音声言語情報処理　4－11
（1994. 12. 16）

4階層共有構造の音素HMM

高橋　敏　　嵯峨山茂樹

NTTヒューマンインタフェース研究所
神奈川県横須賀市武1-2356

あらまし　　音素環境依存モデルは，膨大な数のモデルパラメータを，限られた学習データで推定しなければならないという問題がある．音響モデルを効率良く学習するためには，同じ特性のパラメータは共有し，よいモデルの構造を与える必要がある．本報告では，4階層の共有構造を持つ音素モデルを提案する．4階層の共有とは，1）モデル（allophone環境）レベル，2）状態レベル，3）分布レベル，4）特徴パラメータレベルである．最初の3階層については，既にいくつかの手法が提案されているが，4階層目の特徴パラメータの共有化は，本報告で新たに提案するものである．特徴パラメータレベルの共有化により，多次元混合分布型HMMに含まれる1600個の正規分布の平均値ベクトルを，各次元につき，わずか16個の代表的な平均値の組み合わせで表現しても，認識性能がほとんど劣化しないことがわかった．また，モデルの学習効率を話者適応化実験を通して確かめた．認識時に必要な計算量も削減できることがわかった．

和文キーワード：HMM，共有構造，音素モデル，音声認識

PHONEME HMMS WITH FOUR-LEVEL TIED-STRUCTURE

Satoshi Takahashi　and　Shigeki Sagayama

NTT Human Interface Laboratories
1-2356 Take, Yokosuka-shi, Kanagawa, 238 Japan

Abstract　　One of the problems with context-dependent HMMs is that a large number of model parameters should be estimated using a limited amount of training data. Parameters that have the same property should be tied in order to represent acoustic models efficiently. This paper proposes four-level tied-structure for phoneme models. The four levels include 1) allophone environment level, 2) state level, 3) distribution level, and 4) feature parameter level. Although some techniques have been proposed for the first three levels, feature parameter tying in the fourth level is newly proposed in this paper. We found that feature parameter tying makes it possible to represent 1,600 mean vectors of multivariate Gaussian mixture HMMs by using the combination of 16 representative mean values in each dimension. Experimental results show that feature parameter tying reduces the amount of calculation required for recognition without significant degrading performance. Furthermore, we found that feature parameter tying is also effective for model training, especially in speaker adaptation.

英文 key words: HMM, tied-structure, phoneme model, speech recognition

1．まえがき

統計的モデルの1つであるHMMは，モデルの精密度と認識時の頑健性にトレードオフの関係があることがよく指摘される．モデルの精密度を高めるためにモデルの種類を増やしたり，モデルの規模を大きくしたりすると，推定すべきパラメータ数が増加する．よって，大量の学習データが必要となるが，データ量が十分でないと，モデルパラメータが学習データに依存してしまい，学習データとはほんの異なるテストデータに対しても誤認識を起こしてしまう．逆に，大量の学習データがあっても，モデルパラメータが少ないと十分な認識性能が得られない．これらは，学習データ量とモデルの自由度の問題と言える．

これまでの音響モデル設計における進展を振り返ると，generalized triphoneに代表されるallophone環境の共有化[1][2]，HMnet等の状態の共有化[3][4]，tied-mixture等の基底分布の共有化[5-7]と，構造の共有化が1つの大きな流れであった．すなわち，より自由度の少ない，一般化されたモデルをいかに作るかが大きなテーマであった．

共有構造を追求する利点は2つ考えられる．1つは統計的に同じ性質を持つモデルパラメータに，構造的に連動する仕組みを持たせて，学習効率を高めることができる点である．同じ学習データ量でも共有化によって，より認識率の高いモデルを作成することができるし，また逆に言えば，同じ性能のモデルを得るにも，より少ない学習データ量で済む．もう1つは，総パラメータ数を減らすことによって，認識時の計算量を削減できる点である．

本稿では，これまで提案されている3つのレベルの共有化技術を用い，更に新たに4番目の共有構造として考案した特徴パラメータレベルの共有化を含む4階層の共有構造を持つ音素モデルを提案する．特徴パラメータレベルの共有化は，基底分布の平均値ベクトルの各次元で，平均値（スカラー量）を共有するものである．この4階層の共有構造を含むHMMは，HMMについて考えられる全ての構成要素について共有化を行ったモデルと言ってよい．本モデルを，共有構造のないモデルと，認識性能，学習効率，計算量について比較し，提案するモデルの有効性を確認した．

2．4階層の共有構造

2.1 モデルレベル

この章では，4階層共有モデルにおいて，最も上位の層と思われるものから順に説明する．一番上の層は，モデルレベルの共有である．即ち，異なる音素環境が同一のモデルを共有する．具体的にはtriphone環境において，先行，後続音素が中心音素に及ぼす影響が同じ音素環境モデルが，1つのモデルを共有する．これはgeneralized triphone[1]をはじめ，ほとんどの音素環境依存モデルで行なわれている．従来は，音素環境をクラスタリングするという観点から実行されていたが，これは異なる音素環境間でのモデルの共有関係を見つけ出す手段に他ならない．図1に中心音素が /k/ である場合の例を示す．ここでは，破裂音 /k/ は，先行母音よりもむしろ後続母音によってその特徴が変化すると仮定する．よって図は，後続母音が /a/ であるモデルは1つのモデルを共有できるが，後続母音が /i/ のモデルは共有されないことを示している．これはあくまでも例であるが，物理的な特徴量に従ってモデルの共有関係を調べても，同様の共有化がなされ，効率化が図られるであろう．

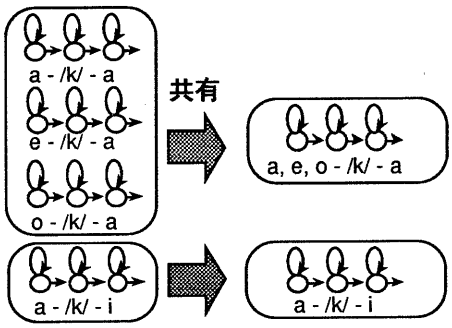

図1　音素環境の共有化

2.2 状態レベル

2番目は状態レベルの共有である．即ち，異なるモデル間で同一の状態を共有する．具体的には，HMMの状態が表現する特徴量分布が，異なる音響モデル間で類似している場合に，1つの状態を共有する．これにより音素環境依存モデルをより少ない数の状態で表現できる．これを実現する方法は2つある．1つは，学習データから得られるすべての音素環境依存モデルを生成した後に，類似した状態をクラスタリングしていく状態マージ方式[4]である．もう1つは1状態1分布から始まり，逐次的に状態を増やしていく状態分割方式[3]である．いずれも最終的には状態ネットワークを生成する．図2は，同じく音素 /k/ の例を示している．仮に，第1状態が /k/ の閉塞部分を，第2状態が破裂部分を表現しているとすると，母音部分への渡りを表現している第3状態は分岐して，それぞれの特徴量分布を表現する．このように，状態をつなぐパスの違いによって，異なる音素環境モデルを表現する．我々はここでは，逐次状態分割法 (Successive State Splitting: SSS) を用いて1層目と2層目を同時に構築する[3]．SSSは尤度最大化基準のもとで，コンテキスト方向と時間方向に状態分割を繰り返し，状態ネットワークを生成する手法である．SSSによって生成された状態ネットワークはHMnetと呼ばれる．

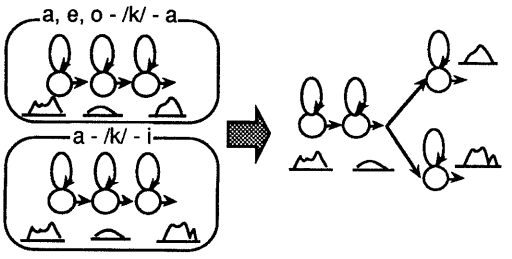

図2　状態の共有化

2.3 分布レベル

3番目は，いわゆるtied-mixture HMM [6][7]（またはsemicontinuous HMM [5]）で代表される基底分布の共有である．即ち，異なる状態間で同一の基底分布を共有する．具体的には，混合分布を構成する分布が，異なる状態間で類似している（平均値ベクトル，共分散行列が類似している）場合に共有化される．基底分布の共有化を行うためには，基本的に2つの手法がある．1つは，状態間で独立な分布を持つモデルを初めに学習し，すべての分布の中から類似した分布に対して部分的に結びの関係をもたせ，再学習する方法である．もう1つは，VQ符合帳のように共通した基底分布のセットを予め用意し，各状態に関係する分布のみを再学習する方法である．ここでは，前者を採用する．基底分布の共有関係を見つけ出すために，分布のクラスタリング手法を用いる．すなわち，すべての状態のすべての分布をクラスタ化し，同じクラスタ内の分布をマージした後，これを代表分布としてクラスタ内の分布が共有するようにする．分布の共有により，より少数の分布で特徴量空間全体を効率良く覆うことができる．図3は，第3状態がそれぞれ3つの混合正規分布で表現されている場合に，ある1つの分布（$N(\mu_3, \sigma_3)$）を2つの状態が共有している例を示している．図では，分布が1次元的に表現されているが，実際は多次元の正規分布であることに注意を要する．

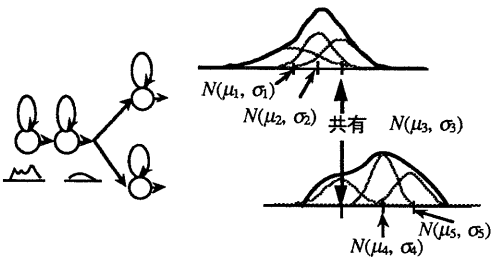

図3　基底分布の共有化

2.4 特徴パラメータレベル

これまで述べた3種類の共有化はすでに提案されたものであるが，これらの延長として最後に考えられる共有化として特徴パラメータレベルの共有化を新たに提案する．特徴パラメータの共有では，異なる基底分布間で同一の特徴パラメータを共有する．具体的には，基底分布の平均値ベクトルの各次元で，平均値（スカラー量）を共有するものである．つまり，多次元の特徴パラメータベクトル（例えば16次のケプストラム）の各次元で，類似した平均値を異なる基底分布間で共有する．

はじめに，特徴パラメータレベルの共有化の可能性について述べる．簡単のため，図4にあるような2次元の特徴量空間を考え，2つの平均値ベクトル μ_1 と μ_2 が存在すると仮定する．2つのベクトルの1次元目の要素 $\mu_{1,1}$ と $\mu_{2,1}$ は，お互いに離れているのでベクトル間の距離は大きく，よって，3層目の平均値ベクトルの共有化ではマージされない．しかし，2次元目の要素 $\mu_{1,2}$ と $\mu_{2,2}$ は近接しているので，特徴パラメータレベルでは共有化できる可能性がある．この様に，3層目の基底分布の共有では，ベクトルのある次元の要素間の距離が離れていると共有化されない可能性があるが，

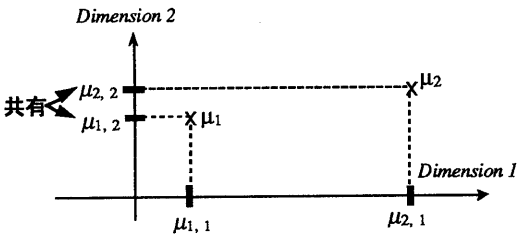

図4　特徴パラメータレベルにおける共有化の可能性

特徴パラメータレベルでは，より多くの共有化が見込める．

また一般に，音素環境依存モデルには多数（例えば1000以上）の分布が存在し，各次元には分布数と同数の平均値が存在する．しかし，ケプストラムなどの特徴パラメータのスペクトル感度（解像度）を考えると，各次元100点以下でも十分に表現可能であると考えられる．たとえ，平均値が各次元でm点にマージされた場合でも，それらが表現できるベクトルの種類はm^P個であり（Pは次元数），かなりの表現能力を保持している．図5に，特徴パラメータレベルの共有化の概念図を示す．黒い点は分布の中心（平均値ベクトル）を表し，それを囲む丸は分布の広さ（分散）を示している．近接する平均値は各次元でマージされるので，結果的に，あらゆる分布の中心（平均

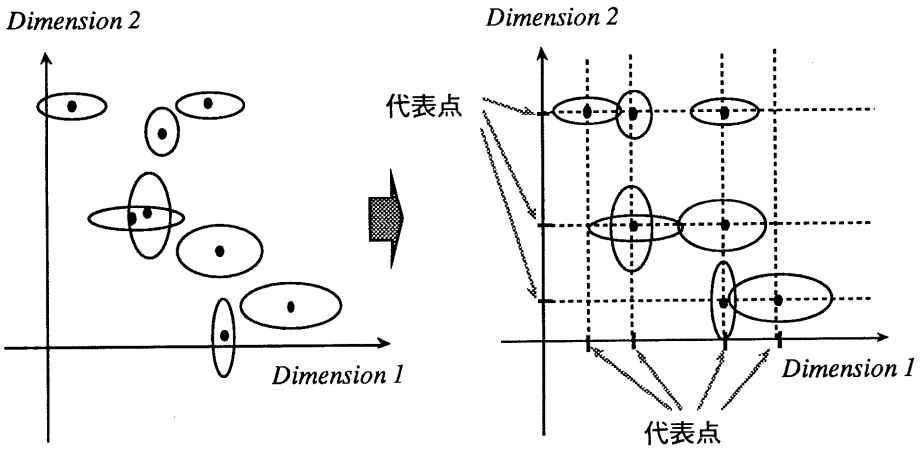

図5　特徴パラメータレベルの共有化

値ベクトル）は，平均値の代表点で決定される格子点に平行移動することになる．このときの移動量が，出力確率密度値に誤差となってあらわれるが，これが認識に影響を与えるかが焦点となる．

次に，計算量の観点から利点を考える．現在の多くのHMMは，多次元無相関正規分布を仮定しているので，時刻 t における入力ベクトル $\mathbf{x}_t=(x_{t,1}, x_{t,2}, \cdots x_{t,P})^T$ の k 番目の分布に対する対数尤度は以下のように計算される．

$$\log p_k(\mathbf{x}_t) = -\frac{P}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{P}\log\sigma_{k,i}^2 - \sum_{i=1}^{P}\frac{(x_{t,i}-\mu_{k,i})^2}{2\sigma_{k,i}^2} \quad (1)$$

ここで，$\mu_{k,i}$ と $\sigma_{k,i}^2$ は k 番目の分布の第 i 次成分の平均値と分散をそれぞれ示す．P は次元数である．入力に関係する第3項目の計算は各次元の計算結果の和になっており，次元間に渡る計算はない．したがって，計算は各次元で独立に考えることができる．平均値を共有すれば，式(1)の第3項目の分子の計算結果 $(x_{t,i}-\mu_{k,i})^2$（入力値と平均値の差の自乗）をテーブルに格納し，それらをモデル間で共有することができるので，認識時の計算量を削減できる．

3．4階層共有構造の生成法

今回の実験で用いた4階層共有モデルの具体的な生成法を説明する．生成のためのフローチャートを図6に示す．

[Step-1]　1層目，2層目の共有を実現するため，状態逐次分割法 (SSS) [3] を用いてHMnetを作成する．1名の話者データを用いて，各状態が単一正規分布で表現された600状態のHMnetを作成する．ちなみに，以下で述べる実験で得られた状態ネットワークには，およそ1700種類のtriphoneモデルが含まれていた．

[Step-2]　上記モデルと同一の状態共有構造で，4名分のHMnetをそれぞれ学習する．

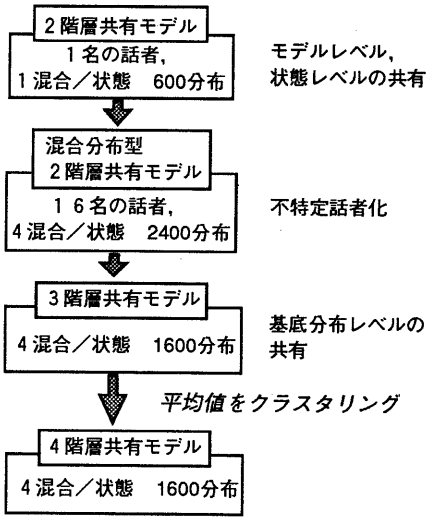

図6　4階層共有モデル生成のための
フローチャート

次に，4つのHMnetの対応する状態を重ね合せ，各状態4混合のHMnetを作成する．これを初期モデルにして16名分の話者データを用いて学習し，不特定話者用HMnetを作成する[8]．

[Step-3]　3層目の基底分布の共有を実現するために，すべての状態の分布（4混合　x 600状態＝2400分布）から1600個の分布クラスタを生成し，クラスタ内の分布を共有化する．クラスタリングの際の距離尺度はKullback divergenceである．クラスタ m の代表分布の平均値 $\mu_{m,i}$，分散 $\sigma_{m,i}^2$（i は次数を示す）は以下のように計算される．

$$\mu_{m,i} = \Big(\sum_{k\in cluster\ m}\mu_{k,i}\Big)/K \quad (2)$$

$$\sigma_{m,i}^2 = \Big(\sum_{k\in cluster\ m}\sigma_{k,i}^2 + \sum_{k\in cluster\ m}\mu_{k,i}^2 - K\cdot\mu_{m,i}^2\Big)/K \quad \cdots\cdots (3)$$

ここで，総和はクラスタ m に属するすべての分布に対してとられる（K はその総数）．よって，各状態の混合数は4混合を保ちながら，総分布数は2400個から1600個に削減される[9]．

[Step-4]　3層目まで実現した共有モデルを

もとに，平均値ベクトルの各次元において1600個ある平均値を n 個 (= 256, 64, 16, 4) にスカラー量子化する．距離尺度はユークリッド距離である．もとの平均値ベクトルはこれら n 個の代表平均値の組み合わせで表現されることになる．なお，3層目，4層目のクラスタリング手法は*k*-means法を用いた．また，共分散行列は共有化していない．

4．評価実験

4.1 データベース

4階層の共有構造を持つHMMの性能，学習効率，認識時の計算量を評価するために，音素認識実験，話者適応化実験，単語認識実験を行った．モデルの学習にはATR重要語5240単語の偶数番目の単語と音素バランス216単語を16名分（全部で45,376単語）用いた．評価には5240単語の奇数番目の単語を4名分用いた．よって，話者，発声内容ともオープンな実験である．音素カテゴリー数は26である．パラメータは16次のケプストラム，16次のΔケプストラム，Δパワー（全部で33次元）である．

4.2 実験結果

各階層モデルの総分布数，平均値クラスタ数，そして平均音素認識率を表1に示す．参考のため，3状態16混合の音素環境独立モデルの性能も示す．2階層目を実現することにより，性能は音素環境独立モデルより3.2%向上した．3層目モデルで若干，認識率が下がったが，分散を含めた総パラメータ数は2/3になっている．この段階で，各次元には，それぞれ1600個の平均値が存在する．これらをクラスタリングして，その結果をもとに共有化し，代表点を256個, 64個, 16個, 4個と減らしていった．注目すべきことは，平均値を各次元16個にまで減らしても，認識性能がほとんど変わらないことである．このことから，従来のモデルでは各次元において冗長な分布成分が数多く存在していると思われる．

4階層モデルの学習効率を話者適応の枠組みの中で調べた．実験では，標準話者から新しい話者への話者適応（1対1話者適応）を行った．標準話者モデルは1名の男性話者の学習データから作成した600状態の単一正規分布のモデルである．これを適応化用単語を用いて通常の最尤推定法で学習し，新しい話者へ適応化した．図7は，

表1　26音素認識実験結果

モデル	状態数	混合数	総分布数	平均値クラスタ数	平均認識率[%]
環境独立モデル	26音素x3	16	1248	1248	84.4
2階層共有モデル	600	4	2400	2400	87.6
3階層共有モデル			1600	1600	86.8
4階層共有モデル				256	86.9
				64	86.9
				16	86.6
				4	84.0

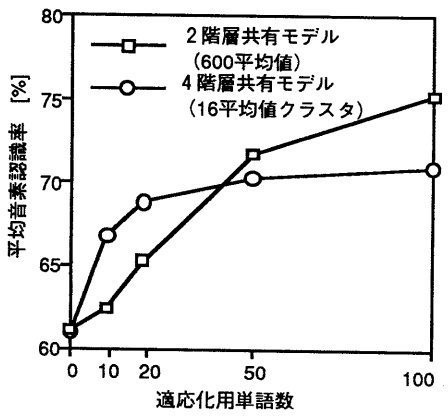

図7　話者適応化における学習曲線
（1対1話者適応）

表2　単語認識における計算回数の比較

モデル	入力値と平均値の差の自乗計算 $(x_{t,i}-\mu_{k,i})^2$ の回数	平均単語認識率 [%]
2階層共有モデル (2400平均値)	1.0	94.0
4階層共有モデル (16平均値クラスタ)	0.009	93.5

標準話者を学習話者の中から1人選定し，新しい話者に4名の評価話者をそれぞれあてはめて実験を行なった際の結果の平均である．折れ線は，適応化用単語を増加したときの平均音素認識率を示している．4階層モデルでは，1つの平均値を多数の平均値ベクトルの要素が共有しているため（1点を，平均600/16≒38個のベクトル要素が共有している），少数の平均値ベクトルが適応化された場合でも多数のベクトル要素が同時に移動する．よって，適応化単語数が少ない場合は，共有構造のないモデルに比べて，認識率が高くなっている．しかし，単語数が多くなると，より自由度の大きい共有構造のないモデルの方が性能が高くなっている．これらは，モデルの自由度と学習データ量から決定されるもので，学習データ量に見合う自由度を備えたモデルが最も性能を発揮する．よってこの実験でも，特徴パラメータレベルの共有化の数を変えれば，適応化速度や認識率が飽和する点も変わると考えられる．

音素認識実験で得られた2階層モデル（2400個の平均値ベクトルを含む）と4階層モデル（1600個の平均値ベクトルを各次元16個の代表平均値の組み合わせで表現）を用いて単語認識を行った場合の認識率と認識時の計算量を表2に示す．語彙はATR5240単語の奇数番目の単語から任意に抜き出した1000単語で，4人の評価用話者がそれぞれ発声した100単語を評価対象とした．計算量は，尤度計算の際に実行された入力値と平均値の差の自乗計算の回数（式(1)の第3項の分子の計算）を比較した．この結果から，部分的ではあるが，尤度計算内の平均値に関わる計算量を大幅に削減できることがわかった．

5．むすび

本稿では，特徴パラメータレベルの共有化を含む4階層の共有構造を提案し，実際にHMMを作成して評価を行った．特徴パラメータレベルの共有化により，モデル全体で1600個ある分布の平均値ベクトルを，各次元16点の代表平均値の組み合わせで表現できることがわかり，認識時の平均値に関わる計算量を削減できることがわかった．また，共有化の少ないモデルよりも，学習効率を高くできる可能性があることを話者適応化実験によって確認できた．

[参考文献]

[1] 嵯峨山，“音素環境のクラスタリング”，音講論 1-5-15 (1987).

[2] K-F Lee, H.-W. Hon, "Large-vocabulary speaker-independent continuous speech recognition using HMM". ICASSP-88, pp. 123-126 (1988).

[3] 鷹見，嵯峨山，“逐次状態分割法(SSS)

による隠れマルコフネットワークの自動生成”，音講論 2-5-13 (1991-10)
[4] S. Young, P. Woodland, "The use of state tying in continuous speech recognition", Eurospeech-93, pp. 2203-2206, (1993).
[5] X. D. Huang, M. A. Jack, "Unified technique for vector quantization and hidden Markov modeling using semi-continuous models", ICASSP-89, pp.639-642 (1989).
[6] J. Bellegarda, D. Nahamoo, "Tied mixture continuous parameter models for large vocabulary isolated speech recognition", ICASSP-89, pp.13-16 (1989).
[7] D. Paul, "The Lincoln robust continuous speech recognition", ICASSP-89, pp.449-452 (1989)
[8] 小坂，鷹見，嵯峨山，“話者混合SSSによる不特定話者音声認識”，音講論 2-5-9 (1992-10)
[9] 鷹見, H. Singer, 大脇，“語意や構文に制約のない音声認識手法における音響モデルの性能評価”，音講論 2-P-21 (1994-3)