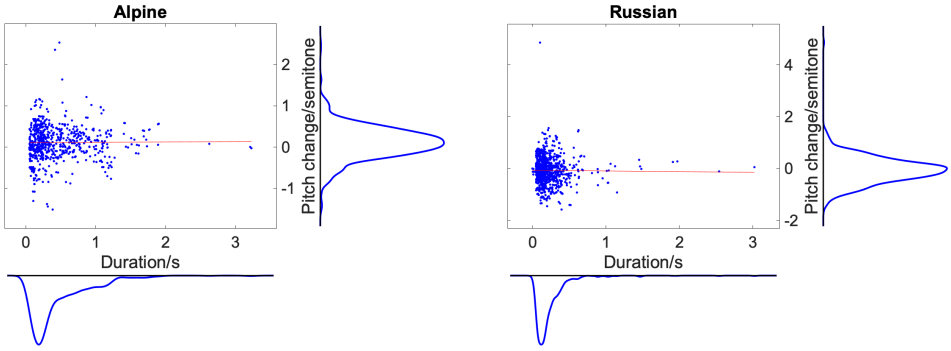


Figure 4.24: Comparative analysis of relationship between pitch change and duration of steady elements

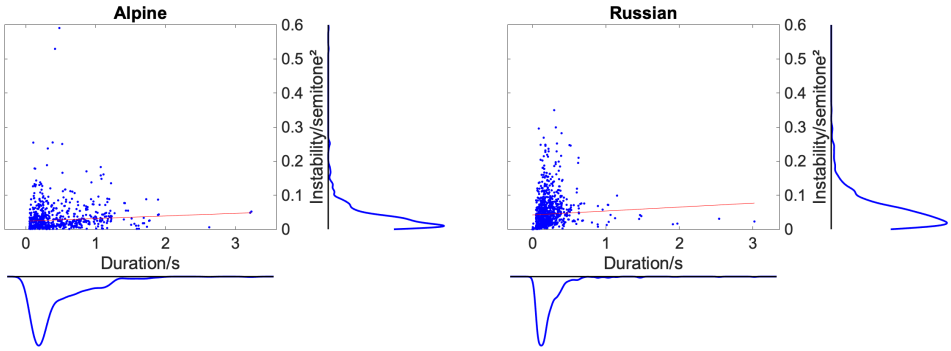


Figure 4.25: Comparative analysis of relationship between instability and duration of steady elements

statistic of $D(22, 51) = 0.506, p < .001$.

The graph on the right portrays the vibrato extent, measured in semitones. Here, the Alpine style exhibits a sharp peak at approximately 0.4 semitones, suggesting a more regulated vibrato extent. Conversely, the Russian style presents a broader, right-skewed curve with a pronounced peak near 0.1 semitones, implying a small yet diverse range in vibrato extent within this data set. The difference in vibrato extent between the two data is statistically significant, with a two-sample Kolmogorov-Smirnov test statistic of $D(22, 51) = 0.622, p < .001$.

Collectively, these visualisations and statistics indicate that the Alpine style tends to employ a quicker and more pronounced vibrato, while the Russian style leans towards a slower and less extensive vibrato.

Figure 4.27 and Figure 4.28 present a statistical analysis of the vibrato rate and extent evolution within a vibrato in Alpine and Russian vocal data through the application of DCT coefficients. Each set of box plots across the two panels corresponds to individual DCT coefficients, ranging from 1st to 7th order, illustrating the central tendency and spread of DCT coefficients of vibrato rates and extents within each vocal tradition.

In Figure 4.27, both Alpine and Russian datasets show a general trend of positive first and second coefficients, indicating an increase in vibrato rate over most of time and an overall concave shape. The Analysis of Variance (ANOVA) conducted on the first and second coefficients did not reveal significant differences between the two groups with a significance level as 0.05: for the 1st coefficient, $F(1, 71) = 1.03, p = 0.314$; for the 2nd coefficient, $F(1, 71) = 0.69, p = 0.407$.

Figure 4.28 illustrates the differences in vibrato extent evolution between Alpine and Russian singing styles. The Alpine data displays relatively stable DCT coefficient values, with minor variance around zero. The median of the 1st DCT coefficient leans slightly towards a positive skew, while the 2nd DCT coefficient shows a mild negative trend. This pattern indicates a subtle concave contour of vibrato extent evolution, characterised by a slight increase after the onset and a decrease towards the end.

In contrast, the Russian samples present a greater degree of variability in the first

two coefficients. The 1st DCT coefficient exhibits a positive skew, implying a decrease in vibrato extent. The 2nd DCT coefficient, displaying a variance closely aligned with the first, suggests a distinct bend in the vibrato extent change. Coefficients 3 through 7 maintain a relatively symmetrical distribution, with the variance progressively decreasing. This pattern suggests that these coefficients capture less noticeable nuances of vibrato extent evolution. The ANOVA yielded $F(1, 71) = 7.78, p = 0.007$ and $F(1, 71) = 0.33, p = 0.569$ for the first and second coefficients, respectively. These results indicate a statistically significant difference in the 1st DCT coefficient between the two singing styles, while the difference in the 2nd DCT coefficient is not statistically significant.

In summary, the two styles exhibit distinct characteristics in the evolution of vibrato rate and extent, with the Alpine style demonstrating more variability in rate and the Russian style showing more variability in extent. However, statistically significant differences were only found in the first DCT coefficient of the vibrato extent evolution.

Figure 4.29 presents an analysis of the regularity of the modulator, instability and slope of the carrier, as well as the duration of modulating elements in both Alpine and Russian singing styles.

The graph on the top left, illustrating the modulator regularity, shows that both Alpine and Russian singers generally produce regular modulator, with peaks around 0.96 (where 1 is the maximum). No significant difference is found with a two-sample Kolmogorov-Smirnov test statistic of $D(22, 51) = 0.242, p > .05$.

The instability measurements, depicted in the figure adjacent to the regularity graph, suggest that the carrier of vibrato in both styles are similar. Again, no significant difference is detected with a two-sample Kolmogorov-Smirnov test statistic of $D(44, 102) = 0.122, p > .05$.

The analysis of the slope of the carrier reveals that both styles predominantly avoid rapid pitch changes. However, the Alpine style exhibits a sharper peak at zero, implying a flatter baseline for modulation. A significant difference is indicated by a two-sample Kolmogorov-Smirnov test statistic of $D(22, 51) = 0.386, p = 0.014$.

The duration analysis, depicted in the figure at the bottom right corner, reveals sim-

ilar bimodal distributions in both styles, albeit with an approximate shift of 0.2 seconds. The Russian style typically centers around a narrower and shorter duration (peaking at approximately 0.4 seconds), while the Alpine style shows a preference for longer durations, peaking at around 0.8 seconds. The difference in duration is significant, as shown by a two-sample Kolmogorov-Smirnov test statistic of $D(22, 51) = 0.504, p < .001$.

Accordingly, these findings highlight distinct stylistic nuances in the use of vibrato. Alpine singers, as represented in the data, display slightly longer and more regular vibrato, with a carrier showing greater stability than that of Russian singers.

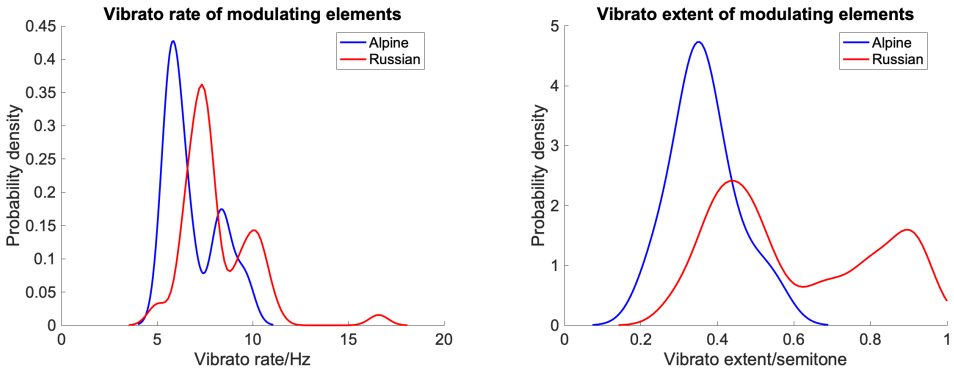


Figure 4.26: Comparative analysis of mean of vibrato rate and extent of modulator

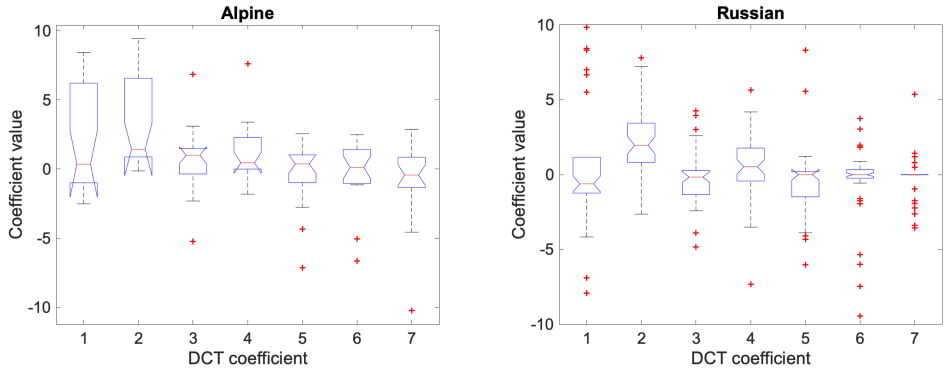


Figure 4.27: Comparative analysis of evolution of vibrato rate. The median is presented by the red line, the interquartile range is captured within the blue boxes, outliers are denoted by red plus signs, and the whiskers extend to capture the range of data points excluding the outliers. Each boxplot corresponds to an individual DCT coefficient. Higher absolute coefficient value indicates the component with higher energy.

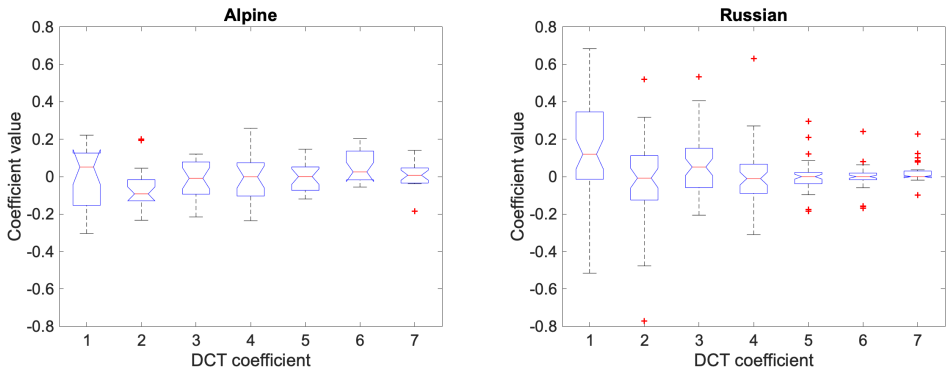


Figure 4.28: Comparative analysis of evolution of vibrato extent. The median is presented by the red line, the interquartile range is captured within the blue boxes, outliers are denoted by red plus signs, and the whiskers extend to capture the range of data points excluding the outliers. Each boxplot corresponds to an individual DCT coefficient. Higher absolute coefficient value indicates the component with higher energy.

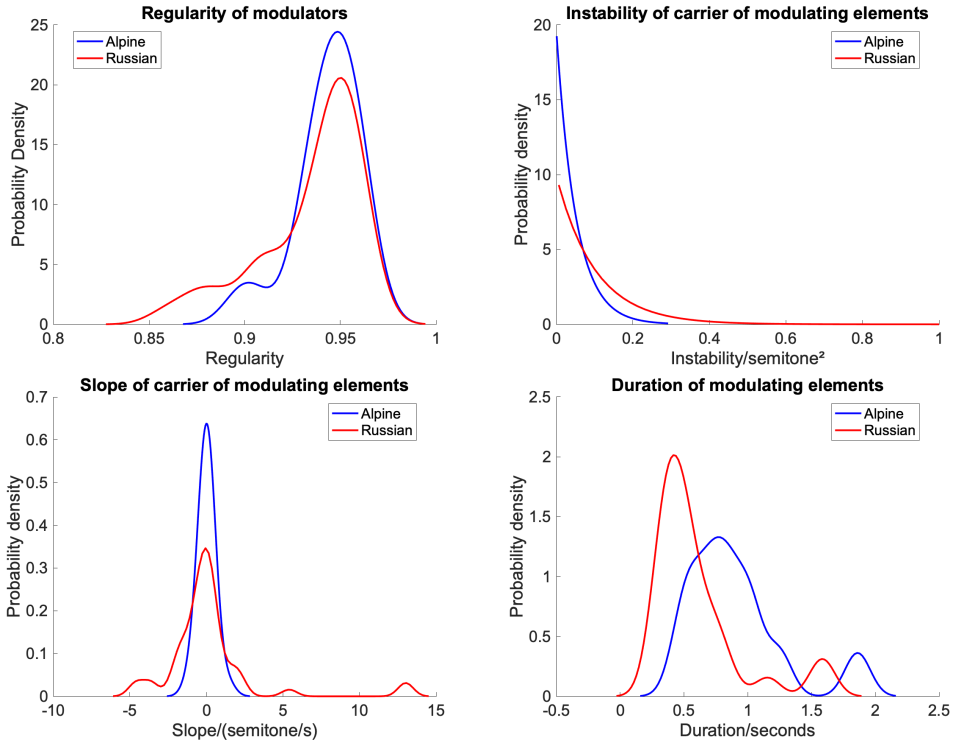


Figure 4.29: Comparative analysis of modulator regularity, carrier properties, and modulating duration

4.5 Transitional Region Analysis

4.5.1 Transitional Region Characterisation

This study presents automated approaches for detecting and identifying musical ornaments in transitional regions, focusing on four main types: glissando, portamento, miscellaneous slides, and mordent. Furthermore, for all pitch slides, including glissando, portamento, and miscellaneous slides, overshoot and preparation are considered. These ornaments are defined in Section 2.4.2 and illustrated in Figure 4.30. For each identified ornament, pitch contour features are extracted.

The criteria for detecting and identifying each type of ornament are defined below. The threshold for distinguishable note pitch differences is set according to Mauch et al. (2015). Subsequently, automated methods are developed based on these predefined rules to detect the four types of ornaments (glissando, portamento, miscellaneous slides, and mordent) as well as two subtypes, overshoot and preparation, which apply to the first three ornaments.

1. **Glissando:** Transitory elements that connect one held region or two held regions of two different notes with pitch difference larger than $\frac{1}{3}$ semitones. These elements are characterised by at least three consecutive elements:

    - A transitory starting element: The initial movement away from the first held region before the steady middle element (marked in red in Figure 4.30a).
    - A steady middle element: A brief region where the pitch momentarily stabilises, similar to a touch note introduced in Section 2.4.2.
    - A transitory ending element: The final movement leading to the second held region, which continues in the same direction as the transitory starting element.

2. **Portamento:** Identified by:

    - Transitory elements that connect two held regions of two different notes with pitch difference larger than $\frac{1}{3}$ semitones.

- The largest slide in the transitory elements should have direction which aligns with the direction of note progression.

3. **Mordent:** Defined by the following characteristics:

   - Transitory elements linking two held regions (this study does not consider mordents consisting a held region),
   - The pitch difference between neighbouring held regions is set to less than $\frac{1}{3}$ semitones.
   - The pitch of the transitory element deviates from the mean pitch of two held regions by more than $\frac{1}{3}$ semitones.

4. **Miscellaneous Slides:** Pitch slides that do not fall under glissando, portamento, or mordent.

5. **Overshoot:** Indicated by the pitch in the glissando, portamento or miscellaneous slides extending beyond the target note.

   - If the transition is from a higher pitch to a lower pitch, the lowest pitch in the transitional region should be lower than the target note's pitch, which is calculated as the median of the pitch in the target note's held region.
   - If the transition is from a lower pitch to a higher pitch, the highest pitch in the transitional region should be higher than the target note's pitch, which is calculated as the median of the pitch in the target note's held region.
   - The correction from the overshoot is defined slides occurring subsequent to the turning point, which is identified as the highest or lowest pitch discussed above.

6. **Preparation:** Indicated by the pitch in the glissando, portamento or miscellaneous slides extending beyond the previous note.

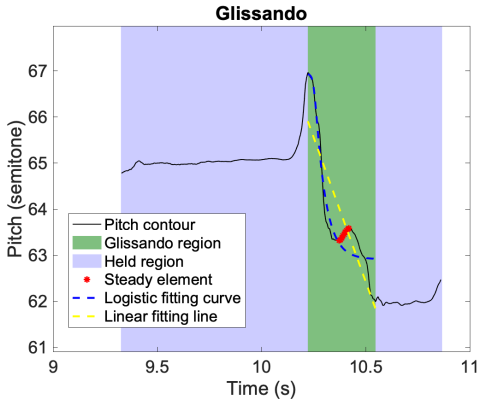
   - If the transition is from a higher pitch to a lower pitch, the highest pitch in the transitional region should be higher than the start pitch of the transition.

- If the transition is from a lower pitch to a higher pitch, the lowest pitch in the transitional region should be lower than the start pitch of the transition.
- The preparation region is defined as slides occurring before the turning point, which is identified as the highest or lowest pitch discussed above.

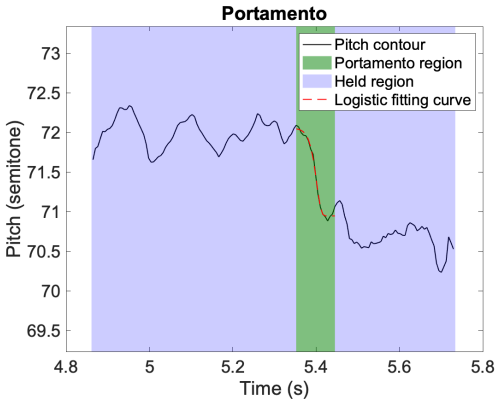
Then multiple features are measured based on the f0 of each ornament. The definition of these features are:

1. **Glissando:**
    - The number of touch notes (short steady notes).
    - Pitch interval of glissando, defined as the pitch difference between the beginning and endpoint.
    - Duration of glissando.
    - The pitch interval of a glissando is segmented into several intervals by touch notes, where each interval is defined as the distance between the median pitch of the touch note and either the boundary pitches of the glissando or the median pitch of the neighbouring touch notes.
    - The time interval is segmented into several intervals by touch notes, where each interval is defined as the time between the median time of the touch note and either the boundaries of the glissando or the median time of the neighbouring touch notes.
    - Duration of each touch note.
    - Slope of glissando. Due to the linear model or logistic model not accurately fitting to the overall slope as the glissando example in Figure 4.30 shows, the slope is calculated directly using:

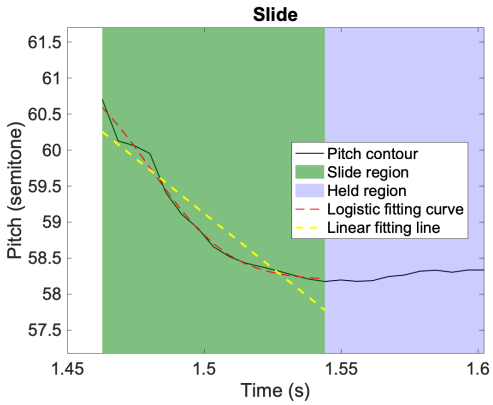
$$\text{Slope} = \frac{\text{Pitch\_interval}}{\text{Duration}} \tag{4.4}$$



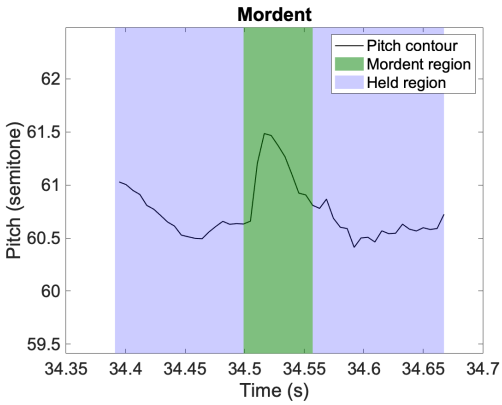
(a) Alpine,
song name: Dr_Braemiser,
time range: 10.22s-10.55s



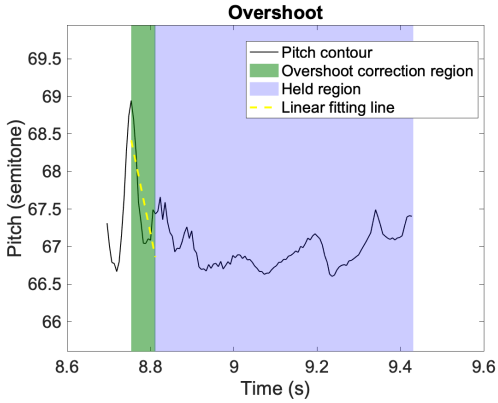
(b) Alpine,
song name: Franz_Lustenberger_Entlebuch,
time range: 5.35s-5.45s



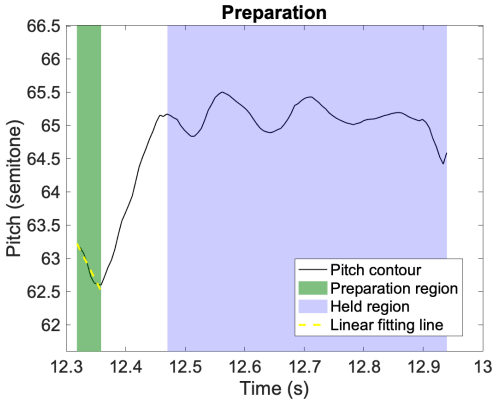
(c) Russian,
song name: Da_po_zoriushke_1,
time range: 1.46s-1.54s



(d) Russian,
song name: Da_po_zoriushke_2,
time range: 34.50s-34.56s



(e) Alpine,
song name: Franz_Lustenberger_Entlebuch,
time range: 8.75s-8.81s



(f) Russian,
song name: Da_po_zoriushke_1,
time range: 12.32s-12.36s

Figure 4.30: Examples of ornaments from different cultures. Each subfigure shows a specific ornament type with the corresponding culture, song name, and time range. Both linear and logistic models are applied to glissandi and slides to determine the optimal approach for measuring slope. Linear fitting is exclusively used for overshoot correction and preparation, as some segments are too brief for the logistic model.

2. **Portamento:** The features of portamento, as estimated using the logistic model (Yang, Chew & Rajab 2015), include the following list. This model has been evaluated as providing the best fit for portamento. The definitions and calculation methods for these features have been introduced in Section 2.5.7.

   (a) Slope

   (b) Duration

   (c) Interval

   (d) Normalised inflection time

   (e) Normalised inflection itch

3. **Miscellaneous Slides:**

   - Interval,
   - Duration,
   - Slope, which is estimated by linear fitting, as this method captures the overall slope better than the logistic model, as illustrated by the slide example in Figure 4.30.
   - Position, which has three categories, head, middle and tail of the pitch contour.
   - Evolution, which is measured by the 1st-7th DCT coefficients, introduced in Section 2.5.7.

4. **Mordent:** The features include duration, as well as the interval between the maximum or minimum pitch value and the mean of the median pitches in the two surrounding held regions.

5. **Overshoot correction region:** The features include duration, interval and slope which is estimated by using linear fitting.

6. **Preparation region:** The features include duration, interval and slope which is estimated by using linear fitting.

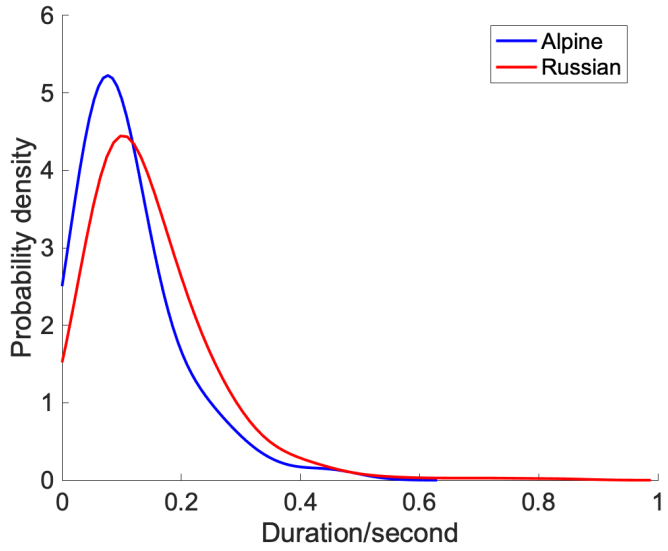


Figure 4.31: Comparison of transitional region durations between Alpine (653 regions) and Russian (1041 regions) datasets

4.5.2 Transitional Region Analysis Results

Overall Transitional Region Feature Distribution

First of all, the duration of transitional regions of Alpine and Russian styles is compared, as shown in Figure 4.31. The duration graph indicates that both styles exhibit a similar distribution, with the probability density peaking around 0.1 seconds and decreasing rapidly as the duration increases. This indicates that transitional regions in both datasets mostly last around 0.1 seconds. However, the difference is significant as shown by a two-sample Kolmogorov-Smirnov test statistic of $D(653, 1041) = 0.215, p < .001$.

To gain a further understanding of the use of transitional regions in connecting or transitioning between held regions, several measures are examined. Table 4.7 presents data on the Transition-Held Ratio (THR), quantifying the ratio of transitional regions to held regions, and includes the counts of both transitional and held regions. The Russian dataset shows a higher Transitional-Holding Ratio (THR). However, the difference is not statistically significant, as evidenced by the chi-squared test statistics, $\chi^2(1, N = 3740) = 0.643, p = 0.423$.

This study further evaluated the characteristics of transitional regions by analysing their distribution by position. Table 4.8 presents the percentages of different transi-

tional region positions relative to held regions and notes. The categories include head, representing transitional regions at the beginning of a note; tail, indicating transitional regions at the end of a note; connect, which refers to transitional regions connecting two held notes; and intra, denoting transitional regions connecting two held regions within a single note. Both cultures predominantly use transitional regions to connect held regions, followed by transitions at the beginning of held regions. Transitional regions at the ends of held notes are less common, and the least frequent are transitional regions with intra position. The chi-square test statistic is $\chi^2(3, N = 1694) = 17.442, p < 0.001$, indicating a significant difference between the Alpine and Russian datasets in terms of the positions of transitional regions.

Moreover, Table 4.9 compares the distribution of ornament types between the Alpine and Russian datasets. The Russian data shows a more use of glissando (2.69%) compared to the Alpine data (1.23%). Both datasets exhibit similar usage of portamento, with the Alpine data at 35.38% and the Russian data slightly lower at 34.97%. The miscellaneous slides category is more prevalent in the Alpine data (61.56%) compared to the Russian data (56.58%). However, the Russian data utilises mordent more frequently (3.07%) than the Alpine data (0.77%). This distribution underscores distinct preferences in ornament types between the two datasets, suggesting a richer use of mordent and glissando in the Russian data, while the Alpine data shows a higher reliance on miscellaneous slides. The chi-square test statistic is $\chi^2(3, N = 1659) = 15.396, p = 0.002$, indicating a significant difference between the Alpine and Russian datasets in terms of the ornament types used in singing.

Finally, Table 4.10 presents the subtype distribution for pitch slides, categorised into three groups: overshoot, preparation, and none. A commonality between the Alpine and Russian data is the majority of instances in both datasets falling under the 'None' category, with 54.60% for Alpine and 54.54% for Russian. The chi-square test statistic is $\chi^2(2, N = 1622) = 1.539, p = 0.463$, indicating no significant difference.

Data	THR	Transitional Region Count	Held Region Count
Alpine	0.80	653	815
Russian	0.85	1041	1231

Table 4.7: Ratio between transitional region and held region counts

Data	Head (%)	Tail (%)	Connection (%)	Intra (%)
Alpine	30.32	27.26	38.13	4.29
Russian	32.18	19.31	41.69	6.82

Table 4.8: Percentage of each position type of transitional region counts

Characteristics of Glissando Distribution

This analysis of glissando focuses on several key features: the number of touch notes, pitch intervals segmented by touch notes, time intervals segmented by touch notes, the duration of each touch note, the overall pitch interval of the glissando, the duration of the glissando, and the slope of the glissando.

Table 4.11 indicates a tendency for the single-touch glissando in both Alpine and Russian vocal data, as they are more prevalent in the counts provided. No significant difference is found between Alpine and Russian as indicated by the chi-squared test result as $\chi^2(2, N = 36) = 0.937, p = 0.626$.

Figure 4.32 compares the probability density of pitch and time intervals for glissando notes segmented by touch notes between Alpine and Russian styles. The two-sample Kolmogorov-Smirnov test statistics are $D(17, 65) = 0.339, p > .05$ for pitch intervals and $D(17, 65) = 0.363, p = 0.043$ for time intervals, indicating no significant difference in pitch intervals but a significant difference in time intervals.

Figure 4.33 compares the distributions of glissando duration and touch note duration in Alpine and Russian vocal data. The left graph illustrates the distributions of touch note durations, while the right graph shows the distributions of glissando durations in Russian and Alpine data. The two-sample Kolmogorov-Smirnov test statistics

Data	Glissando (%)	Portamento (%)	Miscellaneous Slides (%)	Mordent (%)
Alpine	1.23	35.38	61.56	0.77
Russian	2.69	34.97	56.58	3.07

Table 4.9: Distribution of ornament type

Data	Overshoot (%)	Preparation (%)	None (%)
Alpine	9.98	35.41	54.60
Russian	11.82	33.64	54.54

Table 4.10: Subtype of pitch slide distribution for Alpine and Russian data

Data	Single-Touch	Double-Touch	Triple-Touch
Alpine	7	1	0
Russian	20	7	1

Table 4.11: Count of Single-Touch, Double-Touch and Triple-Touch glissando. Single-Touch glissandos represent glides with an intermediate touch note that segments the glide into two parts. Double-Touch refers to glissandos with double touch notes. Triple-Touch refers to glissandos with triple touch notes

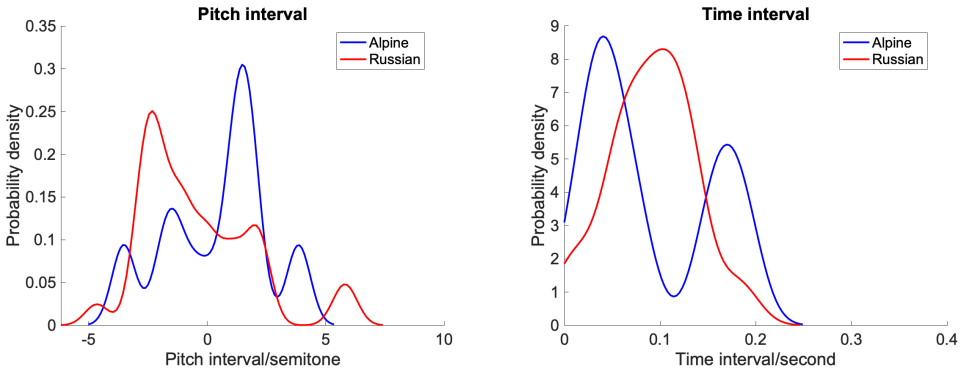


Figure 4.32: Comparative analysis of pitch interval and time interval segmented by touch note

are $D(9, 37) = 0.586, p = 0.008$ for glissando duration and $D(8, 28) = 0.375, p > .05$ for touch note durations, indicating a significant difference in glissando durations but no significant difference in touch note durations.

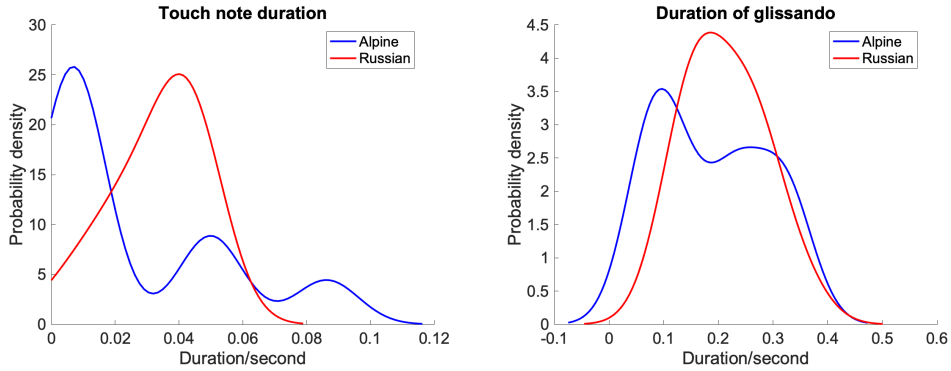


Figure 4.33: Comparative analysis of touch note duration and glissando duration

Figure 4.34, compares pitch interval and slope of glissandos in Alpine and Russian

vocal data. No significant differences are found. The two-sample Kolmogorov-Smirnov test statistics are $D(8, 28) = 0.339, p > .05$ and $D(8, 28) = 0.411, p > .05$.

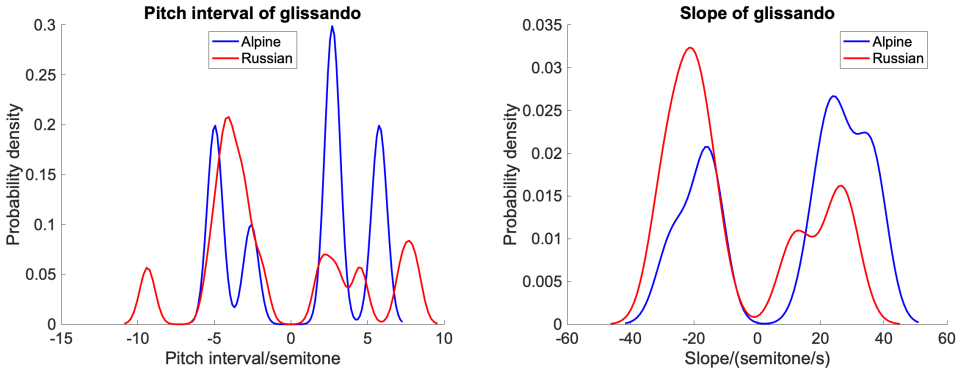


Figure 4.34: Comparative analysis of the interval and slope of glissando

Overall, Alpine singing exhibited greater variation in time intervals segmented by touch notes, as well as in glissando duration. No significant differences in other features were observed. This lack of significance is likely attributable, at least in part, to the limited number of data points.

Characteristics of Portamento Distribution

This analysis of portamento focuses on several features: normalised inflection pitch, normalised inflection time, duration, interval, and slope of portamento. The comparison between the data of Alpine and Russian vocal styles reveals distinct patterns and preferences in their respective distributions.

Figure 4.35 compares the probability density of normalised inflection pitches and times between Alpine and Russian portamento. The inflection pitch graph indicates a very similar downward trend for both styles. The two-sample Kolmogorov-Smirnov test statistic is $D(223, 362) = 0.083, p > .05$, indicating no significant difference in inflection pitch between the two styles. However, the inflection time graph highlights notable differences: the Russian data exhibits a more prominent peak around 0.6, whereas the Alpine data has a broader distribution with concentration approximately from 0.2 to 0.6, suggesting varied inflection timings in Alpine portamento. The two-sample Kolmogorov-Smirnov test statistic is $D(223, 362) = 0.289, p < .001$, indicating a significant difference between the two styles in this aspect.

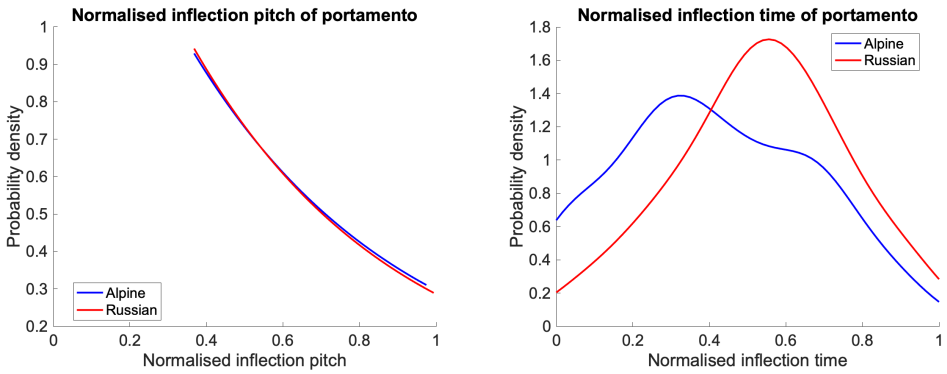


Figure 4.35: Comparative analysis of normalised inflection pitch and time of portamento

Figure 4.36 compares the distribution of duration, interval, and slope of portamento in Alpine and Russian vocal data. The duration graph shows that the mode of the distribution for the Russian data is around 0.1 seconds, while the Alpine data displays a mode at a shorter duration of approximately 0.05 seconds, indicating quicker portamento in the Alpine data. The interval graph reveals that Alpine data features more prominent mode at around -2 semitones than Russian data, while they have very similar distribution of the positive intervals. The slope graph illustrates that the Russian data has higher peaks approximately at 0, while Alpine data presents a broader distribution. These differences are significant, as indicated by the two-sample Kolmogorov-Smirnov test statistics: $D(223, 362) = 0.289, p < .001$ for duration, $D(223, 362) = 0.172, p < .001$ for interval, and $D(172, 338) = 0.244, p < .001$ for slope. In summary, Alpine singing displayed more varied inflection timings and quicker portamento, with a tendency towards larger downward intervals.

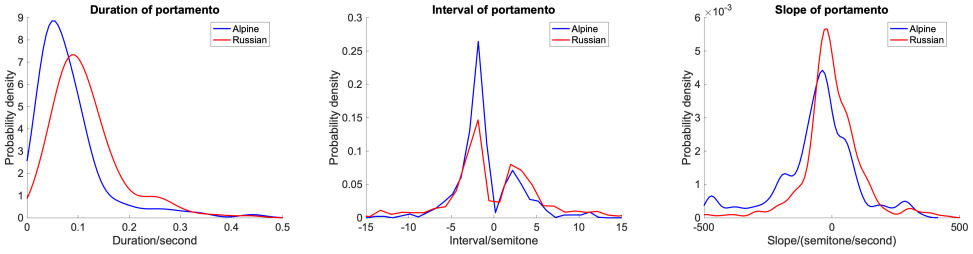


Figure 4.36: Comparative analysis of duration, interval, and slope of portamento. For the interval, the probability density remains above 0 between $-\frac{1}{3}$ and $\frac{1}{3}$ due to the smoothing effect of KDE.

Characteristics of Miscellaneous Slide Distribution

This section analyses the characteristics of miscellaneous slides, focusing on several key features: duration, interval, slope and evolution. The comparison between the data of Alpine and Russian styles reveals similarities and distinct characteristics in their respective distributions.

Figure 4.37 compares the duration, interval, and slope of miscellaneous slides between Alpine and Russian data. The duration graph shows that both Alpine and Russian styles have a mode around 0.1 seconds. However, the Russian style exhibits a lower peak, suggesting a broader duration distribution in Russian slides. This indicates that Russian slides tend to have more variability in duration. The two-sample Kolmogorov-Smirnov test statistics is $D(402, 589) = 0.203, p < .001$, indicating a statistically significant difference.

In the interval graph, both styles show peaks around -1 and 1 semitone. However, the Alpine slides display a more prominent peak around 1 semitone, suggesting a preference for upward slides, while the Russian slides show a broader distribution with a lower peak, indicating a tendency for downward slides. The two-sample Kolmogorov-Smirnov test statistics is $D(402, 589) = 0.133, p < .001$, indicating a statistically significant difference. The slope graph reveals that both styles have modes around 0 semitones/second with similar levels. The two-sample Kolmogorov-Smirnov test statistics is $D(402, 589) = 0.068, p > .05$, indicating no statistically significant difference.

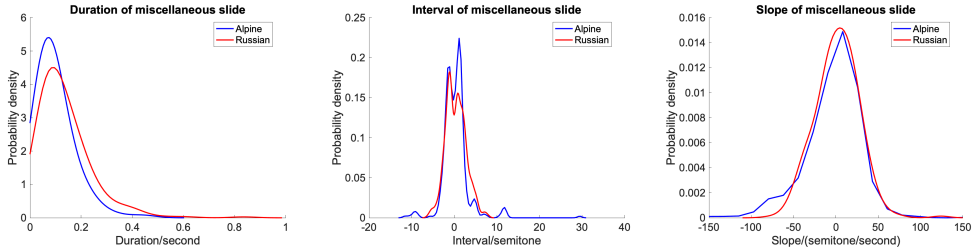


Figure 4.37: Comparative analysis of duration, interval, and slope of miscellaneous slides between Alpine and Russian data

Figure 4.38 compares the distribution of DCT coefficient values (from 1st to 7th) for both Alpine and Russian data. The boxplots illustrate the median, interquartile ranges, and outliers for each DCT coefficient. Both styles show a similar pattern, where the

first coefficient takes the majority of the energy and the second coefficient is the second largest. This suggests that the slides tend to be linear with a slight curvature. The ANOVA yields F(1,456)=0.34,p=0.562 and F(1,456)=0.99,p=0.320 for the first and second coefficients, respectively, indicating no statistically significant difference in the 1st and 2nd DCT coefficient of slides between the two singing styles. Overall, Russian singing shows a broader distribution in the duration of slides, while Alpine singing displayed a tendency for upward slides.

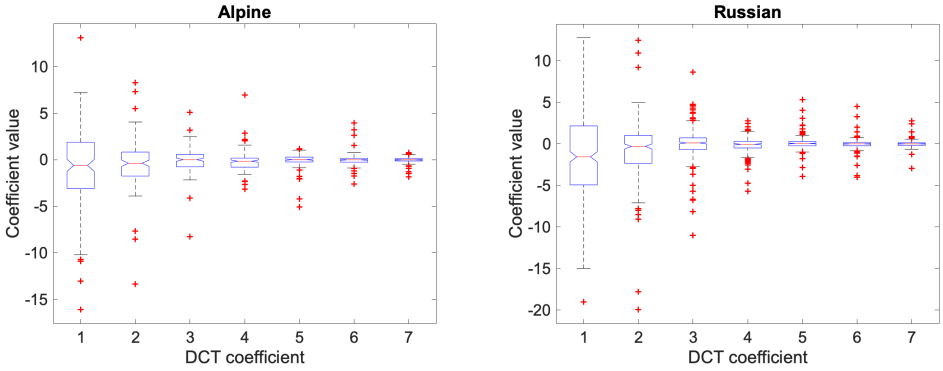


Figure 4.38: Comparative analysis of DCT coefficient values for slides in Alpine and Russian styles. The median is presented by the red line, the interquartile range is captured within the blue boxes, outliers are denoted by red plus signs, and the whiskers extend to capture the range of data points excluding the outliers. Each boxplot correspond to an individual DCT coefficient. Higher absolute coefficient value indicates the component with higher energy.

Characteristics of Mordent Distribution

This section analyses the characteristics of mordents, focusing on the duration and interval. Figure 4.39 compares the probability density of the duration of mordents between Alpine and Russian styles. The two-sample Kolmogorov-Smirnov test statistics is $D(5, 32) = 0.519, p > .05$, indicating no statistically significant difference. (The K-S test can compare distributions of any shape (unimodal, bimodal, etc.) due to its non-parametric characteristic, but with such a small sample size (n=5) for Alpine data, the test's power to detect real differences is limited.)

Figure 4.40 compares the interval of mordents between Alpine and Russian data. The two-sample Kolmogorov-Smirnov test statistics is $D(5, 32) = 0.345, p > .05$, indicating no statistically significant difference. Overall, no significant differences between

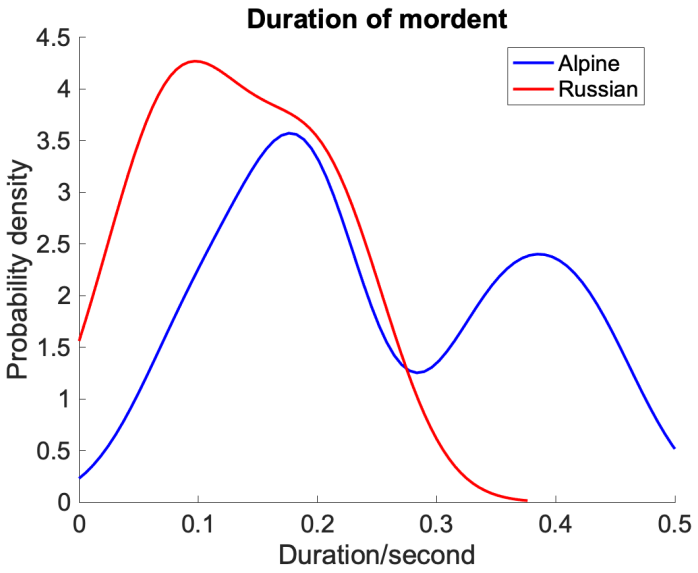


Figure 4.39: Comparison of the duration of mordent between Alpine and Russian data

the datasets regarding mordant features were observed. This lack of significance is likely attributable, at least in part, to the limited number of data points.

Characteristics of Overshoot Correction Distribution

Figure 4.41 compares the duration, interval, and slope of overshoots between Alpine and Russian styles. The duration graph indicates that both styles have similar distribution shapes, with the Russian style having a mode around 0.04 seconds, while the Alpine style has a mode at 0.02 seconds. The Russian style also exhibits a lower probability density at the mode and a broader duration distribution. The interval graph reveals that both styles have a mode around 0 semitones, with the Russian style showing a lower mode and a broader distribution. The slope graph shows a similar pattern between the two styles, though the Russian style exhibits a broader distribution. The two-sample Kolmogorov-Smirnov test statistics are $D(100, 201) = 0.297, p < .001$ for duration, $D(100, 201) = 0.248, p < .001$ for interval, and $D(100, 201) = 0.208, p = 0.005$ for slope, indicating statistically significant differences across all measures. In summary, the Russian style exhibites broader distributions in duration, interval, and slope, indicating greater variability in overshoot correction compared to the Alpine style.

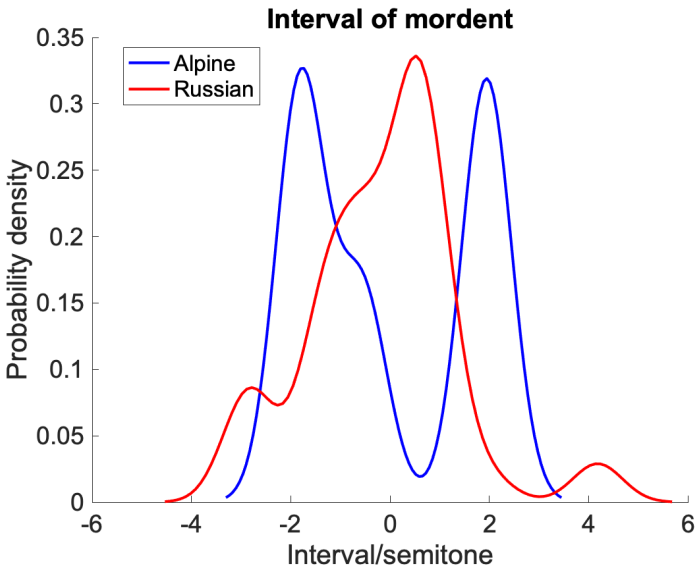


Figure 4.40: Comparison of the interval of mordent between Alpine and Russian data

Characteristics of Preparation Distribution

Figure 4.42 compares the duration, interval, and slope of preparations between Alpine and Russian data. The duration graph indicates that both styles have similar distribution shapes, with the Russian style exhibiting a broader distribution. The interval graph shows that both styles have a mode around 0 semitones, with the Russian style showing a slightly broader distribution. The slope graph shows a similar shape between the two styles, though the Russian style has a mode around -7 semitones per second, while the Alpine style has a mode around 3 semitones per second. The two-sample Kolmogorov-Smirnov test statistics are $D(227, 330) = 0.118, p = 0.043$ for duration, $D(227, 330) = 0.233, p < .001$ for interval, and $D(227, 330) = 0.218, p < .001$ for slope, indicating statistically significant differences across all measures.

In summary, the preparation duration, interval, and slope between Alpine and Russian styles suggest that the Russian data exhibits more flexibility and a slight tendency to prepare for upward pitch slides, while the Alpine style shows a slight tendency to prepare downward pitch slides. The p-values indicate that the differences in preparation duration, interval, and slope are statistically significant.

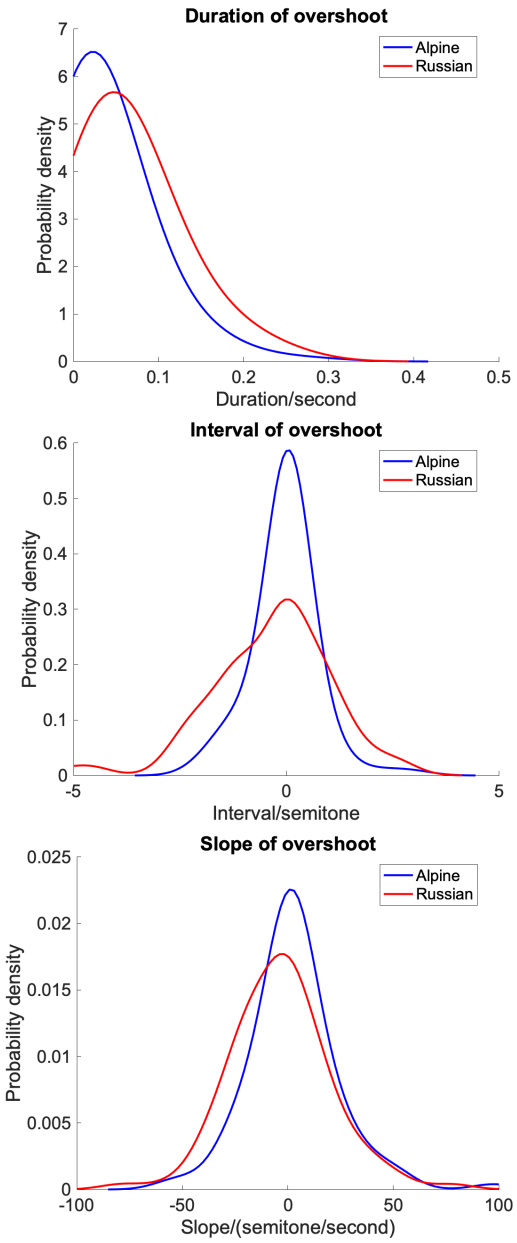


Figure 4.41: Comparative analysis of duration, interval, and slope of overshoot correction between Alpine and Russian data

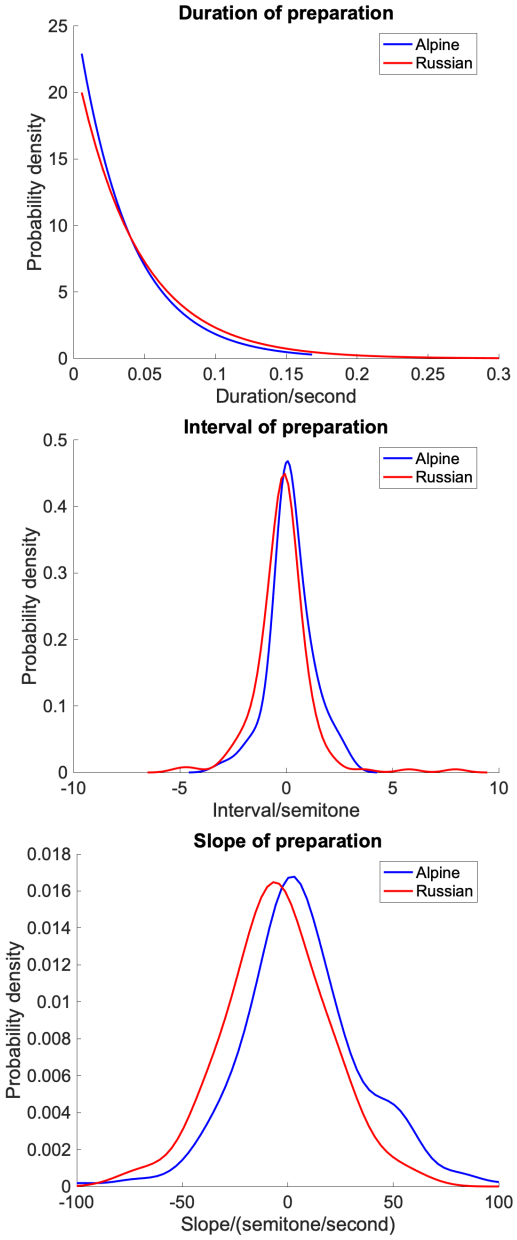


Figure 4.42: Comparative analysis of duration, interval, and slope of preparations between Alpine and Russian styles

4.6 Conclusion

This chapter conducts a comparative analysis of note-level pitch contours in Alpine and Russian singing, using datasets with two versions of note segments transcribed by two experts. The study examines multiple features of both held and transitional regions of musical notes, highlighting differences between these distinct singing styles through visual and statistical comparisons. As this study analyses only 8 Alpine and 10 Russian recordings, the results should be considered as a case study of these particular examples rather than a comprehensive representation of the entire Alpine and Russian vocal traditions.

Key findings include:

1. **Note Type Distribution**: Steady-dominant notes occurred more frequently in the annotated notes. Chi-square tests indicated that variations in annotations within the same dataset are likely due to chance rather than systematic differences in style.

2. **Note Boundary Displacement**: For the Russian dataset, transcribers PP and OV showed remarkably similar annotation patterns, while in the Alpine dataset, transcribers LS and YW displayed significant differences in their note boundary placements, especially highlighting that LS adopts a more flexible annotation approach, whereas YW consistently marks note boundaries relative to transitional regions.

3. **Steady Elements Analysis**: Steady elements in Alpine and Russian data exhibit symmetrical distributions around zero, indicating pitch stability. Alpine data displayed a higher prevalence of level steady elements, suggesting more pitch stability.

4. **Modulating Elements Analysis**: The analysis reveals significant differences in vibrato characteristics between Alpine and Russian singing styles. Alpine singers exhibit quicker vibrato with rate between 6 and 8.5 Hz, and a regulated extent around 0.4 semitones. In vibrato extent evolution, Alpine data is stable around

zero, while Russian data shows greater variability. Both styles produce regular vibrato, but Alpine singers have a more stable carrier and longer modulating elements, with significant differences in the carrier slope and duration.

5. **Transitional Region Analysis**: The Russian data demonstrates a richer use of glissando (2.69% vs. 1.23%), mordent (3.07% vs. 0.77%), and overshoot (11.82% vs. 9.98%), while showing less use of other ornaments compared to Alpine singing. Alpine singing in this dataset tends to place transitional regions more towards the tail of the note, exhibits greater variation in time intervals segmented by touch notes of glissando, and shows longer glissando durations. Alpine singing also displays more varied inflection timings, quicker portamento, and a tendency towards larger downward intervals of portamento. In contrast, Russian singing shows a broader distribution in the duration of slides, a greater variability in overshoot correction, and a slight tendency to prepare for upward pitch slides, while Alpine singing tends towards upward slides and shows a slight preference for preparing downward pitch slides.

The framework is designed to be genre-agnostic in its note-level analysis approach by decomposing complex pitch variations and ornaments into three fundamental pitch contour elements: steady, modulating, and transitory regions. While the framework has demonstrated effectiveness in differentiating characteristics between Alpine and Russian singing styles, its broader applicability across diverse vocal traditions requires further empirical validation.

Overall, the findings highlight broad similarity and nuanced differences in vocal styles between Alpine and Russian singing, based on a framework for computational vocal music analysis using music information retrieval techniques. The manual segmentation approach, while labor-intensive, remains crucial for accurate analysis due to the current limitations of automatic methods. Future work could focus on enhancing automatic segmentation techniques and expanding the analysis to other singing styles and languages.

Chapter 5

Syllable-Level Pitch Contour Analysis

This chapter investigates the realisation of tones in the pitch contours of Chaozhou folk singing, a genre of Chinese folk music originating from Chaozhou, a city in southern China known for its distinctive dialect, which features a greater number of tones compared to Mandarin. In tonal languages, syllables are articulated with specific tones that employ pitch variations to distinguish word meanings. Given the tonal complexity of the Chaozhou dialect, this chapter examines the preservation and modification of these lexical tones in Chaozhou folk singing. This research builds upon an existing Chaozhou folk singing dataset with syllable-level segmentations and tone labels. The study also benefits from the expertise of collaborator who has extensive experience in Chaozhou folk singing research. The objective is to apply computational methods to uncover patterns in syllable-level pitch contours and assess the effects of lexical tones and other contributing factors on the pitch contours in singing.

Section 5.1 introduces the dataset and the factors considered in the analysis. Section 5.2 provides an in-depth examination of the effects of lexical tones on the syllable-level sung pitch contours. Section 5.3 explores the influences of additional factors on tone realisation in Chaozhou folk singing, beyond the effect of tones alone. Finally, Section 5.4 presents the conclusions and briefly discusses potential future research directions in light of the limitations of this study.

5.1 Dataset and Considered Factors

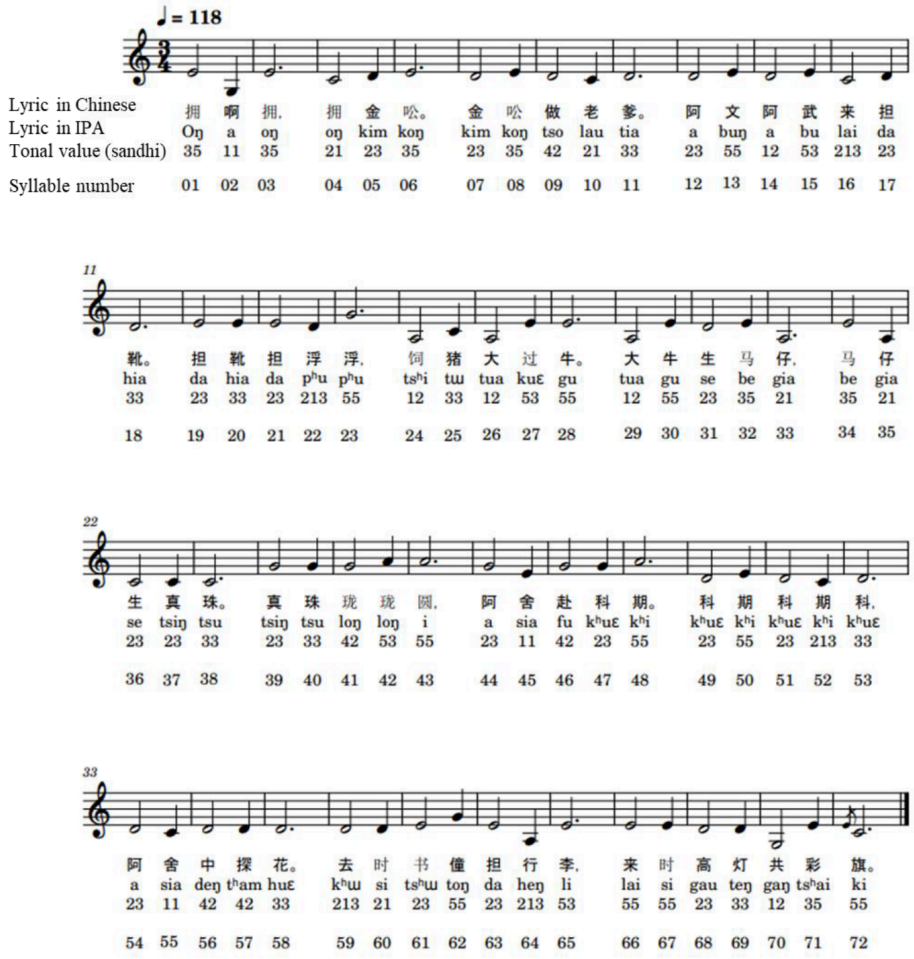


Figure 5.1: Musical score of *Oŋ a oŋ* (Zhang 2024).

The dataset (Zhang 2024) utilised in this study involves a folk song, *Oŋ a oŋ* (translated as 'I am tucking you in, my little baby', 拥啊拥). The musical score is shown in Figure 5.1. The song is performed by thirty-four Chaozhou singers, resulting in 34 recordings of the same song. For each recording, the pitch track was extracted using Praat software (Boersma 1993). An expert of Chaozhou music and dialect, Xi Zhang, then segmented the singing at the syllable level (the irregular pitch contours introduced by consonants are excluded from syllable segments) and manually labelled each syllable's lexical tone by looking up a dictionary (Lin 1995), using Praat software. The sung syllables are consistent across all performances, with a total of 71 annotated

syllables (the last syllable of the song is not annotated).

Multiple variables are considered for singers and syllables in this study, some of which are referenced in Zhang (2024). Among them, four are related to singers:

Gender: The singers' gender distribution is as follows: 22 females and 12 males.

Age: Regarded as a continuous variable, the mean age is 38.12 years (SD = 8.11).

Vocal training background: Three categories are decided: Western bel canto, Chinese folk singing, and non-professionals. Six professional singers (2 females, 4 males) have extensive vocal training in Western bel canto, with a mean training duration of 6.33 years (SD = 1.25). Thirteen professional singers (11 females, 2 males) have extensive vocal training in Chinese folk singing, with a mean training duration of 11 years (SD = 4.56). The fifteen non-professionals (9 females, 6 males) have little or no vocal training, with a mean training duration of 0.67 years (SD = 1.29).

Experience in singing in Chaozhou: This is an ordinal variable and refers to how frequently the singers perform Chaozhou dialect. Participants reported their frequency of singing in Chaozhou in four levels: often (6), sometimes (13), seldom (11), and never (4).

The following factors are identified for each syllable:

Tone Step and Tone Contour: This study examines two variables related to tone: "tone step" and "tone contour". The tone labels are based on the "five-level tone mark" system introduced by Yuen Ren Chao (Chao 1930), where numbers indicate lexical pitch (lower numbers signify lower pitch). The syllables in this song feature ten tones, including three level tones (/11/, /33/, /55/) and seven non-level tones (/35/, /53/, /42/, /213/, /23/, /21/, /12/). The first variable, tone step, categorises tones based on the difference in pitch level. For example, /11/ is categorised as 0 steps, /23/ as 1 step, and /53/ as -2 steps. The tone /213/ is considered a unique step level. This variable is ordinal, ranging from -2 to 2 steps. Tone /213/ falls between 1 and 2 steps. The second variable, tone contour, distinguishes between tones with simple and complex pitch movements. The /213/ tone, which involves multiple pitch changes, is the only complex tone contour. All other tones are simple tone contours, involving only two pitches.

Metrical Structure: There are two levels: strong & long and weak & short. In the song *Oŋ a oŋ*, syllables on a strong beat are sustained longer, while syllables on a weak beat are shorter.

Melodic Interval: The melodic interval indicates the pitch interval between two neighbouring syllables, as referred to in the musical score shown in Figure 5.1, where each syllable corresponds to a note. There are two types of intervals: preceding melodic interval (PMI) and succeeding melodic interval (SMI). Based on existence, size, and direction, there are six levels for each type of interval:

- **None:** For syllables at the beginning or end of a phrase, without a preceding or succeeding interval.
- **Level 0:** Unison.
- **Ascending Level 1:** Minor 2nd, Major 2nd, minor 3rd intervals with ascending direction.
- **Descending Level 1:** Minor 2nd, Major 2nd, minor 3rd intervals with descending direction.
- **Ascending Level 2:** Major 3rd, perfect 4th, perfect 5th, major 6th with ascending direction.
- **Descending Level 2:** Major 3rd, perfect 4th, perfect 5th, major 6th with descending direction.

Citation/Sandhi: This refers to the phenomenon where the tone of a syllable changes based on the tonal context of surrounding syllables. For this variable, there are two levels: tone citation and tone sandhi. The term "citation tone" in Chaozhou refers to tones when single characters are spoken alone(Bao 1999, Lin 1995), while "tone sandhi" refers to the tonal changes that occur when characters are spoken in connected speech (e.g., within words or phrases) (Chen 2000, p. 19).

Vowel Type: This study identifies 42 unique syllables and 17 unique vowels, which are categorised into three vowel types:

- **Monophthong:** {'w', 'a', 'o', 'e', 'i', 'u'}
- **Diphthong:** {'ai', 'au', 'ia', 'ua', 'uɛ'}
- **Nasalised:** {'am', 'aŋ', 'oŋ', 'eŋ', 'im', 'iŋ'}

This categorisation was applied to this song in a simplified approach based on the classifications proposed by Lin (1995) and has been validated by an expert in the Chaozhou dialect.

5.2 Tone Effects on Syllable-Level Sung Pitch Contour

To observe the effects of tone on the pitch trajectories of sung syllables, the pitch contours of different syllable segments with the same tone were averaged to create an overall pitch contour. This allowed for comparison of the overall pitch contour shapes across different tones. The process involved three main steps: first, for each syllable segment, the pitch variation trajectories were calculated by obtaining the difference between the continuous f0 contour and its median; second, the syllable-level pitch variation trajectories were normalised by re-sampling them to 100 points; finally, the common pitch contour shape for each tone was obtained by averaging the re-sampled trajectories of syllables with the same tone.

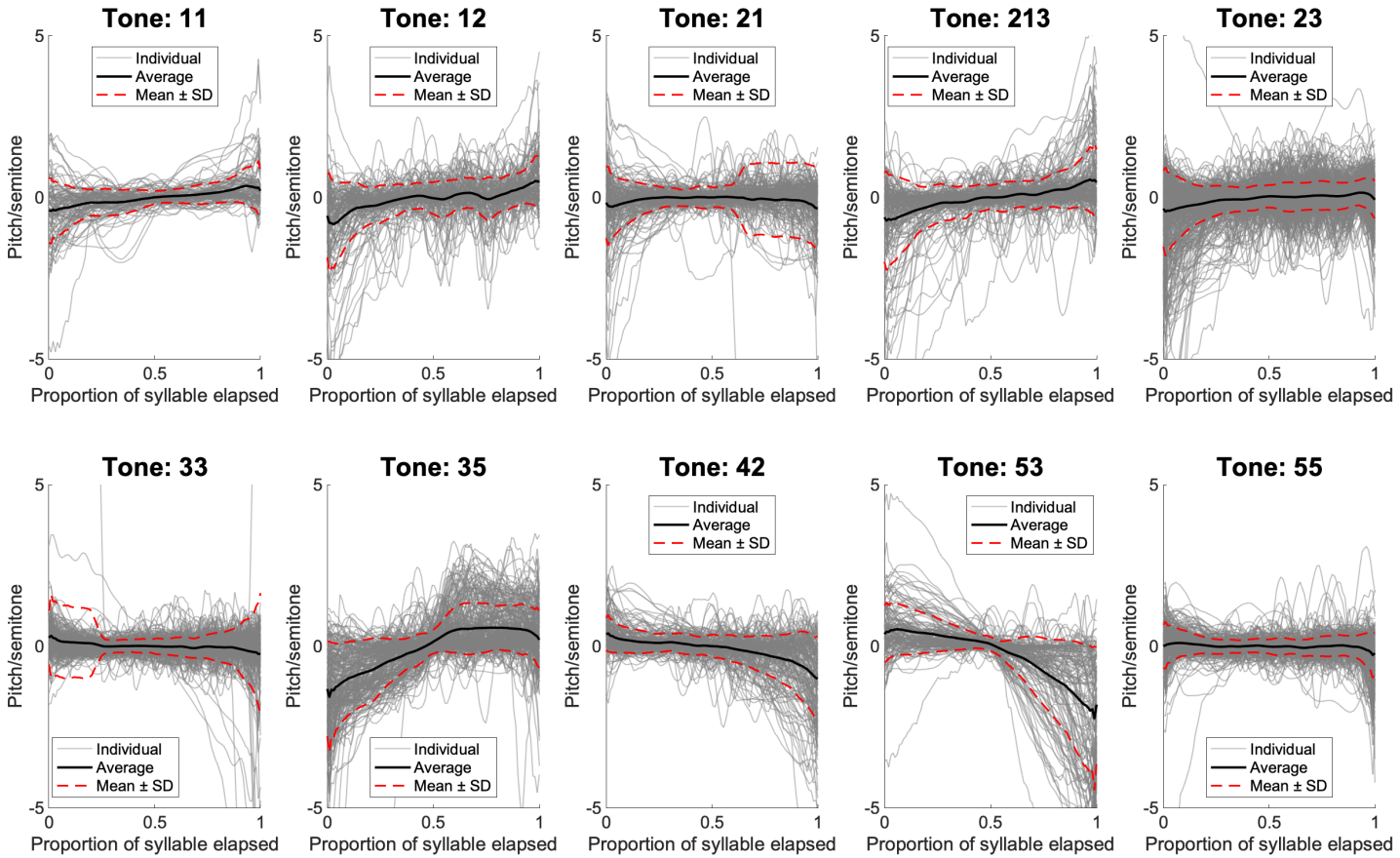


Figure 5.2: Mean pitch variation trajectories of syllables for 10 different tones. Each subplot represents a specific tone and displays the individual pitch variations and their averaged pitch variation over normalised time. The red dashed lines indicate the range of the averaged pitch variations as defined by the standard deviation.

Figure 5.2 compares the sung pitch contours for ten distinct tones (11, 12, 21, 213, 23, 33, 35, 42, 53, and 55). Each subplot presents the individual pitch variations and their averaged pitch variation in semitones relative to the normalised time of the syllable. To assess differences in pitch behaviour across the tones, the pitch changes of the averaged pitch variations are estimated through linear fitting and the standard deviations are obtained to quantify the consistency of individual pitch variations within each tone.

For the three level tones (11, 33, and 55), tones 33 and 55 exhibit average pitch trajectories with minimal pitch change magnitudes of -0.28 semitones and -0.13 semitones, respectively. In contrast, tone 11 shows a slight upward trend, with a pitch change of 0.68 semitones. The rising tones 12 and 23, both characterised by a step size of 1, display an upward pitch trajectory over time, with tone 12 having a larger pitch change magnitude (0.89 semitones) than tone 11, while tone 23 has a smaller pitch change (0.44 semitones). Tone 21, with a step size of -1, presents a relatively level average pitch contour with the smallest pitch change magnitude of 0.003 semitones, resembling the pitch contour shape of tone 55, which has a pitch change of -0.13 semitones. Tones 35, 42, and 53, characterised by two-step scales, exhibit lager pitch changes over time, consistent with their directions. The pitch changes for these tones are 2.03, -0.97, and -2.44 semitones, respectively.

Tone 213, expected to have a distinct contour, instead presents a sliding-up shape rather than the anticipated down-and-up pattern. The pitch change magnitude of tone 213 (1.05 semitones) is smaller than those of the two-step tones 35 and 53, but larger than those of the one-step and level tones. Furthermore, the standard deviations of pitches, indicated by the red dashed lines, tend to be smaller in the middle of the syllable's duration for each tone and larger at the two ends. Although tone 21 exhibits the smallest pitch change magnitude, its standard deviation increases towards the end, becoming greater than that of the level tones 11 and 55.

To assess the significance of tone effects, this study examines the influence of tone step and tone contour on syllable-level sung pitch trajectories. Specifically, it evaluates the effect of tone step on the linear tendency of these pitch trajectories. Additionally,

the study investigates whether the unique tone 213, characterised by its complex tone contour, introduces a greater degree of convex curvature in the sung pitch contour compared to tones with simple contours. The linear tendency and curvature of the pitch variation are quantified using the first and second coefficients of the Discrete Cosine Transform (DCT), respectively. A positive first coefficient corresponds to a negative slope, while a negative first coefficient indicates a positive slope. For the second coefficient, a negative value indicates concave curvature, and a positive value indicates convex curvature. DCT is chosen for its efficiency in simultaneously quantifying linear tendency and curvature without requiring any pre-smoothing. More details about DCT have been provided in Section 2.5.7.

This study hypothesises a significant correlation between tone step and the first DCT coefficient of sung pitch contours in terms of both direction and magnitude, while no significant correlation is expected between tone contour and the second DCT coefficient, based on observations of the averaged pitch variation for tone 213. Linear mixed models (LMMs) are selected for the analysis because they are well-suited for analysing correlated data, such as repeated measures from the same subjects or data points that are temporally close. In this study, multiple DCT measures are obtained from syllables sung by the same singer, and syllables within the same phrase are temporally close. Tone step and tone contour are treated as fixed effects to capture the primary relationships between the predictors and the response variable, while singer and syllable are included as random effects (intercepts only) to account for unexplained variability within clusters or among participants.

The linear mixed models used in this study are described in Equations 5.1 and 5.2 using Wilkinson-Rogers notation.

$$\text{DCT1stCoef} \sim \text{Tone\_step} + (1|\text{Singer}) + (1|\text{Syllable}) \tag{5.1}$$

$$\text{DCT2ndCoef} \sim \text{Tone\_contour} + (1|\text{Singer}) + (1|\text{Syllable}) \tag{5.2}$$

Tone step	Estimate	p-Value	95% CI Lower	95% CI Upper
-1	-4.480	***	-6.460	-2.500
0	-4.304	***	-5.709	-2.900
1	-6.317	***	-7.713	-4.922
Tone 213	-7.515	***	-9.495	-5.535
2	-11.120	***	-12.910	-9.336

Table 5.1: Statistical analysis results for 1st DCT coefficient including estimates, p-values, and 95% confidence intervals for levels of tone step. The stars are used to represent the p-values based on their significance levels: *** for p-values < 0.001, ** for p-values < 0.01, * for p-values < 0.05.

Tone step levels are ranked in the order of -2, -1, 0, 1, tone 213, and 2, with step -2 set as the reference level. The statistical results of tone step effects, presented in Table 5.1, are consistent with the hypothesis and show that all levels of tone step exhibit significant effects. Specifically, holding other fixed effects constant, changing from tone step -2 (reference level) to tone step -1 decreases the 1st DCT coefficient by 4.480 units, indicating that step -1 results in a less steep linear slope compared to step -2. Step 0 decreases the 1st DCT coefficient by 4.304 units, slightly less than step -1, which aligns with Figure 5.2, where the averaged pitch variation in tone 21 exhibits the smallest slope magnitude. The effect of step 2 has the largest absolute effect size, indicating the strongest influence on pitch contour. Steps 1 and tone 213 also show significant effects, though smaller than that of step 2, with each step from 0 to 2 demonstrating progressively larger effects. Additionally, no significant effect is observed for tone contour, suggesting that the spoken pitch contour of tone 213 is not preserved in singing within this Chaozhou folk singing dataset.

5.3 Effects of Other Considered Factors on Syllable-Level Sung Pitch Contour

To account for the potential confounding effects of other factors discussed in Section 5.1 on the influence of tone step on sung pitch contour, a linear mixed model, represented as Equation 5.3, is employed that includes these variables alongside tone step. To properly interpret the effects, it is necessary to set a reference level for categorical variables of interest, which are listed in Table 5.2.

Categorical Variable	Reference Level
Tone step	-2
Vocal training background	non-professional
Experience in singing in Chaozhou	never
Gender of singers	female
Metrical Structure	weak & short
Forwards Melodic Interval	none
Backwards Melodic Interval	none
Vowel type	nasalised
Citation\Sandhi	citation

Table 5.2: Categorical variables and their reference level

$$\begin{aligned} \text{DCT1stCoef} \sim\ & \text{Tone\_step} + \text{Age} + \text{Gender} \\ & + \text{Training\_background} + \text{Experience\_in\_CZ} + \text{Citation\textbackslash Sandhi} \\ & + \text{Vowel} + \text{Metrical\_structure} + \text{PMI} + \text{SMI} \\ & + (1|\text{Singer}) + (1|\text{Syllable}) \end{aligned} \tag{5.3}$$

The statistical results, presented in Table 5.3, show the levels of variables that exhibit significant effects. In addition to the tone step, six other factors are found to have statistically significant effects. To interpret the effects of these factors of interest, two methods are employed.

First, to visualise the influence of these factors on tone realisation in singing, syllable-level pitch variation trajectories, shown in Figure 5.2, are categorised according to the levels of each factor. Figures 5.3 through 5.8 displays the averaged pitch contours for different categories of each factor across specific tones.

Second, to assess the effects of the different levels of the factors illustrated in Figures 5.3 through 5.8, linear mixed models are applied to data for each specific tone. The analysis process was as follows: First, for each tone category, multiple syllables produced by different singers were collected. Then, for each syllable, the 1st DCT coefficient of its pitch contour was calculated. Finally, the six factors of interest, training background, experience in singing in Chaozhou, tone citation sandhi, vowel, PMI, and SMI, are included as fixed effects in the linear mixed model, with syllable and singer

Fixed effect	Coefficient	p-Value	Lower	Upper
Tone_step_-1	-3.582	***	-5.219	-1.945
Tone_step_0	-6.255	***	-7.597	-4.914
Tone_step_1	-4.510	***	-5.578	-3.442
Tone 213	-5.441	***	-6.990	-3.891
Tone_step_2	-12.380	***	-13.740	-11.020
Training_Chinese folk	-0.952	*	-1.833	-0.072
Experience_in_CZ_Often	-1.160	*	-2.206	-0.114
Citation_Sandhi_Sandhi	-2.034	***	-3.158	-0.911
Vowel_Diphthongs	0.809	*	0.102	1.516
PMI_Descending_level_1	1.332	**	0.327	2.337
PMI_Level_0	2.019	***	0.821	3.216
PMI_Ascending_level_1	2.534	***	1.376	3.692
PMI_Ascending_level_2	1.588	*	0.012	3.164
SMI_Descending_level_2	2.357	**	0.751	3.963
Age	0.013	n.s.	-0.021	0.047
Gender_Male	-0.534	n.s.	-1.107	0.038
Training_Bel canto	-0.262	n.s.	-0.876	0.353
Frequency_in_CZ_Seldom	-0.757	n.s.	-1.736	0.223
Frequency_in_CZ_Sometimes	-0.880	n.s.	-1.913	0.152
Vowel_Nasalised	-0.537	n.s.	-1.270	0.196
Metrical_Strong & long	-0.635	n.s.	-1.594	0.325
PMI_Descending_level_2	0.029	n.s.	-1.375	1.434
SMI_Descending_level_1	-0.487	n.s.	-1.835	0.860
SMI_Level_0	-0.927	n.s.	-2.349	0.496
SMI_Ascending_level_1	-1.159	n.s.	-2.622	0.305
SMI_Ascending_level_2	-1.468	n.s.	-3.358	0.422

Table 5.3: Statistical analysis results for 1st DCT coefficient including estimates, p-values, and 95% confidence intervals for levels of fixed effects. The stars indicate significance levels: *** for p-values < 0.001, ** for p-values < 0.01, * for p-values < 0.05, and n.s. for non-significant results ($p > 0.05$).

Tone	Correlated Factors
11	Vowel, PMI, SMI
12	Vowel, PMI, SMI
21	Citation_Sandhi, Vowel, PMI, SMI
213	Citation_Sandhi, Vowel, PMI, SMI
23	Vowel, SMI
33	Vowel, PMI, SMI
35	Citation_Sandhi, Vowel, PMI, SMI
42	Vowel, PMI, SMI
53	Citation_Sandhi, Vowel, PMI, SMI
55	Vowel, PMI, SMI

Table 5.4: Tone and correlated factors

treated as random effects, resulting in Equation 5.4.

$$\begin{aligned} \text{DCT1stCoef} \sim\ & \text{Training\_background} + \text{Experience\_in\_CZ} + \text{Citation\textbackslash Sandhi} \\ & + \text{Vowel} + \text{PMI} + \text{SMI} \\ & + (1|\text{Singer}) + (1|\text{Syllable}) \end{aligned} \tag{5.4}$$

When a factor has only one level within a specific tone, it is removed from the analysis. Additionally, certain factors are excluded due to strong correlations with other factors within the same tone, leading to multicollinearity. Table 5.4 lists the correlated factors that potentially introduce these multicollinearity issues. This is identified by finding factors that have correlation coefficients greater than 0.05 with other factors. To address this multicollinearity issue, different combinations of the correlated factors are tested. For example, for tone 11, Vowel, PMI, and SMI are identified as correlated factors. The analysis then examines all possible combinations of these factors: (Vowel, PMI, SMI), (Vowel, PMI), (Vowel, SMI), (PMI, SMI), (Vowel), (PMI), and (SMI). If a particular combination leads to a model that cannot be fitted due to multicollinearity, it is discarded. For combinations that can be successfully modelled, the factors with significant effects are reported. This systematic approach avoided bias in favour of retaining any particular factor while discarding another.

The effects of these factors on the first DCT coefficient are assessed using linear

mixed models. While some effects were statistically significant, the predominant observation across all visualisations is that the variations are relatively minor when compared to the overarching similarities.

- **Vocal Training Background:** Statistically significant effects are identified for tone 21, where transitioning from 'non-professional background' (reference level) to 'Chinese folk background' results in a reduction of 3.722 units in the 1st DCT coefficient. In tone 42, a shift from 'non-professional' to 'bel canto training' results in a reduction of 2.054 units, while transitioning to 'Chinese folk training' leads to a more substantial reduction of 4.490 units. For tone 53, transitioning from 'non-professional' to 'bel canto background' results in a decrease of 2.860 units. These outcomes, shown in tones 21, 42, and 53 of Figure 5.3, suggest that for these three falling tones (the only falling tones among the ten analysed), singers with professional training exhibit a less pronounced pitch decline at the syllable's end. This reflects a less 'speech-like' style, with subtler tone communication, compared to singers without professional training.

- **Experience in Singing in Chaozhou:** Initial analysis using 'never' as the reference level revealed significant effects only in tone 42, showing an unexpected pattern: transitioning to 'seldom' results in a decrease of 6.624 units, to 'sometimes' results in a decrease of 5.402 units, and to 'often' leads to a decrease of 4.048 units. These results suggest that singers with no experience in Chaozhou singing exhibit the most speech-like singing, tending to realise the falling tone more effectively with a stronger communication of tone at the syllable's end. This pattern contradicts the expectation that greater experience would lead to more tonal realisation.

  Given this unexpected behaviour of the 'never' group, further analyses were conducted using 'often' as the reference level to better understand the experience-dependent patterns across different tones. The results show that in tone 12, transitioning to 'seldom' leads to an increase of 3.150 units; in tone 213, changing to 'seldom' results in an increase of 2.329 units; and in tone 23, moving to 'seldom'

shows an increase of 1.398 units. The visualisation in Figure 5.4 corroborates these findings. These patterns suggest experience-dependent strategies in tonal realisation: for rising tones (12, 23, 213), singers with more experience ('often') tend to start with lower pitch in the sung syllable to realise rising contour more efficiently, while for falling tone (42), they focus on realising the falling contour more effectively at the syllable's end.

- **Tone Citation/Sandhi:** A significant effect is identified for tone 213, where transitioning from 'citation' (reference) to 'sandhi' leads to an increase of 2.857 units in the 1st DCT coefficient. This outcome, which suggests a reduction in pitch slope, contradicts the pattern illustrated for tone 213 in Figure 5.5, indicating potential interactions with other factors influencing the realisation of tone 213 during sandhi.

- **Vowel Type:** Statistically significant effects are detected for tones 11, 213, and 55. In tone 11, the shift from 'monophthong' (reference) to 'diphthong' results in an increase of 1.086 units in the 1st DCT coefficient. In tone 213, transitioning from 'monophthong' to 'diphthong' results in a decrease of 1.456 units, and changes to 'nasalised' vowels leads to a decrease of 4.761 units. In tone 53, the shift from 'monophthong' (reference) to 'nasalised' results in an decrease of 11.067 units. In tone 55, the change from 'monophthong' to 'diphthong' results in a decrease of 2.150 units. These results, visualised in Figure 5.6, suggest that vowel type affects the tone realisation in singing, particularly in tones 11 and 213. Additionally, the effect on tone 213 may interact with the effect of 'tone citation/sandhi'.

- **Preceding Melodic Interval (PMI):** Significant effects are found across several tones. For tone 11, transitioning from 'descending_level_2' (reference) to 'descending_level_1' results in an increase of 1.086 units in the 1st DCT coefficient. In tone 12, the same transition leads to an increase of 1.972 units. For tone 23, transitioning from 'none PMI' (reference) to 'descending_level_1' results in an increase of 1.707 units, while transitioning to 'level_0' results in an increase

of 2.213 units, and to 'ascending_level_1' leads to an increase of 2.669 units. For tone 53, changing from 'ascending_level_1' (reference) to 'ascending_level_2' results in a decrease of 10.918 units. For tone 55, transitioning from 'none' (reference) to 'level_0' results in an increase of 2.203 units, while moving to 'ascending_level_1' leads to an increase of 2.435 units, and to 'ascending_level_2' results in an increase of 1.984 units. With the visualisation in Figure 5.7, these results suggest that the PMI significantly influences the pitch at the onset of the syllable, with smaller descending PMI and ascending PMI generally contributing to a pitch rise, except for the case of tone 53.

- **Succeeding Melodic Interval (SMI):** Changes in SMI significantly influenced various tones. For tone 11, transitioning from the reference level 'ascending_level_1' to 'ascending_level_2' results in a decrease of 1.086 units in the 1st DCT coefficient. In tone 21, shifting from 'none' to 'ascending_level_1' leads to a decrease of 3.244 units. Tone 213 exhibits increases of 2.689 units and 3.024 units when switching from 'level_0' to 'ascending_level_1' and 'ascending_level_2', respectively. In tone 42, a change from 'decending_level_1' to 'ascending_level_1' results in a decrease of 2.892 units. Additionally, for tone 53, transitioning from 'none' to 'descending_level_2' produces an increase of 13.221 units. Lastly, in tone 55, altering from 'none' to 'level_0' leads to a decrease of 2.248 units. With the visualisation in Figure 5.8), these results suggest that the ascending SMI tends to elevate the pitch at the end of the syllable.

Furthermore, when compared to the results presented in Table 5.1, the effect sizes of tone steps -1 (tone 21), 1 (tones 12 and 23), and the tone step in tone 213 become attenuated. This attenuation is likely due to the variance explained by the aforementioned factors, which were previously attributed to tone steps.

An interesting finding is that singers with professional vocal training show less speech-like characteristics in falling tones (53, 21, 42), while those with more Chaozhou singing experience demonstrate more speech-like features, particularly in rising tones (12, 213, 23) and falling tone (42). This raises a question of whether singers with more

Training background	Never	Seldom	Sometimes	Often
Professional	3	8	5	3
Non-professional	1	3	8	3

Table 5.5: Distribution of Chaozhou singing experience across different training backgrounds.

vocal training tend to have less experience in Chaozhou music. However, a Chi-square test of independence between training background and singing experience in Chaozhou style shows only a weak association ($\chi^2 = 3.543$, $p = 0.315$). As shown in Table 5.5, the distribution of singing frequency is similar across training backgrounds. This distribution suggests that professional training and Chaozhou singing experience are largely independent factors. Professional vocal training (both bel canto and Chinese folk) may emphasise controlled, refined pitch release at the syllable's end, resulting in a diminished tonal realisation in sung syllables with falling tones. Experience in Chaozhou singing appears to enhance tonal features by lowering initial pitch for rising tones.

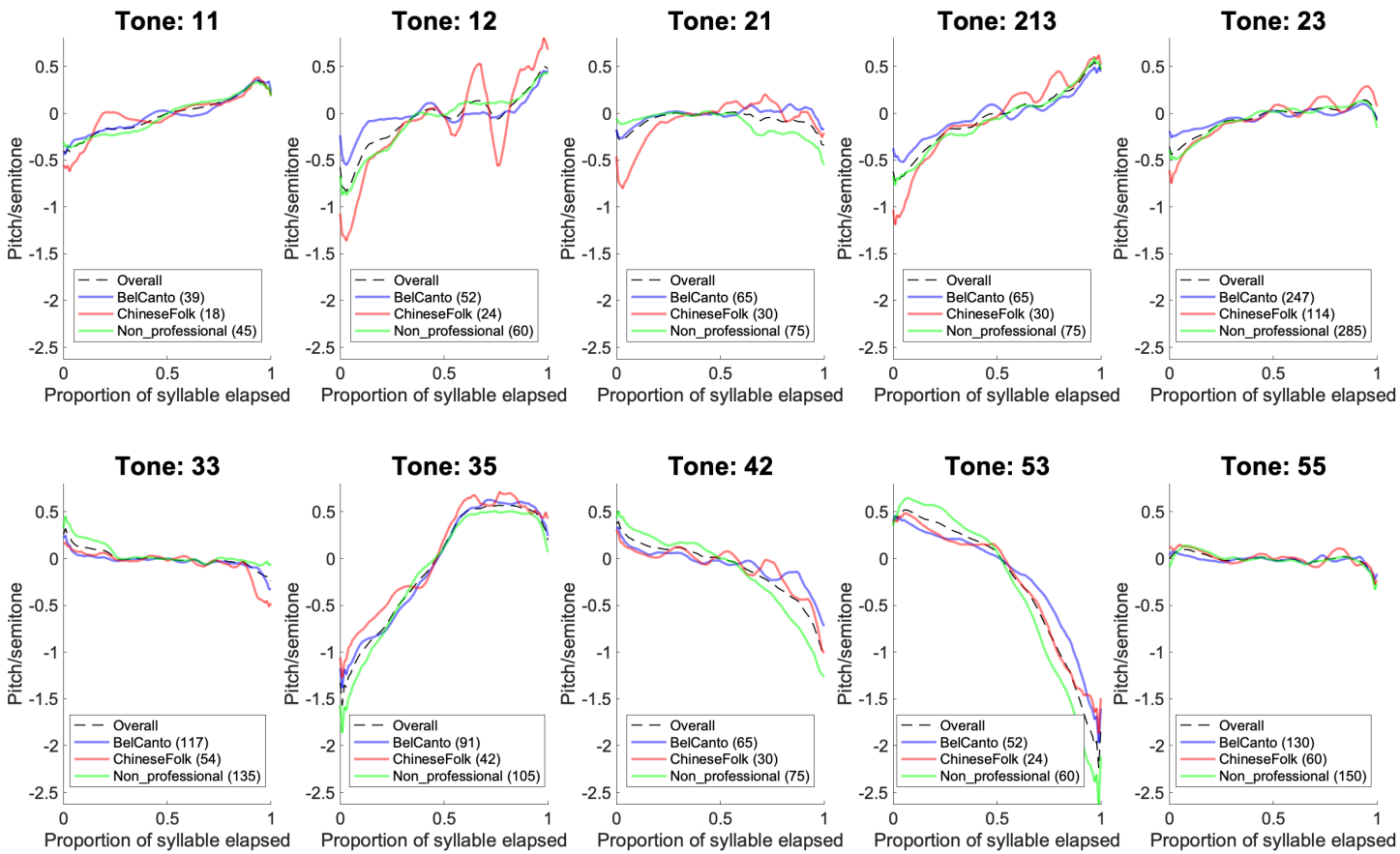


Figure 5.3: Averaged pitch variations split by different vocal training backgrounds. Each subplot corresponds to a specific tone. The dashed black curve represents the overall averaged pitch variation, while the coloured curves correspond to different training background categories. The numbers in the legend indicate the number of individual pitch contours for each category.

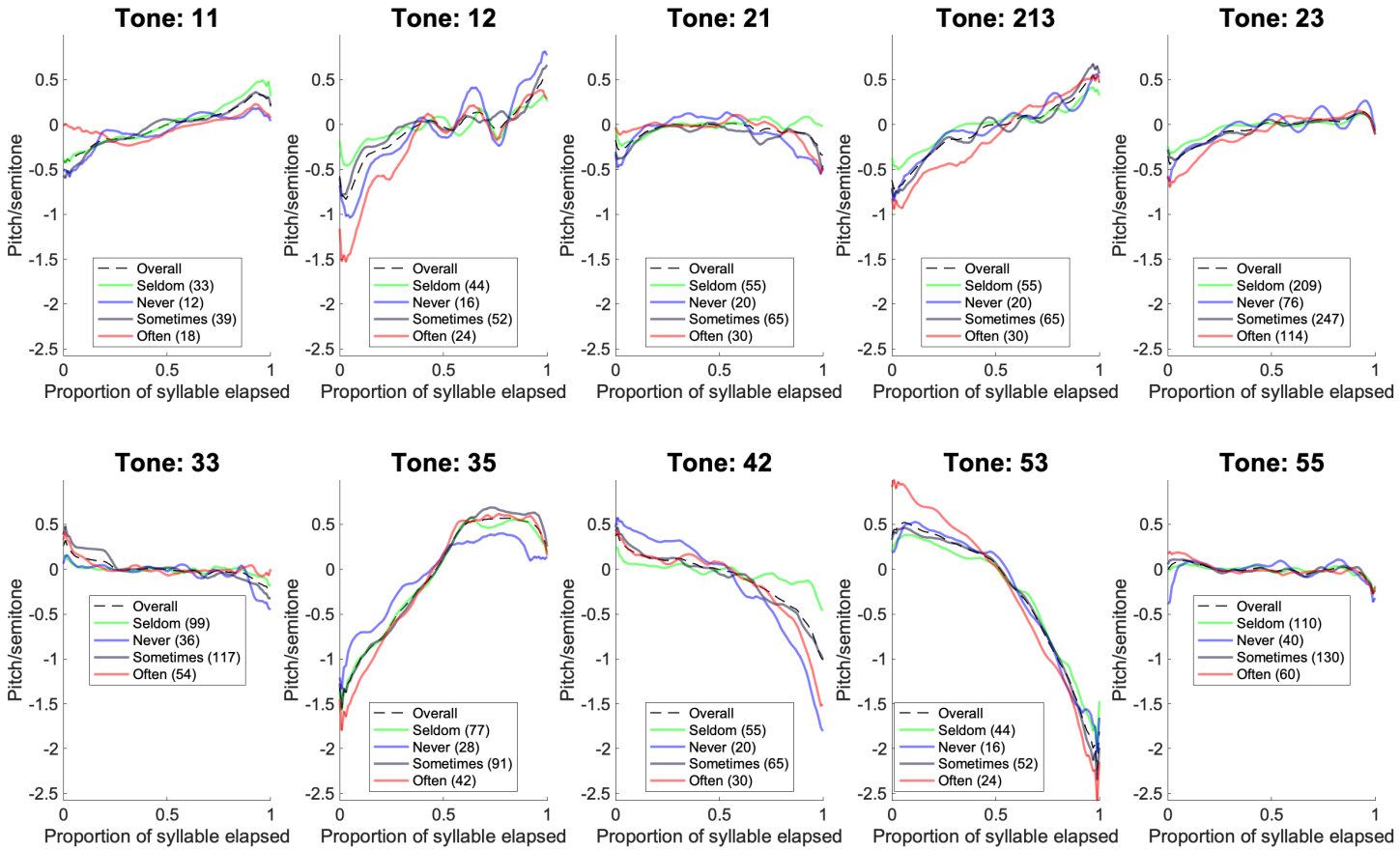


Figure 5.4: Averaged pitch variations split by singing experience in Chaozhou. Each subplot corresponds to a specific tone. The dashed black curve represents the overall averaged pitch variation, while the coloured curves correspond to different singing experience levels. The numbers in the legend indicate the number of individual pitch contours for each category.

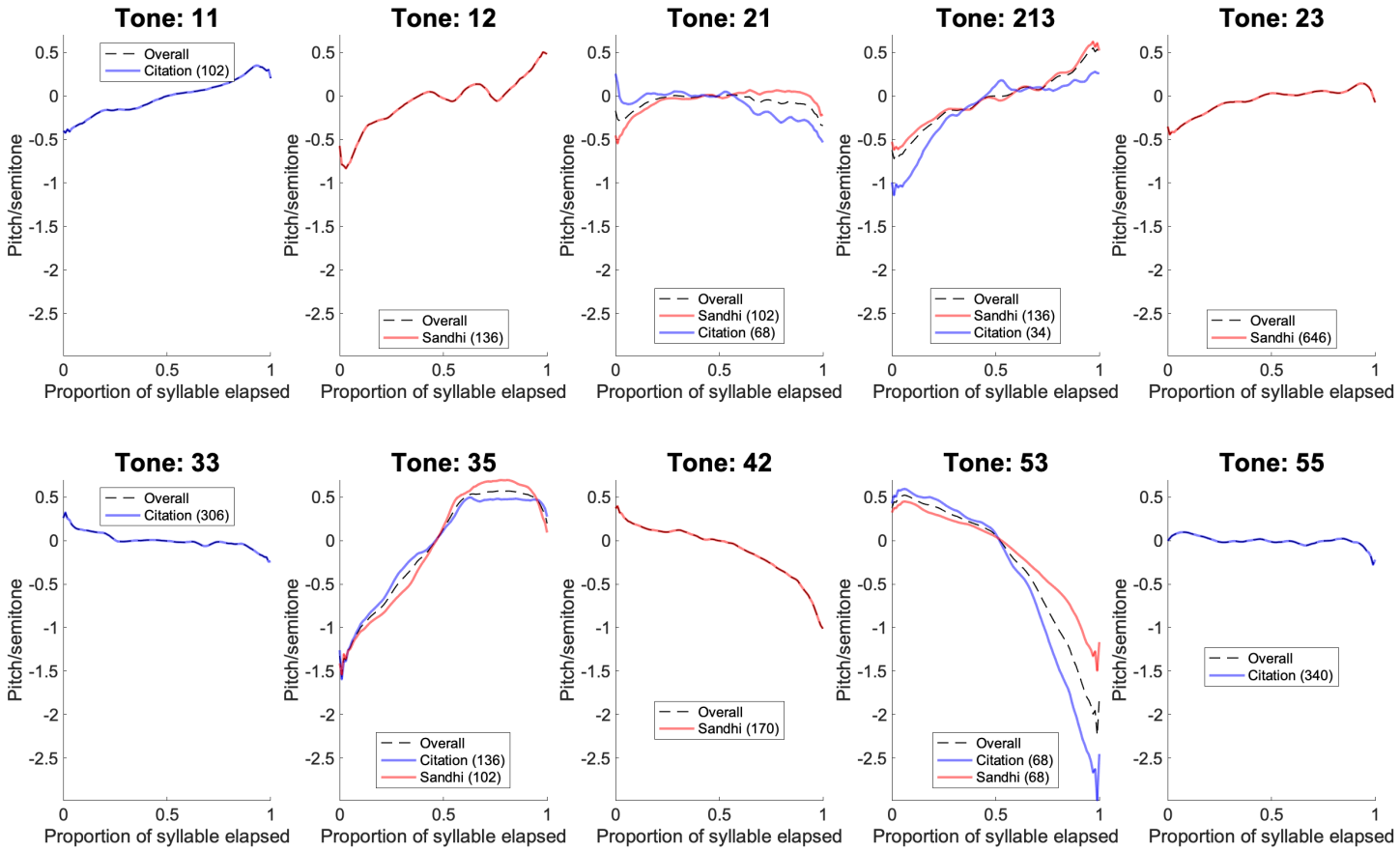


Figure 5.5: Averaged pitch variations split by tone citation and sandhi. Each subplot corresponds to a specific tone. The dashed black curve represents the overall averaged pitch variation, while the coloured curves correspond to tone citation and tone sandhi. The numbers in the legend indicate the number of individual pitch contours for each category.

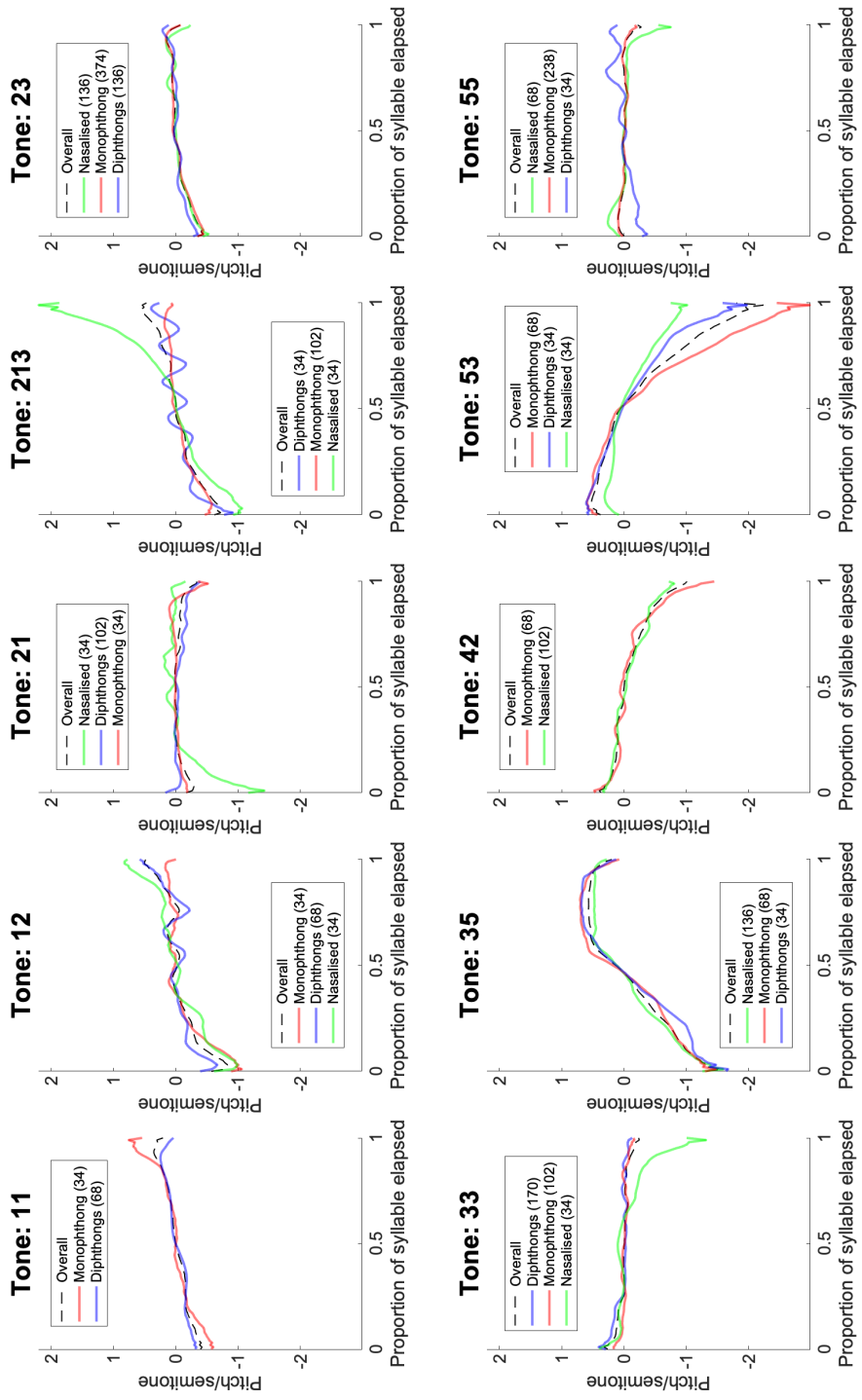


Figure 5.6: Averaged pitch variations split by vowel type. Each subplot corresponds to a specific tone. The dashed black curve represents the overall averaged pitch variation, while the coloured curves correspond to three categories of vowels. The numbers in the legend indicate the number of individual pitch contours for each category.

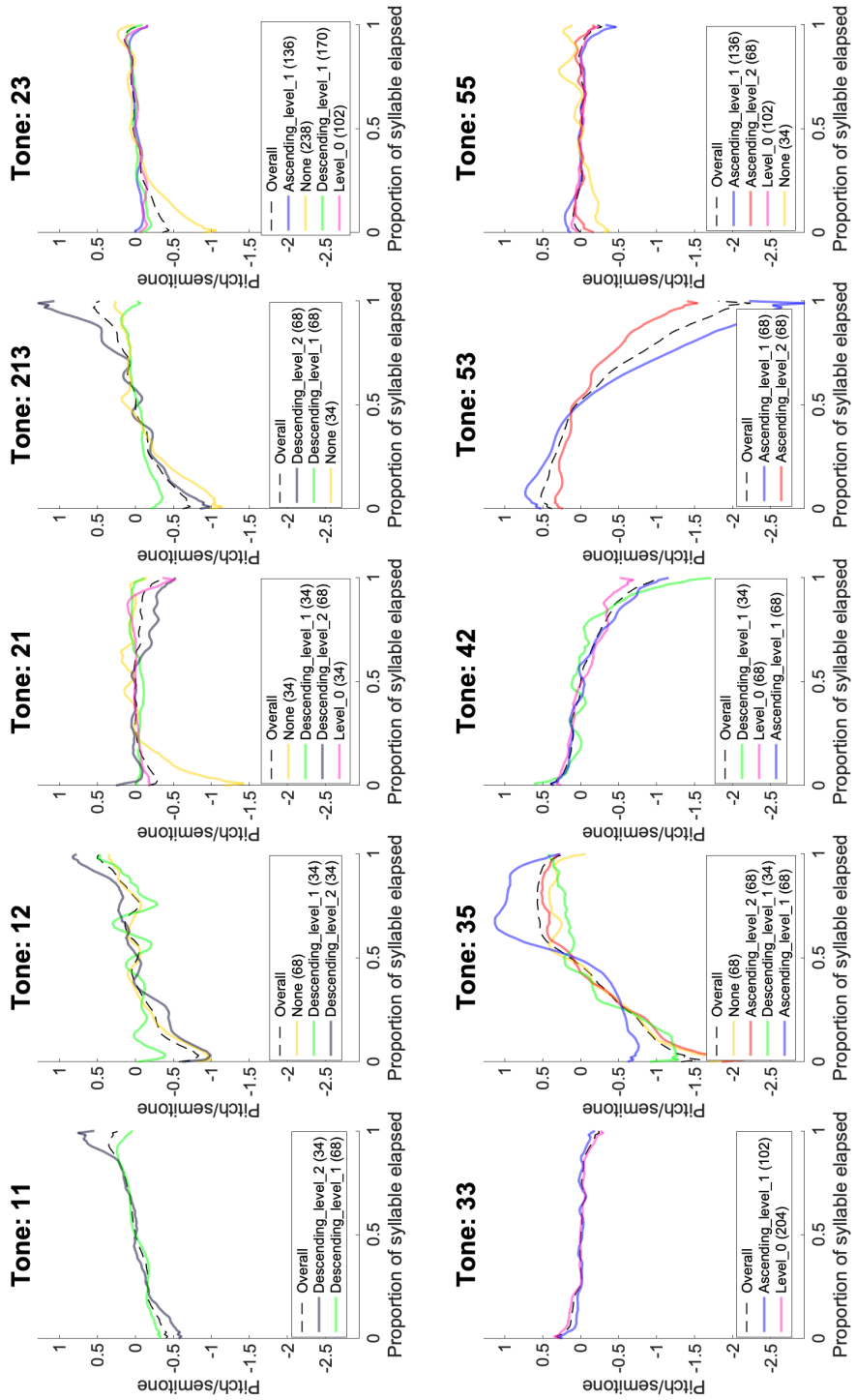


Figure 5.7: Averaged pitch variations split by preceding melodic intervals (PMI). Each subplot corresponds to a specific tone. The dashed black curve represents the overall averaged pitch variation, while the coloured curves correspond to different PMI categories. The numbers in the legend indicate the number of individual pitch contours for each category.

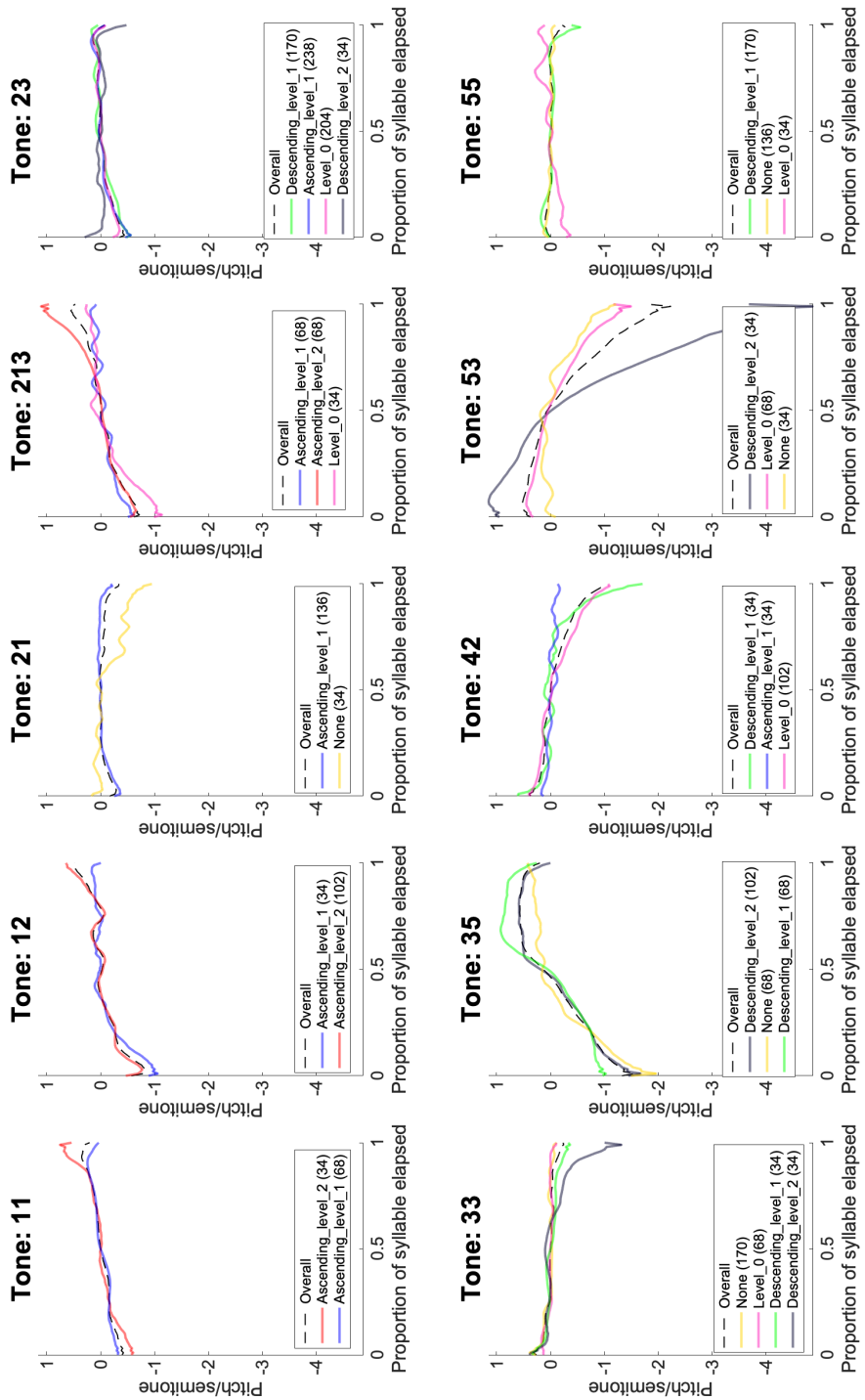


Figure 5.8: Averaged pitch variations split by succeeding melodic intervals (SMI). Each subplot corresponds to a specific tone. The dashed black curve represents the overall averaged pitch variation, while the coloured curves correspond to different SMI categories. The numbers in the legend indicate the number of individual pitch contours for each category.

5.4 Conclusion and Future Work

This chapter has investigated the realisation of lexical tones in Chaozhou folk singing, with a specific focus on pitch contours at the syllable level. By employing the Discrete Cosine Transform (DCT), the overall shape of the pitch contour is captured and parameterised to reflect the linear tendency and curvature of Chinese tones.

The relationship between lexical tones and syllable-level pitch contours in Chaozhou folk music was explored, revealing how tonal variations are preserved or modified in singing. The results indicate that both lexical tone steps and directions significantly influence the linear tendency of the pitch contours, while the curvature of the pitch contour is not significantly affected by the tones. These findings partially align with the traditional Chinese opera principle of "singing according to the syllables" ("YiZiXingQiang"), which has been introduced in Section 2.4.2. While the preservation of tonal direction and step supports this principle, the lack of tonal influence on pitch curvature suggests some deviation from strict syllable-based singing. Additionally, six factors, music training background, experience in singing in the Chaozhou dialect, tone sandhi, vowel type, preceding musical interval, and succeeding musical interval, affect the linear tendency of tone realisation, albeit to a lesser extent than the tones themselves, and exhibit an overall similarity in the manner in which each tone is sung.

The methodological framework developed in this study, using DCT coefficients and statistical analysis of tonal influences, could potentially be adapted to investigate tone-melody relationships in other tonal languages and singing styles. However, the broader applicability of this approach across different genres requires further empirical validation.

Overall, this chapter contributes to the understanding of tone realisation in the Chaozhou folk vocal style, providing insights into the complex interplay between spoken and sung pitch contours. The findings underscore the importance of considering both linguistic and musical factors in the study of Chinese vocal music, paving the way for further research in this interdisciplinary field, particularly when employing a computational approach. However, several limitations of this work have to be discussed to guide future research. First, although the recordings are from 34 singers, the data

is limited to a single song. Expanding the dataset to include more Chaozhou folk songs would provide a broader basis for analysis. Furthermore, it would be valuable to extend the study to other vocal styles in Chinese music, such as traditional Chinese opera or folk music from other regions. Finally, recent advances in technology, such as automatic lyrics transcription (Zhuo et al. 2023), could be utilised to streamline the annotation process.

Chapter 6

Conclusions and Future Perspectives

6.1 Summary

This thesis develops a systematic and computational approach to characterising the pitch aspect of vocal style through pitch contour analysis across diverse musical cultures. It is organised into three case studies: 1. defining and automatically detecting basic pitch contour elements; 2. comparing vocal styles across different cultures by analysing pitch contours in the transitional and held regions of musical notes; 3. investigating lexical tone effects through the characterisation of syllable-level pitch contours. These studies are conducted over a broad range of musical traditions, including Western art music, Jingju, Georgian chants, Russian folk singing, Alpine yodel, and Chinese Chaozhou folk music.

The motivation for this research stems from the need to better understand how pitch contours shape vocal style, addressing gaps in previous studies that often focused on narrow aspects or specific cultural contexts. By employing the proposed automatic pitch contour segmentation method, trained on Jingju pitch contour segments and evaluated using Jingju and Georgian chant data, this thesis provides a systematic framework for pitch contour analysis, enabling the characterisation of both universal and culture-specific expressions in vocal music. This work examines pitch contours at both

the note and syllable levels, depending on the cultural context. The note-level analysis demonstrates the differences and similarities in steady regions and ornaments between Russian folk singing and Alpine yodel, while the syllable-level analysis confirms the effects of lexical tones on the pitch contours of sung syllables in Chinese Chaozhou folk music.

6.1.1 Pitch Contour Segmentation and Characterisation Methods

To characterise complex pitch variations in vocal music, Chapter 3 first defined three fundamental pitch contour region types: steady, modulating, and transitory. To enable the automatic segmentation of the pitch contour into these regions, the concept of the Pitch Contour Unit (PCU) was introduced, which represents discrete segments of the f0 signal delineated by consecutive local peaks and troughs. Positioned between individual frames and notes, PCUs effectively bridge the gap between the excessive granularity of frame-level analysis and the subjective variability inherent in note definition. Duration and extent features are estimated from each PCU to serve as input sequences to a Hidden Markov Model (HMM).

The effectiveness of this method is evaluated not only through the pitch contour segmentation task but also demonstrated in downstream tasks focused on detecting portamento and vibrato in Jingju and steady regions in Georgian vocal datasets, respectively. Comparisons with state-of-the-art methods reveal that the proposed approach either outperforms or matches existing techniques in both frame-level and segment-level evaluations.

6.1.2 Note-Level Pitch Contour Analysis

The purpose of Chapter 4 is to conduct a comparative analysis of note-level pitch contours in Alpine and Russian singing, exploring the differences and similarities in vocal styles between these two cultures. The chapter uses datasets that include two versions of note segments of each culture, each transcribed by different experts. Comparative analyses between the two versions of annotations within each culture were performed to evaluate the reliability of the analyse. It was found that both transcribers in each

culture tend to annotate steady-dominant notes, with chi-square tests indicating that variations in note types between transcribers were not significant. The analysis of note boundary displacements revealed that, in the Russian dataset, transcribers PP and OV showed remarkably similar annotation patterns, while in the Alpine dataset, transcribers LS and YW displayed significant differences in note boundary placements. LS adopted a more flexible annotation approach, whereas YW consistently marked note boundaries relative to transitional regions.

To address the subjectivity in note boundary annotations, the chapter introduces the concepts of "held regions" and "transitional regions" within the annotated notes to characterise and compare vocal pitch contours between the two cultures. Steady and modulating regions, which compose the held regions, are characterised separately. The analysis of held regions reveals that Alpine singing exhibits a higher prevalence of level steady regions, suggesting greater pitch stability compared to Russian singing. Additionally, the study highlights significant differences in modulating characteristics, with Alpine singers demonstrating quicker vibrato and Russian singers displaying greater variability in vibrato extent.

Significant differences are also observed in the characteristics of transitional regions between Russian and Alpine vocal styles. The Russian data shows a richer use of glissando (2.69% vs. 1.23%), mordent (3.07% vs. 0.77%), and overshoot (11.82% vs. 9.98%), while relying less on other ornaments compared to Alpine singing. Alpine singing tends to position transitional regions more towards the tail of the note, exhibits greater variation in time intervals segmented by touch notes, and shows longer glissando durations. Alpine singing also displays more varied inflection timings, quicker portamento, and a tendency towards larger downward intervals. In contrast, Russian singing shows a broader distribution in slide durations, greater variability in overshoot correction, and a slight tendency to prepare for upward pitch slides, while Alpine singing tends towards upward slides and shows a slight preference for preparing downward pitch slides.

This analytical framework, through its decomposition of pitch variations into fundamental elements (steady, modulating, and transitory regions), demonstrates potential

for cross-cultural application. While successful in distinguishing Alpine and Russian singing characteristics, its effectiveness across a broader range of vocal traditions awaits further investigation.

6.1.3 Syllable-Level Pitch Contour Analysis

Chapter 5 aimed to investigate the realisation of lexical tones in Chaozhou folk singing, with a specific focus on syllable-level pitch contours. By employing the Discrete Cosine Transform (DCT), the chapter captured and parameterised the overall shape of the pitch contours, reflecting the linear tendency and curvature of Chinese tones. The analysis revealed that lexical tone steps and directions significantly influence the linear tendency of pitch contours in Chaozhou folk singing, although the curvature was not significantly affected by the tones.

Additionally, the study examined other factors, such as music training background, experience in singing the Chaozhou dialect, tone sandhi, vowel type, and musical intervals, which were found to influence tone realisation, though to a lesser extent than the tones themselves. Notably, in all three falling tones, untrained singers exhibited more 'speech-like' singing style than trained singers. The preceding and succeeding musical intervals were found to affect the pitch contour at the beginning and end of the syllables, respectively.

Overall, this chapter demonstrated a computational methodology to analyse tone realisation in singing and contributed to a deeper understanding of the complex interaction between syllable features, melodic features, and sung pitch contours. The findings emphasised the importance of considering both linguistic and musical factors in the study of Chinese vocal music, paving the way for further interdisciplinary research. The computational approach developed in this study, combining DCT parameterization with statistical analysis of tonal influences, offers a systematic framework that could extend beyond Chaozhou folk music to other tonal languages and singing traditions, though such broader applications require additional empirical studies.

6.2 Future Perspectives

During the development of the methods and the writing of this thesis, several promising research ideas surfaced. Although these ideas could not be pursued within this thesis due to time constraints and practical limitations, they hold significant potential for future exploration. I would like to highlight a few of these concepts that could guide further research on pitch contour and singing style characterisation.

The method established in Chapter 3 had limitations in robustly distinguishing modulating and transitory regions, due to the inability of the HMM to model the similarity between neighbouring PCUs. Several solutions are worth exploring: 1) group two or three connected PCUs as a single token, parameterising the shape of each token using DCT and inputting this sequence into the HMM; 2) apply conditional random fields to the PCU sequence, as thet can model the similarity between neighbouring tokens; 3) fine-tune recent large-scale pre-trained music models, such as Li et al. (2023), using the training data.

Additionally, the note transcription method could be improved from two sources. First, the proposed pitch contour segmentation technique could provide prior information and utilise more note segment data, such as that from the Vocalnotes project (Proutskova et al. 2023), to train a more robust and controllable note segmentation model using HMM. Second, a more robust lyrics transcription system, potentially proposed in the future, could be leveraged to further enhance note segmentation.

Moreover, an application for singing pitch contour analysis is planned. This application will offer an interface to visualise the pitch contour segmentation, characterisation, and statistical results obtained from the methods proposed in this thesis, enabling users to annotate and correct pitches, notes, syllables, and ornaments. Several contributions are promising: 1) By leveraging AI agent frameworks like LangChain,[1], users can interact with computer algorithms using natural language, making the application accessible to those without programming knowledge.

2) In addition to displaying waveforms, piano rolls, pitch curves, and segment bars, embedding musical notation systems, such as the Global Notation System (Killick 2020),

[1] `https://www.langchain.com/`

can provide better visualisation for musicians; 3) This tool will enable musicologists to create more annotations and access more recordings, potentially contributing to future musicology and AI research. On one hand, an expanded data scale can lead to more general and convincing musicological findings; On the other hand, improvements in ornament detection and folk music generation could be achieved with the availability of more recordings and expert annotations, which are currently limited in the field of music. 4) This application could also serve as a platform for musicologists to share data and knowledge, fostering communication on music learning, appreciation, and musicology research, as well as supporting the protection and transmission of intangible cultural heritage.

Bibliography

Anisah, S. (2023), An Analysis of Connotative in English Song Lyrics of Taylor Swift in "Red Album", PhD thesis, UIN Raden Intan Lampung.

Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. (2020), 'wav2vec 2.0: A framework for self-supervised learning of speech representations', *Advances in Neural Information Processing Systems* **33**, 12449–12460.

Bao, Z. (1999), *The Structure of Tone*, Oxford University Press, Oxford.

Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H. & Klapuri, A. (2013), 'Automatic music transcription: Challenges and future directions', *Journal of Intelligent Information Systems* **41**(3), 407–434.

Black, D. A. A., Li, M. & Tian, M. (2014), Automatic identification of emotional cues in chinese opera singing, *in* 'Proceedings of the 13th International Conference on Music Perception and Cognition (ICMPC)', Seoul, South Korea.

Boersma, P. (1993), Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, *in* 'Proceedings of the Institute of Phonetic Sciences', Vol. 17, Amsterdam, Netherlands, pp. 97–110.

Botev, Z. I., Grotowski, J. F. & Kroese, D. P. (2010), 'Kernel density estimation via diffusion', *Annals of Statistics* **38**(5), 2916–2957.

Brown, S. (2017), 'A joint prosodic origin of language and music', *Frontiers in Psychology* **8**, 1894.

Brown, S. & Jordania, J. (2013), 'Universals in the world' s musics', *Psychology of Music* **41**(2), 229–248.

Camacho, A. & Harris, J. G. (2008), 'A sawtooth waveform inspired pitch estimator for speech and music', *The Journal of the Acoustical Society of America* **124**(3), 1638–1652.

Caro Repetto, R., Gong, R., Kroher, N. & Serra, X. (2015), Comparison of the singing style of two jingju schools, *in* 'Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)', Málaga, Spain, pp. 507–513.

Caro Repetto, R., Zhang, S. & Serra, X. (2017*a*), Quantitative analysis of the relationship between linguistic tones and melody in jingju using music scores, *in* 'Proceedings of the 4th International Workshop on Digital Libraries for Musicology', pp. 41–44.

Caro Repetto, R., Zhang, S. & Serra, X. (2017*b*), Quantitative analysis of the relationship between linguistic tones and melody in jingju using music scores, *in* 'Proceedings of the 4th International Workshop on Digital Libraries for Musicology', pp. 41–44.

Chao, Y. R. (1930), ' sistim v "toun-letəz"', *Le Maître Phonétique* **8**(30), 24–27.

Chen, M. Y. (2000), *Tone sandhi: Patterns across Chinese dialects*, Vol. 92, Cambridge University Press.

Clarisse, L. P., Martens, J. P., Lesaffre, M., De Baets, B., De Meyer, H. & Leman, M. (2002), An auditory model based transcriber of singing sequences, *in* 'Proceedings of the International Conference on Music Information Retrieval (ISMIR)'.

Dai, J. (2019), Modelling Intonation and Interaction in Vocal Ensembles, PhD thesis, Queen Mary University of London.

Dai, J. & Dixon, S. (2016), Analysis of vocal imitations of pitch trajectories, *in* 'Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)', pp. 87–93.

Dai, J. & Dixon, S. (2019), 'Intonation trajectories within tones in unaccompanied soprano, alto, tenor, bass quartet singing', *The Journal of the Acoustical Society of America* **146**(2), 1005–1014.

Datta, A. K., Solanki, S. S., Sengupta, R., Chakraborty, S., Mahto, K. & Patranabis, A. (2017), Pitch transition and pitch stability, *in* A. K. Datta, R. Sengupta & S. Chakraborty, eds, 'Signal Analysis of Hindustani Classical Music', Springer, pp. 83–100.

De Cheveigne, A. (2005), Pitch perception models, *in* 'Pitch: Neural Coding and Perception', Springer, pp. 169–233.

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.

de Krom, G. & Bloothooft, G. (1995), Timing and accuracy of fundamental frequency changes in singing, *in* 'Proceedings of the 13th International Congress of Phonetic Sciences (ICPhS)', Vol. 95, pp. 206–209.

De Mulder, T., Martens, J., Lesaffre, M., Leman, M., De Baets, B. & De Meyer, H. (2004), Recent improvements of an auditory model based front-end for the transcription of vocal queries, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 4, IEEE, pp. 257–260.
URL: *http://ieeexplore.ieee.org/document/1326812/*

Demirel, E., Ahlbäck, S. & Dixon, S. (2020), Automatic lyrics transcription using dilated convolutional neural networks with self-attention, *in* 'International Joint Conference on Neural Networks'. arXiv: 2007.06486.
URL: *http://arxiv.org/abs/2007.06486*

Demirel, E., Ahlbäck, S. & Dixon, S. (2021), MSTRE-net: Multistreaming acoustic modeling for automatic lyrics transcription, *in* 'Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)', Online.

Devaney, J. (2011), An Empirical Study of the Influence of Musical Context on Intonation Practices in Solo Singers and SATB Ensembles, PhD thesis, McGill University.

Dibben, N. (1994), 'The cognitive reality of hierarchic structure in tonal and atonal music', *Music Perception* **12**(1), 1–25.

Dixon, S. (2000), On the computer recognition of solo piano music, *in* 'Proceedings of the Australasian Computer Music Conference', pp. 31–37.

Dong, W. (2004), '论润腔', 中国音乐 (4), 62–74.

Downie, J. S., Futrelle, J. & Tcheng, D. K. (2004), The international music information retrieval systems evaluation laboratory: Governance, access, and security, *in* 'Proceedings of the 5th International Society for Music Information Retrieval Conference (ISMIR)'.

Driedger, J., Balke, S., Ewert, S. & Müller, M. (2016), Template-based vibrato analysis in music signals, *in* 'Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)', New York, NY, USA, pp. 566–572.

Dubnowski, J., Schafer, R. & Rabiner, L. (1976), 'Real-time digital hardware pitch detector', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **24**(1), 2–8.

Edmonds, T. J. & Howard, D. M. (2023), 'An investigation in the measurable differences between pitch perception in the voice and pitch perception of external sound sources', *Journal of Voice* .

Fell, M., Cabrio, E., Tikat, M., Michel, F., Buffa, M. & Gandon, F. (2023), 'The wasabi song corpus and knowledge graph for music lyrics analysis', *Language Resources and Evaluation* **57**(1), 89–119.

Fletcher, N. H., Tarnopolsky, A., Lai, J. C. S., Thwaites, S. & Hollenberg, L. C. L. (2001), 'Vibrato in music', *Acoustics Australia* **29**(3), 97–102.

Fu, Z.-S. & Su, L. (2019), Hierarchical classification networks for singing voice segmentation and transcription, *in* 'Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)'.

Ganguli, K. K., Lele, A., Pinjani, S., Rao, P., Srinivasamurthy, A. & Gulati, S. (2017), Melodic shape stylization for robust and efficient motif detection in hindustani vocal music, *in* 'Proceedings of the 2017 Twenty-Third National Conference on Communications (NCC)', pp. 1–6.

Ganguli, K. K. & Rao, P. (2015), Discrimination of melodic patterns in indian classical music, *in* 'Proceedings of the 2015 Twenty-First National Conference on Communications (NCC)', pp. 1–6.

Ganguli, K. K. & Rao, P. (2018), 'On the distributional representation of ragas: Experiments with allied raga pairs', *Transactions of the International Society for Music Information Retrieval* **1**(1).

Garcia, M. (1856), *Garcia' s New Treatise of the Art of Singing (Revised version ed.)*, Boston: Oliver Ditson Company.

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.
URL: *http://www.mitpressjournals.org/doi/10.1162/COMJ_a_00180*

Gong, R. (2018), Automatic Assessment of Singing Voice Pronunciation: A Case Study with Jingju Music, PhD thesis, Universitat Pompeu Fabra.

Gong, R. & Serra, X. (2018), 'Singing voice phoneme segmentation by hierarchically inferring syllable and phoneme onset positions', *arXiv preprint arXiv:1806.01665* .

Gong, R., Yang, Y. & Serra, X. (2016), Pitch contour segmentation for computer-aided jinju singing training, *in* 'Proceedings of the 13th Sound & Music Computing Conference (SMC)', Hamburg, Germany, pp. 172–178.

Grahn, J. A. (2012), 'Neural mechanisms of rhythm perception: Current findings and future perspectives', *Topics in Cognitive Science* **4**(4), 585–606.

Gu, X., Zeng, W., Zhang, J., Ou, L. & Wang, Y. (2023), 'Deep audio-visual singing voice transcription based on self-supervised learning models', *arXiv preprint arXiv:2304.12082* .

Guo, K. (2021), '"润腔"释义', 中国音乐学 .

Haus, G. & Pollastri, E. (2001), An audio front end for query-by-humming systems, *in* 'International Symposium on Music Information Retrieval', pp. 65–72.

Heng, L. & Wang, M. (2022), 'The many facets of musical listening: Auditory perception mechanisms and learned experiences', Timbre and Orchestration Project Reports.
URL: *https://timbreandorchestration.org/writings/project-reports/the-many-facets-of-musical-listening*

Heo, H. & Lee, K. (2017), 'Robust singing transcription system using local homogeneity in the harmonic structure', *IEICE Transactions on Information and Systems* **E100.D**(5), 1114–1123.
URL: *https://www.jstage.jst.go.jp/article/transinf/E100.D/5/E100.D_2016EDP7387/_article*

Herrera, P. & Bonada, J. (1998), Vibrato extraction and parameterization in the spectral modeling synthesis framework, *in* 'Proceedings of the Digital Audio Effects Workshop (DAFx'98)'.

Holzapfel, A., Benetos, E., Killick, A. & Widdess, R. (2022), 'Humanities and engineering perspectives on music transcription', *Digital Scholarship in the Humanities* **37**(3), 747–764.

Howard, D. (1993), 'Real-time visual displays in speech and singing', *Defence science journal. Delhi* **43**(3), 211–221.

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.

Hsuan-Huei Shih, Narayanan, S. S. & Kuo, C.-C. J. (2002), An HMM-based approach to humming transcription, *in* 'Proceedings of the IEEE International Conference on Multimedia and Expo', IEEE, pp. 337–340.
URL: *http://ieeexplore.ieee.org/document/1035787/*

ITC Sangeet Research Academy (2008), 'Introduction to andolan at the ITC sangeet research academy'. Archived December 21, 2008, at the Wayback Machine.
URL: *https://web.archive.org/web/20081221110108/http://www.itcsra.org/alankar/andolan/andol*

Kawase, A. (2017), Quantitative analysis of traditional folk songs from shikoku district, *in* '2017 International Conference on Culture and Computing (Culture and Computing)', pp. 170–177.

Killick, A. (2020), 'Global notation as a tool for cross-cultural and comparative music analysis', *Analytical Approaches to World Music* **8**(2), 235–279.

Kim, J. W., Salamon, J., Li, P. & Bello, J. P. (2018), Crepe: A convolutional representation for pitch estimation, *in* 'Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 161–165.

Koduri, G. K., Gulati, S., Rao, P. & Serra, X. (2012), 'Rāga recognition based on pitch distribution methods', *Journal of New Music Research* **41**(4), 337–350.

Lee, S. W., Dong, M. & Chan, P. Y. (2011), Analysis for vibrato with arbitrary shape and its applications to music, *in* 'Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)'.

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

Li, G. & Li, N. (2006), '关于京剧唱腔的调式探索', 中国戏曲学院学报 **27**(1), 64–66.

Li, Y., Demirel, E., Proutskova, P. & Dixon, S. (2021), Phoneme-informed note segmentation of monophonic vocal music, *in* 'Proceedings of the 2nd Workshop on NLP for Music and Spoken Audio (NLP4MusA)', pp. 17–21.

Li, Y., Yuan, R., Zhang, G., Ma, Y., Chen, X., Yin, H., Xiao, C., Lin, C., Ragni, A., Benetos, E., Gyenge, N., Dannenberg, R., Liu, R., Chen, W., Xia, G., Shi, Y., Huang, W., Wang, Z., Guo, Y. & Fu, J. (2023), 'MERT: Acoustic music understanding model with large-scale self-supervised training', *arXiv preprint arXiv:2306.00107* .

Lin, L. (1995), 新編潮州音字典, 汕頭大學出版社.

Ma, D., Ryant, N. & Liberman, M. (2022), Inferring pitch from coarse spectral features, *in* 'Proceedings of Meetings on Acoustics', Vol. 50, AIP Publishing.

Mainka, A., Poznyakovskiy, A., Platzek, I., Fleischer, M., Sundberg, J. & Mürbe, D. (2015), 'Lower vocal tract morphologic adjustments are relevant for voice timbre in singing', *PLoS One* **10**(7).

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* 'First International Conference on Technologies for Music Notation and Representation (TENOR 2015)', pp. 23–30.

Mauch, M. & Dixon, S. (2014), pYIN: A fundamental frequency estimator using probabilistic threshold distributions, *in* 'Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', Florence, Italy, pp. 659–663.

Mayor, O., Bonada, J. & Loscos, A. (2006), The singing tutor: Expression categorization and segmentation of the singing voice, *in* 'Proceedings of the AES 121st Convention'.

McNab, R. J., Smith, L. A. & Witten, I. H. (1995), Signal processing for melody transcription, Technical report, University of Waikato, Department of Computer Science, Hamilton, New Zealand. Working paper 95/22.

Miao, T., Ji, L. & Guo, N. (1985), 中国音乐词典, 人民音乐出版社.

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

Molina, E., Tardón, L. J., Barbancho, A. M. & Barbancho, I. (2014), 'Sipth: Singing transcription based on hysteresis defined on the pitch-time curve', *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **23**(2), 252–263.

Molina, E., Tardón, L. J., Barbancho, A. M. & Barbancho, I. (2015), 'SiPTH: Singing transcription based on hysteresis defined on the pitch-time curve', *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **23**(2), 252–263.
URL: *http://ieeexplore.ieee.org/document/6837431/*

Mori, H., Odagiri, W., Kasuya, H. & Honda, K. (2004), Transitional characteristics of fundamental frequency in singing, *in* 'Proceedings of the 18th International Congress on Acoustics', Kyoto, Japan.

Müller, M., Rosenzweig, S., Driedger, J. & Scherbaum, F. (2017), Interactive fundamental frequency estimation with applications to ethnomusicological research, *in* 'Proceedings of the 2017 AES International Conference on Semantic Audio', Audio Engineering Society.

Mzhavanadze, N. & Scherbaum, F. (2020), 'Svan funeral dirges (zär): Musicological analysis', *Musicologist* **4**(2), 168–197.

Nakano, T., Goto, M. & Hiraga, Y. (2006), An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features, *in* 'Proceedings of the Ninth International Conference on Spoken Language Processing (ICSLP 2006)'.

Nishikimi, R., Nakamura, E., Fukayama, S., Goto, M. & Yoshii, K. (2019), Automatic singing transcription based on encoder-decoder recurrent neural networks with a

weakly-supervised attention mechanism, *in* 'Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 161–165.

Noll, A. M. (1967), 'Cepstrum pitch determination', *The Journal of the Acoustical Society of America* **41**(2), 293–309.

OpenAI (2023), 'GPT-4 technical report', *arXiv preprint arXiv:2303.08774* .
URL: *https://arxiv.org/pdf/2303.08774.pdf*

Ozaki, Y., McBride, J., Benetos, E., Pfordresher, P. Q., Six, J., Tierney, A. T., Proutskova, P., Sakai, E., Kondo, H., Fukatsu, H. et al. (2021), Agreement among human and annotated transcriptions of global songs, *in* 'Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)'.

Ozaslan, T. H. & Arcos, J. L. (2011), Automatic vibrato detection in classical guitar recordings, Technical Report TR-IIIA-2011-05, Artificial Intelligence Research Institute (IIIA-CSIC).

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

Panteli, M., Benetos, E. & Dixon, S. (2018), 'A review of manual and computational approaches for the study of world music corpora', *Journal of New Music Research* **47**(2), 176–189.

Panteli, M., Bittner, R., Bello, J. P. & Dixon, S. (2017), Towards the characterization of singing styles in world music, *in* 'Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 636–640.

Porter, A., Sordo, M. & Serra, X. (2013), Dunya: A system for browsing audio music collections exploiting cultural context, *in* 'Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)', ISMIR, Curitiba, Brazil, pp. 101–106.

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.

Proutskova, P., McBride, J., Ozaki, Y., Chiba, G., Li, Y., Yu, Z., Yue, W., Crowdus, M., Zuckerberg, G., Velichkina, O. et al. (2023), The VocalNotes dataset, *in* 'Proceedings of the Late Breaking Demo Session at the 24th International Society for Music Information Retrieval Conference (ISMIR 2023)', Milan, Italy.

Proutskova, P., Velichkina, O., McBride, J., Chiba, G., Crowdus, M., Nikolaenko, Y., Ozaki, Y., Shuster, L., Yu, Z., Yue, W., Zuckerberg, G., Killick, A., Li, Y., Phillips, E. & Savage, P. E. (2024), 'Vocalnotes methodology: Framework, challenges and lessons', *Analytical Approaches to World Musics Journal* . In preparation.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. & Sutskever, I. (2023), Robust speech recognition via large-scale weak supervision, *in* 'Proceedings of the International Conference on Machine Learning (ICML)', PMLR, pp. 28492–28518.

Rao, P., Murthy, H. A. & Prasanna, R. M. S. (2023), *Indian Art Music: A Computational Perspective*, Sriranga Digital Software Technologies Pvt. Ltd.

Regnier, L. & Peeters, G. (2009), Singing voice detection in music tracks using direct voice vibrato detection, *in* 'Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', pp. 1685–1688.

Rosenzweig, S., Scherbaum, F. & Müller, M. (2019), Detecting stable regions in frequency trajectories for tonal analysis of traditional georgian vocal music, *in* 'Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)', pp. 352–359.

Rosenzweig, S., Scherbaum, F., Shugliashvili, D., Arifi-Müller, V. & Müller, M. (2020), 'Erkomaishvili dataset: A curated corpus of traditional georgian vocal music for

computational musicology', *Transactions of the International Society for Music Information Retrieval* **3**(1).

Ross, M., Shaffer, H., Cohen, A., Freudberg, R. & Manley, H. (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* 'Proceedings of the Digital Audio Effects Workshop (DAFx'99)', pp. 1–4.

Rossing, T. D. & Sundberg, J. (1984), 'Voice timbre in solo and choir singing: Is there a difference?', *The Journal of the Acoustical Society of America* **76**(S1), S41–S41.

Ryynänen, M. P. & Klapuri, A. P. (2004), Modelling of note events for singing transcription, *in* 'ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing'.

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**(3-4), 405–417.

Sambamoorthy, P. (1958), *South Indian Music, Book 1-3*, Madras: The Indian Music Publishing House.

Sapp, C. S. (2006), 'Mazurka Project plugins for Sonic Visualiser', *Poslední aktualizace* **6**(5).

Seashore, C. E. (1931), 'The natural history of the vibrato', *Proceedings of the National Academy of Sciences* **17**(12), 623–626.

Seashore, C. E. (1937), 'The psychology of music', *Music Educators Journal* **23**(4), 30–33.

Shanghai Art Research Institute & Shanghai Branch of the Chinese Dramatists Association (1981), 中国戏曲曲艺词典, 上海辞书出版社.

Sharma, A. & Salgaonkar, A. (2023), Raga recognition using neural networks and n-grams of melodies, *in* 'Computer Assisted Music and Dramatics: Possibilities and Challenges', Springer, pp. 93–109.

Shen, Q. (1982), '音腔论', 中央音乐学院学报 (4), 13–21.

Shu, T. (2018), '京剧演唱中的嗽音、擞音和颤音', 中国京剧 pp. 64–66.

Singh, Y., Gupta, Y., Patar, S., Saraswat, A. & Biswas, A. (2023), Transcription of indian classical music using convolutional recurrent neural network and CTC loss, *in* 'Proceedings of the 2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)', IEEE, pp. 1–6.

Stoll, G. (1984), 'Pitch of vowels: Experimental and theoretical investigation of its dependence on vowel quality', *Speech Communication* **3**(2), 137–147.

Straehley, I. C. & Loebach, J. L. (2014), 'The influence of mode and musical experience on the attribution of emotions to melodic sequences', *Psychomusicology: Music, Mind, and Brain* **24**(1), 21–34.

Sundberg, J. (1995*a*), Acoustic and psychoacoustic aspects of vocal vibrato, *in* 'Vibrato', Citeseer, pp. 35–62.

Sundberg, J. (1995*b*), 'Vocal fold vibration patterns and modes of phonation', *Folia Phoniatrica et Logopaedica* **47**(4), 218–228.

Sundberg, J., Gu, L., Huang, Q. & Huang, P. (2012), 'Acoustical study of classical peking opera singing', *Journal of Voice* **26**(2), 137–143.

Talkin, D. (1995), A robust algorithm for pitch tracking (RAPT), *in* W. B. Kleijn & K. K. Paliwal, eds, 'Speech Coding and Synthesis', Elsevier, pp. 495–518.

Thompson, W., Bullot, N. & Margulis, E. (2023), 'The psychological basis of music appreciation: Structure, self, source', *Psychological Review* **130**(1), 260–284.

Ventura, J., Sousa, R. & Ferreira, A. (2012), Accurate analysis and visual feedback

of vibrato in singing, *in* 'Proceedings of the 2012 5th International Symposium on Communications, Control and Signal Processing (ISCCSP)', pp. 1–6.

Viitaniemi, T., Klapuri, A. & Eronen, A. (2003), A probabilistic model for the transcription of single-voice melodies, *in* 'Proceedings of the 2003 Finnish Signal Processing Symposium, FINSIG'03', pp. 59–63.

Von Coler, H. & Roebel, A. (2011), Vibrato detection using cross correlation between temporal energy and fundamental frequency, *in* 'Proceedings of the Audio Engineering Society Convention 131'.

Wang, J.-Y. & Jang, J.-S. R. (2021), On the preparation and validation of a large-scale dataset of singing transcription, *in* 'Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 276–280.

Wang, R. (2011), '京剧润腔研究', 戏曲艺术 **3**, 1–11.

Wang, X., Xu, W., Yang, W. & Cheng, W. (2022), Musicyolo: A sight-singing onset/offset detection framework based on object detection instead of spectrum frames, *in* 'Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 396–400.

Wen, X. & Sandler, M. (2008), Analysis and synthesis of audio vibrato using harmonic sinusoids, *in* 'Audio Engineering Society Convention 124', Audio Engineering Society.

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* 'Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 85–88.

Windsor, W. L., Aarts, R., Desain, P., Heijink, H. & Timmers, R. (2000), On time: The influence of tempo, structure, and style on the timing of grace notes in skilled musical performance, *in* 'Rhythm Perception and Production', Swets & Zeitlinger Lisse, pp. 217–223.

Xu, Q., Baevski, A. & Auli, M. (2021), 'Simple and effective zero-shot cross-lingual phoneme recognition using wav2vec 2.0', *arXiv preprint arXiv:2109.11680* .

Yang, J., Huang, Y. & Everett, W. (2017), Callas as violetta: A computer-assisted study on her recorded performances of la traviata, *in* 'Proceedings of the 2017 International Computer Music Conference (ICMC)', Shanghai, China.

Yang, L. (2017), Computational Modelling and Analysis of Vibrato and Portamento in Expressive Music Performance, PhD thesis, Queen Mary University of London.

Yang, L., Chew, E. & Rajab, K. Z. (2015), Logistic modeling of note transitions, *in* 'Proceedings of the International Conference on Mathematics and Computation in Music', Springer, pp. 161–172.

Yang, L., Maezawa, A., Smith, J. B. L. & Chew, E. (2017), Probabilistic transcription of sung melody using a pitch dynamic model, *in* 'Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 301–305.
URL: *http://ieeexplore.ieee.org/document/7952166/*

Yang, L., Rajab, K. Z. & Chew, E. (2013), Vibrato performance style: A case study comparing erhu and violin, *in* 'Proceedings of the International Conference on Computer Music Modeling and Retrieval (CMMR)', Marseille, France.

Yang, L., Rajab, K. Z. & Chew, E. (2017), 'The filter diagonalisation method for music signal analysis: Frame-wise vibrato detection and estimation', *Journal of Mathematics and Music* **11**(1), 42–60.

Yang, L., Rajab, S.-K., Chew, E. et al. (2016), Ava: An interactive system for visual and quantitative analyses of vibrato and portamento performance styles, *in* 'Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)'.

Yang, L., Tian, M., Chew, E. et al. (2015*a*), Vibrato characteristics and frequency histogram envelopes in beijing opera singing, *in* 'Proceedings of the Fifth International Workshop on Folk Music Analysis (FMA)', Paris, France.

Yang, L., Tian, M., Chew, E. et al. (2015*b*), Vibrato characteristics and frequency histogram envelopes in beijing opera singing, *in* 'Proceedings of the Fifth International Workshop on Folk Music Analysis (FMA)'.

Yong, S., Su, L. & Nam, J. (2023), A phoneme-informed neural network model for note-level singing transcription, *in* 'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 1–5.

Zhang, S., Caro Repetto, R. & Serra, X. (2017), Understanding the expressive functions of jingju metrical patterns through lyrics text mining, *in* 'Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)', pp. 612–618.

Zhang, X. (2024), 'Tones shape notes: The realization of lexical tones in chaozhou songs', *Psychology of Music* .

Zhang, X. & Cross, I. (2021*a*), 'Analysing the relationship between tone and melody in chaozhou songs', *Journal of New Music Research* **50**(4), 299–311.

Zhang, X. & Cross, I. (2021*b*), Effect of tone sandhi on singing in chaozhou dialect, *in* 'Proceedings of the Analytical Approaches to World Music Conference (AAWM)', Paris, France.

Zhuo, L., Yuan, R., Pan, J., Ma, Y., Li, Y., Zhang, G., Liu, S., Dannenberg, R., Fu, J., Lin, C. et al. (2023), 'LyricWhiz: Robust multilingual zero-shot lyrics transcription by whispering to ChatGPT', *arXiv preprint arXiv:2306.17103* .