音声言語情報処理 14－12
(1996. 12. 13)

音素決定木に基づく逐次状態分割法による HM-Net の性能改善の検討

堀 貴明　加藤 正治　伊藤 彰則　好田 正紀

山形大学工学部

〒992 米沢市城南 4 丁目 3-16

E-mail: hori@ei5sun.yz.yamagata-u.ac.jp

あらまし　限られた学習サンプルから高精度かつ頑健な音素環境依存モデルを生成するためには，パラメータの共有関係をどのように決定するか，未知の音素環境をどのように扱うかが重要である．このような観点から，我々は音素決定木に基づく逐次状態分割法(Decision Tree-based Successive State Splitting:DT-SSS)を提案し，この手法によって自動生成された HM-Net が高精度かつあらゆる音素環境を表現可能であることを示した[13]．しかし，DT-SSS には時間方向の状態分割が組み込まれておらず，この手法によって生成された HM-Net は SSS の特徴を十分に反映したモデルではなかった．本報告では，DT-SSS の性能改善のために時間方向の状態分割を導入し，様々な初期モデルからの状態分割を試みて，連続音素認識実験により性能を比較する．また，頑健性向上とパラメータ数削減のために，過度に分割が行われた状態の再共有化についても検討する．

キーワード　音声認識，隠れマルコフモデル，コンテキスト依存モデル，隠れマルコフ網，音素決定木

A Study on Improvement of HM-Nets using Decision Tree-based Successive State Splitting

Takaaki HORI, Masaharu KATOH, Akinori ITO and Masaki KOHDA

Faculty of Engineering, Yamagata University

4-3-16 Johnan, Yonezawa 992

E-mail: hori@ei5sun.yz.yamagata-u.ac.jp

Abstract　The important aspects of context-dependent acoustic modeling using a limited training data set for robust speech recognition are how to tie the model parameters and how to handle the unknown contexts. From this point of view, we proposed the Decision Tree-based Successive State Splitting algorithm(DT-SSS), and showed HM-Nets generated with this algorithm had high accuracy and enabled to represent any contexts. But this algorithm was not taken temporal splits into consideration, and therefore did not make the best use of the strong point of SSS. In this paper, we incorporate temporal splits into DT-SSS and generate HM-Nets from various initial models. In continuous phoneme recognition experiments, we show the effects of these improvements.

key words　speech recognition, hidden Markov model, context-dependent model, hidden Markov network, phonetic decision tree

1. はじめに

限られた学習サンプルから高精度かつ頑健な音素環境依存モデルを生成するためには，パラメータの共有関係をどのように決定するか，未知の音素環境をどのように扱うかが重要である．このような観点から，我々は音素決定木に基づく逐次状態分割法(Decision Tree-based Successive State Splitting:DT-SSS)を提案し，この手法によって自動生成された HM-Net が高精度かつあらゆる音素環境を表現可能であることを示した[13].

DT-SSS では，コンテキスト方向の状態分割において音素決定木の質問による yes/no 分割を採用し，かつ単一正規分布からの安定した分布の分割を実現している．これにより，高精度かつあらゆるコンテキストを表現可能な HM-Net の自動生成が可能である．しかしながら，DT-SSS には時間方向の状態分割が組み込まれておらず，この手法によって生成された HM-Net は SSS の特徴を十分に反映したモデルではなかった．

Singer らは，単一分布ベースですべての分割可能な状態の分割を行い，最大尤度を与える状態を選択する形式の ML-SSS を提案した[14]．この手法では，時間方向の状態分割における新たな分布のパラメータを EM アルゴリズムによる繰り返し計算によって求めている．

本報告では，DT-SSS の性能改善のために EM アルゴリズムに基づく時間方向の状態分割を導入し，様々な初期モデルからの状態分割を試みて，連続音素認識実験により性能を比較する．また，頑健性向上とパラメータ数削減のために，過度に分割が行われた状態の再共有化についても検討する．

2. 逐次状態分割法(SSS)

2.1 HM-Net

SSS によって生成される HM-Net は，複数の状態のネットワークとして表される．個々の状態は，それぞれ以下の情報を保有している．

- 状態番号
- 受理可能なコンテキストクラス
- 先行状態・後続状態のリスト
- 自己遷移確率と後続状態への遷移確率
- 出力確率分布パラメータ

HM-Net では，コンテキスト情報が与えられたとき，そのコンテキストを受理することができる状態を先行状態・後続状態のリストの制約内で連結することにより，そのコンテキストに対するモデルを一意に決定することができる．このモデルは自己ループと隣の状態への遷移のみを許す left-to-right モデルと等価であるため，通常の HMM と同様に Baum-Welch アルゴリズムによってパラメータ推定を行うことができる．

2.2 SSS アルゴリズム

SSS は，全コンテキストを表す 1 状態の HM-Net から，状態をコンテキスト方向，または時間方向へ分割することにより，自動的に HM-Net の構造を決定する方法である．SSS における処理の流れを図 1 に示す．

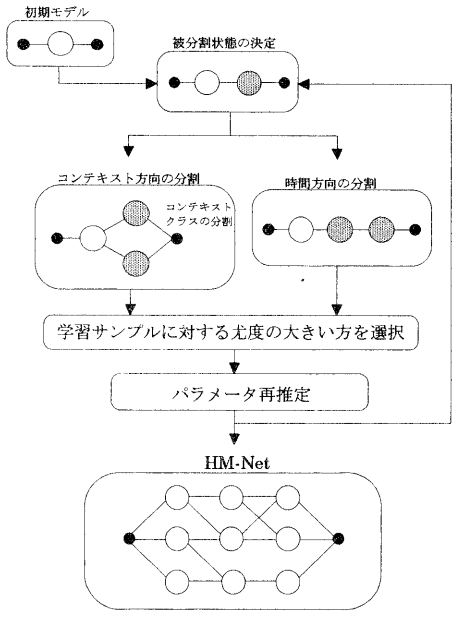

図 1　SSS の流れ

まず初期モデルとして，一つの状態とその状態を始端から終端まで結ぶ一本のパスから成るモデルをすべての学習サンプルから形成する．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いずれか一方に関して行われる．コンテキスト方向への分割時には，パスの分割に伴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．そして，コンテキストクラスの分割も含めたあらゆる分割方法の中で，学習サンプルに対する尤度の総和を最も大きくする分割を採用する．このような状態分割を繰り返すことによって，HM-Net が形成される．

3. 音素決定木に基づく逐次状態分割法

3.1 音素決定木

音素決定木は音素の音響的変動をとらえ，かつ未知コンテキストの音響的特性を予測する方法である．音素決定木は根をコンテキストに独立な単位とする二分木で表され，根から葉に向かってコンテキストクラスの分割が行われる(図 2)．つまり，この木は根から葉に向かうに従ってコンテキスト依存性の強い単位を表す階層構造を持っており，通常葉の部分にモデルを対応付ける．

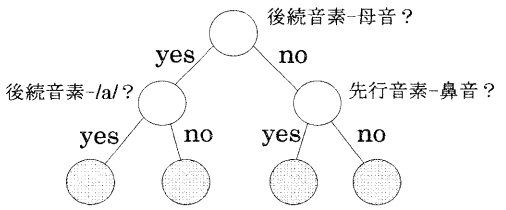

図2 音素決定木

木の各節点には先験的な音響類似性に基づく質問が割り当てられ，yes か no でコンテキストクラスを2分割する．これらの質問は音素環境要因と音素群を対にしたものとなっている．どのようなコンテキストも，木の根から質問をたどることによって必ずどこかの葉に対応するため，未知コンテキストであっても音響的に最も類似した葉に分類されることが期待できる．そのため，未知コンテキストをコンテキスト独立モデル等で代用する必要はなくなる．

3.2 DT-SSS アルゴリズム

DT-SSS は，音素決定木の機構を逐次状態分割法に応用したものであり，コンテキスト方向の状態分割に質問による yes/no 分割を採用している．また，yes/no 各状態の出力確率分布を単一正規分布とし，状態分割時には2混合分布の各々を新たな状態に割り当てるのではなく，新たに分布のパラメータを計算するという改良も加えてある．しかしながら，新たに分布のパラメータを求める際，コンテキスト方向の場合には質問によってサンプルがどちらの状態を通るかが既知であるため，簡単な計算でそれぞれのパラメータを求めることができるが，時間方向の場合には前後に位置する状態にサンプルがどのくらい滞留するかが未知であるため，EM アルゴリズムによる繰り返し計算が必要になる[14].

計算量の点から，従来の DT-SSS では時間方向の状態分割を避け，初期モデルとして3状態の音素環境独立 HMM を用いていた．しかし，これは HM-Net の構造にかなりの制限を与えることとなり，望ましくない．

本報告では時間方向の状態分割を導入する．状態分割時の計算量は増加するが，モデルの最大長を設定することにより時間方向の分割回数をコンテキスト方向の分割回数よりもかなり少なく抑えることができるため，計算量はさほど問題にならないと考えられる．

DT-SSS に時間方向の状態分割を導入したアルゴリズムを次に示す．

Step1:初期モデルの学習

初期モデルを構成し，パラメータを学習する．

初期モデルの構造は任意であるが，各状態の出力確率分布は無相関単一正規分布とする．

Step2:被分割状態の決定

分布の分散の最も大きい状態を式(1)に従って決定する．この式は，従来の SSS と同様に，その状態の推定に用いられた音素サンプル数をも考慮したものとなっている．

$$d_i = n_i \sum_{p=1}^{P} \frac{\sigma_{ip}^2}{\sigma_{Tp}^2} \tag{1}$$

σ_{ip}^2 ：状態 i の分布の分散

σ_{Tp}^2 ：全サンプルの分散(正規化係数)

n_i ：状態 i の推定に用いた音素サンプル数

P ：特徴ベクトルの次元数

Step3:状態の分割

被分割状態 S(m)をコンテキスト方向と時間方向に分割する．時間方向に分割した場合と，各要因を各質問で分割した場合の HM-Net を実際に作成し，その中から学習サンプルに対する尤度の総和が最も大きくなる HM-Net を選択する．

詳細を以下に示す．

Step3.1:サンプルの切り出し

状態 S(m)で受理可能なコンテキストクラスを C(m)とし，その k 番目の要素を c_k とする．まず，S(m)に対応する学習サンプルを Viterbi セグメンテーションによって切り出し，c_k に属するサンプルの平均 μ_k，分散 σ_k^2，総フレーム数 f_k をすべての k について求めておく．

Step3.2:コンテキスト方向(Contextual domain)

C(m)を質問 q について分割した場合の yes と no の分布のパラメータは，Step3.1 で求めた μ_k, σ_k^2, f_k を利用して，次式のように簡単に求めることができる．

$$\mu_{q,yes} = \frac{\sum_{c_k \in Q_q} f_k \mu_k}{\sum_{c_k \in Q_q} f_k} \tag{2}$$

$$\sigma_{q,yes}^2 = \frac{\sum_{c_k \in Q_q} f_k \sigma_k^2 + \sum_{c_k \in Q_q} f_k \left(\mu_k - \mu_{q,yes}\right)^2}{\sum_{c_k \in Q_q} f_k} \tag{3}$$

$$\mu_{q,no} = \frac{\sum_{c_k \notin Q_q} f_k \mu_k}{\sum_{c_k \notin Q_q} f_k} \tag{4}$$

$$\sigma^2_{q,no} = \frac{\sum_{c_k \notin Q_q} f_k \sigma_k^2 + \sum_{c_k \notin Q_q} f_k \left(\mu_k - \mu_{q,no}\right)^2}{\sum_{c_k \notin Q_q} f_k} \quad (5)$$

$\mu_{q,yes}$, $\sigma^2_{q,yes}$: C(m)を質問qによって分割したときのyesの分布の平均と分散

$\mu_{q,no}$, $\sigma^2_{q,no}$: C(m)を質問qによって分割したときのnoの分布の平均と分散

Q_q : 質問qによってyesとなるコンテキストクラス

C(m)を質問qによって分割し，同時にyesとnoの分布を新たな状態S'(m)とS(M)に割り振る．そして，S(m)を通るパスで表現されていた学習サンプルYに対する尤度の最大値Pcを次式のように計算し，Pcを実現するqによって分割されたHM-Netを選択する．

$$P_c = \max_q \left\{ \sum_{c_k \in Q_q} P_m(y_k) + \sum_{c_k \notin Q_q} P_M(y_k) \right\} \quad (6)$$

y_k : c_kに対応するYの部分集合

$P_m(y_k)$: y_kをS'(m)上のパスに割り当てた場合の対数尤度の総和

$P_M(y_k)$: y_kをS(M)上のパスに割り当てた場合の対数尤度の総和

Step3.3:時間方向(Temporal domain)

S(m)を時間方向にコピーし，前に位置する状態をS'(m)，後に位置する状態をS(M)とする．切り出された学習サンプルを用いてS'(m)とS(M)の分布のパラメータをEMアルゴリズムによって推定する．再推定式は次のようになる．

$$\hat{\mu}_s = \frac{\sum_n \sum_t \gamma_s(x_{nt}) x_{nt}}{\sum_n \sum_t \gamma_s(x_{nt})} \quad (7)$$

$$\hat{\sigma}_s^2 = \frac{\sum_n \sum_t \gamma_s(x_{nt}) x_{nt}^2}{\sum_n \sum_t \gamma_s(x_{nt})} - \hat{\mu}_s^2 \quad (8)$$

ただし，s=S'(m), S(M)

μ_s, σ_s^2 : 状態sの分布の平均と分散

x_{nt} : n番目に切り出されたサンプルx_nのt番目の特徴ベクトル

$\gamma_s(x_{nt})$: 状態S'(m)とS(M)からx_nが出力されたときに状態sでx_{nt}が観測される確率

時間方向に分割した際の対数尤度P_tには，Yに対する対数尤度の総和を代入する．

Step3.4:分割方向の決定

$P_c \geqq P_t$ならば，コンテキスト方向に分割したHM-Net，$P_c < P_t$ならば，時間方向に分割したHM-Netを採用する．

Step4:S'(m)をS(m)とし，Mに1を加える．分割の影響が及ぶ範囲でパラメータを再推定する．

Step5:所定の状態数になっていれば終了する．そうでなければ，Step2からStep4を繰り返す．

以上の処理によってHM-Netの構造が決定する．

4.状態の再共有化

トップダウンに分割を進めるクラスタリングでは，一度分割されたクラスタの要素は二度と同じクラスタには含まれない．これは，時折，同じようなクラスタが複数存在する現象を引き起こす．このようなクラスタは再度共有することが望ましい．

音素決定木を構築した後で音響的に近いモデルを再度共有化する方法[9]や状態の分割と融合を繰り返す方法[12]が提案されている．

本報告では文献[9]の評価基準（対数尤度の期待値）を利用し，各時点で最も対数尤度の総和の減少量が少ない状態融合を行い，融合の影響が及ぶ範囲でパラメータ再推定を繰り返す方法を検討する．ただし，状態融合が許されるのは根(初期状態)が同じもの同士とする．

対数尤度の減少量の期待値ΔLは，次のように計算する．

$$\Delta L(s1,s2) = L(s1) + L(s2) - L(s1,s2) \quad (9)$$

$$L(s) = -\frac{1}{2}\left(\sum_{p=1}^{P} \log 2\pi\sigma_{sp}^2 + P\right) \sum_{f \in F} \gamma_s(o_f) \quad (10)$$

$$L(s1,s2) = -\frac{1}{2}\left(\sum_{p=1}^{P} \log 2\pi\sigma_{s1s2p}^2 + P\right) \times \sum_{f \in F} \left\{\gamma_{s1}(o_f) + \gamma_{s2}(o_f)\right\} \quad (11)$$

σ_s^2 : 状態sの分布の分散

σ_{s1s2}^2 : 状態s1とs2の分布を合成した場合の分散

P : 特徴ベクトルの次元数

$\gamma_s(o_f)$: 学習サンプルの第fフレームが状態sで観測される確率

F : 学習サンプルの総フレーム

パラメータ再推定は，状態融合が他の状態に影響を及ぼすことを考慮したものであり，文献[9]では行われ

表 1 音声分析条件

標本化周波数	12kHz
量子化ビット数	16bit
分析フレーム長	32ms
分析周期	8ms
分析窓	ハミング窓
高域強調	$1\text{-}z^{-1}$
特徴パラメータ	1～12 次の LPC メルケプストラム係数と対数パワー，およびその一次と二次の回帰係数(計 39 次元)

ていない．

5 連続音素認識実験

5.1 音声資料・音響分析

特定話者として ATR 音声データ，男性話者 1 名(MHT)が発声した重要語 5240 語を用い，偶数番目を学習用，奇数番目を評価用とする．

不特定話者として音響学会連続音声データベースの男性 30 名が A～J セット(1 セット 50 文，J セットのみ 53 文)の中から 3 セットずつ発声した音素バランス文を用い，A～I セットを発声している男性 20 名の 3000 文を学習用，残りの 10 名の内 J セットを除いた 1200 文を話者 open，タスク closed の評価用，除かれた 6 名の J セット 318 文を話者 open，タスク open の評価用とする．

音声分析条件を表 1 に示す．

5.2 HM-Net の生成

初期モデルは図 3 に示す 4 種類を検討する．④の 29 状態のモデルは Singer らの ML-SSS に用いられていたものである．

時間方向の状態分割は 4 状態までとして，状態分割を 800 状態まで行う． 分割後に分布を特定話者の場合は混合数 2 に再構成して連結学習を 10 回，不特定話者の場合は混合数 4 に再構成して連結学習を 5 回行う．音素カテゴリは 27 音素+無音(sil)とする．

考慮する環境要因は先行，当該，後続音素とする．DT-SSS の質問に用いる音素群は表 2 の 21 種類に各音素のみが yes となる 26 種類を追加した 47 種類とする．ただし，コンテキストクラスの分割を行う際に最適分割を実現する質問が複数存在する場合には，表 2 のより上段にある質問を優先する．

比較のために従来の SSS による HM-Net も作成する．未知コンテキストを代用するコンテキスト独立モデルは，4 状態 3 ループ，混合数 8 の HMM(CI-model)を用いる．また，不特定話者の場合は特定話者の構造を利用する．

5.3 連続音素認識

HM-Net を評価する認識実験として連続音素認識実験を行う．日本語の音節の制約に基づく音素ネットワークを構成し，最適音素系列を求めたときの正解音素系列に対する音素正解率：

$$\text{音素正解率} = \frac{\text{M - I - D - S}}{\text{M}} \times 100\ (\%) \quad (12)$$

によって評価する．ここで，M は評価用サンプルの総音素数，I は挿入誤り，D は脱落誤り，S は置換誤りの数を表す．

6. 実験結果と考察

6.1 特定話者連続音素認識実験

重要語の奇数番目の単語について，DT-SSS において 4 種類の初期モデルを用いた場合の連続音素認識実験を行った．100 から 800 状態までの HM-Net における音素正解率を図 4 に示す．1 状態から従来の SSS によって生成した HM-Net，および，81 状態から時間方向の状態分割を行わない DT-SSS によって生成した HM-Net による結果も併せて示す．

また，800 状態における音素誤り率を図 5 に示す．

表 2 音素決定木の質問に用いる音素群

			a	i	u	e	o	w	y	xy	r	h	f	z	j	s	sh	ts	ch	p	t	k	b	d	g	m	n	N	cl	sil
有声音			○	○	○	○	○	○	○	○	○			○	○								○	○	○	○	○	○		
母音			○	○	○	○	○																							
母音	調音位置	前舌		○		○																								
		後舌	○		○		○																							
	顎の開き	狭		○	○																									
		半狭				○	○																							
		広	○																											
子音	調音方式	半母音						○	○	○	○																			
		摩擦音											○	○	○	○	○	○												
		破擦音																	○											
		破裂音																		○	○	○	○	○	○					
		鼻音																								○	○	○		
子音	調音位置	唇						○												○		○		○		○				
		歯，歯茎											○		○		○		○		○		○		○		○	○		
		口蓋												○		○		○				○		○						
		声門										○																		
無音																													○	○
子音	有声摩擦音													○	○															
	無声摩擦音												○			○	○													
	有声破裂音																						○	○	○					
	無声破裂音																			○	○	○								

*各行の質問に対して○がついている列の音素が yes となる．音素はローマ字表記とし，xy は拗音，cl は促音，sil は無音を表す．

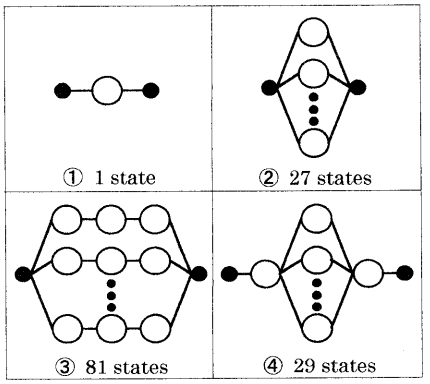

図 3 初期モデル

6.2 状態再共有化 HM-Net による認識実験

特定話者において，DT-SSSによる800,1000,1200状態のHM-Netからそれぞれ状態再共有化を行ったHM-Netによる連続音素認識実験結果を図6に示す．ただし，再共有を行うHM-Netはすべて81状態から分割したモデルである．

また，再共有時のパラメータ再推定を行う場合と行わない場合の比較として，800状態と1000状態からの再共有におけるHM-Netによる連続音素認識実験結果を図7，図8に示す．

6.3 不特定話者連続音素認識実験

不特定話者の文発声に対する連続音素認識実験を行った．男性10名による話者open，タスクclosedの評価実験，および男性6名による話者open，タスクopenの評価実験における音素正解率を図9に示す．初期モデルは27状態と81状態で，時間方向の状態分割を行わない場合は81状態の方を用いた．

また，800状態，1200状態のHM-Net，1200状態から800状態に再共有を行ったHM-Netによる認識結果も併せて示す．

6.4 考察

(1)DT-SSSにおける時間方向の状態分割は，種々の初期モデルからの状態分割を可能とし，認識性能を向上させる．

(2)初期モデルは生成されるHM-Netの認識性能に大きく作用する．認識性能が最も良いのは各音素3状態(81状態)の初期モデルの場合である．これより，初期モデルとして望ましいのは，時間方向にある程度分割が行われ，かつ当該音素間の状態共有を行わないモデルと言える．

(3)図5の音素誤り率を見ると，初期モデルの違いや時間方向の状態分割のあるなしによって特徴が異なる．当該音素間の状態共有を行わない初期モデルの場合は行う場合に比べて置換誤りが少ない．各音素1状態(27状態)の場合は挿入誤りが多いが，これは，時間方向にあまり分割されなかった状態の音素が頻繁に挿入されたことによる．また，時間方向の状態分割は挿入誤りを削減する．

(4)HM-Netの状態再共有は，頑健性向上，パラメータ数削減に有効である．特に，全体の20%から30%の状態削減が効果的である．

(5)状態再共有におけるパラメータの再推定は，状態の融合が他の状態へ及ぼす影響を考慮することができ，有効である．

(6)不特定話者における連続認識実験結果(図 9)より，特定話者と同様に時間方向の状態分割の効果が示された．また，初期モデルも各音素3状態(81状態)が最も良い．しかし，状態再共有に関しては特に大きな効果は見られなかった．これは，800状態と1200状態のHM-Netの間にあまり性能の差がないこと，再共有を進めすぎたことが原因と考えられる．

7. むすび

音素決定木に基づく逐次状態分割法(DT-SSS)の性

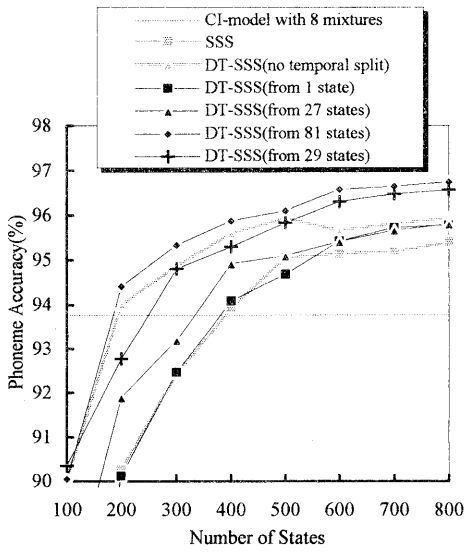

図４　連続音素認識実験結果

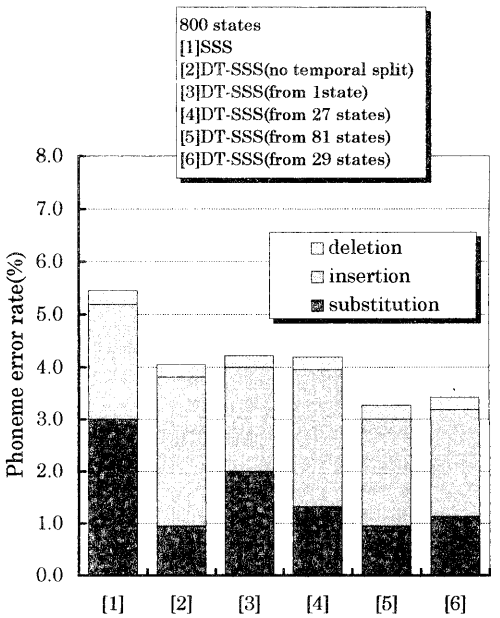

図５　特定話者連続音素認識の誤り率

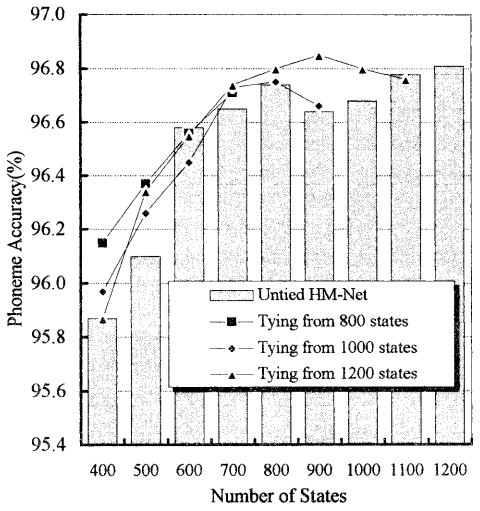

図６　状態再共有化を行った HM-Net による連続音素認識実験結果

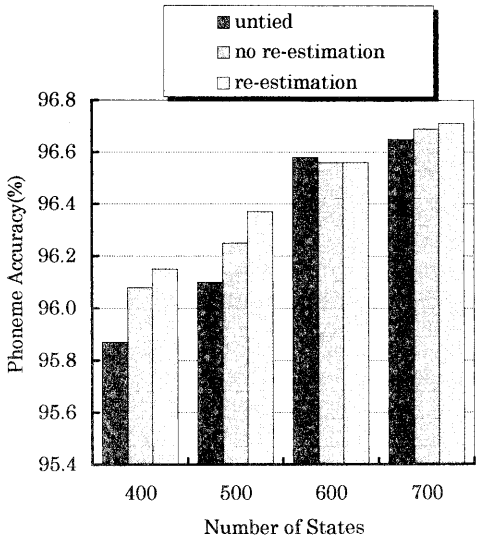

図７　800 状態からの状態再共有におけるパラメータ再推定の有無の比較

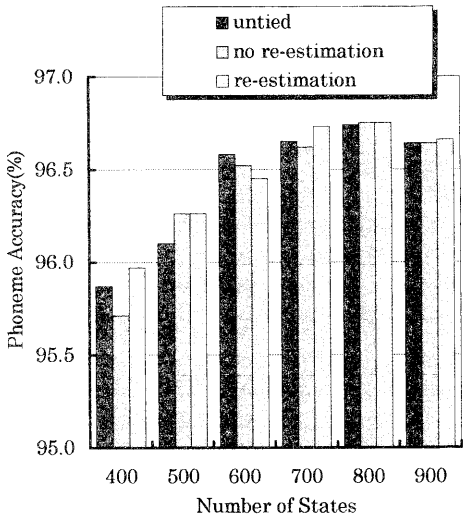

図８　1000 状態からの状態再共有におけるパラメータ再推定の有無の比較

音素決定木に基づく逐次状態分割法(DT-SSS)の性能改善を検討し，時間方向の状態分割と状態再共有化を導入した．特定話者/不特定話者による連続音素認識実験によって性能改善の効果を確認した．今後，DT-SSS による HM-Net のより大規模なタスクへの応用を検討する予定である．

文献

[1]中川聖一：“確率モデルによる音声認識”，電子情報通信学会(1988)

[2]C.-H.Lee, et al.：“Large vocabulary speech recognition using subword units”, Speech Communication 13, pp.263-279(1993).

[3]A.Ljolje：“High accuracy phone recognition using context clustering and quasi-triphonic model”, Computer Speech and Language 8, pp.129-151(1994).

[4]S.Sagayama, et al.：“Estimation of Unknown Context Using a Phoneme Environment Clustering Algorithm”, ICSLP 90, Vol.I, pp.361-364(1990).

[5]鷹見，嵯峨山：“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, pp.2155-2164(1993-10).

[6]S.Hayamizu, et al.：“Description of acoustic variations by tree-based phone modeling”, ICSLP90, pp.705-708(1990).

[7]L.R.Bahl, et al.：“Decision trees for phonological rules in continuous speech”, ICASSP91, pp.185-188(1991).

[8]M-Y.Hwang, et al.：“Predicting unseen triphones with senones”, ICASSP93, Vol.II, pp.311-314(1993).

[9]S.J.Young, et al.：“Tree-based state tying for high accuracy acoustics modeling”, Proc. ARPA Human Language Technology Workshop, pp.307-312(1994).

[10]L.Nguyen, et al.：“The 1994 BBN/BYBLOS speech recognition system”, Proc. ARPA Spoken Language Systems Technology Workshop, pp.77-81(1995).

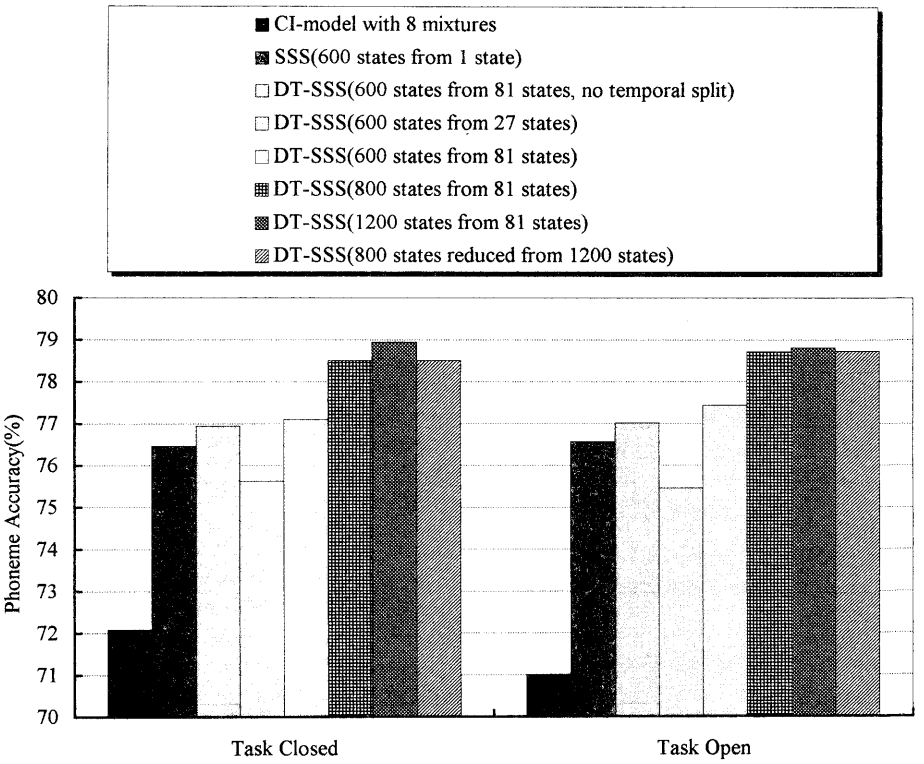

図９　不特定話者連続音素認識実験結果

[11]高橋,嵯峨山：“4 階層共有構造の音素 HMM”，信学技報, SP94-73(1994-12).

[12]鷹見：“状態分割融合法による高効率な隠れマルコフ網の自動生成”，信学論(D-II), J78-D-II, 5, pp.717-726(1995-05).

[13]堀,加藤,伊藤,好田：“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(1996-6).

[14]H.Singer et al.:“Maximum Likelihood Successive State Splitting”，音講論 3-5-12(1996-03)