論文 / 著書情報
Article / Book Information

論題(和文)	話者交代検出を含むオンライン話者適応の検討
Title(English)	
著者(和文)	張 志鵬, 古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会 1999年秋季講演論文集, Vol. , No. 1-1-23, pp. 45-46
Citation(English)	, Vol. , No. 1-1-23, pp. 45-46
発行日 / Pub. date	1999, 9

話者交代検出を含むオンライン話者適応の検討 *

◎張志鵬　古井 貞熙 (東工大)

1. はじめに

音声認識の多くのアプリケーションにおいて、話者の変化を自動的に検出しながら、オンライン（教師なし）で逐次的に話者適応を行うことが有用であると考えられる。われわれはこれまでに、過去の話者に適応化した音素 HMM と、不特定話者用 HMM を並列的に用いて音声認識を行い、話者の交代時点を自動的に検出して、オンライン逐次話者適応を行う方法を提案し、ニュース音声の認識実験によってその有効性を確認した [1]。しかしこの方法には、複数の音素 HMM セットによる認識を並列的に行うため、計算量が大きくなるという問題があった。　本稿では、HMM の代りに 1 状態の混合ガウス分布（GMM）を用いて、話者交代を検出することにより、計算量を削減する方法を提案する。

2. オンライン逐次話者適応法

2.1 GMM の尤度比較による話者変化の検出

テキスト独立形話者認識において、これまで混合ガウス分布モデル（GMM）が広く使われている [2]。ここでは、GMM を話者交代の検出に用いることを試みる。このためには、GMM が話者適応化後の音素 HMM の個人差を十分に表現している必要がある。本研究においては、音素 HMM の話者適応は基本的に MLLR（最大尤度線形変換）法 [3] によって行なわれる。このため、音素 HMM の適応化と並行して、GMM を HMM と同じ変換行列で適応化することにした。ただし、音素 HMM は後述するように 7 つの音素クラスタに分けて変換しているが、GMM は音素独立のため、音素 HMM を 1 クラスタで変換する変換行列を求め、これを用いて GMM を変換する。

2.2 適応化の手順

オンライン逐次話者適応の流れを図 1 に示す。入力音声に対しては、まず不特定話者用（SI）GMM とそれまでの話者に適応化した（SA）GMM に対する尤度を求める。もし (Case 1) SA GMM に対する尤度の方が大きければ、同じ話者が継続していると判断し、その話者に適応化した SA HMM で音声認識する。そのデコーディング結果と入力音声を用いて、音素 HMM と GMM をさらに話者適応化する。もし (Case 2) SA GMM よりも SI GMM に対する尤度の方が大きければ、話者が交代したと判断し、SI HMM で音声認識する。そのデコーディング結果と入力音声を用いて、新たな SA HMM と SA GMM を作る。

ニュース音声などでは、過去に発声したアナウンサーなどが再度発声することがあるので、過去の SA HMM や SA GMM は、総数があらかじめ定めた上限を越えない限り保存しておいて、次の入力音声の話者の候補者に含める。今回の実験では、尤度比較するモデルとして、最新の 2 人の話者と不特定話者の三種類のモデルを使うことにした。

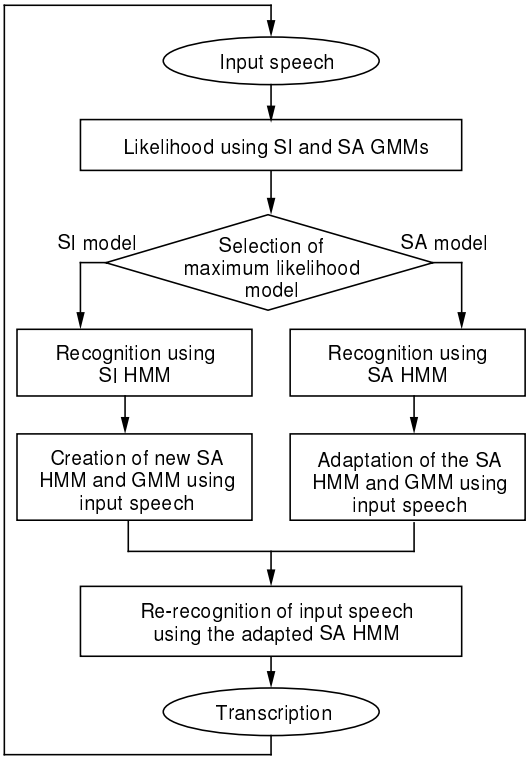

図 1. オンライン適応の流れ (**SI Model** :不特定話者モデル, **SA Model** :話者適応化モデル)

2.3 適応化アルゴリズム

適応手法はまず MLLR [3] 及び MAP [4] によりモデルパラメータの変換行列を求め、その後 VFS [5] により移動ベクトルを平滑化する方法を用いた。音素による違いを考慮して、無音、子音、各 5 母音、計 7 つのクラスタに分類し、各クラスタに対する変換行列を求めた。

3. 認識実験

3.1 音響モデルと言語モデル

今回の実験で用いた SI HMM は、tree-based clustering によって状態共有化を行なった不特定話者文脈依存音素 HMM である。音響特徴量としては 16 次の LPC ケプストラムと正規化対数パワー、及びそ

---

*A Study of On-line Speaker Adaptation Combined with Speaker-Change Detection
By Zhipeng Zhang and Sadaoki Furui (Tokyo Institute of Technology)

れらの一次微分の計 34 次元を用いた。モデルの総状態数は男性が 2106、女性が 2083 である。各状態のガウス分布の混合数はすべて 4 である。混合ガウス分布モデルは同じデータを用いて 64 混合のモデルである。

言語モデルの学習に用いたデータは放送ニュース原稿テキスト 5 年分、約 50 万文である。単語出現頻度上位 2 万語を認識語彙とし、間投詞と読みを考慮した言語モデル [6] を用いた。

3.2 評価用データ

実際に放送されたニュース音声から、スタジオで収録されたクリーンな発話をそれぞれ男女 50 文ずつ抽出した。各評価セットには 5〜6 名の話者の音声が含まれている。

3.3 認識実験結果

bigram を用いる場合と trigram を用いる場合の実験結果を図 2 に示す。図には、適応化を行う前 (baseline)、前回提案した複数の HMM セットによって話者交代を検出する方法 (”HMM”)、今回提案する方法 (”GMM”) による単語誤り率を示す。いずれの評価セットに対しても、適応化により誤り率が低下していることが分かる。”HMM”法に比べ、”GMM”法でも女性では性能の低下はなく、男性でもわずかの低下ですむことがわかる。”GMM” 法により、適応化前に比べて誤り率は男女平均で 10.0%低下している。

次に、逐次的に適応を行わず、入力音声ごとに不特定話者 HMM を用いてオンライン即時適応を行ったときの比較実験を行った。実験結果を図 3 に示す。図には、適応化を行う前 (baseline)、即時適応法 (”instantaneous”)、提案した逐次適応法 (”incremental”) による誤り率を示す。いずれの条件においても即時適応法より逐次適応法のほうが誤り率が低下することが確認された。

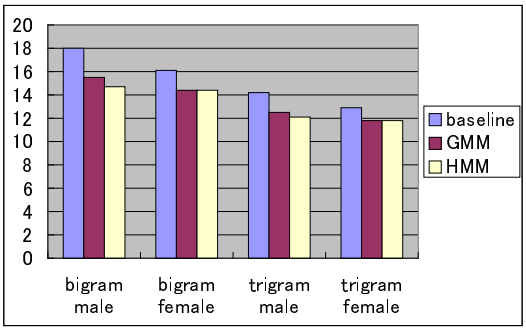

図 2. 提案した逐次適応法の認識結果 (単語誤り率 [%])

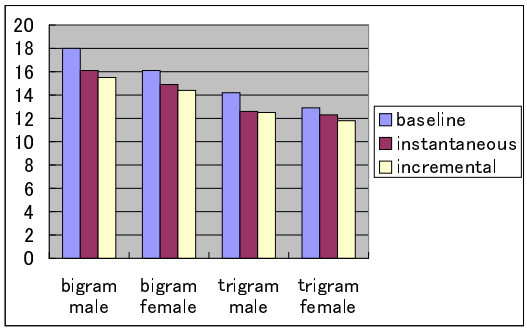

図 3. 提案した逐次適応法と即時適応法の比較実験 (単語誤り率 [%])

4. まとめ

不特定話者用混合ガウス分布モデル (GMM) と、話者に適応した GMM に対する入力音声の尤度比較によって話者交代を検出しながら、MLLR-VFS 法でオンライン逐次話者適応を行う手法を提案した。前回に提案した複数の HMM セットを用いる方法に比べて大幅に計算量を削減しながら、わずかの性能低下ですみ、適応化前に比べて単語誤り率が男女平均で約 10%減少することを確認した。さらに即時適応法との比較によって逐次適応法の効果が確認された。

今後は音声区間の自動切り出し、文の区切りの自動決定、雑音への対処法などと組み合わせていく予定である。

謝辞

ニュース原稿及び音声データを提供して頂いた NHK 放送技術研究所に感謝します。ご助言を頂いた NTT ヒューマンインタフェース研究所の大附克年氏に感謝します。日頃討論頂く東工大の研究室の方々に感謝します。

参考文献

[1] 張 他, 春季音学議論, pp.103-104, 1999-3

[2] 松井 他, 電子情報通信学会論文誌, A Vol.J 77-A No.4 , pp.601-606, 1994-4

[3] C.J.Leggetter et al., Computer Speech and Language, Vol.9, pp.171-185, 1995-9

[4] J.-L.Gauvain et al., IEEETrans. on Speech and Audio Processing,Vol.2,No.2, pp.291-298,1994-4

[5] 大倉 他, 信学論, Vol.J76-D-II, Nol.12, pp.2468-2476, 1993-12

[6] 桜井 他, 春季音学議論,pp.57-58, 1999-3