音声言語情報処理 34－17
(2000. 12. 21)

モノフォンモデルを用いた状態選択に基づく音響尤度計算の高速化

李 晃伸　河原 達也†　鹿野 清宏

奈良先端科学技術大学院大学 情報科学研究科
† 京都大学 情報学研究科
〒630-0101 奈良県生駒市高山町 8916-5
† 〒606-8501 京都府京都市左京区吉田本町

あらまし

音素環境非依存のモノフォンモデルの尤度に基づいてトライフォンの予備選択を行うことで，大語彙連続音声認識における音響尤度計算量を高速化する手法を提案する．フレームごとに全モノフォン状態の尤度を計算し，上位の状態についてのみ，対応するより高い解像度のトライフォンモデルの尤度を計算するとともに，下位の状態に対応するトライフォンについてはモノフォンの尤度をそのまま割り付けることで，非計算状態に対しても信頼度の高い back-off 尤度を与える．比較実験の結果，従来手法であるベクトル量子化に基づくガウス分布選択 (Gaussian selection) と同等の性能を示し，さらに計算するガウス分布数を絞った条件下でも安定した性能が得られた．Phonetic tied-mixture (PTM) モデルとの組合わせで，精度をほとんど落とさずに音響尤度計算コストを全体の約 14% に抑えることができた．

キーワード　大語彙連続音声認識, Gaussian selection, 状態選択, PTM

State Selection using Context-Independent HMM for Fast Likelihood Calculation

Akinobu Lee　Tatsuya Kawahara†　Kiyohiro Shikano

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

Abstract　We address a method to efficiently select Gaussian mixtures for fast acoustic likelihood computation. It makes use of context-independent models for selection and back-off of corresponding triphone models. Specifically, for the k-best phone models by the preliminary evaluation, triphone models of higher resolution are applied, and others are assigned likelihoods with the monophone models. This selection scheme assigns more reliable back-off likelihoods to the un-selected states than the conventional Gaussian selection based on a VQ codebook. Experimental results show that this method can achieves a comparable performance, and works much better under the aggressive pruning condition. Together with the phonetic tied-mixture (PTM) modeling, acoustic matching cost is reduced to almost 14% with little loss of accuracy.

key words　LVCSR, Gaussian selection, state selection, PTM

1 はじめに

近年，高精度な大語彙連続音声認識を実現するためのアプローチとして，大量の混合ガウス分布をもつ状態共有トライフォンなどの大規模な音素環境依存モデルが広く用いられている．しかし，数万個ものガウス分布を定義するためモデルはより詳細になるが，認識を実際に実行する際にはそれらを照合するための計算コストが莫大となる．実際の認識システムにおいては音響尤度計算が認識処理時間の多くを占めることが多い．このため，高精度かつ実時間の認識システムを実際に構築する際には，このような大規模な音響モデルを扱いつつ効率よく音響尤度を計算するための手法が求められる．

Gaussian selection[1] は，大語彙連続音声認識システムで広く用いられている音響尤度計算の高速化手法の一つである．音声の特徴ベクトル空間に対するVQ符合帳に基づいて全ガウス分布をあらかじめクラスタリングしておき，認識実行時にはフレームごとに入力ベクトルに最も近いクラスタに属するガウス分布のみを計算する．尤度の低いであろう分布をあらかじめ除外することで音響尤度計算を効果的に削減できる方法として，様々な拡張も試みられている [2][3][4].

しかし，この特徴ベクトル空間の分割に基づくアプローチの問題点として，選択がHMMの構造と無関係に行われるため，混合分布中のガウス分布が一つも計算されず出力確率が与えられない状態が少なからず存在し，精度の劣化を招くということがある．高速化のためにクラスタを小さくして選択ガウス分布数を少なくするほど，この非計算状態の数は増大し，ミスマッチによる選択誤りの際に最適解の枝刈りを引き起こしやすくなる．そのような状態に対して適当な固定値を与えてflooringすることで誤りを小さくできるが，必ずしも最適な解決策ではない．

本研究では音響尤度計算の高速化手法の一つとして，音素環境非依存のモノフォンHMMに基づいてトライフォンの状態を選択することを提案する．入力フレームごとに全モノフォンHMMを計算し，その上位 k 状態に対応するトライフォン状態のみを再計算する．このとき，非選択のトライフォン状態についても対応するモノフォン状態の確率をback-off確率として付与することで，非選択状態に対しても意味のあるスコア付けが行え，選択誤りに頑健な安定した認識が行える．さらに，予備選択とガウス分布計算の両方にGaussian pruningを導入することで音響尤度計算量をさらに削減できる．なおこの状態単位の選択については子森ら [5] の先行研究があるが，大語彙での評価はなされていない．

以下，第2節でGaussian selectionの仕組みと得失について述べたあと，第3節で提案手法とそのアルゴリズムを示す．第4節では標準的なGaussian selectionと認識実験で比較した結果を示し，第5節でまとめを述べる．

2 音響モデルの計算量削減手法

混合ガウス分布からなる音響HMMにおいては，状態の出力確率は，ガウス分布集合の重み付き線形結合で表される．高い認識精度を得るためには状態あたり十数個から数十個ものガウス分布集合を持つ必要があり，さらに音素環境の依存を考慮したトライフォンモデルでは数千状態が定義されるため，それらの出力確率を全て一つ一つ計算すると計算量が膨大となる．

この音響モデルの尤度計算量を削減するアプローチとしては，モデル間で構造を効率よく共有することでモデル自身のパラメータ数を削減する方法と，モデルは大規模でも認識実行時に一部のガウス分布だけを計算することで実際に計算される分布数を減らす方法に大別できる．前者では，モデル全体でガウス分布集合を重みを変えて共有するtied-mixtureモデルが挙げられる．後者では，入力フレームごとに尤度が高いと見込まれるガウス分布の部分集合を選択し，それらについてのみ計算を行うGaussian selection[1]や，ガウス分布の出力確率の計算過程で動的にガウス分布を計算対象から除外していくGaussian pruning[6]，あるいはガウス分布を階層的に木構造化することで尤度の高い部分集合を決定する手法 [7] などが提案されている．

本研究では，後者の認識実行時に有望な分布のみを計算する枠組みにおいて，頑健かつ安定したガウス分布集合の選択手法と計算手法を提案し，最も一般的なGaussian selection法と比較することで有効性を示す．

なお，探索アルゴリズムにおいてビーム幅を狭く設定したりfast match[8]を用いることで探索の範囲をより限定することで，結果的に認識に必要な音響尤度計算量を減らすことが可能である．しかしこのような探索レベルでの仮説候補の絞り込みだけでは計算コストの削減として不十分であり，上記で述べたような音響モデルのレベルでの尤度計算量の削減手法を組み合わせることで，より高い計算量削減効果を得ることができる．

3 Gaussian Selection

Gaussian selection は，現在の大語彙連続音声認識システムにおいて多く用いられる音響尤度計算の高速化手法の一つである．

ある入力ベクトルがガウス分布の分布の端に位置するとき，そのガウス分布の確率は極めて低くなる．混合分布を基本とする音響モデルにおいては，状態の出力確率は複数のガウス分布の重み付き線形結合で表されるため，そのようなガウス分布の最終的な出力確率への貢献は相対的に小さい．一部の高い確率を与えるガウス分布のみが最終的な出力確率に支配的な影響を持つので，入力ベクトルから遠いガウス分布を無視して入力ベクトル近傍のガウス分布のみ計算しても認識精度に影響はない．

Gaussian selection では，このような入力ベクトル近傍のガウス分布の集合をフレームごとに予備選択することで，計算量の削減を実現する．

3.1 標準的 Gaussian selection 法

標準的な Gaussian selection 法はベクトル量子化に基づくものであり，Bocchieri[1] によって提案された．まずあらかじめ音響特徴ベクトル空間に対する VQ コードブックを作成し，それに基づいて音響特徴量空間を領域分割して全てのガウス分布を各コードワードとの距離に基づいてクラスタリングする．認識時には入力ベクトルを量子化することで最も近いコードワードを選択し，それに対応するクラスタに属するガウス分布集合のみを計算する．実際には，選択誤りの影響を抑えるため，クラスタ間のオーバラップを許してクラスタリングを行う．すなわち，ガウス分布は最も近いコードワードのクラスタのみに属するのではなく，各コードワードベクトルからのユークリッド距離がしきい値以下であるような複数のクラスタに属することとする．特に，クラスタリングの距離としてベクトル要素ごとの分散の平均で重み付けされた距離を用いる手法 [3] を本研究ではベースラインとする．

本手法では 1 入力フレームごとに 1 クラスタ内のガウス分布が計算されるので，1 フレームあたりに計算されるガウス分布数はクラスタの大きさに等しい．クラスタの大きさはクラスタリングの際の距離のしきい値で調節される．このようにクラスタの大きさおよび計算量と認識精度にはトレードオフが存在する．

3.2 問題点

Gaussian selection の枠組みにおける重要な問題は，枝刈りされた状態に対して値が全く与えられないということである．クラスタは音響ベクトル空間全体を最適に分割するという観点から定義され，音響モデル自身の HMM 状態や混合分布の構造を考慮していない．特に，状態の混合分布中のガウス分布が選択されたクラスタに全く含まれなかったときに，その状態の出力確率を与えることができない．このような状態は探索のパス候補から除外され，関連する文仮説も認識過程から結果的に除去されてしまうため，このような状態が多く存在することは，認識精度への影響が大きい．特にこの問題は，より小さなクラスタを用いた場合や，モデルと入力のミスマッチによって選択誤りが生じる場合などに，より顕著に現れる．

そのような非計算状態に対して，多くの場合は適当な固定値を与える discrete flooring がしばしば用いられるが，実際の入力を全く反映しない単純な固定値は根拠に乏しく，安定した認識を行えることを保証するものではない．

この枝刈り状態の出力確率の問題は，音響空間分割に基づく Gaussian selection 法全てに共通の問題である．これまでに，状態ごとに同じクラスタに属するガウス分布数の上限を設定してクラスタリングをより均等に行う手法 [3] などの改善が試みられているが，本質的な解決には至っていない．

4 モノフォンモデルを用いた状態選択に基づく計算量削減手法

前節までの考察より，音響モデルの状態を考慮した選択を行い，かつ未選択の状態に対してもなんらかの意味のある尤度を付与することで，誤選択にも頑健な効率の良い音響尤度計算量の削減が行えると考える．そこで，構造のより単純で荒いモノフォン HMM の状態尤度に基づいて，トライフォンの状態選択と未選択状態への back-off 尤度の付与を同時に行う手法を提案する．

あるモノフォン HMM とトライフォン HMM が音素内で同一の状態構造（状態数および状態間遷移）を持ち，かつ同一コーパスで学習されたとき，両モデル間の対応する状態どうしの出力確率分布には強い相関があると考えられる．すなわち，ある入力に対して高い出力確率を持つモノフォン HMM の状態があるとき，その対応する中心音素のトライフォンの同じ位置の状態も同様に高い出力確率を持つと考えられる．

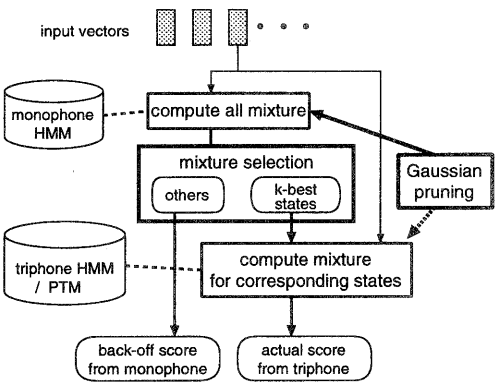

図 1: モノフォンモデルを用いた状態選択と back-off

モノフォンを用いた状態選択の具体的な処理の流れを図1に示す．音響尤度計算のアルゴリズムは，入力フレームごとに以下を繰り返すことになる．

1. モノフォンモデルの全状態の出力確率を計算し，その上位 k 個の状態を選択する．
2. あるトライフォンの状態の尤度を計算する際，その中心音素に対応するモノフォンの尤度が
   - 上位 k 個に入っていれば，トライフォン本来の混合分布の尤度を計算して返す
   - 入っていなければ，そのモノフォン尤度を返す

4.1 特徴

提案手法は，実際の入力に基づく信頼度の高い back-off 尤度を与えるため，より選択量を絞った条件においても精度の劣化は少ないと考えられる．モノフォンモデルは最尤推定に基づいて学習されており，単純な固定値や ad-hoc な計算値よりもトライフォンの近似値としてより適している．

また本提案手法は既存の枠組みから容易に拡張可能である．従来の Gaussian selection では前もって VQ コードブックの学習と大量のガウス分布のクラスタリングが必要であるのに対して，予備選択用のモノフォンモデルはトライフォンの学習過程で容易に得ることができる．また話者適応などによる音響モデルのオンラインの変形や更新にもトライフォン自身と同様の機構で容易に対応できる．

4.2 予備選択における Gaussian pruning の統合

しかし，このようなモデル照合に基づく予備選択は，予備照合のコストが問題となる．より精度の高い選択や back-off を行うには十分な数のガウス分布を持つ大規模なモノフォンモデルが望ましいが，モデルの規模に比例して予備照合のための計算コストは増大し，合計の計算量がかえって増加する可能性がある．

そこで，提案手法の予備選択に Gaussian pruning[9] を統合する．Gaussian pruning は，混合分布内で計算過程に動的な枝刈りを導入してガウス分布集合内の上位 k 位のガウス分布のみを求める手法である．我々はこれまでに phonetic tied-mixture モデルにおいてこの手法を評価し，混合分布の計算量を大きく削減できることを示した [9]．これを予備選択のモノフォンモデル計算に導入することで，予備選択計算のオーバヘッドを抑え，混合数の大きい精度の高いモノフォンの導入を可能にする．

さらに，提案手法では選択後のトライフォン状態の尤度計算にも Gaussian pruning を導入することが可能である．従来の Gaussian selection では状態とは無相関に選択が行われるが，提案手法では状態単位で選択が行われるため，選択された状態内の混合分布計算でさらに Gaussian pruning 導入することで，さらに計算量を削減することができる．

5 評価実験

提案手法を大語彙連続音声認識エンジン Julius [10] に実装して，認識実験による比較および評価を行った．タスクは単語 3-gram を用いた 2 万語の新聞記事読み上げ音声のディクテーションである．音響モデルとして，状態共有トライフォンと phonetic tied-mixture (以下 PTM) モデルの 2 種類を用いる．前者は 2000 状態 16 混合であり，後者は 64 混合からなるコードブックを音素の状態ごとに 129 個定義し，それらを 3000 個の状態間で重みを変えて共有している．なお両者とも性別依存モデルを用いる．これらのモジュールは全て日本語ディクテーション基本ソフトウェア (99 年度版) [11] に含まれている．テストセットは同ソフトウェアの評価 [11] に用いられたものから女性 23 名, 100 文発声を用いる．

提案手法であるモノフォン尤度に基づく状態（すなわち混合分布）選択法を，従来手法の標準 Gaussian selection 法と比較する．以下前者を GMS，後者を SGS と表記する．SGS のクラスタ数は 1119，非選択状態の flooring 値は固定値 (-50.0) を用いる．

表 1: 提案手法の評価

GS method		#Gauss. tri	#Gauss. pre	total %Gauss.	word %Err.
no GS		15772	—	100.00	4.5
SGS	2.1	6672	1119	49.40	4.5
	1.7	4132	1119	33.29	5.2
	1.3	2222	1119	21.18	6.2
	0.9	971	1119	13.25	15.7
GMS	48	6660	690	46.60	5.1
	24	3712	690	27.91	5.9
	8	1468	690	13.68	6.4
	4	824	690	9.60	8.6

音響モデル：状態共有トライフォン
SGS パラメータ：クラスタリングしきい値
GMS パラメータ：選択状態数
tri: 計算されたトライフォン中のガウス分布数
pre: 予備選択の計算コスト

5.1 提案手法 vs. 従来手法

まず，従来手法 (SGS) と提案手法 (GMS) の性能を状態共有トライフォンにおいて比較した．GMS の予備選択には 129 状態,16 混合のモノフォンモデルを用い，計算の際は Gaussian pruning によって各混合分布ごとに最も尤度の高いガウス分布のみを求めた．

選択数を変化させたときのフレームごとのガウス分布の平均計算数と単語認識精度を表 1に示す．SGS については距離しきい値の異なるクラスタを認識に先だって作成し，GMS については認識実行時に選択するモノフォンの状態数を直接指定することで選択数を調節した．正確な計算量の比較のため，実際に計算されたトライフォンのガウス分布のみをカウントし，予備選択の計算量も含めている．すなわち SGS については，クラスタ選択のための計算量をガウス分布の尤度計算量に換算して含んでいる．GMS では予備選択のためのモノフォン HMM の計算量を加えている．表中の “no GS” は予備選択を全く行わない場合の性能であり，このときの音響尤度計算量を 100% とする．

提案手法 (GMS) は従来手法 (SGS) に比べて，十分な量のガウス分布を選択した場合に同等の計算量削減性能を示した．さらに，計算するトライフォンを少なくした場合に，SGS に比べてはるかに安定した性能を示した．ガウス分布の計算量と認識精度を比較したグラフを図 2に示す．提案手法では，選択の結果計算されなかった状態に対して信頼度の高い back-off 尤度が与えられるため，選択のしきい値を厳しくした条件下でも安定して動作することが示された．

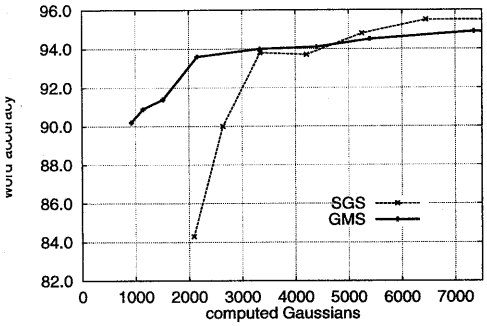

図 2: 選択性能の比較

GMS における予備選択のための計算量のオーバヘッドについては，トライフォンの計算量に対して十分小さかった．予備選択用のモノフォン HMM は 2064 のガウス分布を持つが，Gaussian pruning を導入することで実際の計算量を 690 まで抑えることができ，結果的に小さなオーバヘッドで SGS と同等の予備選択が行えた．

5.2 予備選択用モノフォンの性能比較

次に，予備選択用モノフォンモデルの性能を比較した．様々な混合数のモデル間の計算コストと精度の比較を図 3に示す．混合数の少ないモデルでは back-off の精度が悪化するため，選択数を減らすに従って認識精度が大きく低下している．また選択数を下げると急峻に認識精度が落ち，選択の精度も同様に悪化している．このことから，予備選択においても精度の高いモデルを用いることが有効であることが分かる．なお予備選択の計算量は 16,8,4,2 の各混合モデルでそれぞれ 690, 465, 300, 191 となっており，Gaussian pruning の効果から混合数の多いモデルについても計算量は大きくはない．

5.3 PTM モデルにおける性能

PTM モデルにおける提案手法の性能を表 2に示す．なおここでは選択後のトライフォンの計算にも Gaussian pruning を適用している．ほとんど認識精度を落とすことなくガウス分布計算数を，提案手法を

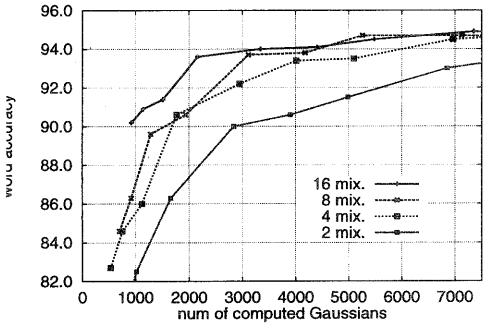

図 3: 予備選択用モノフォンモデルの比較

表 2: PTM モデルにおける提案手法の効果

GS method	#Gauss. tri	#Gauss. pre	total %Gauss.	word %Err.
triphone,2000x16	2644	690	21.14	5.9
PTM,129x64	434	690	13.61	6.0

selection model: 16mix. monophone

非導入時の 20.0% [9] から 13.6% にまで落とすことができ，パラメータ効率の高い PTM モデルにおいても提案手法の有効性が確かめられた．

5.4 システム性能

最後に，提案手法を組み込んだ認識システムの最大性能を測った．音響モデルは PTM を用い，ビーム幅などの探索パラメータを最適化した．テストセットは同じ評価セットの男性分を含め，計 46 人 200 文のサンプルで評価した．結果として，標準的な PC (CPU: 866MHz) の上で実時間処理で女性 93.5%, 男性 90.7% の単語認識精度を達成した．

6 おわりに

高速な音響尤度計算を実現するための，効率の良い HMM 状態の予備選択法を提案した．モノフォンモデルの状態尤度を用いて状態を選択するとともに非選択状態へ back-off 尤度を付与することで，より選択量を絞った高速な認識においても，従来の Gaussian selection 法に比べて精度の落ちが小さく効率の良い認識処理が実現できた．モデルと入力がマッチせず選択が外れやすい条件下においても，頑健な認識が行えると期待できる．PTM モデルと統合した認識システムにおいて，精度をほとんど落とすことなく音響尤度計算量を全体の 14% にまで削減することができた．

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 (99 年度版)」を用いて行われた．

参考文献

[1] E.Bocchieri. Vector quantization for efficient computation of continuous density likelihoods. In *Proc. IEEE-ICASSP*, pp. 692–695, 1993.

[2] K.M.Knill, M.J.F.Gales, and S.Young. Use of Gaussian selection in large vocabulary continuous speech. In *Proc. ICSLP*, Vol. 1.

[3] M.J.F.Gales, K.M.Knill, and S.J.Young. State-based Gaussian selection in large vocabulary continuous speech recognition using HMM's. Vol. 7, No. 2, pp. 152–161, 1999.

[4] D.B.Paul. An investigation of Gaussian shortlists. In *Proc. IEEE workshop on Automatic Speech Recognition and Understanding*, 1999.

[5] 子森康弘, 山田雅章, 山本寛樹, 大洞恭則. 少数分布 hmm による出力確率推定に基づいた効率的な混合連続分布 hmm 音声認識. 電子情報通信学会技術研究報告, SP94-52, 1994.

[6] J.Duchateau, K.Demuynck, and D.Van Compernolle. Fast and accurate acoustic modelling with semi-continuous HMMs. In *Speech Communication*, pp. 24(1):5–17, 1998.

[7] Takao Watanabe, Koichi Shinoda, Keizaburo Takagi, and Ken ichi Iso. High Speed Speech Recognition Using Tree- Structured Probability Density Function. In *Proc. ICASSP*, Vol. 1, pp. 556–559, 1995.

[8] L.R.Bahl, S.V.de Gennaro, P.S.Gopalakrishman, and R.L.Mercer. A fast approximate acoustic match for large vocabulary speech recognition. Vol. 1, No. 1, pp. 59–67, 1993.

[9] A.Lee, T.Kawahara, K.Takeda, and K.Shikano. A new phonetic tied-mixture model for efficient decoding. pp. 1269–1272, 2000.

[10] 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.

[11] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 嵯峨山茂樹, 伊藤克亘, 伊藤彰則, 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エア (99 年度版) の性能評価. 99-SLP-31-2, 2000.