TR-I-0314

教師なしおよび教師あり話者適応手法の検討

A Study on Unsupervised and Supervised Speaker Adaptation

丸山康　　大倉計美
Yasushi MARUYAMA　Kazumi OHKURA

1993.3.3

概要

個人差の内挿による話者適応化法がいくつか提案されている中にファジイ目的関数最小化基準による教師なし話者適応化法があり単語認識実験を通じて短時間学習において効果のあることが報告されている。本研究ではこのファジイ目的関数最小化基準による教師なし話者適応化法と声質変換法から導かれた自乗誤差最小化基準による教師あり話者適応化法の２つの話者適応化法について離散ＨＭＭによる音韻認識実験を通じて比較検討を行った。

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

目次

1 はじめに

不特定話者を対象とした大語彙連続音声認識システムの実現にとって、話者適応化法の確立は重要な課題である。今までに個人差の内挿による話者適応化法がいくつか提案されているなかにファジイ目的関数最小化基準による教師なし話者適応化法があり単語認識実験を通じて短時間学習において効果のあることが報告されている [1]。本研究ではこのファジイ目的関数最小化基準による教師なし話者適応化法と声質変換法から導かれた自乗誤差最小化基準による教師あり話者適応化法 [2] の 2 つの話者適応化法について離散HMMによる音韻認識実験を通じて比較、検討を行なう。

2 話者適応化法

ここでは自乗誤差最小化基準による教師あり適応化法とファジイ目的関数最小化基準による教師なし適応化法について説明する。

2.1 自乗誤差最小化基準による教師あり話者適応化法

本方法では、図 1 に示すように未知話者の特徴空間上にあらかじめ設定された代表点 $\{V_k\}$ $(k = 1, ..., M)$ における標準話者との個人差ベクトル (Δk) を用いて未知話者の任意の特徴ベクトル x_i を次式で変換し、標準話者の特徴空間に写像する。

$$\hat{y}_i = x_i + \sum_{k=1}^{M} W_{ik} \Delta k, \tag{1}$$

$$W_{ik} = d(x_i, V_k)^{-p} / \sum_{r=1}^{M} d(x_i, V_r)^{-p}. \tag{2}$$

但し、p は Δk の寄与を制御するパラメータであり、$p \to \infty$ のとき $\{V_k\}$ による最小距離分類で分割された部分空間毎に対応する Δk のみが加算され不連続な適応となる。また、$p \to 0$ とき Δk の平均が全空間に一様に加算される。ここで、個人差ベクトル (Δk) $(k = 1, \cdot\cdot\cdot, M)$ は、未知話者の学習用サンプルの変換ベクトル $(\hat{y}_i(r))$ とそれに対応付けされた標準話者のベクトル $(y_j(r))$ の差の総自乗誤差に関する目的関数

$$J(C(1), ..., C(N), \Delta 1, ..., \Delta M) = \sum_{r=1}^{N} |y_j(r) - \hat{y}_i(r)|^2. \tag{3}$$

を最小化することによって推定される。ここで目的関数の最小化は、DTW による時間正規化距離 $(C(r))$ の最小化と次の正規方程式による (Δk) の最小化を反復することにより与えられる。

$$\sum_{k=1}^{M} W_{lk} \Delta k = \sum_{r=1}^{N} W_{i(r)l} e_{i(r)j(r)}. \tag{4}$$

$$(l = 1, ...M)$$

ただし、

$$W_{lk} = \sum_{r=1}^{N} W_{i(r)l} W_{i(r)k}. \tag{5}$$

$$e_{i(r)j(r)} = y_{j(r)} - x_{i(r)} \tag{6}$$

こうして変換された未知話者の変換ベクトル $\hat{y}_i$ を、標準話者のコードブックで量子化し、標準話者のHMMで認識する。なお 、 ATR で提案されたコードブック補間法は、今回の入力ベクトル変換法をコードブック変換に用いたものと考えられる。すなわち、代表点としてコードブック内のベクトルそのものを候補として個人差ベクトルを推定し、$p = 1/(F-1)$ とおいて式(1)を用いて推定された個人差ベクトルをコードブック中のベクトルに内挿し更新しているとみなせ、このとき $\sum_{r=1}^{N} w_i(r)k^2$ が 0 か否かによって個人差ベクトルの推定に用いる代表点 $\{\mathrm{V}_k\}$ を決定している。

2.2 ファジー目的関数最小化による教師なし話者適応化法

教師なし適応化では、話者間の音韻に対する正確な対応の情報が利用できないため、適応化に当たり適応の良さを評価する何らかの基準が必要とされる。ここでは、適応化の基準としてファジイ目的関数を用いた教師なし話者適応化法について以下に示す。なお、本方法は単語認識実験において効果のあることが報告されている。本方法では、前節と同様のモデルを用い、図2のように標準話者の特徴空間に設定した M 個の代表点 $\{V_k\}(k = 1, ..., M)$ における未知話者との個人差ベクトル (Δk) を用いて、標準話者のコードブック (Y_j) を次式で内挿し適応化する。

$$\hat{X}_j = Y_j + \sum_{k=1}^{M} W_{ij} \Delta k, \tag{7}$$

$$W_{ij} = d(Y_j, V_k)^{-p} / \sum_{r=1}^{M} d(Y_j, V_r)^{-p}. \tag{8}$$

ここで、個人差ベクトル $(eltak)$ は適応化された大きさ L のコードブック (X_j) を用いて、学習サンプル $x_i(i = 1, ..., N)$ をファジイベクトル量子化する際に、次のファジイ目的関数

$$J_F(U_{i1}, ..., U_{iL}; \Delta 1, ..., \Delta M) \ = \sum_{i=1}^{N} \sum_{j=1}^{L} U_{ij}^F d(x_i, \hat{X}_j) \tag{9}$$

を拘束条件

$$\sum_{j=1}^{L} U_{ij} = 1 \tag{10}$$

$$R = 1/M \sum_{k=1}^{M} |\Delta k|^2 < \eta(E_x^2 - E_y^2). \tag{11}$$

のもとで最小化するように推定する。ここで、η は極端に大きな変化を与えないよう Δk 大きさを制限するパラメータである。また E_y は、Y_j 作成時の学習サンプル自身の平均量子化誤差であり、E_x は、x_i をコードブック Y_j で量子化した際の平均量子化誤差であり、これらから平均的個人差を推定し Δk の大きさを制限する。

代表点数を M に固定した場合、次の手順により目的関数は最小化される。

1. $Deltak = 0$ として U_{ij} について最小化する。

2. 1. の U_{ij} を用いて Δk について最小化する。

   この時、Δk は、正規方程式

$$\sum_{r=1}^{M}(W_{rk} + \lambda\delta_{rk})\Delta_r = \sum_{i=1}^{N}\sum_{j=1}^{L} U_{ij}W_{jk}e_{ij}, \tag{12}$$

$$(k = 1, ..., M)$$

   で与えられる。但し、

$$e_{ij} = Y_j - x_i, \tag{13}$$

$$W_{rk} = \sum_{i=1}^{N}\sum_{j=1}^{L} U_{ij}^{F}W_{jr}W_{jk}. \tag{14}$$

   である。また、R はラグランジェの未定係数の単調減少関数となることから、式(5)の条件を満たさない場合には、λ を増加させて、より小さい R に対する解を求めることができる。

3. 式(7)により Y_j を更新し、目的関数の減少率が δ 以下であれば終了し、それ以外は1. に戻る。

以上に基づく適応化アルゴリズムにおいて、代表点数を1 ～ M まで逐次増加させる段階的最小化手法を用いることによって音韻に共通する個人性から音韻に依存した微細な個人性へと適応化を進めていく。認識時には入力話者の音声を変換コードブックで量子化し、標準話者のHMMを用いて認識を行なう。

3 実験条件

評価実験として離散HMMによる /bdgmnN/ の音韻認識実験を行なった。

3.1 音声データと分析条件、並びに離散ＨＭＭの条件

今回の実験に用いた分析条件は、１２kHzサンプリング、分析窓長２１．３ms、フレーム周期9 ms、ハミング窓、プリエンファシス($1 - 0.97^{-1}$)、１４次ＬＰＣ分析である。実験に用いた話者は、男性２名で１名を標準話者(ＭＨＴ)、他方を未知話者(ＭＡＵ)とした。パラメータとして、１６次ＬＰＣケプストラム係数(CEP)およびその一次回帰係数(*Delta*CEP)、対数パワーの一次回帰係数(ΔLPOW)を用い、コードブックサイズはそれぞれ、２５６、２５６、６４とし、コードブック生成には、可能な２音韻連鎖をすべて含む音韻バランス単語２１６単語の前半１００語を用いた。音素HMMは４状態３ループのものを用い、/bdg/ については語頭と語中の２つのモデルを用いた。音素HMMの学習用データとしては、重要語５２４０単語の偶数番目の単語を、評価用データとして奇数番目の単語を用い、HMMの学習にはハードＶＱを、認識の際にはファジネス１．６、近傍数６のファジイＶＱを用いた。

3.2 適応化条件

自乗誤差最小化基準による教師あり話者適応法の条件としては、DTW は端点フリーの対称型自由度1/3 ～ 3を用い内挿パラメータは、p=1.0 とした。また、未知話者のスペクトル空間に設定する代表点は未知話者の学習サンプルから求めた。また、DTW には、次式に示す各特徴量の線形和により定義される距離値を用いた。

$$D_{ij} = D_{CEPij} + 10.0 D_{\Delta LPOW} + 60.0 D_{\Delta CEPij}.$$

ファジイ目的関数最小化基準による教師なし話者適応化法の条件としては、内挿パラメータ p = 1.0、ファジネス f = 1.5, ノルム制限 η = 1.1、収束いき値 δ = 0.05 とし、また、式(12)において計算量削減のため U_{ij} が 1 / (コードブックサイズ) より大きい項のみを加算した。また、代表点としては、サイズ M のコードブックの各コードに最も近い標準パタンのコードそのものを用いた。そして、適応はそれぞれの特徴量について個別に行なった。ただし、対数パワーの一次回帰係数のコードブックに関しては、標準話者のコードブックで入力話者のサンプルを量子化した際、量子化誤差の増加が見られなかったので標準話者のコードブックを変換コードブックとみなしてそのまま用いた。なお、教師あり、教師なし両方の適応化においてケプストラム、ケプストラムの一次回帰係数、対数パワーの一次回帰係数の代表点数は、ケプストラム、ケプストラムの一次回帰係数では同数、対数パワーの一次回帰係数では、ケプストラムにおける代表点数の1/4 とした。

話者適応用データとして音韻バランス単語２１６語の先頭から、５、１０、２５、１００単語を用いた。

4 実験結果

4.1 教師あり話者適応化法における代表点数の影響

自乗誤差最小化基準による教師あり話者適応化法(以下 INT-VEC と表す)における代表点数と認識率との関係を学習単語１０語の場合について調べた。CEP の代表点数を８、１６、３２、６４、１２８と変化させて認識率を調べた。その結果を図3に示す。ここで横軸は CEP の代表点数、縦軸は6音韻の平均認識率をそれぞれ表す。また、図中の NO-AD は適応なしの場合の認識率である。

4.2 教師なし話者適応化法における代表点数の影響

前節と同様にファジイ目的関数最小化基準による教師なし話者適応化法(以下 MINFZ と表す)についても多段階適応における最大代表点数と認識率の関係を調べた。学習単語１０語の場合について、CEP の最大代表点数を１、２、４、８、１６、３２、と変化させて認識率を求めた。その結果を図4に示す。ここで横軸は CEP の最大代表点数であり、縦軸は6音韻の平均認識率である。また各音韻カテゴリーごとの認識率と代表点数の関係を図5に示す。

4.3 学習単語数による認識率の変化

学習単語数による認識率を調べた。CEP の代表点数を教師あり話者適応化法では学習単語数５、１０、２５、１００語において３２、６４、２５６、２５６点、とし、教師なし適応化

法では学習単語数に関係なく最大１６点 とした。学習単語数と６音韻の平均認識率の関係を図６に示す。横軸は学習単語数、縦軸は６音韻の平均認識率を表す。また、ATR で行われているヒストグラムを用いた話者適応化法 [4] (以下 FZHMAP と表す) と、補間を用いたコードブックマッピング法 [3](以下 INT-CODE と表す) の認識率との比較を行なった。ただし、本実験結果は文献 [3] に示されているものである。

今回評価を行なう２方法においても、HMMの学習にファジネス 1.6、近傍数６のファジイＶＱを用いて文献 [3] における条件と同条件で実験を行った。その他の条件は４．３の実験と同じにした。本条件で学習単語数５、１０、２５語における６音韻の平均認識率を調べた。結果を図７に示す。ここで横軸は学習単語数で縦軸は６音韻の平均認識率である。なお、INT-VEC, MINFZ と FZHMAP, INT-CODE では用いたコードブックが異なっており適応なしにおいて INT-VEC, MINFZ で用いたコードブックの場合、認識率が１％増加している。

5　考察

教師あり話者適応化法で学習単語数を固定した場合、最適な代表点数のあることが分かる。また、学習サンプル間での DTW の距離値は、代表点数の増加に従い減少し続けた。このことよりこれは、代表点数が少ない場合は大まかな個人差しか表現せず、多くなると１代表点当たりに寄与する学習サンプル数が少なくなるので、入力話者の学習用サンプルの内容に依存した個人差が推定されてしまうためだと思われる。

教師なし適応化法では学習単語数を 10 単語に固定した場合、最大代表点数が１点の時が最も認識率が高く、その後認識率は若干低下していく。この結果は信州大学で行なった単語認識実験における結果と異なったものである。音韻カテゴリーごとの認識率を見ると、今回の教師なし適応化法は音韻によって適応の効果が異なり必ずしもすべての音韻に対して効果のあるものではないことが分かる。また、代表点数を増やすことによる認識率の低下は適応化の自由度が増えるに従って極小点に収束する可能性が大きくなることによると思われる。

学習単語数と認識率の関係をみると、教師ありの方法では学習単語数１０語で適応なしに比べ５％近い認識率の向上を示し、本話者適応化法の短時間学習における効果が分かる。教師なし話者適応化法では、学習単語数によらずほぼ一定の認識率を示しているが、適応なしと比較した場合、認識率は下がっており適応が悪影響を与えていることが分かる。しかし、先の考察でも述べた様に、認識率が全体に下がるのではなく効果のある音韻もあり、改善の可能性はあるものと思われる。

ATR で行なわれている２方法 (FZHMAP, INT-CODE) との比較で見ると、今回用いた教師あり話者適応化法 (INT-VEC) は ATR のコードブック補間法 (INT-CODE) とは類似した手法のため同程度の認識率を示している。また、ATR のヒストグラムに基づくコードブックマッピング法 (FZHMAP) より良い性能を示している。これは、FZHMAP は、少数学習単語から作成したコードブックが表す空間しか表現できないことに対して、INT-CODE と INT-VEC は学習単語の表す音声空間の特徴から入力話者の音声空間全体を補間した音声空間を推定しているためである。

6　結び

話者適応化法として自乗誤差最小化基準による教師あり話者適応化法とファジイ目的関数最小化基準による教師なし話者適応化法について, 離散ＨＭＭによる音韻認識実験を通じて比

較、検討をおこなった。その結果、男性-男性間の話者適応実験において、学習単語数１０語で教師ありの方法で 適応なしに比べ約４％の認識率の改善が得られ、今回用いた方法が短時間学習でも有効であることが示された。また、教師なしの方法においては単語認識実験とは異なる傾向を示し新たな検討が望まれる。また、 ATR で提案されたコードブックの補間による話者適応化法と自乗誤差最小化基準による適応化法がほぼ同程度の認識性能であることが確認された。今後の課題としては、各パラメータについての更なる検討、並びにパラメータの決定手法の確立、各音韻カテゴリーごとの検討、教師なし適応化法における認識精度の向上などが上げられる。また今回は男性-男性間における評価だったが男性-女性間での評価、検討も行ないたい。

謝辞

１カ月の間御指導、お世話下さった 大倉 計美さん、村上 仁一さん、 Harald Singer さん、はじめ音声情報処理研究室のみなさん、ならびに慶應義塾大学の中村 悟さん、吉沢 健一郎さんに感謝致します。

参考文献

[1] 松本、山下、西沢 ：“ファジイ目的関数最小化基準によるスペクトルの短時間教師なし話者適応化 ” 音響学会音声研資 SP88-122(1989-1).

[2] 丸山、井上、松本 ：“パワースペクトル包絡の最小自乗写像に基ずく声質変換 ” 音響学会講演論文集 1-7-24 (1989-3).

[3] 服部、嵯峨山 ：“少数学習データを用いたコードブックマッピングによる話者適応化 ” 音響学会講演論文集 1-5-23 (1991)

[4] 中村、花沢、鹿野 ：“ベクトル量子化話者適応のHMM音韻認識への適応 ” 音響学会誌, Vol.45,pp942-949 (1989).

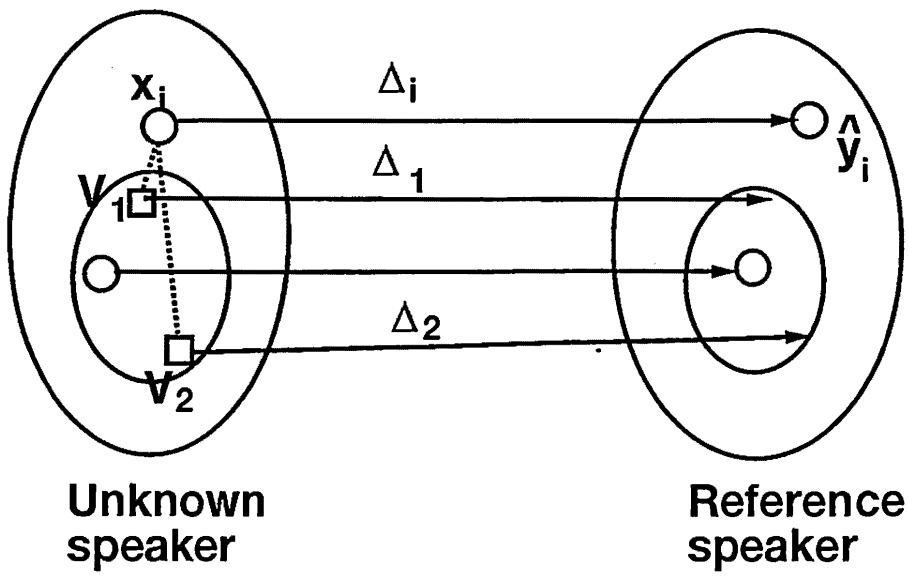

図 1: 教師あり話者適応化法の概念図

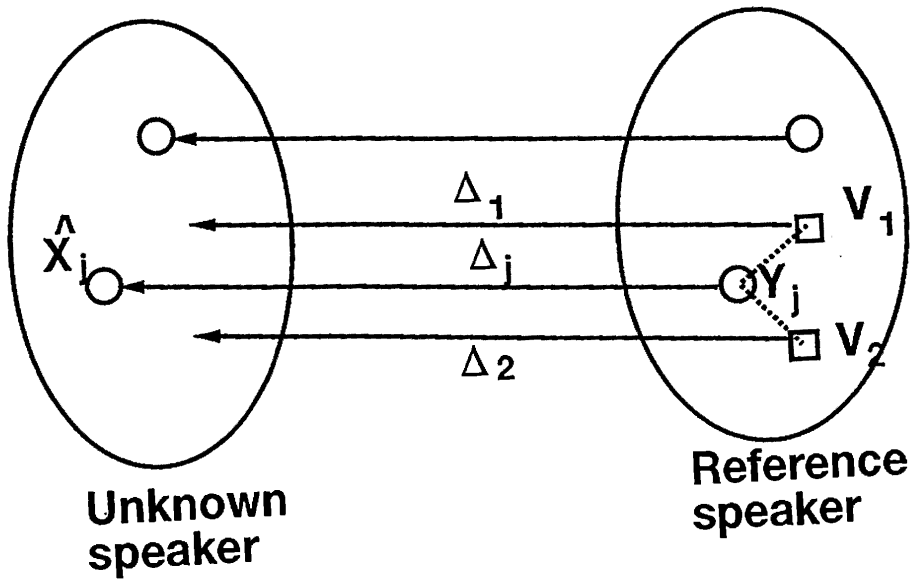

図 2: 教師なし話者適応化法の概念図

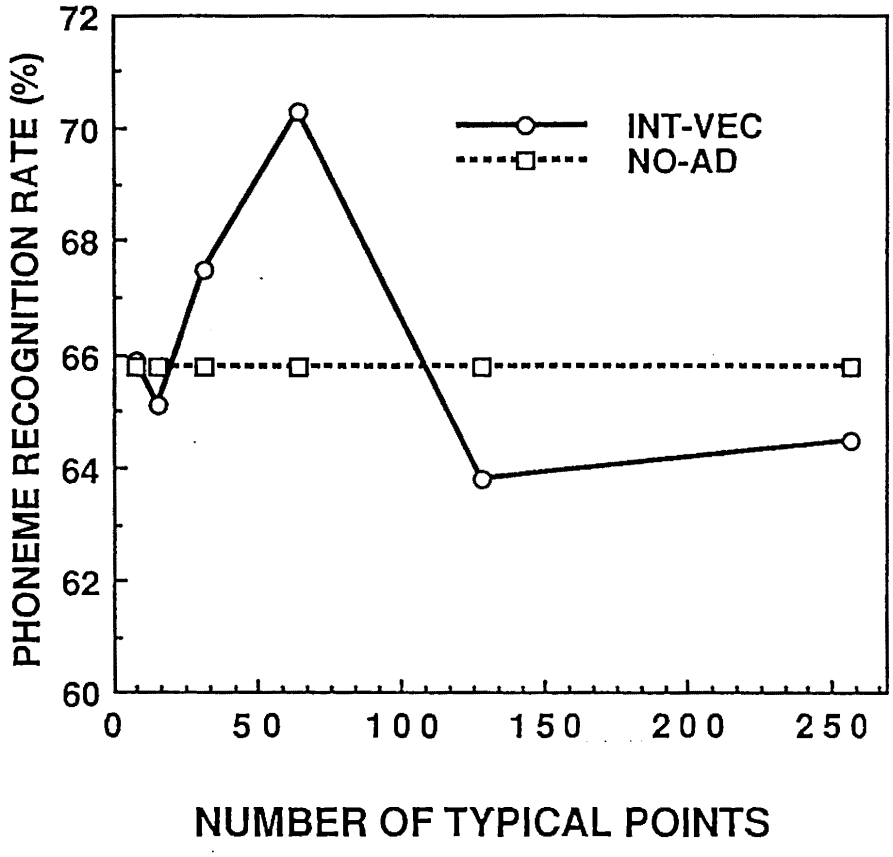

図 3: 代表点と音素認識率の関係

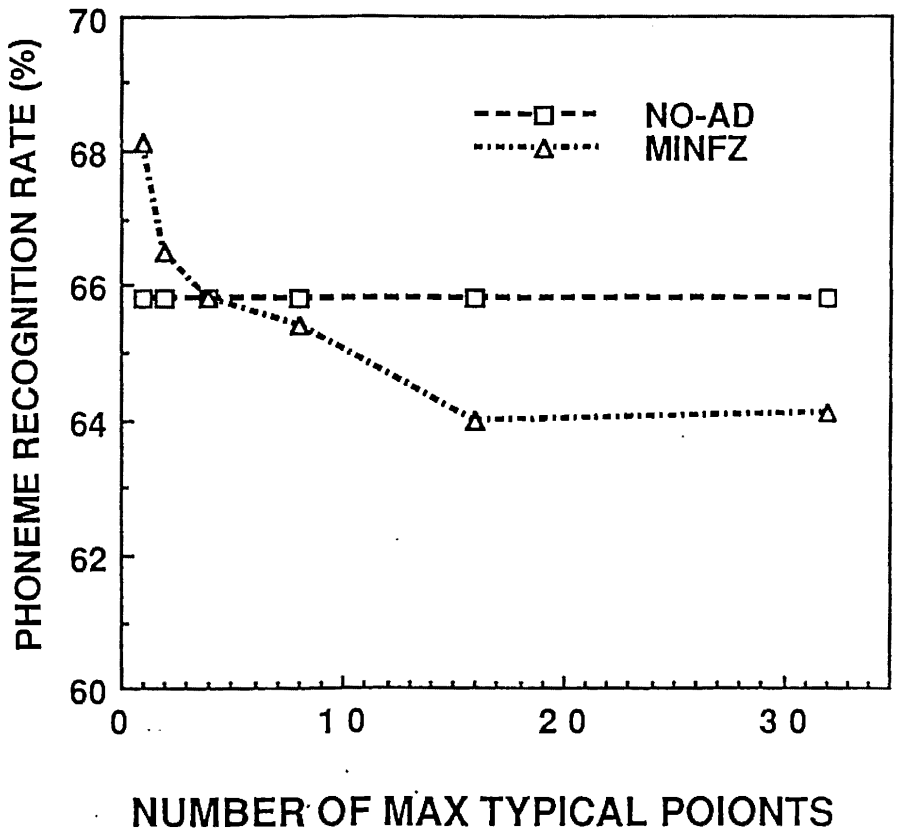

図 4: 最大代表点と音素認識率の関係

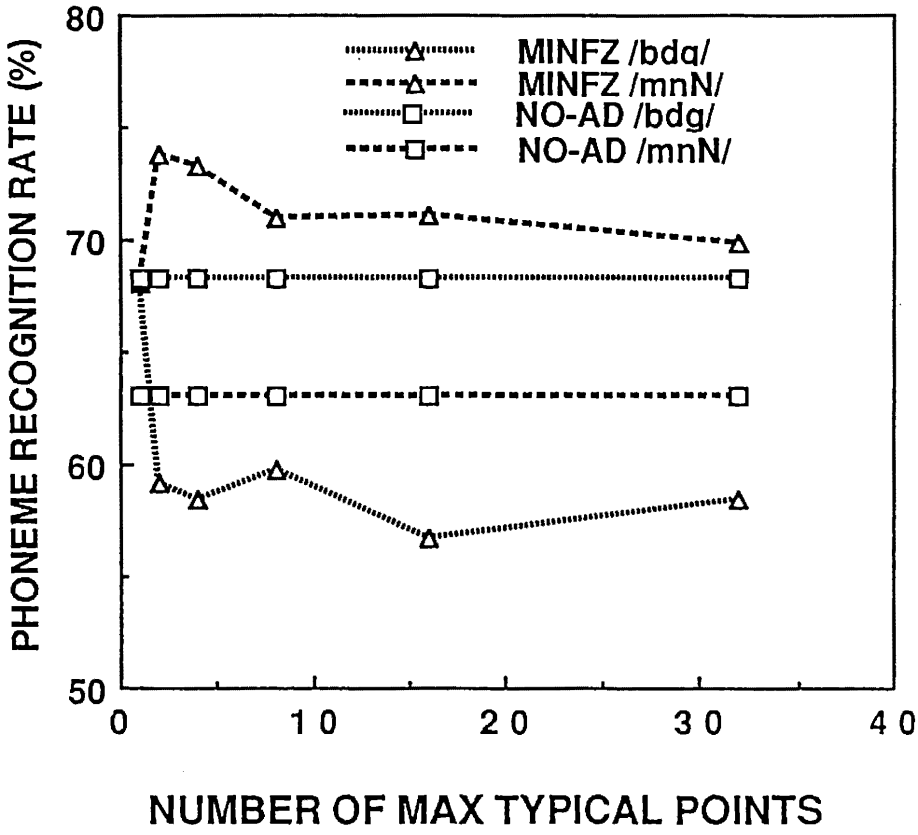

図 5: カテゴリー毎の認識率と代表点の関係

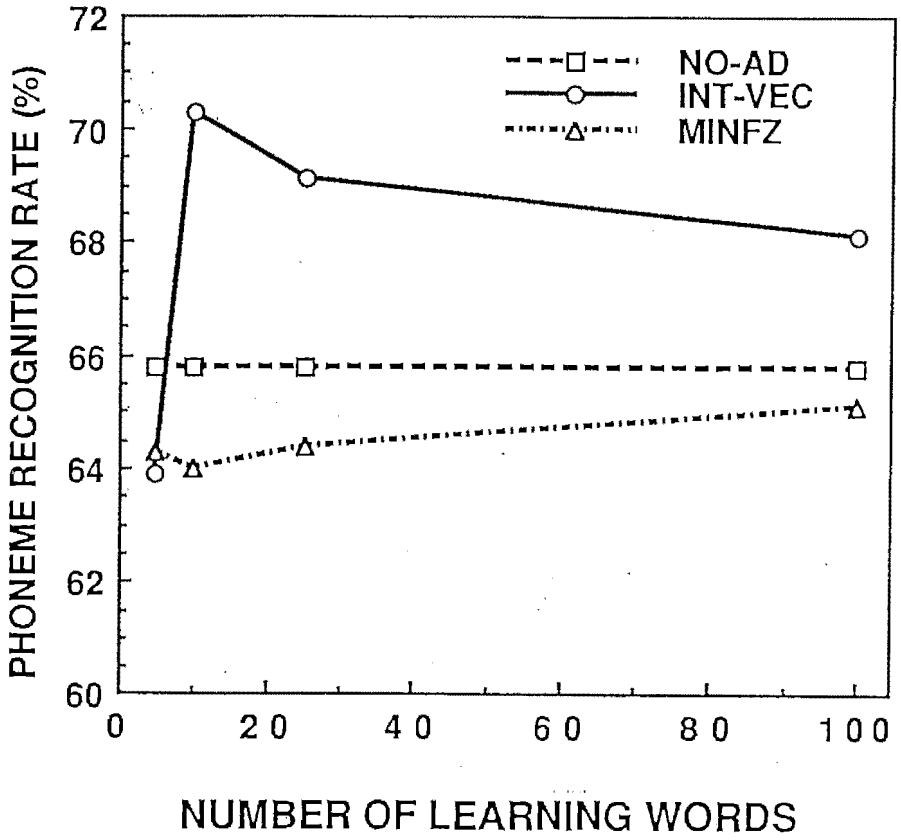

図 6: 学習単語数と音素認識率の関係

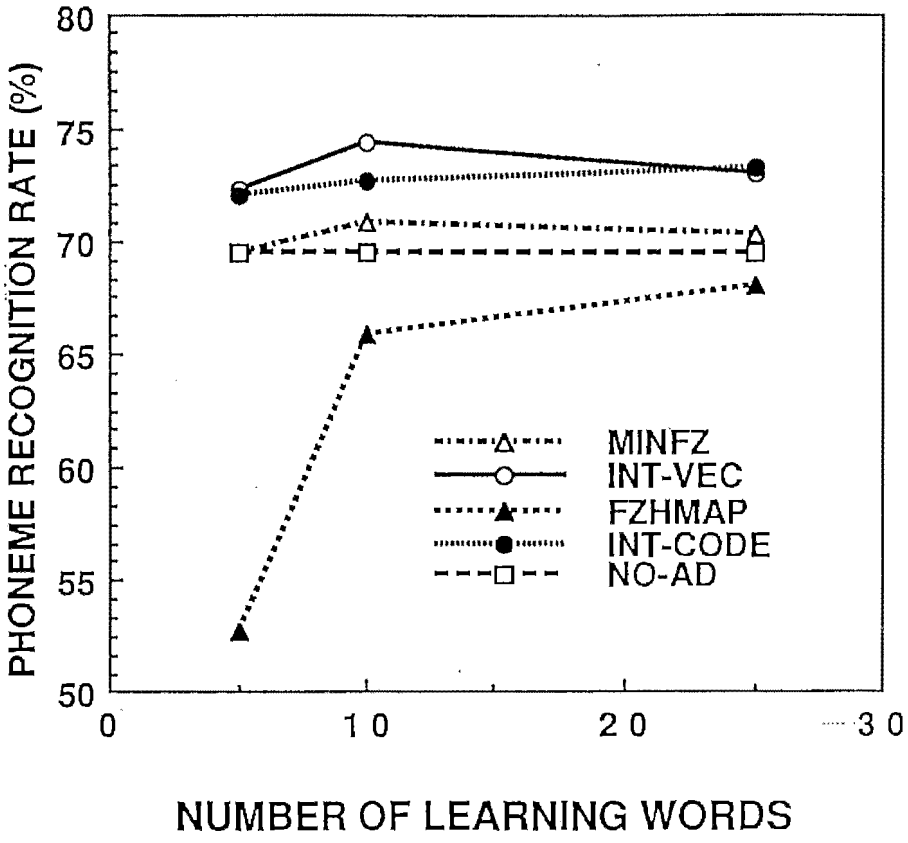
80
75
70
65
60
55
50
PHONEME RECOGNITION RATE (%)
0
10
20
30
NUMBER OF LEARNING WORDS
MINFZ
INT-VEC
FZHMAP
INT-CODE
NO-AD

図 7: 各手法の比較