社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－MUS－74 (13)
2008－SLP－70 (13)
2008／2／9

構造解析を利用した機械学習による演奏表情の解析と付与

山上 信一† 但馬 康宏† 小谷 善行†

本稿では、楽曲の構造が演奏表情に大きく関わると仮定し、楽譜から読み取れる音符の情報と構造の情報を入力として機械学習を用いて演奏表情の学習を試みる。学習が終了した後、構造情報の重要性を調べるために、学習の際に利用した素性と学習の出力の相関関係を求めることで構造情報の有用性について調べた。さらに学習結果を元に楽譜相当のデータに演奏表情の付与を行い、そのデータについて試聴実験を行うことで学習の結果の評価を行った。その結果、オリジナルの演奏には及ばないものの、演奏表情が付与されているという評価を得た。

Analysis and Attach of Musical Expression by Machine Learning using Analysis of Musical Structure

Yamagami Shinichi †Tajima Yasuhiro †Kotani Yosiyuki †

In this paper, we suppose that music expression is influenced by musical structure, so we try learning music expression by machine learning using information from score and the information which get from musical structure as input for machine learning. After learning, to examine the importance of music structure, we investigate a correlation between input and output. And we attach music expression to a score data, and we evaluate it by audition. As the result, it was not match for original performance but it evaluated as expressive music performance.

1. はじめに

人間による音楽演奏は、与えられた楽譜をそのままなぞるものではなく、テンポの緩急や音量の増減など変化に富んだものであり、いわゆる表情を持っている。音楽が芸術として人々を魅了するためにはこの表情が必要不可欠であり、演奏の価値を決めるものと考えられる。

演奏表情を扱った研究は過去にも存在するが、複雑な要因をすべて扱おうとするあまり限られた状況でしか適応できなくなってしまう問題や[2][3]、事例ベースで演奏表情を生成するもの[4]では、旋律の類似度に表情が左右されるため、楽譜上同じフレーズであっても異なる演奏表情を持つ場合などに対応できない。

本研究では、機械学習であるニューラルネットを用いることにより、音楽の演奏データから演奏表情の学習を行い、その結果を音高と音価のみ既知であるような楽譜に相当するデータに適用することで計算機により演奏データを作成した。機械学習では、演奏表情として音符の発音された長さと音量に着目することにより、演奏データとそれに対応する楽譜相当のデータにおいて、どのような状況でどのような演奏表情が付与されていたかを学習した。状況とは演奏表情を調べる対象となる音符の前後の音符の情報や、楽曲の構造解析を行った際に得られる構造に関してのデータなどである。これらの情報はどの楽曲にも存在するものを利用しており、演奏表情の付与されていない音符列のみの楽譜相当のデータや、学習に利用していないデータに関しても、これらの情報を抽出し、学習結果を適用して演奏データを作成できるという利点がある。構造解析に関しては、まったく同じ繰り返しフレーズであっても演奏される際にはその表情が異なることなどから、楽譜上に記載されている目に見える

†東京農工大学
†Tokyo University of Agriculture and Technology

音符情報と同等に演奏表情に大きく関わるものと仮定して実験を行った。

本稿では演奏データからの学習と、学習結果をもとに演奏表情を付与した演奏データに関しての試聴実験について述べる。

2. 本研究における演奏表情の定義

本研究では音楽データとして MIDI データを利用している。そのため、MIDI データから利用できる情報を考慮し、演奏表情を音長と音量の２点に限定し次のように定義する。

- 音長の変化：楽譜上に記載された音価と、演奏の際に実際に演奏された音長の差。
- 音量の変化：音量の増減。実験においては曲ごとに平均音量を計算、もしくは仮定し、その平均音量からの変移。

これら二つを合わせて演奏表情とする。機械学習を行う際にはこれらを教師値とし、MIDI データより得られる各種の情報を入力とすることにより学習を行い、出力はこの演奏表情の増減の多寡とする。

実験に用いたデータは人間による演奏を MIDI データにしたものであり、曲はピアノのバイエルである。このデータからメロディ部分を取り出し、学習の際にはこのデータを用いて実験を行った。

3. 演奏表情学習の原理と演奏表情付与の方法

3.1演奏表情の学習と付与の流れ

演奏表情の学習はニューラルネットを用いて行うが、学習と演奏表情付与の流れは次のようになる。

1. EMG（後述）を用いて演奏データの構造解析を行い、楽曲を木構造で表す
2. 構造解析の結果と演奏データに対する楽譜相当のデータから、教師値を含めニューラルネットへの入力を作成する
3. ２．で述べた情報をニューラルネットへ入力することによりにより学習を行う
4. 楽譜相当のデータから２．で挙げた情報を抽出して学習の終わったニューラルネットに入力し、その出力をもとに演奏表情を付与する

3.2 EMGによる楽曲の構造解析

１節でも述べたように、本研究では楽譜に載っている情報と同様に、楽曲の構造も演奏表情に深く関わると考えている。そこで学習の際のニューラルネットへの入力として、本研究では楽曲の構造解析の結果を利用した。構造解析には EMG[3]を使用した。これは単旋律の楽曲の構造分析の結果を、末端に一つの音符が配置されるような木構造で出力するものである。詳細な解析方法については参考文献に譲るが、本研究ではこの構造上の特徴が演奏表情に深く影響していると仮定し、解析の結果から得られる情報を学習器への入力として利用した。構造解析された結果のモデルを図１に示す。図１のアルファベットにあたる部分が末端となり一つの音符となる。

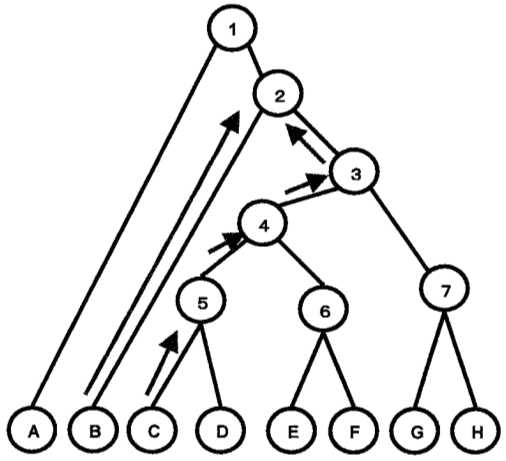

図１EMG による構造解析の結果と構造上の距離のモデル

このように解析された楽曲から、次に挙げる情報を取り出しニューラルネットへの入力とした。

- 音符のあるノードの深さ
- 構造上の距離
- 直後の音符の構造上の距離
- 後続の音符との構造上の距離の差
- 後続の音符との構造上の距離の和

構造上の距離とは、対象となる音符と直後の音符を末端とする最小の部分木のルートまでの距離である。図１の音符 B であれば、音

符 B と直後の音符 C を末端とする部分木のルートであるノード No.2 までの距離であり、この場合の音符 B の構造上の距離は１である。入力には直後の音符の距離も入力している。具体的にここでは音符 C の構造上の距離で４となる。距離の和と差はここで説明した２つの音符の構造上の距離の和と差であり、これは、距離の和が大きくなればなるほど構造の切れ目が大きいことを示し、差が大きくなればなるほど偏った木であることを表している。

構造解析の結果からこれらの５つの情報を取り出したものと、ノードの深さ以外の４つの情報については前後２つの音符に対しても同様に情報を取り出し、計２１個のデータを学習器への入力とした。

3. 3ニューラルネットへの入力情報

ニューラルネットへ入力となる教師値と素性についての説明を行う。

3. 3. 1学習の際の教師値

演奏表情の定義より、音長と音量に関しての教師値を次のように定義する。

- 音長の変化： EMG において構造解析する際に使用した音価を元に、それらが実際の演奏時にどの程度長く、あるいは短く演奏されたかを抽出する。具体的には、音価と同じ長さで演奏された場合に０．５、その曲の中で音価ごとに最も長く発音されたものを探し出し、その音長を教師値では１に、同じく最も短く発音された音符を探し出し、その教師値を０とする。曲ごとに教師値を計算しているので、同じ音価、同じ音長であっても曲が異なれば教師値も異なる可能性がある。

- 音量の変化：演奏データより音符ごとの音量を取り出し、１曲の演奏の平均音量を求める。さらに音量の最大値を１に、最小値を０になるように正規化を行い、教師値が０から１の実数になるように設定する。平均音量と同じ音量で発音された場合は０．５となる。これも、同じ音量であっても曲によって教師値が異なる可能性がある。

3. 3. 2学習器へ入力する素性

ニューラルネットにおいて学習を行う際には、先に挙げた構造に関わるものと、楽譜上から読み取れる情報を入力とする。具体的には演奏表情を学習する対象の音符の次の情報を入力とした。

- 音価（前後３つの音符のも含め計７つ）
- 音高（前後３つの音符のも含め計７つ）
- 対象となる音符の音価と前後３つの音符の音価との差（計６つ）
- 対象となる音符の音高と前後３つの音符の音高の差（計６つ）
- 前３つの音符の音価の平均
- 後ろ３つの音符の音価の平均
- 最初を０、最後を１とした曲内での位置
- 楽曲の始まりに
- 楽曲の終わりに近いか

これらを一つの音符の情報とし、一つの教師値とセットとし、ニューラルネットへの入力とした。入力はここで述べた楽譜上の情報が３１種類と、構造に関するもの２１種類の計５２種類である。

3. 4演奏表情付与の方法

演奏表情の学習結果を利用して楽譜相当のデータに対して演奏表情の付与を行った。教師値を計算する際に演奏データから平均音量など、実際に演奏されたデータから取り出したため楽譜相当のデータにではそれらの情報が取得できない。そこで集められた演奏データから音価ごとの音長の最大値と最小値を求め、さらに音量に関しても最大値と最小値を求め、これらを０、もしくは１とするように正規化を行い、ニューラルネットの出力と対応させた。

ニューラルネットの出力は、３．３．２節挙げた情報を、学習を終えたニューラルネットへ入力することで０から１の出力を得る。

4.　機械学習の結果と試聴実験の結果

4. 1実験概要

実験は、ニューラルネットを用いて演奏表情の学習を行ったものと、さらにその結果を元に楽譜相当のデータへ演奏表情を付与し、試聴実験を行ったものである。

4. 2機械学習の結果

学習は、３１曲用意した曲のうち３０曲を学習に用い、残りの一曲をオープンデータのテスト用とした。

● 実際の演奏と出力の関係

図２図３は、オープンデータによる学習結果のテストを行った際の結果である。図はそれぞれ音長と音量に関するもので、左から右へ時系列になっており、音符一つ一つの値を示している。点線がテストデータの演奏を教師値に対応する値に変換したものであり、波線が学習を終えたニューラルネットにテストデータを入力した際のニューラルネットの出力である。

図２は音長に関してのテストの結果である。

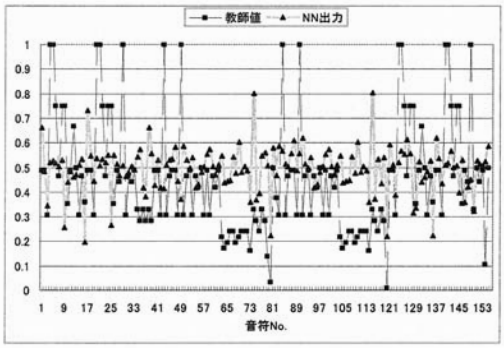

図２　演奏された際の音価と音長の比とそれに対応するニューラルネットの出力

図３は音量に関してのテストの結果である。

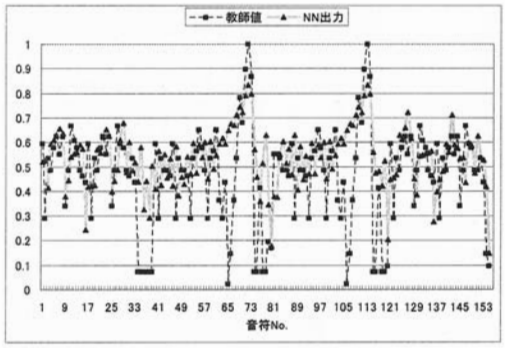

図３　演奏された際の音量とそれに対応するニューラルネットの出力

● 学習器への入力と出力との相関関係

学習の際の入力とニューラルネットの出力の相関関係を調べた。学習を終えたニューラルネットに学習データを入力することで出力が得られるが、この出力と入力の素性との相関係数はそれぞれ次の図のようになった。グラフは左の３１個のデータが音高や音価などの楽譜から抽出した情報であり、３２番以降の２１個のデータが構造に関するものである。

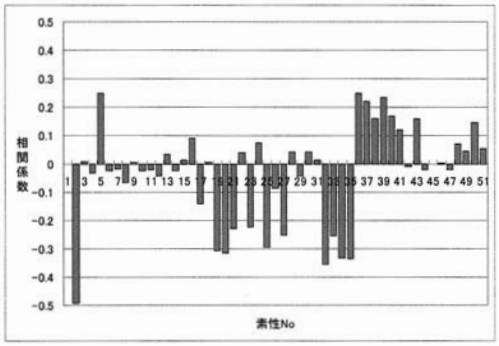

図４　ニューラルネットの出力と学習の際に入力した素性の相関関係（音長）

音長は音価と負の相関があり、その他は構造に関するものが多かった。音価に関しては、音価が長いほど短く演奏される可能性が高くなることを示すという単純なものである。

図５では音量に関しての相関関係を示す

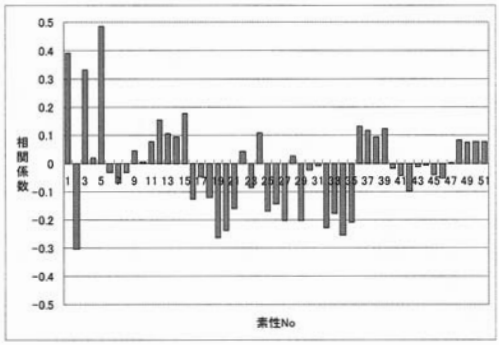

図５　ニューラルネットの出力と学習の際に入力した素性の相関関係（音量）

音量は音高に関連するものと正の相関があり、負の相関に関しては音価に関するものが多かった。

音長と音量について、正負ともに相関係数の大きかったものを表１に示す。

表１　入力した素性とニューラルネットの出力の相関係数（抜粋）

	音長に関する素性	相関係数		音量に関する素性	相関係数
↑正の相関が強い	構造化した際の音符の階層の深さ	０．３１	↑正の相関が強い	直後の音符の音高	０．４８
	直後の音符の音高	０．２４９		音高	０．３９
	直前の音符の構造上の距離	０．２４８		直前の音符の音高	０．３３
	直前の音符との構造上の距離の和	０．２３		構造化した際の音符の階層の深さ	０．２３
	（中略）			（中略）	
	直後の音符との構造上の距離の差	－０．３３０		直前の音価との差	－０．２３
	直後の音符との構造上の距離の和	－０．３３３		直後の音符との構造上の距離の差	－０．２５
	構造上の距離	－０．３５		直前の音価	－０．２６
↓負の相関が強い	音価	－０．４９	↓負の相関が強い	音価	－０．３０

４．３試聴実験

試聴実験用の WEB サイトを作成し試聴実験を行った。試聴実験は、オリジナルの演奏データ、学習結果から演奏表情を付与したデータ、意図的に演奏表情を削除したデータ（音価どおりの発音。常に一定の音量）の３つのデータを用意し、これらを聞き比べた上でこの中から一番自然だと思われる演奏と、一番気に入った演奏を選ばせ、最後に自由記述で感想を書かせた。アンケートを行った結果は次の表のようになった。

表２　試聴実験のアンケート結果（件数）

	オリジナル演奏	演奏表情付与	演奏表情削除
自然だと思った演奏	7	3	2
気に入った演奏	7	4	1

試聴実験の結果、オリジナルの人間による演奏が二つの項目において評価が高かった。

５．　考察

５．１学習結果について

５．１．１演奏表情の学習について

機械学習による演奏表情の学習は、誤差値の観点から見ると学習できているように見える。ただし音長と音量で精度に大きな差があり、概ね学習できているように見える音量に比べ、音長に関しては新たな素性を考えるなどする必要があると考えられる。教師値を音長の比率としていることや、EMG における解析の際に休符を扱わないことや、音価の値が四分音符で４８という非常に小さい値になり、より短い音符などで顕著であるが、教師値のつけ方にも問題点があったと考えられる。

５．１．２学習に用いた素性と学習器の出力

ニューラルネットへ入力した情報は、音高や音価といった基本的な情報に高い相関関係があったのに加え、構造に関するものも高い相関関係があった。音価に関する負の相関が高かった構造上大きな切れ目となる点は、休符を挟むことが多く、このことにより発音長が短くなったと考えられる。これは後続の音符との切れ目を見たものであるが、逆に正の相関が高いのは、前の音符との構造に関する距離であった。前の音符との間に構造上大きな切れ目があると、次の音符は長く発音されやすいということが言えるが、演奏の際に切れ目の後ろでは演奏を再開すると言う意味でアクセントがおかれて演奏されるものと考えられる。

さらに、音長と音量ともに次の音符の音高が高いときに音量や音長が増加しやすいという傾

向にあったが、これは次の音符を意識した演奏であると考えられる。

5.2試聴実験について

試聴実験のアンケートでは、楽曲の好みと自然さの点でオリジナルの演奏の評価が高かった。学習結果を元に演奏表情を付与したものは、学習結果が音高に左右されやすく、音高の上下と音量の増減が連動しがちであった。感想ではその点をしてつまらないというものがあった。逆に演奏表情のないデータについて、演奏が安定しており安心して聴けるという感想もあった。学習データを増やすことで学習の精度が上がる可能性は高いが、人による好みにまで対応できない。そういった点を考慮するために新しいシステムの考案が必要である。

6. まとめ

以上、楽曲の構造解析の結果を利用した演奏表情の解析と付与について述べた。機械学習を用いて演奏表情を学習するという点では、楽譜上から読み取れる情報と楽曲の構造解析をした結果から抽出した情報を利用することで学習を行うことができた。学習の際には音高や音価といった基本的な情報とともに構造解析の結果に高い相関関係があり、演奏表情に楽曲の構造情報が重要であることが示された。

さらに学習の結果を利用して楽譜相当のデータに演奏表情を付与した。試聴実験を行った結果、オリジナルの演奏には及ばないものの、演奏表情のない演奏に比べ楽曲の自然さや好みの点では優位であった。ある程度の演奏表情付けは行われていると考えられる。

参考文献

[1] Widmer G.: Discovering Strong Principles of Expressive Music Performance with the PLCG Rule Learning Strategy, *Proceedings of the 12th European Conference on Machine Learning (ECML'2001)*, Springer Verlag, Berlin, 2001.

[2] Widmer G.: Machine Discoveries: A Few Simple, Robust Local Expression Principles, *Journal of New Music Research*, 31(1), 37-50, 2002.

[3] 鈴木泰山, 徳永健伸, 田中穂積. 事例に基づく演奏表情の生成. 情報処理学会論文誌. Vol.41. No.4. pp.1134 -- 1145. 2000.

[4] 池田剛, 乾伸雄, 小谷善行: 言語クラスEMGを用いた不完全なシーケンスからの構造推定手法, 情報処理学会音楽情報科学研究会報告 SIGMUS52-18, 2003