音素環境独立HMMを用いた混合ガウス分布選択による音響尤度計算量の削減

李 晃伸[†] 河原達也[††] 鹿野清宏[†]

大規模な音響モデルにおいて音響尤度計算量を削減するための，効率の良い混合ガウス分布予備選択法を提案する．従来広く用いられているガウス分布選択法（Gaussian Selection）はVQコードブックに基づいて入力ベクトル近傍のガウス分布集合を決定的に予備選択するが，混合分布中の分布が選択されずに値がまったく得られないHMM状態が多く現れ，認識率の劣化が大きい．本研究では，音素環境独立のモノフォンモデルを用いてトライフォン状態の選択および非選択状態に対する尤度の近似を行う予備選択手法を提案する．モノフォンの状態の尤度をもとにトライフォンを状態単位で選択して計算する一方で，非選択の状態に対してもモノフォンの尤度を近似値として割り付けることで，選択誤りの認識精度への影響を抑えて安定した認識が行える．さらにこの状態選択法にGaussian pruningを導入することで，予備選択のための計算量を抑えて効率の良い音響尤度計算を行う．認識実験より，提案法は従来の標準的なGaussian Selectionと同等の性能を発揮し，とくに選択数をより絞った条件下において認識率の劣化を大幅に抑えられることが示された．最終的にPTMモデルを用いてGaussian pruningと統合することで，認識精度をほとんど落とさずに音響尤度計算量を全体の14%まで削減することができた．

Gaussian Mixture Selection Using Context-independent HMM for Efficient Acoustic Computation

Akinobu Lee,[†] Tatsuya Kawahara[††] and Kiyohiro Shikano[†]

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. It can also incorporate efficient Gaussian pruning at the preliminary evaluation, which offsets the increased size of the pre-selection model. Experimental results show that the proposed method achieves comparable performance as the standard Gaussian selection, and performs much better under aggressive pruning condition. Together with the phonetic tied-mixture (PTM) modeling, acoustic matching cost is reduced to almost 14% with little loss of accuracy.

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，音響モデルや言語モデルの高精度化が進み，ディクテーションなどの大規模なタスクにおいても高い認識精度を達成するに至っている．この進歩には，大量の音声データベースの基盤的整備を背景として，混合ガウス分布やトライフォンなどの詳細かつ大規模な音響モデルが構築可能となったことが大きく寄与している．特に不特定話者を対象とした汎用の認識システムにおいては，性別や年齢などに起因する音声の様々な変動に幅広く対応するために，想定される変動をできるだけ幅広く網羅した大規模な音響モデルが要求される．このように，音響モデルの大規模化と詳細化は今後の音声認識研究においても精度向上のために有効な方策であり続けると見込まれる．

しかし，モデルが巨大かつ詳細になるほど大量のガウス分布が定義され，認識時の音響尤度計算量は著しく増大する．実際，既存の音声認識システムにおいて

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
†† 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

は音響尤度計算が処理時間の多くを占めることが多い．このため，特に大規模な音響モデルを用いて認識を行う際には，効率の良い音響尤度計算手法が重要となる．

ガウス分布予備選択法（Gaussian Selection，以下 GS と表記）は，音響尤度計算量を削減する方法の 1 つとして多くの大語彙連続音声認識システムで採用されている．音響モデル中の全ガウス分布を特徴量空間上の領域ごとにあらかじめクラスタリングしておき，フレームごとにその入力ベクトルの近傍のガウス分布集合のみを計算する．この分布のクラスタリングは，音響特徴ベクトル空間を均一に量子化する VQ コードブックに基づいて行う方法が一般的である．効率良く計算量を削減できる方法であり，これまでに様々な拡張が研究されている[1)~4)]．

しかしながら，このような決定的な枝刈りに基づく高速化のアプローチでは，予備選択の誤りが認識精度の大きな劣化を引き起こす．特に選択の結果ガウス分布がすべて枝刈りされた HMM の状態については，尤度がいっさい与えられず，局所的な音響尤度のみに基づいて認識候補から除外されてしまうことになる．この枝刈りされた状態に対しては，なんらかの固定値を与えてフロアリングすることで誤りをいくぶん解消できるが，単純に人工的な固定値をトップダウンに与える方法は必ずしも適切とはいえない．この問題は決定的な選択を行うすべてのガウス分布予備選択法に共通の問題である．

本論文では，音素環境独立の音響 HMM の状態尤度に基づくガウス混合分布の予備選択手法を提案する．音素環境に独立なモノフォンの HMM の状態の尤度を用いて，対応するトライフォンの状態（およびそれに属する混合分布）を選択する．入力フレームごとにまずモノフォンのすべての HMM 状態の尤度を計算し，その上位 k 位の状態についてのみ，対応するトライフォンの状態の尤度を計算する．k 位以下の選択されなかったトライフォンの状態については，その対応するモノフォンの状態の確率をそのまま割り付ける．この手法は従来の GS と比べて，非選択状態の尤度について実際のモノフォン状態の尤度を割り付けるため，選択の結果除外された状態についても妥当な近似値を与えることができ，選択誤りに頑健である．さらに，混合分布単位で選択することで混合分布の高速計算手法である Gaussian pruning を容易に適用することができ，さらに計算量の削減が見込める．これらの特徴により，選択誤りが発生しやすいより厳しい条件下においても安定した認識性能が見込まれる．

以下，2 章で従来の GS の概要と問題点について述べ，3 章で音素環境独立 HMM を用いた状態単位の混合ガウス分布選択法を提案する．4 章では標準的な GS との認識性能の比較実験を行い，手法の有効性を示す．また予備選択用モノフォンの規模と精度の関連や Gaussian pruning との統合，音響モデルとして PT を用いる際の最終的な認識性能についても評価した結果を示す．最後に 5 章で結論を述べる．

2. ガウス分布予備選択法

2.1 概　　要

ガウス分布予備選択法（GS）は，代表的な音響尤度計算の高速化手法の 1 つである．一般に，ガウス分布は対象ベクトルが分布の中心から遠いときその出力確率は大幅に減衰する．この性質より，混合ガウス分布を用いた音響モデルにおいては，入力ベクトルの近傍にあって出力確率の高いガウス分布のみが最終的な HMM 状態の尤度に支配的影響を持ち，入力ベクトルから遠いガウス分布の値は状態の尤度にほとんど影響しない．よって，認識時に入力ベクトルに対して遠い位置にあるガウス分布をあらかじめ計算から除外することで，認識率に影響を与えずに計算量を削減できる．GS は，この性質を利用して，各入力ベクトルに対する近傍のガウス分布集合をあらかじめ選択する手法である．

標準的な GS 法はベクトル量子化に基づくもので，Bocchieri によって提案された[1)]．あらかじめ音響特徴量空間を均一に量子化する VQ コードブックを作成し，それに従って空間を領域に分割して音響モデル内の各ガウス分布を 1 つ以上の VQ コードブックの要素（コードワード）にクラスタリングする．認識の際には，フレームごとに入力ベクトルをコードブックに従って量子化し，そのコードワードに対応するクラスタ内のガウス分布のみを計算する．音響尤度計算量の削減量はクラスタのサイズに依存し，サイズが小さいほど計算量は少なくなるが選択誤りが生じやすくなる．

2.2 ガウス分布のクラスタリング

ガウス分布のクラスタリングの際は，クラスタ境界における連続性を考慮したクラスタリングを行う[1)]．ガウス分布を単純に最も近い単一のコードワードに属するとした場合，各クラスタは互いに素な状態となる．このとき，クラスタ領域境界近くの入力に対する近傍分布集合は複数クラスタにまたがるため，クラスタ選択によって近傍にあるが計算されないガウス分布が少なからず存在することになる．これを避けるため，ガウス分布は以下の要領でクラスタ間で共有される．あるガウス分布がどのクラスタに属するかは，各領域の

中心（コードワード）との距離のしきい値で決定する．特に本研究では，分散で重み付けしたユークリッド距離に基づくクラスタリング[3]をベースラインとする．ガウス分布 $G(m)$ が平均 μ_m，対角共分散 Σ_m を持つとき，これがあるクラスタ ϕ に属するかどうかは以下の条件式によって決まる．

$$\frac{1}{K}\sum_{k=1}^{K}\frac{(c_\phi(k)-\mu_m(k))^2}{\sqrt{\sigma^2_{avg}(k)\sigma^2_m(k)}} \leq \Theta \qquad (1)$$

ただし K は特徴量ベクトルの次元数，c_ϕ はクラスタの中心ベクトル（コードワード），$\sigma^2_{avg}(i)$ は音響モデル中の全ガウス分布の平均の対角共分散行列の i 番目の対角要素，$\sigma^2_m(i)$ は Σ_m の i 番目の対角要素，Θ は距離のしきい値である．

2.3 問題点

GS による予備選択法は，有望な分布集合をあらかじめ絞り込む枝刈りに基づくアプローチである．しかし，1 フレームごとにその入力ベクトルのみの情報を元に計算対象の絞り込みを行うため，あるフレームにおいては尤度が低いが最終的に言語的制約も含めて最尤仮説候補になるような仮説パスが，局所的に低い音響尤度を持つ場合に途中で失われてしまう可能性がある．特に，選択されたクラスタ内に選択されたガウス分布が 1 つも含まれないような非選択の HMM 状態については，値がまったく得られないため，その時点で必ず枝刈りされてしまう．

このような値が得られない非選択の HMM 状態については，適当な固定値を与えてフロアリングすることである程度精度が改善される．しかし，実際の音声入力の尤度を反映しない単純な固定値を用いることは明らかに不十分であり，高い認識精度を保つことはできない．この予備選択による認識率の劣化と計算量の削減はトレードオフの関係にあり，特にクラスタサイズが小さい場合に認識率の悪化が顕著となる．

この選択誤りとフロアリングの問題は，クラスタリングに基づく GS 法のすべてに存在する．この問題はさらに，より高速な認識を目指して選択ガウス分布数（クラスタサイズ）をより小さくした場合に顕著となる．この問題に対してこれまでに，混合分布中の各ガウス分布がなるべく異なるクラスタに分散して属するように分布をクラスタリングすることで値なしの状態を現れにくくする[3]といった改善手法が提案されているが，根本的な解決には至っていない．

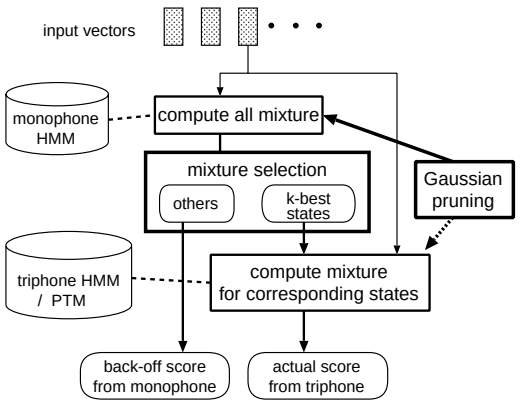

図 1 音素環境独立 HMM を用いた混合ガウス分布選択
Fig. 1 Gaussian Mixture Selection using context-independent HMM.

3. 音素環境独立 HMM を用いた状態単位の混合ガウス分布選択

3.1 状態単位の混合ガウス分布選択法

音素環境依存のトライフォンモデルでのガウス分布選択法において，ガウス分布をクラスタ化するかわりに，より荒い音素環境独立のモノフォンモデルを用いてそのトライフォンモデルとの階層的対応関係を元に HMM 状態単位で選択を行う手法を提案する．**図 1** に手法全体の概要を示す．まず使用するトライフォンモデルについて対応するモノフォンモデルを準備し，音素ごとにモデル間の状態ごとの対応付けを行っておく．このモノフォンモデルは，それを中心音素とするトライフォンと音素ごとに同一の状態数を持ち，かつ同一のデータで学習されていることが必要である．認識実行時には，入力フレームごとにまずすべてのモノフォン状態の尤度を計算し，上位 k 位の状態を決定する．その後，トライフォンの状態のうち対応するモノフォンの状態がその上位 k 状態内に存在するものについてのみ，その本来のトライフォンの混合ガウス分布の出力確率を計算する．対応するモノフォン状態が上位 k 状態にないトライフォン状態については，その対応するモノフォン状態の尤度自身を直接割り付けて近似値とする．この提案手法は，選択を HMM 状態に割り付けられている混合ガウス分布単位で行う．本論文ではこの提案手法を，混合ガウス分布選択法（Gaussian Mixture Selection，以下 GMS）と呼ぶ．

GMS の大きな特徴は，非選択状態に対してもモノフォンの出力確率から妥当な尤度値を割り付けられる点にある．モノフォンモデルは通常の音響モデルと同様に最大尤度基準に従って学習されており，単なる固

定値や他の ad-hoc な計算値によるフロアリングに比べてトライフォンの尤度のより良い近似として働く．トライフォン状態の尤度は最悪でもモノフォンの精度を持つこととなる．このため，高速化のためにより選択数を絞った厳しい条件においても認識精度の劣化は小さく抑えられると見込まれる．

もう 1 つの利点は，実装が簡潔かつ容易なことである．従来の GS 法では，あらかじめ音響特徴量空間全体に対してサイズが数百から数千の最適な VQ コードブックを学習し，かつそれを用いて膨大な数のガウス分布をクラスタリングしておく必要がある．これに対して，GMS では同じ状態数のモノフォンモデルを学習しておくだけでよい．また，実際にはトライフォンの学習の初期過程でそのようなモノフォンが生成されることが多く，特別な学習が不要であるケースが多い．また認識システム上への実装についても，既存の認識システムの Viterbi 処理に少量の変更を加えるだけで実装可能である．

3.2 Gaussian pruning との統合

提案手法の問題点として，予備選択のために新たに 1 フレームごとにモノフォンの全状態の尤度を計算する必要がある点があげられる．予備選択のための計算量が予備選択によって削減される計算量を上回ってしまうと，本手法を用いる実効的な意味が失せることになる．特に，予備選択の精度を向上するには詳細なモノフォンモデルを用いることが有効であると考えられるが，混合分布数の多いモデルを使用すると予備選択のための尤度計算はさらに増大する．

この荒い照合に基づく予備選択のアプローチの先行研究として，小森らによる少数分布 HMM の出力確率推定に基づく手法[5),6)] がある．これは少数分布と多数分布のモノフォン HMM を用いて 2 段階に出力確率を計算する手法であり，計算量を 60%〜70%削減している．ただし文献 5) の実験（6 分布 HMM で選択，上位 10 状態を 24 分布 HMM で再計算）は 1000 語程度の比較的小規模なタスクであったが，それにおいても 1 フレームあたり計算される分布数 690 のうち 450 と約 65%もの計算が予備選択に必要であった．本研究は，トライフォンを用いる数万語の大語彙タスクにおいてこの予備選択のアプローチを適用するものであるが，大量のトライフォンから効率良く選択を行うにはより高精度な選択用モノフォンモデルが必要であり，前処理量の増大から予備選択の効果が弱くなると考えられる．

この問題を解決するため，予備選択の計算に Gaussian pruning[7)] を導入する．Gaussian pruning は混合ガウス分布の出力確率計算量の削減手法の 1 つで，混合分布内の上位 k 個のガウス分布のみを効率良く計算する手法である．具体的には，各次元ごとのベクトル要素の距離を足し込んでいく過程で有望でない分布を動的に枝刈りする．これを予備選択のモノフォンの尤度計算に適用することで，予備選択の計算量を抑えることができ，トータルで効率の良い音響尤度計算を実現する．

さらに，本提案手法では状態単位で予備選択を行うため，選択されたトライフォン状態の混合分布計算にも Gaussian pruning を適用することが可能である．選択後のトライフォン状態の混合分布に対して Gaussian pruning を適用することで，さらに計算量を抑えられる．

このように提案手法と Gaussian pruning を組み合わせることで，状態ごとに適切な尤度を与えながら計算量を削減することができるようになる．すなわち，尤度が低いと見込まれるトライフォン状態に対してはモノフォン状態の分布を，高いと見込まれる状態にはトライフォン自身の混合分布をそれぞれ用いて，さらに Gaussian pruning によってその中の代表的なガウス分布のみを計算する．各状態に対して必ずなんらかの入力に基づく尤度が割り当てられるので，認識誤りの小さい安定した性能が得られる．これに対して従来の GS 法では枝刈りを行うと状態内の分布が 1 つも計算されない値なしの状態が頻出し，全体の精度を劣化させる要因となりやすい．

4. 実験的評価

提案したガウス混合分布選択法（GMS）を認識実験によって評価した．従来の標準的なガウス分布選択法（Standard GS，以下 SGS）と比較し，計算量と精度の比較を行う．各手法をそれぞれ大語彙連続認識エンジン Julius[8)] に実装して評価実験を行った．なお Julius のバージョンは 3.2 である．

4.1 実験条件

タスクは 2 万語の JNAS 新聞記事読み上げコーパスである．言語モデルは単語 3-gram を用いる．音響モデルは状態共有トライフォンモデル，およびトライフォン間で状態位置ごとに分布集合を共有する PTM[9)] を用いる．どちらも性別依存とする．前者は 2000 状態でそれぞれ 16 混合分布を持つ．後者は 43 音素，各 3 状態で計 129 個の分布集合（1 個あたり 64 混合分布）を持ち，それを 3000 状態のトライフォンで重みを変えて共有する．状態共有トライフォンを用いて SGS と GMS の性能比較を行い，PTM は最終的な認識性

表 1 SGS 用クラスタの平均サイズ
Table 1 Average cluster size for SGS.

thres	avg. num of Gauss.
0.9	887
1.1	1503
1.3	2325
1.5	3356
1.7	4588
1.9	5997
2.1	7548
2.3	9198

num of clusters = 1119
total num of Gaussians = 32000

能と計算量の評価に用いる．これらのモジュールはすべて「IPA 1999 年度版日本語ディクテーションツールキット」[10] として入手可能である．テストセットは 23 名の女性話者による 100 文の読み上げ発声である．

実験に先だって，SGS のための VQ コードブックを作成し，それに基づいて状態共有トライフォンのクラスタリングを行った．コードブックのサイズは 1119 である．各距離のしきい値（式 (1) の Θ）に対するクラスタごとの平均ガウス分布数を**表 1** に示す．GMS では予備選択用として 16 混合分布のモノフォンモデルを用いる．これは 43 音素で各音素が 3 状態を持つモデルであり，GMS ではこの合計 129 個の状態から上位 k 個に対応するトライフォン状態を選択することになる．

音響尤度計算量の評価尺度は，1 フレームあたり計算されるガウス分布数の平均とする．この分布計算率 $\%Gauss$ を以下のように定義する．

$$\%Gauss = \frac{G_{sel} + G_{pre}}{G_{full}} \times 100 \ \ (\%) \qquad (2)$$

ただし G_{sel} は予備選択によって選択され実際に計算されたトライフォンのガウス分布数の数，G_{full} は予備選択を行わない場合に計算されるガウス分布数，G_{pre} は予備選択のための計算コストである．予備選択を行わない場合の $\%Gauss$ を 100 とし，値が小さいほど計算量の削減幅が大きいことを表す．ここで，G_{full} はモデルの持つ総ガウス分布数ではない点に注意されたい．実際の認識処理では，モデル内で定義されていても探索過程でビームから外れて実際には計算されないガウス分布が少なからず存在する．本研究では予備選択法による音響尤度計算量の実効的な削減量を測るべく，予備選択を行わない状態でのガウス分布計算数を 100 として評価を行う．

G_{pre} は，SGS ではベクトル量子化の計算量をガウス分布の尤度計算量に換算した値となる．ベクトル量子化については，効率良くコードワードを決定する手法は多く存在するが，ここでは最も単純に入力ベクトルの各コードワードからの距離をすべて計算して最も短い距離のコードワードとする方法をとる☆．この場合コードブックのサイズと G_{pre} は等しくなる．GMS では G_{pre} は選択用モノフォンモデルの計算量であり，Gaussian pruning によって実際に計算された計算量を分布数に換算する．

表 2 状態共有トライフォンにおける手法の比較
Table 2 Comparison of methods with tied-state triphone.

GS method		#Gauss. G_{sel}	#Gauss. G_{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

SGS param.: cluster distance threshold
GMS param.: num of monophone states to select

4.2 GMS vs. SGS

まず提案手法である GMS の性能を従来法の SGS と比較する．音響モデルは状態共有トライフォンを用い，GMS では予備選択用モノフォンの尤度計算にのみ Gaussian pruning を用いる．なお本実験ではトライフォンに対する Gaussian pruning は行わず，GMS でのトライフォンの計算方法は SGS と同一である．SGS の非選択状態に対するフロアリング値は -50.0（常用対数尤度）とした．

両手法において，選択数を変化させたときのガウス分布計算量と認識誤り率を**表 2** に示す．表の GS method の欄の値は選択する分布数を決定するパラメータであり，SGS ではクラスタリングの距離のしきい値，GMS では選択したモノフォン状態数である．予備選択をまったく行わない（no-GS）ときに計算される分布数は 15772 であり，これを 100 として分布計算率 $\%Gauss$ を求めた．さらに計算量と認識精度の関係を**図 2** に示す．

提案した GMS は，十分な量を選択する緩やかな条件下では従来の SGS と同等の性能を示し，さらに選択するガウス分布数を絞り込んだ条件下においてはより安定した性能が得られた．図 2 にあるように，SGS

☆ VQ 処理の高速化法については部分距離探索法（partial distance search）[11] が Gaussian pruning とほぼ同等であり，これを導入した場合 VQ の計算量をほぼ 1/4 に減少できるとされている．

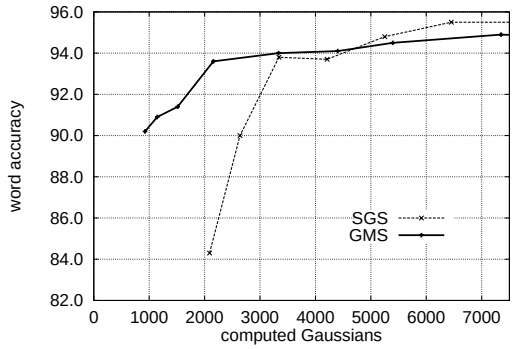

図 2 SGS と GMS の予備選択精度の比較
Fig. 2 Selection performance of SGS and GMS.

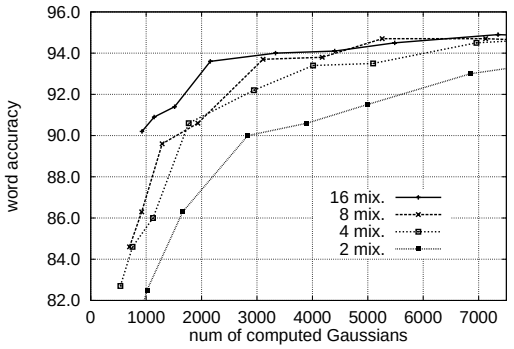

図 3 予備選択用モノフォンの比較
Fig. 3 Comparison of selection models.

では計算ガウス分布数を 3000 近くまで絞り込んだときに認識精度が急峻に悪化するのに対して，GMS の精度劣化は緩やかであり，2000 まで落としても精度が大きく変化することはなかった．これは，GMS では非選択であった状態に対してもモノフォンの尤度を元に信頼性のある値を割り付けるため，選択数を抑えた状況下でも認識率が劣化しにくいことを示している．

予備選択のための計算量についても，GMS の計算量は SGS と比較してより少ない計算量であった．選択に用いたモノフォンモデルは 129 状態，16 混合で計 2064 ガウス分布を持つが，Gaussian pruning を予備選択に導入することで実際に計算される演算量をガウス分布の尤度計算に換算して 690 と約 3 分の 1 に抑えることができた．

4.3 予備選択用モノフォンモデルの比較

次に，様々な精度の予備選択用モノフォンモデルを比較した．混合分布数が 16，8，4，2 の各モノフォンモデルを用いて予備選択を行ったときの，音響尤度計算量と認識精度の変化を**図 3** に示す．認識精度を保つには，混合数の多い高精度なモノフォンが必要であることが確かめられた．なお高精度モノフォンモデルでも実際の計算量の増大は小さく，精度に対する貢献のほうが大きかった．これは Gaussian pruning による効果が高く，詳細なモノフォンであっても予備選択の音響尤度計算量が相対的に抑えられたことによる．実際に各モノフォンモデルの 1 フレームあたりのガウス分布計算数を調べた結果を**表 3** に示す．Gaussian pruning の効果はモデルが詳細であるほど大きくなっていることが分かる．

モデル間の精度の相違について調べたところ，実際に選択される状態集合についてはモデル間に大きな差異が見られず，精度の低いモデルでもほぼ同様の選択が行われていた．このことから，精度の差は主に非選択状態に対する近似尤度の劣化に起因すると考えられる．以上より，予備選択による計算量削減においては，枝刈りされた非選択状態に対しても妥当な近似尤度を与えることが重要であることが示され，提案手法の正当性が確認された．

表 3 予備選択用モノフォンの計算量
Table 3 Computational cost of monophones for selection.

model	total #Gauss.	computed #Gauss.	rate
16mix	2064	690	33%
8mix	1032	465	45%
4mix	516	300	58%
2mix	258	191	74%

表 4 PTM における GMS の効果
Table 4 Effect of GMS on PTM.

method	#Gauss. G_{sel}	#Gauss. G_{pre}	total $\%Gauss$	word %err.
PTM	8192	-	100.00	5.9
+Gprune	1724	-	21.05	5.9
+GMS	434	690	13.61	6.0

selection model: 16 mix. monophone, select 16-best states

4.4 PTM での性能評価

トライフォンモデルとして PTM を用いたときの GMS の性能を**表 4** に示す．予備選択用モノフォンは 16 mix を使用し，上位 16 状態を選択した．まず GMS を行わずに PTM の尤度計算にのみ Gaussian pruning を行うことで，音響尤度計算量は 21.05%にまで削減される．そこから GMS を施すことで，認識率を下げずに計算量を 13.6%まで削減することができた．

最後に，提案手法を用いた認識システムの最大性能を計測した．この実験のみ，テストセットを男女 46 名，200 文に拡張して評価した．種々の探索パラメータは，実時間以下の処理時間で認識精度を最大化するよう最適化を行った．結果を**表 5** に示す．GMS 導

表 5 システム性能
Table 5 System performance.

	word %err.	time (xRT)
without GMS	9.6	1.06
	7.8	1.42
with GMS	7.8	0.99

model: PTM, CPU: Pentium 850 MHz, OS: Linux

入前は実時間処理で 9.6%の単語誤り率であったのに対して，GMS の導入によって処理が高速化できた結果，ビーム幅を広くとることができるようになり，実時間で認識誤り率 7.8%を達成することができた．なお GMS 導入前に同じビーム幅で認識を行った場合の処理時間は実時間の 1.42 倍であったことから，GMS によって実際の全体のデコーディング速度を約 30%高速化できたことが分かる．音響尤度計算のみの改善幅（21.05%から 13.61%）に対して認識処理全体の速度改善が大きかったが，これはモノフォンの尤度が割り当てられる PTM の状態では，混合分布全体の尤度を求めるための各ガウス分布の重み付き尤度和の計算が省略できる効果が大きかったものと考えられる．

5. おわりに

音響尤度計算の高速化のための効率の良い混合ガウス分布予備選択法を提案した．この手法は音素環境非依存のモノフォンモデルの状態の尤度に基づいて対応するトライフォン状態を選択するとともに，非選択状態に対してもモノフォンの尤度を近似尤度として与える．選択されなかった状態に対しても妥当な近似値を与えることで，従来の VQ コードブックに基づくガウス分布選択法に比べ安定して動作し，認識精度の劣化も小さい．PTM モデルや Gaussian pruning との組合せにおいて，提案手法を施すことで，認識精度を下げずに音響尤度計算量を適用前の約 20%から 13.61%と，さらに 32%削減することができた．また実際のデコーディングの処理速度を約 30%高速化することができた．

本研究の成果は，大語彙連続音声認識エンジン Julius バージョン 3.2 の機能として，情報処理学会音声言語情報処理研究会「連続音声認識コンソーシアム」を通じて一般に公開されている．

謝辞 音響モデル，言語モデルは情報処理振興事業協会（IPA）の「日本語ディクテーション基本ソフトウェア 99 年度版」のものを使用した．

参考文献

1) Bocchieri, E.: Vector Quantization for Efficient Computation of Continuous Density Likelihoods, *Proc. IEEE-ICASSP*, pp.692–695 (1993).
2) Knill, K.M., Gales, M.J.F. and Young, S.J.: Use of Gaussian Selection in Large Vocabulary Continuous Speech Recognition using HMM's, *Proc. ICSLP*, pp.470–473 (1996).
3) Gales, M.J.F., Knill, K.M. and Young, S.J.: State-Based Gaussian Selection in Large Vocabulary Continuous Speech Recognition using HMM's, *IEEE Trans. Speech & Audio Process.*, Vol.7, No.2, pp.152–161 (1999).
4) Paul, D.B.: An Investigation of Gaussian Shortlists, *Proc. IEEE workshop on Automatic Speech Recognition and Understanding* (1999).
5) 小森康弘，山田雅章，山本寛樹，大洞恭則：小数分布 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) 小森康弘，山田雅章，山本寛樹，大洞恭則：Rough HMM と Detail HMM を用いた連続 HMM 出力確率計算の高速化，Vol.1-Q-20, pp.135–136 (1995).
7) 李 晃伸，河原達也，武田一哉，鹿野清宏：Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識，電子情報通信学会論文誌，Vol.J83-D-II, No.12, pp.2517–2525 (2000).
8) Lee, A., Kawahara, T. and Shikano, K.: Julius — An Open Source Real-Time Large Vocabulary Recognition Engine, *Proc. European Conf. on Speech Communication and Technology*, pp.1691–1694 (2001).
9) Lee, A., Kawahara, T., Takeda, K. and Shikano, K.: A New Phonetic Tied-Mixture Model for Efficient Decoding, *Proc. IEEE Int'l Conf. Acoust., Speech & Signal Process.*, pp.1269–1272 (2000).
10) Kawahara, T., Lee, A., Kobayashi, T., Takeda, K., Minematsu, N., Sagayama, S., Itou, K., Ito, A., Yamamoto, M., Yamada, A., Utsuro, T. and Shikano, K.: Free Software Toolkit for Japanese Large Vocabulary Continuous Speech Recognition, *Proc. Int'l Conf. on Spoken Language Processing*, Vol.4, pp.476–479 (2000).
11) Bei, C. and Gray, R.M.: An improvement of the minimum distortion encoding algorithm for vector quantization, *IEEE Trans. Communications*, Vol.33, No.10, pp.1132–1133 (1985).

(平成 13 年 11 月 20 日受付)
(平成 14 年 4 月 16 日採録)

李　　晃伸（正会員）

平成 8 年京都大学工学部情報工学科卒業．平成 10 年同大学大学院修士課程修了．平成 12 年同大学院情報学研究科博士課程修了．同年より奈良先端科学技術大学院大学情報科学研究科助手．主として音声認識・理解の研究に従事．博士（情報学）．平成 14 年日本音響学会粟屋潔学術奨励賞受賞．日本音響学会，電子情報通信学会各会員．

河原　達也（正会員）

1987 年京都大学工学部情報工学科卒業．1989 年同大学大学院修士課程修了．1990 年同博士後期課程退学．同年京都大学工学部助手．1995 年同助教授．1998 年同大学情報学研究科助教授．現在に至る．この間，1995 年から 96 年まで 米国ベル研究所客員研究員．1998 年から ATR 客員研究員．1999 年から国立国語研究所非常勤研究員．2001 年から科学技術振興事業団さきがけ研究 21 研究者．音声認識・理解の研究に従事．京都大学博士（工学）．1997 年度日本音響学会粟屋賞受賞．2000 年度情報処理学会坂井記念特別賞受賞．情報処理学会連続音声認識コンソーシアム代表．電子情報通信学会，日本音響学会，人工知能学会，言語処理学会，IEEE 各会員．

鹿野　清宏（正会員）

昭和 45 年名古屋大学工学部電気工学科卒業．昭和 47 年同大学大学院修士課程修了．同年電電公社武蔵野電気通信研究所入所．昭和 59 年～61 年カーネギーメロン大学客員研究員．昭和 61 年～平成 2 年 ATR 自動翻訳電話研究所音声情報処理研究室長．平成 4 年 NTT ヒューマンインタフェース研究所主席研究員．平成 6 年より奈良先端科学技術大学院大学情報科学研究科教授．音情報処理学講座を担当．工学博士．主として音声・音情報処理の研究および研究指導に従事．昭和 50 年電子通信学会米沢賞，平成 3 年 IEEE SP 1990 Senior Award，平成 6 年日本音響学会技術開発賞，平成 12 年情報処理学会山下記念研究賞，平成 13 年 VR 学会論文賞．IEEE，ISCA，音響学会，電子情報通信学会，VR 学会各会員．