音声言語情報処理 29－7
(1999. 12. 20)

不特定話者混合分布HMMにおける Tree-Basedクラスタリングの検討

加藤恒夫† 黒岩眞吾† 清水徹† 樋口宜男†

†KDD研究所
〒356-8502 埼玉県上福岡市大原2-1-15

あらまし Tree-basedクラスタリングは，音素コンテキストを分割条件としてトライフォンの集合に対してクラスタリングを行いHMM状態の共有化を図る有効な手法である．従来の報告では，計算量の点から対象が単一分布HMMに限られていた．しかし，単一分布HMMでは不特定話者の音響的特徴を表現するのに不十分であるため，必ずしも適切なトポロジ（HMM状態の共有関係）が得られていないと考えられる．また所望の混合分布トライフォンを得るためには，tree-basedクラスタリングの後，混合数を倍増する操作と学習を繰り返すため膨大な時間を要する．そこで，本稿では混合分布トライフォンに対して分布のクラスタリングを行いながらtree-basedクラスタリングを行う手法を提案する．本手法により学習時間が1/3程度に短縮され，認識実験では1～2%の音素正解精度の改善を確認した．

キーワード 音声認識，音響モデル，トライフォン，混合分布，クラスタリング

A Study on Tree-Based Clustering for Speaker-Independent Gaussian Mixture HMMs

Tsuneo Kato† Shingo Kuroiwa† Tohru Shimizu† Norio Higuchi†

†KDD R&D Laboratories Inc.
2-1-15, Ohara, Kamifukuoka, Saitama 356-8502

Abstract Tree-based clustering is an effective method to share HMM states by clustering triphones based on phonetic questions. Previous researches on this method have been made on HMMs of single Gaussian output distributions due to computational restrictions. However, single Gaussian HMMs may not be sufficient to create appropriate topology (i.e. HMM state sharing). Furthermore, a significant amount of time is required to obtain Gaussian mixture HMMs for repetitive distribution splitting and embedded training. In this paper, we propose a tree-based clustering for Gaussian mixture HMMs based on distribution clustering. This method achieved 67% reduction on training time and 1-2% improvement in phoneme accuracy.

key words speech recognition, acoustic modeling, triphone, Gaussian mixture, clustering

1. はじめに

大語彙音声認識では，音素などのサブワードHMMを連結して単語を構成している．音素の音響的特徴は前後の音素コンテキストにより大きく変化することが知られ，トライフォンが優れた認識性能を示している．しかし，3音素連鎖で区別されるトライフォンの総数は10,000を超え，学習時には，データ量が不足して頑健性が低下するトライフォンや学習データ中に全く出現しないため推定できないトライフォンが現れる．こうした問題を防ぐため，一般には音響的に近いトライフォンの間でパラメータの共有を図っている．

Tree-basedクラスタリング[1, 2, 3]は，音素コンテキストを分割条件として，中心音素が共通のトライフォンの集合をトップダウンにクラスタリングし，リーフノードに含まれるトライフォンがHMM状態を共有することで，HMM状態あたりの学習データ量を増やす有効な手法である．

Tree-basedクラスタリングでは，学習データに対して尤度等の評価尺度を求め，分割前後の変化量が最大になることを基準に分割条件の選択を行う．通常，この計算には，出力確率分布の平均値，分散と学習データ中の出現頻度のみを利用した近似式が用いられる．混合分布の場合，分布間の重なりを考慮すると計算量が爆発的に増大するため，従来の報告[4, 5, 6, 7]ではいずれも単一分布HMMを対象としている．一方，不特定話者音声認識には“混合”分布状態共有トライフォンが用いられることが多い．これは，不特定話者の音響的特徴を表現するのに単一分布ではパラメータが不十分なためである．このことから，単一分布HMMにおけるtree-basedクラスタリングでは必ずしも適切なトポロジ（ここではHMM状態の共有関係を指す）が生成されていないことが考えられる．さらに，tree-basedクラスタリングの後，所望の混合分布トライフォンを得るには分布を倍増する操作と連結学習を繰り返す必要があるため，膨大な時間を要するという問題がある．そこで，本稿ではK-means法により混合分布HMMを扱えるようにしたtree-basedクラスタリングを提案する．本手法により，トポロジを改善し認識性能の向上を図るとともに，学習時間を短縮する．

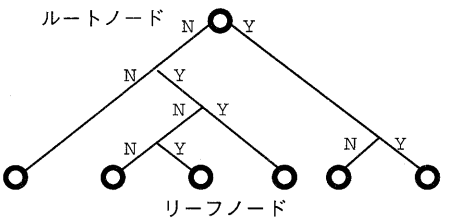

図1　決定木

2. 混合分布HMMにおけるTree-Basedクラスタリング

2.1 従来のTree-Basedクラスタリング

Tree-basedクラスタリングの分割条件は，音声学の知識に基づく先行音素または後続音素に関する二者択一の質問である．例えば，「先行音素が破裂音であるか？」というような質問であり，全てのトライフォンは”Yes”/”No”のいずれかに分類される．

手順としては，まず，中心音素が共通の全てのトライフォンを1つの集合とし，これをルートノードとする．次に，分割条件の1つ1つに従ってノードを仮分割する．それぞれの仮分割に対して尤度等の評価尺度の分割前後の変化量を求め，これが最大となる分割条件を選択してノードを分割する．この分割を繰り返すことにより，図1に示すような決定木を生成する．クラスタリング後，1つのリーフノードに含まれるトライフォンは1つのHMM状態を共有する．以上の操作により，出現頻度が少ないトライフォンにも，十分な学習データを確保したHMM状態を割り当てることが可能になる．また，学習データに出現しないトライフォンも，決定木を辿ることでいずれかのクラスタに含まれるため，HMM状態を共有することができる．

2.2 従来手法の問題点

分割条件の選択には，学習データに対する尤度，エントロピ等の評価尺度を求める必要がある．しかし，計算量を抑えるために，通常はトライフォンを表現するHMMの音響パラメータの平均，分散と学習データ中の出現頻度を用いた近似式を利用する．

Tフレームからなる学習データ$O_t\,(t=1,\cdots,T)$が与えられたとき，ノードS_mに対する尤度$L(S_m)$

は，

$$
\begin{aligned}
& L(S_m) \\
&= \sum_{t=1}^{T} log(N(O_t, \mu_{S_m}, \sigma_{S_m})) \cdot \gamma_t(S_m) \\
&\approx -\frac{1}{2}(K log 2\pi + log|\sigma_{S_m}| + K)\Gamma_m \qquad (1)
\end{aligned}
$$

で求められる．ただし，

$$\gamma_t(S_m) = \frac{\alpha_t(S_m)\beta_t(S_m)}{\sum_i \alpha_t(S_i)\beta_t(S_i)} \qquad (2)$$

$$\Gamma_m = \sum_{t=1}^{T} \gamma_t(S_m) \qquad (3)$$

であり，K はベクトルの次元数，$N(O, \mu, \sigma)$ は，平均ベクトルμ，共分散行列σの正規分布が観測ベクトル O を出力する確率を表す．また，$\alpha_t(S_m)$，$\beta_t(S_m)$ は，時刻 t，ノード S_m における前向き確率，後ろ向き確率を表し，$\gamma_t(S_m)$ は，時刻 t に S_m に存在する確率を表す．したがって，Γ_m はノード S_m の学習データ中の出現フレーム数を表す．

式 (1) では，ノードを単一分布 HMM として表現している．しかし，単一分布では不特定話者の音響的特徴を表現するのに不十分なので，必ずしも適切な分割条件の選択が行われていない，つまり適切なトポロジが生成されていないと考えられる．また，出力する状態共有トライフォンも単一分布 HMM であるが，実際の認識には，認識性能の優れる“混合”分布状態共有トライフォンを用いるのが一般的である．従って，所望の混合数の状態共有トライフォンを得るには，tree-based クラスタリングの後，混合数を倍増する操作と連結学習を何度も繰り返さなければならず，長い時間を要するという問題がある．

2.3 混合分布 HMM における Tree-Based クラスタリング

2.2で述べた問題点に対し，本節では混合分布 HMM に対して tree-based クラスタリングを行い混合分布 HMM を出力することで，性能改善を図るとともに学習ステップを単純化する方法を提案する．2.3.1で tree-based クラスタリングの各ノードにおいてノード中の分布集合から混合分布を生成する方法を説明し，2.3.2でこの混合分布を用いた近似尤度の計算式を示し，2.3.3で手順をまとめる．

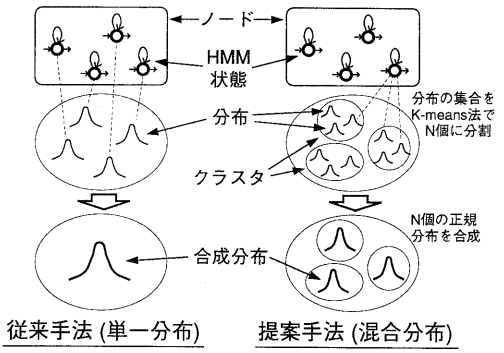

図 2　従来手法と提案手法の違い

2.3.1 K-means 法を用いた混合分布の合成

従来手法が，入力モデル，各ノードにおける近似尤度計算のための中間表現モデル，出力モデルとして，全て単一分布 HMM を用いるのに対し，提案手法では全て混合分布 HMM を用いる．混合分布は，学習データに対してできるだけ大きい尤度を出力するように生成することが望まれる．そこで，ここではノード中の各トライフォンを表現する混合分布をノードを構成する分布の集合としてとらえ，これを K-means 法により N 個にクラスタリングし，分布クラスタ毎に 1 つ分布を生成することで N 混合の分布を構成する．ここで，N はノードに含まれる全トライフォンの混合数の最大値である．従来手法と提案手法との違いを図 2に示す．

分布の K-means クラスタリングには，各分布の学習データ中の出現フレーム数Γが必要となる．Γは，トライフォンの出現フレーム数と分布重みの積で近似する．距離尺度は，次元毎にルートノードの分散で正規化したユークリッド距離を用い，分布クラスタのセントロイドは，各分布の平均ベクトルをΓで重み付けをして求める．各分布クラスタを表現する分布は，対角共分散行列をもつ正規分布を仮定し，式 (4),(5),(6) により平均，分散，分布重みを求める．(以下，この操作を分布の合成と呼ぶ．)

$$\mu_{i,k} = \sum_j \Gamma_{i,j} m_{i,j,k} \Big/ \sum_j \Gamma_{i,j} \qquad (4)$$

$$\sigma_{i,k} = \frac{\sum_j \Gamma_{i,j} v_{i,j,k} + \sum_j \Gamma_{i,j}(m_{i,j,k} - \mu_{i,k})^2}{\sum_j \Gamma_{i,j}} \qquad (5)$$

$$w_i = \sum_j \Gamma_{i,j} \Big/ \sum_i \sum_j \Gamma_{i,j} \qquad (6)$$

ここで，$m_{i,j,k}$，$v_{i,j,k}$，$\Gamma_{i,j}$は i 番目の分布クラスタに含まれる j 番目の分布の第 k 次の平均，分散，出現フレーム数を表し，$\mu_{i,k}$，$\sigma_{i,k}$，w_iは合成分布の平均，分散，分布重みをそれぞれ表す．

また，出力モデルとして，文献 [8] ではリーフノード中の分布のうち分散が最大の 1 つを選択して利用している．しかし，この方法ではリーフノード全体の音響的特徴を十分に表現しているとは考えられないため，提案手法では近似尤度の計算に用いた N 混合の合成分布を出力する．

2.3.2 学習データに対する近似尤度の計算

K-means 法を用いて混合分布の合成を行った後は学習データに対する近似尤度を求める．

混合分布の場合，分布間の重なりを無視せずに 2 つの分布の畳み込み積分を考慮すると計算量が増大する．ここでは，分布間の重なりが十分小さいと仮定し，混合分布の出力確率の和を最大値で近似した式 (7) を用いる．

$$\begin{aligned} L(S_m) &= \sum_{t=1}^{T}[log\sum_{n=1}^{N} w_{m,n}N(O_t,\mu_{m,n},\sigma_{m,n})]\cdot\gamma_t(S_m) \\ &\approx \sum_{t=1}^{T} max[log(w_{m,n}N(O_t,\mu_{m,n},\sigma_{m,n}))\gamma_t(S_{m,n})] \\ &\approx \sum_{n=1}^{N}[\Gamma_{m,n}log(\Gamma_{m,n}) - \frac{\Gamma_{m,n}}{2}(Klog2\pi + K \\ &\quad +log|\sigma_{m,n}|)] - \sum_{n=1}^{N}\Gamma_{m,n}\cdot log(\sum_{n=1}^{N}\Gamma_{m,n}) \qquad (7) \end{aligned}$$

ただし，$\mu_{m,n}$，$\sigma_{m,n}$，$w_{m,n}$，$\Gamma_{m,n}$はノード S_m の n 番目の分布の平均ベクトル，共分散行列，分布重み，出現フレーム数を表す．

2.3.3 Tree-Based クラスタリング手順

混合分布の合成と近似尤度の計算を組み合わせた提案手法の手順を以下に示す．(3) で混合分布の合成とこれを用いた尤度計算を行い，(4) で混合分布 HMM を出力する部分が従来手法と異なる．

(1) 混合分布状態共有無しトライフォンの学習

(2) 初期設定
中心音素が共通のトライフォン集合をルートノードとする．

(3) クラスタリング
末端にあるノード（最初はルートノード）について，尤度上昇分が予め与えた閾値より小さくなるまで i)～iii) を繰り返す．

i) ノードを表現する混合分布を合成し，分割前の近似尤度を計算する．

ii) 各分割条件について分割後の混合分布を合成し，近似尤度を計算する．

iii) 尤度上昇分が最大になるノードと分割条件の組合せにより実際にノードを分割する．

(4) 出力
リーフノード毎に混合分布状態 HMM を出力する．

3. 評価実験

従来手法と提案手法によって学習したモデルの認識性能を比較する場合，混合数を揃えても，その差が分布の生成方法の違いに起因するのか，トポロジの違いに起因するのかわからない．そこで，前者の効果を 3.1で，後者の効果を 3.2で検証する．最後に 3.3では，tree-based クラスタリングに続いて連結学習を行う実際のモデル学習過程において，最終的に性能の改善が得られるか検証する．

3.1 分布の合成による効果の検証

2.3.1で述べた分布の合成による効果を検証するため，単一分布 HMM の tree-based クラスタリングにおいて 2 種類の分布生成方法を比較する．

[実験方法]

単一分布 HMM の tree-based クラスタリングにおいて，リーフノード中の分布のうち分散最大の 1 つを選択する文献 [8] の方法と，式 (4)(5)(6) に従い単一分布の合成を行う提案手法で，2 種類の状態共有トライフォンを出力し，認識性能を比較する．実験に用いた学習データは，男性話者 1,057 名による電話入力の ATR 音素バランス文 9,000 発声である．音響分析条件を表 1に示す．評価用データは，男性話者 30 名による音素バランス文 495 発声を用い，音節タイプライタによる認識結果を音素正解精度 (%Acc) で評価した．

[実験結果]

音素正解精度を表 2に示す．分布を合成することにより音素正解精度が 7.6%上昇した．

表 1　音響分析条件

サンプリング周波数	8,000 Hz
フレーム周期	10 ms
フレーム長	25 ms
プリエンファシス	$1\text{-}0.95z^{-1}$
フィルタバンク数	20
特徴パラメータ	$MFCC$ 1-12 次 (CMS)
	$\Delta MFCC$ 1-12 次
	$\Delta\Delta MFCC$ 1-12 次
	$\Delta power$ 1 次
	$\Delta\Delta power$ 1 次

表 2　単一分布 HMM の tree-based クラスタリングにおける分布生成方法と音素正解精度

分布生成方法	分散最大の分布選択	分布の合成
音素正解精度	61.6 %	69.2 %

3.2　トポロジの改善による効果の検証

次に，混合数の異なる HMM に対する tree-based クラスタリングが生むトポロジの違いが認識性能に現れるか検証する．Tree-based クラスタリングの出力モデルで比較すると，3.1で検証した分布の生成方法による影響が混入してしまうので，ここではクラスタリング結果である決定木と別に用意した境界時刻付き状態ラベルを利用して，状態共有トライフォンの初期学習を行う．こうすることで混合分布の合成による効果を排除したモデルが得られる．

[実験方法]

まず，混合数の異なる HMM に対して tree-based クラスタリングを行う．次に，tree-based クラスタリングの条件毎に生成される決定木に基づいて，別に用意した学習用の境界時刻付き状態ラベルの各状態名（どのトライフォンの何番目の状態か）を参照先の状態名に置換する．これにより学習音声データのレベルで共有化が行われることになる．次に，参照先状態名に置換した境界時刻付き状態ラベルを利用して共通の初期学習方法により同一混合数の状態共有トライフォンを学習する．こうして得られた，学習データと学習手順が共通でトポロジのみ異なる状態共有トライフォンの認識性能を比較することで，トポロジの改善による効果を検証する．ここでは 1，2，4，8 混合の 4 種類の HMM に対して tree-based クラスタリングを行い，その決定木を利用して 4 種類の 4 混合状態共有トライフォンの初期学習を行った．学習データ，音響分析条件，評価方法は，3.1と共通である．

表 3　異なる混合数でクラスタリングを行った結果（トポロジ）に基づく初期学習モデルの音素正解精度

クラスタリング時混合数	1	2	4	8
音素正解精度 (%Acc)	71.3	71.5	71.7	71.6

[実験結果]

音節タイプライタの結果から求めた音素正解精度を表3に示す．上段は tree-based クラスタリングによりトポロジを作成したモデルの混合数を示している．混合分布 HMM における tree-based クラスタリングの方が，僅かであるが音素正解精度が高い．

3.3　連結学習後の認識性能比較

3.1，3.2の結果から，トポロジの改善は僅かであるが，分布の合成に効果があることを確認した．しかし，音声認識に用いるトライフォンは，通常，状態の共有化に続いて連結学習を繰り返して得られる．つまり，tree-based クラスタリングの結果，出力される状態共有トライフォンは，続く連結学習の参照モデルとなる．参照モデルが優れていても，連結学習後の音声認識に用いるモデルが従来手法によって学習したモデルに較べて認識性能が高いとは限らない．そこで，tree-based クラスタリングに引き続き連結学習を繰り返したモデルの認識性能を従来手法によって学習したモデルと比較する．

[実験方法]

従来手法によるモデルは，以下の手順で作成する．まず，単一分布の状態共有無しトライフォンを学習し，tree-based クラスタリングを行う．続いて連結学習を 3 回繰り返す．次に，分布を分割して倍増する操作によって状態あたりの混合数を 2 とした後，連結学習を 3 回繰り返す．さらに分布の倍増操作と 3 回の連結学習を繰り返し，状態あたりの混合数が 4，8 の状態共有トライフォンを順次作成する．

一方，提案手法によるモデルは，混合数が 2，4，8 の状態共有無しトライフォンを学習し，それぞれに対して tree-based クラスタリングを行った後，連結学習を 3 回繰り返し，3 種類の状態共有トライフォンを得る．本実験は，男女両モデルについて行ったが，男声モデルの学習データ，音響分析条件，評価方法は 3.1と共通であり，女声モデルは，学習デー

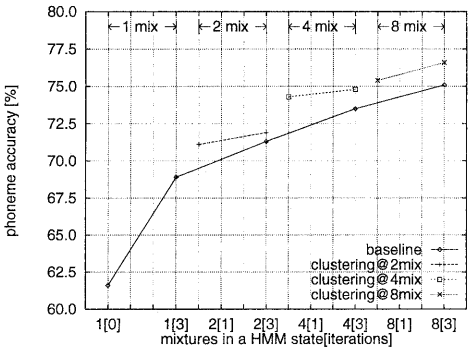

図 3　混合数 [連結学習回数] と音素正解精度 (男声)

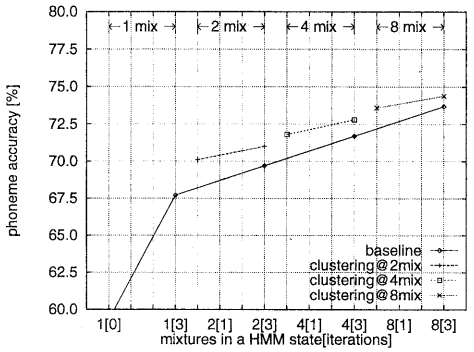

図 4　混合数 [連結学習回数] と音素正解精度 (女声)

タとして 506 名による 10,000 発声を用いた．評価データは女性話者 30 名による 498 発声を用いた．比較は，混合数と連結学習の回数が等しくなる条件で行う．

[実験結果]

図 3に男声モデル，図 4に女声モデルの音素正解精度を示す．横軸は状態あたりの混合数と連結学習回数（括弧内）を表す．図 3，4とも実線が従来手法による学習，破線が混合数 2，4，8 のときの提案手法による学習を表す．破線の左端は，tree-based クラスタリング直後の音素正解精度を表す．まず，提案手法により学習のステップ数が大幅に削減されていることがわかる．8 混合のトライフォンを学習する場合，従来手法では 1，2，4，8 混合において各 3 回，計 12 回連結学習を繰り返すのに対し，提案手法では 8 混合における 3 回だけである．この結果，tree-based クラスタリングに要する時間の増大を含めても学習時間は 1/3 程度に短縮された．音素正解精度は，男女とも全ての混合数において提案手法を用いて学習した場合が高く，連結学習後でも 1～2%優っていることがわかる．男声モデルの提案手法 4 混合の場合 (図 3の 4[3])，従来手法の 8 混合 (8[3],75.1%) に近い 74.9%が得られ，分布数の倍増に匹敵する改善となっている．

4.　まとめ

K-means 法による分布クラスタリングを用いた混合分布 HMM の tree-based クラスタリングを提案し，不特定話者電話音声モデルの学習に適用した．認識実験により，混合分布の合成に効果があり，連結学習を繰り返した後も，提案手法によって学習したモデルの認識性能は従来手法に較べて優れていることを確認した．また，提案手法により学習に要する時間が 1/3 程度に短縮された．

謝辞

本稿に関して熱心な御討論と有益な御助言を頂いた KDD 研究所音声応用グループ各位に、また、本稿執筆の機会を与えて頂いた KDD 研究所村谷所長に感謝します。

参考文献

[1] K.F.Lee et. al.: "Allophone Clustering for Continuous Speech Recognition," Proc.ICASSP 90, pp.749-753 (1990)

[2] L.R.Bahl et. al.: "Decision Trees for Phonological Rules in Continuous Speech," Proc.ICASSP 91, pp.185-188 (1991)

[3] S.J.Young et. al.: "Tree Based State Tying for High Accuracy Modeling," ARPA Workshop on Human Language Technology, pp.307-312(1994)

[4] 篠田浩一、渡辺隆夫: "情報量基準を用いた状態クラスタリングによる音響モデルの作成", 信学技法, SP96-79, pp.9-15 (1996)

[5] W.Chou and W.Reichl: "Decision Tree State Tying Based on Penelized Bayesian Information Criterion," Proc.ICASSP 99, pp.2048-2051 (1999)

[6] R.Singh, B.Raj and R.M.Stern: "Automatic Clustering and Generation of Contextual Questions for Tied States in Hidden Markov Models," Proc.ICASSP 99, pp.117-120 (1999)

[7] W.Reichl and W.Chou: "Decision Tree State Tying Based on Segmental Clustering for Acoustic Modeling," Proc.ICASSP 98, pp.801-804 (1998)

[8] S.Young: "HTKBook", Entropic Cambridge Research Laboratory