TR-IT-0095

MAP-VFS 話者適応方式において 平滑化係数を制御する方法の提案と評価
Proposition and evaluation of parameter smoothing control on MAP-VFS speaker adaptation

門田 暁人 Monden Akito
外村 政啓 Tonomura Masahiro

1995.2.27

音声認識システムに対して少量の学習資料によって話者適応を行う場合、安定した適応結果を得るためには、情報不足を補うことや、音声サンプルの統計的な偏りによる推定誤差の問題を解決することが不可欠である。移動ベクトル場平滑化話者適応法 (VFS)[1] では、モデルパラメータの移動ベクトルの平滑化によって、少量の適応用サンプルに起因する未学習パラメータの補間と適応済みパラメータの補正を同時に実現している。VFS の問題点として、適応データ量がある程度増えてきた場合に、平滑化を行わない場合よりも認識率が低下することがあげられる。本稿では、音響モデルのパラメータごとの適応データ量に応じて平滑化係数を制御する方式を提案し、実験結果によりその有効性を示す。なお、本稿では最大事後確率推定法 (MAP)[3] と VFS を統合した MAP-VFS 話者適応法 [2] を対象とした。

目次

第 1 章

はじめに

不特定多数の話者に対する音声認識の代表的な方法に、不特定話者用音響モデルを用いる方法と、音響モデルを話者適応する方法がある。対話などを通して話者の音声サンプルが得られる場合は、後者の方が有利である。

音声認識システムに対して少量の学習資料によって話者適応を行う場合、安定した適応結果を得るためには、情報不足を補うことや、音声サンプルの統計的な偏りによる推定誤差の問題を解決することが不可欠である。

移動ベクトル場平滑化話者適応法 (VFS)[1] では、“標準話者の音響パラメータ空間から適応話者の音響パラメータ空間への移動ベクトル写像が、比較的なめらかな移動ベクトル場に沿っている”という仮定に基づいて、モデルパラメータの移動ベクトルの平滑化によって少量の適応用サンプルに起因する未学習パラメータの補間と適応済みパラメータの補正を同時に実現している。平滑化においては、各ガウス分布の平均値の移動ベクトルに対しての適応が、他のパラメータの適応に比べて効果が大きいことが知られている。

VFS の問題点として、適応データ量がある程度増えてきた場合に、平滑化を行わない場合よりも認識率が低下することがあげられる。これは、VFS における前述の仮定により、統計的に偏りのない音声サンプルをも平滑化してしまうことによると推測される。本研究では、適応データ量によらずつねに安定した話者適応を行うことを目的とし、適応データ量が増えるにしたがって、音響モデルの各パラメータの推定値の統計的な偏りが解消されていくという仮定に基づいて、パラメータごとの適応データ量に応じて平滑化係数を制御する。

本稿では、最大事後確率推定法 (MAP)[3] と VFS を統合した MAP-VFS 話者適応法 [2] を対象とした。

第 2 章

MAP-VFS アルゴリズム

MAP-VFS アルゴリズムは MAP による移動ベクトルの推定、およびその移動ベクトルを用いた移動ベクトルの補間、移動ベクトルの平滑化、の 3 ステップより構成される MAP と VFS の統合されたアルゴリズムである。

一般に連続分布型 HMM によるモデルに対して少量の適応データにより話者適応を行う場合、ガウス分布の平均値の適応は他のパラメータの適応に比べて効果が大きい [3]。ここでは各ガウス分布の平均値のみの適応を行い、分散値、状態遷移確率および混合ガウス分布の重み係数の適応は行わない。

1 MAP によるパラメータの推定

MAP によるパラメータの推定は、以下の式で表される。

$$v_i^{MAP} = \frac{n_i}{n_i + \tau} v_i^{ML} + \frac{\tau}{n_i + \tau} v_i$$

i	HMnet からの出力分布の番号
v_i	初期モデルの出力分布 i のパラメータ
v_i^{ML}	最尤推定による出力分布 i のパラメータ推定値
v_i^{MAP}	MAP による出力分布 i のパラメータ推定値
n_i	出力分布 i の適応データ量
τ	事前知識と事後知識の相対的なバランスを表す係数

この式より、MAP によるパラメータの推定値 v_i^{MAP} は、初期もモデルのパラメータ v_i と最尤推定による推定値 v_i^{ML} の間を線形補間したものであることがわかる。

2 VFS

VFS は移動ベクトルの補間と平滑化を行う。補間は、学習データが存在しない場合の平滑化であるので、ここでは平滑化についてのみ説明する。

2.1 移動ベクトルの平滑化

HMnet からの各状態からの出力分布を i とし、標準話者モデルの出力分布 i の平均値を v_i とする。平滑化を行わない場合の標準話者の音響パラメータ空間から入力話者の音響パラメー

タ空間への平均値移動ベクトルを Δv_i とし、平滑化を行なった場合の平均値移動ベクトルを $\Delta v_i'$ とすると、$\Delta v_i'$ は以下の式により導かれる。

$$\Delta v_i' = \frac{\sum_{j \in N(i)} \lambda_{i,j} \Delta v_k}{\sum_{j \in N(i)} \lambda_{i,j}}$$

$N(i)$ は出力分布 i の K 近傍にある出力分布の番号であり、$\lambda_{i,j}$ は出力分布 i と出力分布 j の平均値の距離によって決まる重み係数である。本研究では $K = 6$ を用いた。上式で求められた平均値移動ベクトルを元の平均値ベクトルに加算することによって、各出力分布 i に対する平滑化後の出力分布の最終的な平均値ベクトル v' を算出する。なお、状態間の距離は以下の式を用いた。

$$d_{i,j} = \sum_{l=1}^{L} (v_{i,l} - v_{j,l})^2$$

$d_{i,j}$ は出力分布 i と出力分布 j の平均値の距離、L は音響パラメータの次数、$v_{i,l}$ は出力分布 i の出力分布の第 l 次パラメータの平均値である。次に、重み係数 $\lambda_{i,j}$ の算出方法について述べる。

2.2　重み係数の算出

重み係数の $\lambda_{i,j}$ の決定方法の代表的なものとして、ファジー級関数を利用する方法と、ガウス窓関数を利用する方法の2通りがあるが、HMnet における VFS の適応においては、後者の方が優れているという結果が報告されており [4]、本研究ではガウス窓関数を用いた。

この方法では、出力分布 i の K 近傍内にある各出力分布 $j \in N(i)$ の平均値パラメータ v_j に対して、以下の式で算出されるガウス分布型の窓関数を用いた重み付き加算を行う。

$$\lambda_{i,j} = e^{-d_{i,j}/f}$$

f は平滑化係数である。平滑化係数が大きいほど、近傍の出力分布のパラメータの影響が大きくなり、平滑化の度合が大きくなる。

第 3 章

提案する方法

本研究では、学習データ量が増えるにしたがって音響モデルの各パラメータの推定値の統計的な偏りが解消されていくと考えた。パラメータの推定値の偏りが小さい場合には推定誤差が小さいので、平滑化を行うべきではないと考えた。そこで、パラメータごとの学習データ数が増えるにしたがって平滑化係数を減少させるような式（以下、制御式）を考えた。さまざまな制御式が考えられるが、どのような式が最適かはわからないため、単純な線形補間の式を用いて 7 人の話者に対して実験を行った。その式を以下に示す。

$$f_i = f\frac{\alpha}{n_i + \alpha}$$

ここで、f は全てのパラメータに対して共通に与えられる平滑化係数の初期値であり、n_i は i 番目のガウス分布の適応データ量 (フレーム数) を表している。この式により、各パラメータに対する平滑化の強さは適応データ量が増加するに従って弱められていき $n_i \to \infty$ では平滑化を行わない場合と同様の状態に収束することがわかる。また、この時の収束の速さは係数 α によって決められる。本稿では α は実験的に求めた値を使用した。

上の式では、適応フレーム数を元に個々のパラメータに対して平滑化の制御を行っているが、適応文節数を元に全体の平滑化係数を制御する方法も考えられる。しかし、後者の方法では、適応データの音素に偏りがある場合に、適応データが増えるに従って、各パラメータの推定値の偏りが必ずしも解消されていくとは限らない。また、音響モデルの構造 (状態数、混合数など) が変わった場合に、最適な制御の式が変わってしまう可能性が大きい。以上の理由により、適応データの内容によらずつねに安定した話者適応をするためには、適応フレーム数を元に個々のパラメータに対して平滑化の制御を行う方法の方が、優れていると判断した。

1　認識実験

実験には 200 状態の隠れマルコフ網 (HMnet) を使用した。HMnet はモデルパラメータを効果的に共用したコンテキスト依存型の音素 HMM モデルである。初期モデルには不特定話者モデル (285 人分の特定話者モデルから合成することにより作成) を用い、HMnet の各状態のガウス分布の混合数は 5 混合とした。平滑化係数の初期値は $f = 30$ とし、収束の速さを決める係数 α は実験的に求め $\alpha = 0.3$ とした。分析条件、使用パラメータ、適応／認識データを表 3.1 に示す。

$f = 30$ を用いた理由を説明する。平滑化係数を制御しない場合の音素認識率を調べる予備実験を一人の話者に対して行った場合に、適応文節数 N が小さい時 (N=3,5,7) の音素認識率を調べると、平滑化係数をだんだん大きくしていくと認識率は次第に高くなっていったが、あるところで頭うちになった。この時の値が $f = 30$ であった。本研究では、学習データ数が少ない場合に平滑化の効果を最大限に利用するために、$f = 30$ を用いた。

表 3.1: 実験条件

分析条件	
サンプリング周波数 12KHz 20ms ハミング窓、フレーム周期 5ms	
使用パラメータ	
16 次 LPC ケプストラム + 16 次 Δ ケプストラム + log パワー + Δlog パワー	
学習データ	
男性 146 名 + 女性 139 名 (各話者 50 文節)	
適応／認識データ	
話者	男性 4 名 (MAU,MMY,MSH,MTM) 女性 3 名 (FAF,FMS,FYM)
適応データ	256 文節 (SB1 タスク) の文節から順に 取り出した N 個の文節
認識データ	279 文節 (SB3 タスク)

2　実験結果

表 3.2 に MAP、MAP-VFS、および平滑化係数の制御を行った MAP-VFS による話者適応結果を示す。また、図 3.1 に結果のグラフを示す。まず、MAP と MAP-VFS を比較すると適応文節数が 3,5,7 文節のように少ない場合にはほとんどの話者で MAP よりも MAP-VFS の方が認識誤り率は低く VFS が有効であることがわかる。しかし、適応文節数が 20 文節まで増えると MAP と MAP-VFS の効果が 4 名の話者で逆転しており、47 文節以上では全ての話者において MAP-VFS より MAP の方が認識誤り率が低くなっている。この結果より、平滑化は適応データ量が少ない場合には有効であるが適応データ量が多い場合には逆に話者適応性能を劣化させていることが確認できる。次に、上記の MAP および MAP-VFS による話者適応結果と本稿で提案している適応データ量の応じた平滑化係数の制御を行った MAP-VFS による話者適応結果と比較する。適応文節数が 20 文節以下の場合には多くの場合で平滑化係数の制御を行った MAP-VFS が通常の MAP-VFS と同程度、あるいはより低い認識誤り率を示している。一方、47 文節以上では平滑化係数を制御することによって、通常の MAP-VFS より低い認識誤り率が得られ、また MAP に近い認識誤り率を示している。以上の結果より MAP-VFS において平滑化係数を適応データ量に応じて制御することが効果的な平滑化を行う上で有効であることがわかる。

表 3.2: 話者適応結果 – 音素認識誤り率 (%)

上段： MAP、中段： MAP-VFS、下段： MAP-VFS(平滑化係数制御)

話者名	適応文節数							
	適応前	3	5	7	20	47	97	256
MAU	19.2	19.2	17.0	16.1	16.4	11.9	9.7	6.9
		18.1	16.9	16.0	15.7	14.3	12,2	11.0
		18.1	16.3	15.2	15.7	12.3	10.0	6.9
MMY	19.0	19.4	17.7	18.0	16.2	13.7	11.2	9.1
		18.7	17.7	17.5	16.9	15.4	13.4	13.1
		19.2	17.4	17.0	16.5	14.2	11.4	9.3
MSH	24.1	24.2	22.1	21.4	18.3	14.2	10.8	9.9
		23.2	21.8	20.5	19.5	17.7	16.4	15.7
		23.6	20.9	20.3	17.9	14.1	11.5	10.3
MTM	17.5	17.5	16.7	14.9	13.0	9.3	5.7	4.9
		15.2	15.0	14.3	14.0	13.2	11.4	10.1
		15.3	14.7	13.9	12.7	10.2	6.4	5.2
FAF	21.5	20.3	19.8	17.7	17.2	13.0	9.4	7.7
		18.8	16.7	16.2	15.5	15.3	13.9	14.0
		18.7	17.0	17.0	15.5	12.1	9.7	7.7
FMS	21.1	21.4	20.5	20.4	19.4	14.7	11.5	9.0
		20.7	18.2	18.4	17.5	16.0	15.7	15.5
		20.5	18.4	18.3	18.0	14.3	12.0	9.4
FYM	32.6	30.5	29.2	27.2	22.3	19.0	15.4	13.8
		28.6	28.8	27.3	24.7	23.0	21.1	20.2
		28.6	27.4	25.9	21.0	19.2	16.0	14.0
平均	22.0	21.8	20.4	19.4	17.5	13.7	10.5	8.8
		20.5	19.3	18.6	17.7	16.4	15.1	14.5
		20.6	18.9	18.2	16.8	13.8	11.0	9.0

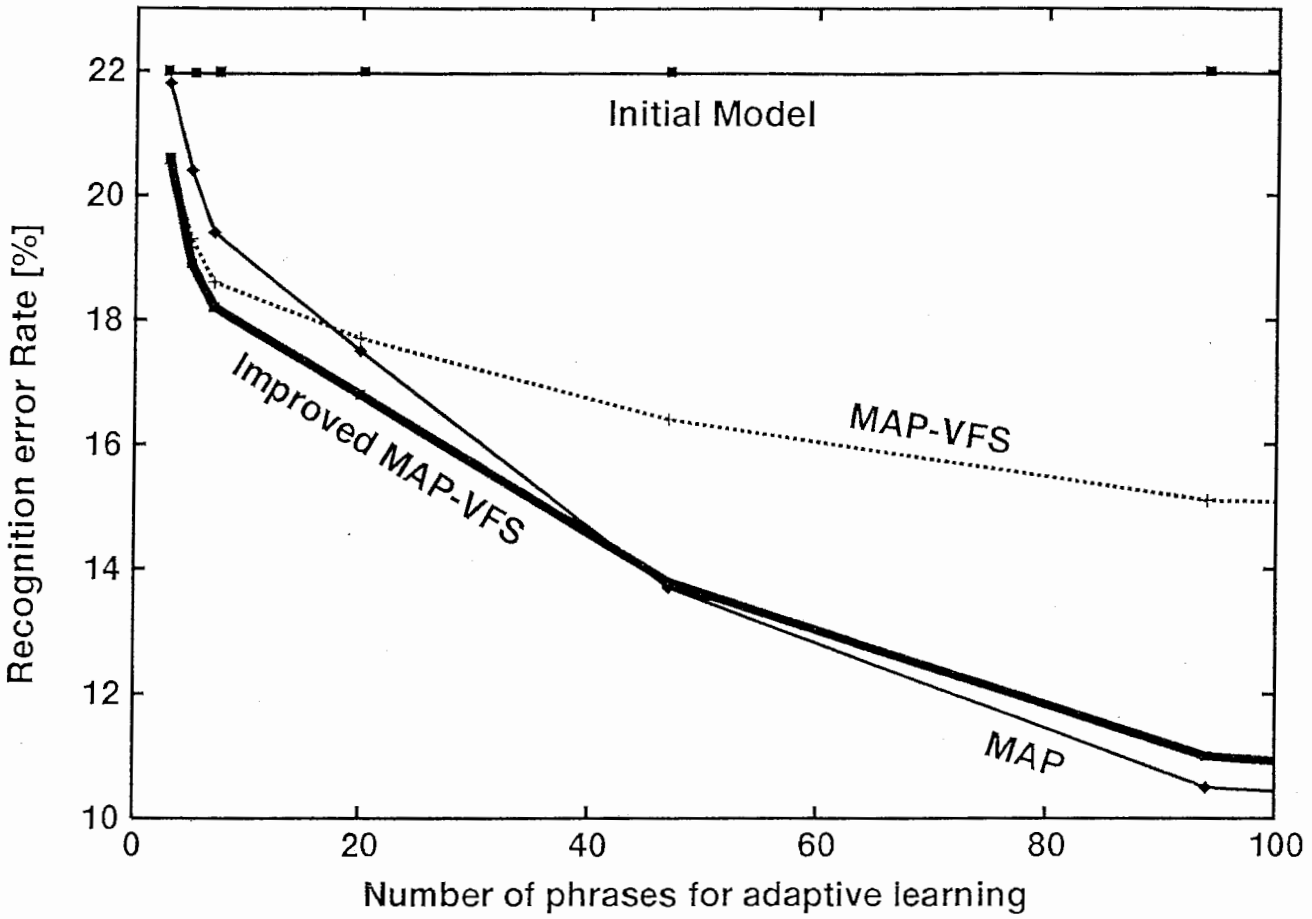


図 3.1: 各話者適応法における適応文節ズ数対音素認識率 (%)

3　最適な制御式の模索

実験結果より、適応文節数が少ない場合には平滑化を行った場合と同等の認識率が得られ、適応文節数が多い場合にも平滑化を行わない場合と同等の認識率が得られることがわかった。また、適応文節数によっては、平滑化を行う場合と行わない場合のどちらの場合よりも認識率が高くなることがわかった。

この実験では、単純な線形補間の式を用いたが、線形補間による平滑化の制御の方法が最適であるという保証はない。そこで、最適な制御式を見つけるために、以下の 5 つの式について実験することにした。

(1)　$f_i = f\frac{\alpha}{n_i+\alpha}$

(2)　$f_i = f\left(1-\frac{n_i}{\alpha}\right)$　　$f_i = 0\ (n_i \geq \alpha)$

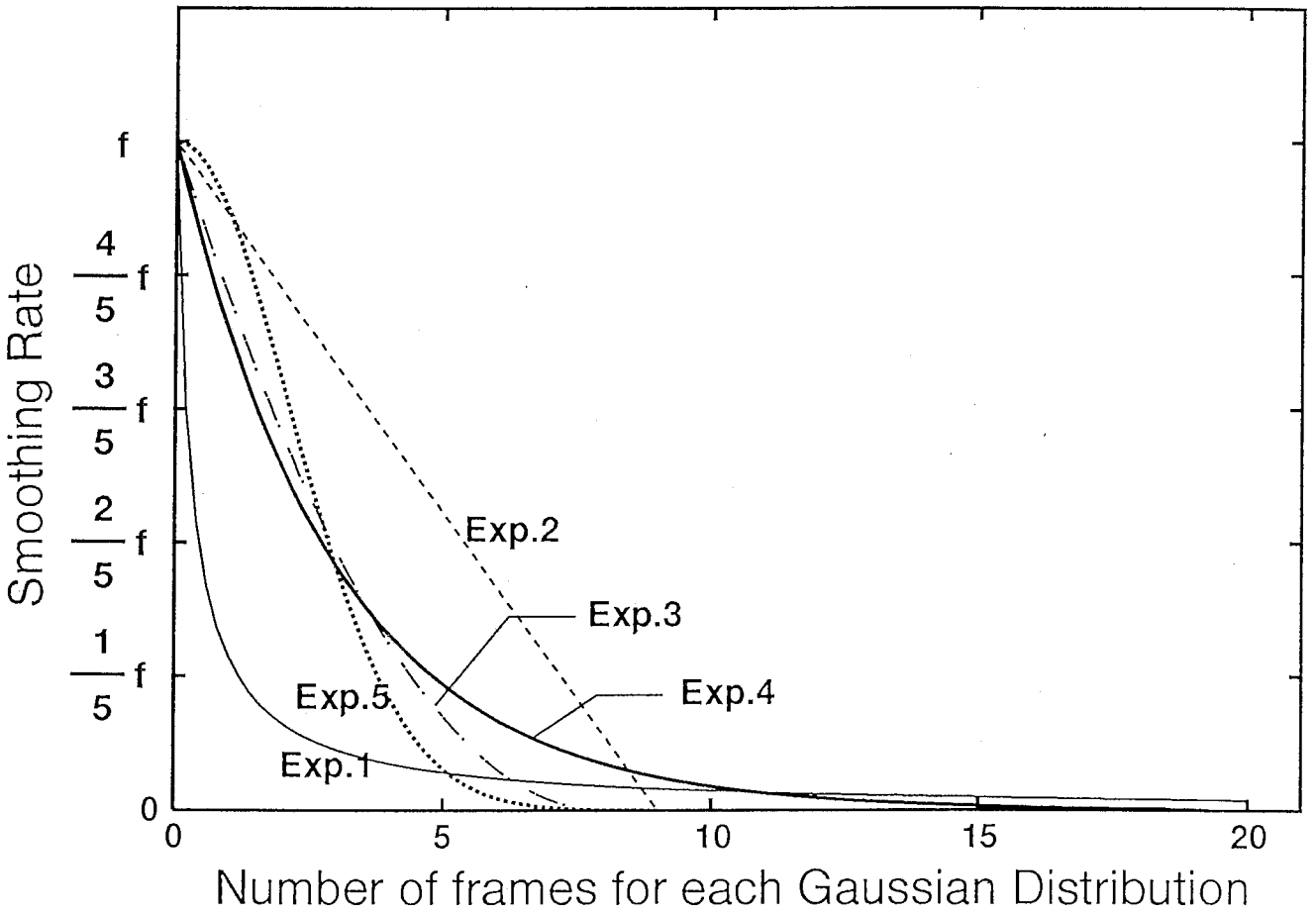
(3)　$f_i = f\left(1-\frac{n_i}{\alpha}\right)^2$　　$f_i = 0\ (n_i \geq \alpha)$

(4)　$f_i = fe^{-\frac{n_i}{\alpha}}$

(5)　$f_i = fe^{-(\frac{n_i}{\alpha})^2}$

それぞれの制御式において、αが大きいほどf_iが0に収束速さが大きくなるが、式ごとに、収束の仕方が違っている。各式において一人の話者(FMS)について最適なαを実験的に求めた。この場合の式をグラフに表したものを図3.2に示す。

図 3.2: 制御式の比較



また、音素誤り率を求めた結果を以下に示す。

・式1

係数	適応文節数							
	3	5	7	20	47	97	256	平均
0.3	20.5	18.4	18.2	18.0	14.3	12.0	9.4	15.8
0.5	20.2	18.4	18.1	17.6	14.7	12.3	9.5	15.8

・式2

係数	適応文節数							
	3	5	7	20	47	97	256	平均
8	20.5	18.5	18.2	18.6	14.6	12.0	9.6	16.0
10	19.7	18.7	18.2	17.3	14.6	12.5	9.8	15.8

式1と大差がなかった。

・式 3

係数	適応文節数							
	3	5	7	20	47	97	256	平均
8	20.5	18.4	18.7	18.4	14.5	11.9	9.5	16.0
10	20.3	18.4	18.5	18.9	14.4	12.2	9.5	16.0

式 1 と比べて $N = 7, 20, 47$ の時の音素誤り率が高い。

・式 4

係数	適応文節数							
	3	5	7	20	47	97	256	平均
3	20.3	18.1	17.9	18.1	14.3	12.0	9.3	15.7
4	20.2	18.1	17.6	17.3	14.8	12.7	9.7	15.8

式 1 と比べて $N = 5, 7$ の時の音素誤り率が低い。

・式 5

係数	適応文節数							
	3	5	7	20	47	97	256	平均
3	20.6	18.4	18.2	18.5	14.7	11.6	9.5	15.9
5	20.3	18.4	18.0	18.1	14.4	12.4	9.5	15.9

式 1 と比べて $N = 20, 47$ の時の音素誤り率が高い。

以上の結果により、話者 FMS においては、平均の音素誤り率は 15.7 ～ 16.0 であり、式による差はほとんどないといえる。

式 1 よりも式 4 を用いた場合の方が音素誤り率が若干低くなっているが、この差が他の話者でも現れるかどうかを確かめるために、7 人の話者について式 4 を用いた場合の実験を行った。平滑化係数の初期値は $f = 30$ とし、収束の速さを決める係数 α は FMS における実験で求めた $\alpha = 3$ を用いた。結果を表 3.3 に示す。また、式 1 と 4 を用いた場合の 7 話者の平均の学習フレーズ数対音素認識誤り率のグラフを図 3.3 に示す。

表 3.3 と図 3.3 より、式 1 を用いた場合より式 4 を用いた場合に平均の音素誤り率が低くなるのは、MTM と FMS の 2 人の話者だけであった。7 話者の平均でみると、式 1 を用いた場合には、$N = 7, 20, 47$ の時の音素誤り率は式 4 の方が少し高いが、$N = 94$ の時の音素誤り率は式 1 の方が少し高い。全般的には大差がないといえる。現状では、式 1 を用いるのが無難と思われる。

なお、最適な α の値は、話者によって若干のばらつきがあるが、誤差の範囲内であると考えられる。

表 3.3: 式 4 を用いた場合の話者適応結果 – 音素認識誤り率 (%)

話者名	適応文節数							
	適応前	3	5	7	20	47	97	256
MAU	19.2	18.3	16.6	16.0	15.7	12.8	10.2	6.9
MMY	19.0	18.8	17.4	17.5	16.5	14.6	11.6	9.2
MSH	24.1	23.3	21.4	21.0	18.6	14.8	12.1	10.1
MTM	17.5	15.0	14.8	13.9	12.6	10.3	6.5	5.2
FAF	21.5	18.9	16.6	16.4	15.8	12.7	10.2	8.0
FMS	21.1	20.3	18.1	17.9	18.1	14.3	12.0	9.3
FYM	32.6	27.8	27.0	25.8	22.7	19.9	16.5	13.9
平均	22.0	20.3	18.8	18.4	17.1	14.2	10.7	8.9

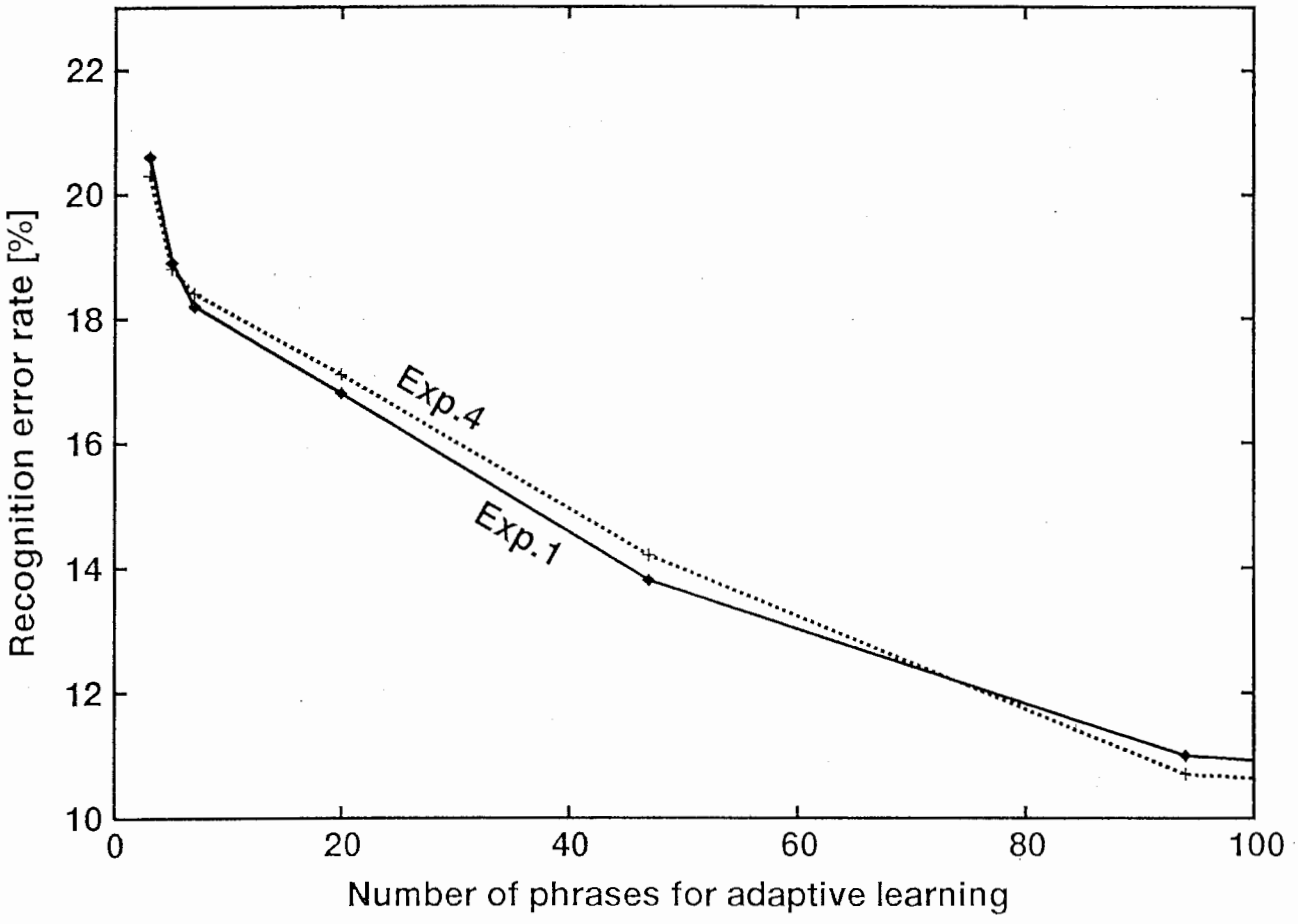


図 3.3: 式 1 と式 4 を用いた場合の音素認識誤り率の比較 (7 話者の平均)

第 4 章

まとめ

本稿では学習データ量に依存せずに常に最適な話者適応性能を得ることを目的として、MAP-VFS 法に適応データ量に応じた平滑化係数の制御を組み込むことを提案し、その有効性を確認した。今後の展望としては、分散などの他のパラメータの適応が考えられる。また今回は、文節ごとに区切って発声されたバランス文を適応データとして用いたが、自由発話データを用いた場合にも提案した方法が有効であるかどうかを確かめる必要がある。

謝辞

本実習にあたり、親切丁寧な御指導と激励をいただきました、外村政啓研究員、並びにATR音声翻訳通信研究所の皆様に感謝致します。また、本実習の機会を与えて下さった、ATR音声翻訳通信研究所の山崎泰弘社長に感謝致します。

参考文献

[1] 大倉計美, 杉山雅英 , 嵯峨山茂樹 “混合連続分布 HMM を用いた移動ベクトル場平滑化話者適応方式” , 信学技報 ,SP92-16 (1992-6)

[2] 外村政啓 , 小坂哲夫 , 松永昭一 “最大事後確率推定法を用いた移動ベクトル場平滑化話者適応方式” , 音響講論 2-8-20,pp.77-78 (1994-10)

[3] C.-H.Lee,C.-H.Lin,B.-H.Juang “A Study on Speaker Adaptation of the Parameters of Continuous Desity Hidden Marcov Models” ,IEEE Trans. on Signal Processing,Vol.39,No.4 (1991)

[4] 鷹見淳一 , 嵯峨山茂樹 “隠れマルコフ網で表現した音素コンテキスト依存モデルのための話者適応” , 信学論 J77-D-II,12,pp.2325-2333 (1994-12)

[5] 外村政啓 , 小坂哲夫 , 松永昭一 , 門田暁人 “MAP-VFS 話者適応における平滑化係数制御の効果” , 音響講論 (1995-3)