Pitch Extraction and U/V Detection of Speech by Cross-Coupling Multi Layered NN with Feedback Architecture

論　文

フィードバック付き多層相互結合形NNによる音声ピッチ抽出およびU/V判定

宮林　穎夫[†]　　船田　哲男[††]

Pitch Extraction and U/V Detection of Speech by Cross-Coupling Multi-Layered NN with Feedback Architecture

Hideo MIYABAYASHI[†] and Tetsuo FUNADA[††]

あらまし　音声の基本的特徴であるピッチ周波数の検出は，音声分析合成を行う上で，最も重要な研究課題の一つである．本論文では，時間連続性や高精度の識別関数を学習する可能性をもつ，出力層からのフィードバック結合や隠れ層内の相互結合を有する3種類の階層形ニューラルネットワークについて，ピッチ抽出およびU/V判定機能を学習させ，その検出性能を比較評価する．実験の結果，出力層からのフィードバック結合や隠れ層内に相互結合を有する階層形ネットワークが，フィードフォワード形と比べて最も検出精度が改善されおり，フィードバック結合や相互結合によって，ピッチ情報やU/V情報の時間連続性，高精度の識別関数が学習されることが明らかになった．

キーワード　ピッチ抽出，U/V判定，ニューラルネット，フィードバック結合，相互結合，時間連続性

1. はじめに

ピッチ周波数は音声の重要な基本的特徴量の一つである．その音声ピッチを正確に抽出することが，音声分析合成系のみならず，音声符号化，音声認識，話者認識等の際に必要となる基本的でかつ重要な課題である．より精度を高めるべく，これまでに研究されてきた代表的なピッチ抽出方法として，波形処理によるものや，相関処理によるもの，スペクトル処理によるもの等があげられるが，まだ決定的な方法は確立されていない[1]．

近年，ニューラルネットワーク（neural network, NN）を応用したピッチ抽出が報告されている[2]～[6]．本研究では，船田らが提案した帯域フィルタ対（band pass filter pairs, BPFP）法[7]によって抽出された音声の特徴から，今回提案するNNを用いてピッチを抽出する．

時間連続性や高精度の識別関数を学習できる可能性をもつ，出力層からのフィードバック結合や隠れ層内の相互結合を有するリカレント構造の3種類の階層形NNのそれぞれについて，ピッチ抽出および有声/無声（voiced/unvoiced, U/V）判定機能を学習させ，その検出性能を比較評価する．実験の結果，フィードバック付き多層相互結合形NN（cross-coupling multi-layered NN with feedback architecture, 以下CCNN-Fと略す）が通常の階層形NNと比べて最も高い検出精度が得られており，出力層・隠れ層間のフィードバック結合と隠れ層内の相互結合によって，ピッチ情報やU/V情報の時間連続性の学習，識別学習が最もよく行われていることが明らかになった．

なお，聴神経から間脳の内側膝状体までの上行する求心性系路において，信号の中継所を経るに従って，周波数分解能がシャープになっていく機構が明らかになっている[12]．回路網レベルでは同一層内でユニット間の興奮性と抑制性の相互結合の存在を示唆する．また，大脳皮質から下行する遠心性系路が見いだされ[12]，入力音に対する適応化のためのフィードバック結合の存在を示唆する．

以下では，まず本研究に用いたNNとその学習ア

[†] 富山商船高等専門学校，新湊市
Toyama National College of Maritime Technology, Shin-minato-shi, 933-02 Japan
[††] 金沢大学工学部，金沢市
Faculty of Engineering, Kanazawa University, Kanazawa-shi, 920 Japan

ルゴリズムについて説明し，次にピッチ抽出システムについて述べた後，性能比較実験の結果について詳述する．

2. ニューラルネットワークの結合形態と学習

本研究に用いた4種類の4階層形NNとその学習アルゴリズムについて順に説明する．

2.1 ネットワークの結合形態

(1) 単純な階層形NN（ネット形態：000）

図1(a)にそのNNの構成を示す．これは従来のフィードフォワード形NNそのものである．

(2) 隠れ層内相互結合形式の階層形NN（ネット形態：010）

図1(b)にそのNNの構成を示す．これは，従来の階層形NNの隠れ層内に1時点前の中間層の状態を保持するための状態層を設け，1時点前の中間層の状態が同一層内の各ユニットに伝達するように時間遅延素子Dと状態層を介して，隠れ層内各ユニット間で相互に結合しあう（自己ユニット間も含む）相互結合形NNを隠れ層内で構成するものである．言い換えれば，相互結合形構成の中間層を有する階層形NNである．

(3) 隠れ層内非相互結合形式のフィードバック付き階層形NN（ネット形態：101）

図1(c)にそのNNの構成を示す．これは，従来の階層形NNの出力層に1時点前の出力層の状態を保持するための状態層を設け，1時点前の出力層の状態が第1中間層の各ユニットに伝達するように時間遅延素子Dと状態層を介して，出力層と隠れ層（第1中間層）間のフィードバック結合を行うものである．

(4) 隠れ層内相互結合形式のフィードバック付き階層形NN（ネット形態：111）

図1(d)にそのNNの構成を示す．本NNは，図1(b), (c)の構成の両方を持ち合わせたものであり，CCNN-Fモデルと称する．本モデルの場合，状態層が全く無くても時間遅延素子Dのみを介してフィードバックおよび相互結合するNNで実現できるが，汎用的に状態層の過去の情報が参照できるように拡張して表現している．

2.2 学習アルゴリズム

本研究で用いた学習アルゴリズムは，バックプロパゲーション学習[8]に基づくものである．フィードバック結合荷重の学習は，状態層を通して1時点前の出力層に現れたパターンを一つの仮の外部入力パターンとみなすことで，通常の誤差逆伝搬法で行える．相互結合荷重の学習は，1時点前の隠れ層の学習誤差が，同一層内から状態層を通して直接の結合経路のみを逆伝搬する近似的な誤差伝搬法で行う．以下，学習アルゴリズムについてCCNN-Fモデルを例にして説明する．

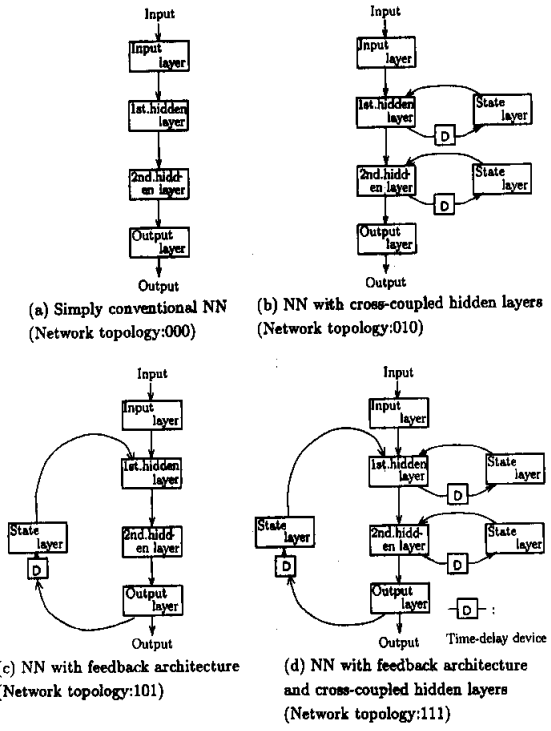

図1 各種の階層形ネットワーク
Fig. 1 Conventional NN and proposed NNs with feedback architecture or cross-coupled hidden layers.

本モデルの入力層，第1中間層，第2中間層，出力層のユニット数をそれぞれ I, $H1$, $H2$, O とする．NNの構成要素である一つのユニットの入出力関係を定式化すると一般に，

$$x_j^{(n+1)}(t) = \sum_i w_{ji}^{(n)} y_i^{(n)}(t) + \theta_j^{(n+1)} \quad (1)$$

$$y_j^{(n+1)}(t) = f(x_j^{(n+1)}(t)) \quad (2)$$

で表される．但し，t は離散的整数ステップで前向きに進むことを意味する．また，一般的に $f(x_j^{(n+1)}(t))$ は次のシグモイド関数で与えられる．

$$f(x_j^{(n+1)}(t)) = \frac{1}{1+\exp(-x_j^{(n+1)}(t)/\kappa)} \quad (3)$$

ここで，κ: シグモイド関数の傾き係数である．

出力層からのフィードバック結合や自層内に相互結合がある第1中間層では，式(1)，(2)は

$$x_j^{(2)}(t)=\sum_{i=1}^{I}w_{ji}^{(1)}y_i^{(1)}(t)+\sum_{m=1}^{O}u_{jm}y_m^{(N)}(t-1)+\sum_{l=1}^{H1}v_{jl}^{(1)}y_l^{(2)}(t-1)+\theta_j^{(2)} \quad (4)$$

$$y_j^{(2)}(t)=f(x_j^{(2)}(t)) \quad (5)$$

となる．ここで，$y_m^{(N)}(t-1)$：1時点前の出力ユニットmの出力値，u_{jm}：1時点前の出力ユニットmから第1中間層jへのフィードバック結合荷重，$y_l^{(2)}(t-1)$：1時点前の第1中間ユニットlの出力値，$v_{jl}^{(1)}$：同一層内ユニットlから第1中間ユニットjへの相互結合荷重である．なお，式(4)の第2項は仮の外部入力と考えれば同式の第1項と同様の取り扱いができる．

第2中間層では，同一層内に相互結合があるだけなので，式(1)，(2)は

$$x_j^{(3)}(t)=\sum_{i=1}^{H1}w_{ji}^{(2)}y_i^{(2)}(t)+\sum_{l=1}^{H2}v_{jl}^{(2)}y_l^{(3)}(t-1)+\theta_j^{(3)} \quad (6)$$

$$y_j^{(3)}(t)=f(x_j^{(3)}(t)) \quad (7)$$

となる．

出力層では式(1)，(2)は

$$x_j^{(N)}(t)=\sum_{i=1}^{H2}w_{ji}^{(3)}y_i^{(3)}(t)+\theta_j^{(N)} \quad (8)$$

$$y_j^{(N)}(t)=f(x_j^{(N)}(t)) \quad (9)$$

となる．

バックプロパゲーション学習とは，次式(10)の2乗出力誤差和$\mathbf{E}$が最小になるように結合荷重を調整することである．

$$\mathbf{E}=\sum_t E(t)=\frac{1}{2}\sum_t\sum_{j=1}^{O}(y_j^{(N)}(t)-y_j^d(t))^2 \quad (10)$$

ここで，$E(t)$：学習期間中の各時刻tにおける2乗出力誤差和，$y_j^d(t)$：時刻tにおける出力ユニットjの目標出力値である．ここでは，時刻ごとに結合荷重を修正すること（逐次修正法）とし，次式(11)，(12)のように，$E(t)$の偏微分値に比例して結合荷重およびバイアス値を修正する．

$$\left.\begin{aligned}\Delta w_{ji}^{(n)}(t)&=-\eta_1\frac{\partial E(t)}{\partial w_{ji}^{(n)}}+\alpha\Delta w_{ji}^{(n)}(t-1)\\ \Delta u_{jm}(t)&=-\eta_1\frac{\partial E(t)}{\partial u_{jm}}+\alpha\Delta u_{jm}(t-1)\\ \Delta v_{jl}^{(n)}(t)&=-\eta_1\frac{\partial E(t)}{\partial v_{jl}^{(n)}}+\alpha\Delta v_{jl}^{(n)}(t-1)\end{aligned}\right\} \quad (11)$$

$$\Delta\theta_j^{(n+1)}(t)=-\eta_2\frac{\partial E(t)}{\partial\theta_j^{(n+1)}}+\alpha\Delta\theta_j^{(n+1)}(t-1) \quad (12)$$

ここで，Δw_{ji}，Δu_{jm}，Δv_{jl}：それぞれの結合荷重の修正量，$\Delta\theta_j$：バイアス値の修正量，η_1,η_2：結合荷重用，バイアス値用の学習率，α:慣性率である．

前式(11)，(12)第1項の偏微分は，更にそれぞれ

$$\left.\begin{aligned}\frac{\partial E(t)}{\partial w_{ji}^{(n)}}&=-\delta_j^{(n+1)}(t)\frac{\partial x_j^{(n+1)}(t)}{\partial w_{ji}^{(n)}}\\ \frac{\partial E(t)}{\partial u_{jm}}&=-\delta_j^{(n+1)}(t)\frac{\partial x_j^{(n+1)}(t)}{\partial u_{jm}}\\ \frac{\partial E(t)}{\partial v_{jl}^{(n)}}&=-\delta_j^{(n+1)}(t)\frac{\partial x_j^{(n+1)}(t)}{\partial v_{jl}^{(n)}}\end{aligned}\right\} \quad (13)$$

$$\frac{\partial E(t)}{\partial\theta_j^{(n+1)}}=-\delta_j^{(n+1)}(t)\frac{\partial x_j^{(n+1)}(t)}{\partial\theta_j^{(n+1)}} \quad (14)$$

のように表現できる．但し，$-\delta_j^{(n+1)}(t)$は

$$-\delta_j^{(n+1)}(t)=\frac{\partial E(t)}{\partial y_j^{(n+1)}(t)}f'(x_j^{(n+1)}(t)) \quad (15)$$

であり，シグモイド関数の微分$f'(x_j^{(n+1)}(t))$は

$$f'(x_j^{(n+1)}(t))=f(x_j^{(n+1)}(t))\times(1-f(x_j^{(n+1)}(t)))/\kappa \quad (16)$$

で与えられる．

ここで，簡単化のために，他ユニットからの誤差逆伝搬は直接の結合経路のみを考えること（近似的学習）とする[9],[10]．式(13)および式(14)右辺の偏微分は，式(4)，(6)，(8)より

$$\left.\begin{aligned}\frac{\partial x_j^{(n+1)}(t)}{\partial w_{ji}^{(n)}}&=y_i^{(n)}(t)\\ \frac{\partial x_j^{(n+1)}(t)}{\partial u_{jm}}&=y_m^{(N)}(t-1)\\ \frac{\partial x_j^{(n+1)}(t)}{\partial v_{jl}^{(n)}}&=y_l^{(n+1)}(t-1)\end{aligned}\right\} \quad (17)$$

$$\frac{\partial x_j^{(n+1)}(t)}{\partial\theta_j^{(n+1)}}=1 \quad (18)$$

となる．

出力ユニット j の誤差 $\delta_j^{(N)}(t)$ は式 (15) より

$$\delta_j^{(N)}(t) = f'(x_j^{(N)}(t))(y_j^d(t) - y_j^{(N)}(t)) \qquad (19)$$

と表される．同様にして，第 2 中間ユニット j の誤差 $\delta_j^{(3)}(t)$ は

$$\delta_j^{(3)}(t) = f'(x_j^{(3)}(t))(\sum_{k=1}^{O} \delta_k^{(N)}(t) w_{kj}^{(3)} + \sum_{l=1}^{H2} \delta_l^{(3)}(t+1) v_{lj}^{(2)}) \qquad (20)$$

第 1 中間ユニット j の誤差 $\delta_j^{(2)}(t)$ は

$$\delta_j^{(2)}(t) = f'(x_j^{(2)}(t))(\sum_{k=1}^{H2} \delta_k^{(3)}(t) w_{kj}^{(2)} + \sum_{l=1}^{H1} \delta_l^{(2)}(t+1) v_{lj}^{(1)}) \qquad (21)$$

と表される．

以上のような式 (11)～(18) を用いた学習則により，すべての結合荷重やバイアス値の修正量を求めることができる．

3. ピッチ抽出システム

本研究に用いたピッチ抽出システムは，BPFP バンク部とピッチ抽出 NN 部および U/V 判定 NN 部からなる．以下にこの三つの部分について順に説明する．

3.1 BPFP バンク部

BPFP バンク部は，サンプリング（0.1 ms 間隔，16 ビット分解能）された音声信号から BPFP 法によってフレーム単位（10 ms 周期，30 ms 長）の音声の特徴を抽出する．BPFP 法とは，対象周波数帯域（本研究では中心周波数 100～250 Hz, 15 Hz 間隔の 11 チャネル分と中心周波数 280～580 Hz, 30 Hz 間隔の 11 チャネル分）の複数の周波数点でパワースペクトルの周波数に関する傾斜とパワーを求め，これを音声の特徴ベクトルとする考え方である．ここでいう傾斜とは，各チャネルの中心周波数に近い高調波周波数がその中心周波数のどちら側にどれだけ離れているかの程度を表す値である．高調波周波数が中心周波数より大きいときは正値，中心周波数より小さいときは負値を示す [7]．

BPFP バンクはこのような 44 次元の特徴ベクトルをフレームごとに出力する．

3.2 特徴ベクトルの正規化

音声信号の振幅に依存しないピッチ抽出および U/V 判定を行うために，前処理部において BPFP バンク部から出力された特徴ベクトルを正規化する必要がある．

スペクトルの周波数に関するパワーに基づく 22 次元特徴ベクトルは，フレームごとにその成分の最大値を求め各成分をその最大値で割ることにより，[0, 1] の範囲に収まるように正規化する．また，元々振幅に依存しない傾きに基づく 22 次元特徴ベクトルについても，フレームごとに各成分が [−1, 1] の範囲に収まるように正規化しておくことにする．

従って，後続のピッチ抽出 NN 部および U/V 判定 NN 部には，音声信号のフレームごとに正規化された 44 次元特徴ベクトルが入力されることになる．

3.3 U/V 判定 NN 部

U/V 判定 NN 部は，BPFP バンクから出力された特徴ベクトルを基にそのフレームの U/V を判定するものである．NN は，入力層が 44 ユニット，隠れ層が 2 層で第 1 中間層 30 ユニットと第 2 中間層 15 ユニット，出力層 1 ユニットで構成する．

NN の学習時には，有声フレームに対して 0.99 を，無声または無音フレームに対して 0.01 を教師信号として与える．また評価時には，出力層の出力値が 0.5 を超えると有声，0.5 以下であれば無声と判断することにする．

3.4 ピッチ抽出 NN 部

ピッチ抽出 NN 部は，U/V 判定 NN 部で判定された有声フレームについて，BPFP バンクから出力された特徴ベクトルを基にピッチ周波数を抽出するものである．NN は U/V 判定 NN 部と同じ構成である．

NN の学習時には，有声フレームに対して正解ピッチを教師信号として与える．また評価時には，抽出されたピッチ周波数が正解ピッチの ±5 %未満の範囲内であればピッチ抽出に成功したと判断することにする．

音声信号のピッチ周波数の上限を 450 Hz, 下限を 50 Hz と考えて，実際に NN の出力層に与える正解ピッチの教師信号は

$$(\text{教師信号}) = \frac{\log((\text{正解ピッチ})/50.0)}{\log(9.0)} \qquad (22)$$

の計算式により [0, 1] の範囲に収まるように変換する．また，NN の出力層の出力値から抽出ピッチを求めるには

(抽出ピッチ)＝50.0×exp((出力値)×log(9.0))　(23)

の計算式により逆変換すればよい．

4. 実験方法

4種類のピッチ抽出NNおよびU/V判定NNを学習し，その検出性能を比較評価した実験の内容について説明する．

4.1 音声資料

実験に用いた音声資料は，研究用連続音声データベース（日本音響学会）より次のようなものを選んだ．

- 学習内話者：男性話者3名（m01, m02, t01）と女性話者3名（m11, m12, t11）の計6名
- 学習外話者：男性話者1名（t03）と女性話者1名（t12）の計2名
- 学習用データ1：学習内話者による発話文5文（a01～a05）の計30文
- 学習用データ2：学習内話者による発話文5文（a06～a10）の計30文
- 評価用学習外データ：学習用データ1で学習する場合は学習内話者の発話文a06～a10の計30文，学習用データ2で学習する場合は学習内話者の発話文a01～a05の計30文，学習外話者による発話文10文（a01～a10）の計20文

4.2 正解ピッチ

学習しようとする音声資料をBPFPバンク部で処理し，128種類（70～451 Hz, 3 Hz間隔）の標準パターンとのマッチングにより，U/V情報とピッチ情報を抽出し，更に，目視で修正を加えたものを正解ピッチとした．なお，無声フレームの場合，この正解ピッチの値を便宜上0とし，教師信号として与えるU/V情報は正解ピッチの値（＝0/＋0）から定めている．

4.3 ピッチ抽出実験

学習用データ1，2の個別の場合と両者合体で学習した場合のNNを用いてピッチ抽出を行い，ネットワーク構成の違いによる抽出性能の差異について比較検討する．但し，正解ピッチの値が0でない有声フレームを対象として，ピッチ抽出NNの評価を行う．4種類のNNの学習において，同一の学習条件（シグモイド関数の傾き係数：0.8，学習率：初期値0.8，減少率0.99，慣性率：初期値0.5，減少率0.99，結合荷重初期値：−0.5～＋0.5，バイアス値初期値：−0.3～＋0.3，学習回数：2000回）を設定した．学習が終了したNNの性能評価項目として，次のものを用いた．

- 正解ピッチと大きく異なる（＞±20%）ピッチを抽出するフレームの割合（gross pitch error, GPE）
- ピッチ抽出に成功するフレームの割合（正解率）

4.4 U/V判定実験

学習用データ1，2の個別の場合と両者合体で学習した場合のNNを用いてUV判定を行い，ネットワーク構成の違いによる判定性能の差異について比較検討する．4種類のNNの学習において，同一の学習条件（シグモイド関数の傾き係数：1.5，学習回数：1000回，その他ピッチ抽出実験と同じ）を設定した．学習が終了したNNの性能評価項目として，次のものを用いた．

- 無声を有声と間違えるフレームの割合（unvoiced-to-voiced error, UVE）
- 有声を無声と間違えるフレームの割合（voiced-to-unvoiced error, VUE）
- 正解フレームの割合（正解率）

5. 実験結果および考察

まず学習用データ1，2の個別で学習した場合のピッチ抽出結果およびU/V判定結果より，フィードバック結合や相互結合による時間連続性の学習効果と識別学習の効果を解析し，次に両者合体で学習した場合のピッチ抽出およびU/V判定結果について説明する．

5.1 識別関数の精度

学習用データ1，2の個別で学習したピッチ抽出NNにおけるピッチ情報の抽出精度のばらつき状況を，GPE値とGPE以外のフレームについて正解ピッチからの絶対偏差の平均/分散値を用いて表1に示す．同表は，学習データ1，2で個別に学習した各NN（以下「学習用データ1NN」，「学習用データ2NN」と略す）を用いて評価用学習外データで評価したものである．同表より，ネット形態：010, 111では学習内話者，学習外話者ともGPEが少なくかつ抽出精度のばらつきが小さいことが読み取れる．図2にピッチ抽出結果の一例を示す．ネット形態ごとの抽出精度のばらつきや抽出誤り状況を視覚的に確認することができる．但し，無音あるいは無声区間については表示していない．

隠れ層内全ユニット間相互結合形式のネット形態：010, 111のNNによるピッチ抽出およびU/V判定の正解率を，それぞれネット形態：000, 101と比較して表2に示す．「学習用データ1NN」，「学習用データ2NN」の両者において，相互結合による効果は学習内話者，学習外話者ともピッチ抽出の場合が顕著にみられる．

表1 各ネット形態におけるピッチ情報の抽出精度のばらつき（学習外データ，単位：平均 [Hz]，分散 [Hz^2]，GPE[%]）

Table 1 Scatter of the accuracy of pitch extraction for each NN (non-training data, a unit: average [Hz], variance [Hz^2], GPE [%]).

平均/分散：GPE を除いた抽出ピッチの誤差の平均/分散値

ネット形態	学習用データ 1NN						学習用データ 2NN					
	学習内話者			学習外話者			学習内話者			学習外話者		
	平均	分散	GPE	平均	分散	GPE	平均	分散	GPE	平均	分散	GPE
000	3.03	16.43	1.48	2.73	12.11	0.63	3.25	17.98	1.47	3.09	12.70	0.54
010	2.85	13.73	1.09	2.62	10.04	0.28	2.94	12.84	0.65	2.74	9.32	0.3
101	2.92	18.09	1.61	2.48	10.46	0.65	3.39	18.19	1.64	3.22	13.67	0.54
111	2.87	14.01	1.28	2.68	11.11	0.32	2.95	15.46	0.51	2.80	10.74	0.20

表2 隠れ層内相互結合有りとした場合の正解率の比較（学習外データ，単位：%）

Table 2 Comparison of correction ability of pitch extraction and U/V detection by using NN with cross-coupled hidden layers (non-training data, a unit: %).

ネット形態：010, 111 の正解率はそれぞれネット形態：000, 101 からの相対値

	ネット形態	学習用データ 1NN		学習用データ 2NN	
		学習内話者	学習外話者	学習内話者	学習外話者
ピッチ抽出	010	+1.95	+1.86	+4.44	+3.84
	111	+0.98	+0.68	+4.77	+4.86
U/V 判定	010	+0.76	−0.03	+0.01	+0.14
	111	−0.06	−0.21	+0.19	+0.65

表3 各ネット形態における U/V 情報の時間連続性の学習効果（学習外データ）

Table 3 Learning effects of time-continuity of U/V for each NN (non-training data).

平均/分散：前フレームからの変化量の平均/分散について U/V 判定結果を正解値で割った値（増加率）

ネット形態	学習用データ 1NN				学習用データ 2NN			
	学習内話者		学習外話者		学習内話者		学習外話者	
	平均	分散	平均	分散	平均	分散	平均	分散
000	1.41	1.38	1.25	1.24	1.37	1.35	1.27	1.25
010	1.21	1.20	1.21	1.20	1.34	1.32	1.29	1.27
101	1.07	1.07	1.10	1.09	1.11	1.11	1.15	1.14
111	1.07	1.07	1.14	1.13	1.14	1.13	1.17	1.16

表4 出力層・隠れ層間フィードバック結合有りとした場合の正解率の比較（学習外データ，単位：%）

Table 4 Comparison of correction ability of pitch extraction and U/V detection by using NN with feedback architecture (non-training data, a unit: %).

ネット形態：101, 111 の正解率はそれぞれネット形態：000, 010 からの相対値

	ネット形態	学習用データ 1NN		学習用データ 2NN	
		学習内話者	学習外話者	学習内話者	学習外話者
ピッチ抽出	101	+0.52	+0.63	−0.21	−0.04
	111	−0.45	−0.55	+0.12	+0.98
U/V 判定	101	+0.29	+0.69	+0.39	−0.26
	111	−0.53	+0.51	+0.57	+0.25

表 5　各ネット形態において，学習用データ 1,2 両者合体で学習した場合のピッチ抽出および U/V 判定結果（学習外話者，単位：%）

Table 5　Results of pitch extraction and U/V detection by using each NN with both data 1 and 2 (non-training, a unit: %).

ネット形態	ピッチ抽出		U/V 判定		
	正解率	GPE	正解率	UVE	VUE
000	93.26	0.56	95.81	4.59	3.89
010	93.93	0.30	96.17	3.19	4.32
101	95.68	0.52	95.51	5.97	3.37
111	96.63	0.28	96.24	4.18	3.45

表 6　各種方法による U/V 判定およびピッチ抽出結果（学習外話者，単位：%）

Table 6　Comparison with other methods of U/V detection and pitch extraction (non-training, a unit: %).

各種方法	U/V 判定			ピッチ抽出		システム全体		備考
	正解率	UVE	VUE	正解率	GPE	正解率	UVE+VUE+GPE	
ケプストラム	86.12	7.69	18.58	78.15	1.07	84.26	14.48	しきい値＝2.0
残差波形自己相関	93.63	11.26	2.66	88.36	4.48	88.52	8.92	しきい値＝0.15
NN	95.81	4.59	3.89	90.98	0.26	92.90	4.34	ネット形態：000
CCNN-F	96.24	4.18	3.45	94.19	0.08	94.90	3.81	ネット形態：111

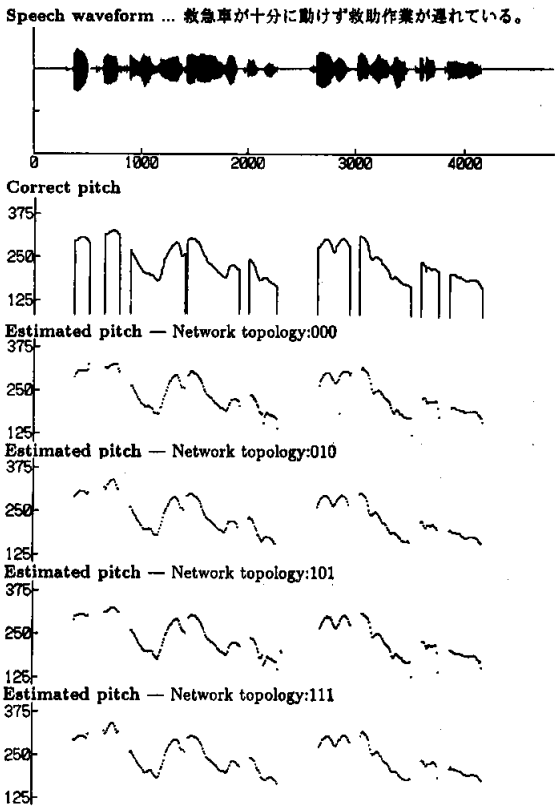

図 2　学習用データ 1 で学習した場合のピッチ抽出結果の例
(t12-a05, 学習外女性話者)

Fig. 2　An example of pitch extraction for female speaker by using each NN trained with data 1(t12-a05,non-training speaker).

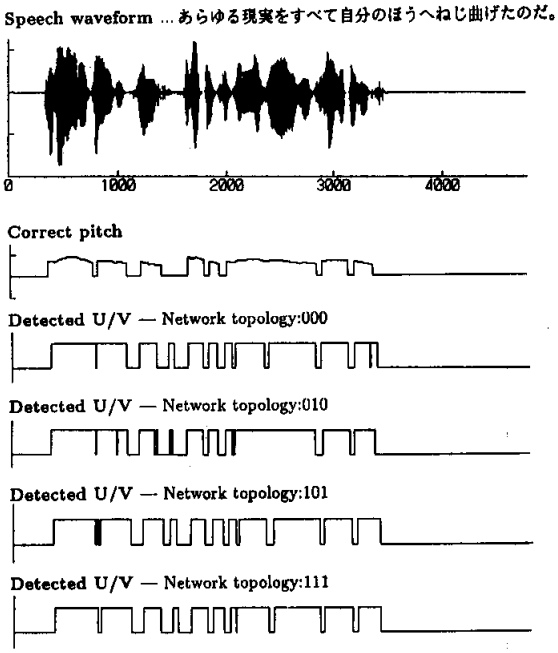

図 3　学習用データ 2 で学習した場合の U/V 判定結果の例
(t03-a01, 学習外男性話者)

Fig. 3　An example of U/V detection for male speaker by using each NN trained with data 2 (t03-a01, non-training speaker).

以上のことから，隠れ層内に相互結合をもつ NN の学習は特にピッチ抽出において識別関数の精度を向上させる効果があるといえる．これは，相互結合によってピッチ変動情報を平滑化して少ない学習データ数で学習の汎化性をもたらすためである [13]．

5.2 時間連続性の学習効果

学習用データ 1，2 の個別で学習した U/V 判定 NN における U/V 情報の時間連続性の学習効果を表 3 に示す．同表の平均/分散値は，2 値化された U/V 判定値と正解値のそれぞれについて前フレームからの変化量の平均/分散を求め，判定値の平均/分散を正解値のそれらで割った値（増加率）を表す．同表より，ネット形態：101, 111 では学習内話者，学習外話者とも平均/分散の増加率が小さいこと，すなわち U/V 判定値の変化回数が正解値の回数に近いことが読み取れる．U/V 判定結果の一例を図 3 に示す．ネット形態ごとの判定の誤り状況を視覚的に確認することができる．

フィードバック付きのネット形態：101, 111 の NN によるピッチ抽出および U/V 判定の正解率を，それぞれネット形態：000, 010 と比較して表 4 に示す．フィードバック結合による効果は「学習用データ 1NN」と「学習用データ 2NN」とでは異なるが，学習外話者について，「学習用データ 1NN」，「学習用データ 2NN」を通して平均的に捕え直すと，ピッチ抽出，U/V 判定ともわずかであるがフィードバック結合による改善効果はあると考えてよい．

以上のことから出力層から隠れ層へのフィードバック結合をもつ NN は，特に U/V 情報のような時系列パターンの時間軸方向の連続性を学習し，U/V 境界で発生しやすい U/V 情報のチャタリングを減少させる効果がある．しかし，時間連続性の学習効果は直接正解率の向上に寄与しない [13]．

5.3 ピッチ抽出および U/V 判定結果

学習用データ 1，2 の両者合体で学習した各 NN によるピッチ抽出および U/V 判定結果を表 5 に示す．同表より，ピッチ抽出，U/V 判定ともネット形態：111 すなわち CCNN-F が最もよく改善され，ネット形態：000 と比較して前者は正解率約 3.4%増加，GPE 約 0.3%減少，後者は正解率約 0.4%増加，UVE，VUE 共約 0.4%減少したことがわかる．

なお，ネット形態：101 のピッチ抽出 NN やネット形態：010 の U/V 判定 NN において，ネット形態：000 と比較して正解率が改善されているのは，合体学習により学習データ数が増加したためである．CCNN-F の場合，学習データ数の多少にかかわらず安定して最も高い検出精度が得られている [13]．

5.4 他の方法との比較

学習用データ 1，2 の両者合体で学習した NN（ネット形態：000）法による U/V 判定およびピッチ抽出結果を，他の代表的な手法であるケプストラム法や LPC 残差波形の自己相関法 [11] と比較して表 6 に示す．この場合，ピッチ平滑化などの後処理を行わないでフレームごとに抽出された値を直接比較する．同表より，本 NN 法による正解率，誤り率（UVE+VUE, GPE）は共に他の 2 法より上回っていることがわかる．U/V 判定結果を参照し，有声と判定されたフレームのみでピッチ抽出するようなシステム全体で考えた場合，他の 2 法より，システム全体の正解率は約 4.4%以上の増加，誤り率（UVE+VUE+GPE）は約 4.6%以上の減少となり，本 NN 法の優位性が明らかである(注 1)．なお，ケプストラム法や残差波形自己相関法による場合は U/V 判定のしきい値をどう定めるかが結果に大きく影響するが，ここではシステム全体の正解率が最も大きくなるように設定した．

また，CCNN-F（ネット形態：111）法は前述の単純な通常の NN 法と比較して，システム全体で正解率 2%増加の 94.9%，誤り率約 0.5%減少の 3.8%であり，今回提案した CCNN-F モデルによって更に改善されたといえる．

6. むすび

本研究では，BPFP バンクと 4 種類の階層形 NN を用いて，ネットワーク構成の違いによるピッチ抽出および U/V 判定性能の比較実験を行った．実験結果から次の知見が得られた．

- BPFP バンクと通常の階層形 NN を用いた方法は，他の代表的な手法であるケプストラム法や LPC 残差波形の自己相関法と比べて，かなり有効なピッチ抽出法である．
- 出力層・隠れ層間のフィードバック結合と隠れ層内の相互結合を有する CCNN-F はロバスト的で最も高い検出精度が得られており，通常の階層形 NN と比べて，2%正解率が増加し約 0.5%誤り率（UVE+VUE+GPE）が減少し，フィードバック結合と相互結合の相乗効果である．
- 隠れ層内に相互結合をもつ NN の学習は特にピッチ抽出において識別関数の精度を向上させる効果がある．
- 出力層から隠れ層へのフィードバック結合をも

(注 1)：ここでシステム全体の正解率とは，UVE，VUE，および正解ピッチの ±5%未満の範囲を逸脱したピッチ抽出エラーのいずれにも該当しないフレームを正解フレームとし，全フレーム数に対する割合で算出している．

つNNは，特にU/V情報の時間連続性を学習する．

今後の課題として，本方法によるピッチ抽出結果を音源とした合成音での聴取実験や，雑音などで劣化した音声のピッチ抽出実験を行い，今回提案したCCNN-Fモデルの有効性を確認する必要がある．また，隠れ層内の相互結合経路数の増加と識別関数の精度向上の関係を追求する必要がある．更に，識別関数の学習が十分でないと思われるので，隠れ層内の直接経路の1時点前の誤差逆伝搬までしか考慮していない近似的学習アルゴリズムの改良が必要となる．

文　献

[1] 古井貞熙，音響・音声工学，近代科学社，1992.

[2] E. Barnard, R.A. Cole, M.P. Vea, and F.A. Allcva, "Pitch detection with a neural-net classifier," IEEE Trans., Signal Processing, vol.39, no.2, pp.298–307, Feb. 1991.

[3] H. Martinez-Alfaro and J.L. Contreras-Vidal, "A robust real-time pitch detector based on neural networks," Proc. Int. Conf. ASSP, Toronto, Canada, pp.521–523, 1991.

[4] T. Ghiselli-Crippa and A. El-Jaroudi, "A fast neural net training algorithm and its application to voiced-unvoiced-silence classification of speech," ibid., pp.441–444.

[5] A. Ogihara and K. Fukunaga, "A correcting method for pitch extraction using neural networks," IEICE, Trans. Electron., vol.E77-A, no.6, pp.1015–1022, June 1994.

[6] 宮林頼夫，船田哲男，"ネットワーク構成の違いによる有声/無声判定NNの性能比較，" 信学技報，vol.DSP94-107, Jan. 1995.

[7] 船田哲男，鈴木達也，"帯域フィルタ対バンクによる音声ピッチ抽出，" 信学論(A)，vol.J72-A, no.3, pp.466–474, March 1989.

[8] D.E. Rumelhart and J.L. McClelland (ed.), Parallel distributed processing, MIT Press, 1988.

[9] K. Doya and S. Yoshizawa, "Adaptive neural oscillator using continuous-time back-propagation learning," Neural Networks, vol.2, pp.375–385, 1989.

[10] 銅谷賢治，"ニューラルネットワークによる振動パターンの記憶，" コンピュートロール，vol.29, pp.52–62, 1990.

[11] J.D. Markel and A.H. Gray, Jr., Linear prediction of speech, Springer-Verlag, Berlin, 1976.

[12] 三浦種敏監修，新版聴覚と音声，電子情報通信学会，1980.

[13] 宮林頼夫，船田哲男，"ピッチ抽出NNにおけるフィードバックおよび層内相互結合の効果の検討，" 信学技報，vol.SP95-41, July 1995.

(平成7年7月12日受付，12月18日再受付)

宮林　頼夫　(正員)

昭44富山大・工・電気卒．昭46同大大学院修士課程了．同年日本重化学工業入社．平1富山商船高専・情・助教授，現在に至る．ニューラルネットワーク，音声情報処理，並列処理に関する研究に従事．情報処理学会，計測自動制御学会各会員．

船田　哲男　(正員)

昭41金沢大・工・電子卒．昭46名大大学院博士課程了．工博．昭46金沢大・工・講師．生体信号処理，音声情報処理の研究に従事．現在，同大教授．共著「情報科学の基礎」，共著「数値解析の基礎」など．IEEE，日本音響学会，日本ME学会，情報処理学会各会員．