5S-04

ディープラーニングによる自動採譜システムの開発

土村　貫太†　小松　貴大‡
福井工業高等専門学校　電子情報工学科

1. 研究背景

近年、動画･音楽配信サービスの普及により、音楽の入手経路が増えている。一方、音楽検索の際はタイトル・アーティストなどを参照して既知の音楽を探す手段が多く、未知の音楽検索には向いておらず、自分の好みの音楽を探し出すのは難しい。テンポや使用楽器、曲調から音楽を検索できれば好みの音楽を探しやすくなるが、一般的に音楽ファイルは wav や mp3 といった形式で、この形式から音楽の特徴を抽出することは困難である。そこで、音楽ファイルを楽譜に変換することができれば、楽曲にジャンルや長調・短調、主旋律などのタグ付けを自動で行う事ができ、音楽検索の可能性がさらに高まると考えた。本研究では、wav ファイルを midi ファイル(SMF)に変換する自動採譜システムの開発を目的とした。また、その手段として Deep Learning を用いた。

2. 概要

2.1　開発環境

本研究では Python 3.7 と Keras、NumPy、Pandas、Pretty Midi、Scikit-learn を使用した。midi の音源は windows10 標準搭載の音源である Microsoft GS Wavetable SW Synth を使用した。

2.2　システムの概要

wav ファイルを、あらかじめ音階を分類するように学習した分類モデルに入力すると自動で採譜が行われ、midi ファイル(SMF)が出力される。

2.3　音楽ファイルの形式

今回扱う wav ファイルと midi ファイルはフォーマットが大きく異なる。wav ファイルは音の波形を一定周期で標本化・量子化したもので、mp3 や wma、aac などは wav ファイルを圧縮したものである。演奏した音楽や音声をそのまま収録するのに向いている一方、非圧縮の wav ファイルはサイズが非常に大きくなってしまうという欠点がある。midi ファイルは演奏の命令のみが入っているファイルで、データサイズは非常に小さいが、再生される音楽は音源に依存する。midi シーケンサなどのソフトウェアを使うことで、人間が直感的に理解できる楽譜を表示できる。

Music transcription by use of deep learning
† Tsuchimura Kanta, National Institute of Technology, Fukui College‡
‡ Komatsu Takahiro, National Institute of Technology, Fukui College

2.4　深層学習について

機械学習には様々なアルゴリズムがあり、脳の情報処理ネットワークをコンピュータ上で模型化したものをニューラルネットワーク(Neural Network, NN)と呼ぶ[1]。入力層から情報を受け取ったニューロンは、受け取った情報に重みを付けて次の層へ伝達していき、その過程で与えられた情報について「特徴量」を見出す。そしてその特徴量から結果を得るのが NN の目的である[1]。深層学習（Deep Learning）は NN を用いた機械学習の一種で、NN を多層化した DNN（Deep Neural Network）を用いる。

2.5　音階分類

音階を分類するためには、DNN に対して、それぞれの音の波形がどのような特徴を持っているか判断しやすいデータを与える必要がある。しかし、音の波形そのまま（時間領域）では特徴を見出すのは困難である。そこで、音の波形を高速フーリエ変換(FFT)することで周波数領域に落とし込み、その音が持つ基準周波数を DNN に与える。基準周波数とは音階が持つ固有の周波数のことで、A4 を 440Hz として決められている。図 1 に、A4 を鳴らしたときの周波数スペクトルを楽器ごとに比較した図を示す。

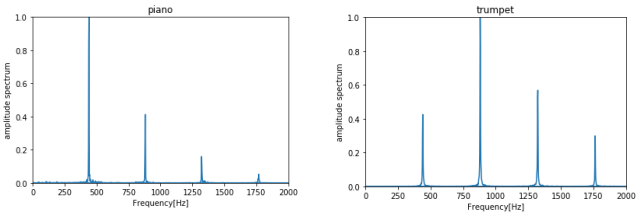

図 1　周波数スペクトルの比較
（左：ピアノ A4　右：トランペット A4）

図 1 より、基準周波数とその定数倍の周波数のスペクトル強度が高くなっており、同じ音階でも楽器によって最大強度となる周波数が違うことが分かる。

2.6　音階分類モデルの作成

自動採譜を実現するため、DNN による音階分類モデルを作成する。まず、1 音のみを出力する midi ファイルを作成し、それを wav ファイルに変換後、FFT し、周波数に対するスペクトル強度を得る。そのスペクトル強度を[0:1]で正規化し配列に与え、配列の最後尾にはその音階を示すラベルを付与する。これが、1 つの音を示す教師データとなる（図 2）。この処理を教師データとして与える音の数だけ行い、DNN に与える 2 次元配列を作成し DNN に与える。与えた教師データの 95%を訓練データ、5%を検証データに分割し学習させることで音階分類モデルが作成

される。図 3 に、今回使用した DNN の構造を示す。

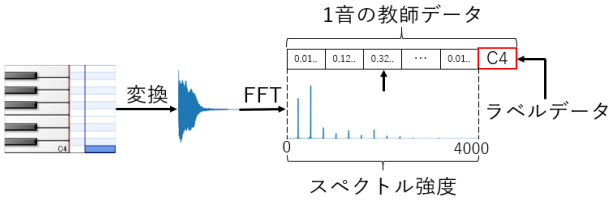

図 2　教師データ作成の流れ図

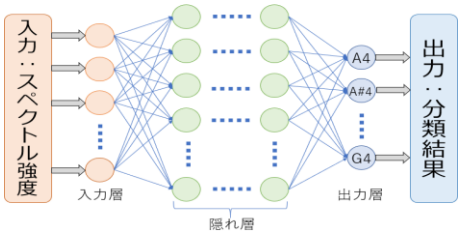

図 3　DNN の構造

2.7　自動採譜

採譜したい wav ファイルを一定間隔でサンプリング・FFT し、その結果得られたスペクトル強度を音階分類モデルに与え、得られた結果を基に midi データを出力する。図 4 に自動採譜手順の図を示す。

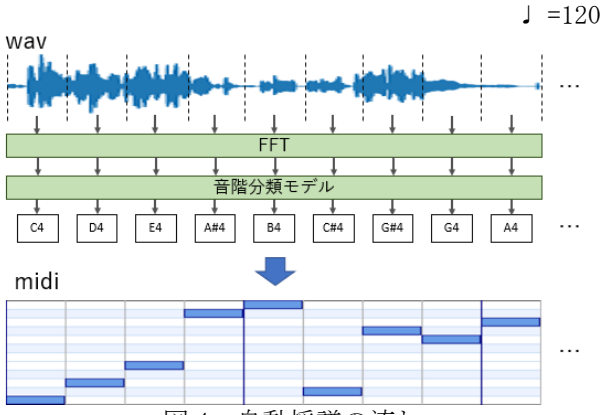

図 4　自動採譜の流れ

3. 実験結果

音階分類の精度確認実験の結果を以下に示す。

3.1　単音分類

ピアノ 4 種類、トランペット 4 種類の計 8 種類の楽器を用いて C4～B4 の 12 音階を 1 音鳴らし、音階と楽器を分類させる実験を行った。12 音階を分類させると同時に、楽器がピアノかトランペットかを分類させるため、出力層数は 12×2=24 個とした。音の長さ(10 段階)とホワイトノイズ(20 段階)を変化させてデータ数を水増ししたため、総学習データ数は 8×12×10×20=19200 となった。この学習データを与えて学習させた分類モデルに、無加工の検証データ 24 個（ピアノとトランペットそれぞれ 12 音階）を与え音階と楽器を分類させた結果、いずれも 100%の精度で分類することができた。ただし、ここで与えた検証データの演奏楽器は学習に用いた 4 種類の内 1 つのみなので、より正確な検証のためには 4 種類の楽器全てで試す必要があった。これは後の実験(3.2、3,3)でも同様である。

3.2　和音分類（同じ楽器の組み合わせ）

2 つの音を同時に鳴らし、その 2 和音の音階を分類させる実験を行った。使用する楽器はピアノ 8 種類で、同一種類のピアノで和音を構成した。2 和音の組み合わせは ${}_{12}C_2$=66 パターンである。また、1 音のみ鳴らした場合も分類できるようにするため、出力層数は 66+12=78 個とした。音の長さ(10 段階)とホワイトノイズ(20 段階)を変化させてデータ数を水増ししたため、総学習データ数は 124800 個となった。この学習データを与えて学習させた分類モデルに、無加工の検証データ 78 個(2 和音 66 個+1 音 12 個)を与えて分類させた。その結果を表 1 に示す。なお、2 和音の場合は 2 音の音階がどちらとも一致した場合に限り正答とする。

表 1　3.2 分類結果

一致数	割合[%]
72/78	92.3

正確に採譜できなかった音はすべて和音で、その数は 6 つだった。採譜できなかった音は全体的に周波数が低い傾向にあった。

3.3　和音分類（違う楽器の組み合わせ）

2 つの楽器が違う音を同時に鳴らし、その 2 和音の音階と楽器を分類させる実験を行った。楽器はピアノ 4 種類、トランペット 4 種類を使用した。1 つの音がとりうるパターンは 12×8=96 個あるため、2 和音の組み合わせは ${}_{96}C_2$=4560 パターンである。1 音のみの場合も考慮すると合計 4560+96=4656 パターンの音が存在する。ホワイトノイズ(40 段階)を変化させてデータ数を水増ししたため、総学習データ数は 186240 個となった。2 和音を構成する音の音階(12 個)と楽器(ピアノかトランペットかの 2 種類)を分類するため、出力層数は ${}_{24}C_2$+24=300 個となる。この学習データを与えて学習させた分類モデルに、無加工の検証データ 300 個を与えて分類させた。その結果を表 2 に示す。

表 2　3.3 分類結果

	一致数	割合[%]
音階・楽器一致数	77/300	26
音階一致数	283/300	94
楽器一致数	81/300	27

表 3 より、音階分類の精度は 94%と高いが、楽器分類の精度は 27%と低いことが分かる。そのため、音階と楽器が共に一致した割合も低い。また、楽器が一致してない音の多くはピアノとトランペットを同時に鳴らした音だった。

4. 参考文献

[1]: 斎藤　康毅(2016)『ゼロから作る　Deep Learning』　O'Reilly Japan