062017　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

計算機工学講座C　松浦 祥悟

1　はじめに

話者適応は認識する話者の音声を学習データとして利用し認識精度を向上させる手法である．話者適応にはすでに様々な学習方法が提案されているが [1][2]，話者適応に用いる学習データが少ない場合，認識精度が向上するとは限らない．また，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あまり考察されていない．

そこで本研究は，学習データ内の各音素の数に着目し，各音素の数による認識精度の変化を調査すると共に，認識精度を低下させずに話者適応を行う手法として「混合 HMM」を用いる手法を提案し，認識実験により評価する．また，混合 HMM の実験結果から，学習データの各音素数に偏りを持つ学習データを提案し，認識精度の向上を試みる [3][4].

2　混合 HMM

混合 HMM は，話者適応における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 を組み合わせて作成する．話者適応において，学習データ内の音素数が少ない音素ほど，話者適応 HMM の信頼性が低いと考えられる．よって，学習データ内の音素数が少ない音素に不特定話者 HMM を用い，音素数が多い音素に話者適応 HMM を用いる．

具体的には，音素数の基準を n 個とした場合，学習データ内の音素数が n 個未満の音素に不特定話者 HMM を用い，n 個以上の音素に話者適応 HMM を用いる．このようにして作成した混合 HMM を「n 個未満混合 HMM」と呼ぶ.

3　評価実験

本研究では，不特定話者 HMM と話者適応 HMM と混合 HMM を用いて単語音声認識を行い，認識精度を比較する．

3.1　音素 HMM の作成

不特定話者の音素 HMM の作成から話者適応の音素 HMM の作成までの手順を以下に示す.

step1. 不特定話者の学習データを用いて Baum-Welch 学習を行い，「不特定話者 HMM」を作成する．

step2. 「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．

3.2　混合 HMM の作成条件

本研究の実験は，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ため，不特定話者 HMM と話者適応 HMM を使い分ける音素数の基準を 10 個，20 個，30 個とする．

3.3　評価データと学習データ

ATR 単語発話データベース Aset(1 話者につき 5,240 単語) の男女各 10 名の音声データを使用する．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．不特定話者に用いる学習データは，男女別に，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．話者適応に用いる学習データは，認識する話者の奇数番号から，各音素の出現割合が評価データと同程度になるように選択した 164 単語と 82 単語の 2 種類を用いる．

3.4　実験条件

評価実験は，Aset の男女各 10 名のうち男性話者 3 名，女性話者 3 名で行う．実験には単語音声認識ツールの HTK[5] を使用する．作成する HMM は，連続型 HMM とする．特徴パラメータに MFCC を，共分散行列に Diagonal-covariance を使用する．その他のパラメータは HTK のデフォルトのパラメータを使用する．

4　実験結果

4.1　不特定話者

表 1 に不特定話者 HMM を用いた単語音声認識の実験結果を示す．表中の括弧内の分母は認識話者の評価単語数，分子は誤認識した単語数を示す.

表 1　不特定話者の誤り率

	不特定話者
6 話者 平均	11.17% (1756/15720)

4.2　話者適応

表 2 に話者適応 HMM を用いた単語音声認識の実験結果を示す.

表 2　話者適応の誤り率

	164 単語 話者適応	82 単語 話者適応
6 話者 平均	13.48% (2119/15720)	33.08% (5201/15720)

実験より以下の結果を得た.

(1)　164 単語の話者適応と 82 単語の話者適応の共に，不特定話者より認識精度が低い.

(2)　82 単語の話者適応は，不特定話者と比較して認識精度が大きく低下している.

4.3　混合 HMM

表 3 に，混合 HMM を用いた単語音声認識の実験結果を示す.

表 3　混合 HMM の誤り率

	164 単語 学習データ	82 単語 学習データ
10 個未満 混合 HMM	9.99% (1570/15720)	14.13% (2221/15720)
20 個未満 混合 HMM	8.84% (1390/15720)	11.92% (1874/15720)
30 個未満 混合 HMM	8.84% (1390/15720)	10.74% (1689/15720)

実験より以下の結果を得た.

(1)　164 単語の学習データでは，全ての混合 HMM において不特定話者と話者適応よりも認識精度が

高い．

(2) 82 単語の学習データでは，30 個未満混合 HMM において不特定話者と話者適応よりも認識精度が高い．

(3) 164 単語の学習データと 82 単語の学習データの共に，30 個未満混合 HMM が最も認識精度が高い．

4.4 考察

82 単語の話者適応の認識精度が大きく低下していることと，30 個未満混合 HMM が最も認識精度が高く，20 個未満，30 個未満となるにつれて認識精度が低下していることから，話者適応において，音素数が少ない音素を多く含む学習データほど，認識精度が低下するといえる．

5 音素数に偏りを持つ学習データ

4 章の実験結果より，あらかじめ音素数の少ない音素を削除し，音素数の多い音素を増やした学習データを作成することで，更に認識精度が向上すると考えられる．本章では，このように音素数に偏りを持つ学習データの作成を行い，認識精度の評価を行う．

5.1 音素数に偏りを持つ学習データの作成

偏りを持つ学習データは，評価データ内の音素の出現率が 2% 未満の子音を含む単語を，2% 以上の子音を含む単語に選択し直して作成する．3 章と同様に 164 単語の学習データ，82 単語の学習データを作成する．

図 1 に評価データと 3 章で利用した 164 単語の学習データと音素数に偏りを持つ 164 単語の学習データの音素の出現率を降順にソートした分布を示す．なお 82 単語学習データもほぼ同様の分布となる．

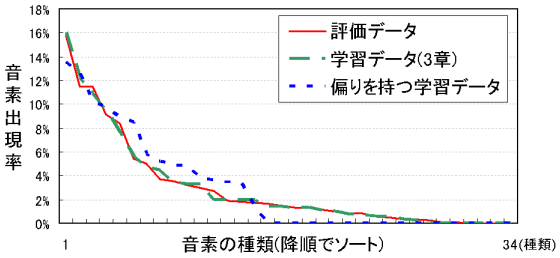

図 1 164 単語の学習データ内の音素の出現率

5.2 実験結果

偏りを持つ学習データを用いて作成した話者適応 HMM の単語音声認識の実験結果を表 4 に，混合 HMM の単語音声認識の実験結果を表 5 に示す．

表 2 と表 4, 表 3 と表 5 の比較より，偏りを持つ学習データを用いることで，話者適応と全ての混合 HMM において認識精度が向上することが確認できた．

本章で作成した学習データは，混合 HMM の比較を行うために 30 個未満の音素を含んでいるが，より詳細な条件を用いて学習データを作成することで，更に認識精度を向上させることができると考えている．

6 考察・母音と子音による認識精度の違い

本研究で作成した混合 HMM は子音と母音を区別せずに作成している．本章では，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 を用いた混合 HMM を作成し，認識精度の違いを調査する．学習データは 5 章で作成した 82 単語を使用する．表 6 にそれぞれの認識結果を示す．

表 4 偏りを持つ学習データを用いた話者適応の誤り率

	164 単語 話者適応	82 単語 話者適応
6 話者 平均	9.14% (1437/15720)	13.47% (2117/15720)

表 5 偏りを持つ学習データを用いた混合 HMM の誤り率

	164 単語 学習データ	82 単語 学習データ
10 個未満 混合 HMM	9.14% (11437/15720)	11.16% (1754/15720)
20 個未満 混合 HMM	8.52% (1340/15720)	10.25% (1612/15720)
30 個未満 混合 HMM	8.52% (1340/15720)	9.84% (1547/15720)

表 6 82 単語適応の混合 HMM の誤り率

	話者適応 HMM を用いた音素		
	k	k s	k s r t
6 話者 平均	10.45% (1643/15720)	10.71% (1683/15720)	11.40% (1792/15720)
	a u	a i u e	a i u e o
6 話者 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 とした時のみ不特定話者より認識精度が向上した．これより，母音は子音より改善が困難であるといえる．よって学習データ量がより少量となった場合，子音のみを用いて混合 HMM を作成するなどの対策が必要と考えている．

7 おわりに

本研究では学習データに含まれる各音素の数に注目し，話者適応 HMM と不特定話者 HMM を組み合わせた混合 HMM を提案し，実験より混合 HMM の有効性を確認した．また，混合 HMM の実験結果を元に，音素数に偏りを持つ学習データを作成し，話者適応と混合 HMM において認識精度が向上することを確認した．

今後の課題として，単語数がより少ない学習データでの認識精度向上と，MLLR や MAP 推定などの話者適応を用いた場合の認識精度の調査が挙げられる．

参考文献

[1] 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).

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

[3] 松浦, 村上, 池原, 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, 電子情報通信学会技術研究報告, 音声・思考と言語・福祉情報, SP2007-182, pp.87-91, (2008).

[4] 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.227-228, (2007).

[5] S.Young, P.Woodland and G.Evermann, HTK Book, Cambridge University Engineering Department, (2002).