002

TR-IT-0238

ガウス混合分布再構成法の適用領域に関する検討
Restructuring Gaussian mixture pdfs in speaker-independent HMM -A study on its effectiveness for several conditions-

杉田 洋介 Sugita Yosuke
中村 篤 Nakamura Atsushi

1997.9.12

既学習のガウス混合分布型の表現力向上を、混合分布の再構成によって図る手法として、ガウス混合分布再構成法が提案されている。本報告では、再構成法の適用領域を明らかにすることを目的として、

- 性別非依存モデル
- Decision tree clustering を用いて作成したモデル

に対して本手法を適用し、単語単位、及び音素単位の連続音声認識実験によって、その効果を確認する。

目 次

第 1 章

はじめに

隠れマルコフモデル (HMM) による音響モデルと単語 N-gram などの言語モデル、及び時間同期ビーム探索の組合せによる連続音声認識においては、音響モデルの表現力不足に因る音響尤度の局所的落ち込みにより、正解の単語系列が枝刈りされたり、複数認識候補中の低い序列に位置してしまうようなことがしばしば起こる。この音響尤度の局所的落ち込みは、多数話者の多様な発音を取り扱う不特定話者音声認識や、発音の崩れが顕著な自然発話を対象とした音声認識において特に多く見られ、これらによる悪影響を克服することは高精度な不特定話者・自然発話音声認識のために極めて重要である。

このような問題に対しては、実際の音響現象に基づいて音響モデルとしての HMM に何らかの改善を図る必要がある。そこで音声認識用の音響モデルの高度化の一貫として、既学習のガウス混合分布型 HMM の表現力向上を、音声サンプルを用いた混合分布の再構成、具体的にはコンポーネントの追加と共有によって図るという手法が提案されている [1]。追加するコンポーネントは他のガウス混合分布の中から選択し、追加元と追加先の両ガウス混合分布で共有する。これにより HMM 全体としてのガウス分布を増大させることなく、個々のガウス混合分布の表現力の向上を図ることができる。この手法に関しては、学習用サンプル、テストデータとして男性話者による自然発話音声、初期モデルとして ML-SSS アルゴリズム [4] によって作成した HMnet を、言語モデルとして可変長単語 N-gram[9] を用いた音声認識の組合せによってその効果が確認されている。

本報告では、本手法に対して、種々の条件下でその効果を確認し、その適用領域を明らかにする。具体的には以下の二つの実験を行なう。

- 性別非依存の音響モデルを再構成し、連続音声認識実験によって本手法の効果を確認する。
- 初期モデルの生成アルゴリズムとして Decision tree clustering を用い、異なるアルゴ

リズムにおける効果の差異を確認する。

本報告の構成は以下の通りである。まず第2章では、ガウス混合分布再構成法の概要について述べる。第3章では、性別非依存のモデルに対して音声認識実験を行ない、その結果を示す。第4章では、初期モデル生成アルゴリズムの違いについて述べ、Decision tree clustering による初期モデルに対して音声認識実験を行ない、その結果を示す。さらに第5章で実験結果に対する考察を行ない、今後の課題を述べる。

第 2 章

ガウス混合分布再構成法の概要

2.1 ガウス混合分布再構成を中心とした処理の流れ

Baum-Welch アルゴリズムなどの一般的な手法によって、あらかじめ初期 HMM が作成されているものとする。

音声サンプルに基づき、初期 HMM に対してガウス混合分布の再構成を行なう。再構成後の HMM に対して、パラメータを再推定し、最終的な HMM とする。ガウス混合分布の再構成、及びパラメータ再推定においては、基本的に初期モデルの作成に用いた音声サンプルをそのまま用いる。従って本処理のために、新たに音声サンプルを用意する必要はない(図 2.1)。

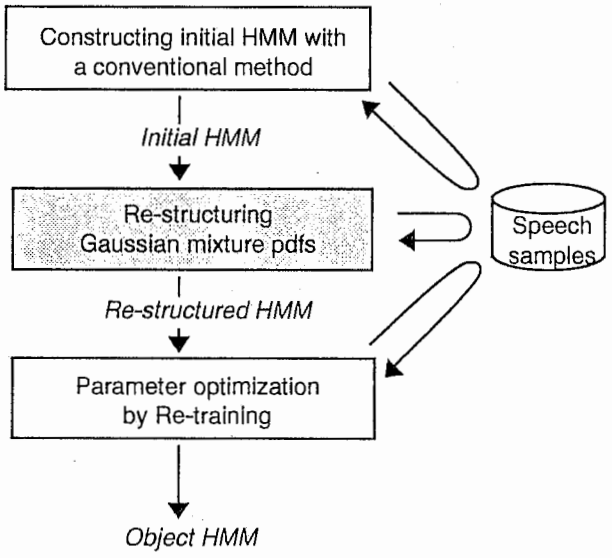

図 2.1: 分布再構成を中心とした処理の流れ

2.2 コンポーネント追加に関する考え方

提案手法の本質は、音声サンプルに対して初期 HMM が起こすフレーム単位の識別誤りの傾向を考慮して、ガウス混合分布のコンポーネントの追加を行なうことである。このフレーム誤りとは、初期 HMM と音声サンプルによる Viterbi alignment によって得られた、各時刻の特徴ベクトルとガウス混合分布の対応において、以下の条件を満たす場合を指す。

$$g_t \neq \underset{\gamma \in \Gamma}{argmax}\{P(o_t|\gamma)\} \tag{2.1}$$

但し、

o_t:	時刻 t における特徴ベクトル
g_t:	Viterbi alignment によって時刻 t に割り当てられたガウス混合分布
Γ:	初期 HMM のガウス混合分布の集合
$P(o\|g)$:	分布 g から特徴ベクトル o が出力されることに対する尤度

以下、フレーム誤りを事象 E で表し、E の余事象を E^C と書くことにする。

あるガウス混合分布についてのフレーム誤りが、ある音響現象を照合の対象としたときに頻繁に起こるならば、このガウス混合分布は、実際の音声認識において、正解経路上で当該音響現象との照合を行なう際に音響尤度の落ち込みを起こしやすい。

今、時刻 t について以下のガウス分布の集合を考える。

$$x_t = \underset{\xi \in \Xi}{argmax}\{P(o_t|\xi)\} \tag{2.2}$$

Ξ: 初期HMMの全体のガウス分布の集合

x_t は本来 HMM 全体の中のいずれかのガウス混合分布のコンポーネントであるが、ここでは特徴ベクトル o_t に対して最大の音響尤度を与える単独のガウス分布として扱う。

このとき、系列 $\{x_t\}$ は、特徴ベクトル系列 $\{o_t\}$ に対して尤度最大を与える、Ξ の元 ξ の系列 (ガウス分布の最尤系列) となる。同様に、特徴ベクトル系列 $\{o_t\}$ に対して Viterbi 経路を与える、Γ の元 γ の系列、即ちガウス混合分布の Viterbi 系列 $\{g_t\}$ を考える。$\{g_t\}$ と $\{x_t\}$ について、各々の元の出現頻度を分析することにより、条件付きフレーム誤り確率分布、$P(E,\xi|\gamma)(\gamma \in \Gamma, \xi \in \Xi)$ が得られる。ある ξ について、$P(E,\xi|\gamma)$ が大きい値を持つならば、ガウス混合分布 γ は ξ の近傍の音響現象との照合を行なう際に音響尤度の落ち

込みを起こしやすいと言える。そこで、そのフレームに対して最大の音響尤度を与えるガウス分布 ξ を γ のコンポーネントとして新たに追加することにより、尤度の落ち込みを抑止することができると考えられる。

2.3 ガウス混合分布再構成の手順

追加したコンポーネントは、その分布が元々属していたガウス混合分布と共有する。以下の手順で再構成を実行することにより、コンポーネントの追加、共有が実現される。

1. 初期 HMM と音声サンプルの間で Viterbi alignment を実施し、Γ の元の系列であるガウス混合分布の Viterbi 系列 $\{g_t\}$、及び Ξ の元の系列であるガウス分布の最尤系列 $\{x_t\}$ をそれぞれ求める (図 2.2)。

2. ステップ 1 で得た系列から、ガウス混合分布 γ とガウス分布 ξ の全ての組合せについて、条件付きフレーム誤り確率 $P(E,\xi|\gamma)$ を算出する (図 2.3)。

3. 全てのガウス混合分布 γ についてステップ 4 を実行する。

4. 全てのガウス分布 ξ に付いてステップ 5 を実行する。

5. $P(E,\xi|\gamma)$ があらかじめ定めた閾値を越える場合、ξ を γ の新しいコンポーネントとして追加する (図 2.4)。追加したコンポーネントは、γ と ξ が元々属していたガウス混合分布の間で共有する (図 2.5)。

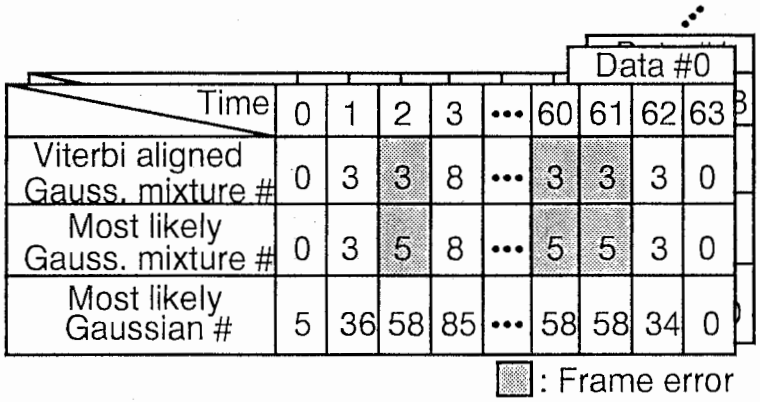

図 2.2: Viterbi 系列と最尤系列

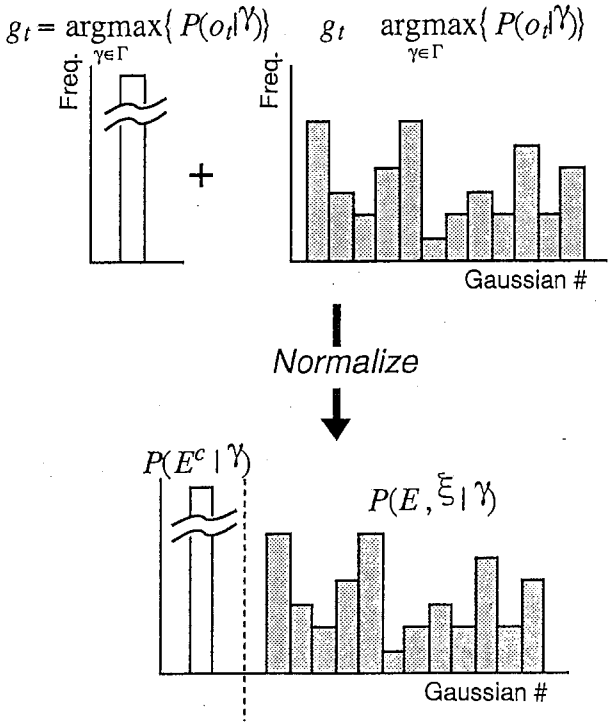

図 2.3: フレーム誤り確率の算出

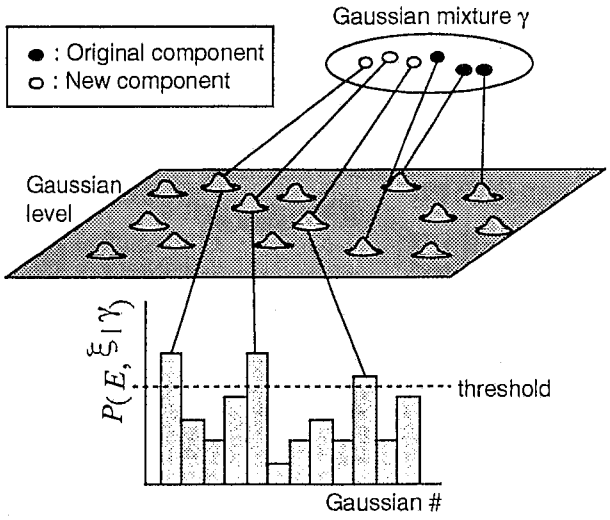

図 2.4: 誤り確率に基づくコンポーネント追加

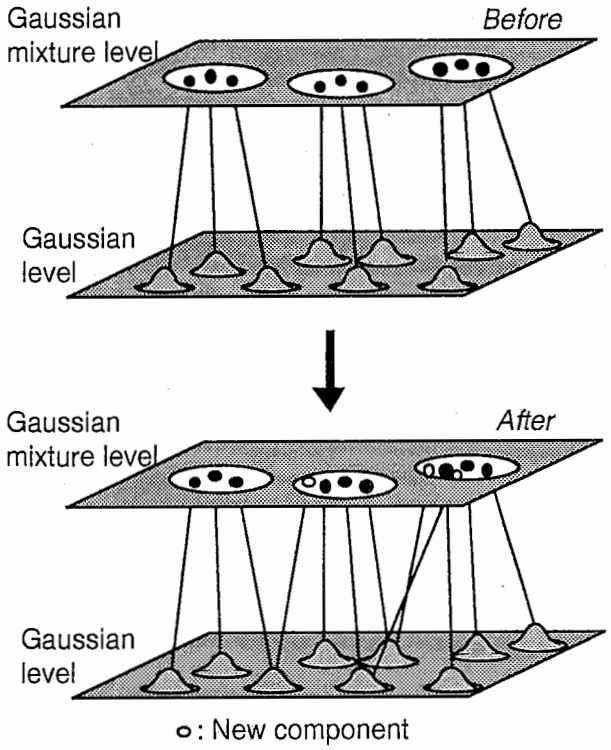

図 2.5: 追加コンポーネントの共有

2.4 パラメータの再推定

初期 HMM に対するガウス混合分布再構成の後、尤度最大、尤度比最大等の基準により、以下のパラメータを再推定する。

- ガウス分布の平均
- ガウス分布の分散
- ガウス混合分布の混合重み
- 状態遷移確率

ガウス分布の平均、分散、及び状態遷移確率については、初期 HMM の値をそのまま初期値として用いる。またガウス混合分布の混合重みについては、フレーム誤り確率、及びコンポーネントの追加実行の閾値を考慮して、次のように初期値を定める。

まずガウス混合分布 γ にガウス分布 ξ が新たなコンポーネントとして追加された場合の混合重み初期値を、条件付きフレーム誤り確率の値をそのまま用いて、

$$\widetilde{W}_{\xi}^{\gamma} = P(E, \xi|\gamma) \tag{2.3}$$

とする。コンポーネントが追加されたことにより、初期 HMM に元々含まれていたコンポーネントの混合重みに対しても新たな初期値が必要となる。これらを以下により与える。

$$\widetilde{W}_{\xi}^{\gamma} = \left(P(E^C|\gamma) + P(E, \Xi_{\rho}^{\gamma}|\gamma)\right) \bullet W_{\xi}^{\gamma} \tag{2.4}$$

但し、

W_{ξ}^{γ} : 初期HMMにおけるガウス分布ξのガウス混合分布γにおける混合重み

ここで Ξ_{ρ}^{γ} は、コンポーネント追加実行の閾値が ρ の時に、ガウス混合分布 γ に対するコンポーネント追加の対象とならないガウス分布の集合であり、以下によって与えられる。

$$\Xi_{\rho}^{\gamma} = \{\xi | P(E, \xi|\gamma) < \rho; \xi \in \Xi\} \tag{2.5}$$

第 3 章

話者非依存モデルへの適用

本手法に関しては、前記したように ML-SSS によって作成した男性話者用不特定話者 HMnet に関して実験を行ない、効果が確認されている。話者非依存（男女混合）モデルにおいてもその効果を確認するため、以下の実験を行なった。

3.1 ガウス混合分布の再構成

第 2章で述べた手法により、既学習 HMM のガウス混合分布再構成を行なった。初期モデルにおいて、コンポーネントの追加実行の閾値 ρ を 0.01 から 0.05 まで変化させ、混合分布の再構成を行なった。

初期 HMM としては、 ATR Travel Arrangement Corpus の男性 100 名、女性 130 名による自然発話音声を学習用音声サンプルとして作られた HMnet を用いた。初期 HMM に用いた HMnet についての条件を表 3.1、図 3.1に示す。

表 3.1: 音響分析条件

Sampling freq.	12 kHz
Quantization	16 bit liner
Pre-emphasis	1-0.97z^{-1}
Window	20 ms Hamming
Frame shift	10 ms
Feature Vector	log-power + 16-order LPC-Cep. + Δlog-power + 16-orderΔCep.

```
・801 states speaker independent HMnet
  800 states for state-shared allophone HMMs
   (Triphone-context-dependent HMMs)
  1 state for silence HMM
・Acoustical units: Japanese 25 phoneme + silence
・Mixture size: 5 mixture/state
・Covariance type: Diagonal
```

図 3.1: HMnet の構造に関する条件

3.2 連続音声認識実験

前記の初期 HMM を再構成することによって得られた HMM(以後 Restructured-HMM) を用いて連続音声認識実験を行なった。実験条件を以下に示す。

連続音声認識器 マルチパス探索と単語グラフ出力を特徴とする連続音声認識器 (ATRlattice r04r04)[8]

言語モデル 可変長単語クラス N-gram、分離クラス数：500[9]

単語辞書 語彙数：6579

テストデータ ATR Travel Arrangement Corpus(S1,S2,S4)、44 話者 [5]

3.2.1 閾値による比較

コンポーネント追加実行の閾値を 0.01 から 0.05 まで変化させ、Restructured-HMM と、初期 HMM をそのまま用いた場合とそれぞれ認識率を比較した。連続音声認識器のビーム幅は 85,85、言語重みは 8.0,8.0 とした。

図 3.2に結果を示す。グラフの横軸は、各閾値における再構成後の出力分布の混合数の総数とする。閾値と混合数の対応を表 3.2に示す。

表 3.2: 閾値毎の混合数の総和

閾値	0.05	0.04	0.03	0.02	0.01
混合数	4130	4296	4784	6332	13594

Restructured-HMM における単語 Accuracy 、単語 %Correct は、いずれも閾値を大きくすると一時的に認識率が下がり、再び上がるという傾向が見られた。

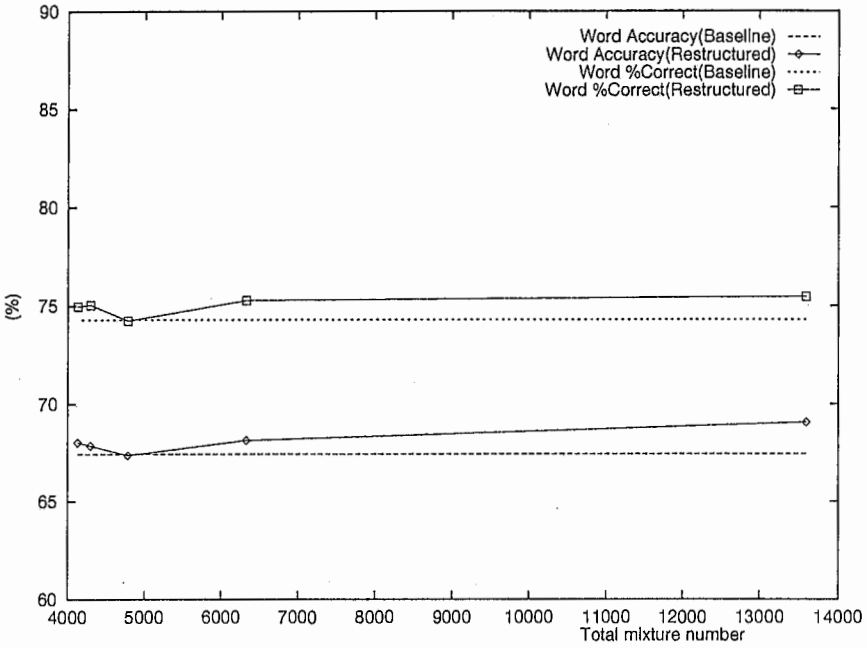

図 3.2: 再構成後の音声認識結果

3.2.2 単語認識実験

先の実験において最も効果があった閾値による Restructured-HMM を用いて、単語認識実験を行ない、初期 HMM の場合と比較した。初期 HMM を用いた音声認識において、単語 Accuracy が最大になるように以下の順序でビーム幅、及び言語尤度重みを変化させた。

1. 探索ビーム幅を変化させる
2. 第 1、第 2 パス探索用の言語重みを同時に変化させる
3. 第 2 パス探索用の言語重みのみを変化させる

初期 HMM に対する最適設定 (ビーム幅 80,80、言語重み 8.0) から、第 2 パス探索用の言語重みのみを変化させた際の認識結果を図 3.3に示す。単語 Accuracy、単語 %Correct いずれについても Restructured-HMM が初期 HMM を上回っていることがわかる。

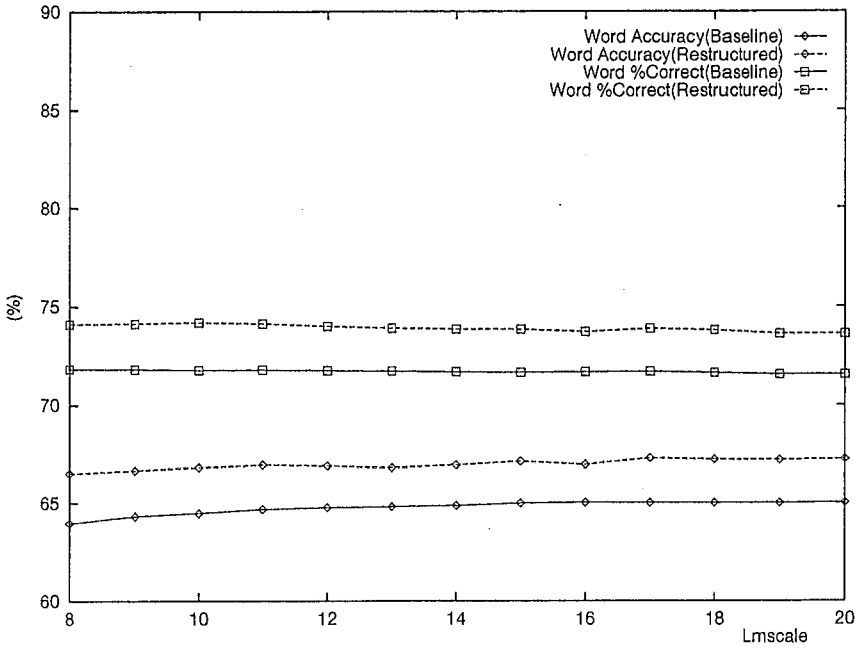

図 3.3: 単語認識結果の比較

初期モデルに対する最適設定での認識率の最大値と、そのときの再構成後のモデルの認識率を表 3.3に示す。またこの時の対話毎の認識率を付録 B に示す。

表 3.3: 認識率の最大値 (%)

	Accuracy	%Correct
初期モデル	65.02	71.84
再構成後のモデル	66.96	74.10

3.2.3 音素認識実験

単語認識と同様に、最も効果があった閾値による Resutructured-HMM と初期 HMM を用いて音素認識実験を行ない、ビーム幅ごとに認識率を比較した。

図 3.4に認識結果を示す。再構成による認識率の上昇はあまり見られず、ビーム幅によっては認識率の降下も見られた。本手法では追加したコンポーネントを他の音素と共有することになるので、それらの音素の識別は難しくなる。このことが音素認識において悪影響を及ぼしていると考えられる。

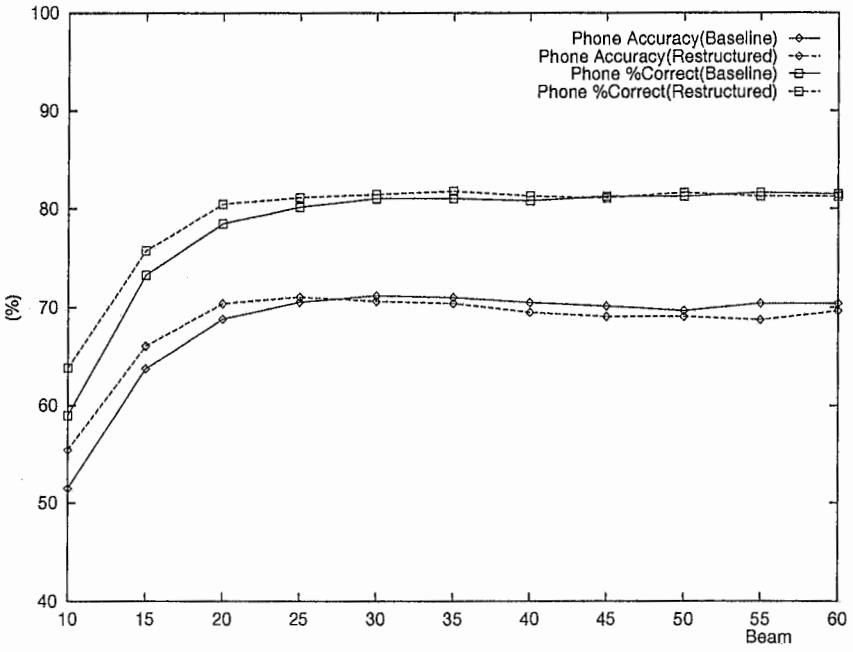

図 3.4: 音素認識結果の比較

第 4 章

Decision tree clustering による初期モデルへの適用

初期モデルとして ML-SSS による HMnet を用いた実験においては、本手法の有効性を確認することができた。この結果から、状態を分割することによって表現不能になった音響現象を、再構成によって再び表現可能にすることができたと考えられる。

さらに、異なるアルゴリズムで生成された音響モデルに対して再構成を行ない、効果の差異を確認する。まず実験に用いた二つの音響モデル生成アルゴリズムについて以下に示す。

4.1 隠れマルコフモデル (HMM) による音響モデル

隠れマルコフモデル (Hidden Markov Model:HMM) は現在音声認識に広く利用されており、認識性能や頑健性の点で優れた手法の一つである。最近では、音素の音響パターンを変動させる要因となる先行音素や後続音素などの音素環境に依存するようなモデルが用いられるようになってきた。

本報告で用いたガウス混合分布再構成法は、音声認識用音響モデル高度化のための一手法であり、音響モデルとしての HMM の表現力向上を図ることを目的としている。この初期モデルの生成法として ML-SSS アルゴリズムを用いた場合においては、本手法の効果が確認されているが、異なる生成法での効果も確認する必要がある。本論文では異なる初期モデル生成アルゴリズムとして Decision tree clustering を用いた。この二つのアルゴリズムについて、簡単に説明する。

4.1.1 逐次状態分割法

逐次状態分割法 (Successive State Splitting:SSS) は、全音素環境を表す一状態の HMnet(Hidden Markov Network) から、状態を音素環境方向、または時間方向へ分割すること

により、自動的に HMnet 構造を決定する方法である。ここで生成される HMnet とは、複数の HMM の状態をネットワーク状に連結したモデルであり、状態共有のない通常の HMM を完全に包含した、より柔軟性の高い HMM の表現方法となっている。状態の分割を適切に行なうことによって、状態共有構造を持った HMnet を生成する (図 4.1)。

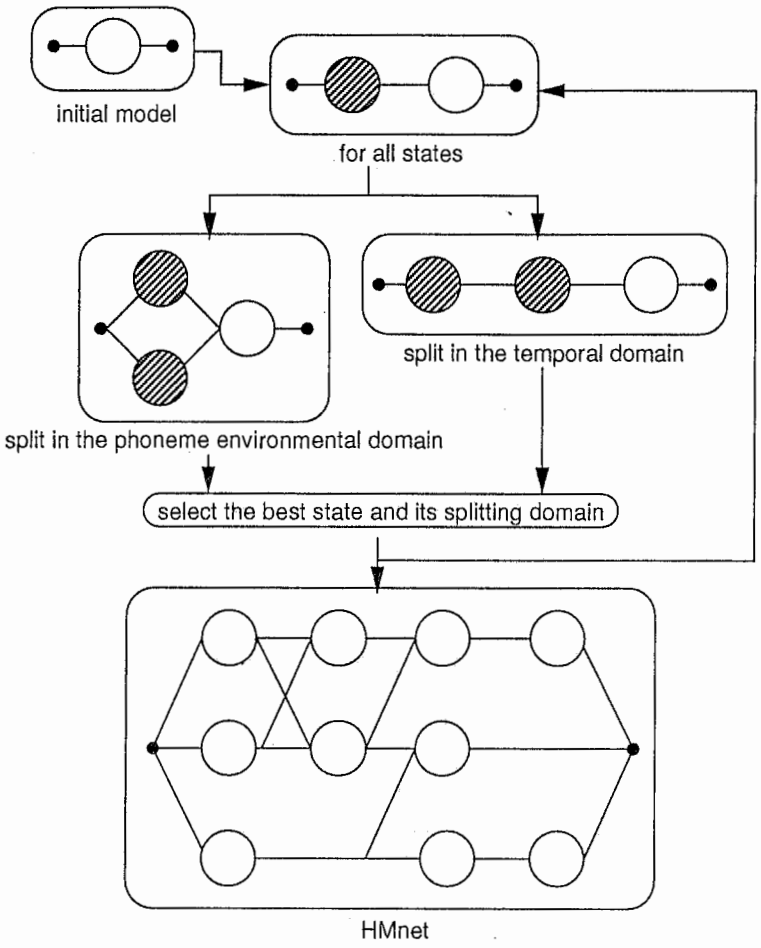

図 4.1: SSS の原理

本報告における実験では、HMnet の生成アルゴリズムとして SSS の改良型アルゴリズムである ML-SSS(Maximum Likelihood SSS)[4] を用いた。通常の SSS では、最良の分割状態が実際の分割の前に選択される。即ちある時点で存在している全状態のうち、音響パラメータ空間上で最も大きく広がった出力確率分布を持つ状態を分割することが、全体の出力尤度を向上させるために最も効果的であるという仮定に基づき分割を行なう。しかしこの基準によって行なわれた分割が、必ずしも最良の選択であるとは限らない。ML-SSS では、分割対象の状態と要因を、分割による尤度増加の期待値に基づいて決定することにより、こ

の問題の改善が図られている。

4.1.2 Decision tree clustering

Decision tree clustering は、SSS とは逆に HMM の全ての音素モデルを用意しておき、それらを音素決定木によってクラスタリングし、最終的に同一クラスに属する状態を融合することによって、状態共有構造を得る方法である (図 4.2)。音素決定木は音素の音響的変動をとらえ、かつ未知音素環境の音響的特性を予測する方法であり、二分木で表され、各ノードにおいて yes/no の質問が与えられる。この質問によって各モデルを音響的に最も類似した部分に分けていくことができる。決定木における各ノードに与える質問は、逐次クラスタリングによる尤度の増加を最大にするようなものが選ばれる [7]。

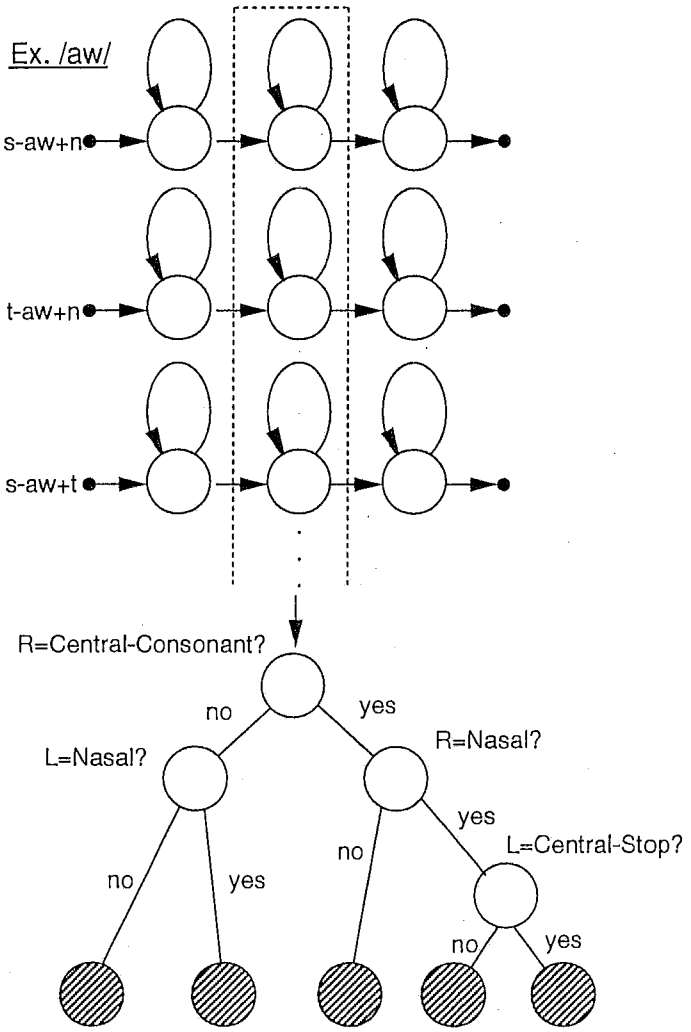

図 4.2: 音素決定木

4.2 Decision tree clustering による Triphone HMM 作成

不特定話者の音声サンプルから、HTK を用いて Decision tree clustering による初期 HMM を作成した [6]。音素決定木における各ノードに与える質問は、文献 [3] で用いられているものを使用した。

4.3 連続音声認識実験

初期モデル生成アルゴリズムとして Decision tree clustering を用い、モデルを再構成し、ML-SSS による場合と同様の実験条件下で音声認識実験を行なった。初期モデルの条件は図 4.3に示す。

```
・802 states speaker independent HMnet
  801 states for state-shared allophone HMMs
   (Triphone-context-dependent HMMs)
  1 state for silence HMM
・Acoustical units: Japanese 25 phoneme + silence
・Mixture size: 5 mixture/state
・Covariance type: Diagonal
```

図 4.3: HMnet の構造に関する条件

ML-SSS によるモデルと同条件で比較を行なうため、作成された初期 HMM に対してポーズセグメント単位で再学習を行なった。また第 3章で用いた音声認識器での使用を可能にするために、得られた HMM(HTK v2.0 形式) に対してフォーマット変換を施した。

4.3.1 閾値による比較

図 4.4に音声認識結果を示す。SSS による初期モデル同様の傾向が見られ、閾値 0.02 以外では認識率が向上した。

表 4.1: 閾値毎の混合数の総和

閾値	0.05	0.04	0.03	0.02	0.01
混合数	4124	4301	4726	6347	13453

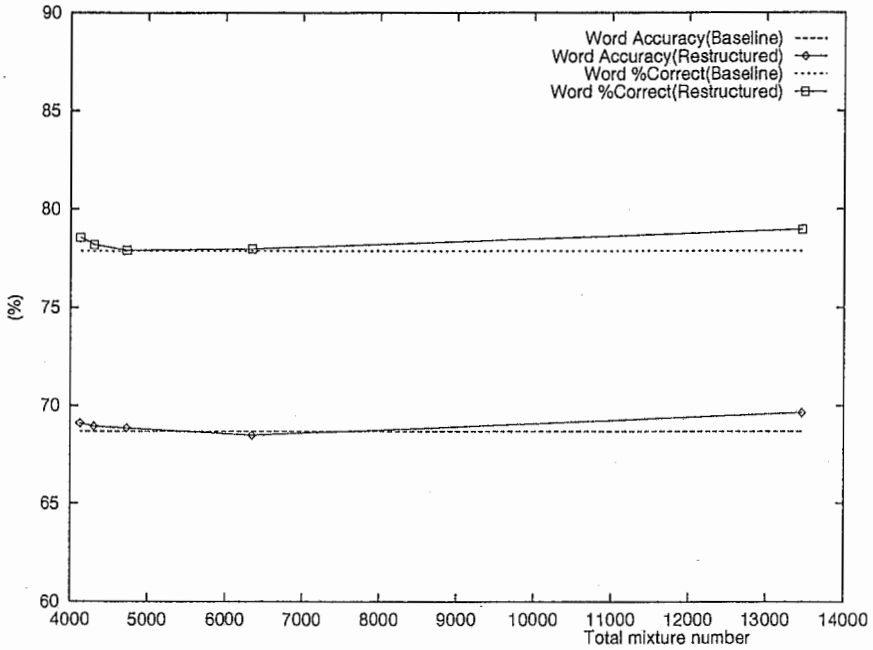

図 4.4: 再構成後の音声認識結果

4.3.2 単語認識実験

初期 HMM における最適設定 (ビーム幅 85,85、言語重み 8.0) に対する単語認識結果を図 4.5に示す。Decision tree clustering によるモデルを再構成した場合でも、いくらか認識率は上がった。但し SSS によるモデルに比べるとその改善の度合は小さかった。

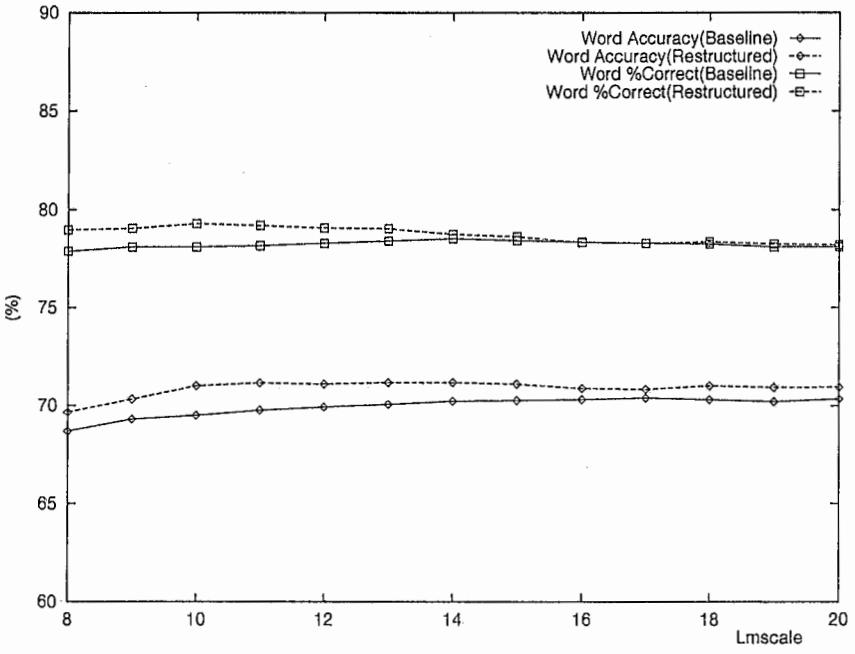

図 4.5: 単語認識結果の比較

初期モデルに対する最適設定での認識率の最大値と、そのときの再構成後のモデルの認識率を表 4.2に示す。またこの時の対話毎の認識率を付録 B に示す。

表 4.2: 認識率の最大値 (%)

	Accuracy	%Correct
初期モデル	70.38	78.53
再構成後のモデル	70.82	78.75

4.3.3 音素認識実験

音素認識では SSS の場合と同様にあまり効果はなく、初期モデルにおけるピークにおいて、再構成後の方が認識率が低くなるという結果になった。

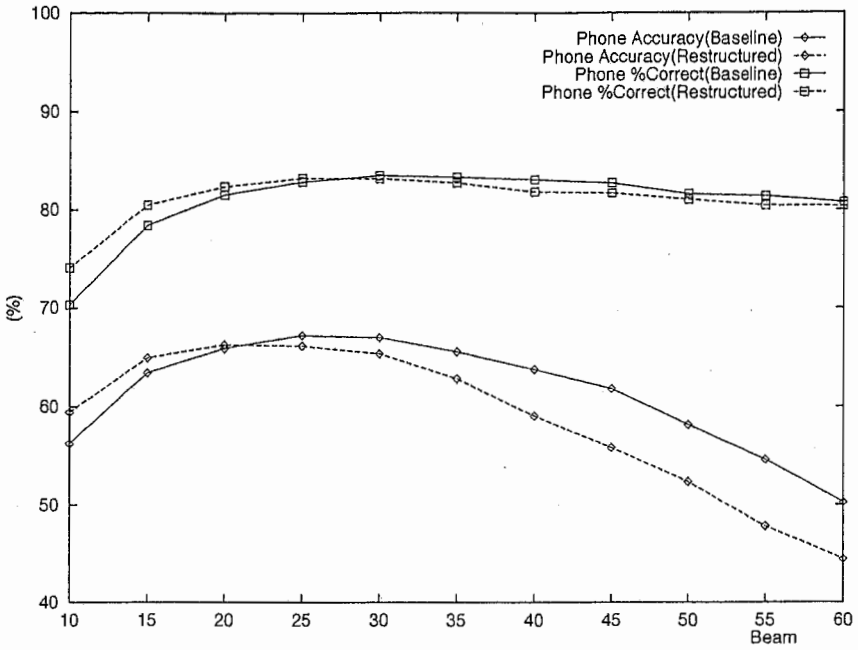

図 4.6: 音素認識結果の比較

第 5 章

おわりに

5.1 まとめ

音響モデルの高度化を目的としたガウス混合分布再構成法において、その効果の適用領域を調べるため種々の条件下で実験を行なった。

まず性別非依存のモデルに対して本手法を適用し、認識率が向上することを確認した。即ち、本手法は既に効果が確認されている男性話者モデルのみではなく、性別非依存のモデルに対しても有効であると言える。

次に異なる初期モデル生成アルゴリズムにおいて、本手法の効果を確認するため、Decision tree clustering によるモデルを用い、実験を行なった。ML-SSS による音響モデルにおいては、HMnet 作成過程の逐次状態分割により表現不能になった音響現象を、再び表現可能にすることができたと考えられる。Decision tree clustering は、SSS とは逆にあらゆる状態を融合していく方法である。このアルゴリズムで生成されたモデルに対する実験においても、再構成による認識率の向上が見られた。このことから、本手法が必ずしも分割型のアルゴリズムによるモデルに対してのみ有効な手法ではなく、融合型のアルゴリズムによるモデルに対しても有効であることが確認された。但し SSS によるモデルに比べて、Decision tree clustering によるモデルに対する認識率の向上の度合は小さかったことから、本手法は分割型のアルゴリズムによるモデルに対してより有効であると考えられる。

また音素認識実験においてはあまり効果が得られなかったが、本手法では追加したコンポーネントを他の音素モデルのガウス混合分布と共有することになるので、共有した音素の識別は困難になる。単語認識においては、言語的制約から多くの場合この問題が認識率にあまり影響を及ぼさないが、音素認識においては識別困難になった音素が認識率の向上を妨げたと思われる。音素認識率に改善が見られないのにもかかわらず、単語認識率に改善が見られたということは、ガウス混合分布再構成法が、強い言語制約の下での音声認識において効

果のある手法であるということを示している。

5.2 今後の課題

閾値を大きくするとコンポーネントと追加が起こりにくくなるので、認識率は落ちると考えられたが、実際は一度下がった後再び上昇するという傾向がみられた。閾値の変化に対する本手法の効果の変化は予想以上に複雑であり、さらに詳細な実験が必要であると考えられる。

謝辞

本研究を進めるにあたり、御指導いただいた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深く感謝致します。また同じ実習生として共に学び、励ましあった杉村耕司君どうもありがとう。さらに実務訓練の機会を与えて下さった早稲田大学の白井克彦教授及び ATR 音声翻訳通信研究所の山本誠一社長に心から感謝致します。

参考文献

[1] 中村篤,“ガウス混合分布の再構成による不特定話者音響モデルの改善”, 信学技報 SP97-19(1997-06)

[2] 鷹見淳一, 嵯峨山茂樹,“逐次状態分割法による隠れマルコフ網の自動生成”, 信学論 (D-II),J76-D-II,10(1993-10)

[3] 堀貴明, 加藤正治, 伊藤彰則, 好田正紀,“音素決定木に基づく逐次状態分割法による HM-Net の検討”, 信学技報 SP96-22(1996-06)

[4] M.Ostendorf,H.Singer,“HMM topology design using maxmum likelihood successive state splitting”,Computer Speech and Language 11(1997)

[5] H.Singer,M.Tonomura,Q.Huo,J.Ishii,T.Fukada,M.Schuster,“Baseline Acoustic Models for the Spoken Lnaguage Database (SDB/SLDB)”,TR-IT-0206,Interpreting Telecomunications Research Labpratories,ATR(1997-03)

[6] C.Sanderson,H.Singer,“Japanese Phoneme Recognition Experiments on ATR's Travel Task (SDB) using HTK v2.02”,TR-IT-0223,Interpreting Telecomunications Research Labpratories,ATR(1997-05)

[7] S.Young,J.Jansen,J.Odell,D.Ollason,P.Woodland,“The HTK BOOK”,Entropic Cambridge Rsearch Laboratory

[8] 清水徹, 山本博史, 政瀧浩和, 松永昭一, 匂坂芳典,“大語い連続音声認識のための単語仮説数削減”, 信学論 (D-II),J79-D-II,12(1996-12)

[9] H.Masataki,YSagisaka,“Variable order N-gram generation by word class splitting and consecutive word grouping”,Proc. ICASSP-96(1996)

付録

A　男性話者モデル、女性話者モデルへの適用実験

初期 HMM 生成アルゴリズムとして ML-SSS を用いた場合の、男性話者モデル、女性話者モデルにおける連続音声認識実験結果を以下に示す。

A.1　閾値による比較

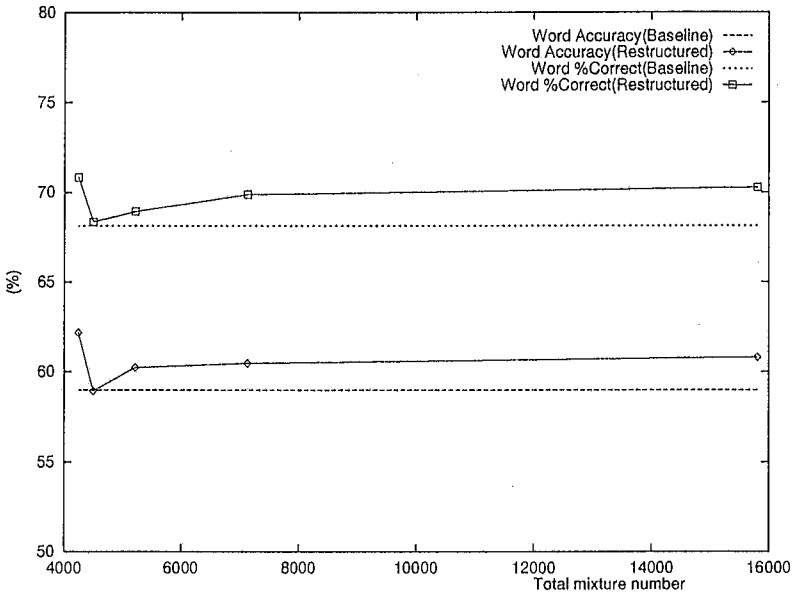

図 A.1: 男性話者

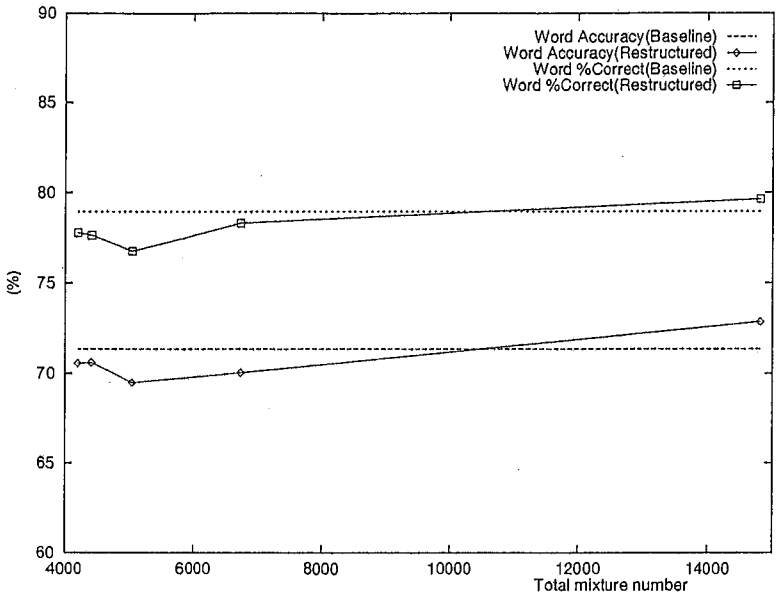

図 A.2: 女性話者

A.2 単語認識実験

男性話者: ビーム幅 = 90,90、第一パス用言語重み = 8.0

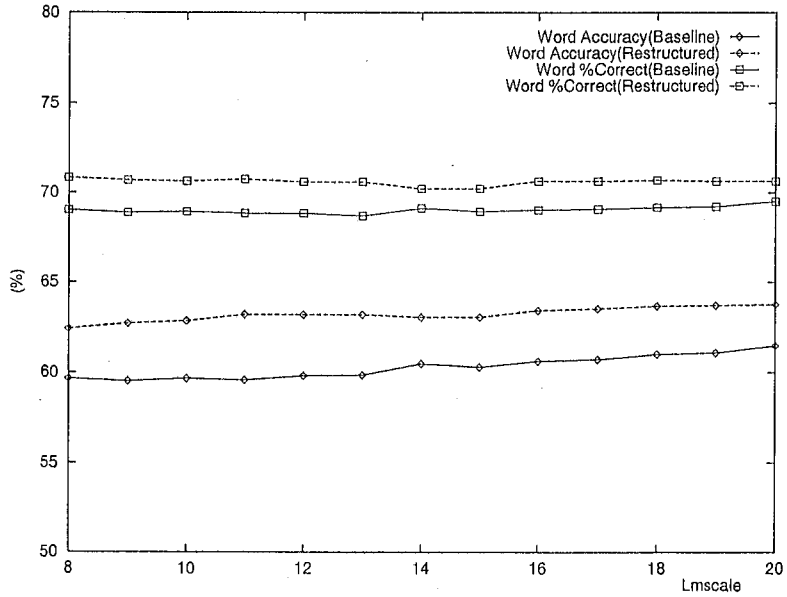

図 A.3: 男性話者

女性話者: ビーム幅 = 85,85、第一パス用言語重み = 8.0

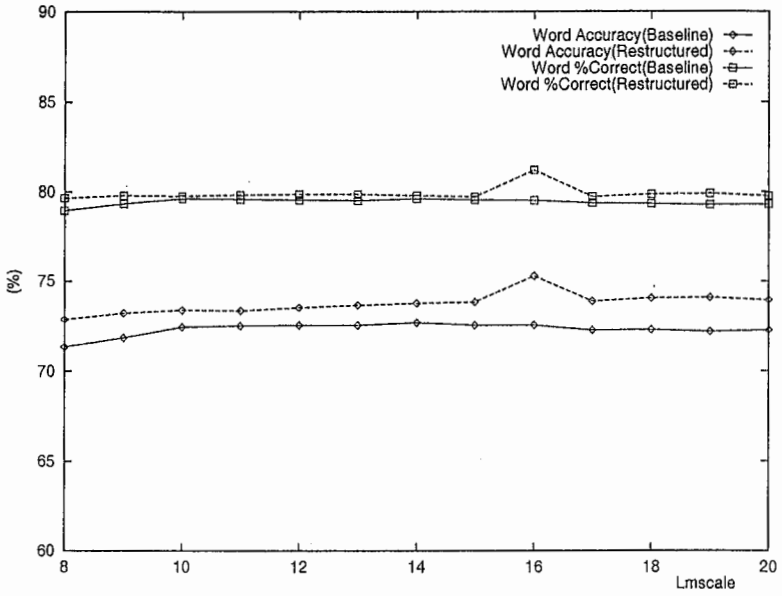

図 A.4: 女性話者

A.3 音素認識実験

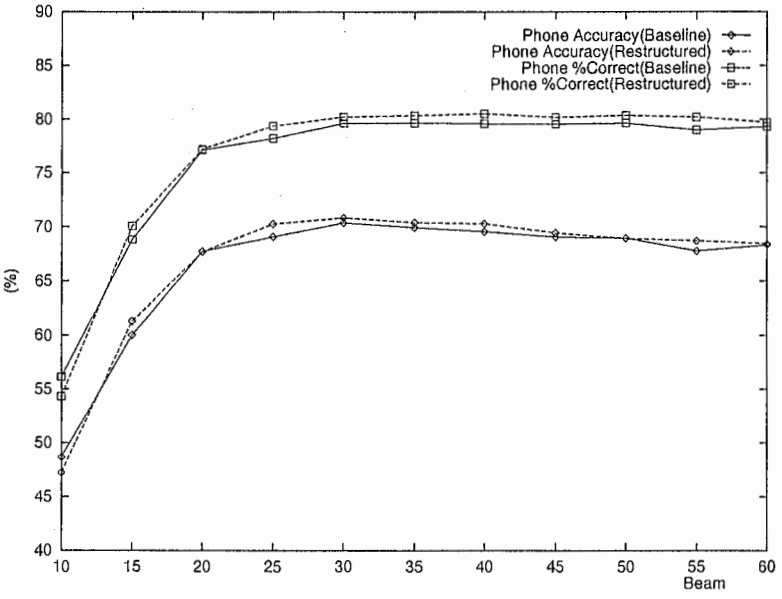

図 A.5: 男性話者

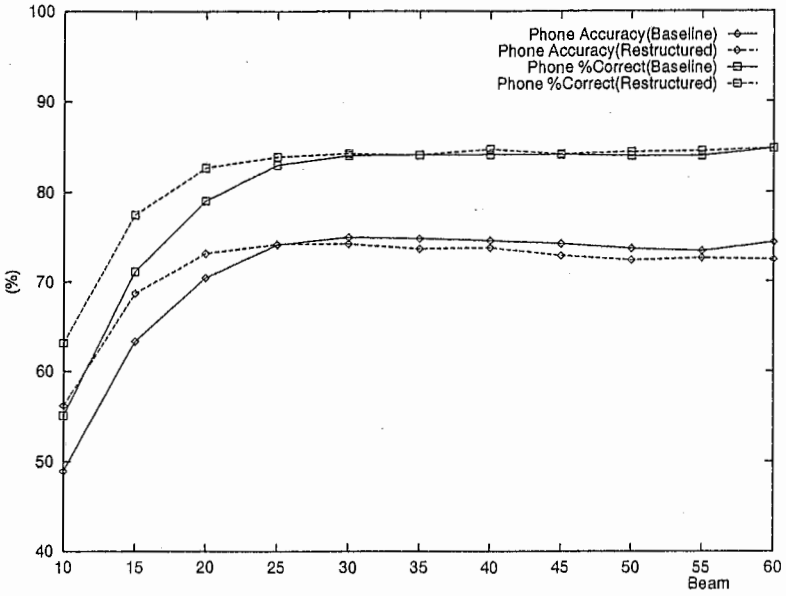
Phone Accuracy(Baseline)
Phone Accuracy(Restructured)
Phone %Correct(Baseline)
Phone %Correct(Restructured)
100
90
80
70
60
50
40
(%)
10
15
20
25
30
35
40
45
50
55
60
Beam

図 A.6: 女性話者

B　再構成による対話毎の認識率の比較

B.1　SSS によるモデル

表 B.1: 対話毎の認識率 (%)

	初期 HMM						Restructured-HMM					
Conv	Acc	Cor	Word	Ins	Del	Sbt	Acc	Cor	Word	Ins	Del	Sbt
TAC70015_A	64.40	76.27	118	14	4	24	72.88	79.66	118	8	6 f	18
TAC70016_A	60.37	65.09	106	5	15	22	59.43	62.26	106	3	15	25
TAC70017_A	84.81	89.87	79	4	2	6	82.27	88.60	79	5	3	6
TAC70019_A	79.69	80.45	133	1	9	17	78.94	80.45	133	2	9	17
TAC70021_A	69.10	80.48	123	14	6	18	70.73	82.92	123	15	2	19
TAC70022_A	54.71	59.74	159	8	16	48	59.74	69.18	159	15	6	43
TAC70023_A	72.34	82.97	141	15	0	24	63.82	81.56	141	25	0	26
TAC70101_A	86.52	90.78	141	6	0	13	91.48	93.61	141	3	1	8
TAC70102_A	78.44	88.62	167	17	0	19	80.23	86.82	167	11	1	21
TAC70103_A	69.87	72.28	83	2	7	16	71.08	74.69	83	3	5	16
TAC70201_A	76.92	82.05	156	8	2	26	74.35	82.69	156	13	1	26
TAC70202_A	51.38	55.24	181	7	26	55	53.03	56.90	181	7	25	53
TAC70203_A	78.14	83.60	183	10	3	27	75.40	80.87	183	10	2	33
TAC70301_A	76.86	81.34	134	6	5	20	82.83	85.82	134	4	5	14
TAC70303_A	87.28	88.98	118	2	3	10	82.20	84.74	118	3	6	12
TAC70304_A	54.02	59.77	87	5	5	30	63.21	71.26	87	7	1	24
TCC70103_A	64.51	80.64	93	15	3	15	70.96	84.94	93	13	2	12
TCC70109_A	63.02	70.58	119	9	7	28	60.50	64.70	119	5	8	34
TCC70201_A	37.86	43.68	103	6	16	42	40.77	47.57	103	7	12	42
TCC70212_A	36.17	45.21	188	17	19	84	46.27	52.12	188	11	16	74
TCC70307_A	67.87	70.30	165	4	7	42	68.48	72.12	165	6	3	43
TCS70004_B	56.36	58.91	314	8	40	89	56.05	57.96	314	6	37	95
TCS70010_A	34.37	43.75	128	12	22	50	41.40	55.46	128	18	15	42
TCS70013_A	71.05	84.21	114	15	2	16	71.05	82.45	114	13	3	17
TCS70020_A	72.72	81.06	132	11	2	23	76.51	83.33	132	9	0	22
TCS70023_A	64.98	73.15	257	21	11	58	68.09	78.98	257	28	9	45
TCS70025_A	80.71	82.14	140	2	8	17	82.14	84.28	140	3	6	16
TCS70028_A	73.63	83.63	110	11	3	15	80.00	84.54	110	5	2	15
TCS70034_A	55.49	62.82	191	14	14	57	56.54	65.44	191	17	15	51
TCS70047_A	61.48	69.62	135	11	10	31	66.66	71.11	135	6	11	28
TCS70055_A	56.84	64.21	190	14	15	53	62.63	66.84	190	8	24	39
TCS70059_A	79.00	83.00	100	4	2	15	78.00	83.00	100	5	3	14
TCS70070_A	51.76	60.00	85	7	7	27	50.58	60.00	85	8	8	26
TCS70074_A	75.49	81.37	102	6	4	15	78.43	83.33	102	5	3	14
TCS70082_A	57.89	68.42	171	18	5	49	57.89	73.09	171	26	2	44
Total	65.02	71.67	4946	329	300	1101	66.96	73.69	4946	333	267	1034

B.2 Decision tree clustering によるモデル

表 B.2: 対話毎の認識率 (%)

	初期 HMM						Restructured-HMM					
Conv	Acc	Cor	Word	Ins	Del	Sbt	Acc	Cor	Word	Ins	Del	Sbt
TAC70015_A	70.33	77.96	118	9	6	20	69.49	77.96	118	10	7	19
TAC70016_A	67.92	70.75	106	3	10	21	68.86	70.75	106	2	11	20
TAC70017_A	79.74	84.81	79	4	1	11	81.01	89.87	79	7	1	7
TAC70019_A	77.44	79.69	133	3	5	22	81.95	84.96	133	4	4	16
TAC70021_A	66.66	80.48	123	17	2	22	73.17	84.55	123	14	3	16
TAC70022_A	50.94	64.77	159	22	3	53	54.08	66.66	159	20	7	46
TAC70023_A	68.08	85.81	141	25	1	19	73.75	87.94	141	20	0	17
TAC70101_A	92.90	96.45	141	5	1	4	92.90	95.74	141	4	1	5
TAC70102_A	76.64	86.22	167	16	2	21	80.23	88.02	167	13	4	16
TAC70103_A	89.15	91.56	83	2	0	7	84.33	86.74	83	2	3	8
TAC70201_A	82.05	88.46	156	10	1	17	80.12	86.53	156	10	1	20
TAC70202_A	65.19	70.71	181	10	9	44	71.82	72.37	181	1	15	35
TAC70203_A	75.40	85.24	183	18	6	21	73.77	81.42	183	14	7	27
TAC70301_A	79.85	84.32	134	6	6	15	79.85	85.82	134	8	5	14
TAC70303_A	87.28	91.52	118	5	3	7	84.74	89.83	118	6	3	9
TAC70304_A	57.47	74.71	87	15	0	22	55.17	70.11	87	13	1	25
TCC70103_A	66.66	86.02	93	18	0	13	70.96	89.24	93	17	2	8
TCC70109_A	68.90	71.42	119	3	4	30	68.06	73.10	119	6	3	29
TCC70201_A	48.54	55.33	103	7	10	36	41.74	49.51	103	8	11	41
TCC70212_A	47.87	56.38	188	16	14	68	51.06	61.70	188	20	10	62
TCC70307_A	73.93	75.75	165	3	1	39	73.93	74.54	165	1	4	38
TCS70004_B	59.55	65.92	314	20	31	76	61.14	66.24	314	16	36	70
TCS70010_A	66.40	71.09	128	6	5	32	60.93	66.40	128	7	5	38
TCS70013_A	79.82	85.96	114	7	3	13	79.82	89.47	114	11	2	10
TCS70020_A	56.81	74.24	132	23	1	33	61.36	77.27	132	21	1	29
TCS70023_A	72.37	81.71	257	24	6	41	71.59	80.54	257	23	8	42
TCS70025_A	85.00	86.42	140	2	3	16	78.57	85.71	140	10	2	18
TCS70028_A	89.09	93.63	110	5	0	7	81.81	90.90	110	10	0	10
TCS70034_A	59.68	70.68	191	21	12	44	58.11	69.10	191	21	15	44
TCS70047_A	76.29	80.74	135	6	9	17	71.11	75.55	135	6	10	23
TCS70055_A	67.36	75.78	190	16	8	38	63.68	71.57	190	15	12	42
TCS70059_A	72.00	86.00	100	14	1	13	77.00	88.00	100	11	0	12
TCS70070_A	56.47	63.52	85	6	4	27	56.47	64.70	85	7	6	24
TCS70074_A	91.17	96.07	102	5	0	4	95.09	98.03	102	3	1	1
TCS70082_A	69.59	80.70	171	19	2	31	76.60	81.28	171	8	4	28
Total	70.38	78.28	4946	391	170	904	70.82	78.28	4946	369	205	869

C　初期モデル生成アルゴリズムの違いによる連続音声認識実験結果の比較

ML-SSS による初期モデルと、HTK を用いた Decision tree clustering による初期モデルの認識率の比較を示す。

C.1　単語認識実験

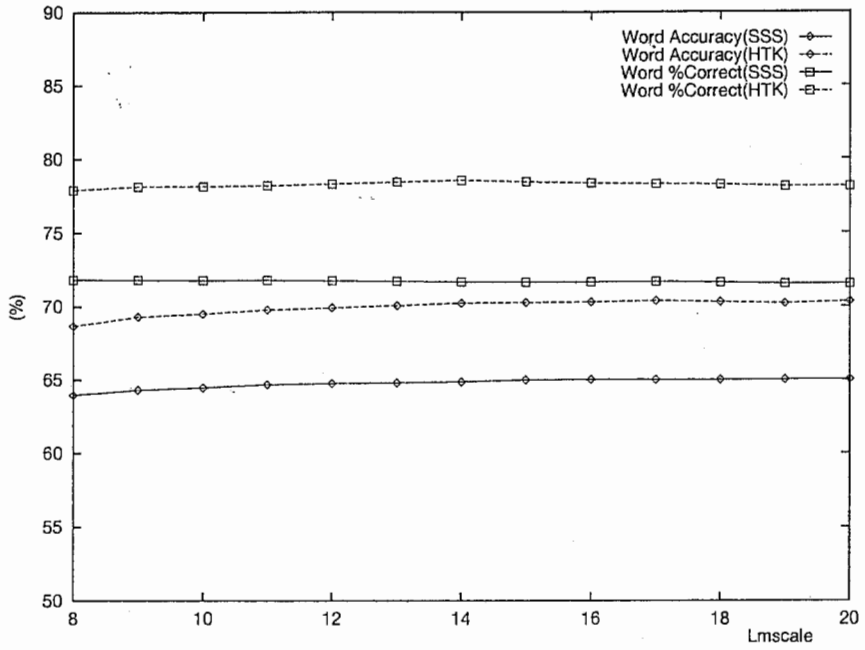

図 C.1: 単語認識結果

最適設定でのそれぞれのモデルの認識率の最大値を示す。

表 C.3: 認識率の最大値 (%)

	Accuracy	%Correct
SSS	65.00	71.84
HTK	70.38	78.53

C.2 音素認識実験

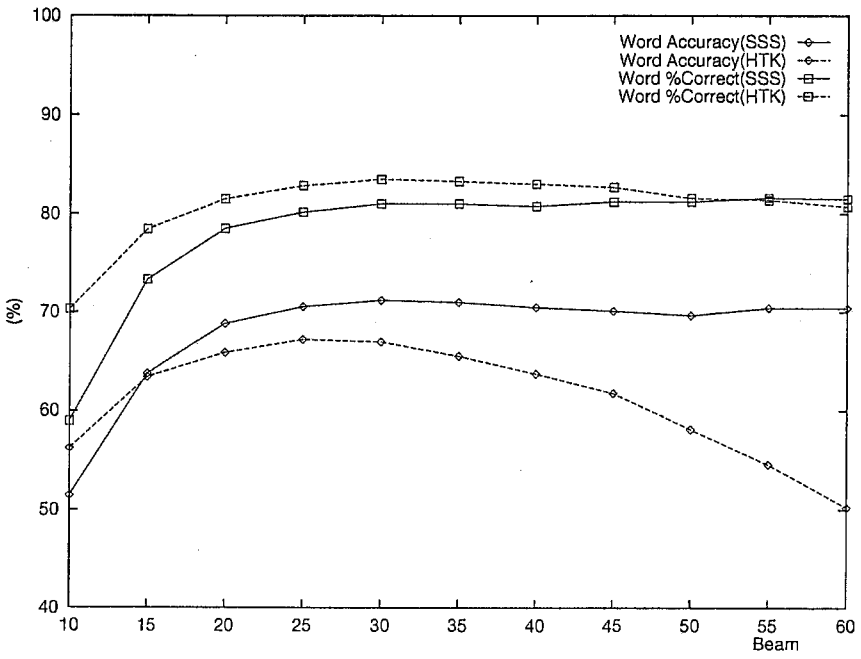

図 C.2: 音素認識結果

最適設定でのそれぞれのモデルの認識率の最大値を示す。

表 C.4: 認識率の最大値 (%)

	Accuracy	%Correct
SSS	71.18	81.62
HTK	67.18	83.30