機械学習のための深層学習による非線形特徴変換

2016 年 9 月

北海道科学大学大学院

丹羽 孔明

目次

表目次

図目次

第 1 章

序論

1.1　現代社会における機械学習の位置付け

機械学習はコンピュータシステムが経験によって, 環境への適応や目標達成の方法を自己で発見するための方法論である. トム・ミッチェルは著書『Machine Learning』(Mitchell 1997) の冒頭において “機械学習の分野では, コンピュータプログラムが経験によって自動的に改善していくにはどうしたらいいかというテーマを掲げています.” と述べている. これを簡単に表現すると, 人間や他の生物のように経験から学習するコンピュータシステムの構築を目指す研究領域と言える. 一般には人工知能研究の中の一つの研究分野として周知され, 近年のマスメディアなどで人工知能研究と呼ばれているものは, 実際には機械学習を指すことが多い.

この機械学習では, データの表現, 規則, 知識あるいは特定の環境や状況下での行動を自動的に獲得するための方法論が体系化されている.

従来, 自動制御ではコンピュータ式あるいは機械式にかかわらず, 制御に関わる全ての要因は技術者の手で設計されていた. 例えば, 1955 年に東芝の自動式電気釜 (つまり電気炊飯器) が販売されたが, 自動で電源が切れるように機械式の自動制御機構が組み込まれていた. これはバイメタル技術を利用したサーモスタッド式の機構が採用されていたと言われている. また, このような自動制御のルーツについては, より古典的にオートマタやからくり人

形, 機械式時計にまで遡ることができるだろう. マイクロコンピュータが実用化されるに至り, 純粋な機械式自動制御からセンサとコンピュータを組み合わせた自動制御が主流となったが, コンピュータ上での情報処理や自動制御に要するデータの表現や制御規則と手順は技術者の手によって設計されていた.

しかし, 近年においては自動制御や情報処理における制御対象や達成目標が多様かつ複雑化し, 人の手による設計には限界が見え始めてきた. また, 専門家であっても適切なデータの表現や目標達成までの制御手順に対して, 明確な設計を持たせることが困難な課題が発生するようになっている. 例えば, 宇宙空間や惑星の探索機の制御などはこのような課題の最たるものであろうし, 近年話題になっている自動運転に関連する課題も同様であろう. さらには, 人間では処理が不可能な程に大規模なデータを分析することで, 有益な情報が得られることが知られてきた. この背景には, コンピュータの高性能化やセンサ技術あるいは他の工学技術の発展, ネットワーク技術の発展とこれに伴う大量の情報資源の収集が容易に可能となったこと, さらにはコンピュータによる自動制御や情報処理に対する要求の増加が考えられるだろう.

このような傾向は, 画像認識や音声認識など高次元データからのパターン認識に関連する分野やロボティクス分野での自動制御においては, 既知の問題でもあった. 例えば, 高次元データの認識課題ではデータから抽出する特徴量が認識精度を左右する. このため, どのように特徴量を設計し, これを抽出する特徴抽出器を設計するかが研究領域の中心にあった. つまり, 高次元のデータをどのようなデータの表現に写像し変換するかが重要な課題であったが, ここに機械学習の方法論が取り入れられるようになった. また, ロボティクスでの自動制御においても, より良い制御方法の探求やセンサからの高次元な情報を処理するために機械学習の方法論が導入されている. このように, 人の手による各種要因の設計や情報処理には課題が見えてきたが, これに対応するために機械学習の方法論が用いられてきた.

現代においては, 工業生産ラインや交通・通信・情報インフラでの自動制御や管理と運用のシステム, 数多く身の回りにある電子製品の中に, このよ

うな研究成果が応用されている．例えば, 身近なものをあげると, 自律的に稼働し掃除するロボットやスマートフォンに搭載されている各種の認識技術, より高度になったコンピュータ制御式の家電製品やディジタルカメラの各種自動補正機能などには機械学習の方法論が活用されている．また, 先に述べた自動運転や自律ロボット, 大規模データの分析, この他にも電子制御式の義肢装具や医療従事者を支援するシステムにも機械学習の導入が見られる．このように機械学習は現代社会において中核となる研究領域となりつつある．

このような状況下において近年, 深層学習 (Deep learning) と呼ばれる方法論が提唱された．この深層学習が従来の方法論を圧倒する性能や興味深い特性を示し, 多くの研究者や技術者の関心と社会の人工知能技術 (機械学習) への期待感を高めている．

深層学習は多層構造の人工ニューラルネットワーク (ANN: Artificial Neural Network) を用いた機械学習の方法論である．多層パーセプトロンを始めとして, 多層構造の学習モデルは古くから知られていたが, 計算機の性能や技術的な問題点により積極的に用いられてはいなかった．

しかし近年において, 多層構造の学習モデルに存在した技術的課題を解決する方法論が提案された．また, 計算機の性能の爆発的な進歩により, 計算処理の高速化と大規模かつ高次元なデータ資源を活用することが可能となり, 多層人工ニューラルネットワークの性能が, 従来よりも引き出されるようになってきた．これにより, 画像処理や音声処理あるいは予測処理の分野で驚異的な成果を収め, 未だその性能向上や応用の裾野が広がり続けている．特に 2016 年 3 月に開催された囲碁の対局において, Google DeepMind によって開発された AlphaGo がハンディキャップのない試合で人間のプロ囲碁棋士に勝利したことは衝撃的なニュースとなった．

1.2　音と生物と機械学習

社会性を持つ生物にとって音は重要な情報伝達媒体である．社会性を持つ昆虫や動物は他の個体と協調して, 自身の能力以上の仕事をこなすことがで

きる. 例えば, 複数の個体で餌を運んだり, 連携して狩りをしたり, あるいは移動経路を作ることもある. このとき, 彼らの多くは音や振動を利用してコミュニケーションをとり, 周辺環境で何が起こっているのか認識することに役立てることもある. また, 人間の音楽活動のように直接的なコミュニケーション用途ではなく, 音のパターンを高度に組み合わせ何らかの表現活動を通じた間接的なコミュニケーション手段とすることもある. 音や音のパターンに対して何らかの反応と行動を示し, またその音のパターンを自ら生成するといった 能力は, 社会性と音を感知する器官を持つ生物にとって欠かせないものである.

1.2.1 音と工学的応用

他方, 群知能 (Swarm Intelligence)(Beni and Wang 1993) やスワームロボティクス (Swarm Robotics)(Ohkura et al. 2010; ahin 2004; 大倉和博 et al. 2011) という研究領域がある. 社会性を持つ生物をモデルとして, 複数の自律したシンプルなロボットが協調してタスクを達成する方法論や, それらの群れが持つ知的な振る舞いを理解しようという領域である. ここで扱われるタスクは Fig. 1.1 に示すように, 協調して対象を探したり, 対象物を目標まで移動させたり, 通常は通行不可能な場所に道を設置したりといったものがある. これらのタスクは災害現場などでは特に必要とされる応用領域であろう.

通常, 自律的なロボットシステムは視覚ベースの認識メカニズムを備えている. しかし, スワームロボティクスでは各ロボットがシンプルな構造かつ, 局所的な情報を観測して行動する. 各個体が局所的な情報に基づいて独立に行動するため, 一部の個体に異常や故障が発生したとしてもシステム全体は正常に稼働し続ける. このように群れ全体のシステムとしては高いロバスト性を持つものの, この特性を維持するためには各個体に搭載するセンサ系や個体間の通信手段が問題となる.

社会性の生物は音を媒体としてコミュニケーションや状況認識を行う能力を有する. そして, たとえ一部の個体が取り除かれたとしても群れ全体は正

常に維持され活動を続ける．つまり, 音という媒体が局所的な性質を持つ情報源であり, これを介してコミュニケーションや状況認識の能力が発揮されていることが示唆される．これをシンプルなロボットの個体に取り入れることは効果的である．

視覚ベースの認識システムを有する場合にも, 聴覚ベースの認識システムは有効に働く．視覚ベースの認識システムを有する自律的ロボットシステムにおいても, 死角となる位置の情報を観測することはできない．例えば, 壁の向こうの状況を把握することは困難であるし, ガス漏れのように通常のカメラでは捉えられない場合もある．このようなとき, 聴覚ベースの認識システムは視覚ベースの認識システムをカバーすることができる．

1.2.2 音と情報処理的応用

さらに, 音楽情報処理など, 高度な音のパターンの認知と生成を要する研究領域もある．音楽情報処理は音楽についてのあらゆる事象を情報処理の観点から理解し, あるいは何らかのシステムの実現に応用しようという分野である．この領域の中には, 音にまつわる人間の生理的あるいは心理的な反応を再現しようという領域や, それを情報処理の枠組みで捉え, 自然な音声の合成や音楽を自動的に生成しようという試がなされている．

音や音楽は数学や計算幾何学との関連性が深い対象であるが, 生理学や心理学とも深い関連性を持つことが知られている．例えば．音楽を聴いているとき, 心理的には気分が落ち着いたり, 逆に興奮を誘発したり, 生理的には心拍数や呼吸数, 体温や血圧に変化を与えることが知られ, 音楽療法のように音楽が人へ与える作用を応用して, 心身の回復の助けとしようという試みも確立されている．また, 音楽を聴いて満足感を感じると神経伝達物質であるドーパミンが分泌されることが V. Salimpoor, R. Zatorre らの研究 (Salimpoor et al. 2011) により明らかにされている．このように人間は音楽に対して心理的あるいは生理的な反応を示すことがわかっているが, 音を検知する器官を持つ他の生物も音のパターンに対しては何らかの生理的反応を示すだろう．

1.2.3 機械学習と深層学習によるモデルの提案

これら音に関連する生物の反応や音楽ように高次の音のパターンを取り扱う能力は, 工学分野の応用や情報処理分野の応用としても重要な対象問題である.

このためには音響信号から何らかの音響イベントを検出し認識する方法論を要する. ここで, コップを落として割ったときの音やピアノの打鍵と発生した音のパターンなど, 何らかの事象に付随して発生する音や音のパターンを音響イベントという. これまで, 音響イベント検出や認識の研究では特徴量を人手で設計し, 隠れマルコフモデル (Hidden Markov Model; HMM) やサポートベクタマシン (Saport Vector Machine; SVM) あるいは他の認識モデルとともに用いることが主流であった. 特に, 音響信号の特徴量としてはメル周波数ケプストラム係数 (Mel-Frequency Cepstrum Coefficients; MFCC), 認識部分には HMM を基本として拡張を施したモデルは多い.

近年では深層学習と呼ばれる多層人工ニューラルネットワークの手法が, 画像や音声あるいは時系列予測といった対象問題で大きな成果を上げている. 深層学習の方法論の重要な特性の一つは, 人の手を介さずに観測データから適切な特徴抽出器を自動的に形成することである.

これまでの深層学習の成果を考慮すると, 音に関するデータを記号化せず, 時系列に並んだ周波数成分として表現された音響信号, あるいは音響信号の波形領域そのものを適切に処理することができるだろう. これは音のデータを一旦記号情報に変換し, この記号情報を基に行動や出力を決定するのではなく, 音の入力から記号を介さずに出力を決定するモデルが実現できる可能性を示唆する. 例えば, 近年の深層学習による画像生成のように楽曲の音響信号を直接生成するモデルやロボットが感知する音から記号を介さずに行動決定を行うモデルが考えられるだろう.

そこで, 本研究においては, 音や音のパターンに対して何らかの反応と行動を示し, またその音のパターンを自ら生成するといった活動を機械学習の方法論によってモデルにすることを目指す. モデルは深層学習の方法論を

取り入れ, 記号を介さずに音の入力から音響信号レベルでの音のパターン生成や行動決定を行うことを考える. Deep Q-Network[Mnih et al. (2013); mnih2015] や小鳥のさえずり行動獲得などの生理学的な研究領域での発見 [KOJIMA (2012); 西川淳 2007] より, モデルは強化学習の Actor-Critic 法に深層学習のメカニズムを取り込むことで実現できそうである. 本研究が目指す第一の段階は, 提案する聴覚ベースのシステムを搭載したシンプルなロボットが Fig. 1.2 に示すようなタスクを達成したり, 人間が音楽として許容可能な音のパターンを出力することである.

ここで, 入力となる音響信号について特徴量の設計と時間軸方向の依存性の解決が第一の課題となる. 本論文ではこの研究の一環として, まずは音響信号の取り扱いに関する領域に深層学習を取り入れた方法論を構築するための幾つかの検討と提案を行う.

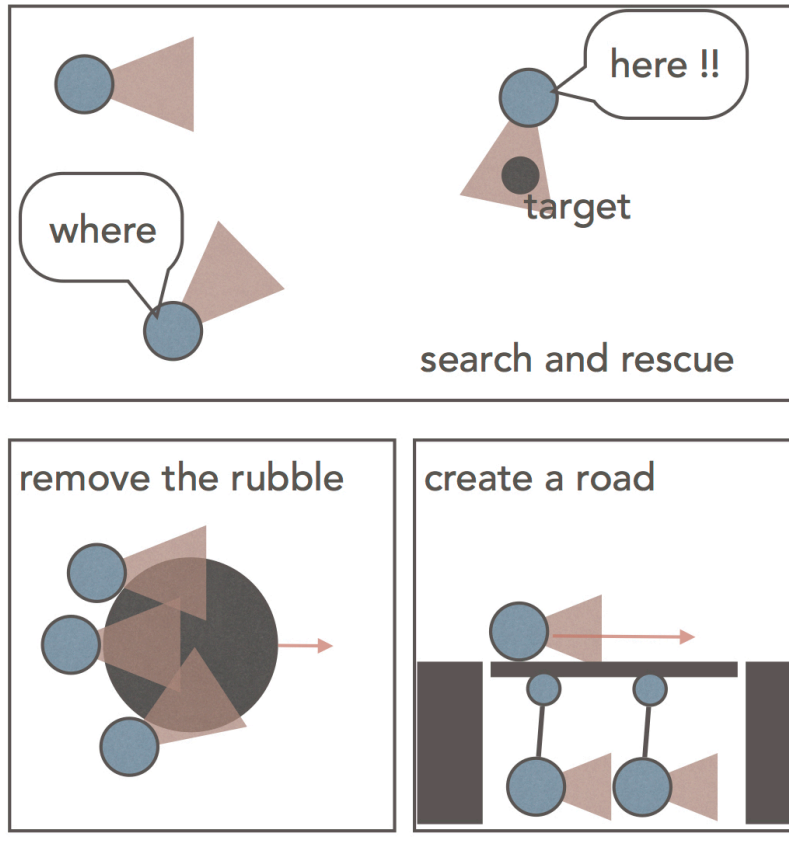


Fig.1.1: Example of tasks in the Swarm robotics

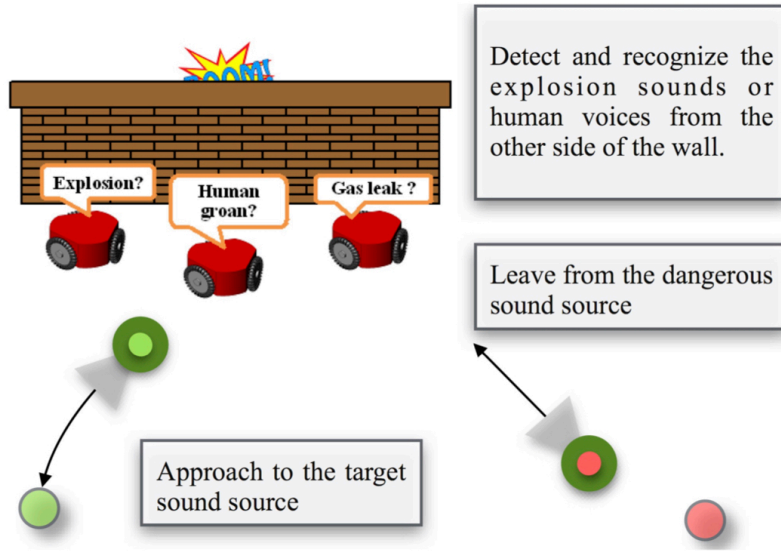


Fig.1.2: Tasks of Our first target

1.3 本論文の構成

第 1 章　序論

一般に, 機械学習は人工知能研究の一つの領域として捉えられている研究の領域である. この機械学習は簡単に表現をすると, あるコンピュータ制御されるシステムが人間や他の生物のように経験から学習し, 周囲の環境に適応するにはどうすれば良いかというテーマを掲げている. 現代社会においては, 自動制御や情報処理において機械学習の重要性が増し, 人工知能研究への期待感も大きい. この章では, この機械学習が現代社会でどのような位置にあるのかを最新の深層学習まで簡単に紹介するとともに, 本研究のテーマとその対象とする音を観測状態とした深層学習と強化学習によるモデルについて触れる.

第 2 章　概念モデルと関連研究

この章では, 本研究のテーマとする対象についての概念モデルを提示し, 本論文で対象とする課題の範囲について規定する. 提案する概念モデルは Actor-Critic 型の強化学習に深層学習のメカニズムを取り入れ, 記号を介さずに与えられた音響信号から直接的に行動を決定するようなモデルとなる. そこで, 深層学習とその基礎となる人工ニューラルネットワーク (Artificial Neural Network; ANN) 研究のこれまでの歩みに触れ, また, 強化学習の代表的な方法論について述べる.

第 3 章　深層学習による時間領域の信号波形ベースのモデル

従来の信号処理では, 時間領域で表現される信号波形を周波数分析を行い, 時間-周波数領域の情報に変換する. この周波数成分の時系列データについて, 各種の分析を経て特徴量の設計や特徴抽出を行い, この特徴量を分類器などの入力としてタスクの達成を目指す. しかし, 深層学習のメカニズムであれば, 時間領域の信号波形より直接的に特徴抽出や時間方向の依存性を解決可能な多層ニューラルネットワークの学習が可能であるかもしれない. この章では, 制約ボルツマンマシン (Restricted Boltzmann Machine; RBM) と Conditional RBM と呼ばれる生成モデルによる多層ニューラルネットワークを提案し, 時間領域の信号波形の時系列データを対象とした予測と信号波形の復元を課題としてモデルの検証を行う.

第 4 章　深層学習と聴覚フィルタおよび Echo State Network によるモデル

音響信号を観測して出力を決定するシステムは, 雑音のある環境下においても適切な出力ができるように雑音に対してのロバスト性を求められる. そこで, 深層学習のメカニズムと聴覚フィルタによる周波数分析を用いることで, 特徴抽出や時間方向の依存性に加え, 雑音環境下においてもロバスト性を有する多層ニューラルネットワークの学習を試みる. この章では, 自己符号化器 (Auto Encoder; AE) とエコーステートネットワーク (Echo State Network; ESN) による多層ニューラルネットワークに聴覚フィル

タによる周波数分析を加えたモデルを提案し,IEEE AASP Challenge の D-CASE challenge に含まれる OL subtask(Giannoulis et al. 2013; IEEE n.d.; Stowell et al. 2015) を課題としてモデルの検証を行う. この OL subtask は, オフィスで頻繁に発生する音響イベントについて, 雑音を含む実環境下での多クラス分類課題とその課題のための音響信号データセットである.

第 5 章　人間の作曲活動をモデル化した自動作曲システム

音楽を自動的に生成するシステムは自動作曲システムと呼ばれており, 記号を介さずに与えられた音響信号から直接的に行動を決定するようなモデルの対象問題として, モデルが達成すべき課題をいくつも内包している. この章では, 本研究の概念モデルについて, 自動作曲を対象問題として Actor-Critic の強化学習部分の構築と学習に用いる報酬と呼ばれるスカラーな量の設計を検討する. 本論では, まず初期の段階として, Actor は記号情報を用いたネットワークを用いて, 記号情報の出力を行うコンポーネントとする. 具体的には, Actor は音楽における音高-音価 (音の高さ-音の長さ) の組をノード, 各ノード間を遷移の確率を持った有向のリンクで接続したネットワークとし, ある状態の遷移確率に従ってネットワーク上のノードの音高-音価を順次出力する. Critic は Actor の出力パターンの音を評価し, Actor の持つネットワークの構造をある目標に向かって修正するコンポーネントと定義する. これらを学習させるために与える報酬の設計についても, 合わせてこの章で検討と検証を行う.

第 6 章　結論

この章では, 各章の概要と実験結果をまとめ, 本論文の結論として各章の総括を述べる.

参考文献

Beni, G., & Wang, J. (1993). Swarm intelligence in cellular robotic systems. In *Robots and biological systems: Towards a new bionics?* (pp. 703–712). Springer.

Giannoulis, D., Benetos, E., Stowell, D., Rossignol, M., Lagrange, M., & Plumbley, M. D. (2013). Detection and classification of acoustic scenes and events: An ieee aasp challenge. In *Applications of signal processing to audio and acoustics (wASPAA), 2013 iEEE workshop on* (pp. 1–4). IEEE.

IEEE. (n.d.). IEEE aASP challenge: Detection and classification of acoustic scenes and events. http://c4dm.eecs.qmul.ac.uk/ scenesevententschallenge/, Acsessed 17 May 2016.

KOJIMA, S. (2012). 小鳥のさえずり学習の神経機構: 大脳基底核経路と強化学習モデル. **比較生理生化学**, *29*(2), 58–69.

Mitchell, T. M. (1997). Machine learning (international edition). *Computer Science Series. McGraw-Hill, New York*.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Ohkura, K., Yasuda, T., & Matsumura, Y. (2010). Coordinating the adaptive behavior for swarm robotic systems by using topology and weight evolving artificial neural networks. In *Evolutionary computation*

(cEC), 2010 iEEE congress on (pp. 1–8). IEEE.

Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A., & Zatorre, R. J. (2011). Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature neuroscience*, *14*(2), 257–262.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., & Plumbley, M. D. (2015). Detection and classification of acoustic scenes and events. *Multimedia, IEEE Transactions on*, *17*(10), 1733–1746.

ahin, E. (2004). Swarm robotics: From sources of inspiration to domains of application. In *Swarm robotics* (pp. 10–20). Springer.

大倉和博, 保田俊行, & 松村嘉之. (2011). 構造進化型人工神経回路網による swarm robotics のための適応的協調行動の生成. **日本機械学会論文集** *C* **編**, *77*(775), 966–979.

第 2 章

概念モデルと関連研究

2.1 本研究のテーマと本論の対象とする課題

社会性を持つ生物は音や音のパターンに対して何らかの反応を示し, またその音のパターンを自ら生成する能力を有している. 我々は本研究において, この活動を機械学習の方法論においてモデルにすることを目指している. 具体的には, 音楽や音声に関するデータを記号化せず, 時系列に並んだ周波数成分として表現された音響信号, あるいは音響信号の波形領域そのものを深層学習のメカニズムにより概念学習を行う. そして, そこから新しい楽曲の生成やロボットが感知する音響信号からの行動決定などに応用する. モデルの概念は, 強化学習の Actor-Critic 法に深層学習のメカニズムを取り込むことで表現する. ここで, モデルを構築するための第一の課題は, 音響信号について, 特徴量の設計と時間方向の依存性解決となる.

まず, 音響信号の特徴量の設計が一つ目の課題である. 音響信号はシンプルな時間領域の 1 次元シーケンスデータながらも, 高次の情報を内包する媒体である. このため, 画像に関するタスクと同様に音響信号に関する各種のタスクは, 適切な特徴量を設計して与えなければ, タスクの達成が上手くいかないことは周知されている. 時間領域の波形信号, あるいは時間-周波数領域の時系列に並んだ周波数成分の何れにしても, まずは特徴量の取得を要する.

次に, 音響信号の時間方向の依存性を解決することが 2 つ目の課題である. 音の時間領域の波形信号, あるいは時間-周波数領域の時系列に並んだ周波数成分の何れもが, 次の状態の予測や信号から情報を取り出すために, 過去の状態履歴を参照しなければならない. このように, 音は時間方向について過去の状態に依存性を持つ, N 次マルコフモデル (N-Order markov model) に属する情報媒体である. 強化学習の枠組みで取り扱うためには, 単純マルコフモデル (1 次マルコフモデル) に状態を近似することが望ましい. 特に, 音響信号から得られた特徴量が過去の状態の履歴を内包する設計でない限りは, この課題を別途解決する方法論を要する.

そこで, 我々は深層学習のメカニズムに基づいた方法論に基づき, 多層ニューラルネットワークによって, これらの課題を解決するモデルを提案する. 具体的には, 深層学習に見られる特徴量の自動獲得に関する特性を音響信号の特徴量空間への変換 (特徴抽出) に, リカレント型ニューラルネットワークのメカニズムを時間軸方向の依存性解決に用いることを検討する. ここで, 両者ともに人工ニューラルネットワークによる方法論であることに着目し, 深層学習の事前学習の概念を応用すると, それぞれを独立して学習させたとしても, 一つの多層ニューラルネットワークに統合可能であることがわかる. 本論ではここで述べたように, 深層学習のメカニズムを音響信号の特徴抽出と時間軸方向の依存性解決に用いることを課題の範囲とする.

2.2 本研究の概念モデル

ここで, 本研究で提案するモデル全体の概念を示す. モデルの概念は Fig. 2.1 に示すように, 強化学習の Actor-Critic 法に深層学習のメカニズムを取り込むことで表現する. モデルは Actor-Crotic 法と同様に, 現在の環境の状態から将来に渡る報酬を予測する Critic と, 状態と価値関数の予測から行動を決定する Actor を持つ. これに加えて, 深層学習のメカニズムによる音響信号の特徴抽出と時間方向の依存性を解決するためのモジュールから構成される. ここで, 特徴抽出を行う部分をモジュール A, N 次マルコフモデルを単純マルコフモデルに近似する部分をモジュール B と呼ぶこととする. また,

モジュール A とモジュール B は独立したモジュールと見立てて学習を進めるが, 事前学習の概念を応用し, 一つの多層ニューラルネットワークとして動作するように設計する. 提案モデルは状態観測を行うセンサ系から Actor と Critic への間に, センサ系から得られた音響信号を処理する多層ニューラルネットワークが介在する構造となる.

また, 提案モデルのエージェントの処理フローについて簡単に説明する. エージェントは時刻 t に環境から音響信号を状態 s'_t として観測する. 状態 s'_t は深層学習のメカニズムにより特徴ベクトル v_t を介して, N 次マルコフモデルから 1 次マルコフモデル (単純マルコフモデル) の状態 s_t に近似される. エージェントはこの状態 s_t で行動 a_t を取り, 報酬 r_t を得る. このとき, 環境の状態はエージェントの行動に従って, 決定的あるいは確率的に状態 s'_{t+1} へ遷移する. 状態 s_t は一つ前の時刻 $t-1$ の状態のみに依存するよう単純マルコフモデルへ近似されるので, Actor と Critic の学習は後述する標準的な手法がそのまま適用できる. また, 人工ニューラルネットワークによって. Actor と Critic の双方を関数近似するとき, モデル全体を一つの多層ニューラルネットワークに統合することができるだろう.

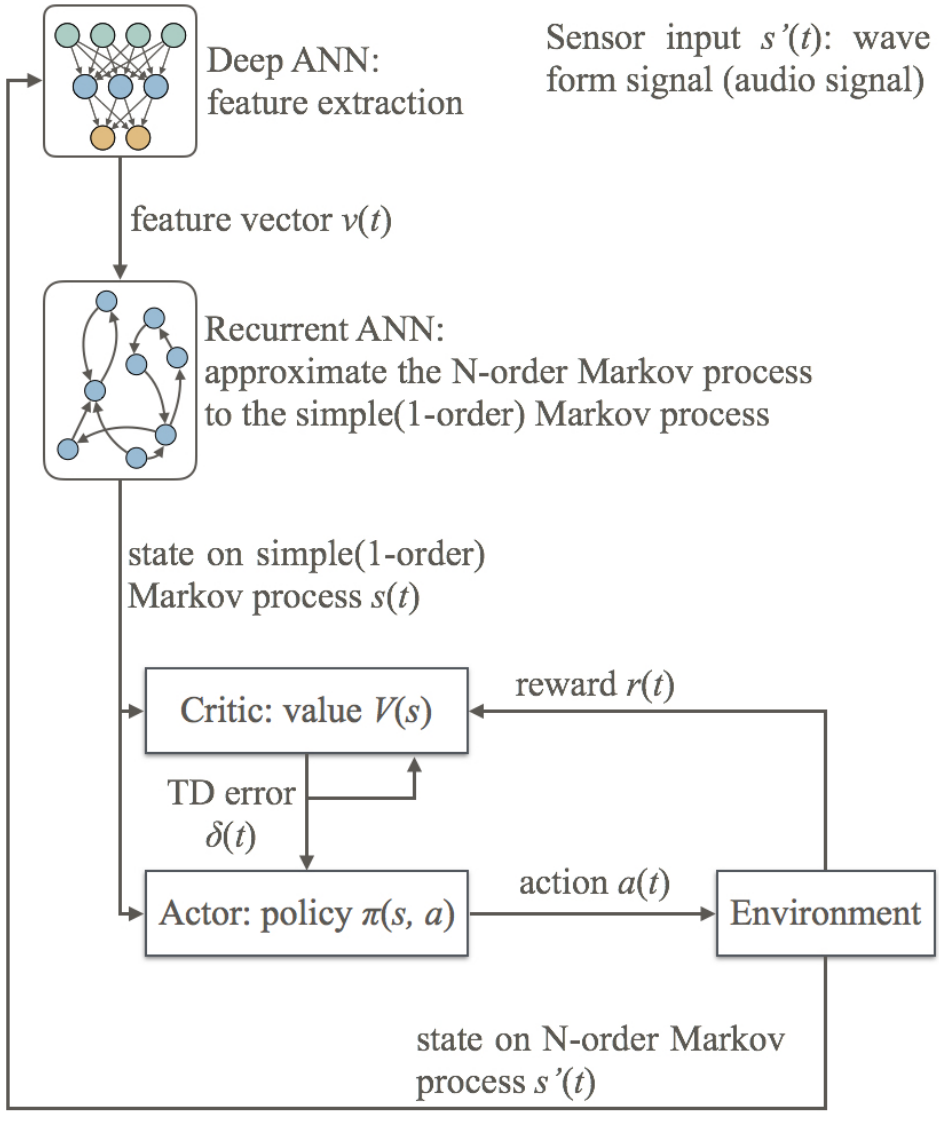


Fig.2.1: Approach of our Actor-Critic model include Deep learning and Recurrent ANN

2.3 深層学習

2.3.1 深層学習の概要

深層学習は画像処理や音声処理あるいは予測処理の分野で驚異的な成果を収め, 未だその性能向上や応用の裾野が広がり続けている. 深層学習とは, 多層構造の人工ニューラルネットワーク (ANN: Artificial Neural Network) を用いた機械学習の方法論である. 多層パーセプトロンを始めとして, 多層構造の学習モデルは古くから知られていたが, 計算機の性能や技術的な問題点により積極的に用いられてはいなかった.

しかし近年において, 多層構造の学習モデルに存在した技術的課題を解決する方法論が提案された. また, 計算機の性能の爆発的な進歩により, 計算処理の高速化と大規模かつ高次元なデータ資源を活用することも可能となった. これらの要因が重なり, 多層人工ニューラルネットワークの性能が, 従来よりも引き出されるようになってきた. この結果, 従来の方法論を圧倒する性能や興味深い特性を示すことがわかり, 多くの研究者や技術者の関心と社会の人工知能技術への期待感を高めている.

深層学習は多層構造による変数間の複雑な関連性により, 高い柔軟性と表現能力を持つ学習モデルである. この多層構造がもたらす有益な特性として, 多重なエンコードを介した特徴の抽出・変換が知られている. 生の, あるいは最低限の前処理を施した観測データを各層で逐次的にエンコードしていくことにより, 適切なデータの表現が自動的に獲得される. つまり, 人の設計や仮説によらない特徴量の抽出・変換器である.

2.3.2 人工ニューラルネットワーク研究のこれまで

高度な情報処理を行う機械学習の方法論構築を目指して, 生物の神経ネットワークをモデルとした人工ニューラルネットワークの研究が長年に渡って行われてきた. しかし, その道のりは平坦ではなく, これまでにも 2 度の流行とその後下火になるという状況があった. 現在の深層学習は人工ニューラ

ルネットワークの研究にとって 3 度目の流行期と言える.

生物の神経ネットワーク

生物の脳には多数のニューロンが存在し, それぞれが他のニューロンと結合して複雑なネットワークを形成している. 個々のニューロンの構造と動作はシンプルである. ニューロンは Fig. 2.2(古川正志 et al. 2012, p. 154) に示すように, 本体である細胞体とそこから伸びる軸索, そして樹状突起からなる. 軸索は細胞体から出力される信号を他のニューロンへ伝達し, 樹状突起は他のニューロンからの信号を受信する. また, あるニューロンの軸索と他のニューロンの樹状突起間の結合をシナプスという. 自分自身の樹状突起に対して自身の軸索が結合することはないが, 他のニューロンを介して再度自身に結合するようなループした結合構造を持つことはある.

ニューロンはシナプスを介して他のニューロンからの信号を受け取る. このとき細胞体の膜電位が上がったり下がったりするが, 電位の上昇と下降はシナプスの結合に付随した特性である. 電位が上がるシナプス結合を興奮性, 下がるシナプス結合を抑制性という. また, どのくらい電位が変動するのかも個々のシナプス結合により差がある. このような活動の中で細胞体の膜電位が一定の値を超えると, その細胞体から電気的なインパルス信号が出力される. これを神経の発火や活性化といい, インパルス信号は活動電位と呼ばれる. このとき, 神経発火が起こる膜電位の境界値を閾値という.

ニューロン間を次々と伝搬するインパルス信号とその相互作用によって, 脳では各種の情報処理がなされる. 各シナプス間の結合強度は, 外界からの刺激や他のニューロンの活動に影響を受け, 機能的・構造的な変化が常に起きている. これをシナプスの可逆性という. この可逆性が生体においての記憶や学習など高次の機能の基盤となり, 人工ニューラルネットワークの方法論を構築するための重要な手掛かりにもなっている.

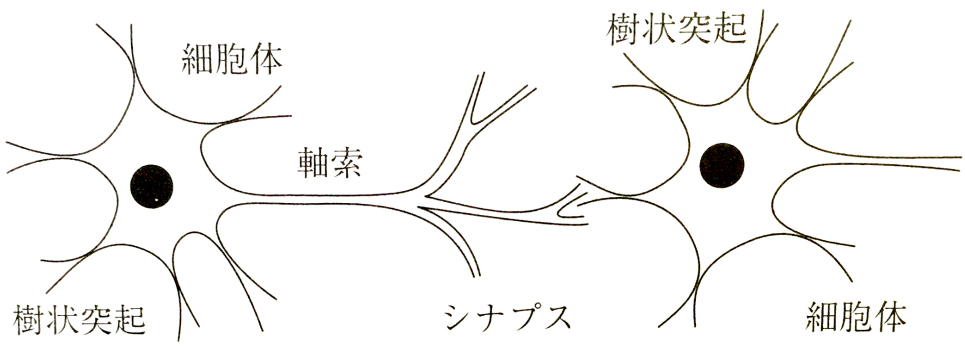


Fig.2.2: Neuron structure of the living body (古川正志 et al. 2012, p. 154)

形式ニューロン

ニューロンの最初の数理モデルは McCulloch と Pitts により提案され (McCulloch and Pitts 1943), 形式ニューロンと呼ばれている. モデルは Fig. 2.3 に示すようにシンプルな構造である. ニューロンは重み付きの複数の入力信号 (x) を受け取り, この総和 (u) がある値を超えると信号を出力 (活性化) する. これを数式で表現すると次のシンプルな式となる.

$$y_j = f(u_j) = f(\sum w_{ji} x_i) \tag{2.1}$$

ここで, y_j : 出力信号, u_j : ニューロンの膜電位, h_j : ニューロンの閾値 (バイアスと呼ばれることもある), w_{ji} : i 番目の入力信号に対する結合加重, x_i : i 番目の入力信号 $(x_1, x_2, \ldots, x_i)$, $f(\cdot)$: ニューロンの活性化関数である. McCulloch と Pitts による形式ニューロンでは, 活性化関数は 0 か 1 を出力する次のステップ関数が採用されている.

$$f(u) = \begin{cases} 0 & (u < h_j) \\ 1 & (u > h_j) \end{cases} \tag{2.2}$$

ニューロンの数理モデルには, その活動をより詳細なモデルとした

Hodgkin-Huxley モデルや, これを簡略化した FitzHugh-Nagumo モデルといった他の数理モデルも提案されている. しかし, 実用的に用いられる人工ニューラルネットワークモデルの多くは, この形式ニューロンを基に拡張を加えたり活性化関数を変更したニューロンモデルが利用される.

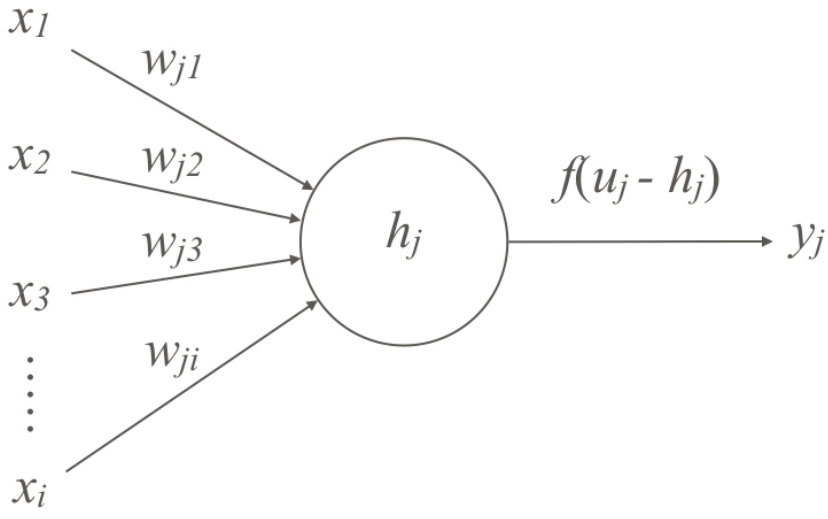


Fig.2.3: Formal neuron

パーセプトロン

その後, Rosenblatt により形式ニューロンを応用したパーセプトロン (Rosenblatt 1958) が提案された. パーセプトロンは形式ニューロンを複数組み合わせ, 教師データを用いた勾配降下法 (Gradient Descent) により学習を行うモデルである. ここで, 学習とは結合重み w のパラメータを決めるための更新処理をいう. パターン認識に基づいて分類問題を解くシンプルなネットワークありながら, 学習能力を持つモデルであった.

Rosenblatt の提案したモデルは S(sensory layer), A(associative layer), R(response layer) と呼ばれる 3 層からなっている. 信号は $S \to A \to R$ と伝搬される. $S \to A$ 間の重みランダムに決定され, パーセプトロンの本質で

ある学習は $A \to R$ 間のみで行われる. このように $A \to R$ 間の学習しか定義されていないため, 実質的には入力層と出力層を持つ 2 層の階層型ネットワークモデルであった. この 2 層のパーセプトロンは単純パーセプトロンと呼ばれている.

このパーセプトロンの登場が人工ニューラルネットワーク研究の第一の流行期であった. しかし, 単純パーセプトロンは線形分離可能な問題しか学習できないという限界が指摘され (Minsky and Papert 1969), 下火となった.

多層ネットワークと誤差逆伝搬法

1980 年代に入り, Fig. 2.4 のように多層にしたパーセプトロンを誤差逆伝搬法で学習させる方法が提案された (Rumelhart et al. 1986, 1988). なお, 誤差逆伝搬法自体はそれ以前にも, 異なる名称で幾度も再発見と発表がなされている. また, 多層の順伝搬型人工ニューラルネットワークは基本的にこの多層パーセプトロンのことをいう. これにより, 単純パーセプトロンの型分離不可能な問題が解けないという課題が解決された.

線型分離不可能な問題が解けるようになると, 人工ニューラルネットワークを用いた多くの他クラス分類問題が試みられるようになった. 例えば, Fig. 2.5 のような画像セットから特徴量を抽出し, Fig. 2.6 のように順伝搬型人工ニューラルネットワークに与えて分類する課題などである. これにより人工ニューラルネットワークの研究が大きな広がりを見せた. これが 2 度目の流行期と言われているが, これも 90 年台後半には次第に低調になっていった.

多層パーセプトロンと誤差逆伝搬法での学習による 2 度目の流行期が低調になっていった要因は, 大きく二つあると言われている. 一つは, 誤差逆伝搬法による多層ネットワークの学習には勾配消失問題と呼ばれる課題があり, 2 層や 3 層より深いネットワークの学習が困難であったことである. もう一つは, モデルや学習のパラメータが多く, それぞれがどのような影響を与えるかは経験的なノウハウに依存し, 体系化された理論がなかったことと言われている (Simard et al. 2003).

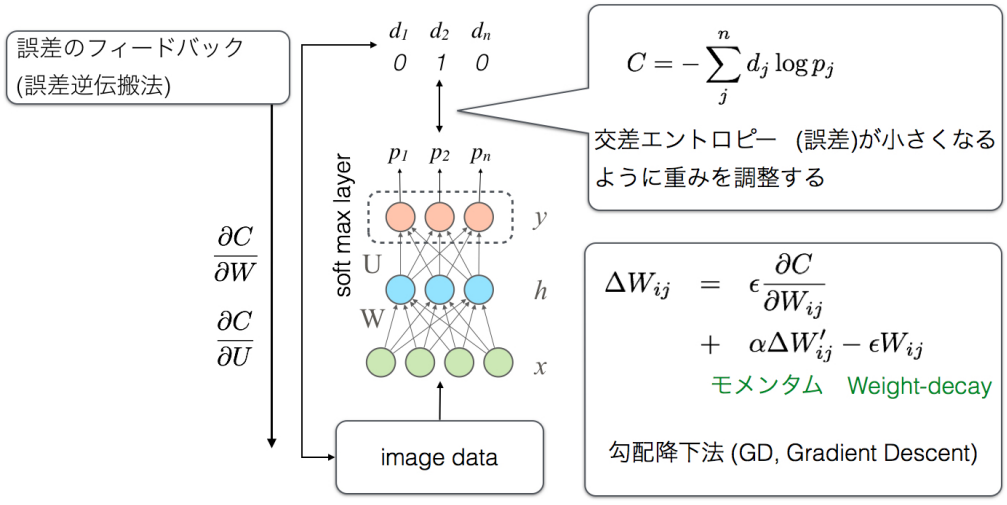


Fig.2.4: Multilayer perceptron and Back propagation

Fig.2.5: Image dataset

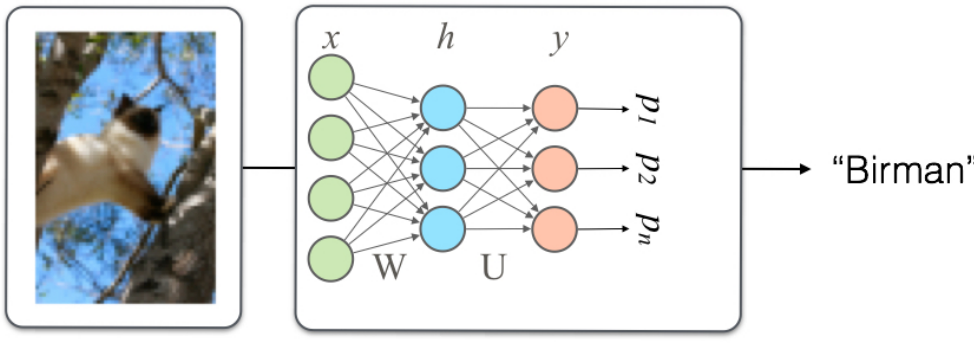


Fig.2.6: Classification with the Multilayer perceptron

勾配消失問題

誤差逆伝搬法による誤差の勾配は線形の計算で逆伝搬される．これにより, 各層の重みが大きいと誤差の勾配が各層への伝搬に伴い急速に大きくなるため, 誤差の勾配が発散する．また, 逆に各層の重みが小さいと, 急速に消失し 0 になる．いずれの場合でも, 特に入力層に近い層において重みの更新が適切に行われず学習が困難となる．

このような問題が勾配消失問題 (vanishing gradient problem) と呼ばれる．この発散や消失は 2 層や 3 層程度の浅いネットワークでは, さほど問題とはならない．しかし, Fig. 2.7 に示すような 4 層以上の深いネットワークの学習においては, 学習を妨げる大きな障害なっていた．この勾配消失問題をどのように解決するかが, 人工ニューラルネットワーク研究の重要な課題であった．

この問題を解決するために提案されたのが後述する事前学習である．事前学習はこれまで困難であった多層のニューラルネットワークの学習を可能とし, また多層のニューラルネットワークが想像以上に高い性能を示したことで, 再び多くの研究者の関心を引くこととなった．

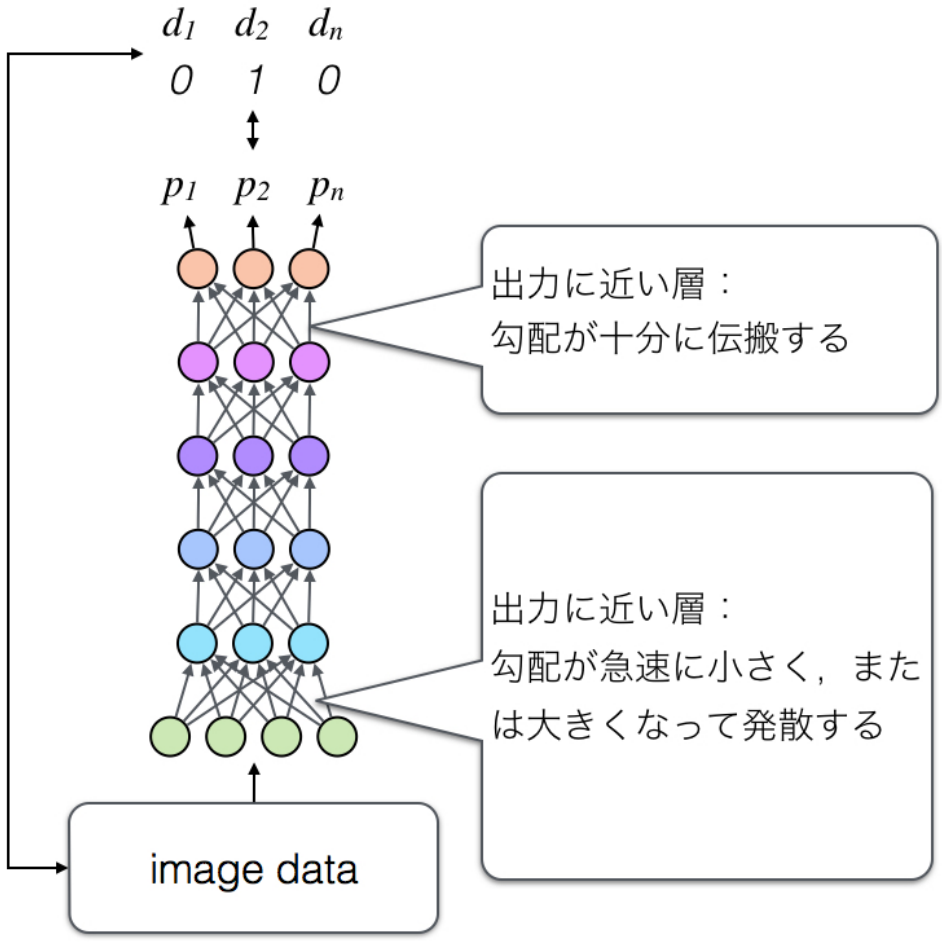


Fig.2.7: Gradient loss problem in a deep neural network

2.3.3 多層ニューラルネットワークの事前学習

多層ネットワークの事前学習は Hinton らにより提案され, この事前学習が現在の深層学習につながるブレークスルーとなった.

当時, Hinton らの研究していたディープビリーフネットワーク (Deep Berief Network; DBN) においても, 多層のモデルの学習は困難な課題であった. DBN は一般的な多層の人工ニューラルネットワークと同様に, 多

層構造を持ったグラフィカルモデルである．このモデルは確率的な振る舞いで記述され, データの確率分布を学習してデータを生成することのできる生成モデルであった．一般的な人工ニューラルネットワークで採用される誤差逆伝搬法とは異なる原理の方法論を用いてモデルの学習がなされるが, 多層構造のモデルの場合には多層ニューラルネットワーク同様に学習が困難であった．

Hinton らはこれを解決するために, 多層構造のモデルを小さなモデルに分割し, 入力層に近い方から順番に学習させるという手法を提案した．具体的には Fig. 2.8 に示すように, 多層構造のモデルを 2 層構造のモデルを積層したものとして捉え, 各層を制約ボルツマンマシン (Restricted Boltzmann Machine; RBM) を用いて教師なしの訓練をさせる．制約ボルツマンマシンは Fig. 2.9 の (c) に示すように, 結合に制限を加え完全 2 部グラフで表現されるボルツマンマシンの特殊な形のモデルである．訓練が終了した層のパラメータは固定され, その層の出力値が次の層の訓練データとして与えられる．これにより多層構造であっても DBN の学習が行えることが示された．

学習された DBN は多層の人工ニューラルネットワークへと転換が可能であった．これは, DBN と多層の人工ニューラルネットワークとの間に動作上の差異はあっても, 構造的な差異はなかったためである．このため, Hinton らの提案した手法で DBN を学習し, その後に分類用の層を追加した上でモデル全体を誤差逆伝搬法で学習させるという方法が試みられた．この結果, 多層の人工ニューラルネットワークであっても, 勾配消失や過適合といった既存の問題が解決されることが判明した．

後に RBM を用いた DBN ではなく, より単純な自己符号化器 (Auto Encoder; AE) を用いても, Hinton らの提案した手法に応用し, 同様の効果を得られることが明らかになった．

自己符号化器は非常にシンプルな教師なし学習の方法論である．このモデルは Fig. 2.9 の (a), (b) に示すように, 入力層, 隠れ層, 出力層を持つ 3 層構造の人工ニューラルネットワークを用いる．また, 入力層から隠れ層の流れを符号化 (encode), 隠れ層から出力層への流れを復号化 (decode) という．通常, 人工ニューラルネットワークは分類結果のラベルを教師信号に用いて,

出力を分類結果のラベルに近づけるように誤差逆伝搬法を用いてモデルを学習させる. 自己符号化器はラベルの代わりに入力信号自信を教師信号に用いて, 出力を入力と同様となるように誤差逆伝搬法を用いてモデルを学習させる. 通常の人工ニューラルネットワークを用いていながらも, 教師なし学習の方法論であったため, RBM の代わりに各層の学習に適用することができた. このとき, 各層のパラメータは自己符号化器の符号化部分のみを用いる.

このように, 目的とする多層人工ニューラルネットワークの学習において, 事前に層ごとの学習を行ってパラメータの良い初期値を求める方法論が事前学習 (Pre-Training) と呼ばれるようになった. また, 多層人工ニューラルネットワークの場合は, 事前学習後に目的に応じて教師信号を与え, ネットワーク全体を学習し調整 (Fine-Tuning) する. これにより, 学習が可能となった多層人工ニューラルネットワークが幾多の課題において, 従来手法を圧倒する性能を示し, 深層学習の概念が確立されることとなった.

しかしながら, 現在では事前学習は積極的には用いられなくなってきている. この背景には, $f(x) = max(0, x)$ で定義される ReLu 関数 (Fig. 2.9 の (b)) のような勾配消失問題の発生が抑制されるような活性化関数の提案, 学習中にニューロンユニットをランダムに無効化するドロップアウト (Drop out, Fig. 2.9 の (a)) のような過学習を抑制するテクニックの発見, といった要因がある. さらには, ある問題についての大規模なデータが利用可能になり, これを高速に処理可能なパワーを持つ計算資源に加え, 近年では慎重に初期値を決定し, データを適切に正規化することでも多層の人工ニューラルネットワークが上手くいくことがわかってきている. また, 現在の深層学習においては, もともと多層構造の学習が成功していた畳み込みネットワークを用いた応用が主流である. このため, 事前学習は分類器と独立した多層の特徴抽出器を学習させるといった用途での利用が多くなっている.

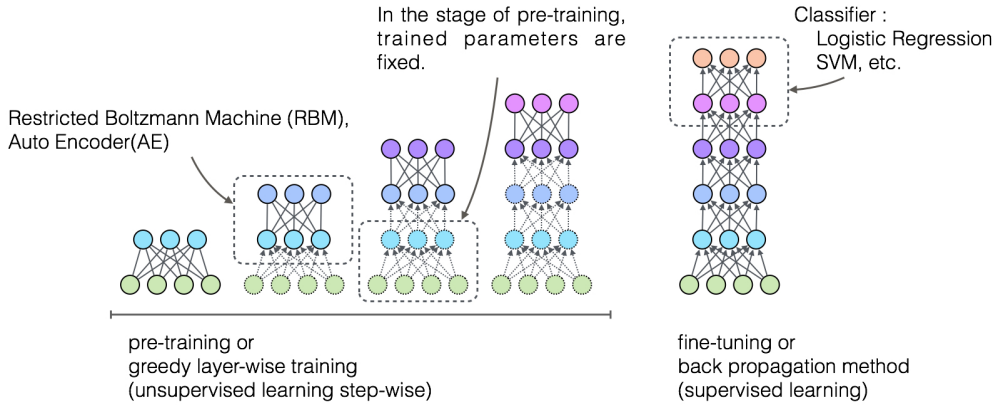


Fig.2.8: The Pre-training and the Fine-tuning

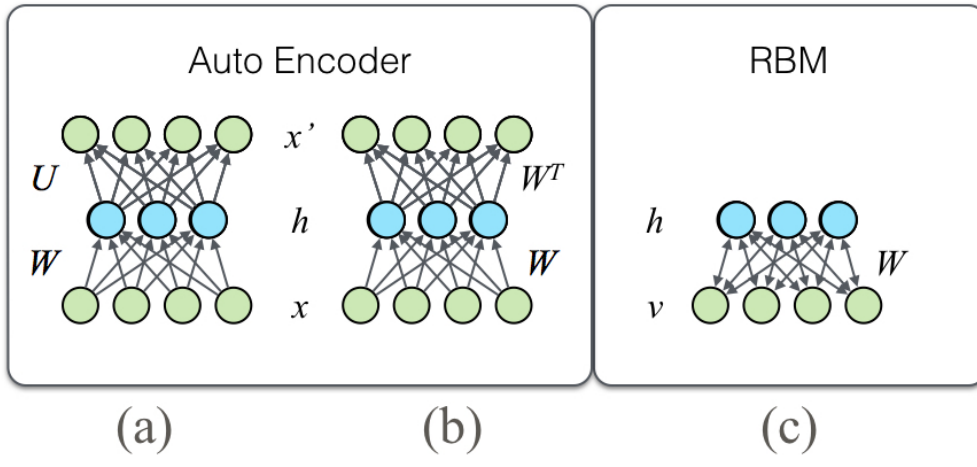


Fig.2.9: Auto Encoder and Restricted Boltzmann Machine

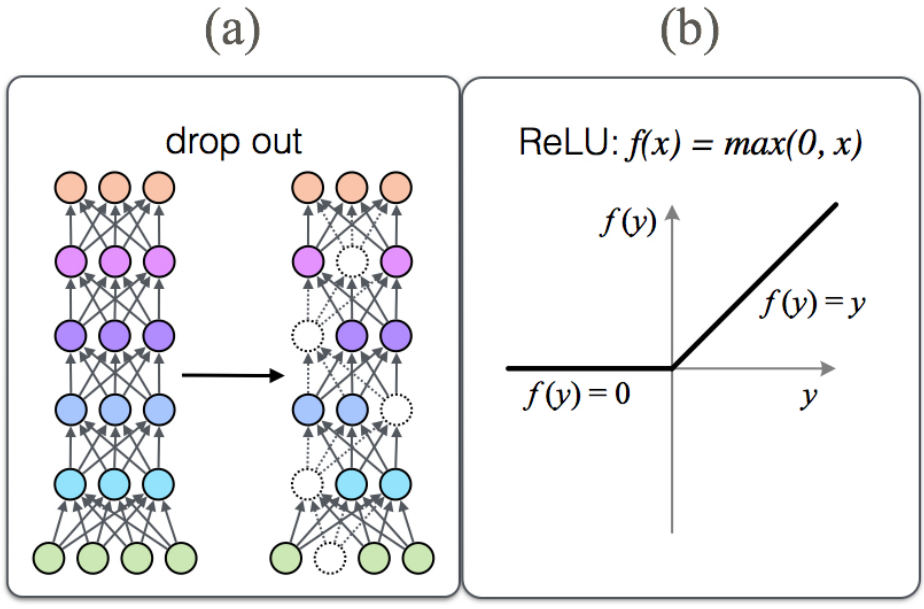


Fig.2.10: Drop out and ReLU function

2.3.4 畳み込みニューラルネットワーク

各層の間が全結合, つまり各層間が密に結合された多層ニューラルネットワークは学習が困難であった. そのような中, 例外的に多層構造のモデルが学習できていた方法論がある. 画像を対象とする畳み込みニューラルネットワーク (Convolutional Neural Network; CNN) である. この畳み込みニューラルネットワークは 80 年台後半には 5 層構造の多層ニューラルネットワークの学習に成功している.

多層ニューラルネットワークでは勾配消失問題や過適合が主だった課題であったが, この他にも移動や歪みに対する柔軟性と汎用性にかけるといった課題もあった. 例えば, 手書き文字画像の認識タスクを仮定する. 各層の間が全結合である多層ニューラルネットワークでは, 入力の各ユニットがそれぞれの画素に対応することになる. このため, 固定的な画像に対しては正しい応答を行うように学習することができたが. 全く同じ文字の形であって

も, 画素の位置がズレると入力の誤差が出力に影響を及ぼし, 正しい応答が得られないことがある.

畳み込みニューラルネットは, Fukushima らのネオコグニトロンに LeCun らが額誤差伝搬法による学習を取り入れる形で実現したモデルである.

ネオコグニトロンは, 1960 年ごろに猫の脳で発見された単純型細胞や複雑細胞の働きをヒントとして構築されたパターン認識システムである. ネオコグニトロンは順伝搬型の階層構造のモデルであったが, 単純型細胞や複雑細胞の働きを再現するため, Fig. 2.11(岡谷貴之 2015, p. 80)(b), (c) のように各層間の結合は制限されていた. このように単純型細胞と複雑細胞間の受容野の局所性, つまり局所受容野が再現されているモデルである.

この単純型細胞と複雑細胞間の働きはシンプルな挙動をしているが, その特性は非常に重要である. 単純型細胞と複雑細胞間の結合は局所性を持つ. このため, Fig. 2.12(岡谷貴之 2015, p. 81) に示すように, 前の層の信号を局所的・限定的に受け取り, これが上位の層において統合・反映される. この構造上の特性が画像の位置ズレや歪みに対してロバストな性能を発揮する.

この局所性を持つ結合構造は多層ニューラルネットワークの学習において, 有用な特性でもある. 畳み込みニューラルネットはネオコグニトロンを誤差逆伝搬法により学習可能にした拡張であるため, このような構造と特性はそのまま受け継がれている. 多層ニューラルネットの誤差逆伝搬法による学習では勾配消失や過適合の問題があることを述べた. しかし, 畳み込みニューラルネットのように局所性を持つ結合のとき, 不要な誤差の伝搬が抑制される. これに加え, 多層ニューラルネットの持つ自由度が強制的に制限されるため, 多層構造であっても勾配消失やか適合の問題を回避できた. このように, 局所的な結合であることが勾配消失や過適合といった問題に対しての一つの回答となっていた.

現在の深層学習において, 認識タスクや関数近似には畳み込みニューラルネットをより多層にしたり, 拡張を加えたり, 時系列を扱うことのできるモデル (Long short-term memory;LSTM など) と組み合わせたりして用いることが主流となっている.

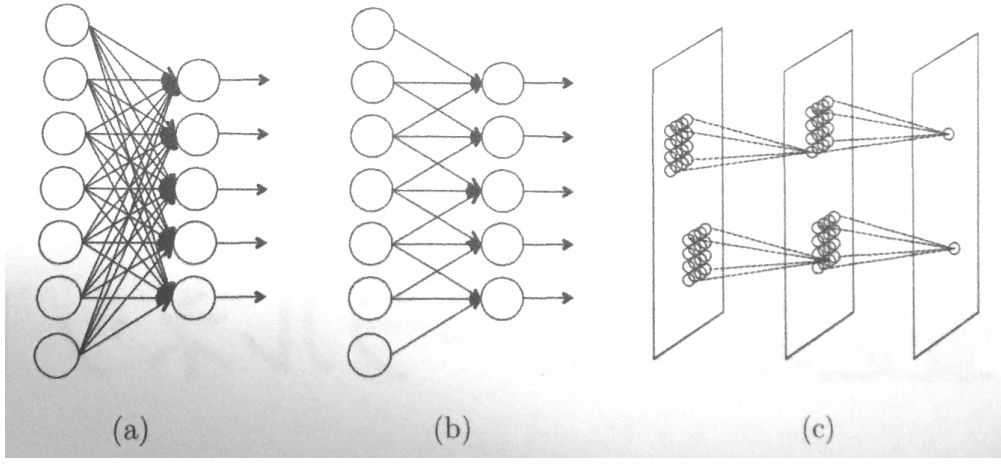


Fig.2.11: Standerd ANN and Convolutional Neural Network (岡谷貴之 2015, p. 80)

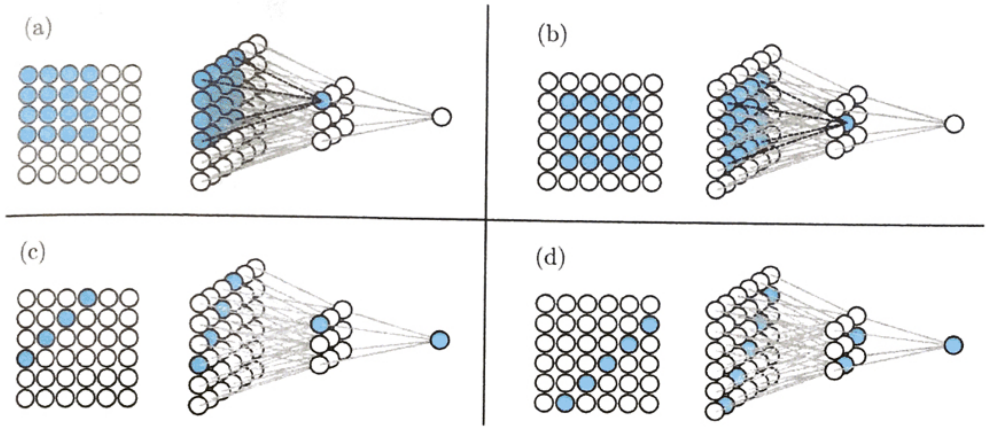


Fig.2.12: Behavior of the Simple cell and the Complex cells in the Local receptive field (岡谷貴之 2015, p. 81)

2.4 強化学習

2.4.1 強化学習の概要

強化学習は学習を行う自律的主体 (エージェント) が試行錯誤を通じて, 環境に適応する機械学習の方法論である. 教師付き学習 (Supervised learning) とは異なり, 状態入力に対応する正しい行動の出力である教師が与えられることはない. この代わりに, 報酬 (reward) というスカラーな量の情報を手掛かりに学習をする. このとき, 報酬にはノイズや与えられるタイミングに遅れがある. エージェントは環境から観測したある状態でとった行動が正しいか否かを判断するが, これには困難が伴う. 報酬のノイズや遅れの問題により, 行動を実行した直後の報酬 (即時報酬) のみでは, 行動が正しかったかどうかを判断できないためである.

このような条件において学習を行うためには, 報酬の予測が重要な鍵となる. 強化学習において, エージェントは現在の環境の状態から将来に渡る報酬を予測する価値関数と, 状態と価値関数の予測から行動を決定する行動規範を定義する関数から構成される. ここで, 行動規範を定義する関数は方策 (policy) と呼ばれている. これらは試行錯誤の中で, できるだけ正確な報酬予測を行うように価値関数の学習を進め, 行動規範を定義する関数は得られる報酬を増加させる行動を出力するように学習される. これにより, エージェントは将来的に得られる報酬の最大化を目的として, 観測される状態に適した行動出力への対応を学習する.

強化学習は何らかの制御を要する分野において注目を集めている. 実世界にある多くの制御問題において不確実性の取り扱いは大きな課題であるが, 強化学習では不確実性のある環境を扱うために多種の実問題へ応用ができる. また, 報酬遅れと確率的な状態遷移を伴う環境下において, これらを内包した制御規則を学習できる. これにより, 目標達成時に適した報酬を与えるように課題を設定することで, 目標の到達方法はエージェントが試行錯誤的な学習によって自動的に獲得する. 制御プログラム設計の自動化や省力化,

ハードコーディングするよりも優れた解が得られる可能性, 自律性や想定外の環境変化への対応といった, システムの設計者にとって厄介な課題解決の方法論となる.

例えば, 宇宙や深海のように通信が物理的に困難な環境で活動する探査機の制御は代表的な例だろう. このような環境で起こりうる環境変化やアクシデントについて, システムの設計者があらゆる事態を想定した制御を設計することは困難であるし, 大変な重労働でもある. このように制御方法がシステム設計者にとって自明でない課題において, 強化学習は最大の能力を発揮する.

2.4.2 価値関数と TD 誤差

強化学習において学習を行うためには, 報酬の長期予測や報酬予測の誤差を利用する. ここで, 報酬の長期予測や報酬予測誤差を定義する. 強化学習のエージェントは時刻 t に環境から観測される状態 s_tで行動 a_tを取り, 報酬 r_tを得る. このとき, 環境の状態はエージェントの行動に従って, 決定的あるいは確率的に状態 s_{t+1} へ遷移する.

状態価値関数

ある観測状態 s_tから見た長期の報酬予測は

$$V(s_t) = E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots] = E[\sum_{k=0}^{\infty} \gamma^k r_{t+k}] \tag{2.3}$$

と定義され, これを状態価値関数と呼ぶ. 状態価値関数は状態 s_tから将来的にどの程度の報酬が得られるか, 報酬の期待値を表現する. γは報酬の減衰率で $0 \leq \gamma \leq 1$ の間の値をとる. これは割引率と呼ばれ, 将来の報酬をどの程度重要視するかを決定する. 割引率を 1 とすると減衰のない報酬の期待値を表し, 0 とすると直近の報酬のみを重視することを表す.

状態価値関数について $V(s_t)$ と $V(s_{t+1})$, 2 つの価値関数の間には

$$V(s_t) = E[r_t + \gamma V(s_{t+1})] \tag{2.4}$$

の関係性が成り立つべきことになる. この $V(s_t)$ と $V(s_{t+1})$ の差

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{2.5}$$

は, 長期的な報酬の予測について誤差を表し, 状態価値関数の学習に用いられる. このδ_tは TD 誤差 (temporal difference error; TD error) と呼ばれ, 当初の予測と比較し, どれだけ多い, あるいは少い報酬が得られたかを示す.

状態価値関数の更新式は

$$V(s_t) \leftarrow V(t_s) + \alpha\delta_t \tag{2.6}$$

となる. ここで, αは $1 \geq \gamma > 0$ の値をとる学習率である. 方策が定常である場合には, これを TD 学習あるいは TD(0) 学習と呼ぶ.

行動価値関数

状態価値関数は状態についての報酬を予測するものであるが, これと同様にある観測状態 s_tにおいて行動 a_tをとった時点から見た長期の報酬予測は

$$Q(s_t, a_t) = E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots] = E[\sum_{k=0}^{\infty} \gamma^k r_{t+k}] \tag{2.7}$$

と定義され, これを行動価値関数と呼び, 予測される報酬の値は Q 値と呼ばれる. ここで, 状態 s_tで取り得る行動の集合 $A_t = \{a_{t1}, a_{t2}, \dots, a_{tn}\}$ と表すと, $a_t \in A_t$である. これにより, 状態 s_tにおいて最大の Q 値を持つ行動 a_tが最適な行動となり, $\max Q(s_t, a_t)$ と定義される.

行動価値関数についても $Q(s_t, a_t)$ と $Q(s_{t+1}, a_{t+1})$, 2 つの価値関数の間には

$$Q(s_t, a_t) = E[r_t + \gamma \max Q(s_{t+1}, a_{t+1})] \tag{2.8}$$

の関係性が成り立つ．この $Q(s_t, a_t)$ と $Q(s_{t+1}, a_{t+1})$ の差

$$\delta_t = r_t + \gamma \max Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \tag{2.9}$$

が行動価値関数の学習に用いられる．こちらも, TD 誤差と同様に当初の予測と比較し, どれだけ多い, あるいは少い報酬が得られたかを示す．

行動価値関数の更新式は

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\delta_t \tag{2.10}$$

となる．これを Q 学習と呼ぶが, 状態 s_tにおいて選択される行動 a_tは, ある行動選択方式 (探索戦略) によって行動価値関数のみから決定されるため, 方策 off 型 (off-policy) TD 学習と呼ぶこともある．

Actor-Critic 法

Actor-Critic 法のエージェントは Fig. 2.13 に示すように, 現在の環境の状態から将来に渡る報酬を予測する Critic と, 状態と価値関数の予測から行動を決定する Actor から構成される．Critic は前述した式 2.3 の状態価値関数である．また, Actor は Critic とは独立した方策を定義する関数で, ある観測状態においての行動を出力する．

Actor, Critic ともに学習には TD 誤差を用いる．Critic は状態価値関数であるので TD 学習と同様の式 2.6 を用いて, 正確な報酬予測を行ため TD 誤差を 0 に近づけるように学習される．また, Actor は得られる報酬を増大させるため, TD 誤差が正の方向に大きくなる行動を選択するように学習される．例えば, $\pi_t(s, a) = p(s_t, a_t)$ を Actor が状態 s_tで行動 a_tをとる確率だとすると

$$p(s_t, a_t) \leftarrow p(s_t, a_t) + \beta\delta_t \tag{2.11}$$

として方策の学習を行う．ここで, βはステップサイズ変数, δ_t は式 2.5 の TD 誤差である．

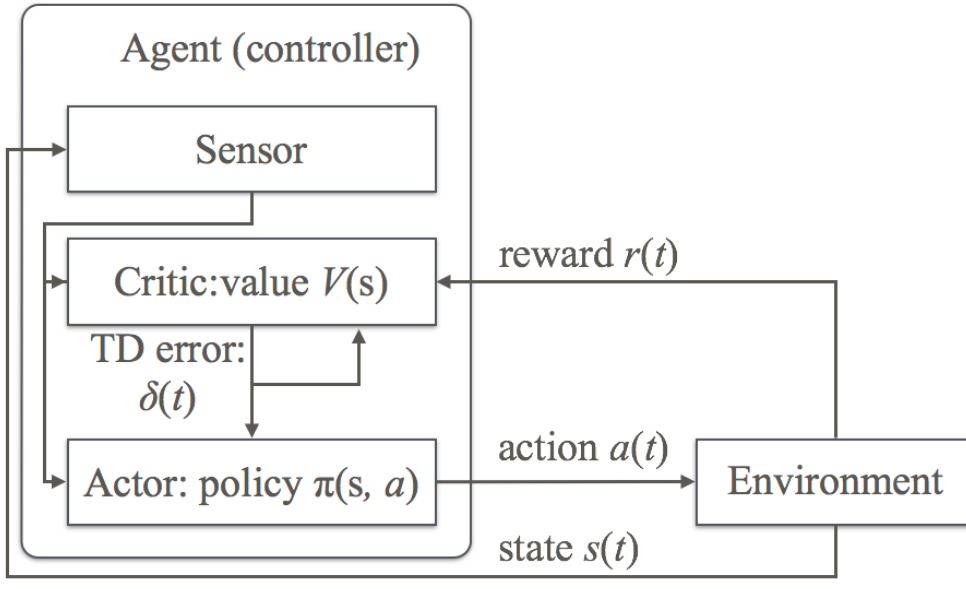


Fig.2.13: Actor-Critic

参考文献

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115–133.

Minsky, M., & Papert, S. (1969). Perceptrons.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Rumelhart, D. E., McClelland, J. L., Group, P. R., & others. (1988). *Parallel distributed processing* (Vol. 1). IEEE.

Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Null* (p. 958). IEEE.

古川正志, 川上敬, 渡辺美知子, 木下正博, 山本雅人, & 鈴木育男. (2012). **メタヒューリスティクスとナチュラルコンピューティング**. コロナ社.

岡谷貴之. (2015). **機械学習プロフェッショナルシリーズ　深層学習**. 講談社.

第 3 章

深層学習による時間領域の信号波形ベースのモデル

3.1 はじめに

2 章で述べた概念モデルは. 音楽や音声に関するデータを記号化せず, 時系列に並んだ周波数成分として表現された音響信号, あるいは音響信号の波形領域そのものを深層学習のメカニズムにより概念学習を行い, そこから新しい楽曲の生成やロボットが感知する音響信号からの行動決定などの応用を目指している. このモデルは強化学習の Actor-Critic 法に深層学習のメカニズムを取り込むことで表現するが, モデル構築の第一の課題は音響信号について, 特徴量の設計と時間方向の依存性解決となる.

本研究では. このモデルを実現するために, まず, 深層学習のメカニズムを取り入れ, 音響信号の特徴抽出と時間軸方向の依存性の解決を行うことのできる多層ニューラルネットワークの構築を検討する. これは, 2 章で述べた概念モデルのモジュール A とモジュール B からなる多層ネットワークの構築に相当する.

本節では, 制約ボツルマンマシン (Restricted Boltzmann Machine; RBM) と Conditional RBM を用いて, 純粋な深層学習の方法論のみの多層ニューラルネットワークの構築を行う.

3.2 提案手法

3.2.1 提案モデル

Fig. 3.1 は本節において提案するモデルの構成である. モデルは特徴抽出に用いる RBM と時間依存の解決に用いる Conditional RBM から構成される. これらは, RBM → Conditional RBM の順で学習を進めるが, 時間依存を考慮しないモデルと時間依存を考慮したモデルを組み合わせた多層ニューラルネットワークと見なすことができる. このモデルにおいての最大の特徴は, 音響信号の前処理について, 振幅の範囲の補正と窓関数を適用することのみを行なう, つまり, 時間領域の信号波形から直接的に生成モデルの獲得を試みることにある.

従来の信号処理においては, 時間領域の信号波形を周波数分析により, 時間-周波数領域に投射し, 周波数成分のシーケンスデータとした上で, さらに特徴量の設計を行う. しかしながら, 時間領域の信号波形そのものから適切に特徴量を抽出できる方法論が確立できれば. 幾つかの処理ステップを省略し計算コストを削減するなど, 有用な方法論が提案できるだろう.

RBM および Conditional RBM はともに生成モデルである. このため, 学習した実音響信号の時系列予測を伴う信号の復元を行い, この精度を見ることでモデルの評価とできるだろう. 精度よく予測を行い元の信号系列を復元するためには, 特徴抽出器とする音響信号の生成モデルの獲得と, 音響信号の時間依存性を内包し内部状態が多層ニューラルネットワーク内に獲得されているとこを要するためである. つまり, 音響信号の時系列予測と信号の複合が精度よく行われるとき, このモデルは目標とする課題の解決ができることを示唆する.

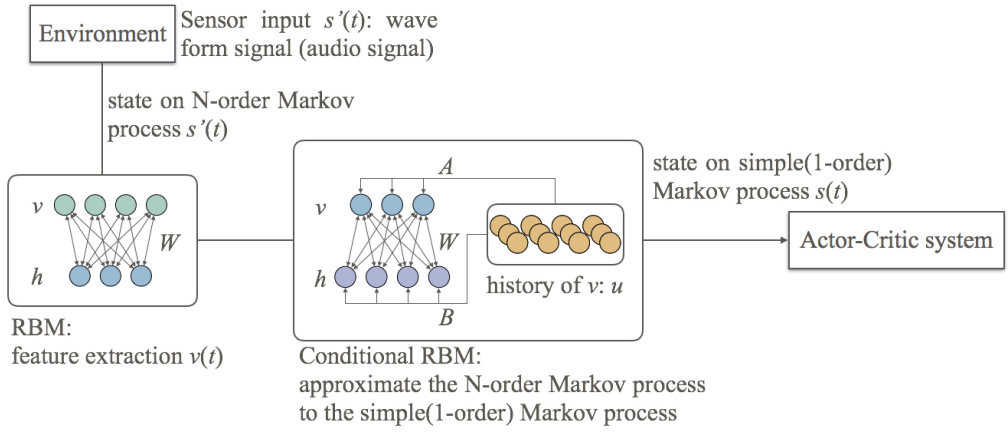


Fig.3.1: Proposal model

3.2.2 Boltzmann Machine

ボルツマンマシン (Boltzmann Machine) は 1980 年代に提案された, それぞれのニューロンユニット間に双方向な結合を持つ相互結合型の人工ニューラルネットワークである. マルコフ確率場 (Markov random field) と呼ばれる無向グラフで表現された確率的なグラフィカルモデルの一種で, ネットワークの挙動を確率的に記述できる.

しかし, ボルツマンマシンは計算量爆発と呼ばれる問題を抱えていたため, これまで機械学習のモデルとして積極的に用いられることはなかった. 学習に要する勾配の計算において, モデルのユニット数 M に対して 2^M の組み合わせの総和をとる期待値の計算項が含まれていたためである. このため, 十分に小さなユニット数のモデルのみ計算可能であり, 実用上大きな課題が残されていた.

しかし 2000 年代初頭に, 制約ボルツマンマシン (RBM: Restricted Boltzmann Machine) と呼ばれる制限のついた構造のボルツマンマシンと, これに対する効果的な近似学習のアルゴリズムが提案された. この制約ボルツマンマシンを用いたディープビリーフネット (Deep Belief Nets) の研究と成功が, 深層学習の概念を構築する重要な第一歩となった.

3.2.3 Restricted Boltmann Machine (RBM)

制約ボルツマンマシン (Restricted Boltmann Machine; RBM)(Freund and Haussler 1994; Hinton et al. 2006) は観測データセットより, 確率変数セット間の依存関係を記述する無効グラフィカルモデルである. Fig. 3.2(a) に示すように, モデルは完全 2 部グラフで表現された結合に制約のあるボルツマンマシンであり, 可視変数 (あるいは観測変数)v は隠れ変数 (あるいは潜在変数)h へリンク結合を持つ. 可視変数と隠れ変数の結合分布 $p(v, h)$ はエネルギー関数 $E(v, h)$ を介して

$$p(v, h) = \frac{1}{Z} e^{-E(v,h)} \tag{3.1}$$

と定義され, 可視変数の確率密度 $p(v)$ は

$$p(v) = \sum_{h} p(v, h) \tag{3.2}$$

となる. RBM の学習の目標はこの $p(v)$ の最尤推定となる. ここで, $Z = \sum_{v,h} e^{-E(v,h)}$ は規格化定数 (あるいは分配関数) である. この規格化定数については ALS(Annealed Importance Sampling)(Neal 2001) によって推定することができる.

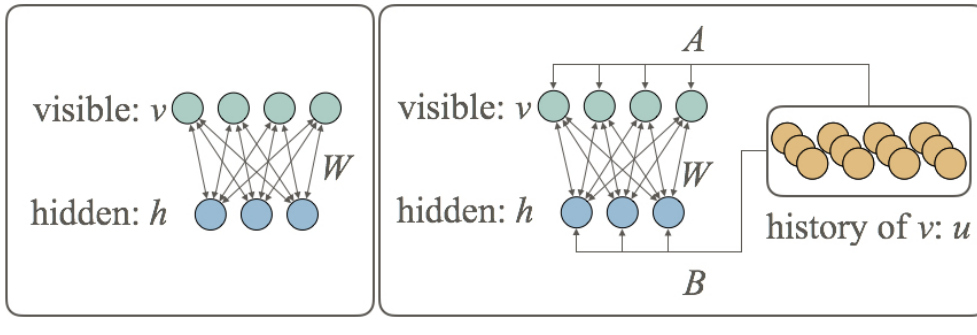


Fig.3.2: Restricted Boltzmann Machine (RBM) and Conditional RBM

Binary visible units and binary hidden units

$v = [v_1, \ldots, v_I], v_i \in \{0, 1\}$ and $h = [h_1, \ldots, h_J], h_j \in \{0, 1\}$ と各変数が 2 値を取る基本的なRBMのエネルギー関数 $E(v, h)$ は

$$E(v, h) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j v_i W_{ij} h_j \tag{3.3}$$

と定義される. v_i, h_j は i 番目の可視変数の状態と j 番目の隠れ変数の状態, a_i, b_j はそれらのバイアスであり, W_{ij} は変数間の結合加重である. RBM は可視変数同士あるいは隠れ変数同士のように同じ層に属する変数間には結合を持たず, これにより, 可視変数と隠れ変数のアクティベーションは相互に条件付き独立となる. 従って, 各変数のアクティベーションを求める条件付き確率 $p(v_i|h)$ と $p(h_j|v)$ はシンプルな関数によって表現される.

$$p(v_i = 1|h) = sigmoid(a_i + \sum_j h_j W_{ij}) \tag{3.4}$$

$$p(h_j = 1|v) = sigmoid(b_j + \sum_i v_i W_{ij}) \tag{3.5}$$

ここで, $sigmoid(x) = 1/(1 + exp(-x))$ は標準シグモイド関数 (standard sigmoid function) である.

Gaussian visible units

自然画像やメル周波数ケプストラム係数 (Mel-Frequency Cepstral Coefficients; MFCC) などの実数値データのとき, 可視変数と隠れ変数がともにバイナリ (2 値) を取る標準的な RBM は, 与えられるデータに適したモデルではない. ここで, 実数値データの分布のモデル化には, 可視変数をガウス分布に従うように拡張したモデル (gaussian-binary model) を適用することができる (Cho et al. 2011; Hinton and Salakhutdinov 2006). $v = [v_1, \ldots, v_I], v_i \in R$ and $h = [h_1, \ldots, h_J], h_j \in \{0, 1\}$ である gaussian-binary モデルのエネルギー関数 $E(v, h)$ は

$$E(v,h) = \sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_j b_j h_j - \sum_i \sum_i \frac{v_i}{\sigma_i} W_{ij} h_j \tag{3.6}$$

と定義され, このときの条件付き確率 $p(v_i|h)$ は

$$p(v_i|h) = N(v|a_i + \sum_j h_j W_{ij}; \sigma_i^2) \tag{3.7}$$

となる．$N(x|\mu;\sigma^2)$ は平均μ分散σ^2 のガウス分布に従う確率密度関数である．また, σ_i^2は i 番目の可視ユニットについての分散を表すパラメータである．

Contrastive Divergence Learning Algorithm

CD_k(Contrastive Divergence)(Hinton 2002) アルゴリズムは, 対数尤度の勾配を近似するための高速計算アルゴリズムである．訓練データが与えられると, RBM のモデルパラメータ $\theta = \{W, a, b\}$ は $p(v)$ の最尤学習によって推定される．このとき, 対数尤度を最大化するモデルパラメータは, 一般に確率的勾配法 (Stochastic Gradient Descent; SGD) を用いて決定する．この対数尤度の勾配は RBM のエネルギー関数 $E(v,h)$ を介して

$$\frac{\partial \ln L(\theta)}{\partial \theta} = -\sum_h p(h|v) \frac{\partial \ln E(v,h)}{\partial \theta} + \sum_{v,h} p(v,h) \frac{\partial \ln E(v,h)}{\partial \theta} \tag{3.8}$$

と与えられる．しかし, 計算コストが指数関数的に増大するため, この勾配を厳密に計算することは困難である．この要因は式 3.8 の第二項の $p(v,h)$ にある．第一項の $p(h|v)$ はデータ依存の成分であり, RBM のネットワーク構造の制約によって容易に計算できるが, 第二項の $p(v,h)$ は定義したモデル依存の総和のために全ての v, h に対する総和を要する．

そこで, CD アルゴリズムでは可視変数と隠れ変数が互いに条件付き独立となる性質を利用し, k ステップのギブスサンプリング (Gibbs sampling) と条件付き確率 $p(v|h)$ と $p(h|v)$ を用いて対数尤度の勾配を近似する．具体

的には, $p(v|h)$ と $p(h|v)$ を用いて, $v^{(0)} \to h^{(0)} \to v^{(1)} \to h^{(1)} \to \cdots \to v^{(k)}$ と可視変数と隠れ変数を交互に k 回サンプリングしていく. これにより与えられる勾配は

$$\frac{\partial \ln L(\theta)}{\partial \theta} = -\sum_h p(h|v^{(0)}) \frac{\partial \ln E(v,h)}{\partial \theta} + \sum_h p(h|v^{(k)}) \frac{\partial \ln E(v^{(k)},h)}{\partial \theta} \tag{3.9}$$

となり, 第二項の計算が可能となる. これにより, モデルパラメータ $\theta = \{W, a, b\}$ について, 各パラメータの更新に用いられる勾配はそれぞれ

$$\frac{\partial \ln L(\theta)}{\partial W_{ij}} = p(h_j|v^{(0)}) v_i^{(0)} - p(h_j|v^{(k)}) v_i^{(k)} \tag{3.10}$$

$$\frac{\partial \ln L(\theta)}{\partial a_i} = v_i^{(0)} - v_i^{(k)} \tag{3.11}$$

$$\frac{\partial \ln L(\theta)}{\partial b_j} = p(h_j|v^{(0)}) - p(h_j|v^{(k)}) \tag{3.12}$$

と求められる. ここで, $v^{(0)}$は訓練データそのもの, $v^{(k)}$は条件付き確率 $p(v|h)$ と $p(h|v)$ を用いた k ステップのギブスサンプリングによって得られる可視変数 v の値である. 基本的には k 回のサンプリングで打ち切るため, CD_k アルゴリズムと呼ばれるが, $k \to \infty$ のとき, 理論上は正確な解に収束すると言われている.

3.2.4 Conditional RBM

Conditional RBM(CRBM)(Taylor and Hinton 2009) は時系列な状態を持つ確率変数セット間の依存関係を記述できるように, RBM を拡張したモデルである. Fig. 3.2(b) に示すように, CRBM は RBM 同様に可視変数 v は隠れ変数 h へ双方向 (無向) のリンク結合を持つ. これに加え, 可視変数 v について, N ステップの状態履歴を保持する変数 $u =$

$[v_{t-1}, v_{t-2}, \ldots, v_{t-N}]$ が可視変数 v および隠れ変数 h へ有向のリンク結合を持つ. ここで, 式を簡単に表現するために 変数 v, h はそれぞれ, 時刻 t の v_t, h_t を表すこととする. CRBM は可視変数の過去の状態をバッファとして考慮することで, 時系列な状態を学習し, また, その時系列データの生成を可能とした RBM の拡張の一つである.

この CRBM において, 可視変数の状態履歴 u が与えられたとき, 可視変数と隠れ変数の結合分布 $p(v, h|u)$ はエネルギー関数 $E(v, h|u)$ を介して

$$p(v, h|u) = \frac{1}{Z} e^{-E(v,h|u)} \tag{3.13}$$

と定義され, 可視変数の確率密度 $p(v|u)$ は

$$p(v) = \sum_h p(v, h|u) \tag{3.14}$$

となる. Z は規格化定数である.

$v = [v_1, \ldots, v_I], v_i \in R$ and $h = [h_1, \ldots, h_J], h_j \in \{0, 1\}$ である実数値入力の CRBM において, 可視変数の状態履歴 u があるときのエネルギー関数 $E(v, h|u)$ は

$$E(v, h|u) = \sum_i \frac{(v_i - \hat{a}_i)^2}{2\sigma_i^2} - \sum_j \hat{b}_j h_j - \sum_i \sum_i \frac{v_i}{\sigma_i} W_{ij} h_j \tag{3.15}$$

と与えられる. ここで W は可視層と隠れ層間の無向接続の重み行列であり, $\hat{a}$ と $\hat{b}$ ダイナミックバイアスと呼ばれ, それぞれ可視変数と隠れ変数のバイアスである. このダイナミックバイアスは 可視変数の状態履歴 u を用いて

$$\hat{a}_i = a_i + \sum_k A_{ki} u_k \tag{3.16}$$

$$\hat{b}_j = b_j + \sum_k B_{kj} u_k \tag{3.17}$$

と与えられる. ここで, A は可視変数の状態履歴変数 u と可視変数 v の結合加重, 同様に B は u と隠れ変数 h の結合加重である. このとき, 状態履歴変数 u の下の条件付き確率 $p(v|h,u)$ と $p(h|v,u)$ は

$$p(v_i|h,u) = N(v|\hat{a}_i + \sum_j h_j W_{ij}; \sigma_i^2) \tag{3.18}$$

$$p(h_j = 1|v,u) = sigmoid(\hat{b}_j + \sum_i v_i W_{ij}) \tag{3.19}$$

となる. ここで, $sigmoid(x) = 1/(1+exp(-x))$ は標準シグモイド関数 (standard sigmoid function), $N(x|\mu;\sigma^2)$ は平均μ分散σ^2 のガウス分布に従う確率密度関数である. また, σ_i^2は i 番目の可視ユニットについての分散を表すパラメータである.

CRBM のモデルパラメータは RBM と同様に CD_k アルゴリズムににによって推定される. 条件付き確率 $p(v_i|h,u)$ と $p(h_j|v,u)$ および k ステップのギブスサンプリング (Gibbs sampling) を用いて与えられる対数尤度の勾配は

$$\begin{aligned}\frac{\partial \ln L(\theta)}{\partial \theta} = \quad & -\sum_h p(h|v^{(0)},u)\frac{\partial \ln E(v,h|u)}{\partial \theta} \\ & +\sum_h p(h|v^{(k)},u)\frac{\partial \ln E(v^{(k)},h|u)}{\partial \theta}\end{aligned} \tag{3.20}$$

と与えられる. 従って, 各パラメータの更新に用いる勾配は次のようになる.

$$\frac{\partial \ln L(\theta)}{\partial W_{ij}} = p(h_j|v^{(0)},u)v_i^{(0)} - p(h_j|v^{(k)},u)v_i^{(k)} \tag{3.21}$$

$$\frac{\partial \ln L(\theta)}{\partial A_k i} = v_i^{(0)}u_k - v_i^{(k)}u_k \tag{3.22}$$

$$\frac{\partial \ln L(\theta)}{\partial B_{kj}} = p(h_j|v^{(0)},u)u_k - p(h_j|v^{(k)},u)u_k \tag{3.23}$$

$$\frac{\partial \ln L(\theta)}{\partial a_i} = v_i^{(0)} - v_i^{(k)} \tag{3.24}$$

$$\frac{\partial \ln L(\theta)}{\partial b_j} = p(h_j|v^{(0)}, u) - p(h_j|v^{(k)}, u) \tag{3.25}$$

3.3 RBM による時間領域の信号波形の特徴抽出と信号の復元

3.3.1 実験設定

本節で提案するモデルでは, 音響信号の周波数分析を行わずに, 信号波形を直接取り扱うことを提案している. 時間領域の信号波形はある時間区間で区切り, その時間区間内の信号波形を内包する幾つものフレームとする. 提案モデルを実現するためには, モデルに用いる RBM が各フレームの内包する時間領域の信号波形から生成モデルを獲得し, また隠れ変数の出力パターンが信号波形の特徴を表現できなければらなない. そこで, まず, 事前実験として信号波形のデータを用いて RBM を訓練し, この結果を評価する.

モデルの訓練データ

訓練データには実環境の音響信号データを用意することが最も望ましいと考えられる. しかし, 信号波形は単純な信号の合成により複雑な混合音の信号であっても表現することが可能である. 実際に, 幾つかの周波数分析手法では基本的にこの仮定の基で周波数分析が行われている. このことから, 実信号の代わりに周期の異なる大量のサイン波を学習に用い, 訓練データセットとすることで, 実環境音の代替データとしながらも多くの混合音に対する汎化性能の獲得が期待できる.

そこで, まずは周波数の異なる複数のサイン波を大量に用意し, RBM による信号波形の生成モデルの獲得を試みる. サイン波のデータは 100Hz から 8kHz まで 1Hz ごとに周波数を設定した 7900 パターンのデータを用い

る. このとき. サンプリングレートは 16kHz とし, 振幅幅は $-5.0 \leq x \leq 5.0$ とした. また, 信号波形データのサンプル数は 512 サンプルとし, これにガウス窓を適用した.

また, 訓練後のモデルを評価するために, 実在の楽曲の信号波形を与え, モデルの出力と復元の精度を確認する. ここで, 楽曲には The Beatles の “Let It Be” を用い, また, この “Let It Be” によって訓練したモデルを別途用意して比較対象とする. 音響信号はサンプリングレート 16kHz, モノラルの信号として再サンプリングし, 振幅の絶対値が最大の値を用いて, 振幅の範囲を $-5.0 \leq x \leq 5.0$ に補正した. また, 信号のデータを 160 サンプル (10msec) でシフトさせながら, 512 サンプルごとに切り出してガウス窓を適用した.

RBM のパラメータ設定

RBM のモデルは Gaussian-Binary 型のモデルを用い, モデルのパラメータは Table 3.1 に示す通りに設定し, この訓練には CD_k アルゴリズムをギブスサンプリングのステップ数を 1 として用いた. また, 各バイアスと重み行列は 平均 0, 分散 0.1^2 の正規分布に従う乱数により初期化した.

Table 3.1: Parameter of RBM

parameter name	parameter value
visible layer size	512
hidden layer size	512
learning rate	0.0001
mini-batch siz	100
learning epoch cycle	1000

3.3.2 結果と考察

信号波形の再構成

訓練後のモデルに実在の楽曲の信号波形を与えモデルの評価を行う. ここでサイン波で訓練したモデルを “model 1”, “Let It Be” で訓練したモデルを “model 2” とする. 評価用に用いる “Let It Be” の分割した信号波形について, その各フレームをモデルに与えたときの隠れ層の出力パターンが Fig. 3.3 である. なお, 図は視認性を考慮し, 各値について絶対値を取り, その最大値を用いて上限が 1 となるように補正した上で作図している. また, 入力信号に対してギブスサンプリングを介して得られた可視変数 v, つまり再構成された信号を Fig. 3.4 にしめす. こちらも同様に信号の絶対値のうち最大のものを利用し, $-1.0 \leq x \leq 1.0$ となるように補正している.

隠れ変数の出力は model 1, model 2 ともに同様のパターンを示しており, 周波数の異なる大量のサイン波を用いることで, 実音響信号の代替とできる可能性が示されている. また, 入力信号と隠れ変数の出力パターンを比較すると, 入力信号の強度に応じて出力強度が強くなっている様子が見られる. さらに, 入力信号と再構成された信号を比較すると, いずれも再構成後にはノイズが加わっているものの良好な精度で信号が復号されていることがわかる. 特に, サイン波を用いて訓練した model 1 は実際の楽曲の信号を用いて訓練した model 2 よりも, 元の信号を再現できている. しかしながら, これについてはオーバーフィッティングが懸念されるため, 追加の実験により確認が必要であろう. これにより, たとえモデルに与える音響信号が実環境のデータでなくとも, 周波数の異なる多数のサイン波のデータを用いることで, 音響信号の生成モデルが獲得できることが示唆され, また隠れ層の出力が特徴量として機能することが示唆される.

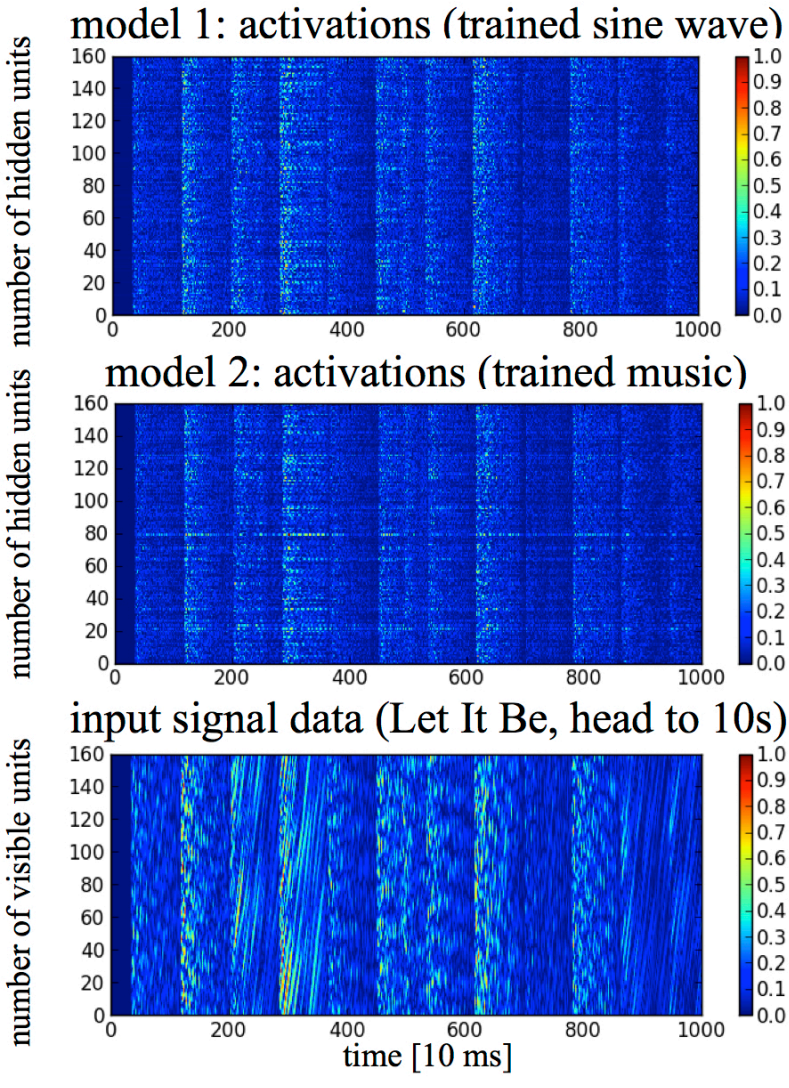


Fig.3.3: Feature vector patterns: input data is "Let It Be"

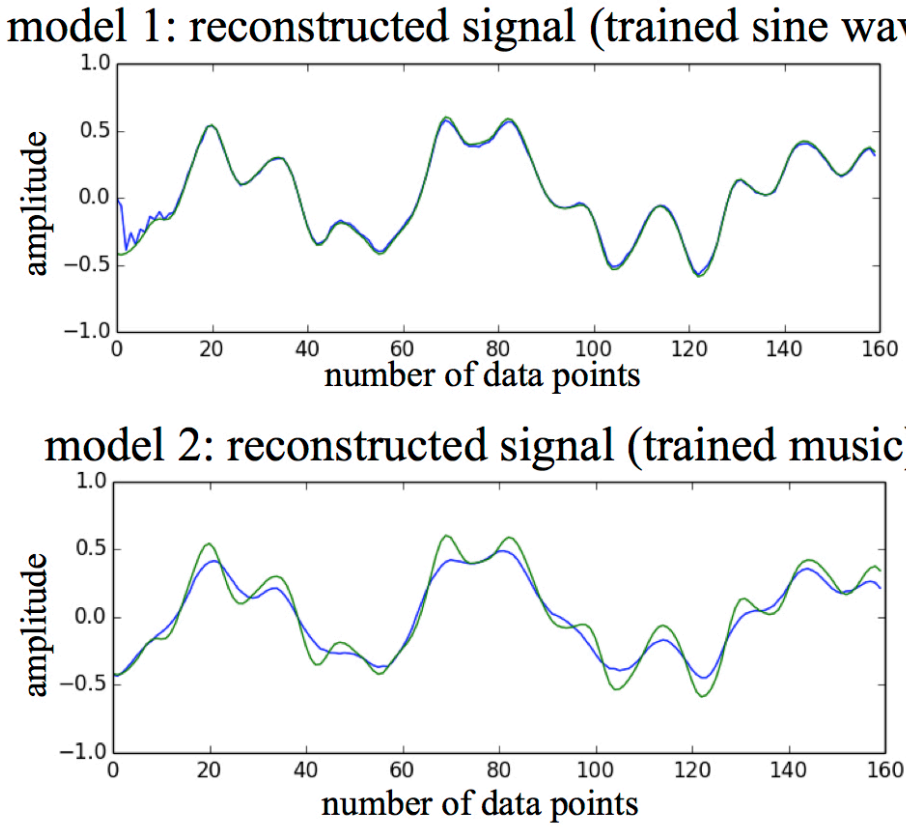


Fig.3.4: Reconstracted wave form data given ”Let It Be”: green lien is original wave form, blue line is output from model

データ生成モデルとしてのノイズ耐性

RBM はグラフィカルモデルの一つであるため, 学習したデータの確率分布は新しいデータの生成に用いることができる. 具体的には, 隠れ層の変数に任意の値をセットし, 可視層の変数の値を計算すれば良い. このように, 特徴抽出器としての用途以外にもデータの生成モデルとして用いることができる.

例えば, Actor-Ctritic 法によるシステムにおいて, Actor の出力を RBM の隠れ層の値とするシステムを想定する. このとき, 出力として得られる値には学習時の誤差によりノイズが加わることが懸念される. このためデータ

の生成モデルとしてのロバスト性について, 特徴ベクトルにランダムな値を加算しノイズの影響を調べる.

モデルの隠れ層から得られる特徴ベクトルからランダムに 32,64,128,256 個の値を選択し, これに $-0.3 \leq x \leq 0.3$ の範囲のランダムな値を加算する. この値を隠れ層の変数にセットして可視層の変数の状態を再計算し信号の再構成を行う. 特徴ベクトル, ノイズおよびノイズを加えた特徴ベクトルから再構成した信号のサンプルが Fig. 3.5 である. 隠れ層の変数の 50% にあたる 256 の変数に対してランダムな値を加算したときにおいても, 元の信号が判別可能な程度に再構成された信号が得られている. ここで, 再構成された信号に注目すると, 特徴ベクトルに対するノイズは再構成後の信号において, 元の信号の復元が不可能になるのではなく, 信号に加わるノイズの強度として現れることが確認された.

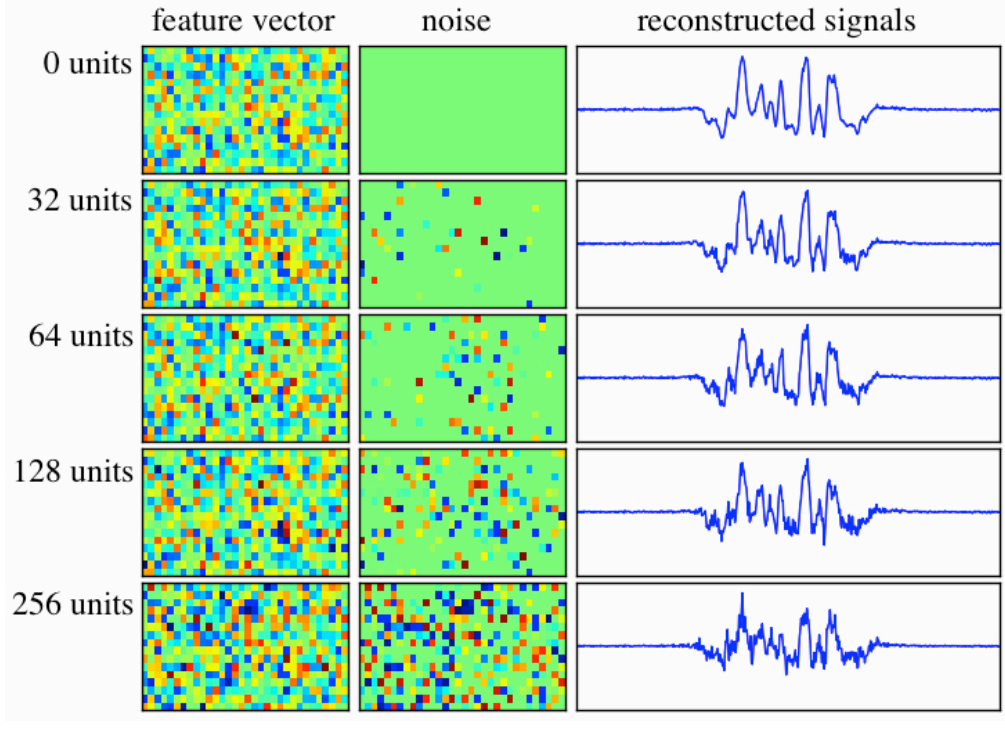


Fig.3.5: Sample of reconstructed signal

3.4 RBM および Conditional RBM による時間領域の信号波形の時系列予測

3.4.1 実験設定

本節で提案する Fig. 3.1 に示すモデルについて, これを構成し評価を行う. 各コンポーネントは, 制約ボルツマンマシン (Restricted Boltzmann Machine; RBM) による特徴抽出器と Conditional RBM から構成されるモデルとなる. ここで, RBM の隠れ層 h と CRBM の入力層 v を共有するものとすると, 入力から出力までを RBM と CRBM からなる多層ニューラルネットとすることができる.

このモデルについて, 実音響信号の時系列予測を伴う信号の復元を行い, この予測精度と信号の復元の誤差を見ることでモデルの評価を行う.

3.4.2 データセット

モデルの訓練用データセットには, 実在の楽曲の音響信号を用いた. 楽曲はそれぞれ, The Beatles の"Let It Be", Vivaldi の" 和声と創意への試み" より" 四季" として知られている" 春", " 夏", " 秋", " 冬" の各曲である. " 四季" の 4 曲はそれぞれ第一から第 3 までの 3 つの楽章で構成されるため, 各楽章を 1 曲と扱うと計 12 曲分のデータとなる. "Let It Be" は 1 曲分すべて, 240 秒の音響信号データを用い, また, "四季" の各 12 曲については, 冒頭 0 秒から 15 秒までの時間区間で切り出し, これを順番に並べて結合した 180 秒間の音響信号データとして編集したものを用いた.

これらの音響信号は 16kHz / 16bit, 1channel(モノラル信号) となるように再サンプリングし, 絶対値が最大の値を用いて, $-5.0 \leq x \leq 5.0$ の範囲に収まるように補正した. 補正後の信号波形は, 16,000 サンプル (1000ms; 1 秒) の時間区間で分割し, この 16,000 サンプルの区間一つを 1 単位のフレームとした. これについて, 窓関数の適用は行っていない.

3.4.3 モデルのパラメータ設定

モデルを構成する RBM と CRBM のパラメータは, それぞれ Table 3.2 と Table 3.3 に示す通りに設定した. 可視変数と隠れ変数間の結合加重は 平均 0, 分散 0.1^2 の正規分布に従う乱数により初期化した. このとき, パラメータの数値の範囲は $-0.5 \leq x \leq 0.5$ となる. 可視変数の状態履歴を保持する変数 (状態履歴変数) から可視変数間, 状態履歴変数から隠れ変数間の結合加重と各バイアスの初期値はそれぞれ 0 とした. 各コンポーネントは CD_k 法により訓練を行い, ギブスサンプリングのステップ数は 1 と設定した. また, モデルの構成は Fig. 3.6 に示す.

Table 3.2: Parameter of RBM

parameter name	parameter value
visible layer size	1600
hidden layer size	200
type of visible laye	gaussian unit
type of hidden laye	binary unit
learning rate	0.0001
mini-batch siz	100
learning epoch cycle	6000

Table 3.3: Parameter of Conditional RBM

parameter name	parameter value
visible layer size	200
hidden layer size	200
history of visivle layer	1400 (7 histories)
type of visible laye	binary unit
type of hidden laye	binary unit

parameter name	parameter value
mini-batch siz	100
learning epoch cycle	6000

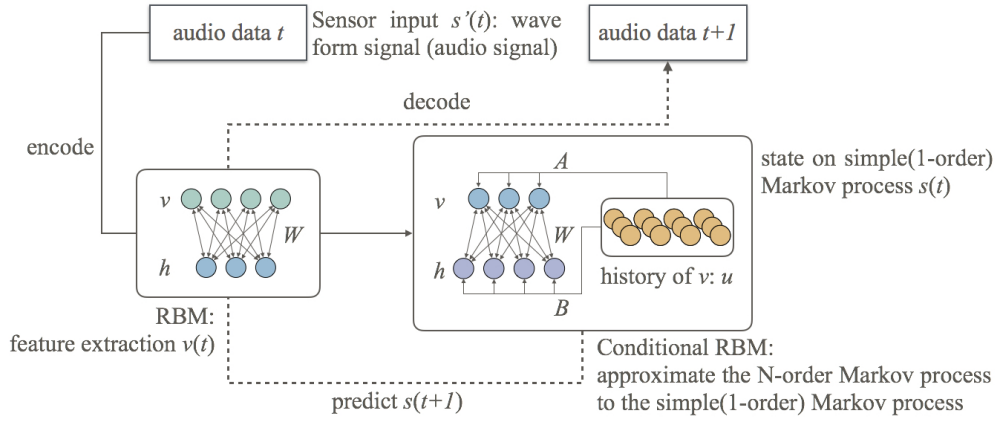


Fig.3.6: Model structure

3.4.4 結果と考察

歌唱を含む楽曲を訓練データとしたケース

まず, The Beatles の “Let It Be” の 240 秒間の信号波形について, モデルに学習させた結果を述べる.

訓練後のモデルに冒頭 0 秒から 10 秒間までを与え, その後の信号波形を予測させた. 信号波形について結果の比較のために, 訓練に用いた信号波形と予測により得られた信号波形をそれぞれ短時間フーリエ変換 (short-time Fourier transform; STFT) により周波数分析を行った. この周波数分析結果のパワースペクトログラムが Fig. 3.7 であり, 上段が訓練用の音響信号, 下段がモデルにより得られた音響信号である.

各信号波形の周波数分析結果より, モデルから予測と復元を介して得られた信号波形は訓練データとした音響信号を比較的良く予測・復元できていることがわかる. パワースペクトル上では, 復元された音響信号には訓練用に

用いた信号にはない周波数成分が全域にわたって加わっている．これは, 白色ノイズなど非定常, 非周期的なランダムな成分が信号に乗った時に見られる現象である．つまり, 復元された信号波形に何らかのノイズが含まれていることを示す．これは, 3.3 節の事前実験結果において, 信号波形を描画したグラフである Fig. 3.4 に見られるような, 信号波形を復元する段階での誤差が本節の実験においても見られることを示唆する．

ここで, 被験者による音響信号の聴取実験を行った．対象の音響信号はモデルにより予測・復元された “Let It Be” の音響信号であり, 対象者には, モデルにより予測・再構成された音響信号であるということ以外は一切の説明を行っていない．この結果, 予想通りに白色ノイズが観測されたが, その強度は人間の聴覚上では気にならないほどに弱く, 原曲がはっきりと判別できる程度であった．また, この聴取実験により白色ノイズらしき成分が含まれている以外は, 原曲の音響信号を忠実に予測・復元できていることが確認された．この結果において興味深い点は, ある被験者において, ノイズの原因がイヤホンの故障や劣化が原因であると誤認したケースが見られたことである．この被験者の回答を考慮すると, 提案モデルは訓練い用いた信号波形の時系列データを忠実に記憶・生成できる生成・記憶系のモデルが獲得されていることを示唆する．

これらの結果より, 訓練パラメータの調整, あるいは予測・復元後の信号に対してノイズ成分を除去する方法論を適用によって, より忠実に時間領域の信号波形を予測・復元できる生成モデルの実現が示唆されているものと考えられる．

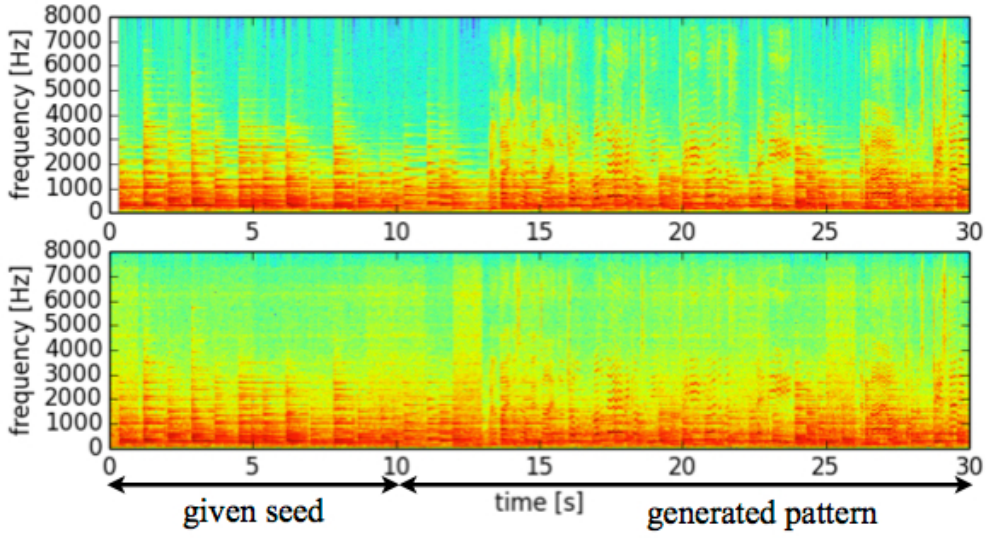


Fig.3.7: Power spectrogram of the original music (“Let It Be” by The Beatles), and the reconstructed music from predicted audio signal by proposul model. Upper image is spectrogram of the original audio signal (training data), and lower image is spectrogram of the reconstructed music from predicted audio signal by proposul model. Where parameters forshort-time Fourier transform (STFT): window size is 512 samples, overlap size is 0 samples, and window function is hamming window.

歌唱を含まない楽曲を訓練データとしたケース

次に, Vivaldi の” 和声と創意への試み” より” 四季” として知られている” 春”, ” 夏”, ” 秋”, ” 冬” の各曲を編集した 180 秒間の信号波形について, モデルに学習させた結果を述べる.

前項の実験結果 A 同様に, 訓練後のモデルに冒頭 0 秒から 30 秒間までを与え, その後の信号波形を予測させた. 信号波形について結果の比較のために, 訓練に用いた信号波形と予測により得られた信号波形をそれぞれ短時間フーリエ変換 (short-time Fourier transform; STFT) により周波数分析を行った. この周波数分析結果のパワースペクトログラムが Fig. 3.8 であり, 上段が訓練用の音響信号, 下段がモデルにより得られた音響信号である.

モデルに与える予測のトリガーとなる初期値の信号波形の時系列データについて, 冒頭 0 秒から 30 秒間よりも短時間であった場合に, 予測処理がうまく行われないことを確認している. また, モデルより得られた音響信号について, 約 7 秒間ほど時間軸上で後方にズレが生じた. これにより, 学習させる信号波形データについて, 予測処理のために初期に与える時間区間が訓練対象の音響信号によって定常ではない可能性が示唆され, また, 信号波形の予測処理を行う上で, 後続の信号波形を予測するためにトリガーとなる内部状態の存在が懸念される.

この他の結果については, 前項の実験結果 A と同様の傾向を示した. しかしながら, Fig. 3.8 より, The Beatles の “Let It Be” を訓練データとした実験結果 A と比較すると, 予測・復元された信号にはより強い強度のノイズが含まれていることがわかる. 被験者を対象とした聴取実験においても, ノイズの強度が強いことを確認しているが, 対象の原曲を聞き取れないレベルのノイズ強度ではないことも確認されている. このような傾向は, 訓練に用いる音響信号の信号波形のデータセットに応じて, 適したモデルパラメータの存在が示唆されているものと考えられる.

これにより, 時間領域の信号波形を提案モデルで学習させる場合に, 訓練対象の音響信号データセットについて, 信号波形のデータに依存せず, 良くモデルを学習させるためのパラメータの推定を要することが, 新たな課題として見えてきた.

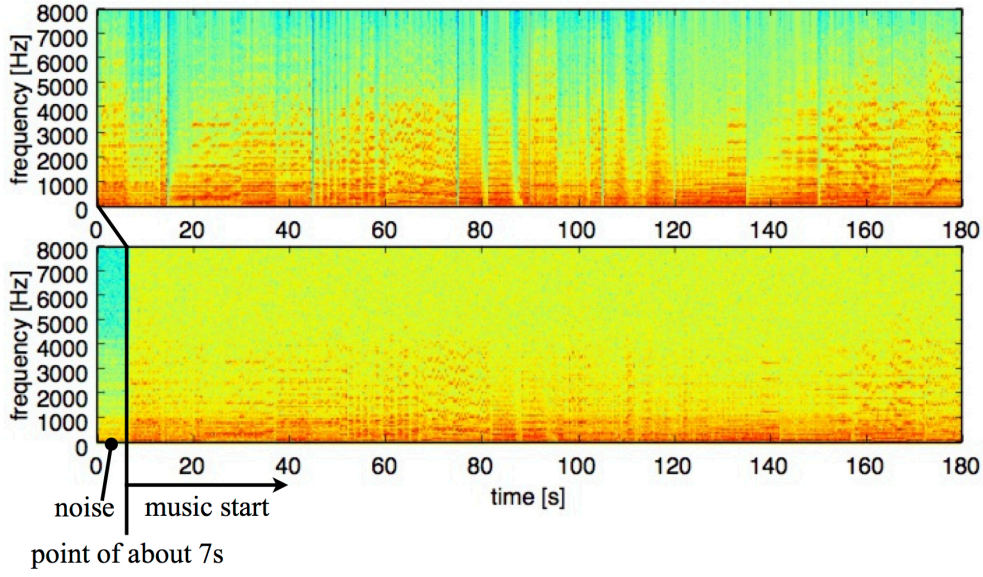


Fig.3.8: Power spectrogram of the original music (fragments of 180 seconds in "The Four Seasons"), and the reconstructed music from predicted audio signal by proposul model. Upper image is spectrogram of the original audio signal (training data), and lower image is spectrogram of the reconstructed music from predicted audio signal by proposul model. Where parameters forshort-time Fourier transform (STFT): window size is 512 samples, overlap size is 0 samples, and window function is hamming window.

参考文献

Cho, K., Ilin, A., & Raiko, T. (2011). Improved learning of gaussian-bernoulli restricted boltzmann machines. In *Artificial neural networks and machine learning–ICANN 2011* (pp. 10–17). Springer.

Freund, Y., & Haussler, D. (1994). Unsupervised learning of distributions of binary vectors using two layer networks. Computer Research Laboratory [University of California, Santa Cruz].

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, *14*(8), 1771–1800.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527–1554.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, *11*(2), 125–139.

Taylor, G. W., & Hinton, G. E. (2009). Factored conditional restricted boltzmann machines for modeling motion style. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1025–1032). ACM.

第 4 章

深層学習と聴覚フィルタ, 及び Echo State Network によるモデル

4.1 はじめに

2 章で述べた概念モデルは. 音楽や音声に関するデータを記号化せず, 時系列に並んだ周波数成分として表現された音響信号, あるいは音響信号の波形領域そのものを深層学習のメカニズムにより概念学習を行い, そこから新しい楽曲の生成やロボットが感知する音響信号からの行動決定などの応用を目指している. このモデルは強化学習の Actor-Critic 法に深層学習のメカニズムを取り込むことで表現するが, モデル構築の第一の課題は音響信号について, 特徴量の設計と時間方向の依存性解決となる.

本論では. このモデルを実現するために, まず, 深層学習のメカニズムを取り入れ, 音響信号の特徴抽出と時間軸方向の依存性の解決を行うことのできる多層ニューラルネットワークの構築を検討する. これは, 2 章で述べた概念モデルのモジュール A とモジュール B からなる多層ネットワークの構築に相当する.

本節では, 積層型自己符号化器 (stacked Auto Encoder; sAE) と

ESN(Echo Stete Network) からなる多層ニューラルネットワークに加え, 周波数分析器として聴覚フィルタを導入したモデルの検討を行う.

4.2 提案手法

4.2.1 提案モデル

Fig. 4.1 は本節において提案するモデルの構成である. モデルは特徴抽出に用いる sAE と時間依存の解決に用いる ESN に加え, 生物の聴覚系の特性を考慮した周波数分析を行う聴覚フィルタバンクから構成される. 与えられた時間領域の信号は聴覚フィルタバンクによって, 周波数成分のシーケンスデータとして得られる. 各コンポーネントの学習は sAE $\rightarrow$ ESN の順で進めるが, 時間依存を考慮しないモデルと時間依存を考慮したモデルを組み合わせた多層ニューラルネットワークと見なすことができる. このモデルにおいての特徴は, 音響信号の処理について, 生物の聴覚抹消系で行われている周波数分析の特性を取り入れていること, また ESN の導入による時系列の扱とリカレントニューラルネットワークの訓練の簡易化が可能である点であろう.

IEEE AASP Challenge に設けられている実音強信号を用いたイベント検出タスクを提案モデルで実行し, この結果を見ることでモデルの評価とする. 特に精度よくイベント検出ができるとき, 音響信号の特徴量の自動獲得や時間依存性を解決できていることが示される.

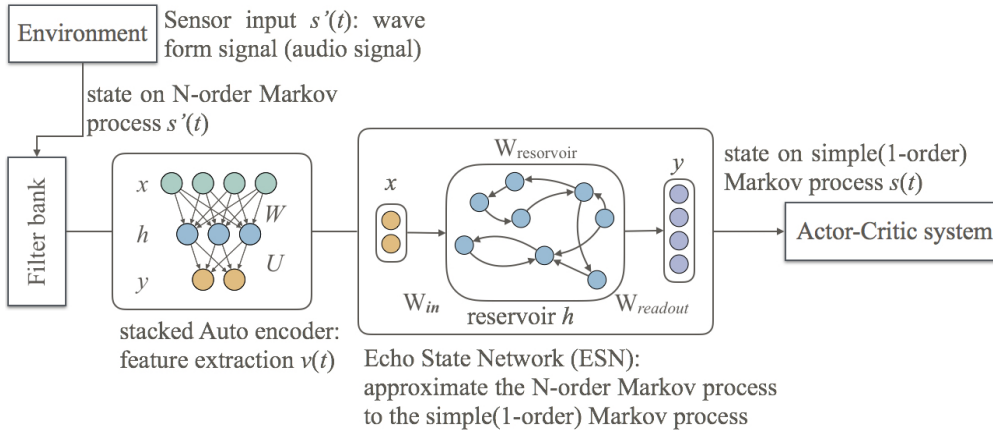


Fig.4.1: Proposal model

4.2.2 Auto Encoder

自己符号化器 (Auto Encoder; AE) はシンプルな構成の人工ニューラルネットワーク (Artificial Neural Network) による教師なし学習の方法論である. データをよく表現する特徴の自動的な獲得や深層学習の事前学習に用いられる.

Fig. 4.2 に従来の 3 層の順伝搬型 ANN と AE を示す. なお, 図中左が順伝搬型 ANN, 右が AE である. AE の構造は標準的な 3 層構造の ANN と全く同様の構造を持つが, 決定的な差異は教師信号に与える信号にある. 通常, 一般的な ANN では 目標 d に対して出力 y'の誤差が最小になるようにネットワークを訓練する. これに対し, AE では 目標 d の代わりに入力 x を目標として出力 x'との誤差が最小となるようにネットワークを訓練する. つまり, 通常の多層ニューラルネットワークが出力 y'と教師信号 d の誤差

$$E = \frac{1}{2}||d - y'||^2 \tag{4.1}$$

を最小化するようにネットワークを訓練するのに対し, AEでは多層ニューラルネットワークの出力 x'と入力 x の誤差

$$E = \frac{1}{2}||x - x'||^2 \tag{4.2}$$

を最小化するようにネットワークを訓練する．ここで, 入力層から隠れ層のブロックを符号化 (encode) や符号化器 (encoder), 隠れ層から出力層のブロックを復号化 (decode) や復号化器 (decoder) という．

このように入力を復元するようにネットワークを訓練することで, 隠れ層のベクトルとしてデータの良い表現が自動的に獲得される．一般には入力層より隠れ層のユニット数を少なく設定した, 砂時計型のネットワーク構造がよく用いられる．隠れ層のユニット数を少なくすることで, より少ないユニットでデータの表現を行うように訓練されるため, 高次元のデータを次元を落とした空間で表現する写像が得られる．これは, 実質的には主成分分析と等価の情報処理となる．

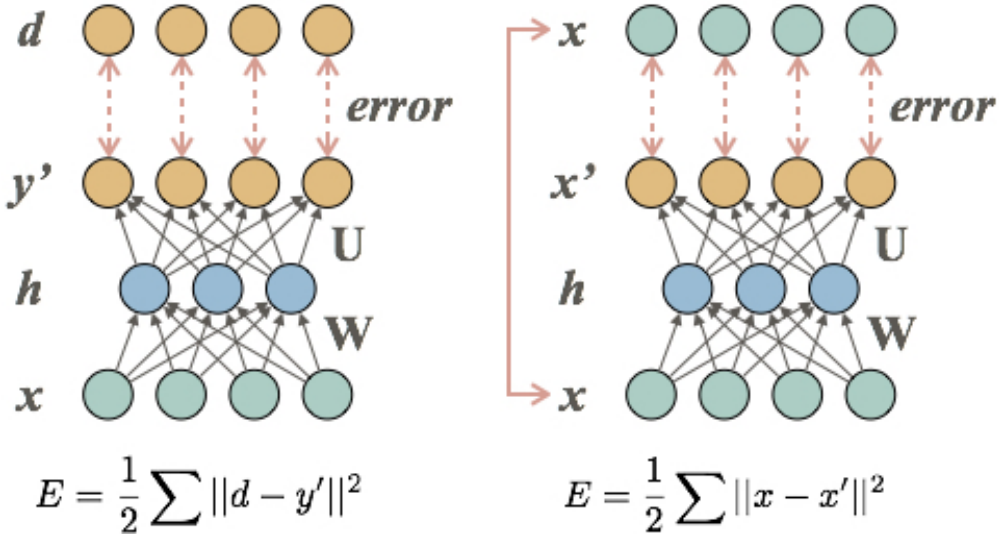


Fig.4.2: Standerd ANN (left) and Auto Encoder (right)

4.2.3 Echo State Network

リカレントニューラルネットワーク (RNN) の学習は単純な構造を持つフィードフォワード (順伝搬型) ニューラルネットワークと比べ, 訓練は著し

く困難である. しかし, RNN は時系列データといった, データが時間や順番に依存する場合, つまり N 次のマルコフモデル (N order markov model) に強く, 記憶を表現出来ることから多くの研究がなされてきた.

エコーステートネットワーク (Echo State Network; ESN) も RNN の一つではあるが, 再帰的な結合を持つ (リカレントな) 部分をブラックボックスとし, 中間層 → 出力層間の結合加重のみを学習の対象とする. RNN となっている中間層をブラックボックスとして学習対象としないことで, モデルの状態遷移や時間依存を考慮した勾配を用いる必要がなく, 学習は中間層 → 出力層間の訓練のみであるので, ESN の訓練自体が実質的には 2 層のフィードフォワードネットワークの訓練として扱うことができる. これにより, 計算量コストを大きく削減するとともに, RNN 特有の状態遷移の取り扱いが容易になっている.

基本的な構造の ESN は Fig. 4.3 に示すように, 入力層 x, リザーバと呼ばれる中間層 h, および出力層 y からなる. 中間層では, 各ニューロンの間の結合は疎になるようにランダムに設定され, その結合加重もまた乱数により設定される. 一般的な ESN では中間層を訓練の対象とはしないので, 中間層のモデルパラメータはこれ以降変化しない. 中間層のニューロンは自身への結合や他のニューロンを介した再帰的な結合を許容するので, 中間層自信がリカレントニューラルネットワークの構造を持つ. 入力層はこの中間層へ入力データを適用し, 出力層では中間層の状態を読み出して出力目標との対応をとるように訓練される.

ここで, 中間層のリーザバの状態は入力として与えらえるデータと自身の前の状態にのみ依存して決定される. この ESN において, ある時刻 t の中間層 (リザーバ) の状態は

$$h_t = f(W_{resorvoir} h_{t-1} + W_{in} x_t + a) \tag{4.3}$$

と与えられる. ここで, x : 入力ベクトル, h : リザーバ内ユニットの状態ベクトル, a : リザーバ (中間層) のバイアス, W_{in} : 入力層とリザーバ間の結合加重, $W_{resorvoir}$: リザーバ内のユニット間の結合加重, $f(x)$: ニューロンの活

性化関数であり, ここでは $f(x) = max(0, x)$ の $ReLU$ 関数としている. また, 中間層から出力層へのベクトルは 時刻 t の中間層の状態より

$$y_t = f(W_{readout} h_t + b) \tag{4.4}$$

となる. ここで, h : リザーバ内ユニットの状態ベクトル, y : 出力ベクトル, b : 出力層のバイアス, $W_{readout}$: リザーバと出力層間の結合加重, $f(x)$: ニューロンの活性化関数であり, ここでは $f(x) = max(0, x)$ の $ReLU$ 関数としている.

中間層 $\rightarrow$ 出力層間の結合加重はフィードフォワードネットワーク同様に確率的勾配降下法 (stochastic gradient descent; SGD) を用いて訓練される. このとき, あらかじめ入力データを順番に中間層へ適応し, 中間層の状態をバッファとして貯めておく. これにより, 中間層の状態集合 $H = [h_t, h_{t-1}, \ldots]$ と教師信号の集合 $D = [d_t, d_{t-1}, \ldots]$ についての時系列データの学習をフィードフォワードネットワーク同様に, 例えば

$$E = \frac{1}{2}||h_t - d_t||^2 \tag{4.5}$$

として平均二乗誤差 (mean squared error; MSE) の最小化問題などとして解くことができる.

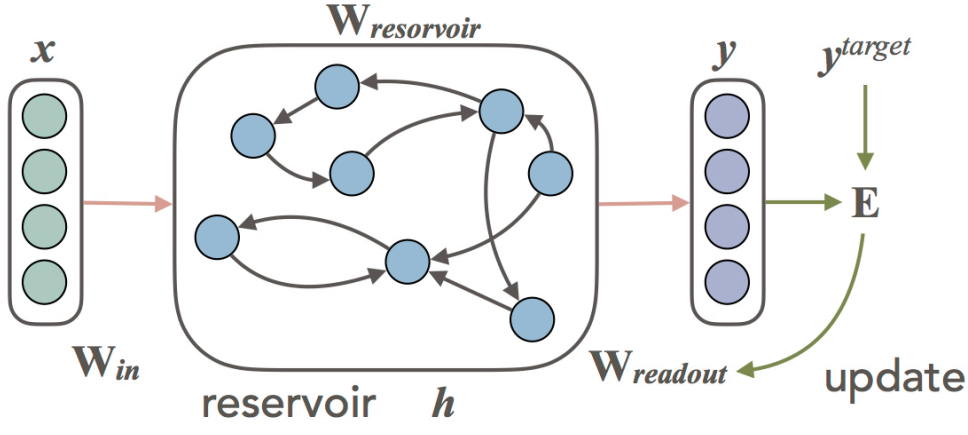


Fig.4.3: Structure of Echo State Network

4.2.4 聴覚フィルタ

聴覚フィルタは聴覚の周波数分析について, その生理的機能を一連の帯域通過フィルタと捉えたモデルである.

聴覚系の研究においては, 聴覚の末端には蝸牛と呼ばれる蝸牛に似た形の器官があり, この蝸牛にある基底膜が音の周波数分析を行う器官であることがわかっている. まず, 耳に伝わった振動 (音響信号) は外耳から中耳を経て内耳に到達する. 内耳の中には蝸牛があり, この蝸牛内の基底膜に音の振動が伝達される. 基底膜上には, ある特定の周波数に共振する有毛細胞, 及び有毛細胞と接触することで神経インパルスを脳へ伝達する神経細胞からなる神経系がある. この基底膜上にある伝達神経系は, 複数の周波数に対応した幾つもの伝達神経が周波数順に並んでいることが確認されている. このように基底膜での機械的な振動を介して聴覚の末端では周波数分析が行なわれている.

信号処理の分野には周波数分析手法の一つとしてフィルタバンクを用いた分析がある. フィルタバンク (Filter bank) はバンドパスフィルタを周波数順に並べたアレイである. フィルタバンクによる周波数分析では, 各バンドパスフィルタごとにフィルタの通過帯域に従って入力信号から特定の周波数帯域成分が取り出され, 複数のコンポーネントに分割される. このとき, コンポーネントの数はフィルタバンクの持つフィルタの数と同数となる.

基底膜の様子をこのような信号処理の観点から捉えると, 基底膜の位置ごとに帯域幅と中心周波数の異なるバンドパスフィルタが並んでいるアレイとしてモデル化できる. つまり, 多数の帯域通過フィルタからなるフィルタバンクとして定式化することができる. このように聴覚の特性をフィルタとして捉えた個々のフィルタは聴覚フィルタ (Auditory filter), これを周波数帯域順に並べて基底膜のモデルとして表現したフィルタバンクは聴覚フィルタバンク (Auditory filter bakn) と呼ばれている.

聴覚の生理反応では, 低い周波数帯では帯域幅が狭く周波数分解能が高くなり, 逆に高い周波数帯では帯域幅が広く周波数分解能が低くなる. このた

めにフーリエ変換とは異なり, 周波数の分析結果が対数スケールで表現されることになる. 聴覚フィルタは, このような聴覚の特性を適切に処理できるように設計されている.

ガンマトーンフィルタ

ガンマトーンフィルタは聴覚のフィルタ関数を近似するために導入されたフィルタである. 聴覚のフィルタ関数を近似するモデルとしてroex(rounded exponential) フィルタが Patterson によって導入された. この roex フィルタはフィルタ形状の中心周波数に対する非対称や音圧による変化といった, 聴覚の特性がよくモデル化されていたため広く用いられていたが, 周波数領域においての重み関数であるためにインパルス応答を持たない. このため聴覚抹消系の時間的なフィルタは表現できなかった. そこで, roex フィルタをさらに近似するフィルタ関数として, インパルス応答のあるガンマトーンフィルタ (gammatone filter) が導入された.

このガンマトーンフィルタは中心周波数 f_cにおける短形帯域幅 $ERB_N = 24.7 \times (\frac{4.37 f_c}{1000} + 1)$ を用いて次のように定義される.

$$g(t) = at^{n-1} exp(-2\pi b ERB_N(f_c)t) cos(-2\pi f_c t + \phi) \tag{4.6}$$

ここで,t : 時間 $(t > 0)$, f_c : 中心周波数, a : 振幅, b : 係数, ϕ : 位相である. また, 中心周波数によって変化する帯域幅のフィルタが等間隔に並ぶように ERB_N番号 $(ERB_{Nnumber}) = 21.4 \log 10(\frac{4.37 f_c}{1000} + 1)$ が定義されている. このガンマトーンフィルタの形状はインパルス応答を周波数領域に射影することで確認でき, Fig. 4.4 に示すフィルタ形状となる.

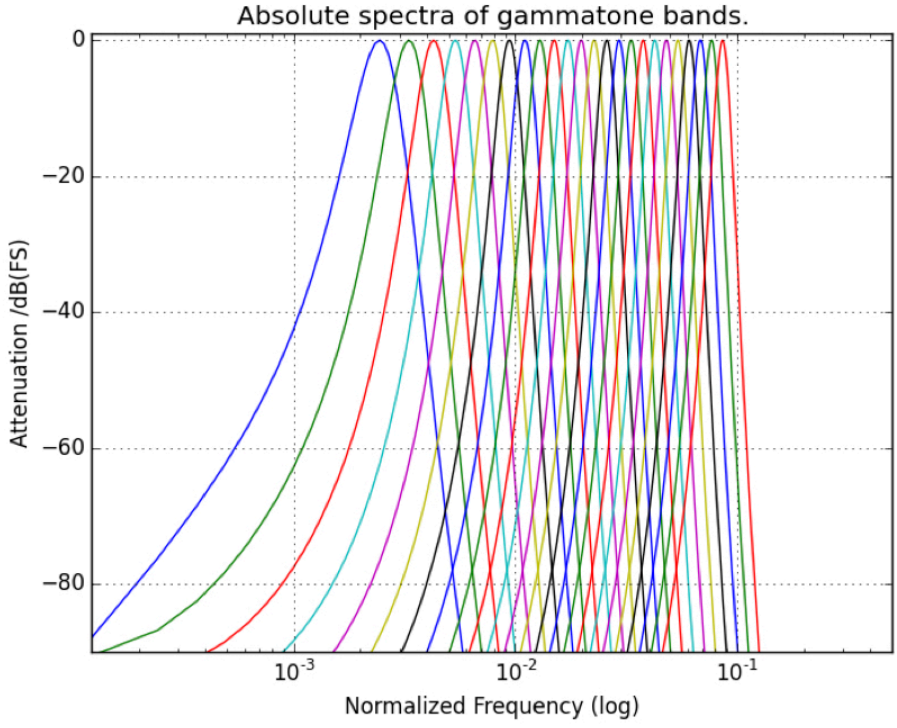


Fig.4.4: Filter shape of gammatone filters (gammatone filter bank)

ガンマチャープフィルタ

ガンマチャープフィルタはガンマトーンフィルタを拡張し, 聴覚の特性をより良く表す聴覚フィルタのモデルとして, Irino と Patterson らにより提案された (Irino and Patterson 1997). 生物の聴覚応答はフィルタ系として捉えると, 周波数特性が中心周波数に対して非対称なフィルタ形状を持つが, 標準的なガンマトーンフィルタはこの非対称性を近似できない. ガンマチャープフィルタは信号処理の最適性の観点の考察を取り入れ, 解析的に導出された関数系であるが, 周波数変化 (*chirp*) 系数 c を入力音圧の関数とすることで, 非対称でレベル依存性のあるフィルタ形状を良く近似できるように改良されている.

ガンマチャープフィルタもガンマトーンフィルタと同様に, 漸近周波数 f_r における短形帯域幅 $ERB_N = 24.7 \times (\frac{4.37 f_r}{1000} + 1)$ を用いて次のように定義

される.

$$g_c(t) = at^{n-1}\exp(-2\pi bERB(f_r)t) \times \cos(2\pi f_r t + c\ln t + \phi) \qquad (4.7)$$

ここで, t : 時間 $(t > 0)$, f_r : 漸近周波数, c : 周波数変化 $(chirp)$ 系数, ϕ : 位相系数である.

4.3 聴覚フィルタバンクと RBM による音響信号からの特徴抽出

4.3.1 実験設定

ESN の特性より, 聴覚フィルタを通して得られる周波数成分のシーケンスデータは, 教師なし学習の方法論で訓練を行わなければならない. 深層学習において, 教師なしには事前学習に用いられる AE か RBM を用いることが一般的である. そこで, 事前実験として, 聴覚フィルタを通して得られる周波数成分のシーケンスデータの学習を RBM により試みる.

聴覚フィルタによる音響信号の周波数分析

まず, ガンマチャープフィルタを用いた次の手順で音響信号の処理を行い, 周波数分析結果のデータについて訓練用と評価用のデータを得る.

1. まず, 短時間フーリエ変換 (sshort-time Fourier transform; STFT) により周波数成分が並んだシーケンスデータ (スペクトログラム) を得る, ここで, 4096 サンプル (92.8ms) のハニング窓を用い, 2048 サンプル (46.4ms) のオーバーラップとした
2. スペクトログラムから有効範囲の振幅スペクトル (2048 次元) を取り出し, 聴覚フィルタバンクの各フィルタを周波数領域に変換したものを用いて重み付けを行う, ここで, 各フィルタの中心周波数は 10Hz から 10kHz の間で対数スケール上に等間隔となるように 128 枚分設定した

3. フィルタバンク分析で得られた各値を平均 0, 分散 1 となるように正規化する
4. これを時刻 10 ごとにまとめ, これを 1 つのフレーム単位として全体を分割する

このガンマチャープフィルタによる周波数分析後の訓練用データを用いて, RBM を訓練する. このとき, 訓練後の RBM では隠れ層出力パターンが周波数成分の特徴ベクトルを表現する.

データセット

訓練用のデータセットには公開されている MIDI ファイルのセット (“Classical piano midi page” n.d.; Poliner and Ellis 2007) より, MIDI ファイルから音響信号を録音して用いる. 録音時した音響信号は, サンプリング周波数を 44.1 kHz, チャンネル 1 のモノラル信号に再サンプリングし, また, 冒頭より 60 秒間を切り出した. RBM の訓練には, データセット中の Testing セットを用い, 結果の評価には Validation セットより 6 曲を用いた.

RBM のモデルパラメータの設定

RBM のモデルは Gaussian-Binary 型のモデルを用い, モデルのパラメータは Table 4.1 に示す通りに設定し, この訓練には CD_k アルゴリズムを用いる.

Table 4.1: Parameter of RBM

parameter name	parameter value
visible layer size	1280
hidden layer size	16
learning rate	0.001
learning epoch cycle	500

4.3.2 実験結果と考察

まず, 訓練後のモデルに対し, 入力データを与え, ギブスサンプリングを介して再構成された可視層の出力が Fig. 4.5 である. 入力のスペクトルと再構成されたスペクトルを見ると, 再構成されたスペクトルは比較的よく元のデータを復元できていることがわかる. 特に, 各信号の発生タイミングは比較的忠実に表現されているが, ここの周波数成分については, 信号強度が低下していたり, 失われている周波数成分が見られる.

次に訓練後のモデルに対し, 入力データを与え, これにより得られた隠れ層の出力パターンを Fig. 4.6 に示す. この出力パターンが聴覚フィルタより得られたスペクトルの特徴ベクトルのシーケンスである. 各パターンを比較すると, いずれにおいてもニューロンの発火が頻繁である箇所と, 全く発火しないニューロンの存在が確認される. また, 時間軸に従ってパターンを見ると, 何らかの発音タイミングをコーディングしているらしきパターンが確認される.

これらの結果より, 聴覚フィルタと RBM を用いた教師なし学習による方法論は, 一定の成果を得られたものの, RBM から出力される特徴ベクトルのシーケンスが各データ間で大きく類似している点を考慮しなければならない. これについては, パラメータの調整等の追加の実験により解決する可能性があるものの, 現段階において検出すべき音響イベントについて, 誤判断の要因になる可能性を否定できない. また, 現状の RBM においては, このパラメータの設定が経験的ノウハウに依存する部分も多く, どのパラメータがどの結果に影響するかが不透明であることも問題である.

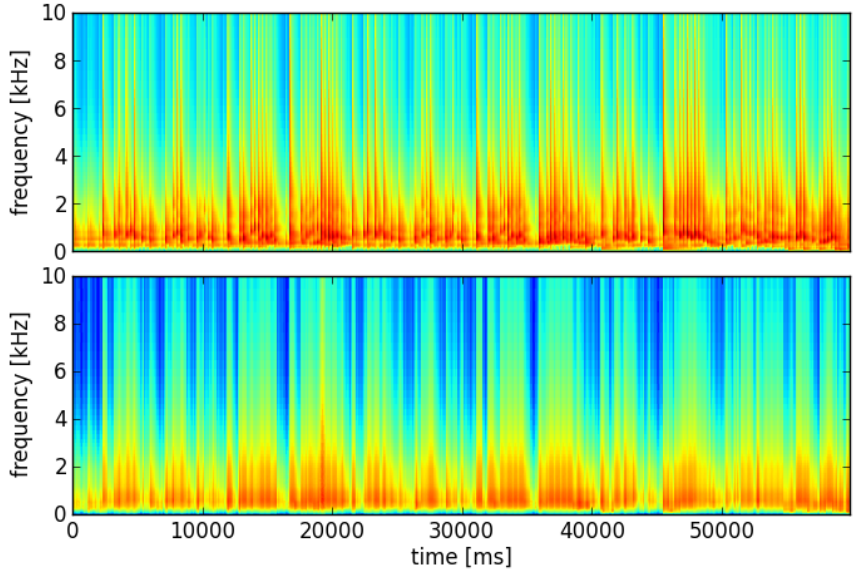


Fig.4.5: Input spectrogram (top) and reconstracted spectrogram with rbm (down)

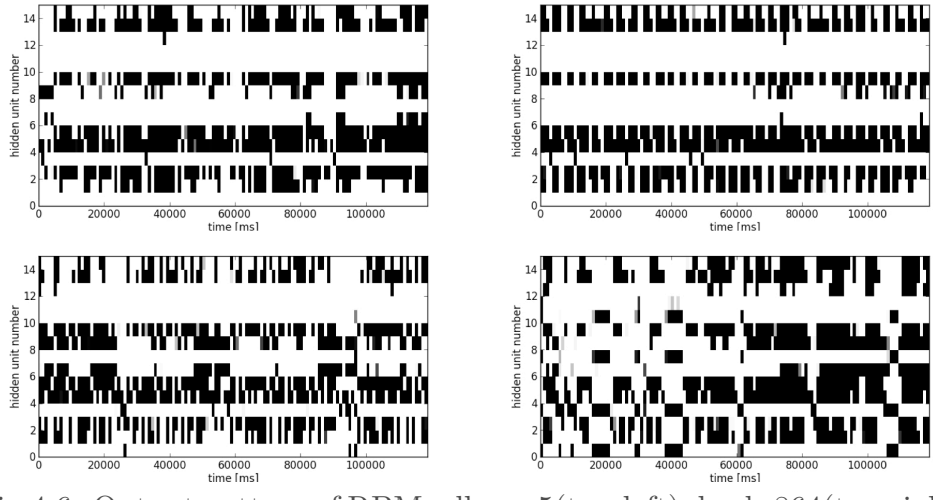


Fig.4.6: Output pattern of RBM: alb_eps5(top-left), bach_864(top-right), bor_ps5(down-right), deb_clai(down-right)

4.4 周波数成分の時間的入力バッファと畳み込みニューラルネットワークによる音響イベント検出

4.4.1 実験設定

エコーステートネットワークを取り入れたモデルによる実験の前に, 時間領域の入力バッファを持ったモデルを用いて, D-CASE challenge (Giannoulis et al. 2013; IEEE n.d.-a; Stowell et al. 2015) の OL subtask についての事前実験を行う.

時間-周波数領域で表現されたスペクトルのパッチ

2 次元の時間-周波数領域で表現される音響信号のスペクトログラムをある時間単位で分割したスペクトルのパッチを作成し, これを畳み込みのメカニズムを取り入れたモデルにより分類問題を解く. このスペクトルのパッチは静止画像同様に時間-周波数領域において, 2 次元で表現されたデータである. ここで, 画像においての x 軸が時間軸に相当するが, これが時間方向の入力バッファとなるため, スペクトルのパッチは時間軸方向の入力バッファ $[x_t, x_{t-1}, \ldots, x_{t-n}]$ を持つことと等価である. これにより, 畳み込みのメカニズムにおいて, 時系列のデータを取り扱うことが可能となる.

音響信号は, 44.1 kHz/16 bits かつ 1 channel(モノラル信号) となるように再サンプリングし, 10ms のフレームシフトで 512 サンプルのハミング窓を用いた STFT により周波数分析を行う. これにより得られる時間-周波数領域のスペクトログラムより, 有効範囲である 256 次元の振幅スペクトル成分のみを取り出した. この後, 50ms のフレームシフトで順次, スペクトログラムから 100ms の範囲を切り出し, 周波数軸が 256 次元, 時間軸が 100ms オーダーのスペクトルのパッチを作成した. モデルへの入力は, この 256 次元 $\times$ 100ms オーダーのスペクトルのパッチであり, 出力は分類対象の各クラスである.

データセット

モデルの訓練と評価には D-CASE challenge の OL subtask より, 訓練に訓練用データセット (Training set) を評価に開発用データセット (Development set) を用いた. これらは, オフィス環境下で発生する代表的な 16 のカテゴリーの音を録音したデータセットである. また, 学習用データと開発用データはともに, エアーコンディショナーの稼働音ように, ノイズのある実環境下での録音がおこなわれている. このため, モデルには雑音環境に対してのロバスト性が要求される. 本実験では, フレームベースの分類の枠組みを採用し, モデルは 50ms ごとに音響イベントのクラスを判断する.

モデルのパラメータ設定

事前実験用のモデルは畳み込みのメカニズムを持った代表的な多層ニューラルネットワークを用意した. 一つは, 畳み込みニューラルネットワーク (Norouzi et al. 2009) による多層ニューラルネットワークであり, もう一つは, Convolutional Deep Belief Nets (CDBN)(Lee et al. 2009) による多層ニューラルネットワークである. CDBN は畳み込みのメカニズムを取り入れ拡張した Convolutional RBM により構成される DBN である. 各モデルのネットワーク構造は 3 つの畳み込み層, 2 つの Max プーリング層, そして, 1 つの全結合層からなり, ネットワーク構造は比較のため, 同一の構造となるよう, どちらのモデルも Table 4.2 に示す設定とした. また, 比較のために乱数によりランダムにクラスを分類するモデルも用意した.

モデルの訓練について, CDBN の事前学習においては CD_k 法を用い Convolutional RBM を順次訓練する. このとき, ギブスサンプリングのステップ数は 1 ステップとし, 学習率は全ての層において 0.0001 と設定して CRBM のパラメータの推定を行った. また, CDBN の分類器の訓練と CNN の訓練については Adam(Kingma and Ba 2014) と呼ばれる訓練方法を適用し, また過学習を抑制するために早期停止の手法を取り入れた.

Table 4.2: Setting of each layers on CNN and CDBN.

Layer name	Filter size ($w \times h$)	Outout map size	Stride	Function
data		$257 \times 10 \times 1$		
conv 1	5×2	$253 \times 9 \times 16$	1×1	
pool 1	2×2	$127 \times 5 \times 16$	2×2	ReLu
conv 2	5×2	$123 \times 4 \times 16$	1×1	
pool 2	2×2	$62 \times 2 \times 16$	2×2	ReLu
conv 3	5×2	$58 \times 1 \times 16$	1×1	
full 1		$1 \times 1 \times 17$		Soft-max

4.4.2 実験結果と考察

実験結果において, 各モデルにはオーバーフィッティング (Over-fitinb; 過適合) の傾向が見られる. Fig. 4.7 と Fig. 4.8 はモデルの学習曲線であり, それぞれクロスエントロピー (cross entropy) の誤差の平均値 (mean loss) とモデルによる分類の正答率 (accuracy) である. これによると, 訓練データに対しては誤差が順調に降下し, 正答率も上がっているが, 評価用データに対しては, 誤差が上昇し, また正答率も降下する傾向が確認される. これはモデルが過適合を起こす際の典型的な学習曲線である. また, 実測の値は異なるものの, CDBN, CNN のどちらのモデルも同様の傾向を示している.

訓練後のモデルによる OL subtask 中の各音響イベントのクラス分類結果についても結果は良好ではない. Fig. 4.9 に各クラスごとの F 値を示すが, 各モデルの F 値はランダムにクラスを分類する場合に比べて有意差は見られない.

我々は, これらの結果とこれまでの深層学習の成果より, 入力データの表現に要因があるのではないかと考えている. 近年では、深層学習より STFT により得られるスペクトログラムから, より良い特徴を抽出することが可能であることが知られている. しかしながら, すべての場合において, 効果的で

あるとは限らないことが示唆される.

ここで, 生物の聴覚系に着目すると, 聴覚の抹消系においても周波数分析が行われていることが明らかにされているが, その仕組みと得られる周波数成分の表現はフーリエ変換により得られるものとは若干異なる. 生物の聴覚機構を考慮した聴覚フィルタによる周波数分析を行った場合, より良い結果が得られるものと我々は推測する.

また, 本実験において時間のバッファが 100ms であったことも, 要因の一つと考えられる. 時間領域の波形信号について, 10ms ごとにフレームをシフトさせながら周波数分析を行った場合, 100ms の時間範囲は x_tから x_{t-9} の 10 個のフレーム分となるが, ある音響イベントを認識するためには, より深い時間領域の依存性が存在する可能性も考えれられる. これにより, 入力バッファとして確保する時間軸方向のフレーム長が, より長くなくてはならないことが考えられる. しかし, 音響信号は各イベントごとに発音時間長の異なる可変長なデータであるため, 音響信号処理の観点からは, 入力バッファ長を単純に増やすという手法は適切な解法とはならないだろう.

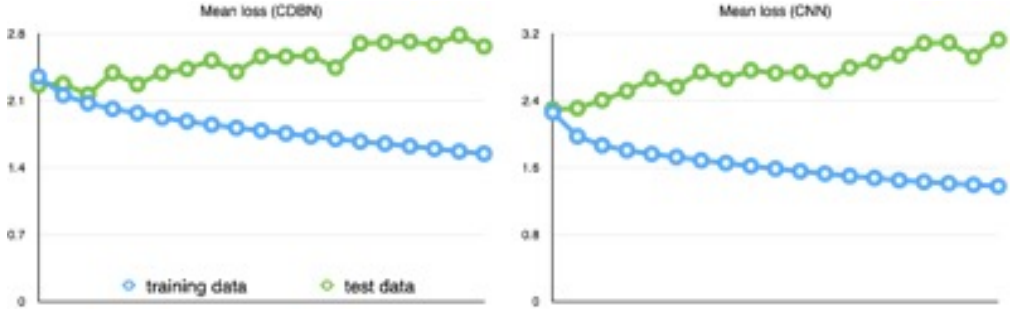


Fig.4.7: Learning curve (mean loss of cross entropy).

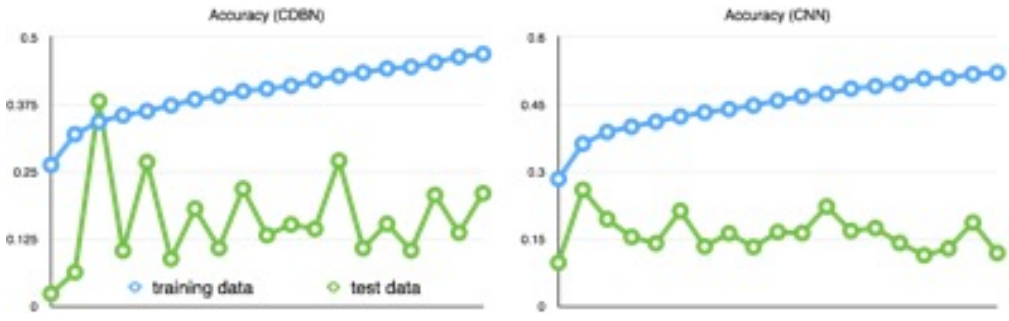


Fig.4.8: Learning curve (accuracy).

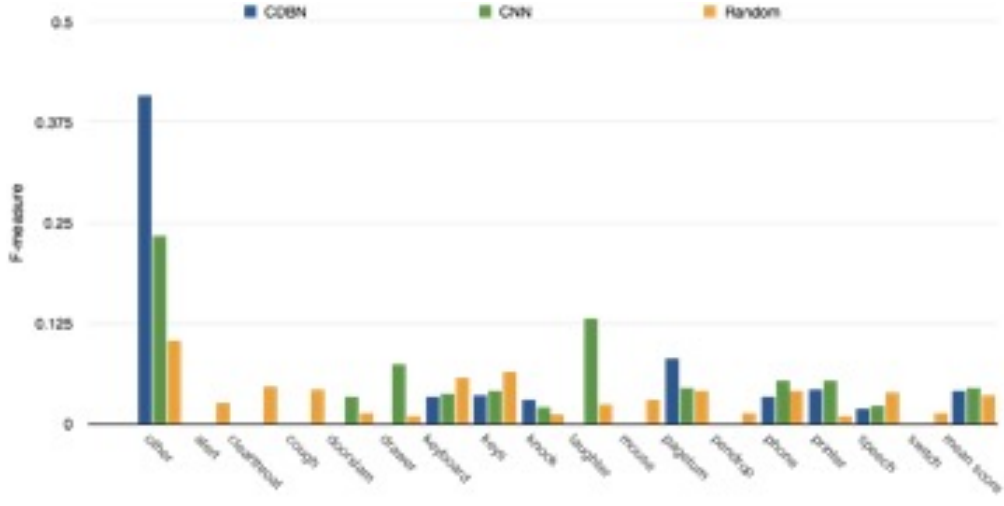


Fig.4.9: F-measure score of each category.

4.5 Auto Encoder および Echo State Network による音響イベント検出

4.5.1 実験設定

本節で提案する Fig. 4.1 に示すモデルについて, これを構成し評価を行う. 各コンポーネントは, ガンマトーンフィルタによる周波数分析器と sAE による特徴抽出器, そして ESN から構成されるモデルとなる. ここで, sAE

の出力層 y と ESN の入力層 x を共有するものとすると, 入力から出力までを sAE と ESN からなる多層ニューラルネットで処理できることになる.

このモデルについて, IEEE AASP Challenge に設けられている D-CASE challenge のうち, OL subtask を実行しモデルの評価を行う.

Office Live subtask in D-CASE challenge (IEEE AASP Challenge)

音響イベント検出と分類のためのタスクセットとして, IEEE AASP Challenge には音響信号を対象とした D-CASE challenge が設けられている. Office Live subtask (OL subtask) はこの D-CASE challenge のタスクの一つである. オフィス環境で発生する日常的な音を対象としたイベントの高精度な検出と分類がタスクの達成目標である.

イベント検出のためのデータセットは, development (開発用), training (訓練用), testing (テスト用) のサブセットで構成ている. この内, 開発とテストのデータセットは, オフィス環境で発生する日常的な音響イベントの 1 分間の録音データで構成されている. これには次の音響イベントが含まれる：door knock, door slam, speech, laughter, keyboard clicks, objects hitting table, keys clinging, phone ringing, turning page, cough, printer, short alert-beeping, clearing throat, mouse click, drawer, switches.

データセット

モデルの訓練と評価には D-CASE challenge の OL subtask より, 訓練に訓練用データセット (Training set) を評価に開発用データセット (Development set) を用いた. これらは, オフィス環境下で発生する代表的な 16 のカテゴリーの音を録音したデータセットである. また, 学習用データと開発用データはともに, エアーコンディショナーの稼働音ように, ノイズのある実環境下での録音がおこなわれている. このため, モデルには雑音環境に対してのロバスト性が要求される.

音響信号は 44.1 kHz/16 bits かつ 1 channel(モノラル信号) となるように再サンプリングし, さらに,40Hz ～ 19kHz 間に対数スケール上で等幅な中心周波数を持つ, 80 枚のフィルタで構成したガンマトーンフィルタにより周

波数分析を行った．ここで, 評価基準には, 10ms ごとにイベントを判定するフレームベースの枠組みに従った．

モデルのパラメータの設定

モデルについて, sAE と ESN のパラメータはそれぞれ Table 4.3 と Table 4.4 に示す通りに設定し, 訓練には Adam アルゴリズムを用いた．ここで, 時刻ステップ t でのモデルへの入力は, 80 の周波数ビンを持つスペクトルデータについて 10ms 分の範囲とし, この 10ms の範囲を 1 フレームの単位とする．また, モデルの構成は Fig. 4.10 に示す．

Table 4.3: Parameter of Stacked Auto Encoder

parameter name	parameter value
input (1st) layer size	800
hidden (2nd) layer size	100
output (3rd) layer size	200
learning rate	0.001
learning epoch cycle	3000

Table 4.4: Parameter of Echo State Network

parameter name	parameter value
input layer size	200
hidden (reservoir) layer size	2000
output layer size	16
learning rate	0.001
learning epoch cycle	3000

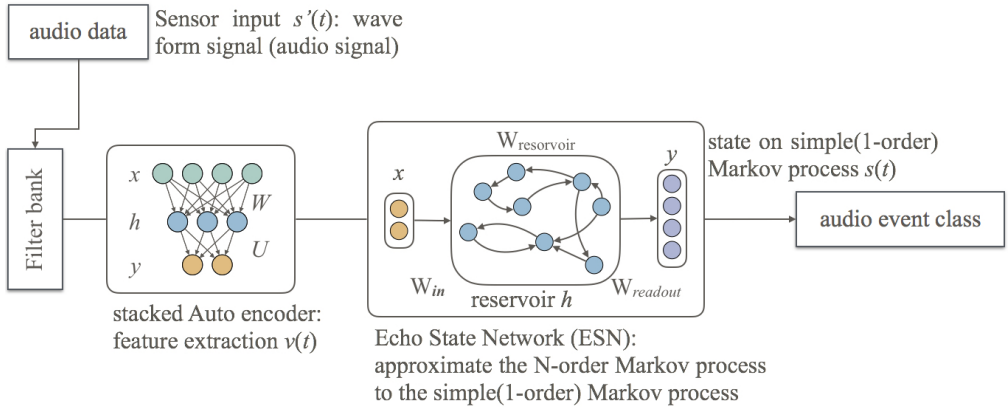


Fig.4.10: Model structure

4.5.2 実験結果と考察

まず, Table 4.5 に 2013 年度の D-CASE challenge のスコアと我々の提案手法によるスコアを示す. 提案手法 (表中の Proposal) は Baseline となるシステムよりも高い精度を示したが, 最高スコアのモデルを超える結果は得られていない.

また, Fig. 4.11 に各音響イベントごとのスコアを示すが, こちらでは各音響イベントごとの F-measure に大きな差異が認められる. マウスのクリック音とスイッチの音は全く認識できておらず, アラートやペンが落ちた音, ページを捲る音およびノック音のスコアといった, 幾つかのクラスのスコアが極端に低い. これとは対照的に, スコアの高いイベントは発音時間が長い, あるいは音量が大きいといった特徴が見られる. これらを比較すると, 短い発音時間や低い音量ではスコアが低いと見られるが, これはつまり提案モデルが入力音の音量や発音時間に依存してしまっていると考えられる. この要因には音響信号の分割やノイズの影響によ, 短い発音時間や低い音量の場合には判断に必要な入力を得られていない可能性が考えられるだろう. また, 時間ー周波数軸上で観測される音について, 時間軸あるいは周波数軸の位置シフトに対応できていないことも考えられる. さらに, 各音響イベントについて音量を正規化する必要性もあるだろう.

今回のアプローチでは, AE の特徴抽出機能がノイズや他の要因に対してもロバストに働くと予想していた. しかし, この予想が覆される結果が確認された. この OL subtask は訓練, 開発, テスト, 全てのデータセットがノイズ環境下で録音されている. 2013 年度の challenge において, 高いスコアを記録しているモデルにはこのノイズを除去する工程や各音響信号を正規化する組み込まれている. これについて, AE の学習時にノイズを付加する Denoising AE であれば, よりノイズに頑強なモデルが得られた可能性があり, また位置のシフトに対応出来る畳み込みアーキテクチャの導入も精度向上につながりそうである. さらには, 正規化層の導入も検討するべきであろう.

これらの結果より, 現段階の提案手法では音響イベントの検出と分類に関して十分に高精度なモデルを学習できないことが示唆された. しかしながら, 先の事前実験の結果と比較した場合には, 大きな精度の向上が見られる. これにより, 提案手法をベースに改良を加えることで, より高精度なモデルの学習が可能であることも期待される. このためには, 本実験において何が認識精度向上のためのボトルネックとなっていたのかを正確に判断するための追加実験を要するだろう.

Table 4.5: Score of OL subtask in D-case challenge (2013)(IEEE n.d.-b) and our aproach

model	F (%)	Pre (%)	Rec (%)	AEER
CPS	3.82	9.15	3.05	2.116
DHV	26.0	19.84	45.28	3.128
GVV	31.94	61.78	22.29	1.084
NR2	34.66	37.15	34.96	1.885
NVM_1	40.85	59.90	32.90	1.115
NVM_2	42.76	61.15	34.28	1.102
NVM_3	45.50	57.23	38.80	1.102
NVM_4	42.86	50.79	37.79	1.360
SCS_1	53.02	59.89	48.28	1.167

model	F (%)	Pre (%)	Rec (%)	AEER
SCS_2	61.52	66.18	57.83	1.016
VVK	43.42	68.14	32.60	1.001
Baseline	10.72	12.13	10.56	2.590
Proposal	30.17	36.19	25.88	none

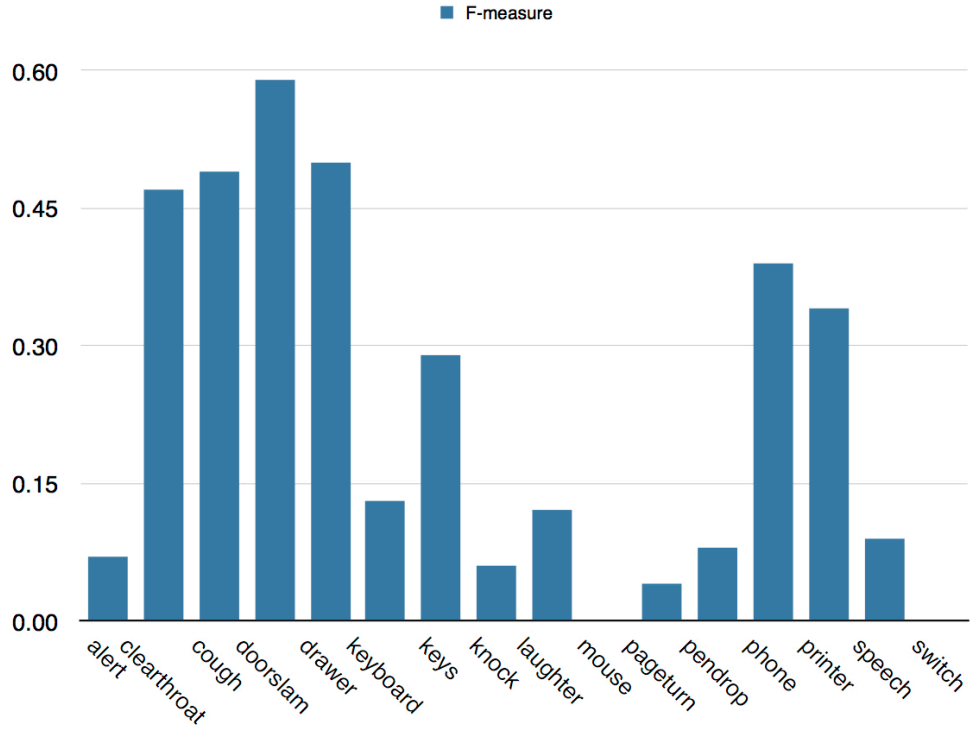


Fig.4.11: F-measure of classified each audio event

参考文献

Classical piano midi page. (n.d.). http://www.piano-midi.de/, Accessed 17 May 2016.

Giannoulis, D., Benetos, E., Stowell, D., Rossignol, M., Lagrange, M., & Plumbley, M. D. (2013). Detection and classification of acoustic scenes and events: An ieee aasp challenge. In *Applications of signal processing to audio and acoustics (wASPAA), 2013 iEEE workshop on* (pp. 1–4). IEEE.

IEEE. (n.d.-a). IEEE aASP challenge: Detection and classification of acoustic scenes and events. http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/, Acsessed 17 May 2016.

IEEE. (n.d.-b). IEEE aASP challenge: Detection and classification of acoustic scenes and events; event detection - office live results. http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/resultsOL.html, Acsessed 17 May 2016.

Irino, T., & Patterson, R. D. (1997). A time-domain, level-dependent auditory filter: The gammachirp. *The Journal of the Acoustical Society of America*, *101*(1), 412–419.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical

representations. In *Proceedings of the 26th annual international conference on machine learning* (pp. 609–616). ACM.

Norouzi, M., Ranjbar, M., & Mori, G. (2009). Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning. In *Computer vision and pattern recognition, 2009. cVPR 2009. iEEE conference on* (pp. 2735–2742). IEEE.

Poliner, G. E., & Ellis, D. P. (2007). A discriminative model for polyphonic piano transcription. *EURASIP Journal on Applied Signal Processing*, *2007*(1), 154–154.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., & Plumbley, M. D. (2015). Detection and classification of acoustic scenes and events. *Multimedia, IEEE Transactions on*, *17*(10), 1733–1746.

第 5 章

人間の作曲活動をモデル化した自動作曲システムの構築に関する研究

5.1　はじめに

5.1.1　コンピュータサイエンスと音楽

1956 年に L. Hiller と L. Isaacson らによって「ILLIAC 組曲」(Hiller and Isaacson 1979) が発表された. 最初期のコンピュータ「ILLIAC」を用いた数値計算によって, 初めて作曲された音楽である. 音楽や作曲技法と数値計算や数学は古くから関連性のある古典的な領域にあったが, コンピュータによって作曲がなされたことは, 他の研究領域に対しても大きな可能性を示した.

コンピュータによる数値計算結果を作曲に取り入れることは, コンピュータ支援作曲やアルゴリズム作曲 (Algorismic Composition) と呼ばれ, 現代音楽においては代表的な作曲技法として認知されている. これらは, 音楽家・作曲家が自身の音楽を表現するための技法という側面が強く, 一般の人々の音楽制作支援あるいは音楽の新しい楽しみを感じるためのシステムではな

かった.

現代ではコンピュータの数値計算は音楽のあらゆる場面で活用されている. 作曲の支援を行うシステム, アイディアとなるフレーズを自動生成するシステム, VOCALOID, コンピュータによる録音やマスタリング・ミキシング, Apple 社の GarageBand といった音源と一体化した統合作曲環境, さらにこれからも多種のシステムが開発されるだろう. これらは音楽制作環境の提供や作曲の支援, 新しい音源の提供あるいは新しい音楽の楽しみ方といった, 作曲技法とは異なる観点からコンピュータが利用されている.

自動作曲システムは「ILLIAC 組曲」やアルゴリズム作曲の系譜に連なり, コンピュータ・システムによって人間の作曲活動を再現しようとする研究領域である. 確率や統計, 機械学習といった人工知能研究, 音響信号処理, 自然言語処理の方法論が用いられ, 研究成果は先述した音楽の制作支援や音楽の新しい楽しみを提供するシステムの数々に応用されている.

自動作曲システムは人工知能研究の対象課題としての側面が興味深いが, 社会的には作曲支援や音楽の新しい楽しみ方を提供することの意義が強い. 例えば, 楽器や発声器官により, 試行錯誤的に楽曲を制作することは作曲方法として珍しいことではないが, これは音楽や楽器演奏の経験者においてである. 初級者や作曲になじみのないものには敷居が高いが, 作曲技法や音楽理論の習得もまた困難が伴う. 制作過程において, 初級者に困難である部分をコンピュータ・システムが代替し, 音楽制作を誰でも楽しめるものとしようということが, 自動作曲研究の社会的な意義の一つとなる.

自動作曲システムを使用した作曲は, 例えば Fig. 5.1 に示すような手順となる. 基本的にはシステムが楽曲のパーツを生成し, ユーザがこれを選択して組み合わせたり編集したりする. ここで, バックエンドにあるアルゴリズムが, システムに用意されているパラメータの値から楽曲のパーツを自動生成する. システムが音楽制作過程に介入することで, 音楽制作における敷居を低くすることが期待できる.

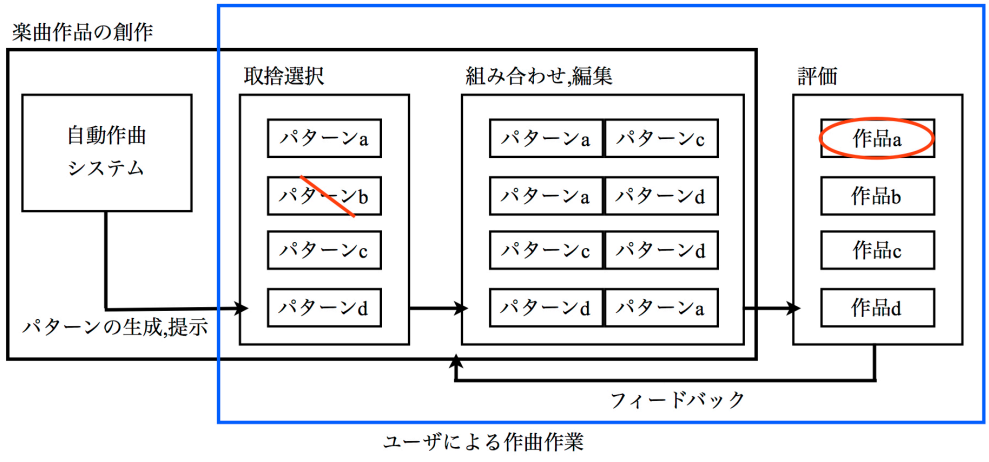


Fig.5.1: Conventional automatic composition system

5.1.2 提案システムの概要

本章では作曲活動全体の代替・再現を行う新しいタイプの自動作曲システムを提案する．システムは Fig. 5.2 に示すように, システムは音列パターンについて, 生成と評価のモジュールを持ち, これら 2 つのモジュールは互いにフィードバックを与えながら繰り返し稼働する．システムの全体の枠組みは機械学習分野における強化学習の方法論をイメージしてもらいたい．システムがこれら 2 つのモジュールを持つ理由は, 作曲活動が生成と評価の反復を繰り返しながら行われる試行錯誤的な活動と考えられるからである．

システムの主要な目的は大きくは 2 つである．まず, 人工知能研究の対象課題として自動作曲システムを開発すること, 次に, マルチメディアコンテンツ制作の支援システムとして自動作曲システムを開発することである．

特に, 近年において個人でのコンテンツ制作と公開が容易になっているが, 素材の選定については種々の制約がある．クリエイティブ・コモンズ (CC n.d.; CCJP n.d.) といったライセンスによる対応も取られているが, 自動作曲システムのようにコンテンツ素材を自動生成することのできるシステムは有効な意義を持つだろう．

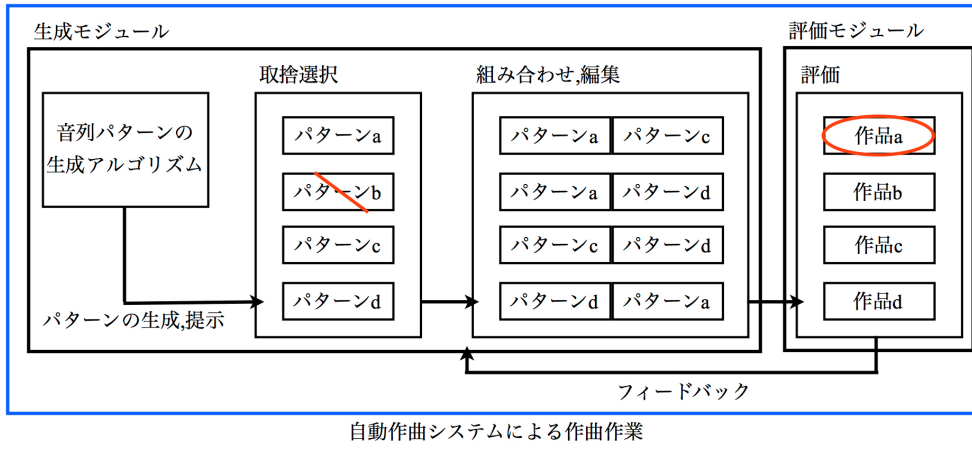


Fig.5.2: Proposal of automatic composition system

5.1.3 本章の構成

本章では, 提案システムの生成と評価のそれぞれについて, 検討を重ねた方法論を述べる. まず, 続く第 2 節において自動作曲システムに関する先行研究を紹介し本章の位置付けを示す. 第 3 節においては提案システムの概要と生成と評価それぞれのモジュールの方法論についてを述べ, 第 4 節では今後の展望と課題について述べるとともに本章をまとめる.

5.2 関連研究

5.2.1 音楽情報処理

音楽情報処理は, 音楽のあらゆる側面を情報処理の枠組みで扱う分野である. ここでは, 音楽情報処理から幾つかのテーマと先行研究を紹介する.

計算機が音楽を探す (音楽検索・音楽推薦)

このテーマではコンピュータが音楽を探すという課題を扱う. つまり, 音楽検索や音楽推薦に関する研究テーマである. 年々大量の楽曲が生み出され

続け, 個人ユーザが全ての音楽を知る, あるいは探すことは困難である. まだ知らないが好みに合う音楽や良い音楽をユーザに提示し, 新しい音楽との出会いを提供することが大きな目的となる.

人の心拍数を利用した検索システム (野地保 et al. 2010), コンテンツベースの推薦システム (Logan 2004), コンテンツベースと協調フィルタリングのハイブリッド型推薦システム (吉井和佳 et al. 2006; 梶克彦 et al. 2004), などが提案されている代表的なシステムである.

計算機が音楽を理解する (音楽理解)

人間にとって音楽の認知処理は比較的意識せず行なわれている. 音の高さ, メロディーの判断, リズムへの追従, ジャンルの分類, この他にも多種の活動がある. しかし, コンピュータにとってこれらの情報処理は困難な課題となる.

そこで, 音楽理解に関する活動をコンピュータに再現させようとする研究がこのテーマにおいて扱われる範囲である. 混合した音をそれぞれの音に分離する音源分離 (後藤真孝 and 村岡洋一 1994), 隠れマルコフモデルによるリズム推定 (大槻知史 et al. 2002; 齋藤直樹 et al. 1999) やピッチ推定 (三輪多恵子 et al. 1998; 亀岡弘和 et al. 2003), が代表的である. コンピュータに音楽を理解させようとするこのテーマの研究は, 音楽情報処理の根幹をなす領域である.

計算機が歌う (音声合成)

初音ミク (Crypton n.d.) というソフトウェアが爆破的な人気と知名度を獲得している. これは YAMAHA が開発した VOCALOID(YAMAHA n.d.; 剣持秀紀 2012; 剣持秀紀 and 藤本健 2015) という, 歌声合成エンジンがバックエンドにある商用の音源ソフトウェアである. これまで電子楽器の音源は実在の楽器やシンセサイザによる合成音のみであったが, VOCALOID の開発により人の歌声が電子楽器上で扱えるようなった. このテーマでは, 歌声合成システムの研究 (吉田由紀 and 中嶌信弥 1999; 酒向慎司 et al. 2004), さらに自然な人の歌声に近づけるための研究 (中野倫靖 et al. 2008) が進ん

でいる.

計算機が音楽を創作する (自動作曲)

音楽情報処理において, 自動作曲はコンピュータが音楽を創作するテーマになる. 音楽の創作活動をコンピュータ上で再現することを目指す, 人工知能研究にも近い領域である. 例えば, 商用ソフトウェアで有名な Band-in-a-Box(e-frontier n.d.) には, 作曲支援として自動的にメロディや伴奏を生成する機能が提供されていように, このテーマの研究成果は音楽制作を支援するシステム自体や音楽制作環境の支援機能として応用されることがある.

Orpheus(嵯峨山茂樹 et al. 2012; 深山覚 et al. 2008) は自動作曲研究の成果としてよく知られているシステムの一つである. 日本語の歌詞と用意されているパラメータを調整することで, 歌唱曲を自動生成するシステムである. システムの方法論としては, 日本語歌詞の韻律 (イントネーション) と確率モデルを用いている. 他の自動作曲の方法論には, 和声論に基づいたシステム (三浦雅展 and 江村伯夫 2012), 遺伝的アルゴリズムを用いたシステム (山田拓志 and 椎塚久雄 1998), などが提案されている.

5.2.2 MIDI

MIDI(Music Instrument Digital Interface) は, コンピュータによる音楽処理のための通信規約である. コンピュータ制御された音楽機器間において, 演奏データの通信をするために策定されたが, 後にコンピュータ上で音楽を扱うためのデータ・フォーマットとして, 標準的に利用されるようになった.

MIDI は各機器間において, MIDI message と呼ばれるコマンドを送受信する. 例えば, 「時刻 T に, 音 S を強さ V で発音する」というな内容のイベントメッセージが通信され, 適宜処理される. MIDI ファイルの実態は, この MIDI message を格納したコマンドの集合である.

MIDI において音高と音価はそしてヴェロシティは基本的なパラメータであるが, これらは離散化して扱われる. まず, 音高は 128 段階に分割され, これを 7bit で表現し 0 から 127 の整数値が割り振られる. この割り当てら

れた数値はノートナンバーと呼ばれ, 1 増減するごとに音高が半音づつ増減する. 次に, 音価は 4 分音符分の解像度を表す tick time と呼ばれる単位時間が用いられる. 例えば, 4 分音符の解像度を 96tick とした場合には, 8 分音符が 48tick, 2 分音符が 192tick となる. さらに, ヴェロシティは 128 段階に分割され,0 から 127 の整数値が 7bit で割り当てられる. ここで, 音高とは音の高さ, 音価は音の発音時間のことを言い, ヴェロシティは音の大きさのパラメータである.

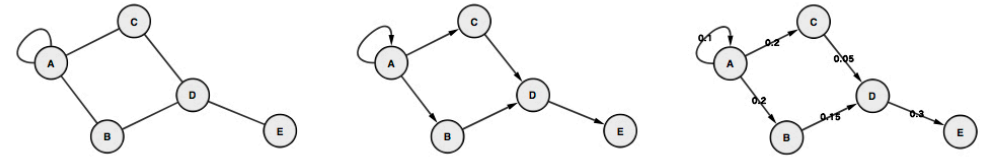
5.3 生成モジュール

生成モジュールは音符の遷移情報から構成したネットワークを用いて, 音列のパターンを生成する. このネットワークは既存の楽曲を用いて構築するため, 不快感を与える遷移のリンクが存在せず, 構築に用いた楽曲の特徴が反映されていることが期待出来る.

5.3.1 複雑ネットワーク

複雑ネットワークはグラフ理論の一つであるが, 特に大規模かつ複雑なグラフ構造による情報の表現や可視化の方法論を扱う. 複雑ネットワークの典型的な例は, 人間関係のネットワークや WWW(World Wide Web) であろう. 例えば, ノードを人としてノード間のリンクを関係の有無で結合する, そして親密度によってリンクに重みをつけると友人関係をネットワークで表現することができる.

グラフを構成する基本要素はノード及び, ノード間のリンク結合を表すエッジである. エッジには向きと重みのパラメータがあり, これらの有無によって Fig. 5.3 のように, いくつかの種類に分類される.



(a) 無向グラフ　(b) 有向グラフ　(c) 重み付き有向グラフ

Fig.5.3: Graph type

5.3.2 音符遷移ネットワーク

楽譜は音楽や演奏の情報を記述する方法の一つである. 現代では五線譜上に音符と呼ばれる記号を配置して, 音楽の情報記述する方法が最も一般的である. 楽譜では音の情報について, 音高を五線譜上の縦の配置場所, 音価を音符の種類によって決定し, 音の強さやリズムなどの付加情報は他の表記が用いられる. 記号的に音楽を解釈すると, 五線譜に配置された音符の 2 次元的な時系列パターンと考えられる. 以降, 特に前置きを置かない限りは音符を音高と音価のセットを指す言葉として用いる.

この五線譜上の音符の関連性は主に次の 2 つとなる.

1. ある音符 A から音符 B へ遷移する
2. ある音符 A と音符 B が同時に発音する

このとき, 同時刻に発音する音符の組みを纏めて一つの音符 (記号) と解釈すると, 楽譜は基本的には音符の遷移関係のみを記述していることがわかる.

音符遷移のネットワークは五線譜上の音符の関連性を整理し, グラフで表現したものである. 音符をノード, 音符間の遷移を向きを持ったリンクの結合で表す. また, 同じ遷移のパターン数, つまり頻度を重みとして設定する. ここで, ノード間の同じ向きを持つリンクは, その総和と個々の頻度を用いて遷移の確率を表現できる.

例えば, Fig. 5.4 は「蛍の光」のメロディーパートを用いて, 音符の遷移

関係をネットワークで表現したものである. 各ノードに与えられている数値は後述する音符遷移ネットワークの構築の項において, 設定している音高と音価を表す数値である.

このようなネットワークで楽譜を表現すると, 音列パターン生成を確率的な移動経路探索, あるいは移動経路生成の問題として扱うことができる. このとき, 移動経路のノードが表す音符を記録して行けば, 自然と楽譜として記述されることになる.

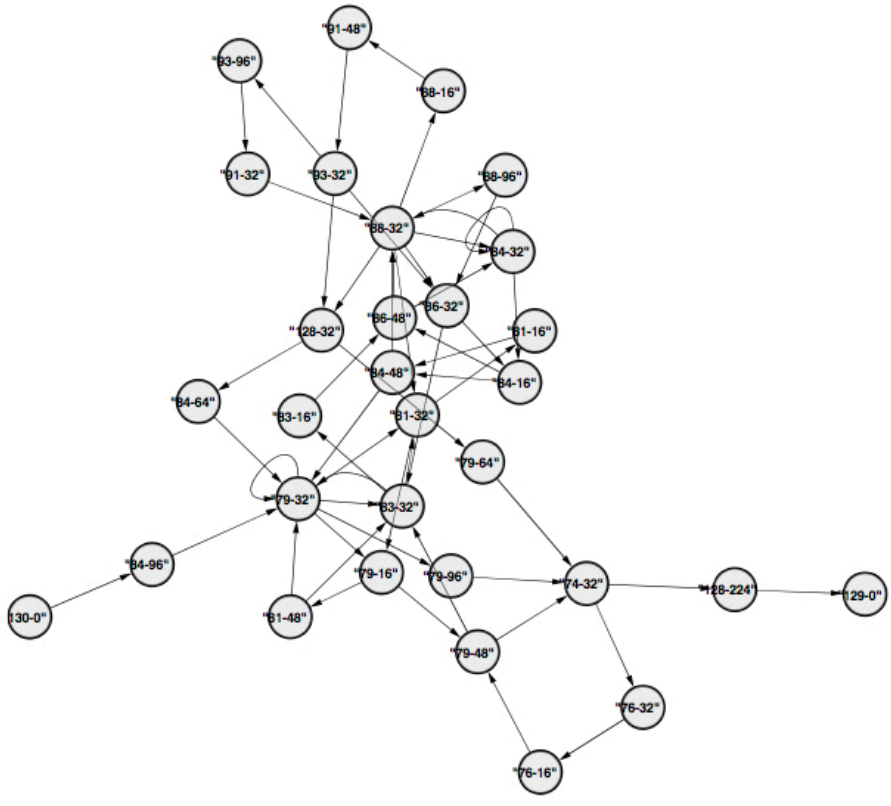


Fig.5.4: Network sample of created from ”Hotaru no hikari”

5.3.3 音列パターンの生成

音符遷移ネットワークの構築

音符の遷移ネットワークの獲得は, 何らかの機械学習手法を用いることによっても可能であると思われるが, ここでは既存の楽曲をネットワークで表現することとする. ノードが表現する音高と音価は MIDI に準拠して次のように設定し, ヴェロシティは考慮しないものとする.

1. ノードは音高と音価のセットを表す. 音高はノートナンバーである 0 から 127 の整数値とし, 休符には 128 のナンバーを与える.
2. 音価は 128 分音符の tick time を基準単位とし, 128 分音符の 1 から全音符の 128 までの整数値を与える.
3. 開始記号として音高が 129 の整数値かつ音価 0, 終端記号として音高が 130 の整数値かつ音価 0 のノードをそれぞれ用意する.

これらの数値設定について, 音高の対応表を Table 5.1, 音価の対応表を Table 5.2 に示す. ここで, 音高は楽器の中で最も広い音域を持つピアノに準拠する.

Table 5.1: note-number list

音高	ナンバー	音高	ナンバー	音高	ナンバー	音高	ナンバー
		C2	36	C4	60	C6	84
		C#2	37	C#4	61	C#6	85
		D2	38	D4	62	D6	86
		D#2	39	D#4	63	D#6	87
		E2	40	E4	64	E6	88
		F2	41	F4	65	F6	89
		F#2	42	F#4	66	F#6	90
		G2	43	G4	67	G6	91
		G#2	44	G#4	68	G#6	92

音高	ナンバー	音高	ナンバー	音高	ナンバー	音高	ナンバー
A0	21	A2	45	A4	69	A6	93
A#0	22	A#2	46	A#4	70	A#6	94
B0	23	B2	47	B4	71	B6	95
C1	24	C3	48	C5	72	C7	96
C#1	25	C#3	49	C#5	73	C#7	97
D1	26	D3	50	D5	74	D7	98
D#1	27	D#3	51	D#5	75	D#7	99
E1	28	E3	52	E5	76	E7	100
F1	29	F3	53	F5	77	F7	101
F#1	30	F#3	54	F#5	78	F#7	102
G1	31	G3	55	G5	79	G7	103
G#1	32	G#3	56	G#5	80	G#7	104
A1	33	A3	57	A5	81	A7	105
A#1	34	A#3	58	A#5	82	A#7	106
B1	35	B3	59	B5	83	B7	107
						C8	108

Table 5.2: duration-number list

音価	整数値	音価	整数値
全音符	128		
2 分音符	64	付点 2 分音符	96
4 文音符	32	付点 4 分音符	48
8 文音符	16	付点 8 分音符	24
16 文音符	8	付点 16 分音符	12
32 文音符	4	付点 32 分音符	6
64 文音符	2	付点 64 分音符	3
128 文音符	1		

経路選択による音符情報の出力

生成モジュールが出力する音列パターンは, 音符の遷移ネットワーク上の確率的な経路選択により得られる. 経路選択は開始記号のノードから終端記号のノードに到達するまで継続し, あるノードから移動する次のノードは各ノード間のリンクに付与されている確率に従いサンプリングする. このとき, 一度通った経路を再び通過するような重複した経路の選択は許すこととする.

検証

提案手法による音列パターン生成について検証を行った. 音符の遷移ネットワーク構築には, 楽譜集「ピアノで弾く風の谷のナウシカ」より次の 6 曲分のメロディーパートを用いた. これらのメロディーパートは複数の音高が同時刻に発音されることはないので, 単音の場合のみ考慮したネットワークとなっている.

1. 風の谷のナウシカ
2. 戦闘
3. 風の谷への道
4. はるかな地へ
5. 虫愛づる姫
6. 鳥の人

Fig. 5.5 には音符の遷移ネットワークを示す. ここで, ノードの大きさはそのノードが表す音符の出現頻度を示している.

まず, Fig. 5.6 に提案手法によって得られた楽譜より 4 小節分を示すが, 第 1 小節が全休符のため実質的には 2 小節程度である. 以降は楽譜上の 2 小節単位の塊をフレーズと呼ぶ. このように, 1 フレーズ程度の小さい断片においては比較的違和感を与えない, つまり音楽らしい音列のパターンを得ることができることが確認された.

楽曲はいくつかのフレーズのシーケンスデータと捉えられる．楽曲上にはいくつかの構造的な共通項や典型的なパターンがある．あるフレーズやフレーズのパターンが一定期間繰り返されたり, 時間的に離れた場所で再使用されたりするが, 楽曲の構造には次のようなパターンが典型的に見られる．

1. フレーズ x をそのまま再度使用する
2. フレーズ x を若干変更し, フレーズ x' を作成して使用する
3. 一連のフレーズパターン X をそのまま再度使用する
4. 一連のフレーズパターン X の内, 一部のフレーズを変更した X' を作成して使用する
5. 全く新しいフレーズ y やフレーズパターン Y を作成し使用する

一般的な楽曲はこれらがバランス良く用いられることによって, 成り立っている．

次に開始記号から終端記号に着くまでの全ての経路から作成した一つの楽譜を Fig. 5.7 と 5.8 に示す．先に述べたように, 音楽らしいフレーズがいくつか確認されるが, これが全体の中でも少数であることがわかる．また, 全体を一つの楽曲と見たとき, 楽曲上に見られる典型的な構造が再現されていないため, このまま演奏すると違和感や不快感を感じるだろう．

これらの要因は, この生成手法が実質的には音楽を単純マルコフ連鎖であるとした経路選択を行っているためであろう．グラフのノードとそのリンク結合, 遷移確率は既存楽曲を用いてネットワークを構築している．このため, フレーズ程度の短い経路においては, 単純マルコフ連鎖であっても音楽的な整合性のとれたパターンが生成されやすい．しかし, 移動経路が長くなれば, 単純マルコフ連鎖の基で音楽的な整合性を取ることが困難となる．これは, 音楽自体の性質としてある音遷移の決定が過去の幾つかの状態に依存するためである．つまり, 音楽においては N 階マルコフ連鎖と仮定する方がより自然である．これにより, 生成手法について N 階マルコフ連鎖を仮定したモデルに更新することで改善される可能性が示唆される．

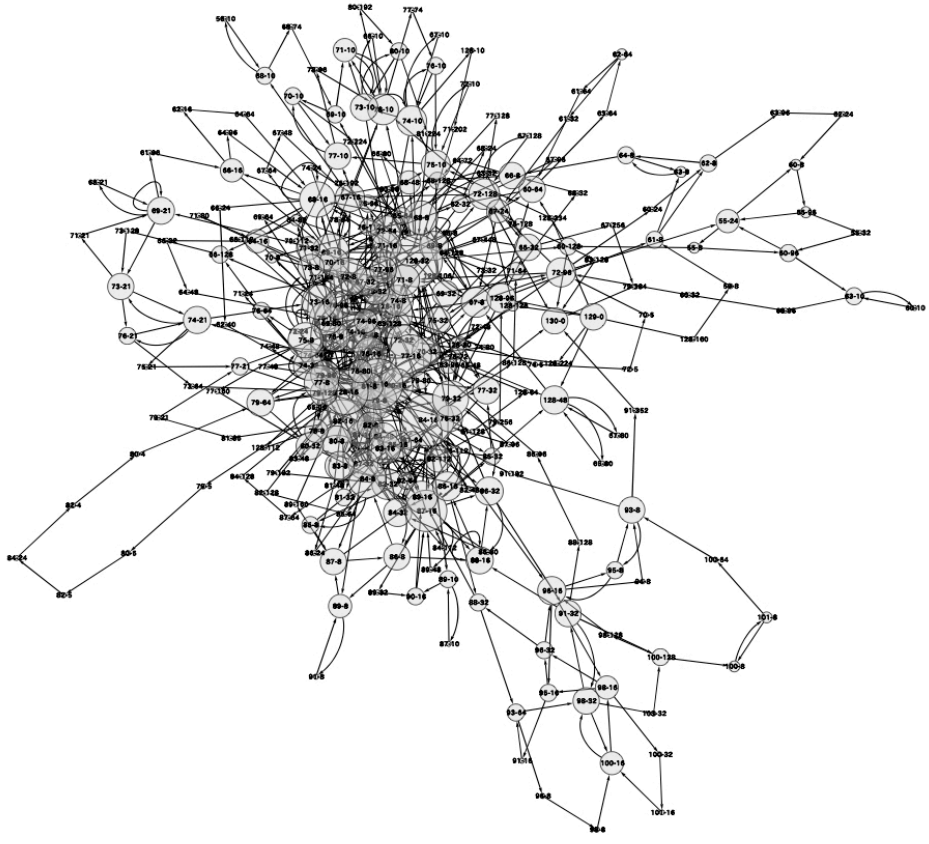


Fig.5.5: Network sample of created each musics

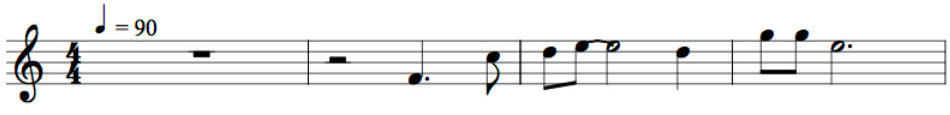


Fig.5.6: Pattern of phrase in generated music score

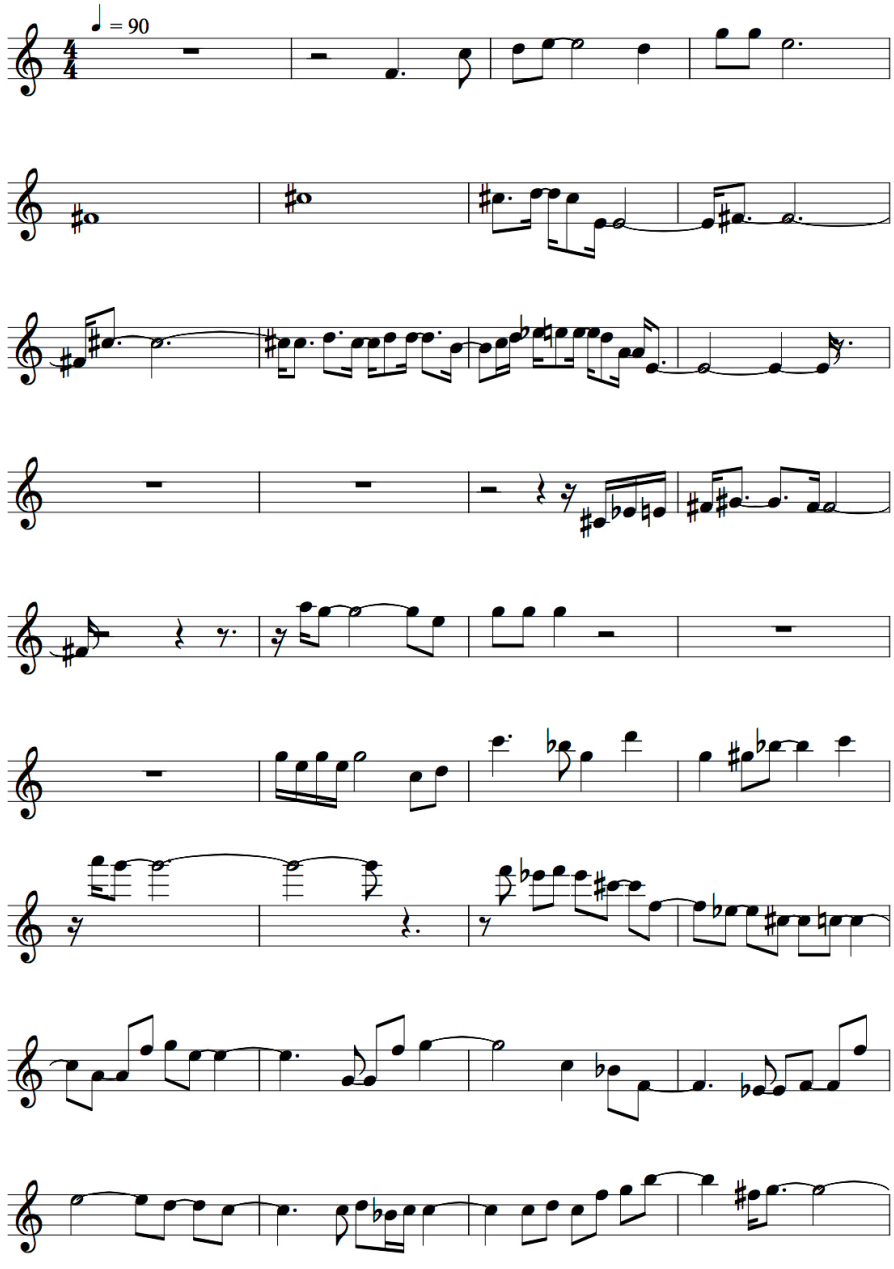


Fig.5.7: Pattern of generated music score

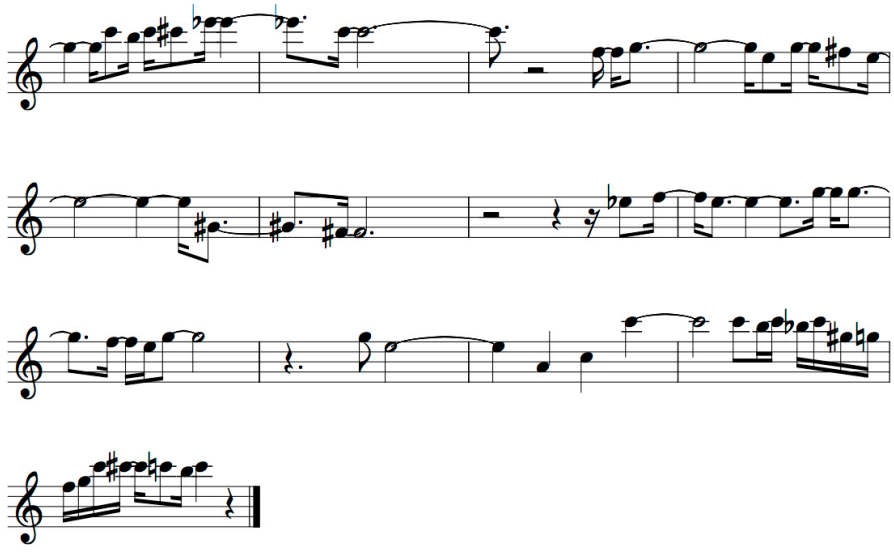


Fig.5.8: Pattern of generated music score

5.4 評価モジュール

評価モジュールでは生成モジュールにより得られた音列パターンの評価を行う. 評価モジュールによる評価は音楽の完成度の判断と生成モジュールへのフィードバックとして用いる. これにより不快感や違和感を与えるような出力を抑制し, 生成物の完成度向上に寄与させる. この評価器はリカレント型の人工ニューラルネットワークを用いて構成する. 本章では最も標準的な Elman net(Elman 1990) と呼ばれるモデルを採用した.

5.4.1 音楽と予想

ある個人の音楽に関連する嗜好はその個人の音楽的経験や, 音楽とともに体験した経験の積み重ねによって形成される. 現代社会ではいたるところで音楽が流れているが, その内容は文化圏やコミュニティあるいは個人の行動

範囲によってある種の偏りを持つ. 例えば, 好んで観るテレビ番組やその合間の CM で用いられる音楽, 通学や通勤時に聞こえる街頭の音楽, ある国や時代, コミュニティや友人間で流行している音楽, ある複数の個人間でこれらに共通項を持つことはよく知られている. 推薦システムの方法論にはこのような関連性を利用する手法も提案されているほどである. このような背景により, ある文化圏やコミュニティ, 行動範囲を共にする者同士の間では, 類似した音楽的バックグラウンドと嗜好を獲得しやすい.

このように獲得された音楽的経験を背景に, 人間は予測処理をしながら音楽を聴取していると Meyer は指摘している (Meyer 2008). これによると, ある音楽的経験から獲得した予測モデルに従って, 次にどのような音が発音されるか, メロディはどのように展開されるか, といった期待を持っていることになる.

この予測処理には許容される範囲の予測誤差があることも指摘され, これは音楽の認知処理や情動喚起に関わっていると言われている. 予測が外れ続けるような音のパターンは音楽と認知せず, その一方で予測が全て当たるような音のパターンは情動喚起が起こらないか低い. これに従うと, 良い音楽の条件はある音楽的経験による予測を満たしつつも, 許容可能な予測誤差の範囲を逸脱しない程度に予測を外すパターンを含むことであると考えられる.

5.4.2 12 平均律

人間の感じる音の高さはある振動の周波数に対応し, 音楽の分野ではこの周波数と音高記号との対応を音律と呼ぶ. 12 平均律は音高と周波数の対応を決定する手法の一つであり, 現代では多くの楽器や電子楽器がこれを採用している. 例えば, ピアノやギターの調律は基本的には 12 平均律に従った調整が行われている. 音律を簡単に捉えるとピアノの各鍵盤と打鍵したとき弦が振動する周波数の対応関係となる.

人間の聴覚の特性として, ある音高の 2 倍の周波数を持つ音が 1 オクターブ高い同じ音高として聞こえることがわかっている. 12 平均律ではこの特

性に基づいて 1 オクターブを一つの区間とし, 各音高の周波数比を決定する. 具体的には区間内の周波数比が同じ比率になるように 12 分割し, 基準音の周波数を用いて音高に固有の周波数を対応させる. ある音の周波数を f_i, 半音 n 個分シフトした音の周波数を f_{i+n}とすると, その周波数比は次式で求められ, 差異が半音の隣り合う音の周波数比, つまり $f_i : f_{i+1}$は常に $\sqrt[12]{2}$ となる.

$$f_i : f_{i+1} = 1 : 2^{n/12} \tag{5.1}$$

5.4.3 時系列予測による評価モデル

音楽と予測・期待感は前に述べたように密接な関係にある. そこで, この音楽を聴いているときの期待感, つまり予測処理に着目した評価モデルを検討する. 例えば, 予測の正答率を評価指標として用いることが可能であれば, ある正答率に当てはまる音列パターンを生成モジュールから得られるようにフィードバックをかけることができる. ここでは, 特に音楽のメロディーに焦点を当てることとする.

音価への平均律の適用

音楽を予測処理の枠組みで捉えると, 記号による絶対的な表現よりも, 何らかの相対的数値を用いる方が都合が良い. 音高は平均律を用いて記号による表現から相対的な実数値表現とできるが, 音価についても同様の表現であることが望ましい. そこで, 音価の記号へ 12 平均律の手法を適用し, 相対的な実数値で表現する.

まず, 全音符から 32 分音符までの区間を音高の 1 オクターブ区間に見立てる. この区間中には付点音符を含めると音価を表現する記号が 11 種類ある. ここで 32 分音符を 1, 全音符を 2 として, この区間を均等な比率になるように 11 分割しする. これにより, 音価の記号を音高の平均律と同様の実数値表現として扱うことができる. ある音価を l_i, n 個分シフトした音価を l_{i+n}とすると, その比は次式で求められ, 差異が 1 の隣り合う音価の比, つ

まり $l_i : l_{i+1}$は常に $\sqrt[11]{2}$ となる.

$$l_i : l_{i+1} = 1 : 2^{n/11} \tag{5.2}$$

入力信号と教師信号

Elman net への入力, 及び教師信号は実数値で表現した相対的な音高と音価の値となる. これはつまり, 1 時刻前の音高・音価からの変化量である. ここで, 教師用の音楽データはあらかじめ前述の平均律の手法を用いて, 記号から変化量の推移を算出し時系列データとしておく. そして, ある時刻 t の値を入力とし, 次の時刻 $t+1$ の値を出力とするようにモデルを訓練する.

ANN の構成

予測処理による評価モデルは Elman net を採用し, また上述したように, 音高・音価の変化量を予測するモデルとして訓練される. このモデルのネットワーク構造は Fig. 5.9 に示すように構成し, 各層のパラメータは Table 5.3 に示すように設定をした.

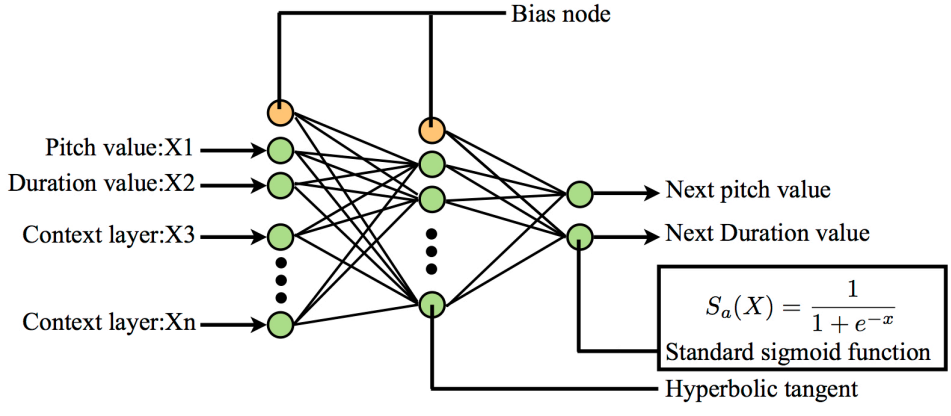


Fig.5.9: Evaluation model of time-series prediction by Elman net

Table 5.3: Parameter of predict model for evaluation

parameter name	parameter value
input layer size	2 + 1(bias node)
hidden layer size	32 + 1(bias node)
context layer size	32
output layer size	2
learning rate	0.1
learning epoch cycle	50000

検証

提案手法による音列パターンの評価が有効であるか検証を行った. モデルの学習にはクラシック曲より次の 3 曲を学習用データとして用いた.

A) パヘルベルのカノン
B) トッカータとフーガニ短調
C)「四季」第 1 番 春 ホ長調

また, 評価用データとしてクラシック曲より次の 3 曲を用意した.

D) 前奏曲 平均律クラヴィーア曲集 第 2 巻 プレリュードとフーガ第 1 番より
E) アリア ゴドルベルク変奏曲より
F) メヌエット ト長調

まず, モデルが予測処理を行うことができるように学習されたかを確認する. 学習用データを再度用いて予測処理を行った結果を Fig. 5.10 に示す. 図中のラベル A, B, C はそれぞれ前述の学習用データ A, B, C に対応する. また, 上段が音高, 下段が音価であり, 赤の破線が実データ, 緑の線が予測データである. 図より, 各楽曲に対して学習結果が良好であることが確認される. ラベル B と C は教師信号と出力値のグラフに重なりや増減のに

ついて同様の結果が見られ, 特に良好である.

次に, 未学習のデータについて予測処理の結果を見る. 評価用データをモデルに与え, 予測を行った結果を Fig. 5.11 に示す. 図中のラベル D, E, F はそれぞれ前述の評価用データ D, E, F に対応する. また, 上段が音高, 下段が音価であり, 赤の破線が実データ, 緑の線が予測データである. 図より, 未知のパターンに対しては正答率が低いことが確認でき, ラベル E 及び F の音高は特に誤差が大きい. しかしながら, ラベル D の音高, ラベル E と F の音価は誤差自体はあるものの予測値の推移傾向が類似したパターンを示している.

提案手法は音楽と音楽的経験による予測の関係性に着目し, 予測の正答率や予測傾向の類似性を評価指標として用いることを目的とする. 今回の結果では未知のパターンに対して, 同ジャンルの音曲でありながら大きな予測誤差が確認された. しかしながら, 実データと予測データの推移傾向が類似している点があることは無視できない要因である. また, 学習用データについては良好な結果であることから, 以下の要因を示唆していることが考えられる.

1. 学習用データの種類や量に不足があり, 十分な音楽的背景をモデルが獲得できていない.
2. 学習用データと評価用データが同ジャンルに分類されながらも差異が大きく, 学習用データから評価データの予測を十分に行うことができない問題設定である.

予測モデルを用いた評価についてその可能性が否定されたとは考えられないが, 学習データ・評価データの選定を含め, 再度の検討を要する.

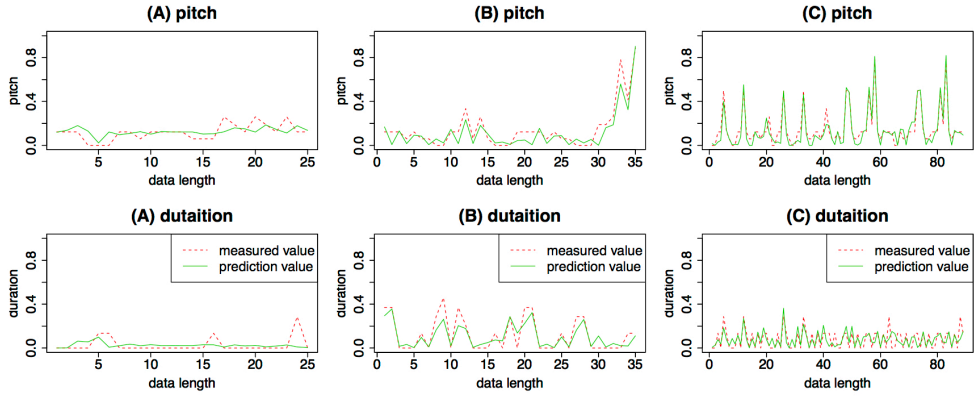


Fig.5.10: Prediction of known music

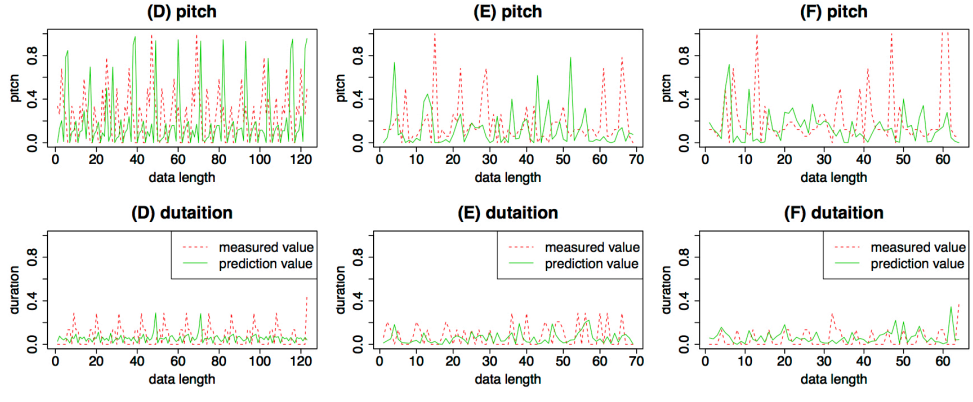


Fig.5.11: Prediction of unknown music

5.4.4 音楽とドーパミン

音楽を聴いているとき, 心理的あるいは生理的な反応を示すことがよく知られている. 例えば, 心理的には気分が落ち着いたり, 逆に興奮を誘発したり, 生理的には心拍数や呼吸数, 体温や血圧に変化を与えることが知られている. このため, 音楽療法のように音楽が人へ与える作用を応用して, 心身の回復の助けとしようという試みも確立されている. これについて音楽を聴いて満足感を感じると, 神経伝達物質であるドーパミンが分泌されること

が V. Salimpoor, R. Zatorre らの研究により明らかにされた (Salimpoor et al. 2011).

彼らは「背筋がゾクッとするような興奮 (ここでは満足度や満足感と呼ぶうこととする)」を感じさせる音楽に反応して分泌されるドーパミン量を, 音楽の満足度に関連する心拍数, 呼吸数や皮膚コンダクタンスなどの変化とともに測定した. この結果, 満足度の高い音楽ではドーパミンの分泌が活発になり, 分泌量は情動喚起や満足度に相関があることが示唆された. また, ドーパミンの分泌活動には 2 種類あり, 曲調の高まりを予感・期待 (anticipation) し気分が高まっていく過程と, クライマックスを経験 (experience) して感極まったとき, つまり予感・期待と経験の 2 つ局面のがあることも明らかにしている. さらには, 満足感を得る音楽をこれから聴く, という期待感だけでもドーパミンが分泌されることも発見された.

ドーパミンは報酬と最も関わりが深い神経伝達物質である. 一般的には, おいしい食事といった気持ち良い・心地よいと感じる体験をしているとき分泌が活性化し, タバコのようにドーパミン分泌に作用のあるものへの依存性形成に関与していると考えられている. これについて, 大脳基底核と呼ばれる部位では, このドーパミンが報酬自体や期待される報酬を表現していることが発見され, 生物の報酬を伴う行動の学習に関与していることがわかっている. 好きな音楽を何度も聴く, 聴きたくなるという心理にもドーパミンが関与している可能性は否定できないことであろう.

もし, 音楽を聴いているときのドーパミンの分泌を模擬できるようなモデルを構築することができれば, 自動作曲のみならず, 推薦システムや作曲支援システムといった他の音楽情報処理分野の研究領域においても有効な方法論を提案できることだろう.

5.4.5 生理反応による評価モデル

前述のように音楽を聴いて満足感を得ているときには, 神経伝達物質のドーパミンが活発に分泌されることが明らかにされている. ドーパミン分泌量を報酬や期待報酬といったスカラーな量と見なすと, 機械学習分野の強化

学習の方法論が適用できる．強化学習の方法論には, 状態の価値, つまり報酬や期待報酬の関数系を人工ニューラルネットワークで関数近似する手法がある．音楽を聴くときのドーパミン分泌の活動について, 音楽のある時刻の音を状態として与えて報酬あるいは予測報酬を出力する関数系とすると, 同様の手法が適用できそうである．

SMF の実時間表現

人工ニューラルネットワークで報酬関数を近似することを考える．ここで, モデルへ与えるデータは音高と時間の 2 次元領域のデータであることが望ましい．ある楽曲の中で報酬を与える時刻を決定するために, 実時間上でなければ人手での作業に困難が伴うためである．

そこで, MIDI ファイルに置いて最も一般に用いられている標準フォーマット (Standerd Midi FIle; SMF) から, 音高と音価の情報を音高-時間の 2 次元データに変換する．MIDI ファイル上では時間 (デルタタイム) の単位に tick が用いられる．これを 4 分音符あたりの時間分解能とテンポ情報を用いて 1 tick あたりの実時間を求める．テンポが BPM(Beats Per Minute) 120, つまり 60 sec に 120 個の 4 分音符, 4 分音符の分解能が 480 tick とすると 1 tick あたりの実時間 $t[msec/tick]$ は次式となり, 4 分音符あたりの実時間は 500 msec となる．

$$t = \frac{60/120}{480} \times 1000 = 1.0416[msec/tick] \tag{5.3}$$

また, 一般に楽器は打鍵後の発音時間中に音量が自然減衰する．そこで, これを考慮して発音時刻に 1.0, 以降 100ms ごとに 0.025 減衰するパラメータを音高に付与する．

上記二点の処理を加えたデータを SMF から作成し, これを以降スペクトログラム表現と呼ぶこととする．Fig. 5.12 にこのスペクトログラム表現のサンプルを示す．

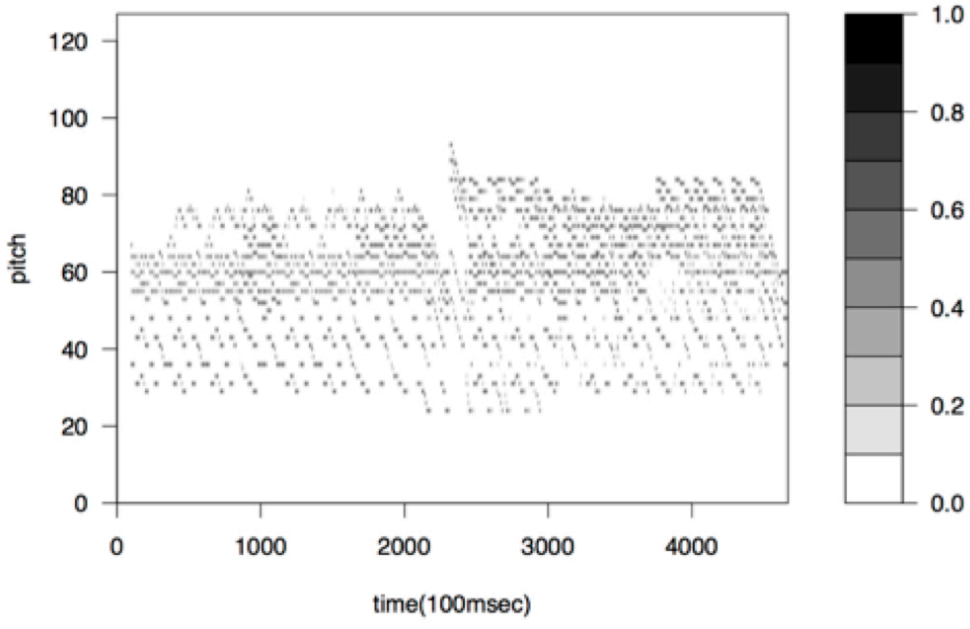


Fig.5.12: Sample of SMF spectrogram-pattern

ドーパミン分泌の模擬データ

ある楽曲に対応する実際のドーパミンの反応データを用意することはできない. そこで, これを模擬するための教師用データを仮定する.

音楽聴取時のドーパミン分泌の反応は報酬と期待報酬と捉えることができると述べた. これはそれぞれ, 体感と予感・期待に対応する. ここで, 体感については直接の報酬と考えられることから, 対象の楽曲を実際に聞いて条件に当てはまる場所, つまり聴取者が好む箇所で任意に報酬を設定できる. 問題は予感・期待を表す反応であるが, ここでは体感として設定した報酬を前に 15 sec シフトさせ, かつ下限値を 0.2 として報酬の 0.8 倍の値を仮定した.

この報酬, 期待報酬データのサンプルを Fig. 5.13 に示す. 赤の実線が体感に対応する直接の報酬を表し, 緑の実線が予感・期待に対応する期待報酬を表す.

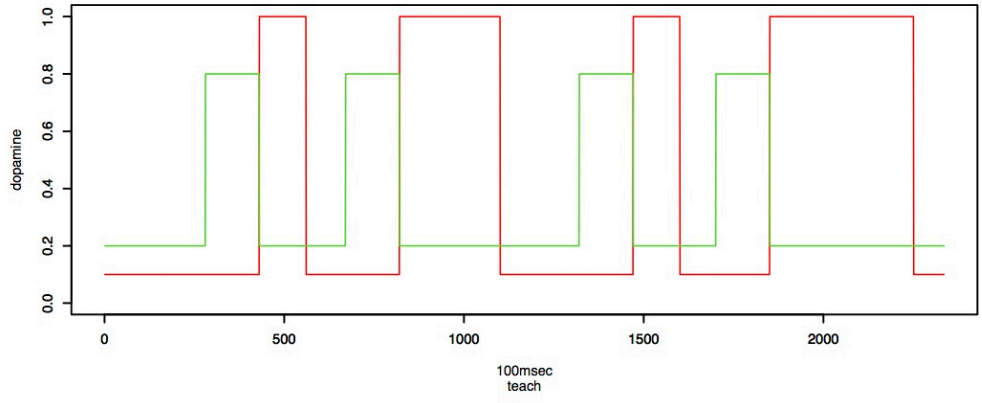


Fig.5.13: Sample of dopamine value pattern

入力信号と教師信号

Elman net への入力および教師信号は, 時刻 t において実数値で表現した各音高の音量のベクトルと報酬と予測報酬を表すベクトルである. つまり, 先に述べたスペクトログラム表現を入力に, これに対応するドーパミン分泌の模擬データを教師とする. 入力を時刻 t の各音高の音量の実数値ベクトルとし, 出力を時刻 t の報酬と予測報酬を表す実数値ベクトルとする. つまり, 音楽聴取時のドーパミン分泌の反応を再現するような関数を近似するモデルとして訓練される.

ANN の構成

生理的反応による評価モデルは Elman net を採用し, また上述したように, ある時刻においての報酬と期待報酬を予測する関数近似モデルとして訓練される. このモデルのネットワーク構造は Fig. 5.14 に示すように構成し, 各層のパラメータは Table 5.4 に示すように設定をした.

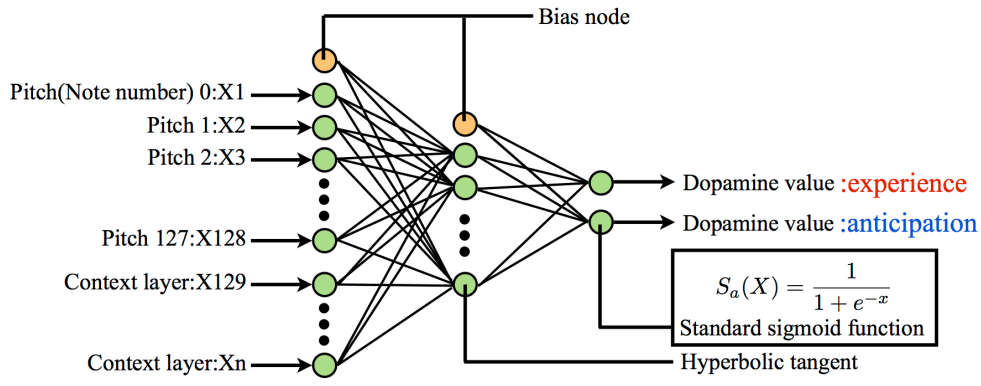


Fig.5.14: Evaluation model of dopamine value by Elman net

Table 5.4: Parameter of evaluation model by dopamine

parameter name	parameter value
input layer size	128 + 1(bias node)
hidden layer size	64 + 1(bias node)
context layer size	64
output layer size	2
learning rate	0.01
learning epoch cycle	1500

検証

提案手法による音列パターンの評価が有効であるか検証を行った. モデルの学習には The Beatles の「 Let It Be 」を用いた. 教師用信号を作成した者が特に好んでいた楽曲かつ, 楽曲の構造がシンプルなためである. また, 未知のパターンに対する反応を見るために次の楽曲を用意した. モデルがよく学習されているならば, 同じアーティストと異なるアーティストの楽曲間に何らかの差異が観察されるはずである.

a) I Me Mine (by The Beatles)

b) Maggie Mae (by The Beatles)

c) Mass in B minor (by J. S. Bach)

訓練終了後のモデルに「Let It Be」, および各未知パターンのスペクトログラム表現データを与えた際の出力結果を Fig. 5.15 から Fig. 5.18 に示す. ここで, 図の上段がモデルの出力, 中段が実データ, 下段がスペクトログラム表現データである. また, 赤の実線が体感を表す報酬の数値, 緑の線が予感・期待を表す期待報酬の数値である.

まず, 「Let It Be」のスペクトログラム表現データを与えた結果, モデルの出力信号は振動が見られるものの教師信号によくフィッティングしていることがわかる. これにより, 既知のデータについては報酬と期待報酬の関数を近似できることが確認された.

次に, 評価用の各楽曲のスペクトログラム表現データを与えた結果についてである. 結果 B, C の同じアーティストの楽曲では既知の楽曲より誤差を持つものの, 実データと同様の傾向を示している. 特に体感に対応する報酬を表す数値の推移は, 実データとして仮定した報酬値よりも, 既知の楽曲を背景にした報酬箇所を示しているようにも考えられる.

結果 D においてはモデルの出力値が激しい振動を見せている. これは, 既知の楽曲と全く異なるタイプの楽曲であるためとも考えられる. しかしながら, 既知の楽曲を背景としたときに, 報酬となるような箇所が断片的に含まれていると考えることもできるだろう.

これら結果 B から D のグラフを比較すると, 既知の楽曲と同じアーティスト異なるアーティストについて, モデルの出力の様子は明らかに異なる. この差異がどのような要因に起因するかを特定することは現段階では困難である. これについて, より多くの楽曲を背景に持つモデルを訓練させ, 複数のジャンルの未知曲を与えるなど, 追加の実験を要するだろう. しかし, 現段階においても, 出力傾向の差異を適切に処理することができれば, ある一つの楽曲を基準とした類似性の評価や推薦システムへの応用など, 楽曲の評価や応用が期待できるだろう.

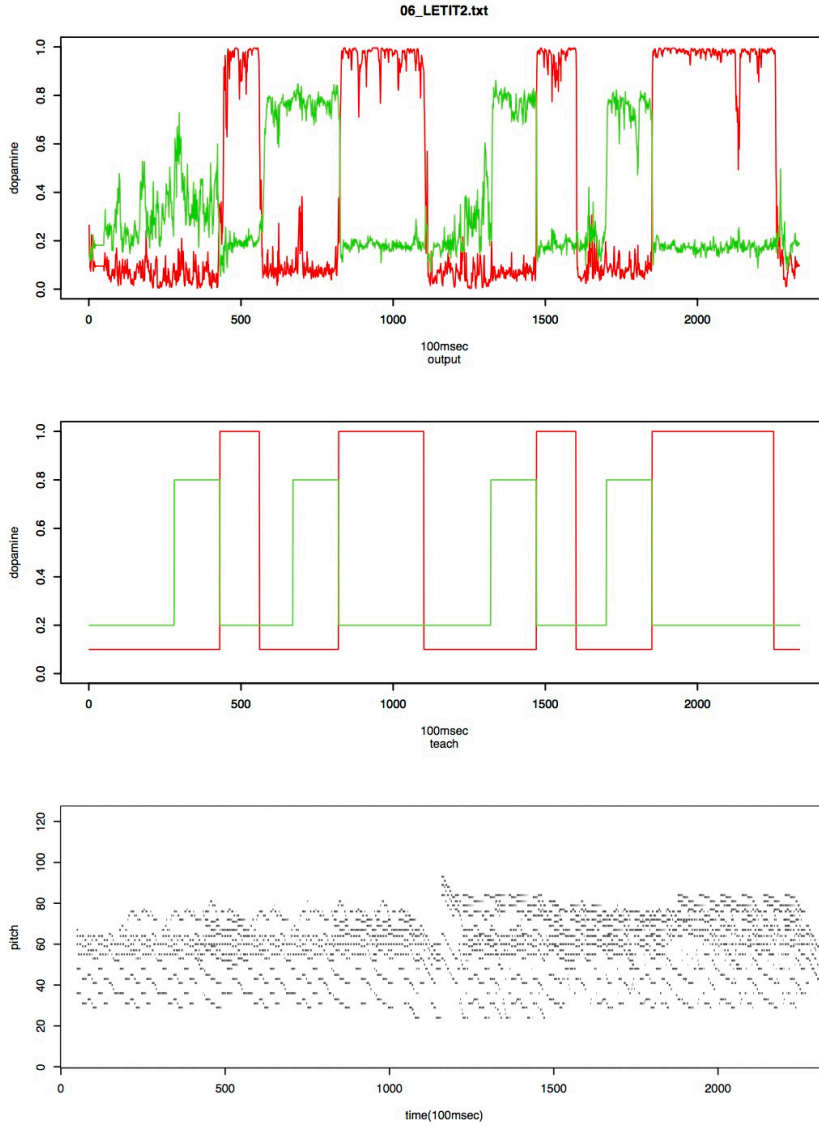


Fig.5.15: Result A: Let It Be

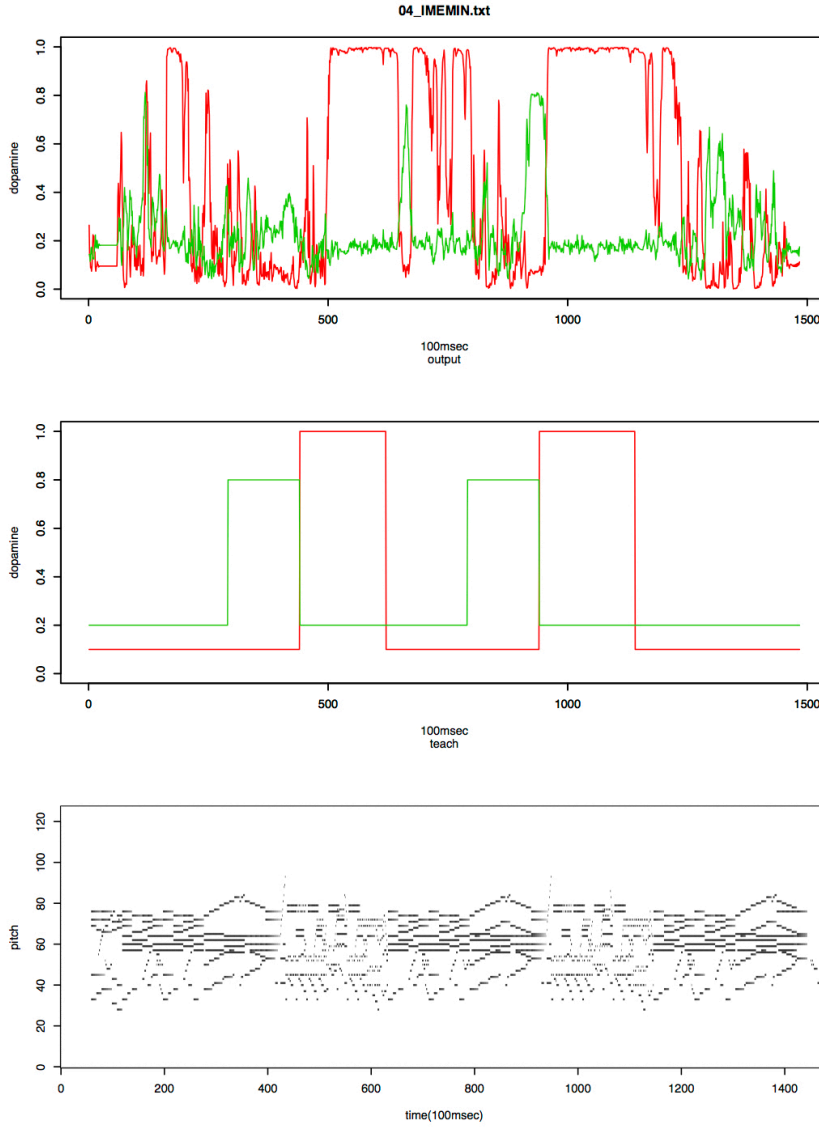


Fig.5.16: Result B: I Me Mine (by The Beatles)

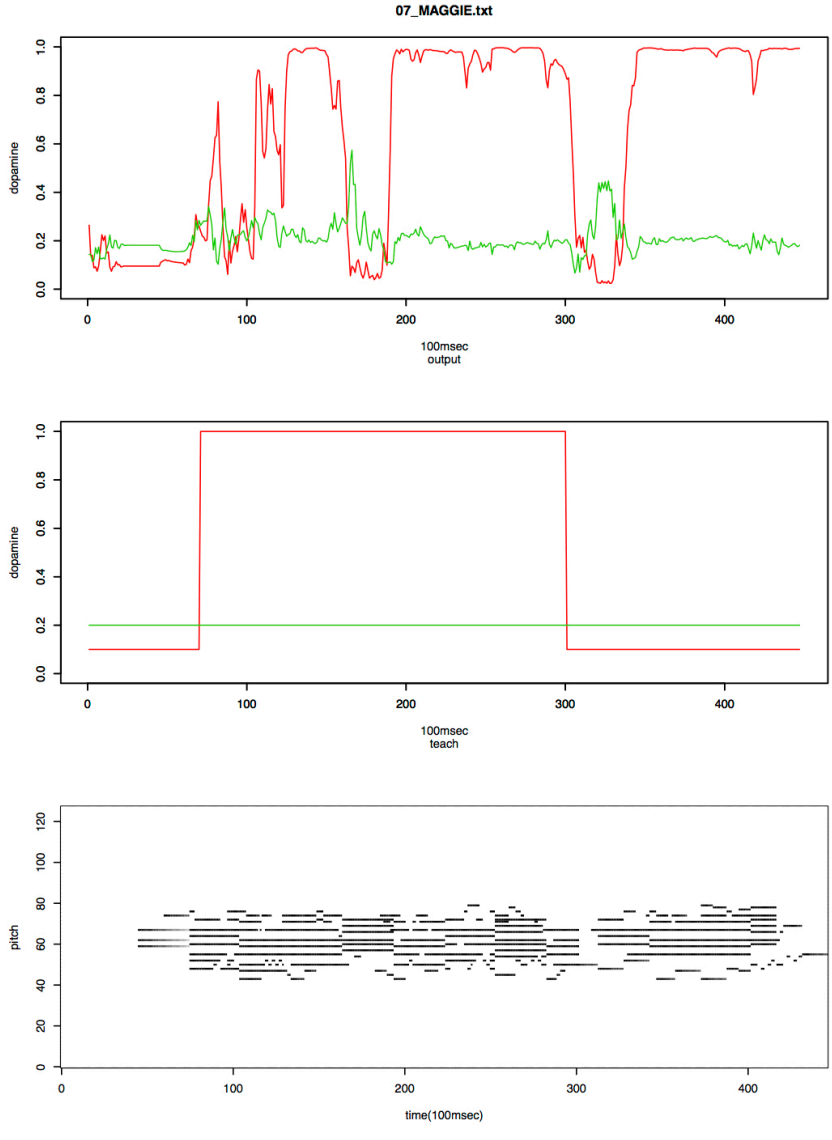


Fig.5.17: Result C: Maggie Mae (by The Beatles)

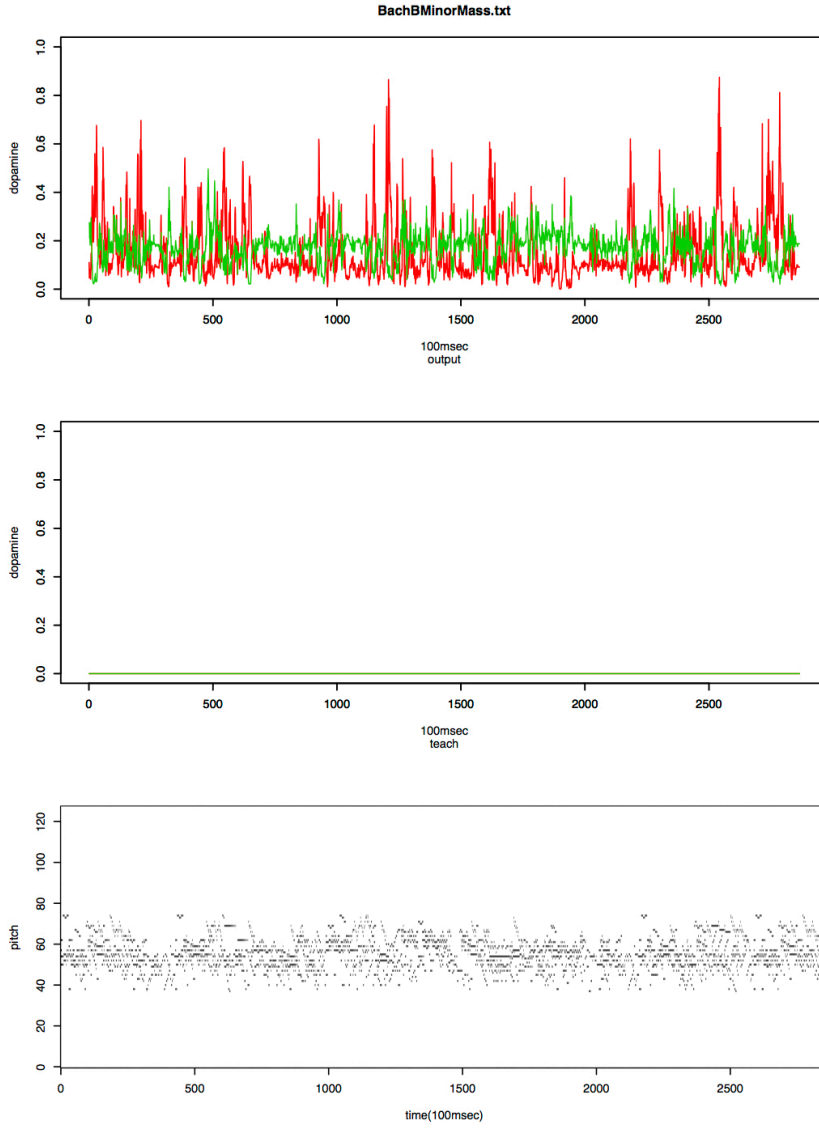


Fig.5.18: Result D: Mass in B minor (by J. S. Bach)

5.5 まとめ

本章では自動作曲システムについて, 生成と評価のモジュールを持つ新しいタイプのシステムを提案した. これは, 作曲活動全体の代替・再現を目指すものであり, 人工知能や機械学習の研究分野の側面と, 社会的には昨今のマルティメディア・コンテンツの素材に関連する緒問題への対応を目的とする. この自動作曲システムについて, 生成と評価の方法論の検討を本章で述べた.

まず, 生成については音楽を複雑ネットワークで表現し, その経路選択として新しい音のパターンを生成することを提案した. この結果, フレーズ程度の断片的なパターンは実用できると考えるが, 楽曲全体については追加の検討や他の方法論の導入を要するだろう.

次に, 評価については予測と生理的な反応に着目した 2 種類の方法論を提案した. いずれもリカレント型人工ニューラルネットワークを用いた方法論である.

片方は楽曲について音列の予測を行い予測誤差を評価に用いることを提案した. この結果, 未知のパターンでは同ジャンルの楽曲でありながら大きな予測誤差が確認された. しかし, 実データと予測データの推移傾向が類似している点から, 学習データ・評価データの選定を含め再度の検討を要する.

また, もう片方は楽曲に関連するドーパミン分泌についての反応を報酬と捉え, 強化学習における状態価値の関数近似手法を応用することとした. 既知の楽曲を背景として, 未知の楽曲中の報酬箇所を示すようにも考えられる結果が得られたが, これをより確かなものとするためには追加の実験を要する. しかし, モデルからの出力傾向の差異を適切に処理することができれば, 現段階においても, ある一つの楽曲を基準とした類似性の評価や推薦システムへの応用などが期待できるだろう.

今後は追加検討とともに, 生成と評価のモジュールを相互フィードバックを持つ一つのシステムとして, 自動作曲システムのプロトタイプを構築する.

参考文献

CC. (n.d.). Creative commons. https://creativecommons.org, Accessed 17 May 2016.

CCJP. (n.d.). Creative commons japan. https://creativecommons.jp, Accessed 17 May 2016.

Crypton. (n.d.). VOCALOID 初音ミク. http://www.crypton.co.jp/mp/pages/prod/vocaloid/index.jsp, Accessed 17 May 2016.

e-frontier. (n.d.). Band-in-a-box. http://www.biab.mu, Accessed 17 May 2016.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Hiller, L. A., & Isaacson, L. M. (1979). *Experimental music; composition with an electronic computer*. Greenwood Publishing Group Inc.

Logan, B. (2004). Music recommendation from song sets. In *ISMIR*.

Meyer, L. B. (2008). *Emotion and meaning in music*. University of chicago Press.

Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A., & Zatorre, R. J. (2011). Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature neuroscience*, *14*(2),

257–262.

YAMAHA. (n.d.). VOCALOID. http://jp.yamaha.com/products/music-production/ software/vocaloid/, Accessed 17 May 2016.

三浦雅展, & 江村伯夫. (2012). 和声理論に基づいた作・編曲システム (< 特集> 音楽制作と情報処理の友好関係). **システム/制御/情報: システム制御情報学会誌**, *56*(5), 213–218.

三輪多恵子, 田所嘉昭, & 斎藤努. (1998). くし形フィルタを利用した採譜のための異楽器音中のピッチ推定. **電子情報通信学会論文誌 D**, *81*(9), 1965–1974.

中野倫靖, 後藤真孝, & others. (2008). VocaListener: ユーザ歌唱を真似る歌声合成パラメータを自動推定するシステムの提案. **情報処理学会研究報告**, *2008*(50), 49–56.

亀岡弘和, 西本卓也, 篠田浩一, 嵯峨山茂樹, & others. (2003). ハーモニッククラスタリングによる多重音の基本周波数推定アルゴリズム. **情報処理学会研究報告**, *SIGMUS50-5*, 37–43.

剣持秀紀. (2012). 歌声合成システム vOCALOID の開発 (< 特集> 音楽制作と情報処理の友好関係). **システム/制御/情報: システム制御情報学会誌**, *56*(5), 244–248.

剣持秀紀, & 藤本健. (2015). **ボーカロイド技術論歌声合成の基礎とその仕組み**. ヤマハミュージックメディア.

吉井和佳, 後藤真孝, 駒谷和範, 尾形哲也, 奥乃博, & others. (2006). ユーザの評価と音響的特徴との確率的統合に基づくハイブリッド型楽曲推薦システ

ム. **情報処理学会 音楽情報科学研究会 研究報告**, *2006*(90), 45–52.

吉田由紀, & 中嶌信弥. (1999). 歌声合成システム: CyberSingers. **情報処理学会研究報告. *SLP,* 音声言語情報処理**, *99*(14), 35–40.

大槻知史, 齋藤直樹, 中井満, 下平博, 嵯峨山茂樹, & others. (2002). 隠れマルコフモデルによる音楽リズムの認識. **情報処理学会論文誌**, *43*(2), 246–255.

山田拓志, & 椎塚久雄. (1998). 遺伝的アルゴリズムを用いた自動作曲について. **情報処理学会研究報告*.[*音楽情報科学*]***, *98*(96), 7–14.

嵯峨山茂樹, 酒向慎司, 堀玄, & 深山覚. (2012). 確率的手法による歌唱曲の自動作曲 (< 特集> 音楽制作と情報処理の友好関係). **システム/制御/情報*:* システム制御情報学会誌**, *56*(5), 219–225.

後藤真孝, & 村岡洋一. (1994). 打楽器音を対象にした音源分離システム. **電子情報通信学会論文誌** *D*, *77*(5), 901–911.

梶克彦, 平田圭二, 長尾確, & others. (2004). 状況と嗜好に関するアノテーションに基づくオンライン楽曲推薦システム. **情報処理学会研究報告**, *127*, 33–38.

深山覚, 中妻啓, 米林裕一郎, 酒向慎司, 西本卓也, 小野順貴, et al. (2008). Orpheus: 歌詞の韻律に基づいた自動作曲システム. **情報処理学会研究報告**, *2008*, 179–184.

酒向慎司, 宮島千代美, 徳田恵一, 北村正, & others. (2004). 隠れマルコフモデルに基づいた歌声合成システム. **情報処理学会論文誌**, *45*(3), 719–727.

野地保, 荻野正, & 児山佳大. (2010). 心拍数を使った音楽検索システムの検討 (次世代経営情報技術, 一般). **電子情報通信学会技術研究報告*.* *SWIM,***

ソフトウェアインタプライズモデリング, *110*(184), 35–39.

齋藤直樹, 中井満, 下平博, 嵯峨山茂樹, & others. (1999). 隠れマルコフモデルによる音楽演奏からの音符列の推定. **平成 *11* 年情報処理学会音楽情報科学研究会資料**, 27–32.

第 6 章

結論

社会性を持つ昆虫や動物は他の個体と協調して, 自身の能力以上の仕事をこなすことができる. 例えば, 複数の個体で餌を運んだり, 連携して狩りをしたり, あるいは道のない場所に移動経路を作ることもある. 彼らのこのような協調作業において個体間のコミュニケーションは欠かせない能力であり, 多くは空気の振動, つまり音を媒介とすることが多い. 彼らはこの音を利用してコミュニケーションを取り, 音についての高度な感覚器官をもつ種では周辺環境で何が起こっているのか認識することが可能である. また, 人間においては音楽活動のように直接的なコミュニケーション用途ではなく, 音のパターンを高度に組み合わせ, 何らかの表現活動を通じた間接的なコミュニケーション手段とする活動も見られる. このとき, 対象の楽曲に合わせて多様な生理的反応を示すことが多くの研究により明らかにされ, また脳の活動としても報酬を司る神経系の中に音楽反応して活動する種類のものが観測されている. 音や音のパターンに対して何らかの反応と行動を示し, またその音のパターンを自ら生成するといった 能力は, 社会性と音を感知する器官を持つ生物にとって欠かせないものである.

このような生物の行動や能力について, ある環境への適応を伴うモデルとして再現する際には, 機械学習の方法論がよく用いられる. 従来では, 画像や音響信号のような現実の環境から観測させる高次元のデータを伴うタスクを適切に処理するためには困難な課題が山積していた. しかし, 近年になって

確立された深層学習という機械学習の方法論の一つがこの問題を解決しつつある.

深層学習は多層構造の人工ニューラルネットワーク (ANN: Artificial Neural Network) を用いた機械学習の方法論である. 多層構造による変数間の複雑な関連性により高い柔軟性と表現能力を持つ学習モデルであり, この多層構造がもたらす有益な特性として多重なエンコードを介した特徴の抽出・変換が知られている. つまり, 生の, あるいは最低限の前処理を施した観測データより, 適切なデータの表現が自動的に獲得されるような, 人の設計や仮説によらない特徴量の抽出・変換器である. 画像処理や音声処理あるいは予測処理の分野で驚異的な成果を収め, 画像や音響信号といった高次の情報源を適切に扱わなければならないタスクにおいても幾つかの課題が解決され, 幾つもの成果が報告されている.

本研究において, 我々は冒頭で述べたような社会性生物の持つ音を手掛かりとした行動を表現するモデルの構築を行っている. 具体的には, 音楽や音声など音の音響信号に関するデータを記号化せず, 時系列に並んだ周波数成分, あるいは時間領域の信号波形そのものを深層学習のメカニズムにより概念学習を行う. そして, ロボットが感知する音響信号からの行動決定や新しい楽曲の生成などに応用する. 記号を中間に介さず, 入力とする音響信号から直接的に行動や生成といった出力を行うモデルである. このための概念モデルとして, 強化学習に分類される Actor-Critic 法に深層学習のメカニズムを取り込んだモデルを提案し, これの構築を行っている.

強化学習の方法論において, 環境からの状態観測と観測されたデータの処理は特に重要な課題である. ここで音響信号を観測データとしたとき, その特徴量の設計と時間方向の依存性解決が提案モデルを構築するための第一の課題となる. 本論においては, 本研究の概念モデルを提示し, 音響信号の特徴量の設計と時間方向の依存性の解決を行うために深層学習によるメカニズムを取り入れた多層ニューラルネットワークのモデルについて, 検討とその検証を行った. ここで, 本論の各章の総括を次に述べる.

第 1 章　序論

一般に, 機械学習は人工知能研究の一つの領域として捉えられている研究の領域である. この機械学習は簡単に表現をすると, あるコンピュータ制御されるシステムが人間や他の生物のように経験から学習し, 周囲の環境に適応するにはどうすれば良いかというテーマを掲げている. 現代社会においては, 自動制御や情報処理において機械学習の重要性が増し, 人工知能研究への期待感も大きい. この章では, この機械学習が現代社会でどのような位置にあるのかを最新の深層学習まで簡単に紹介するとともに, 本研究のテーマとその対象とする音を観測状態とした深層学習と強化学習によるモデルについて触れた.

第 2 章　概念モデルと関連研究

この章では, 本研究のテーマとする対象についての概念モデルを提示し, 本論文で対象とする課題の範囲について規定した. 概念モデルは Actor-Critic 型の強化学習に深層学習のメカニズムを取り入れ, 記号を介さずに与えられた音響信号から直接的に行動を決定するようなモデルである. 入力とする音響信号データを処理する多層ニューラルネットワークがセンサ系と ACtor, Critic の間に介在する構造とし, 多層ニューラルネットワークは特徴抽出を行う層と時間依存性を解決する層を持つ. 音響信号の特徴量の設計と時間方向の依存性は, 深層学習のメカニズムにより自動的に学習され, この概念が多層ニューラルネットワーク上に獲得される. このような概念モデルを提案し, 多層ニューラルネットワークの方法論の検討と構築についてを本論で扱う課題の範囲として規定した.

第 3 章　深層学習による時間領域の信号波形ベースのモデル

従来の信号処理では, 時間領域で表現される信号波形を周波数分析を行い, 時間-周波数領域の情報に変換する. この周波数成分の時系列データについて, 各種の分析を経て特徴量の設計や特徴抽出を行い, この特徴量を分類器などの入力としてタスクの達成を目指す. しかし, 深層学習のメカニズ

ムであれば, 時間領域の信号波形より直接的に特徴抽出や時間方向の依存性を解決可能な多層ニューラルネットワークの学習が可能であるかもしれない. この章では, 制約ボルツマンマシン (Restricted Boltzmann Machine; RBM) と Conditional RBM と呼ばれる生成モデルによる多層ニューラルネットワークを提案し, 時間領域の信号波形をある区間で分割した時系列データをモデルの訓練データとして訓練データの予測と信号波形の復元精度を実験により確認した.

ここで, まずは信号波形を入力として単体の RBM のみを学習させ, 精度よく信号波形を復元できる生成モデルが学習可能であること, その信号波形の特徴が RBM の隠れ変数の出力パターンとして表現されることを事前実験により明らかにした.

これに基づき, 訓練後の提案モデルより予測と復元を介して得られた音響信号について, 周波数分析と被験者による聴取実験によって評価を行い, 提案モデルが信号波形から直接的に特徴を抽出し, 信号波形の時系列データを忠実に記憶できることが示された. しかしながら, 周波数分析結果, 聴取実験のどちらにおいても, 元の信号波形に白色ノイズと類似した特性を持つ雑音が混入してしまうことも明らかに成ったが, 聴取実験の結果からは雑音の強度は弱く, その要因は学習誤差やモデルの各種パラメータの影響であることが推測される.

これらより, 本章での提案モデルは時間領域の信号波形をある区間で分割した時系列データについて, その特徴量の自動獲得と時間方向の依存性を解決可能な多層ニューラルネットワークを学習可能であることが示された.

第 4 章　深層学習と聴覚フィルタおよび Echo State Network によるモデル

音響信号を観測して出力を決定するシステムは, 雑音のある環境下においても適切な出力ができるように雑音に対してのロバスト性を求められる. そこで, 深層学習のメカニズムと聴覚フィルタによる周波数分析を用いることで, 特徴抽出や時間方向の依存性に加え, 雑音環境下においてもロバスト性を有する多層ニューラルネットワークの学習を試みた. この章では, 自己符号化器 (Auto Encoder; AE) とエコーステートネットワーク (Echo State

Network; ESN) による多層ニューラルネットワークに聴覚フィルタによる周波数分析を加えたモデルを提案し,IEEE AASP Challenge の D-CASE challenge に含まれる OL subtask を課題としてモデルの検証を行った．この OL subtask は, オフィスで頻繁に発生する音響イベントについて, 雑音を含む実環境下での多クラス分類課題とその課題のための音響信号データセットである．

ここで, まず ESN のように時間依存を解決できるリカレント型ニューラルネットを用いず, 聴覚フィルタによる周波数成分で得られた周波数成分について, 時間方向の入力バッファをとって畳み込みニューラルネットワーク (Convolutional Neural Network; CNN) を学習させ, 音響イベントの多クラス分類課題の精度が乱数によるクラス分類と比較し有意差が見られないことを事前実験により確認した．

これに基づき, 訓練後の提案モデルにより音響イベントの多クラス分類課題を行った．この結果, 2013 年度の OL subtask における各チームの結果のうち, 上位のモデルには届かなかったものの, Baseline モデルの精度を超え, F 値にして 30.17 % の分類精度を記録した．また, これは先の事前実験の結果に用いたモデルよりも高い分類精度であり, 時間方向の入力バッファによるモデルよりもリカレント型ニューラルネットワークのメカニズムを取り入れることが, 音響信号の時間依存性の解決に有効であることが示唆された．ここで, 精度が向上しない主要な要因は実環境下における雑音の影響や, 振幅スペクトログラムについての時間軸方向へのズレをモデルが適切に吸収できなかった可能性などが考慮される．

これらより, 課題が残るものの, 本節の提案手法においても, 音響信号の特徴量の自動獲得と時間方向の依存性の解決に加え, 雑音環境下でのロバスト性を獲得できる多層ニューラルネットワークを学習できる可能性が示されたが, 雑音環境下でのロバスト性についてはさらなる検討と実験を要する．

第 5 章　人間の作曲活動をモデル化した自動作曲システム

音楽を自動的に生成するシステムは自動作曲システムと呼ばれており, 記号を介さずに与えられた音響信号から直接的に行動を決定するような

モデルの対象問題として, モデルが達成すべき課題をいくつも内包している. この章では, 本研究の概念モデルについて, 自動作曲を対象問題として Actor-Critic の強化学習部分の構築と学習に用いる報酬と呼ばれるスカラーな量の設計について検討と検証を行った.

本論では, まず初期の段階として, Actor は記号情報を用いたネットワークを用いて, 記号情報の出力を行うコンポーネントとした. 具体的には, Actor は音楽における音高-音価 (音の高さ-音の長さ) の組をノード, 各ノード間を遷移の確率を持った有向のリンクで接続したネットワークとし, ある状態の遷移確率に従ってネットワーク上のノードの音高-音価を順次出力する. Critic は Actor の出力パターンの音を評価し, Actor の持つネットワークの構造をある目標に向かって修正するコンポーネントと定義した. ここで, 問題設定として探索する状態空間に制限を加えるために, このネットワークの初期の構造は既存の楽曲から構成し, ノードの追加は不可とすることとした.

この Actor について, まずは Critic を介さずに独立に動作させ, 出力となるノードのパターンを得たところ, 音楽において定義される 2 小節程度の断片では, 人間に不快感を与えないような音楽的なパターンを得られた. しかし, 長期的なパターンでは音楽によく見られる構造的性質や幾つかの条件を満たさず, 人間に不快感を与えるパターンが出力され, 本研究で提案するように, 長期的な状態履歴への依存を解決するコンポーネントが不可欠であることが示された.

また, Critic により学習を行う段階において, 音楽という題材ではどのように報酬を設計することが適切か新たな課題が発生した. 音楽は報酬遅れを伴う問題の典型例である. ある楽曲の評価は制作後に聴取した人間によって与えられるが, 制作段階で逐次それが得られるケースは多くはなく, 制作段階で製作者自身が聴取して評価することになる. これを一つのシステムと見立てると, 報酬を与えるための系がシステムの外部のみにあるのではなく, 外部の報酬系とは独立した報酬系がシステムの内部に内包されていることが示唆される.

そこで, この報酬の設計について自身の出力と予測との誤差による評価の

モデルと, 音楽に関して明らかにされているドーパミン神経系 (報酬を司る脳の神経系) の働きを模倣するような評価のモデルを提案し, それぞれ検証を行った. これらのモデルは, それぞれリカレント型ニューラルネットワークを用いた教師あり学習によって構成した.

予測との誤差による評価のモデルは, 音楽的な経験を背景に, 人間は予測処理をしながら音楽を聴取しているとの Meyer による指摘などを基に構築を行った.

この結果, モデルの訓練に用いたデータの予測結果は訓練データとよく合致し, 未知のデータに対しても, 訓練データとよく類似する部分については実データと類似する予測傾向が得られ, また類似しない部分については予測を大きく外すという傾向が確認された. 音楽的な経験を背景とした予測処理を表現するようなモデルが獲得されている可能性が考慮できるが, 現段階ではさらなるモデルの検討と発展を要する.

ドーパミン神経系の働きを模倣するような評価のモデルは, V. Salimpoor, R. Zatorre らの研究により明らかになった音楽に関連するドーパミン神経系の働きをヒントにして構築を行った. V. Salimpoor, R. Zatorre らの研究では, 満足度の高い音楽ではドーパミンの分泌が活発になり, 分泌量は情動喚起や満足度に相関があることが示唆され, ドーパミンの分泌活動には期待と経験の 2 つ局面のがあることも明らかにしている. ここで, 実際の神経系の実データを得ることは現実的ではないため, 体感や期待に対応するスカラーな値を人の手で仮定した模擬データを設定し, モデルの訓練を行った.

この結果, 同じアーティストの曲と異なるアーティストの曲では, 出力パターンに著しい差異が見られた. 特に未知のデータに対しては, 体感に対応する報酬を表す数値の推移は模擬データとして仮定した報酬値よりも, 既知の楽曲を背景にした報酬箇所を示しているようにも見られる出力傾向が得られている. 体感に値するデータを何らかの手法でスカラー量として定義することが達成できれば, 十分に報酬の関数として機能することが示唆される.

謝辞

本博士論文は, 筆者が北海道科学大学 (旧名：北海道工業大学) 大学院工学研究科 電気工学専攻の修士課程及び博士後期課程在籍中に, 川上研究室において行った研究を総括したものです.

本研究を学位論文として纏めるにあたり, 主査並びに副査として懇切丁寧なご指導とご鞭撻を賜りました北海道科学大学工学部情報工学科　川上敬教授, 北海道科学大学未来デザイン学部メディアデザイン学科 木下正博教授, 三田村保教授, 並びに北海道科学大学工学部情報工学科 和嶋雅幸教授に深く御礼申し上げます.

川上敬先生には, 私が学部生の頃から修士課程そして博士課程と, 長い間にわたって大変お世話になりました. 学会での口頭発表、論文の投稿に際しては、それぞれの文面のご確認と詳細なご助言をいただきました. また, 学会や研究会をはじめとして, 多くの成長する機会を与えて下さいました. 頂いた機会を通じ, 様々な方々との出会いや多くの経験を得たことは何事にも代えがたいものです. このように研究を進め, 博士論文を書けるようになったのは, 川上敬先生にご指導を賜り, 道を示していただいたからに他なりません. 改めて心より御礼を申し上げます.

北海道科学大学工学部情報工学科助教 大江亮介先生には, 論文だけでなく研究活動全体において多くのご助言とご忠言をいただきました. 特に博士論文執筆の終盤ではご多用の中にもかかわらず, ご指導をいただき深く感謝しております.

北見工業大学工学部情報システム工学科助教 岩館健司先生には, 学会な

どでお会いした際に私の疑問に幾度もお付き合いいただき, その中の幾つかが研究や本博士論文執筆の際のアイディアやヒントとなりました. 北海道科学大学工学部情報工学科, 並びに創生工学部情報フロンティア工学科の先生方には, 私の研究活動に関しまして日常業務における様々なご配慮を頂きました. そして, 同研究室の先輩方と同期・後輩諸氏には, 普段より議論や雑談を通じて, アイディアや思考の整理を助けていただき, また良い刺激を与えていただきました.

お世話になりました先生方, そして支えて下さった多くの方々のご助言・ご支援・ご協力・励ましに対しまして, 深く感謝申し上げます.