A Unified Generative and Discriminative Approach to Automatic Chord Estimation for Music Audio Signals(Abstract_要旨)

AUTHOR(S):

Wu, Yiming

（ 続紙 1 ）

京都大学	博士（情報学）	氏名	Yiming Wu
論文題目	A Unified Generative and Discriminative Approach to Automatic Chord Estimation for Music Audio Signals（音楽音響信号に対する自動コード推定のための生成・識別統合的アプローチ）		

（論文内容の要旨）

This thesis describes a principled statistical approach to automatic chord estimation (ACE) for music audio signals. The chord is an important mid-level representation of polyphonic music that lies between the musical intentions of composers and the actual musical sounds. The ACE task has thus been one of the fundamental research topics in the field of music information retrieval (MIR).

There are two major approaches to ACE. The *discriminative* approach is based on a chord classification model implemented by a deep neural network (DNN) that directly estimates a chord sequence from a music signal. The performance of this approach is bounded by the amount of annotated data used for supervised training. In contrast, the *generative* approach is based on a probabilistic latent variable model that represents the generative process of a music signal from a chord sequence. To make the inverse problem analytically solvable, only classical models such as hidden Markov models (HMMs) have mainly been used.

To overcome these limitations, in this thesis two techniques are proposed for neural ACE. First, a sufficient amount of external MIDI data are used for learning to extract chroma vectors from music signals. Second, a unified generative and discriminative approach is taken for linking a deep classification model to a deep generative model through a latent chord sequence following a language model in a variational autoencoding manner. This enables semi-supervised training of the classification model regularized by the generative and language models.

Chapter 2 reviews related work on ACE. Specifically, hand-crafted and data-driven feature representation methods for ACE, generative and discriminative approaches to ACE, and multi-task learning methods for joint estimation of multiple musical elements are introduced. The evaluation metrics are also explained.

Chapter 3 proposes a DNN-based chroma vector extraction method that predicts the existence of each of twelve pitch classes in higher (vocal), lower (bass), and wider (accompaniment) frequency ranges. Specifically, a DNN is trained in a supervised manner by using a large amount of pairs of MIDI-formatted musical scores and the corresponding synthesized music signals. It has experimentally been shown that a chord classification model trained with the proposed chroma vectors outperforms one trained with the hand-crafted chroma vectors.

Chapter 4 proposes a unified generative and discriminative approach to ACE based on amortized variational inference (AVI). Specifically, a deep generative model that represents the generative process of chroma vectors from discrete chord labels following a Markov language model and continuous latent features following a Gaussian distribution is formulated. Given chroma vectors as observed data, the posterior distributions of the latent labels and features are predicted with deep classification and recognition models, respectively. These three models form

a variational autoencoder (VAE) and are trained jointly in a (semi-)supervised manner. It has experimentally been shown that the performance of ACE can be improved even under a fully supervised condition thanks to the regularization of the classification model based on the language model of chord labels and the generative model of chroma vectors.

Chapter 5 applies the VAE-based regularized training to multi-task learning for joint chord and key estimation. Specifically, key labels are introduced as additional latent variables and a hierarchical generative model of keys, chords, and chroma vectors reflecting the typical process of music composition is formulated. The separated, shared, and hierarchical architectures of the chord and key classification models and the uniform, Markov, and autoregressive architectures of the language model have comprehensively been investigated. Through the comparison of different combinations, It has experimentally been shown that the VAE-based multi-task learning improves chord estimation as well as key estimation.

Chapter 6 concludes this thesis and briefly discusses future work. The main remaining issues of this study are the limited expression capability of chroma vectors and the limited performance improvement obtained by the VAE-based semi-supervised training. Higher-dimensional feature representation learning, sophisticated temporal modeling, language modeling at a beat or score level, and latent variable disentanglement are discussed.

（続紙 ２ ）

（論文審査の結果の要旨）

音楽情報処理分野の基盤技術として、音楽の三大要素（リズム・メロディ・ハーモニー）のうち、ハーモニーと密接に関連するコードを解析する技術が不可欠である。本論文は、音声認識/合成・自然言語処理・画像処理など関連分野における深層学習技術の発展に着想を得て、自動コード推定システムを構成する主な要素である音響特徴量抽出部およびコード分類部に対し、大規模データによる事前学習、生成的・識別的アプローチの統合による半教師あり学習およびマルチタスク学習をそれぞれ導入することにより、コード推定精度の向上に取り組んだ研究をまとめたものである。主な成果は以下の通りである。

1. クロマベクトル抽出器の事前学習：自動コード推定では、コード分類器の入力特徴量として、1オクターブ中の各音高クラスの存在度を表す12次元のクロマベクトルを用いると良いことが経験的に知られている。従来、倍音成分の混入を低減したクロマベクトルを計算するため、古典的な信号処理技術が利用されていた。これに対して、本研究では、大量に入手可能なMIDIデータと、そこから合成した音響信号を用いた一種の自動採譜を事前タスクとして、特徴量抽出器を教師あり学習する方法を考案した。実験の結果、提案法で抽出されたクロマ特徴量を用いることでコード推定精度が向上することを確認した。

2. コード分類器の半教師あり学習：深層ニューラルネットワーク(DNN)に基づく自動コード推定では、音響信号（クロマベクトル）と正解コード系列のペアデータを用いてコード分類器を教師あり学習する必要があったが、大規模な学習データを準備することは困難であり、コード推定精度には限界があった。この問題に対して、コード系列に対してマルコフ言語モデルを仮定したうえで、クロマベクトル系列からコード・潜在特徴系列を推定する分類・認識モデルと、逆にクロマベクトル系列を推定する生成モデルを組み合わせた変分自己符号化器(VAE)を構成することにより、コードアノテーションを持たない音響信号を併用した半教師あり学習を実現した。実験の結果、言語モデルと生成モデルによる分類モデルの正則化により、コード推定精度が向上することを確認した。

3. コード・キー分類器のマルチタスク学習：従来のコード・キー推定においては、両者を独立したタスクとみなすか、コードとキーは対等関係にあると仮定した上で、分類器の教師あり学習が行われていた。これに対して、本研究では、作曲時におけるコードのキーに対する依存関係に着想を得て、クロマベクトル系列からコード系列を推定し、そこからさらにキー系列を推定する分類モデルと、逆に、キー系列からコード系列およびクロマベクトル系列を順に推定する生成モデルを組み合わせたVAEのマルチタスク学習を考案した。実験の結果、コード・キー推定精度がともに向上することを確認した。

以上のように本論文は、音楽要素（コード）と音響信号との相互変換に関する深い洞察と、最新の深層学習の技術動向に関する深い理解に立脚した、生成的・識別的アプローチの統合に基づく複数音楽要素の同時解析法を提示したもので、学術上・実用上寄与するところが少なくない。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和3年8月23日に論文とそれに関連した内容に関する口頭試問を行った結果、合格と認めた。なお、本論文のインターネットでの全文公表について支障がないことを確認した。

要旨公開可能日：　　年　　月　　日以降