Vol.2010-SLP-82 No.12
2010/7/24

トップダウン及びボトムアップ手法に基づく音韻 HMM のクラスタリング

宮垣諒一[†]　　川端豪[†]

大語彙連続音声認識において，個々のトライフォン HMM を学習するための十分なデータを確保することは難しい．この問題を解決するために，様々なクラスタリング技術が研究されている．あるクラスタに属するすべてのトライフォンのデータを使って，一つの音韻 HMM を学習することによって，モデルの信頼性を向上させるという考え方である．クラスタリングの手法には，音韻文脈によってクラスタを分割していくトップダウン的な手法と，音響的に類似するトライフォンをまとめるボトムアップ的手法がある．本報告では両手法を併用することの有効性を，CSJ 講演データを用いて検討する．

Clustering of Phone HMMs based on Top-down and Bottom-up Approaches

Ryoichi Miyagaki[†]　and Takeshi Kawabata[†]

For large vocabulary speech recognition, it is difficult to collect sufficient training data for each triphone HMM. To cope with this problem, various clustering techniques have been researched. A HMM is trained using all triphone data belonging to a cluster. There are two clustering approaches, top-down and bottom-up methods. The top-down method divides a cluster into two small clusters based on the phonetic decision tree. The bottom-up method merges several clusters into a large cluster based on their acoustic similarity. Experimental results show that the combination of both approaches is effective for CSJ lecture tasks.

1. はじめに

近年，計算機の高速化，大規模音声コーパスの整備などに伴い，大語彙連続音声認識に関する研究が盛んに行われている．大語彙連続音声認識においては，音韻を単位として HMM(Hidden Markov Model)を学習することが一般的である．さらに音韻の音響的特徴は先行及び後続音韻の影響を大きく受けるため，3 音韻の連鎖（トライフォン）を単位とする HMM を学習することがよく行われている．

しかし，トライフォンには膨大な種類が考えられるため，学習データ中に十分な個数のトライフォンが含まれない状況が起こりうる．学習データに現れないトライフォンは原理的に認識できず，また出現回数が少なければ HMM の信頼性が低下するという問題が起きる．

この問題を解決するために，複数のトライフォンを集めてクラスタを作り，互いに代用することで見かけ上の学習データを増やす手法が研究されてきた．音韻文脈に基づくトップダウン的な情報を用いてクラスタリングを行う TBC（Tree Based Clustering）が有名である[1][2][3][4]．TBC では HMM を構成する状態ごとにクラスタリングを行う．ある状態に対し，同じ中心音韻を持つトライフォン全体を大きな一つのクラスタと考え，音韻文脈や学習される HMM の統計的信頼性を手がかりにクラスタを分割していく．一つのクラスタに含まれるトライフォンはその状態を共有する．この操作を HMM の各状態について行う．

TBC クラスタの数と認識率に関する報告があり，ある状態数で精度がピークに達するが，その後は悪化することが知られている[5]．これは，文脈的に近くても音響的に類似しないトライフォンが同じクラスタに分類されることによって，精度に悪影響を及ぼす可能性を示唆している．そこで，本報告ではトップダウン的クラスタリングと，音響的に類似するトライフォンをまとめるボトムアップ的クラスタリングを組み合わせることによって，より高精度の音響モデルを学習する．

2. トップダウンとボトムアップに基づく音韻 HMM クラスタリング

2.1 クラスタリングの必要性

大語彙連続音声認識では音韻（phoneme）単位で HMM を作ることが一般的である．例えば，連続音声中の「秋」という単語を認識するためには，/a/，/k/，/i/という各音韻の HMM を連結して確率を計算すればよい．しかし，音韻の音響的特徴は前後の音韻によって大きく影響を受ける．同じ音韻/k/でも「秋（/a k i/）」の/k/と「駅(/e k i/)」の/k/では先行音韻が異なるので舌及び顎の位置が異なり，音響的に違いが現われる．

[†] 関西学院大学　理工学研究科
School of Science and Technology, Kwansei Gakuin University

そこで，音韻環境を考慮した認識単位として，3 音韻連鎖（トライフォン）が用いられる．

認識単位としてトライフォンを用いる際の問題点は，その種類の多さである．例えば日本語の場合，音韻の種類数を仮に 40 とすると，トライフォンの総数は 64000（40×40×40）もの数になってしまう．このためトライフォンあたりの学習データが少なくなり HMM の統計的信頼性が低下する．最悪の場合，学習データに現れなかったトライフォンについては学習が行えない．

そこで，いくつかのトライフォンをグループ化（クラスタリング）し，一つの HMM を学習することによってモデル数を削減することが考えられる．例えば，トライフォンを音韻文脈の類似性に従って段階的に分割するトップダウン的考え方に基づくグループ化手法が研究されている．また，音響的に類似するトライフォンをボトムアップ的にまとめていく考え方もある．トップダウン的音韻クラスタリングにおいては，先行・後続音韻の場合分けでクラスタリングを行うので，音響的に類似したトライフォンが同じクラスタ内にあるとは限らない．このことが認識精度に悪影響を及ぼす可能性がある．

本研究では，音響的に類似するトライフォンを探し，それらを一つのグループとするボトムアップ的音韻クラスタリング手法を導入する．ボトムアップ手法のみでは出現頻度の低い，または出現しないトライフォンに対応することができないため，トップダウン的手法とボトムアップ的手法を組み合わせる．

2.2 Tree Based Clustering

Tree Based Clustering(TBC)は，先行・後続音韻を考慮し音響的に類似した環境を木構造で構成してトップダウン的にクラスタ化を行う手法である[1][2][3][4]．TBC の分割条件は，音声の特徴に基づく先行音韻･後続音韻に関する 2 者択一の質問である．たとえば「先行音韻が破裂音であるか？」などというような質問があり，すべ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，まず中心音韻が共通の全てのトライフォンを一つの集合とし，これをルートノードとする．次に，分割条件の一つ一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学習される HMM の尤度など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って決定木を生成する．最終的なリーフノードがクラスタになる．

図 1 に，中心音韻/a/に対し，先行音韻あるいは後続音韻の音韻論的類似性に基づいて場合分けを進める手続きを示す．クラスタリング後，1 つのリーフノード（最末端のノード）に含まれるトライフォンは 1 つの HMM の状態を共有する．以上の操作により，出現頻度の低いトライフォンに対しても十分な学習データを確保できる．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代用 HMM を与えることができる．

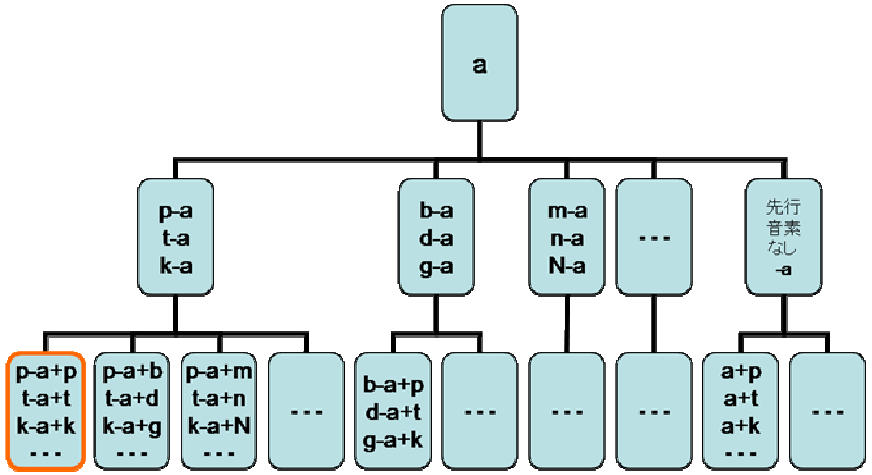

図 1　Tree Based Clustering の概念図

2.3 ボトムアップクラスタリングの手法

ボトムアップ的にクラスタ化を行う手法について説明する．まず学習データに出現する全てのトライフォン間の音響的距離を求める．そして距離の近いものを一つのクラスタとして扱うことで，十分なデータ数を用いて HMM を学習することができる．以下，順に距離の計算方法とクラスタの生成方法について説明する．また，5 状態の音韻 HMM の第 3 状態に注目しクラスタを作る場合と，第 2 状態・第 3 状態・第 4 状態ごとにクラスタを生成する場合と，2 種類の検討を行う．

トライフォン間の音響的距離の計算法について説明する．すべてのトライフォンに対する HMM を混合数 1 で初期学習しておく．出力確率に関するパラメータベクトルは，次節に述べる各特徴量に対する平均と分散である．

中心音韻が同じトライフォンを集め，パラメータベクトルの各要素 i ごとに分散を求める．この値を σ_i^2 とする．あるトライフォンのある状態に対するパラメータベクトルを $\vec{x}$,別のトライフォンの同じ状態に対するパラメータベクトルを $\vec{y}$ とする．両者の距離を，(1)式に基づく分散正規化距離で与える[6].

Vol.2010-SLP-82 No.12
2010/7/24

$$d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{\sigma_i^2}} \qquad (1)$$

なお，n はパラメータベクトルの次元数である．

次に，この距離に基づくボトムアップクラスタの生成手法について述べる．本報告においては各トライフォンに対する音韻 HMM として初期状態と最終状態を含めて 5 状態を設定する．クラスタリングの対象として，

1. 第 3 状態のみ
2. 第 2 状態・第 3 状態・第 4 状態の各々

の 2 種類を検討する．

まず前者についていえば，同じ中心音韻を持つすべてのトライフォン HMM の第 3 状態のパラメータベクトルに注目し，(1)式に基づいてすべてのトライフォン間の距離を計算する．全体の中で最も距離の小さい 2 つのトライフォンを見つけ，一つのクラスタに統合する．この統合操作を指定の回数繰り返す．

後者についてもこれと同じように，第 2・第 3・第 4 状態について，状態ごとに別々にクラスタリングを行う．これによって HMM の状態共有も第 2・第 3・第 4 状態について別々に設定する．

2.4 トップダウンとボトムアップの併用

TBC で分類したクラスタと，距離を計算しボトムアップ的に生成したクラスタを組み合わせて，音響モデルを構築する．基本的な概念を図 2 に示す．図の最上段が TBC のルートノードであり，音韻文脈に基づく木構造に従ってクラスタが形成される．一方，音響的に基づくボトムアップのクラスタを生成し，これに加えていく．図中に丸で囲まれたクラスタがボトムアップ的に生成されたクラスタを表している．このトップダウン及びボトムアップで生成された両方のクラスタを併用する．この結果，一つのトライフォンがトップダウンクラスタとボトムアップクラスタの両方に重複して含まれることもありうる．

設定したクラスタに基づいて音響 HMM の学習を行う．クラスタリングに用いた初期学習 HMM においては混合数 1 としたが，音声認識の精度を考え，音響モデルの学習における混合数は 16 とする．学習に際し，複数のクラスタに重複して含まれるトライフォンをどのように扱うかという点で 2 つの考え方がある．一つ目は基本的に各トライフォンが一つのクラスタのみに含まれるようにするという考え方である．ボトムアップのクラスタ内のトライフォンが TBC で分類されたクラスタ内にも存在すれば，トップダウンのクラスタ内から除外して学習を行う．二つ目は各トライフォンが複数のクラスタに含まれてもよいという考え方である．除外せずに学習を行うことで一つのトライフォンで 2 つの状態を学習することができる．また両者を比べることで，トップダウンのクラスタから除外されたトライフォンの有効性を検討する．後の節で実験的に両者の優劣を比較する．

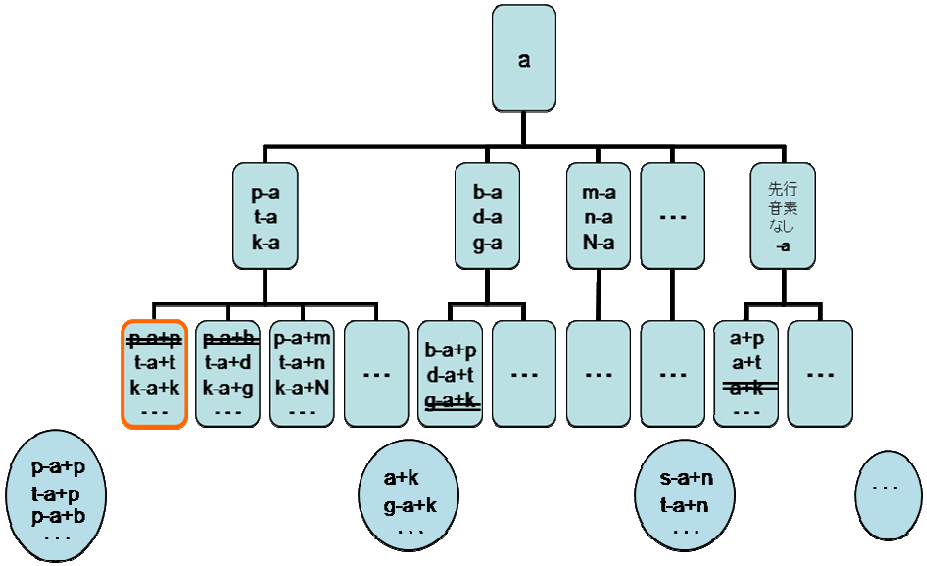

図 2　トップダウンとボトムアップの併用

3. 評価実験

3.1 実験条件

学習データとしては，日本語話し言葉コーパス（CSJ）[7] の中から，学会講演 100 セットを選択し用いる．ただし，話者はすべて女性を選んだ．各講演は 10 分から 20 分程度である．特徴量は 12 次元 MFCC と対数エネルギー，及びそれらのデルタ項，デルタデルタ項の計 39 次元である．トライフォンの混合数は 16 とする．学習ツールとして，HTK[8] を用いる．

評価データには学習データに含まれない学会講演 10 講演分の女性話者データを使用する．音声認識デコーダは Julius[9] を用いる．その他実験条件を表 1 に示す．

Vol.2010-SLP-82 No.12
2010/7/24

表 1　実験条件

基本周波数	16kHz
分析窓	Hamming 窓
分析窓長	25ms
フレーム周期	10ms

認識結果は正解文章に対する単語単位での単語正解精度(Acc)，単語誤り率(WER)について（2）式に定義し，集計した．ここで正解文の単語数を N，認識結果における置換誤り単語数を S，挿入誤り単語数を I，脱落誤り単語数を D とする．

$$ACC = \frac{N-S-D-I}{N} * 100.0 \qquad WER = \frac{S+D+I}{N} * 100.0 \qquad (2)$$

3.2　トップダウン的に生成したクラスタ（TBC）に基づく音響モデルを用いた音声認識実験

まず，トップダウン的手法によって生成したクラスタに基づく音響モデルの性能を把握するために，TBC で様々な数のクラスタを生成し，音響モデルを学習して音声認識の性能を調べた．実験結果を図 3 及び表 2 に示す．

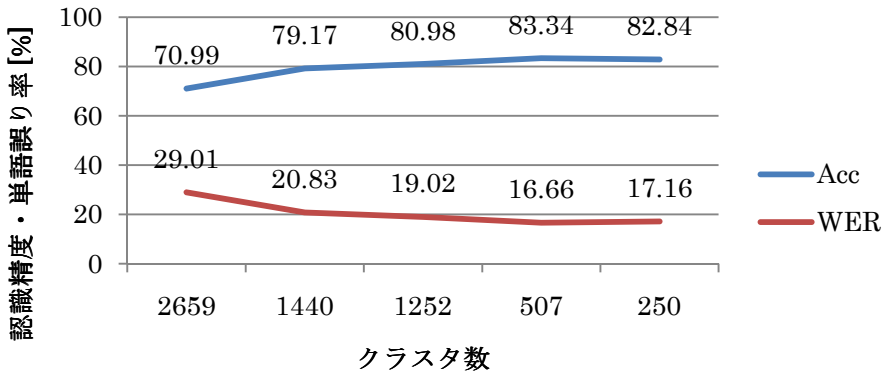

図 3　トップダウン的に生成したクラスタ（TBC）に基づく音響モデルを用いた音声認識実験の結果

表 2　トップダウン的に生成したクラスタ（TBC）に基づく音響モデルを用いた音声認識実験の結果

クラスタ数	2659	1440	1252	507	250
Acc(%)	70.99	79.17	80.98	83.34	82.84
WER(%)	29.01	20.83	19.02	16.66	17.16

実験結果より，Acc・WER 共に状態数が 507 のところで認識精度のピークに達し，それ以降スコアは低下する．これより TBC によって状態数が約 500 までは状態の共有化が上手く行われているが，その後は音響的に類似しない状態まで共有化がされ認識精度に影響されているのではないかと考えられる．そこで次の実験では，過剰に共有化された状態数約 250 の場合に，ボトムアップクラスタリングで作成したクラスタを加えることで過剰な共有化を緩和し，認識精度の向上を目指す．

3.3　TBC に第 3 状態のみボトムアップ的に作成したクラスタを加えた音響モデルを用いた音声認識実験

TBC で作成したクラスタに，音響的に類似するようボトムアップ的に作成したクラスタを加えた音響モデルを作成して実験を行う．前節で述べたように，中心音韻ごとに各トライフォン同士の距離を計算し，距離の最も近い 2 つのトライフォンを見つけリンクを張り，これを指定された回数繰り返すことでクラスタを作成する．ここでは，5 状態の音韻 HMM の第 3 状態のみで類似度を計算し，トライフォンのクラスタリングを行った．距離の計算結果例を表 3 に示す．表 3 は中心音韻 /a/ のトライフォンを集め，その中で距離の近いものを上位いくつか集めたものである．ここで距離の近いトライフォンが，TBC で分類された同じクラスタ内に存在しないことが観察された．逆に，距離の遠いものが TBC の同じクラスタにある例も見受けられた．これより TBC では音響的特徴が類似するものがすべて正確に分類されていないとこが分かり，音韻文脈的に共有化されたものをボトムアップ法によって緩和することでより認識精度が向上する音響モデルが構築できるのではないかと考えられる．なお，ボトムアップのクラスタ内のトライフォンが TBC で分類されたクラスタ内にも存在すれば，トップダウンのクラスタ内から除外して学習を行う．

クラスタを生成するために今回は距離の近いトライフォン上位 200 個・100 個・50 個にリンクを張ることでグループを作った．図 4 は中心音韻 /a/ の場合で，リンク数 50 の場合の生成されたクラスタの図である．実験は TBC で認識結果に低下が見られた状態数 250 の音響モデルに新しくボトムアップ的に作成したクラスタを加えて学習

Vol.2010-SLP-82 No.12
2010/7/24

表 3　距離計算結果例（中心音韻 /a/ の場合）

状態	トライフォン		距離
3	"g-a"	"b-a"	1.678089
3	"d-a+i"	"t-a+i"	1.761916
3	"d-a"	"t-a"	1.838190
3	"s-a+i"	"t-a+i"	1.852989
3	"s-a+n"	"t-a+n"	1.869259
3	"g-a+i"	"k-a+i"	1.875878
3	"g-a"	"k-a"	1.877973
3	"y-a"	"ny-a"	1.946553
3	"d-a+q"	"d-a+t"	2.041517
3	"s-a+N"	"t-a+N"	2.110002

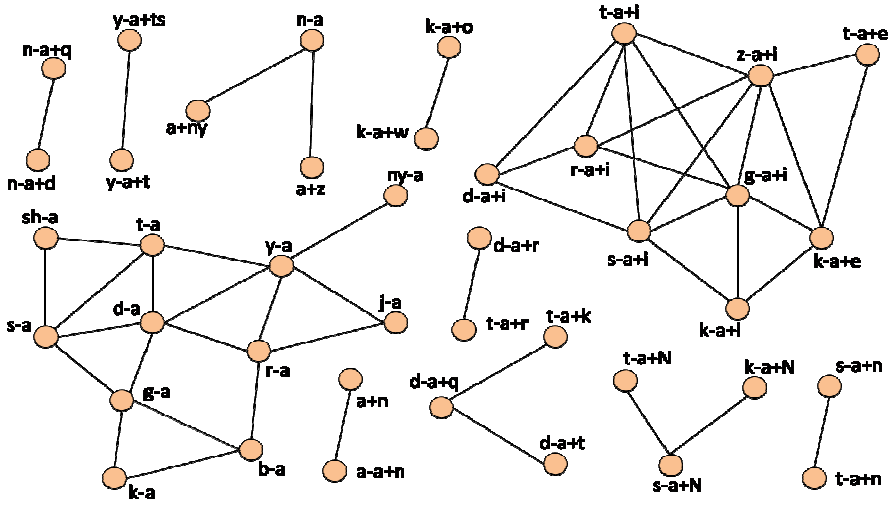

図 4　ボトムアップ的に作成したクラスタ

表 4　TBC に第 3 状態のみボトムアップ的に作成したクラスタを加えた音響モデルを用いた音声認識実験の結果

リンク数	200	100	50
クラスタ数	250+132=382	250+169=419	250+183=433
Acc(%)	83.49	83.9	83.09
WER(%)	16.51	16.1	16.91

表 5　TBC に第 2・第 3・第 4 状態をボトムアップ的に作成したクラスタを加えた音響モデルを用いた音声認識実験の結果

クラスタ数	250+711=961	250+280=530
Acc(%)	83.68	84.07
WER(%)	16.32	15.93

を行い，認識実験を行った．クラスタを追加したことにより状態数も変動する．表 4 に認識結果を示す．

TBC のみでの実験結果ではクラスタ数が 507 の時に認識性能は最高であったが，今回の結果ではその結果と同程度あるいはそれ以上の認識性能が得られた．またクラスタを作る際のリンク数によっても認識結果は変動することから，さらに優れたクラスタ生成方法があることが示唆されている．

3.4　TBC に第 2・第 3・第 4 状態をボトムアップ的に作成したクラスタを加えた音響モデルを用いた音声認識実験

3.3 では類似度を HMM 状態の真ん中第 3 状態のみで計算を行いトライフォンのクラスタリングを行った．次に第 2，第 3，第 4 状態それぞれ別に距離計算を行い，状態ごとにクラスタリングを行う．クラスタ生成として，リンク数 50 としてクラスタ化した．またこの時状態数が 711 個に増えたが，TBC の状態数と比較するため状態数を 507 に近づけるようボトムアップクラスタ内の個数が一定個以上のみを用いることで状態数を 530 個に削減した．認識結果を表 5 に示す．

表 5 をみると，状態数 961 では Acc，WER 共に TBC の状態数 1252 よりも良い結果となっている．また TBC で状態数 507 の時に認識結果がピークであったが，ボトムアップを加え状態数を 530 まで削減した結果をみると Acc，WER 共にスコアは良くなっている．また，前節での HMM 状態の真ん中第 3 状態のみで類似度計算してクラスタ

表 6　一つのトライフォンがトップダウン及びボトムアップの
クラスタに重複して含まれる場合の音声認識実験の結果

クラスタ数	250+711=961	250+280=530
Acc(%)	83.74	84.07
WER(%)	16.26	15.93

リングする方法よりも精度は向上しているため，状態別ボトムアップクラスタリングの有効性が確認できた．

3.5 一つのトライフォンがトップダウン及びボトムアップのクラスタに重複して含まれる場合の検討

ここまでの実験においては, ボトムアップのクラスタ内のトライフォンが TBC で分類されたクラスタ内にも存在すれば，トップダウンのクラスタ内から除外して学習を行ってきた．本稿では，一つのトライフォンがトップダウン及びボトムアップのクラスタに重複して含まれる場合の検討を行う．

3.4 までは各トライフォンが一つのクラスタのみに含まれるようにして学習を行っていた. つまり 1 つのトライフォンで 1 つの状態しか学習を行っていなかった. 次に，各トライフォンが複数のクラスタに含まれても除外せずに複数のクラスタを学習できるようにする．これによって重複するトライフォンがクラスタにどのように影響するか観察する．認識結果を表 6 に示す．リンク数・状態数の条件は 3.4 と同じである．

実験の結果，前節の状態別ボトムクラスタリングを組み合わせた結果と同程度か，それ以上の性能が得られることがわかった．このように，TBC やボトムアップ的にクラスタを生成する場合，そのクラスタの性能を向上させるために役に立つトライフォンがあることが示唆された．

4. おわりに

音声認識の枠組みでは，音韻単位で HMM を学習するのが一般的であり，特に前後の音韻環境を考慮したトライフォンモデルを単位とする手法が主流である．トライフォンには膨大な種類が考えられ，すべてのトライフォンを学習するために必要なデータ数を確保することが困難である．学習データに現れないトライフォンは認識できず，また出現回数の少ないトライフォン HMM の信頼性が低下するという問題が起きる．この問題を解決するために，トップダウン的に音韻文脈に基づいて複数のトライフォンを集めクラスタを作り，互いに代用することで見かけ上の学習データを増やす手法が考えられてきた．しかしこの方法では先行・後続音韻の場合分けでクラスタリングを行うので，音響的に類似したトライフォンが同じクラスタ内にあるとは限らない．そこで本研究では音響的に類似するトライフォンを探し，それらを一つクラスタとするボトムアップ的音韻クラスタリング手法をトップダウン的手法に組み合わせる方式を提案した．

まずトライフォン間の距離を計算したが，距離の近いトライフォンが TBC で分類された同じクラスタ内に存在しない, また距離の遠いものが TBC の同じクラスタにあることが観察された．今回は 5 状態の音韻 HMM の第 3 状態のみ，また第 2・第 3・第 4 各状態で類似度を計算しボトムアップクラスタリングを行った．作成したクラスタを TBC の音響モデルに加え認識実験を行った結果，両者とも TBC のみでの実験結果よりも良い認識結果が得られた．またクラスタを作る際のリンク数によっても認識結果は変動することから，さらに優れたクラスタ生成方法があるのではないかと考えられる．さらに一つのトライフォンがトップダウン及びボトムアップのクラスタに重複して含まれるよう学習を行う場合の検討も行ったが，この場合でも先程の結果と同程度かそれ以上の結果が得られた．

参考文献

1) S.J.Young, J.J.Odell, P.C.Woodland : Tree-Based State Tying for High Accuracy Acoustic Modelling , ARPA Workshop on Human Language Technology, pp.307-312 (1994)
2) 嵯峨山 茂樹 : 音素環境クラスタリングの原理とアルゴリズム, 電子情報通信学会技術報告, SP87-86, pp. 1-8 (1987)
3) Sung-Ill Kim, Tetsuro Kitazoe : Continuous Speech Recognition Using Tree based State Tying , 情処学研報, 98-SLP-20-16 (1998)
4) 鹿野清広, 伊藤克亘, 他, 著: 音声認識システム, オーム社出版
5) 渡部晋治, 佐古淳, 中村篤 : ベイズ的音声認識 VBEC を用いた音響モデル構造の自動決定, 音響学会春季講演論文集, 1-8-6, pp.11-12 (2004)
6) 岡登洋平, 速水悟, 板橋秀一 : クラスタリングによる HMM 間の距離尺度の検討, 信学技報, SP94-16, pp.15-20 (1994)
7) 前川喜久雄 :「日本語話し言葉コーパス」付属ドキュメント 「日本語話し言葉コーパス」の概観(2004)
8) S. Young, et. al. : The HTKBook , Entropic Cam- bridge Research Laboratory
9) 河原達也, 李晃伸 : 連続音声認識ソフトウェア Julius, 人工知能学会誌, Vol.20, No.1, pp41-49(2005)