音声言語情報処理 29－8
(1999. 12. 20)

Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識

李 晃伸　河原 達也　武田 一哉†　鹿野 清宏‡

京都大学 情報学研究科
†名古屋大学 工学研究科
‡奈良先端科学技術大学院大学 情報科学研究科
〒606-8501 京都府京都市左京区吉田本町
†〒464-8603 愛知県名古屋市千種区不老町
‡〒630-0101 奈良県生駒市高山町 8916-5

あらまし 大語彙連続音声認識のための新たな phonetic tied-mixture (PTM) モデルを提案する．このモデルは monophone モデルの各状態が持つ数十個のガウス分布集合を triphone の対応する状態に割り当て，重みのみを変えて共有することで合成する．通常の状態共有 triphone に比べて音響空間を効率よく表現でき，また巨大なコードブックを要する従来の tied-mixture モデルよりも学習が容易である．JNAS の 2 万語の新聞記事読み上げタスクにおいて評価した結果，triphone での最大性能と同等の 7.0 % の単語誤り率をより少ないパラメータ数で達成した．また処理効率の面においても，音響スコア計算に用いるガウス分布を上位 3% にまで削減しても精度がほとんど低下しなかった．いくつかのガウス分布の足切り計算 (Gaussian pruning) 手法を提案および比較した結果，最終的に音響尤度計算を約 5 分の 1 にまで削減できた．

キーワード triphone, tied-mixture, PTM, 大語彙連続音声認識, Gaussian pruning

Phonetic Tied-Mixture Model for LVCSR

Akinobu Lee　Tatsuya Kawahara　Kazuya Takeda†　Kiyohiro Shikano‡

Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
†Nagoya University, Nagoya 464-8603, Japan
‡Nara Institute of Science and Technology, Ikoma 630-0101, Japan

Abstract A phonetic tied-mixture (PTM) model for efficient large vocabulary continuous speech recognition is presented. It is synthesized from context-independent phone models with 64 mixture components per state by assigning different mixture weights according to the shared states of triphones. Mixtures are then re-estimated for optimization. The model achieves a word error rate of 7.0% at 20k-word dictation of newspaper corpus, which is comparable to the best figure by the triphone of much higher resolutions. Compared with conventional PTMs that share Gaussians by all states, the proposed model is easily trained and reliably estimated. Furthermore, the model enables the decoder to perform efficient Gaussian pruning. It is found out that computing only two out of 64 components does not cause any loss of accuracy. Several methods for the pruning are proposed and compared, and the best one reduced the computation to about 20%.

key words triphone, tied-mixture, phonetic tied-mixture, Gaussian pruning

1 はじめに

近年の大語彙連続音声認識の研究においては，大量の音声コーパスの整備を背景としたコンテキスト依存の音素モデルなどの大規模なモデル化のアプローチが成功を収めている．しかし異なる環境ごとに大量のコーパスを整った形で収集するのは多大な労力を要する上，適用の分野によっては大量のデータ収集そのものが困難な場合もある．したがってパラメータをHMM内で効率よく共有することで全体のパラメータ量を減らすことが音響モデルの効率の良い学習にとって重要である．

今日もっとも広く用いられているパラメータ共有方法は，HMMの状態を音響的な類似度あるいはトップダウンなコンテキストの類似性のルールに従ってクラスタリングする方法である．これは状態共有triphone モデルと呼ばれる．また別のアプローチとしては，音響空間全体を単一の大きなガウス分布の集合としてコードブック化し，全ての状態はそのコードブックを参照しつつ重みのみを変えて持たせる方法がある．これは混合分布結合（tied-mixture，以下TM）モデルと呼ばれている．そしてさらにこのTMの拡張として，音素ごとにコードブックを作成して，中央音素が同じtriphoneの間で共有する音素混合分布結合（phonetic tied-mixture，以下PTM）モデル[1][2]が提案されている．

TMやPTMモデルは状態共有triphoneに比べて，音響空間全体を少ないガウス分布で表現するのでより信頼度の高いパラメータ推定が行えるという利点がある．しかし従来のTMやPTMでは単一のコードブックで広い範囲をカバーする必要があるため，コードブックは巨大になり，特に分布数の多い高精度なモデルを構築するのは容易ではない．これに対して本稿では，音素内の状態単位で分布集合を結合することで高精度でかつ学習が容易な新たなPTMを提案する．

また，通常の triphone モデルに対する TM および PTM モデルのもう一つの利点は，分布集合が多くの状態で共有されるため，ガウス分布の出力確率計算において足切りの機構を導入しやすいことである．そこで本稿では，効果的なガウス分布計算の足切り (Gaussian pruning) のアルゴリズムについてもいくつか提案し，比較を行う．

2 音響モデルの状態共有と分布結合

状態共有 triphone と分布結合ベースの TM および PTM を，音響特徴量空間の適切なモデル化という観点から比較する．なお，本稿ではHMMの出力分布のモデルとして対角共分散の混合ガウス分布のみを考える．

音素のコンテキストに依存して状態を定義するtriphoneモデルでは，状態を音響的な類似性，すなわち出力分布の類似性に従ってクラスタリングし，クラスタ内で混合分布を共有することで安定した学習を行うのが一般的である．しかしこのような状態単位の共有では不十分であると考えられる．分布全体が似通っている場合は共有がうまく行われるが，部分的な共有が行えないため，分布の一部のみが重なるような場合に，その重なった部分は各状態ごとに独立に学習され，音響空間上で多重に定義されてしまう．このような冗長な分布の増大はパラメータの増大を招き，学習が不安定になる．

一方TMモデルでは，音響空間全体を効率的に表現するガウス分布集合がまず定義され，各状態についてその集合を参照しながら各ガウス分布に対する重みのみ個々に推定する．triphoneにあったような分布の冗長性は原理上なく，より少ないパラメータで音響空間全体を表現できるため学習の信頼性は高い．

しかしTMモデルはtriphoneに比べて，これまで十分な認識性能を示していない．これはtriphoneが数千の状態と数万のガウス分布を持つことが可能であるのに対し，TMではそのような数のガウス分布からなる1つのコードブックを最適に学習するのは極めて困難であるため，全体のガウス分布数はtriphoneよりも大幅に少なく抑えられてしまい，十分な識別能力を得ることができないためである．

このTMの特徴はそのまま従来のPTMにもあてはまる．音素単位でコードブックを作成するため音響空間は音素の数だけ細分化されるが，それでも1つあたり数百個のガウス分布からなるコードブックを安定して学習するのは容易ではない．

3 状態単位でコードブックを共有 Phonetic Tied-Mixture モデル

これらの考察に基づいて，本稿では音素内の状態単位でコードブックを共有するPTMを提案する．同じ中央音素を持つtriphoneの各状態どうしでコードブックを共有することで，冗長なガウス分布をマー

ジして効率よくパラメータを共有できる．また通常のPTMと比べると，中心音素が同じtriphone HMMの集合内では音素の各状態ごとに独立したコードブックを持つので，それらをmonophoneモデルの各状態の分布から安定して与えることができる．ただし，中心音素が同じtriphone HMMは状態数が全て同じであるという仮定が必要である．

3.1 構成

提案したPTMは，monophoneモデルと状態共有triphoneモデルを合成することにより構成できる．モデルの構成と作成の流れを図1に示す．monophoneの各HMMの状態ごとに学習された混合分布は，それぞれコードブックとして対応する中央音素を持つtriphoneの状態に割り付けられる．元のtriphoneで共有されていなかった状態についてはコードブックのみを共有して別々の重みを持たせ，triphoneで共有されていた状態については重みも共有する．この後，コードブックと重みを再学習して最適化する．

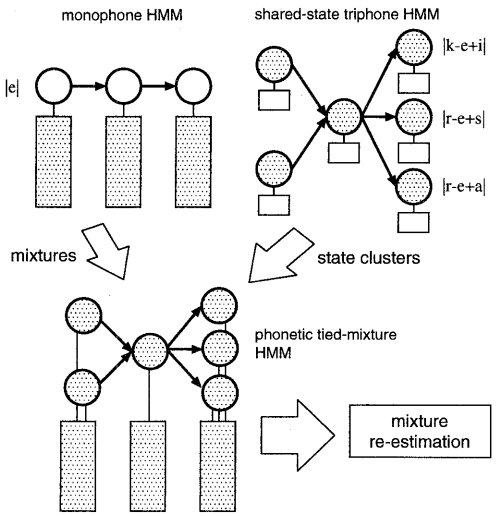


図1: 状態単位のコードブックをもつPTM HMMの作成

この作成過程は単純で実装が容易である．各状態のコードブックは，標準的なmonophoneの学習法の枠組みで徐々に混合数を増やしながら学習できる．これによって分布集合を着実に高い信頼度で推定できることができる．また状態レベルのクラスタリングと各分布の重みを変えた形での共有を組み合わせることで，モデルのコンテキスト依存性を表現しつつ効率のよいパラメータ共有が行える．

3.2 構築手順

具体的な構築の手順は以下の通りである．

1. 状態ごとに混合分布を持つ通常の monophone HMMを学習する．混合分布の大きさはここでは64とした．
2. 同じ音素セットを用いて状態共有triphone HMMを学習する．目的は状態レベルの共有構造を得ることのみであるので，ここで学習に用いる出力分布は単一ガウス分布で十分である．なお状態のクラスタリングは中心音素が同じ triphone HMM どうしでのみ行い，かつ各triphoneの状態数は中心音素が同じ triphone HMM 内で同じである必要がある．
3. monophoneの混合分布をtriphoneの対応するそれぞれの状態に割り当てる． triphone間で共有されていない状態については，分布集合のみを共有し独立した重みを持たせる．
4. 分布集合と重みの両方を再学習して，モデル全体での最適化を図る．

4 Gaussian Pruning 手法

大規模な音響モデルを用いる大語彙連続音声認識システムでは，計算するガウス分布数が膨大となる．そのため音響マッチングの計算量が認識における処理量の多くを占める場合が多い．高速なデコーディングを実現するためには音響マッチングの回数を削減することが重要である．

この音響スコア計算の高速化手法の1つとして，低い尤度になるであろうガウス分布を計算途中で除外するガウス分布の足切り (Gaussian pruning) がある．入力ベクトルに対するあるガウス分布の尤度は，入力ベクトルの各次元ごとに対数確率を加えていく際に単調減少する．したがってそのガウス分布の尤度があるしきい値を下回るかどうかを全ての各次元要素を計算し終える前に判断できる．実際，ある入力に対して大きな影響を持つガウス分布はコードブックの上位数個であるので，足切りによる精度の低下は小さいと見込まれる．

ここで，この Gaussian pruning の手法をいくつか提案する．ここでの目的は，ある混合分布において上位 k 個のガウス分布を求めつつ，他の余分なガウス分布の計算をできるだけ抑えることにある．以下に，提案する 4 つの手法を述べる．

k-best vector threshold
尤度の上位 k 個のガウス分布集合を常に保持しておき，その k 位の尤度を足切り値とする．あるガウス分布の尤度を各次元ごとに計算中にこのしきい値を下回ったら，その分布は現在の k 位以内に入らないことが確定するためその時点で足切りされる．最後まで計算され上位 k 以内に入れば，その上位 k 個の集合としきい値を更新する．

k-best vector threshold, previous best
上記と同じだが，直前のフレームで上位 k 個に残っていた分布集合を最初に計算する．入力ベクトルは連続したフレームではゆるやかに変化すると見込まれるため，あるフレームで上位だったガウス分布がその直後のフレームでも高い尤度になると期待できる．これによって足切り値の初期値が真の上位 k 位に近くなり，足切り性能が上がると期待される．

vector threshold with heuristic estimation
ガウス分布の各次元において，現在の尤度に未計算の次元のそれまでの最大値を加えることで最終的な尤度の上限を推定し，それを基に足切りを行う．最大値を 0 とおくと，上記の手法と等しくなる．各次元の最大値は直前フレームのベスト k 個のガウス分布であらかじめ計算しておく．

scalar threshold (dimension-independent)
ガウス分布の尤度そのものではなく，計算過程の各次元ごとに足切り値を定める．足切り値は，あらかじめ直前フレームのベスト k 個の各次元での最大値を求めておき，そこから一定の幅の値とする．

前者の 2 つの手法は，残りを計算しても確実に上位 k 個に入らないことで足切りを行うので，最終的に誤りなく k ベストのガウス分布集合が求められることが保証されている．これに対して後者の 2 つは安全でない方法であり，より高い足切り性能が得られる可能性があるが，最大値による未計算部のスコア推定が実際のスコアよりも悪かった場合やスコアの幅が小さ過ぎる場合に，計算の過程で本来の k ベストが失われる可能性がある．

5 評価実験

提案した PTM モデルの認識精度と処理効率を評価した．64 混合分布の monophone モデルと 3000 状態の triphone モデルから性別非依存の PTM モデルを合成して作成した．43 個の各音素ごとに 3 状態の HMM を用意するので，合計で 129 個のコードブックを持つことになる．

タスクは日本音響学会の新聞記事読み上げコーパスの 20k 語の認識である．言語モデルの単語 3-gram および辞書は「日本語ディクテーション基本ソフトウェア」 1998 年度版 [3] のものを使用する．デコーダは我々の開発した A* 探索に基づく 2 パスデコーダ Julius[4] を用いる．バージョンは 2.3a で，第 1 パスで単語間の音素環境依存性を扱う [5]．比較対象の音響モデルは，「日本語ディクテーション基本ソフトウェア」 1998 年度版に含まれている性別非依存の状態共有 triphone モデルを使用する．状態数は全て 2000 であるが 1 状態ごとの混合分布数は 4, 8 ,16 の 3 種類を用意した．なお本実験で作成した PTM はこれらと同じ学習セットを用いて学習されている．

テストセットは男女各 23 名の話者による計 200 文の発声を用いる．認識実験は Sun ワークステーション上で行った．CPU は UltraSPARC 300MHz である．

5.1 モデルの比較

提案モデルと既存の triphone モデルの単語認識精度を比較した．結果を表 1 に示す．なお表には各モデルの状態数と総ガウス分布数も示している．PTM は総混合分布数が同じ triphone よりもはるかに高い精度を示し，4 倍の分布数を持つ triphone と同等の性能を得ることができた．この認識精度はこのテストセットでほぼ最良の結果である．これによって提案した PTM は通常の triphone に比べて少ないパラメータで同等の性能を得ることができることが示された．表 1 の「PTM,synthesized」と「PTM,re-trained」については，前者は構築の最終段階で分布の重みのみを学習したモデルであり，後者は分布集合自身についても再学習を行ったモデルである．後者のほうがより高い精度が得られたことから，重みだけでなく分布自身の再学習が有効であることがわかる．これは主として monophone と triphone で音素のアラインメント

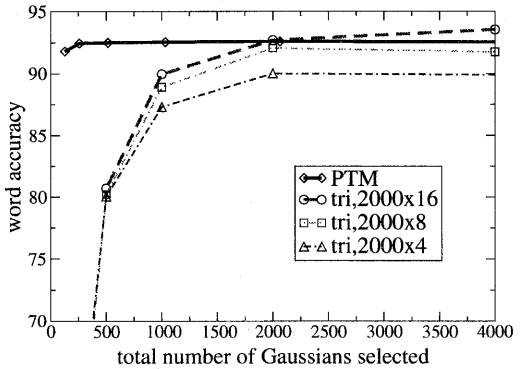
表 1: モデルの精度比較

	state× mix. size	total G. #	word accuracy
triphone	2000×16	32000	93.8
	2000× 8	16000	92.7
	2000× 4	8000	90.8
PTM,synthesized	129×64	8256	92.3
PTM,re-trained	129×64	8256	93.0

G. #: number of total mixture Gaussians
beam width = 1500

が異なるためと考えられる.

次に, Gaussian pruning のそれぞれのモデルに対する影響を調べた. 計算する総ガウス分布数を制限していった際の各モデルの単語認識精度の変化を図2に示す. なおモデルは PTM,re-trained を使用している. また足切りは, PTM では各コードブックごとに独立して行っているが, triphone では全ガウス分布をまとめて足切りを行っている. これは状態ごとに足切りを行うと計算するガウス分布数を状態数 (ここでは2000) より下げることができないためである. また実験の利便性のため, ここでは表1のときよりも狭いビーム幅を用いている.



(beam width = 800)

図 2: ガウス分布計算数制限による認識精度の変化

提案した PTM では, 計算するガウス分布数を全体の 3% (64分布のうちの2分布) まで削減しても認識精度はほとんど低下せず, 大幅な計算量の削減が可能であることがわかった. 一方, triphone では Gaussian pruning があまり有効に働かず, 十分な精度を得るには 2000 個以上のガウス分布を計算する必要があった.

これらの結果から, 我々が提案した PTM は通常の triphone よりも高い認識性能を持ちながら, Gaussian pruning がうまく働くため音響スコアの計算量も削減できる優れたモデルであることが示された.

5.2 Gaussian pruning 手法の比較

次に, 提案した各 Gaussian pruning 手法の PTM における性能を評価した. ここでは, コードブックごとに64個の中から上位2個のガウス分布を計算する際の音響スコア計算量を比較する. 音響スコア計算量は実際に計算された次元ごとのガウス分布距離の計算数の合計の, 全体 (ベクトル次元数 × 総ガウス分布数) に対する割合とする. なおここではベクトルは 25 次元である.

表 2: Gaussian pruning 手法の比較

pruning methods	Gaussian distance components computed	word acc.
k-best	59 %	92.5
k-best,previous-best	52 %	92.5
heuristic	36 %	92.3
scalar	21 %	92.2

beam width = 800, 2-best selected

表2に各手法の計算量と認識精度を比較する. 1番目の vector-threshold の方法では計算するガウス分布距離数を全体の 59% に抑えられた. そして直前フレームの k ベストのガウス分布から足切りの初期値を求める2番目の方法では, 52% にまで改善された. これらの方法では上位 k 個のガウス分布が必ず誤りなく得られるため, 足切りによる精度の低下なしに音響尤度計算のコストを約半分に抑えられる方法として, 単純かつ有用な方法であるといえる.

ヒューリスティックな尤度推定を用いる手法で, 全体の 36 % にまで計算量を削減できるが, この方法では未計算部の誤推定による足切り誤りが若干生じた. そして各次元ごとに独立した足切り値を設定する scalar threshold の方法によって最もよい結果が得られた. 計算量は全体の 21% にまで抑えられ, 認識精度の低下もわずか 0.3% であった. しかしこの方法で

は足切りの性能は各次元ごとの最大値からの幅のパラメータに依存するので，最大の効果を得るにはかなりのパラメータ値の最適化が必要である．

6 第1パスの monophone tree 化

本稿で提案した PTM モデルは，monophone の混合分布を音素ごとに割り当てることで構築される．したがって構築された PTM の内部で triphone だけでなく monophone も合わせて定義し，その混合分布を triphone と共有することで，monophone と triphone のハイブリッドモデルを作成することが可能である．これを用いて第1パスで monophone，第2パスで triphone を適用することで，さらに効率の良いデコーディングが行えると考えられる．monophone はコンテキスト独立なので，第1パスで monophone を使うことでコンパクトな木構造化辞書を作成することができ，特に単語間のコンテキスト依存を無視できることで処理量を抑えられるためである．なお同じ分布集合を共有するので，第1パスで計算した出力確率のキャッシュをそのまま第2パスの triphone の計算で利用でき，余分なガウス分布の計算が生じない．

表 3: triphone と monophone の木構造化辞書の比較

lexicon	state #	acc. (acc1)	time(×RT)
triphone	173251	92.2 (82.2)	4.5
monophone	128188	90.2 (76.8)	4.4

acc1: accuracy on the preliminary pass
CPU: UltraSPARC 300MHz

この結果を表3に示す．monophone による木構造化辞書の圧縮は速度の改善には結びつかなかった．また同じビーム幅でより多くの候補を残せる分，小さいビーム幅においては monophone 化辞書のほうが精度の低下が小さいと予想されたが，これも大差はなかった．逆に第1パスでのエラーの増大により，最終的な認識精度への影響が大きかった．この結果から，初期のパスにおいても高精度の音響モデルを用いることが重要であることが確かめられた．

最後に，探索のパラメータを最適化してより小さなビーム幅を用いることで，実時間の 2.3 倍の処理速度で 90.4% を達成することができた．

7 おわりに

状態単位の混合分布共有を用いた新しい PTM モデルを提案した．このモデルは monophone モデルの混合分布と triphone モデルの状態クラスタから合成して作成される．この構築方法は単純かつ実装が容易であり，パラメータの学習も高い信頼度で行うことができる．

このモデルは単語誤り率 7.0% を達成した．この値は数倍の数のパラメータを持つ状態共有 triphone での最良な値にほぼ等しい．さらに Gaussian pruning によって，分布集合ごとに 64 個のうち上位 2 つだけを計算に用いても認識精度の低下は見られなかった．ベクトルの各次元ごとに足切りのしきい値を独立に設定する pruning 手法によって，音響マッチングの計算量は約 20% にまで削減できた．

謝辞 本研究は，情報処理振興事業協会 (IPA) の「日本語ディクテーションの基本ソフトウェアの開発」の援助を受けて行われた．ご協力いただいた関係各位に感謝します．

参考文献

[1] G.Zavaliagkos, J.McDonough, D.Miller, A.El-Jaroudi, J.Billa, F.Richardson, K.Ma, M.Siu, and H.Gish. The BBN BYBLOS 1997 large vocabulary conversational speech recognition system. In *Proc. IEEE-ICASSP*, pp. 905–908, 1998.

[2] A.Sankar. A new look at HMM parameter tying for large vocabulary speech recognition. In *Proc. IEEE-ICASSP*, pp. 2219–2222, 1998.

[3] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 伊藤克亘, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏. 日本語ディクテーション基本ソフトウェア (98年度版) の性能評価. 情報処理学会研究報告, 99-SLP-26-6, 1999.

[4] 李晃伸, 河原達也, 堂下修司. 単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識. 電子情報通信学会論文誌, Vol. J82-D-II No.1, pp. 1–9, 1999.

[5] 李晃伸, 河原達也. 大語彙連続音声認識エンジン Julius における A* 探索法の改善. 情報処理学会研究報告, 99-SLP-27-6, 1999.