音声言語情報処理 19－7
(1997. 12. 11)

# 音素毎の話者クラスタリングに基づく話者適応法

○阿部 俊朗†, 鈴木 基之‡, 牧野 正三‡, 阿曽 弘具†

東北大学大学院工学研究科 † 東北大学大型計算機センター ‡
(〒980-77 東北大学工学部電気・情報系)†
(〒980-77 東北大学大型計算機センター)‡

あらまし 木構造話者クラスタリングによる話者適応法は認識対象となる話者がモデルに存在しない場合でも柔軟に対応できる手法である。しかしこの方法はどの音素も話者の個人性を同じものとして捉えている。そこで我々は音素毎に話者の個人性を考慮した話者適応法を提案する。認識実験をおこなった結果、音素による違いを考慮しない方法によりもよい結果を得た。これは、各音素毎に最適な話者クラスタを選択できたことが有効に働いたためと考えられる。

キーワード 不特定話者音声認識, 話者適応, 木構造話者クラスタリング, SSS-free

# A speaker adaptation based on speaker clustering considering phoneme characteristics

Toshiaki Abe†, Motoyuki Suzuki‡, Shozo Makino‡ and Hirotomo Aso†

Graduate School of Engineering, Tohoku University†
Computer Center, Tohoku University‡
(Faculty of Engineering, Tohoku University 980-77 Japan)†
(Computer Center, Tohoku University 980-77 Japan)‡

Abstract The tree-structured speaker clustering as a speaker adaptation method can select the model which is most similar to a target speaker. In the previous works, this method was not considered about speaker difference dependent on phoneme class. In this report, we propose a speaker adaptation method based on speaker clustering taking difference dependent on phoneme class into account. The experimental results tells us that the new method gives a better performance, because it selects most similar model for every phoneme class.

**英文 keywords** speaker-independent speech recognition, speaker adaptation, tree-structured speaker clustering, SSS-free

## 1. はじめに

現在音声認識の分野において、隠れマルコフモデル ( HMM ) に基づいた統計的手法が幅広く研究されている。しかし HMM を用いても話者や語彙の制約がない場合、高精度な認識を達成するにはまだ多くの問題を残している。そのひとつに話者の個人性による音響的特徴の違いの問題がある。不特定話者モデルは、多数の話者の音声情報を用いて学習するので話者の個人性による音声パターンの変動をある程度吸収することができる。しかしそれは話者の特徴を幅広くとっているためで、認識対象である話者の特徴を端的に表わしているわけではなく、話者の個人性を十分に捉えたモデルとは言えない。

そこで話者の個人性の問題を改善するために、認識をおこなう前に対象となる話者の情報を利用して、事前に用意した不特定話者モデルを認識対象の特徴にあうようにモデルを再構成する話者適応法が研究されてきた。なかでも、"木構造話者クラスタリングによる話者適応法" [5](小坂、1993) はごく少量の話者音声で高精度の話者適応が可能であるといわれている。このモデルは木構造の各ノードに HMnet が配置されるようになっている。木構造の上層においては不特定話者モデルの特徴を持ち、下層では特定話者モデルの特徴を持っている。また認識時には認識対象が学習で使用した話者の特徴に近ければ下層のモデルを選択し、そうでなければ上層のモデルを選択するので木構造を構成する話者に含まれない話者においても柔軟に適応することが可能である。

しかしこの手法では話者の個人性はどの音素についても同じものとして捉えて話者適応をおこなっている。一方話者の個人性は母音や鼻音にあると言われ、すべての音素の個人性を一律的に扱うより、音素毎に個人性を考慮した話者適応を考えた方が精度の高い話者適応が可能であると考えられる。

本研究では、音素を単位に話者の個人性を考慮した話者適応手法を提案し、その性能を認識実験により比較、検討する。

## 2. 音素毎の個人性を考慮した木構造話者クラスタリング

音素毎の個人性を考慮した木構造話者クラスタリングのアルゴリズムを以下に示す。

### 2.1 学習アルゴリズム

(1) **多数の特定話者モデルの作成**

1 人が発声した多量の学習音声を用いて初期 HMnet を作成する。この HMnet に VFS [4] をおこなうことで、近似的に多数の話者の特定話者 HMnet を生成する。ここで学習に VFS を用いるのは、比較的少量の音声で学習ができることまた構造の同じ HMnet を得ることができるためである。

(2) **各音素毎に木構造話者クラスタの作成**

HMnet の中の各音素に対応する部分を抜き出す。これを音素 HMnet と呼ぶことにする。HMnet を音素 HMnet と区別するために全音素 HMnet と呼ぶ。各々の音素 HMnet すべてに次の手続きで木構造クラスタを作成する。

i) クラスタを構成する特定話者音素 HMnet において距離 $D_p(M_i, M_j)$ が最大となる 2 つのモデル $M_i$ と $M_j$ を求める。ただし $D_p(M_i.M_j)$ は各々の特定話者 HMnet の音素 $p$ に対応する状態の出力分布の Bhattacharyya 距離の総和である。

ii) 各特定話者モデルを距離 $D_p$ を用いて $M_i$、$M_j$ に近いほうに分類し、クラスタを 2 つのサブクラスタに分割する。

iii) サブクラスタに含まれる話者情報を用いて、クラスタの代表となる HMnet を作成する。各状態の出力確率は式 (1)を用いて合成した。これはクラスタ内の各話者の HMnet の出力確率の重み付き和である。

$$b^{(i)}(x) = \sum_s \frac{n_s^{p(i)}}{\sum_s n_s^{p(i)}} N(\boldsymbol{\mu}_s^{(i)}, \boldsymbol{\Sigma}_s^{(i)}) \quad (1)$$

ただし $i$ は状態番号、$s$ はその話者クラスタに含まれる話者、$n_s^{p(i)}$ は話者 $s$ の HMnet の状態 $i$ に対応する音素 $p$ のサンプル数を表わす。

iv) クラスタに含まれる話者数が 1 になるまで ii) で分割されたサブクラスタにおいても同様に i) ～ iii) をおこなう。

また全音素 HMnet においても同様の方法により木構造話者クラスタを作成する。

### 2.2 話者適応アルゴリズム

認識対象となる話者の音声を使用して、各音素毎に最適な話者クラスタを選択する。適応音声中に存在する各音素に関して、その音素の木構造話者クラスタを用いて以下の手順で木構造を探索してモデルを選ぶ。

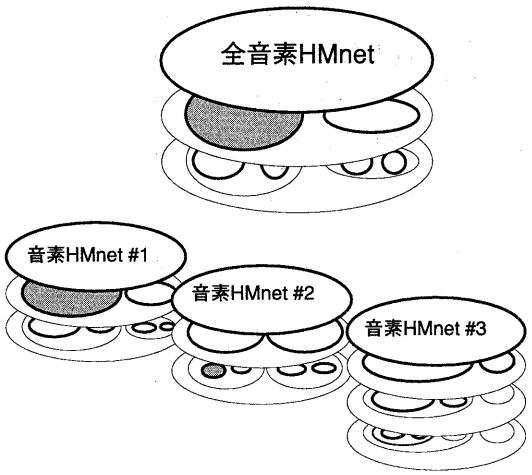


図 1. 音素毎の木構造話者クラスタ

(1) クラスタを最上層にマークし、このクラスタの HMnet に適応音声を与え尤度を求める。

(2) マークされたクラスタに属するサブクラスタの HMnet に適応音声を与え、最尤となる HMnet を持つサブクラスタを選択する。

(3) 選択されたサブクラスタにマークし、クラスタにサブクラスタが存在しなくまるまで (2) を続ける。

(4) (1)〜(3) の処理により全ての階層で 1 つのクラスタが選択される。その全てのクラスタの HMnet の中から更に最尤となる HMnet を選択し、これを認識に用いる。

ただし適応用音声に存在しない音素に関しては全音素 HMnet で構成した木構造クラスタから、適応音声に含まれる全ての音声サンプルを用いて同様の方法により最尤となる HMnet をもつ話者クラスタを選択し、その音素に対応する HMnet の一部を抜きだし認識に用いる。

図 1を例に説明する。いま適応用音声に音素 #1 と音素 #2 が含まれ、音素 #3 については存在しなかったとする。このとき適応用音声に存在しなかった音素 #3 は全音素 HMnet で選択された話者クラスタの 音素 #3 に対応する HMnet の情報 (出力確率) を使用する。

提案した手法では、木構造を音素毎に作っているため、音素毎に話者の個人性が異なっていても、その個人性を忠実に表現できるような木構造クラスタを構成することができると期待される。また、話者クラスタを各音素毎に選択するために、認識対象となる話者の特徴により合ったモデルを選択することが期待できる。

## 3. 実験

音素毎に個人性を考慮した木構造話者クラスタと考慮しない木構造話者クラスタを用いた話者適応について木構造を構成する話者が少ない場合と多い場合について認識実験をおこなった。実験条件は表 1 のとおりである。

### 3.1 木構造を構成する話者が少ない場合の認識実験

まず木構造を構成する話者数が比較的少ない場合に関して実験をおこなった。初期 HMnet は ATR 日本語音声データベース音韻バランス 503 文章から男性 1 名が発話した 400 文章を用いて学習した。また 50 文章の音声で VFS を用いて 8 人分の HMnet を作成した。認識には男性 2 人と女性 2 人について 50 文章を用いた。木構造話者クラスタを選択する適応音声の量は 1〜5 文章と変化させて認識実験をおこなった。

認識結果を図 2に示す。図中の “音素毎” のデータが音素の個人性を考慮した木構造話者クラスタ、“HMnet” のデータが考慮していない木構造話者クラスタを用いて話者適応をおこなった場合のそれぞれの子音・母音の認識率を示している。また、音素毎にどのような木構造を構成しているか、どのノードを選択したかその結果の一部を図 3に示す。このとき認識対象の話者は男性 ( MSH ) であり 1 文章を使って適応をおこなったものである。またここで出てくる話者の頭の 1 文字目が M は男性を F は女性を示している。

認識結果を見ると母音に関しては、2%程度の認

| | |
|---|---|
| 初期 HMnet | 状態数 210 |
| | 特定話者モデル |
| | SSS-free[3]により学習 |
| 各状態の分布 | 単一ガウス分布 |
| | 対角共分散行列 |
| パラメータ | logpow,cep(16),Δlogpow |
| | Δcep(16) からなる |
| | 34 次元ベクトル |
| 分析条件 | サンプリング周波数 12kHz |
| | 16bit 量子化 |
| | 20ms ハミング窓 |
| | フレーム周期 5ms |

表 1. 実験条件

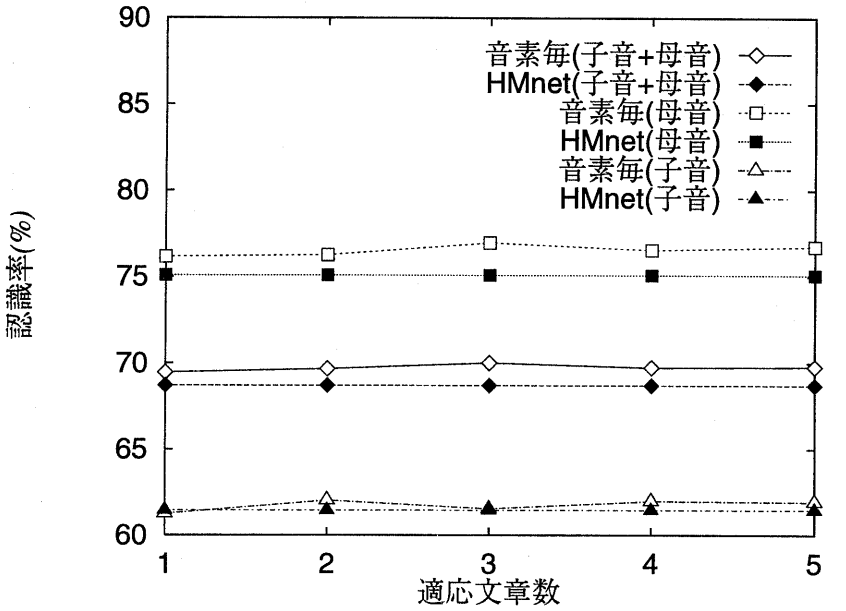


**図 2. それぞれの木構造クラスタを用いた話者適応の音素認識実験結果**

識率の向上がみられた。また子音については認識率にほとんど差が見られなかった。音素毎に木構造話者クラスタの構造を見てみると、ほぼどの音素においても上層の方は男性と女性への分割のように同じような分割がおこなわれ、下層にいくにしたがって次第に異なる分割をしてゆく。図 3の /sh/ 以外がその例である。全音素 HMnet で構成した木構造話者クラスタと比較すると 2 層目までは同じ構成であり、男性と女性とにクラスタが分割されている。また2層より下の階層では各々異なる分割になっている。

しかし、上層での分割も男女には分かれず、他のものと全く異なる構造を持つ音素 HMnet も得られた。図3の /sh/ がその例である。その理由は、その対象となった音素について、初期 HMnet から VFS で不特定話者モデルを作成するときの学習データが少なかったことにあると思われる。その音素認識率を調べると音素の個人性を考慮していない木構造話者クラスタを用いた方が良い傾向にあることがわかった。これは VFS に与えられた音声の量が少ないため本来の音素の特徴とは異なった音素モデルが生成されたためであると思われる。その結果、木構造話者クラスタは音素の個人性を反映したものとはいい難くなってしまう。これらのことは音素の個人性を考慮せずに木構造を生成すれば起こらないことであるので、音素毎に木構造を作成したことがが悪く働いたと考えられる。

次に、適応時に選択されるクラスタがどのような話者から構成されているかを、全音素 HMnet での木構造話者クラスタと、音素毎の木構造話者クラスタについて比較した。子音の場合は、音素毎の木構造から選択された話者クラスタは、全音素での木構造話者クラスタにおいて選択された話者クラスタに類似する場合が大半となっていた。すなわち、双方ともほぼ同じ話者から構成されているクラスタ、あるいは片方がそのようなクラスタと木構造上で隣接したクラスタであった。これが子音に認識率の改善が見られない原因である考えられる。

一方母音の場合はそれぞれの音素で異なる話者クラスタを選択していて共通性というものはあまり見られなかった。これは音素毎に話者の個人性を考慮して木構造話者クラスタを作成したことに効果があることを示していると思われる。

### 3.2 木構造を構成する話者が多い場合の認識実験

次に木構造を構成する話者が多人数の場合について実験をおこなった。実験条件は表 1であり、初期 HMnet も少人数での実験と同じ条件である。また 50 文章の音声で VFS を用いて 50 人分の HMnet を作成した。認識は男性 2 人と女性 2 人について 50 文章でおこなった。適応音声の量は 1 文章である。ただし、認識対象となる話者、木構造を構成する話

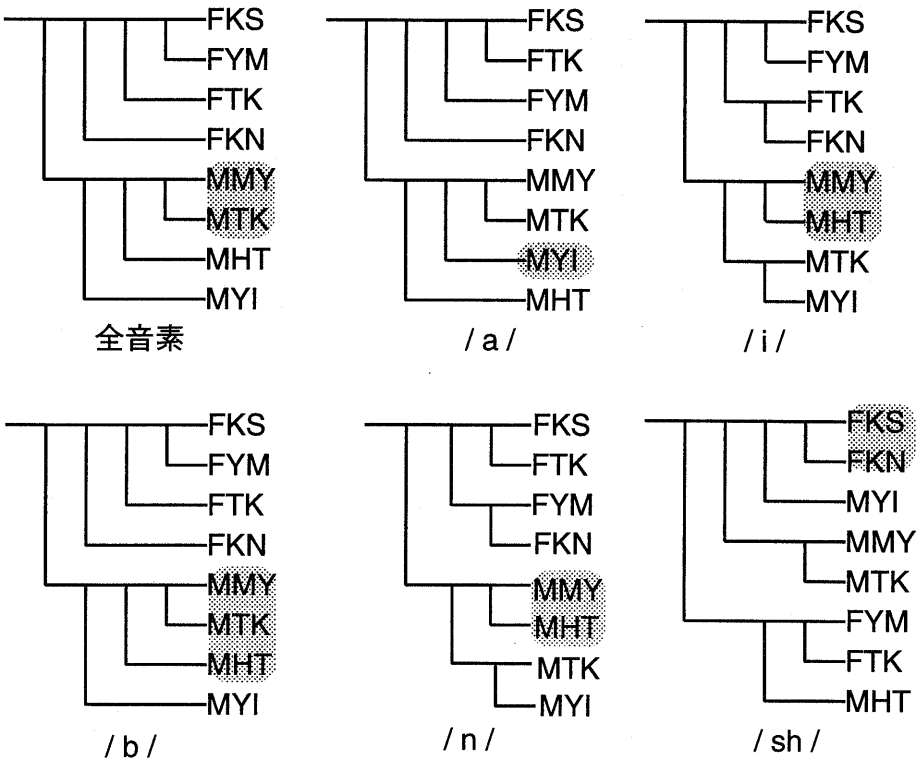


**図 3. 話者 MSH における音素毎の木構造クラスタ**

| | 音素毎 (%) | HMnet(%) |
|---|---|---|
| 平均 | 60.66 | 58.87 |

**表 2. 多人数で構成される木構造話者クラスタによる認識率**

者は少人数での実験で使ったものと異なるので少人数でおこなった場合の認識率とこの認識率を直接比較することはできない。

認識結果を表2に示す。この結果より、少人数の場合と同様に多人数で木構造を構成した場合でも音素毎に話者性を考慮したモデルのほうが良い認識結果を得た。少人数の実験と同じように木構造の構造を調べてゆくと、少人数の場合では上層の方は男性と女性への分割のように同じような分割がおこなわれ、下層にいくにしたがって次第に異なる分割をしていったが、多人数で構成した木構造話者クラスタの場合には、上層のほうにおいても異なる分割がおこなわれているものが増える傾向にある。

適応時に選択された話者クラスタがどのような話者で構成されているかを調べたところ、少人数の場合と同じような傾向がみられた。つまり、子音において選択された話者クラスタは互いに類似する場合が多く、母音の場合はそれぞれの音素で異なる話者クラスタを選択していて共通性というものはあまり見られなかった。

## 4. まとめ

音素毎に話者の個人性を考慮した木構造話者クラスタリングに基づく話者適応法を提案した。そのモデルの認識性能を評価するために実験をおこなった。木構造を構成する話者が少ない場合でも多い場合でも従来法よりもよい認識性能を得ることができた。これは音素毎に最適な話者クラスタを選択することが有効に働いている。

## 参考文献

[1] 中川聖一：“確率モデルによる音声認識”，電気情報通信学会 (1988)

[2] 鷹見淳一、嵯峨山茂樹：“逐次状態分割法による隠れマルコフ網の自動生成”，信学論、J76-D-II,No.10,pp.2155-2164(1993)

[3] M. Suzuki, S. Makino, A. Ito, H. Aso and H. Shimodaira : "A New HMnet Construction Algorithm Requiring No Contextual Factors", IEICE Trans. Inf. & Syst., E78-D, No.6, pp. 662–668 (1995).

[4] 大倉計美, 杉山雅英, 嵯峨山茂樹："混合連続分布 HMM を用いた移動ベクトル場平滑化話者適応方式", 信学技報, SP 92-16(1992)

[5] 小坂哲夫, 鷹見淳一, 嵯峨山茂樹："木構造話者クラスタリングを用いた話者適応", 電子情報通信学会論文誌 A,Vol.J77-(A), No.2, pp. 103–111(1994)

[6] 阿部 俊朗, 鈴木 基之, 牧野 正三, 阿曽 弘具: "音素毎の話者クラスに基づく話者適応法" , 日本音響学会平成 9 年度秋季研究発表会, 1-1-14 (1997-9)