氏　　名　Zhe Zhang

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2531 号

学位授与の日付　2024 年 9 月 27 日

学位授与の要件　複合科学研究科　情報学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Generative Models for Cross-Modal Music Generation

論 文 審 査 委 員　主　査　高須　淳宏
情報学コース　教授
相澤　彰子
情報学コース　教授
神門　典子
情報学コース　教授
武田　英明
情報学コース　教授
佐藤　真一
国立情報学研究所　コンテンツ科学研究系　教授
浜中　雅俊
理化学研究所　革新知能統合研究センター
音楽情報知能チーム　チームリーダー

# Summary of Doctoral Thesis

Name in full Zhe Zhang

Title Generative Models for Cross-Modal Music Generation

Music is an essential medium of human expression and culture, deeply impacting emotional and social interaction regardless of language barrier. With the development of artificial intelligence (AI) techniques, the music creation process is also evolving, particularly through the advent of generative models. Music generation models aim to mimic human music creation behaviors to produce novel and coherent music. Specifically, cross-modal music generation is a more complicated and creative process, requiring a sophisticated design of architectures to make the system learn from the music data distribution and be aware of musical domain knowledge in various music modalities. However, a significant challenge in this task is to bridge the gap between the different modalities of music, such as audio, lyrics, and symbolic music, due to their complex and unclear inter-modality relationships. Existing research of cross-modal music generation work often adopts advanced techniques but without carefully considering how to bridge the gap between music modalities, resulting in degraded generation or limitations in application.

Generative models for cross-modal music generation is an important topic due to the potential to integrate different modalities of musical expression via deep learning techniques. By applying elaborated methods to address the technical issues in bridging the modality gaps, the models are prospective to generate high-quality, plausible, and diverse music content. This task not only pushes traditional boundaries of music creation but also enhances accessibility, allowing individuals without extensive musical training to engage in music production.

Aiming at bridging the modality gaps in generative models for cross-modal music generation, this dissertation further identifies several specific technical issues that can prospectively overcome the modality gap from different aspects: i) Integration of musical domain knowledge; ii) Training objectives for generating musical contents; iii) Controllable and interactive generation; iv) Diverse generation; and v) Multi-modal information in music datasets. Based on different cross-modal music generation tasks, this dissertation proposes novel techniques to address the above issues, making contributions to building more intelligent cross-modal music generation systems that can overcome the gaps between modalities.

We first propose a controllable lyrics-to-melody generation network that is able to generate realistic melodies from lyrics in user-desired musical style. To model the dependencies of music attributes across multiple sequences, inter-branch memory

fusion (Memofu) is proposed to enable information flow between multi-branch stacked LSTM architecture, thus integrating the musical domain knowledge. Reference style embedding (RSE) is proposed to improve the quality of generation as well as control the musical style of generated melodies. Human interaction is also feasible in the inference stage to steer the generation process. Sequence-level statistical loss (SeqLoss) is proposed to help the model learn sequence-level features of melodies given lyrics, which is a tailored objective for generating musical content.

Next, we propose a controllable melody-to-lyrics model that is able to generate syllable-level lyrics with user-desired rhythm. A prior attention mechanism is proposed to enhance the controllability and diversity of lyrics generation while incorporating syllable structure information of lyrics. An explicit n-gram (EXPLING) loss is proposed to train the Transformer-based model to capture the sequence dependency and alignment relationship between melody and lyrics and predict the lyrics sequences at the syllable level. Experiments and evaluation metrics verified that our proposed model is capable of generating higher-quality lyrics than previous methods and the feasibility of interacting with users for controllable and diverse lyrics generation.

Finally, we address a more fundamental issue regarding the gap between music modalities: the scarcity of multi-modal information in music datasets. To overcome this, we present a modularized dataset construction pipeline designed to create aligned audio-MIDI-lyrics (AUMILY) datasets, leveraging the capabilities of the latest techniques and pre-trained models. This pipeline processes collected songs at different alignment levels, resulting in three subsets tailored for various tasks in cross-modal music generation. The modularized construction pipeline is the first of its kind to support the integration of multi-track audio stems, symbolic music alignment, and synchronized lyrics. It offers a flexible framework that allows researchers to create new datasets from scratch according to specific needs, such as music genre and lyrics language. Additionally, the pipeline can be updated to incorporate advanced techniques, continuously improving the dataset quality. By providing a robust and adaptable method for generating richly informative multi-modal music datasets, this pipeline significantly contributes to bridging the modality gaps in cross-modal music generation tasks. It also opens new avenues for broader research on generative models, facilitating advancements in music AI by enabling more sophisticated and nuanced interactions between different musical modalities.

（別紙 Separate Sheet）

Results of the Doctoral Thesis Defense

# 博士論文審査結果

Name in Full
氏　　名　Zhe Zhang

Ｔｉｔｌｅ
論文題目　Generative Models for Cross-Modal Music Generation

本学位論文は，「Generative Models for Cross-Modal Music Generation」と題し，全６章から構成されている．

まず第１章で，本研究が扱うクロスモーダル音楽生成問題の背景および問題の定義，解くべき課題について述べた後に本論文の主な貢献を示している

第２章では，本研究の要素技術である生成モデル，モダリティと楽曲生成，クロスモーダル楽曲生成についての関連研究をまとめている．続く３つの章で本博士論文の主たる貢献を述べている．

第３章では，歌詞からメロディへのクロスモーダル楽曲生成問題を論じている．与えられた歌詞から作曲を行う場合に膨大なメロディが候補となりうる．そこで，作曲者の意図にあったメロディを生成するための楽曲スタイルの埋め込み表現法を提案している．本章では，メロディを構成する音の高さや長さ，休符の長さなどを関連づけられるように拡張した LSTM と楽曲スタイルを表すモデルを統合し，敵対的学習を行うことで得られるメロディ生成モデルを提案している．評価用データを用いた既存手法との比較実験および被験者実験により，本論文で提案した生成モデルによって多様な作曲が可能であることが示されている．

第４章では，与えられたメロディから作詞を行う方法を提案している．本章では，MIDI 形式で与えられるメロデイに対するエンコーダと歌詞を生成するデコーダから構成される Transformer を提案している．ここでは，音素のレベルで作詞を制御する事前アテンション機構を提案し，デコーダに組み込むことでメロディに整合する歌詞を生成することを可能にしている．

第３章および第４章で述べられた研究を通して，楽曲のさまざまなモダリティに対応した楽曲生成モデルの学習のためには大規模マルチモーダル楽曲データセットが重要であることが判明した．そこで，第５章では，MIDI，歌詞，オーディオ信号で表された楽曲データベースを構築するためのシステムを提案している．このシステムは，楽曲オーディオデータベースと歌詞データベースから，歌詞，MIDI，オーディオトラックを対応づけたデータを生成するソフトウェアモジュール群より構成されており，楽曲生成モデル用の大規模学習データを生成できることが示されている．

第６章では本論文のまとめと残された課題について述べている．

公開発表会では博士論文の章立てに従って発表が行われた．その後に行われた論文審査会および口述試験では，審査委員から，多くの層から構成される楽曲生成モデルの各層と歌詞生成制御機能との関係，現在利用可能な楽曲生成モデル学習用データと本論文で提案するシステムによって構築されるマルチモーダル楽曲データベースの特性の違いなどについて質疑があり，出願者は適切に回答した．

質疑応答後に審査委員会を開催し審査委員で議論を行った．博士論文審査の結果，出願者は情報学分野の十分な知識と研究能力を持つと認められた．本博士論文は、歌詞，MIDI，オーディオ信号な

どさまざまなモダリティの音楽生成技術を提案している．本研究が提案する音楽生成過程の制御技術は，モダリティ間で成り立つ多様な関係を制御することを可能にするもので，音楽生成の分野の発展に貢献する学術的価値の高い研究と認められた．

また、本学位論文の成果は，国際学術雑誌論文２件，査読付き国際会議論文１件として発表されており，社会的な評価も得ている．以上の理由により，審査委員会は本学位論文が学位の授与に値すると判断した．