TR-IT-0187

ビタビ学習法による音響モデル作成時間の高速化
Rapid acoustic model training based on the Viterbi algorithm

宮本 宗易
Miyamoto Muneyasu

深田 俊明
Fukada Toshiaki

1996.9.20

現在、音響モデルの学習にはForward-Backward学習を用いていたが、今回はビタビ学習によりモデル学習の時間短縮を図るとともに、それぞれの音響モデルの性能評価を示す。従来の学習法に比べてビタビ学習は学習時間が高速化でき、認識性能を比較してみても連結学習とほぼ同程度の性能が得られた。認識の性能評価は、音素タイプライタ認識と単語認識の2種類について評価を行ない、混合数や学習単位が認識性能に対する影響の評価結果を示す。

目次

第 1 章

はじめに

音声認識システム全体の性能を向上させる上で、認識性能の良い音響モデルの開発は必要不可欠な課題である。現在、ATR では不特定話者自然発話音声認識システムにおける高性能な音響モデルの作成のために、多数話者の大規模自然発話音声データベースの収録を行なっている。一般に、学習データが増えると高精度な音響モデルが作成できるが、その一方で学習に要する時間が増大するという問題が生じる。

現在 ATR では、音響モデルの学習を行なう際に、一般に広く用いられている Forward-Backward(以下、F-B と省略する) 学習を用いている。将来、音声データが更に増大していくことを考えると、この F-B 学習による音響モデルの学習は、膨大な時間を要することから、学習時間の短い学習方法の必要性が高まっている。例えば、一定の時間で学習を行なうということを考えた場合、(1) 学習速度が早い方法を用いて多くの学習データを利用する方法、(2) しっかりした学習方法で少ない学習データを利用する方法の 2 通りが考えられる。

音響モデルの学習方法としては、F-B 学習の他に Viterbi 学習 [1] があり、一般に学習時間は F-B 学習よりも短い。また、Viterbi 学習により音響モデルを作成している研究機関も少なくない [2] [3] [4]。そこで本報告では、F-B 学習及び Viterbi 学習の学習時間や性能比較について検討を行なった結果について報告する。

以下、第 2 章では、F-B 学習と Viterbi 学習ついて説明を行ない、2 つの学習方法の違いを述べる。次に、第 3 章では、これら 2 つの学習方法により学習した音響モデルを用いて、自然発話音声認識実験を行なった結果について述べる。また、第 4 章では、Viterbi 学習を用いて、学習の繰り返し回数、初期値、トポロジー、混合数、状態数、学習データ量を変化させた場合の認識性能への影響について述べる。

第 2 章

原理

2.1 Forward - Backward 学習

前向き処理及び後向き処理を行ない、そしてt番目のフレームが存在し得ることのできるすべての状態iについて存在確率$\gamma_t(i)$を求め、最後にモデルパラメータの更新を行なう。この繰り返しがF-Bの基本的な学習方法である。

まず、前向き処理では、t番目のフレームの特徴ベクトルが状態iに存在する確率$\alpha_t(i)$を定義し、以下の式により求める。

1) 初期化

$$\alpha_1(i) = \begin{cases} 1 & \text{i=1} \\ 0 & \text{otherwise} \end{cases} \quad (1 < i < S) \tag{2.1}$$

2) 帰納

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(\mathbf{O}_{t+1}) \tag{2.2}$$

初期化では、第1フレーム目が1番目の状態に1の確率で存在することを表し、2番目以降の状態から始まらないことを仮定している。帰納での計算は、t番目のフレームが到達可能なN個の状態からt+1番目のフレームが状態jに到達する確率を求めている。同様に、後向き処理でもt番目のフレームの特徴ベクトルが状態iに存在する確率$\beta_t(i)$を定義し、以下の式より求める。

1) 初期化

$$\beta_T(i) = \begin{cases} 1 & \text{i=S} \\ 0 & \text{otherwise} \end{cases} \quad (1 < i < S) \tag{2.3}$$

2) 帰納

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\mathbf{O_{t+1}}) \beta_{t+1}(j) \tag{2.4}$$

前向き処理と同様、初期化では最終フレームが最終状態へ1の確率で到達することを意味し、最

終状態以外の状態で終了しないことを仮定している。帰納での計算は、t 番目のフレームが状態 i に存在する確率を、$t+1$ 番目のフレームが到達できるすべての状態 j のそれぞれについて、i から j への遷移確率 a_{ij}、状態 j の出力確率 $b_j(\mathbf{O}_{t+1})$、状態 j 以降の全フレームを考慮にいれた存在確率 $\beta_{t+1}(j)$ を乗算し、そしてそれらの総和を求めている。

これら2つの存在確率から、あるフレーム以外のフレームを考慮し、かつそのフレームが状態 i の j 番目の混合分布に存在する確率 $\gamma_t(i,j)$ を次式より求める。

$$\gamma_t(i,j) = \left[\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \right] \left[\frac{c_{ij}\mathcal{N}(\mathbf{o}_t, \mu_{ij}, \mathbf{U}_{ij})}{\sum_{m=1}^{M} c_{im}\mathcal{N}(\mathbf{o}_t, \mu_{im}, \mathbf{U}_{im})} \right] \tag{2.5}$$

この $\gamma_t(i,j)$ を用いて、以下の重み、平均、分散の再推定式からモデルパラメータの更新をする。

重み：
$$\bar{c}_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(i,j)}{\sum_{t=1}^{T}\sum_{j=1}^{M} \gamma_t(i,j)} \tag{2.6}$$

平均：
$$\bar{\mu}_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(i,j) \cdot \mathbf{o}_t}{\sum_{t=1}^{T} \gamma_t(i,j)} \tag{2.7}$$

分散：
$$\bar{\mathbf{U}}_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(i,j) \cdot (\mathbf{o}_t - \mu_{ij})(\mathbf{o}_t - \mu_{ij})^t}{\sum_{t=1}^{T} \gamma_t(i,j)} \tag{2.8}$$

2.2 Viterbi 学習

前向き処理は2.1と同様に行なうが、それ以降の処理が異なり、Viterbi学習法ではバックトレースを行なってViterbiアライメントを求め、最後にモデルパラメータの更新を行なう。F-B学習法と同様に、この計算の繰り返しにより学習を行なっている。

バックトレースでは、前向き処理により求められた$\alpha_t(i)$に従ってベストスコアのパスを求めるため、$\delta_t(i)$及び$\psi_t(i)$なる量を定義し、以下の式により求める。

1) 初期化

$$\delta_1(i) = \alpha_1(i) = \begin{cases} 1 & \text{i=1} \\ 0 & \text{otherwise} \end{cases} \quad (1 < i < S) \tag{2.9}$$

$$\psi_1(i) = 0 \tag{2.10}$$

2) 帰納

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(\mathbf{O}_t) \tag{2.11}$$

$$\psi_t(j) = \arg \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] \tag{2.12}$$

ここで、$\delta_t(i)$はt番目のフレームにおいて、それより前の全てのフレームのそれぞれが常にベストスコアの状態をたどり、かつ現在のフレームが存在することのできるN個の状態でベストスコアをもつ状態iの存在確率を表している。また、$\psi_t(i)$は$\delta_{t-1}(j)$の状態を値として持ち、t番目のフレームでのベストスコアパスを表している。

このベストスコアパスを用いて、t番目のフレームが状態iのj番目の混合分布に存在する確率$\gamma_t(i,j)$を次式より求める。

$$\gamma_t(i,j) = P_t(j) \cdot \left[\frac{c_{ij} \mathcal{N}(\mathbf{o}_t, \mu_{ij}, \mathbf{U}_{ij})}{\sum_{m=1}^{M} c_{jm} \mathcal{N}(\mathbf{o}_t, \mu_{im}, \mathbf{U}_{im})} \right] \tag{2.13}$$

$$P_t(j) = 1 \text{ or } 0 \tag{2.14}$$

この$\gamma_t(i,j)$からF-B学習の再推定式と同じ式によりモデルパラメータの更新を行なう。Viterbi学習がF-B学習と異なる点は、$\gamma_t(i,j)$の右辺第1項目がViterbi学習では、1もしくは0であるということである。

第 3 章

実験

3.1 実験条件

3.1.1 学習条件

再学習前の音響モデルは、尤度最大化基準逐次状態分割法 (ML-SSS) [5] により生成された男性不特定話者モデルの Topology に 1 状態 10 混合の無音モデルを付け加えて使用した。その初期音響モデルは状態数は 400 に固定し、初期値は各学習データを時間方向に状態数分だけ分割し、それぞれを混合数と同数にクラスタリングして与え初期出力確率値を計算したもの (VQ) を利用した。そして、繰り返し回数は 10 回、話者数は 175 話者に固定した。実験ではポーズ句単位 (Embedded) とラベル単位 (Label)[1]の 2 種類の学習データを用いて、そのそれぞれ対して Forward-Backward(F-B) 学習と Viterbi 学習のアルゴリズムを使い、その全てに対して混合数を 1,3,5,10 として再学習を行なった。ここで、実験に用いた学習データを表 3.1に示す。なお、データベースとして Spontaneous Speech Database [6] を用いた。

表 3.1: 学習データ

話者数	男性 175 人 (175 対話)
発声数	2,060 発声
発声時間	約 116 分

3.1.2 認識条件

認識評価として、音素タイプライター認識実験と単語認識実験の 2 種類で評価し、音素正解率と単語正解率の比較を行なった。音素タイプライター認識では日本語の 26 音素を対象とした音素単位での評価を行ない、単語認識では辞書の単語総数 6635 語を用いて単語単位で評価を行なった。音素タイプライター認識では第 1 パス及び第 2 パスのビーム幅を 300,000 及び,300,000

[1]対話ごとに MAP/VFS により話者適応モデルを作成し、これを用いて Viterbi アライメントをとった自動ラベルを用いた。

とし、単語認識では 500,000,500,000 とした。テストデータは、学習データのデータベースのうち学習で用いなかった表3.2のデータを使った。表3.3に、テストデータの話者の内訳を表記する。

表 3.2: テストデータ

話者数	男性7人 (7対話)
タスク数	2 (SL2, SL3)
発声数	7発声
発声時間	約4.3分

表 3.3: 認識実験に用いた対話の内訳

タスク	対話番号
SL2	TAS12026
	TAS32007
	TAS32009
	TAS32016
SL3	TAS12001
	TAS32008
	TAS32010

3.1.3 分析条件

学習データは、16kHz でサンプリングされた自然発話音声データを 12kHz にダウンサンプリングしたものを使用した。フレーム長は 20 msec、フレームシフトは 10 msec、特徴ベクトルは 16 次 LPC ケプストラム、パワー、16 次 ΔLPC ケプストラム、Δ パワーの合計 34 次元とした。

3.2 それぞれの再学習法における時間比較

音響モデルの学習時間を Embedded 単位の F-B 学習時間に対する比として表 3.4に示す (() 内は絶対時間 [秒])。この表で Embedded 単位の学習は F-B 学習と Viterbi 学習を混合数で比較すると、混合数が少ないほど Viterbi 学習の時間比が小さく、約 3 倍の高速化が図られている。これらの 2 種類の学習方法の認識率については **3.3.1**に示す。Label 単位の学習では混合数が多いほど Embeddec 単位の F-B 学習に対する時間比が小さくなっている。

表 3.4: 再学習に要した時間の Embedded 単位の F-B 学習に対する比

単位 / 学習方法	混合数			
	1	3	5	10
Emb/F-B	1(26747)	1(46901)	1(66756)	1(117820)
Emb/Viterbi	.2244	.3115	.3417	.3637
Label/F-B	.0533	.0504	.0494	.0469
Label/Viterbi	.0728	.0670	.0460	.0342

Label 単位の学習の 1,3 混合では、F-B 学習と Viterbi 学習の時間比が逆転しているのがわかる。このことが、学習の計算によるものでないことを示すため規模の小さい学習データで実験を行なった。その結果を表 3.5に示す。規模の小さい学習データ (small type) は 20 発声の 6.4 分を用いた。同表の Label 単位での F-B 学習と Viterbi 学習の時間比から Viterbi 学習の方が速いことがわかる。quantify で調べた結果、メモリの allocate 及び free で Viterbi の方が大幅に時間を要していたが原因ははっきりとわかっていない。

表 3.5: Embedded 単位の F-B 学習に対する他の学習方法の時間比 (small type)

単位 / 学習方法	混合数			
	1	3	5	10
Emb/F-B	1(657.1)	1(1216.1)	1(1675.2)	1(2530.1)
Emb/Viterbi	.2560	.3437	.3857	.4829
Label/F-B	.0728	.0637	.0645	.0705
Label/Viterbi	.0504	.0482	.0476	.0535

3.3 認識実験

ATR の連続音声認識システム [7] を用いて、**3.1.2**の条件で自然発話音声の認識を行なった。

3.3.1 音素タイプライタ認識

4 種類の方法で学習されたモデルを ATRlattice により音素タイプライタ認識を行なった結果を表 3.6 と図 3.1 に示す。この結果から Embedded 単位と Label 単位の両方で F-B 学習と Viterbi 学習にほとんど差がないことが分かる。**3.2**で示したように Viterbi 学習の学習速度のほうが速いので、Viterbi 学習を行なった方が非常に効率的である。

表 3.6: 音素正解率 % (平均 (SL2/SL3))

単位 / 学習方法	混合数			
	1	3	5	10
Emb/F-B	66.41(66.54/66.26)	66.73(67.07/66.37)	67.75(66.75/68.79)	68.35(67.60/69.12)
Emb/Viterbi	65.87(65.86/65.88)	66.59(66.65/66.54)	67.51(66.81/68.24)	68.78(68.23/69.35)
Label/F-B	66.68(67.28/66.04)	69.32(69.50/69.12)	69.50(69.97/69.01)	70.31(71.61/68.96)
Label/Viterbi	66.65(66.75/66.54)	69.50(70.03/68.96)	69.50(69.76/69.24)	70.18(71.24/69.07)

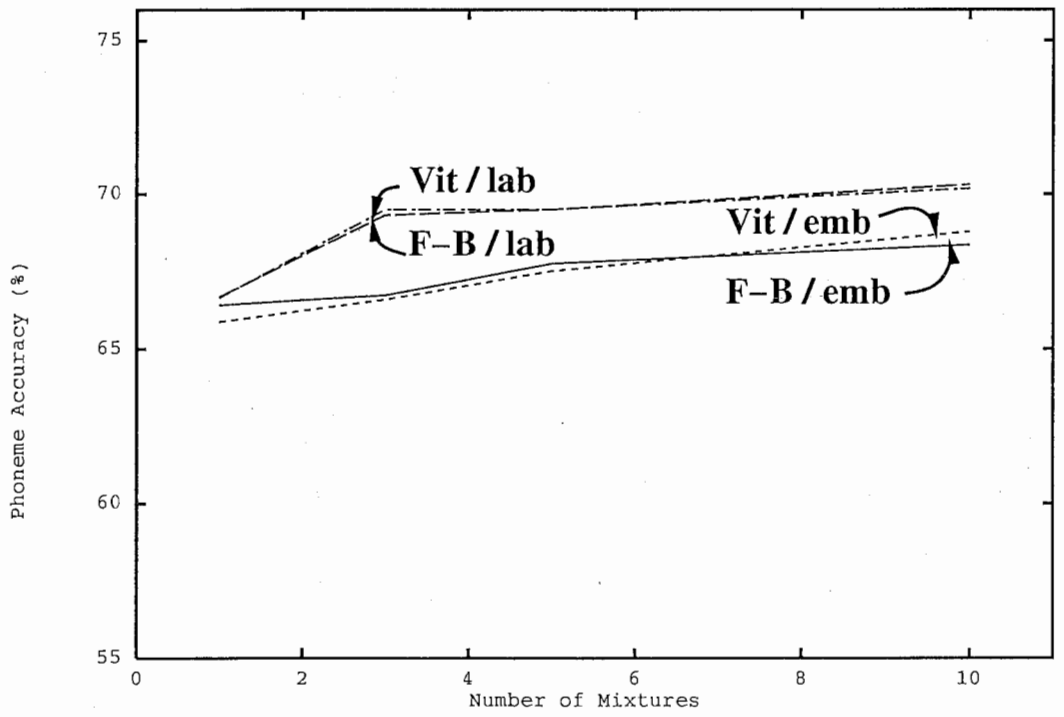

図 3.1: 音素タイプライタ認識の結果

3.3.2 単語認識

3.3.1で使ったモデルと同じを用いてATRlatticeにより、単語認識を行なった結果を表3.7と図3.2に示す。同図から学習方法にはほとんど影響なく同程度の認識率が得られることが分かる。表からは言語モデルがcloseであるタスクSL3が、言語モデルopenであるタスクSL2に比べてはるかに良いこともいえる。

表 3.7: 単語正解率 % (平均 (SL2/SL3))

単位 / 学習方法	混合数			
	1	3	5	10
Emb/F-B	17.58(11.97/23.40)	18.37(15.64/21.20)	19.06(14.48/23.80)	23.77(18.92/28.80)
Emb/Viterbi	16.21(13.71/18.80)	18.47(17.95/19.00)	22.59(18.15/27.20)	21.32(15.44/27.40)
Label/F-B	18.86(16.02/21.80)	21.32(15.25/27.60)	27.41(20.85/34.20)	27.60(24.90/30.40)
Label/Viterbi	18.17(15.64/20.80)	20.33(15.83/25.00)	28.09(21.81/34.60)	25.74(25.10/26.40)

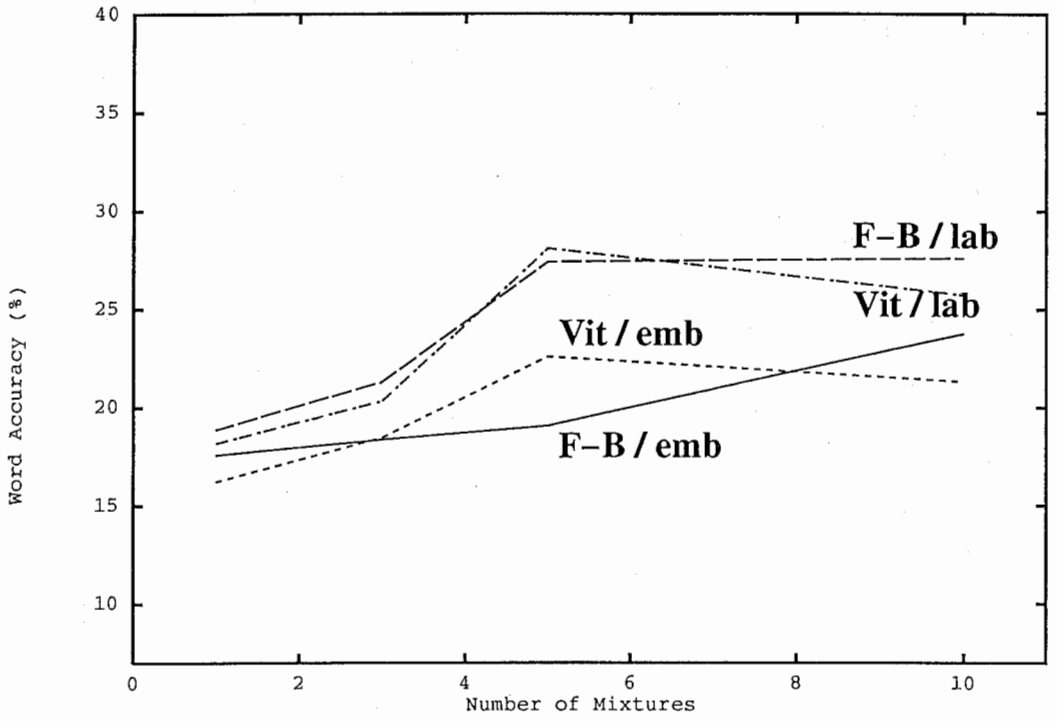

図 3.2: 単語認識の結果

第 4 章

考察

音響モデルを作成する要因としては、学習方法などのほかに繰り返し回数、初期値、混合数、topology 作成に用いる話者数、状態数、再学習データ量など様々である。ここでは、それらが音響モデルの認識率にどのように影響を与えるか考察する。表 4.1に、各実験のパラメータの条件を示す。

表 4.1: 各実験のパラメーター一覧表

	状態数	混合数	学習方法	学習単位
繰り返し回数	400	5	Vit	Emb,Lab
初期値	400	1,3,5,10	Vit	Emb
混合数	400	1,3,5,10,20,50	Vit	Lab
topology	400	1,3,5,10	Vit	Emb
状態数	200,400,600	1,3,5,10	Vit	Emb
再学習データ量	400	1,3,5,10	Vit	Emb
	繰り返し回数	再学習データ	初期値	topology
繰り返し回数	1,2,3,4,5,10,15,20	175	VQ	175
初期値	10	175	VQ,Lab	175
混合数	10	175	VQ	175
topology	10	175	VQ	81,120,175
状態数	10	175	VQ	175
再学習データ量	10	175,372	VQ	175

4.1 繰り返し回数による認識性能の影響

音響モデルの学習は、繰り返し回数を 1,2,3,4,5,10,15,20 と変化させ、混合数は 5 混合のみで、Embedded と Vitervi 単位の Viterbi 学習を行なった。他の要因は **3.1.1**の条件と同じとした。そのモデルで音素タイプライタ認識と単語認識を行なった結果を表 4.2及び 4.3と図 4.1及び 4.2に示す。音素タイプライタ認識では、これらの結果から繰り返し回数が 5 回までは認識率が急速に上昇し、それ以上になると認識率の上昇は緩やかになり、飽和状態になっていることがいえる。そのことから繰り返し回数は 10 回前後で十分であると考えて良い。単語認識では、認識率の上下が大きいが同じことがいえる。

表 4.2: 繰り返し回数の違いによる認識率の比較 (音素タイプライタ認識)

音素正解率 (%)　　平均 (SL2/SL3)

繰り返し回数	学習単位	
	Embedded	Label
1	60.34(59.95/60.76)	67.16(66.60/67.75)
2	64.04(64.01/64.06)	67.86(67.70/68.02)
3	66.27(65.38/67.20)	68.35(68.81/67.86)
4	67.24(66.97/67.53)	69.18(69.66/68.68)
5	67.78(67.02/68.57)	68.48(67.92/69.07)
10	67.51(66.81/68.24)	69.50(69.76/69.24)
15	68.02(67.65/68.41)	70.12(70.82/69.40)
20	68.43(67.97/68.90)	69.96(70.29/69.62)

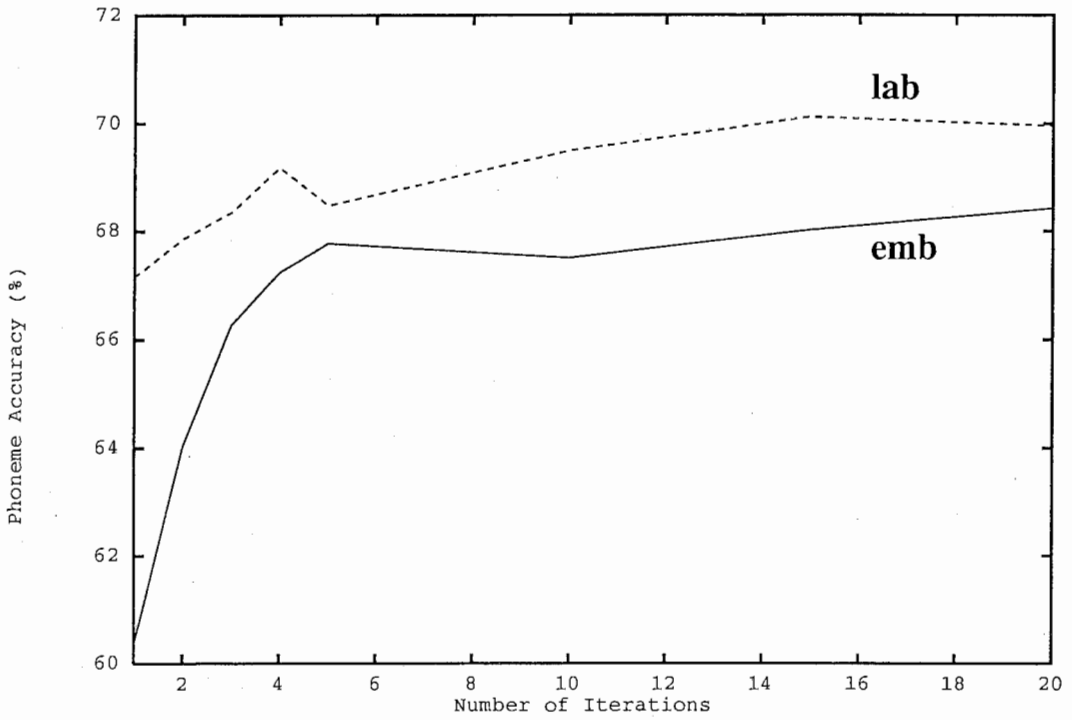

図 4.1: 繰り返し回数による影響 (音素タイプライタ認識)

表 4.3: 繰り返し回数の違いによる認識率の比較 (単語認識)

単語正解率 (%)　　平均 (SL2/SL3)

繰り返し回数	学習単位	
	Embedded	Label
1	16.56(14.48/19.00)	19.45(13.51/25.60)
2	18.17(10.62/26.00)	23.48(17.18/30.00)
3	21.91(15.83/28.20)	25.93(19.69/32.40)
4	20.92(15.44/26.60)	25.64(21.81/29.60)
5	22.20(19.11/25.40)	24.07(19.50/28.80)
10	22.59(18.15/27.20)	28.09(21.81/34.60)
15	20.92(13.51/28.60)	26.92(22.39/31.60)
20	24.46(21.04/28.00)	26.42(18.92/34.20)

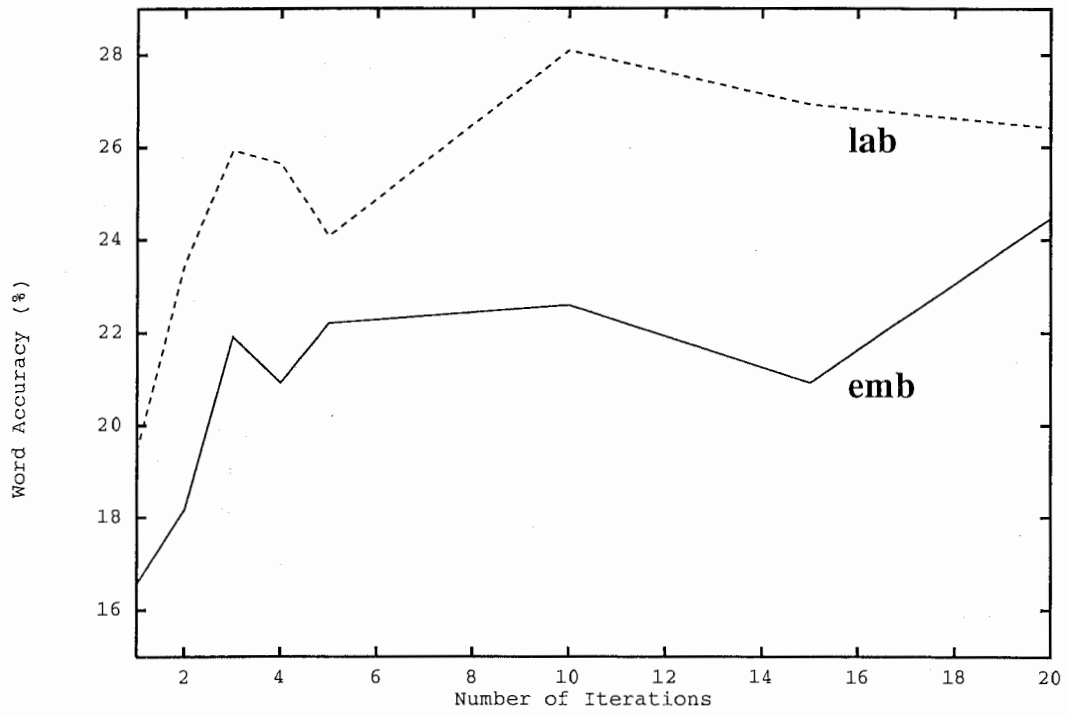

図 4.2: 繰り返し回数による影響 (単語認識)

4.2 初期値に対する認識性能の影響

音響モデルの初期値を VQ(初期モデル 1) 及び **3**章の Label 単位の Viterbi 学習済みのモデル (初期モデル 2) の 2 種類とし、他の要因は **4.1**と同様 **3.1.1**の条件で再学習を行なった。そのモデルの認識結果を表 4.4及び 4.5と図 4.3及び 4.4に示す。音素タイプライタ認識では、図 4.3より明らかに初期モデルが Label 学習済みであるほうが 2 〜 3% ほど認識率が高い。一方、単語認識でも音素タイプライタ認識と同じことがいえる。5 混合では初期値が VQ であるときとほとんど同じであったが、3,10 混合のときは VQ に比べて大幅に改善されていることが分かる。このことから、初期値の選択は重要であり認識性能に大きく影響を与えることがいえる。

表 4.4: 初期モデルを変えた場合の認識性能の比較 (音素タイプライタ認識)

混合数	音素正解率 (%) 平均 (SL2/SL3)	
	初期モデル 1	初期モデル 2
1	65.87(65.86/65.88)	67.16(66.12/68.24)
3	66.59(66.65/66.54)	69.75(70.08/69.40)
5	67.51(66.81/68.24)	70.64(70.98/70.28)
10	68.78(68.23/69.35)	71.01(71.03/71.00)

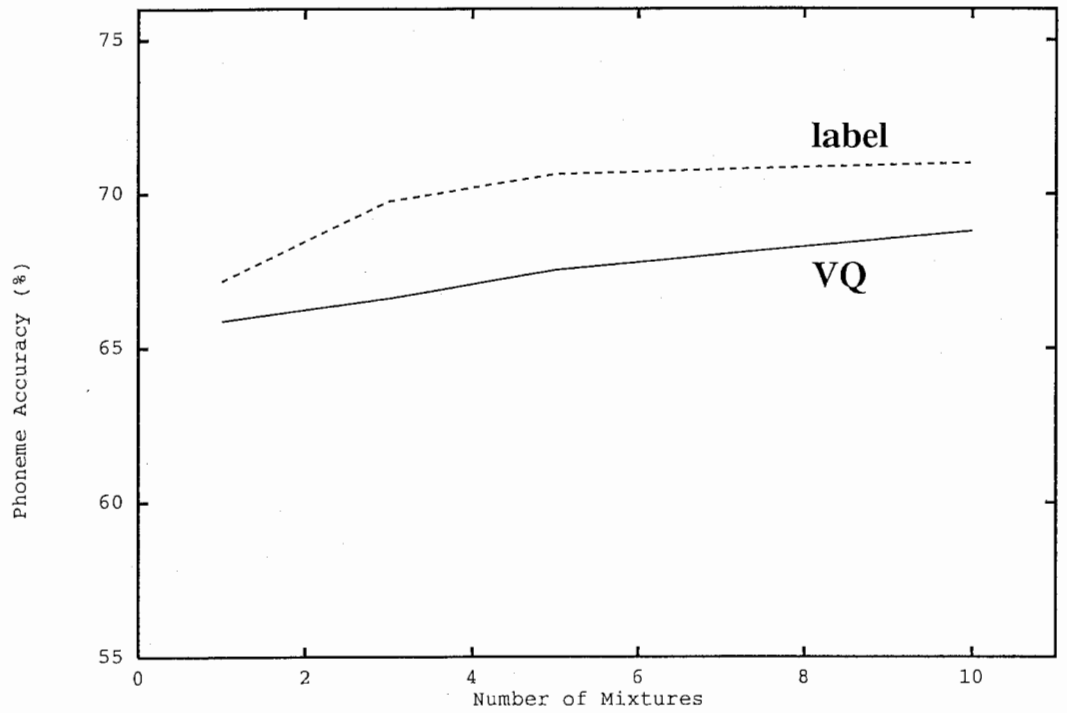

図 4.3: 初期値による影響 (音素タイプライタ認識)

表 4.5: 初期モデルを変えた場合の認識性能の比較 (単語認識)

混合数	単語正解率 (%) 平均 (SL2/SL3)	
	初期モデル 1	初期モデル 2
1	16.21(13.71/18.80)	17.58(12.16/23.20)
3	18.47(17.95/19.00)	29.57(24.52/34.80)
5	22.59(18.15/27.20)	20.53(16.02/25.20)
10	21.32(15.44/27.40)	28.88(23.55/34.40)

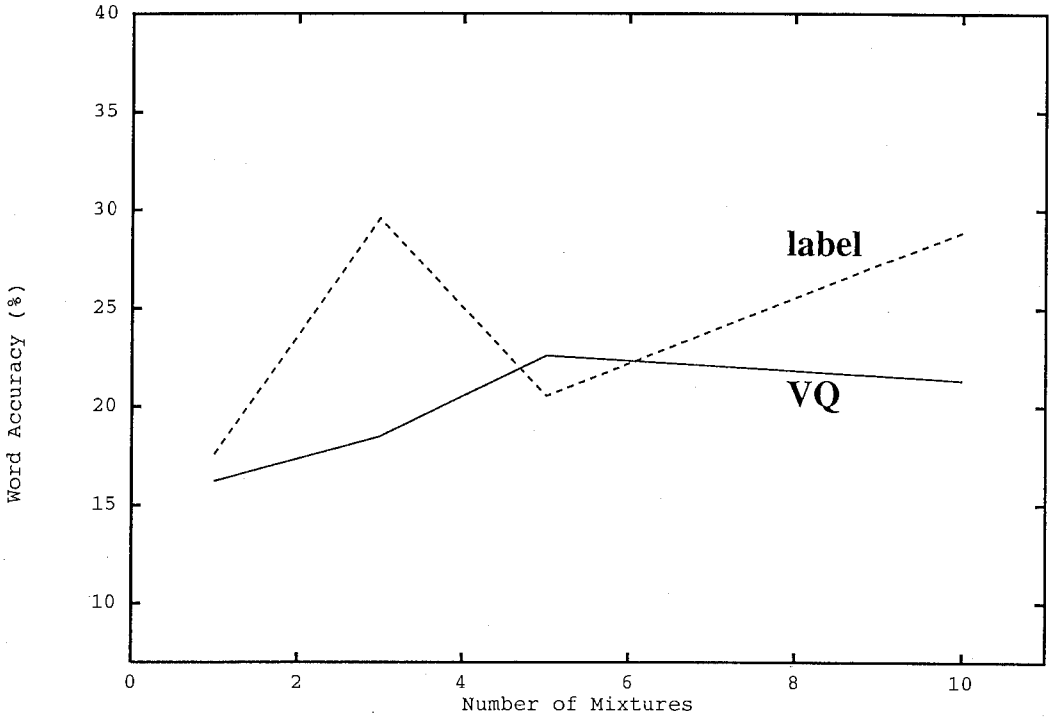

図 4.4: 初期値による影響 (単語認識)

4.3 混合数の増加による認識性能の影響

学習条件の混合数を更に 20 及び 50 混合と増やしたときの認識結果を表 4.6と図 4.5及び 4.6に示す。音素タイプライタ認識では、混合数が 10 を越えると認識率が飽和していることが分かる。また、単語認識では、混合数を 10 混合以上に増やすことによって認識率が下がってしまう傾向にあり、混合数を増やし過ぎても逆効果である。原因としては学習データ量に対して混合数が多過ぎたと考えられる。

表 4.6: 混合数を増加させたときの認識率

混合数	認識方法 (%) 平均 (SL2/SL3)	
	音素正解率	単語正解率
1	66.75(66.54/66.65)	18.17(15.64/20.80)
3	70.03(68.96/69.50)	20.33(15.83/25.00)
5	69.76(69.24/69.50)	28.09(21.81/34.60)
10	71.24(69.07/70.18)	25.74(25.10/26.40)
20	71.56(71.82/71.69)	24.66(19.50/30.00)
50	70.45(70.12/70.29)	18.47(12.71/23.40)

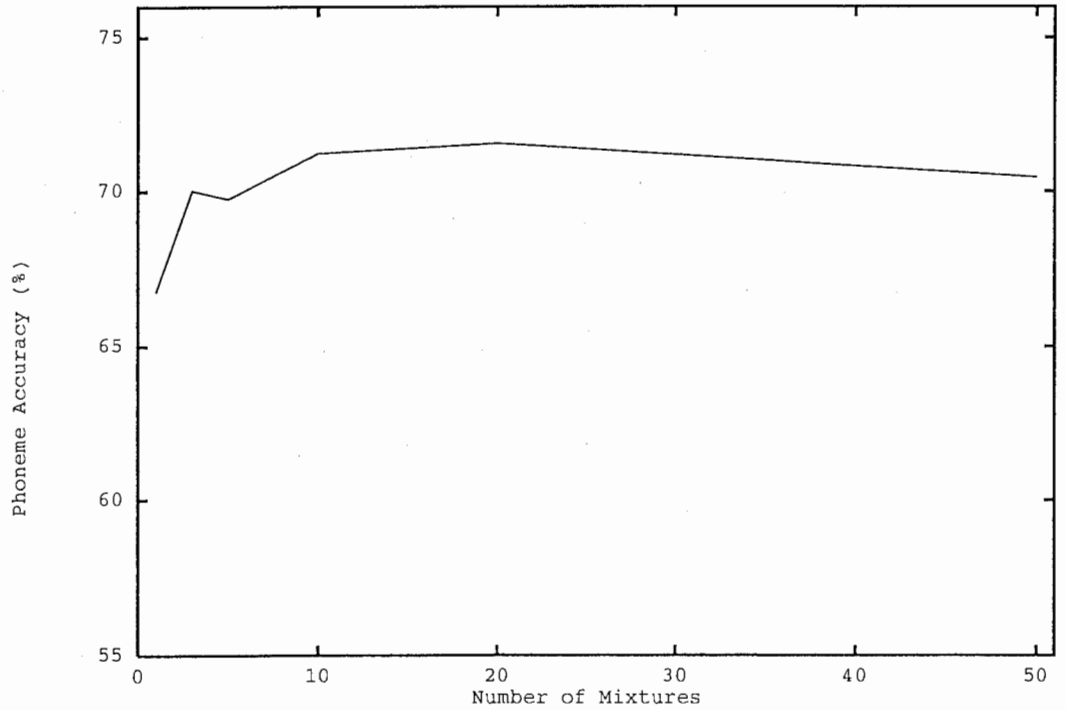

図 4.5: 混合数による影響 (音素タイプライタ認識)

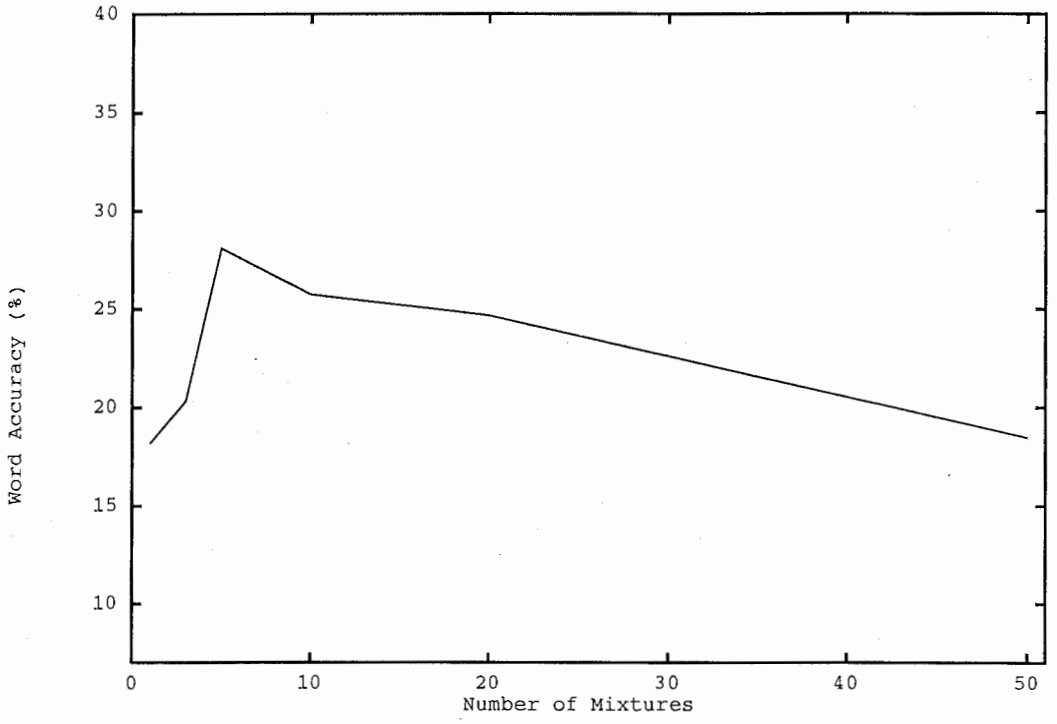

図 4.6: 混合数による影響 (単語認識)

4.4 異なる話者数で作成された Topology による認識性能の影響

Topology を 175 話者のデータで作成していたもの以外に 81 話者と 120 話者のデータで作成した Topology も使って認識性能の影響を調べた。なお、再学習はすべて 175 話者のデータで行なった。その結果を表 4.7及び 4.8と図 4.7及び 4.8に示す。結果より 175 話者で作成された Topology で学習を行なうよりも 81 話者、120 話者で作成された Topology で学習したものの方が良い結果が得られた。これは、音素タイプライタ認識、単語認識のいずれにもいえることで、特に単語認識では 120 話者で作成された Topology が最も良い結果となった。考えられるとしたら、81 話者と 120 話者の Topology は同じタスクのデータベースより作成し、175 話者の Topology は 120 話者のタスクに別の 55 人のタスクを加えたもので作成していることがあげられる。

表 4.7: 異なるデータで topology を作成したモデルによる影響 (音素タイプライタ認識)

音素正解率 (%)　　平均 (SL2/SL3)

混合数	データの種類		
	M081T081.10M4S27SI	TM120.10M4S27SI	AM175.10M4S27SI
1	67.21(67.70/66.70)	66.49(67.18/65.77)	65.87(65.86/65.88)
3	68.64(68.50/68.79)	69.40(69.82/68.96)	66.59(66.65/66.54)
5	70.34(70.55/70.12)	68.99(68.97/69.01)	67.51(66.81/68.24)
10	69.67(69.39/69.95)	71.36(71.08/71.66)	68.78(68.23/69.35)

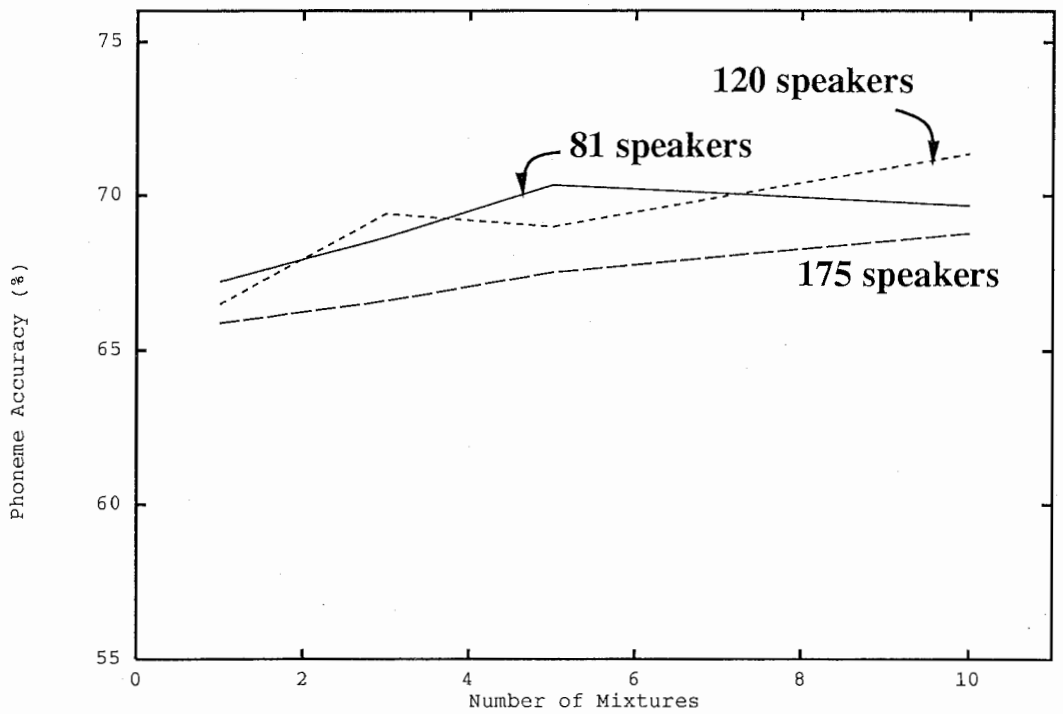

図 4.7: Topology による影響 (音素タイプライタ認識)

表 4.8: 異なるデータで topology を作成したモデルによる影響 (単語認識)

単語正解率 (%)　　平均 (SL2/SL3)

混合数	データの種類		
	M081T081.10M4S27SI	TM120.10M4S27SI	AM175.10M4S27SI
1	21.91(17.57/26.40)	26.13(25.10/27.20)	16.21(13.71/18.80)
3	28.78(27.03/30.60)	31.43(29.73/33.20)	18.47(17.95/19.00)
5	25.64(21.24/30.20)	33.60(32.43/34.80)	22.59(18.15/27.20)
10	30.06(27.41/32.80)	34.28(33.98/34.60)	21.32(15.44/27.40)

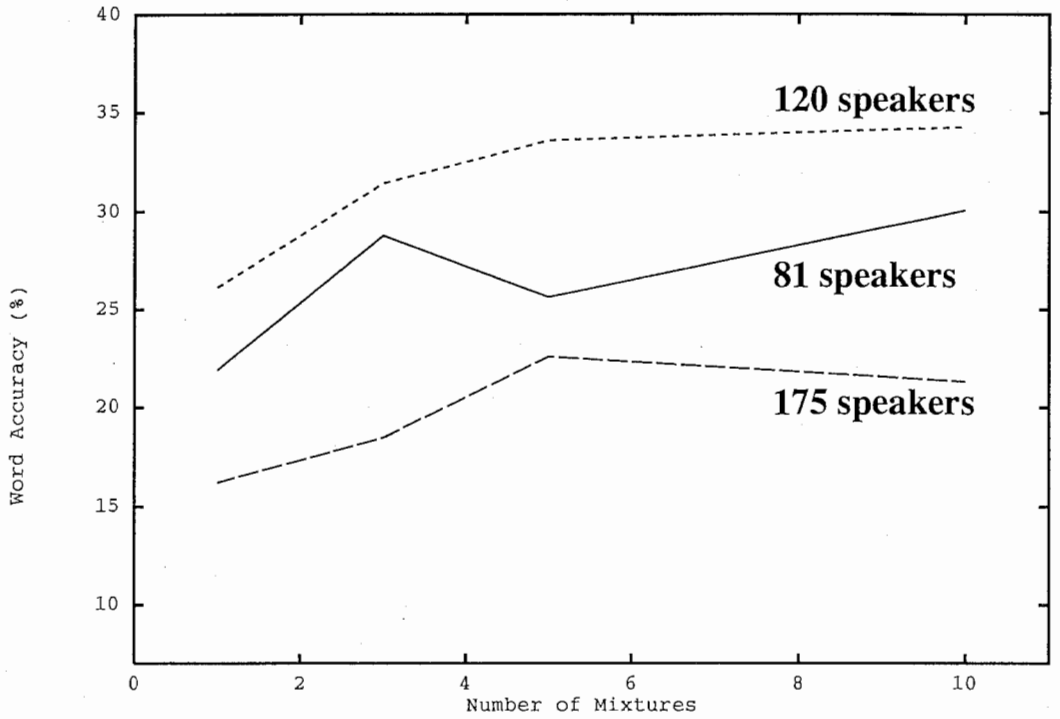

図 4.8: Topology による影響 (単語認識)

4.5 Topology の状態数に対する認識性能の影響

3.1.1の条件で Topology の状態数を 200,600 と変化させて認識実験を行なった結果を表 4.9及び 4.10と図 4.9及び 4.10に示す。音素タイプライタ認識及び単語認識のどちらも状態数が増加することによって大幅に認識率が改善されている。600 状態でも認識率が飽和するようすが見受けられず、更に状態数を増加させることにより認識率の改善も行なわれると考えられる。

表 4.9: Topology の状態数による影響 (音素タイプライタ認識)

音素正解率 (%)　　平均 (SL2/ SL3)

混合数	状態数 200	400	600
1	55.60(56.89/54.27)	65.86(65.88/65.87)	69.13(69.39/68.85)
3	56.22(59.53/52.78)	66.65(66.54/66.59)	70.91(71.03/70.78)
5	57.79(60.95/54.49)	66.81(68.24/67.51)	70.80(71.03/70.56)
10	57.70(60.47/54.82)	68.23(69.35/68.78)	71.15(70.92/71.38)

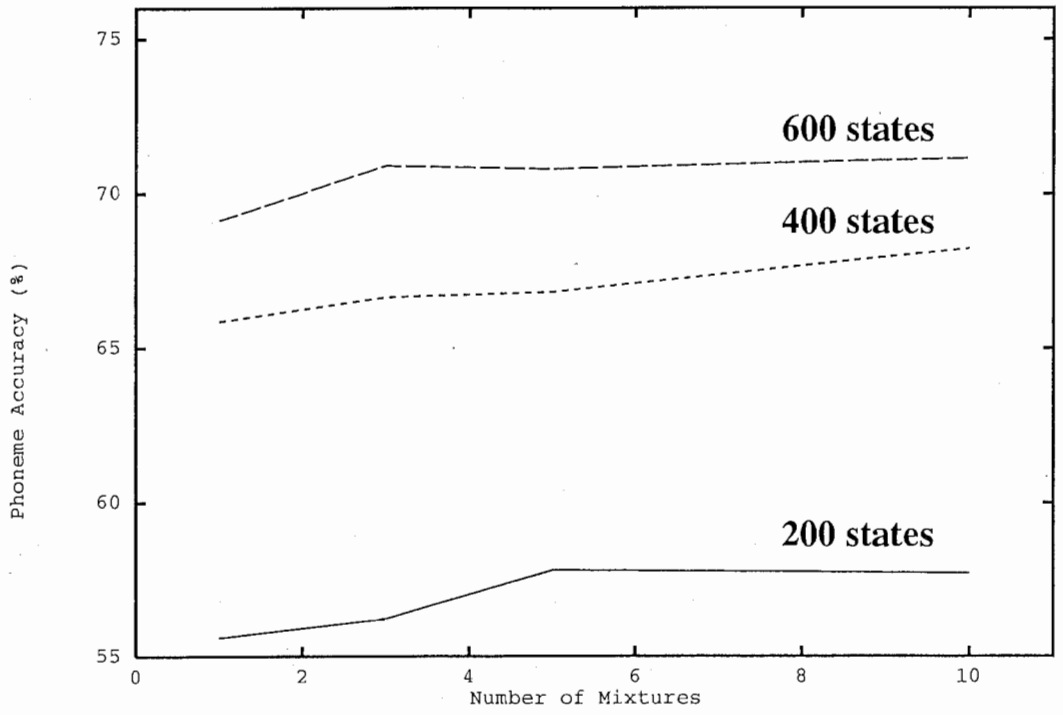

図 4.9: 状態数による影響 (音素タイプライタ認識)

表 4.10: Topology の状態数による影響 (単語認識)

単語正解率 (%)　　平均 (SL2/ SL3)

混合数	状態数		
	200	400	600
1	7.27(5.79/ 8.80)	16.21(13.71/18.80)	25.05(18.34/32.00)
3	8.94(7.34/10.60)	18.47(17.95/19.00)	27.90(20.08/36.00)
5	12.48(13.13/11.80)	22.59(18.15/27.20)	28.78(23.94/33.80)
10	9.82(6.76/13.00)	21.32(15.44/27.40)	28.29(23.17/33.60)

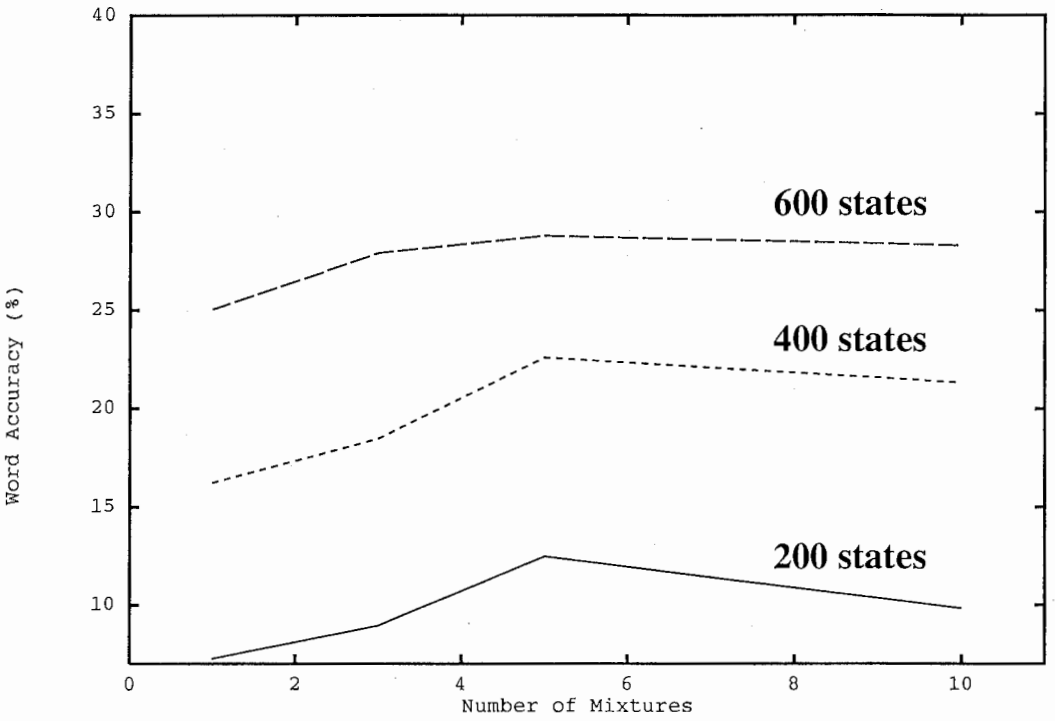

図 4.10: 状態数による影響 (単語認識)

4.6 学習データ量による認識性能の影響

学習条件はそのままで、学習データ量を 175 話者から 372 話者へ増加した場合の影響を調べた結果、表 4.11及び 4.12と図 4.11及び 4.12のようになった。音素タイプライタ認識では、学習データ量が増加してもほとんど認識率は同じであった。しかし、単語認識では 2 〜 3% の改善が見られた。

表 4.11: 学習データ量 (話者数) による違いに対する影響 (音素タイプライタ認識)

混合数	音素正解率 (%) 平均 (SL2/SL3)	
	175 話者	372 話者
1	65.87(65.86/65.88)	65.89(65.65/66.15)
3	66.59(66.65/66.54)	67.81(67.02/68.63)
5	67.51(66.81/68.24)	69.07(68.60/69.57)
10	68.78(68.23/69.35)	68.94(68.39/69.51)

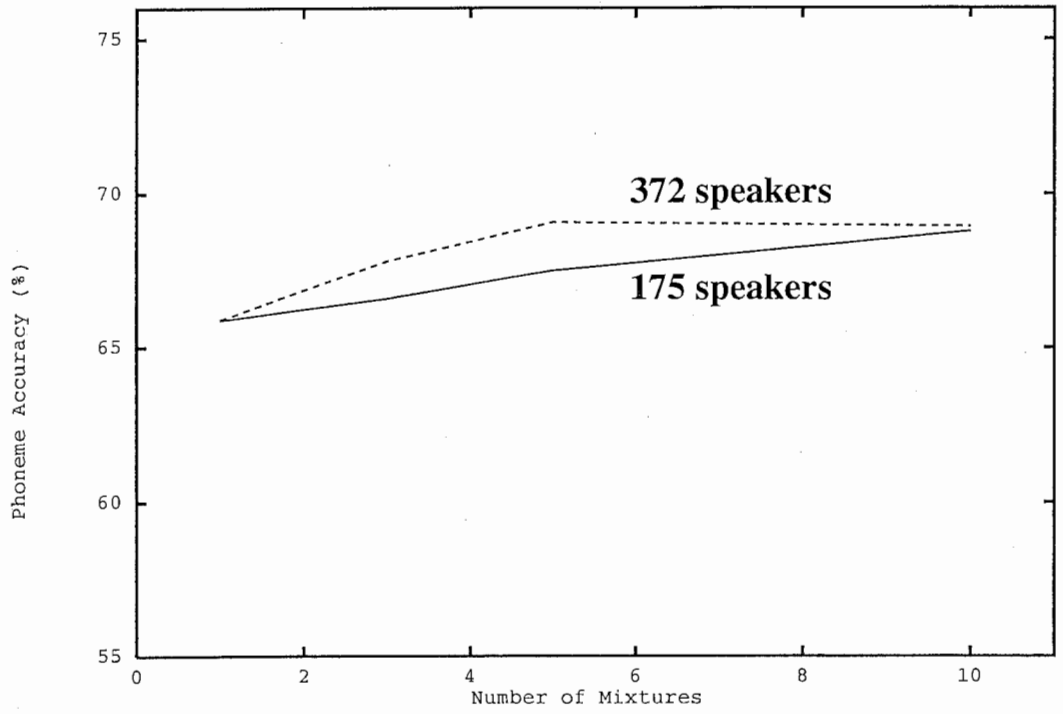

図 4.11: 学習データ量による影響 (音素タイプライタ認識)

表 4.12: 学習データ量 (話者数) による違いに対する影響 (単語認識)

混合数	単語正解率 (%) 平均 (SL2/SL3)	
	175 話者	372 話者
1	16.21(13.71/18.80)	14.64(8.49/21.00)
3	18.47(17.95/19.00)	22.89(19.88/26.00)
5	22.59(18.15/27.20)	25.34(22.01/28.80)
10	21.32(15.44/27.40)	25.05(22.78/27.40)

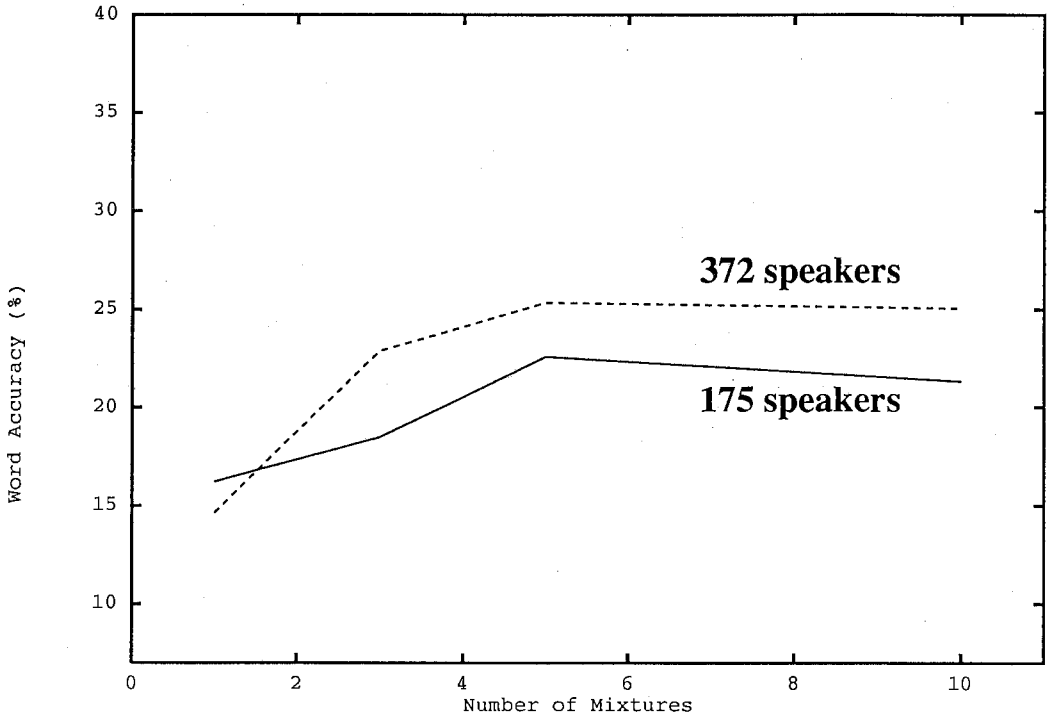

図 4.12: 学習データ量による影響 (単語認識)

第 5 章

結論

本稿はF-B学習とViterbi学習の音響モデルの学習時間及び性能比較を行ない、更にViterbi学習を用いて種々のパラメータ(繰り返し回数、初期値、混合数、Topology、状態数、学習データ量)を変化させた場合の実験を行なった。その結果、Viterbi学習はF-B学習の学習時間と比較して約3～6倍高速化された。そして、認識性能においては両者とも同等の性能を示した。このことから、Viterbi学習で短時間に音響モデルを作成することが可能となった。また、Label単位の学習はEmbedded単位で学習したモデルより認識率が高く、学習速度も大幅に速い。

Viterbi学習を用いた種々の実験では、状態数を増加させた場合に認識率が大幅に上昇する結果が得られ、更に状態数を増加させてもまだ認識率改善の余地が十分にあると予測される。今回考察することの多かったTopologyや状態数については、多岐にわたっての詳細な実験が残っており、認識実験のテストデータが少量であることから、大量のテストデータによる認識評価をする必要がある。

謝辞

本研究を進めるにあたり、Viterbi学習のプログラムを作成していただいたDimitry Rtischev研究員、及び音響モデルを提供していただいた外村政啓研究員に深く感謝いたします。また、御指導、御支援いただいた、Harald Singer研究員、匂坂芳典室長を始めとするATR音声翻訳通信研究所第一研究室の皆様に深く感謝いたします。さらに実務訓練の機会を与えて下さった奈良先端科学技術大学院大学の鹿野清宏教授及びATR音声翻訳通信研究所の山崎泰弘社長に心から感謝いたします。

参考文献

[1] H.Ney, V.Steinbiss, R.Haeb-Umbach, B.-H.Tran, and U.Essen : "An Overview of The Philips Research System for Large Vocabulary Continuous Speech Recognition," International Journal of Pattern Recognition and Artificial Intelligence, Vol.8, No.1, pp.33–70 (1994).

[2] 高木啓三郎, 篠田浩一, 服部浩明, 渡辺隆夫 : "雑音環境の変動を考慮した話者適応化", 信学技報 SP95-100, pp.45–52 (1995.12).

[3] 茂木直子, 柴田和紀, 堀内元, 松本弘 : "制限つき重回帰モデルによる話者適応の検討", 日本音響学会講演論文集, 1-5-22, pp.51–52 (1996.3).

[4] 中川聖一, 越川忠 : "最大事後確率推定法を用いた連続出力分布型 HMM の適応化", 日本音響学会誌, **49**, pp.721–728 (1993).

[5] Harald Singer, Mari Ostendorf : "Maximum Likelihood Successive State Splitting," *Proc. ICASSP-96*, pp.II-601–604 (1996).

[6] Atsushi Nakamura, Shoichi Matsunaga, Tohru Shimizu : "Japanese Speech Databases for Robust Speech Recognition," *Proc. ICSLP-96*, (1996). (to appear)

[7] 清水徹, 山本博史, 松永昭一, 匂坂芳典 : "単語グラフを用いた自由発話音声認識", 信学技報 SP95-88, pp.49–54 (1996.12).

付録

A　データベース一覧表

今回の実験で、学習に用いたデータベースの一覧表を次の表 A.1に示す。

表 A.1: 再学習データベース

SXC20076	SXC20021	SXC20016	SXC20006	SXC20099	SXC20012	SXC20036
SXC20049	SXC20008	SXC20038	SXC20045	SXC20058	SXC20048	SXC20011
SXC20110	SXC20014	SXC20009	SXC20061	SXC20069	SXC20040	SXC20057
SXC20054	SXC20019	SXC20035	SXC20059	SXC20052	SXC20023	SXC20024
SXC20011	SXC20046	SXC20039	SXC20007	SXC20050	SXC20026	SXC20068
SXC20028	SXC20056	SXC20003	SXC20076	SXC20079	SXC20105	SXC20010
SXC20001	SXC20042	SXC20063	SXC20075	SXC20043	SXC20012	SXC20017
SXC20041	SXC20030	SXC20043	SXC20029	SXC20107	SXC20025	TAC70030
TCC60200	TCS70042	TCS70024	TCC70208	TCS70070	TCS70022	TAS70011
TAC70029	TCS70033	TAC60186	TCC70312	TCS70037	TCC70303	TSC71010
TCC70210	TAC71015	TSC71012	TCC70202	TCS70076	TAC00002	TCC70308
TCC71015	TCS70050	TAC00003	TSC71001	TCS70074	TCC71031	TCC70212
TAS70013	TCS70034	TCC22072	TCC70310	TAC71009	TCS70026	TCS70015
TCS70003	TCS70001	TCC71032	TCC71002	TCS70055	TSC71008	TCS70014
TCC70307	TCC71008	TCC71038	TCS70027	TCC70506	TCC71033	TCC71006
TCC71017	TCC70305	TCC70108	TAC70027	TAC71007	TCS70075	TCC70502
TAC60241	TCC71035	TCS70062	TAC60298	TAC71008	TCS70054	TAC71013
TSC71001	TAS33014	TCC71034	TCC70403	TCC70099	TAC00003	TSC71003
TAS70010	TCC70201	TCS70018	TCC70312	TCS70041	TSC71009	TAC71010
TAS70004	TAC70028	TCC70206	TCS70032	TAS70009	TAC60312	TCS70080
TCC70313	TCC71037	TAS70003	TAC71002	TAC60367	TCC70098	TCC70311
TCS70040	TCS70056	TCS70002	TCS70063	TAC60228	TCS70029	TAC00001
TCC70501	TSC71011	TAC71004	TCC70404	TCC70205	TCS70060	TCS70038
TCC70092	TCC70110	TCC71020	TCC70204	TCS70061	TCS70084	TAC71001
TCC70109	SMC75002	SMC75007	SMC75009	SMC75003	SDC71006	SDC71005

B Running マシンの CPU 使用時間の誤差

4 台の Running マシン (atrh15,23,25,30) について学習時間にどれだけ差があるか調べた結果を表 B.1に示す。実験条件は、Embedded 単位の F-B 学習で行ない、繰り返し回数は 10 回、初期値は VQ、混合数は 3 混合のみ、Topology は 175 話者で作成したもの、状態数は 400、学習データ量は 175 話者で行なった。更に、この実験を Running マシンそれぞれについて 5 回ずつ行なった。結果より、Running マシンの CPU 使用時間の誤差は最大で約 1.7% しかなかった。よって、Running マシンによる学習時間への影響はないと考えて良い。

表 B.1: マシンに対する CPU 時間の誤差 (sec)

	実験回数					
マシン名	1	2	3	4	5	平均
atrh15	4088.9	4065.4	4066.6	4065.3	4063.5	4069.9
atrh23	4094.1	4106.0	4084.2	4105.2	4086.2	4095.1
atrh25	4070.0	4070.8	4074.7	4065.5	4068.7	4069.9
atrh30	4097.2	4095.2	4128.6	4133.7	4108.2	4112.6

C 372 話者のデータによる学習の初期値の影響

4.2とまったく同じ実験条件で、学習データのみ 372 話者に変えて学習を行なった。表 C.1及び C.2にそれぞれ音素タイプライタ認識及び単語認識の結果を示す。ただし、初期値を VQ としたときの結果は **4.6**の表 4.11及び 4.12を参照。音素タイプライタ認識では、初期値を label 単位の学習後を初期値にすることによって、いずれも認識率が改善されている。また、学習データ量が増えたことによって認識率が若干改善された。単語認識でも、同じことがいえて学習データ量が増えたことにより認識率が上昇した。

表 C.1: label 単位の学習後を初期値とした場合 (音素タイプライタ認識)

混合数	音素正解率 (%) 平均 (SL2/SL3)	
	175 話者	372 話者
1	67.16(66.12/68.24)	66.81(67.02/66.59)
3	69.75(70.08/69.40)	70.72(71.03/70.39)
5	70.64(70.98/70.28)	71.88(71.35/72.43)
10	71.01(71.03/71.00)	72.41(72.61/72.21)

表 C.2: label 単位の学習後を初期値とした場合 (単語認識)

混合数	単語正解率 (%) 平均 (SL2/SL3)	
	175 話者	372 話者
1	17.58(12.16/23.20)	17.19(12.55/22.00)
3	29.57(24.52/34.80)	26.33(22.20/30.60)
5	20.53(16.02/25.20)	29.08(25.29/33.00)
10	28.88(23.55/34.40)	29.86(25.68/34.20)