Vol.2017-MUS-115 No.33
2017/6/17

Encoder-decoderモデルと Stacked bidirectional LSTMに基づく 和声解析の検討

保利 武志[1,a)]　中村 和幸[1,b)]　嵯峨山 茂樹[1,c)]

概要：音楽の生成過程において和声進行は重要な役割を果たしており，楽曲解析や自動作曲・編曲システム，多重音解析，音楽情報検索など様々な分野においてそのモデル化の研究が盛んに行われている．従来の研究ではルールベースに，あるいは音高パタンを観測系列とした Hidden Markov Model (HMM) や Probabilistic Context-free Grammar (PCFG) などの確率モデルによる階層的なモデル化が行われてきたが，近年では Deep learning を用いて直接的に学習，推定する手法も検討されており，特に時系列データを扱う Long Short-term Memory (LSTM) を用いた手法が注目され始めている．本研究では事前処理として定 Q 変換を施したパワースペクトルに対し，Specmurt 分析を用いて倍音成分の抑制及び基本周波数を強調し，クロマベクトルへと変換した特徴量系列を入力とした LSTM ベースの Deep Neural Network による学習を行う．特に，系列データを前向きと後ろ向きの双方向に学習する bidirectional LSTM に基づき，これに翻訳モデルとして用いられる Encoder-decoder LSTM へと拡張した RNN など複数の RNN による和声の推定を行う．各 RNN によるコード推定精度の比較実験を行った結果，bidirectional な構造と Encoder-decoder モデルによる和声認識への有効性が示唆された．

Investigation of Chord Analysis : Based on Encoder-decoder Model and Stacked Bidirectional LSTM

Hori Takeshi[1,a)]　Nakamura Kazuyuki[1,b)]　Sagayama Shigeki[1,c)]

1. はじめに

和声は特に西洋音楽においてその基盤ともなる重要な要素である。近年では芸術音楽の変革に伴い，調性のない無調音楽やフリージャズなども盛んとなり，従来の調に支配された楽曲構成から離れた多種多様な構造，旋法による作曲も数多く行われているが，現在でも広く知られている音楽や我々がよく耳にするポピュラー音楽など，大多数の音楽は未だ調性による主従構造を持つものが多く，和声の解析は楽曲構造や生成過程などを解析する上で重要な意義を持つ．

和声構造を想定した確率モデルを導入した研究としては，旋律に対する自動和声付け [1] や自動編曲 [2] など直接的に作曲に関わるものから，近年の電子媒体による楽曲検索の需要等から飛躍的に進歩を遂げている音楽情報検索 [3]，また多重音解析や音源分離 [4], [5]，自動採譜 [6] など実に多岐に渡るものが挙げられる．これらは，音楽音響信号や MIDI 信号，楽譜等を入力として，調やコードなどによる階層構造を持つ音楽の生成モデルに対する逆問題として定式化されることが多い．

このような和声認識のモデルは音声認識における言語モデルに相当するものとしてモデル化されている．言語モデルとは例えば”私”の次は”は”が続きやすいといった単語の連接を N-gram 等を用いて確率的に記述したもので，和

[1] 明治大学
Meiji University, Nakano, Tokyo 164–8525, Japan
[a)] cs51003@meiji.ac.jp
[b)] knaka@meiji.ac.jp
[c)] sagayama@meiji.ac.jp

声においても”I-IV-V-I”という進行に代表されるように，コード間の連接に偏りを持つことがよく知られており，言語モデルと近似したモデルが適用できる．音声認識分野ではこれまで，観測される音響特徴量系列 (メル周波数ケプストラム係数 (MFCC) などが主によく用いられる) に対し，言語モデルと音響モデル (音素単位の left to right Hidden Markov Model (HMM) が用いられることが多い) の 2 つの確率モデルによる尤度最大化問題として定式化することで，最も尤もらしい単語 (列) の探索を行う他，近年では HMM と Deep Neural Network (DNN) を組み合わせた DNN-HMM による音響モデルによる推定精度向上の研究がなされており [7], [8]，また，HMM を用いない Long Short-Term Memory Reccurent Neural Network (LSTM, LSTM-RNN) による End to End な学習による認識の研究なども盛んに行われ，飛躍的な精度の向上へと繋がっている．LSTM は系列データの深層学習手法として代表的なニューラルネットワークであり，前後の文脈によって語彙が制限されるような言語モデルや，同様に HMM でモデル化されてきた時間的な方向性をもつ音響モデルに対してもよく合致しており，音声認識において LSTM が用いられ始めたのは昨今のニューラルネットワーク研究の隆盛を鑑みれば自然な潮流であると言える．

和声認識もまた言語モデルと同型の問題として捉えることが可能であるならば，これまで HMM[9] や Probabilistic Context-free Grammar (PCFG) [10] 等でモデル化されてきた音楽モデル に対して，LSTM をはじめとした Deep learning の手法を適用することで精度が向上することが期待できる．

本研究では，WAVE 形式の音源を入力として和声を推定する LSTM-based な複数の DNN について比較評価し，和声モデルのためのネットワーク構造を検討する．

2. 和声認識

人間は訓練を積むことで旋律 (音高情報) を元に調やコードを認知することが可能であるが，これは演奏されている音楽の旋律に対して音の響き (高さ) を感じると同時に，音楽的知識に基づいた推定を行うことにより実現していると考えられる．同様にして，観測として得られる音楽音響信号を元に機械が音高を推定するためには，観測信号の基本周波数成分を抽出し可能な限り正確な音高情報を得ると同時に，ある観測信号系列に対応するコードや，コードの連接しやすさ等を学習する必要がある．

本研究では図 1 に示すように，録音 CD 等の WAVE 音源による音楽音響信号を入力として和声の推定を行う．音高が明確な MIDI 信号とは異なり，楽器演奏による音楽音響信号には倍音成分が多く含まれることから基本周波数を解析・推定することは難しい．そこで，本研究では Specmurt 分析に基づく倍音成分の抑制及び基本周波数の強調を行う．また，コードの推定において多くの場合，オクターブに跨る同一ピッチクラスの情報 (例えば A4:440.0Hz と A5:880.0Hz など) は等価な情報としてみなせるため，これをさらにクロマベクトル化した特徴量系列を入力として用いる．

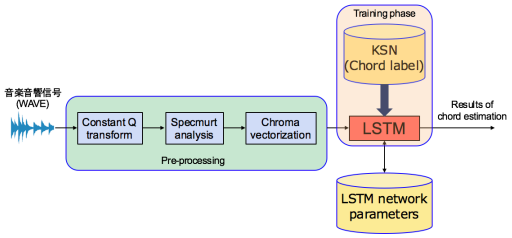

図 1 本研究におけるブロックダイアグラム．音楽音響信号に対し Specmurt 分析とクロマベクトル化を事前処理として用いた特徴量系列を入力とし，LSTM で学習した後．得られたネットワークパラメータを用いてコードを推定する．

2.1 Specmurt 分析

Specmurt 分析は調波構造を持つ楽音の短時間パワースペクトルに対する多重ピッチ解析手法として，高橋らにより提案された [11]．Specmurt 分析では，調波構造の倍音パワー比について，基本周波数に依らず共通であるとする仮定をおくことにより，通常の線形周波数領域において複数のスペクトルは周波数方向に線形伸縮の関係であるのに対し，対数周波数領域に変換することによって線形シフトな関係となることを利用する．これにより，対数周波数 x の領域における入力インパルス関数 (理想的には，多重音に含まれる全ての基本周波数のみにピークが立つようなインパルス関数) を $u(x)$，共通の調波構造を $h(x)$ とすると，観測されるパワースペクトル $v(x)$ はこれらの畳み込み，

$$v(x) = u(x) * h(x) \tag{1}$$

として記述できる (図 2)．従って，$v(x), u(x), h(x)$ の (逆) フーリエ変換を $V(y), U(y), H(y)$ とおくと，

$$U(y) = \frac{V(y)}{H(y)} \tag{2}$$

となるので，

$$u(x) = FT\left[\frac{V(y)}{H(y)}\right] \tag{3}$$

とすることで，観測スペクトル (と既知とした共通調波構造) から基本周波数分布 $u(x)$ を求めることができる．なお，$FT[\cdot]$ は (逆) フーリエ変換を表す．

ここで，Specmurt 分析は周波数軸を対数変換するが，一般に窓関数を通した短時間フーリエ変換 (STFT) により得られるスペクトルは線スペクトルではなく幅を持つ．この広がりは周波数に依らず一定であるため，そのまま対数周波数領域へと置換すると周波数に反比例して狭くなり、

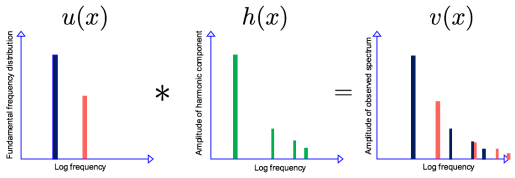

図 2 Specmurt 分析の概略図．共通調波構造が既知であれば，基本周波数分布を推定できる

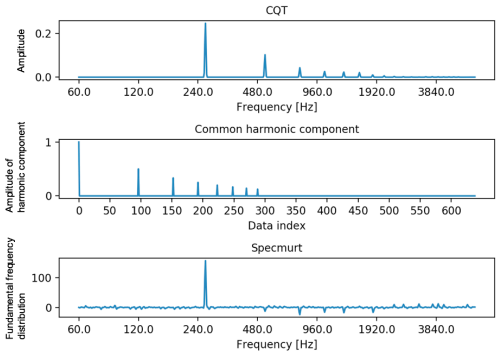

図 3 ピアノの単音 (C4) のスペクトルに対する，1/f の減衰特性倍音比を持つ共通調波構造及び Specmurt 分析結果．

表 1 Specmurt 分析の実験条件

フレームシフト	10ms
解析最低周波数	60Hz
解析最高周波数	6000Hz
周波数分解能	12.5cent
考慮する倍音数	8

畳み込み演算が厳密には成り立たない．そのため，通常はウェーブレット変換や定 Q 変換 (CQT)[12] を用いることによって，対数周波数領域でも広がりが一定となるような処理を施す。本研究においては，定 Q 変換を行った．図 3 に RWC 研究用音楽データベース [13] の楽器音データベースを基に作成したピアノの C4 単音に対する Specmurt 分析結果を示す．共通調波構造パタンは周波数のベキ乗 $f^{-1.0}$ の減衰特性倍音比を持つモデルを仮定した．また，実験条件を表 1 に示す．なお，周波数分解能 12.5cent とは 1 オクターブを周波数 bin96 点で分割するのに等しい．

図 3 が示すように，Specmurt 分析を行うことによって，基本周波数 (261.6Hz) の倍音成分 (523.2Hz 等) のパワーが抑制され，基本周波数が強調されているのがわかる．

2.2 Chrome vector(クロマベクトル)

コードの推定においては，C4 や C5，D4 などの詳細な音高情報ではなく，C や D といったピッチクラスのパワーが主要な手がかりとなる． クロマベクトル [14], [15] はパワースペクトルを半音ごとの 12 次元のピッチクラスに割り当てたもので，ピッチクラスを k，時刻 (フレーム) を t とすると，クロマベクトル $c(k,t)$ は，

$$c(k,t) = \sum_{i=0}^{I-1} H(12i+k,t), \quad k = 0, \cdots, 11 \tag{4}$$

で与えられる．ただし，$H(i,t)$ はスペクトログラムの周波数ビン i と時刻 t におけるパワー，I は取得するオクターブ数を表す．本研究では Specmurt 分析によって得られる基本周波数分布に対して，負の値に対して 0 にクリッピングした上でクロマベクトル化を行う．また，得られたクロマベクトルは正規化処理を施すことによって時間フレームごとの偏りを緩和する．

3. LSTM による学習

前章で述べたように，事前処理として，CQT で得られた対数周波数軸におけるスペクトログラムに対して Specmurt 分析を行い倍音成分抑制及び基本周波数を強調した上で，正規化したクロマベクトルの時系列データを入力とする．また，入力系列との時間整合を行ったコード進行系列を正解ラベルとし，LSTM によって学習を行う．

近年では LSTM ユニットを複数重ねた stacked LSTM や，入力の系列データを前向きと後ろ向きの両方向に重ねた bidirectional LSTM[16] による認識精度の向上が報告されており，本研究でも bidirectional LSTM をベースとした複数の LSTM による推定精度の比較を行う．

また，通常の LSTM の構造上，入力系列の時系列は明示的であるのに対し，出力の時系列は入力系列に対する結果として得られるものであり，明示的に教師データとして反映されてはいない．これに対し，日本語と英語の翻訳モデルなどに用いられる Encoder-decoder LSTM (ED-LSTM)[17] では，入力系列長と出力系列長が異なる場合にも対応できると同時に，出力側の時系列 (英語の文章など) も同時に学習する必要があることから，入力系列のみではなく出力系列の時系列情報も Decoder 部において明示的に与えられるネットワークである．

本研究では上記の bidirectional LSTM に加え，ED-LSTM に基づき，Encoder 部に bidirectional LSTM を組み合わせた stacked bidirectional ED-LSTM(図 4) と，Encoder 部の bidirectional LSTM によって推定した出力結果を Decoder 部の入力として bidirectional に学習する stacked bidirectional encoder bidirectional decoder LSTM (biED-biD-LSTM)(図 5) を提案し，その評価を行う．

4. 実験

学習には RWC 研究用音楽データベースのクラシック音楽データベースの主調が長調である器楽曲から 8 曲を入力データとし，我々が作成した KSN 和声ラベルデータを正解ラベルとして用いた [18]．本データベースの和声進行は，高度な和声学を修めた音楽大学学生や専門家により，我々が提案する KS 表記法 (KS notation : KSN) で記述された

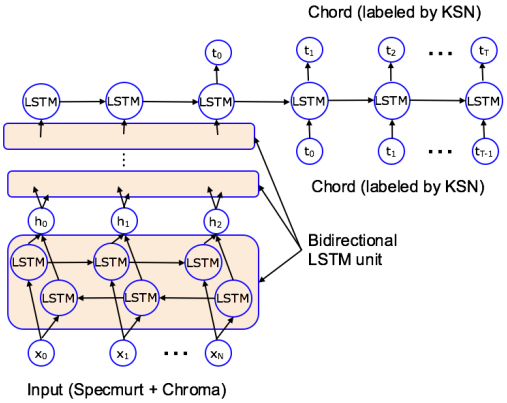

図 4 stacked bidirectional ED-LSTM．Encoder 部を bidirectional にネットワーク化し，複数層を積み重ねた構造を持つ．

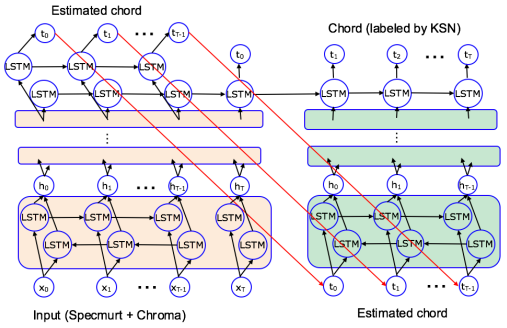

図 5 stacked biE-biD-LSTM．Encoder 部で推定したコードを Decoder 部の入力として用いることで，Decoder 部にも bidirectional 構造が適用可能．

表 2 実験条件

サンプリング周波数	16kHz
フレームシフト	10ms
解析最低周波数	60Hz
解析最高周波数	8000Hz
周波数分解能	12.5cent
共通調波構造倍音比	$f^{-1.5}$
考慮する倍音数	10
テンポ	120bpm
調	C-Majar
学習用楽曲数	8
積層数	3
学習回数	10000

ものである．本稿では V 度のみセブンスを区別し，また，借用和音等も KSN の表記に従い区別する．これは，本モデルによってコードの推定と同時に転調を識別可能か検討するためである．学習データ内に現れたコード数は 67 であった．

また，学習データを効率的に用いるために，上記データベースの MIDI データを C-Majar へと移調処理を施し，WAVE 変換を行ったデータに対して解析を行った．音響解析や学習に用いたパラメータは表 2 に示す通りである．

学習は 2 小節ごとにデータを分割して行い，学習に使用しなかった他の器楽曲データ 2 曲を用いて推定精度を算出した．評価実験結果を表 3 に示す．Encoder-decoder モデルを用いた (2)，(3) の手法の方が高い認識率を記録したが，これは入力の特徴量系列に加えて記号的なコードの時系列情報を学習に加えたためであると思われる．

表 3 実験結果

	(1)biLSTM	(2)biEDLSTM	(3)biEbiDLSTM
コード認識率	35.2	40.3	41.8

一方で全体的に先行研究 [9] と比較して認識率が低くなったが，これはフレームシフトを細かくとったことで入力の系列長が長くなり学習が進みづらいことが挙げられる．また，今回使用したクラシック曲は転調 (借用和音) も多く，その正解ラベルの数が与える学習や推論への影響を考慮すれば，学習データの充実を図ることで精度の向上が期待できる．

5. おわりに

本研究では音楽音響信号から和声を推定するための LSTM 及び解析手法について述べた．事前処理として Specmurt 解析による倍音抑制・基本周波数強調とクロマベクトル化を行い，3 種の LSTM で学習して比較評価した結果，bidirectional な構造や Encoder-decoder モデルがコードの認識に有効であることが示唆された．今後は，今回事前処理として用いた Specmurt だけではなく，NMF などを用いた処理も検討し，その有効性を検討したい．また，和声の認識をモデルとして併用した自動作曲や編曲，多重音解析，自動採譜など様々な応用に取り組んで行きたい。

参考文献

[1] 川上隆，中井満，下平博，嵯峨山茂樹ほか：隠れマルコフモデルを用いた旋律への自動和声付け，情報処理学会研究報告音楽情報科学 (MUS)， Vol. 2000, No. 19 (1999-MUS-034), pp. 59–66 (2000).

[2] Hori, G., Kameoka, H. and Sagayama, S.: Input-output HMM applied to automatic arrangement for guitars, *Information and Media Technologies*, Vol. 8, No. 2, pp. 477–484 (2013).

[3] Schedl, M., Gómez, E., Urbano, J. et al.: Music information retrieval: Recent developments and applications, *Foundations and Trends® in Information Retrieval*, Vol. 8, No. 2-3, pp. 127–261 (2014).

[4] Peeters, G.: Chroma-based estimation of musical key from audio-signal analysis., *ISMIR*, pp. 115–120 (2006).

[5] Mauch, M. and Dixon, S.: Approximate Note Transcription for the Improved Identification of Difficult Chords., *ISMIR*, pp. 135–140 (2010).

[6] Sigtia, S., Benetos, E. and Dixon, S.: An end-to-end neural network for polyphonic piano music transcription, *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, Vol. 24, No. 5, pp. 927–939 (2016).

[7] 河原達也ほか：音声認識の方法論に関する考察—世代交代に向けて—，研究報告音声言語情報処理 (SLP)，Vol. 2014, No. 3, pp. 1–5 (2014).

[8] Dahl, G. E., Yu, D., Deng, L. and Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 1, pp. 30–42 (2012).
[9] 丸尾智志，吉井和佳，糸山克寿，後藤真孝ほか：コード制約付き NMF を用いた音高推定に基づくコード認識，第 77 回全国大会講演論文集， Vol. 2015, No. 1, pp. 421–422 (2015).
[10] Kameoka, H., Ochiai, K., Nakano, M., Tsuchiya, M. and Sagayama, S.: Context-free 2D Tree Structure Model of Musical Notes for Bayesian Modeling of Polyphonic Spectrograms., *ISMIR*, Vol. 2012, pp. 307–312 (2012).
[11] 高橋佳吾，西本卓也，嵯峨山茂樹ほか：対数周波数逆畳み込みによる多重音の基本周波数解析，情報処理学会研究報告音楽情報科学 (MUS)， Vol. 2003, No. 127 (2003-MUS-053), pp. 61–66 (2003).
[12] Brown, J. C.: Calculation of a constant Q spectral transform, *The Journal of the Acoustical Society of America*, Vol. 89, No. 1, pp. 425–434 (1991).
[13] 後藤真孝，橋口博樹，西村拓一， 岡隆一ほか：RWC 研究用音楽データベース: 研究目的で利用可能な著作権処理済み楽曲・楽器音データベース，情報処理学会論文誌，Vol. 45, No. 3, pp. 728–738 (2004).
[14] Fujishima, T.: Realtime Chord Recognition of Musical Sound: a System Using Common Lisp Music., *ICMC*, pp. 464–467 (1999).
[15] Bartsch, M. A. and Wakefield, G. H.: To catch a chorus: Using chroma-based representations for audio thumbnailing, *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, IEEE, pp. 15–18 (2001).
[16] Graves, A. and Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures, *Neural Networks*, Vol. 18, No. 5, pp. 602–610 (2005).
[17] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to sequence learning with neural networks, *Advances in neural information processing systems*, pp. 3104–3112 (2014).
[18] Kaneko, H., Kawakami, D. and Sagayama, S.: Functional harmony annotation database for statistical music analysis, *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR): Late Breaking session* (2010).