TR-I-0145

TDNNの構造の音韻認識率、シフトイバリアント性への影響

南 泰浩† 沢井 秀文
Yasuhiro MINAMI † Hidefumi SAWAI

1990.2

概要

TDNNを連続音声認識に用いる場合、TDNNの連続音声に対する音韻認識率の向上が非常に重要となってきた。本報告ではTDNNの連続音声に対する音韻認識率をHMMと比較し、TDNNの連続音声に対する性能せを調べた。この結果、TDNNはカテゴリ数の大きな連続音声の認識に対して、HMMに比べて弱いことが確認された。本報告ではさらに、TDNNの構造の変化による音韻認識率とシフトインバリアント性の比較を行い、TDNNの連続音声に対する音韻認識率の向上の可能性を示す。

本報告は、学外実習生南泰浩(慶応大学)が行った実習の報告書である。

ATR Interpreting Telephony Research Laboratories
ATR 自動翻訳電話研究所

† Keio Univ.
† 慶応義塾大学

1．　はじめに

ＡＴＲでは従来より、ＴＤＮＮ（Ｔｉｍｅ-Ｄｅｌａｙ Ｎｅｕｒａｌ Ｎｅｔｗｏｒｋｓ）を用いた音声認識を行ってきた[1]。ＴＤＮＮは特定話者の／ｂｄｇ／のタスクに対して９８．６％と高精度の音韻認識率を達成することが確認された[1]。しかし、ＴＤＮＮを用いた連続音声認識システムを構築する過程で様々な問題が生じてきた。我々が構成したシステムは従来ＡＴＲにおいて構成した連続音声認識用システムのＨＭＭ（Ｈｉｄｄｅｎ Ｍａｒｋｏｖ Ｍｏｄｅｌ）－ＬＲの音声処理部（ＨＭＭ部）をＴＤＮＮの処理部に置き換えたものである[2]。このシステムを用いて連続音声認識を行った結果、ＨＭＭ－ＬＲに比べ文節認識率が低いことが確認された。この原因はＴＤＮＮの連続音声に対する音韻認識率の低下であると考えられる。

そこで本報告では２章でＴＤＮＮの文節発声データに対する音韻認識率をＨＭＭと比較することによって、ＴＤＮＮの連続音声に対する性能を調べる。そして、３章、４章でＴＤＮＮの構造による音韻認識や、シフトインバリアント性を調べ、ＴＤＮＮの改良の可能性を調べる。

2．　ＴＤＮＮとＨＭＭの音韻認識率の比較

まず、ＴＤＮＮが連続音声に対してどのような音韻認識率を示すかをＨＭＭと比較する。以下に実験の条件を示す。３．章以降で述べる実験結果はすべてこの章で述べる実験条件と同じものを使用した。

以下に比較実験を行うタスク、認識に用いた音韻、学習に用いた音韻について述べる。

2．1　比較対象タスク

比較に用いたタスクは以下の小カテゴリ（3カテゴリ）と中カテゴリ（6カテゴリ）と大カテゴリ（18カテゴリ）の3種類である。

（1）／b d g／

（2）／b d g m n N／

（3）全子音（18子音）

18子音は／b／,／d／,／g／,／p／,／t／,／k／,／m／,／n／,／N／,／s／,／sh／,／h／,／z／,／ch／,／ts／,／r／,／w／,／y／であり音韻区間は以下の条件を満たすもののみを用いた。

（1）／ch／は／chi／のみから切り出す。

（2）／sh／は／shi／のみから切り出す。

（3）／z／は／za／,／ze／,／zu／,／zo／のみから切り出す。

（4）／h／は／ha／,／hi／,／he／,／ho／のみから切り出す。

（5）その他はラベルに従いその音韻の存在する区間を切り出す。

以上のラベルの表記はＡＴＲデータベースの表記法(3)に従い、音韻はこのラベルにより切り出したものである。

2．2　認識対象音韻

認識データとしては発声様式の異なった以下の4種類のデータから切り出した音韻を用いる。認識に用いた音韻の数は全ての手法で同じ数とした。

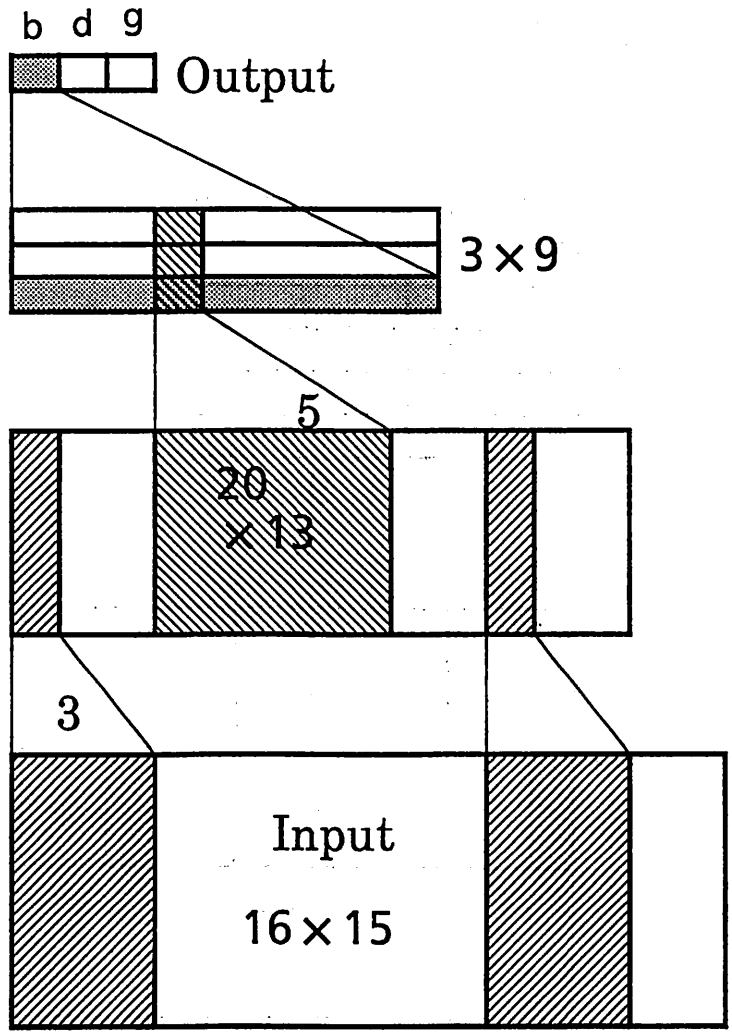

図1 /bdg/用TDNN

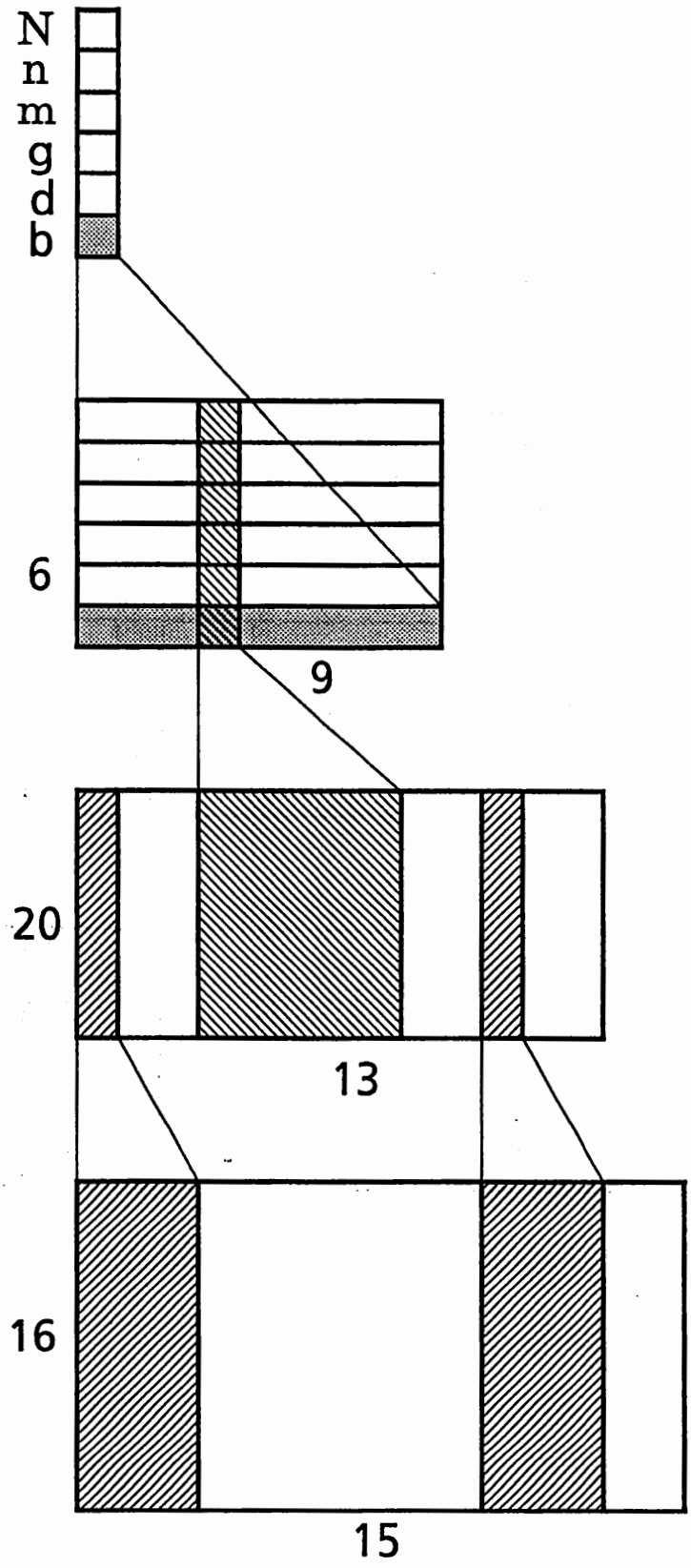

図2 /bdgmnN/用TDNN

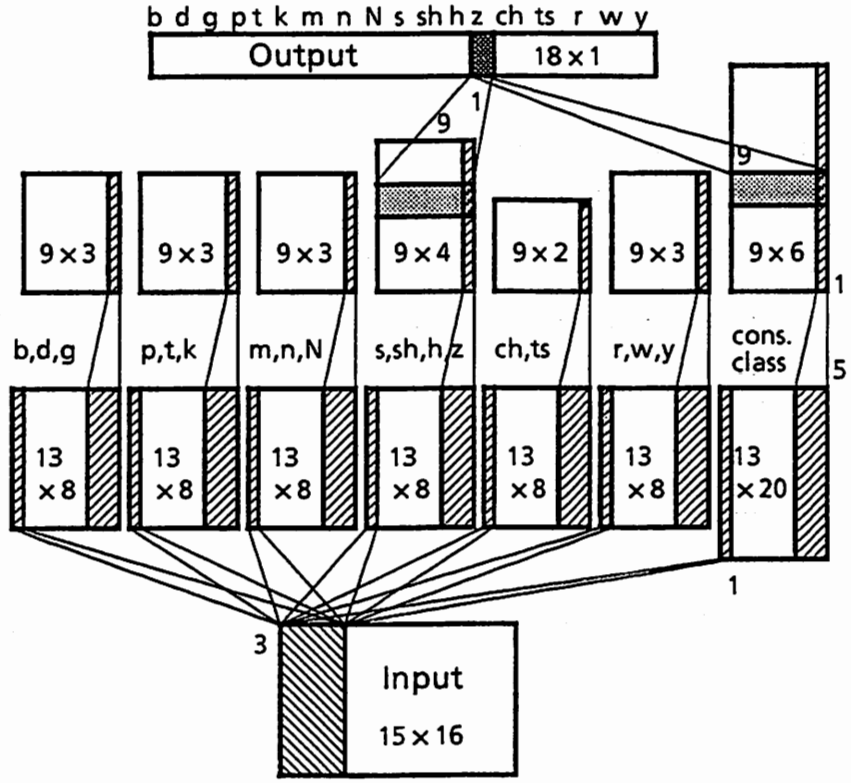

図3　18子音用TDNN

（１）５２４０単語の奇数番目のデータ（以下単語発声とする）（５．６８モーラ／ｓｅｃ）
（２）複合語を許さない文節発声（以下短い文節発声とする）（７．１４モーラ／ｓｅｃ）
（３）文節発声（７．７２モーラ／ｓｅｃ）
（４）句切り指定を行わない発声（以下自由発声とする）（９．５６モーラ／ｓｅｃ）

２．３　学習音韻

学習音韻は重要５２４０単語の偶数番目のデータから切り出した音韻を用いた。単語発声から切り出したデータだけで、どれだけ連続音声認識に対して有効であるかを調べるため学習データには文節発声や自由発声データを用いていない。

２．４　発声話者

発声はアナウンサ１名の発声した音声で特定話者の認識を行った。

２．５　認識率の計算

ここで求めた認識率は各音韻毎に認識率を求め、その値を各音韻の個数に対応する重みを掛けて総和したものを全体の認識率とした。

２．６　ＴＤＮＮ

ＴＤＮＮを用いた／ｂｄｇ／を認識するアーキテク

チャは図1のようになる[1]。また、／ｂｄｇｍｎＮ／を識別するアーキテクチャを図2に示す(1)。

１８子音に対しては図3に示すアーキテクチャを用いる[4]。これは図1に示したネットワークを多数カテゴリを識別するように構成したものである。このアーキテクチャは１８音韻（１８子音／ｂ／,／ｄ／,／ｇ／,／ｐ／,／ｔ／,／ｋ／,／ｍ／,／ｎ／,／Ｎ／,／ｓ／,／ｓｈ／,／ｈ／,／ｚ／,／ｃｈ／,／ｔｓ／,／ｒ／,／ｗ／,／ｙ／）を識別する。このＴＤＮＮは６個の子音グループ内を識別するサブネットワーク”ｂｄｇ”,”ｐｔｋ”,”ｍｎＮ”,”ｓｓｈｈｚ”,”ｃｈｔｓ”,”ｒｗｙ”と子音グループ間を識別するネットワーク”ｃ－ｃｌａｓｓ”のモジュールで構成される。以上で述べた3つのアーキテクチャをＢａｃｋ－Ｐｒｏｐａｇａｔｉｏｎ学習則[5]を用いて学習することにより、他カテゴリの出力値を考慮（側抑制）するようなネットワークを構成することができる。この際、同じ値を持つように設定された重みは、この設定を崩さないように学習される。１８子音の学習は各モジュールに分割することなく、一括に学習した。

２．７　ＴＤＮＮにおける分析

ＴＤＮＮで用いる1フレームの音声データは１２ＫＨｚでサンプリングし、２５６ポイントのハミング窓で、ＦＦＴを計算した後、１０ｍｓ毎に１６次のメル尺度のフィルタバンクを通し、１５フレーム内で平均０、±１に正規化したものである。

２．８　複数コードブックを用いるＨＭＭ

本実験ではＨＭＭとして複数のコードブックを利用するＨＭＭを用いる。

ここで用いたパラメータは以下の３種類出ある。

（１）スペクトル　ＷＬＲ距離尺度による２５６個のベクトルからなるコードブック。

（２）パワー　ＰＷＬＲのパワー項の距離尺度による６４個のスカラからなるコードブック。

（３）スペクトルの動的特徴　ＬＰＣケプストラム係数の１次から１６次までの時間方向の線形回帰係数、のユークリッド距離による２５６個のベクトルからなるコードブック。

２．９　ＨＭＭにおける分析

分析は音声データを１２ＫＨｚでサンプリング後、窓長２１．３ｍｓのハミング窓で、フレーム周期３ｍｓごとに切り出し、高域強調後１２次のＬＰＣ分析を行った。

２．１０　比較結果

表１から表１２に比較結果を示す。

２．１１　比較実験に対する検討

ＨＭＭとＴＤＮＮでは大カテゴリの識別を行ったときに、音韻認識率の差が大きく現れる。この中で最も特徴的なのはＴＤＮＮが自由発声に発声様式が近づくにつれて認識率が極端に低下することである。

表1　単語発声データに対する累積音韻認識率(/bdg/)

順位	1	2	3
HMM	95.6%	99.4%	100%
TDNN	98.8%	99.9%	100%

表2　文節発声データに対する累積音韻認識率(/bdg/)

順位	1	2	3
HMM	86.8%	97.8%	100%
TDNN	90.8%	94.0%	100%

表3　短い文節発声データに対する累積音韻認識率(/bdg/)

順位	1	2	3
HMM	87.5%	96.8%	100%
TDNN	91.3%	95.0%	100%

表4　自由発声データに対する累積音韻認識率(/bdg/)

順位	1	2	3
HMM	80.8%	92.8%	100%
TDNN	85.3%	91.8%	100%

表5　単語発声データに対する累積音韻認識率(/bdgmnN/)

順位	1	2	3	4	5
HMM	89.5%	98.3%	99.5%	99.8%	99.9%
TDNN	95.1%	98.6%	99.7%	99.8%	100%

表6　文節発声データに対する累積音韻認識率(/bdgmnN/)

順位	1	2	3	4	5
HMM	73.7%	90.1%	97.0%	98.4%	99.6%
TDNN	72.5%	88.8%	94.6%	98.0%	99.2%

表7　短い文節発声データに対する累積音韻認識率(/bdg/mnN)

順位	1	2	3	4	5
HMM	75.4%	89.6%	95.5%	98.3%	99.3%
TDNN	75.7%	89.3%	95.2%	97.2%	99.1%

表8　自由発声データに対する累積音韻認識率(/bdgmnN/)

順位	1	2	3	4	5
HMM	64.6%	83.5%	91.8%	96.6%	99.4%
TDNN	64.7%	83.2%	90.4%	94.5%	97.9%

表9　単語発声データに対する累積音韻認識率(18子音)

順位	1	2	3	4	5
HMM	93.1%	98.4%	99.5%	99.7%	99.8%
TDNN	96.2%	98.9%	99.6%	99.8%	99.9%

表10　文節発声データに対する累積音韻認識率(18子音)

順位	1	2	3	4	5
HMM	79.8%	92.2%	96.6%	97.9%	99.0%
TDNN	72.4%	84.3%	90.5%	93.1%	94.9%

表11　短い文節発声データに対する累積音韻認識率(18子音)

順位	1	2	2	3	2
HMM	81.4%	91.7%	95.9%	98.0%	98.8%
TDNN	76.2%	86.3%	91.5%	94.7%	96.6%

表12 自由発声データに対する累積音韻認識率(18子音)

順位	1	2	2	3	2
HMM	71.6%	86.2%	92.4%	95.3%	97.2%
TDNN	56.6%	70.9%	78.5%	84.3%	87.8%

3．　ネットワークの構造による認識率の違い（18子音用）

ネットワークの構造によってどのように認識率が変化するかを調べるために以下の基本的な2種類の実験を行った。ここで用いた音韻は2．で求めたものと同じものを用いた。

3．1　ネットワークのモジュール構成

TDNNとしてどのような構成がよいのかを調べるために、以下の2つのTDNNを用いて実験を行った。

（1）分散型TDNN

このネットワークを図4に示す。

（2）集中型TDNN

このネットワークを図5に示す。

表13に図3の従来型のTDNNの各発声様式に対する累積音韻認識率を示す。以上の2つのTDNNの各発声様式に対する累積認識率を表14と表15に示す。学習にはDYNET[6]を用いた。

● 検討

分散型も集中型も図3のTDNNに比べ認識率が低い。これは図3のクラス分けを行うネットワークが有効なのではないかと考えられる。

3．2　出力層と中間層の間の接続法

出力層と中間層の接続がどのようにシフトインバリアントに影響を及ぼすかを見極めるため出力層と中間

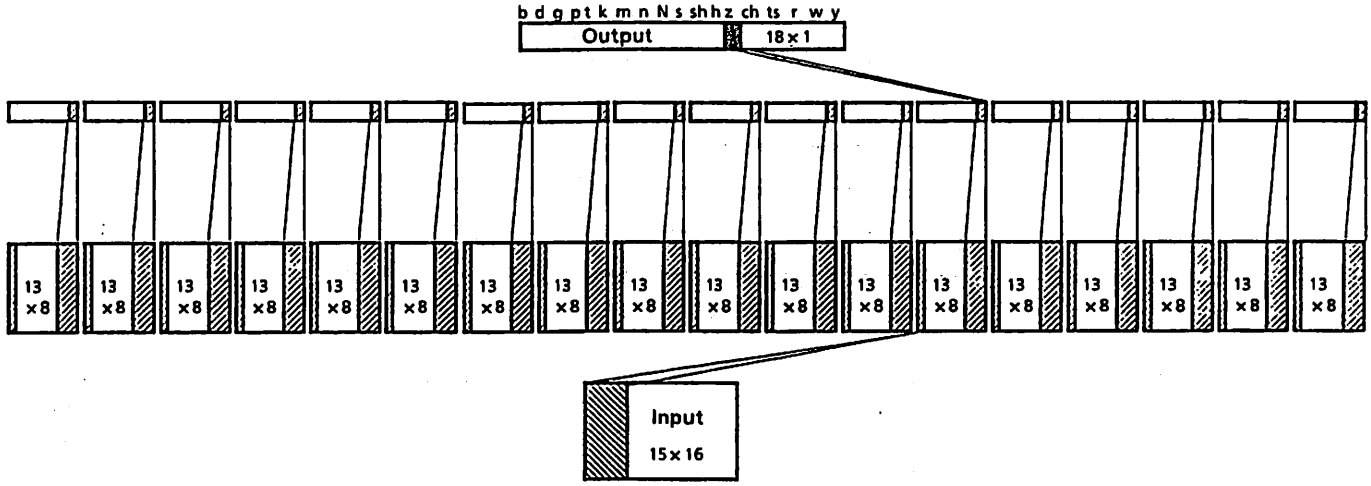

図 4　分散型18子音用TDNN

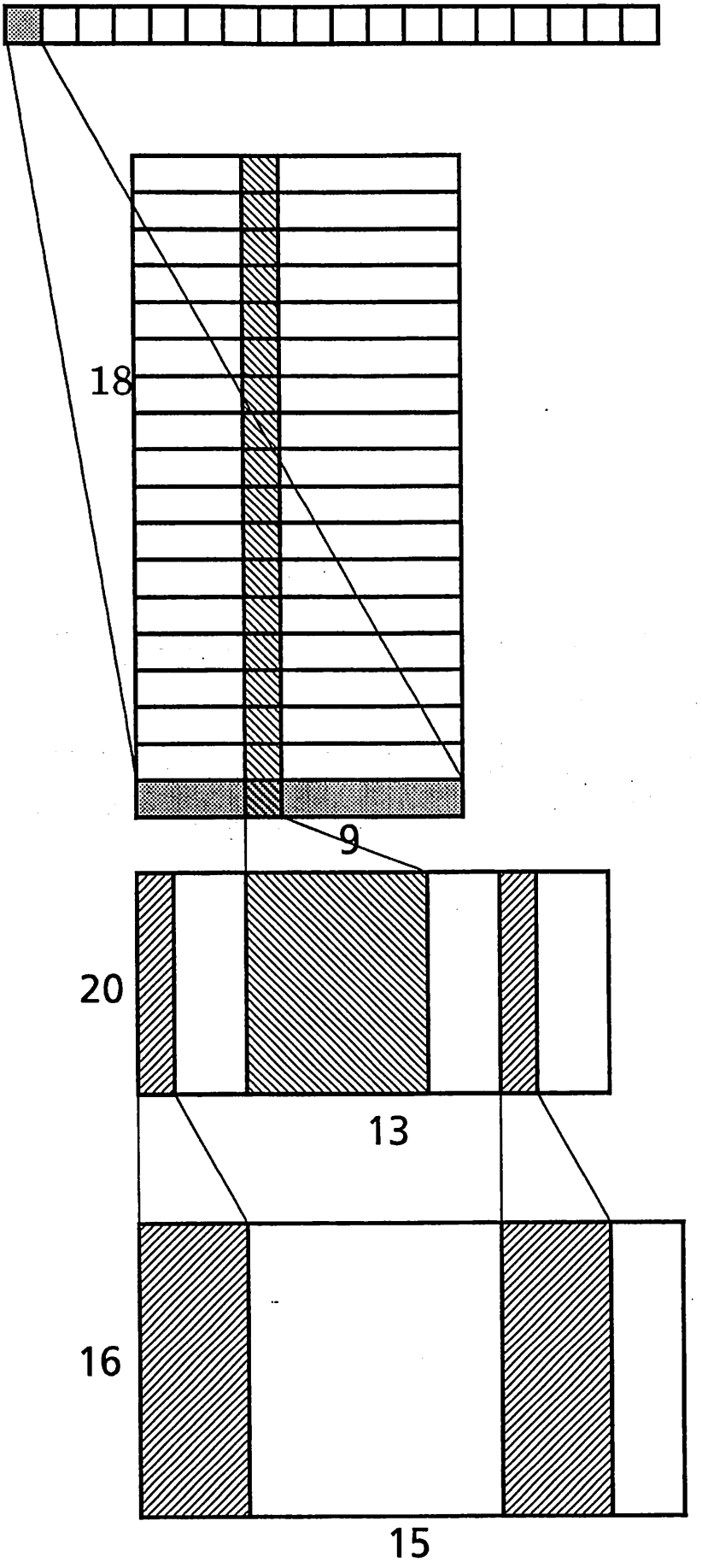

図5 集中型TD18子音用TDNN

表13 図3のTDNNの累積認識率

順位	1	2	3	4	5
単語発声	96.2%	98.9%	99.6%	99.8%	99.9%
文節発声	72.4%	84.3%	90.5%	93.1%	94.9%
短い文節発声	76.2%	86.3%	91.5%	94.7%	96.6%
自由発声	56.6%	70.9%	78.5%	84.3%	87.8%

表14 分散型TDNNの累積認識率

順位	1	2	3	4	5
単語発声	96.9%	99.2%	99.6%	99.8%	99.9%
文節発声	66.4%	82.3%	88.4%	91.5%	93.7%
短い文節発声	69.7%	83.8%	89.5%	92.8%	94.5%
自由発声	49.9%	67.9%	77.4%	83.0%	86.9%

表15　集約型TDNNの累積認識率

順位	1	2	3	4	5
単語発声	95.5%	98.7%	99.3%	99.6%	99.8%
文節発声					
短い文節発声	71.5%	84.5%	89.1%	92.3%	94.3%
自由発声	50.0%	65.9%	%	%	%

層との間の接続法を変化させたものを用いて、シフトインバリアントの効果を調べた。ここでは3．2．1，3．2．2に示す2種類の接続方法に対して評価を行った。この2つ接続法に対するシフトインバリアント性を図3のTDNNと比較するために、図3のTDNNのシフトインバリアント性について表16に示す。この表はTDNNの入力を10ms（1フレーム）毎にシフトし、結果を表にしたものである。認識に用いた音韻の発声様式は単語発声である。

3．2．1　自由な接続

図3に示すTDNNの中間層と出力層との接続はカテゴリ毎に全て同じ重みでつながっていた。ここではこの重みを全て自由にして、音韻の認識率とシフトインバリアント性を調べた。表17，18に結果を示す。表17は各発声様式に対する音韻認識率を示し、表18は単語発声した音韻のシフトに対する認識率を示す。この学習にはDYNETを用いた。

● 検討

この表17と表13を比較するとかなりの音韻認識率の向上がみられる。しかし、表18と表16を比較するとシフトインバ性に対しては従来のTDNNの方がよいということが分かる。このことから従来のTDNN出力層と中間層の間の結合法はシフトインバリアント性には有効であると考えられる。

3．3　2　分割接続

前節では全ての重みを自由にしていたが、シフトイ

表16 図のTDNNのシフトインバリアント性能(単語発声)

シフト幅(ms)	1	2	3	4	5
-30	61.2%	79.3%	88.0%	92.5%	95.1%
-20	82.8%	93.3%	96.7%	98.4%	99.2%
-10	93.2%	97.9%	99.2%	99.6%	99.8%
0	96.2%	98.9%	99.6%	99.8%	99.9%
10	93.0%	98.5%	99.4%	99.8%	99.8%
20	78.5%	92.0%	96.8%	98.7%	99.3%
30	60.0%	78.1%	87.2%	91.8%	94.3%

表17 出力層と中間層間の重みを自由にしたTDNNの累積認識率

順位	1	2	3	4	5
単語発声	95.7%	98.9%	99.6%	99.8%	99.8%
文節発声	78.7%	90.4%	94.1%	96.2%	97.2%
短い文節発声	82.2%	92.1%	94.9%	96.4%	97.7%
自由発声	65.2%	78.6%	84.3%	88.3%	90.9%

表18 出力層と中間層間の重みを自由にしたTDNNの
シフトインバリアント性能(単語発声)

シフト幅(ms)	1	2	3	4	5
-30	33.5%	47.4%	56.4%	63.6%	69.5%
-20	52.9%	68.4%	76.4%	81.5%	85.5%
-10	84.7%	94.1%	96.7%	98.0%	98.8%
0	96.7%	99.1%	99.7%	99.8%	99.9%
10	78.6%	88.5%	92.8%	95.0%	96.6%
20	44.8%	57.3%	63.9%	68.8%	73.0%
30	28.9%	42.4%	48.5%	53.2%	58.2%

ンバリアント性を高めるため、ここではこの重みを前部４個と後部の５個とに二分割、前部の重みは全て同じ値とし、後部の重みも全て同じ重みとした。表１９，２０に結果を示す。表１９は各発声様式に対する音韻認識率を示し、表２０は単語発声の音韻の位置ずれ（シフト）に対する認識率を示す。この実験の学習にはＤＹＮＥＴを用いた。

● 検討

表１９は表１７と表１３からこの方式が従来型のＴＤＮＮと出力層と中間層の間の重みを自由にしたＴＤＮＮの中間程度の認識率を示すことが分かる。表２０は表１８と表１６からシフトインバリアント性についいてもこの手法が２つのＴＤＮＮの間にあることが確認された。

４．　ネットワークの構造による認識率の違い（／ｂｄｇｍｎＮ／）

１８子音の学習は非常に多くの時間を必要とするのでここでは、６音韻（／ｂｄｇｍｎＮ／）のＴＤＮＮを用いる。本章では、３．章で述べた手法を考慮にいれて、様々な手法を提案し、その効果を調べる。ここで各手法に入る前に図２のＴＤＮＮの各発声様式に対する音韻認識率と、シフトインバリアント性について表２１と表２２に示す。ここで用いた音韻は２．で求めたものと同じものを用いた。この実験の学習にはＤＹＮＥＴを用いた。

４．１　周波数方向の重みをぼかす手法

表19 出力層と中間層間の重みを中心で2つに分けたTDNNの累積認識率

順位	1	2	3	4	5
単語発声	96.7%	99.1%	99.7%	99.8%	99.9%
文節発声	77.9%	90.5%	94.3%	96.0%	96.9%
短い文節発声	80.8%	91.1%	95.0%	96.7%	98.2%
自由発声	62.0%	79.2%	85.2%	89.1%	92.0%

表20 出力層と中間層間の重みを中心で2つに分けたTDNNの
シフトインバリアント性能(単語発声)

シフト幅(ms)	1	2	3	4	5
-30	46.2%	63.6%	73.6%	79.9%	84.2%
-20	75.7%	87.7%	92.4%	94.9%	96.4%
-10	93.0%	98.0%	99.1%	99.5%	99.7%
0	95.7%	98.9%	99.6%	99.8%	99.8%
10	92.0%	97.4%	98.7%	99.3%	99.7%
20	68.8%	81.5%	87.8%	91.9%	94.5%
30	38.5%	51.2%	60.4%	67.9%	73.7%

表21　図2のTDNNの累積認識率(/bdgmnN/)

順位	1	2	3
文節発声	72.5%	88.8%	94.6%
短い文節発声	75.7%	89.3%	95.2%
自由発声	64.7%	83.2%	90.4%

表22　図2のTDNNの
シフトインバリアント性能(/bdgmnN/)(短い音節発声)

シフト幅(ms)	1	2	3
-30	44.9%	66.8%	86.7%
-20	60.3%	82.8%	92.6%
-10	72.5%	88.5%	94.7%
0	75.7%	89.3%	95.2%
10	74.1%	88.7%	94.4%
20	69.2%	86.1%	93.2%
30	56.6%	79.0%	89.7%

重みの周波数方向に対する敏感性を軽減するために重みを学習させるときに周波数方向の隣の重みの変分量をこの重みに３０％程度加えて学習させた。この効果によって周波数方向に対する位置ずれを吸収できると考えた。この結果を表２３に示す。この実験の学習にはＤＣＰを用いた。

● 検討

表２３の結果と表２１の結果を比較すると、この手法があまり有効でないことが分かる。原因としては、周波数領域がすでに１６次元に情報が落とされてぼけているなどが挙げられる。

４．２　出力層と中間層の間の重みを自由にする手法

３．２で行ったように／ｂｄｇｍｎＮ／の識別に対しても出力層と中間層の接続がどのようにシフトインバリアントに影響を及ぼすかを見極めるため出力層と中間層との間の接続を自由にしたものを用いて音韻の認識と、シフトインバリアントの効果を調べた。この結果を表２４、表２５に示す。表２４は各発声様式による、音韻認識率である。表２５は各シフト（ｍｓ）に対する音韻認識率である。この実験の学習にはＤＹＮＥＴを用いた。

● 検討

表２４、表２５と表２１と表２２と比較すると次のことが分かる。１８子音と同様この場合も音韻認識率はかなりよい値を示しているが、シフトインバリアント性はかなり悪化する。出力層と中間層の間の重みがシフトインバリアントに重要であることが確認された。

表23 周波数方向に重みをぼかす
TDNNの累積認識率(/bdgmnN/)

順位	1	2	3
文節発声	69.8%	87.4%	94.3%
短い文節発声	70.2%	85.7%	92.2%
自由発声	59.9%	82.0%	91.6%

表24 出力層と中間層間の重みを自由にした
TDNNの累積認識率(/bdgmnN/)

順位	1	2	3
文節発声	85.1%	94.4%	96.9%
短い文節発声	83.1%	93.6%	97.0%
自由発声	80.8%	93.1%	97.5%

表25 出力層と中間層間の重みを自由にしたTDNNの
シフトインバリアント性能(/bdgmnN/)(短い音節発声)

シフト幅(ms)	1	2	3
-30	19.7%	34.4%	43.7%
-20	33.7%	55.7%	67.0%
-10	67.6%	85.9%	92.1%
0	83.1%	93.6%	97.0%
10	71.1%	89.4%	94.5%
20	36.6%	60.7%	77.8%
30	22.4%	33.4%	51.2%

4．3　入力層と中間層の間の重みを自由にする手法

入力層と中間層間の重みを自由にして、入力層と中間層の間の接続がどのくらいシフトインバリアントに影響を及ぼすかを調べる。実験結果を表26、表27に示す。表26は各発声様式による、音韻認識率である。表27は各シフト（ms）に対する音韻認識率である。この実験の学習にはDYNETを用いた。

● 検討

表26、表27と表21と表22と比較すると次のことが分かる。音韻認識率はかなりよい値を示しているが、シフトインバリアント性はかなり悪化する。入力層と中間層の間の重みがシフトインバリアントに重要であることが確認された。表27を表25と比較すると、シフトインバリアントには中間層と出力層の接続の方が効果あると考えられる。この実験の学習にはDYNETを用いた。

4．4　出力層と中間層の間の重みに窓を掛ける手法

出力層の直前の中間層は、各音韻毎9個のユニットを持っている。このユニットの周波数方向での一番外側の値は出力層に対してあまり有効に作用していない。そこで、出力の直前の中間層と出力層との間の重みに図6に示すような窓を掛けることを行った。この効果による各発声様式による認識率の比較を表28に示す。また、このときのシフトインバリアント性がどの様になるかを表29に示す。この実験の学習にはDYNETを用いた。

● 検討

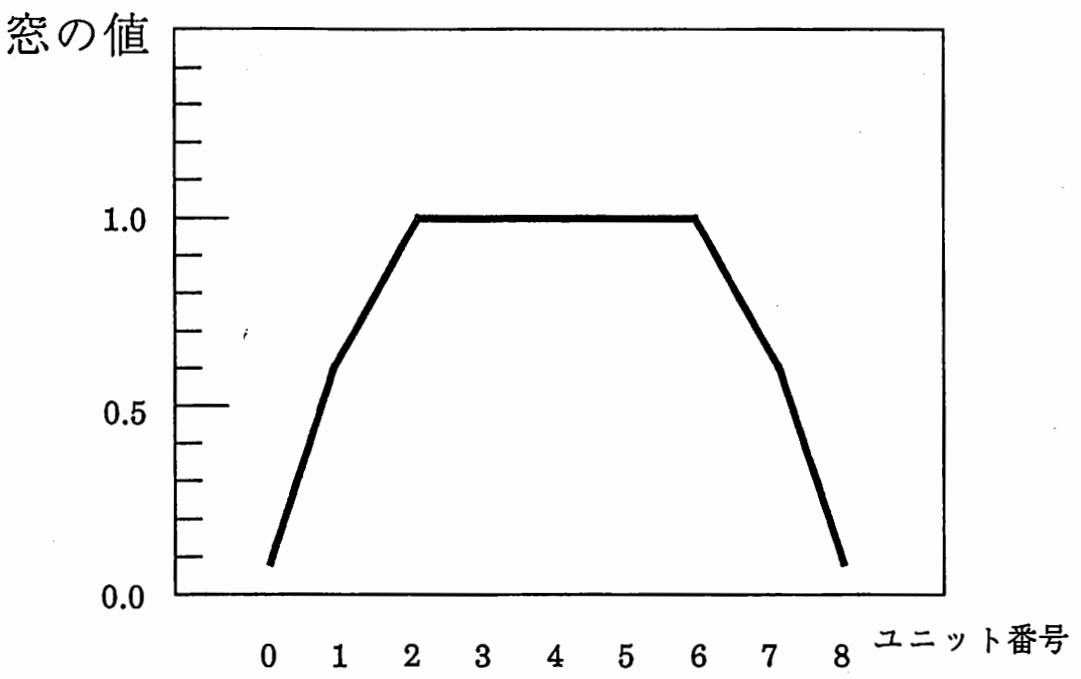

図6　出力層と中間層の間の結合に掛ける窓の値

表26 入力層と中間層の重みを自由にした
TDNNの累積認識率(/bdgmnN/)

順位	1	2	3
文節発声	84.7%	94.9%	98.4%
短い文節発声	84.9%	94.6%	97.7%
自由発声	78.8%	92.0%	96.3%

表27 入力層と中間層の重みを自由にしたTDNNの
シフトインバリアント性能(/bdgmnN/)(短い音節発声)

シフト幅(ms)	1	2	3
-30	30.0%	48.7%	64.3%
-20	44.0%	61.6%	74.9%
-10	71.7%	87.0%	94.0%
0	84.9%	94.6%	95.7%
10	79.1%	91.8%	96.8%
20	52.7%	75.4%	87.4%
30	28.3%	49.4%	68.3%

表２８、表２９と表２１と表２２と比較すると次のことが分かる。音韻認識率はかなりよい値を示しており、シフトインバリアント性も比較的よい結果となっている。このことは出力層と中間層の間の重みと中間層の出力値をうまく設定すると認識率を向上させることが可能であるということを示している。言い換えると今のＴＤＮＮの出力層と中間層の間の接続には改良の余地があると考えられる。

４．５　出力層と中間層の間の重みを固定する手法

４．４で調べた結果から出力層の直前の中間層の出力はあまりよく学習されていないのではないかと考え出力層と最終の中間層の間の重みをすべて０．５に固定し、出力層の閾値の値もすべて０に固定して学習を行った。このことにより出力層に接続する中間層の全てのユニットが０．５または－０．５になるように学習される。この効果による各発声様式による認識率の比較を表３０に示す。また、このときのシフトインバリアント性がどの様になるかを表３１に示す。

● 検討

表３０、表３１と表２１と表２２と比較すると次のことが分かる。音韻認識率は向上を示しており、シフトインバリアント性も向上している。

このネットワークの学習は出力層に接続する中間層に０．５または－０．５を直接教えることに相当する。この学習で著者が期待したのは次のことである。この中間層の１つのユニットの下には入力層の７フレームの窓内のユニットすべてがつながっている。もしこの窓内にユニットに対応する音韻の特徴が存在するので

表28 出力層と中間層の間の重みに窓をかけた
TDNNの累積認識率(/bdgmnN/)

順位	1	2	3
短い文節発声	79.2%	92.5%	95.9%
自由発声	69.8%	86.4%	93.1%

表29 出力層と中間層間の重みに窓をかけたTDNNの
シフトインバリアント性能(/bdgmnN/)(短い音節発声)

シフト幅(ms)	1	2	3
-30	37.8%	62.5%	82.2%
-20	57.2%	81.7%	92.2%
-10	72.6%	90.3%	95.0%
0	79.2%	92.5%	95.9%
10	80.8%	93.2%	96.3%
20	74.8%	89.5%	95.1%
30	59.1%	81.0%	90.8%

表30 出力層と中間層の間の重みを固定した
TDNNの累積認識率(/bdgmnN/)

順位	1	2	3
文節発声	76.0%	92.4%	96.2%
短い文節発声	77.3%	90.7%	95.4%
自由発声	69.8%	88.4%	94.7%

表31 出力層と中間層の間の重みを固定したTDNNの
シフトインバリアント性能(/bdgmnN/)(短い音節発声)

シフト幅(ms)	1	2	3
-30	59.9%	80.2%	90.0%
-20	68.1%	85.7%	92.9%
-10	73.3%	89.7%	89.7%
0	77.3%	90.7%	95.4%
10	76.8%	91.3%	95.5%
20	75.0%	89.2%	94.9%
30	71.8%	87.2%	93.8%

あれば、このユニットが０．５になり、そうでなければ－０．５になってほしいということである。つまり少しでも音韻を識別する特徴が入力層７フレームの窓内にあれば、それを見つけるようにネットワークを学習させるということになる。

４．６　クラス分けネットワークを付加する手法

３．の実験に於て、クラス分けのネットワークが効果があるのではないかと考えられる。そこで図に示すような／ｂｄｇ／と。、／ｍｎＮ／とを分けるクラス分けネットを用いて、各発声様式の違いによる音韻の認識率を表３２に示す。この実験の学習にはＤＹＮＥＴを用いた。

● 検討

３．の実験とこの実験からクラス分けのネットワークが有効であると考えられる。

５．　まとめ

本研究を通していろいろな実験を行った。この過程のなかで以下の４つのことがいえる。

（１）クラス分けネットの有効である。

（２）出力層と中間層の接続の改良が必要である。

（３）従来のＴＤＮＮの入力層と中間層の間の接続がシフトインバリアントに有効である。

（４）従来のＴＤＮＮの出力層と中間層の間の接続がシフトインバリアントに有効である。

謝辞

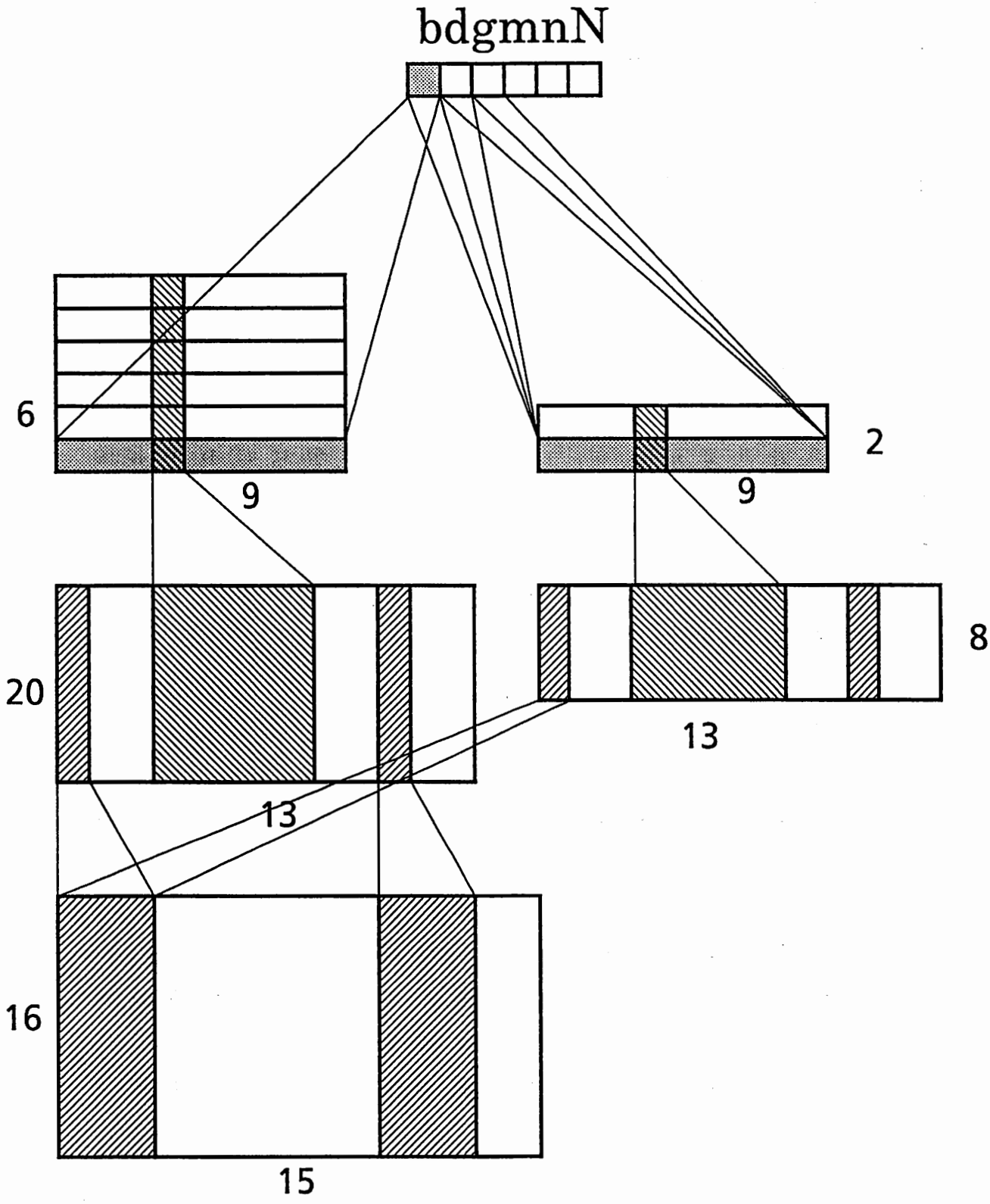

図7 クラス分けネットを持つ/bdgmnN/用TDNN

表32　クラス分けネットを付加した
TDNNの累積認識率(/bdgmnN/)

順位	1	2	3
短い文節発声	80.7%	92.5%	96.6%
自由発声	67.6%	86.9%	94.3%

本研究の機会を与えて頂いたＡＴＲ自動翻訳電話研究所樽松明社長に深謝致します。また、熱心に御討論頂いた音声情報処理研究室前室長の鹿野清宏博士、主任研究員の川端豪博士、田村震一氏をはじめとする音声情報処理研究室の皆様に感謝致します。

参考文献

(1)Ａ．ワイベル：”時間遅れ神経回路網（ＴＤＮＮ）による音韻認識”、信学会、信学技法SP87-100(1987.12)

(2)南泰浩、宮武正典、沢井秀文、鹿野清宏：”ＴＤＮＮ音韻スポッティングと拡張ＬＲパーザを用いた文節音声認識”,音響講論集　３－１－１１(1989.10).

(3)武田一哉、匂坂芳典、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,44,10,pp.747-754(昭63.10).

(4)A. Waibel, H. Sawai and K. Shikano,"Consonant Recognition by Modular Construction of Large Phonemic Time-Delay Neural Networks ",ICASSP'89,pp.112-115(1989).

(5)D.E.Rumelhart and J.L.McClelland :"Parallel Distributed Processing;Explorations in the Microstructure of Cognition",Chap.8,vol.II,MIT Press,Cambridge,MA(1986).

(6) P.Haffner, H.Sawai, A.Waibel and K.Shikano: "Fast Back-Propagation Learning Methods for Large Phonemic Neural Networks"、音響講論集1-6-14(1989.3)または P.Haffner, A.Waibel, H.Sawai and K.Shikano:"Fast Back-Propagation Learning Methods for Neural Networks in Speech" ATR Tech. Report TR-I-0058(1988.11).