論　文

# 音韻ラベルを用いないHMM評価法とそれを用いた連続音声認識用HMMの評価

正員　南　泰浩[†]　正員　松岡　達雄[†]　正員　鹿野　清宏[††]

# Phoneme HMM Evaluation Algorithm without Phoneme Labeling Applied to Continuous Speech HMM Evaluation

Yasuhiro MINAMI[†], Tatsuo MATSUOKA[†] *and* Kiyohiro SHIKANO[†], *Members*

**あらまし**　本論文では，音韻ラベルに基づいて切り出された音声データに対する音韻認識率による音韻HMMの評価の代わりに，音韻ラベルを用いないHMM評価法を提案する．この評価法は，音韻ラベルによる切り出しを行う必要がないため，音韻ラベルの付けられていない音声データベースを用いても，音韻認識率による音韻HMMの評価を行うことができる．更に，本論文では大規模の不特定話者連続音声データベースを用いて音韻HMMの連結学習を行い，学習話者の増加に対するHMMの音韻認識率を，提案するラベルを用いないHMM評価法を用いて評価し，本手法の有効性を示した．

**キーワード**　HMM，音韻認識，音韻ラベル

## 1. まえがき

不特定話者大語い音声認識には，話者によるゆらぎの吸収に優れているHidden Markov Model（HMM）を音韻単位で用いる手法が有効である．不特定話者大語い連続音声認識を高精度で実現するためには，HMMの認識性能を向上させることが重要である．このためには大量のデータを用いて，話者のゆらぎやコンテキストのゆらぎをHMMの中に反映することが必要である．これまでにデータ量と音声認識の性能に関して，アメリカではCMU[(1)]とBBN[(2),(3)]において，日本では，NTTデータとSRI[(4)]の共同研究においての報告がある．

まず，CMUでの報告では，109人の話者がそれぞれ40文を発声した音声データを用いて学習を行った結果が報告されている[(1)]．次にBBNでは，12人の話者がそれぞれ600文を発声した音声データを用いて学習を行った結果を報告している[(2),(3)]．更にNTTデータとSRIでは，110人の話者が84文を発声したデータを用いて，単語認識率での評価を行っている[(4)]．これらの結果では，学習データ量が認識結果に大きく影響することが報告されている．

本論文では，日本音響学会連続音声認識データベースを用いて学習した音韻HMMを，音韻認識率の観点から評価した[(5)]．このデータベースは64人の話者がそれぞれ約150文を発声した音声を収録している．

従来，学習された音韻HMMを評価する手法としては音韻ラベルに基づいて切り出した音韻に対する認識率を計算する手法や，あるタスクを設定して単語認識や文認識率で評価する手法[(6)]が主であった．しかし，前者では，音韻ラベルを音声データベースに付けるのに莫大な労力が必要である，後者では，タスクに依存しやすい，評価するのに手間がかかる，などの問題点がある．ここではこのような問題点を解決するために，音韻ラベルを必要としない音韻HMMの評価方法を提案し，この手法に基づいて音韻認識率の評価を行う．これからの説明では，音韻ラベルを必要としないHMM評価法を，ラベルなし評価法と呼ぶことにする．本論文では，まず**2.**で不特定話者HMMの作成法について述べる．**3.**では本論文で提案する音韻ラベルを用いないHMM評価法について述べる．**4.**では，**4.1**で従来の音韻ラベルに基づいて切り出したデータを用いる評価法を用いた学習話者数に対する音韻認識率について述べる．**4.2**では，まず，同一の音声データに対して，音韻ラベルを用いないHMM評価法と従来の音韻ラベルに基づいて切り出したデータを用いる評価

† NTTヒューマンインタフェース研究所，武蔵野市
NTT Human Interface Laboratories, Musashino-shi, 180 Japan

法の両方を用いて，不特定話者 HMM の評価を行う．この実験により音韻ラベルを用いない HMM 評価法の有効性を調べる．更に，音韻ラベルの付与されていない日本音響学会連続音声データベースで，音韻ラベルを用いない HMM 評価法により，不特定話者 HMM の学習話者数に対する音韻認識率について述べる．

## 2. HMM の学習

### 2.1 HMM の種類と構造

用いた音韻は以下の 43 種類である．摩擦音 (s, sh, h, z)，破擦音 (ch, ts)，破裂音 (p, t, k, b, d, g)，鼻音 (m, n, N)，半母音 (w, y, r)，母音 (a, i, u, e, o)，長母音 (aa, ii, uu, ee, oo)，二重母音 (ei, ou)，拗音 (sy, hy, zy, cy, py, ky, by, gy, my, ny, ry)，無音，促音に伴う無音．

すべての HMM は，4 状態 3 ループのモデルとした．各状態の出力確率分布は，四つのガウス分布からなる連続混合ガウス分布である．各ガウス分布の共分散行列は対角成分のみ存在する．本報告で用いた音声データの分析条件と特徴パラメータを表 1 に示す．

### 2.2 連結学習 HMM の構成

本論文では音韻のラベル付けをされていないデータベースを用いて学習を行った．データベース中の各文には発声内容を示すテキストが添付されている．連結学習[7],[8]では，このテキストに基づいて音韻モデルを連結し文モデルを作成する．音韻モデルの連結は，図 1 のように，音韻モデルの最終状態を次の音韻の初期状態として行った．連結したモデルの両端と句読点の部分には，無音モデルを付け加えた．文モデルを学習文ごとに作り，Baum-Welch アルゴリズム[9]によって HMM パラメータの学習を行う．最後にモデルを分解して各音韻ごとの HMM を作成する．連結学習での HMM の学習繰り返し回数は 5 回とした．

表 1 分析条件と特徴パラメータ

| | |
|---|---|
| サンプリング | 12 KHz<br>16 bits |
| 分析条件 | ハミング窓,フレーム長32ms<br>フレームシフト 8 ms<br>高域強調、16次LPC分析 |
| 特徴パラメータ | LPCケプストラム(16次)<br>Δケプストラム(16次)<br>Δパワー |

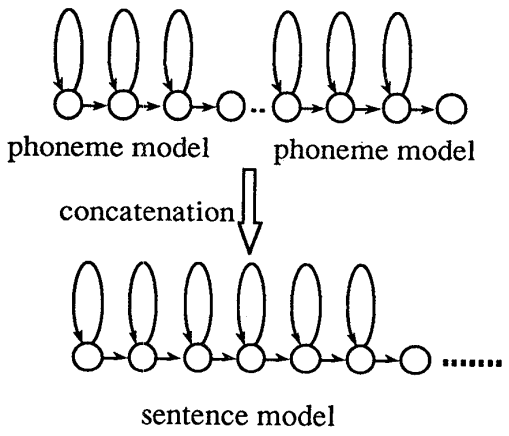



図 1 HMM の連結学習
Fig. 1 Concatenated HMM training.

### 2.3 初期モデル

学習用の連続音声データベースは長い文を含んでいる．長い文で連結学習を行う場合，初期値をランダムな値から始めると，Forward Backward アルゴリズムの過程でオーバフローが起こる可能性がある．そこで初期モデルとして，音韻ラベルの付与されているデータベースを用いて，不特定話者の HMM を作成した．初期モデルを作成するのに用いたデータを表 2 に示す．このデータから音韻ラベルによって切り出した音韻を用いて，初期モデルの学習を行った．但し，by, my, py は学習サンプルの数が少ないので，ATR の音韻バランス 216 単語のデータベースから切り出した音韻も学習サンプルに加えて，学習を行った．データベース中の二つ以上の音韻が融合して分割できない音韻 (融合している音韻) は学習に用いていない．学習繰返し回数は，連結学習と同様に 5 回とした．

### 2.4 連結学習に用いたデータベース

連結学習に用いたデータは，音響学会の連続音声データベースである．データベースには，一人当り約 150 文の連結音声データが 64 人分収録されている．発声内容は ATR 音韻バランス 503 文である．学習には，このデータベースの中から発声誤りを除いた文章を用い，話者数を 2 人から 64 人と変化させてモデルを作成した．学習に用いた人数と文の数を表 3 に示す．

表 2 初期モデルに用いたデータ

| | |
|---|---|
| 話者 10人 | 男性 5 名：MAU, MMS, MNM, MSH, MTM<br>女性 5 名：FAF, FFS, FKN, FMS, FSU |
| データ | ATR音声データベースの区切り指定を行なわない発話 (S.C) |
| 文の数 | 115 文 |

表 3　学習話者数と文の数

| 話者数（男性／女性） | 文数 |
|---|---|
| 2（1／1） | 300 |
| 4（2／2） | 600 |
| 8（4／4） | 1200 |
| 16（8／8） | 2400 |
| 32（14／18） | 4801 |
| 64（30／34） | 9603 |

## 3. ラベルなし評価法

### 3.1 音韻ラベルを用いない HMM 評価法

ラベルなし評価法では，評価用データと発声された内容を示すテキストを用いて，以下のように音韻認識率を求める．

まず，テキストから図 2 の 1 番上のモデルのような正しい文モデルを作成する．この文モデルを用いて文全体の出力確率を計算する．今，枠で囲われた音韻 HMM に注目し，この音韻 HMM の評価を行うことにする．この注目した音韻を図 2 の下側のように，他の音韻に置き換えた文モデルを音韻モデルの数だけ作成し，出力確率を計算する．その後，正しい文モデルも含めて，今まで求めた出力確率の中で最も高いものを認識結果とする．例えば図 2 で，もし，1 番上の文モデルが最も高い出力確率を示した場合には，正しく /r/ を認識できたことになり，2 番目の文モデルが最も高い出力確率を示した場合には，/r/ を /a/ と誤認識したことになる．このような操作を文の最初の音韻から最後の音韻まで評価すれば，1 文中のすべての音韻の認識結果が得られる．更に，データベース全体について行えば，データベースに対する音韻認識率を求めることができる．

音韻ラベルを利用しないで音韻 HMM を評価する方法として，Viterbi 法によりセグメンテーションを行って音声を切り出し，このデータに対して音韻認識を行う方法がある．これに対し，ラベルなし評価法は，音声の切り出しは行わず，文モデル中の評価しようとする音韻 HMM 以外の音韻 HMM を固定し，評価しようとする音韻 HMM だけを種々の音韻 HMM に取り替えて，文の出力確率を計算する．この出力確率によって音韻 HMM の評価を行う．このため，ラベルなし評価法は連続音声中のある音韻 HMM が他の音韻 HMM へ置換するしやすさを求める手法であると考えられる．

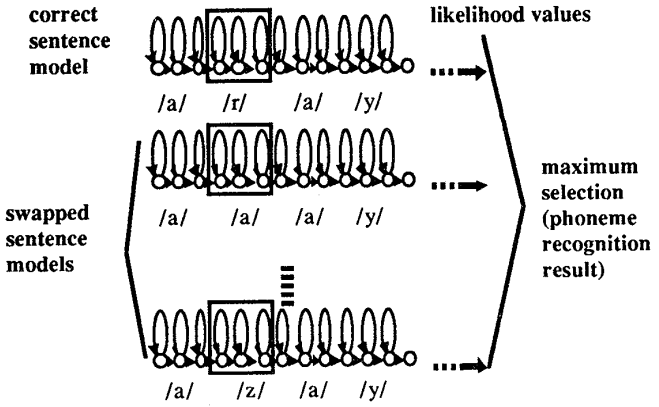



図 2　音韻ラベルを用いない HMM 評価法
Fig. 2　Phoneme HMM evaluation algorithm without phoneme labeling.

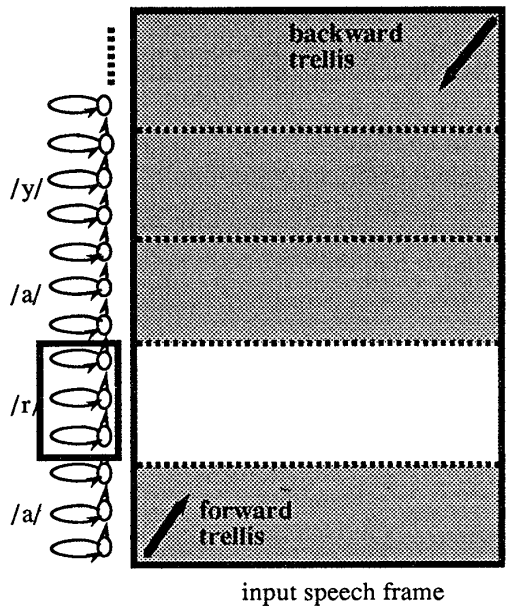



図 3　前向きトレリスと後ろ向きトレリスを用いた効率的な計算法
Fig. 3　Efficient evaluation algorithm using forward and backward trellises.

### 3.2 ラベルなし評価法の効率化

3.1 で述べた手法は，文中のすべての音韻に対してその音韻を他の音韻に入れ替えて，HMM のトレリスを計算しなければならないので，非常に多くの計算量を必要とする．しかし，HMM の学習で用いられる Baum-Welch アルゴリズム中の，Forward のトレリス計算結果および Backward のトレリス計算結果を再利用することで，効率的な計算が可能である．最初に正しい文モデルを用いて，Forward と Backward の出力確率を計算し，途中結果を（文 HMM の状態数）×（フレーム数）のマトリクスに保存しておく．図 3 は縦軸に文の HMM の状態，横軸に入力のフレームを示

した．今，枠で囲った音韻の認識を行う場合を考える．すでにこの音韻の一つ前の状態までは，Forwardによる出力確率がマトリクス上に保存されている．またこの音韻の一つ後の状態までは，Backwardによる出力確率がマトリクス上に保存されている．このため音韻モデルを入れ替えたときに新しく計算し直す所は，網かけのない所だけである．ここで一つ前の音韻の最終状態を初期値にして，評価している音韻の次の音韻の初期状態までトレリスを計算して，Backwardによる出力確率とつなぎあわせれば全体の出力確率を計算できる．

## 4. 認識実験

認識実験には以下の三つのデータベースを用いた．

（a） ATRの国際会議問合せ文[10]

（b） ATRの音韻バランス文[11]

（c） 音響学会連続音声データベースの対話読上げ文[12]

音韻認識による評価に用いた話者数と文の数をそれぞれ表4，表5，表6に示す．（a）の国際会議のデータベースのテキストは，初期モデルに用いたデータのテキストと同じであり，（b）の音韻バランス文のテキストは，連結学習に用いたデータのテキストと同じである．（c）のテキストは音韻バランス文とは全く独立のものである．また，（a）と（b）はともにATRで収録されたものであり，発声者はアナウンサー・ナレーター等でかつ，マイクロホン，録音環境等も厳しく管理されている．（a）の国際会議の問合せ文を発声した話者は，初期モデルの学習に使用していない話者である．（a），（b）のデータには音韻ラベルが添付されているが，（c）には音韻ラベルは添付されていない．

表4 ATRの国際会議問合せ文

| | |
|---|---|
| 話者10人 | 男性5名：MHT, MMY, MTK, MTT, MXM<br>女性5名：FKN, FKS, FTK, FYM, FYN |
| データ | ATR音声データベースの区切り指定を行なわない発話（SC） |
| 文の数 | 115文 |

表5 ATRの音韻バランス文

| | |
|---|---|
| 話者6人 | 男性2名：M001, M002<br>女性4名：F001, F002, F003,F004 |
| データ | ATR音声データベース（Cセット）の音韻バランス503文中のAセット |
| 文の数 | 50文 |

表6 対話読上げ文

| | |
|---|---|
| 話者13人 | 男性7名：can0001, etl0005, fuj0003 kdd0006, mac0001, nec0011, ric0002<br>女性6名：ecl1010, hit1003,ntd1001 sha1002, son1003, wsd1001 |
| データ | 対話読み上げ文<br>各話者1対話を評価に用いる。（計13対話） |

### 4.1 ラベルに基づいて切り出した音声に対する音韻認識

ここでは，従来より行われている，ラベルに基づいて切り出した音声による音韻認識率を求めた．（a）国際会議問合せ文と（b）ATRの音韻バランス文の両方のデータベースを用い，連結学習に用いた話者数と音韻認識率の関係を求めた．認識に用いた音韻は23音韻（s,sh,h,z,ch,ts,p,t,k,b,d,g,m,n,N,r,w,y,a,i,u,e,o）である．また音韻ラベルが融合している音韻は，評価には使用していない[13]．

結果を図4，図5に示す．図4は国際会議問合せ文に対する音韻認識率であり，図5は音韻バランス文に対する音韻認識率である．初期モデルの二つのデータベースに対する認識率を図中に点線で示す．

図4，図5のどちらの音韻認識結果とも，人数の増加によって認識率が向上することがわかる．しかし，人数が32人を超えると認識率が飽和している．このことから今回用いたHMMは，ATRのデータベースに関して，32人程度の学習データがあれば十分であると考えられる．図の認識結果を初期モデルの認識率と比較すると，国際会議問合せ文に対しては，初期モデルの方が高い認識率を示しているが，逆に音韻バランス文に対しては学習話者32人，64人の連結学習HMMの方が高い認識率を示している．これは学習音韻と認識音韻のコンテキストの違いであると考えられる．初期モデルはATRの国際会議問合せ文から作成したものであり，連結学習HMMはATRの音韻バランス文から作成したものである．このため二つのモデルは，学習時と同一なコンテキストのデータに対して高い認識率を示すと考えられる．

初期モデルはどちらのデータベースに対しても連結学習HMMの認識率と比べ比較的高い認識率を示している．これらの原因は学習に用いたデータと評価に用いたデータの録音環境の違いや，話者の発声様式の差などが考えられる．この原因以外にも次のようなこ

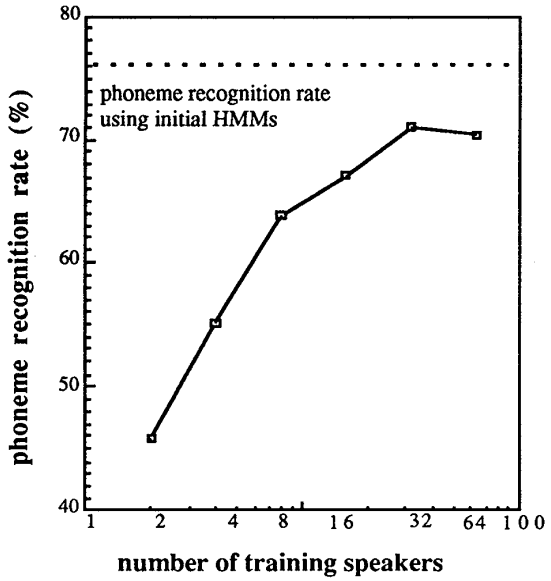



図 4　学習話者数に対する国際会議問合せ文中の音韻認識率(音韻ラベルに基づいて切り出した音韻を用いた評価法)

Fig. 4　Number of training speaker vs. phoneme recognition rate for ATR conference registration database (phoneme HMM evaluation algorithm with phoneme labeling).

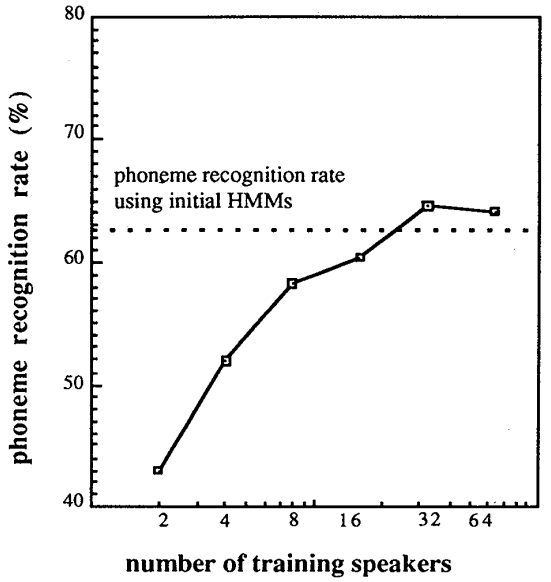



図 5　学習話者数に対する音韻バランス文中の音韻認識率(音韻ラベルに基づいて切り出した音韻を用いた評価法)

Fig. 5　Number of training speaker vs. phoneme recognition rate for ATR phoneme balanced database (phoneme HMM evaluation algorithm with phoneme labeling).

とが考えられる．初期モデルでは音韻ラベルに基づいて切り出された音韻を学習時に使っている．このため初期モデルが音韻ラベルに基づいて切り出された音韻に特殊化され，音韻ラベルに基づいて切り出された音韻に対して強いモデルになっていると考えられる．また，初期モデルでは，学習時も認識時も融合している音韻を使っていない．これに対して連結学習モデルでは音韻特徴がくずれた，あるいは全く欠落している音韻も学習に利用されている．これらの音韻は認識時には出現しない．このため認識率の低下が起きると考えられる．

### 4.2　ラベルなし評価法による音韻認識実験

（1）　ATR の音韻バランス文(b)の評価実験

**3.** で新しく提案した評価法を使った音韻認識実験を行った．ATR の音韻バランス文(b)の評価を行った．評価音韻は **4.1** と同じ 23 音韻である．

ここでは ATR の音韻バランス文に付加された音韻ラベルの情報を利用して，二つの評価法を比較できるように，ラベルなし評価法では融合化している音韻を評価から除いた．**4.1** のラベルを用いた場合の評価法とラベルなし評価法での認識結果は評価している音韻の数が違うため正確な比較ができない．これはラベルを用いた手法では，融合している音韻を除いていたのに対して，音韻ラベルなし評価ではテキストに音韻が融合化しているかどうかの情報がないので，融合している音韻も認識の対象となるからである．図 6 に実験結果を，音韻ラベルを用いた場合の結果と合わせて示す．

ラベルなし評価法による音韻認識結果は，学習人数の増加に伴い認識率が向上し，32 人を過ぎると認識率が飽和する傾向を示している．これは従来の音韻ラベルによる評価法と同じ傾向であることがわかる．この結果から，音韻ラベルを付けられていない音声データに対しても，ラベルなし評価法を用いれば，音韻認識率による評価ができることがわかる．

（2）　音響学会連続音声データベースの対話読上げ文(c)の評価実験

ラベルなし評価法を用いて，音響学会連続音声データベースの対話読上げ文の評価を行った．話者は各収録機関の中から一人ずつ 13 人を選んだ．読上げ文については，各話者ごとに別の対話を選んだ．話者中には学習データである音韻バランス文を発声している話者もいる．

図 7 に話者の平均の認識率と学習話者数との関係を示す．この平均値には学習データを発声している 3 名の認識率も加えた．この結果では人数が 32 人を超えて

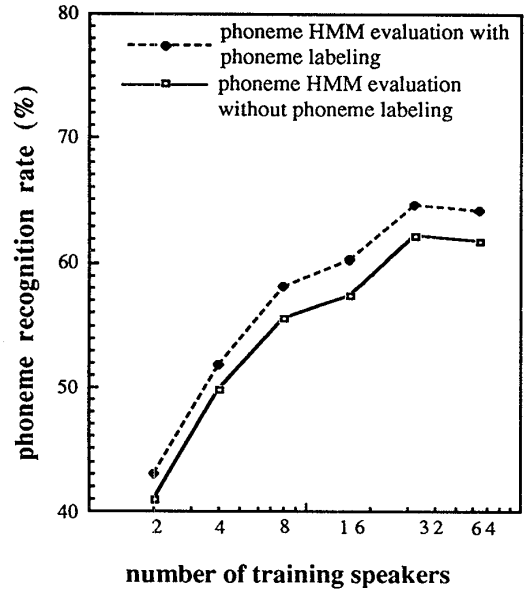



図 6 音韻ラベルに基づいて切り出した音韻による HMM 評価法と音韻ラベルを用いない HMM 評価法による音韻認識率（音韻バランス文）

Fig. 6 Phoneme recognition rate using phoneme HMM evaluation algorithms with phoneme labeling and without phoneme labeling (for phoneme balanced database).

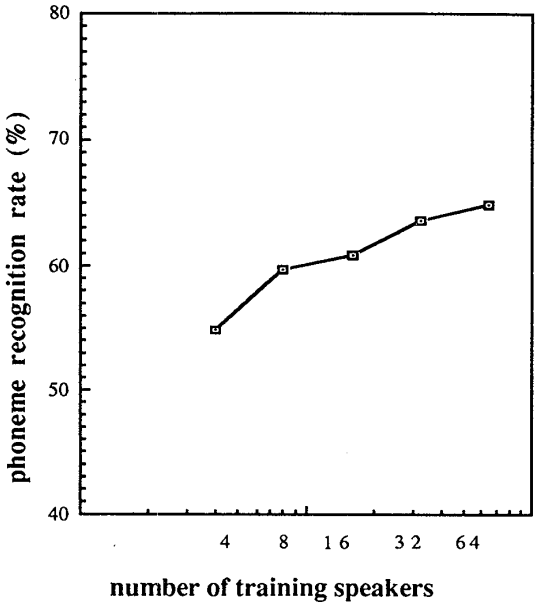



図 7 学習話者数に対する対話読上げ文中の音韻認識率

Fig. 7 Number of training speaker vs. phoneme recognition rate for ASJ read dialog database.

も認識率の飽和は見られない．しかし，認識率の上昇は，人数の増加に伴って少なくなっており，飽和傾向にあることがわかる．ATR のデータベースと音響学会のデータベースでの実験結果の違いは主に，録音環境の違い，話者の発声スタイルの違いによるものと考えられる．

## 5. むすび

本論文では，従来までの音韻ラベルに基づく音韻認識率の評価に代わる，新しいラベルなし評価法を提案した．この評価法は，音韻ラベルを必要とせずに，音韻 HMM の音韻認識率を調べることができる．不特定話者連続音声データベースを用いて HMM の連結学習を行い，この新しい評価法を用いて，学習話者の増加に対する HMM の音韻認識率の評価を行った．

具体的には，ATR の連続音声データベース，音響学会のデータベースを使って，学習話者数の増加に対する音韻認識率の変化を評価した．この結果，ラベルなし評価法は，音韻ラベルを利用しないでも音韻 HMM を評価できることが確認された．本手法を用いれば，データベースに音韻ラベルを付けるという，膨大な労力を必要としないで，音韻 HMM の評価ができる．

ラベルなし評価法の今後の課題として次のようなことがあげられる．ラベルなし評価法でも，音韻環境依存型のモデルを評価するときに，評価する音韻だけでなく，前後の音韻のことも考慮にいれなければならない．また，ラベルなし評価法は連続音声認識での音韻 HMM の置換のしやすさを評価したものであるが，挿入や脱落などの評価も行う必要がある．今後これらの問題点を解決していく予定である．

**謝辞** 日頃御指導頂くヒューマンインタフェース研究所古井特別研究室古井室長，有意義な議論をして下さった古井特別研究室の方々に感謝致します．

本実験では，日本音響学会連続音声認識データベースおよび ATR の日本語音声データベースを用いた．

### 文　献

(1) Lee K-F., Hon H-W. and Hwang M-Y.: "Recent Progress in the SPHINX Speech Recognition System", Proc. of the DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, Inc. (Feb. 1989).

(2) Kubala F. and Schwartz R.: "A New Paradigm for Speaker-Independent Training", Proc. ICASSP-91, pp. 833-836 (May 1991).

(3) Kubala F. and Schwartz R.: "A New Paradigm for Speaker-Independent Training and Speaker Adaptation", Proc. of the DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, Inc. (June 1991)

(4) Shirotsuka O., Arima I. and Kawai G.: "Spoken Japanese Sentence Recognition Based on Hidden Markov Models", 音響講論集, 1-5-61 (1991-10).

(5) 南　泰浩，松岡達雄，鹿野清宏："不特定話者連続音声データベースによる連結学習HMMの評価"，信学技報，**SP91**-113 (1992-01).
(6) Haung X.: "Minimizing Speaker Variation Effects for Speaker Independent Speech Recognition", Proc. Speech and Natural Language Workshop, pp. 191-196 (Feb. 1992).
(7) Lee K-F. and Hon H-W.: "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM", ICASSP '88, pp. 123-126 (April 1988).
(8) 丸山活輝，花沢利行，川端　豪，鹿野清宏："HMM音韻連結学習を用いた英単語音声の認識"，信学技報，**SP88**-119 (1988-01).
(9) Levinson S. E., Rabiner L. R. and Soundhi M. M.: "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition", The Bell System Technical Journal, **62**, 4 (April 1983).
(10) 武田一哉，匂坂芳典，片桐　滋，阿部匡伸，桑原尚夫："研究用日本語音声データベース利用解説書"，ATRテクニカルレポート，TR-I-28 (1988).
(11) 阿部匡伸，匂坂芳典，梅田哲夫，桑原尚夫：研究用日本語音声データベース利用解説書(連続音声データ編)"，ATRテクニカルレポート，TR-I-166 (1990).
(12) 田中和世，ほか："Intelligent Speech Processing System"，日本情報処理開発協会 (1992-06).
(13) 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"，音響誌，**44**，10，pp. 747-754 (1988-10).

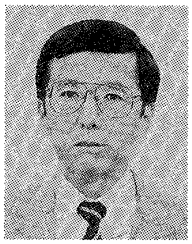
(平成5年7月6日受付，9月22日再受付)

**南　泰浩**

昭61慶大・理工・電気卒，平3同大大学院博士課程了．同年日本電信電話(株)入社．現在ヒューマンインタフェース研究所研究主任．音声認識の研究に従事．日本音響学会会員．

**松岡　達雄**

昭57早大・理工・電子通信卒．昭59同大大学院修士課程了．同年日本電信電話公社(現NTT)入社．横須賀電気通信研究所においてディジタル電話端末の通話系構成法の研究に従事．昭62より，NTTヒューマンインタフェース研究所においてニューラルネット，HMMによる音声認識の研究に従事．平4～5 AT&T Bell研究所 (Murray Hill) において客員研究員としておもに話者適応化の研究に従事．現在，NTTヒューマンインタフェース研究所古井特別研究室主任研究員．日本音響学会，IEEE各会員．



**鹿野　清宏**

昭45名大・工・電気卒．昭47同大大学院修士課程了．同年電電公社武蔵野電気通信研究所入所．昭59～61カーネギーメロン大客員研究員．昭61～平2 ATR自動翻訳電話研究所音声情報処理研究室長．現在，NTTヒューマンインタフェース研究所主席研究員．工博．主として音声認識の研究に従事．昭50本会米沢賞，平3 IEEE SP 1990 Senior Award受賞．IEEE，音響学会，情報処理学会各会員．