E-017

複数音の協和性を表現可能な特徴量を用いたCNN-LSTMコード進行推定法
Harmonic Representation for CNN-LSTM Automatic Chord Estimation

伊藤 威 † Tsuyoshi Ito
荒井 秀一 † Shuichi Arai

1. はじめに

楽曲信号からのコード進行推定 [1] は, 国際音楽情報検索学会[1]が開催するコンテスト:MIREX[2]で取り扱われている重要な課題である. コード (和音) は非常に抽象的な概念であるため, その推定は困難である. 困難さの要因は主に３つあると考えられ, それらを図１のような例を用いて説明する. １つ目は推定区間によっては, 図下部に示す正解とは異なるコードが観測される点である. 図中 (A) 部では正解の Dm ではなくコードFが見える. ２つ目は図中 (B) のように, 正解のコード構成音に無い音 (経過音) が含まれる点である. 最後に図中 (C) のようにコードの部分的構成音のみで演奏されうる点. したがって, コード進行推定は区間分割とその区間でのコードを同時に推定する課題であると言える.

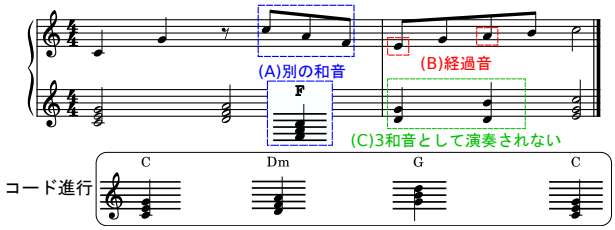

図 1: コード進行推定の難しさ

2. 従来研究

楽曲信号からのコード進行推定は, 特徴抽出過程と推定過程に分けることができる. まず, 楽曲信号から “CQT(Constat-Q-Transform)” などを用いて特徴抽出を行う. その後, “CNN(Convolutional Neural Network)”, “RNN(Recurrent Neural Network)” や “Transformer” などの深層学習の手法で推定するのが近年の主流である [2].

3. 提案

推定モデルが深層学習に置き換わり大きな成果を上げたものの, 特徴量には依然としてCQTが用いられている. しかしながら, CQT は対数周波数領域の特徴量であるため倍音を十分に表現できない. これは, コードの推定において, 構成音の協和を十分に考慮できないことを意味している. なぜなら, 図２のように, 複数構成音の倍音の周波数成分が一致することで初めて協和を表現できるからである. この関係をも明示的に入力特徴量として与えることで, 複数楽器の音高が同時に鳴り響いている状況でも推定が有利に進むと考えられる.

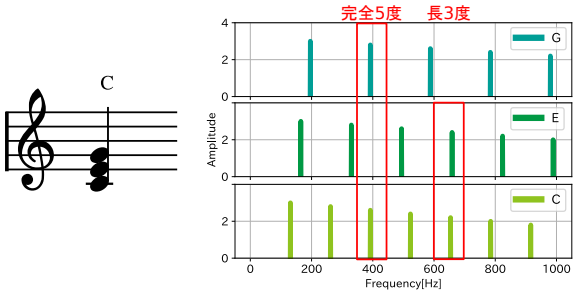

図 2: コードの構成音に含まれる倍音とそれらの関係

そこで本稿では図３に示すように協和性を表現可能な特徴量を用いたコード進行推定法の提案をする.

3.1. 複数音の協和性を表現可能な特徴量

この特徴量は平均律上の各音高に, それらが持つ整数倍音を結合して対数・線形の成分を同時に表現可能にした. 図３中 (A) の楽曲信号より切り出された l 分のサンプルに, 式 (1) の対数変換 (CQT) と線形変換 (FFT) でそれぞれのスペクトルを計算する. この処理で得られた図３中 (B) の対数周波数瓶 f_k の整数倍の周波数成分を, 図３中 (C) の線形スペクトルより計算する. 最後に, 図３中 (D) のように各音高に対する整数倍音の成分を結合して特徴量が完成する. なお, 式 (2)-(5) の具体的な値は 4.1 項で説明する.

$$X^{cqt}[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} w[n,k]x[n] \exp(-j\frac{2Q\pi}{N[k]}n) \quad (1)$$

$$K = b\log_2 \frac{f_{max}}{f_{min}} \quad (2) \qquad f_k = f_{min}(2^{\frac{k}{b}}) \quad (3)$$

$$N[k] = \frac{SR}{f_k}Q \quad (4) \qquad Q = (2^{\frac{1}{b}} - 1)^{-1} \quad (5)$$

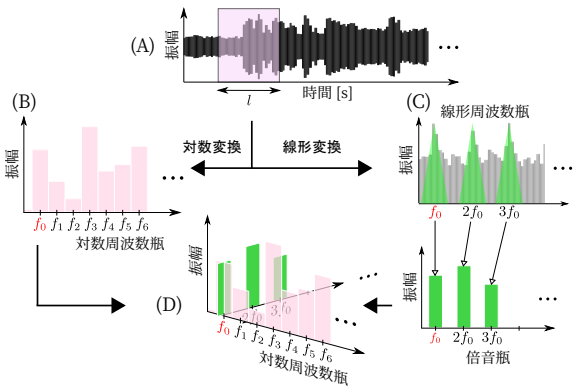

図 3: 特徴抽出の流れ

†東京都市大学 総合理工学研究科
[1]https://ismir.net/
[2]https://www.music-ir.org/mirex/

3.2.CNN-LSTM コード進行推定法

代表的なコードノートである Maj は “完全 5 度” と “長 3 度”, Min は “完全 5 度” と “短 3 度” の音程の組み合わせである．これらが図 4 中の第 1 小節のように, 和音として演奏されていれば局所的な区間だけでコードの推定が可能だが, 第 2-3 小節のようにコードの構成音が分解されて演奏された場合は, そのコードノートであると断定出来る音程を観測するまではコードが推定できない．そこで, 局所的な周波数方向に音高・音程を CNN で抽出し, それらを LSTM(Long Short Term Memory) を用いて周辺のコードノートの素片を用いた推定を目的としたネットワークを構築する．

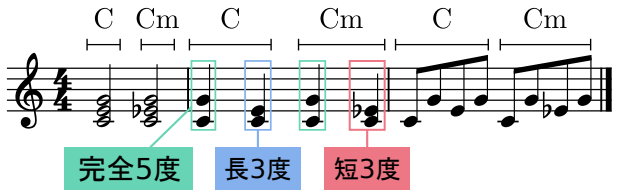

図 4: Maj と Min の違い

図 3 中 (D) の各フレームのスペクトルをネットワークへの入力とする．このとき倍音方向の情報は画像の RGB レイヤーのように積み重なっており, これをチャンネル方向に入力する．CNN への入力チャンネルが 6 である理由は, Min に含まれる “短 3 度” の音程が, 基音の 5 倍音と 6 倍音に現れるからである．先頭の畳み込み層のカーネルサイズ (17, 1) は, Maj や Min などの 3 和音 (トライアド) の音程に対応する周波数帯域のフィルタリングを目的としている．

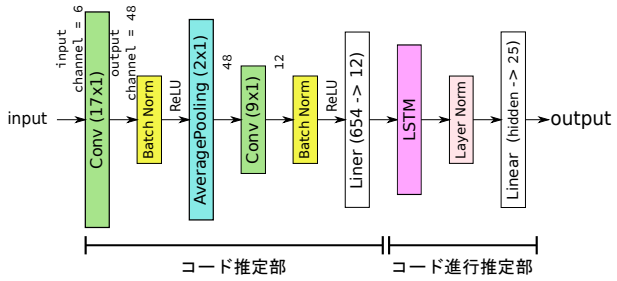

図 5: 提案するネットワーク

4. 実験及び結果

本稿では楽曲信号のある区間が, 各基音の “Maj” と “Min”, “ノーコード” の計 25 種類のコードノートのいずれかであるかを推定し, WCSR(Weighted Chord Symbol Recall) で評価する．実験で用いた楽曲は “the Beatles”, “Queen”, “Robbie Williams”, “RWC Pop Data”, “Zweieck” からの計 383 曲である．コード進行推定用のアノテーションデータは Isophonic[3]等で公開されているものを用いた．結果を比較するため、ベースラインとなる手法は同一条件下で再度学習し評価した．

4.1. 各種パラメータ

入力特徴量はサンプリングレート 22,050Hz の楽曲信号を用い, サンプル長 8192, フレーム周期 1024 で分析した．加えて, CQT の計算に用いる変数は $f_{min} = 32.703$(音名:C1), $f_{max} = 2093.005$(音名:C7), $SR = 22,050$, $b = 24$ と設定し, 整数倍音は 6 倍音まで結合した．ネットワーク学習時の最適化手法は Adagrad を用いた．

4.2. 結果

表 1: ベースラインとの性能比較

model	WSCR	#Params (M)
CNN_{2016}[3]	76.7	0.97
CRNN_{2017}[4]	77.0	0.44
BTC_{2019}[2]	77.5	3.18
Ours_{2021}	**78.8**	**0.19**

表 1 より, 本稿で提案した手法が従来研究よりも高い性能を示し, それに加えて訓練すべきネットワークパラメータ数も少ないことが分かった．一般的に, アーキテクチャの規模を大きくすると汎化能力が向上するとされているが, そのためにはパラメータを十分に更新するだけのデータが必要である．少ないパラメータで性能が維持できることは、データの増強が困難なコード進行推定において有用である．

5. おわりに

複数音の協和性を表現可能な特徴量を用いた CNN-LSTM ベースのネットワークによるコード進行推定を行った結果, 従来手法を上回る性能を少ないパラメータで実現できた．これにより, 提案した特徴量の基本的な有効性が確認できた．

参考文献

[1] Johan Pauwels, Ken O’hanlon, Emilia Gómez, and Mark B Sandler. 20 Years of Automatic Chord Recognition From Audio. *The International Society for Music Information Retrieval (ISMIR) 2019.*

[2] J Park, K Choi, S Jeon, Dokyun Kim, and Jonghun Park. A Bi-directional Transformer for Musical Chord Recognition. *ISMIR 2019.*

[3] F. Korzeniowski and G. Widmer. A fully convolutional deep auditory model for musical chord recognition. *ISMIR 2016.*

[4] B. McFee and J.Bello. Structured training for large-vocabulary chord recognition. *ISMIR 2017.*

[3] http://isophonics.net/datasets