音楽情報科学　39－18
（2001. 2. 23）

ニューラルネットワークを用いた複数楽器の音源同定処理

村瀬 樹太郎　　中西 正和

慶應義塾大学大学院 理工学研究科 計算機科学専攻

juta@nak.ics.keio.ac.jp

あらまし　音源同定とは、認識された音が何の楽器の音なのかを判別することであり、自動採譜を行うための重要な処理の一つである。本研究では、複数の楽音に対して、誤差逆伝搬法によって学習したニューラルネットワークを用いて音源同定を行った。音の性質は、基本周波数成分とその倍音成分の関係である調波構造に現れるため、これを特徴量として利用した。実験結果から、調波構造が音源同定に重要な特徴量であることを確認し、ニューラルネットワークが音源同定に有効であることが示せた。また、整数倍の倍音成分のみを特徴量として用いているため、ネットワークとしてシンプルなもので実現できた。

キーワード　音楽情報処理、音源同定、ニューラルネットワーク、 調波構造

Sound Source Identification Process of Polyphony using Neural Network

Jutaro Murase　Masakazu Nakanishi

Graduate School of Science and Technology, Keio University

juta@nak.ics.keio.ac.jp

Abstract　The sound source identification is to distinct what instrument the recognized sound is played by. It is one of the most important process in automatic transcription. In this paper, we have approached the sound source identification of polyphony using the neural network learned by back propagation. As for the property of sound, since it appears in harmonic structure, a harmonic relationship of the fundamental frequency component, we have used it as the feature value. As a result, we have confirmed that the harmonic structure is an important feature value, and the neural network was found to be effective for sound source identification. Moreover, since we have used only the harmonics as the feature values, a simple network was realized.

key words　Music Information Processing, Sound Source Identification, Neural Network, Harmonic Structure

1 はじめに

音楽を聴く上で、音の重なりを味わうことは最も重要な要素の一つである。ピアノのように一つの楽器で高音と低音の重なりを楽しむこともあれば、オーケストラのように様々な楽器の音色の重なりを楽しむこともある。そこで我々人間は、曲を聴けば主旋律と伴奏を聴き分けることができ、また、重なっている音を聴き分け、音程や楽器の認識が可能である。音楽を聴いてそれを楽譜として表すことは、音楽的知識や聴音能力が必要で、一般的な人にとって簡単なことではない。そこで、計算機を用いて採譜を自動的に行うことができれば、特別な能力のない人でも採譜ができるようになる。その中で、音が重なっている場合に、重なっている音を聴き分けたり、何の楽器であるか認識することが必要となってくる。

計算機を用いて採譜を自動的に行うことを自動採譜といい、音楽情報処理の代表的な研究分野である。音源同定とは、認識された音が何の楽器の音なのかを判別することであり、自動採譜を行うための重要な処理の一つである。

音の波形は、振幅と位相が時間的に緩やかに変化する正弦波の和で構成されている。音の知覚において重要な特徴は、主として振幅情報に含まれている。以上の理由により音の性質を調べるときは、周波数スペクトルなどのスペクトルに関連した特徴に変換して扱うことが多い。

また、音源分離および音源同定を行う際に使用できる音の特徴量としては、調波構造 (基本周波数とその倍音から構成される音)、音の立ち上り、音の立ち下がり、AM (Amplitude Modulation)、FM (Frequency Modulation)、音色、音源方向などが考えられる。

音は倍音成分からなるため、調波構造に注目することは有効であり、様々な研究が行われている [1][2][3]。調波構造の情報だけでは不十分なため、音の立ち上りや音の立ち下がりなどの情報を統合する場合が多い [4][5][6][7][8]。

木下らは、文献 [9] で、周波数成分が重なったときの特徴に合わせて特徴量を分類し、それに応じて重なりのある周波数成分の特徴量を適応的に変化させ、音源同定処理を行っている。この手法では、知識ベースで特徴量をテンプレートとして持っているため、テンプレートにない特徴量が抽出された場合の処理が困難になる。また、テンプレートマッチングではテンプレートが多ければ多いほど、情報量が増え、計算量も膨大になってしまうという問題点がある。

文献 [10] では、各音に対応したサンプリング周波数を使用した、くし形フィルタを用いたピッチ推定法が提案されている。異種楽器和音は、入力和音に対してくし形フィルタを縦続接続することによって、単一楽器成分に分離することができる。この分離波形の形状から楽器を推定する。このとき、楽器音の時間波形 (減衰、振動など) の特徴は保持され、その比較には音量の時間変化 (エンベロープ) を用いている。問題点として、採譜対象音域が狭いこと、同属楽器の推定ができないことなどがあげられる。

音声認識の分野において，視覚情報と聴覚情報を統合することによって、聴覚情報の認識率が向上することも報告されている [11]。複数のマイクロホンを用いて集音し、音源の位置情報、時間差情報や強度差情報を用いる手法がある [1][12][13]。人間は、モノラル音響信号よりも、ステレオ音響信号の方がそれぞれの楽器音を個別に認識しやすいこと (両耳受聴効果) を利用している。文献 [14] においても聴覚情報だけでなく、視覚情報を用いた音源分離方法も提案され、実験結果から正確な音源方向を与えることによって認識率が上がり、視覚と音響の両方の情報を融合することの有効性が報告されている。

ニューラルネットワークは、クラスタリングを得意とする。音楽情報処理の分野においても、ニューラルネットワークの特性をいかせないかと、研究に用いられ始めている [15]。自己組織化マップ (Self-Organizing Map; SOM) は、競合作用に基づいた学習方式のニューラルネットワークである [16]。入力に他次元から 2 次元への非線形射影を施す。その際に入力データのもとの空間における相互関係を保持したままマッピングを行うため、データの分布状態を素直に反映したマップを作ることができる。文献 [17] では、単音を SOM を用いて、聴覚モデルと組み合わせて、クラスタリングを行っている。ここでは、12 種の楽器を音色の違いで単音をマップ上に分類している。文献 [18] では、ピアノの和音の分解を SOM を用いて行っている。SOM は教師信号無しで学習できることと、クラスタリングを行った結果が視覚化されることが特徴であるが、最終的な判断は人為的に行われる。また、複数楽器を同時に鳴らした場合に、ニューラルネットワークを用いて音源分離の研究報告例はない。

2 本研究の手法

2.1 調波構造の特徴

音源同定を行う際、最も重要な情報は調波構造にある。これは、楽器によってその音に含まれる倍音成分の割合が異なるという性質に基づく。一つの楽器におけるいくつかの音の整数倍の倍音成分の割合を以下の図に示す。図 1 は piano、図 2 は saxophone、図 3 は flute の結果である。音響データは生演奏を収録した CD(Compact Disc) から取り出したものである。横軸は倍音数で、1 が基本周波数にあたる。縦軸はパワーを正規化したものである。

同じ時刻の成分を線で結んである。図から楽器によってその割合が異なることがわかる。

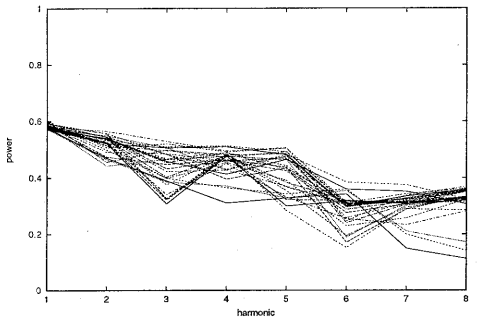

図 1: piano の倍音成分

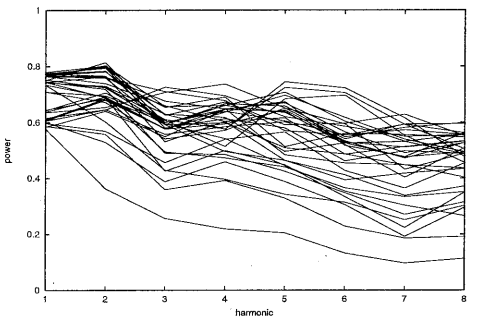

図 2: saxophone の倍音成分

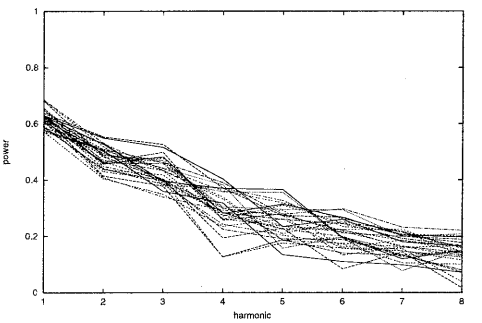

図 3: flute の倍音成分

複数音が同時に鳴った場合においては、単音ごとに分離することによって、一つ一つの音の倍音成分の情報が取り出せると推測される。そこで、取り出した単音ごとの倍音成分の割合を比べることにより音源同定ができると考えられる。

本研究では、単音成分に分離し、分離後の情報を用いて、ニューラルネットワークによる音源同定の手法を提案する。

2.2 本研究のシステムの流れ

本システムは、周波数解析モジュールとクラスタリングモジュールの二つのモジュールで構成される。

音響データは、DAT(Digital Audio Tape)-Link [19] を用いてデジタル化を行ったものを使用する。

本システムの処理の流れを図 4 に示す。

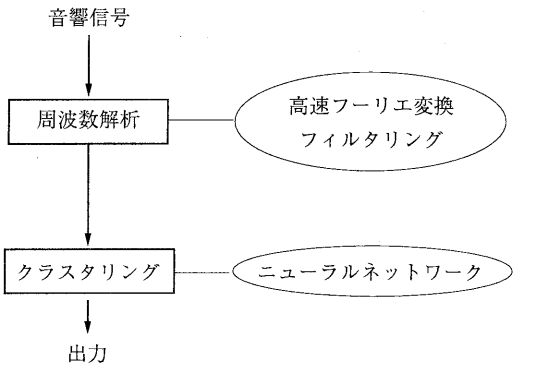

図 4: システムの流れ

2.2.1 周波数解析モジュール

デジタル化したデータをハミング窓で短区間に分割し、高速フーリエ変換を行う。その結果である短区間パワーの極大値から基本周波数を抽出する。図 5 に、3 和音に対して倍音成分を取り除いて基本周波数のみを抽出した結果を示す。横軸は時間を表し、縦軸は周波数を表す。

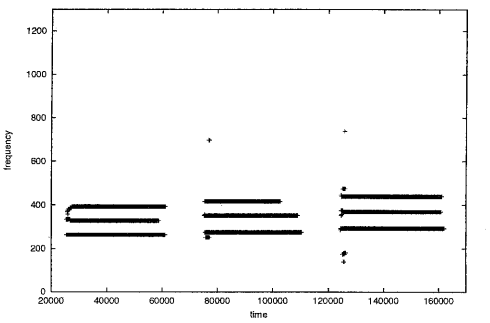

図 5: 周波数解析の結果

その抽出した基本周波数は平均律表[20]を用いてそれぞれの音名に割り当てる。

複数の基本周波数が抽出された場合は，それぞれの基本周波数成分のみにフィルタリングを行って単音ごとに分離をする。これは、先に求めた基本周波数からその倍音成分のみを抽出することによって実現する。図6に、2音を同時に鳴らした場合のフィルタリングの結果を示す。横軸は周波数を表し、楯軸はパワーを表す。

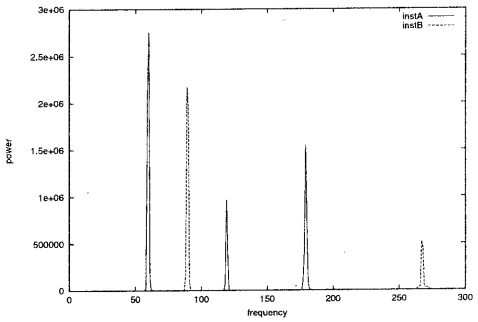

図 6: フィルタリングの結果

ニューラルネットワークの入力とするために、単音ごとに得られたその倍音成分のパワーは、パワーの上限値 $C_{\max}$ および下限値 $C_{\min}$ を用いて、以下の式(1)で正規化される。

$$x_i' = \frac{\log_{10} x_i - \log_{10} C_{\min}}{\log_{10} C_{\max} - \log_{10} C_{\min}} \qquad (1)$$

x_i は i 番目の倍音パワー値を表し、x_i' は正規化後のパワー値となる。$i = 1$ の場合は基本周波数にあたる。

2.2.2 クラスタリングモジュール

クラスタリングは調波構造に注目して行う。3層構造のニューラルネットワークで誤差逆伝搬法による学習をする。これに単音の調波構造情報を学習させ、複数楽音に関する調波構造をテストデータとし音源同定を行う。データとして、周波数解析によって単音ごとに分離した結果であるそれぞれの倍音成分のパワー値を用いる。

ニューラルネットワークの入力として、式(1)で求めた正規化後のパワー値を与える。出力として出た値の中で最も値が大きいものをそのデータの楽器とする。

3 実験および評価

3.1 実験条件

DAT-Link を用いて、音響データを標本化、量子化をしてデジタル化し、短区間に分割して分析を行った。実験環境を表1に示す。

表 1: 実験環境

標本化周波数	24,000 Hz
量子化	16 bit
周波数解析手法	高速フーリエ変換
窓関数	ハミング窓
フレーム長	4,096 point (170.7 msec)
フレーム周期	256 point (10.7 msec)
周波数分解能	11.7 Hz

短区間で得られた8倍音までの正規化したパワー値を一つのデータとし、ニューラルネットワークの入力として与えた。式(1)において、正規化する際のパワーの上限値 $C_{\max}$ を 100,000,000、下限値 $C_{\min}$ を 0.01 とした。出力ユニット数は学習させた音源の数で、学習終了条件は誤差が 0.0001 未満とした。ニューラルネットワークの仕様を表2に示す。

表 2: ニューラルネットワークの仕様

入力ユニット数	8
隠れユニット数	60
出力ユニット数	2 または 3
学習効率 η	0.01
慣性項 α	0.9

データとして、YAMAHA の Music Synthesizer S80 から音響信号を採取した。シンセサイザーの仕様を表3に示す。

表 3: S80 の仕様

キーボード鍵盤数	88
音源方式	AWM2、モジュラーシンセシス プラグイン システム
Voice Wave ROM	24 MByte

3.2 予備実験

音響データは生音から取ることが理想である。予備実験として、生演奏を収録した CD から採取したデータでの実験を行った結果、75 % の認識率を得た。

しかし、生音や生演奏を収録した CD などからデータを採取することは困難であり、また、実験としての幅広いデータを取ることができない。よって、実験環境として前節に述べたようにシンセサイザーから音響データを採取した。

3.3 ニューラルネットワークに用いたデータ

3.3.1 学習用データ

学習用データとして 四つの楽器を用いた。データを表4に示す。音数は鳴らした音の数を表し、総データ数はデータとして用いた短区間の数を表す。

表 4: 学習データ

楽器名	音数	総データ数
piano	6	63
clarinet	8	58
violin	8	66
trumpet	8	52

また、学習データとして使用した楽器のそれぞれの倍音成分の割合を以下に示す。データはシンセサイザーから採取した。図7、図8、図9、図10は、それぞれ piano, clarinet, violin, trumpet を表す。音の鳴り始めから鳴り終わりまでのどの部分の短区間情報かは限定しない。同じ短区間のものを線で結んである。

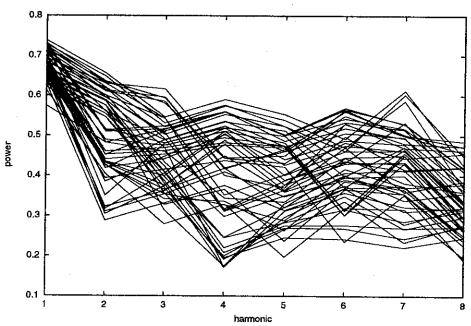

図 7: piano の倍音成分

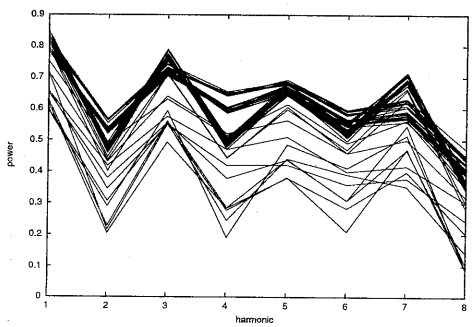

図 8: clarinet の倍音成分

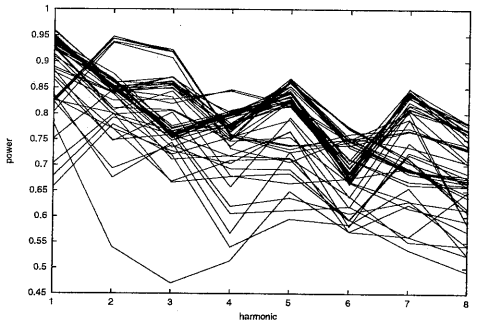

図 9: violin の倍音成分

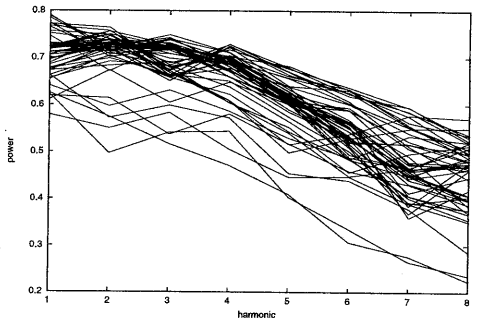

図 10: trumpet の倍音成分

3.3.2 テスト用データ

単音のみのデータでは、学習データとテストデータは違う音の高さのものを用いた。表5にテストデータを示す。

音程とは、二つの音の高低差を意味し、度数で表現する。複数の音程を縦に重ねたものを和音、コードという。和音の一番低い音、元になる音のことをルート (root) と呼び、それから順に上に 3 度、5 度と呼ぶ。長三和音は、ルートから長 3 度 (Major 3rd) と完全 5 度 (Perfect 5th) でできている [21][22]。

表 5: テストデータ

楽器名	音数	総データ数	音程
piano	5	43	-
clarinet	5	47	-
violin	5	66	-
trumpet	5	49	-
piano + clarinet	11	502	完全 4 度
piano + trumpet	6	254	完全 4 度
piano + violin	6	138	完全 5 度
clarinet + trumpet	6	170	完全 4 度
piano + piano + clarinet	4	141	長三和音
piano + trumpet + trumpet	5	171	長三和音
piano + piano + violin	4	162	長三和音
piano + clarinet + violin	4	189	長三和音

3.4 実験結果

2 楽器を学習させた場合と 3 楽器を学習させた場合の結果を学習データごとにそれぞれ表 6 と表 7 に示す。学習させた楽器の組み合せを太字で表す。

3.4.1 2 楽器を学習させた場合

表 6: 認識率 (2 instruments)

楽器名	正解数	総データ数	認識率 [%]
piano + clarinet			
piano	37	43	86.0
clarinet	47	47	100.0
piano + clarinet	488	502	97.2
piano + piano + clarinet	141	141	100.0
piano + trumpet			
piano	27	43	62.8
trumpet	39	49	79.6
piano + trumpet	247	254	97.2
piano + trumpet + trumpet	131	171	76.6
piano + violin			
piano	43	43	100.0
violin	56	66	84.8
piano + violin	93	138	67.4
piano + piano + violin	127	162	78.4
clarinet + trumpet			
clarinet	47	47	100.0
trumpet	43	49	87.8
clarinet + trumpet	149	170	87.6

3.4.2 3 楽器を学習させた場合

表 7: 認識率 (3 instruments)

楽器名	正解数	総データ数	認識率 [%]
piano + clarinet + trumpet			
piano + clarinet	401	502	79.9
piano + trumpet	241	254	94.9
clarinet + trumpet	104	170	61.2
piano + piano + clarinet	122	141	86.5
piano + trumpet + trumpet	99	171	57.9
piano + clarinet + violin			
piano	43	43	100.0
clarinet	47	47	100.0
violin	63	66	95.5
piano + clarinet	431	502	85.9
piano + violin	94	138	68.1
piano + piano + clarinet	97	141	68.8
piano + piano + violin	123	162	75.9
piano + clarinet + violin	173	189	91.5

3.4.3 認識率一覧

以上の結果をまとめ、学習データごとの認識率を表 8 に、テストデータごとの認識率を表 9 に、同時に鳴らした音数ごとの認識率を表 10 に示す。

表 8: 学習データごとの認識率

楽器名	正解数	総データ数	認識率
piano + clarinet	713	733	97.2
piano + trumpet	444	517	85.9
piano + violin	319	409	76.1
clarinet + trumpet	239	366	89.8
piano + clarinet + trumpet	967	1238	78.1
piano + clarinet + violin	1071	1288	83.2

表 9: テストデータごとの認識率

楽器名	正解数	総データ数	認識率
piano	150	172	87.2
clarinet	141	141	100.0
violin	119	132	90.2
trumpet	82	98	83.7
piano + clarinet	1320	1506	87.6
piano + trumpet	488	508	96.1
piano + violin	187	276	67.8
clarinet + trumpet	253	340	74.4
piano + piano + clarinet	360	423	85.1
piano + trumpet + trumpet	240	342	70.2
piano + piano + violin	250	324	77.2
piano + clarinet + violin	173	189	91.5

表 10: 音数ごとの認識率

同時に重なった音の数	正解数	総データ数	認識率
1 音	492	543	90.6
2 音	2248	2630	85.5
3 音	1023	1278	80.0

4 考察

4.1 認識率

表 10 より、8 割以上の認識率を得ることができた。この値は先行研究 [9] と同等以上である。

高速フーリエ変換を行う際、音の波形は短区間では定常であるとして、短区間に区切り分析を行った。しかし、音は定常ではなく、時間と共に変化するため、同じ音でも短区間によって情報のばらつきが生じる。このばらつきが、ニューラルネットワークにおいて認識率を下げる原因となる。特にパワーは、音の鳴り始めから鳴り終わりまで常に変化するので、ばらつきが多く生じると考えられる。

学習させるものや楽器の組み合せによって、認識率が異なることが表 8 や表 9 からわかる。

また、ニューラルネットワークの入力として、周波数解析の結果を用いているため、その周波数解析の精度が認識率に影響を与える原因の一つだと考えられる。

4.2 倍音成分の重なり

二つの音があるときの音程と周波数比の関係を表 11 に示す。音程は音程名の都合上 C major scale を基にしたものとする。また、piano と clarinet の 2 音を同時にならしたとき、音程の違いにおける認識率を併記する。

周波数比は、音どうしがどこで重なるかを表す。例えば二つの音が完全 5 度の関係にある場合、周波数比は 2 : 3 で、低い音の 3 倍音が高い音の 2 倍音と重なる。本システムでは 8 倍音までの情報をニューラルネットワークに入力として与えているが、表から、ある 1 オクターブに

おいて異なる 2 音を同時に鳴らした場合、約半数は 8 倍音までに重なることがわかる。

表 11: 音程と周波数比の関係

音程	比率	データ数	認識率 [%]
完全 1 度	1:1	-	-
短 2 度	15:16	30	100
長 2 度	8:9	38	92
短 3 度	5:6	36	97
長 3 度	4:5	16	100
完全 4 度	3:4	8	100
増 4 度	5:7	36	100
完全 5 度	2:3	32	84
短 6 度	5:8	24	100
長 6 度	3:5	18	100
短 7 度	5:9	28	96
長 7 度	8:15	22	100
完全 8 度	1:2	-	-

倍音成分の重なり方は大きく分けて 2 通りある。

- あるの単音の n 倍音成分が、別の単音の基本周波成分と重なる場合
- あるの単音の m 倍音成分が、別の単音の n 倍音成分と重なる場合

基本周波成分のパワーは、通常その音の倍音成分の値の中で一番大きい。そのため、あるの単音の n 倍音成分が別の単音の基本周波成分と重なった場合、両方のパワー値が加算され、その n 倍音成分は基本周波成分の影響を大きく受け、認識が困難であると考えられる。本研究ではこの場合の実験は行っていない。

基本周波ではない倍音どうしが重なる場合、表 11 から完全 5 度の関係にあるときが認識が困難であることがわかる。また、実験結果において、完全 5 度やそれを含む長三和音の認識率が他と比べ低いものとなった。

本研究では、正規化の際に常用対数を用いた。これによって、二つの値が重なったときに、小さい方の値だったものが大きい方の値に吸収され、大きい方の値としてでてくる。したがって、音源の組み合せ方により認識率が若干異なることになる。

4.3 従来の手法との比較

従来手法では特徴量などを決定する際に人為的な閾値を加えているものが主であった。また、テンプレートマッチングは情報量を多く必要としてしまい、計算機への負荷が問題であった。

本研究では、調波構造のうち整数倍の倍音成分のみに注目したことによって、少ない情報量で音源同定が可能であることを示した。また、人為的な閾値をともなわないシンプルなニューラルネットワークによって実現することができた。

音源同定において、学習していない音の高さのものに関して認識できることから、ニューラルネットワークが未学習のデータに対しても，学習したデータを基に判別できるという汎用性が見られた。また、ニューラルネットワークのクラスタリング能力が高いため、短区間情報のばらつき、つまり、入力のばらつきがあるにもかかわらず認識ができた。

5 結論および今後の課題

5.1 結論

本研究では、調波構造に注目し、ニューラルネットワークを用いて複数楽音の音源同定を図った。実験結果から、調波構造のうち整数倍の倍音成分が音源同定に重要な特徴量であり、少ない情報量でも音源同定が可能なことを確認できた。また、誤差逆伝搬法による学習を行ったニューラルネットワークが音源同定に有効であることが示せた。情報量が少ないため、シンプルなネットワークで実現できたといえる。

5.2 今後の課題

認識率の向上のためには、倍音成分のみの調波構造だけでは不十分で、別の特徴量と組み合わせることが必要である。例えば、パワーの時間変化などが考えられる。ニューラルネットワークとして時系列を扱うことができるリカレントニューラルネットワークなどがあげられる。また、2 つの単音が同じ音程 (完全 1 度) やオクターブの関係にあるときの分離を可能にしなくてはいけない。

参考文献

[1] T. Nakatani, H. G. Okuno, and T. Kawabata: Auditory Stream Segregation in Auditory Scene Analysis with a Multi-Agnet System, In Proc. of 12th International Conference on Artifical Intelligence(AAAI-94), pp.100–107, 1994.

[2] 奥乃 博, 中谷智広: マルチエージェントシステムによる音響ストリーム分離, システム情報制御学会, Vol.41, No.8, pp.309–315, 1997.

[3] 後藤真孝: 音楽音響信号を対象としたメロディーとベースの音高推定, 電子情報通信学会論文誌, Vol.J84–D–II, No.1, pp.12–22, 2001.

[4] 小野徹太郎, 斎藤英雄, 小沢慎治: 自動採譜のための遺伝的アルゴリズムによる混合音推定, 計測自動制御学会論文集, Vol.33, No.5, pp.417–423, 1997.

[5] G. J. Brown, and M Cooke: Perceptual Grouping of Musical Sounds: A Computational Model, Journal of New Music Research, Vol.23, pp.107–132, 1994.

[6] 柏野邦夫, 木下智義, 中臺一博, 田中英彦: 音響情景分析の処理モデル OPTIMA における単音の認識, 電子情報通信学会論文誌, Vol.J79-D-II, No.11, pp.1751-1761, 1996.

[7] 柏野邦夫, 木下智義, 中臺一博, 田中英彦: 音響情景分析の処理モデル OPTIMA における和音の認識, 電子情報通信学会論文誌, Vol.J79-D-II, No.11, pp.1762-1770, 1996.

[8] 柏野邦夫, 村瀬 洋: アンサンブル実演奏の自動アンミキサ, 情報処理学会音楽情報科学研究会報告, 98-MUS-24-5, pp.33-40, 1998.

[9] 木下智義, 坂井修一, 田中英彦: 周波数成分の重なりに適応処理を用いた複数楽器の音源同定処理, 電子情報通信学会論文誌, Vol.J83-D-II, No.4, pp.1073-1081, 2000.

[10] 三輪多恵子, 田所嘉昭, 斎藤 努: 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, Vol.J81-D-II, No.9, pp.1965-1974, 1998.

[11] G. J. Wolff: Sensory Fusion: Integrating Visual and Auditory Information for Recognizing Speech, In Proc. of IEEE International Conference on Neural Networks, Vol.2, pp.672-677, 1993.

[12] J. Blauert: Spatial hearing, MIT Press, 1983.

[13] 三輪明宏, 守田 了: 能動環境におけるステレオ音楽音響信号を用いた 3 重奏に対する音源分離, 電子情報通信学会論文誌, Vol.J84-D-II, No.1, pp.23-30, 2001.

[14] Y. Nakagawa, H. G. Okuno, and H. Kitano: Using Vision to Improve Sound Source Separation, Proc. of IJCAI-99 Workshop on Computational Auditory Scene Analysis, pp.99-107, 1999.

[15] P. M. Todd, and D. G. Loy: Music and Connectionism, The MIT Press, 1991.

[16] T. Kohonen: Self-Organization Maps, Springer-Verlag, 1995.

[17] P. Cosi, G. D. Poli, and G. Lauzzana: Auditory Modelling and Self-Organizing Neural Netwoks for Timbre Classification, Journal of New Music Research, Vol.23, pp.71-98, 1994.

[18] F. Carreras, M. Leman, and M. Lesaffre: Automatic Harmonic Description of Musical Signals using Schema-based Chord Decomposition, Journal of New Music Research, Vol.28, No.4, pp.310-333, 1999.

[19] DAT-Link, http://www.tc.com/

[20] T. Kientzle: A Programmer's GUIDE TO SOUND, Addison Wesley Developers press, 1997.

[21] 井桁 学: ギタリストのための学典, リットーミュージック, 1995.

[22] Music Theory for Contemporary Music, http://www.kittysound.com/