Internal Use Only

002

TR-IT-0204

話者正規化音声データによるHMnet生成の検討

A Study on HMnet generation Using Speaker Normalized Speech Data

花井 建豪　　石井 純
Kengo Hanai　　Jun Ishii

1997.2.25

自然発話を対象とした不特定話者音声認識では，自然発話音声データは話者毎に発声内容が異なり，音素環境にばらつきが生じている．このような音声データを用いて逐次状態分割法(SSS)によって隠れマルコフ網(HMnet)を作成した場合，話者の違いの要因での状態分割と音韻コンテキストの違いによる要因の状態分割がの両方が生じる．本来HMnetは音韻コンテキストによる変動を表現するべきものであるので，話者性の違いによる分割は認識率低下の原因の一つとなっていると考えられる．本稿では重回帰写像モデルを用いた話者正規化を行ない，話者性を取り除いた音声データによってHMnetを作成することで，音韻コンテキストの違いによる変動を正確に表現することを検討し，音素認識実験によって評価した結果を示す．

©ＡＴＲ音声翻訳通信研究所

©ATR Interpreting Telecommunications Research Laboratories

もくじ

1 はじめに **1**

2 ML-SSS による HMnet の生成 **2**

1 ML-SSS 2

2 自然発話音声データを用いた場合の ML-SSS の問題点 3

3 重回帰写像モデルを用いた話者正規化 **5**

1 重回帰写像モデル 5

2 重回帰写像モデルを用いた話者正規化による HMnet の作成 9

4 実験 **12**

1 実験条件 12

2 実験結果 16

2.1 認識結果 16

2.2 allophone の数 20

3 実験結果の考察 22

5 まとめ **23**

参考文献 25

A Training Data の仕様 **26**

1 各音素の音素数と時間 26

B 音素認識実験結果 **30**

1 実験結果 30

2 allophone の数 37

第 1 章

はじめに

我々は自然発話音声データを対象にした不特定話者音声認識を，音素環境依存モデルである隠れマルコフ網 (HMnet) を基本とした音声認識システムによって検討している．

自然発話音声データの特徴としては，話者毎に発声内容が異なり，音素環境に話者毎のばらつきが生じていることが挙げられる．このような音声データにを用いて逐次状態分割法 (Successive State Splitting: SSS)[1][2] によって不特定話者の隠れ HMnet を作成した場合，音韻コンテキストの違いに起因する分割に加えて，話者間の変動によっても分割されてしまう．HMnet は音素環境依存モデルの共有化構造を表現するためのものであるので，SSS では音韻コンテキストの違いのみの影響によって状態分割を行なうべきである．従って，話者間の変動要因による状態分割は認識率低下の原因の一つになっていると考えられている．このような話者間の変動を減少させる方法として話者正規化による方法 [3][4][5][6][7][8][9][10][11] があり，不特定話者モデルの作成に対する有効性が報告されている．

本稿では重回帰写像モデルに基づいた方法 [12][13][14] によって話者正規化を行ない，話者性を取り除いた音声データによって HMnet を生成することで，音韻コンテキストの違いによる変動を正確に表現することを検討した．

以下 2 章では最尤逐次状態分割法 (maximum likelihood SSS: ML-SSS)[2] による HMnet の生成について述べ，3 章では重回帰写像モデルを用いた話者正規化の方法について述べる．さらに 4 章では，実験方法とその結果，そして考察を示し，5 章でまとめと今後の課題を述べる．

第 2 章

ML-SSS による HMnet の生成

1 ML-SSS

一般に，連続発声された音声においては，前後の音素のコンテキストのみならず，発話速度やピッチなども含めた広い意味での「音素環境」によって音素パターンの変形，変動が著しく生じる．このような音声パターンの変動は，音声認識が困難な大きな理由である．

そこで，このような音素パターンの変動の情報を積極的に音声認識に利用して，認識単位として音素環境依存の音素モデルを用いることが，認識性能の高い音声認識のためには有効である．すでに，いくつかの連続音声認識システムでは，音素環境情報を活用して認識性能を向上させており，その有効性が広く認められている．

この音素環境依存モデルをトップダウン的な知識ではなく，モデルを作成するためにデータの特徴により自動的に生成する方法として ML-SSS アルゴリズム [2] が提案されてた．本研究でもこの ML-SSS アルゴリズムによって作成された隠れマルコフ網 (HMnet) を用いている．以下に ML-SSS アルゴリズムについて概説する．

ML-SSS では音素の特徴空間上に割り当てられた確率的定常信号源（状態）の間の確率的な推移を表現した確率モデルに対して，尤度最大化の基準に基づいて個々の状態をコンテキスト方向または時間方向へ分割するといった操作を繰り返す．この操作により逐次的にモデルの精密化が行なわれる．これよりモデルの単位の決定とそのモデルの構造決定，および各状態のパラメータ推定を共通の評価基準の下で同時に実現することができる． ML-SSS における処理の流れを図 2.1に示し，この図に従って ML-SSS の原理を説明する．

まず初期モデルとして，ただ 1 つの状態と，その状態を始端から終端まで結ぶ 1 本のパスからなるモデルをすべての音声サンプルから形成し，この状態を分割することから始める．ある時点における状態の分割は，パスの分割を伴うコンテキスト方向，あるいはパスの分割を伴わない時間方向のうちのいずれか一方に関して行なわれる．特にコンテキスト方向への分割時には，パスの分割にともなってそれぞれのパスに割り当てられるコンテキストクラスも同時に分割される．実際の分割方法としては，コンテキストクラスの分割方法も含めてその時点で可能

なすべての分割法の中から，音声サンプルに適応した場合の尤度の総和が最も大きくなるものを採用する．

このような状態分割を繰り返すことによって，少ない状態数で高い尤度を達成することのできる効率の良いモデルが生成される．以上の方法で得られるモデルをここでは HMnet と呼ぶ．

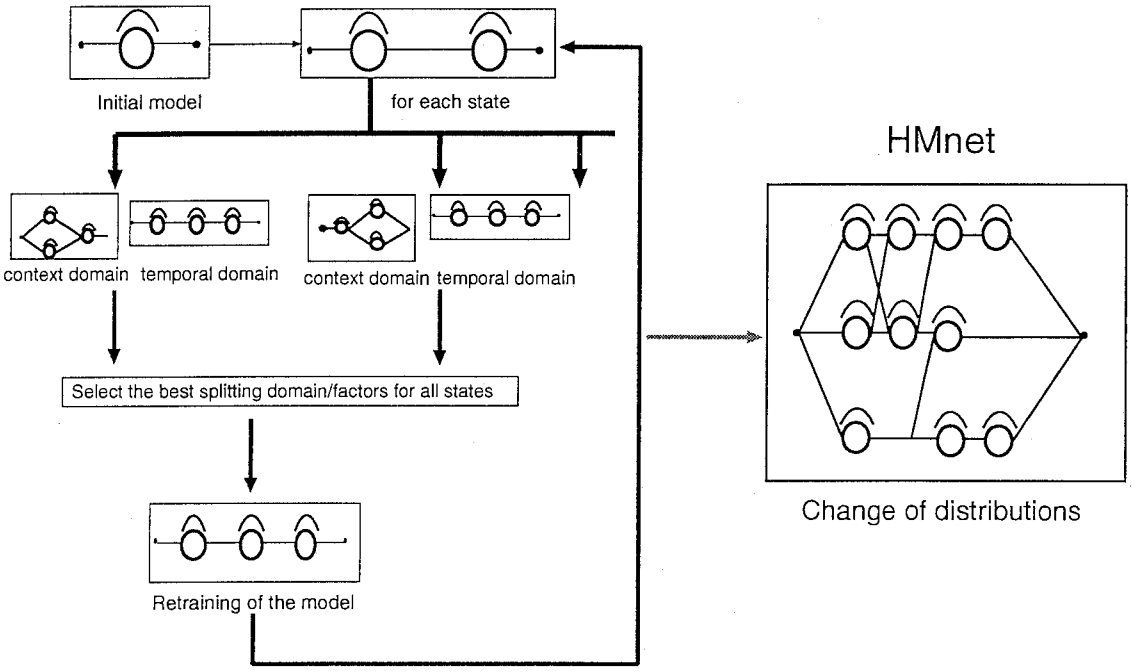


図 2.1: ML-SSS による HMnet の生成

2 自然発話音声データを用いた場合の ML-SSS の問題点

自然発話音声データは話者毎に発声内容が異なり，音素環境にばらつきが生じる．このような音声データを用いて，ML-SSS を行なった場合，音韻コンテキストによる変動と話者性の違いによる分割が生じる．本来 HMnet は音韻コンテキストの違いのみの影響によって状態分割を行なうべきである．例えば，図 2.2のように /a/k/a, o/ を表現するモデルがあった場合に，学習データでは，/a/k/a/ は，ある一人の話者である speaker 1 でしか発声されておらず，また，/a/k/o/ は，speaker 2 でしか発声されていないならば，ML-SSS では，話者 speaker 1,

speaker 2 の話者性の違いが要因となって状態分割を行なうことになる．このような話者性の違いによる分割は認識率低下の原因の一つになると考えられる．

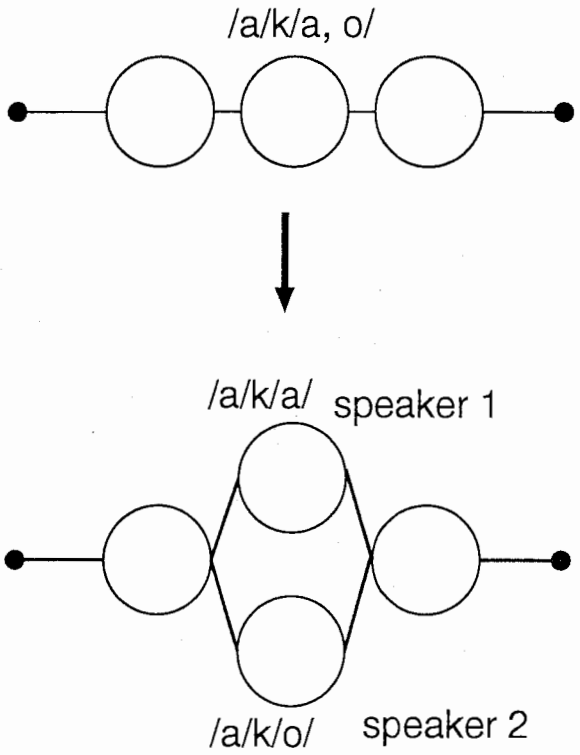


図 2.2: 話者性の違いによる状態分割

第 3 章

重回帰写像モデルを用いた話者正規化

1　重回帰写像モデル

重回帰写像モデル [12][13][14] による適応は初期モデルの k 番めのガウス分布の平均ベクトル μ_k(次元数 n) を式 (3.1) によって，適応モデルの平均ベクトル $\hat{\mu}_k$ に変換することで行なわれる．

$$\hat{\mu}_k = A_c \mu_k + b_c \tag{3.1}$$

ここで A_c は $n \times n$ の変換マトリクス，b_c は n 次元の定数項ベクトルである．次に重回帰写像モデルによる話者適応のブロック図を図 3.1に示す．

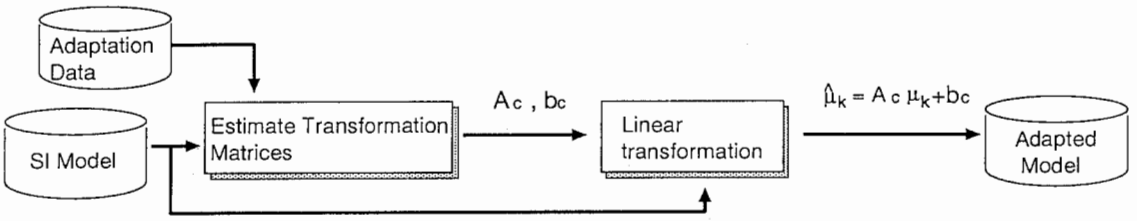


図 3.1: 重回帰写像モデルによる話者適応のブロック図

ここでは，適応後のモデルにおけるガウス分布 k の出力確率密度関数を式 (3.2) と仮定し，尤度最大を基準に A_c， b_c を推定する maximum lilelihood linear regression(MLLR)[12] を用いる．

$$b_k(\mathbf{o}_t) = \frac{1}{(2\pi)^{n/2}|\Sigma_k|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{o}_t - A_c\mu_k - b_c)^T \Sigma_k^{-1}(\mathbf{o}_t - A_c\mu_k - b_c)\right] \tag{3.2}$$

ここで， $\mathbf{o}_t$ は時間 t における観測ベクトル， Σ_k はガウス分布 k の対角共分散行列 $diag(\sigma_{k_1}^2, \ldots, \sigma_{k_n}^2)$ を表している．

A_c， b_c の推定は式 (3.3) に示されている Baum の補助関数の最大化によって行なわれる．

$$Q(\lambda, \bar{\lambda}) = \sum_{\theta \in \Theta} \mathcal{F}(O, \theta|\lambda) \log(\mathcal{F}(O, \theta|\bar{\lambda})) \tag{3.3}$$

ここで， O は適応データの特徴ベクトルの系列

$$\mathbf{o}_1, \ldots, \mathbf{o}_T \tag{3.4}$$

を， θ は状態の系列

$$\theta = (\theta_0, \theta_1, \ldots, \theta_T) \tag{3.5}$$

を表している． λ， $\bar{\lambda}$ は適応前後のモデルパラメータであり， $\mathcal{F}(O, \theta|\lambda)$， $\mathcal{F}(O, \theta|\bar{\lambda})$ は状態系列 θ に対する適応前後の尤度である． $a_{i,j}$ を状態 i から j への遷移確率， K_{θ_t} を時間 t における状態の混合成分の集合， c_k を混合成分の重みとすれば，状態系列 θ に関する尤度は下式で表される．

$$\begin{aligned} \mathcal{F}(O, \theta|\lambda) &= a_{\theta_T, N} \prod_{t=1}^{T} a_{\theta_{t-1}, \theta_t} \left[\sum_{k \in K_{\theta_t}} c_k b_k(\mathbf{o}_t) \right] & (3.6) \\ &= \sum_{k \in K_{\theta_1}} \cdots \sum_{k \in K_{\theta_T}} \left[a_{\theta_T, N} \prod_{t=1}^{T} a_{\theta_{t-1}, \theta_t} c_k b_k(\mathbf{o}_t) \right] & (3.7) \end{aligned}$$

従って，補助関数は

$$Q(\lambda, \bar{\lambda}) = \sum_{\theta \in \Theta} \mathcal{F}(O, \theta|\lambda) \log \mathcal{F}(O, \theta|\bar{\lambda}) \tag{3.8}$$

$$Q(\lambda, \bar{\lambda}) = \sum_{\theta \in \Theta} \mathcal{F}(O, \theta|\lambda) \sum_{k \in K_{\theta_1}} \cdots \sum_{k \in K_{\theta_T}} f(\bar{a}, \bar{b}, \bar{c}) \tag{3.9}$$

$$f(\bar{a}, \bar{b}, \bar{c}) = \left[\log \bar{a}_{\theta_T N} + \sum_{t=1}^{T} \log \bar{a}_{\theta_{t-1}\theta_t} + \sum_{t=1}^{T} \log \bar{b}_k(\mathbf{o}_t) + \sum_{t=1}^{T} \log \bar{c}_k \right] \tag{3.10}$$

と再定義され，変換行列の推定は出力確率にのみ依存するので

$$Q(\lambda, \bar{\lambda}) = constant + \sum_{\theta \in \Theta} \mathcal{F}(O, \theta|\lambda) \sum_{k \in K_{\theta_1}} \cdots \sum_{k \in K_{\theta_T}} \sum_{t=1}^{T} \log \bar{b}_k(\mathbf{o}_t) \tag{3.11}$$

と簡単化できる．

さらに，共有化されたガウス分布の集合 K における補助関数を Q_K は下でで定義される式となる．

$$Q_K(\lambda,\bar{\lambda}) = constant + \sum_{k\in K}\sum_{t=1}^{T}\mathcal{F}(O,k_t=k|\lambda)\log\bar{b}_k(\mathbf{o}_t) \tag{3.12}$$

今，$\gamma_k(t)$ を，時刻 t にガウス分布 k でデータが観測される期待値

$$\gamma_k(t) = \frac{1}{\mathcal{F}(O|\lambda)}\mathcal{F}(O,k_t=k|\lambda) \tag{3.13}$$

であるとすると式 (3.12) は

$$Q_k(\lambda,\bar{\lambda}) = constant + \mathcal{F}(O|\lambda)\sum_{k\in K}\sum_{t=1}^{T}\gamma_k(t)\log b_k(\mathbf{o}) \tag{3.14}$$

さらに，式 (3.2) で定義された出力確率密度関数により，

$$Q_K(\lambda,\bar{\lambda}) = constant - \frac{1}{2}\mathcal{F}(O|\lambda)\sum_{k\in K}\sum_{t=1}^{T}\gamma_k(t)[n\log(2\pi)+\log|\Sigma_k|+h(\mathbf{o}_t,k)] \tag{3.15}$$

$$h(\mathbf{o}_t,k) = (\mathbf{o}_t - A_c\mu_k - b_c)^T\Sigma_k^{-1}(\mathbf{o}_t - A_c\mu_k - b_c) \tag{3.16}$$

に書き換えられる．

補助関数が最大値を示す変換係数 A_c，b_c は補助関数を A_c，b_c で偏微分し，共有化クラス K において零とすることで得られる．

$$\begin{aligned}\frac{d}{dA_c}Q(\lambda|\bar{\lambda}) &= \mathcal{F}(O|\lambda)\sum_{t=1}^{T}\sum_{k\in K}\gamma_k(t)\Sigma_k^{-1}\\ &\times[\mathbf{o}_t-(A_c\mu_k+b_c)]\mu_k' = 0\end{aligned} \tag{3.17}$$

$$\begin{aligned}\frac{d}{db_c}Q(\lambda|\bar{\lambda}) &= \mathcal{F}(O|\lambda)\sum_{t=1}^{T}\sum_{k\in K}\gamma_k(t)\Sigma_k^{-1}\\ &\times[\mathbf{o}_t-(A_c\mu_k+b_c)] = 0\end{aligned} \tag{3.18}$$

従って，式 (3.17)，(3.18) から，変換行列 A_c の p 行目 $a_{c_{p,i}}$，及び定数項 b_c の p 番目の要素 b_{c_p} は下式で与えられる．

$$\begin{bmatrix} a_{c_{p,1}} \\ \vdots \\ a_{c_{p,n}} \\ b_{c_p} \end{bmatrix} = \begin{bmatrix} g_{1,1}^{(p)} & \cdots & g_{1,n+1}^{(p)} \\ \vdots & & \vdots \\ g_{n,1}^{(p)} & & g_{n,n+1}^{(p)} \\ g_{n+1,1}^{(p)} & \cdots & g_{n+1,n+1}^{(p)} \end{bmatrix}^{-1} \begin{bmatrix} z_1^{(p)} \\ \vdots \\ z_n^{(p)} \\ z_{n+1}^{(p)} \end{bmatrix}, \tag{3.19}$$

ここで，

$$g_{i,j}^{(p)} = \sum_{k\in K}\sum_{t=1}^{T}\gamma_k(t)\frac{\mu_{k_i}\mu_{k_j}}{\sigma_{k_p}^2}, 1\le i,j\le n \tag{3.20}$$

$$g_{i,n+1}^{(p)} = g_{n+1,i}^{(p)} = \sum_{k\in K}\sum_{t=1}^{T}\gamma_k(t)\frac{\mu_{k_i}}{\sigma_{k_p}^2}, 1\le i\le n \tag{3.21}$$

$$g_{n+1,n+1}^{(p)} = \sum_{k \in K} \sum_{t=1}^{T} \gamma_k(t) \frac{1}{\sigma_{k_p}^2} \tag{3.22}$$

$$z_i^{(p)} = \sum_{k \in K} \sum_{t=1}^{T} \gamma_k(t) \frac{\mathbf{o}_{t_p} \mu_{k_i}}{\sigma_{k_p}^2}, 1 \leq i \leq n \tag{3.23}$$

$$z_{n+1}^{(p)} = \sum_{k \in K} \sum_{t=1}^{T} \gamma_k(t) \frac{\mathbf{o}_{t_p}}{\sigma_{k_p}^{(p)}} \tag{3.24}$$

但し，μ_{k_i} は平均ベクトルの i 番目の要素，σ_{k_p} は対角共分散行列の (p,p) 要素，$\mathbf{o}_{t_p}$ は時刻 t の入力ベクトルの p 番目の要素を表している．

2 重回帰写像モデルを用いた話者正規化による HMnet の作成

不特定話者自然発話音声データに対して話者正規化を行ない，音韻コンテキストの変動要素だけを抽出し，この正規化された音声データによって HMnet の共有化構造の決定を行なうことで，HMnet の識別性能の向上を試みた．

図 3.2は提案する話者正規化方式の概念図である．また図 3.3は M 人の話者正規化音声データによる HMnet 作成のブロック図である．以下に作成手順を述べる．

1. 初期モデルを不特定話者モデルとし，各話者毎に話者適応を行なう．

2. 話者適応モデルを用い，viterbi アルゴリズムによって音声データのセグメンテーションを行なう．

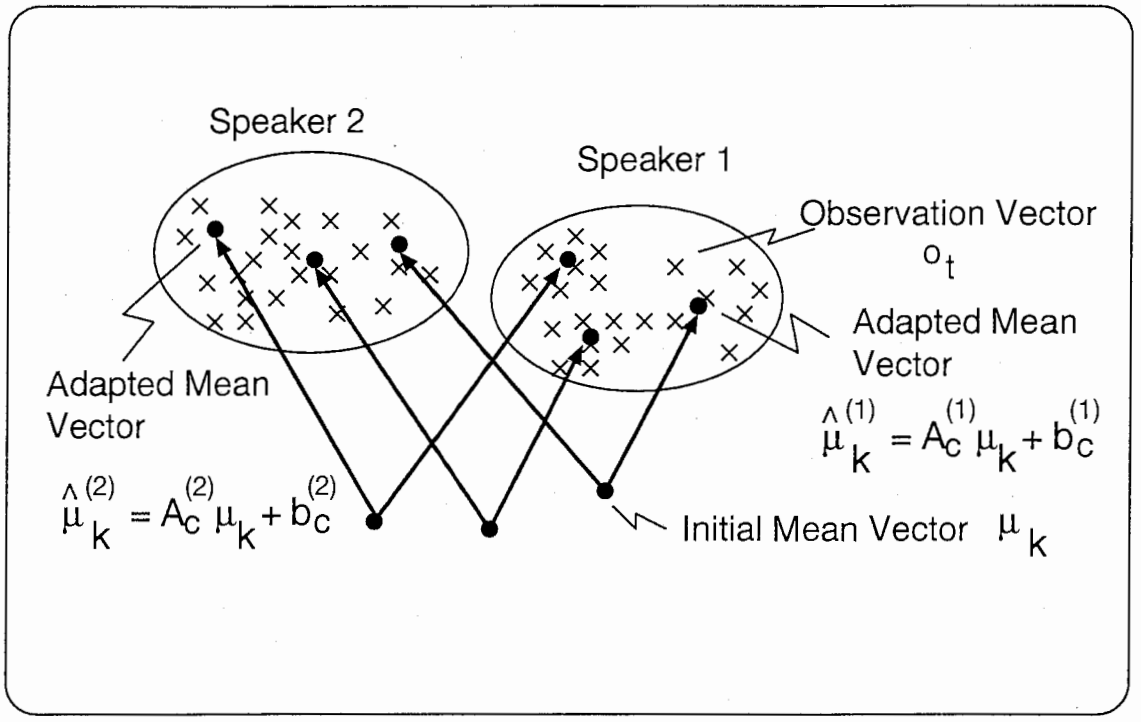
3. 下式で音声データ正規化を行なう．

$$\hat{\mathbf{o}}_t = \mathbf{o}_t^{(m)} - \hat{\mu}_k^{(m)} + \mu_k$$

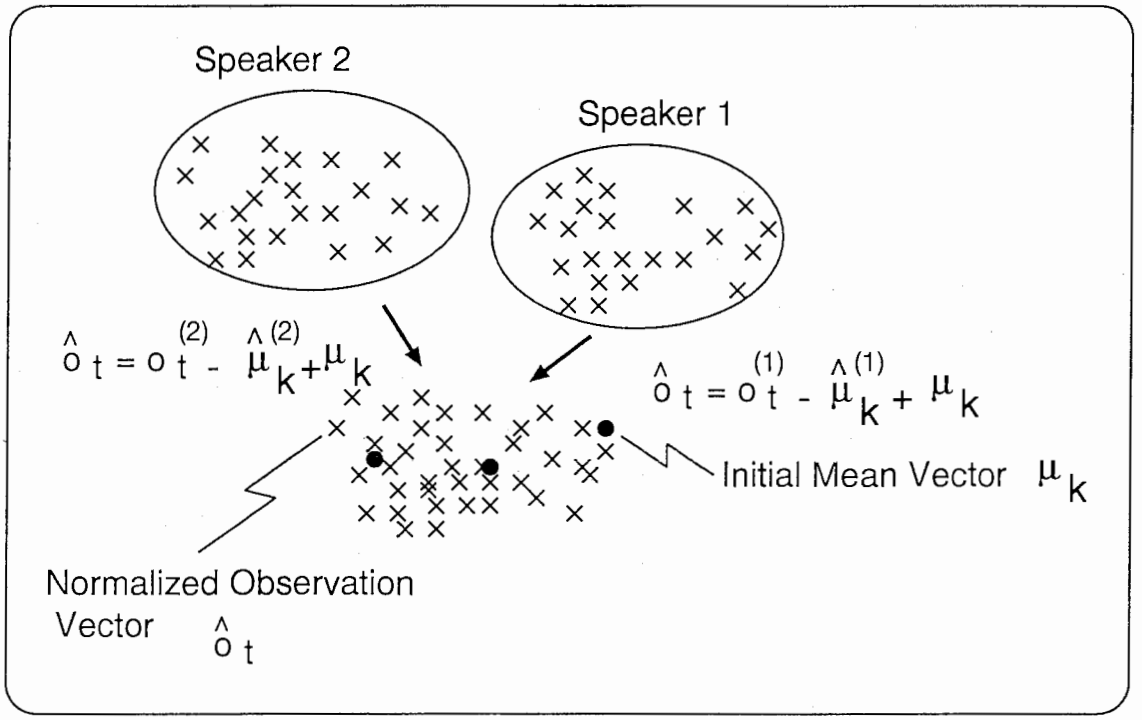
   ここで，$\hat{\mathbf{o}}_t$ は正規化音声データ，$\mathbf{o}_t^{(m)}$ は話者 m の正規化前の音声データ，μ_k，$\hat{\mu}_k^{(m)}$ は $\mathbf{o}_t^{(m)}$ が観測される適応前後のガウス分布の平均ベクトルである．

4. 正規化音声データによって ML-SSS を行ない HMnet を生成する．

5. 正規化無しの音声データを用い，連結学習によって再学習を行なう．



(a) MLLR Procedure



(b) Speaker Normalization Procedure

図 3.2: 話者正規化の概念図

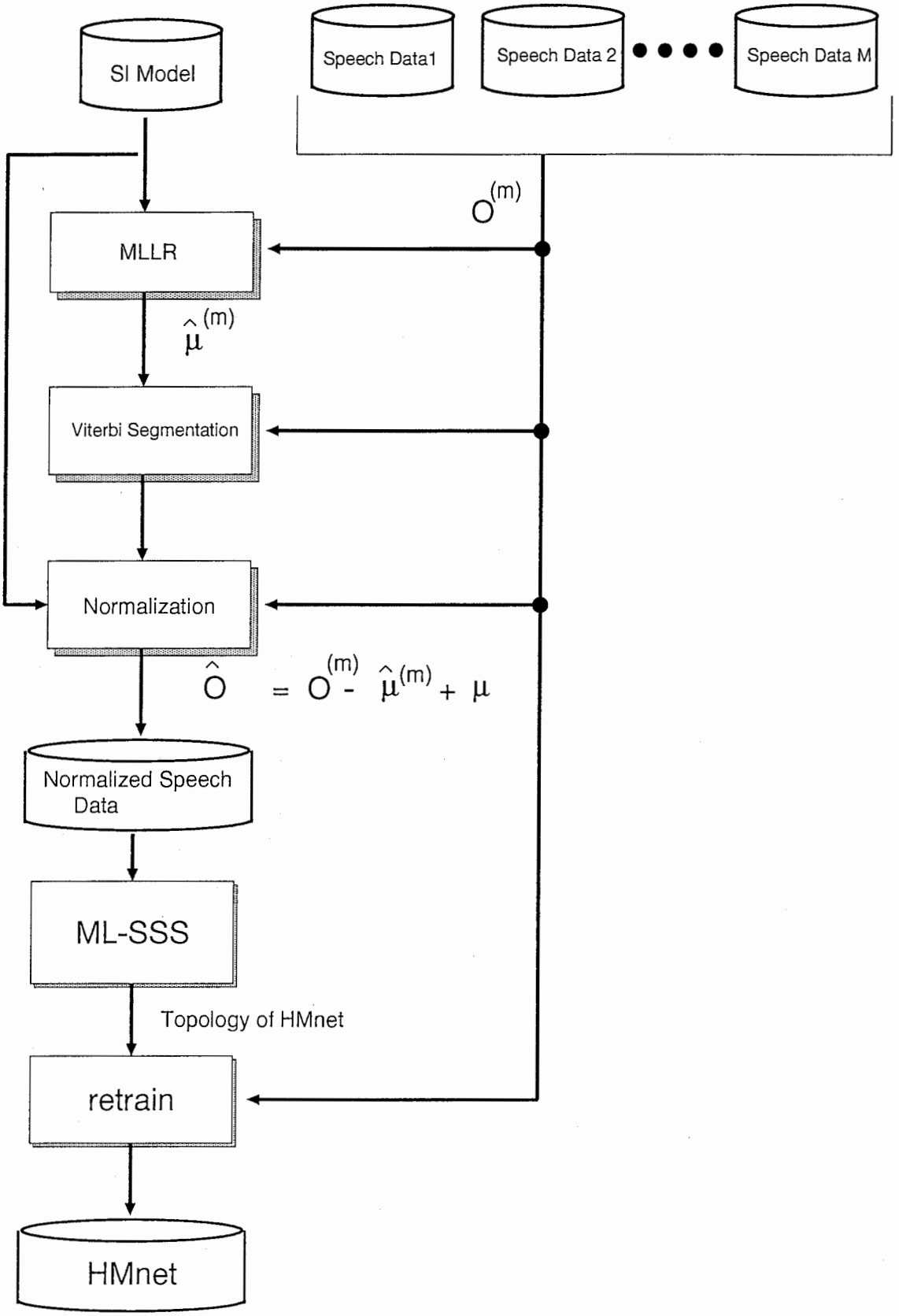


図 3.3: 話者正規化音声データによる HMnet の作成

第 4 章

実験

1　実験条件

評価実験として 26 音素を対象とした言語制約を用いない音素タイプライター型の音素認識実験を行なった．表 4.1に音響分析条件，また，使用した音声データを表 4.2に，表 4.3に全発声数，全音素数，全発声時間を示す．用いた音声データは自然発話音声データ（旅行会話タスク）である．規準モデルとして正規化なしで作成した 1000 状態 1 混合 HMnet の不特定話者 (SI) モデルと context independent (CI) モデル (1 状態 1 混合 / 音素) を使用した．この時，正規化は MLLR の行列 A の対角成分のみを用い，全てのガウス分布を共有化 (クラス数 = 1) して推定することで行なった． HMnet の初期構造は各音素 3 状態の音素環境非依存 HMM の両端の状態を共有化した 27 状態の HMnet (図 4.1) を用いた．学習話者は男性話者 99 人 (T_M_0099)，女性話者 131 人 (T_F_0131) とし， ML-SSS で生成した状態数 400， 600， 800， 1000，それぞれ混合数 1， 5(1 状態 10 混合の無音モデルを付加) の合計 8 個の HMnet に対して認識実験を行なった．認識実験は男性のみの場合，女性のみの場合，男女不特定で行なった．またテストデータは男性 18 人，女性 24 人 (S1 S2 S4) であり，学習データとの話者の重複はない．

表 4.1: 音響分析条件

サンプリング周波数	12kHz
フレーム長	20ms(ハミング窓)
フレーム周期	10ms
特徴ベクトル	16 次 LPC ケプストラム +16 次 Δ ケプストラム + 1 次 log パワー +1 次 Δlog パワー

表 4.2: 音声データ

学習データ	自然発話音声データ（旅行会話タスク）
HMnet	男性 99 人 (T_M_0099)，女性 131 人 (T_F_0131) で学習 状態数 400, 600, 800, 1000 混合数 1, 5 + 1 状態 10 混合無音モデル
正規化規準モデル	・context independent モデル (1 状態 1 混合 / 音素) ・1000 状態 1 混合 HMnet (正規化無し音声データで学習)
テストデータ	男性 18 人，女性 24 人の自然発話音声データ (テストセット: S1 S2 S4)

表 4.3: 音声データの発声数，音素数，発声時間

	Training data			Test data		
	発声数	音素数	発声時間 (sec)	発声数	音素数	発声時間 (sec)
男性	1237	48976	3292.110	224	7590	528.495
女性	1725	68057	4855.250	327	12422	885.365
合計	2962	117033	8147.360	551	20012	1413.860

正規化

- MLLR の行列 A の対角成分のみを推定し，正規化を行なっている．
- 共有化クラス = 1.

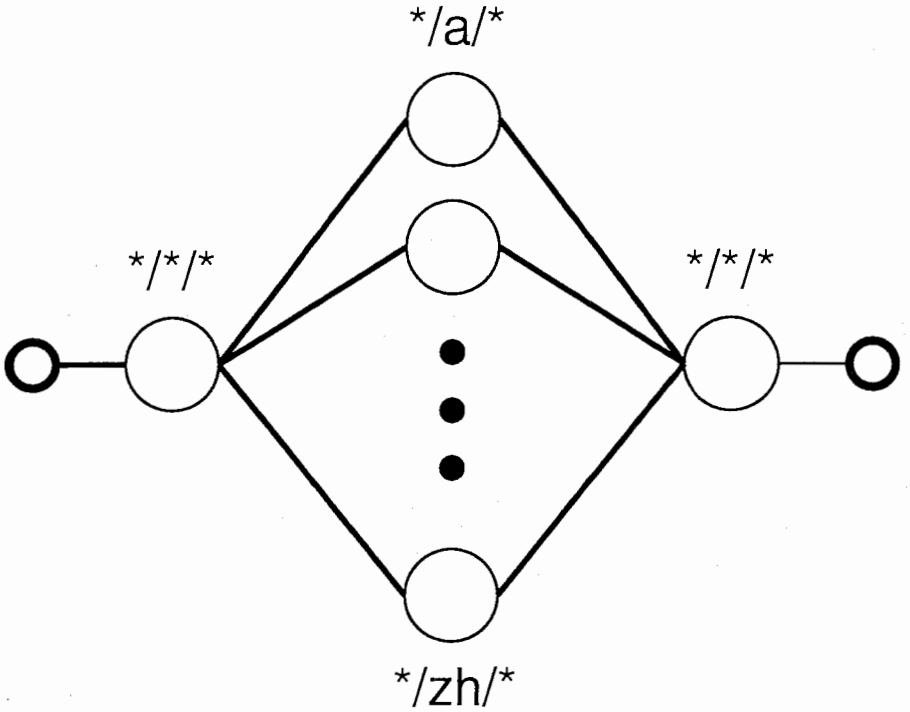


図 4.1: HMnet の初期構造

認識率は以下の式で定義する．

$$recognition\ rate = \frac{Total\ phonemes - insertion - deletion - substitution}{Total\ phonemes}$$

また，学習に用いた音声データの音素数を図 4.2に，男性と女性の音素の割合を図 4.3に示している．この図から，男女間で音素の比率はそれほど変わらないことが確認でき，発話様式は違いが少ないことが分かる．

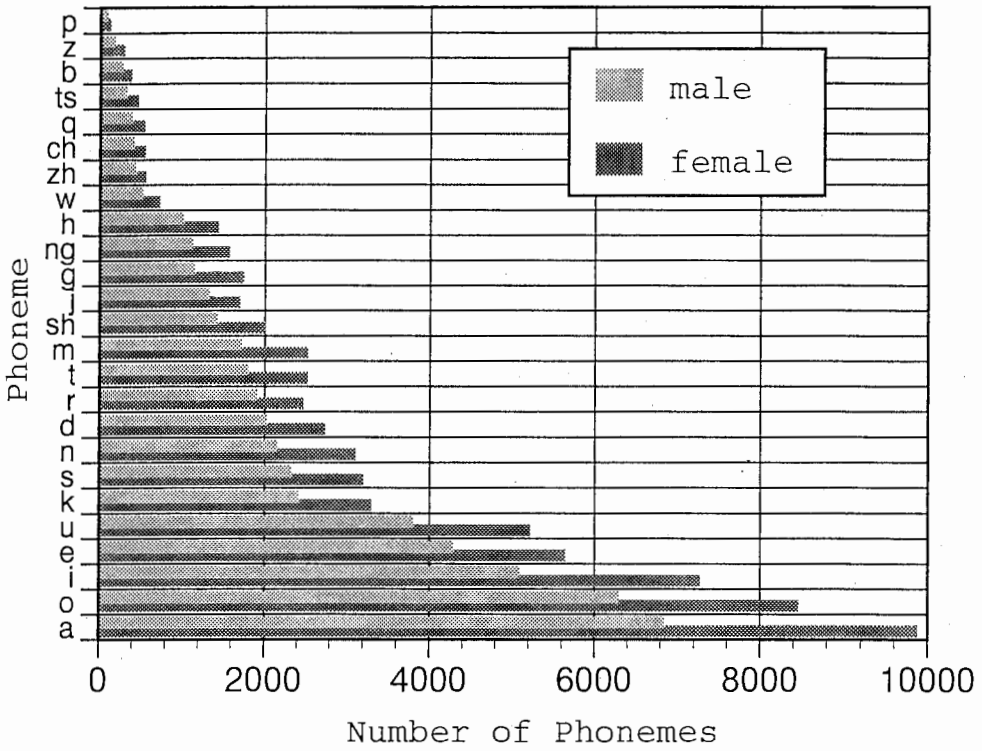


図 4.2: 音素の数

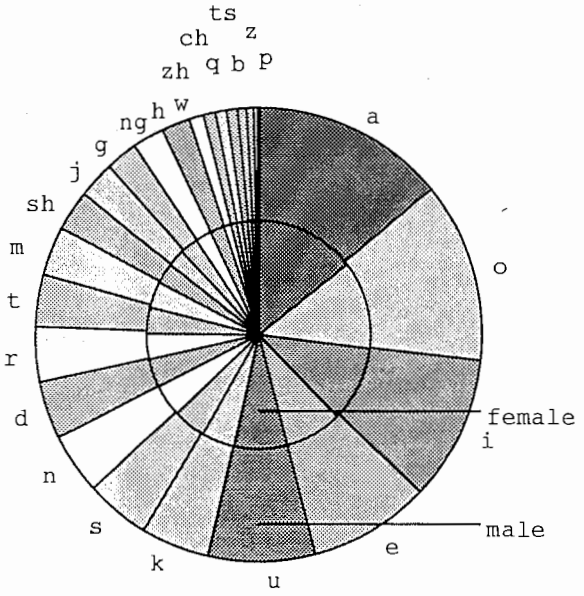


図 4.3: 音素の割合

2 実験結果

2.1 認識結果

実験結果を図 4.5～図 4.6に示す．また図の説明を図 4.4に示す．実験結果から，男性話者では認識率の向上が得られた．特に 400 状態の時はおよそ 4% もの認識率の向上が見られる．女性話者では認識率が正規化を行なうことにより逆に低下してしまった．特に 1000 状態 1 混合 HMnet 規準モデルとして正規化した場合は劣化が大きく，正規化が悪影響を及ぼしていると考えられる．男女不特定話者の場合は正規化しない場合とそれほど変わらない実験結果が出てきた．

表 4.4: 図の説明

1-best	:	尤度最大の認識結果
network	:	ビーム幅を 30 にした場合の network 内における認識結果
mixture	:	混合数
conventional	:	従来法の認識実験
Normalize with HMnet(1000states)	:	1000 状態 1 混合 HMnet を規準モデルとし正規化した場合の結果
Normalize with CI	:	context independent モデル (1 状態 1 混合 / 音素) を規準モデルとし，正規化した場合の結果

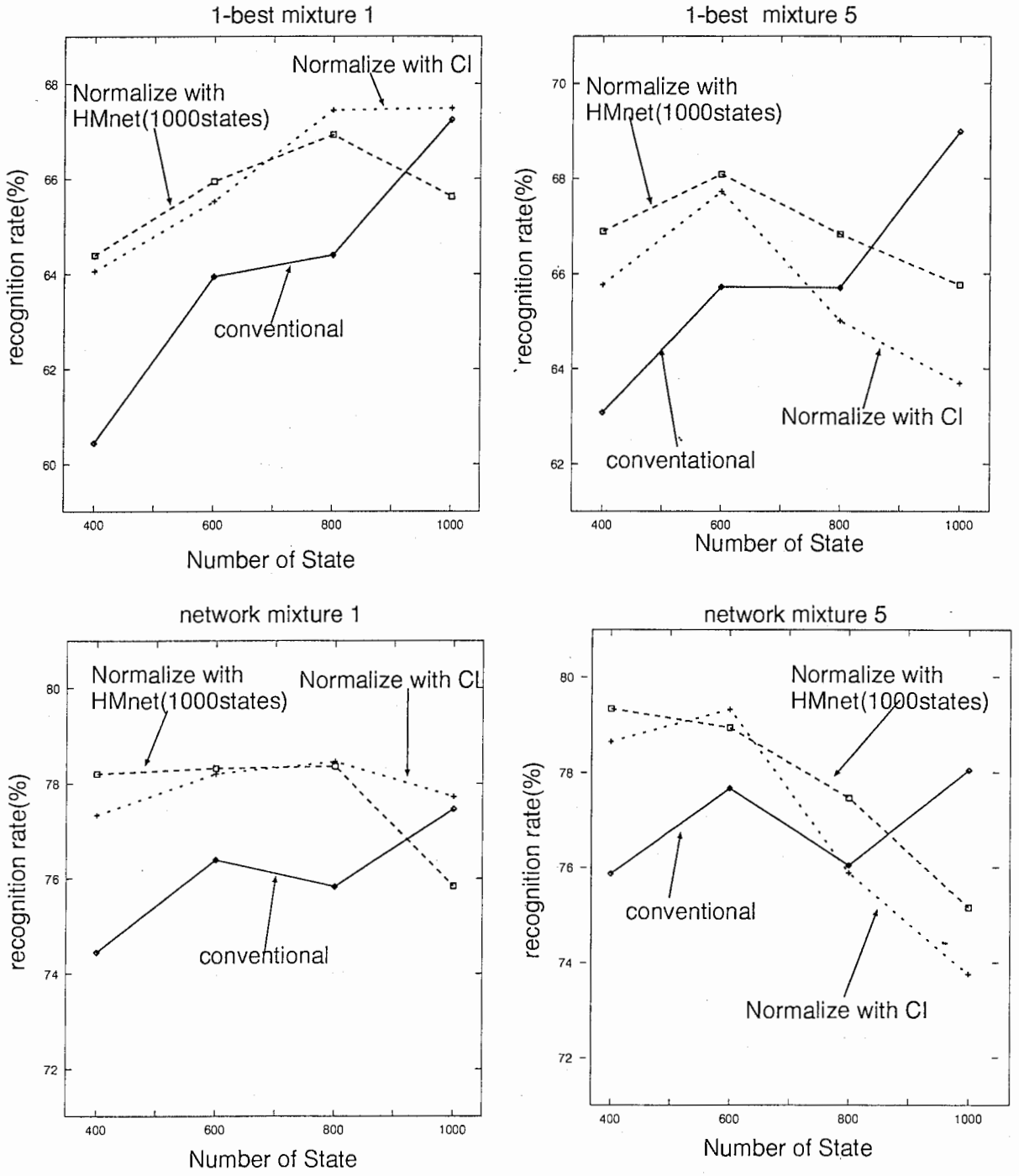


図 4.4: 男性話者の認識率

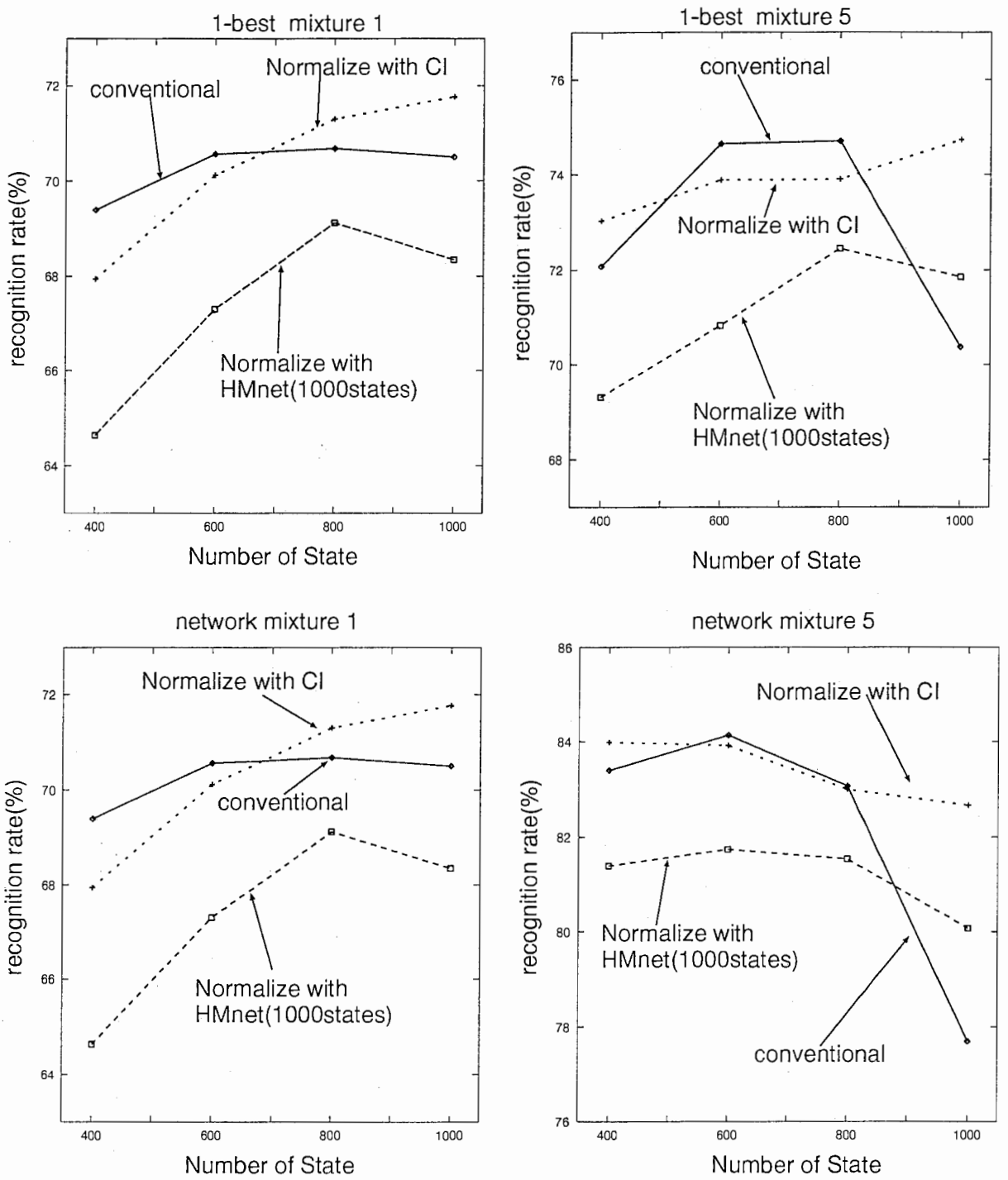


図 4.5: 女性話者の認識率

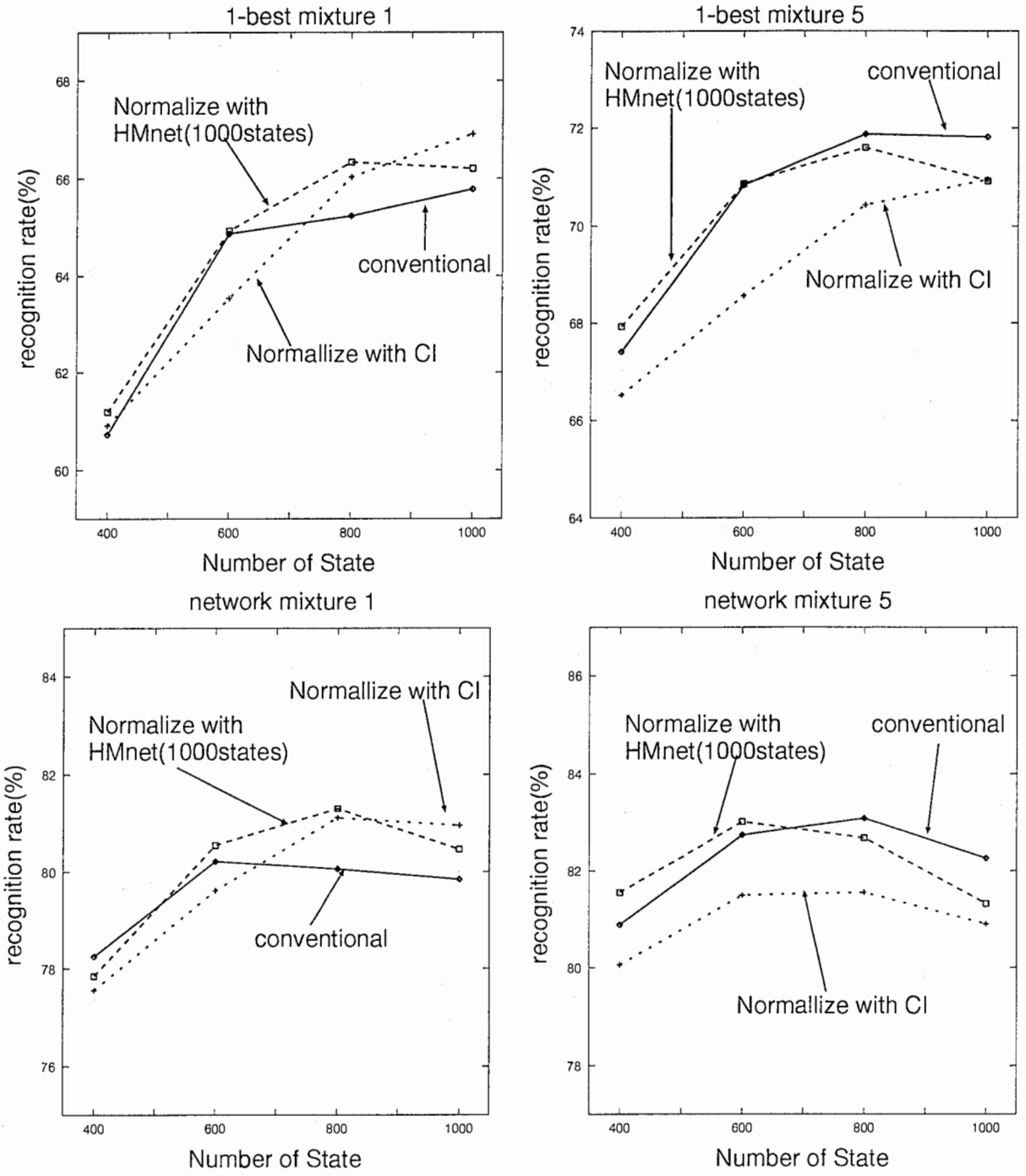


図 4.6: 男女不特定の認識率

2.2 allophone の数

次に HMnet で表現できる triphone の数である allophone の数について調べてみた.

男性話者の場合は従来法に比べ, allophone の数が増えていることがわかる. これにより, triphone の表現性が良くなり, それが認識率向上につながっていると考えられる. 女性話者の場合は allophone の数は従来法と変わらない結果となった. 男女不特定話者の場合は, allophone の数は従来法に比べて減少している. これは, 話者性による変動が少なくなったために allophone の数が減ったのだと考えられる. しかし, これは今回の実験では認識結果に反映されなかった.

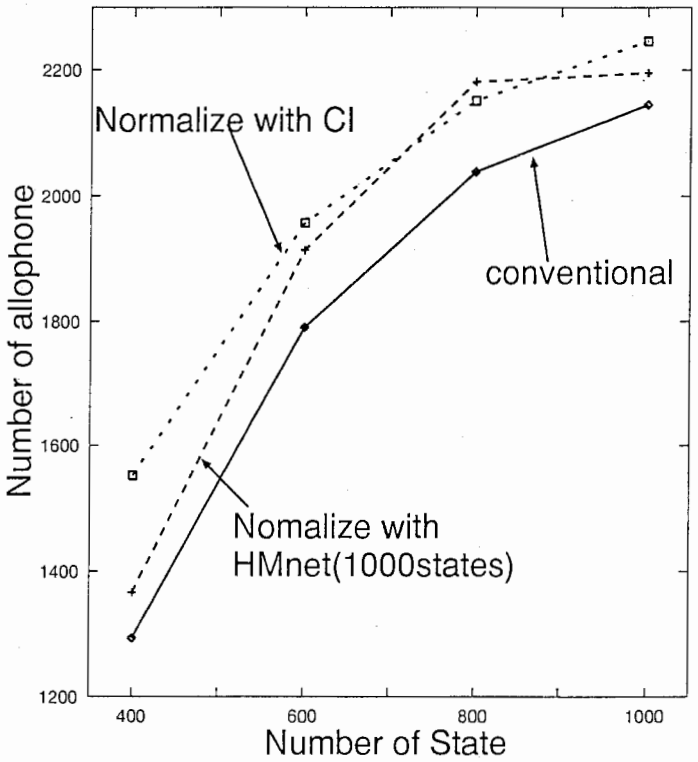


図 4.7: 男性話者 HMnet の allophone 数

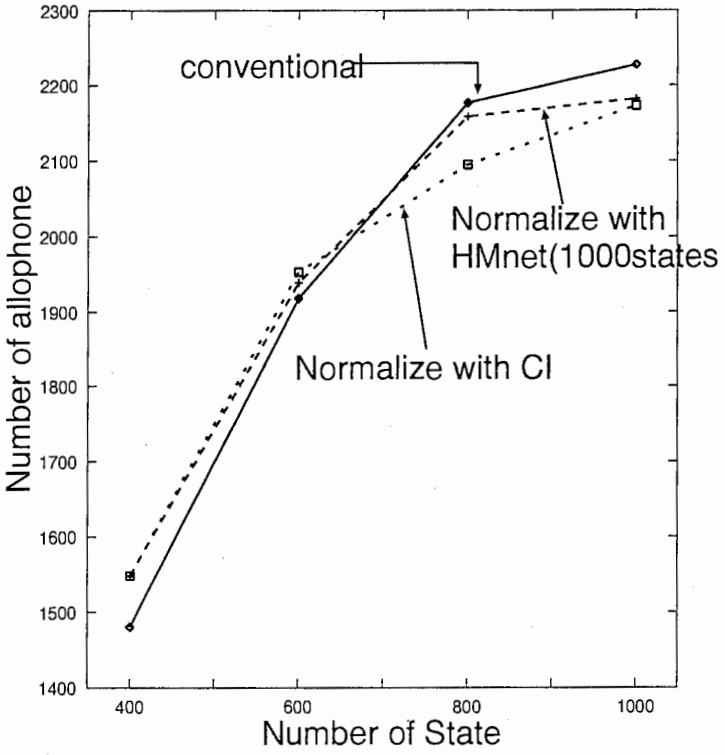


図 4.8: 女性話者 HMnet の allophone 数

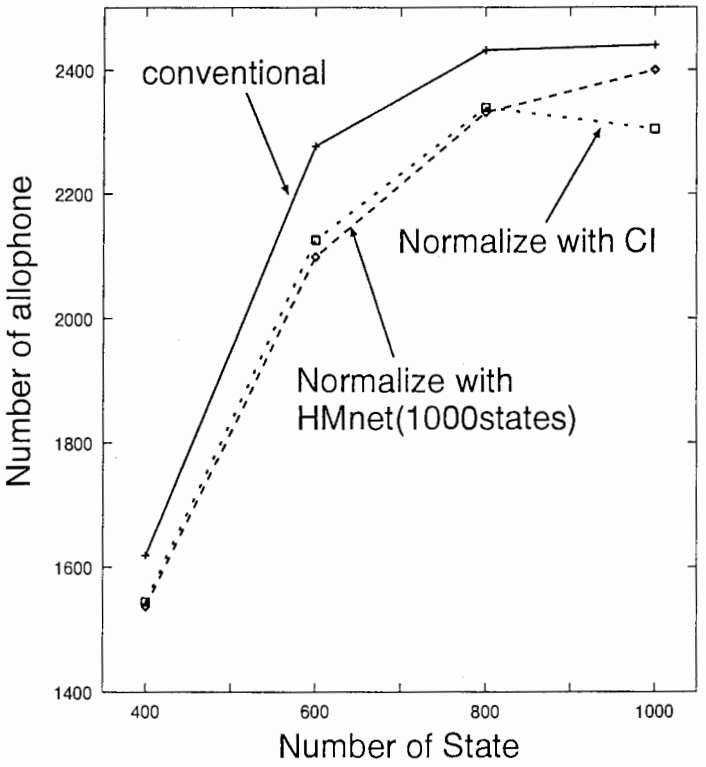


図 4.9: 男女不特定話者 HMnet の allophone 数

3 実験結果の考察

ここで今回行なった実験の結果をまとめたものを表 4.5に示す.

表 4.5: 実験結果のまとめ

	(1)	(2)	(3)
男性話者	良	良	増
女性話者	悪	悪	同
男女不特定話者	同	同	減

(1) : context independent(1 状態 1 混合 / 音素) モデル を用いた場合

(2) : 1000 状態 1 混合 HMnet を用いた場合

(3) : allophone の数

これより，男性話者は話者性の違いによる変動が大きく，正規化の効果があると考えられる．また女性話者は話者性の違いによる変動が小さく，正規化の効果がなく，むしろ，今回用いた MLLR 法は線形変換であるので悪影響を及ぼしているのではないかと考えられる．

第 5 章

まとめ

本稿では，自然発話音声データを対象とした不特定話者の HMnet 生成のために，重回帰写像モデルに基づく話者正規化方式によって学習音声データから話者性を取り除くことを検討し，音素認識実験で評価を行なった．

結果は男性話者のみで学習し，認識実験をおこなったものは認識率が向上し，正規化の有効性が現れた．女性話者のみで学習し，認識実験を行なった場合，特に 1000 状態 1 混合の HMnet を正規化したものは，従来法に比べて認識率が低下してしまい，検討の余地を残した．最後に男女不特定話者で学習し，認識実験を行なったものは従来法とあまり変わらない結果となった．今後は，共有化クラス数を増やした場合の検討が必要であると考えられる．

謝辞

本実習にあたり，親切丁寧な御指導と激励を頂きました，石井 純研究員，匂坂 芳典第一研究室室長，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 山崎 泰弘社長，豊橋技術科学大学 中川 聖一教授に感謝致します．

参考文献

[1] J. Takami and S. Sagayama: "A Successive State Splitting Algorithm for Efficient Allophone Modeling," *Proc. of ICASSP 92*, pp. 573-576 (1992)

[2] Harald Singer, Mari Ostendorf: "Maximum Liklyhood Successive State Splitting," *Proc. of ICASSP 96*, pp. 601-604 (1996)

[3] Tasos Anastasokos, John McDonough, Richard Schwartz, John Makhoul: " A Compact Model for Speaker-Adaptive Training" *Proc. of ICSLP 96*,pp.1137-1140(1996)

[4] 岩橋 直人: "適応を考慮した識別器の設計法" 音講論集 **3-2-6**(1995.9)

[5] 羅 平, 尾関 和彦: "アフィン変換を用いた音声特徴量の正規化" 音講論集 **2-8-2**(1996.3)

[6] J. Ishii and M. Tonomura "Speaker normalization and adaptation based on linear transformation," *Proc. of ICASSP 97* (to appear)

[7] Mazin G Rahim, Biing-Hwang Juang: "Signal Bias Removal for Robust Telephone Based Speech Recognition in Adverse Envionments," *Proc. of ICASSP 94*, pp. 445-448 (1994)

[8] Li Lee, Richard C. Rose: "Speaker Normalization Using Rfficient Frequency Warping Procedures," *Proc. of ICASSP 96*, pp. 352-356 (1996)

[9] Ellen Eide, Herbert Gish: "A Parametric Approch to Vocal Tract Length Normalization," *Proc. of ICASSP96*, pp. 346-348 (1996)

[10] Steven Wegmann, Don McAllaster, Jeremy Orloff, Barbara Peskin "Speaker Normalization on Conversational Telephone Speech," *Proc. of ICASSP 96*, pp. 339-341 (1996)

[11] Alejandro Acero, Xuedong Huang: "Spealer and Gender Normalization for Continuous-Densuty Hidden Markov Models," *Prov. of ICASSP 96*, pp. 342-345 (1996)

[12] C. L. Leggetter and P. C. Woodland: "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, Vol. 9, pp. 171-185 (1995)

[13] S. J. Cox and J. S. Bridle: "Unsupervised speaker adaptation by probabilistic spectrum fitting," *Proc. of ICASSP 89*, pp. 294-297 (1989)

[14] 茂木, 柴田, 堀内, 松本："制限付重回帰モデルによる話者適応の検討," 音学講論, **1-5-22** (1996.3)

付録 A

Training Data の仕様

1　各音素の音素数と時間

表 A.1～表 A.3に今回用いた音声データの仕様を示す.

表 A.1: 男性の音声データ (T_M_0099) の仕様

	Training Set 1237 utterances				Test Set 224 utterances			
音素	数	(%)	時間 (sec.)	(%)	数	(%)	時間 (sec.)	(%)
a	6830	(13.95)	554.595	(16.85)	1044	(13.75)	84.790	(16.04)
i	5073	(10.36)	298.700	(9.07)	807	(10.63)	50.060	(9.47)
u	3785	(7.73)	208.395	(6.33)	611	(8.05)	36.930	(6.99)
e	4275	(8.73)	346.905	(10.54)	688	(9.06)	60.260	(11.40)
o	6276	(12.81)	434.770	(13.21)	880	(11.59)	71.665	(13.56)
ng	1107	(2.26)	82.430	(2.50)	193	(2.54)	13.425	(2.54)
k	2411	(4.92)	161.400	(4.90)	353	(4.65)	22.510	(4.26)
j	1308	(2.67)	70.280	(2.13)	173	(2.28)	9.470	(1.79)
zh	426	(0.87)	33.290	(1.01)	73	(0.96)	5.320	(1.01)
z	161	(0.33)	13.030	(0.40)	24	(0.32)	1.805	(0.34)
d	2016	(4.12)	87.750	(2.67)	309	(4.07)	12.900	(2.44)
m	1705	(3.48)	97.975	(2.98)	269	(3.54)	15.495	(2.93)
g	1140	(2.33)	52.155	(1.58)	205	(2.70)	10.005	(1.89)
ch	402	(0.82)	34.725	(1.05)	77	(1.01)	6.680	(1.26)
r	1901	(3.88)	76.300	(2.32)	294	(3.87)	11.315	(2.14)
sh	1409	(2.88)	99.575	(3.02)	211	(2.78)	15.335	(2.90)
ts	311	(0.64)	28.675	(0.87)	63	(0.83)	5.810	(1.10)
s	2313	(4.72)	240.455	(7.30)	377	(4.97)	39.085	(7.40)
b	257	(0.52)	16.410	(0.50)	37	(0.49)	2.855	(0.54)
q	374	(0.76)	37.745	(1.15)	69	(0.91)	7.045	(1.33)
t	1790	(3.65)	110.140	(3.35)	277	(3.65)	15.700	(2.97)
w	503	(1.03)	25.810	(0.78)	72	(0.95)	3.995	(0.76)
n	2136	(4.36)	100.475	(3.05)	335	(4.41)	15.625	(2.96)
p	66	(0.13)	5.345	(0.16)	3	(0.04)	0.170	(0.03)
h	1001	(2.04)	74.780	(2.27)	146	(1.92)	10.245	(1.94)
total	48976		3292.110		117033		8147.360	

表 A.2: 女性の音声データ (T_F_0131) の仕様

	Training Set 1725 utterances				Test Set 327 utterances			
音素	数	(%)	時間 (sec.)	(%)	数	(%)	時間 (sec.)	(%)
a	9862	(14.49)	913.295	(18.81)	1802	(14.51)	171.950	(19.42)
i	7254	(10.66)	432.160	(8.90)	1325	(10.67)	78.035	(8.81)
u	5205	(7.65)	333.475	(6.87)	919	(7.40)	53.835	(6.08)
e	5628	(8.27)	476.675	(9.82)	988	(7.95)	84.300	(9.52)
o	8442	(12.40)	633.490	(13.05)	1615	(13.00)	119.495	(13.50)
ng	1560	(2.29)	110.220	(2.27)	253	(2.04)	17.750	(2.00)
k	3285	(4.83)	227.435	(4.68)	620	(4.99)	43.085	(4.87)
j	1685	(2.48)	85.575	(1.76)	339	(2.73)	16.910	(1.91)
zh	541	(0.79)	41.895	(0.86)	113	(0.91)	7.555	(0.85)
z	277	(0.41)	19.875	(0.41)	58	(0.47)	4.035	(0.46)
d	2730	(4.01)	119.700	(2.47)	483	(3.89)	21.465	(2.42)
m	2514	(3.69)	167.665	(3.45)	451	(3.63)	30.625	(3.46)
g	1731	(2.54)	82.630	(1.70)	286	(2.30)	12.765	(1.44)
ch	531	(0.78)	49.625	(1.02)	104	(0.84)	9.135	(1.03)
r	2459	(3.61)	89.580	(1.85)	446	(3.59)	15.745	(1.78)
sh	1992	(2.93)	160.495	(3.31)	380	(3.06)	30.090	(3.40)
ts	449	(0.66)	40.560	(0.84)	83	(0.67)	7.115	(0.80)
s	3189	(4.69)	318.205	(6.55)	581	(4.68)	59.205	(6.69)
b	362	(0.53)	20.540	(0.42)	55	(0.44)	3.835	(0.43)
q	523	(0.77)	44.660	(0.92)	99	(0.80)	8.465	(0.96)
t	2511	(3.69)	157.525	(3.24)	481	(3.87)	31.655	(3.58)
w	705	(1.04)	38.330	(0.79)	118	(0.95)	5.665	(0.64)
n	3096	(4.55)	167.740	(3.45)	568	(4.57)	32.265	(3.64)
p	105	(0.15)	7.000	(0.14)	14	(0.11)	0.785	(0.09)
h	1421	(2.09)	116.900	(2.41)	146	(1.92)	10.245	(1.94)
total	68057		4855.250		12422		885.365	

表 A.3: 男女不特定の音声データ (T_M_0099, T_F_0131) の仕様

	Training Set 2962 utterances				Test Set 551 utterances			
音素	数	(%)	時間 (sec.)	(%)	数	(%)	時間 (sec.)	(%)
a	16692	(14.26)	1467.890	(18.02)	2846	(14.22)	256.740	(18.16)
i	12327	(10.53)	730.860	(8.97)	2132	(10.65)	128.095	(9.06)
u	8990	(7.68)	541.870	(6.65)	1530	(7.65)	90.765	(6.42)
e	9903	(8.46)	823.580	(10.11)	1676	(8.37)	144.560	(10.22)
o	14718	(12.58)	1068.260	(13.11)	2495	(12.47)	191.160	(13.52)
ng	2667	(2.28)	192.650	(2.36)	446	(2.23)	31.175	(2.20)
k	5696	(4.87)	388.835	(4.77)	973	(4.86)	65.595	(4.64)
j	2993	(2.56)	155.855	(1.91)	512	(2.56)	26.380	(1.87)
zh	967	(0.83)	75.185	(0.92)	186	(0.93)	12.875	(0.91)
z	438	(0.37)	32.905	(0.40)	82	(0.41)	5.840	(0.41)
d	4746	(4.06)	207.450	(2.55)	792	(3.96)	34.365	(2.43)
m	4219	(3.60)	265.640	(3.26)	720	(3.60)	46.120	(3.26)
g	2871	(2.45)	134.785	(1.65)	491	(2.45)	22.770	(1.61)
ch	933	(0.80)	84.350	(1.04)	181	(0.90)	15.815	(1.12)
r	4360	(3.73)	165.880	(2.04)	740	(3.70)	27.060	(1.91)
sh	3401	(2.91)	260.070	(3.19)	591	(2.95)	45.425	(3.21)
ts	760	(0.65)	69.235	(0.85)	146	(0.73)	12.925	(0.91)
s	5502	(4.70)	558.660	(6.86)	958	(4.79)	98.290	(6.95)
b	619	(0.53)	36.950	(0.45)	92	(0.46)	6.690	(0.47)
q	897	(0.77)	82.405	(1.01)	168	(0.84)	15.510	(1.10)
t	4301	(3.68)	267.665	(3.29)	758	(3.79)	47.355	(3.35)
w	1208	(1.03)	64.140	(0.79)	190	(0.95)	9.660	(0.68)
n	5232	(4.47)	268.215	(3.29)	903	(4.51)	47.890	(3.39)
p	171	(0.15)	12.345	(0.15)	17	(0.08)	0.955	(0.07)
h	2422	(2.07)	191.680	(2.35)	387	(1.93)	29.845	(2.11)
total	117033		8147.360		20012		1413.860	

付録 B

音素認識実験結果

1　実験結果

insertion, deletion, substitution も記した結果を表 B.3～表 B.6に示す.

表 B.1: 男性話者認識率 (1-best)

手法	状態数	混合数	Total	ins.	del.	sub.	認識率 (%)
conventional	400	1	7988	633	681	1108	60.440661
	600		7988	621	633	1626	63.945919
	800		7988	619	630	1596	64.384076
	1000		7988	642	595	1380	67.238358
	400	5	7988	749	572	1628	63.082123
	600		7988	766	544	1428	65.723585
	800		7988	772	537	1430	65.711067
	1000		7988	793	426	1259	68.978468
Normalize with HMnet(1000states)	400	1	7988	592	740	1512	64.396595
	600		7988	609	650	1462	65.936405
	800		7988	662	567	1414	66.912869
	1000		7921	640	788	1295	65.623027
	400	5	7988	717	648	1281	66.875313
	600		7988	730	623	1197	68.077116
	800		7930	884	495	1252	66.822194
	1000		7028	721	630	1055	65.765509
Normalize with CI	400	1	7988	593	755	1523	64.058588
	600		7988	634	653	1468	65.510766
	800		7988	660	592	1349	67.438658
	1000		7988	651	599	1348	67.476214
	400	5	7988	725	657	1352	65.773660
	600		7988	743	589	1247	67.714071
	800		7756	889	516	1308	65.020629
	1000		7801	874	644	1314	63.696962

表 B.2: 男性話者認識率 (network)

手法	状態数	混合数	Total	ins.	del.	sub.	認識率 (%)
conventional	400	1	7988	399	534	1108	74.449174
	600		7988	364	479	1043	76.389584
	800		7988	372	529	1030	75.826239
	1000		7988	443	505	853	77.453681
	400	5	7988	507	432	988	75.876314
	600		7988	484	443	857	77.666500
	800		7988	534	440	940	76.039059
	1000		7988	552	364	838	78.042063
Normalize with HMnet(1000states)	400	1	7988	351	588	803	78.192288
	600		7988	366	559	807	78.317476
	800		7988	424	502	803	78.355033
	1000		7921	415	720	779	75.836384
	400	5	7988	477	502	672	79.331497
	600		7988	496	533	654	78.930896
	800		7930	610	423	754	77.465322
	1000		7028	507	581	658	75.156517
Normalize with CI	400	1	7988	351	619	841	77.328493
	600		7988	405	546	791	78.192288
	800		7988	431	498	792	78.455183
	1000		7988	412	520	848	77.716575
	400	5	7988	502	492	712	78.642964
	600		7988	502	468	682	79.318978
	800		7756	640	449	781	75.889634
	1000		7801	624	551	872	73.759774

表 B.3: 女性話者認識率 (1-best)

手法	状態数	混合数	Total	ins.	del.	sub.	認識率 (%)
conventional	400	1	13026	886	904	2197	69.391985
	600		13026	1023	757	2055	70.558882
	800		13026	1041	768	2012	70.666360
	1000		13026	1030	864	1949	70.497467
	400	5	13026	1123	686	1829	72.071242
	600		13026	1156	609	1538	74.643022
	800		13026	1217	554	1525	74.696760
	1000		12871	1120	1106	1587	70.375262
Normalize with HMnet(1000states)	400	1	13026	955	1026	2625	64.639951
	600		13026	1016	934	2309	67.303854
	800		13026	1093	817	2113	69.115615
	1000		13026	1087	872	2164	68.347920
	400	5	13026	1174	655	2168	69.315216
	600		13026	1204	666	1929	70.835253
	800		13026	1311	539	1739	72.447413
	1000		13026	1217	675	1775	71.848610
Normalize with CI	400	1	13026	951	985	2239	67.948718
	600		13026	1007	820	2066	70.113619
	800		13026	1004	762	1973	71.295870
	1000		13026	1063	693	1921	71.771841
	400	5	13026	1082	682	1749	73.030861
	600		13026	1142	596	1664	73.883003
	800		13026	1198	630	1573	73.890680
	1000		13026	1152	563	1576	74.735145

表 B.4: 女性話者認識率 (network)

手法	状態数	混合数	Total	ins.	del.	sub.	認識率 (%)
conventional	400	1	13026	526	650	1288	81.083986
	600		13026	576	610	1151	82.058959
	800		13026	643	624	1204	81.030247
	1000		13026	661	732	1183	80.224167
	400	5	13026	740	474	949	83.394749
	600		13026	747	456	863	84.139413
	800		13026	814	420	972	83.064640
	1000		12871	800	1016	1056	77.686271
Normalize with HMnet(1000states)	400	1	13026	556	854	1533	77.406725
	600		13026	636	747	1340	79.095655
	800		13026	673	695	1268	79.763550
	1000		13026	689	767	1381	78.220482
	400	5	13026	728	492	1205	81.383387
	600		13026	781	508	1090	81.736527
	800		13026	892	429	1084	81.536926
	1000		13026	834	589	1173	80.070628
Normalize with CI	400	1	13026	560	763	1215	80.515891
	600		13026	613	630	1124	81.828650
	800		13026	641	668	1111	81.421772
	1000		13026	687	585	1169	81.260556
	400	5	13026	675	466	945	83.985874
	600		13026	728	444	923	83.916782
	800		13026	775	496	945	82.987870
	1000		13026	799	464	996	82.657761

表 B.5: 男女不特定話者認識率 (1-best)

手法	状態数	混合数	Total	ins.	del.	sub.	認識率 (%)
conventional	400	1	21014	1764	1605	4885	60.721424
	600		21014	1750	1497	4135	64.871038
	800		21014	1794	1403	4109	65.232702
	1000		21014	1844	1368	3979	65.779956
	400	5	21014	1877	1443	3527	67.416960
	600		21014	1899	1234	2995	70.838489
	800		21014	1848	1191	2871	71.875892
	1000		21014	1958	1103	2860	71.823546
Normalize with HMnet(1000states)	400	1	21014	1765	1683	4708	61.187780
	600		21014	1799	1475	4098	64.918626
	800		21014	1814	1328	3933	66.331969
	1000		21014	1833	1302	3969	66.193966
	400	5	21014	1955	1409	3375	67.930903
	600		21014	1927	1266	2930	70.862282
	800		21014	1979	1165	2823	71.604645
	1000		21014	1989	1171	2950	70.924146
Normalize with CI	400	1	21014	1669	1647	4898	60.911773
	600		21014	1854	1457	4350	63.543352
	800		21014	1897	1312	3932	66.017893
	1000		21014	1851	1255	3849	66.903017
	400	5	21014	1963	1444	3630	66.512801
	600		21014	2105	1212	3287	68.573332
	800		21014	2318	1103	2792	70.433996
	1000		21014	2246	1096	2761	70.957457

表 B.6: 男女不特定話者認識率 (network)

手法	状態数	混合数	Total	ins.	del.	sub.	認識率 (%)
conventional	400	1	21014	947	1208	2416	78.247835
	600		21014	950	1054	2153	80.217950
	800		21014	1010	1098	2082	80.060912
	1000		21014	1047	1092	2097	79.842010
	400	5	21014	1191	1063	1762	80.888931
	600		21014	1169	887	1571	82.740078
	800		21014	1171	887	1496	83.087465
	1000		21014	1286	879	1563	82.259446
Normalize with HMnet(1000states)	400	1	21014	976	1301	2378	77.848101
	600		21014	996	1048	2043	80.551061
	800		21014	1004	990	1935	81.302941
	1000		21014	1033	1036	2038	80.455887
	400	5	21014	1170	1043	1663	81.555154
	600		21014	1208	903	1459	83.011326
	800		21014	1222	893	1524	82.682973
	1000		21014	1276	908	1740	81.326735
Normalize with CI	400	1	21014	895	1375	2447	77.553060
	600		21014	1045	1109	2132	79.604073
	800		21014	1031	1002	1936	81.112592
	1000		21014	1101	1001	1901	80.950795
	400	5	21014	1213	1124	1853	80.060912
	600		21014	1362	887	1639	81.498049
	800		21014	1535	832	1510	81.550395
	1000		21014	1559	877	1576	80.907966

2 allophone の数

allophone の数を音素毎に表したものを表 B.7〜表 B.9に示す.

表 B.7: allophone の数 (男性)

音素	正規化なし				1000 状態 1 混合 HMnet				CI モデル			
	状態数				状態数				状態数			
	400	600	800	1000	400	600	800	1000	400	600	800	1000
a	302	377	394	387	355	407	419	417	410	410	420	426
i	228	296	322	342	148	289	309	303	215	324	312	319
u	43	118	156	173	77	203	259	276	74	220	247	251
e	150	275	300	321	181	270	335	318	253	276	329	338
o	365	378	382	379	389	389	374	353	388	377	368	372
ng	8	57	74	75	16	54	92	95	24	53	95	102
k	20	29	31	48	42	43	48	47	45	47	48	48
j	26	35	35	34	20	24	28	31	15	33	37	38
zh	2	5	14	21	2	3	12	14	2	3	12	14
z	1	3	10	16	1	3	5	18	1	5	6	18
d	14	21	23	23	6	21	23	23	6	16	16	23
m	22	30	36	36	18	28	35	35	18	28	28	29
g	9	13	28	32	6	12	24	30	6	10	22	29
ch	2	3	6	18	2	6	12	12	2	6	12	12
r	25	31	33	36	12	31	31	39	10	27	34	36
sh	14	14	16	16	12	14	16	16	8	15	16	16
ts	1	2	3	5	1	1	4	8	1	1	3	3
s	24	28	28	28	26	28	28	28	28	28	28	28
b	1	2	20	24	1	5	24	24	1	4	20	24
q	3	12	32	32	1	7	12	17	7	8	22	26
t	9	24	24	25	15	24	28	28	9	18	22	24
w	4	6	6	6	4	6	6	6	4	5	5	6
n	15	18	39	38	16	29	39	37	15	30	36	37
p	1	1	1	1	1	1	2	2	1	2	2	2
h	5	11	25	29	15	15	15	18	10	11	11	25
total	1294	1789	2038	2145	1548	1939	2158	2181	1553	1957	2151	2246

表 B.8: allophone の数 (女性)

	正規化なし				1000 状態 1 混合 HMnet				CI モデル			
	状態数				状態数				状態数			
音素	400	600	800	1000	400	600	800	1000	400	600	800	1000
a	353	409	457	446	360	410	416	411	340	373	407	396
i	177	272	364	355	232	288	384	383	249	311	367	356
u	126	205	228	268	82	169	236	241	85	180	207	266
e	313	331	343	342	312	324	345	325	277	338	307	313
o	314	364	385	376	361	390	368	365	372	411	387	380
ng	16	32	51	73	16	51	65	78	16	34	66	83
k	36	36	41	45	18	28	31	32	42	42	42	43
j	14	31	37	37	15	31	34	34	16	24	36	36
zh	4	6	14	14	3	7	14	14	1	6	12	12
z	1	4	8	8	1	4	10	10	1	3	8	12
d	14	21	24	2	10	21	24	24	16	24	24	24
m	15	21	22	26	15	19	22	29	18	22	26	29
g	9	24	26	26	12	30	33	33	15	28	34	34
ch	1	4	5	6	1	4	5	6	1	3	5	5
r	2	42	42	42	7	34	35	37	4	30	31	33
sh	8	14	16	16	10	16	16	16	8	14	14	14
ts	1	1	1	1	1	1	1	2	1	1	1	1
s	28	32	32	31	31	31	31	31	28	32	31	31
b	1	4	6	12	1	4	6	12	1	4	6	12
q	4	14	14	14	3	9	12	20	3	12	12	15
t	4	10	12	14	15	20	20	23	15	19	20	21
w	4	4	6	6	3	6	6	6	4	4	6	7
n	27	25	27	28	28	28	28	28	24	25	29	29
p	1	1	1	1	1	1	1	1	1	1	1	1
h	8	10	15	16	10	13	15	20	10	12	16	19
total	1620	2278	2431	2439	1548	1939	2158	2181	1548	1953	2095	2172

表 B.9: allophone の数 (男女不特定)

音素	正規化なし				1000 状態 1 混合 HMnet				CI モデル			
	状態数				状態数				状態数			
	400	600	800	1000	400	600	800	1000	400	600	800	1000
a	412	467	458	458	363	434	472	455	383	448	453	441
i	179	343	382	377	222	332	357	353	174	340	359	347
u	97	261	292	294	92	193	248	302	120	173	205	207
e	321	381	402	381	224	362	390	379	254	375	399	366
o	378	448	431	421	410	433	423	409	373	451	441	409
ng	16	58	90	98	6	36	66	84	20	40	100	111
k	36	43	43	43	37	42	43	43	36	42	48	48
j	24	28	30	34	21	27	31	34	24	36	37	39
zh	2	5	12	18	3	6	18	18	3	6	14	14
z	1	2	5	10	1	3	5	10	1	1	15	17
d	8	14	25	25	14	14	18	24	8	14	23	23
m	24	30	33	35	24	28	41	41	24	28	28	41
g	8	24	28	28	12	24	28	27	12	15	28	27
ch	1	1	4	4	1	1	1	4	1	1	7	14
r	10	24	32	32	10	30	36	36	15	24	34	33
sh	8	13	15	15	8	12	13	16	8	12	13	15
ts	2	2	3	3	1	1	2	2	2	2	2	3
s	32	32	32	31	32	32	32	31	32	32	32	31
b	1	6	6	8	1	6	6	8	1	6	6	8
q	5	21	23	26	4	11	12	25	3	4	12	14
t	6	12	20	26	12	18	27	28	12	21	27	27
w	4	5	5	7	4	6	6	6	4	5	5	7
n	34	42	41	39	24	36	35	36	24	36	33	37
p	1	1	2	2	1	1	2	3	1	1	2	3
h	10	15	17	24	10	10	20	26	9	12	16	22
total	1620	2278	2431	2439	1537	2098	2332	2400	1544	2125	2339	2304