3N-04

変分オートエンコーダを用いた多重音解析システムの性能評価

森口　寛生†　大村　英史†　桂田　浩一†

東京理科大学　理工学部情報科学科†

1. はじめに

複数の楽器音が発音している中で個々の音源を抽出する処理を多重音解析[1]という．多重音解析には，入力データを頻出パターンとその重みの 2 つに分解することができる NMF（Non-Negative Matrix Factorization）[2]が広く用いられている．NMF を用いた多重音解析では，各楽器の音高に対応したスペクトルを頻出パターンとして与え，重みが最大となる音高を解析結果として得る．しかし NMF は一つのスペクトルパターンのみを教師データとして用いるため解析性能は十分とは言えない．そこで本発表では DVAE (Denoising Vatiational AutoEncoder) [3]を用いた多重音解析システムを提案する．DVAE を用いることで楽器別の音高の生成モデルを分布として学習することができることから， NMF と比べて良好な解析結果が得られると考えられる．本稿では DVAE を用いた多重音解析を，NMF を始めとする多重音解析法と比較する．

2. DVAE を用いた多重音解析システム

本研究では多重音解析システムを 2 ステージで構築した．1 ステージ目で DVAE を用いて音響特徴を抽出し，2 ステージ目で多層ニューラルネットワーク（以下多層 NN）を用いて音高を同定する．音響特徴の抽出と音高判定という異なる問題を 2 ステージに分割することで，それぞれの問題を効率的に解くことができる．

2.1. DVAE

AutoEncoder はニューラルネットワークの一種で，元データ x から潜在変数 z へ符号化した上で z から x へ複号化する恒等変換機を学習することができる．VAE は AutoEncoder の潜在変数 z に正規分布(=生成モデル)の仮定をすることで z により有意な情報を与えたものである．

VAE は一般的に図 1 のように構築される．第 1 層への入力は第 2 層で符号 1 に符号化される．第 3 層では符号 1 を生成するための生成モデルの平均と分散がモデル化される．第 4 層には第 3 層の生成モデルに基づいてサンプリングしたデータが与えられ，第 5 層で再び符号 1 に複号した後に

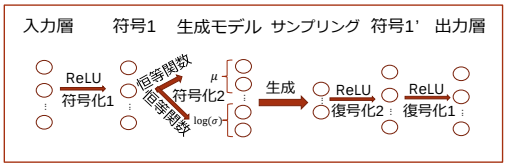

図 1:VAE の各層の概略

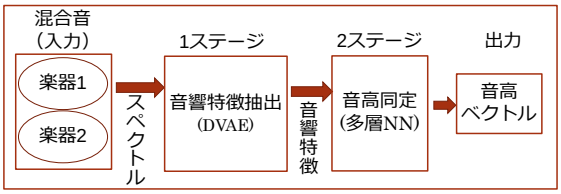

図 2:提案システムの構成概要図

第 6 層で第 1 層のデータを複号した値が出力される．第 1 層にノイズを付加したデータを与え，第 6 層に与える教師データをノイズのないデータにすることにより，ノイズを除去可能な DVAE として学習することも可能である．

2.2. システム概要

本システムでは 2 つの楽器音からなる混合音から一方の楽器音の音高を判定する．システムの概要を図 2 に示す．図に示すように第 1 ステージへの入力は楽器音 1，楽器音 2 の混合音のスペクトルである．第 2 ステージへの入力は図 1 の 4 層にあたる生成モデルをサンプリングしたものを用いる．第 2 ステージの出力は楽器音 1 の音高を one-hot ベクトルで表したものである．

3. 実験と評価

提案手法の性能評価のため 3 つの実験を行う．はじめに提案手法と NMF との比較を行う．次に第 2 ステージの入力に DVAE の第 4 層以外の層を用いた方法と提案手法とを比較する．最後に DVAE 以外の様々なオートエンコーダを用いた場合と提案手法とを比較する．

3.1. 実験の概要

本実験では解析対象の楽器音として“RWC 研究用音楽データベース：楽器音”を用いた．すべての楽器音はサンプリング周波数 44100Hz で収録されている．楽器 1，楽器 2 共にそれぞれの単音の定常部分を用意し，それを楽器 2 が鳴らな

Evaluation of Multipitch Analysis Using Variational AutoEncoder
†Hiroki Moriguchi, Hidefumi Ohmura, Kouichi Katsurada (Tokyo University of science)

表 1:NMF との比較

	NMF	提案手法
認識率(P-B)	96.9	94.0
認識率(P-C)	63.3	91.6

い時も含めてすべての組み合わせでデータセットを作成する．（例えば楽器 1 が 88 音，楽器 2 が 40 音なら 88×(40+1)=3608 通り）その 7 割を学習に用いて 3 割を検証データとした．
システムへの入力は定常部分を 4096 点(約 0.093 秒)で FFT して鏡部分を切り捨てた 2049 次元のものを使う．DVAE に用いる生成モデルの個数を楽器 1 と楽器 2 の音域数の合計の 4 倍とした（楽器 1 が 88 音で楽器 2 が 40 音ならば((88+40)×4＝512）．第 3 層では生成モデルの平均と分散がそれぞれ学習されるため，第 3 層のノード数はこの 2 倍(1024)となり，で第 4 層のノード数が生成モデルの個数(512)となる．第 3 層の生成モデルから第 4 層のデータを生成する際には，1 つのデータに対して 1 つサンプリングしたものを用いた．第 2 層と第 5 層のノード数は 1000 とした．学習回数は 500 エポックでとした．

第 2 ステージの多層 NN は 4 層で構築した．入力のノード数は第 1 ステージの第 4 層のノード数で，出力のノード数は楽器 1 の音高数である．中間層のノード数はそれぞれ 600,300 とした．学習回数は 400 エポックとした．以後楽器名の略称として P をピアノ,クラリネットを C，ベースを B で表す．

3.2. NMF との比較実験

NMF と提案手法を比較する．NMF は半教師ありを用いて教師部分（頻出パターンを表す行列）に楽器 1 のスペクトルを用いた．音高の識別結果はアクティベーション行列（重みを表す行列）が最大値になっている列とした．NMF は 1000 エポックで学習させた．

表 1 に結果を示す．表の(P-B),(P-C)はそれぞれ楽器音 1 がピアノ，楽器音 2 がベースもしくはクラリネットであることを示す．表 1 に示すとおり，楽器 2 にクラリネットを用いた時の NMF の認識制度は低く，提案手法が大きく上回った．これはピアノとクラリネットの定常部分のスペクトルが似ていて NMF では 2 つの楽器を分離できなかったためであると考えられる．一方提案手法では生成モデルを分布として学習できたため，スペクトルが類似していても楽器 1 の音高を同定することができた．

3.3. 音高同定器への入力の違いによる比較

本実験では第 2 ステージの多層 NN への入力を

表:2 ステージ目への入力比較

	第 4 層	第 3 層	出力層
認識率(P-C)	91.6	89.3	85.4

表 3: ほかのオートエンコーダとの比較

	DVAE	DAE	VAE	AE
認識率(P-C)	91.6	89.2	73.5	73.3

DVAE の第 3 層，出力層にした場合と，第 4 層を用いた場合とを比較した．表 2 より第 4 層を入力とした場合が最も性能が良いことが分かる．第 4 層では生成モデルに基づいて最も低次元に圧縮できていたため，多層と比較して良好な結果が得られたと考える．

3.4. 他のオートエンコーダとの比較

DVAE の優位性を示すため，デノイジングを行わない VAE，デノイジングを行う通常のオートエンコーダ(DAE)，一般的なオートエンコーダ(AE)を用いた時と比較した．なお DAE と AE は 5 層で構築し，2，4 層のノード数は 1000，第 3 層のノード数は DVAE の生成モデルの層のノード数と同じとした．表 3 よりデノイジングを行う DVAE と DAE の結果が他の二つの性能を大きく上回っていることがわかる．これはデノイジングによって楽器 2 を適切に除去できたからであると考えられる．また，DVAE の結果が DAE の結果を上回ったことより，生成モデルが有効に働くことが確認できた．

4. まとめ

本研究ではデノイジング変分オートエンコーダを用いた多重音解析の性能を評価した．実験の結果，多重音解析で一般的に用いられる NMF，および他の種類のオートエンコーダと比較して良好な結果を得ることができた．今後は 3 重音以上の多重音解析に取り組みたい．

参考文献

[1] 亀岡弘和，後藤真孝，“多重音解析と自動採譜，”情処学論， vol.50, no.8, pp.711–716, 2009.
[2] P. Smaragdis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in IEEE Workshop on Applications of Signal Process. Audio Acoust., New Paltz, NY, 2003, pp. 177–180.
[3] D. J. Im, S. Ahn, R. Memisevic, and Y. Bengio. Denoising criterion for variational auto-encoding framework. In arXiv preprint arXiv:1511.06406, 2015.