TR-I-0144

# TDNN音韻スポッティングと予測LRパーザを用いた大語彙単語音声認識

# Large Vocabulary Spoken Word Recognition Using Time-Delay Neural Network Phoneme Spotting and Predictive LR-Parsing

南 泰浩† 沢井 秀文 宮武 正典* 鹿野 清宏**

Yasuhiro MINAMI † Hidefumi SAWAI Masanori MIYATAKE Kiyohiro SHIKANO

1990.2

## 概要

時間遅れ神経回路網（ＴＤＮＮ）による音韻スポッティング法と予測ＬＲパーザを用いた大語彙の単語音声認識システムを提案する。これはニューラルネットを用いて大語彙の単語音声認識をする最初の試みである。単語中の音韻予測には予測ＬＲパーザを利用する。予測ＬＲパーザが予測した音韻とＴＤＮＮによる音韻スポッティング結果とをＤＰマッチングの手法を用いて照合を行う。男性話者１名の発声した重要語５，２４０語の内、学習に用いていない２，６２０語を重要語の項目全てを対象とする特定話者の単語音声認識を行った結果、第１位の認識率で９２．６％、第２位までの認識率で９７．６％、第５位までの認識率で９９．１％の高認識率を達成した。



† Keio Univ.
† 慶応義塾大学
* 現在 三洋電機-情報通信システム研究所
** 現在NTTヒューマンインタフェース研究所

目次

## 1.はじめに

現在の計算機の進歩により、大規模並列(Massively Parallel)計算の可能性が大きくなるとともに、ニューラルネットワークの研究が再び盛んになってきた。特に多層型のネットワークはBack-Propagationというアルゴリズムにより各種の分野で実用的なシステムの構築を可能とした。アレクッス・ワイベルらは多層型の時間遅れ神経回路網(Time-Delay Neural Network、以下TDNNと略す)という音韻認識用のアーキテクチャを提案した[1][2]。TDNNは特定話者の音韻認識に対して高い認識率を達成した。現在、著者らはTDNNを連続音声に適用することを目的とした研究を行っている[3]。TDNNを連続音声に適用するための音韻認識方法としては音韻セグメンテーション手法と組み合わせた音韻認識[4]と音韻スポッティング法[5]の2つが考えられる。これまでに、連続音声認識に適した音韻スポッティングのためのTDNNのアーキテクチャを構築し、予備的な実験を通してその有効性を示した[5]。また、音韻スポッティングの性能を向上させるための効果的な学習法を確立し、単語中の音韻を98.0%の高い確率でスポッティングできることを示した[6][7]。本報告では連続音声認識への前段階として音韻スポッティング法を利用した大語彙の単語音声認識への適用法を提案し、その認識性能について報告する。これはニューラルネットを用いた大語彙の認識としては初めての試みである。本認識システムでは、単語中の音韻系列の予測には予測LRパーザを利用する。即ち、予測LRパーザは予め作成しておいたLRテーブルを参照しながら次につながる音韻を予測する。音韻スポッティング結果との照合は、DPマッチング法を用いて予測LRパーザにより予測された音韻系列の評価を行う。

## 2.TDNNによる音韻スポッティング法

### 2.1 全音韻スポッティング用TDNNの構造

図1に日本24(18子音/b/, /d/, /g/, /p/, /t/, /k/, /m/ ,/n/ ,/N/ ,/s/ ,/sh/ ,/h/ ,/z/ ,/ch/ ,/ts/ ,/r/ ,/w/ ,/ y/ +5 母音/a/ ,/i/ ,/u/ ,/e/ ,/o/ + 無音)を識別するアーキテクチャを示す[5][8]。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"に母音グループ内識別用サブネットワーク"aiueo" と無音判別用ネットワーク"speech/silence"のモジュールで構成される。これ

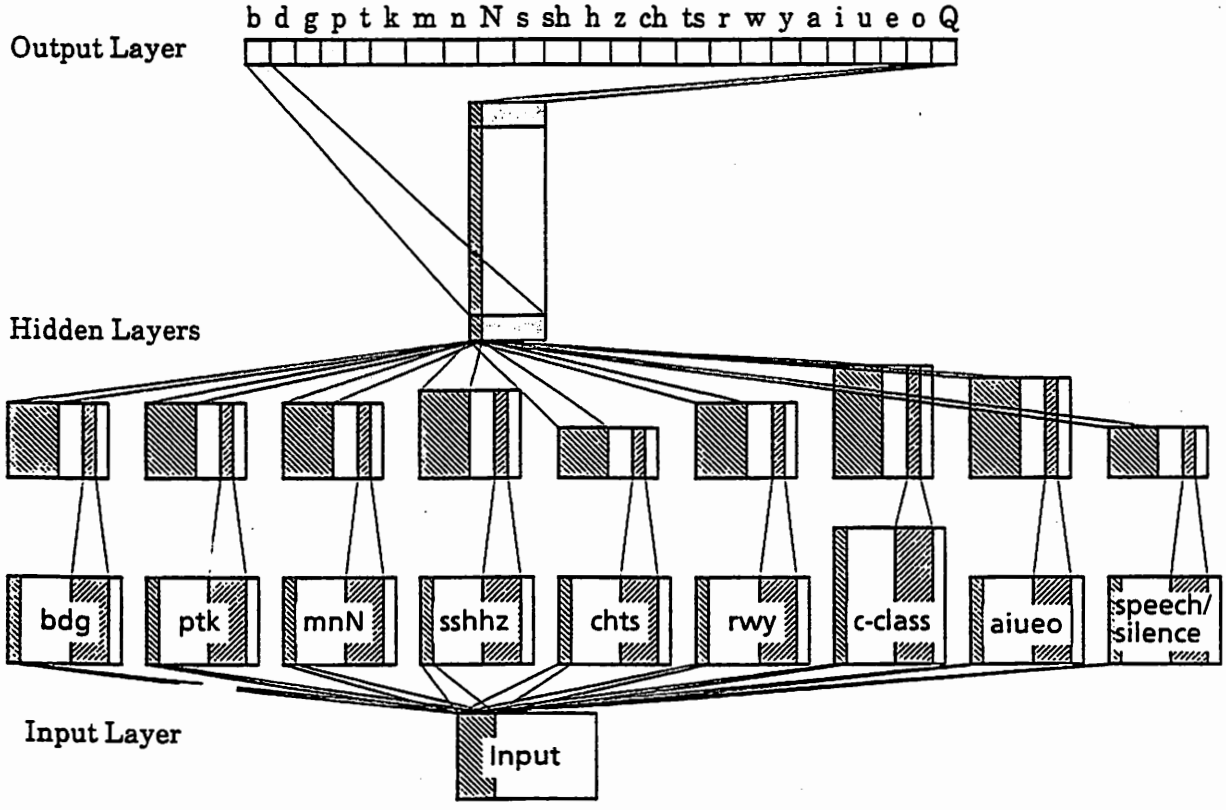


Fig. 1 An Architecture of TDNNs for Spotting all Phonemes
図1 全音韻スポッティング用TDNN

らのサブネットワークは24ユニットを持つ隠れ層3で統合された構成となる。

## 2.2 学習用音韻データ

TDNNの学習用音韻データは、男性話者1名の発声した重要語5240単語の偶数番目の単語中より視察ラベルに基づいて切り出し、24音韻のカテゴリに分類したサンプルである。切り出し方法は音韻のデータの位置によるかたよりをなくすため、図2に示すように各音韻区間毎に複数の位置から学習サンプルを切り出した。切り出し位置は音韻の中心より20msec毎に前後にずらして150msecの区間を抽出した。ただし、音韻の境界部分のデータは使用しない。1カテゴリ当りのサンプル数は上述した手法で切り出し、ランダムに1,000個選んだものであるが、1,000に満たないサンプルについては同じサンプルを重複して用いている。各サンプル毎に16次×15フレーム(150msec)のFFTメル・スペクトラムの特徴を求め、平均値ゼロ、最大値+1、最小値-1に正規化したデータをTDNNへの入力データとした[1]。

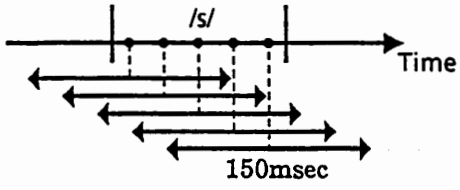


図2　学習用データの抽出方法(例:音韻区間/s/)

## 2.3 TDNNの学習

学習方法にはBack-Propagation学習則[9]を利用している。しかし、大規模なネットワークでのBack-Propagationはかなりの学習時間を必要とする。そこで、Back-Propagationの高速化アルゴリズム[10]を利用した。この手法は急峻な誤差空間を用い、重み係数を頻繁に変化させたり、ステップサイズやモーメンタムをできるだけオーバシュートしないようにスケールアップしたり、誤差の十分小さなデータに対しての処理を省いたりする手法である。ここではこのアルゴリズムを用い、図2の全てのモジュールを分割して学習することなく一度に学習を行っている。

## 2.4 音韻スポッティングの連続発声音声への適用法

図3に2.1で述べたTDNNを用いた連続音声中の音韻スポッティング法について示す。図中の下部の枠はTDNNの入力層を示し、上部の枠は出力層を示す。隠れ層は表記を省略している。図3に示すように音韻スポッティング出力は入力層を1フレームづつずらして音韻認識を行った結果として得られる。この音韻スポッティング法は音韻のセグメテーションをする必要がなく、ネットワークをスキャンするだけで音韻スポッティング結果を得られる特徴がある。この手法を用いて単語中の音韻スポッティングを行った結果の例を図4に示す。下段は入力音声のスペクトログラム、上段は音韻スポッティング出力を示す。横軸は時間を示し、下段の縦軸は周波数を上段の縦軸は24音韻の種類を示している。

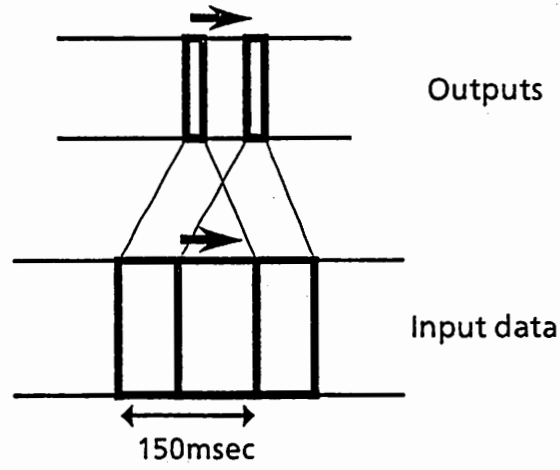


図3　連続音声への適用方法

## 3.予測LRパーザによる単語予測

TDNNのスポッティング結果を単語と照合する手法としては各種の手法が考えられるが、文章などへの応用を考え、ここでは拡張LR構文解析法[11]を用いる。拡張LR構文解析法は従来のLR構文解析法[12]では対処できなかった曖昧な構文に対しても対応できるようにしたものである。さらに北らはこの拡張LR構文解析を連続音声に適用するために予測LRパーザを提案した[13]。本章ではLR構文解析法、拡張LR構文解析法、予測LRパーザについて述べ、最後に予測LRパーザの大語彙単語音声認識への適用方法について述べる。

## 3.1 LR構文解析法

LR構文解析法は従来プログラミング言語の分野でよく知られた構文解析法である[12]。この構文解析法は言語処理の分野でよく取り扱われる文脈自由文法のうちの大部分を解析できる手法である。LR構文解析法は入力された記号を見ながら、順次構文解析を行って

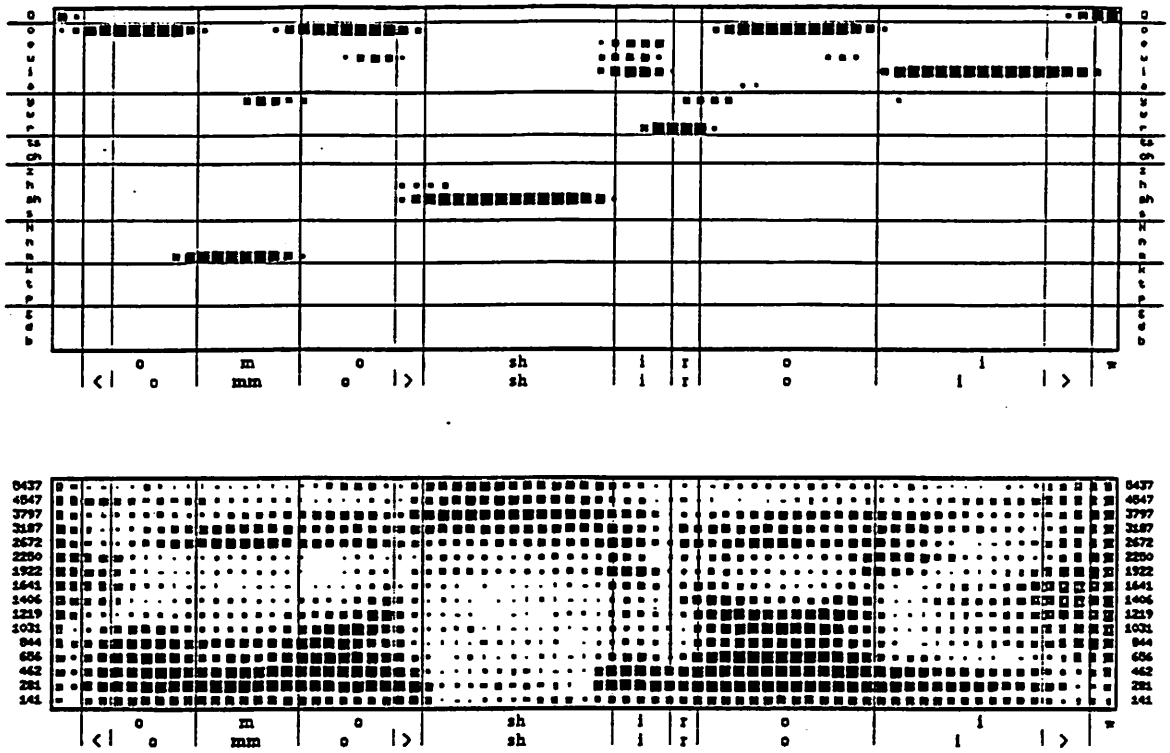

図4 スポッティング結果の例(例:/omoshiroi./)

いく方法で、バックトラックなどの処理を行わないため高速な処理を実現できる。LR構文解析法は状態を保存する状態スタックと呼ばれるFirst In First Out(FIFO)のデータ保存領域と、動作表、行先表という2つの表から成り立つ。これらの表は文脈自由文法の構文より生成される。LRパーザはこれらの表とスタックと入力された記号を見ながら処理を行う。動作表は入力記号と自分自身の状態から動作を決定するものである。動作にはshift、reduce、accept、errorの4種類がある。shiftはパーザの状態をスタックの先頭に入れる動作であり、reduceは状態スタックを文法規則によってまとめる動作である。またaccept、errorはそれぞれ解析完了、解析不能を表す。

### 3.2 拡張LR構文解析法

拡張LR構文解析法はLR構文解析法では実現できなかった曖昧な構文解析を実現する。LR構文解析の表の動作は一つの欄に一つの動作のみ記述していたが、拡張LR構文解析では複数の動作を記述することが可能である。複数の動作を並列に行うことで文法の曖昧性に対応する。

### 3.3 予測LRパーザ

3.1及び、3.2で述べたLRパーザは構文を解析するものであった。予測LR構文解析法はこれらの構文解析手法を連続音声認識に応用できるように改良したものである。予測LR構文解析法は解析の途中で次に来る音韻を予測しながら構文解析を行う。音韻の系列はあらかじめ文脈自由文法で記述され、LRテーブルに変換される。予測はこのLRテーブルを利用することで簡単に実現できる。構文解析の途中で、ある状態に予測LRパーザがあるとすると、次にこの予測LRパーザに入力される可能性がある音韻はLRテーブルを参照することによって調べることができる。予測LRパーザはこれらの音韻を予測音韻と判定する。

### 3.4 予測LRパーザによる単語認識

前節までに示した拡張LR構文解析法では文法を記述していたが、これを図5に示すように単語中の音韻の系列を記述しただけの構文に対しても利用できる。本論文ではこのような単語辞書を用いて認識を行う。

## 4. TDNNによる音韻スポッティング法と予測LRパーザを用いた単語認識システム

(1) S -> WORD

(2) WORD -> a i k a w a r a z u

(3) WORD -> a i s a ts u

(4) WORD -> a i s u r u

図5 単語辞書の例

図6にTDNNによるスポッティング法と予測LRパーザを用いた単語認識システム(以下TDNN-LRと記す)の基本構成を示す。まず、音声入力は2.で述べたTDNN音韻スポッティングを用いて図4上段のような出力に変換される。この出力と単語との照合は予測LRパーザによって文法規則にしたがって処理される。文法規則はあらかじめ文脈自由文法によりLRテーブルとして登録される。予測LRパーザは、いまで処理された音韻系列から次の音韻系列を予測する。この際に複数の予測音韻候補が出現する場合がある。このとき予測LRパーザは解析を並列化して処理する。この予測された音韻とTDNNによるスポッティング結果とをDPマッチングにより照合し、予測された音韻の評価を行う。この操作を入力音声中の音韻が無くなるまで繰り返す。しかし、全ての予測された音韻を評価するためには非常に多くの計算量を必要とする。そこで、ここでは評価された音韻系列の内、上位B候補を残すビームサーチを用いる。このBの値をビーム幅と呼ぶ。

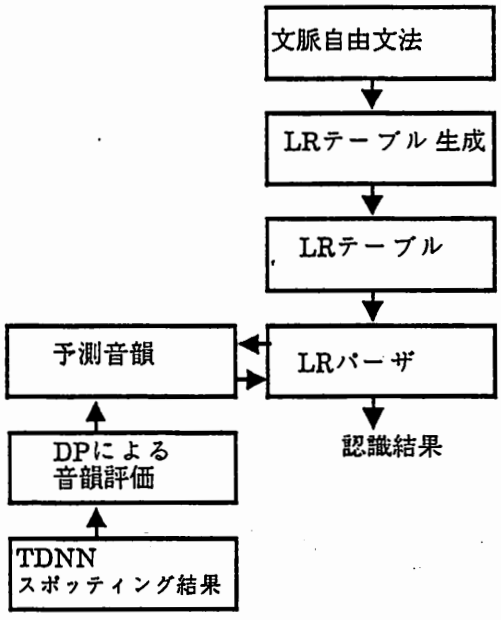


図6 TDNN-Lの構成

## 4.1 予測音韻の評価

予測された音韻の評価は、予測された音韻とTDNNによるスポッティング結果との間でDPマッチングにより行なわれる。DPを行うためには標準パターン長と、標準パターンと入力パターン間の距離または尤度を決定する必要がある。標準パターン長はあらかじめ求めておいた学習単語中の平均の音韻継続時間長を用いる。DPマッチングの尤度としては予測された音韻のTDNNスポッティング出力の対数値を用いる。この場合、累積尤度が最大になるようにDPを行う。使用するDPパスは図7に示すような標準パターン側を基準としたものである。DPパスとしては様々な種類のものが考えられるが、TDNNスポッティング出力が1フレーム分脱落しても対応できるように、また入力音韻の継続時間の長い部分を過度に評価しないために傾き1/2~2に制限したこのDPパスを使用した。音韻照合は具体的には以下のように実現される。

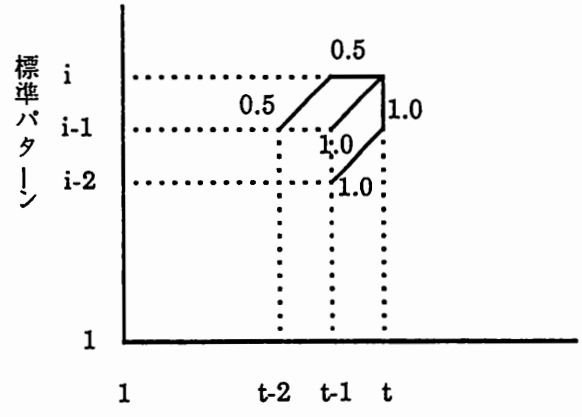


図7 認識に用いたDPパス

[記号の定義]

N :未知音韻スポッティング結果の最大フレーム長

M :照合される音韻の平均フレーム長

j :測された音韻

p(t,j) :フレームtでの音韻jのスポッティング出力値 (1<p(t,j)≦1)

D0(t) :尤度の累積値保存用テーブル0

D1(t) :尤度の累積値保存用テーブル1

[漸化式の計算]

$Q(0,t)=D0(t)(t=1,...,N)$

$Q(1,t)=D1(t)(t=1,...,N)$

初期値 $Q(1,1)=p(1,j)$ ,その他は全て0

$(t=1,..,N,i=1,..,M)$について以下を実行する

$$Q(i,t)=\max[Q(i-1,t-1)+\log(p(t,j)),$$
$$Q(i-2,t-1)+\log(p(t,j))+\log(p(t,j)),$$
$$Q(i-1,t-2)+0.5\times\log(p(t-1,j))$$
$$+0.5\times\log(p(t,j))]$$

$D0(t)=Q(M-1,t)(t=1,...,N)$

$D1(t)=Q(M,t)(t=1,...,N)$

このD0(t),D1(t)が次の音韻のDPマッチングの初期値となる。以上の処理は予測された音韻に対して両端点フリーのDPマッチング手法である。この処理を続けることによって、音韻系列を積み上げてDPマッチングを行うことになる。最終的には積み上げられた音韻系列の中でビーム幅内にある尤度の最も大きな音韻系列を認識結果とする。

## 5.認識実験

認識実験には、男性話者1名の発声した音韻バランス216単語と重要語5,240語の内、学習用の音韻サンプルの抽出に用いていない奇数番目の2,620語を利用した。重要語の認識では評価用の単語数および、認識対象語彙数を共に、100、500、2,620、5,240と変化させた。但し、5,240語の場合には、評価用の単語数は2,620語である。100単語は、音韻バランス単語216語[14]と重要語の奇数番目に含まれる単語の共通単語を用いた。また500語は、評価用の2,620語からランダムに選択したものである。LRパーザのビーム幅は100とした。音韻バランス単語の認識率を表1に、重要語の認識率を図8に示す。音韻バランス216単語の認識では第1位の認識率97.2%、第2位~5位までの認識率は共に99.5%となった。図8の重要語に対する認識率では、横軸に認識対象語彙数、縦軸に第n($1\leqq n\leqq 5$)位までの累積認識率を示した。重要語5,240の認識において対象語彙数を5,240とした場合、第1位の認識率では92.6%、第2位までの認識率では97.6%、5位までの認識率では99.1%となった。また、第1位認識率は認識対象語彙数が100、500、2,620、5,240と変化するのに伴い、100%、98.0%、95.0%、92.6%と変化する。しかし、第5位までの認識率はどの場合も99.1%以上の認識率を達成している。

## 6.検討および考察

表1 音韻バランス216単語の認識率

| 候補順位 | 累積正当数 | 累積認識率(%) |
|---|---|---|
| 1 | 210 | 97.2 |
| 2 | 215 | 99.5 |
| 3 | 215 | 99.5 |
| 4 | 215 | 99.5 |
| 5 | 215 | 99.5 |

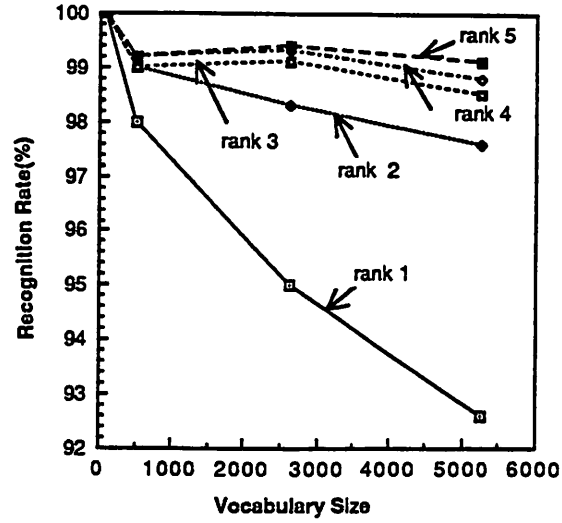


図8 重要語5240単語の認識率

音韻バランス216単語での認識誤りは"chaNto"(2位)、"rejaa"(5位までに認識されない)、"neage"(2位)、"popyuraa"(2位)、"kaNpyou"(2位)、pyuupyuu"(2位)の6単語だけである。この原因は"p"の置換誤り(例、"kaNpyou"→"kaNbyou")、及び"n"が"m"に置換した(例、"neage"→"megane")ことなどがあげられる。置換された"p"や"n"などはほとんど発火せず、これが最大の誤認識の原因となっている。これは、図1のTDNNの側抑制(lateral inhibition)が効きすぎているためと考えられる。より柔軟な音韻スポッティング法を実現することにより、このような致命的な誤りを解消する必要がある。図8から、第1位の認識率が単語数の増加に伴い、低下することがわかる。しかし、第5位までの認識率は単語数が増加してもほとんど変わらない。これはTDNNのスポッティングの精度が評価用単語2,620語中の音韻に対し、98.0%[6][7]とかなりよいためと考えられる。重要語5,240単語中の誤認識は以下の3種類が非常に多い。

(1) 語頭の"t"や"k"の挿入。(例:"aisuru"→"taisuru")

(2) 短い単語を長い単語と認識する。(例:"aa"→"hanahada")

(3) 促音と無声破裂(破擦)音に伴う閉鎖区間との混同。(例:"itai"→"iQtai")

(1)は語頭が不安定でTDNNが"t"や"k"を挿入しやすいためにおきる。(2)はDPパスの制限以上の入力に対応できないことが原因である。(3)は促音と無声破裂(破擦)音に伴う閉鎖区間の継続時間長の違いをDPマッチングの手法の中に取り入れていないからである。

## 7.まとめ

本論文ではTDNNによる音韻スポッティング法と構文解析法の一種である予測LRパーザを利用した大語彙単語音声認識システムについて述べた。予測LRパーザは文法に従って、ある時点での音韻系列から次の音韻を予測し、音韻の参照パターン長は学習用の単語の統計から予め求めた平均音韻長を用い、予測された音韻とTDNN音韻スポッティング結果とをDPマッチングの手法により照合する。このシステムで重要語5,240語を対象として認識した結果、第1位認識率で92.6%、第2位までの認識率で97.6%、第5位までの認識率で99.1%を達成した。このことによりTDNNによる音韻スポッティング法を用いた音声認識の可能性が開けた。本システムは単語認識だけでなく構文を変えることにより連続音声認識に適応可能である[3]。今後の課題としては、発声方法や発声者の違いに対応できる柔軟なネットワークの構築や学習方法や評価方法の開発等が挙げられる。

謝辞

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。日頃ご指導頂く慶応義塾大学中川正雄教授に深謝致します。また、熱心に御討論頂いた音声情報処理研究室主任研究員の川端豪博士、客員研究員のアレックス・ワイベル博士をはじめとする音声情報処理研究室の皆様、データ処理研究室の北研二氏に感謝致します。

文献

[1]A.Waibel,T.hanazawa,G.Hinton,K.Shikano and K.Lang :"Phoneme recognition using time-delay neural networks",IEEE Trans. on ASSP,pp328-339,vol.37.No.3(March 1989). または ATR Tech. Report TR-I-0006(1987.10).

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

[3]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).

[4]小森康弘、畑崎香一郎、田中孝明、川端豪、鹿野清宏:"スペクトログラム・リーディング知識とニューラルネットワークを用いた音声認識エキスパートシステム",信学技法SP89-33(1989).

[5]沢井秀文、宮武正典、アレックス・ワイベル、鹿野清宏:"連続音声認識のための時間遅れ神経回路網を用いた音韻/音節スポッティング",信学論D-II,vol.J72-D-II,No.8,pp.1151-1158(1989.8).

[6]宮武正典、沢井秀文、鹿野清宏:時間遅れ神経回路網(TDNN)による音韻スポッティングの改良",音響講論集,1-1-25(1989.10).

[7]宮武正典、沢井秀文、鹿野清宏:時間遅れ神経回路網(TDNN)のための音韻スポッティングの効果的学習法", ATR Tech. Report TR-I-0103 (1988.8).

[8]宮武正典、沢井秀文、鹿野清宏:"連続音声中の音韻スポッティングのためのTDNN構成法", 信学技法SP89-32(1989.6).

[9]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).

[10] P.Haffner,H.Sawai,A.Waibel,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 ,K.Shikano: "Fast Back-Propagation Learning Methods for Neural Networks in Speech" ATR Tech. Report TR-I-0058 (1988.11).

[11]M.Tomita:"Efficient Parsing for Natural Language :A Fast Algorithm for Practical Systems" ,Kluwer Academic Publishers(1986).

[12]A.V.Aho and J.D.Ullman:"Principles of Compiler Design",Addison-Wesley (1977).

[13]北研二、川端豪、斎藤博昭:"HMM音韻認識とLRパーザを用いた文節認識", 信学技報SP88-88(1988.10).

[14]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)