話者選択型音声認識の可能性について*

○松浦祥悟，村上仁一，池原悟 (鳥取大)

1　はじめに

現在，複数の話者の音声を 1 つの HMM に学習する不特定話者音声認識は，認識率を向上させることが困難な状況とされている [1]．そこで本研究は，従来の手法とは異なる,「複数の特定話者の HMM を選択的に用いる」話者選択型の音声認識 (以下，話者選択) の検討を試みる．

まず，話者選択と不特定話者の精度の比較を行う．次に，話者選択 HMM と不特定話者 HMM に話者適応を行った場合の，有効性の比較を行う．これらの 2 つの視点から，話者選択の可能性について検討する．また，認識結果より，話者適応の精度を向上させるための考察を行う．

2　話者選択型音声認識

話者選択型音声認識では，複数の特定話者 HMM を作成し，認識する話者に適切な HMM を選択する．次に，選択した HMM を用いて認識を行う．HMM の選択は，認識率の最も高いものを選択する．

10 話者 (A，B，…，J) のデータが存在し，話者を A として認識する時の流れを図 1 に示す．

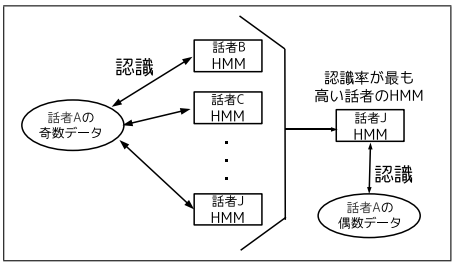

Fig. 1　話者選択型音声認識

まず，各話者の HMM を使用して，話者 A の認識を行う．その中で最も認識率の高かった話者 J の HMM を選択する．選択した話者 J の HMM を使用して，話者 A の認識を行い，認識率を求める．

3　話者適応

話者適応は，認識する話者のデータを利用して認識率を向上させる手法である．話者適応の手法には，MLLR 法 [2] や MAP 推定 [3] などがあるが，本研究では以下の 2 種類の最も基本的な話者適応を使用して，話者選択と不特定話者の比較を行う．

手法 1　話者選択 HMM 及び不特定話者 HMM に，認識する話者のデータを用いて連結学習を行う．

手法 2　話者選択 HMM 及び不特定話者 HMM の学習データに，認識する話者のデータを加えて，HMM を作成する．

4　評価実験

4.1　実験データ

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．学習データは，9 話者の奇数番号 (1 話者につき 2,620 単語) を使用し，評価データは，残りの 1 名の偶数番号 (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．

話者適応に用いるデータは，認識する話者の奇数番号のデータから，音素の出現割合が変わらないように選択した 328・82・21 単語を使用する．

4.2　実験条件

実験条件を表 1 にまとめる．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．特徴パラメータの混合分布数を表 2 にまとめる．

Table 1　音響パラメータ

基本周波数	16kHz
混合分布数	母音・撥音・無音 4mixture 子音 2mixture
特徴パラメータ	MFCC
共分散行列	Diagonal-covariance
HMM	半連続型 HMM

Table 2　混合分布数

特徴パラメータ 混合分布数 A	MFCC 1024 Δ MFCC 1024 対数パワー，Δ対数パワー 64
特徴パラメータ 混合分布数 B	MFCC 256 Δ MFCC 256 対数パワー，Δ対数パワー 32

4.3　実験結果

4.3.1　話者選択と不特定話者の比較

話者選択と不特定話者の実験結果を表 3 に示す．

Table 3　話者選択と不特定話者の実験結果

混合分布数	話者選択	不特定話者
A(1024 1024 64)	85.18% (13390/15720)	88.32% (13884/15720)
B(256 256 32)	83.60% (13142/15720)	86.93% (13665/15720)

実験の結果，不特定話者の認識率は，話者選択の認識率より高いことがわかった．また，混合分布数 A の話者選択の実験は，話者ごとに調べてみると，話者 mau が 89.85%，話者 mmy が 86.18%，話者 mnm が 79.43%と話者間でばらつきが大きい結果となった．

*The possibility of the speaker selection for speech recognition. by MATSUURA Syougo，MURAKAMI Jin'ichi and IKEHARA Satoru(Tottori University)

4.3.2 話者適応法・手法 1

不特定話者 HMM 及び話者選択 HMM に，認識する話者のデータを連結学習させた実験結果を表 4 に示す．

Table 4　話者適応法，手法 1 の実験結果

話者選択			
混合分布数	328 単語	82 単語	21 単語
A(1024 1024 64)	93.15% (14643/15720)	65.64% (10313/15720)	21.39% (3363/15720)
B(256 256 32)	92.75% (14581/15720)	85.67% (13466/15720)	60.02% (9435/15720)
不特定話者			
混合分布数	328 単語	82 単語	21 単語
A(1024 1024 64)	94.32% (14789/15720)	78.50% (12341/15720)	15.20% (2389/15720)
B(256 256 32)	92.10% (14478/15720)	88.66% (13937/15720)	57.25% (9025/15720)

話者選択の認識率は，328 単語と 82 単語において，不特定話者の認識率より低い．また，適応単語数が 21 単語の場合に，認識率が大きく減少している．

4.3.3 話者適応法・手法 2

不特定話者及び話者選択の学習データに，認識する話者のデータを加えて学習し，HMM を作成する実験結果を表 5 に示す．

Table 5　話者適応法，手法 2 の実験結果

話者選択			
混合分布数	328 単語	82 単語	21 単語
A(1024 1024 64)	91.28% (14349/15720)	88.55% (13920/15720)	85.99% (13518/15720)
B(256 256 32)	88.94% (13982/15720)	86.49% (13596/15720)	84.32% (13255/15720)
不特定話者			
混合分布数	328 単語	82 単語	21 単語
A(1024 1024 64)	89.31% (14040/15720)	88.74% (13950/15720)	88.60% (13928/15720)
B(256 256 32)	87.67% (13782/15720)	87.20% (13707/15720)	87.26% (13717/15720)

話者選択の認識率は，328 単語の話者適応において，不特定話者の認識率より高く，82 単語と 21 単語の話者適応において，不特定話者の認識率より低い結果となった．

5　考察

5.1　話者選択の可能性

話者選択と不特定話者を比較すると，不特定話者の認識率の方が良い結果となった．また，話者適応の結果を見ると，328 単語・82 単語・21 単語の全ての適応おいて，不特定話者の認識率が良い結果となった．

本研究の結果から，話者選択を用いる音声認識は困難であると考えている．

5.2　話者適応の手法について

手法 1 は，21 単語の適応において認識率が非常に低い．手法 2 は，21 単語の適応において手法 1 より認識率が高いが，328 単語の適応では認識率が低い．

よって話者適応の手法は，適応に用いるデータなどの条件によって，変更する必要がある．

5.3　話者適応による認識率の向上

話者適応の精度を向上させるために，本研究の誤り分析を行った．不特定話者音声認識の実験結果から誤認識の多かった音素を調べ，誤認識の多い音素 HMM を特定話者の音素 HMM と置き換えることで，認識率がどのように変化するかを調べた．誤認識の多い音素は，母音は “u”，“i”，子音は “k”，“r”，“m” であった．1 つの音素の置き換え実験と複数の音素の置き換え実験を行った．なお、この実験では連続型 HMM を使用している．

実験の結果を表 6 に示す．図の “iu” の表記は “i” と “u” の HMM を特定話者の HMM に置き換えて実験を行っている．置き換えを行う前の不特定話者の認識率は，88.83%であった．

Table 6　音素を置き換えた実験結果

	置き換えた音素			
	i	u	k	m
平均	87.23% (13712/15720)	87.93% (13823/15720)	89.61% (14087/15720)	88.17% (13860/15720)
	iu	aiueo	kr	gm
平均	86.65% (13621/15720)	90.05% (14155/15720)	90.49% (14220/15720)	88.59% (13927/15720)

“aiueo” 全ての置き換えと，“k”，“kr” の置き換えの場合のみ認識率が向上した．その他の実験では，置き換えた音素の認識率は向上したが，他の音素の誤認識が増加し，全体の認識率が下がった．全体の認識率を向上させるためには，誤認識が多い音素を改善するだけでは，効果がないと考えられる．今後，“k” と “kr” の置き換えや “aiueo” 全ての置き換えのように，全体の認識率を向上させることができる適応データを選択して，より精度の高い話者適応を行うことが課題となる．

6　終わりに

本研究は，話者選択の可能性について，不特定話者と比較を行い，更に話者適応を行った場合の有効性を調査した．両者において，不特定話者の方が良い結果となり，話者選択を用いた音声認識は難しいことがわかった．今後は， 不特定話者の誤り分析を元に適応に用いるデータを選択し，認識率の調査を行う．

参考文献

[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 (2006-03).

[2] 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.

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