社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

話者適応における学習データ内の音素数と認識精度の考察

松浦　祥悟†　　村上　仁一†　　池原　　悟†

† 鳥取大学工学部知能情報工学科　〒 680-8552 鳥取市湖山町南 4-101
E-mail: †{s022048,murakami,ikehara}@ike.tottori-u.ac.jp

あらまし　本研究は，学習データに含まれる音素数に着目し，音素数により話者適応 HMM と不特定話者 HMM を組み合わせた混合 HMM を作成する．また，話者適応の認識精度がより高くなるように，各音素の数に偏りを持たせることで学習データを作成する．実験の結果，作成した混合 HMM と偏りを持たせた学習データを用いて，164 単語の話者適応において 2.65%，82 単語の話者適応において 1.33%の改善が得られた．

キーワード　話者適応 不特定話者 学習データ 連続型 HMM

The consideration between number of phone in training sets and recognition accuracy for speaker adaptation

Shougo MATSUURA†, Jin'ichi MURAKAMI†, and Satoru IKEHARA†

† Faculty of Engineering,Tottori University　Minami 4-101,Koyama,Tottori,680-8552,Japan
E-mail: †{s022048,murakami,ikehara}@ike.tottori-u.ac.jp

Abstract　This study pay attention to number of phones in training sets. We make a hybrid HMM which mixed speaker adaptation HMM and independent HMM by number of phones. In addition, we make training sets by given bias to number of phones to obtain the high performance in speaker adaptation. As a result of experiments, we have improvement of 2.65% and 1.33% for 164 words training sets and 82 words training sets in speaker adaptation using a hybrid HMM and made training sets.

Key words　speaker adaptation，speaker independent speech recognition，training data sets，continuous HMM

1. はじめに

現在，不特定話者音声認識には複数話者の音声を 1 つの HMM に学習する手法 [1] や，複数の話者を選択的に用いる話者選択型 [2] などの手法がある．しかし，不特定話者の認識精度では不十分である．そこで認識精度を向上させる手法として，認識する話者のデータを利用する話者適応が挙げられるが，認識する話者のデータを大量に収集することは困難であり，限られたデータでより効果的に話者適応を行う必要がある．

話者適応には MLLR [4]，MAP 推定 [5] などの様々な手法がある．しかし，学習データ量が少ない状況では認識精度が向上するとは限らない．また，話者適応に用いる学習データ内に含まれる音素数が，認識精度に与える影響についてはあまり考察されていないようである．そこで本研究では，学習データ内の音素数に着目し，認識精度を低下させず，より効果的な話者適応を行う手法について検討する．

まず，音素数によって話者適応 HMM と不特定話者 HMM を組み合わせた混合 HMM を作成する．混合 HMM は，話者適応の学習データにおいて，数が多い音素は信頼性が高いと考えて話者適応 HMM を使用し，数が少ない音素は信頼性が低いと考えて不特定話者 HMM を使用する．

また，話者適応の学習において，より効果的な学習データについて考察する．数の多い音素をより多く，数の少ない音素をより少なく偏りを持たせることで，話者適応 HMM の精度が良くなると考え，学習データを作成する．本研究では基本となる学習データと，偏りを持たせた学習データを用いて認識精度を調査する．

2. 話者適応

2.1 話者適応 HMM の作成

本研究は，連結学習を用いて，教師あり話者適応を行う．具体的には，不特定話者の音素 HMM を初期モデルとして，話者適応の学習データを用いて連結学習を行い，話者適応の音素 HMM を作成する．不特定話者の音素 HMM の作成から話者適応の音素 HMM 作成までの手順を図 1 に示す．

(1)　不特定話者の学習データを用いて学習を行い，不特定話者 HMM を作成する．
(2)　不特定話者 HMM に話者適応の学習データを用いて連結学習を行い，話者適応 HMM を作成する．
(3)　話者適応 HMM を用いて認識し，評価する．

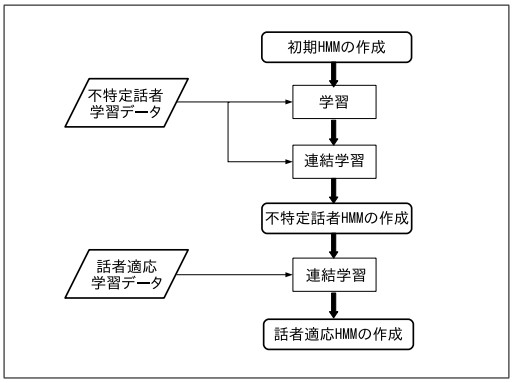

図 1　話者適応 HMM の作成手順

2.2　話者適応 HMM と不特定話者 HMM の利用

話者適応において認識精度の低下を防ぐために，話者適応 HMM と不特定話者 HMM を組み合わせた混合 HMM を用いて認識を行う．混合 HMM は，学習データ内に含まれる音素数が一定数未満の音素に不特定話者 HMM を用いて，その他の音素に話者適応 HMM を用いて作成する．また今回の実験は，学習データ内の音素数の変化による認識精度の違いを調査する．そこで不特定話者 HMM を用いる条件を，学習データ内に含まれる音素の数が 10 個，20 個，30 個未満の音素の 3 種類とし，3 種類の混合 HMM を作成する．

3.　評 価 実 験

本研究では，不特定話者 HMM，話者適応 HMM，混合 HMM のそれぞれを用いて単語音声認識を行い，認識精度の調査を行う．

3.1　実 験 条 件

実験には単語音声認識ツールの HTK [3] を使用する．データベースとして ATR 単語発話データベース Aset の男女各 10 名を使用する．不特定話者 HMM を作成する学習データは，認識する話者以外の 9 話者の奇数番号データ (1 話者につき 2,620 単語) を使用する．話者適応の学習データは，認識する話者の奇数番号データを条件によって選択した単語を使用する．評価データは，認識する話者の偶数番号データ (2,620 単語) を使用する．本研究は，男性話者 mau，mmy，mnm，女性話者 faf，fms，ftk の男女計 6 名を認識する話者として実験を行い，認識率の平均を求める．

3.2　評価データの音素の分布

本研究で使用する評価データの音素は，母音は “u”，“a”，子音は “k” が特に多く，続いて “r”，“s”，“m”，“t” が多い．

3.3　話者適応に用いる学習データ

本研究で用いる話者適応の学習データは，データベースの奇数番号から 164 単語，82 単語を選択して作成する．82 単語の学習データは，作成した 164 単語の学習データの単語数を半分にして作成する．また，164 単語，82 単語の学習データそれぞれに基準の違うリスト 1，リスト 2 を作成する．ここで，2 つのリスト内に含まれる全音素の音素数の合計は同程度とする．作成する際の選択基準を以下に示す．

リスト 1

- 各音素数の割合が評価データの割合と同程度

リスト 2

- 音素数が上位の子音をリスト 1 より多く含むように選択
- リスト 1 の音素で割合が 2%未満 (10 個) となった音素を削除

評価データと作成した学習データのリスト 1，リスト 2 の音素の出現率を降順にソートした分布を以下に示す．164 単語の学習データの分布を図 2，82 単語の学習データの分布を図 3 に示す．

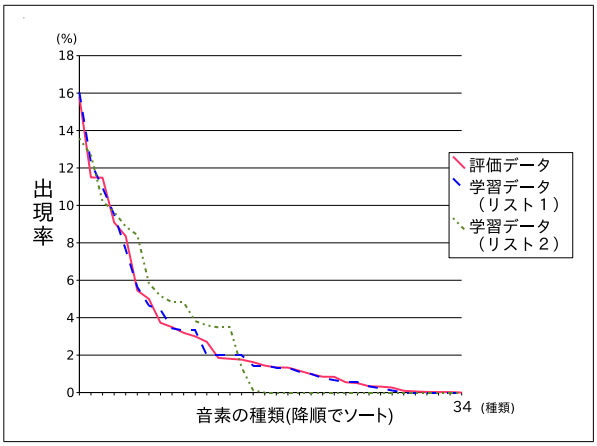

図 2　164 単語学習データ内の音素数の出現率

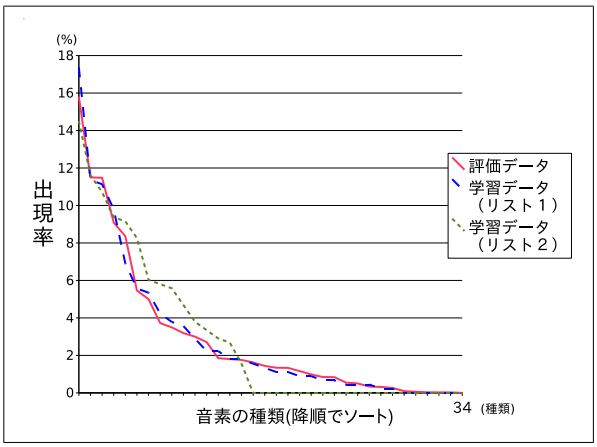

図 3　82 単語学習データ内の音素数の出現率

図 2，3 中の，音素の出現率が 10%より上に分布している音素は母音である．母音は学習データ作成の際に偏りを持たせなかったため，3 種類のデータはほぼ同じ分布となっている．リスト 1 の分布は評価データの分布とほぼ重なっている．リスト 2 はリスト 1 と比較して，音素数が上位の子音の出現率が高く，出現率が 2%以下の子音は，ほぼ存在しない．

3.4 特徴パラメータ

まず不特定話者音素 HMM を作成する．不特定話者 HMM の特徴パラメータは MFCC を，共分散行列は Diagonal-covariance を使用する．その他の特徴パラメータの実験条件を表 1 に示す．

表 1　特徴パラメータの実験条件

基本周波数	16kHz
分析窓	Hamming 窓
分析窓長	25ms
フレーム周期	10ms
音響モデル	3 ループ 4 状態 (連続分布型)
stream 数	3
特徴	12 次 MFCC+Δ 12 次 MFCC
パラメータ	+対数パワー+Δ対数パワー (計 26 次)
	(母音・撥音)
連続型	MFCC 10, Δ MFCC 10,
HMM の	対数パワー 4, Δ対数パワー 4
初期モデル	(その他の子音)
混合分布数	MFCC 4, Δ MFCC 4,
	対数パワー 2, Δ対数パワー 2

4. 実験結果

4.1 不特定話者音声認識

不特定話者音声認識の認識結果を表 2 に示す．表中の括弧内の分母は認識話者の評価データ数，分子は認識できた単語数を示す．

表 2　不特定話者音 HMM の単語音声認識の誤り率

	不特定話者
男性話者	11.16% (877/7860)
女性話者	11.18% (879/7860)
平均	11.17% (1756/15720)

4.2 話者適応

164 単語の話者適応の認識結果を表 3 に，82 単語の話者適応の認識結果を表 4 に示す．

表 3　164 単語話者適応 HMM の単語音声認識の誤り率

	リスト 1	リスト 2
男性話者	13.33% (1048/7860)	9.07% (713/7860)
女性話者	13.63% (1071/7860)	9.21% (724/7860)
平均	13.48% (2119/15720)	9.14% (1437/15720)

表 4　82 単語話者適応 HMM の単語音声認識の誤り率

	リスト 1	リスト 2
男性話者	33.79% (2656/7860)	13.82% (6774/7860)
女性話者	32.38% (2545/7860)	13.12% (1031/7860)
平均	33.08% (5201/15720)	13.47% (2117/15720)

実験より以下の結果を得た．

(1)　リスト 2 を用いた方が，164 単語，82 単語の両方の話者適応においてリスト 1 を用いるよりも認識精度が高い．

(2)　164 単語の話者適応において，リスト 1 を用いた場合は不特定話者より認識精度が低くなり，リスト 2 を用いた場合は不特定話者より認識精度が高い．

(3)　82 単語の話者適応は，リスト 1，リスト 2 の共に不特定話者よりも認識精度が低い．

4.3 話者適応 HMM と不特定話者 HMM の混合 HMM

164 単語適応における混合 HMM の認識結果を表 5 に，82 単語適応における混合 HMM の認識結果を表 6 に示す．

表中の “10 個未満混合 HMM” は，164 単語の話者適応を行い，学習データ内の音素数が 10 個未満の音素を不特定話者 HMM と入れ換えて作成した混合 HMM を用いた認識実験である．

表 5　164 単語適応における混合 HMM の単語音声認識の誤り率

	10 個未満混合 HMM	20 個未満混合 HMM	30 個未満混合 HMM
リスト 1			
男性話者	10.27% (807/7860)	9.47% (744/7860)	9.47% (744/7860)
女性話者	9.71% (763/7860)	8.22% (646/7860)	8.22% (646/7860)
平均	9.99% (1570/15720)	8.84% (1390/15720)	8.84% (1390/15720)
リスト 2			
男性話者	9.07% (713/7860)	8.89% (699/7860)	8.89% (699/7860)
女性話者	9.21% (724/7860)	8.15% (641/7860)	8.15% (641/7860)
平均	9.14% (11437/15720)	8.52% (1340/15720)	8.52% (1340/15720)

表 6　82 単語適応における混合 HMM の単語音声認識の誤り率

	10 個未満 混合 HMM	20 個未満 混合 HMM	30 個未満 混合 HMM
リスト 1			
男性話者	14.62% (1149/7860)	12.72% (1000/7860)	11.37% (894/7860)
女性話者	13.64% (1072/7860)	11.12% (874/7860)	10.11% (795/7860)
平均	14.13% (2221/15720)	11.92% (1874/15720)	10.74% (1689/15720)
リスト 2			
男性話者	11.46% (901/7860)	10.75% (845/7860)	10.42% (819/7860)
女性話者	10.86% (853/7860)	9.76% (767/7860)	9.26% (728/7860)
平均	11.16% (1754/15720)	10.25% (1612/15720)	9.84% (1547/15720)

実験より以下の結果を得た．

(1)　30 個未満混合 HMM を用いた場合，164 単語，82 単語の共に不特定話者 HMM や話者適応 HMM よりも認識精度が高い．

(2)　リスト 2 の学習データを用いた方が，全ての条件で認識精度が高い．

(3)　164 単語，82 単語の混合 HMM の共に，30 個未満混合 HMM の認識精度が最も高く，混合 HMM の基準の音素数が少なくなるほど認識精度が下がる．

(4)　164 単語の学習データを用いた実験では，学習データ内に 20 個以上 30 未満の範囲内に音素が存在しなかったために，20 個未満と 30 個未満の混合 HMM の認識精度が同じとなった．

話者適応 HMM と混合 HMM を用いた単語音声認識実験より，学習データ内の音素数が上位の音素は話者適応 HMM を用いて，下位の音素は不特定話者 HMM を用いて作成した混合 HMM の有効性が得られた．また，音素数に偏りを持たせて作成したリスト 2 の学習データが有効であるとわかった．

5. 考　　察

5.1　話者適応による認識精度の低下

164 単語の学習データを用いた話者適応において，認識精度はリスト 1 を用いた場合に不特定話者より低く，リスト 2 を用いた場合に不特定話者より高かった．

表 7 に 164 単語・リスト 1 の学習データを用いた話者適応の実験において，学習データ内の音素数が少なく，適応前と適応後で認識精度が低下した音素の例を示す．

表 7　164 単語話・リスト 1 者適応 HMM の音素誤り率

学習データ内 音素数	音素 n (18 個)	音素 y (13 個)	音素 ch (9 個)
不特定話者 平均	6.32% (106/1678)	1.89% (20/1056)	4.59% (41/894)
話者適応 平均	7.39% (124/1678)	4.17% (44/1056)	7.72% (69/894)

結果からリスト 1 は，学習データ内に数が少ない音素が多数存在するため，認識率が低下したと考えている．これに対し，リスト 2 は数の少ない音素をあらかじめ省いて作成しているため，認識精度が低下していない．また，82 単語の学習データを用いた話者適応は，単語数が減ったことにより，データ内の数の少ない音素が増え，リスト 1，リスト 2 の共に認識精度が低下した．以上より，話者適応に用いる単語数が減少した場合は，学習データの単語をより詳細な条件で選択することで，認識精度が改善できると考えている．

5.2　音素による認識精度の違い

学習データ内の音素数がほぼ同様の音素について，話者適応を行った場合の認識精度の違いを調査する．82 単語・リスト 2 の学習データの音素数が 20 個以上 30 個未満の範囲に，ほぼ同じ数の音素が 4 種類ある．この 4 種類の音素は 30 個未満混合 HMM では不特定話者 HMM を用いて，20 個未満混合 HMM では話者適応 HMM を用いている．4 種類の音素の 30 個未満混合 HMM と 20 個未満混合 HMM の音素認識精度を表 8 に示す．

表 8　82 単語・リスト 2 混合 HMM の音素誤り率

学習データ内 音素数	音素 s (27 個)	音素 r (26 個)
20 個未満 HMM 平均	0.44% (14/3174)	8.41% (382/4542)
30 個未満 HMM 平均	0.50% (16/3174)	4.38% (199/4542)

学習データ内 音素数	音素 t 25(個)	音素 m (25 個)
20 個未満 HMM 平均	5.91% (161/2724)	2.93% (85/2898)
30 個未満 HMM 平均	5.10% (139/2724)	3.80% (110/2898)

各音素の認識精度の結果から，同数の音素数を用いて学習しても，音素によって認識精度に差があることがわかる．本研究の条件では 6 話者全ての実験で，音素 “r” の適応後の認識精度が悪くなった．また，音素 “s” は不特定話者において高い認識率を得ているため改善が難しいことがわかる．

これらの結果は，用いるデータベース，学習方法などによって変化すると考えられる．予め認識精度の傾向をとらえ，学習データに反映することで認識精度を改善できると考えている．

5.3　母音と子音による認識精度の違い

4.3 節で用いた混合 HMM は子音と母音を区別せずに作成している．本節では，子音のみ話者適応 HMM を用いた混合 HMM と，母音のみ話者適応 HMM を用いる混合 HMM を作成し，認識精度の違いを調査する．話者適応を行う際の学習データは 82 単語・リスト 1 を用いる．

5.3.1 子音のみ話者適応 HMM の利用

子音の音素数が上位の音素のみ話者適応 HMM を用い，その他の音素は不特定話者 HMM を用いて作成した混合 HMM の認識精度を求める．実験の結果を表 9 に示す．

表 9 子音のみ話者適応 HMM を利用して単語音声認識した場合の誤り率

		話者適応 HMM を用いた音素	
	不特定話者	k	k s
男性話者	13.33% (1048/7860)	10.33% (812/7860)	10.62% (835/7860)
女性話者	13.63% (1071/7860)	10.57% (831/7860)	10.79% (848/7860)
平均	13.48% (2119/15720)	10.45% (1643/15720)	10.71% (1683/15720)

子音 “k” のみ話者適応 HMM を用いた場合に高い認識精度が得られた．しかし，子音 “s” は 5.2 節で示したように改善が難しく，今回の実験の場合は認識精度が低下した．

5.3.2 母音のみ話者適応 HMM の利用

母音のみ話者適応 HMM を用い，その他の音素は不特定話者 HMM を用いて作成した混合 HMM の認識精度を求める．実験の結果を表 9 に示す．

表 10 母音のみ話者適応 HMM を利用して単語音声認識した場合の誤り率

	話者適応 HMM を用いた音素		
	a u	a i u e	a i u e o
男性話者	11.50% (904/7860)	11.81% (927/7860)	11.21% (881/7860)
女性話者	11.47% (901/7860)	10.85% (853/7860)	9.94% (781/7860)
平均	11.48% (1805/15720)	11.32% (1780/15720)	10.57% (1662/15720)

母音は子音に比べ学習データ内のデータ量が多い．しかし，母音全てを話者適応 HMM を用いた場合のみ認識精度が向上し，他の条件では認識精度が低下した．従って，一部の母音のみを用いて全体の認識精度の改善は難しいといえる．

話者適応に用いた 164 単語，82 単語の学習データでは母音全てにおいて十分な量を学習できた．しかし，更に少数の学習データを用いる場合は，子音のみを用いて混合 HMM を作成した方がよいと考えている．

5.4 特定話者音声認識との比較

特定話者は，一般的に学習データが少なくても比較的高い認識精度が得られる．そこで，話者適応に用いた 164 単語・82 単語を特定話者の学習データとして特定話者 HMM を作成し，混合 HMM の認識精度と比較を行う．

リスト 2 の学習データは含まれる音素の種類が少ないため，実験には 164 単語・82 単語の共にリスト 1 を用いる．特定話者の実験結果を表 11 に示す．参考として 2,620 単語で学習した特定話者の結果を同時に示す．

表 11 特定話者 HMM の単語音声認識の誤り率

学習データ量	2,620 単語	164 単語	82 単語
男性話者	4.62% (363/7860)	15.62% (1228/7860)	35.54% (2794/7860)
女性話者	4.56% (359/7860)	16.46% (1294/7860)	35.08% (2757/7860)
平均	4.59% (722/15720)	16.04% (2522/15720)	35.31% (5551/15720)

結果より，164 単語・82 単語の両方において，話者適応が有効であった．また，特定話者の認識精度が 164 単語から 82 単語の間で大きく低下しているため，認識する話者の音声が少ないほど話者適応が有効である．

5.5 話者適応の手法の検討

本研究における話者適応は，最も単純な手法である連結学習を用いた．話者適応には他にも MLLR，MAP 推定を始め多くの手法が存在する．しかし，学習データ内の音素数が多いほど認識精度が向上し，音素数が少ないほど認識精度が低下する傾向は変わらないと考えている．よって本研究で用いた混合 HMM と，偏りを持たせた学習データは，話者適応の手法によらず有効であると考えている．

6. おわりに

本研究では学習データに含まれる音素数に注目し，音素数が上位の音素は話者適応 HMM を用い，下位の音素は不特定話者 HMM を用いる混合 HMM を作成し，認識精度の調査を行った．また，話者適応の認識精度がより高くなるように，数の多い音素をより多く，数の少ない音素をより少なくなるように偏りを持たせた学習データ (リスト 2) を作成した．実験の結果，不特定話者 HMM と比較して，認識精度が 164 単語・リスト 2 の混合 HMM で 2.65%，82 単語・リスト 2 の混合 HMM で 1.33%改善し，混合 HMM と偏りを持たせた学習データ (リスト 2) の有効性を示した．

今後の課題として，より単語数が少ない学習データでの認識精度の調査が挙げられる．また，MAP 推定，MLLR を含め，他の話者適応に本研究の手法を適応して，認識精度の向上を目指す．

文　献

[1] 堀田, 村上, 池原, アクセントを用いた同音異義語の不特定話者音声認識, 電子情報通信学会技術研究報告, SP2005-195, pp. 65-70, (2006).

[2] 松浦, 村上, 池原, 話者選択型音声認識の可能性について, 日本音響学会 2007 年秋期研究発表会, 3-Q-9, pp. 134, (2007).

[3] HTK *Ver*3.2 *reference manual*, Cambridge University, (2002).

[4] C. Leggetter and P. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, Computer Speech and Language, Vol. 9, pp. 171-185, (1995).

[5] G. Zavaliagkos, R. Schwartz and John McDonough, Maximum a posteriori adaptation for large-scale HMM recognizers, Proc. ICASSP-96, pp. 725-728, Detroit, (1995).