Thesis Proposal

Towards Artificial Musicians: Modeling Style for Music Composition, Performance, and Synthesis via Machine Learning

Shuqi Dai
Jan 2024

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee
Roger B. Dannenberg, *Chair*
Chris Donahue
Junyan Zhu
Julius O. Smith (Stanford University)
Gus Guangyu Xia (MBZUAI)

Thesis proposal
Submitted for the degree of Doctor of Philosophy in Computer Science

Keywords: Computer Music, Music Artificial Intelligence, Generative Artificial Intelligence, Music Generation, Symbolic Music Composition, Expressive Performance Control, Audio Synthesis, Singing Synthesis, Instrumental Synthesis, Music Structure, Music Analysis, Music Style Transfer, Zero-shot Timbre Transfer, AI Ethics

Abstract

The field of Artificial Intelligence Generative Content (AIGC) is increasingly delving into music content creation. However, three fundamental and intricate challenges persist in understanding and creating music: (1) multi-modal music representations, (2) highly complex and logical music structure, and (3) personalized and stylistic music preferences. This thesis tackles these three challenges by focusing on a practical application: creating virtual musicians or "re-creating" existing musicians.

The thesis creates artificial musicians across different music creation levels and representation modalities. (1) For symbolic music composition, I combine music domain knowledge with machine learning models to compose melodies, harmonies, and bass lines while preserving specific styles. (2) Expressive performance control, highly crucial in music creativity but often ignored, is achieved through diffusion models, generating pitch envelopes, dynamics, and playing techniques, capturing the unique performance styles of singers and instrumentalists. (3) Acoustic audio synthesis involves synthesis from scratch and transferring timbres of vocals and instruments, including zero-shot vocal and instrumental synthesis of unseen targets. These layers converge to model composition and musicianship across multi-modal music representations.

The thesis emphasizes music domain knowledge in stylistic and personalized music modeling, and delves into music structure analysis to elevate generation quality. I further discuss the applications of the technologies in this thesis in areas such as music therapy, music education, the theory development of non-Western music, and human-computer interactive live performance. Ethical and legal implications of AI music are also explored, foreseeing its fusion with the future music industry.

The proposal outlines technical foundations and design frameworks for the three music creation levels, rationalizes technology choices, presents achievements, and offers solutions for pending tasks. The contribution, success criteria, future prospects, and research schedule are discussed.

Contents

1 Introduction

My favorite singer is Teresa Teng, yet she passed away shortly after my birth. I grew up listening to her soulful melodies, and I have always dreamt of attending her live concerts, performing with her, and hearing her new songs. This is the initial motivation for this thesis: to create an artificial Teresa and bring her back.

In the current surge of generative artificial intelligence, the fusion of deep learning and past computer music technologies has illuminated the path to creating artificial musicians. Recent advancements showcase promising results in AI-generated music, such as large music representation models [39], large text-to-audio music generation models [1, 26, 61, 112], automatic music generation for video [40], virtual singers [85, 86, 142, 144], and so on. However, as a musician, I realize that these models still fall short of the music created by professional musicians. Human creativity in music remains largely unexplored. For instance, What makes music resonate deeply within us? Why does certain music transcend time, stirring memories across generations? How does music influence our emotions, and how do musicians convey their messages and emotions through music? What interplay of musical elements shapes our perception, and what truly distinguishes music styles and genres at their essence? Beyond these deep questions, even basic tasks like beat tracking and transcription — achievable by many without musical training — fall short in today's algorithms. I don't believe AI can replace humans, but I do believe that we need a better understanding of the limitations and potential of AI in music. Thus, in this thesis, I aim to explore human creativity, with music as my way in and computer science as my tool.

To create artificial musicians, there are three basic modules of the proposed framework: symbolic music composition, expressive performance control, and sound synthesis. These three modules correspond to the intrinsic *multi-level*, *multi-modal* character of music representation: music can be read, listened to, or performed, and it all depends on whether we are relying on *score* (the top-level, abstract representation), *sound* (the bottom-level, concrete representation), or *control* (the intermediate representation) [28, 36]. This is very different from image or language representation. So far, information saliently contained in different levels of music representation cannot be freely converted and manipulated across these modalities, and no end-to-end system can elegantly deal with all levels of music representation together. Consequently, most studies only focus on a certain level/modality of music representation. In this thesis, I aim to cover all three music creation levels and music representation modalities.

A fundamental challenge in music creation is the highly complex and logical repetition and structure of music. The nature of repetition and structure in music is still not well understood, and much remains to be explored with music information retrieval techniques. In this thesis, I use "structure" to refer broadly to organizing principles in music, which are generally hierarchical and include sections, phrases, and various kinds of patterns. As a basic indicator of music structure, repetition includes not just music content within repeat signs but also approximate repetitions at different time scales. Music relies heavily on repetition to create internal references, coherence, and structure. In music generation, many researchers rely on deep learning models to capture music structure and organizing principles implicitly from data. However, repetition, especially long-term repetition structure, does not seem to emerge automatically in deep music generation. We will see in this thesis that phrase structure, song structure, and other elements of music are intertwined. Thus, we need a better understanding of repetition and structure if we want machines to compose or even just listen to music in a more human way, delving into the essence of human creativity in music.

Another significant problem is personalized and stylistic music preferences, which are intertwined with our cultural backgrounds, emotional states, and personal music histories, making them deeply personal and varied across individuals. Recent practice in stylistic music generation mainly focuses on machine learning of general musical rules or style, which tends toward generic musical output with no support for personalization. The number of favorite and familiar songs that a person can provide is insufficient for deep learning approaches, which need large amounts of training data. However, every piece of music has its own distinctive abstract qualities, for example, structure, melodic contour, rhythmic pattern, chord progression, bass line pattern, performance control, sound timbre, etc. These abstract qualities might vary a lot from piece to piece, even within a music genre and in works by the same musician. By focusing on distinctive musical qualities within the constraints of general rules of music, I hope to introduce stylistic models that are able to capture compositional, performance, and timbre styles from examples and imitate them in a new piece.

Data scarcity stands as a formidable obstacle. Across every task in music research, there is a persistent lack of large-scale, high-quality, diverse, and annotated datasets, especially in the day of deep learning and generative AI. To pursue the aimed task, I collected a high-quality multilingual singing dataset with style transfer [33], and helped collect a pop song dataset for symbolic music composition [128], and annotated structure labels for the POP909 dataset [30].

The fusion of AI with music introduces critical ethical and legal quandaries. Issues range from intellectual property rights, given the ambiguous ownership of AI-generated compositions, to concerns about authenticity and the potential erosion of genuine artistic value. Furthermore, the looming threat of job displacement for musicians and composers, potential biases in AI models sidelining underrepresented music styles, and privacy concerns related to personalized AI music experiences underscore the need for careful navigation. As AI becomes more intertwined with the musical landscape, addressing these implications is paramount to preserving human creativity in music. I will address the current challenges and bring discussions to this thesis.

In all, this thesis tackles all the above challenges by focusing on this practical application: creating virtual musicians or "re-creating" existing musicians. The basic ideas for each module are: (1) For symbolic music composition, I combine music domain knowledge with machine learning models to compose melodies, harmonies, and bass lines while preserving specific styles. (2) Expressive performance control, highly crucial in music creativity but often ignored, is achieved through diffusion models, generating pitch envelopes, dynamics, and playing techniques, capturing the unique performance styles of singers and instrumentalists. (3) Acoustic audio synthesis involves syntheses from scratch and transferring timbres of vocals and instruments, including zero-shot vocal and instrumental synthesis of unseen targets.

Apart from creating artificial musicians, I further discuss the applications of the technologies in this thesis in areas such as music therapy, music education, the theory development of non-Western music, and human-computer interactive live performance.

Throughout this thesis, my work focuses on popular music for multiple reasons. The original motivation for much of this work originated in the idea of constructing music for rhythmic auditory stimulation (RAS) for Parkinsons's disease patients, and therapists often use popular music. In addition, the idea of creating virtual models of popular artists suggests working on popular styles of music. Secondly, most people have familiarity and knowledge of popular music styles. Therefore, it is easier to compare, discuss, and evaluate popular music than other music. It is also the case that popular music has many conventions that seem to simplify the music generation problem. This is not to say that truly great popular music is easy to make, but convincing machine-composed "popular" songs exist, suggesting this is a productive place to start.

The minimum achievement for this thesis should be the completion of the design, implementation, and evaluation of the system. The built system should be able to compose stylistic symbolic music, generate expressive performance control, and synthesize singing and instrumental sound in good quality. An insightful discussion on the ethical and legal implications of AI music should be provided. Once I achieve the minimum requirement, I plan to improve the system's quality further and apply the technologies to potential applications.

The proposed thesis will be organized as follows. Each chapter is described in greater detail in the corresponding sections of this proposal, as follows:

Chapter 2 will describe a computational study of music repetition structure, including algorithms on hierarchical music structure analysis [30], structural influences on various music elements [27], how repetition and structure develop over time as revealed with human perception, and implications for automatic music composition.

Chapter 3 will describe two directions toward symbolic music composition with structure, style, and control. [32] introduces a statistical machine-learning model that can capture and imitate the structure, melody, chord, and bass style from a given example seed song. [31] combines deep learning with music domain knowledge to create a full-length melody guided by long-term repetitive structure, chord, melodic contour, and rhythm constraints.

Chapter 4 will outline work on stylistic singing performance control, including models generating performance timing, F0 curves, and amplitude (loudness) envelopes. In Section 4, I discuss current progress on instrumental performance control and design for introducing style and structure information.

Chapter 5 will present work on expressive singing synthesis and zero-shot singing synthesis with unseen speech targets. In Section 5, I discuss the current progress on instrument synthesis and plans for further improvement.

Chapter 6 will discuss potential applications of the technologies in this thesis, such as interactions with other Generative AI fields, music therapy, music education, music perception, the theory development of non-Western music, and human-computer interactive live performance.

Chapter 7 will consider the ethical and legal issues from a historical perspective of the music industry, compare them with current challenges of AI music, compare music with other art areas, and discuss potential technical support.

Section 8 of this proposal presents a schedule for the completion of this thesis.

Section 9 conclude this thesis proposal.

2 Computational Study of Music Repetition and Structure

Repetition and structure have a significant place in music theory, but the structure hierarchy and its influences are often ignored in both music analysis and music generation. Chapter 2 will give an overview of related work 2.1, and describe novel algorithms based on repetition to extract music structure hierarchy from a MIDI data set of popular music and show its effectiveness through evaluation 2.2. Then, I will introduce new data-driven approaches to estimate and validate structural influences in music 2.3. Results show that the automatically detected hierarchical repetition structures reveal significant interactions between structure and harmony, melody, rhythm, and predictability. Different levels of hierarchy interact differently, providing evidence that *structural hierarchy* plays an important role in music beyond simple notions of repetition or similarity. Structure and repetition also influence songs' use of limited vocabulary so that individual songs do not follow general statistics of song collections. Moreover, in 2.4, I analyze that over the course of a song. Repetition is not random, but follows a general trend as revealed by cross-entropy, which is potentially aligned with human perception results in EEG experiments.

All the above findings offer challenges as well as opportunities for deep-learning music generation and suggest new formal music criteria and evaluation methods. Thus, in 2.5, music from recent music generation systems is analyzed and compared to human-composed music in two popular music datasets (Chinese and American), often revealing striking differences from a structural perspective.

Finally, I will summarize and discuss the future possibilities 2.6 of this chapter in highlighting roles that structure can play in music analysis, music similarity, music generation, music evaluation, and other Music Information Retrieval tasks.

2.1 Related Work

2.1.1 Repetition and Structure in Music

Repetition is a key element of music structure. Repetition is one of the three commonly used principles for segmenting music, along with novelty at segment boundaries and homogeneity within segments [97]. People have developed a variety of segmentation and section detection methods based on repetition with acoustic features[34, 102]. Repetition becomes especially useful in segmenting symbolic music or lead sheet representations where timbre and texture may be lacking [29].

Music form and structure, including repetition, is also a major focus of Music Theory [12, 69, 119]. Apart from music theory, repetition also plays an important role in music expectation and prediction [64, 96]. Studies of repetition and structure are common in Music Psychology [88]. For example, listening experiments with reordered Classical and Popular music have shown that listeners are rather insensitive to restructuring, but these results are subtle and somewhat ambiguous [110]. [100] conducts four listening experiments to illustrate the effects of pitch and temporal contributions to musical phrase determination. [80] exposes the relationship between tonal structure and tension-resolution patterns by qualitatively analyzing musical tension ratings for two piano pieces from Mendelssohn and Mozart.

2.1.2 Computational Analysis of Music Structure

Computational analysis of musical form has long been an important task in Music Information Retrieval (MIR). Large-scale structure in music, from classical sonata form to the repeated structure in pop songs, is essential to music analysis as well as composition. Schenkerian analysis, a reduction technique that also aims to uncover musical structure, has been implemented by [89], and the automated reduction has achieved convincing results in recognizing the variation in ten pieces by Mozart. [55] describes a tool for Generative Theory of Tonal Music (GTTM) analysis that matches closely the analyses of musicologists. [2] use unsupervised learning to segment Mozart string quartets. They analyzed the classical sonata form structure from a dataset of Mozart's string quartets and discovered that unsupervised learning emits better section boundaries than manually set parameters. The structure analysis of [49] performs structural analyses using homogeneity, repetitiveness, novelty, and regularity. Our work builds on the idea of extracting structure by discovering repetition.

Identifying hierarchical structure is likely to play a role in music listening. [52] employs ideas from Natural Language Processing (NLP) and performs Combinatory Categorical Grammar parsing to obtain a hierarchical structure of chord sequences. [90] state that advances in the theory of tree structures in music will depend on clarity about data structures and explicit algorithms. [66] propose a two-step segmentation algorithm for analyzing music recordings in predefined sonata form: a thumb-nailing approach for detecting coarse structure and a rule-based approach for analyzing the finer substructure. [5] analyzes music structure in different levels of resolutions based on graph theory and multi-resolution community detection. We present a detailed algorithm for segmenting music into phrases and deriving a higher-level sectional structure starting with a symbolic representation.

Segmentation of music audio is a common MIR task with a substantial literature. [34] survey audio segmentation techniques based on repetition, textural similarity, and contrast. [4] perform audio music segmentation based on timbre and rhythmical properties. However, MIDI has the advantage of greater and more reliable rhythmic information along with the possibility of cleanly separating melody. Many chord recognition algorithms exist, e.g. [91] use a semi-Markov Conditional Random Field model. References to melody extraction from MIDI can be found in [67] who use maximum likelihood and Dynamic Programming. [111] presents an efficient algorithm for spotting matching melodic phrases, which relates to our algorithm for segmentation based on matching sub-segments of music. [87] proposes a music segmentation evaluation measure considering over- and under-segmentation. [21] develop a geometric approach to discover inexact intra-opus patterns in point-set representations of piano sonatas. My work in Section 2.2 introduces new methods for the analysis of multi-level hierarchy in MIDI.

2.1.3 Implications of Music Structure in Music Generation

There are many deep learning models for music generation [10, 39, 59, 103, 109], however, capturing repetition and long-term dependencies in music still remains a challenge. One mainstream approach is to model distribution of music via an intermediate representation (embedding), such as Variational Auto-Encoders (VAE) [109, 130], Generative Adversarial Networks (GANs) [139] and Contrastive Learning [53, 130]. Due to their fixed-length representation and short-length output, it is difficult to exhibit long-term structure. Another popular trend is to use sequential models such as LSTMs and Transformers [59, 103, 123] to generate longer music sequences, but they still struggle to generate repetition and coherent structure on long-term time scales. Some recent works introduce explicit structure planning for music generation, which shows that using

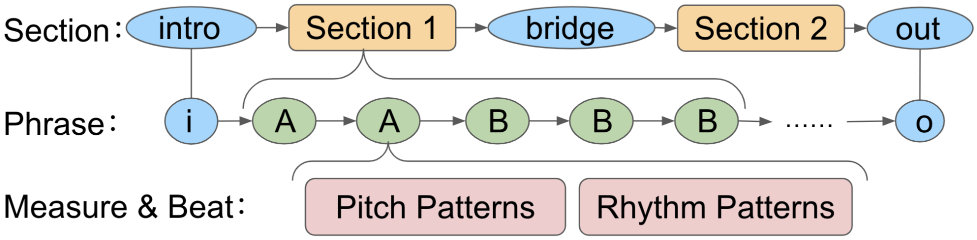


Figure 1: Structure hierarchy in pop music.

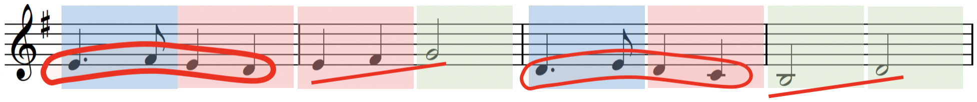


Figure 2: Repeated motives in a phrase in *Yankee Doodle*.

structure information leads to better musicality [22, 31, 93].

Current evaluation methods for music generation rarely consider repetition and structure. Deep music generation systems [59, 60] use objective metrics such as negative log-likelihood, cross-entropy and prediction accuracy to compare generated music with ground-truth human-composed music. But these metrics do not precisely correspond to human perception and are not reliable for musicality. Another trend is using domain-knowledge [17] and musical features [32, 46, 118, 137] such as pitch class, pitch intervals, and rhythm density to evaluate music statistically. However, most of them ignore even short patterns, and none evaluate music structure. In contrast, I offer quantitative and objective methods to evaluate music repetition and structure.

2.2 Analysis of Repetition and Structure Hierarchy

Music structure is hierarchical (Figure 1). It contains multiple levels of repetitions, ranging from low-level pitch and rhythm motives (patterns) to higher-level phrases (analogous to sentences) and sections (analogous to paragraphs). This section introduces novel algorithms to extract repetition structure at all levels from a MIDI data set of popular music, with a statistical exploration of hierarchical structure in real data, which is a combination of two of my previous papers [29][27].

I began the study by developing a method to identify low-level structure "phrases" in popular songs. Given input consisting of a chord and melody sequence for each song together with its time signature (obtained from MIDI pre-processing), the algorithm outputs a repetition structure of phrases, resulting in high agreement with human judgment. Next, I discovered a simple way to infer higher-level structure 'sections' from this phrase-level structure. I will introduce the data collection and processing, design motivation, structure representation, details of the algorithm, and some evaluation results in the full version of the thesis.

There is at least another level of repetition below phrases (Figure 1). For example, in the first 8-measure phrase of the chorus in *Yankee Doodle* (Figure 2), the first and second half repeat elements of rhythm and interval. The colored boxes show repeated rhythm patterns, and the red lines point out repeated pitch contours. Through algorithmic analysis, I found abundant evidence for repetition within phrases. For example, rhythm patterns show a clear repetition structure within phrases; the vocabulary of rhythm and pitch patterns within a song or phrase is also very limited compared to the whole dataset, implying pitch sequence repetitions within the phrase level.

I further characterize the hierarchical structure in the datasets with statistical studies to show the distribution and trend.

Unlike traditional music theory with case-by-case human analysis, I explore these problems with a data-driven approach. For training and testing, I use a Chinese pop song dataset POP909 [128], which has 909 pop song performances in MIDI, and an American pop song dataset PDSA [6], in MusicXML, which has 348 American pop songs originating from 1580 to 1924. I use only songs in 4/4 time to simplify analysis.

2.3 Structural Influence

Can repetition structure be considered orthogonal to melody, harmony, and rhythm generation? If so, we can simply generate music note-by-note, ignoring structure, and then impose structure by repeating generated sequences. However, if there is significant interaction between structure and other facets of music, then structure is an integral part of music analysis, music modeling, and music generation. My findings show the latter is the case.

Another intuition for this structural influence study is that, we could have used any number of ways to form higher-level structure (sections), but we wanted an objective procedure that is independent of musical features (e.g., "sections end on a long tonic note"). Our choice is supported by the finding of interactions between sections, melody, harmony and rhythm that are not explained by interactions at the phrase level, suggesting that the section structure is not just an arbitrary construction. On the other hand, we suspect there are even better constructions in terms of matching human analyses or consistency with musical features.

Thus, in this section, I introduce new data-driven approaches to estimate and validate structural influences in music, based on my previous papers [29][27]. Results show that the automatically detected hierarchical repetition structures reveal systematic interactions between repetition structure and melody, rhythm, harmony, and predictivity. Different levels of hierarchy interact differently, providing evidence that *structural hierarchy* plays an important role in music beyond simple notions of repetition or similarity.

For example, chords at the ends of phrases differ from chords in the middle of phrases. Furthermore, the final chords in sections differ significantly from final chords in phrases that are not at the ends of sections. This does not mean that “good music” must reflect a structural hierarchy, but at least this finding offers insight into how music generation might be improved, and it raises questions for further study.

I further study how musical structure has evolved over decades of popular music writing.

2.4 Repetition and Structure Over Time with Perception

Music is not repeated randomly. After seeing different levels of hierarchy in Section 2.3, we ask: Is there a schema for repetition? How does repetition play out over time?

Huron suggests that if music is to manipulate prediction through repetition, it makes sense to repeat some of the music early on [64]. This affords immediate pleasure from successful prediction rather than delaying until all novel material is exhausted. POP909 supports this hypothesis: More

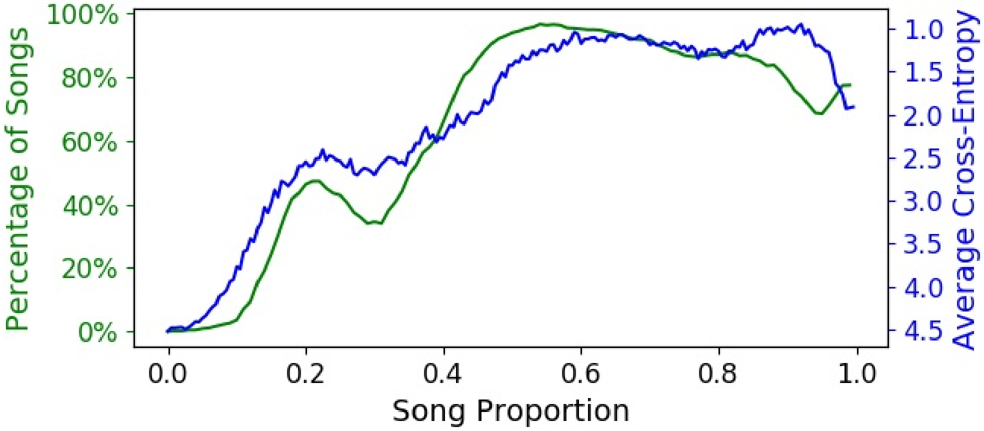


Figure 3: **Left** (green): Percentage of POP909 songs that have phrase repetitions at different song locations. **Right** (blue): Average prediction cross-entropy using variable-order Markov models on diatonic pitches in POP909 songs over time, and note the y-axis is inverted.

than 50% of the phrases repeat immediately, and almost all phrases repeat within a quarter of the song.

In Section 2.2, we have seen that most rhythmic and melodic patterns in a phrase are repetitions. At the song level as well, using the phrase repetition labels in POP909, we found that for 79% of songs, 15% to 35% of their duration consists of new material and the rest is repetition.

Returning to our consideration of structure over time: How does surprise vary with structure? We might expect less surprise at the ends of sections to give a sense of completion or resolution. Figure 3 Left shows a histogram percentage of phrase-level repetition over the course of a song. We see a relatively low repetition rate in the first 1/10 of the song. The repetition rate sharply increases as we progress to the first 1/5 of the song because, after introducing the initial materials, most songs repeat them. There is a noticeable drop around the quarter-way point where many songs introduce new material. In the second half, almost everything is a repetition or variation of what came in the first half. Finally, the graph shows novel material is often introduced near the end. In Figure 3 Right, we use the variable-order Markov model to calculate the average cross-entropy over time on melody pitches. To show the correspondence, we flip the vertical cross-entropy axis. Note the similarity between the trends in repetition and cross-entropy.

From these results, it is clear that repetition is not random but follows a plan in which novelty is revealed, presumably to enhance the enjoyment or effect of the music. This is perhaps surprising because other research shows that music can be substantially rearranged without destroying positive impressions [45]. Whether this organization matters to listeners or simply reflects composers' intentions requires further research.

2.5 Implications For Automatic Music Composition

One application of our studies is to gain insight into deep-learning models for automatic music composition. We can apply the analyses in this Chapter to melodies generated from deep learning models. Through these case studies, we can characterize repetition and structure from deep music generation and compare them to human-composed songs. To be clear, I am not claiming that any particular structure is *necessary* or even *good*. My goal is to illuminate possibilities and better understand both real and generated music. We then discuss our results and point out new directions and ideas for future work in deep music generation.

The detailed content in this section is from my previous work [27]. I studied repetition and structure in deep learning generated melodies to answer three questions: 1) do melodies have multiple levels of repetition and structure? 2) do they have song-specific vocabulary and common patterns? 3) how does cross-entropy vary over time? I used two deep music generation models: One is a VAE model using representation learning [130], chosen because it uses contrastive learning to generate longer sequences (8 bars) than other VAE models. The other model is Music Transformer [59].

The results show that rather than learning and reproducing general statistics of datasets, we need to learn how songs strategically diverge from background or stylistic norms to create interest, surprise, and individuality. It is particularly interesting that phrases can be better predicted by relatively short phrases within the same song than by large amounts of training data from other songs. It seems plausible that songs of the same artist or same sub-genre may be more predictive than songs in general. My findings also reinforce previous findings that using structure in MusicFrameworks [31] results in better human evaluations.

Examples in the case study also suggest that we can compare generated music to real music using measures of structure, repetition, and entropy. Matching these measures is not guaranteed to make music "better," but we should not simply ignore clear objective differences. We would at least expect differences to be small when the task is to imitate a style or genre. We can also speculate that these measures are relevant to listener preferences even if they do not tell a complete story.

This work focuses on popular music, where repetition and hierarchical structure are relatively easy to study compared, for example, to large symphonic works, which show greater variation, development, and orchestration. Nevertheless, we are encouraged by results using variable-order Markov models which make relatively few assumptions on the underlying music. We hope our work will lead to future explorations of Classical and other music.

2.6 Conclusion

It should be no surprise that structure, repetition, pitch, rhythm, harmony and entropy are all strongly connected and interdependent. We have offered new ways to explore these connections objectively, using a data-driven approach without relying on subjective human analyses.

Among our findings are that within-song and within-phrase vocabulary and repetition are not a reflection of more general background statistics from a collection of songs. Instead, songs and phrases gain "individuality" through more repetition and smaller vocabulary. This has important implications for machine learning and music generation systems.

There are clear differences between measurements of real songs and those of many music generation systems, suggesting that there are important gaps to fill through new research. We hope that this work will inspire further research in the roles played by repetition and structure in music as well as methods to learn repetition and structure.

3 Symbolic Music Composition With Structure, Style, and Control

Many practices have recently been presented in symbolic music composition. While stylistic music generation using deep learning techniques has become mainstream, these models still struggle to generate music with high musicality, different levels of music structure, and controllability. In addition, more application scenarios, such as music therapy, require imitating more specific musical styles from a few given music examples, rather than capturing the overall genre style of a large data corpus.

Thus, I introduce music domain knowledge into symbolic music composition to address the above requirements that challenge current models. I will develop two approaches in this chapter. The first one [32] is a statistical machine learning model that is able to capture and imitate the structure, melody, chord, and bass style from a given example seed song. The second approach [31] introduces *MusicFrameworks*, a hierarchical music structure representation, and a multi-step generative process to create a full-length melody guided by long-term repetitive structure, chord, melodic contour, and rhythm constraints.

Experiments of both approaches show that, with the help of music domain knowledge, especially music structure representations, I am able to factor symbolic music composition into sub-problems, which allow simpler models, require less data, and achieve high musicality and controllability.

3.1 Related Work

Automation of music composition with computers can be traced back to 1957 [56]. Long before representation learning, musicians looked for models that explain the generative process of music[57]. Early music generation systems often relied on generative rules or constraint satisfaction [23, 24, 25, 57]. Subsequent approaches replaced human learning of rules with machine learning, such as statistical models [113] and connectionist approaches [9]. Now, deep learning has emerged as one of the most powerful tools to encode implicit rules from data [10, 54, 59, 60, 83].

One challenge of music modeling is capturing repetitive patterns and long-term dependencies. There are a few models using rule-based and statistical methods to construct long-term repetitive structure in classical music [22] and pop music [32, 46]. Machine learning models with memory and the ability to associate context have also been popular in this area and include LSTMs and Transformers [59, 63, 103, 123], which operate by generating music one or a few notes at a time, based on information from previously generated notes. These models enable free generation and motif continuation, but it is difficult to control the generated content. StructureNet [93], PopMNet [132], and Rachmaninoff [22] are more closely related to my works in that they introduce explicit models for music structure.

Another thread of work enables a degree of controllability by modeling the distribution of music via an intermediate representation (embedding). One such approach is to use Generative Adversarial Networks (GANs) to model the distribution of music [42, 51, 139]. GANs learn a mapping from a point z sampled from a prior distribution to an instance of generated music x and hence represent the distribution of music with z. Another method is the Autoencoder, consisting of an encoder transforming music x into embedding z and a decoder that reconstructs music x again from embedding z. The most popular models are Variational Auto-Encoders (VAE) and their variants [72, 74, 108, 120, 129, 138]. These models can be controlled by manipulating the embedding, for example, mix-and-matching embeddings of different pitch contours and rhythms [13, 108, 138]. However, a high-dimensional continuous vector has limited interpretability and thus is difficult for a user to control; it is also difficult to model full-length music with a simple fixed-length representation. In contrast, my approach in Section 3.3 uses a hierarchical music representation (i.e., *music framework*) as an "embedding" of music that encodes long-term dependency in a form that is both interpretable and controllable.

For stylistic music composition, rule-based methods and deep-learning systems both have their pros and cons. Models that define style- and theory-related rules as constraints [23, 24, 25, 57] offer fast and controllable generation but may sacrifice creativity and struggle with rule extension for new styles. Meanwhile, the utilization of deep learning, particularly sequence learning models, requires a huge amount of training data and computation power, and so far, has only been used for general music models because it would be difficult to find enough data to model a specific style. Recent advancements involve representation learning through VAEs, offering more opportunities for capturing a specific song style, for example ''style transfer", but in music composition, disentanglement of style and the other musical content, as well as musicality, still remains a problem.

Two systems highly related to my work in Section 3.2 are discussed here. The first one is [46], which created an algorithmic composition system for popular music using probabilistic methods guided by music theory. Users can specify input settings like phrase length, chord transition matrix, and metrical salience histogram. The algorithm outputs a piece of music in MIDI format with melody, chord progression, and simple phrases. They create a structure indicating phrase durations, accents within phrases, and phrase repetitions. Chords are then selected using a Markov chain. Melody generation is based on ten different functions that estimate the conditional probability of the next pitch and rhythm based on the previous note, current chord, and accent locations, which are selected during the design of the overall structure. Humans rated the musical output slightly lower than human compositions, but not significantly. This system generates a generic popular music style, but cannot represent style variations, personalize music, or model bass style. I adapt some of the ideas in Section 3.2, for example, using theory and statistics to help capture and apply abstract qualities of style.

IDyOM ([104]) is another probabilistic system based mainly on variable-order Markov models for adapting to and predicting musical sequences. IDyOM has been used to estimate information content of musical sequences, but it does not seem to contain sufficient constraints for its prediction to be used for music composition.

Evaluation of computer-composed music is often informal and ad-hoc. It is common to use human ratings based on short listening examples, and often comparisons are made to simple baseline models to show that improvement is achieved by new methods. [59] use 180 comparisons to rate three techniques as well as excerpts from a human-composed dataset on the basis of "which one is more musical" and using a Likert scale. [46] compared machine- to human-composed songs along several dimensions using Likert scales. I believe my approach in Section 3.2 is the first study to evaluate the degree to which computers can successfully compose music by imitation.

3.2 Personalised popular music composition using imitation and structure

3.2.1 Overview

Recent practice in stylistic music composition in the symbolic domain mainly focuses on machine learning of general musical rules or style, which tends toward generic musical output with no support for personalization. The number of favorite and familiar songs that a person can provide is insufficient for deep learning approaches, which need a large amount of training data. However, every piece of music has its own distinctive abstract qualities, for example, structure, melodic contour, rhythmic pattern, chord progression, bass line pattern, etc. These abstract qualities might vary a lot from piece to piece, even within a music genre and in works by the same composer. By focusing on distinctive musical qualities within the constraints of general rules of music, I hope to create more enjoyable music and come closer to capturing elements of "style" or character within a limited set of examples.

To avoid the problem of "smoothing over" distinctive attributes, and to further my goal of imitation, I need to "learn" from a single song called the *seed song*. However, I cannot rely *completely* on the seed song because that would result in too much similarity. An overly similar song sounds as if the original were being played with mistakes, or else the original was simply plagiarized. This is perhaps related to the "uncanny valley" phenomenon [94]. In addition, the inverted-U model of preference for music [7, 14, 116] suggests that the pleasantness (hedonic value) of a music piece increases as its novelty increases (familiarity diminishes), and will decrease once the novelty reaches a certain level (Figure 4). Therefore, it is important to control how similar the new song should be in order to achieve both stylistic similarity and creative differences. This is a new challenge for the field of automatic music composition.

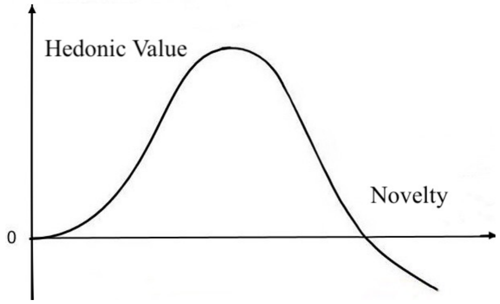


Figure 4: Inverted-U relationship: the Wundt Curve originally suggested by Wundt and later adapted by [7], and the linking of favorability to familiarity/time curve by [116].

Another significant issue in music generation is music structure, particularly at the level of phrases and sections. Longer-term structure was at the heart of many early works on music generation, but largely ignored in more recent generation systems based on neural networks and sequence learning. The resulting failure of many systems to exhibit interesting long-term behavior has now been widely recognized and is receiving renewed attention. In this work, I consider longer-term and higher-level music structure formed through section repetitions. For example, a common pattern in popular music songs is `ABABB`, where letters stand for sections and repeated letters represent an approximate repetition of a section. In addition to repetition structure, there can be a common rhythm in A and B, a repetition in melody contour from one phrase to the next, etc. Here, I consider both higher-level repetition structure and lower-level repetition music pattern structure in my representation and generation process.

I will introduce a stylistic music generation model that is able to capture melody, chord, and bass style from a single pop song and imitate them with structure information in a new complete piece. Like "music structure," *music style* is a general term that can refer to almost any aspect of music. Definitions are further complicated by the multi-level, multi-modal character of music representation—music can be notated and read, performed, and listened to, and each of these modalities has aspects of style [28, 36]. Style refers to general characteristics, but these can be seen at the level of music genre, sub-genre, composers, and individual compositions.

In this work, *music style* refers to information at the symbolic music notation level, including rhythm, pitch, and dynamics, and my approach focuses on imitating single songs. Often, style at the level of a single song is *not* representative of more general styles, and some might not even use "style" to refer to characteristics of a single song. Furthermore, there is no clear distinction between style at the levels of individual song and the more general genre level. For example, by imitating a Chinese pop song, we are at least likely to compose a mostly-pentatonic melody, and the result can be recognizable as Chinese popular music. Thus, song imitation is likely to discover and use *some* more general elements of musical style. I also note that imitation of performance, orchestration, and production (especially in pop music) are important aspects of style and perceived similarity, but I leave these to future work.

My work will offer three main contributions. First, I will offer methods that generate likable music overall. If music is not likable, the fact that it is a successful imitation that listeners find *similar* to something likable is a small consolation. Second, I will produce music that listeners recognize as *similar* to the seed songs. Thus, I will be able to learn enough from a single input song to form an imitation, even when seeds vary from Chinese Pop to Western Pop songs. Finally, one must ask if this whole enterprise of imitation to create likable songs is valid to begin with. By making imitations, do I improve the preference for my machine-created songs? I will show through correlation that an increased preference for the seed song predicts an increased preference for my generated imitation. Thus, my original idea to create personalized music by imitation shows promise.

This approach should be able to use imitation without large training sets to enhance listener preferences for generated music. In addition, it generates complete customized songs of any length and contains a logical, hierarchical music structure, as opposed to generating a few bars of music or long rambling sequences lacking in longer-term structure. We will also see how individual models for stylistic melody, chord, and bass generation can be combined to create hybrid styles, e.g., creating a song with melody style from song A, chord style from song B, and bass style from song C.

For training and testing data, this work uses 13 MIDI songs labeled with an accurate analysis of structure, melody, chord, and bass. While there is certainly some degree of subjectivity in labeling ([95]), my representation is more that of a simplified lead sheet than a detailed analysis, so label ambiguity is unlikely to have a significant impact. I will introduce the detailed collection, representation, and pre-processing steps of the data in the thesis.

I will briefly introduce the proposed method and experiment design in the following sections. In the thesis, I will describe this system in detail, including the data preparation process, system design details, as well as both objective and subjective experiments to evaluate the model output. I believe there are many ideas that can inform future systems, serve as a baseline for comparison, and offer insights into music perception and cognition.

3.2.2 Proposed Method

I use statistical machine learning methods to imitate the styles of melody, chord progression, bass and structure from an input seed song. The system framework is shown in Figure 5. The system takes one seed song as input. After some data pre-processing steps, I feed each part from the seed into a corresponding generation module.[1] I also feed the results from structure and chord progression modules into the other modules as inputs. The tempo is set to the same as the tempo of the seed song input. Finally, I combine the newly generated stylistic melody, chord accompaniment, and bass MIDI tracks and synthesize them to obtain audio output. Each module is introduced accordingly as follows.

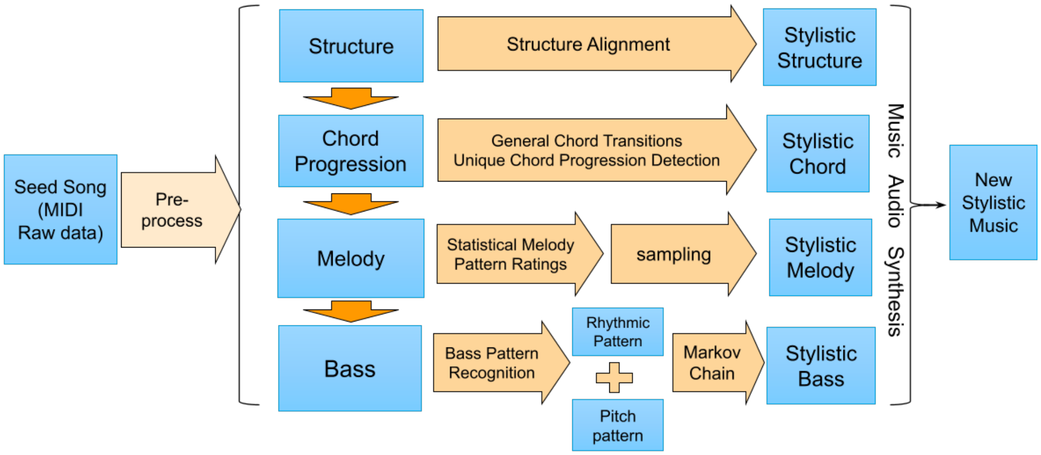


Figure 5: Framework of stylistic music generation system.

Structure Alignment and Generation I will introduce three ways to generate new structures: (1) copy the seed song structure; (2) generate according to a specification string such as “AABABC” from the user and treat each section as 8 bars, and here “A”“B” indicates different sections; (3) generate random structures, which includes selecting from a collection of typical structures. This gives flexibility in generation. Since the new structure can be different from the structure of the seed song, I need to align each new section to an appropriate seed song section. For example, in Figure 6, I want to imitate the seed song shown at the top using the new structure shown below it. A good alignment will imitate the seed’s intro style in the new intro, imitate seed’s first chorus in the new chorus, etc.

Stylistic Chord Generation To generate convincing chord progressions while imitating the harmonic style of the seed song, we combine statistical features from general popular music data set[2], seed song statistics, and distinctive chord progressions from the seed song. Within each

[1] Seed songs for different modules can be different, e.g., taking the melody from seed A and bass from seed B.
[2] https://github.com/tmc323/Chord-Annotations

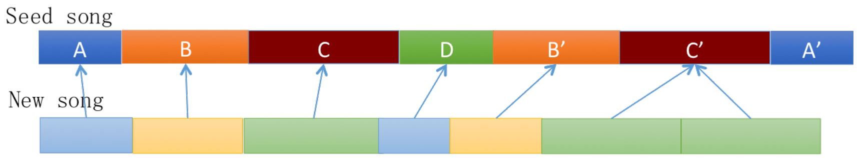


Figure 6: An example of ideal structure alignment.

section, we generate chord progressions in time order, selecting one or more successive chords at each step. Section endings are treated specially so that we can impose a stylistically consistent harmonic resolution to the section.

Stylistic Melody Generation From music theory, rhythm and pitch interact to form melody. Thus, I design a number of melody style rating functions to evaluate the suitability of the next note given previous notes in the context of a chord progression. I generate melody in time order, note-by-note, using rating functions to estimate the probability of each possible next note and make a weighted random choice. Thus, I use a statistical sampling method. To avoid outliers, I generate 30 candidate melody sequences and pick the one with the best rating as the stylistic melody output.

More formally, I have $n = 15$ rating functions $P_i(x, y)$ where x is the next note (pitch and duration pair), and y is the context including the note position in the current bar and phrase, the previous notes, the chord progression, and the seed song. I treat these functions as independent probabilities, multiplying them together to form a weight for each note:

$$w_x = \prod_{i=1}^{n} P_i(x, y) \tag{1}$$

The stylistic rating functions used to extract and represent melody style features are inspired by [46], who use similar methods to generate popular music. I extend these functions and use seed song statistics to encode specific melody style qualities from the input seed songs, and I develop additional rating functions that emphasize long-term structure.

Stylistic Bass Generation I represent bass style using patterns and generate the stylistic bass section-by-section. Each bar of bass is represented by a rhythm pattern that denotes onset times. For each section, the output bass rhythm copies the rhythms of the first and last bars of the section in the seed song, and every other bar uses the most frequent bass rhythm pattern. Selection of the bass note pitch follows consideration of both chord tone and rhythmic positions.

3.2.3 Experiment Setting

I conduct both objective and subjective evaluations of this stylistic music generation system. In the 13 collected MIDI songs, three of them are used in the training stage for parameter tuning. The other ten songs, five Western pop songs and five Chinese pop songs, are used as seed songs for evaluation. One application of my system is to imitate favorite pop songs, so I want to imitate songs that are already popular and familiar. Hopefully, results will show that my stylistic music generation system is able to create music with both high musicality and similarity to the seed song. I will also further explore the factor of familiarity in symbolic music composition.

For the objective evaluation, I first use a paired t-test to show that my stylistic imitations have a higher similarity to seed songs than non-imitations. Then, I use another paired t-test to show that there is no significant difference in estimated probabilities (objective rating functions) between the original seed songs and the generated imitations. If all the above assumptions hold, then I can claim that my approach is able to produce music that is similar to human-composed music in terms of my objective rating functions.

I also conduct human listening evaluations for the melody, chord, and bass modules in my system and for the combination of all three. For each module, listeners are presented with ten pairs of 30-second music fragments in a random order. Each pair consists of a seed song and either a stylistic imitation or a non-imitation of another seed song. Participants rate each pair for similarity and personal preference using a Likert scale from 1 to 5. To ensure unbiased assessments, especially in the melody and bass studies, I control for non-melodic elements by deriving them from a third song, which is designed to isolate the evaluated component (e.g., melody) from influencing factors like chord progression. In the chord study, only simple block chords were presented. Audio versions were produced using MIDI and software synthesizers, with parameters adjusted to minimize effects of timbre, pitch, and tempo differences, maintaining identical settings for each pair and aligning them closely with the original seed song settings.

3.3 Controllable deep melody generation via hierarchical music structure representation

Generating a customizable full piece of music with consistent long-term structure remains a challenge in deep learning. Thus, I aim to explore automatic melody composition with multiple levels of structure awareness and controllability in deep learning. My focus is on (1) addressing structural consistency inside a phrase and on the global scale, and (2) giving explicit control to users to manipulate melody contour and rhythm structure directly.

My solution, *MusicFrameworks*, is based on the design of hierarchical music representations I call *music frameworks* inspired by Hiller and Ames [57]. A music framework is an abstract hierarchical description of a song, including high-level music structure such as repeated sections and phrases, and lower-level representations such as rhythm structure and melodic contour. The idea is to represent a piece of music by music frameworks, and then learn to generate melodies from music frameworks. Controllability is achieved by editing the music frameworks at any level (song, section, and phrase); I also present methods that generate these representations from scratch. *MusicFrameworks* can create long-term music structures, including repetition, by factoring music generation into sub-problems, allowing simpler models and requiring less data.

In the generation process, I first organize the full melody with section and phrase-level structure. To generate melody in each phrase, I generate rhythm and basic melody using two separate transformer-based networks, and then generate the melody conditioned on the basic melody, rhythm, and chords in an auto-regressive manner. To customize or add variety, one can alter chords, basic melody, and rhythm structure in the music frameworks, letting my networks generate the melody accordingly. Additionally, I introduce new features to encode musical positional information, rhythm patterns, and melodic contours based on musical domain knowledge.

Evaluations of the *MusicFrameworks* approach include objective measures to show expected behavior and subjective assessments. I compare human-composed melodies and melodies generated under various conditions to study the effectiveness of music frameworks. I summarize the contributions as follows: (1) devising a hierarchical music structure representation and approach called *MusicFrameworks* capable of capturing repetitive structure at multiple levels, (2) enabling controllability at multiple levels of abstraction through music frameworks, (3) a set of methods that analyze a song to derive music frameworks that can be used in music imitation and subsequent deep learning processes, (4) a set of neural networks that generate a song using the *MusicFrameworks* approach, (5) useful musical features and encodings to introduce musical inductive biases into deep learning, (6) comparison of different deep learning architectures for relatively small amounts of training data and a sizable listening test evaluating the musicality of my method against human-composed music.

3.3.1 Experiment Design

Objective Evaluation I first examine whether music frameworks promote controllability. I aim to show that given a basic melody and rhythm form as guidance, the model can generate a new melody that follows the contour of the basic melody, and has a similar rhythm form. Also, I follow methods in Section 2.3 to test if the generated melody exhibits similar structure-related distributions to that of the POP909 dataset.

Subjective Evaluation I conduct a listening test to evaluate the generated songs. To avoid listening fatigue, I present sections lasting about 1 minute and containing at least 3 phrases. I randomly select six sections from different songs in the validation set as seeds and then generated melodies based on conditions 1–6 presented in Table 1. To render audio, each melody is mixed with the original chords played as simple block triads via a piano synthesizer. For each section and each condition, I generate at least two versions.

In each rating session, a listener first enters information about their music background and then provides ratings for six pairs of songs. Each pair is generated from the same input seed song using different generation conditions (see Table 1). For each pair, the listener answers: (1) whether they heard the songs before the survey (yes or no); (2) how much they like the melody of the two songs (integer from 1 to 5); and (3) how similar are the two songs' melodies (integer from 1 to 5). I also embed one validation test in which a human-composed song and a randomized song are provided to help filter out careless ratings.

4 Expressive Performance Control with Style and Structure

Music fundamentally relies on performance. Performance is where musicians interpret scores with their personal styles and emotions. In AI-generated music, expressive performance control is often ignored. It encompasses critical music elements often missing in scores and symbolic music composition, such as timing, dynamics, pitch, playing techniques, and timbre control. These performance controls are key to making generated music sound natural. As my advisor Roger

	R.Melody	Basic Melody	Rhythm
0	copy	copy	copy
1	gen	copy	copy
2	gen	gen	copy
3	gen	without	copy
4	gen	copy	gen with BRF
5	gen	copy	gen without BRF
6	gen	gen	gen with BRF

Table 1: Seven evaluation conditions. Group 0 is human-composed. R.Melody: realized melody; gen: generated from our system; BRF: Basic Rhythm Form; copy: directly copying that part from the human-composed song; without: not using music frameworks.

Dannenberg often emphasizes, "No good control, no good synthesis." This is comparable to a masterful violin that, regardless of its high quality, will sound vastly different in the hands of a professional compared to a novice, underscoring the importance of skilled performance control.

To generate good expressive performance control, I focus on three performance elements: timing, pitch, and dynamics. My experiments involve both vocal and instrumental performances. In addition, I aim to invent models that can produce performance controls in diverse styles. I will further look into the impact of music structure on music performance.

4.1 Related Work

Expressive performance controls can be categorized into timing, pitch, dynamics, and timbre control [36, 81]. Timing, crucial for mood and style conveyance, often involves rhythm and tempo variations. Performance timings are different from score timings with regular note durations in beats. Studies [18] highlight how performance timing affects musical expression. Most studies on the generation of expressive performance timing [115, 134] focus on piano, as MIDI format is easy for representing and modeling timing control. Recent practices in piano performance timing control [134] have shown promise using traditional machine learning methods. These methods model deviations between note onsets in performance timing and the original score timing. In my thesis, I adopt a similar approach for modeling timing onset deviations but leverage deep learning architectures to incorporate more style controls.

Pitch is another fundamental parameter in expressive performance. While instruments like the piano produce discrete pitches, voices and many instruments such as strings and woodwinds have continuous pitch variation during performance. This is typically analyzed using Fundamental Frequency (F0), representing continuous F0 pitch curves, which are closely tied to playing techniques like vibrato, glissando, and ornaments. Prior research has explored modeling F0 curves to enhance expressiveness. For instance, in Ning Hu's thesis [58], she used traditional machine learning models to generate expressive F0 curves, producing natural vibrato controls for brass instruments. Furthermore, [68] uses a neural network with bi-directional LSTM to generate F0 curves from scores for instruments. These studies collectively underline the critical role of F0 curves in enhancing the expressiveness and emotional depth of musical performances, whether live or synthesized. In my work, I use deep learning models to generate F0 curves from scores for both singing and instrumental synthesis with style and structure controls.

Dynamics, which involve the loudness and softness of notes, also play a significant role in expressive performance. For example, Juslin and Timmers [70] emphasize how dynamic variations can dramatically alter the listener's perception of music. In expressive performance control, researchers often use amplitude envelopes (curves) extracted from audio performance to represent dynamics control. Clynes [19, 20] suggested that amplitude envelopes should be controlled based on music context. Dannenberg and Derenyi [38] designed *Spectral Interpolation Synthesis* where amplitude envelopes were generated and modified using machine learning techniques to create realistic performance in trumpet. Ning [58] further extended that work and experimented on a performance model with more natural amplitude control. More recently, the integration of deep learning algorithms, as seen in the work of [133], has enabled dynamic control, mimicking the subtle amplitude fluctuations found in live performances. In this thesis, I will further explore generating amplitude envelopes from scores with deep-learning models.

In addition, timbre control, often linked with performance techniques, is crucial for musical expression. However, the lack of performance technique labels in current datasets hinders models from generating timbre control directly from scores. In my model for singing performance, a subset of the opera singing data includes technique labels, which I integrate into the system to model timbre control based on these techniques. For other singing styles and instrumental performances, I leave the timbre control to the model and let it figure out suitable techniques automatically.

Style modeling in music performance seeks to capture and replicate the unique expressive characteristics that define different musical genres and individual musicians. Widmer and Goebl [50] focused on understanding and modeling the distinctive styles of famous pianists with machine learning techniques to analyze. Another study by Dixon, Goebl, and Widmer [41] explored the automatic detection of performance style in recordings, showcasing the potential of AI in distinguishing and learning diverse musical expressions. More recent advances involve deep learning models, as seen in the work of Oore et al. [92, 99], which generated stylistically diverse music performances using neural networks. I incorporate different music styles in the singing performance control, such as pop, opera, rock, children's songs, and traditional Chinese music.

Like symbolic music composition, structure significantly impacts musical expression. For example, musicians often sing identical phrases with the same lyrics and notes differently depending on their placement within a song. However, few studies have looked at the relationship between music structure and expressive performance control. Time permitting, I aim to further investigate the analysis and generation of expressive performance control in relation to structure, though this is not a requisite component of the thesis.

4.2 Data collection

For singing performance control and synthesis, there has been a persistent lack of publicly accessible data, particularly concerning the diversity of languages and performance styles. Thus, I collected SingStyle111 [33], a large studio-quality singing dataset with multiple languages and different singing styles, and presented singing style transfer examples. It features 111 songs performed by eight professional singers, spanning 12.8 hours and covering English, Chinese, and

Italian. It incorporates different singing styles, such as bel canto opera, Chinese folk singing, pop, jazz, and children. Specifically, 80 songs include at least two distinct singing styles performed by the same singer. All recordings were conducted in professional studios, yielding clean, dry vocal tracks in mono format with a 44.1 kHz sample rate. I segmented the singing voices into phrases, providing lyrics, performance MIDI, and scores with phoneme-level alignment. I also extracted acoustic features such as Mel-Spectrogram, F0 curves, and amplitude envelopes.

Together with SingStyle111, I also used other public datasets for singing research, including Opencpop [127], M4Singer [141], Children Song Dataset [16], VocalSet [48, 131], and PopCS [85]. I processed them following the same methods for SingStyle111. Detailed representation and processing approaches will be discussed in the thesis.

For instrumental performances modeling and synthesis, on the one hand, I organized and re-processed existing public clean monophonic instrumental dataset, including Bach10 [43], URMP [82], CBF dataset [125], CCOM-HuQin [143], Filosax [3], EEP, and ViolinEtudes [11]. On the other hand, my advisor Prof. Roger Dannenberg performed and recorded stylistic Trumpet performances. The combination spans a wide range of instruments, covering strings, woodwinds, and brass instruments. Features such as mel-spectrogram, F0 curves, and amplitude envelopes are extracted and aligned with scores and audio recordings.

4.3 Singing Performance Control with Style and Structure

I use diffusion models to generate expressive timing, F0 curves, and amplitude envelopes from the symbolic score and lyrics input. Input context includes the symbolic score (note pitch, onset, and duration), lyrics (language category, phone ID, word boundary indication), positional features (order of notes), and style tokens (singer identity, style genre, technique, emotion, dataset). The pipeline first generates expressive timing from the input context and then passes the generated timing as an additional input condition to the generation models of F0 curves and amplitude envelopes.

Model Architecture Each generation module shares a similar model architecture (as shown in Figure 7). It is inspired by DiffWave [76] and uses a diffusion process with a modified WaveNet model [98] as the backbone. All the residual layers share the same projection of the input context but with an extra non-shared convolutional layer. All the generating targets have been normalized to be closer to the Gaussian distribution before being input into the model.

Experiment Design I use both objective and subjective methods to evaluate the models. Since currently there is no explicit performance control modeling for singing voice yet, I do not have a baseline and thus directly compare the generated results with the Ground Truth. For objective evaluation, I compute the MSE loss. For subjective evaluation, I use the synthesizers in the next section to generate a singing according to the generated performance control. I also plan to do ablation studies to see the effects of different style token controls.

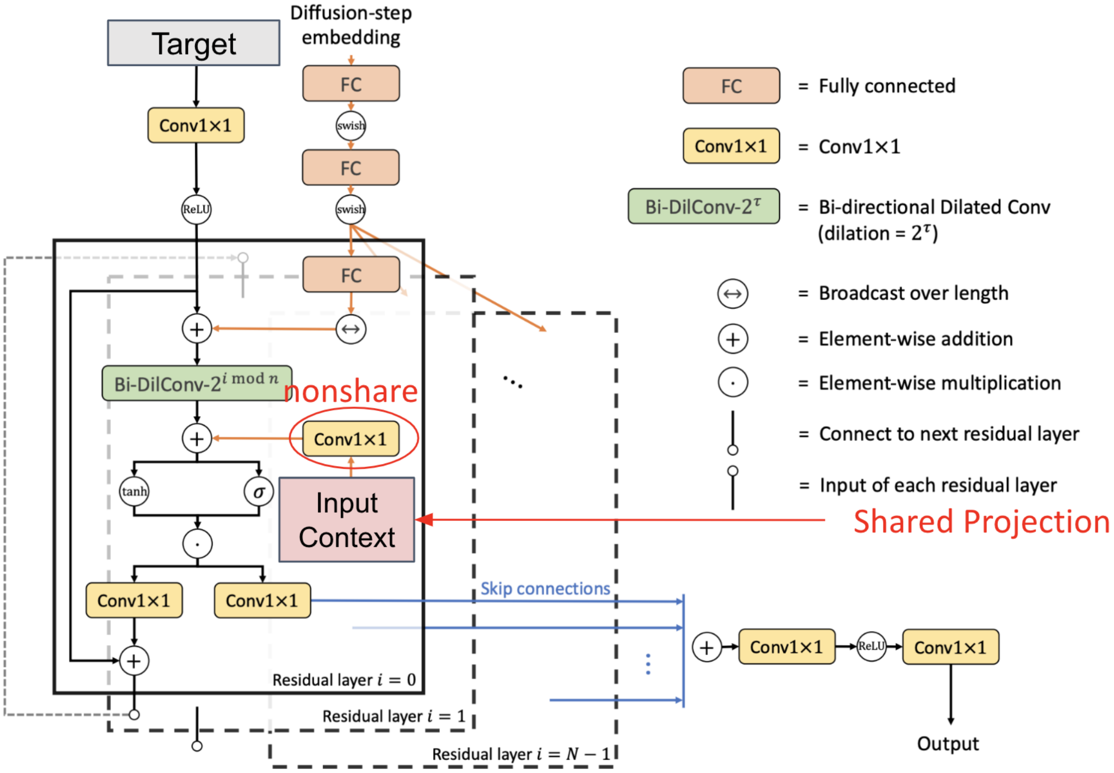


Figure 7: Architecture of the proposed methods. "Target" varies according to the generating content. For example, in expressive timing control, the target is a sequence of onset deviation; for generating F0 curves, the target is a framewise F0 sequence.

4.4 Instrumental Performance Control with Style and Structure

Instrumental performance control will be focused on strings, brass, and woodwinds because piano and other percussion instruments have much less freedom in pitch control. The plan is to apply similar approaches used in singing to these instruments, except that they do not need to encode lyrics. In addition, I also do some experiments using a Transformer and VAE architecture instead of a diffusion process for F0 and Amplitude control.

5 Singing and Instrumental Synthesis with Style Control

Music synthesis, including singing voice and instrumental synthesis, is a long-standing area of research that has lasted for almost a century. Recently, advanced machine learning and deep learning technologies [8, 77, 98, 140] have accelerated the development in this area and largely improved audio synthesis quality. However, several challenges still remain difficult for music: (1) lack of high-quality data with annotations, especially for public and large datasets; (2) high fidelity synthesized audio without artifacts; (3) expressive and natural sound with expressive performance control; (4) realize timbre style control and transfer even with unseen targets.

To address the above issues, I extend my work on performance control (Section 4) to audio synthesis. For data scarcity, I collected SingStyle111 and re-processed available singing and instrumental datasets (as discussed in Section 4.2). In the synthesis process, I design an acoustic model that enables explicit style and performance controls to generate high-quality mel-spectrograms for singing. Meanwhile, I modify the state-of-the-art speech synthesis vocoders to generate high-quality singing audio from mel-spectrograms. To further explore synthesizing with unseen voices and timbre style controls, I integrate audio codecs [78], a highly efficient audio representation, into the synthesis architecture. It can take only 5 seconds of unseen speech voice and synthesize realistic singing using this voice together with score and lyrics input. Similar approaches will be applied to instrumental synthesis as well.

5.1 Related Work

5.1.1 Singing Voice Synthesis (SVS) and Singing Voice Conversion (SVC)

Voice synthesis can be traced back to the 1930s in Bell Labs [44], and researchers have developed successful traditional vocal synthesizers such as VOSIM and FOF that are widely used in the industry. In recent years, the development of machine learning and deep learning techniques in audio and Text-To-Speech (TTS) synthesis, such as Wavenet [98], deep acoustic models (FastSpeech [107]), neural vocoders (MelGAN [77], HifiGAN [75], BigVGAN [79], DiffWave [76]) and audio Encodec [78, 140], has also become mainstream in singing voice synthesis and conversion.

A widely used architecture in current TTS and SVS practices is a two-step synthesis process. As shown in Figure 8, an SVS system typically handles two types of inputs: (1) score with lyrics, or (2) performance data including F0 curves and amplitude envelopes. Score-based SVS systems process symbolic inputs like scores and lyrics, generating performance controls implicitly within the model. Meanwhile, some SVS systems [85] take lyrics and ground-truth performance controls, such as the actual F0 curves, as input. SVS systems generally comprise two key components: an acoustic model that transforms input into an acoustic representation from input, and a vocoder that synthesizes the final audio output from this representation. The acoustic representation could be standard formats like spectrograms or mel-spectrograms, or more specialized pre-trained representations like audio Encodec codes [65] and harmonic representations in DDSP [142], among other predefined features and templates.

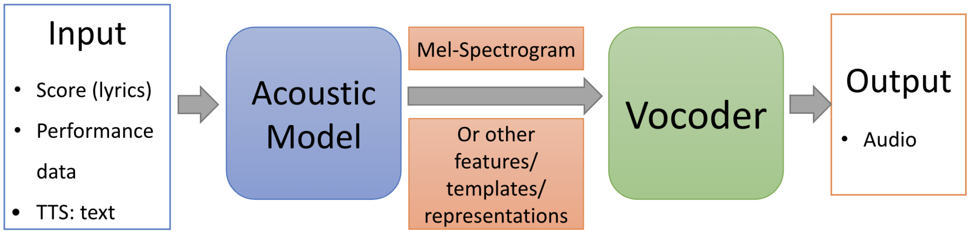


Figure 8: A commonly used architecture in SVS.

There are several common practices for acoustic model development. First is transformer-based models such as FastSpeech2 [107]. It enhances the transformer architecture with three implicit variance adaptors, controlling pitch, energy, and duration within the model. Additionally, WaveNet and FFT (Fast Fourier Transform)-based methods are widely utilized. These models are often effectively paired with GANs [144] or diffusion processes [85] to enhance their performance.

Recent advancements in vocoders, driven by deep learning, have marked significant progress beyond traditional methods like Griffin-Lim [105]. Modern vocoders such as Parallel-WaveGAN [136] and HiFiGAN [75] utilize GAN frameworks and employ multi-period and multi-scale loss functions, operating across both time and frequency dimensions to produce high-quality audio. BigVGAN [79] extends HifiGAN's capabilities, integrating periodic activation functions and anti-aliased multi-periodicity composition, yielding high-quality speech and music synthesis but requiring extensive training data. Diffwave [76], leveraging a Denoising Diffusion Probabilistic Model (DDPM) [117] with a WaveNet backbone, offers ease of training but is slower in both training and inference phases and can lose high-frequency components in the output. RefineGAN [135] creates a waveform template from F0 curves for phase information, subsequently refining it with Mel-spectrograms to produce the final audio waveform. To improve pitch sensitivity in singing vocoders, some studies [144] have incorporated quantized F0 curves as additional conditions. In this section, Diffwave and BigVGAN are utilized as vocoders, processing inputs of mel-spectrograms and quantized F0 curves.

Singing Voice Conversion, recent advancements have significantly enhanced the ability to alter a singer's voice while retaining the song's integrity. One example is the SO-VITS-SVC model (SoftVC VITS Singing Voice Conversion) [3], which integrates a speech encoder SoftVC [122] into a speech synthesizer VITS [73], and achieves realistic results for pre-trained singers. In the general voice conversion, the latest models like VALL-E [126] and NANCY [15] are able to transfer unseen voice targets in speech. However, these models are still insufficient for zero-shot singing voice synthesis or transfer, mainly due to data scarcity and the more complex nature of singing compared to speech. The zero-shot singing synthesis work in this section is largely inspired by previous works in speech voice conversion and synthesis, including speech content encoders [106], speaker encoders [124], and the Descript audio codec [78].

For stylistic singing voice synthesis, most SVS systems do not have explicit control of singing genre styles or techniques. Instead, [62, 144] some of them have control over choosing different singers and their corresponding singing timbre. In this section, I decided to explicitly integrate style genre control into the model, as well as singing techniques control in the opera genre. These controls are first used in performance models in the previous section and serve as conditions in the acoustic model in this section.

5.1.2 Instrumental Synthesis

Instrumental synthesis has evolved through a series of significant methodologies over the past century. For instance, Spectrum Modeling Synthesis (SMS) [114] is an influential model in this field by tracking harmonic series with peaks in the sound and assuming the leftover is noise. By decomposing harmonics and noises, SMS allows for detailed reconstruction and manipulation of instrumental sounds, including non-harmonic instruments. Later works further extended this idea and even applied advanced deep learning techniques, such as DDSP [47].

[3] https://github.com/svc-develop-team/so-vits-svc

Another promising approach is Spectrum Interpolation Synthesis (SIS) [38], which offers a comprehensive framework integrating performance control with audio synthesis. It assumes that every note is nearly harmonic, and uses performance controls (F0 curves and Amplitude envelopes) to determine the harmonic series, and then utilizes wavetables to resynthesize. The spectrum itself is stored as a two-dimensional array of the relative amplitudes of the harmonics, and a continuous spectrum is calculated by interpolating among four neighboring spectra. This approach was further extended in [58] and achieved realistic brass instrumental synthesis results. However, in-harmonic components of the sound, such as noisy attacks, could not be directly inferred from the model but spliced using additive synthesis.

The emergence of deep learning techniques has brought new possibilities into this area. MIDI-DDSP [133] integrates discrete controls in MIDI (such as timing, pitch, and velocity) with the DDSP framework to synthesize instrument performances. The Control-Synthesis Approach [68] uses a bi-directional LSTM model to generate pitch and loudness curves for performance control and input them into a synthesizer with a noise model from the DDSP library. These works show capacities with deep learning techniques but do not achieve high-quality synthesized results. Recently, foundation models using huge model sizes and training data have become a new direction. Large-scale models like MusicGen [26], SoundStorm [8], and SoundStream [140] have broadened the scope of possibilities in audio and music synthesis. However, high-quality instrumental synthesis with detailed performance control still remains challenging and will be the focus of this section.

5.2 Expressive Singing Synthesis with Style Control

Generative models have been the focus of a plethora of recent research in deep learning with applications such as image generation [71], text-to-image, 3D assets, or video generation [84]. The disruption brought about by diffusion models has also affected recent advances in audio-related generative models for speech and music. Diffusion models have the benefit of being highly expressive models with the drawback of slow sampling speed. Therefore, they are applicable in creative use cases where real-time generation is not a requirement, which makes them the ideal choice for the task of singing voice synthesis (SVS).

Unlike Text-To-Speech (TTS) Synthesis, Singing Voice Synthesis (SVS) from scores and lyrics encounters distinct challenges. Firstly, singing encompasses a broader pitch frequency range than speech, and current models still struggle to generate pitch-sensitive results, especially for high-frequency components. Secondly, singing exhibits diverse timbre textures, such as breathiness, chest voice, and head voice, which are typically not a focus or even excluded for most TTS systems in speech synthesis. Moreover, singing demands a higher level of mastery in various techniques and performance control compared to speech. Another significant challenge is data scarcity; unlike speech, which benefits from abundant, well-processed data for training and testing, singing data is relatively limited. Therefore, in contrast to many previous SVS systems that adapt methods from TTS, my research concentrates on addressing the unique characteristics and challenges inherent in singing.

Singing performance is highly multimodal. Given the same lyrics and musical score, there are multiple ways a performer can sing the input score. In musical performance, a performer can creatively decide how to interpret the musical score and render the performance. I handle this multimodality by utilizing fine-grained or dense conditioning of the generative models, such as providing explicit stylistic performance controls such as F0 curves and amplitude envelopes into the synthesis process.

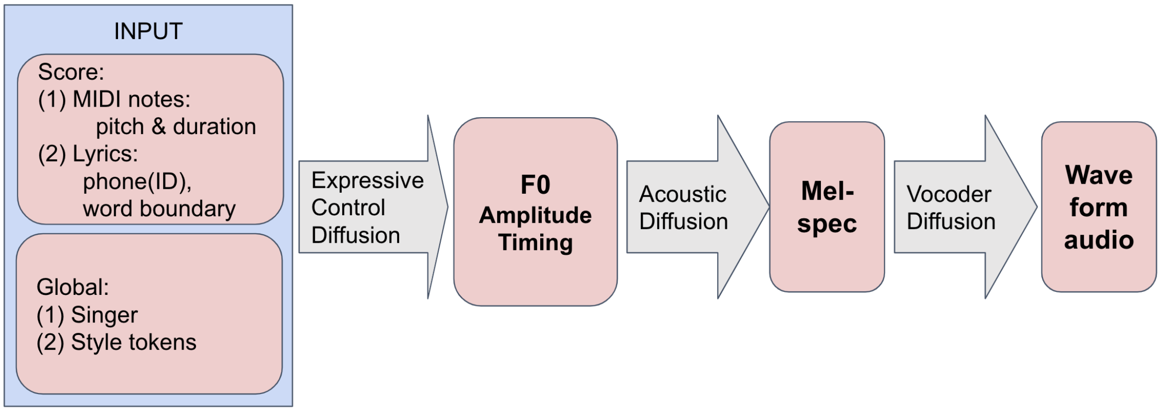


Figure 9: Pipeline for my SVS system.

This work takes score, lyrics, style tokens, and singer ID as input and generates expressive and realistic singing. It involves a cascade of diffusion models following architecture in Figure 7. As shown in Figure 9, the pipeline involves (1) performance control models, including timing, F0 curves, and loudness curves (discussed in Section 4.3); (2) an acoustic model that generates the mel-spectrograms conditioning on performance control signals; (3) a DiffWave vocoder to generate the waveform from mel-spectrograms and F0 curves. This section will focus on discussing the acoustic model and vocoder.

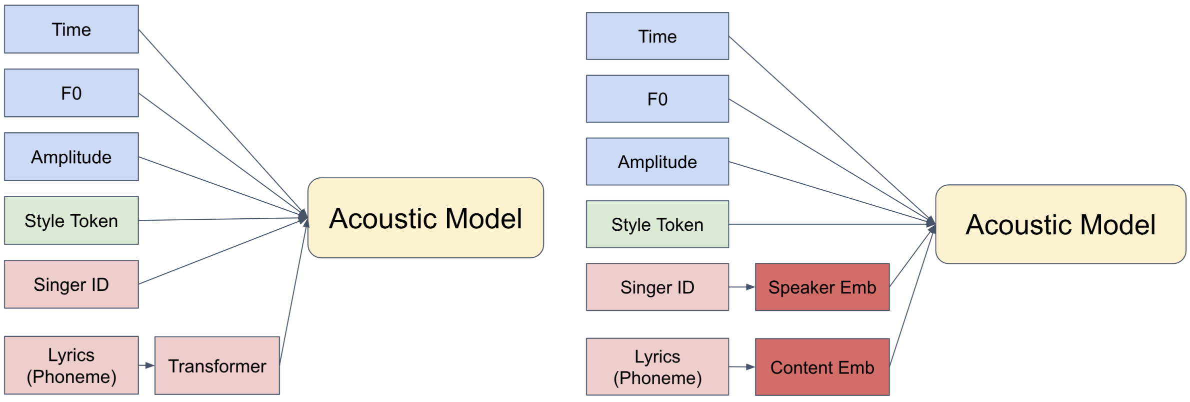
I collected a new singing dataset for this work as discussed in Section 4.2 and Section 5.1.1, and used them in my SVS model. This also involves various processing steps applied to the provided scores for the other datasets to achieve a unified data format. I also utilize a shared phoneme set based on IPA and pinyin to achieve multilingual SVS. To the best of our knowledge, my work is the first system for multilingual SVS. The system can generate Chinese, Korean, English, and Italian singing.

The input of the SVS system is the same as described in Section 4.3. Apart from performance control signals, the acoustic model (Figure 10a) also takes style tokens, singer ID, and lyrics as input conditions in the diffusion process. The style tokens include style genre ID, dataset ID, emotion ID, and singing technique ID. The lyrics processed as phoneme ID are projected with a transfermer encoder. As discussed in Section 5.1.1, I add F0 curves to the DiffWave vocoder as additional input conditions to improve the vocoder synthesis quality.

The experiment uses DiffSinger [85] as the baseline model to compare the generated singing with the proposed methods. I plan to conduct both objective and subjective evaluations. Ablation studies are also conducted to test the effectiveness of style control and performance control signals.

5.3 Zero-shot Singing Synthesis with Unseen Speech Target

This work takes 5-second speech audio of the target (an unseen target voice excluded in training data), together with score, lyrics, and style as input, output a realistic singing using the target



(a) Inputs for the acoustic model in SVS (b) Inputs for the acoustic model in zero-shot SVS

Figure 10: Inputs for the acoustic model in the SVS and zero-shot SVS systems.

voice timbre. This work is much harder than previous SVS systems since (1) the speech and singing voices of the same person might already be very different; (2) the pitch range of input score might vary too much from the target speech; (3) modeling voices in such a zero-shot setting itself is very challenging.

I employ two distinct approaches to address this problem. First, I replace the singer ID in the SVS system with a generalized speaker embedding derived from audio. Concurrently, the Lyrics transformer embedding is substituted with a pre-trained content embedding. This modification enhances the model's capacity to better disentangle aspects such as style, lyrics content, and singer voice timbre, thereby enabling zero-shot control over voice timbre. For the speaker embedding, I use Resemblyzer [124] trained on speech data, coupled with a newly developed pre-trained model for converting lyrics into content embeddings. The effectiveness of this approach has been somewhat limited, potentially due to a paucity of singers in the dataset. To improve results, I am considering the integration of speech data for mixed training.

The second strategy is inspired by the methodologies used in VALL-E [126], transitioning the acoustic model's output from a mel-spectrogram to an audio codec. Additionally, the content embedding is adapted to a codec representation that excludes singer information. I fine-tuned an existing codec model (not published yet within Adobe) on singing data to acquire these content embeddings. This method has shown promising results in zero-shot voice timbre control, but aspects like singing quality and pronunciation accuracy across different languages require further refinement.

5.4 Expressive Instrument Synthesis

In instrumental synthesis, I intend to utilize a similar approach in Section 5.1, where performance control is integrated into the synthesizer, coupled with the implementation of a diffusion process within the acoustic model. For the vocoder, a BigVGAN model will be employed. Initial results have demonstrated promising outcomes for the acoustic model and vocoder across various instruments. However, high-frequency components in the generated audio still remain challenging, warranting further investigation. Furthermore, I plan to apply the techniques from Section 5.2 to instruments, exploring the feasibility of transferring timbres between different instruments in a zero-shot manner. The experimental framework will incorporate both objective and subjective evaluation methods, and the findings will be benchmarked against results from SIS and MIDI-DDSP systems.

6 Applications

I will discuss potential applications of the technologies in the thesis, encompassing areas such as integration with other areas in generative AI, music therapy, music education, music perception, human-computer interactive performance, and theory development of Non-Western music.

6.1 Generative AI and Combination With Other Generative Areas

The technologies in this thesis on symbolic music composition offer tools for generative music, serving as assistants in composition. They enable whole-song writing with a more organized structure, as well as various control options and imitations for different music elements, which might facilitate interactive AI-human composition and foster personalized styles.

Expressive singing voice synthesis models in Sections 4 and 5 can generate high-quality artificial singers with capabilities in various singing styles and languages, and can be employed to create singing demos for composers within Digital Audio Workstation (DAW) software. Zero-shot singing synthesis, which operates on short and unseen speech inputs, enables the creation of singing voices for individuals with limited singing abilities, democratizing singing for all. Additionally, these models hold significant potential in interactive and social media platforms, where creators can use their own voices to generate singing tracks for their videos. It may further be combined with vision technologies such as HeyGAN [4] to generate singing facial expressions.

Instrumental performance control and synthesis technologies hold the potential to greatly enhance Digital Audio Workstation (DAW) software, offering more realistic performances compared to existing sampling libraries. This advancement could potentially revolutionize the sampling industry. Additionally, these technologies can be applied in film and game music production, as well as virtual reality experiences, to create high-quality sound demos and even replace human performances, thereby reducing costs significantly. Content creation could also benefit from these developments.

The study of music repetition structures may also enhance other fields. For instance, a study [101] suggested that pretraining on music data can boost model performance in language tasks, likely due to the complex and logical structure of music aiding the learning of language structures. Additionally, my research found that pretraining on singing data enhances prosody and pitch generation in speech synthesis.

6.2 Music Therapy

In Section 3.2, a key motivation for my work on personalized music lies in its application to Music Therapy, particularly Rhythmic Auditory Stimulation (RAS) [121], which assists individuals with

[4] https://www.heygen.com

Parkinson's disease in enhancing gait and mobility. With millions affected by this condition, the challenge lies in finding music that not only adheres to RAS treatment criteria, like beat strength and tempo, but also aligns with a patient's musical preferences. The limited availability of therapists further hinders the effective implementation of RAS. By generating customized music, we can address this gap, creating therapy music tailored to both the patient's music preferences and health requirements. Similar methods can be adapted for other music therapy applications like Rhythmic Auditory Cueing (RAC), which can aid in enhancing safe walking practices among the elderly.

Furthermore, in Musicokinetic Therapy (MKT), where music is employed to facilitate movement and kinesthetic engagement, AI-composed music can adapt in real-time to the movements of patients. By leveraging motion-tracking technologies and models in my thesis, the music can dynamically respond to the pace, rhythm, and intensity of a patient's movements. This adaptive feature is crucial in MKT, as it ensures that the musical tempo and rhythm are in sync with the patient's physical capabilities and therapy goals. Such synchronization between music and movement can significantly improve motor coordination and rhythmic synchronization in patients, leading to better therapy outcomes.

AI-composed music in the thesis also presents significant potential in Music Stimulation by providing a tailored auditory experience that aligns with specific therapeutic goals. In Music Stimulation, where music is used to elicit cognitive or emotional responses, algorithms in my thesis can help create compositions that better target specific music preferences. Moreover, my zero-shot singing synthesis work introduces a potential for incorporating patients' or their relatives' voices, which might enhance the effectiveness and personal relevance of the stimulation music.

6.3 Music Perception

In *Sweet Anticipation: Music and Psychology of Expectation* [64], David Huron introduces a cognitive mechanism on how anticipation and expectation significantly influence emotional responses in music. He poses the idea that much of the emotional response to music arises from the interplay between expectation and the subsequent fulfillment or violation of these expectations. He also discusses how composers manipulate musical elements like melody, harmony, and rhythm to create or defy expectations, thereby eliciting various emotional and psychological responses from listeners. This work inspired my analysis in Section 2.4, where algorithmic methods are employed to analyze expectations and surprise in relation to music repetition structures and predictions. The study in Section 2 could further help extend the perception analysis in a more data-driven manner, and even compare with EEG and MEG results in human experiments.

6.4 Music Education

Expressive singing voice synthesis and conversion technologies in this thesis present novel opportunities in music education. They enable students to access a variety of vocal style demonstrations, enriching their grasp of diverse singing techniques, especially when live demonstrations are unavailable. Additionally, a common challenge in music lessons, such as Solfege, is the difficulty students face in adapting techniques to their own voices. The zero-shot singing synthesis model introduced in this thesis can address this by providing tailored demonstrations suited to individual vocal characteristics, facilitating more effective learning.

Symbolic music composition technologies could enhance music education by offering interactive and adaptive learning tools. On the one hand, students can engage with automatically generated compositions that illustrate specific theoretical concepts, from chord progressions to complex harmonic structures, aiding in a deeper understanding of music theory. On the other hand, imitation algorithms can suggest improvements and alternative approaches to students' compositions, encouraging creative experimentation. If combined with personalized learning design, these technologies can adapt to a student's skill level, presenting challenges that are optimally aligned with their learning curve and fostering a more effective and engaging learning environment.

Technologies in section 2 about music structure studies could help in music education to automatically identify and analyze various structural elements of music, such as phrases, sections, and motifs. This is particularly beneficial for beginners, as it simplifies the visualization of music forms, making it easier for them to comprehend the underlying structure of compositions. When combined with the score following techniques [35], it could visualize the structure components during listening and study.

6.5 Theory Development of Non-Western Music

The data-driven analysis of music (for example, the study of music structure and other elements) could advance music theory by verifying existing music concepts, developing new music features and insights for traditional music theory, and even helping create and develop theories for non-western music. Non-western music from various global regions possesses distinct characteristics that set them apart from the Western musical system. However, many of them lack their own music theories comparable to that of the West. For example, traditional Chinese music boasts a rich history and high artistic merit, yet faces challenges due to its primarily oral transmission and lack of formalized music theory. Recent compositions of traditional Chinese music have adopted Western music theory, risking the loss of traditional instrumental techniques and artistic uniqueness. The evolution of Western music theory itself, a process spanning centuries, underscores that such development is gradual and intricate. In this context, computer music analysis methods, particularly the data-driven approach discussed in this thesis, present a promising avenue for advancing the theoretical frameworks of non-Western music traditions.

6.6 Human-Computer Interactive Live Performance

The Human Computer Music Performance (HCMP) system [37] enables interactive performances between computers and musicians, coordinating both human and artificial performers. These artificial performers can rehearse, follow human players based on scores, and even improvise [134]. Integrating stylistic music composition technologies from this thesis will enhance the system's capabilities in mimicking and personalizing the interactive performance experience. Additionally, the incorporation of stylistic singing and instrumental synthesis promises to enrich the musical experience within HCMP. By embedding generative music AI technologies, along with video generation and robot control, the HCMP system moves closer to realizing the vision of creating artificial musicians capable of composing and performing music collaboratively.

7 Ethical and Legal Implications of AI Music: Challenges and Discussion

I believe that the aim of copyright protection is not for protection itself but for better invention and distribution. In this chapter, I plan to look back on the ethical and legal issues in the history of the music industry, compare them with current challenges of AI music, compare differences in ethical issues between AI music and other AI areas, discuss potential technical support and potential impact on the music industry. Here are the main points I would like to cover in this chapter:

- Ethical and legal issues in the history of the music industry, including traditional concepts of music copyrights and royalties, especially the past experience with recordings and samples.
- Challenges posed by AI-generated music: (1) Who owns the rights? (2) Transparency and interoperability of AI music to creators and customers; (3) Potential impacts on revenue models for artists and composers.
- Compare ethical issues in music to other types of arts: In which way is music different from some other concerns? Especially what is novel/specific to music. This will also help people in the general AI area understand the challenges in AI music.
- Authenticity and Artistic Value: defining genuine artistic expression in the age of AI. The value of human creativity in music (e.g., even the AI chess is much better than human, human chess is still popular). Potential for AI to overshadow or diminish human musical contributions.
- Potential impact on the music industry. How AI music could potentially change the way people create, perform, listen to, and transmit music, bringing potential job displacement and industry evolution. The changing landscape: Roles at risk in an AI-dominated music industry. Opportunities for new roles and collaborations between AI and human artists, for example, AI as a tool for music interactively.
- Bias, Diversity, and Representation. The impact of training data on AI music output. Risks of sidelining underrepresented music styles. The imperative for diversity and inclusivity in AI music models.
- Privacy Concerns in Personalized AI Music Experiences. Data collection and usage for tailored music experiences. Transparency, consent, and user rights in an AI-driven music ecosystem.
- Potential technical support for helping solve the above challenges. For example, music watermark, music similarity comparison and better fingerprint algorithms.

My plan for this chapter is:

- Survey on existing papers on AI music ethics and similar topics in general AI ethics. Get a sense of what is missing in current literature, as well as the writing style and workload of such a chapter.

- Setup a goal: will this chapter be a broad overview? Or an in-depth discussion towards one direction of the above main points?
- Write and revise an outline of problems that need discussion.
- Schedule multiple rounds of discussions with experts.
- Summarize the discussion results and write the paper.

8 Schedule

Nov 2023 - Feb 2024	Propose thesis Chapter 4.2 Instrumental Performance Control
Feb 2024 - May 2024	Chapter 5.3 Instrumental Synthesis Chapter 7 AI Ethics and Legalty
June 2024 - July 2024	Thesis write-up
July 2024	Thesis defense

9 Conclusion

I am proposing a scheme for creating artificial musicians across three different music creation levels and representation modalities: symbolic music composition, expressive performance control, and music audio synthesis. Meanwhile, this work targets two major challenges in current music creation: integration of hierarchical music repetition structure and personalized music style, achieved by integrating music domain knowledge into the modeling process. In addition, I will discuss the potential applications of these technologies, as well as the ethical and legal implications of AI music.

So far, I have already finished works in Chapter 2 (computational study of music repetition and structure) and Chapter 3 (symbolic music composition with structure, style and control). Expressive performance control and audio synthesis for singing voice are also in the finishing stage. I will be continuing the exploration of instrumental performance control and synthesis in the next few months. At the same time, I am organizing discussions on AI music ethics among musicians, lawyers, people in the music industry, AI ethics experts, and music technologists, in order to gain more insights into Chapter 7.

If time allows, I will explore more application scenarios described in Chapter 6. For example, combining this work with our previous Human Computer Performance System to make an interactive performance demo with artificial musicians. Moreover, I would like to explore more insights for this work in depth.

In all, I believe the topic is very interesting, the experimental approach is clear, and I should be able to finish the thesis project within the scheduled time frame.

10 Reference

[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023. 1

[2] P. Allegraud, L. Bigo, L. Feisthauer, M. Giraud, R. Groult, E. Leguy, and F. Levé. Learning sonata form structure on mozartś string quartets. *Transactions of the International Society for Music Information Retrieval*, 2, Dec. 2019. 2.1.2

[3] and others. Filosax: A dataset of annotated jazz saxophone recordings. 2021. 4.2

[4] L. Barrington, A. B. Chan, and G. Lanckriet. Dynamic texture models of music. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1589–1592, 2009. 2.1.2

[5] J. Berardinis, M. Vamvakaris, A. Cangelosi, and E. Coutinho. Unveiling the hierarchical structure of music by multi-resolution community detection. *Transactions of the International Society for Music Information Retrieval*, 3(1):82–97, 2020. 2.1.2

[6] David Berger and Chuck Israels. *The Public Domain Song Anthology*. Aperio, Charlottesville, Mar 2020. ISBN 978-1-7333543-0-1. doi: 10.32881/book2. 2.2

[7] Daniel Ellis Berlyne. *Aesthetics and Psychobiology*. New York: Appleton-Century-Crofts, 1971. 3.2.1, 4

[8] Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*, 2023. 5, 5.1.2

[9] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. *arXiv preprint arXiv:1206.6392*, 2012. 3.1

[10] Jean-Pierre Briot, Gaëtan Hadjeres, and François Pachet. Deep learning techniques for music generation. *Springer*, 10, 2019. 2.1.3, 3.1

[11] Nazif Can Tamer, Pedro Ramoneda, and Xavier Serra. Violin etudes: a comprehensive dataset for f0 estimation and performance analysis. In *in Proceedings of the 23nd International Society for Music Information Retrieval Conference (ISMIR)*, 2022. 4.2

[12] William E. Caplan. *Classical Form: A Theory of Formal Functions for the Instrumental Music of Haydn, Mozart, and Beethoven, Revised Edition*. Oxford University Press, 2000. 2.1.1

[13] Ke Chen, Cheng-i Wang, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Music sketchnet: Controllable music generation via factorized representations of pitch and rhythm. In *Proc. of the 21st International Society for Music Information Retrieval Conference*. ISMIR, 2020. 3.1

[14] Anthony Chmiel and Emery Schubert. Back to the inverted-u for music preference: A review of the literature. *Psychology of Music*, 45(6):886–909, 2017. 3.2.1

[15] Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Lee, Hoon Heo, and Kyogu Lee.

Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in Neural Information Processing Systems*, 34:16251–16265, 2021. 5.1.1

[16] Soonbeom Choi, Wonil Kim, Saebyul Park, Sangeon Yong, and Juhan Nam. Children's song dataset for singing voice research. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2020. 4.2

[17] Ching-Hua Chuan and Dorien Herremans. Modeling temporal tonal relations in polyphonic music through deep networks with a novel image-based representation. In *Proc. of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2.1.3

[18] Martin Clayton, Rebecca Sager, and Udo Will. In time with the music: the concept of entrainment and its significance for ethnomusicology. In *European meetings in ethnomusicology.*, volume 11, pages 1–82. Romanian Society for Ethnomusicology, 2005. 4.1

[19] Manfred Clynes. Secrets of life in music: Musicality realised by computer. In *Proc. Intl. Comouter Music Conf., 1984*, pages 225–232, 1984. 4.1

[20] Manfred Clynes and Edward C Carterette. Music, mind, and brain: The neuropsychology of music edited by manfred clynes. *The Journal of the Acoustical Society of America*, 75 (4):1308–1309, 1984. 4.1

[21] T. Collins, A. Arzt, S. Flossmann, and G. Widmer. Siarct-cfp: Improving precision and the discovery of inexact musical patterns in point-set representations. In *Proceedings of the International Society for Music Information Retrieval*, pages 549–554, 2013. 2.1.2

[22] Tom Collins and Robin Laney. Computer-generated stylistic compositions with long-term repetitive and phrasal structure. *Journal of Creative Music Systems*, 1(2), 2017. 2.1.3, 3.1

[23] David Cope. *Computers and musical style*, volume 6. Oxford University Press Oxford, 1991. 3.1

[24] David Cope. *The Algorithmic Composer*. AR Editions, Inc., 2000. 3.1

[25] David Cope. *Computer models of musical creativity*. MIT Press Cambridge, 2005. 3.1

[26] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre D'efossez. Simple and controllable music generation. *ArXiv*, abs/2306.05284, 2023. URL https://api.semanticscholar.org/CorpusID:259108357. 1, 5.1.2

[27] S. Dai, H. Yu, and R. Dannenberg. What is missing in deep music generation? a study of repetition and structure in popular music. In *Proceedings of the International Symposium on Music Information Retrieval*, 2022. 1, 2.2, 2.3, 2.5

[28] Shuqi Dai, Zheng Zhang, and Gus G Xia. Music style transfer: A position paper. In *Proceedings of 6th International Workshop on Musical Metacreation*, 2018. 1, 3.2.1

[29] Shuqi Dai, Huan Zhang, and Roger B. Dannenberg. Automatic analysis and influence of hierarchical structure on melody, rhythm and harmony in popular music. In *in Proc. of the 2020 Joint Conference on AI Music Creativity (CSMC-MuMe 2020)*, 2020. 2.1.1, 2.2, 2.3

[30] Shuqi Dai, Huan Zhang, and Roger B. Dannenberg. Automatic analysis and influence of hierarchical structure on melody, rhythm and harmony in popular music. In *Proc. of the*

2020 Joint Conference on AI Music Creativity (CSMC-MuMe), 2020. 1

[31] Shuqi Dai, Zeyu Jin, Celso Gomes, and Roger B Dannenberg. Controllable deep melody generation via hierarchical music structure representation. In *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.*, 2021. 1, 2.1.3, 2.5, 3

[32] Shuqi Dai, Xichu Ma, Ye Wang, and Roger B. Dannenberg. Personalized popular music generation using imitation and structure. *arXiv preprint arXiv:2105.04709*, 2021. 1, 2.1.3, 3, 3.1

[33] Shuqi Dai, Siqi Chen, Yuxuan Wu, Ruxin Diao, Roy Huang, and Roger B. Dannenberg. Singstyle111: A multilingual singing dataset with style transfer. In *in Proc. of the 24th Int. Society for Music Information Retrieval Conf.*, 2023. 1, 4.2

[34] R. Dannenberg and M. Goto. *Music Structure Analysis from Acoustic Signals*, volume 1, pages 305–331. Springer Verlag, 2009. doi: 10.1007/978-0-387-30441-0_21. 2.1.1, 2.1.2

[35] Roger B Dannenberg. An on-line algorithm for real-time accompaniment. In *ICMC*, volume 84, pages 193–198, 1984. 6.4

[36] Roger B Dannenberg. Music representation issues, techniques, and systems. *Computer Music Journal*, 17(3):20–30, 1993. 1, 3.2.1, 4.1

[37] Roger B Dannenberg. Human computer music performance. In *Dagstuhl follow-ups*, volume 3. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012. 6.6

[38] Roger B Dannenberg and Istvan Derenyi. Combining instrument and performance models for high-quality music synthesis. *Journal of New Music Research*, 27(3):211–238, 1998. 4.1, 5.1.2

[39] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 1, 2.1.3

[40] Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, and Shuicheng Yan. Video background music generation with controllable music transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2037–2045, 2021. 1

[41] Simon Dixon, Werner Goebl, and Gerhard Widmer. Real time tracking and visualisation of musical expression. In *International Conference on Music and Artificial Intelligence*, pages 58–68. Springer, 2002. 4.1

[42] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. *Proc. of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. 3.1

[43] Zhiyao Duan, Bryan Pardo, and Changshui Zhang. Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2121–2133, 2010. 4.2

[44] Homer Dudley. The vocoder—electrical re-creation of speech. *Journal of the Society of Motion Picture Engineers*, 34(3):272–278, 1940. 5.1.1

[45] Z. Eitan and R. Y. Granot. Growing oranges on mozart's apple tree: "inner form" and aesthetic judgment. *Music Perception: An Interdisciplinary Journal*, 25(5):397–417, 2008. 2.4

[46] Anders Elowsson and Anders Friberg. Algorithmic composition of popular music. In *Proc. of the 12th Int. Conference on Music Perception and Cognition and the 8th Triennial Conf. of the European Society for the Cognitive Sciences of Music*, pages 276–285, 2012. 2.1.3, 3.1, 3.2.2

[47] Jesse Engel, Chenjie Gu, Adam Roberts, et al. Ddsp: Differentiable digital signal processing. In *International Conference on Learning Representations*, 2019. 5.1.2

[48] Behnam Faghih and Joseph Timoney. Annotated-vocalset: A singing voice dataset. *Applied Sciences*, 12(18):9257, 2022. 4.2

[49] S. Go, N. Ryo, N. Eita, and Y. Kazuyoshi. Statistical music structure analysis based on a homogeneity-, repetitiveness-, and regularity-aware hierarchical hidden semi-markov model. In *Proceedings of the International Symposium on Music Information Retrieval*, 2019. 2.1.2

[50] Werner Goebl, Elias Pampalk, and Gerhard Widmer. Exploring expressive performance trajectories: Six famous pianists play six chopin pieces. In *Proceedings of the 8th international conference on music perception and cognition*, pages 505–509. Causal Productions Sydney, Australia, 2004. 4.1

[51] I. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, B. Xu, David Warde-Farley, S. Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *ArXiv*, abs/1406.2661, 2014. 3.1

[52] M. Granroth-Wilding. *Harmonic analysis of music using combinatory categorial grammar*. PhD thesis, Univ. of Pennsylvania, 2013. 2.1.2

[53] Gaëtan Hadjeres and Léopold Crestel. Vector quantized contrastive predictive coding for template-based music generation. *arXiv preprint arXiv:2004.10120*, 2020. 2.1.3

[54] Gaëtan Hadjeres, François Pachet, and Frank Nielsen. Deepbach: a steerable model for bach chorales generation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1362–1371. JMLR. org, 2017. 3.1

[55] M. Hamanaka, K. Hirata, and S. Tojo. Musical structural analysis database based on gttm. In *Proceedings of the International Symposium on Music Information Retrieval*, 2014. 2.1.2

[56] Lejaren Hiller and Leonard Isaacson. *Experimental Music: Composition with an Electronic Computer*. McGraw-Hill, New York, 1959. 3.1

[57] Lejaren Hiller, Charles Ames, and Robert Franki. Automated composition: An installation at the 1985 international exposition in tsukuba, japan. *Perspectives of New Music*, 23(2): 196–215, 1985. ISSN 00316016. 3.1, 3.3

[58] Ning Hu. *Automatic Construction of Synthetic Musical Instruments and Performers*. PhD thesis, Carnegie Mellon University, 2013. 4.1, 5.1.2

[59] C.-Z. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. Dai,

M. Hoffman, M. Dinculescu, and D. Eck. Music transformer. *arXiv preprint arXiv:1809.04281*, 2018. 2.1.3, 2.5, 3.1

[60] Cheng-Zhi Anna Huang, Tim Cooijmans, Adam Roberts, Aaron Courville, and Douglas Eck. Counterpoint by convolution. In *Proc. of the 18th Int. Society for Music Information Retrieval Conf.*, Suzhou, China, 2017. 2.1.3, 3.1

[61] Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*, 2022. 1

[62] Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3945–3954, 2021. 5.1.1

[63] Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Generating music with rhythm and harmony. *arXiv preprint arXiv:2002.00212*, 2020. 3.1

[64] David Huron. *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press, Cambridge, MA, 2006. 2.1.1, 2.4, 6.3

[65] Ji-Sang Hwang, Sang-Hoon Lee, and Seong-Whan Lee. Hiddensinger: High-quality singing voice synthesis via neural audio codec and latent diffusion models. *arXiv preprint arXiv:2306.06814*, 2023. 5.1.1

[66] N. Jiang and M. Müller. Automated methods for analyzing music recordings in sonata form. In *Proceedings of the International Society for Music Information Retrieval*, pages 595–600, 2013. 2.1.2

[67] Z. Jiang and R. Dannenberg. Melody identification in standard midi files. In *Proceedings of the 16th Sound & Music Computing Conference*, pages 65–71, 2019. 2.1.2

[68] Nicolas Jonason, Bob Sturm, and Carl Thomé. The control-synthesis approach for making expressive and controllable neural music synthesizers. In *2020 AI Music Creativity Conference*, 2020. 4.1, 5.1.2

[69] Olivier Julian and Christophe Levaux, editors. *Over and Over: Exploring Repetition in Popular Music*. Bloomsbury Academic, 2018. 2.1.1

[70] Patrik N Juslin and Renee Timmers. Expression and communication of emotion in music performance. *Handbook of music and emotion: Theory, research, applications*, pages 453–489, 2010. 4.1

[71] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 5.2

[72] Lisa Kawai, Philippe Esling, and Tatsuya Harada. Attributes-aware deep music transformation. In *Proceedings of the 21st International Society for Music Information Retrieval Conference, ISMIR 2020*. ISMIR, 2020. 3.1

[73] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021. 5.1.1

[74] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3.1

[75] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020. 5.1.1, 5.1.1

[76] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020. 4.3, 5.1.1, 5.1.1

[77] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32, 2019. 5, 5.1.1

[78] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *arXiv preprint arXiv:2306.06546*, 2023. 5, 5.1.1, 5.1.1

[79] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *The Eleventh International Conference on Learning Representations*, 2022. 5.1.1, 5.1.1

[80] M. Lehne, M. Rohrmeier, D. Gollmann, and S. Koelsch. The influence of different structural features on felt musical tension in two piano pieces by mozart and mendelssohn. *Music Perception: An Interdisciplinary Journal*, 31(2):171–185, 2012. 2.1.1

[81] Alexander Lerch, Claire Arthur, Ashis Pati, and Siddharth Gururani. Music performance analysis: A survey. In *in Proc. of the 20th International Society for Music Information Retrieval Conference*, 2019. 4.1

[82] Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 21(2):522–535, 2018. 4.2

[83] Feynman Liang. Bachbot: Automatic composition in the style of bach chorales. *University of Cambridge*, 8:19–48, 2016. 3.1

[84] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 5.2

[85] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11020–11028, 2022. 1, 4.2, 5.1.1, 5.1.1, 5.2

[86] Peiling Lu, Jie Wu, Jian Luan, Xu Tan, and Li Zhou. Xiaoicesing: A high-quality and integrated singing voice synthesis system. *arXiv preprint arXiv:2006.06261*, 2020. 1

[87] H. Lukashevich. Towards quantitative measures of evaluating song segmentation. In *ISMIR*, pages 375–380, 2008. 2.1.2

[88] Elizabeth Hellmuth Margulis. *On Repeat: How Music Plays the Mind*. Oxford University Press, 2013. 2.1.1

[89] A. Marsden. Recognition of variations using automatic schenkerian reduction. In *Proceedings of the International Symposium on Music Information Retrieval*, 2010. 2.1.2

[90] A. Marsden, K. Hirata, and S. Tojo. Towards computable procedures for deriving tree structures in music: Context dependency in gttm and schenkerian theory. In *Sound & Music Computing Conference*, 2013. 2.1.2

[91] K. Masada and R. C. Bunescu. Chord recognition in symbolic music: A segmental crf model, segment-level features, and comparative evaluations on classical and popular music. *Transactions of the International Society for Music Information Retrieval*, 2018. 2.1.2

[92] Nicholas Meade, Nicholas Barreyre, Scott C Lowe, and Sageev Oore. Exploring conditioning for generative music systems with human-interpretable controls. *arXiv preprint arXiv:1907.04352*, 2019. 4.1

[93] Gabriele Medeot, Srikanth Cherla, Katerina Kosta, Matt McVicar, Samer Abdallah, Marco Selvi, Ed Newton-Rex, and Kevin Webster. Structurenet: Inducing structure in generated melodies. In *Proc. of 19st Int. Conference on Music Information Retrieval Conf., ISMIR*, pages 725–731, 2018. 2.1.3, 3.1

[94] Masahiro Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970. 3.2.1

[95] Daniel Müllensiefen, David Lewis, Christophe Rhodes, and Geraint Wiggins. Treating inherent ambiguity in ground truth data: Evaluating a chord labelling algorithm. In *8th International Conference on Music Information Retrieval (ISMIR)*, 2007. 3.2.1

[96] Eugene Narmour et al. *The analysis and cognition of melodic complexity: The implication-realization model*. University of Chicago Press, 1992. 2.1.1

[97] O. Nieto, G. J. Mysore, C. C. Wang, J. B. L. Smith, J. Schlüter, T. Grill, and B. McFee. Audio-based music structure analysis: Current trends, open challenges, and applications. *Transactions of the Int. Society for Music Information Retrieval Conf.*, 3(1):246–263, 2020. 2.1.1

[98] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016. 4.3, 5, 5.1.1

[99] Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. This time with feeling: Learning expressive musical performance. *Neural Computing and Applications*, 32:955–967, 2020. 4.1

[100] C. Palmer and C. Krumhansl. Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing, and familiarity. *Perception & Psychophysics*, 41 (6):505–518, 1987. 2.1.1

[101] Isabel Papadimitriou and Dan Jurafsky. Learning music helps you read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839, 2020. 6.1

[102] Jouni Paulus, Meinard Muller, and Ansii Klapuri. Audio-based music structure analysis. In *Proc. of the 11th Int. Society for Music Information Retrieval Conf.*, pages 625–636, 2010. 2.1.1

[103] Christine Payne. Musenet. *OpenAI, openai.com/blog/musenet*, 2019. 2.1.3, 3.1

[104] Marcus T. Pearce. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. PhD thesis, Department of Computer Science, City University of London, UK, 2005. 3.1

[105] Nathanaël Perraudin, Peter Balazs, and Peter L Søndergaard. A fast griffin-lim algorithm. In *2013 IEEE workshop on applications of signal processing to audio and acoustics*, pages 1–4. IEEE, 2013. 5.1.1

[106] Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, David Cox, Mark Hasegawa-Johnson, and Shiyu Chang. Contentvec: An improved self-supervised speech representation by disentangling speakers. In *International Conference on Machine Learning*, pages 18003–18017. PMLR, 2022. 5.1.1

[107] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*, 2020. 5.1.1, 5.1.1

[108] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In *Proceedings of the 35th International Conference on Machine Learning*, 2018. 3.1

[109] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In *Proc. of the International conference on machine learning*, pages 4364–4373. PMLR, 2018. 2.1.3

[110] Jonathan J. Rolison and Judy Edworthy. The role of formal structure in liking for popular music. *Music Perception: An Interdisciplinary Journal*, 29(3):269–284, 2012. 2.1.1

[111] P-Y. Rolland. Discovering patterns in musical sequences. *Journal of New Music Research*, 28(4):334–350, 1999. 2.1.2

[112] Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Scholkopf. Moûsai: Text-to-music generation with long-context latent diffusion. 2023. URL https://api.semanticscholar.org/CorpusID:264439679. 1

[113] Walter Schulze and Brink Van Der Merwe. Music generation with markov models. *IEEE Annals of the History of Computing*, 18(03):78–85, 2011. 3.1

[114] Xavier Serra and Julius Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24, 1990. 5.1.2

[115] Zhengshan Shi. Computational analysis and modeling of expressive timing in chopin's mazurkas. In *ISMIR*, pages 650–656, 2021. 4.1

[116] W Sluckin, DJ Hargreaves, and AM Colman. Novelty and human aesthetic preferences. *Exploration in animals and humans*, pages 245–269, 1983. 3.2.1, 4

[117] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 5.1.1

[118] Bob L Sturm and Oded Ben-Tal. Taking the models back to music practice: Evaluating generative transcription models built using deep learning. *Journal of Creative Music Systems*, 2:32–60, 2017. 2.1.3

[119] Jay Summach. The structure, function, and genesis of the prechorus. *Music Theory Online*, 17(3), October 2011. 2.1.1

[120] Hao Hao Tan and Dorien Herremans. Music fadernets: Controllable music generation based on high-level features via low-level feature modelling. In *Proc. of 21st International Conference on Music Information Retrieval, ISMIR*, 2020. 3.1

[121] Michael H Thaut, Gerald C McIntosh, Ruth R Rice, Robert A Miller, J Rathbun, and JM Brault. Rhythmic auditory stimulation in gait training for parkinson's disease patients. *Movement disorders: official journal of the Movement Disorder Society*, 11(2):193–200, 1996. 6.2

[122] Benjamin van Niekerk, Marc-André Carbonneau, Julian Zaïdi, Matthew Baas, Hugo Seuté, and Herman Kamper. A comparison of discrete and soft speech units for improved voice conversion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6562–6566. IEEE, 2022. 5.1.1

[123] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2.1.3, 3.1

[124] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE, 2018. 5.1.1, 5.3

[125] Changhong Wang, Emmanouil Benetos, Elaine Chew, et al. Cbf-peridb: a chinese bamboo flute dataset for periodic modulation analysis. 2019. 4.2

[126] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023. 5.1.1, 5.3

[127] Yu Wang, Xinsheng Wang, Pengcheng Zhu, Jie Wu, Hanzhao Li, Heyang Xue, Yongmao Zhang, Lei Xie, and Mengxiao Bi. Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis. *arXiv preprint arXiv:2201.07429*, 2022. 4.2

[128] Ziyu Wang, Ke Chen, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, Guxian Bin, and Gus Xia. Pop909: A pop-song dataset for music arrangement generation. In *Proc. of 21st Int. Conference on Music Information Retrieval Conf.*, 2020. 1, 2.2

[129] Ziyu Wang, Dingsu Wang, Yixiao Zhang, and G. Xia. Learning interpretable representation for controllable polyphonic music generation. In *Proc. of 21st International Conference on Music Information Retrieval, ISMIR*, 2020. 3.1

[130] Shiqi Wei and Gus Xia. Learning long-term music representations via hierarchical contextual constraints. In *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.*,

2021. 2.1.3, 2.5

[131] Julia Wilkins, Prem Seetharaman, Alison Wahl, and Bryan Pardo. Vocalset: A singing voice dataset. In *ISMIR*, pages 468–474, 2018. 4.2

[132] Jian Wu, Xiaoguang Liu, Xiaolin Hu, and Jun Zhu. Popmnet: Generating structured pop music melodies using neural networks. *Artificial Intelligence*, 286:103303, 2020. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2020.103303. 3.1

[133] Yusong Wu, Ethan Manilow, Yi Deng, Rigel Swavely, Kyle Kastner, Tim Cooijmans, Aaron Courville, Cheng-Zhi Anna Huang, and Jesse Engel. Midi-ddsp: Detailed control of musical performance via hierarchical modeling. In *International Conference on Learning Representations*, 2021. 4.1, 5.1.2

[134] Gus Guangyu Xia. Expressive collaborative music performance via machine learning. 2016. 4.1, 6.6

[135] Shengyuan Xu, Wenxiao Zhao, and Jing Guo. Refinegan: Universally generating waveform better than ground truth with highly accurate pitch and intensity responses. *arXiv preprint arXiv:2111.00962*, 2021. 5.1.1

[136] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203. IEEE, 2020. 5.1.1

[137] Li-Chia Yang and Alexander Lerch. On the evaluation of generative models in music. *Neural Computing and Applications*, 32(9):4773–4784, 2020. 2.1.3

[138] Ruihan Yang, Dingsu Wang, Ziyu Wang, Tianyao Chen, Junyan Jiang, and Gus Xia. Deep music analogy via latent representation disentanglement. In *20th International Society for Music Information Retrieval Conference*, 2019. 3.1

[139] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2.1.3, 3.1

[140] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021. 5, 5.1.1, 5.1.2

[141] Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, et al. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. *Advances in Neural Information Processing Systems*, 35:6914–6926, 2022. 4.2

[142] Yongmao Zhang, Heyang Xue, Hanzhao Li, Lei Xie, Tingwei Guo, Ruixiong Zhang, and Caixia Gong. Visinger 2: High-fidelity end-to-end singing voice synthesis enhanced by digital signal processing synthesizer. *arXiv preprint arXiv:2211.02903*, 2022. 1, 5.1.1

[143] Yu Zhang, Ziya Zhou, Xiaobing Li, Feng Yu, and Maosong Sun. Ccom-huqin: an annotated multimodal chinese fiddle performance dataset. *arXiv preprint arXiv:2209.06496*, 2022. 4.2

[144] Zewang Zhang, Yibin Zheng, Xinhui Li, and Li Lu. Wesinger 2: fully parallel singing voice synthesis via multi-singer conditional adversarial training. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1, 5.1.1