ニューラル･ファジー学習法による TDNN-LR連続音声認識システムの性能向上

An Improvement of TDNN-LR Continuous Speech Recognition System Using A Neural Fuzzy Training Approach

小森　康弘、　アレックスワイベル†、　嵯峨山　茂樹

Yasuhiro KOMORI, Alex H. Waibel † and Shigeki SAGAYAMA

ATR自動翻訳電話研究所

ATR Interpreting Telephony Reseach Laboratories

†カーネギーメロン大学

† Carnegie Mellon University

あらまし

本稿では、新しい音素識別ニューラルネットワークのためのファジー学習法を提案する。提案するファジー学習法は、バックプロパゲーション法により実現されるが、ニューラルネットワークの教師信号の与え方が従来の学習法の1または0を与える方法と異なる。ファジー学習法では、学習サンプル毎に全音素クラスに対する類似性を教師信号として与える。この類似性は、学習サンプルとそのサンプルに最も近い各音素クラスに属するサンプルとの距離により求められる。ファジー学習法の有効性を示すために、ATRの文節音声データベースを用いて、/b,d,g,m,n,N/識別、18子音識別、さらにTDNN-LRを用いた文節認識実験を行った。いずれの実験においても、ファジー学習法により改善がみられた。特にTDNN-LR(約500ルール)を用いた文節認識実験では、ファジー学習法により、第1位認識率が70.9%から81.7%に改善され、第5位累積認識率は91.7%から95.7%に改善された。

ABSTRACT

In this paper, we propose a new fuzzy training approach for phoneme identification neural networks. The fuzzy training approach is realized by back-propagation algorithm, but differs in the point of how to give the target values to the neural networks from the conventional training approach which target values are defined as 0 or 1. In the proposed fuzzy training approach, the target values are defined as how likely the input phoneme is to every phoneme class. This likelihood is computed according to the distance between the input datum and its nearest datum which belongs to each class in the training data set. To show the effectiveness of the fuzzy training approach, /b,d,g,m,n,N/ identification and 18-consonant identification experiments were performed using the ATR phrase database. TDNN-LR phrase recognition experiment was also performed on the same database using 25-phoneme identification TDNN. Improvements can be observed on every experiments. Particulary, TDNN-LR phrase recognition result drastically improved from 70.9% to 81.7% on top-1 candidate and from 91.7% to 95.7% on top-5 candidates using the proposed fuzzy training approach using the LR-paser of about 500 rules.

1. はじめに

近年の音声認識研究は、計算機の急激な進歩により多量のデータ処理ができるようになり、多量データを用いた手法が盛んに研究されている。その中でも強力な学習アルゴリズムであるバックプロパゲーション法が開発されたfeed-forward識別型ニューラルネットワーク[1-2]は、高い音素識別性能を示している。Alex Waibelにより提案されたTDNN[3]は、音素識別性能が高く、時間方向へのずれも吸収する能力があるなどの特徴を持ち、音声認識に向いたモデルとして知られている。

しかし、TDNNなどの識別型ニューラルネットワークによる音声認識研究が進むにつれ、バックプロパゲーション法により学習したfeed-forward識別型ニューラルネットワークでは、発話速度の差などに対するロバスト性が不十分であることが判明してきた[4-8]。これは、教師信号に識別クラスの属性の情報のみを0あるいは1の2値信号で与えるために、各クラスに対する近さ(類似性)が出力値に現れないのが原因と考えれる。更に、従来の学習法では、学習データに対して過学習が起こりやすく、学習対象の音声空間を急峻に判別してしまう問題等が報告されている[4-8]。この結果、第2位以下の候補の出力値が0付近に抑えら

れ累積認識率が向上しなかったり、学習時と異なる発話速度の音素識別性能が大きく低下する問題が生じている。この累積認識率が向上しない問題は、連続音声認識を実現する際、言語情報等のトップダウン処理との結合において情報の欠落が起こり、致命的な誤りを起こす原因となる。

この問題解決のために幾つかの研究が行われている。鈴木らは、平均曲率を用いた評価基準により過学習を回避する方法[4]を提案している。南らはTDNN出力を積分した出力値の平滑化[5]により、累積認識率の向上をねらっている。川端のk-近傍内挿学習法[6]では、学習サンプルとその近傍サンプルを用いてサンプル間とそれらの出力値間の制約条件を導入することで過学習を抑えようと試みている。鷹見らは、複数個の相対判定型ニューラルネットワーク[7]を提案し、多数決の効果による出力値の平滑化を行い認識率の向上を目指している。また、中村らの音素フィルタ・ニューラルネットワーク[8]では、直接音素の分類を避け恒等写像の誤差により、累積認識率の向上を目指している。

本稿では、学習サンプル毎に全音素クラスに対する類似性を教師信号として与える識別型ニューラルネットワークのためのファジー学習法[9](Neural Fuzzy Training)を提案する。提案するファジー学習法は、学習サンプルとそのサンプルに最も近い各音素クラスに属するサンプルとの距離により各音素クラスへの類似度を求め、それを0から1までの連続値で表し、この類似度を教師信号として与えバックプロパゲーション法により学習する方法である。提案するファジー学習方法は、従来の0,1の2値学習とこの点で異なる。

以下、音素識別ニューラルネットワークのためのファジー学習法の基本的な考え方を述べる。加えて、その有効性を示すために/b,d,g,m,n,N/と18子音の識別実験およびTDNN-LRを用いた文節認識実験を行ったので報告する。

2. ニューラルネットワークのファジー学習法

図1に従来学習法および提案するファジー学習法の基本的な考え方を示す。従来学習法は、学習サンプルに属している音素クラスに1を、その他の音素クラスに0を教師信号として与え、バックプロパゲーション法により学習を行う。図1に示す■では、教師信号は$target_{01}=\{1,0,0\}$になる。この方法に対して、提案するファジー学習法は、学習サンプルの音素クラス属性を教師信号として0または1で与えるのではなく、ファジーメンバシップ関数のように0から1の間の値を用いてその音素クラス属性らしさで与える。つまり、サンプル毎に全音素クラスに対する類似度を求め教師信号として与え、バックプロパゲーション法により学習を行う。学習サンプルの全音素クラスに対する類似度は、各音素クラス内のサンプルにうち学習サンプルに最も近いサンプルとの遠近に応じて求めることができる。つまり、学習サンプルとある音素クラス内のサンプルのうち学習サンプルに最も近いサンプルが遠いときは、学習サンプルはこの音素クラスにはあまり類似しておらず、逆に近いときには、学習サンプルはこの音素クラスに類似していると考えることができる。このとき、学習サンプル空間上での遠近を定義するために距離d(例:ユークリッド距離)を導入する。さらに、距離dを類似度に変換するために類似度変換関数$f(d)$を定義する。学習サンプル空間上での遠近を類似度として反映するには、類似度変換関数$f(d)$をなんらかの単調減少関数で表現すれば良い。このように定義すれば、遠いものの類似度は小さい値に、近いものの類似度は大きい値に変換される。例えば、図2に示す$f(d)=exp(-\alpha\cdot d^2),\ \alpha>0$である。図1に示す●では、教師信号は$target_{FZ}=\{f(d_A),f(d_B),f(d_C)\}$となる。但し、提案するファジー学習法は、全学習サンプルに対して音素クラス毎に学習サンプルに最も近いサンプル求めるための膨大な距離計算N·(N-1)/2(Nは全サンプル数)を必要とする。

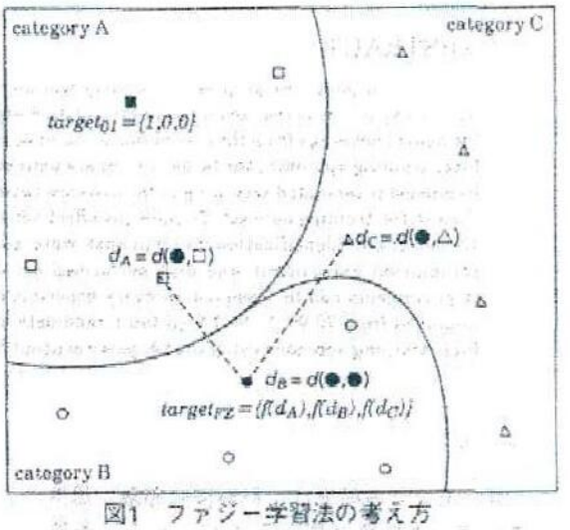


図1 ファジー学習法の考え方

3. /b,d,g,m,n,N/と18子音の識別実験

提案するファジー学習法の有効性を示すために日本語6音素(/b,d,g,m,n,N/)と日本語18子音(/b,d,g,p,t,k,ch,ts,s,sh,h,z,m,n,N,r,w,y/)の音素識別実験を行った。

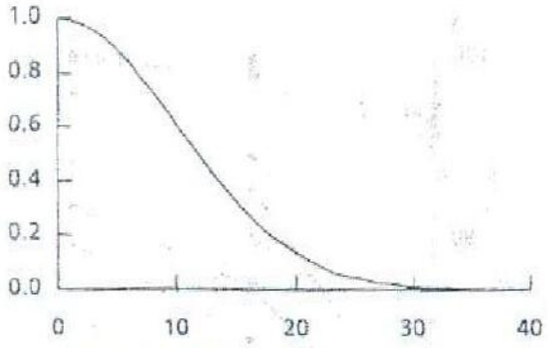
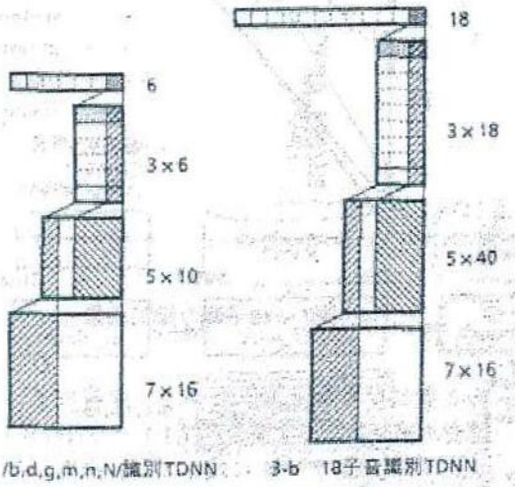


図2　類似度変換関数　$f(d) = exp(-0.005 \times d^2)$

3.1　実験条件

学習には、ATRデータベース[10]5,240単語の偶数番単語(5.7モーラ/秒)からラベルを用いて切り出した音素を用いた。学習サンプルは、/b,d,g,m,n,N/の学習では1音素クラス当たり最大500の計1,857サンプル、日本語18子音の学習では1音素クラス当たり最大250の計3,638サンプルを用いた。評価には、発話速度に対する耐性を考慮してATRデータベースから5,240単語の残りの奇数番単語(5.7モーラ/秒)、文節発声(7.1モーラ/秒)および文発声(9.6モーラ/秒)のデータからラベルを用いて切り出した音素を用いた。

実験に用いた音素識別ニューラルネットワークを図3-a (/b,d,g,m,n,N/識別TDNN)、図3-b (18子音識別TDNN) に示す。feed-forward 型の4層からなるTDNN(入力層7フレーム)を用い、各音素の終端を中心とする70msのデータを入力として用いた。入力音声の分析条件を表1に示す。



3-a　/b,d,g,m,n,N/識別TDNN　　3-b　18子音識別TDNN

図3　子音識別TDNNの構成

表1　入力音声の分析条件

sampling	12kHz
shift	10ms
window	21.3ms Hamming
FFT	256 points
パラメータ	16次元メルスケール
正規化	10ms毎にまとめ ±1.0の間に 平均値0.0として正規化

比較評価は、2つの従来学習法と提案するファジー学習法の3種の方法を用いて行った。2つの従来学習法は、バックプロパゲーション法のerror関数に1) M.S.E. (mean square error: ε^2, εは誤差) を適用したものと2) McClelland error ($log(1+\varepsilon^2)$, εは誤差) を適用したものである。3) ファジー学習法のバックプロパゲーション法のerror関数には、M.S.E. を用いた。McClelland errorは、誤差εを強調して伝搬する学習方法であり、ニューラルネットワークの識別クラス数が大きいときに学習効率が非常に良い。

サンプル間の距離は入力7フレームデータのユークリッド距離dを用い、ここでは、類似度変換関数として$f(d)=exp(-a \cdot d^2)$, $a=0.005$を用いた。定数aは、第2位以下の音素クラスの類似度が適当に反映するように定めた。尚、認識のときに用いるTDNNの重み係数は、学習の際の重み係数更新100回中、評価用の奇数番単語から切り出した音素の識別率が最大となる重み係数を選んだ。

3.2　実験結果

●学習速度

各学習方法に対する/b,d,g,m,n,N/識別TDNNの学習速度を図4に、日本語18子音識別TDNNの学習速度を図5に示す。横軸にTDNNの重み係数の学習更新回数を、縦軸に学習データに対する音素識別率を示し、学習速度とした。識別音素クラス数が6と少ない/b,d,g,m,n,N/識別TDNNの学習に関しては、いずれの学習方法においても差が現れない。しかし、識別音素クラス数が18と多い日本語18子音識別TDNNの学習においては、予想通り従来学習法にバックプロパゲーション法のerror関数にMcClelland errorを適用した方法が、従来学習法にM.S.E.を適用したときよりかなり速く学習が進むことが分かる。識別音素クラス数がさらに多くなればこの差はさらに広がると思われる。学習速度の観点からは、従来学習法にはMcClelland errorを適用した方法が良いと考えられる。一方、提案したファジー学習法は、従来学習法にMcClelland errorを適用したときと同程度の速度を示

した。従って、教師信号が求まってさえいれば、識別音素クラスが多いときでもかなり速い学習が期待できる。しかし、ファジー学習法では、教師信号である各音素クラスへの類似度を求めるためにかなりの計算量を必要とする欠点がある。

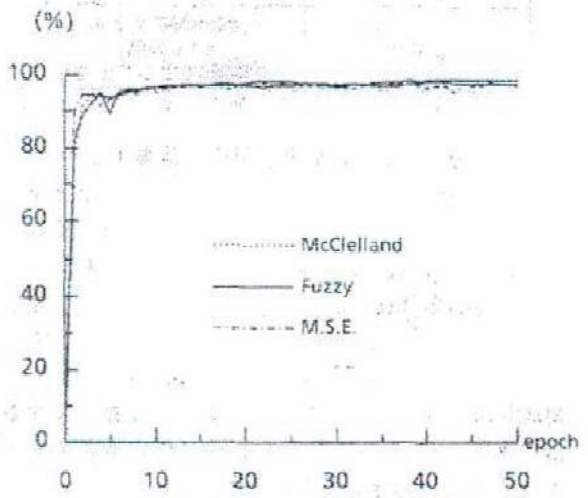


図4　/b,d,g,m,n,N/の学習速度

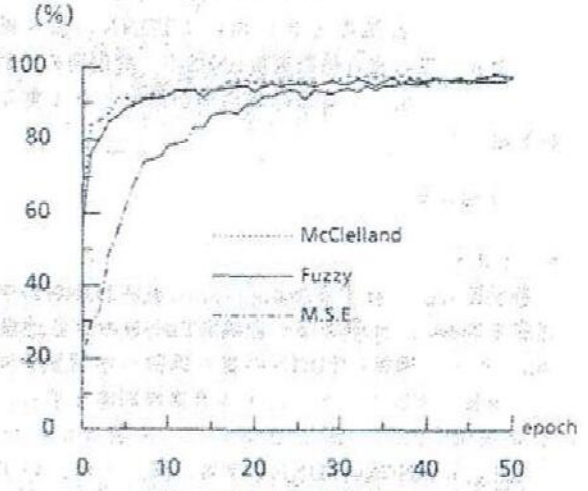


図5　18子音の学習速度

● 識別結果

各発声データ(単語発声学習用データ、単語発声・文節発声・文発声発声評価用データ)に対する/b,d,g,m,n,N/の識別結果を図6に、日本語18子音の識別結果を図7に示す。縦軸に音素識別率、横軸に音素認識候補数を示す。学習用データおよびそれと発話速度が差がない単語発声データ(5.7モーラ/秒)においては、いずれの識別実験においても音素識別性能はほとんど変わらなかった。しかし、学習用データと発話速度が差が異なる文節発声(7.1モーラ/秒)および文(9.6モーラ/秒)発声データに対しては、提案したファジー学習法は第1位及び第N位累積識別結果がとも従来学習方法より向上している。特に、文発声データにおける第N位累積識別結果が顕著に改善されている。連続

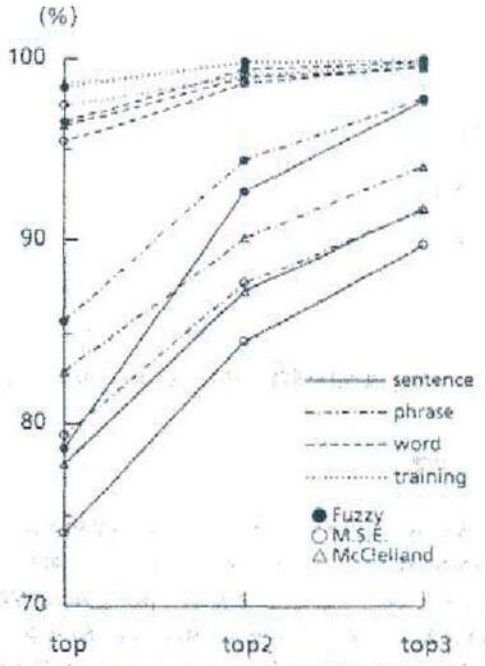


図6　/b,d,g,m,n,N/の識別結果

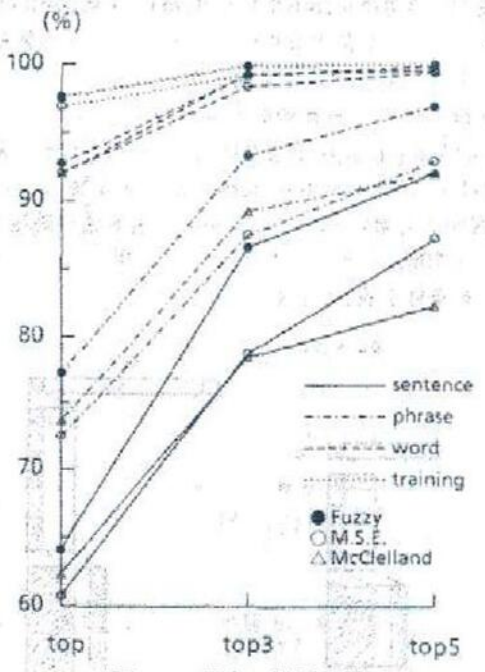


図7　18子音の識別結果

音声(文節-文)認識では、第1位の識別結果だけではなく、この第N位累積識別結果も言語情報との結合において非常に重要である。従って、提案するファジー学習法によるこの改善は、言語情報との結合した文節-文認識においてかなり有効であると考えられる。

一方、文発声データに対する日本語18子音の識別結果をみると、識別音素クラス数が多いとき速く学習が進むMcClelland errorを従来学習法に適用したときの方が、M.S.E.を従来学習法に適用したときより第5位累積識別結果が劣化していることが分かる。これは、McClelland errorは、誤差εを強調して伝搬することによりニューラルネットワークの出力値を0または1に強引に近づけてしまい、第2位以下の音素の出力値がほとんど0になってしまうのが原因と考えられる。この結果より、McClelland errorを従来学習法に適用した方法は、言語情報との結合した文節-文等の連続音声認識においては、識別音素クラス数が多いときにニューラルネットワークが速く学習できても、あまり有効な方法ではないと考えられる。

従来学習方法では、M.S.E.およびMcClelland errorのいずれのerror関数を適用した場合でも学習速度や累積識別率に問題が生じる。一方、提案したファジー学習法には、これらの問題はないものの、学習サンプル数が多い場合には、全学習サンプル間の距離計算を行って教師信号を求めるため、莫大な計算量を必要とする欠点がある。

4. TDNN-LRを用いた文節認識実験

音素識別実験に加え、25音素識別TDNNと拡張LR-parser[11]を統合したTDNN-LR連続音声認識システム[12]を用いた文節音声認識実験を行った。

4.1 TDNN-LR連続音声認識システム

TDNN-LR連続音声認識システムのブロック・ダイアグラムを図8に示す。本システムは、言語情報を利用しながら音素照合を行う音声認識と言語解析が同時に進行する効率的かつ高精度な音声認識処理系である。文脈自由文法(context-free grammar)で記された文法規則は、LR-table generatorによりあらかじめLR-table(動作表と行先表)に変換される。LR-parserはこのLR-tableに従って、言語情報の制約によりある状態から連結し得る次の音素を提示し、その音素の照合要求を出す。音素照合は、DPマッチング(Dynamic Programming)により実現される。DPマッチングのためのスコアは、照合要求を出された音素とあらかじめTDNNを用いてフレーム毎に音素識別を行ったTDNN音素スポッティング[13]結果により求まる値の対数で定義され、その値が照合音素のスコアとなる。DPマッチングの傾斜制限は1/2~2とする。DPパスは、基本軸が参照パターン側である非対称型DPパスを用いる。標準パターン(照合音素パターン)の長さは、学習用の大語彙単語から切りだした音素の平均継続時間長を用いる。音素継続時間長制御は、音素のDPマッチング後に平均音素継続時間長μと実際に照合された区間長tとのずれに応じたペナルティをガウス分布から求め、累積スコアに掛け合わせることで行う。音素継続時間長のペナルティには、計算式$p(t)=exp(-(t-\mu)^2/(2\pi\sigma^2))$を用いる。尚、文節認識時の音素継続時間長パラメータは、単語発声データから求めた音素継続時間長パラメータを文節用に変換したものを用いた。

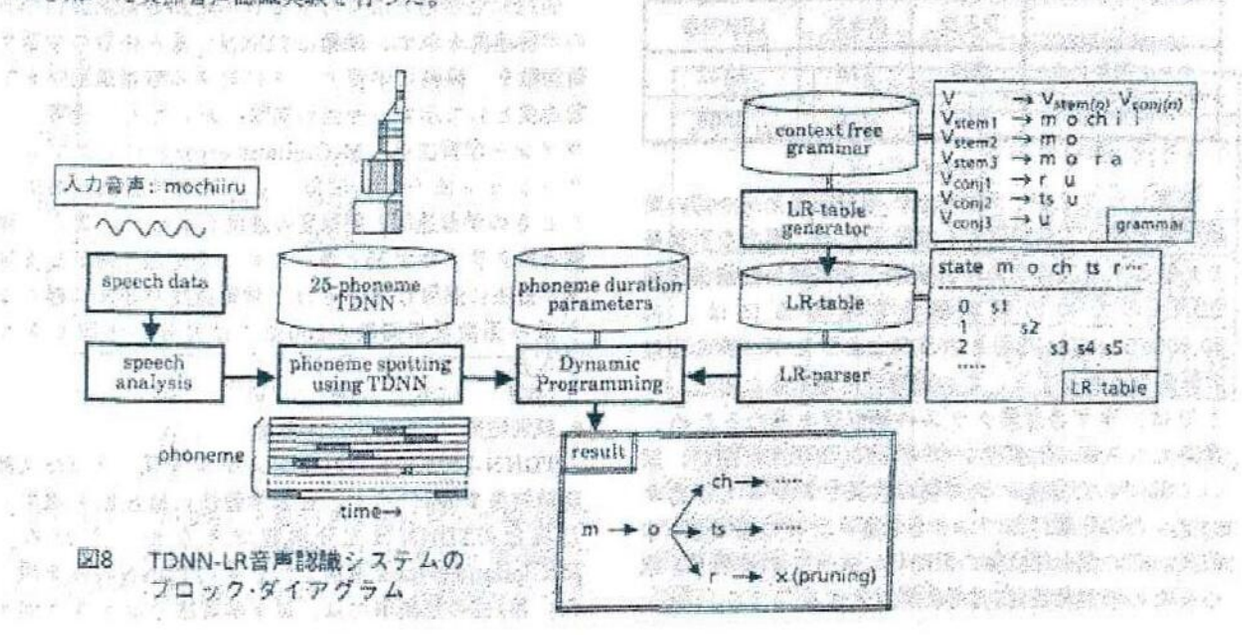


図8 TDNN-LR音声認識システムのブロック・ダイアグラム

4.2 実験条件

25音素識別TDNNの学習には、前述の実験と同様ATRデータベース5,240単語の偶数番単語からラベルを用いて切り出した音素を用い、学習サンプルには、1音素クラス当たり最大2,000の38,321サンプルを用いた。音素を切り出す位置は、前述の実験と異なり、図9に示すように1つの音素から複数の学習サンプルを切り出す。音素中心のサンプルおよび前後の音素境界からサンプルの中心が15ms内側にあるものとそれらに加えて、この前後の両サンプルの内側でかつ音素の中心から前後方向に15msづつずれた全てのサンプルを用いた。TDNN-LR連続音声認識システムの文法として、表2に示す2種類の文法: 1)一般的文法と2)タスク向き文法を用い、ATR文節データベース中の278文節音声を用いて評価を行った。ニューラルネットワークは、図10に示すfeed-forward型25音素識別TDNNを用いた。入力層は前述の実験と同様7フレームで、入力音声の分析条件も表1と同様である。

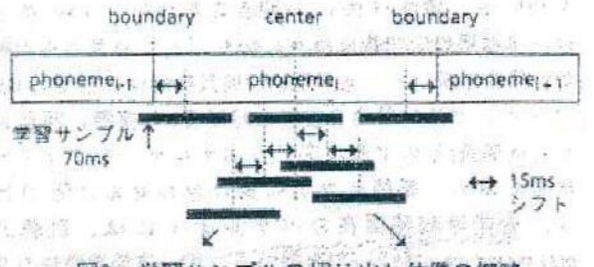


図9 学習サンプルの切り出し位置の概略

表2 文法の大きさ(文節認識用)

	規則数	語彙数	LR状態数
タスク向き文法	607	275	1341
一般的文法	1672	1035	4866

提案したファジー学習法は、全学習サンプル間の距離計算を行って教師信号を求めるため、莫大な計算量を必要とする欠点がある。実際、本実験の25音素識別TDNNのための教師信号を求めるには、約50,000(50,000-1)/2回という莫大なサンプル間の距離計算量を必要とする。この問題に対処するために、ここでは、まず各音素クラスの類似度を求めるため、音素クラス毎に代表サンプルを200個任意に選び、続いて選ばれた音素クラス毎の代表サンプルと学習サンプルとの距離に基づいて各音素クラスの類似度を求める。この類似度計算の簡略化により、教師信号を求めるための計算量は約1/5に削減される。

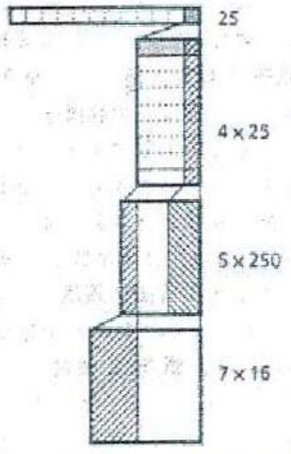


図10 25音素識別TDNNの構成

比較評価は、音素識別実験と同様に2つの従来学習法とファジー学習法の3種法を用いて行った。2つの従来学習法は、バックプロパゲーション法のerror関数に1) M.S.E.(学習速度のみ)を適用したものと2) McClelland errorを適用したものである。3) ファジー学習法のバックプロパゲーション法のerror関数には、M.S.E.を適用した。

サンプル間の距離も入力7フレームデータのユークリッド距離dを用い、類似度変換関数として$f(d)=\exp(-\alpha \cdot d^2), \alpha=0.005$を同様に用いた。認識ときに用いるTDNNの重み係数は、学習更新回数100回目のものを用いた。

4.3 実験結果

● 学習速度

図11に各学習方法に対する日本語25音素識別TDNNの学習速度を示す。横軸にTDNNの重み係数の学習更新回数を、縦軸に学習データに対する音素識別率を学習速度として示す。今回の実験においても、提案したファジー学習法は、McClelland errorをバックプロパゲーション法のerror関数として従来学習法に適用したときの学習速度と同程度の速度を示した。また、識別音素クラス数が25と多いため、予想通りM.S.E.を従来学習法に適用したときは、学習速度が非常に遅くなり重み係数更新回数が100回では充分に学習できなかった。

● 識別結果

TDNN-LR連続音声認識システムを用いた278文節認識結果を表3に示す。従来学習法にM.S.E.を適用した場合のTDNNは充分学習できなかったため、McClelland errorを適用したときのTDNNのみを用いた。第1位の認識率では、従来学習法ではタスク向き

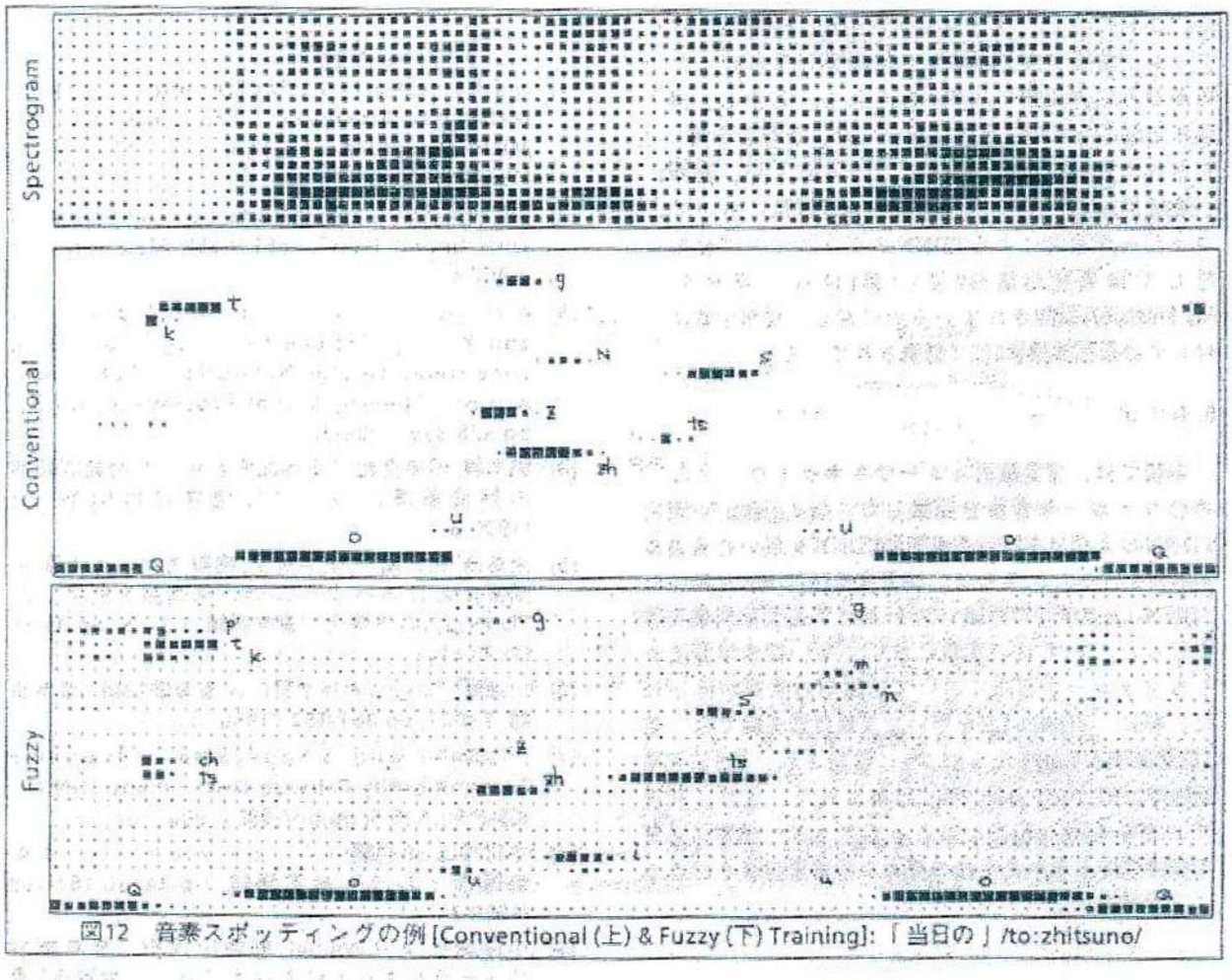


図12　音素スポッティングの例[Conventional (上) & Fuzzy (下) Training]:「当日の」/to:zhitsuno/

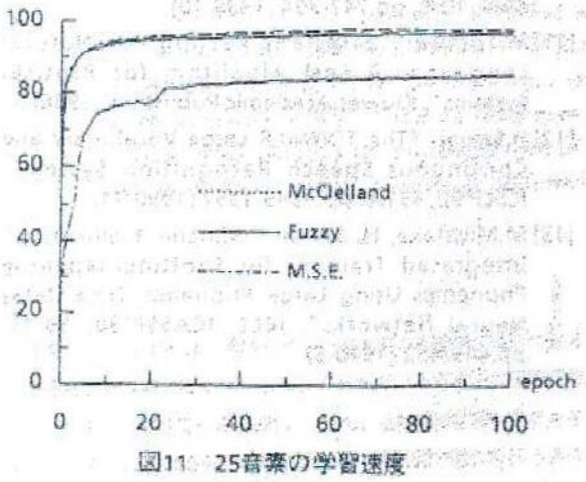


図11　25音素の学習速度

文法において70.9%、一般向き文法において64.7%であった。これに対して提案したファジー学習法では、それぞれ81.7%、70.9%とかなり向上した。また、累積認識率もタスク向き文法、一般向き文法のいずれにおいても4%~5%の向上がみられ、ファジー学習法の有効性が確認された。

表3　TDNN-LRを用いた278文節認識率(%)

文法	学習法	top-1	top-3	top-5
タスク向き	従来学習法+McClelland	70.9	86.3	91.7
約500rules	Fuzzy学習法+M.S.E.	81.7	92.8	95.7
一般向き	従来学習法+McClelland	64.7	80.9	85.3
約1500rules	Fuzzy学習法+M.S.E.	70.9	87.1	91.0

従来学習法により学習されたTDNNとファジー学習法により学習されたTDNNによる音素スポッティングの例「当日の」/to:zhitsuno/を図12に示す。従来学習法による音素スポッティング結果では、致命的な誤認識につながりやすい脱落誤り(/ts/, /n/)が良くみられ

る。これに対し、ファジー学習法による音素スポッティング結果では、脱落誤りはあまりみられず、そのかわりに挿入誤りが多くみられる。しかし、挿入誤りが起こっている区間においては、正解の音素も出力されており致命的な誤りにはなりにくい。実際、一般向き文法を用いたTDNN-LR音声認識において、ファジー学習法によるTDNNスポッティング結果に対しては累積認識第2位に(第1位は「当日を」/to:zhitsuo/)認識されているのに対し、従来学習法に対しては累積認識第5位で認識されている。

5. むすび

本稿では、音素識別ニューラルネットワークのためのファジー学習法を提案した。/b,d,g,m,n,N/識別TDNNおよび日本語18子音識別TDNNを用いた音素識別実験を行った。さらに、25音素識別TDNNを用いたTDNN-LR連続音声認識システムによる文節認識実験を行った。いずれの実験においても、従来学習法よりもファジー学習法において認識率の改善が得られた。特に、TDNN-LRを用いた文節認識実験では、第1位認識率が70.9%から81.7%に改善され、第5位累積認識率は91.7%から95.7%に改善された。また、代表サンプルを用いることによってファジー学習に必要な教師信号を求めるための莫大な計算量削減もできることを示した。

謝辞

本研究の機会を与えて下さったATR自動翻訳電話研究所樽松明社長、前音声情報処理研究室鹿野清宏前室長、討論をして頂いた新旧ATRの諸氏およびCMUの諸氏に感謝します。加えて、高速ニューラルネットワーク学習プログラム"Dynet"を提供して下さったPatrick Haffner氏、"TDNN-LR"を提供して下さった鷹見淳一氏、甲斐充彦氏に深謝いたします。

文献

[1] D.E.Rumelhart, J.L.McClelland: "Parallel Distributed Processing ; Explorations in the Micro Structure of Cognition", MIT Press, (1986).

[2] R.P.Lippmann: "An Introduction to Computing with Neural Nets", IEEE ASSP Mag. pp.4-22 (1987-4).

[3] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano and K.Lang: "Phoneme Recognition Using Time-Delay Neural Networks", IEEE Trans. Acoust., Speech, Signal Processing, vol.37, pp.328-339 (1989-3).

[4] 鈴木紳, 河原英紀: "平均曲率を用いた神経回路網の評価基準について", 信学技報 NC89-103 (1990-3).

[5] 南泰浩, 田村露一, 沢井秀文, 鹿野清宏: "入力層・中間層におけるベクトルの近傍情報を利用したTDNN出力の平滑化", 音学講論, 1-3-18, pp.35-36 (1990-3).

[6] 川端豪: "k-近傍内挿学習による音韻認識", 音学講論, 2-p-21, pp.161-162 (1990-3).

[7] J.Takami and S.Sagayama: "Phoneme Recognition by Pairwise Discriminant TDNNs", ICSLP'90, S16.5, pp.677-680 (1990-11).

[8] 中村雅己, 田村露一: "ニューラルネットによる音韻フィルタ", 音学講論, 2-p-24, pp.167-168 (1990-3).

[9] 小森康弘, A.H.Waibel, 嵯峨山茂樹: "音素識別ニューラルネットにおけるファジー学習法", 音学講論, 1-5-15, pp.33-34 (1991-3).

[10] 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巻, 10号, pp.747-754 (1988-10).

[11] M.Tomita: "Effecient Parsing for Natural Language: A Fast Algolithm for Pratical Systems", Kluwer Academic Publishers (1986).

[12] H.Sawai: "The TDNN-LR Large-Vocabulary and Continuous Speech Recognition System", ICSLP'90, S31.4, pp.1349-1352 (1990-11).

[13] M.Miyatake, H. Sawai, Y.Minami, K.Shikano: "Integrated Training for Spotting Japanese Phonemes Using Large Phonemic Time-Delay Neural Networks", IEEE, ICASSP'90, S8.10, pp.449-452 (1990-5).