論文 / 著書情報
Article / Book Information

論題(和文)	ニュース音声認識の話者適応法の検討
Title(English)	
著者(和文)	張 志鵬, 桜井 直之, 古井 貞熙, 大附 克年
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会1999年春季講演論文集, Vol. , No. 3-1-4, pp. 103-104
Citation(English)	, Vol. , No. 3-1-4, pp. 103-104
発行日 / Pub. date	1999, 3

ニュース音声認識の話者適応法の検討 *

◎張志鵬　桜井 直之　古井 貞熙 (東工大)　大附 克年 (NTT ヒューマンインタフェース研究所)

1. はじめに

話者適応技術は、音声認識において話者の個人差の問題に対処する重要な手段である。放送ニュース音声の認識においては、話者の交代が頻繁に起こり、しかも未知の話者の音声が入力されるので、オンラインの教師なし適応を行うことが必要である。本稿では、尤度比較により話者境界を自動的に検出しながら音素モデルを適応する、オンライン即時 ·逐次型話者適応手法を提案する。

2. 話者検出に基づく話者適応法

2.1 ニュース音声の特徴

ニュース音声にはスタジオのアナウンサーによる発声だけでなく、中継先の記者や VTR 映像にあわせて原稿を読み上げた発声など様々な話者の発声が含まれている。この話者適応に関しては、次のことを考慮することが必要かつ有効と考えられる。

(1) 事前に話者情報を得ることができないので、オンライン即時型適応が必要。(2) 同じ話者が複数の文を続けて発声することが多いので、逐次型適応が有効。(3) 話者の交代情報を得ることができないので、自動的に検出することが必要。本研究では、このような観点から、話者境界を自動的に検出しながらオンライン即時 ·逐次型教師なし話者適応を行う方法について検討した。

2.2 尤度比較による新話者検出

不特定話者の音素モデルを尤度最大化規準で特定の話者に適応化した場合、同じ話者の異なる音声に対するそのモデルの尤度は、不特定話者のモデルの尤度よりも高くなると期待される。逆に、新しい話者の音声の声質がそれ以前の話者の音声と異なる場合には、新しい話者の音声は、以前の話者に適応化したモデルよりもむしろ不特定話者のモデルに適合すると考えられる。従って、適応化モデルと不特定話者モデルの尤度を比較することによって話者境界を検出し、高い尤度を示すモデルを用いて、新しい話者に適応させるのが適当と考えられる。そして、同じ話者が複数の文を継続して発声していると判定される間は、そのモデルを逐次適応化して行くことにより、認識性能が向上すると予想される。さらに、新しい話者が検出された後でも、ニュース音声の場合は、以前のアナウンサーが再度発声することが考えられるので、ある程度の数の話者に適応したモデルをそれぞれ保存しておき、活用するのが適当であろう。このような適応化法はニュース音声認識だけでなく、対話システム、会議など、話者交代を伴う多くの場合に使えると考えられる。今回の実験では、計算時間を考慮し、尤度比較するモデルとして、直前の話者、現在の話者、および不特定話者の三つのモデルを使うことにした。オンラインの適応の流れを図 1 に示す。

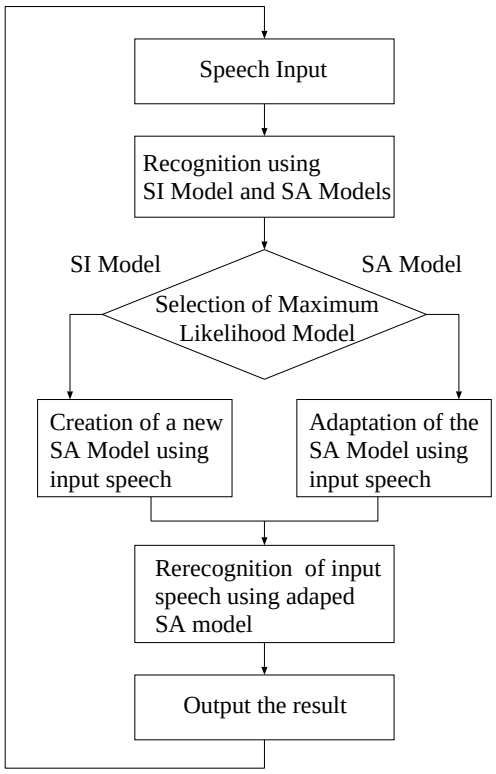

図 1. オンライン適応の流れ (**SI Model** :不特定話者モデル, **SA Model** :話者適応化モデル)

2.3 適応手法

適応手法はまず MLLR [1] 及び MAP [2] によりモデルパラメータの変換行列を求め、その後 VFS [3] により移動ベクトルを平滑化する方法を用いた。すべての音素を共通の行列で変換する場合と、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用いる場合について検討した。

3. 認識実験

3.1 音響モデル

今回の実験で用いた音響モデルは、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。

---

*A Study of Speaker Adaptation for Automatic Transcription of Japanese Broadcast-News Speech

By Zhipeng Zhang,Naoyuki Sakurai, Sadaoki Furui (Tokyo Institute of Technology) and Katsutoshi Ohtsuki (NTT Human Interface Laboratories)

3.2 言語モデル

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(1992年7月から1996年5月), 約50万文である。形態素解析システムJTAGを用いて形態素に分解し，その形態素を単語として単語bigram , trigramを学習した。単語出現頻度上位2万語を認識語彙とした。

3.3 評価用データ

1996年7月に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(clean)をそれぞれ男女(male,female)50文ずつ抽出した。(以下m/c,f/cと略す) 各評価セットには5〜6名の話者の音声が含まれている。

3.4 認識実験結果

二種類の言語モデル[4]を用いた実験を行った。LM1は読みを考慮した言語モデル、LM2はさらに間投詞対策として、形態素に読点を含めた言語モデルである。これらの言語モデルによる認識結果を用いて、教師なしの適応化を行った。まずLM1の結果の男女各50文をそれぞれ用いて適応実験を行なった。その結果を図2に示す。図には、適応化を行う前(baseline)、音素を1クラスタで適応化した場合(1 cluster)、および7クラスタ(7 clusters)の結果が示してある。言語モデルとして、bigramを用いる場合とtrigramを用いる場合の結果を示した。全ての評価セットにおいて、適応化により単語正解精度が向上していることが分かる。1 clusterの場合、bigramでは平均10.8%, trigramでは平均9.6%誤り率が減少している。7 clustersを用いるとさらに認識性能が向上し、1 clusterの場合に比べて、bigramでは平均6.6%, trigramでは平均5.5%誤り率が減少している。7 clustersで適応化した場合、適応化前に比べて、誤り率が男女平均で約15%減少している。

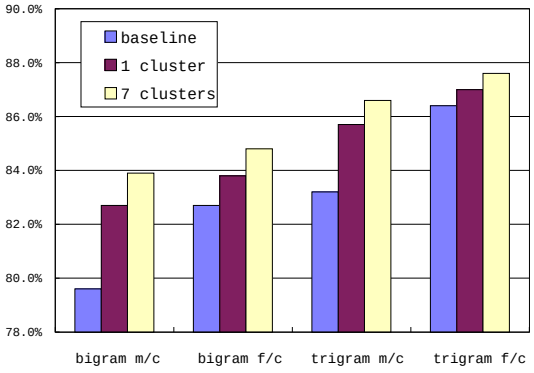

図 2. LM1での適応結果(単語正解精度[%])

LM2による実験結果を図3に示す。1 clusterの場合、bigramモデルにおいて平均7.8%, trigramでは平均4.2%誤り率が減少している。7 clustersを用いると、 1 clusterの場合よりもbigramで平均7.0%, trigramで平均5.2%誤り率が減少している。7 clustersで適応化した場合、適応化前に比べて、誤り率が男女平均で約12%減少している。

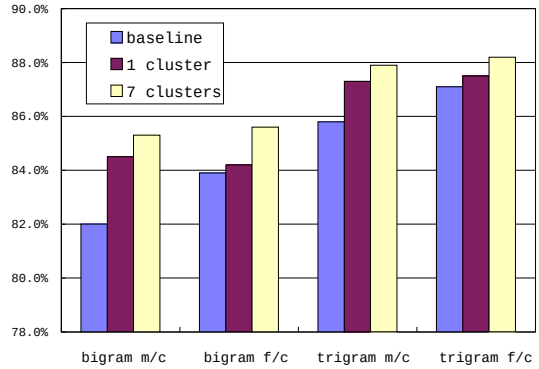

図 3. LM2での適応結果(単語正解精度[%])

参考のための比較実験として、正しい話者境界を与え、そこで適応の初期モデルとして不特定モデルを強制的に用いる実験を行った。LM2での適応実験を行なった結果を図4に示す。1 clusterの場合の男女平均認識率は、話者境界を自動検出した場合と強制的に与えた場合でほぼ同じであるが、7 clustersの実験では自動検出の方が平均3.6%誤り率が小さい。声質の似ている話者のモデルを初期モデルとして適応する方法が、不特定話者モデルを初期モデルとして使うより有効だと言える。

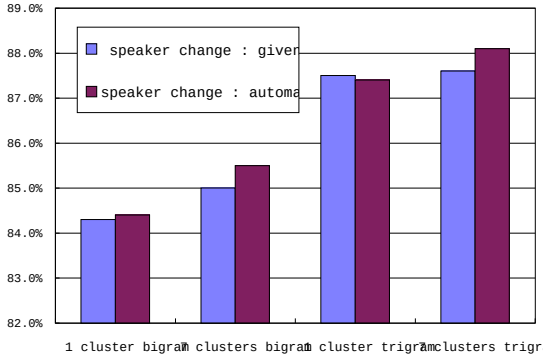

図 4. 話者境界を与えた場合の適応結果との比較 (m,f/c, 単語正解精度[%])

4. まとめ

ニュース音声の特徴を用いた、尤度比較による新話者検出に基づく教師なしオンライン即時 ·逐次型話者適応化法を検討した。音素クラスタを用いるMLLR-MAP-VFS法で単語誤り率が、男女平均で約12〜15%減少することを確認した。

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

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

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

[4] 桜井 他, 春季音学議論, 1999-3