論文 / 著書情報
Article / Book Information

論題	単語音声認識における能率的学習法
著者	古井貞熙
掲載誌/書名	研究実用化報告, Vol. 29, No. 7, pp. 1277-1290
発行日	1980,

801.314 : 801.4 : 534.322 : 374.1

単語音声認識における能率的学習法*

古井貞熙**

あらまし　不特定多数の話者が発声した多数の語彙を対象とした単語音声認識システムにおいて高い認識率を得るために，発声者が変わるたびごとに少数の音声サンプルを用いて，能率的に学習を行う方法について検討した．本方式では，音韻単位の標準パタンを用いて，少数の学習サンプルから，それらの音韻の全部あるいは一部のスペクトルを抽出し，これにもとづいて，全部の語彙に適応した全音韻の標準パタンを推定する．男性・女性それぞれ30名が発声した10数字音声，および男性30名が発声した67単語音声の認識実験を行った結果，10数字に対しては5単語による学習で，全単語による学習と変わらぬ認識率が得られた．また，10数字に対しては3単語，67単語に対しては12単語程度の学習で，学習なしの場合よりも大幅に認識率を向上させることができることがわかった．

1 まえがき

音声の機械認識の研究は1950年頃から行われており，最近では文章音声の認識の研究もさかんに行われるようになってきたが，単語音声認識の研究は，現在でも，近い将来に実現の可能性がある研究課題として極めて重要である．その中で，特定話者が発声した限られた語彙を対象とした認識装置はすでに作られ，一部で実用に供されているが，研究の目標は，不特定多数の話者が発声した多数の語彙を高い精度で認識することにある．語彙が極めて小さい場合には，発声者が変わるたびごとに全部の語彙を発声してもらって，機械をその話者の声に適応させること(学習)が可能であるが，語彙数が数十単語以上になって，しかも発声者がしばしば変わる場合には，全部の語彙による学習は，極めて非現実的である．

このような問題に対処する方法としては，声の個人差に対する学習は全く行わずに，直接的に，すべての人の声に対応できるような特徴抽出法と認識論理を構成する方法[11][12]，声の種類に対応できるように，各単語について複数の標準パタンを用意しておく方法[3][4]，発声者が変わるたびごとに，少数の単語あるいは短い文章を発声してもらって能率的に学習を行う方法等が考えられる．第1の方法が最も理想的と言えるが，現在のレベルで高い認識精度を得ることは難しく，第2，第3の方法に比べて認識論理が複雑になりがちである．第2の方法は現在研究が進められており，一つの現実的な解決策といえるが，何種類くらいの標準パタンを用意すれば十分であるか，等の問題点がある．

本論文は，上述の第3の方法，すなわち少数の学習サンプルを用いて能率的に学習を行う新しい方法の提案と，その方法による実験結果について述べたものである．この種の方法としては，これまでに，音韻単位の標準パタンを用いて，学習サンプルからそれらの音韻の全部あるいは一部を学習し，不足している音韻があればそれを推定する方法[5]，学習サンプルとして5母音と撥音のみを発声してもらい，それらにもとづいて，多数の話者の平均的な標準パタンをずらす方法[6]，母音学習サンプルから，個人の特徴を声道長と声帯波スペクトルの概形として推定し，平均的な標準パタンをその個人に適応化する方法[7]などが試みられているが，まだ満足すべき結果は得られていない．本論文で述べる方法は，上述の文献(5)の方法の改良と言うことができる．

2 音声認識システム

本研究で用いた単語音声認識システムのブロック図を

* A Training Procedure for Spoken Word Recognition Systems. By Sadaoki FURUI.
この研究は斎藤特別研究室で行われたものである．

** 武蔵野電気通信研究所研究専門調査員

図1に示す[5][9]．システムには，音韻を単位とする標準パタンと，各単語を音韻の系列で表現した単語辞書が蓄えられている．いくつかの音韻は，調音結合によるスペクトルの変動に対処するため，複数個の標準パタンを有している．

入力音声はしゃ断周波数 3.4 kHz の低域通過フィルタを通ったのち 8 kHz, 11 bit で標本化および量子化され，15 ms ごとに区分化される．以下この各区分をフレームとよぶ．各フレームのディジタル波形から，N 次までの相関係数を抽出する．ここでは $N=10$ とした．入力音声の j 番目のフレームと，i 番目の音韻 x_i の標準パタンとのスペクトルの類似度を，対数尤度尺度 $l_j(x_i)$ で表現する．対数尤度尺度 $l_j(x_i)$ は，式(1)に示すように，入力音声の相関係数 $\zeta_\tau(j)$ と音韻標準パタンの線形予測係数（LPC）$a_\tau(x_i)$ から計算される．ただし，τ は時間遅れを示す．音韻標準パタンは，線形予測係数の相関関数 $A_\tau(x_i)$ の形でシステムに蓄えておく．

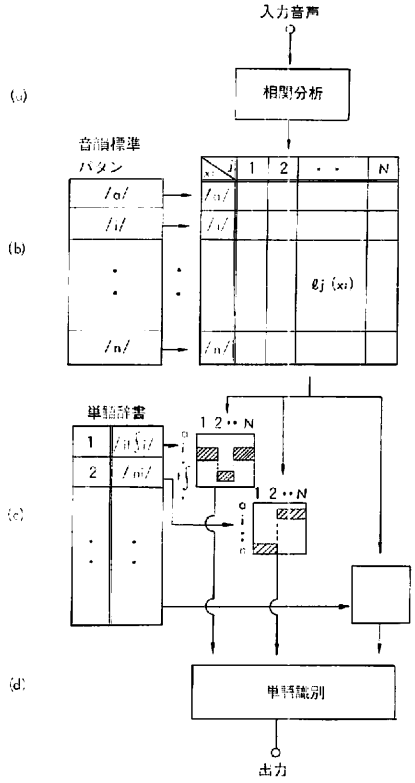
$$\left.\begin{aligned} l_j(x_i) &= -\log\left\{\sum_{\tau=-N}^{N} A_\tau(x_i)\zeta_\tau(j)\right\} \\ A_\tau(x_i) &= \sum_{k=0}^{N-\tau} a_k(x_i)a_{k+\tau}(x_i) \end{aligned}\right\} \quad (1)$$

計算された対数尤度尺度 $l_j(x_i)$ を，入力音声フレーム番号を列，音韻標準パタンを行とする尤度行列に配列する．単語の識別はこの尤度行列と単語辞書を用いて，次の手順にしたがって行う．

各候補単語ごとに，その音韻系列にしたがって尤度行列の要素を入力音声の全フレームに渡って合計し，この値（尤度和）をその単語との類似度とする．このとき，音韻系列中の各音韻には，1つあるいは複数の入力音声フレームが対応づけられるので，入力音声のどこからどこまでをどの音韻に対応づけるかに関しては，非常に多くの組合せが考えられるが，この選択は，各音韻についてあらかじめ定めている最大と最小の長さの範囲内で，入力音声全体として尤度和が最大になるように行う．この演算は，DP（ダイナミックプログラミング）の手法を用いて，容易に行うことができる．このようにして尤度和を各候補単語について計算し，最大の尤度和を示す単語を認識結果として出力する．

音韻標準パタンは，認識に先だつ学習の段階において，学習用音声サンプルを用いて，認識の手順と同様にして自動的に各音韻区間を検出し，その区間内での平均スペクトルを計算することによって作成する．詳しくは文献(8)を参照されたい．

本研究では，10数字と67空港名の2組の語彙を認識対象に用いた．



(a) 入力音声波の相関係数抽出
(b) 入力音声の各区分と音韻標準パタンとの対数尤度行列の計算
(c) 入力音声と候補単語との尤度和の計算
(d) 最大尤度和を有する単語の選択
図 1 単語音声認識システムのブロック図

3 少数単語による能率的学習法

3.1 原　　理

認識対象語彙のすべてを用いずに，少数の音声サンプル（学習サンプル）から各音韻の標準パタンを作成しよ

うとすると，通常次のような二つの問題点に対処する必要が生ずる．

(1) 学習サンプルに含まれない音韻の標準パタンの作成法．

(2) 学習サンプルから抽出された音韻のスペクトルが有する調音結合によるスペクトルの片寄りの除去．

ここでは，この両者を同時に解決する方法を検討した．基本的な考え方としては，少数の学習サンプルから，それらに含まれる音韻のスペクトルを抽出し，この値をもとに，全部の認識対象語彙を学習に用いたときに抽出されるであろう全音韻のスペクトルを推定する．このとき，学習サンプルから抽出されなかった音韻のスペクトルを推定するだけでなく，抽出された音韻のスペクトルも変換する．

学習サンプルとしては，原理的には認識対象と異なる単語や短文章などを用いてもよいが，今回の実験では，認識対象の一部の単語を用いることにした．この方法は，上述 (1) の処理を精密化し，さらに (2) の処理をとり入れたという点で，文献 (5) の方法の改良といえる．以下，認識対象語彙の一部の単語を用いる学習法を"部分単語学習"，全部の単語を用いる学習法を"全単語学習"とよぶ．

3.2 予備学習—重回帰分析—

実際の学習サンプルを用いた部分単語学習に先だって，各音韻のスペクトルに関する統計的な分析を行う．この段階を"予備学習"とよぶ．

後に学習および認識を行うべき音声の話者とは全く独立に，一定の数（20～30名程度）の話者（"予備学習話者"とよぶ）に，全部の語彙を一回ずつ発声してもらう．これらの各話者ごとに，これらの音声を用いて，二組の音韻標準パタンを作成する．そのうちの一組は，後に部分単語学習に用いる語彙のみから作成し，一組は，すべての語彙から作成する．各音韻の標準パタンを作成する方法は，前章に述べたとおりで，各音韻に属すると判定されたフレームにおけるスペクトルの平均化は，相関関数の平均化によって行う．平均化された相関係数からPARCOR係数を抽出し，これをさらにarctanh変換して対数断面積比（LAR）とする．

次に，部分単語学習用の語彙から作成された音韻標準パタンと，全部の語彙から作成された音韻標準パタンとの関係をもとめるため，重回帰分析[20]を行う．前者の標準パタンに関して，音韻 j の l 次の LAR を v_{jl} で表わし，後者の標準パタンに関して，音韻 i の m 次の LAR を u_{im} で表わすとき，$U_i=(u_{i1}, u_{i2}, \cdots, u_{iN})'$ と $V_j=(v_{j1}, v_{j2}, \cdots v_{jN})'$ の間に，すべての話者に共通に，

$$U_i=\Phi_{ij}V_j+E_{ij} \qquad (2)$$

の関係が成り立つものと仮定する[10]．ここで Φ_{ij} は $N\times N$ の行列であり，E_{ij} は残差である．筆者らは以前に音声波に含まれる個人性情報と音韻性情報に関するいくつかの統計的分析を行ったが[14][15]，その結果は式 (2) に示したような仮定の可能性を示している．

変換行列 Φ_{ij} の各要素は，全ての予備学習話者の音声サンプルを用いた重回帰分析により推定することができる．このとき同時に，部分単語学習用の語彙から抽出された音韻 j の LAR ベクトル V_j を用いて，式 (2) にしたがって推定した音韻 i の LAR ベクトル $\hat{U}_i$ の l 次の要素と，実際に全部の語彙から抽出された音韻 i の LAR ベクトル U_i の l 次の要素との重相関係数 R_{ijl} を計算しておく．行列 Φ_{ij} は i, j のすべての組合せについて計算し，重相関係数 R_{ijl} は i, j, l のすべての組合せについて計算する．さらに，各音韻 i について，各予備学習話者の音声サンプルから抽出した LAR ベクトル U_i を平均して，ベクトル $\bar{x}_i$ とする．

後述するように，このようにして予備学習の段階で計算された行列 Φ_{ij}，重相関係数 R_{ijl} および平均値ベクトル $\bar{x}_i$ を用いて，部分単語学習を行う．

3.3 部分単語学習

各話者は認識に先だって，認識対象語彙のうちの，あらかじめ定められているいくつかの単語を一回ずつ発声する．これらの音声サンプルを用いた部分単語学習と認識の手順を図2のブロック図に示した．

各学習サンプルの音声波は，まず相関係数の時系列に変換され，この時系列は前述のようにして自動的に各音韻区間に区分化（セグメンテーション）される．各音韻ごとに，全学習サンプルの，その音韻に該当する全フレ

ームの平均相関係数を計算し，これをさらに LAR パラメータに変換する．このようにして直接的に学習サンプルから抽出された各音韻の LAR を用いて，次の式(3)より，すべての音韻（学習サンプルから抽出された音韻および抽出されなかった音韻のすべて）の標準パタンを推定する．

$$\hat{x}_{ik}=(r/\sum_{j\in X_L} w_{ij})\sum_{j\in X_L} w_{ij}\Phi_{ij}x_{jk} +(1-r)y_{ik},\ i\in X \qquad (3)$$

ただし，

X：標準パタンの全音韻の集合

X_L：学習サンプルから抽出される音韻の集合 ($X_L\subset X$)

$\hat{x}_{ik}$：話者 k の推定される音韻 i の N 次元 LAR ベクトル

x_{jk}：話者 k の学習サンプルから抽出される音韻 j の N 次元 LAR ベクトル

である．

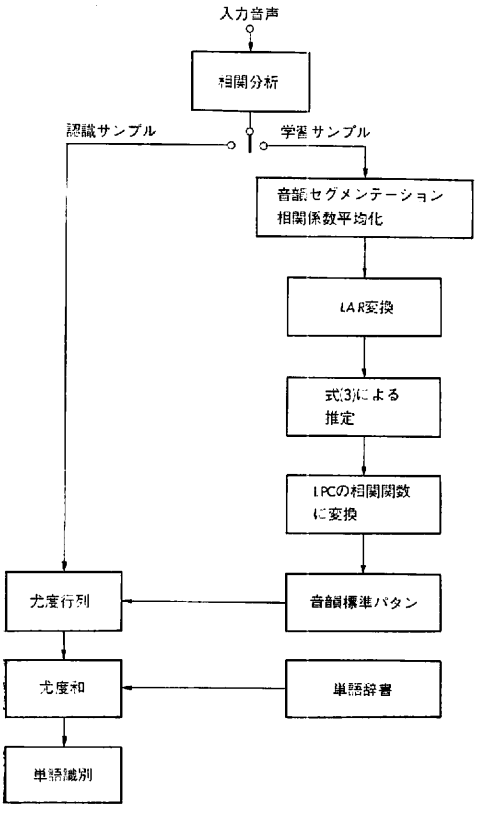


図 2　部分単語学習と認識のブロック図

式(3)は二つの項から成っている．第1項は，予備学習の段階で計算された $N\times N$ の変換行列 Φ_{ij} を用いて，学習サンプルから抽出された種々の音韻スペクトルと，推定すべき音韻標準パタンを関連づける項である．この中には，次式で定義される重み係数 w_{ij} が含まれている．

$$w_{ij}=(1/N)\sum_{l=1}^{N} R_{ijl}^{2} \qquad (4)$$

ここで，R_{ijl} は Φ_{ij} と同時に計算された重相関係数である．

式(3)の第2項は，今推定しようとしている音韻 i が学習サンプルに含まれている場合はそのスペクトル，含まれていない場合は予備学習段階で抽出された多数話者の音韻 i の平均スペクトルと，音韻 i の推定標準パタンとを直接的に関連づける項である．これらの第1項と第2項の混合比を決定する重み係数 r は，0と1の間で実験的に設定される．y_{ik} は次式のように定義される．

$$y_{ik}=\begin{cases} x_{ik} & : i\in X_L \\ x_{i'k} & : i\notin X_L,\ i'\in X_L \\ \bar{x}_i & : i\notin X_L,\ i'\notin X_L \end{cases} \qquad (5)$$

ここで i' は，調音結合によるスペクトルの変動に対処するために用意してある，音韻 i と同じ音韻の異なる標準パタンを示す．$\bar{x}_i$ は，予備学習で抽出された多数話者の音韻 i の平均 LAR ベクトルである．

LAR の領域で式(3)に示すような変換を行うのは，LAR は，このような変換を行っても線形予測係数等のようにパラメータが不安定になる心配がなく，パラメータの微小変化に対するスペクトル感度がほぼ一様で[16]，個人性情報を表現するパラメータとしても LAR が有効である[17]等の理由による．

推定された各音韻の標準パタンは，LAR から線形予測係数の相関関数に変換され，蓄えられる．

4　10数字音声による認識実験

4.1　実験方法

男性・女性各30名（無作為に抽出した研究所員）

が発声した10数字音声の認識実験により，ここで検討している学習方法の有効性の評価を行った．各話者の学習サンプルとして，10数字のうちの3単語または5単語を用い，各話者の認識サンプルとして，各数字を10個ずつランダムな順序で含む100サンプルを用いた．学習に用いる3単語としては，2，3および9を，5単語としては，0，2，3，5および9を用いた．これらの学習用単語は，学習を行わずにすべての話者に共通の（平均）の標準パタンを用いたときに比較的認識誤りの多い単語を選んだ．

標準パタンのすべての音韻および学習サンプルに含まれる音韻は図3に示すとおりである．/a/，/i/，/u/は，調音結合の影響を考慮して，それぞれに二種類の標準パタンを用いる．/itʃi/の/tʃ/，/saN/の/s/のような無声子音は，継続時間が短かく，10数字音声の認識ではあまり大きな役割りを果たさないので，標準パタンを用意していない．12種類の音韻のうち，5単語学習では10種類が，3単語学習では6種類が抽出され，これらのスペクトルをもとにして，式(3)により，すべての音韻の標準パタンが推定される．

Φ_{ij}，w_{ij}および$\bar{x}_i$は，学習および認識用音声を発声する話者とは独立の男性・女性それぞれ29名の話者の音声を用いた予備学習により，あらかじめ設定した．これらの値は，男性用と女性用に別々に用意した．

比較のために，全く学習を行わずに，本人を含まぬ多数話者の平均の標準パタンを用いる場合，すなわち$\hat{x}_{ik}=\bar{x}_i$とした場合（学習なし）の認識実験，$r=0$，すなわち式(3)の第1項による推定を行わずに，学習サンプルから抽出された値あるいは多数話者の平均値をそのまま用い，$\hat{x}_{ik}=y_{ik}$とした場合の認識実験，および全単語学習の場合の認識実験をあわせて行った．すなわち，学習方法に関して，次のような4種類の条件による認識実験を行った．

（i）　部分単語学習（推定あり）
$\hat{x}_{ik}=$式(3)

（ii）　部分単語学習（推定なし）
$\hat{x}_{ik}=y_{ik}$

（iii）　学習なし
$\hat{x}_{ik}=\bar{x}_i$

（iv）　全単語学習

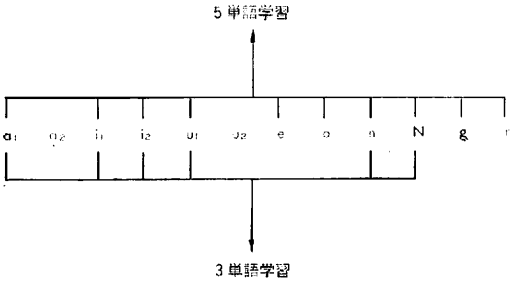


図3　10数字音声の認識に用いられる全音韻と学習サンプルに含まれる音韻

4.2 総合結果

重み係数を$r=0.5$としたときの，30名の話者および10数字に関する平均認識誤り率を図4に示す．図5には，重み係数rの値と，部分単語学習（推定あり）の場合の平均認識誤り率の関係を示す[12]．図5には男性話者の結果のみを示した．図4の結果から，10数字の半分にあたる5単語を用いて，本研究で検討した部分単語学習（推定あり）を行えば，男性・女性の音声に関してそれぞれ0.9%と2.1%の誤り率が得られ，これらの値はほぼ全単語学習の場合の結果に等しいことがわかる．3単語のみによる学習の場合でも，誤り率は1.4%（男性）と3.3%（女性）であり，学習なしの場合よりも約1.5%低下している．推定なしの部分単語学習の場合には，3単語学習の場合のように一部の音韻しか抽出されないときに，大きな誤りを生ずる．

図4には，以前に声帯波と声道長の近似的正規化を試みたときの結果[7]をあわせて示してあるが，その結果と比較すると，今回の方法で3単語学習を行った場合の方が，ややよい結果を示している．また図5の結果から，重み係数rの値は0.5程度が最適で，$r=0$の場合だけでなく，$r=1$としてy_{ik}の項を用いない場合も，誤り率が増加することがわかる．

全般的に男性よりも女性の場合の方が誤り率が大きくなっている理由は明らかでないが，女性の基本周波数が男性のそれよりも2倍程度大きいために，同じ周波数帯域に含まれる高調波の数が男性の1/2程度しかなく，このために女性のスペクトル包絡の方がやや不安定な傾向が

あると考えることもできる.

4.3 単語別・話者別結果

図6は，5単語学習の場合の男性と女性の単語別の結果を，学習なしおよび全単語学習の場合の結果と対比して示したものである．この結果を見ると，推定を行うことにより，推定なしに比較して，学習単語に含まれなかった 4，6，7 等の誤り率が大きく低下し，全単語学習の場合と類似した単語間の誤りの分布を示すことがわかる．この傾向は男性と女性に共通している．図3からもわかるように，5単語学習の場合は，推定を行わなくても，学習サンプルから直接的にほとんどの音韻が抽出され，不足するものは，a_2 と u_2 のみであるから，この場合の推定効果はほぼ節3.1で述べた(2)の調音結合の影響の除去の効果であると言えよう．

図7は，3単語学習の場合の男性と女性の単語別結果を示したものである．推定を行うことにより，学習単語に含まれなかった単語のうちの 7，8，0 および学習単語に含まれていた9の誤り率が大きく低下している．推定

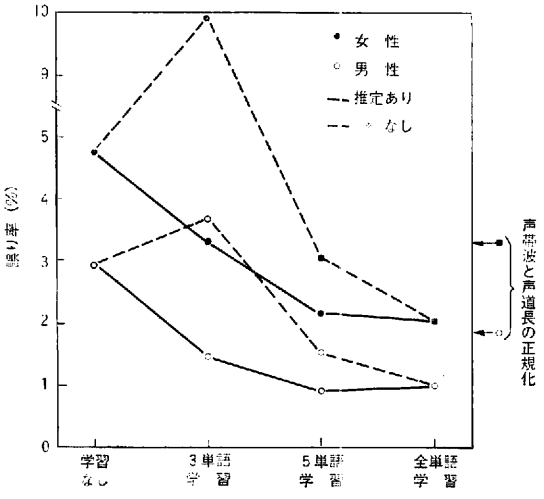


図 4　10数字音声の平均認識誤り率 ($r=0.5$)

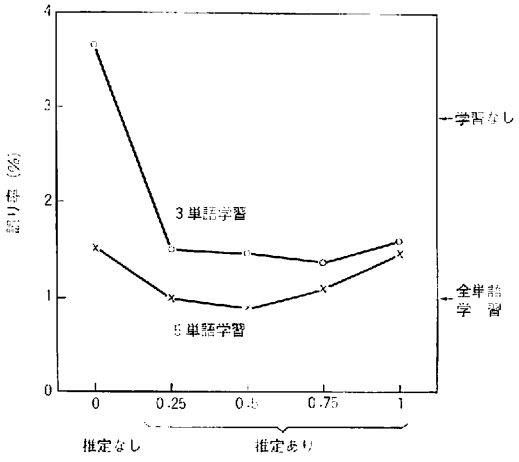


図 5　重み r と認識誤り率の関係（男性）

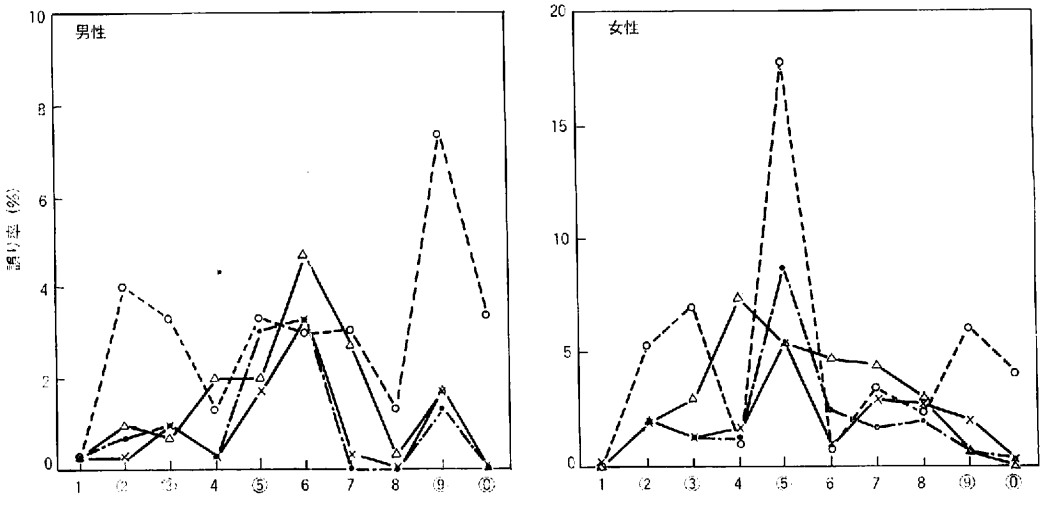
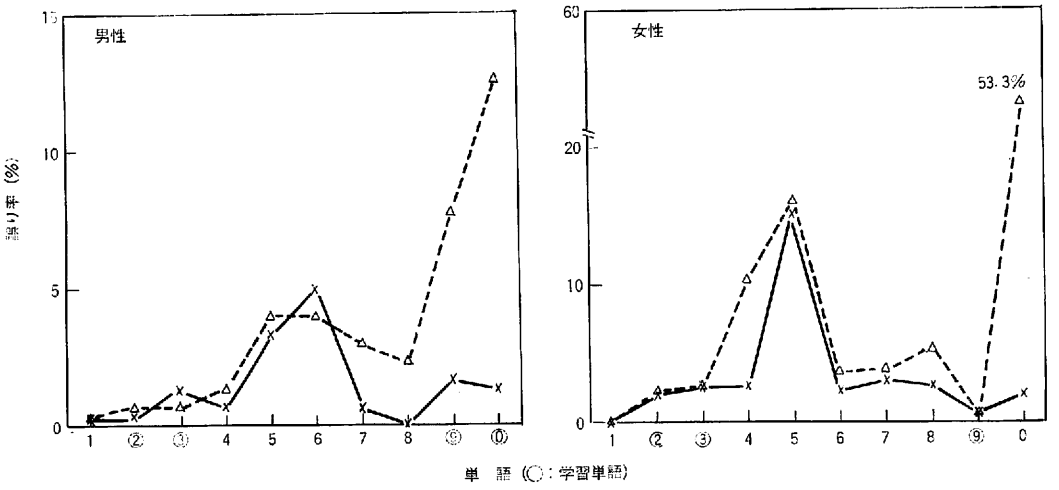


図 6　単語別認識誤り率（5単語学習）



－×－ 3単語・推定あり　　--△-- 同・推定なし

図 7 単語別認識誤り率（3単語学習）

なしのときの認識誤りは，ほとんど9と0および0と2の間でおこっている．推定による改善は，主として音韻eの推定効果によるものと考えられる．

図8は，30名の男性話者のうち，種々の実験条件で平均的に誤りの多い12名の話者について，話者別の誤り率を示したものである．5単語学習の場合の結果を見ると，話者ごとに見ても，総合的な結果の場合と同様に，推定ありの部分単語学習により，全単語学習の場合と極めて類似した結果を得ることができることがわかる．また，学習なしの場合に比べて他の3条件の場合は，話者による誤り率のばらつきが小さくなっていることがわかる．

3単語学習の場合は，推定なしの部分単語学習を行うと，学習なしの場合よりもかえって誤り率が大きくなってしまう話者を多く生ずることがわかる．しかし，推定を行えば，3単語学習でも誤り率は大きく低下して，話者によるばらつきが小さくなり，全単語学習の場合と類似した傾向を示すようになる．

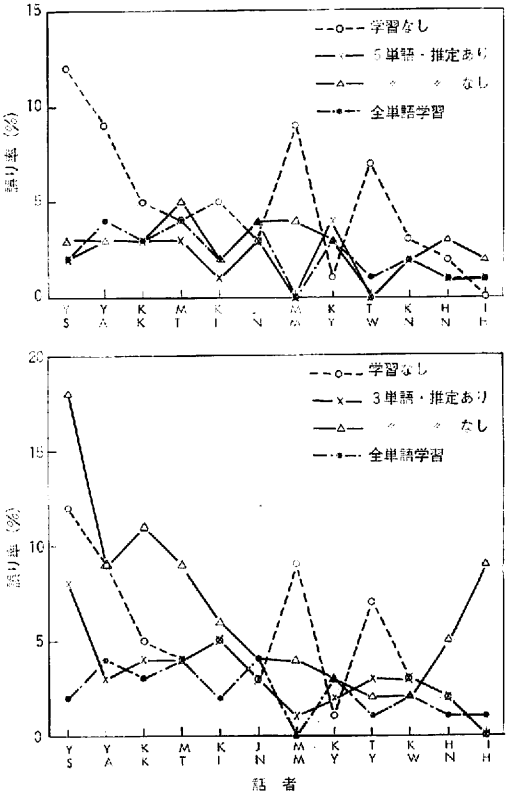


図 8 話者別誤り率（平均的に誤りの多い話者）

4.4 考　　察

図9に，参考のために，推定を行うときの重み係数 w_{ij} の実際の値を示した．5単語学習の場合の結果を(a)に，3単語学習の場合の結果を(b)に示した．これらの値は，予備学習の段階で，男性29名の音声サンプルを用いた統計的分析によって得られたものである．

全体として，学習サンプルから抽出される音韻のうち，推定すべき音韻と同じ音韻に対する重み（$i=j$）が最も大きく，その音韻が学習サンプルから抽出されない場合は，その音韻と同じ種類の音韻で調音結合に対する対

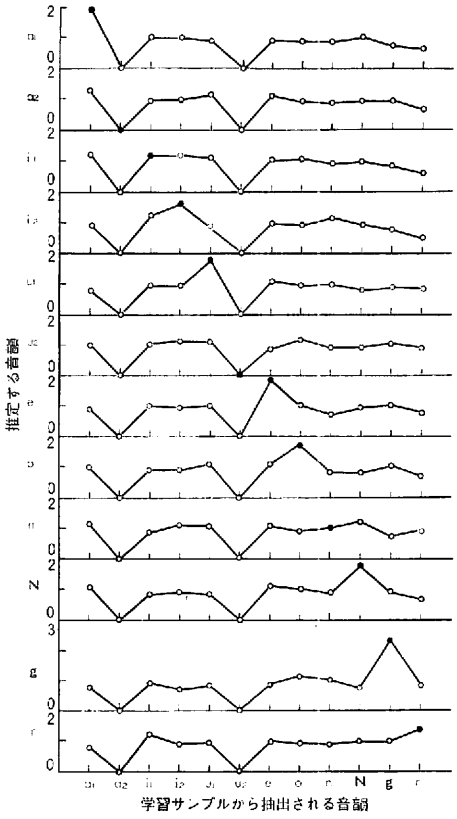


図 9 (a)　重み係数 w_{ij} (5単語学習・男性)

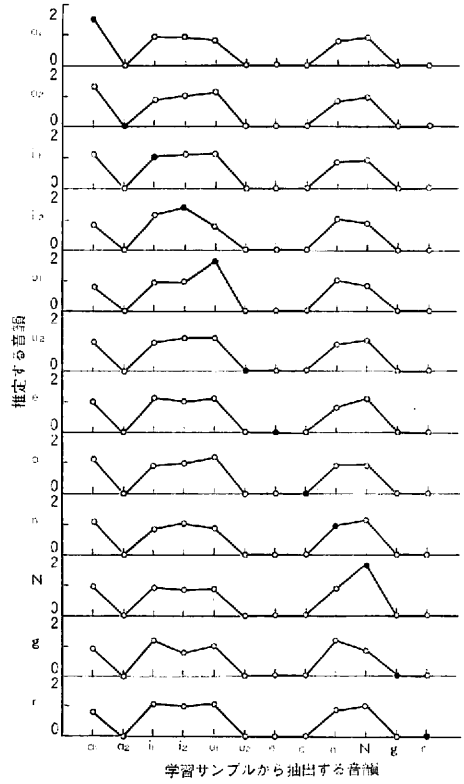


図 9 (b)　重み係数 w_{ij} (3単語学習・男性)

策として用意してある音韻があれば，その重みが最も大きくなっている．また，音韻 g に対する他の音韻からの推定の重み (5単語学習の場合) は，それ自身からの場合に比して極めて小さくなっており，音韻 g の個人差の構造が，他の音韻とかなり異なっていることを示している．女性の場合も，上述の男性の場合と同様の傾向が得られる．

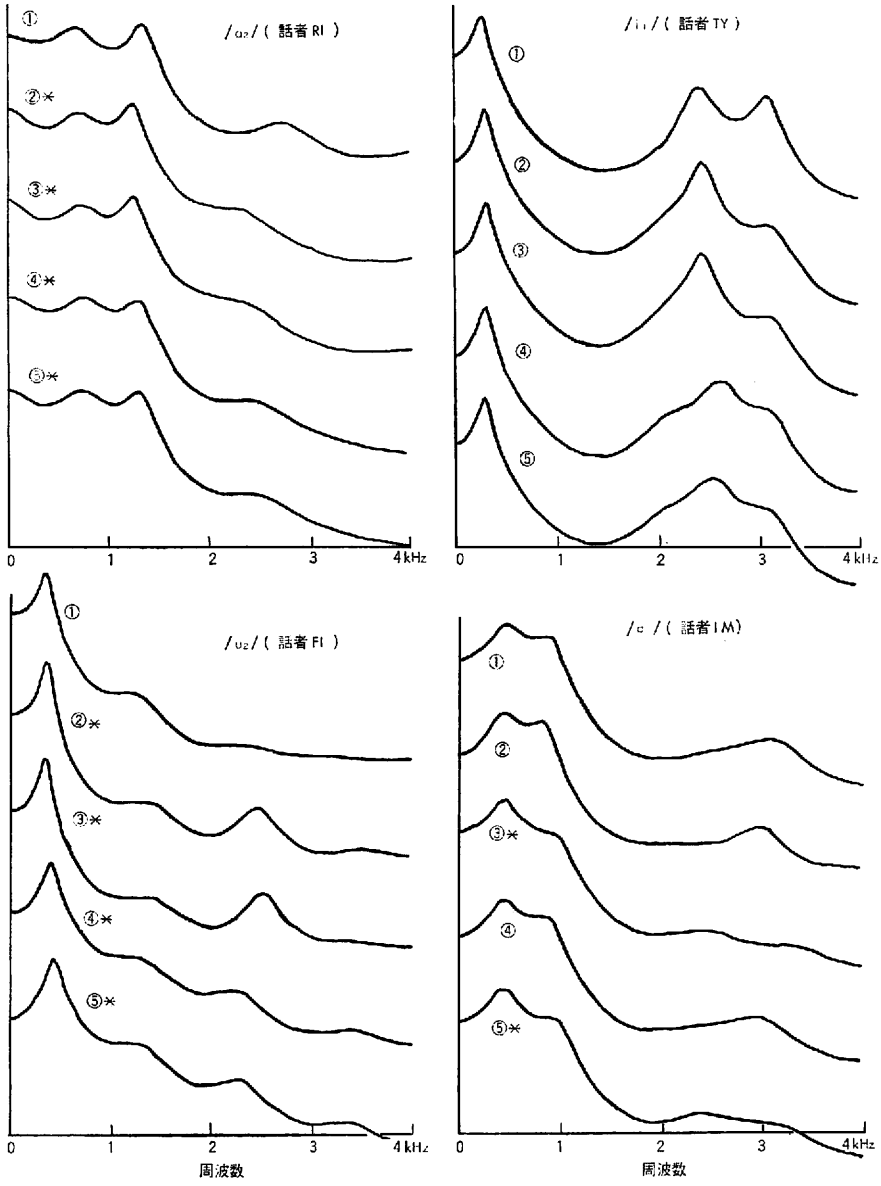
図 10 に，種々の学習条件によって作られた音韻標準パタンのスペクトル包絡の比較を，4種類の例によって示した．これらのスペクトル包絡は，10次の全極形モデルによって表現したものである．図中の * 印は，その学習条件の場合に，その音韻が学習サンプルから直接的に抽出できないことを示している．これらの結果から，推定ありの部分単語学習を行うことにより，その音韻が学習サンプルに含まれていなくても，全単語学習の場合に類似したスペクトル包絡を有する標準パタンを得ることができることがわかる．

以上の10数字音声を用いた認識実験により，ここで検討している部分単語学習法が，少数の音声サンプルから能率的に個人差学習を行う方法として有効であることが確認された．次章では，この方法を実際に多数語彙の認識系に適用して，その有効性を調べる．

5　67単語音声による認識実験

5.1　実験方法

男性30名が発声した空港名67単語音声の認識実験を行った．認識対象単語は表1に示すとおりで，部分単語学習には，表に示す12単語または25単語を用いた．この認識系で用いる音韻標準パタンは，休止区間を含めて42種類で，表2に示すとおりである．10数字音声の場合と同様に，調音結合を考慮して，同じ音韻でも標準パタンを複数個用意してあるものがあるが，この実験では，簡単のためにすべて独立と考え，推定式 (3)



①全単語学習　②5単語学習・推定なし　③3単語学習・推定なし　④5単語学習・推定あり
⑤3単語学習・推定あり　　＊学習サンプルに含まれない音韻

図 10　各学習方法による音韻標準パタンのスペクトル包絡の例

中の y_{ik} を指定する式 (5) の中の i' に相当する音韻の存在は考慮しなかった．したがって式 (5) は次のように変形される．

$$y_{ik}=\begin{cases} x_{ik} : i\in X_L \\ \bar{x}_i : i\notin X_L \end{cases} \qquad (5')$$

変換行列 Φ_{ij}，重み係数 w_{ij} および平均値ベクトル $\bar{x}_i$ は，学習および認識用音声を発声する話者とは独立の25名の男性話者の音声を用いた予備学習によって設定した．重み係数 r の値は0.5とした．

部分単語学習に用いる単語は，10数字音声の認識の場合と同様に，主として学習なしの場合に誤認識の多い単語から選び，誤認識が特に多くなくても発声時間の長い（多くの音韻を含む）単語を追加した．学習単語に含まれる音韻と含まれない音韻は表3に示すとおりで，

表 1　67 単語音声の認識における認識対象単語

○：25 単語学習
◎：12 単語学習

1	札　幌	○	18	米　子		35	沖永良部	○	52	紋　別	
2	旭　川	○ ◎	19	出　雲	○	36	山　形		53	中標津	
3	女満別		20	徳　島	○ ◎	37	仙　台		54	沖　縄	○
4	稚　内		21	高　松	○ ◎	38	八丈島		55	久米島	○
5	釧　路	○	22	高　知	○ ◎	29	大　島		56	南大東	
6	帯　広	○	23	広　島	○	40	三宅島	○	57	宮　古	
7	函　館		24	宇　部		41	富　山	○ ◎	58	多良間	
8	秋　田		25	松　山	○	42	金　沢		59	石　垣	
9	青　森	○	26	大　分		43	福　井	○	60	与那国	○ ◎
10	八　戸		27	福　岡		44	鳥　取		61	千　歳	
11	花　巻		28	宮　崎		45	名古屋		62	小　松	
12	東　京		29	鹿児島		46	北九州	○ ◎	63	大　村	◎
13	大　阪		30	種子島		47	長　崎		64	那　覇	
14	南紀白浜	○	31	屋久島		48	熊　本		65	三　沢	
15	新　潟	○	32	喜界島		49	福　江	○ ◎	66	利　尻	○ ◎
16	岡　山		33	奄美大島		50	佐　渡	○ ◎	67	奥　尻	○ ◎
17	隠　岐		34	徳之島		51	壱　岐				

表 2　67 単語音声の認識に用いられる音韻標準パタン

分　類	数	種　類
母音	9	a i_1 i_2 i_3 u_1 u_2 u_3 e o
母音のわたり	6	ia io iu ai oi ui
有声子音	13	b d g z_1 z_2 r_1 r_2 w N m n_1 n_2 ŋ
無声子音	13	p t ts tʃ k_a k_i k_u k_e k_o h_1 h_2 f s
休止区間	1	*
計	42	

25 単語学習の場合は全単語学習の場合と同じ数の音韻が抽出されるが，12 単語学習の場合は，学習サンプルから直接的に抽出されないものが 8 個生ずる[11]．

各話者が 67 単語を，同じ順序で通して 5 回発声したうち，1 回目の全単語あるいはその一部を学習サンプルとして用い，2～5 回目の音声を認識サンプルに用いた．

5.2　総合結果

各学習条件の場合の，30 名の話者による 67 単語の全認識誤り率を図 11 に示す．25 単語学習の場合は，推定なしでも全単語学習の場合（1.4%）に近い誤り

表 3　学習単語に含まれる音韻と含まれない音韻

a	i_1	i_2	i_3	u_1	u_2	u_3	e	o	ia	io	iu	ai	oi
●	●	×	●	●	●	●	●	●	●	●	●	●	●
ui	b	d	g	z_1	z_2	r_1	r_2	w	N	m	n_1	n_2	ŋ
●	○	●	○	○	○	●	○	●	×	●	●	●	○
p	t	ts	tʃ	k_a	k_i	k_u	k_e	k_o	h_1	h_2	f	s	*
×	●	●	●	●	●	●	○	●	○	●	●	●	●

× 全単語学習でも含まれない可能性のあるもの
○ 25単語には含まれるが，12単語には含まれないもの
● 25単語にも12単語にも含まれるもの

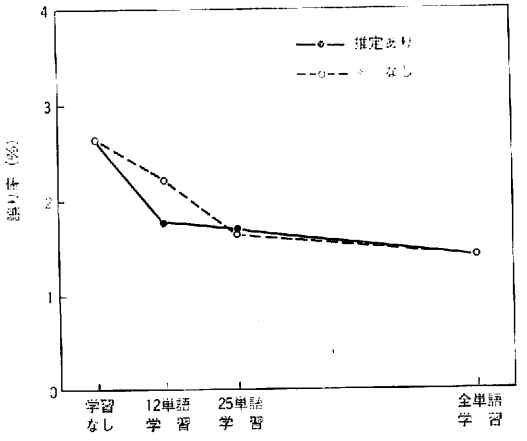
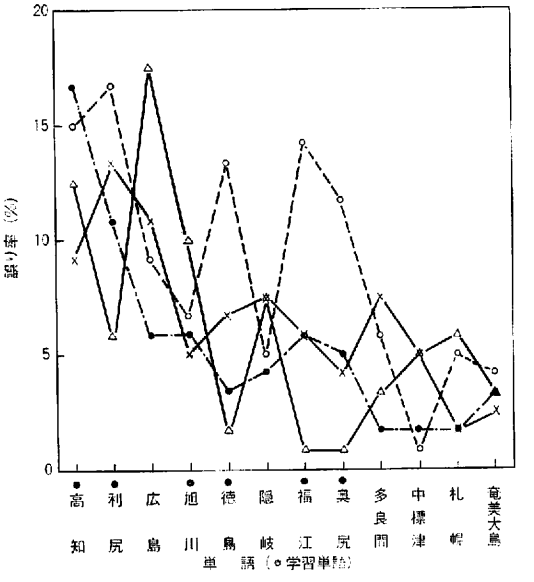


図 11　67単語の平均認識誤り率



--○-- 学習なし
－×－ 12単語・推定あり
－△－ 〃　〃　なし
－●－ 全単語学習

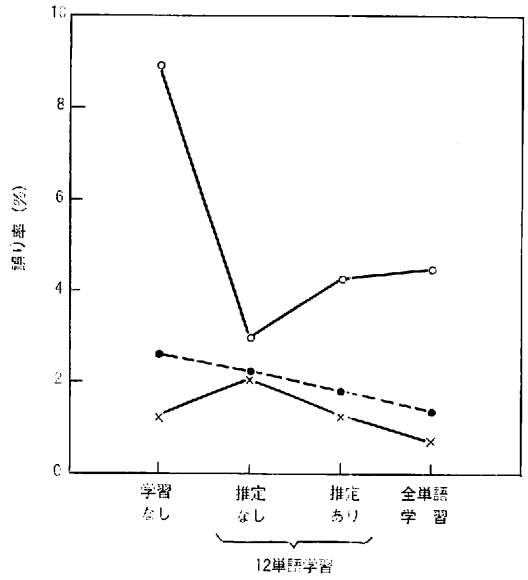
図 12　単語別認識誤り率（平均的に誤りの多い単語）

率（1.6%）が得られてしまうので，推定の効果が見られないが，67単語のほぼ1/5弱に相当する12単語学習の場合は，推定によって，誤り率が2.2%から1.8%に改善されており，学習なしの場合の誤り率（2.6%）に比べると，少ない発声の手間で，認識率をかなり向上させることができることがわかる．

5.3　単語別・話者別結果

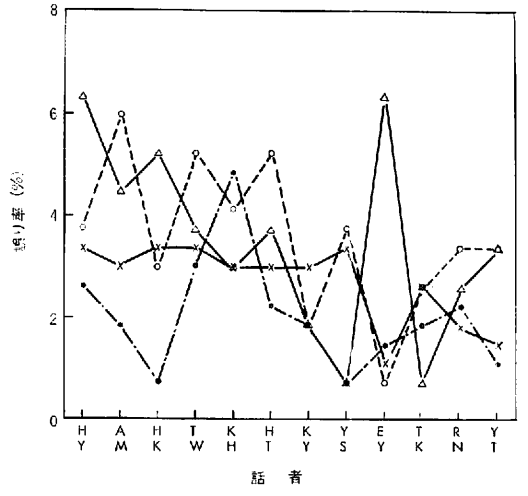
図12は，種々の学習条件で平均的に誤りの多い12種類の単語（図に示すように，このうち6種類の単語は学習単語に含まれている）について，12単語学習（推定あり・なし），学習なし，および全単語学習における誤り率を示したものである．図13には，この4種類の学習条件の場合について，学習単語に含まれる単語（もちろん，学習サンプルと認識サンプルは異なる）の誤り率と，学習単語に含まれない単語の誤り率を別々に平均化して示した．全単語について平均化した誤り率もあわせて示してある．

これらの結果から，推定なしの12単語学習を行うと，学習なしの場合に比較して，学習単語に含まれる12単語の誤り率は大きく低下するが，学習単語に含まれない55単語の誤り率は逆にわずかながら増加してしまうことがわかる．これに対して，推定を行うと，学習単語に含まれる単語の誤り率はやや増加するが，学習単語に含まれない大多数の単語の誤り率が低下し，全体としての平均誤り率が低下する．推定を行ったときの学習単語に含ま



○ 学習単語に含まれる単語
× 〃 含まれない単語
● 全単語の平均誤り率

図 13　学習単語に含まれる単語の誤り率と含まれない単語の誤り率（12単語学習）



--○-- 学習なし
－×－ 12単語・推定あり
－△－ 〃 〃 なし
－●－ 全単語学習

図 14　話者別誤り率（平均的に誤りの多い話者）

れる単語と含まれない単語の誤り率はいずれも，全単語学習の場合のこれらの単語の誤り率にほぼ等しい．

以上の結果から，ここで検討している部分単語学習

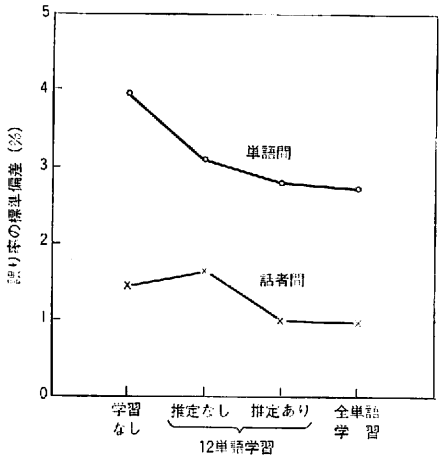


図 15　認識誤り率の単語間・話者間標準偏差

法は，学習単語の認識率を多少犠牲にして，全体としての認識率を上げる方向に働いていると考えられる．学習単語に含まれる特有の調音結合の影響を除去して，全体としての最適値，すなわち全単語学習による音韻標準パタンの推定を行っている以上，当然とも言える．

図 14 は，種々の実験条件で平均的に誤りの多い 12 名の話者について，上述と同じ 4 種類の学習条件における誤り率を示したものである．図 15 には，各学習条件における誤り率の単語間および話者間における分布の標準偏差を示した．学習なしの場合は，単語間，話者間のいずれにおいても誤り率の分布の広がりが大きいが，推定ありの部分単語学習を行うと，いずれの分布の広がりも小さくなって，全単語学習の場合とほぼ同程度になる．これらの様子は図 12 と図 14 にも見られる．

5.4 考　　察

以上の 67 単語音声を用いた認識実験により，ここで検討した学習法が，実際の多数語彙の認識系においても有効であることがわかった．ここで認識対象とした 67 単語の語彙は，全単語を学習した場合と全く学習を行わない場合の認識率の差が比較的小さいので，ここで検討した学習法の効果が，認識率（誤り率）の絶対値の上ではあまり大きく現れていないが，もう少し学習効果の大きい認識対象について実験を行えば，さらに明確な効果が現れるものと期待される．

6 むすび

不特定話者を対象とした多数語彙の音声認識システムにおいて高い認識率を得ることを目標として，各話者ごとに少数の音声サンプルによって能率的に学習を行う方法を提案し，10数字音声と67単語音声を用いた認識実験により，その有効性を評価した．ここで実験を行ったシステムでは，各話者は認識対象語彙のうちの一部の単語を一回ずつ，学習用音声として発声する．この音声サンプルは相関係数の時系列に変換され，自動的に音韻を単位とするセグメンテーションが行われる．次に，各音韻について，全学習サンプル中の，その音韻にラベルづけされた全フレームの平均相関係数が計算され，対数断面積化パラメータに変換される．このようにして抽出された各音韻の対数断面積化パラメータに対して，あらかじめ予備学習によって得られている変換規則を適用し，全認識対象語彙に適応した全音韻の標準パタンを推定する．

この変換規則は，学習サンプルから抽出される音韻の特徴パラメータと，推定すべき各音韻標準パタンの特徴パラメータとに関する重回帰分析によって決められる．この分析に用いられる音声を発声する複数の話者（予備学習話者）は，学習および認識用の音声サンプルを発声する話者とは全く独立に選ばれるので，ここで検討したシステムは，擬話者不依存システム（quasi-speaker independent system），あるいは話者依存・不依存ハイブリッドシステム（hybrid speaker dependent/speaker independent system）とよぶことができる[13]．

男性および女性それぞれ30名が発声した10数字音声を用いた認識実験の結果，5単語を学習に用いれば，全単語を用いた場合と等しい認識率を得るのに十分であることが明らかとなった．また，3単語による学習でも，学習を行わない場合に比べて，誤り率を大きく低下させることができることがわかった．5単語学習の場合の平均認識率は，99.1%（男性）と97.9%（女性）であり，3単語学習の場合の平均認識率は98.6%（男性）と96.7%（女性）であった．

次に，男性30名が発声した67単語音声を用いた認識実験を行った結果，12単語による学習が，少ない発声の手間で高い認識率を得る手段として有効であることがわかった．このときの平均認識率は98.2%であった．この結果は，ここで検討した学習方法が，実際の多数語彙の認識システムに適用可能であることを示している．

この学習方法の特徴は，学習サンプルから抽出されない音韻の標準パタンを推定するだけでなく，抽出された音韻の標準パタンに関しても，それのもつ特有の調音結合の影響を取除いて，自動的に，学習サンプルとして用いられなかった単語に含まれるスペクトルにも適応するように変換する機能を有している点にある．この学習方法で用いる学習用音声サンプルは，原理的には，認識対象語彙と全く異なる単語や短文章でもよい．

今後検討すべき事項としては次のようなものがある．

（i） 学習単語あるいは学習用短文章の最適選択，

（ii） 推定式に含まれる重み係数の最適化，

（iii） 他の多数語彙の認識系や連続音声の認識系への適用，

（iv） 認識と同時に学習を行う，教師あり[18]あるいは教師なし[14][15][19]の学習方法との組合せによる改良．

謝辞

御指導・ごべんたつを頂いた旧斎藤特別研究室長に感謝する．好田調査役および長島研究主任には，データ収集・認識プログラムなどの面で，多大なお世話になった．また，研究室の他の諸氏には熱心な討論をして頂いた．心からの謝意を表わしたい．

文献

（1） 千葉・亘理・渡辺：不特定話者を対象とした単語声音認識システム，信学会情報全大，No. 219，1977.

（2） 三輪・牧野・松岡・城戸：オンライン単語音声自動認識装置，信学会研資，PRL 76-8，1976.

（3） L. R. Rabiner et al.：Speaker-Independent Recognition of Isolated Words Using Clustering Techniques, IEEE Trans. ASSP-27, No. 4, p. 336, 1979.

（4） R. L. Davis：Application of Clustering to the Recognition of Spoken Connected Digits, Proc. 4 th IJCPR, p. 1025, 1978.

(5) 好田・斎藤：部分的な学習サンプルによる単語音声の認識，信学会研資，PRL 72-87，1972.
(6) 坂井・中川・林：音韻スペクトルの個人差の予備学習による限定語彙単語音声の認識，信学会研資，EA 75-61，1976.
(7) 古井：音声認識における個人差の学習法について，音学会研資，S 75-25，1975.
(8) 好田・橋本・斎藤：数字音声の機械認識系，信学論，**55**-D，No. 3，p. 186，1972.
(9) 長島・好田：音韻の標準パタンを用いた単語音声の認識，音学春季講論，4-2-7，1976.
(10) 古井：単語音声識認における個人差の能率的学習法，音学秋季講論，3-1-19，1977.
(11) 古井：多数語彙単語音声認識における学習の能率化，音学春季講論，4-1-12，1978.
(12) 古井：単語音声認識における学習の能率化，音学会研資，S 77-43，1977.
(13) S. Furui：A Training Procedure for Isolated Word Recognition Systems，IEEE Trans. ASSP-28，2，p. 129，1980.
(14) 古井：音声波に含まれる個人性と音韻性情報の分離，音学秋季講論，2-2 10，1974.
(15) 古井：音声波に含まれる声質情報と韻質情報の分離，音学会研資，S 74-17，1974.
(16) R. Viswanathan et al.：Quantization Properties of Transmission Parameters in Linear Predictive Systems，IEEE Trans. ASSP-23，3，p. 309，1975.
(17) 古井・板倉：単語の統計的パラメータによる話者認識，信学論，**56**-A，No. 11，p. 717，1973.
(18) B. T. Lowerre：Dynamic Speaker Adaptation in the Harpy Speech Recognition System，1977 ICASSP，23. 2，1977.
(19) 中川・白方・三樹・坂井：個人差の教師なし学習機能を持つ60都市名の実時間音声認識，音学会研資，S 77-11，1977.
(20) 奥野・久米・芳賀・吉沢：多変量解析法，第2章，p. 25，1971，日科技連.

(1980，3，25 受付)