TR-I-0098

ベクトル量子化話者適応の時間遅れ神経回路網による音韻認識への適用

VQ-based Speaker Adaptation Applied to Time Delay Neural Network Phoneme Recognition

中村　哲，鹿野清宏
Satoshi NAKAMURA, Kiyohiro SHIKANO

1989. 8

概要

ベクトル量子化話者適応アルゴリズムを時間遅れ神経回路網(TDNN)を用いた音韻認識へ適用する。TDNNへの適用にあたってはTDNNの入力パラメータの比較、ベクトル量子化を用いた場合のTDNNの構成の検討を行う。入力パラメータとしては、FFT、LPC分析によるスペクトルとケプストラム、自己相関係数の比較を行う。また、ベクトル量子化を用いた場合のTDNNの構成としてパラメータ入力、ベクトル量子化の符号を入力とするTDNNを検討し話者適応化を適用する。音韻バランス216単語、重要語5240単語、国際会議申し込みに関する会話文を用いて有声破裂音/b,d,g/の認識実験を男女計4名の話者について行った。この結果、(1)TDNNの入力パラメータの形式については、周波数領域のパラメータが優れている。周波数領域で表現されていれば、FFTでも、LPCの分析方法でも大差がない。(2)ベクトル量子化をTDNNに適用する場合、符号列入力形式のTDNNよりもパラメータを入力とするTDNNが優れている。ベクトル量子化としては、ファジィベクトル量子化を用いることでベクトル量子化による認識率の劣化を改善できる。(3)ベクトル量子化話者適応アルゴリズムをTDNNに適用した結果、男性間と男女間の平均で78.7%の認識率が得られ、話者適応化が有効に行えることが明かとなった。

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

目次

1. まえがき

近年、多層神経回路網の研究がバックプロパゲーション学習[1-3]や、計算機の進歩などを契機に盛んになり、種々の分野に適用され有効性が確かめられている。音声認識の分野においても各種の神経回路網モデルが提案され試みられている[4-7]。本論文では、これらのうち文献(7)で提案された時間遅れ神経回路網(以下TDNNと呼ぶ)に注目する。TDNNは、一定長の時間周波数パターンを入力とし、4層の神経回路網を構成するものである。また、時間方向に一定の単位で同じ重みを持つ構成になっているので時間方向の不変性と神経回路網のフリーパラメータの削減を同時に実現し、学習も比較的容易にできる特徴がある。また、実際に音声へ適用した場合、特定話者の音韻単位の認識率は従来提案されている方法よりかなり高いことが文献(7)において報告されている。しかし、TDNNへの入力パラメータ表現方法の最適性、不特定話者への対処方法については検討されていない。不特定の話者の認識については、TDNNが学習に多くの学習データを必要とするため未知話者のTDNNをそのつど学習することは困難であり、また不特定の話者でTDNNを学習する場合は識別境界が曖昧になって認識率が劣化することが予想される。このため、未知話者の音声を標準話者に話者適応化を行って写像したのち認識を行うのが実際的であると考えられる。本論文では、筆者らが既に提案したベクトル量子化話者適応化アルゴリズム[8-11]をTDNNに適用し、不特定話者への対処を試みる。また、話者適応化アルゴリズムの適用に際し、TDNNの入力パラメータの比較、ベクトル量子化を行った場合のTDNNの構成法の検討を行う。

ベクトル量子化話者適応アルゴリズムは、ベクトル量子化を音声の離散表現とみなして各話者の音声を有限の離散表現で表し、これらの離散点の対応付けを学習により求め話者適応化を行う方法である。筆者らはファジィベクトル量子化[15,16]を導入することにより高精度化を行い[11]、HMMモデルによる音韻認識に有効であることを示した[11]。このアルゴリズムは、入力音声を標準話者の音声に変換する前処理と位置づけられるので、TDNNに対しても容易に適用が可能である。本論文では、ベクトル量子化話者適応アルゴリズムのTDNNへの適用方法の検討を中心にTDNNにおける入力パラメータの評価、ベクトル量子化の適用、発声様式を変化させた場合の評価、異話者の音声入力時の評価と話者適応化の適用について評価実験を行い、ベクトル量子化話者適応化の適用の有効性を明らかにする。

2. TDNNの入力パラメータの評価

2.1. TDNNの構成

TDNNの構成を図1に示す。図は、文献(7)において提案された有声破裂音/b,d,g/の音韻認識のための4層構成のTDNNである。入力は、メルスケールで配置された16チャンネルのフィルタバンク出力の15フレーム(150msec)分であり、入力層は、16×15の240ユニットにより構成される。第一隠れ層は、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、第二隠れ層は、周波数方向に9ユニット、時間方向に3ユニットの27ユニットにより構成される。また、入力層から第一隠れ層への接続は入力の3フレーム(48ユニット)が第一隠れ層の1フレーム(8ユニット、実際には時間軸のフレームとは一意に対応しないが便宜上フレームと呼ぶ)につながっており、入力層を1フレームずつシフトする毎に第一隠れ層の次のフレームに接続するようになっている。シフト不変性とフリーパラメータの削減の目的で入力層の3フレームから第一隠れ層の1フレームまでの接続における重みは、シフトしても等しくなるように設定される。第一隠れ層から第二隠れ層への接続も同様に第一隠れ層の5フレーム(40ユニット)が第二隠れ層の1フレーム(3ユニット)につながっており、シフトに際して重みが等しくなるように設定する。この第二隠れ層の3ユニットは、識別する3つの音韻に対応している。最終段の第二隠れ層から出力層への接続は出力層のユニットからそれぞれの音韻に対応する第二隠れ層のユニットへ9フレーム分全部に同じ重みで接続されている。このようにして時間方向へのシフト不変性を実現する構成となっている。

2.2 入力パラメータ

音声認識における音声の特徴表現方法については、従来から多くの研究がなされて各種のパラメータ表現方法が提案されている。しかし、実際には認識方法に依存して最適なパラメータが決められる事が望ましい。文献(7)では、聴覚特性を考慮してメルスケールのフィルタバンクの出力が用いられているが、他のパラメータに対しては検討が行われていない。そこで、TDNNの入力パラメータの比較を行い、TDNNにおける音声の特徴の抽出のし易さについての検討を行う。本論文では、次に示す6種類のパラメータに関する比較を行なう。

(1) FFT-Mel 16ch フィルタバンク
FFTを行ったのち、図1に示されるメルスケールの16chフィルタバンクの出力を求める。

(2) LPC-Mel 16ch フィルタバンク
14次LPC分析を行った後、LPCパラメータから256点FFTでLPCスペクトル包絡を求め、図1に示されるメルスケールの16chフィルタバンクの出力を求める。

(3) LPC- 16ch フィルタバンク
14次LPC分析を行った後、LPCパラメータから256点FFTでLPCスペクトル包絡を求め、リニアスケールの16chフィルタバンクの出力を求める。

(4) LPC ケプストラム
14次LPC分析を行った後、16次までのLPCケプストラムを求める。

(5) LPC-Mel ケプストラム

16次までのLPCケプストラムを求めた後、64次まで外挿を行った後、次式で示すBilinear Transform[12]により16次までのメルケプストラムを求める。

$$H(z)=\frac{1-a\cdot z^{-1}}{z^{-1}-a} \qquad (a=0.4)$$

(6)自己相関係数

16次までの自己相関係数を求める。

2.3 分析条件及び実験条件

分析条件を表1に示す。分析周期は、2フレームを平均することにより10msec分析周期とみなす。1、2、3については、フィルタバンクの出力の対数をとったのち、15フレームのスペクトルの最大値と最小値が-1,+1になるように正規化を行う。認識実験は、日本語の有声破裂音/b,d,g/を用いて行なう。TDNNの学習は、5240単語のデータベースの偶数番目を用いて行い、奇数番目を用いて認識実験による評価を行なう。学習、認識実験とも音韻は目視によって与えられた境界を用い、クロージャーの終点がTDNNの中心になるように入力する。話者は、男性2名(MAU,MHT)、女性2名(FSU,FFS)である。また、TDNNの学習は、バックプロパゲーション法であり、文献(13,14)で提案されているアルゴリズムを用いて高速化をはかる。

2.4 結果

表2に実験結果を示す。表は、話者4名の特定話者での音韻認識率を表し、表中closedは学習データに対する認識率を示し、openは評価用未知データに対するものである。学習データに対しては、ほとんど100%の認識率が得られている。学習時のバックプロパゲーションは、50回まででほとんど収束するが表の値は、500回中の最も高い認識率を示す。

実験の結果、次のことが明かとなった。

- 4名の話者の平均認識率によるパラメータの比較では、LPC-Mel スペクトル、FFT-Mel スペクトル、LPC ケプストラム、LPC-Mel ケプストラム、自己相関係数の順に認識率が高い。
- (1)(2)(3)の周波数領域のパラメータと(4)(5)(6)の時間領域のパラメータの比較では、周波数領域の方が認識率が高い。
- 周波数軸のスケーリングに関しては、周波数領域ではメルスケールの方が、ケプストラムではリニアスケールの方が有声破裂音/b,d,g/に対し認識率が高い。
- FFTのスペクトルとLPCのスペクトルの比較では0.2%LPCスペクトルの方が認識率が高い。

以上のことから、TDNNの入力パラメータとしては周波数領域のパラメータが適していることがいえる。FFTとLPCスペクトルの比較では、有声破裂音/b,d,g/に対する音韻認識率ではLPCの方が若干高い認識率であったものの有意な差ではなかった。入力層から出力層へ任意の写像が実現できればどのパラメータも同じ情報を持つので同じ認識率になると予想されるが、本実験ではTDNNの隠れ層と隠れ層のユニットの数が不十分で任意の写像が実現されず、このため音韻がカテゴリーとして比較的まとまった領域を構成する周波数領域のパラメータが優位な認識率を示すものと考えられる。

3. 話者適応化

本章では、ベクトル量子化話者適応アルゴリズムの概略とTDNNへの適用について述べる。

3.1 ベクトル量子化話者適応アルゴリズム[8-11]

ベクトル量子化話者適応アルゴリズムはベクトル量子化を音声の離散表現と見なし、各話者の音声を有限の離散表現で表し、これらの離散点の対応付けを学習により求め話者適応化を行う方法である。図2にこのアルゴリズムの構成を示す。基本アルゴリズムは次の2つのステップにより構成される。

(1)学習

1) 未知話者の学習用音声を用いてベクトル量子化コードブックを生成する。
2) 未知話者のコードブックを用いて未知話者の学習用音声のベクトル量子化を行う。
3) 予め準備した標準話者の学習単語の標準パターンとDTWを行い、同一単語間における最適パスを求める。
4) DTWの最適パスにしたがってコードベクトルの対応回数を求め、対応付けヒストグラムを求める。
5) 対応付けヒストグラムの値を重みとして、新しく標準話者の空間に写像されたコードベクトルを求める。
6) 5)のコードベクトルを未知話者のコードベクトルとして、ステップ1)-5)を繰り返し、スペクトル歪が収束するまで繰り返す。最終的に得られた5)に於けるコードベクトルより成るコードブックを変換コードブックと呼ぶ。

(2)適応化

1) 未知話者の入力音声を未知話者のコードブックでベクトル量子化する。
2) スペクトルパラメータを直接用いる認識方式や音声合成における声質変換では、入力音声をベクトル量子化したのち学習で求めた変換コードブックを用いて標準話者の空間にスペクトルを写像することにより話者適応化を行う。

離散HMMモデルではベクトル量子化の符号列を入力とするので未知話者の音声をベクトル量子化した後、学習で求めたコードベクトルの対応付けヒストグラムを用いて標準話者のベクトル量子化の符号列に変換するすることにより話者適応化を行う。さらに、スペクトル、パワーを独立に取り扱うセパレートベクトル量子化[9]、入力ベクトルを既存のコードベクトルの凸結合で表すことによりベクトル量子化歪を低減するファジィベクトル量子化、ファジィベクトル量子化で精密に表された入力ベクトルを離散点のコードベクトルの対応付けを頼りに連続的に標準話者の空間に写像するファジィマッピング、ファジィベクトル量子化のファジィ級関数を確率と見なして精密なコードベクトルの対応付けを求めるファジィヒストグラムを用いた話者適応化[10]を適用する。

3.2 ベクトル量子化による影響

図2に示されるように話者適応化はベクトル量子化を経て実現されるが、ベクトル量子化をTDNNへ適用するにあたり次の2通りの接続方法を比較する。

(a)TDNN-1

入力をベクトル量子化し、復号化したスペクトルを求め通常のTDNNへ入力する方法。

(b)TDNN-2

入力をベクトル量子化し、ベクトル量子化の符号列を、符号列を入力とするように設計したTDNNへ入力する方法。

符号列入力のTDNNを構成することにより、符号列レベルで話者適応化された入力の認識が可能になる。ベクトル量子化話者適応アルゴリズムでは、HMMへの適用において符号列の写像を行なう場合により効果的であることが示されているが[11]、TDNNにおいても写像された符号列を入力とすることで話者適応の認識精度を改善できることが期待できる。本節では、まず上記2種類の方法の特定話者認識における認識性能の劣化の程度、および改善方法について述べる。

3.2.1 方法(a) [スペクトル入力]

方法(a)では、1で述べた通常のTDNNであるTDNN-1を用いる。入力パラメータは(2)のLPCスペクトルを用いる。また、ベクトル量子化には、通常のベクトル量子化およびファジィベクトル量子化を用いる。ファジィベクトル量子化は次の(1)式のX'により入力ベクトルを量子化する方法でファジィ級関数は(2)式のように求められる[15,16]。

$$X' = \frac{\sum_{i=1}^{k} u_i^m \cdot v_i}{\sum_{i=1}^{k} u_i^m} \qquad (1)$$

$$u_i = \frac{1}{\sum_{j=1}^{k} \left(\frac{d_i}{d_j} \right)^{\frac{1}{m-1}}} \qquad (2)$$

ここで、u_iは級関数の値であり

$u_i \in [0,1]\ \forall_i$

$d_i = \| x - v_i \|^2$　　$\|.\|^2$はユークリッド距離

v_iはコードベクトル

mはファジィネス　$m = 1.6$

kはk-近傍数の値で6として用いる。

本論文では通常のベクトル量子化とファジィベクトル量子化を区別するために通常のベクトル量子化をhardベクトル量子化と呼ぶ。また、ベクトル量子化は、コードブックサイズを256とし、音韻バランス216単語の前半100単語から5000フレームを抽出してコードブックを生成して用いる。本節では、次の5つの比較実験を行なう。実験条件は1章と同じで、有声破裂音/b,d,g/に対する3名の話者の特定話者の実験とする。

(3-1) ベクトル量子化を用いず入力のスペクトルそのままで学習し、認識する。

(3-2) ベクトル量子化を用いないデータで学習し、hardベクトル量子化を通して得られたスペクトルをTDNNへ入力し認識する。

(3-3) ベクトル量子化を用いないデータで学習し、ファジィベクトル量子化を通して得られたスペクトルをTDNNへ入力し認識する。

(3-4) hardベクトル量子化を通して得られたスペクトルを用いて学習し、hardベクトル量子化を通して得られたスペクトルをTDNNへ入力し認識する。

(3-5) ファジィベクトル量子化を通して得られたスペクトルを用いて学習し、ファジィベクトル量子化を通して得られたスペクトルをTDNNへ入力し認識する。

表3に結果を示す。認識率は評価データに対するものである。この結果、次のことが明かとなった。

- ベクトル量子化を用いずに学習したTDNNに対しベクトル量子化を通したスペクトルを入力するとベクトル量子化のスペクトル歪により認識率が約10%低下する。hardベクトル量子化とファジィベクトル量子化の比較では、話者により若干異なるが平均でファジィベクトル量子化の方が高い認識率が得られる。

- ベクトル量子化を通したデータでTDNNの学習を行い認識を行った場合は、認識率が大幅に改善される。hardベクトル量子化を用いた場合約7%、ファジィベクトル量子化を用いた場合約8%の改善が得られる。ファジィベクトル量子化を用いた場合は(3-1)に比べて約2%の差となる。

以上のことから、ベクトル量子化を用いる場合はベクトル量子化を通したパラメータを用いてTDNNを学習すること、ファジィベクトル量子化を適用することにより認識率の劣化をかなり抑えられることが示された。

3.2.2 方法(b) [符号列入力]

方法(b)では、TDNNの入力にベクトル量子化の符号列を用いる。そのため、本論文では、1フレームの特徴量の次元をTDNN-1の16から257とするTDNN-2を構成する。これは、ベクトル量子化のコードブックの大きさの256に、単語内で正規化したパワーの項を加えたものである。第一隠れ層、第二隠れ層、出力層の構成はTDNN-1とまったく同じで入力層のみ異なる構成となっている。TDNN-2の評価を行うため3名の話者を用いて特定話者の音韻認識実験を行なう。実験条件は、1章と同じである。また、ベクトル量子化としては、hardベクトル量子化とファジィベクトル量子化を用いる。

実験結果を表4に示し、結果を次にまとめる。

- ベクトル量子化を用いない場合との比較では、hardベクトル量子化の場合8.5%、ファジィベクトル量子化の場合3.8%の劣化となる。
- 方式(1)のTDNN-1との比較では、TDNN-2の方が認識率が低くhardベクトル量子化で4.5%、ファジィベクトル量子化で約2%の差がある。
- TDNN-2においてもファジィベクトル量子化は効果的でhardベクトル量子化に比べ4.7%認識率が改善される。

以上のことから符号列入力のTDNN-2の場合は、TDNN-1に比べて2%低い認識率となることがわかった。TDNN-2では入力が0または1の離散値であるため、学習時において汎化が起こりにくいことが認識率の低い原因と思われる。この問題は、表4に示されるようにファジィベクトル量子化を用いることによりかなり改善されている。

3.3 発声速度変化による影響

本節では、TDNNの構成方法の比較、およびベクトル量子化を通すことによる影響を、話者内変動、特に発声速度が変わった場合に対する汎化能力という観点から検討する。比較はTDNN-1、TDNN-2に対してベクトル量子化を用いない場合、ファジィベクトル量子化を行った場合の比較を行う。TDNNの学習はいずれの場合も単語から切り出した音韻を用いる。評価用のデータとしては、次の3通りの発声様式のデータ[20]から切り出された音韻を用いる。

(1) 単語発声(実験1と同じ)

(2) 国際会議に関する会話文を自立語+付属語程度の単位に区切って発声したもの(ATRデータベースDSB)

(3) 国際会議に関する会話文を、実際の会話と同じように連続に発声したもの(ATRデータベースDSC)

(2)(3)の区切り方の例を次に示す。

(2)もしもし･こちら･通訳･電話･国際･会議･事務局です。

(3)もしもし･こちら通訳電話国際会議事務局です。

実験条件は、1章と同じで、話者3名についての評価を行なう。表5に実験結果を示す。参考に話者1名に関するHMMモデルの音韻認識率を示す。HMMモデルは文献(17-19)に示される3状態で、スペクトル、パワー、スペクトルの動的特徴のマルチコードブックで、音韻の継続長制御を行ったものである。

実験結果を、次にまとめる。

- (2)のDSBを入力した場合、ベクトル量子化を用いないTDNN-1(noVQ)で4.6%認識率が低下する。ファジィベクトル量子化を用いたTDNN-1(FZVQ)で約8%、TDNN-2(FZVQ)で約10%低下する。さらに、(3)のDSCを入力した場合、TDNN-1(noVQ)で約10%、TDNN-1(FZVQ)で約11%、TDNN-2(FZVQ)で約14%の低下となる。
- ベクトル量子化による認識率は、(2)のDSBを入力した場合の劣化が大きい。
- TDNN-1とTDNN-2の比較では、いずれの場合もTDNN-1の方が認識率が高く(2)のDSBの場合4.0%、(3)の場合4.7%の差がある。
- 1名の話者におけるHMMとの比較では、平均でTDNN-1はHMMより2.4%認識率が高く、TDNN-2はHMMより1.0%低い認識率となる。

以上より、発声様式の変化から受ける影響は非常に大きく、特にベクトル量子化の場合には(2)の単語発声から会話文発声にかわることによる発声様式の変化が大きく影響することがわかった。

TDNN-2はその構成上HMMモデルと類似しているがHMMモデルの方が高い認識率を示している。一定の学習データ量ではフリーパラメータ数が多いほど学習が難しく、汎化の能力が低くなる。TDNN-2の場合はフリーパラメータの数が多い上に、符号列を入力とする構成のため汎化が起こり難くなっており、この影響が発声様式の変化に非常に顕著に現れている。

3.4 話者適応化

3.1に述べた話者適応化を行う。実験条件は1章と同じで、3名の話者を用いる。男性話者1名MAUを未知話者とし、男性話者MHT、女性話者FSUを2名の標準話者として、話者適応化無しの場合、話者適応化を行った場合、標準話者の特定話者認識の場合の比較実験を行う。また、TDNN-1,TDNN-2について、hardベクトル量子化、ファジィベクトル量子化を用いた場合の比較を行う。ベクトル量子化話者適応アルゴリズムにおける特徴量としては、LPCのパラメータを用い、話者適応のための学習には音韻バランス216単語の前半100単語を用いる。

表6に結果を示す。参考に、文献(11)の方法を用いて同じデータについてHMM話者適応化を行った結果を示す。ただし、TDNN-1における「話者適応無し」および「特定話者」は、ベクトル量子化を一切用いずに未知話者の音声を標準話者の音声で学習したTDNN-1へ入力するもの、TDNN-2の場合はファジィベクトル量子化を通した未知話者の音声をファジィベクトル量子化を通して学習したTDNN-2へ入力した場合の結果である。

結果を次にまとめる。

- TDNN-1とTDNN-2の比較では、TDNN-1が認識率が高くhardベクトル量子化の場合18.3%、ファジィベクトル量子化の場合13.5%の差となる。
- TDNN-1のベクトル量子化手法の比較では、ファジィベクトル量子化を用いた方が男性間と男女間の平均で約2%認識率が高い。
- ファジィベクトル量子化を用いて学習したTDNN-1を用いた場合が認識率が最も高く話者適応なしの場合に比べて、17.4%の改善が得られる。また、HMM話者適応化との比較では、HMM話者適応化の方が男性間と女性間の平均で0.8%高い程度でほぼ同等の認識率が得られる。

TDNN-2の認識率の低さの原因としては、3.3節の発声速度変化で述べた汎化能力の低さがあげられる。異話者間の認識においては、話者内の発声変動よりもかなり大きな相違が存在すると考えられるので汎化能力の低いTDNN-2は、話者適応無し、話者適応化のいずれの場合も他の方式よりもかなり低い認識率を示している。逆に、TDNN-1は汎化能力が高く話者適応なしの場合に最も高い認識率を示している。

以上の結果から、ファジィベクトル量子化を通して学習したパラメータ入力のTDNN-1を用いて話者適応を行うことにより有声破裂音に対し78.7%の音韻認識率が得られ、ベクトル量子化話者適応アルゴリズムのTDNNへの適用の有効性が明かとなった。

4. まとめ

本論文では、ベクトル量子化話者適応アルゴリズムを時間遅れ神経回路網(TDNN)へ適用した。適用にあたってTDNNの入力パラメータの表現形式、およ

びベクトル量子化を行う場合のTDNNの構成、改善方法について検討を行なった。有声破裂音/b,d,g/に関する評価実験を行い、次のことが明かとなった。(1)TDNNの入力パラメータの形式については、音韻のカテゴリー化が容易である周波数領域のパラメータが優れている。周波数領域で表現されていれば、FFTでも、LPCのパラメータでも大差がない。周波数領域では、聴覚特性に近いメルスケールを用いた方が僅かに優れている。(2)ベクトル量子化をTDNNに適用する場合、符号列入力形式のTDNNよりもパラメータを入力とするTDNNが優れている。ベクトル量子化としては、ファジィベクトル量子化を用いることでベクトル量子化の導入による認識率の劣化を改善できる。(3)ベクトル量子化話者適応アルゴリズムをTDNNに適用した結果、パラメータ入力のTDNNが優れており男性間と男女間の平均で78.7%の認識率が得られた。これらの結果から、TDNNに対しベクトル量子化話者適応が有効に適用できることが明らかとなった。今後の課題としては、スペクトルの時間変化の線形回帰係数やパワー変化量などの付加情報を用いた認識率の改善、話者による発声様式の適応化などを行ってさらに話者適応化の改善を行う予定である。

文献

(1) D.E.Rumelhart,J.L.McClelland, Parallel Distributed Processing ; Explorations in the Microstructure of Cognition, Vol. I and II. Cambridge, MA : M. I. T. Press, 1986

(2) R.P.Lippmann, "An introduction to computing with neural nets," IEEE ASSP Mag., vol. 4, Apr. 1987.

(3) D.E.Rumelhart, G.E.Hinton, R.J.Williams, "Learning representations by back-propagating errors," Nature, vol.323, pp.533-536, Oct. 1986

(4) H.Bourlard, C.J. Wellekens, "Multilayer perceptrons and automatic speech recognition," in Proc. IEEE Int. Conf. Neural Networks, June 1987.

(5) R.P.Lippmann, B.Gold, "Neural-net classifiers useful for speech recognition," in Proc. IEEE Int. Conf. Neural Networks, June 1987.

(6) D.J.Burr, "A neural network digit recognizer," in Proc. IEEE Int. Conf. Syst., Man, Cybern., Oct. 1986

(7) A.H.Waibel, "Phoneme Recognition Using Time-Delay Neural Networks," 信学技報SP-87100(1987-12)

(8) K.Shikano,K.F.Lee,R.Reddy, "Speaker Adaptation through Vector Quantization," ICASSP' 86

(9) 中村、鹿野、"ベクトル量子化を用いたスペクトログラムの正規化," 信学技報 SP-17(1987-06)

(10) 中村、鹿野、“ファジィベクトル量子化を用いたスペクトログラム正規化の検討,” 信学技報SP87-123 (1988-02)

(11) 中村、花沢、鹿野、“ベクトル量子化話者適応アルゴリズムのHMM音韻認識による評価,” 信学技報SP88-106(1988-12)

(12) A.V.Oppenheim, D.H.Johnson, “Discrete Representation of Signals,” Proceeding IEEE, 1972, 60 NO.6, pp. 681-691

(13) P.Haffner, A.Waibel, K.Shikano, “Fast Back-Propagation Learning Methods for Neural Networks in Speech,” 音講論集2-P-1(1988-10)

(14) P.Haffner, “Dynet, a Fast Program for Learning in Neural Networks,” ATR Technical Report TR-I-0059

(15) E.Ruspini, “Numerical Methods for Fuzzy Clustering,” Inf.Sci.,Vol.2(1970)

(16) H.P.Tseng, M.J.Sabin, E.A.Lee, “Fuzzy Vector Quantization Applied to Hidden Markov Modeling,” ICASSP’87

(17) 花沢、川端、鹿野, “Hidden Markovモデルを用いた日本語有声破裂音の識別,” 音講論集1-5-10(1987-10)

(18) 花沢、川端、鹿野, “HMM音韻認識におけるセパレートベクトル量子化の検討,” 音講論集2-P-21(1988-10)

(19) 花沢、川端、鹿野, “HMM音韻認識におけるモデル継続長の制御手法,” ATR Technical Report TR-I-0050

(20) 武田、匂坂、片桐、阿部、桑原、“研究用日本語音声データベース利用解説書,” ATR Technical Report TR-I-0028

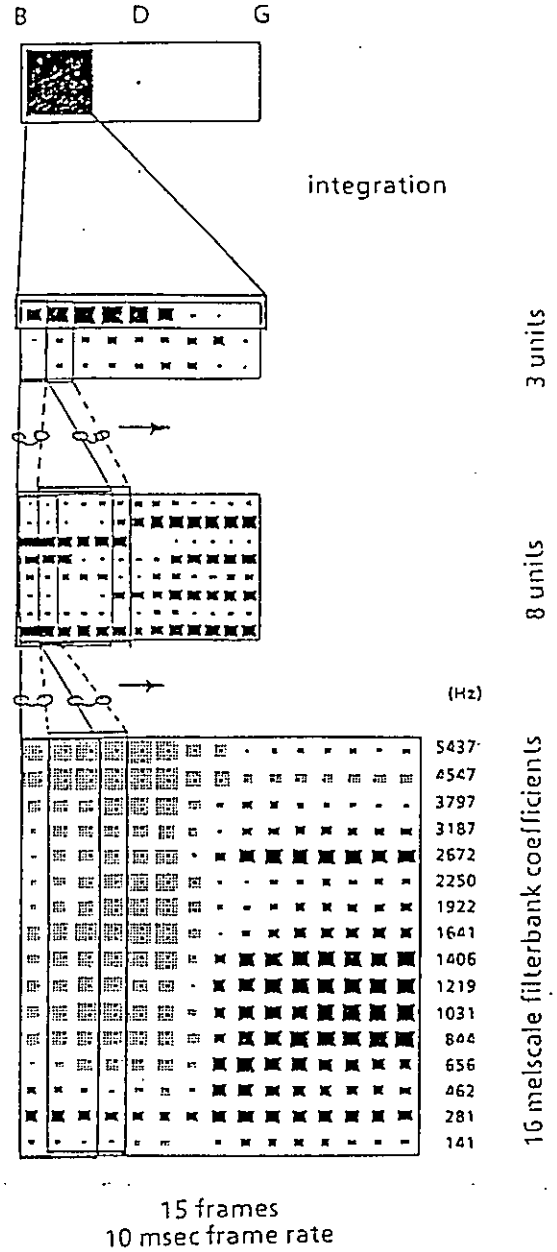

Fig.1. Architecture of TDNN

図1.TDNNの構成

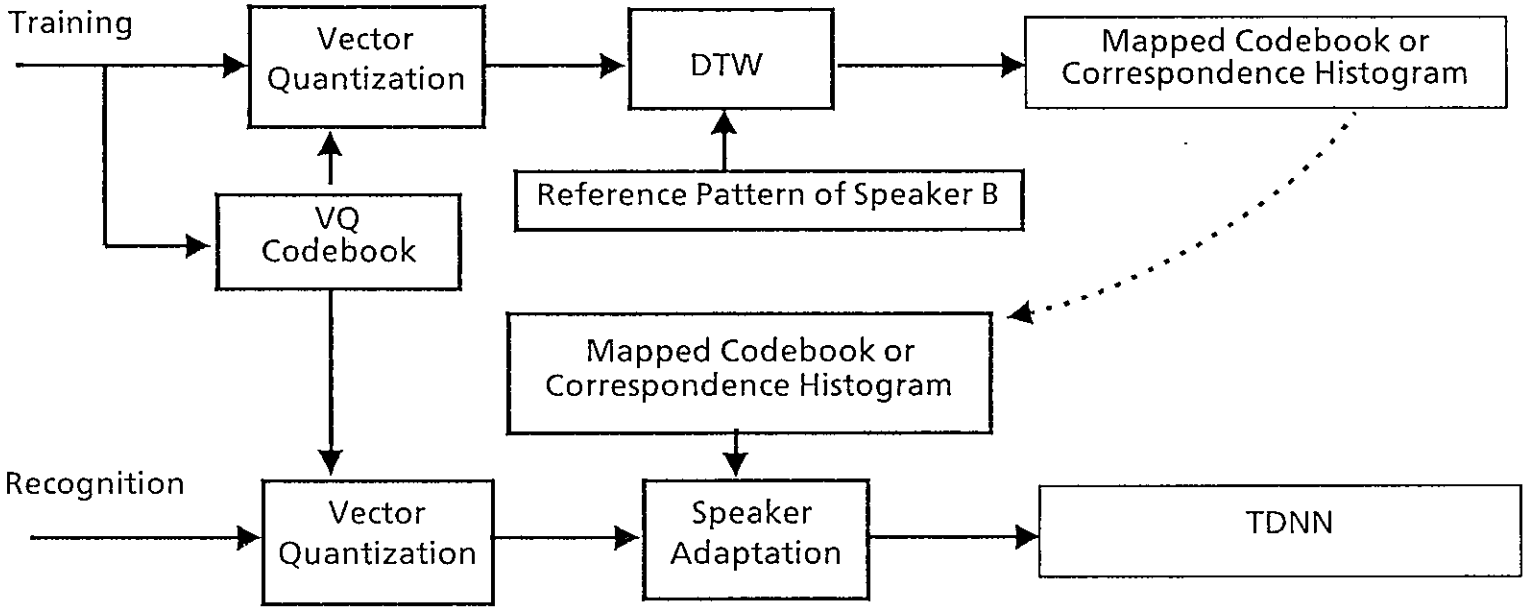

Fig.2. Block diagram of TDNN speaker adaptation
図2. TDNN話者適応化のブロック図

表1 分析条件

Table 1. Analysis Conditions

Sampling Frequency	12 kHz
Window Function	Hamming
Window Length	21.3 msec
Analysis Interval	5 msec
Analysis Method	FFT 256 points
	LPC 14 order

表2. 各種入力パラメータに対する音韻認識率 /b,d,g/

Table 2. Phoneme recognition rates vs. input parameters on TDNN /b,d,g/

	(1)FFT Mel-BPF		(2)LPC Mel-BPF		(3)LPC BPF		(4)LPC Cepstrum		(5)LPC Mel-Cepstrum		(6)Autocor-relation	
	closed	open	closed	open	closed	open	closed	open	closed	open	closed	open
MAU	99.4	98.9	100	99.1	99.9	98.0	99.9	97.9	100	97.6	99.6	95.6
MHT	99.5	99.5	100	99.2	99.7	99.1	99.4	99.0	99.1	99.3	99.9	98.7
FSU	100	95.8	100	95.4	100	95.4	100	95.1	99.7	95.0	100	86.5
FFS	100	95.8	100	97.1	100	95.9	100	95.5	99.1	93.8	100	94.5
Average	99.7	97.5	100	97.7	99.9	97.1	99.8	96.9	99.5	96.5	99.9	93.8

表3. ベクトル量子化の適用時の音韻認識率 /b,d,g/

Table 3 Phoneme Recognition Rates vs. VQ Method for /b,d,g/

	(3-1)	(3-2)	(3-3)	(3-4)	(3-5)
Training	no VQ	no VQ	no VQ	Hard VQ	Fuzzy VQ
Testing	no VQ	Hard VQ	Fuzzy VQ	Hard VQ	Fuzzy VQ
MAU	99.1	87.0	90.7	95.8	97.9
MHT	99.2	83.5	81.6	97.6	97.9
FSU	95.4	90.9	90.7	88.2	92.1
Average	97.9	87.1	87.7	93.9	96.0

表4. TDNN-2の音韻認識率

Table 4 Phoneme recognition rates using TDNN-2

	no VQ	TDNN-1		TDNN-2	
		Hard VQ	Fuzzy VQ	Hard VQ	Fuzzy VQ
MAU	99.1	95.8	97.9	93.1	95.9
MHT	99.2	97.6	97.9	95.7	96.9
FSU	95.4	88.2	92.1	79.5	89.6
Average	97.9	93.9	96.0	89.4	94.1

表5. 各種発声速度における音韻認識率 /b,d,g/

Table 5 Phoneme recognition rates for various speaking speed /b,d,g/

	(1)Word				(2)Sentence DSB				(3)Sentence DSC			
	TDNN-1 no VQ	TDNN-1 FZVQ	TDNN-2 FZVQ	HMM	TDNN-1 no VQ	TDNN-1 FZVQ	TDNN-2 FZVQ	HMM	TDNN-1 no VQ	TDNN-1 FZVQ	TDNN-2 FZVQ	HMM
MAU	99.1	97.9	95.9	96.0	92.0	90.6	86.0	87.8	84.6	84.8	82.1	-
MHT	99.2	97.9	96.9	-	95.6	92.7	85.0	-	92.9	93.3	81.6	-
FSU	95.4	92.1	89.6	-	92.9	80.0	80.3	-	85.6	76.8	76.8	-
Average	97.9	96.0	94.1	-	93.5	87.8	83.8	-	87.7	85.0	80.3	-

表6. 話者適応化音韻認識率 /b,d,g/

Table 6 Speaker Adaptation Results /b,d,g/

	TDNN-1				TDNN-2				HMM			
	NO Adap	Adap hard VQ	Adap FZVQ	Depen-dent	NO Adap	Adap hard VQ	Adap FZVQ	Depen-dent	NO Adap	Adap hard VQ	Adap FZVQ	Depen-dent
MAU→ MHT	80.0	82.3	84.3	99.2	33.1	60.3	64.8	96.9	79.2	80.0	83.1	97.6
MHT→ FSU	42.6	71.3	73.0	95.4	30.8	56.7	65.5	89.6	38.0	76.2	75.9	95.5
Average	61.3	76.8	78.7	97.3	32.0	58.5	65.2	93.3	58.6	78.1	79.5	96.6