3Q-03

ニューラルネットワークを用いた和音の教師なし学習の可能性の検討

石塚　匠†　　櫻井彰人‡

慶應義塾大学大学院理工学研究科†　　慶應義塾大学理工学部‡

1. はじめに

音楽音響信号に対する自動和音認識は音楽情報処理における主要なタスクの一つである．認識結果を直接的に楽譜として利用する目的のほか，楽曲推薦のための特徴量として用いるなど様々な応用が存在することがその理由である．

統計的機械学習的なアプローチによって和音認識を行うためには，大量のデータセットが必要である．しかし正解ラベル（コードネーム等）の付与には高度な専門的知識と多大な手間が必要である上に，楽曲の権利の処理が難しいため，手軽に入手可能な正解ラベル付きデータセットの量およびジャンルが限られている．

2. 提案手法の全体像

本稿では，正解ラベルの無いデータセットすなわち楽曲の音響信号のみを用いて，和音に関する何らかの中間表現を機械学習によって得る方法を提案する．なお，具体的にコードネームを得ることは目的としていない．

まず，楽曲のスペクトログラムの各時刻におけるスペクトルを入力としてオートエンコーダ(auto-encoder)[1]を学習させる．次に，このオートエンコーダによって得られた中間層の値を入力として別のオートエンコーダを学習させる（積層オートエンコーダ）．最後に，積層オートエンコーダの最も上層のニューロンの振る舞いを観察する．

スペクトログラムへの変換法およびオートエンコーダの更新アルゴリズムは，独自に作成した．3節で詳細を述べる．

3. 提案手法の詳細

3.1. ERB周波数スペクトログラムへの変換

一般的に音響信号波形からスペクトログラムへの変換にはSTFTあるいは定Q変換が用いられるが，得られるスペクトルの周波数解像度がヒトの聴覚と異なるという問題点がある（STFTは高音域の解像度が高すぎ，定Q変換は高音域の解像度が高すぎる）．

これを解決するべく，本稿では複素ガボールフィルタで構成されたフィルタバンクを用いた．複素ガボールフィルタは正弦波とガウス窓の積で表され，次の式により定義される．

$$g_s(t) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{t^2}{2\sigma^2}\right)\exp(2\pi i f_0 t)$$

ここでf_0は中心周波数，σはガウス窓の標準偏差である．f_0とσは等価矩形帯域幅(ERB)[2]とERB-rateを用いて決定した．更に，フィルタの出力の絶対値の理論上の最大値が1になるように，フィルタの係数を定数倍した．

このようにして作成されたフィルタバンクには以下に挙げる利点がある．

- ERBを用いているためヒトの聴覚に近い周波数解像度（Q値）のスペクトルが得られる．
- 複素フィルタであるため，聴覚を模倣する代表的なフィルタであるガンマトーンフィルタ等と違い位相のずれに強い．
- ガウス窓を用いているため，時間-周波数の不確定性のトレードオフが優れている．
- 出力値が区間$[0,1]$に入るため，シグモイド関数を活性化関数とするオートエンコーダへそのまま（スケーリング等をせずに）入力する事ができる．

このフィルタバンクを用いてステレオ信号をチャンネル毎にパワースペクトログラム（周波数ビンの個数は166）へ変換し，その和をとることでモノラルのスペクトログラムとした．

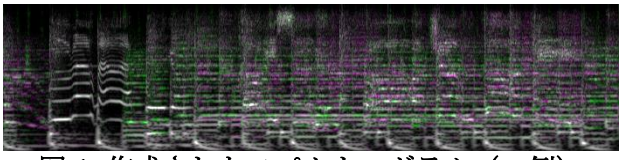

図1 作成されたスペクトログラム（一例）

3.2. Iダイバージェンス規準オートエンコーダ

オートエンコーダは，ニューラルネットワークを使用した次元削減法であり，3層ニューラルネットワークにおいて出力層と入力層を再現するように学習させる．

オートエンコーダの目的関数をLとすると

$$L(W_1, W_2, \boldsymbol{b}, \boldsymbol{c}) = \sum_i D\left(\boldsymbol{v}_0, g\left(W_2 f\left(W_1 \boldsymbol{v}_0^{(i)} + \boldsymbol{c}\right) + \boldsymbol{b}\right)\right)$$

である．ここで，$\boldsymbol{v}_0$は入力ベクトル，$W_1, W_2, \boldsymbol{b}, \boldsymbol{c}$

A study of unsupervised learning for music chords using neural networks
†Takumi Ishizuka (Keio University Graduate School)
‡Akito Sakurai (Keio University)

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.

は学習可能なパラメータである．

本稿では非線形関数f, gにはシグモイド関数を，誤差関数$D(x, y)$にはIダイバージェンス[3]を用いた．また，$W_2^{\mathrm{T}} = W_1 := W$とした(tied-weight)．このとき，確率的勾配降下法によるパラメータ更新式を近似的に非常に簡単な形で記述することができる．

$$
\begin{aligned}
&\boldsymbol{h}_0 = \sigma(\boldsymbol{c} + W\boldsymbol{v}_0) \\
&\boldsymbol{v}_1 = \sigma(\boldsymbol{b} + W^{\mathrm{T}}\boldsymbol{h}_0) \\
&\boldsymbol{h}_1 = \sigma(\boldsymbol{c} + W\boldsymbol{v}_1) \\
&W \leftarrow W + \varepsilon\left(\boldsymbol{h}_0\boldsymbol{v}_0^{\mathrm{T}} - \boldsymbol{h}_1\boldsymbol{v}_1^{\mathrm{T}}\right) \\
&\boldsymbol{b} \leftarrow \boldsymbol{b} + \varepsilon(\boldsymbol{v}_0 - \boldsymbol{v}_1) \\
&\boldsymbol{c} \leftarrow \boldsymbol{c} + \varepsilon(\boldsymbol{h}_0 - \boldsymbol{h}_1)
\end{aligned}
$$

図 2　Iダイバージェンスオートエンコーダのパラメータ更新式

この更新式（図 2）は，平均場近似を用いたRBM（制限付きボルツマンマシン）の更新式[1]と同一になっている．紙面の都合上詳細な導出過程は省くが，以下に概略を述べる．

$\boldsymbol{v}_0 \ll 1$のとき，シグモイドクロスエントロピー誤差はIダイバージェンスを近似する．また，ある程度学習が進んだ場合において，シグモイドクロスエントロピー誤差を誤差関数とするオートエンコーダの更新式と平均場近似を用いたRBMの更新式は，微小項を除いて等しい．ゆえに，Iダイバージェンスを誤差関数とするオートエンコーダの更新式を，平均場近似を用いたRBMの更新式で代用することができる．

本稿では上記の更新式を用いるオートエンコーダを2段に積み重ね，166→60→40と次元削減を行った．

4.　データセットおよび学習

学習用のデータセットには，J-POPを10曲用いた．スペクトログラムのフレームシフトは23msで，フレーム幅は46msである．

学習に用いたマシンはLet's note CF-S10で，第1層と第2層を合わせて1日程度計算した．

5.　評価

5.1.　評価条件

MIDI音源のピアノ音色を用いて，和音およびそれを構成する単音を含む音響信号を作成し，評価用データとした．その楽譜とスペクトログラムを図4に示す．

この評価用データを学習済みの積層オートエンコーダに入力した場合における，最も上層の40個のニューロンの振る舞いを観察した．

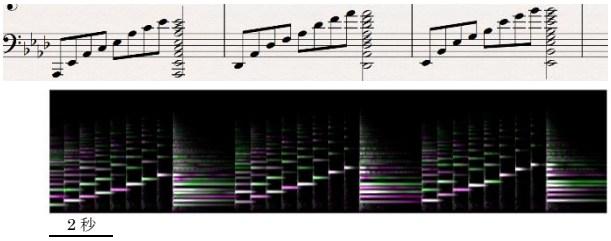

図 4 評価に用いた音響信号

5.2.　結果・考察

最も上層の40個のニューロンの中で，和音に対して選択的に活性化するニューロンが存在することが確認された．特に顕著な3つのニューロンのアクティベーションを図3に示す．

図 3 ニューロンのアクティベーション

横軸は時刻，縦軸は個々のニューロン．白はニューロンが強く活性化していることを表す．

図3の青枠と赤枠の部分を比較することにより，これらのニューロンは単音に対しては強く反応せず，かつ，和音に対しては強く反応していることがわかる．ゆえに，和音に関する何らかの中間表現をオートエンコーダが教師なし学習により獲得した可能性を示していると考えられる．

6.　おわりに

本稿ではERB周波数スペクトログラムおよびIダイバージェンス規準オートエンコーダの簡潔な更新式を提案し，それを用いて和音に関する何らかの中間表現を獲得する方法を提案した．今後は，これが真に教師なし学習の結果であるのか否かを確かめる予定である．例えば，和音を含まないデータセットや異なるジャンルのデータセットを用いた場合と比較することなどが考えられる．

参考文献

[1] Bengio, Yoshua. "Learning deep architectures for AI." *Foundations and trends® in Machine Learning* 2.1 (2009): 1-127.

[2] Moore, Brian CJ, and Brian R. Glasberg. "Suggested formulae for calculating auditory - filter bandwidths and excitation patterns." *The Journal of the Acoustical Society of America* 74.3 (1983): 750-753.

[3] Lee, Daniel D., and H. Sebastian Seung. "Algorithms for non-negative matrix factorization." *Advances in neural information processing systems*. 2001.

Copyright ©2016 Information Processing Society of Japan.
All Rights Reserved.