音声言語情報処理　34－15
（2000．12．21）

充足統計量と話者距離を用いた音韻モデルの教師なし学習

芳澤伸一*　馬場朗*　松浪加奈子**　米良祐一郎**　山田実一**　鹿野清宏**
*　（財）イメージ情報科学研究所
**　奈良先端科学技術大学院大学　情報科学研究科
*〒630-0101　奈良県生駒市高山町 8916-12
**〒630-0101　奈良県生駒市高山町 8916-5
yosizawa@nara.image-lab.or.jp

あらまし　充足統計量と話者距離を用いた音韻モデルの教師なし学習法を提案する。提案法では、（1）発声話者に音響的に近い話者を選択し、（2）選択された話者の充足統計量を用いて発声話者に適応した音韻モデルを作成する。充足統計量の計算は適応処理の前にオフラインで行う。提案法では少量の発声文章で適応処理が行われる。また、充足統計量を利用することにより短時間で適応処理が行われる。話者クラスタリングによる方法と比較すると、提案法では発声話者のデータによりオンラインで動的に話者クラスタを決定するため、適切な話者クラスタを獲得することができる。認識実験により、少量の発声文章により適応を行った場合、MLLR より高い認識率を獲得できることを示す。

キーワード　音韻モデル、話者適応、充足統計量、教師なし学習

Unsupervised training based on the sufficient HMM statistics from selected speakers

Shinichi Yoshizawa*　Akira Baba*　Kanako Matsunami**　Yuichiro Mera**
Miichi Yamada**　Kiyohiro Shikano**
* Laboratories of Image Information Science and Technology Ikoma, 630-0101, Japan
** Nara Institute of Science and Technology, Ikoma, 630-0101, Japan
yosizawa@nara.image-lab.or.jp

Abstract　This paper describes an efficient method of unsupervised training. This method is based on (1) selecting a subset of speakers who are acoustically close to a test speaker, and (2) calculating adapted model parameters according to the previously stored sufficient HMM statistics of the selected speakers' data. In this method, only a few unsupervised test speaker's data are required. Also, by using the sufficient HMM statistics of the selected speakers' data, a quick training can be done. Compared with a pre-clustering method, the proposed method can obtain a more optimal speaker cluster because the clustering result is determined according to test speaker's data on-line. Experiment results show that the proposed method attains better improvement than MLLR from the speaker-independent model. Moreover the proposed method utilizes only one unsupervised sentence utterance, while MLLR usually utilizes more than ten supervised sentence utterances.

key words　acoustic model, speaker adaptation, sufficient statistics, unsupervised

1 はじめに

充足統計量と話者距離を用いた音韻モデルの教師なし学習法を提案し、発声話者に適応した音韻モデルを構築する。本論文では、少数の適応用の発声データにより短時間に話者適応を行うことを目的とする。

さまざまな話者適応アルゴリズムが提案されている。特定話者(-like)モデルもその一つである。特定話者(-like)モデルでは、発声話者のデータ、もしくは音響的特徴が近い話者のデータを用いて発声話者に適応した音韻モデルを構築する。この方法により高い認識率を獲得することができる。しかし、音韻モデルの学習に時間がかかるため、即時性が要求されるもとでの適応（音韻モデルの作成）が困難である。

上述の問題を解決するため、話者クラスタリングによる方法が提案されている[1][2][3]。この方法では、学習データを音響的に近い話者集合（話者クラスタ）に分割し話者集合ごとに音韻モデルをオフラインで学習する。そして、発声話者の発声データが入力されると発声話者に適した音韻モデルが選択され認識が行われる。適応段階ではモデル選択を行うのみであるため即時に適応が行われる。この方法では、事前にどのような話者クラスタを構築しておくかが高い認識率を獲得するための鍵となる。

一方、適応アルゴリズムでよく利用されているものとして MLLR[4][5][6]がある。MLLR と話者クラスタリングを組み合わせた方法[1]も提案されている。MLLR は高い認識率を獲得することができる。しかし、MLLR では比較的多くの音素ラベル付きの発声データを用いて適応を行うため、適応処理時間が長くなる。また、発声話者は比較的多くの文章を発声しなくてはならない。

本論文では、新しい話者適応モデルの学手法を提案する。提案法では、充足統計量と話者距離を用いて音素ラベルなしの発声データにより音韻モデルの構築を行う。提案法では、（1）発声話者に音響的特徴が近い話者を選択し、（2）選択された話者の充足統計量を用いて発声話者に適応した音韻モデルを作成する。充足統計量の計算は適応処理の前にオフラインで行う。提案法では少量の発声文章で適応処理が行われる。また、充足統計量を利用することにより短時間で適応処理が行われる。話者クラスタリングによる方法[1][2][3]と比較すると、提案法では発声話者のデータによりオンラインで動的に話者クラスタを決定するため、適切な話者クラスタを獲得することができる。認識実験により、少量の発声文章により適応を行った場合、MLLR より高い認識率を獲得できることを示す。

2 充足統計量と話者距離による教師なし学習法

提案法の概念図を図1に示す。提案法は3つのステップから構成されている。第1ステップでは、話者ごとの HMM に関する充足統計量を計算し蓄積する。第2ステップでは、GMM (Gaussian mixture model) により表現された話者モデルを用いて、発声話者に音響的特徴が近い話者を選択する。第3ステップでは、第2ステップにおいて選択された話者の充足統計量を用いて発声話者に適応した音韻モデルを作成する。以下に各ステップについて詳細に述べる。

2．1 充足統計量

本節では第1ステップについて述べる。本ステップでは、話者ごとの HMM に関する充足統計量を計算してシステムに蓄積する。この処理はオフラインで行われる。ここで充足統計量とは、HMM などの音韻モデルにおける平均、分散、E-M カウントなどの統計量のことである。HMM の充足統計量を話者ごとのデータを用いて独立に計算する。充足統計量は、EM アルゴリズムにより不特定話者モデルから一回学習することにより計算される。

2．2 話者クラスタリング

本節では第2ステップについて述べる。本ステップでは、GMM (Gaussian mixture model) により表現された話者モデルにより、発声話者に音響的特徴が近い話者を選

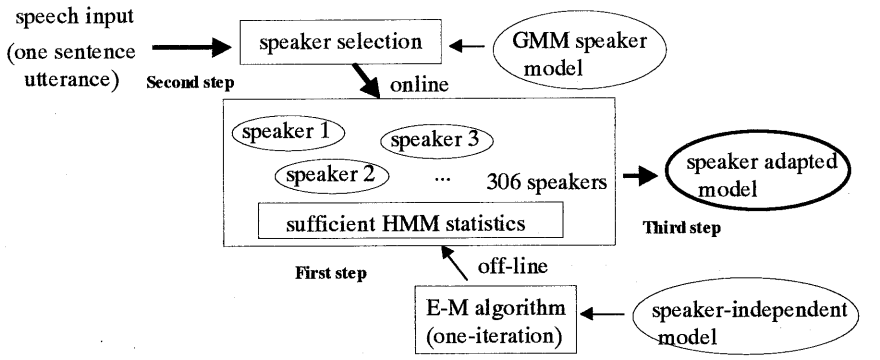

Figure 1: Blockdiagram of the proposed method based on speaker selection and sufficient HMM statistics

択する。GMM は混合正規分布で表現した観測データの確率モデルである。発声話者に適応した音韻モデルを構築するためには、発声話者に音響的に近い話者のデータのみを用いて音韻モデルを作成することは重要である。

本論文では、話者選択モデルとして 64 混合の GMM を用いる。GMM は音韻独立に作成する。発声話者とシステムに蓄積された話者との音響的な距離を測るため、発声データに対する GMM の尤度を用いる。尤度が高い話者ほど発声話者に音響的に近い話者である。GMM は構造が単純であるため、少量の発声データにより適切に発声話者に近い話者を選択することができる（本論文では、適応用の発声データとして一文章の発声を用いている)。第 2 ステップでは尤度の高い上位 N 人の話者を選択する。

話者クラスタリングによる方法[1][2][3]と比較すると、提案法では発声話者のデータによりオンラインで動的に話者クラスタを決定するため、適切な話者クラスタを獲得することができる。

2．3　話者適応モデル

本節では第 3 ステップについて述べる。本ステップでは、前節により選択された話者の充足統計量を用いて発声話者に適応した音韻モデルを作成する。

充足統計量を利用することで短時間に適応処理を行うことが可能になる。充足統計量をオフラインで計算して、適応処理では簡単な統計処理計算で音韻モデルを作成することができるためである。また、提案法は変換行列に基づいた MLLR[5][6]のような構造的制約はない。

第 3 ステップで、発声話者に音響的に近い話者の充足統計量を用いて、発声話者の音韻モデルを統計処理計算により作成する。上記の処理は、発声話者に音響的に近い話者のデータを用いて、不特定話者モデルから 1 回学習して音韻モデルを作成することと同等である。

3　認識実験

提案法の有効性を示すために認識実験を行う。認識実験には日本音響学会による JNAS データベース[7]を利用する。JNAS データベースは 306 人の話者データより構成されており、各話者は約 200 文章の発声データをもつ。サンプリング周波数 16kHz、16bit のデータである。特徴量として、窓シフト長 10ms で分析した 12 次元の MFCC (Mel-frequency cepstrum coefficient)とデ

ルタケプストラム、デルタパワーを用いる。特徴量抽出において CMN (Cepstrum mean normalization) 処理がほどこされている。20k の新聞記事より構築した言語モデルとデコーダとして Julius を用いる。

3種類の音韻モデルにより評価を行う。2種類のモノフォーンモデル（16 混合、64 混合）と PTM (Phonetic Tied Mixture)モデル[8]を音韻モデルとして用いる。PTM モデルは、状態共有トライフォーンにしたがう混合重みをもつ、文脈非依存 HMM である。43 音素の HMM により評価を行う。評価に用いる PTM モデルは、全体として 2500 状態から構成され、各状態ごとに 64 混合の正規分布をもつ。一般に PTM モデルはモノフォーンモデルと比較して高い認識率を獲得できる。

46 話者の発声を評価用データとして用いる。評価用データは学習用データに含まれていない。提案法において、話者適応モデルは発声話者（評価話者）を除く話者の充足統計量により作成される。提案法では適応データとして音素ラベルなしの一文章の発声を用いる。

ベースラインモデルとして不特定話者モデルを用いる。ベースラインモデルにおける単語誤り率（置換誤り、挿入誤り、脱落誤りを含む）は、18.5%（モノフォーン 16

Table1. Comparison with MLLR

method		proposed method	MLLR		speaker-independent model
		unsupervised	supervised		
# of sentence utterances		1	10	50	-----
word error rate	monophone model (16 Gaussians)	14.9%	15.6%	13.8%	18.5%
	monophone model (64 Gaussians)	10.9%	12.6.%	12.0%	13.5%
	PTM (Phonetic Tied Mixture)	8.3%	9.0%	7.6%	10.0%

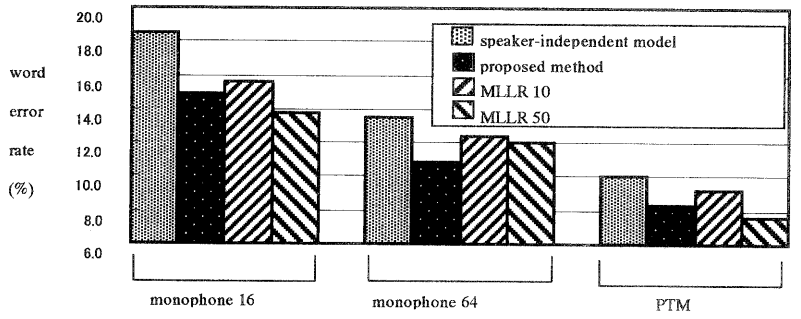

Figure 2: Comparison with various models

混合)、13.5% (モノフォーン 64 混合)、10.0% (PTM) であった。

代表的な適応アルゴリズムである MLLR [5]での適応結果を表 1 に示す。

提案法による結果を図 3 に示す。図 3 には第 2 ステップでの選択話者数と単語誤り率の関係が示されている。図 3 より、最小の単語誤り率は、14.9% (モノフォーン 16 混合)、10.9%(モノフォーン 64 混合)、8.3% (PTM) である。今回の実験では、PTM モデルにおける適応処理時間は、提案法は MLLR より約 3 倍 (発声データ 10 文章)、約 16 倍 (発声データ 50 文章) 速かった。これらの結果を表 1 と図 2 にまとめる。

4 考察

表 1、図 2 の結果より、提案法は 10 文章の発声を用いた MLLR より低い単語誤り率を獲得できることがわかる。MLLR では低い単語誤り率を獲得するために 10 文章以上の発声が必要である。適応処理時間は、PTM モデルにおいて提案法は MLLR より速い。適応処理時間の差は適応文章数が多くなるほど大きくなる。この結果より提案法は少量の発声文章により適応を行う場合に特に有効であることがわかる。また、提案法では音素ラベルなし、MLLR では音素ラベル付きの発声により適応を行う。音素ラベル付きの発声データを獲得する手段として書き上げ文章を読み上げることが考えられる。この操作は任意の文章を発声する場合と比較して労力は大きい。

第 2 ステップにおける選択話者数について考察する。図 3 に選択話者数と単語誤り率の関係を示す。単語誤り率が最小の選択話者数は、20 人(モノフォーン 16 混合)、40 人(モノフォーン 64 混合)、80 人(PTM)であ

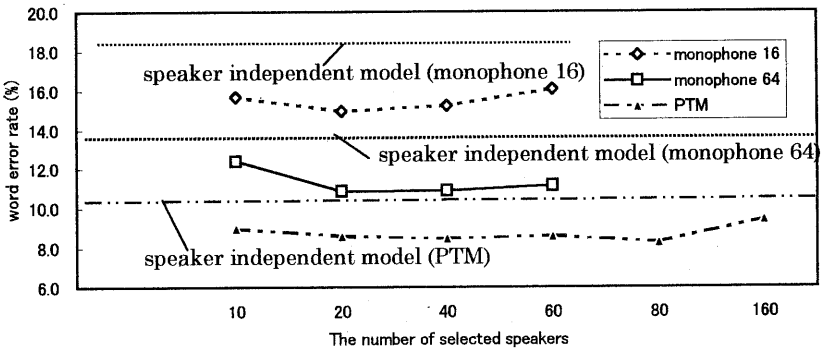

Figure 3: Word error rate for the proposed method

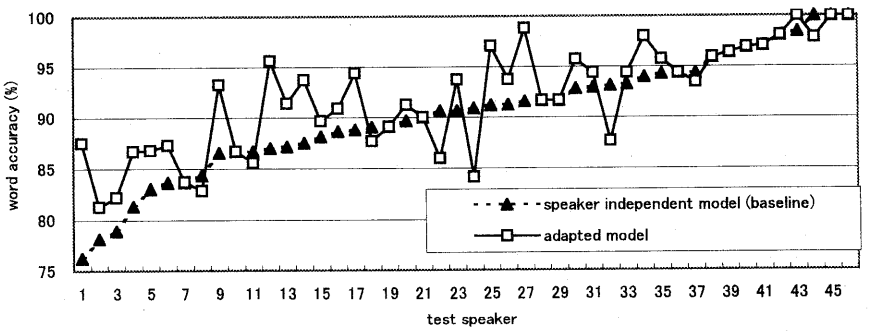

Figure 4. Improvement of word accuracy for each speaker using PTM model

る。この結果より音韻モデルが複雑になるにしたがい学習に必要な話者数が増加していることがわかる。

また、話者ごとの認識率の改善について考察する。図4にPTMモデル（選択話者数80人）における各話者に対する認識率の改善を示す。横軸は話者を適応前の認識率が悪い順に並べかえたものである。結果より適応前に低い認識率であった話者の認識率が特に改善されることがわかる。また最低認識率が大きく改善されている。

5 むすび

充足統計量と話者距離を用いた音韻モデルの教師なし学習法を提案した。提案法では、（1）発声話者に音響的特徴が近い話者を選択し、（2）選択された話者の充足統計量を用いて発声話者に適応した音韻モデルを作成する。提案法では少量の発声文章で適応処理が行われる。また、充足統計量を利用することにより短時間で適応処理が行われる。話者クラスタリングによる方法[1][2][3]と比較すると、提案法では発声話者のデータによりオンラインで動的に話者クラスタを決定するため、適切な話者クラスタを獲得することができる。認識実験により、少量の発声文章により適応を行った場合、MLLRより高い認識率を獲得できることを示した。

謝辞　本研究は、NEDO：新エネルギー・産業技術総合開発機構の援助を受けて行われた。ご協力いただいた関係各位に感謝します。

参考文献

[1] M.Padmanabhan, L.R.Bahal, D.Nahamoo, M.A.Picheny, "SPEAKER CLUSTERING AND TRANSFORMATION FOR SPEAKER ADAPTATION IN LARGE-VOCABULARY SPEECH RECOGNITION SYSTEM", Proceedings of the ICASSP, pp.701-704, 1995.

[2] 佐藤庄衛,世木寛之,尾上和穂,今井亨,田中秀樹,安藤彰男, "話者クラス音響モデルのための学習データの自動選択手法",信学技報,SP2000-11,pp.9-15,2000.

[3] 加藤恒夫,黒岩眞吾,清水徹,樋口宜男, "多数話者電話音声データベースを用いた話者クラスタリング",信学技報,SP2000-10,pp.1-8,2000.

[4] Yuqing Gao, Mukund Padmanabhan, Michael Picheny, "SPEAKER ADAPTATION BASED ON PRE-CLUSTERING TRAINING SPEAKERS", Proceedings of the Eurospeech, pp.2091-2094, 1997.

[5] C.J.Leggetter, P.C.Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language, vol.9, pp.171-185, 1995.

[6] M.J.F.Gales and P.C.Woodland, "Mean and variance adaptation within the MLLR framework", Computer Speech and Language, vol.10, pp.249-264, 1996.

[7] Katunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, Kiyohiro Shikano, and Shuichi Itahashi, "JNAS:Japanese speech corpus for large vocabulary continuous speech recognition research", The Journal of the Acoustical Society of Japan (E), Vol.20,No.3,pp.199-206,1999 .

[8] Akinobu Lee, Tatsuya Kawahara, Kazuya Takeda, Kiyohiro Shikano, "A NEW PHONETIC TIED-MIXTURE MODEL FOR EFFICIENT DECODING", Proceedings of the ICASSP, pp.1269-1272, 2000.