音楽情報科学 44－7
（2002．2．16）

Neural Networkを用いた音源学習と音源同定

佃　卓磨　　甲藤　二郎

早稲田大学大学院理工学研究科

{tsukuda,katto}@katto.comm.waseda.ac.jp

あらまし　学習波形データをwavelet変換しその周波数分布を誤差逆伝搬で学習させる効率のよい方法を検討する．出力層と目標値を勝者ユニット方式で音源識別できるように設定し，結合係数を目標値に合わせて制御するリセットアルゴリズムを提案する．その結果，1回の誤差逆伝搬の計算量は増えるが全体の学習効率は向上することが示せた．また単音学習させたものを和音に適用させ，認識率と処理速度が向上する方法を検討する．2和音から1音をNeural Networkで識別し，仮音源分離を行い残りの1音を識別するアルゴリズムを提案する．その結果，音源同定の認識率は向上し，処理速度も向上することが示せた．

キーワード　音源同定，自動採譜，音源分離，ニューラルネットワーク

Sound Source Learning and Identification Using Neural Network

Takuma Tsukuda　　Jiro Katto

Graduate School of Science and Engineering, Waseda University

{tsukuda,katto}@katto.comm.waseda.ac.jp

Abstract　In this paper, we investigate an effective learning way for back propagation using frequency spectrums which are transformed from PCM data with wavelet transform. We suggest a winner unit algorithm that output layers and target values are set so as to identify a sound source, and a reset algorithm that coupling coefficient is controlled according to the target values. As a result, we have confirmed that the total learning efficiency is improved though the calculated amount for each back propagation step increases slightly. We also investigate a way that improves the recognition rate and the processing speed by applying Neural Network learned with single sound waves to chordal sound waves. We suggest an algorithm that Neural Network identifies a single sound wave from mixed two sound waves, and after one wave is separated from the mixed wave, the other wave is identified by the same Neural Network. As a result, we have confirmed that the recognition rate of the sound source identification and the processing speed are improved.

Key words　sound source identification, automatic music transcription, sound source separation, neural network

1. はじめに

近代技術発展の背景の一つに波動に関する研究成果が挙げられる．地震，脳波，通信などその応用分野は幅広い．そしてその波形のほとんどは複数の波形が混合されたものである．また日常に耳にする音声，音響ストリームも同様である．これらの波形は短時間または非定常状態の場合があり，その解析は一筋縄でいかない場合が多い．

音響ストリームに対する研究は盛んに行われている[1)~4)]が，Neural Network を用いた研究は数が少ない．文献[5)]ではパワースペクトルの主成分分析を行い，スペクトル数 128 を累積寄与率 90%で 35 に削減して Neural Network で学習させ音源同定を行っている．また文献[6)]では周波数成分の調波構造を特徴量とし，整数倍の倍音成分だけを用いて学習を行い音源同定を行っている．

本研究では周波数軸方向と時間軸方向の特徴が得られる wavelet 変換を用いたスペクトルを Neural Network で学習させる．しかし入力数が多くなると学習しにくくなるので，学習効率を上げるために Reset アルゴリズムと勝者ユニットアルゴリズムを提案する．また学習完了後は 2 和音から 1 音を Neural Network で識別し，その後重み係数を求めて仮音源分離を行い残りの 1 音を識別する音源同定アルゴリズムを提案する．

2. 提案手法

2.1 音源学習

2.1.1 Neural Network

誤差逆伝搬の Neural Network の構成は入力層 L_{in}，隠れ層 L_{hid}，出力層 L_{out} とし，教師信号の目標値を T とする．それぞれのユニットの数を N_{in}, N_{hid}, N_{out} とする．入力層数 N_{in} は PCM 波形を wavelet 変換した周波数スペクトルの数である．

また Neural Netowrk の終了条件は，誤差逆伝搬の目標値 T と出力層 L_{out} の 2 乗平均誤差 err が誤差閾値 TH_{err} を下回ったときとする．また極小解に陥った場合を考えて逆伝搬の回数 N_{BP} は N_{BPth} を限度とする．

2.1.2 音源学習従来手法の問題点

文献[8)]の Neural Network の学習方法について簡単に述べる．出力層数 N_{out} は学習させるテンプレートの数を N_t としたとき

$$2^{N_{out}-1} < N_t \leq 2^{N_{out}} \qquad (1)$$

をみたす N_{out} とした．また目標値の設定は 0 か 1 を割り当て，学習テンプレート k 番目の目標値は交番 2 進法（グレイ符号）の k 番目とした．

学習終了後の識別方法は出力層の値をある閾値 TH_{out} を境に 0 か 1 に離散化し，グレイ符号の割り当てをもとに音源の識別を行っていた．しかし学習させる際に誤差が 0 に収束せず極小解に陥りやすかった．図 **1** がその例である．横軸を学習回数 N_{BP}，縦軸を誤差 err とした．学習回数を重ねることにより誤差は減少傾向にあるが，誤差が 1.0 付近で収束してしまっている場合が多かった．

2.1.3 Reset アルゴリズム

極小解から抜け出すために，また極小解に陥りにくくするために Reset アルゴリズムを提案する．これは図 1 のように誤差 err が 1.0 付近に収束してしまうことに注目した．あるテンプレート $\mathbf{A}_m$ のある出力層 $L_{out}(k)$ の誤差 $err_{A(k)}$ が 1.0 に近い値で，他は全て 0 に近い値であった．つまり出力層の一ケ所だけが完全に誤って学習をしてしまい，他はきちんと学習をしていたということである．この一ケ所の誤りを訂正することができれば極小解から抜け出すことが可能であると考えた．

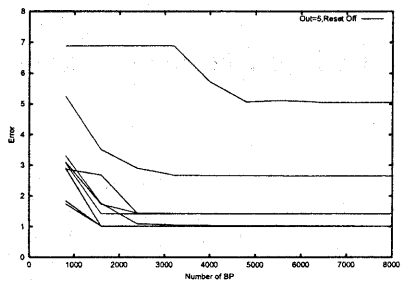

図 1 従来手法 Reset Off $N_{out} = 5$

ある出力層 $L_{out}(k)$ の値は図 **2** に示すような太線の結合係数 w_{ij}, w_{jk} を経由して値が決定される．ここで w_{ij} は入力層から隠れ層の結合係数，w_{jk} は隠れ層から出力層の結合係数である．

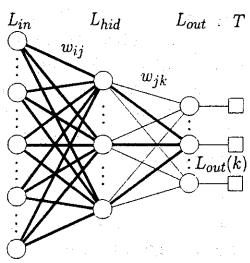

図 2 リセットアルゴリズム

あるテンプレート $\mathbf{A}_m$ のときの出力層 $L_{out}(k)$ が誤って学習を続けていると判断されたとき，隠れ層から出力層の結合係数 w_{jk} を全て 0 に Reset する．その他の結合係数は変化させない．w_{jk} が Reset された直後は $L_{out}(k)$ に入力される値は 0 であり，sigmoid 関数を通過することにより 0.5 と出力される．こうすることにより誤差が 1.0 に近かったものが 0.5 となり極小解から抜けだせるようになる．つまり Reset アルゴリズムは

$$\begin{cases} \text{if} & |T_K - L_{out}(K)| > 1.0 - \varepsilon \\ \text{then} & w_{jK} = 0 \end{cases} \tag{2}$$

である．ここで極小正数 $\varepsilon \approx 0$，T_K は K 番目の目標値の値である．

2.1.4 勝者ユニットアルゴリズム

式 (1) による出力層の数 N_{out} はテンプレートの数 N_t がちょうど 2 のべき乗であればいいが，実際はそうでない場合が多いので N_{out}[bit] 全てが割り当てられない．そのため Neural Network で識別を行おうとすると，目標値として割り当てられていない識別結果を出す場合がある．また出力層が閾値 TH_{out} で識別されるため，この閾値近辺の値で識別結果が異なる場合がある．これらの問題点を解決するために勝者ユニットアルゴリズムを提案する．出力層の数 N_{out} とテンプレートの数 N_t を等しくし，k 番目のテンプレート $\mathbf{A}_k$ のときの目標値 T は，k 番目を 1 とし他は 0 とする．つまり

$$\begin{cases} T_k = 1 \\ T_j = 0 \quad (j \neq k \quad 0 \leq j < N_{out}) \end{cases} \tag{3}$$

として学習を行う．

次に識別時について述べる．従来手法では閾値 TH_{out} により 0 か 1 に離散化したが，提案手法では出力層 L_{out} の中で最大の値をもつユニット $L_{out}(m)$ を勝者ユニットとしテンプレート $\mathbf{A}_m$ と判断する．従来の閾値処理では同時に 1 と出力するユニットが複数存在する場合や全く存在しない場合があり正しく識別出来ないことがあった．

2.2 音源同定

2.2.1 音源重み係数計算

ある 2 つの和音周波数分布 $\mathbf{M}$ が a, b を重み係数として

$$\mathbf{M} = a\mathbf{A} + b\mathbf{B} \tag{4}$$

で表されるとする．ここで $\mathbf{A}$,$\mathbf{B}$ は既知周波数分布のうちのいずれかであり、a, b は未知とする。

また内積値が $|\mathbf{A}|^2 = |\mathbf{B}|^2 = |\mathbf{M}|^2 = 1$ となるように正規化する．

ここで $\mathbf{A}$,$\mathbf{B}$ と $\mathbf{M}$ の内積を計算し，内積値を $\mathbf{M}\cdot\mathbf{A} = P_{MA}$ と表し行列表示すると

$$\begin{pmatrix} P_{MA} \\ P_{MB} \end{pmatrix} = \begin{pmatrix} P_{AA} & P_{BA} \\ P_{AB} & P_{BB} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \tag{5}$$

である．逆行列を用いると

$$\begin{pmatrix} a \\ b \end{pmatrix} = C \begin{pmatrix} 1 & -P_{AB} \\ -P_{AB} & 1 \end{pmatrix} \begin{pmatrix} P_{MA} \\ P_{MB} \end{pmatrix}$$

$$C = \frac{1}{1 - {P_{AB}}^2} \tag{6}$$

となり、重み係数 a, b を求めることができる．

2.2.2 仮音源分離

式 (6) により求めた a, b を用いて仮音源分離を行う．あるテンプレート $\mathbf{A}_i$,$\mathbf{B}_j$ に対して求めた a_i, b_j から

$$\mathbf{D}_i = \mathbf{M} - a_i\mathbf{A}_i \tag{7}$$

を計算する．もしスペクトル成分が負になれば 0 とする．

そして $\mathbf{A}_i$ が減算された $\mathbf{D}_i$ を内積正規化し

$$\mathbf{D}_i \approx \mathbf{B}_j \tag{8}$$

が成り立てば入力音源 $\mathbf{M}$ には $\mathbf{A}_i$ と $\mathbf{B}_j$ が含まれていることがわかり音源同定可能である．

2.2.3 音源同定従来手法の問題点

文献[7] の音源同定方法について簡単に述べる．図 3 にあるように，音源候補は総当たりで内積の計算を行い，その中から内積値の高いものをいくつか選んでいた．しかし，内積値は僅差でありどの程度絞り込むかの判断が難しかった．また従来手法での仮音源分離は，時間領域で相互相関を用いて位相を合わせながら減算を行っていたので計算時間がかかっていた．

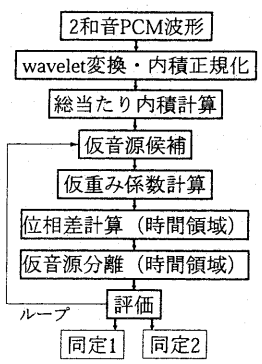

図 3 音源同定フローチャート（従来手法[7]）

2.2.4 音源同定提案手法

図 4 に音源同定提案手法のフローチャートを示す．

2 和音の PCM 波形を wavelet 変換しその後内積正規化を行う．そして学習済み Neural Network に入力し，2 和音のうち 1 音 $\mathbf{A}_i$ を読み取り同定 1 とする．次に仮音源候補 $\mathbf{B}_j$ を仮定して重み係数 a_i, b_j を計算する．そして式 (7) により $\mathbf{D}_i$ を求め，正規化したのちに Neural Network に入力する．Neural Network が識別した音源名と仮音源候補 $\mathbf{B}_j$ が一致したら $\mathbf{B}_j$ を同定 2 とする．

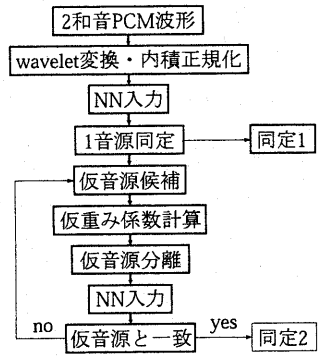

図 4　音源同定フローチャート（提案手法）

3. 評価実験

3.1 実験条件

音源テンプレートは MIDI 楽器サンプラーの出力を 16[bit]，44.1[kHz]，88,200[sample] で取り込んだ．サンプラーであれば PCM 波形のサンプリング時間を確認することができる．

wavelet 変換を行う区間は 5,512〜22,050[sample] とし，時間方向に $D_t = 4$ 分割，周波数方向に $D_f = 120$ 分割とした．これは 1[octave] を 12 分割とすると 10[octave] に相当し，22.8[Hz] から 22,050[Hz] までをカバーする．wavelet 関数は複素 Gabor 関数としたが，変換後は実部と虚部の二乗平均をスペクトルとしているため，複素関数である必要はない．

テンプレートの数 N_t は 25 とし，Guitar が 110〜440[Hz] の $N_{t_g} = 9$ 音源，Flute は 440〜880[Hz] の $N_{t_f} = 8$ 音源，Sax は 262〜523[Hz] の $N_{t_s} = 8$ 音源とする。

Neural Network の入力層数 N_{in} は周波数スペクトル数で $N_{in} = D_t \times D_f = 480$ である．また学習終了条件は $TH_{err} = 0.01$, $N_{BPth} = 8000$ とした．

3.2 Reset アルゴリズム実験

第 2.1.3 項にあるアルゴリズムを用いて Neural Network の学習実験を行った．出力層の数 N_{out} は式 (1) により 5 である．隠れ層の数 N_{hid}，学習効率 η，慣性項 α，学習させるテンプレートの順番をパラメータ群としたランダムサーチの実験である．ここでの wavelet の窓区間は 22,050[sample] である．

図 5 に誤差 err の減衰状況を示す．図 1 に比べて誤差が 0 に収束するようになったのが確認できる．しかし Neural Network は Reset on によりスパルタ的に学習させられたので，Reset off よりも誤差が大きくなってしまったものもある．

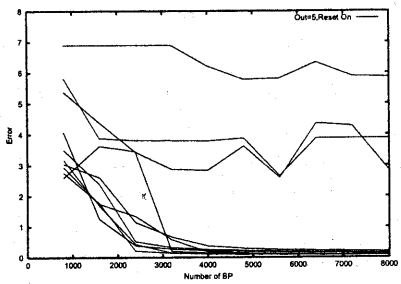

図 5　提案手法 Reset On $N_{out} = 5$

次に多数のパラメータ群から学習が完了した数とその割合を**表 1** に示す．学習が完了したとは $err < 1$ となったときとする．Reset On は Reset Off に比べて約 10 倍効率良く学習できたといえる．

3.3 勝者ユニットアルゴズム実験

第 2.1.4 項の実験を第 3.2 項と同じ環境で行った．出力層の数 N_{out} は 25 であり，隠れ層の数 N_{hid} も 25 以上の値である．**図 6** にその結果を示す．

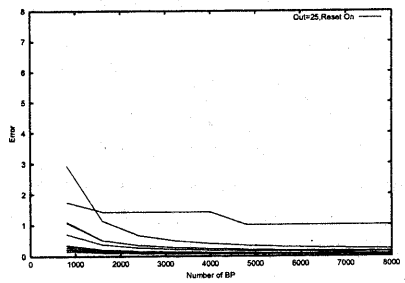

図 6　提案手法 Reset On $N_{out} = 25$

図 1，図 5 と比べて誤差は小さい値から始まっており，より多くのものが 0 に収束しているといえる．

表 1 では出力層の数が多くなった分一回の誤差逆伝搬の計算量は増大したが，誤差が 0 に収束する速度がが速くなり総合的に学習効率がよくなったといえる．

表 1　学習完了パラメータ群の割合

N_{out}	Reset	群数	学習数	学習完了率 [%]
5（従来）	Off	1561	68	4.4
5（従来）	On	905	362	40.0
25（提案）	On	592	512	86.5

3.4　音源同定実験

第 2.2.4 項の音源同定実験を行った．学習させたデータを 0.5:0.5 の割合で位相差を 0 として加算したのち wavelet 変換を施す．この 2 和音からどちらか一方を正しく読み取れたものの数とその割合を表 **2** に示す．wavelet 窓区間は 22,050[sample] であり，組み合わせの総数は ${}_{25}C_2 = 300$ である．

この表 2 の結果から，従来手法のグレイ符号を割り当てたものに対して勝者アルゴリズムを用いた提案手法は 77.7[%] から 94.0[%] へと認識率の向上が確認できた．

表 2　2 和音からの 1 音同定（同定 1）

N_{out}	1 音同定数	1 音認識率 [%]
5（従来）	233	77.7
25（提案）	282	94.0

次に仮音源分離を行い残りの 1 音の同定処理を行う．表 **3** に 2 和音同定実験結果を示す．wavelet 窓区間は 5,512[sample] である．認識率は従来手法が 73.0[%] に対し，提案手法は 93.5[%] へと向上することが出来た．

表 3　2 和音同定実験結果

		和音数	同定数	認識率 [%]
従来	全体	600	438	73.0
提案	同定 1	300	282	94.0
	同定 2	300	279	93.0
	全体	600	561	93.5

3.5　重み係数実験

ここでは位相差を 0 とし重み係数を変化させた和音の音源同定実験を行う．重み係数を 0.1:0.9〜0.9:0.1 まで変化させたときの認識率を図 **7** に示す．iden 1,2 は同定 1,2 を意味し，全体の認識率 total は iden 1 と iden 2 の平均となる．

同定 1 は重み係数が大きいほど認識率が高くなるが，逆に同定 2 は低くなってしまう．そして全体的には認識率は一定であり，重み係数の違いの影響は少ないといえる．

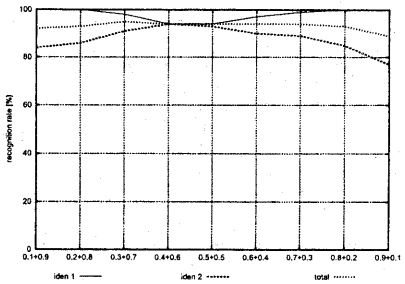

図 7　重み係数と認識率

3.6　位相差実験

ここでは重み係数を 0.5:0.5 とし位相をずらした和音の音源同定実験を行う．位相差を −400〜 +400[sample] までとしたときの認識率を図 **8** に示す．

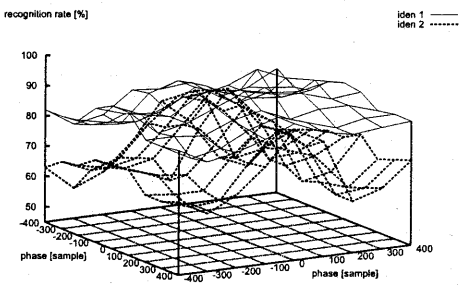

図 8　位相差と認識率

同定 1 は位相差が大きくても認識率はあまり変わらず位相差に強くなっている．一方同定 2 は位相差が大きくなると認識率は低くなり位相差に弱くなっている．

次に図 8 において認識率の最も低かった位相差条件のとき，計算で求めた重み係数とその正誤答数を図 **9** に示す．位相差は $(x, y) = (-400, -300)$[sample] のときであり，そのときの同定 1 の認識率は 75.7%，同定 2 は 55.0%，全体では 65.3%であった．weight a, b はそれぞれ同定 1,2 の重み係数であり，right,wrong はそれぞれの重み係数のときの正答，誤答である．

理想的なグラフは重み係数が 0.5 のときにピークとなるものであるが，計算で正しく読み取れていないことがわかる．また重み係数 a が 0.5 より大きいところに分布し，b はその逆となっている．そして a の right と wrong の差は大きいが b はほぼ同数となっており，a の認識率が高く b は低いことがわかる．

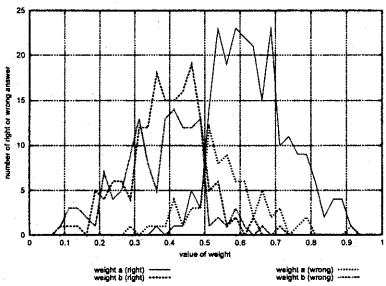

図 9　重み係数と正誤数

3.7　処理時間比較

ここでは従来手法と提案手法の計算処理時間の比較を行う．**表 4** に処理時間結果を示す．学習時間は提案手法の方が約 3.5 倍長くなっている．また音源同定の処理速度は約 3.5 倍に向上したといえる．

表 4　処理時間の比較

	学習時間 [分]	同定時間 [分]
従来手法	2	116
提案手法	7	33

4. 考　　察

4.1　学 習 効 率

表 4 より学習時間について考察する．従来に比べ学習時間は長くなったが，これは出力層と隠れ層の数が増えたことにより結合係数の数も約 5 倍に増えたからである．しかし隠れ層の数などのパラメータをランダムサーチした場合，表 1 より提案手法の方が約 20 倍優れているので，総合的に考えると約 $1/3.5 \times 20 = 5.6$ 倍に学習速度が向上したといえる．

4.2　認　識　率

まず同定 1 の認識率向上について考察する．表 2 では Neural Network が 2 和音から 1 音の読みとりが可能であることを示している．これは図 7 からも重み係数によらず 9 割以上の認識率が得られている．また図 8 からも位相差によらず 8 割以上の認識率が得られている．これらの要因として 2 和音から判別しやすいどちらかを判別しているからである．

一方表 3 から同定 2 は同定 1 より認識率が下がっている．また図 7，図 8 からも同様のことがいえる．同定 2 は同定 1 で読み取りにくかった方の判別になるため認識率は下がる．また図 7 より重み係数の影響は少ないが図 8 より位相差の影響は大きいといえる．

図 9 では同定 1 の係数 a が 0.5 より大きくなっており，内積相関が高いことを示している．これは Neural Network でも認識率が高いのと共通している．一方同定 2 の係数 b は相関が小さく認識率も下がってしまうと考えられる．

4.3　音源同定処理時間

表 4 より音源同定処理時間が向上したことについて考察する．まず仮音源候補において，従来手法は複数挙げていたが提案手法では 1 音源をほぼ確定したことにより，それ以降のループ処理の回数を減らすことができた．また仮音源分離において，従来手法は時間領域で位相差を検出したが提案手法では周波数領域で位相差を丸め込んで計算した．これにより位相差検出のための相互相関の計算時間を省くことができた．

5. おわりに

本研究では音源周波数分布を Neural Network で効率良く学習させるために Reset アルゴリズムと勝者ユニットアルゴリズムを提案し，その有効性を示した．また音源同定も勝者ユニットアルゴリズムにより 2 和音から 1 音の認識率が高いことを示し，仮音源分離後の Neural Network の音源同定可能性を確認した．その結果全体の認識率は向上し，その処理時間も高速化されることを示した．

参 考 文 献

1) 柏野 邦夫, 木下 智義, 中臺 一博, 田中 英彦:音楽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,D-II,Vol.J79-D-II,No.11,pp.1762-1770,1996 年 11 月
2) 柏野 邦夫, 村瀬 洋:アンサンブル実演奏の自動アンミキサ, 信学技報 SP97-104(1998-02)
3) 木下 智義, 坂井 修一, 田中 英彦:特徴量に注目した複数楽器の演奏における音源同定処理, 音楽情報科学 29-8(1999.2.19)
4) 北原 鉄朗, 後藤 真孝, 奥乃 博:音高による音色変化に着目した音源同定手法, 音楽情報科学 40-2(2001.5.23)
5) 秋田 真彦, 増山 繁, 船橋 賢一:ニューラルネットワークと主成分分析を用いた自動採譜の検討, 信学技報 NC92-81(1992-12)
6) 村瀬 樹太郎, 中西 正和: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, 音楽情報科学 39-18(2001.2.23)
7) 佃 卓磨, 甲藤 二郎:テンプレートマッチングを用いた音源同定に関する一検討, 2001 年電子情報通信学会総合大会 D-14-10
8) 佃 卓磨, 甲藤 二郎:信号解析とニューラルネットワークを用いた音源同定, 情報処理学会第 63 回全国大会 2Q-6