TR-I-0178

時間遅れ神経回路網を用いた 不特定話者の音韻認識

Speaker-Independent Phoneme Recognition Using Time-Delay Neural Networks

中村 悟* 沢井 秀文
Satoru Nakamura and Hidefumi Sawai

1990.9.7.

概要

時間遅れ神経回路網(TDNN)を基本構造として、特定話者用及びマルチスピーカー用のネットワークを不特定話者の音素認識に適用し、その能力を比較検討した。各ネットワークの学習に用いる話者数を6人および12人とし、有声破裂音/b, d, g/の3音韻を用いて実験を行ったところ、学習に6話者及び12話者を用いた場合の話者オープンの認識率は、最高でそれぞれ92.1%と95.6%であった。

またネットワークの構造として、Modular TDNN のように話者に対応するモジュールを用いてネットワークを構成することは、学習回数の軽減という面において有効であり、同じ程度のキャパシティーをもつSingle TDNNと比較すると、認識率は若干上回った。これは限られたネットワークのキャパシティーを有効に利用しているためと考えられる。また、一方でSingle TDNNの隠れ層のキャパシティーを更に増すことによっても、Modular TDNNを上回る認識率を得ることができた。

*慶応大学理工学部
*Faculty of Science and Technology, Keio University
ATR 自動翻訳電話研究所
ATR Interpreting Telephony Research Laboratories

目　次

1. はじめに

近年、音声認識の分野において、ニューラルネットワークの応用が活発に行われてきている[1,2,3]。特にTDNN(Time-Delay Neural Network)により、/b,d,g/の音韻認識において高い性能が示されて以来[1]、TDNNを基本構造とする18子音認識用のネットワーク[2]や23音韻認識用[9]のネットワーク、マルチスピーカーの音韻認識を行うネットワークが多数提案されてきた[4-6]。

本報告では、これらのネットワークを不特定話者の音韻認識に拡張するため、種々のネットワークの識別能力を比較検討する。比較を行うネットワークは、各ネットワークの基本構造であるSingle TDNN、Modular TDNN、SID(Stimulus Identification)ネットワーク、Meta-Piネットワークである。最後にあげたMeta-Piネットワークはマルチスピーカーの音韻認識において、最も認識率の良いネットワークとして報告されている[4-6]。

また本報告では、Meta-Piネットワークを不特定話者認識へ拡張する目的でModular Speaker Identification TDNNと呼ぶ構造を提案する。各ネットワークにおいて、6人及び、12人の話者の発声音韻/b, d, g/を学習させ、そのときのClosed Speaker、Open Speakerの音韻認識率を求めた。その結果をもとに、いくつかの実験を行いネットワーク能力を比較検討する。

2. ネットワークの構造

● Single TDNN[1]

各ネットワークの基本構造となるSingle TDNNの構造をFig.1に示す。Fig.1は、有声破裂音/b, d, g/の音韻認識のための4層TDNNである。TDNNへの入力は、メルスケールで配置された16チャネルのフィルタバンク出力15フレーム分(150msec)であり、入力層は16×15の240ユニットよりなる。第1隠れ層は、周波数方向に8ユニット、時間方向に13ユニットの104ユニット、第2隠れ層は、周波数方向に識別音韻に対応する3ユニット、時間方向に9ユニットの27ユニットである。入力層から第1隠れ層への結合は、入力層の3フレーム分が第1隠れ層の1フレーム分とフルコネクションをもち、第1隠れ層から第2隠れ層への結合は、第1隠れ層の5フレーム分が第2隠れ層の1フレーム分とフルコネクションをもつ。これらの結合は、下層のフレームが1フレーム分シフトするごとに、上層の次のフレームに結合し、シフトしたときの結合は、タイドコネクションになっている。これは、入力に対してシフトトレランスをもたせる目的によるものである。最後に、第2隠れ層から出力層への結合は、シフトトレランスの目的から、出力層の各ユニットへ対応する9個のユニットから全て同じ重みで、結合している。

● Modular TDNN[6]

Modular TDNNの構造は、Fig. 2に示すような、Single TDNNを並列に並べたモジュール構造である。Modular TDNNは、特定話者に対して学習されたSingle TDNNの重みを初期値としてネットワーク全体を再学習する2-stage learningにより構成される。

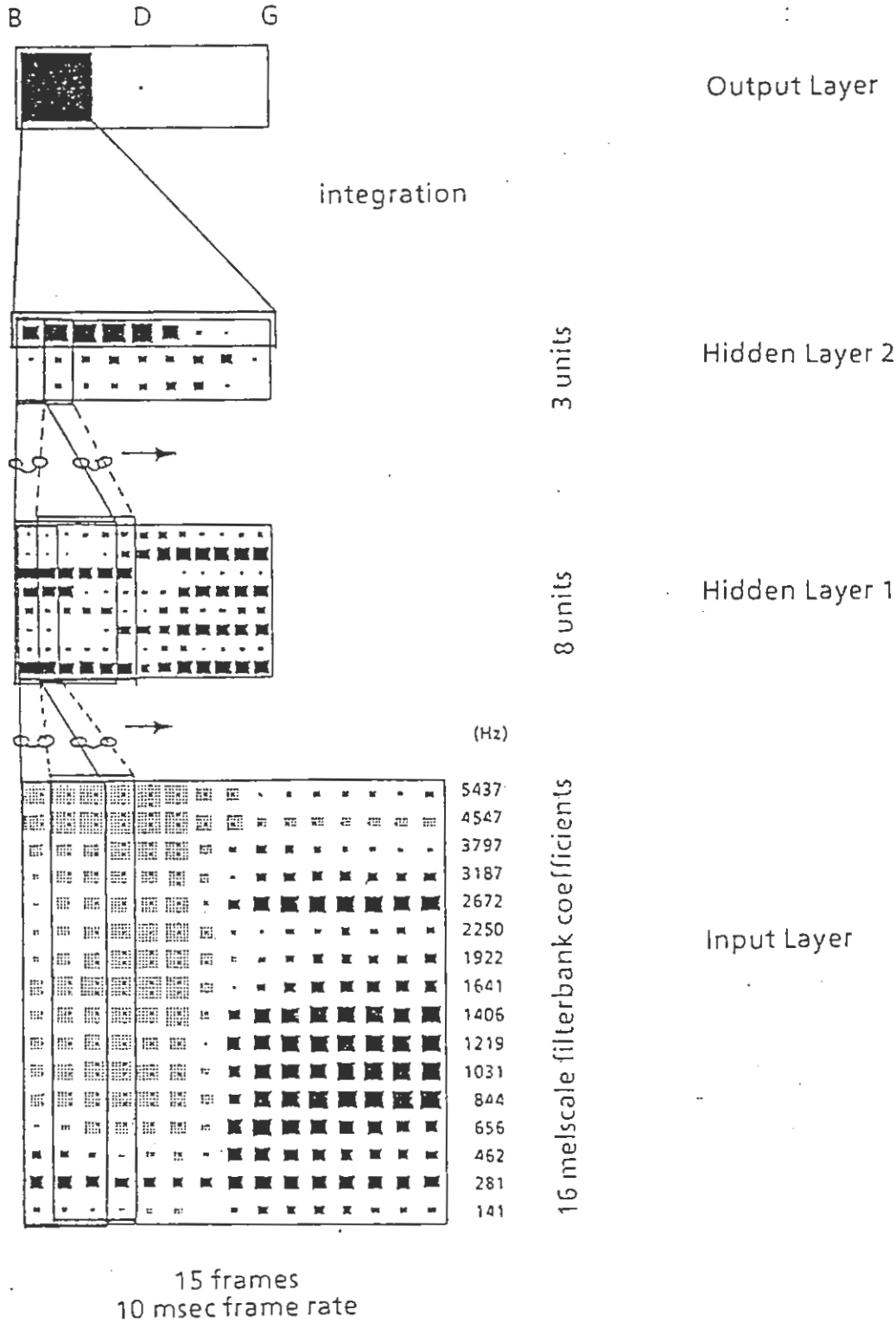

Fig. 1. Single TDNN の構造[1]

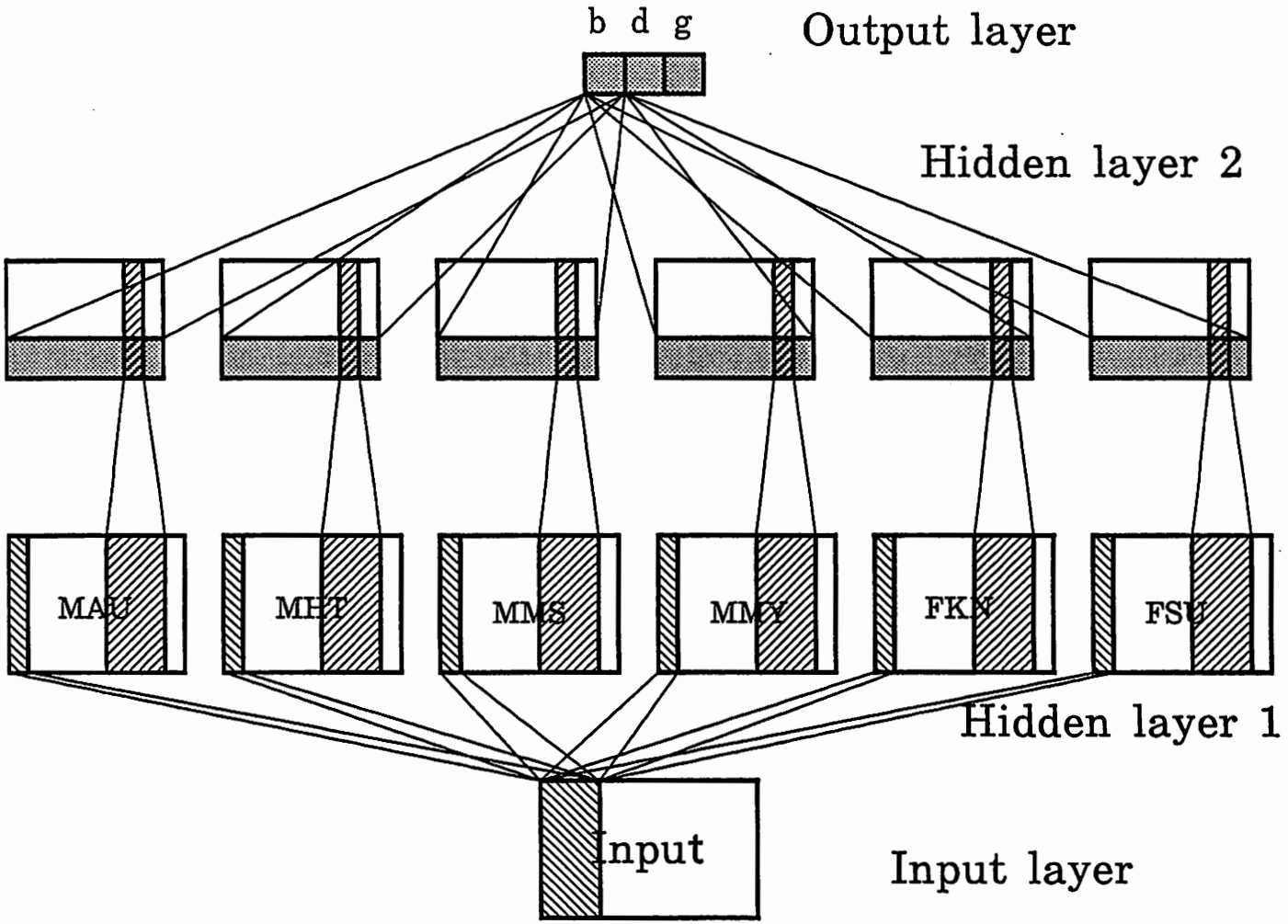

Fig. 2. Modular TDNN の構造

● SID (Stimulus Identification) network[5]

SID network の構造をFig.3 に示す。Mudular TDNN と同じように各話者に対するモジュールを用い、更に、学習済みの話者認識用のネットワークを付加した構造になっている。話者認識用のネットワークは、与えられた入力が、どの話者に対応する入力であるかを特定し、その話者に対応するモジュールの出力を、全体の出力とする。その他にも、SID network の出力を決定するために、話者認識用のネットワークの出力に比例して、各話者の出力を結合することによりネットワークの全体の出力を得る方法がある。これら2つの方法は、各々winner-take-all, proportional conbination と呼ばれる[5]。

このネットワークの特徴は、各モジュールが完全に独立に動作することである。

●Meta-Pi network[5]

Meta-Pi ネットワークの構造は、Fig.4 に示すようにSID networkと同じ構造で、各話者に対するモジュールとその出力を線形結合させるMeta-Pi ネットよりなる。SID network との違いは、各モジュールを合成した後に、チューニングを行う点である。出力ユニットにおける誤差は、バックプロパゲーションによりMeta-Pi ネットのみに逆伝播され、各話者用ネットのモジュールの重みは固定される。

Meta-Pi ネットワークの各モジュールは、3 つのTDNNを並列に並べた構造("Multiple TDNN")になっている。それぞれのネットワークは、異なる3つの目的関数により学習が行われ、その出力が最終的に加算により統合される("3-Way Arbitrated")[7]。その3 つの目的関数は、従来のMSE(mean-squared-error)に加え，CE(cross entropy)とCFM (classification figure-of-merit)である。Meta-pi network の n 番目の音韻に対する出力O_nを式(1)に示す。

$$O_n = \frac{1}{\mu} \sum_k \rho_{kn} M\pi_k \qquad (1)$$

$$\mu = \sum_k M\pi_k \qquad (2)$$

ここで、ρ_{kn}はk番目の話者に対するモジュールのn番目の音韻に対する出力値であり、$M_{\pi k}$はMeta-piネットのk番目の話者に対する出力値である。またμは、ネットワーク全体の出力を0から1の間に正規化するために、式(2)に示されるように、$M_{\pi k}$の和になっている。

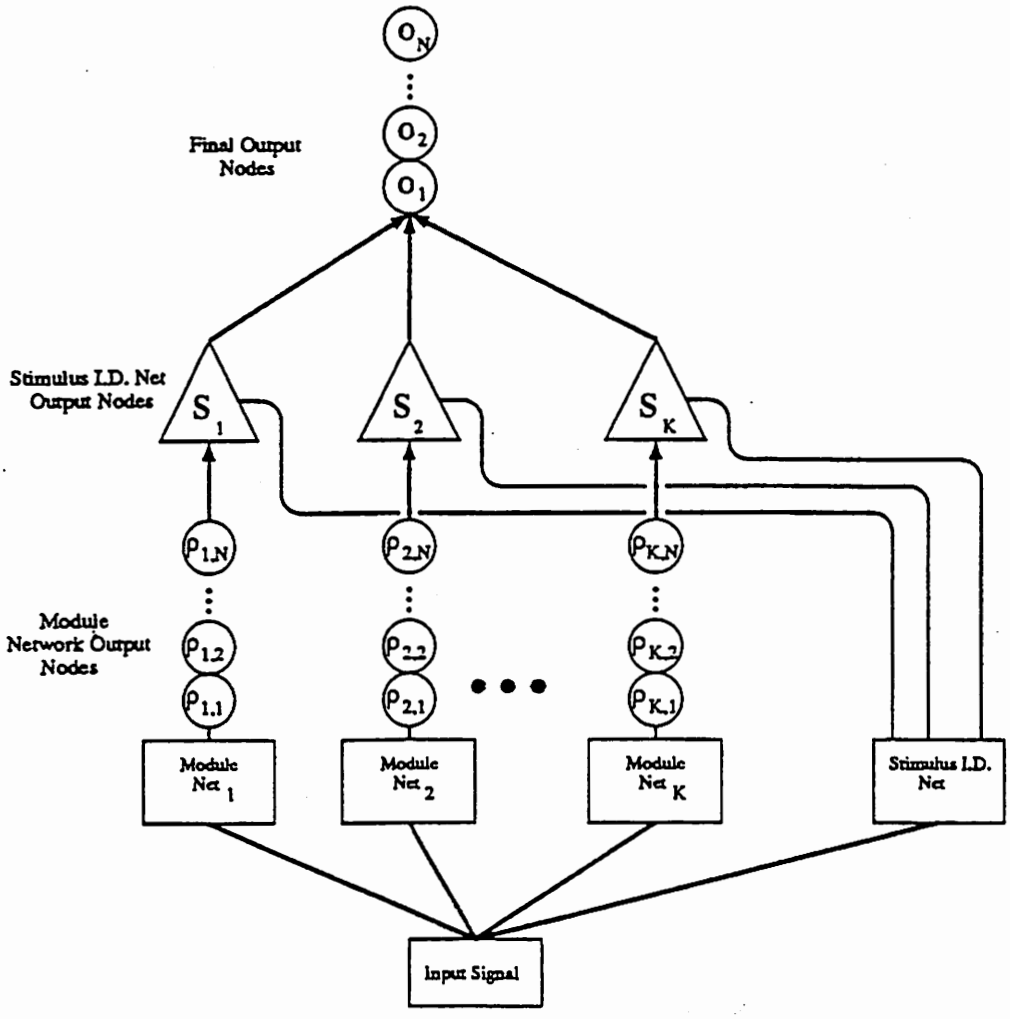

Fig. 3. SID network の構造[5]

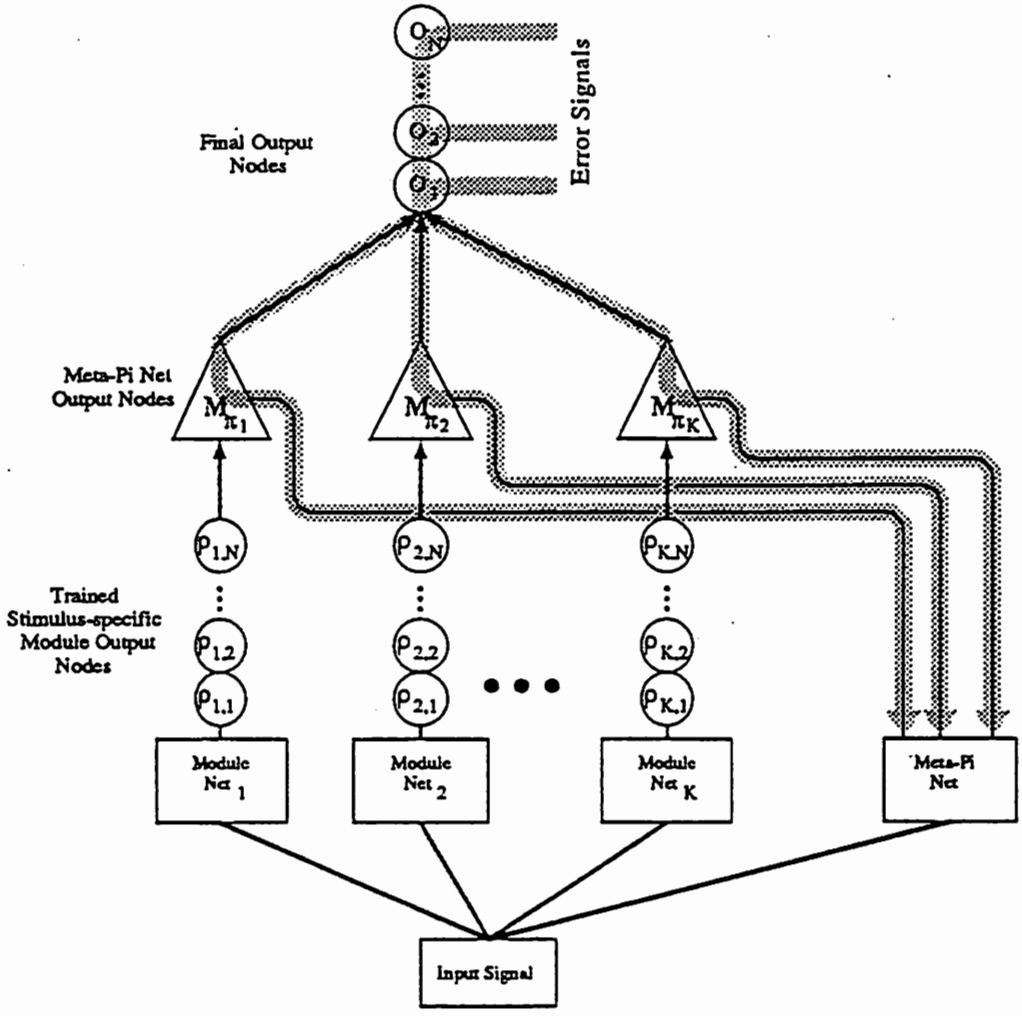

Fig. 4. Meta-Pi network の構造[5]

式(1)からも分かるようにこのネットワーク全体の出力値は、各話者のモジュールの出力値とMeta-Piネットの出力値の積で与えられる。ここで、Fig. 5で示されるような積和ユニットの一般化デルタルールを、簡単に述べる[8]。Fig. 5において、i番目とj番目のユニットの出力の積が、k番目のユニットに伝播する。そのときの、ユニットiからユニットkへの重みW_{ki}と、ユニットjから、ユニットkへの重みW_{kj}は等しく、W_{kij}で示すことにする。

積和ユニットの出力値O_kと誤差δ_kは、一般的に次式のようになる。

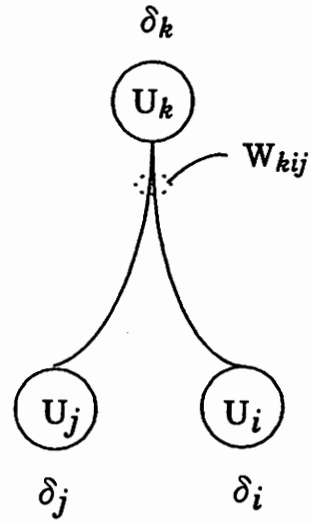

Fig. 5. 積和ユニット

$$O_k = f_k(\sum_i W_{ki} \prod_j O_{ij}) = f_k(\sum_{i,j} W_{kij} O_i O_i) \qquad (3)$$

$$\delta_i = f_k^{'}(net_i) \sum_{j,k} \delta_k W_{kij} O_j \qquad (4)$$

同様にして、Meta-Pi network における Meta-Pi net の出力ユニットの誤差信号δが計算できる。各話者に対応するMeta-Pi netのk番目の出力値を$M\pi_k$とすると、

$$\delta_k = - f_k^{'}(net_k) \frac{\partial E}{\partial M\pi_k}$$

$$= - f_k^{'}(net_k) \frac{\partial E}{\partial O} \frac{\partial O}{\partial M\pi_i}$$

$$= (1 - M\pi_k) \cdot M\pi_k \cdot \frac{1}{\mu} \sum_n \{(D_n - O_n) \cdot (\rho_{kn} - O_n)\} \qquad (5)$$

また、各話者のモジュールをチューニングした場合の各話者のモジュールの出力層に逆伝播される誤差信号は、次式のように求まる。

$$\delta_i = (1 - \rho_{kn}) \cdot \rho_{kn} \cdot \frac{M\pi_k}{\mu} (D_n - O_n) \qquad (6)$$

● Modular Speaker I.D. network

Modular Speaker I.D. network と名付たネットワークの構造は、Fig.6.に示すように、Modular TDNN に対して、話者性を付加した形になっている。この構造は、18子音認識に用いられたようなネットワークの構造[2]になっている。このネットワークとModular TDNN の認識率を比較することにより、特にマルチスピーカーの音韻認識において、話者性がどの様に利用されているかを探ることが出来るであろう。

3. 実験条件とデータ

各ネットワークの学習用と評価用の音声データベース[10]は、アナウンサー又は、ナレータの発声した重要語5240単語、男性8名、女性8名の計16名分を用いた。このうち、各話者の偶数番目の2620単語を学習用として、奇数番目の2620単語を評価用として用いた。認識実験に用いた音韻は、/b,d,g/の 3 音韻で、単語中から切り出された各音韻は、5ms毎に256ポイントのハミング窓を掛けてＦＦＴを行い、10ms毎に16次元のメルスケールのFFT出力に変換したものである。ネットワークへの入力は、この16次元のメルスケールFFT出力15フレーム分を±1.0に正規化したものである。各話者に対して、各音韻/b,d,g/の個数は、学習用データと評価用データともに、200個程度であった。

各ネットワークについて、話者6名（男性4名、女性2名）を用いて学習を行った場合と、その2倍の12人（男性6名、女性6名）を用いて学習を行った場合の Closed Speaker と Open Speaker の音韻認識率を求めた。Closed Speaker の音韻認識率は、学習に 用いた話者の評価用データを用い、各話者に対し認識率を求め、その平均値を全体の認識率とした。また、Open Speaker の認識率は、2つの実験において、同じ4名（男性2名 、女性2名）を用い、Closed Speaker の場合と同じく各話者の認識率の平均値を全体の認識率とした。　各実験に用いた具体的な話者名を Table 1 に示す。

Table 1. 各実験において用いた話者名

	学習用話者 (Closed Speaker)	不特定話者認識用話者 (Open Speaker)
6 話者学習時	MAU MHT MMS MMY FKN FSU	MNM MTK FYM FYN
12 話者学習時	MAU MHT MMS MMY MNM1 MSH FKN FSU FFS FKS FMS FTK	MNM MTK FYM FYN

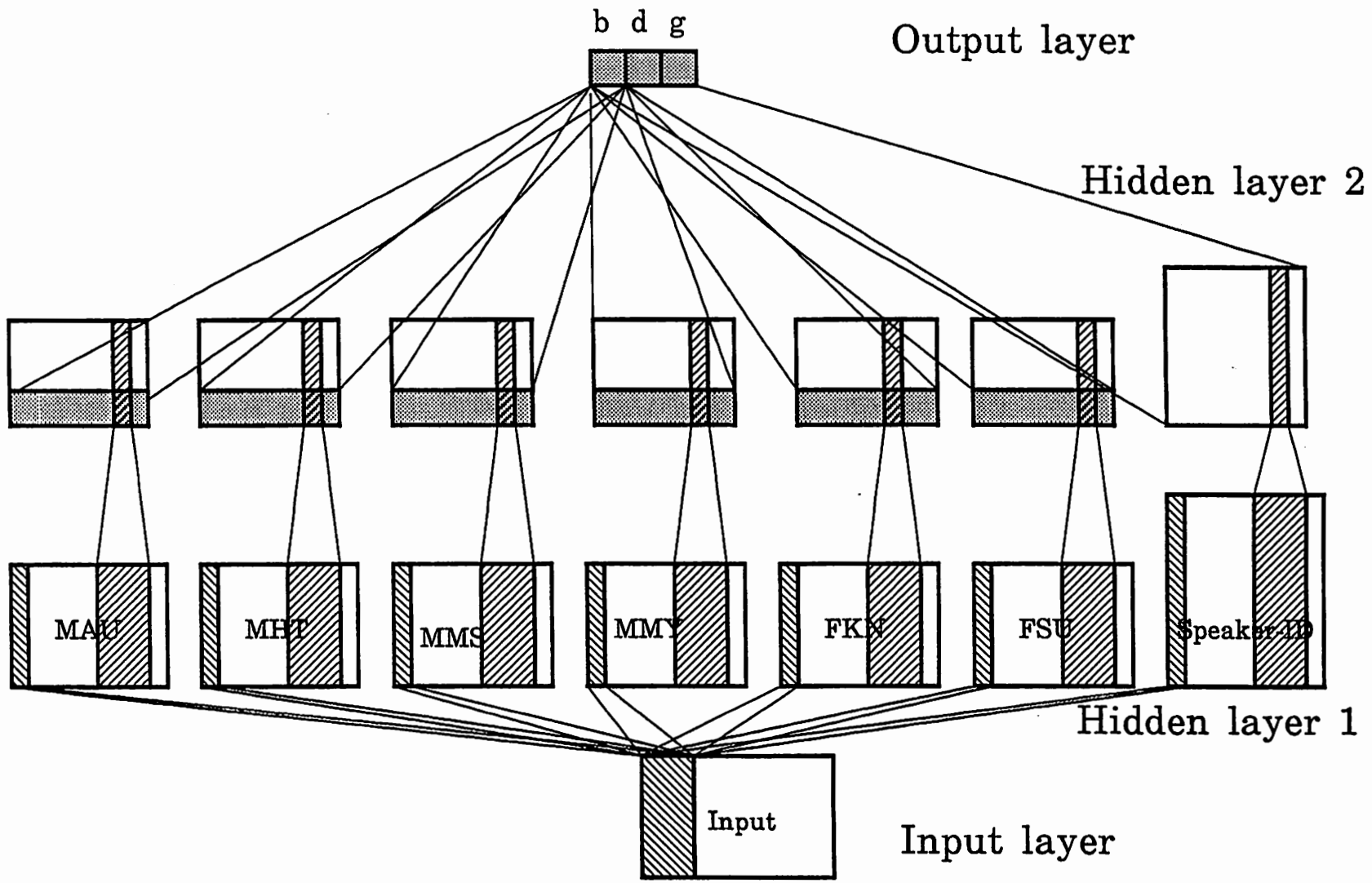

Fig. 6. Modular Speaker ID network の構造

4. 結果

Table 2. Closed Speaker の音韻認識率(%)
(6話者で学習 : MAU, MHT, MMS, MMY, FKN, FSU)

	MAU	MHT	MMS	MMY	FKN	FSU	平均 (%)
Single TDNN	91.30	96.40	91.30	94.10	93.70	92.10	93.15
Modular TDNN	97.30	99.10	96.60	97.00	97.50	98.50	97.67
SID network	95.90	98.89	96.47	96.79	94.76	95.81	96.44
Meta-Pi network	96.24	98.99	97.05	96.79	95.66	96.94	96.94
Modular Speaker I.D. network	97.30	98.90	97.20	96.80	96.50	96.30	97.17

各ネットワークに対して6話者の有声破裂音/b, d, g/の3音韻を学習させたときのClosed Speaker と Open Speaker に対する認識率をTable 2 と Table 3 に、また、12話者の3音韻を学習させたときのClosed Speaker と Open Speaker に対する認識率をTable 4 と Table 5 に示す。Table 2 と Table 4 では、共通に用いた6話者の認識率を示し、Table 2 ではClosed Speaker 6名の平均値を、Table 4 ではClosed Speaker 12名の平均値を全体の認識率として示している。Table 3 と Table 5 では、共通に用いた、4名のClosed Speaker の認識率と、その平均値を示す。

これらのTableより、ほとんどの結果においてModular TDNN と Modular Speaker I.D. network の認識率は同等であり、最も高いものであった。6話者学習時は、Mudular TDNN により、Closed Speaker に対して97.7%,Open Speaker に対して92.1%、12話者学習時には、Modular Speaker I.D. network により、Closed Speaker に対して97.3%、Open speakerに対して95.6%であった。また、学習話者を6人から12人に増したことにより、Closed Speaker の認識率とopen Speaker の認識率の差が、5.6%から1.7%に減少した。この二つのネットワークの構造的な違いは、話者認識用のサブネットが付加されているか否かという違いであるが、認識率にはそれ程差はなく、話者情報を与える有効性はあまりないといえる。

Meta-Pi network の認識率は、SID netowork をチューニングしたことによりSID netoworkの認識率を少し改善はしたものの、Modular TDNN と Modular Speaker I.D. network の認識率を下回る結果となった。文献[6]で報告さ

Table 3. Open Speaker の音韻認識率(%)
(6話者で学習 : MAU, MHT, MMS, MMY, FKN, FSU)

	MNM	MTK	FYM	FYN	平均 (%)
Single TDNN	82.40	90.80	80.60	86.50	85.07
Modular TDNN	90.70	95.20	88.10	94.20	92.05
SID network	81.30	77.30	79.84	81.35	79.95
Meta-Pi network	82.77	79.16	82.63	83.26	81.95
Modular Speaker I.D. network	88.60	93.60	84.90	91.70	89.67

れているMeta-Piの認識率は、今回の実験の認識率を上回るものであったが、それは各モジュールがMultiple TDNNとして構成されていた要因によるものが強いと考えられる。Meta-Pi networkはマルチスピーカーの音韻認識用のネットワークであるため、当然のことながらOpen Speakerに対しては認識率は低いものであった。

5. 検討

ここで、実験結果をTable 6.にまとめる。
Modular TDNNの学習は、2段階に分けて行われたものであった。つまり、はじめに各モジュールは、各話者ごとに学習が行われ、次に、それらのモジュールを並列に並べることにより、Modular TDNNを構成し、再学習を行ったわけである。このような2段階の学習を行うことにより、学習回数をある程度軽減できるという利点はあるものの、Modular TDNNのチューニングとして行われる再学習の際に、話者性が有効に利用されているのかという点は興味深い。
そこで、このように学習を2段階に分けて行った場合と、ランダム値から学習を行った場合の実験結果をTable 7に示す。

Table 7に示されるように、2つの学習法による認識率は、ともに2段階の学習を行った方が良く、その差はClosed Speakerの場合0.4%、Open Speakerの場合0.8%であった。これは、各話者用のモジュールを初期値として用いることによって、安定して収束するためといえる。この差は、学習話者や認識音韻数が増えるに従い収束性という点に顕著に現れると考えられる。また、これはModuar Speaker I.D networkにおいてもいえることである。

Table 4. Closed Speaker の音韻認識率(%)
(12話者で学習 : MAU, MHT, MMS, MMY, FKN, FSU
MNM1, MSH, FFS, FKS, FMS, FTK)

	MAU	MHT	MMS	MMY	FKN	FSU	平均 (%)
Single TDNN	—	—	—	—	—	—	—
Modular TDNN	95.70	98.70	97.50	97.40	96.80	97.60	97.09
SID network	94.53	97.47	94.33	95.35	96.03	95.81	95.39
Meta-Pi network	94.53	97.53	94.79	95.51	95.87	95.81	95.44
Modular Speaker I.D. network	96.50	98.40	96.80	98.60	97.80	97.40	97.32

それでは比較を行ったネットワークにおいて、高い認識率を得る要因は何であるのだろうか。構造とキャパシティーの要因を比較するために、次のような実験を行った。

元来のTDNNは、特定話者の音韻認識に対して、最適と考えられるキャパシティーをもつもので、そのTDNNをそのまま、前に示した6話者の Multi-Speaker の問題に、適用しても、明らかにこの問題に対してキャパシティー不足であると考えられる。そこで、基本構造とする Single TDNN のキャパシティーを増やすことにより、6人の話者を学習 し、どの程度の能力が得られるかを確かめた。hidden layer 1 のユニット数を増すこと によりSingle TDNN のキャパシティーを増した時のClosed Speaker及び Open Speakerの音韻/b,d,g/の認識率を Table 8 に示す。

Table 8. をみると、hidden layer 1のユニット数が増すにつれ、Closed Speakerと Open Speaker に対する音韻認識率はともに改善していることが分かる。この結果から、Single TDNNでも、hidden layer 1 のユニット数を増すことによりネットワークのキャパシティーを増やせば、Modular TDNN と同等の能力を得ることが出来るといえる。

以上のことから、3音韻/b, d, g/の音韻認識の認識率の面において、Modular TDNNのように話者毎のModuleによりネットワークを構成することに、有効性はあまり認められず、単にネットワーク全体のキャパシティーを増加させるはたらきをするといえる。しかしながら、学習時間、学習回数の面においては、Fig.9に示すように、モジュール化して、2段階学習を行った方が、安定して素早く極小解に収束することがわかる。但し、識別すべき音韻を増やした場合には、音韻カテゴリー毎にModule 構造をとることは有効であると報告され

Table 5. Open Speaker の音韻認識率(%)
(12話者で学習 : MAU, MHT, MMS, MMY, FKN, FSU
MNM1, MSH, FFS, FKS, FMS, FTK)

	MNM	MTK	FYM	FYN	平均 (%)
Single TDNN	—	—	—	—	—
Modular TDNN	98.30	95.10	91.40	97.20	95.50
SID network	81.45	85.08	84.44	92.32	85.82
Meta-Pi network	81.30	85.40	85.08	91.69	85.87
Modular Speaker I.D. network	97.40	93.70	93.30	97.80	95.55

ているように、学習話者数や認識音韻数が増すにつれ、2段階学習やModule構造は学習の収束性や安定化という点においては有効であろう。

一般に限られた学習データを用いてネットワークの学習を行った場合、ネットワークのキャパシティーが学習データに対して過剰であると、学習データに対してオーバーチューニングを起こし、未学習データに対する認識率が低下する。今回の実験でSingle TDNN のキャパシティーを過剰と思われる程度まで増やして認識実験を繰り返し行ったわけであるが、飽和状態には近付いたもののオーバーチューニングの現象はみられなっかた。これは学習データの数が評価用のデータの特徴、つまりは、有声破裂音/b, d, g/の特徴をとらえるのにある程度十分であるためと考えられる。極論すれば、無限のキャパシティーをもつネットワークを用いて、無限のデータを学習すれば限りなく高い認識率を得ることが出来るといえるが、現実問題として不可能である。データ数を増やすことにより、認識率が改善されることは、Modular TDNN の6話者と12話者の実験結果を比較することによりあきらかである。同様に、ネットワークのキャパシティーを増加させることにより認識率が改善されることは、Table 7.が示している。逆にこのことは、多次元空間における音韻の分布は複雑に入り組んで存在していることを示している。

Table 6. 各ネットワークの6,12話者学習時のClosed, Open Speakerに対する音韻/b, d, g/の認識率(%)

Architecture	Number of Connections 学習話者: 上段: 6話者 --- 下段: 12話者	Recognition Rates (%)			
		6 Training Speakers		12 Training Speaker	
		Closed	Open	Closed	Open
Single TDNN	6,233	93.2	85.1	—	—
Modular TDNN	37,383 74,763	97.7	92.1	97.1	95.5
SID (Stimulus I.D.network)	48,468 98,636	96.4	80.0	95.4	85.8
Meta-pi network	48,468 98,636	96.9	82.0	95.4	85.9
Modular Speaker I.D. TDNN	48,483 98,564	97.2	89.7	97.3	95.6

6. 結論

TDNNを基本構造とする特定話者及びマルチスピーカーの音韻認識を行うネットワークを、不特定話者の音韻認識/b, d, g/タスクに適用し，その能力を比較検討した。

不特定話者の3音韻/b, d, g/の認識実験において最も認識率の高いものは、12話者学習時のModular Speaker I.D. TDNNの結果で、95.6%と高いものであった。今回の実験において、特定話者の認識率は平均で97.9%、Closed speakerに対しては97.7%であり、不特定話者の認識率は特定話者の認識率にその差2.3%に迫る結果となった。また、学習データ数及びネットワークのキャパシティーを2倍程度に増やすことにより、不特定話者の誤り率は7.9%から4.4%に減少した。

Modular TDNN のように二段階学習を行うことは、学習回数の軽減という面において有効であり、同じ程度のキャパシティーをもつSingle TDNNと比較すると、認識率は若干上回り、限られたキャパシティーを有効に利用していることが分かった。また、Single TDNNでも、ある程度十分なデータを用いればキャパシティーを増すことによって、高い認識率が得られることが分かった。

Table 7. Modular TDNN の学習を2段階に分けた場合と、ランダム値から行った場合の、3音韻/b, d, g/の認識率(%)

Learning Method	Recognition Rates (%)	
	6 training speakers	
	Closed	Open
2 Stage Training	97.7	92.1
Learn From Random Weight	97.3	91.3

Table 8. Single TDNN のHidden layer 1 のユニット数を可変にしたときの音韻/b, d, g/ の認識率(%)

Unit number of Hidden layer 1	Number of Connections	Recognition Rates (%)	
		Closed (6名平均)	Open (4名平均)
8	6233	93.2	85.1
16	12409	94.6	86.3
24	18585	95.4	89.2
32	24761	95.9	91.2
40	30937	97.4	91.1
48	37113	97.4	91.5
56	43289	97.6	93.8
64	49465	—	—

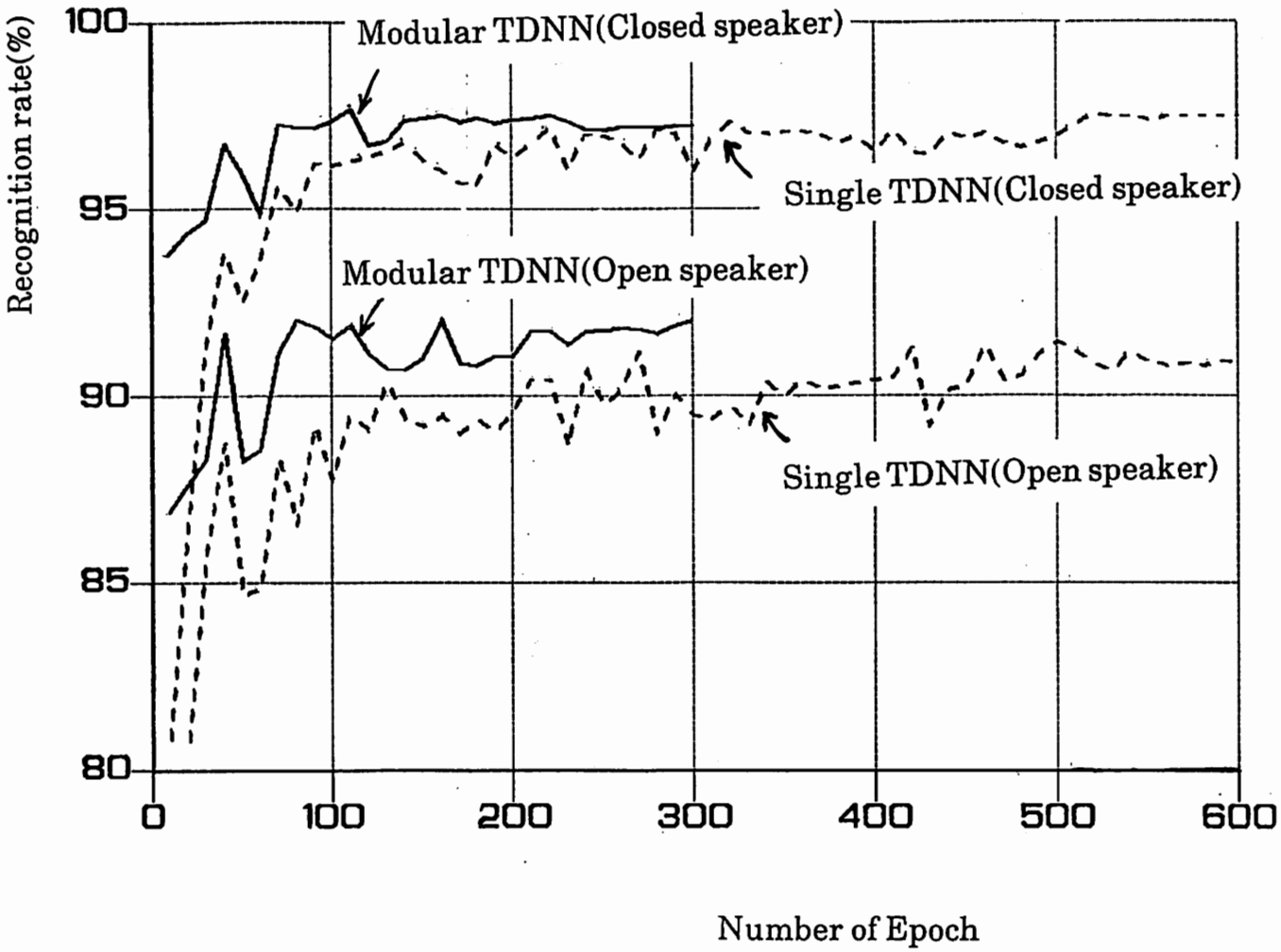

Fig.7 学習回数に対するClosed SpeakerとOpen Speakerの認識率(%)
6話者学習時のModular TDNN(実線)とSingle TDNN(点線)

今後の課題

今回の実験では、3音韻の認識実験を行ったが、識別すべき音韻数を18、24と増やした場合に、各ネットワーク能力がどのようになるか、また、そのときの認識率と、HMMなどによる認識率と比較検討を行うことにより、TDNNの能力を明らかにする必要があるであろう。

謝辞

研究の機会を与えて下さったＡＴＲ自動翻訳電話研究所の樽松明社長に深く感謝いたします。また、直接御指導して下さいました沢井秀文氏、ならびに、音声情報処理研究室、データ処理研究室の皆様に感謝致します。

参考文献

[1] A. Waibel, T Hanazawa, G. Hinton, K. Shikano, and K.J. Lang, "Phoneme Recognition Using Time-Delay Neural Networks", IEEE Trans. Acoustics, Speech and Signal Processing, March 1989.

[2] Waibel, A., Sawai, H., Shikano, K., "Consonant Recognition by Modular Construction of Large Phonemic Time-Delay Neural Networks", Proceedings of the 1989 IEEE International Conference on Acoustics, Speech and Signal Processing, May, 1989, pp.112-115, May, 1989.

[3] Matsuoka, T., Hamada, H. and Nakatsu, R., "Syllable Recognition Using Integrated Neural Networks", Int. Joint Conf. on Neural Networks, Proceeding of IJCNN-89, vol. 2, pp251-258, June 1989.

[4] Hampshire, J., Waibel, A., "The Meta-Pi Network: Connectionist Rapid Adaptation for High-Performance Multi-Speaker Phoneme Recognition", Proceedings of the 1990 IEEE International Conference on Acoustics, Speech and Signal Processing, S3.9, pp164-168 1990.

[5] Hampshire, J., Waibel, A., "The Meta-Pi Network: Building Distributed Knowledge Representations for Robust Pattern Recognition", Carnegie Mellon University Technical Report CMU-CS-89-166, August, 1989.

[6] Hampshire, J., Waibel, A., "Conectionist Architectures for Multi-Speaker Phoneme Recognition", Carnegie Mellon University Technical Report CMU-CS-89-167, August, 1989.

[7] Hampshire, J., Waibel, A., "A Novel Objective Function for Improved Phoneme Recognition Using Time-Delay Neural Networks", Carnegie Mellon University Technical Report CMU-CS-89-118, March, 1989.

[8] Rumelhart, D. E., McClelland, J. L., et al., "Parallel Distributed Processing", vol. 1. Cambridge, MA.: MIT Press, 1987.

[9] Sawai, H., Waible, A., et al, "Parallelism, Hierarchy, Scaling in Time-Delay Neural Networks for Spotting Japanese Phonemes/CV-Syllables". Int. Joint Conf. on Neural Networks, Proceeding of IJCNN-89, vol.2, pp81-88, June 1989.

[10] 武田一哉、勾坂芳典、片桐滋、桑原尚夫、"研究用日本語音声データベースの構築"、音響学会誌、44巻10号、pp747-754、1988.