TR-I-0203

話者重畳型HMMによる文節認識

Speech Recognition Using Supplemented HMM

服部 浩明　中村 哲†

Hiroaki HATTORI　Satoshi NAKAMURA

概要

話者適応を行う際に問題となる話者間の調音様式の違いに対処するための手法として、複数話者HMM学習をもちいる方法(話者重畳型HMM)を提案する。話者重畳型HMMは複数の学習用話者の音声を静的なスペクトルの写像により一名の基準話者の空間へ写像し、それらの写像された音声を用いてHMM学習を行うものである。これにより、静的には一人の話者でありながら、動的には複数の話者の調音様式を含むHMMが得られる。男女各一名を未知話者とし、男性6名を標準話者として評価実験を行い本手法の有効性の検討を行った。日本語23音韻を用いた実験では第一候補で71.3%、第三候補までの累積認識率で93.2%と、従来の静的なスペクトルマッピングのみの場合に比べ、それぞれ0.7%と1.5%の改善が得られた。また、日本語279文節を用いた評価実験では、従来法と組み合せることにより第一候補で74.9%、第五候補までで96.2%の文節認識率を得た。とくに、累積文節認識率では特定話者の98.9%に近い認識率が得られ、本方式の有効性が確認された。

ATR自動翻訳電話研究所
ATR Interpreting Telephony Research Laboratories

†Sharp 中央研究所
Sharp Central Research Laboratory

目次

1 はじめに

近年HMMを用いた音声認識手法の研究が盛んである。HMMは音声特徴量の変動を確率的統計的に記述できる性質を持っており、大量の学習用データが用意できる場合にはモデルを精度よく推定することができる。しかしながら、すべての話者について大量の学習用データを用意するのは実際的ではなく、不特定話者認識や話者適応化の技術が必要とされる。何ら前処理を行うことなく誰の声でも認識する不特定話者音声認識は、マンーマシンインターフェイスとしての音声認識を考えた場合、最も望ましい形態である。しかし、大規模なタスクの不特定話者の連続音声を十分な精度で認識することは非常に難しく、不特定話者の音声認識は1000語程度のタスクに対して実験システムが研究されている段階である[1]。一方、話者適応化を用いる音声認識は少数の語彙により適応化できるならば、十分に実際的である。また、話者適応は、認識環境の変化をも含んだ適応化として捉えることができ、伝達経路の周波数特性や環境雑音の変化、あるいは、音声の経時変化やlombard効果等の発声変形にも対処できる。本稿では、HMMを用いた音声認識における話者適応方式について述べる。

異なる話者間での音声の違いは静的なスペクトル形状の違いと、動的な時間構造の違いに大別される。前者は発声器官の構造の違いに起因するものであり、後者は発声様式の違いによる。前者の静的な違いに関しては種々の検討がなされている[2~13]。筆者らは話者間の静的なスペクトル形状の適応化方法として、ファジイベクトル量子化を用いた教師あり学習による適応化方法を既に提案した[4~7]。この方法は未知話者と標準話者の音声をDTWを用いて対応づけを行い、コードブック間の対応付けヒストグラムを求めることにより、未知話者から標準話者への写像関数を求める方法である。この方法により、男女間のようにスペクトル形状の著しく異なる話者間においても良好な結果を得ている。しかしながら、特定話者の認識率とは依然、差が存在する。また、未知話者の音声スペクトルの適応化に加えて、HMMを標準話者空間に写像された未知話者のデータを用いて学習した場合、つまり、未知話者の調音様式を取り込んだ場合には、さらに認識率が向上する[7]。このことは、静的なスペクトルの対応付けではまだ不十分であり、後者の話者間の動的な違いに対処する必要があることを意味する。

いままでに、動的な違いを考慮した話者適応方式として、DTWを用いた認識方式において複数の時間構造をマルチテンプレートとしてもつ方式が提案されている[2,15,20]。ここでは、HMMの音声のゆらぎを確率的に表現できる特性に着目し、標準話者のHMMに複数の話者の時間構造をもたせるための学習方法を提案する。これにより、テンプレート数を増やすことなく話者間の調音様式の違いに対処できる可能性がある。

本稿では、はじめに静的なスペクトルの適応方法およびHMMにおける話者適応化について述べる。次に、複数話者学習を用いたHMM学習について述べると

ともに、日本語全音韻および文節発声された音声を用いた評価実験の結果について述べる。

2　ベクトル量子化を用いた話者適応化

話者適応化方式は学習音声の発声内容が既知である教師あり適応化方式と、発声内容が未知である教師無し適応方法がある。どちらの方法にも一長一短があるが、本稿では教師あり適応化法を用いる[4~7]。図1に話者適応化法のブロック図を示す。

2.1　対応づけヒストグラムを用いたスペクトル適応化

未知話者の発声した学習用音声は、標準話者の発声した同じ発声内容の音声と対応付けを行い、標準話者への写像関数を求める。以下に静的なスペクトル適応化の手続きを示す。

1)未知話者の学習用音声を用いてベクトル量子化コードブックを生成する。

2)未知話者のコードブックを用いて未知話者の学習用音声のベクトル量子化を行う。

3)あらかじめ準備した標準話者の学習単語の標準パターンとDTWを行い、同一単語間における最適パスを求める。

4)DTWの最適パスにしたがって未知話者のコードベクトル$V^{(A)}{}_i$と標準話者のコードベクトル$V^{(B)}{}_j$の対応回数を求め、対応付けヒストグラムh_{ij}を求める。

5)対応付けヒストグラムの値を重みとして、新しく標準話者の空間に写像されたコードベクトルを求める。

$$V_i^{(A)} \rightarrow V_i^{(A\rightarrow B)} = \sum_{j=1}^{c} h_{ij} \cdot V_j^{(B)} \;/\; \sum_{j=1}^{c} h_{ij} \qquad (1)$$

ここで、

$V^{(A)}i$: 未知話者Aのコードベクトル

$V^{(B)}j$: 標準話者Bのコードベクトル

$V^{(A\rightarrow B)}i$: 写像後のコードベクトル

h_{ij}: 対応づけヒストグラム

c: コードブックのサイズ

6) 5)のコードベクトルを未知話者のコードベクトルとして、ステップ1)-5)を繰り返し、スペクトル歪が収束するまで繰り返す。最終的に得られた5)におけるコードベクトルよりなるコードブックを変換コードブックと呼ぶ。

話者適応化は、未知話者の音声Xをベクトル量子化し、コードベクトル$V^{(A)}{}_i$とした後、対応する変換コードブックのコードベクトル$V^{(A\rightarrow B)}{}_i$を出力することにより行なう。ファジィベクトル量子化を用いる場合には、未知話者と標準話者の特徴空間において、1対1にコードベクトルが対応し、ファジィ級関数が保存されることを仮定することにより同様に行える。このような写像をファジィマッピ

ングと呼んでいる[6]。(2)式に未知話者の音声Xを標準話者の空間に写像する場合を示す。

$$X \rightarrow X' = \frac{\sum_{i=1}^{k} \left(u_i^{(A)}\right)^m \cdot V_i^{(A)}}{\sum_{i=1}^{k} \left(u_i^{(A)}\right)^m} \rightarrow X'' = \frac{\sum_{i=1}^{k} \left(u_i^{(A)}\right)^m \cdot V_i^{(A \rightarrow B)}}{\sum_{i=1}^{k} \left(u_i^{(A)}\right)^m} \quad (2)$$

ここで、
$u_i^{(A)}$:ファジイ級関数
m:ファジネス
k:近傍数
である。

2.2 HMMにおける話者適応

前節でのべた適応法により、未知話者の音声を標準話者の空間へ写像することができる。この写像された音声を、標準話者のコードブックで再度量子化することで、標準話者の音声により訓練された離散型HMMを用いることができる。しかし、この方法は効率が悪いだけでなく、量子化歪をも受けてしまい、好ましくない。そこで、ここでは対応づけヒストグラムh_{ij}をファジイVQを用いたHMMにおけるファジイ級関数と見なすことにより、標準話者のHMMの適応化を行う[6]。ファジィVQを用いたHMMは、離散モデルのHMMにおける学習データの不足を補うための方法として提案されている[14]。以下に出力確率を適応化するためのアルゴリズムを示す。

1)ベクトル量子化による話者適応法を用い、ヒストグラムh_{ij}を用いて未知話者の入力コードベクトル$V^{(A)}{}_i$を標準話者の空間点$V^{(A \rightarrow B)}{}_i$に写像したと仮定する。
2)1)で得られたヒストグラムh_{ij}を写像された点$V^{(A \rightarrow B)}{}_i$から標準話者のコードベクトル$V^{(B)}{}_j$のファジィ級関数とみなす。
3)コードベクトル$V^{(A)}{}_i$の出現確率ω_iを(3)式を用いて計算し、認識を行う。

$$\omega_i = \sum_{j=1}^{c} h_{ij} \cdot O_j \quad (3)$$

ここで、
Oj;標準話者のコードベクトル$V^{(B)}{}_j$の出現確率。
である。入力ベクトルをファジイVQした場合には(3)式は次式のように書ける。

$$\omega = \sum_{j=1}^{c} \left(\sum_{i=1}^{k} u_i \cdot h_{ij} \right) \cdot O_j \quad (4)$$

この方法によりベクトル量子化を重複して行う事なく効率的に話者適応化が実現できる。

3 複数話者音声によるHMM学習[7]

上で述べた方法により、未知話者の音声を標準話者空間へ写像し、標準話者の

HMMを用いて認識することができる。ここではさらに、話者間の調音様式の違いに対処するための複数話者音声を用いたHMM学習について述べる。

2.1節で述べたフレーム毎のコードブックマッピングによる写像は静的なものであり、写像された音声は、動的には影響を受けず、写像前の話者の特徴を保っている。そこで、複数の標準話者の音声を一人の基準となる標準話者(以下、このように静的な特徴の基準となる話者を基準標準話者と呼ぶ。)にそれぞれ写像した音声の集まりを考えると、このような集まりは静的には基準標準話者であるが、動的には複数の標準話者の特徴を含んでいるものと思われる。したがって、この音声集合を用いて学習を行うことにより、複数の話者の動的特徴を含むHMMを得ることができる。また、見かけ上基準標準話者の学習用データが増えるので、学習用データの少ないモデルの推定精度の向上も期待される。このようにして学習されたHMMを話者重畳型HMMと呼ぶことにする。図2にこの様子を示す。以下に話者重畳型HMMを推定するアルゴリズムを示す。

1)複数の標準話者の話者適応用学習データを用意し、話者ごとに基準標準話者への対応づけヒストグラムを求める。

2)1)で求めた対応づけヒストグラムを用いて複数の標準話者のHMM学習用データを基準標準話者の空間へ写像する。

3)写像された全学習用データを用いてHMMの学習を行う。

認識時には2.2節で述べた認識方式をそのまま用いることができる。

4 評価実験

話者重畳型HMMの有効性を確認するために、日本語全音韻および日本語文節を用いて評価を行った。男女各一名を未知話者、それと異なる男性話者2名を基準標準話者とし、4通りの組合せについて話者適応実験を行った。話者重畳型HMMを求める際の標準話者としては基準標準話者を含む男性6名を用いた。

音響分析は、12kHzサンプリング、分析窓長21.3ms、LPC14次で行った。分析フレーム間隔は日本語全音韻では3ms、文節認識では9msで行った。コードブック作成及び話者適応ヒストグラム作成にはには音韻バランス216単語の前半100単語を用いた。コードブックはWLR、パワー、差分ケプストラムの三種を用い、コードブックサイズはそれぞれ256、64、256とした。ファジィVQを行う際の近傍数は6、ファジィネスは1.6とした。

4.1 日本語音韻認識実験

日本語18子音および日本語5母音を対象として認識実験を行った。HMMは撥音以外の子音は図3.aの4状態3ループモデル、母音および撥音は図3.bの2状態1ループモデルを用いた。また、/ch,ts,p,t,k,b,d,g/については語頭と語中で別のモデルとした。モデルの学習には重要語5240単語中の偶数番目の単語に含まれる音韻を用い、評価には重要語の奇数番目の単語に含まれる音韻を用いた。学習、評価と

もに音韻の切り出しは、視察により付与されたラベルを用いた[17]。

表1に第三位までの累積認識率を示す。平均では話者重畳型HMMを用いることで第一位で71.3%、第三位までで93.2%の音韻認識率が得られ、従来の静的な話者適応方式に比べそれぞれ0.7%、1.5%の改善が得られた。

個々の話者の組合せについて見ると、未知話者がmau、基準標準話者がmnmの場合に効果が大きく、第一位で3.0%、第三位までで2.2%の改善が得られた。一方、未知話者がfsu、基準標準話者がmhtの場合には、かえって第一位での認識率は低下している。これは、未知話者と基準標準話者の調音様式が異なる場合には、話者重畳型HMMによりその違いがうまく吸収されるが、逆に調音様式が類似している場合には、複数話者の調音様式を取り込むことで、特徴がぼやけてしまい、認識誤りが増えたものと考えられる。

音韻ごとの第一位での認識率および第三位までの累積認識率をそれぞれ図4、図5に示す。音韻別に見ると、有声、無声破裂音の認識率が向上している。これらの音韻では破裂から母音へのわたり区間での音響的特徴の時間変化パターンが識別のための重要な意味をもち、話者による調音様式の違いが現れやすい。そのため、このような音韻の場合には話者重畳型HMMの複数話者の調音様式を持つ効果が大きい。また、/p/の認識率の向上が大きいのは、/p/の学習サンプルが少なく、話者重畳型HMMを用いることにより、調音様式の差が吸収されただけでなく、見かけ上のサンプル数の増加によるモデルの推定精度の向上も寄与していると考えられる。一方、母音や摩擦音等のスペクトル形状そのものが識別にとって重要である音韻や、/m,n/、/k,h/のようにスペクトルの微妙な差異をもちいて識別を行わなければならない音韻に関しては、話者重畳型HMMを用いることにより、その特徴がぼやけてしまい、認識率が低下している。しかし、これらの音韻での誤りは致命的なものではなく、第三位までの累積認識率で見た場合には、あまり低下していない。

4.2 日本語文節認識実験

HMM-LR文節認識システム[18、19]を用いて話者重畳型HMMの性能評価を行った。評価に用いたデータは国際会議の問合せを想定した25会話文中の279文節である。文法の複雑性を表2に示す。HMMは母音、長母音、撥音は2状態1ループモデル、それ以外の音韻は4状態3ループモデルを用いた。HMMの学習には日本語5240単語中に含まれる音韻を用いた。ただし、/py/については重要語5240単語中に1サンプルしか存在しないため、音韻バランス216単語中の7サンプルも併せて用いた。継続時間長制御はモデルの状態毎に継続時間長分布を求め、認識時に各状態に留まった時間長に応じたペナルティを与えることで行った。

表3に文節認識率を示す。平均では話者重畳型HMMを用いることにより第一位での文節認識率で75.2%と、従来の静的な話者適応方式に比べ1.0%の改善が得られた。

話者の組合せについて見ると、基準標準話者がmnmの場合には、未知話者が

fsuの場合には10.5%、mauの場合には2.5%認識率が改善しているが、基準標準話者がmhtの場合には逆に、未知話者がfsuの場合には2.2%、mauの場合には6.8%低下している。これは、先の音韻認識実験と同様に、調音様式が異なる場合には、話者重畳型HMMによりその違いがうまく吸収されたが、調音様式が類似している場合には、複数話者の調音様式を取り込むことで、特徴がぼやけてしまい、認識誤りが増えたものと考えられる。

個々の音韻について調べた結果、基準標準話者により異なる傾向が見られた。基準標準話者がmhtの場合には、話者重畳型HMMを用いることで、どちらの未知話者においても/n/を/m/へ誤る場合が増え、認識順位が低下したものが多かった。未知話者がfsuの場合の第一位の認識率の低下の殆どは、この誤りが原因である。未知話者がmauの場合には、それに加えて、/k, r/等の音韻での尤度が低下したために認識順位が低下した場合が見られた。基準標準話者がmnmの場合には、話者重畳型HMMにより/k,r,b,g/等の音韻での誤りが改善された。特に、/k/での誤りの改善が著しく、mnmと未知話者では/k/の調音様式が大きく異なるものと思われる。

以上の様に話者重畳型HMMは音韻や話者の組合せにより効果が異なる。話者重畳型HMMは複数話者の音声を用いて学習することで、調音様式を取り込むだけでなく、静的な特徴をも取り込んでしまう。これはスペクトル写像に誤差が存在するかぎり、ある程度は避けられない。そのために、ある未知話者と基準標準話者の組合せにおいて、音韻の静的な特徴の違いが大きいか調音様式の違いが大きいかによって、話者重畳型HMMの効果は異なってくる。

これに対処する一つの簡単な方法として、動的な変化により特徴づけられる音韻に対してのみ話者重畳型HMMを用いる方法が考えられる。このような音韻では、静的なスペクトルにより特徴づけられる音韻に比べ、話者による調音様式の違いが現れやすいと考えられる。音韻表4に示すのは、破裂音、半母音、拗音についてのみ話者重畳型HMMを用いた場合の累積文節認識率である。話者の組合せによる差は、表3の場合に比較して小さくなっており、平均で第一位での認識率は74.9%、第5位までででは96.2%の累積認識率を得、従来法のみを用いた場合に比べ、それぞれ0.7%と1.0%の改善となっている。

5 おわりに

話者間の調音様式の違いに対処するための方法として、複数話者の音声を用いて学習を行う話者重畳型HMMを提案した。日本語23音韻を用いた実験では第一位で70.6%、第三位までの累積認識率で93.2%が得られた。また、日本語279文節を用いた評価実験では、従来法と組み合せることにより第一位で74.9%、第五位までで96.2%の文節認識率が得られ、本方式の有効性が確認された。

本報告では、従来の静的なスペクトル写像方法と話者重畳型HMMを組み合せて用いた。しかし、ある音韻に対してどちらの方法を用いるべきかは、一般的

に決められるものではなく、未知話者と基準標準話者の組合せによって決定しなければならない。また、話者適応を行うという意味においては、必ずしも全ての標準話者を用いる必要はなく、未知話者の調音様式を吸収し得るに十分なだけの話者を標準話者として用いればよい。したがって、話者適応を行う際に、音韻ごとにどの標準話者の調音様式を取り込むかを決定することができれば、さらに高精度の話者適応を実現できる可能性がある。

今後は、より多くの話者を用いて評価を行うとともに、未知話者の調音様式を積極的に用いる方法を検討する予定である。

謝辞

研究の機会をあたえてくださった樽松社長に深謝いたします。また、プログラムを提供して頂いた花沢氏、御討論頂いた音声情報処理研究室の皆様に感謝致します。

参考文献

(1) K-F.Lee, H.W.Hon, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM," Proc. ICASSP 88 S3.7 (1988)

(2) K.Shikano,K-F.Lee,R.Reddy,"Speaker Adaptation through Vector Quantization," Proc. ICASSP 86, 49.5 (1986)

(3) 中島邦夫、高橋真哉,"大語彙音声認識における話者適応化法," 音講論集1-1-6(1983-10)

(4) 中村哲、鹿野清宏,"ベクトル量子化を用いたスペクトログラムの正規化," 音響学会誌 44, 595-602 (1988)

(5) 中村哲、鹿野清宏,"ファジィベクトル量子化を用いたスペクトログラム正規化," 音響学会誌 45, 107-114 (1989)

(6)中村哲、花沢利行、鹿野清宏、"ベクトル量子化話者適応アルゴリズムのHMM音韻認識による評価",信学技報SP88-106(1988,12)

(7)中村哲、服部浩明、鹿野清宏、"複数話者HMM学習を用いた話者適応化の音韻認識による評価"、音講論集2-p-15,(1989,3)

(8) 新美康永、小林豊,"ベクトル量子化のコードブックの話者適応化,"音講論集2-5-13(1987-10)

(9) K.Choukri,G.Chollet,Y.Grenier,"Spectral Transformations through Canonical Correlation Analysis for Speaker Adaptation in ASR," Proc. ICASSP 86, 49.9 (1986)

(10)山下泰樹、松本弘,"単語音声認識におけるベクトル量子化誤差を利用した話者適応," 信学技報SP87-118 (1988.01)

(11)古井貞熙,"スペクトル空間のクラスタ化に基づく教師なし話者適応化法," 音講論集2-2-16(1988-3)

(12)R.Shwartz,Y.Chow,F.Kubala,"Rapid Speaker Adaptation Using a Probalistic Spectral Mapping," Proc. ICASSP 87 15.3 (1987)

(13)M.Nishimura,K.Sugawara,"Speaker Adaptation Method for HMM-Based Speech Recognition," Proc. ICASSP 88 S5.7 (1988)

(14)M.Feng,F.Kubala,R.Schwartz,J.Makhoul, "Improved Speaker Adaptation Using Text Dependent Spectral Mappings," Proc. ICASSP 88 S3.9 (1988)

(15)古井貞熙,"音声スペクトルの動的特徴を用いた単語音声認識"、信学技報S84-65(1984)

(16) H.P.Tseng, M.J.Sabin, E.A.Lee ,"Fuzzy Vector Quantization Applied to Hidden Markov Modeling," Proc. ICASSP 87 15.5 (1987)

(17)武田一哉、匂坂芳典、片桐滋、"音声データベース構築のための音韻ラベリング"、音講論集2-p-20(1988,10)

(18)北研二、川端豪、斎藤博昭、"HMM音韻認識と予測LRパーザを用いた文節認識"、信学技報SP88-88(1988,10)

(19)花沢利行、中村哲、川端豪、鹿野清宏、“ベクトル量子化話者適応アルゴリズムのHMM文節認識による評価”、音講論集2-p-18(1989,3)
(20)古井貞熙、”マルチテンプレートと教師なし話者適応化による音声認識“、電子情報通信学会論文誌A Vol.J72-A、No.10、pp1476-1483(1989-10)

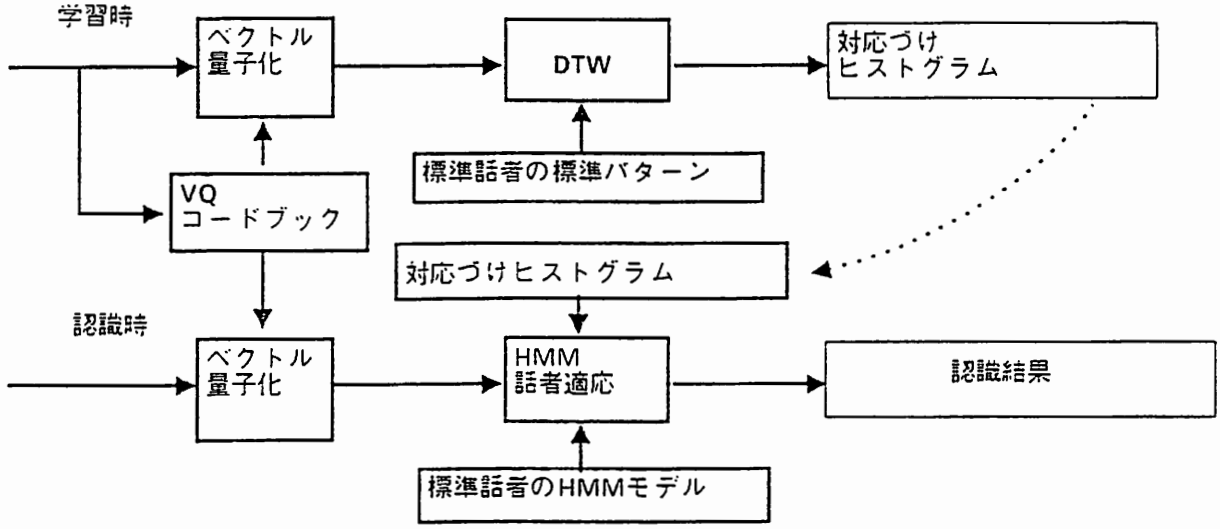


図1　HMM話者適応化のブロック図

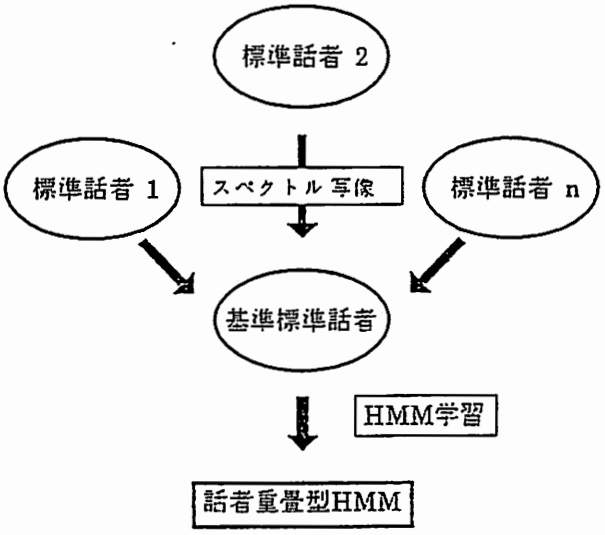
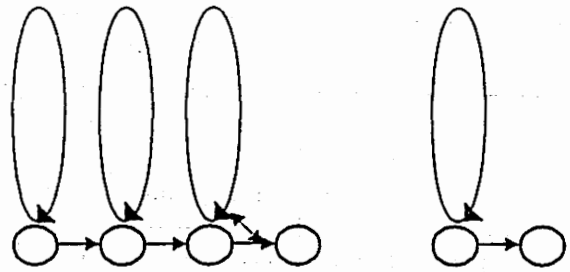


図2　話者重畳型HMMの学習



a)4状態3ループモデル　　b)2状態1ループモデル

図3　モデルの構造

表1　音韻累積認識率　(%)

	スペクトル適応化					話者重畳型HMM					特定話者		
	fsu ↓ mht	mau ↓ mht	fsu ↓ mnm	mau ↓ mnm	平均	fsu ↓ mht	mau ↓ mht	fsu ↓ mnm	mau ↓ mnm	平均	fsu	mau	平均
1	70.0	70.8	70.7	70.9	70.6	67.9	72.4	71.0	73.9	71.3	90.1	90.3	90.2
~2	85.2	86.8	85.6	86.1	85.9	86.7	87.9	87.0	88.4	87.5	96.1	96.6	96.3
~3	91.9	92.4	90.9	91.8	91.7	92.8	93.1	92.8	94.0	93.2	98.3	97.9	98.1

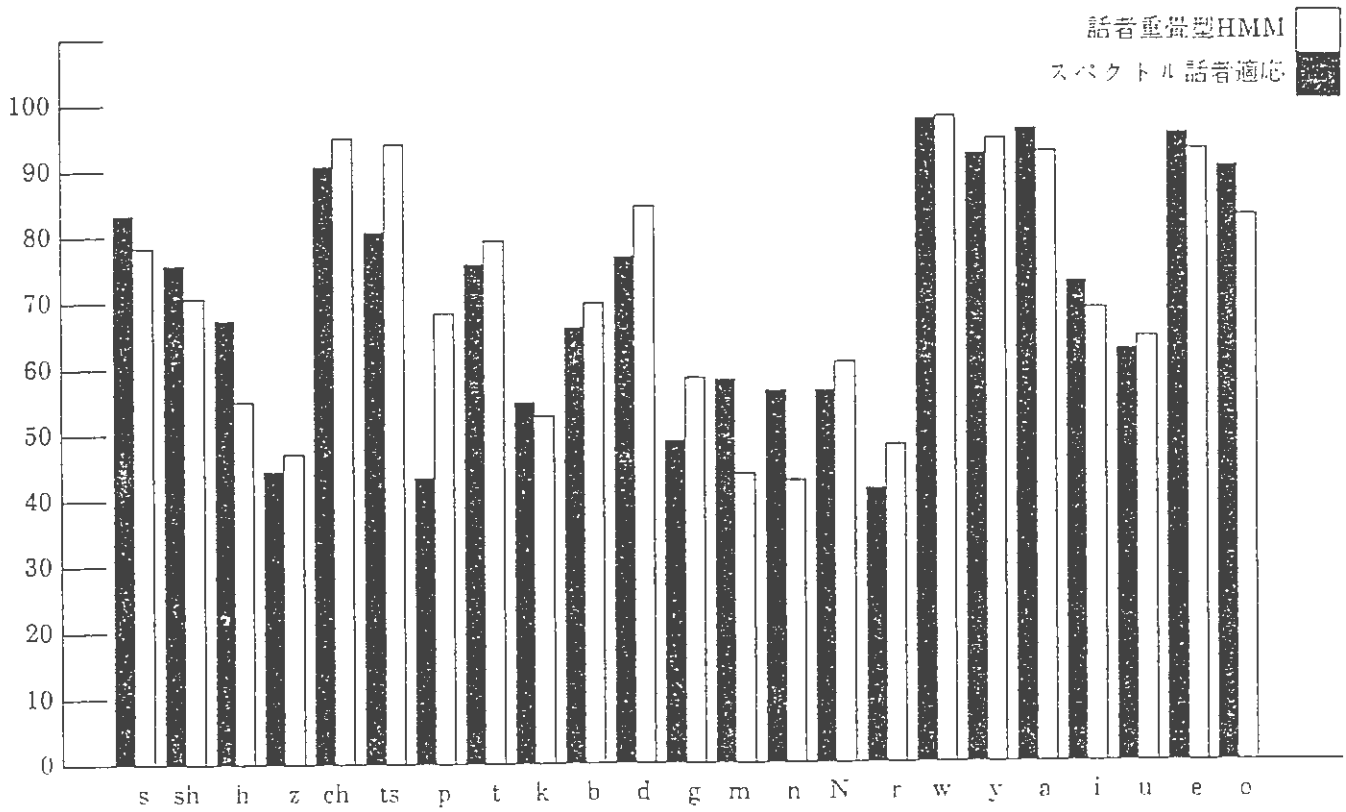

話者重畳型HMM
スペクトル話者適応
100
90
80
70
60
50
40
30
20
10
0
s sh h z ch ts p t k b d g m n N r w y a i u e o

図4　音韻認識率(%)

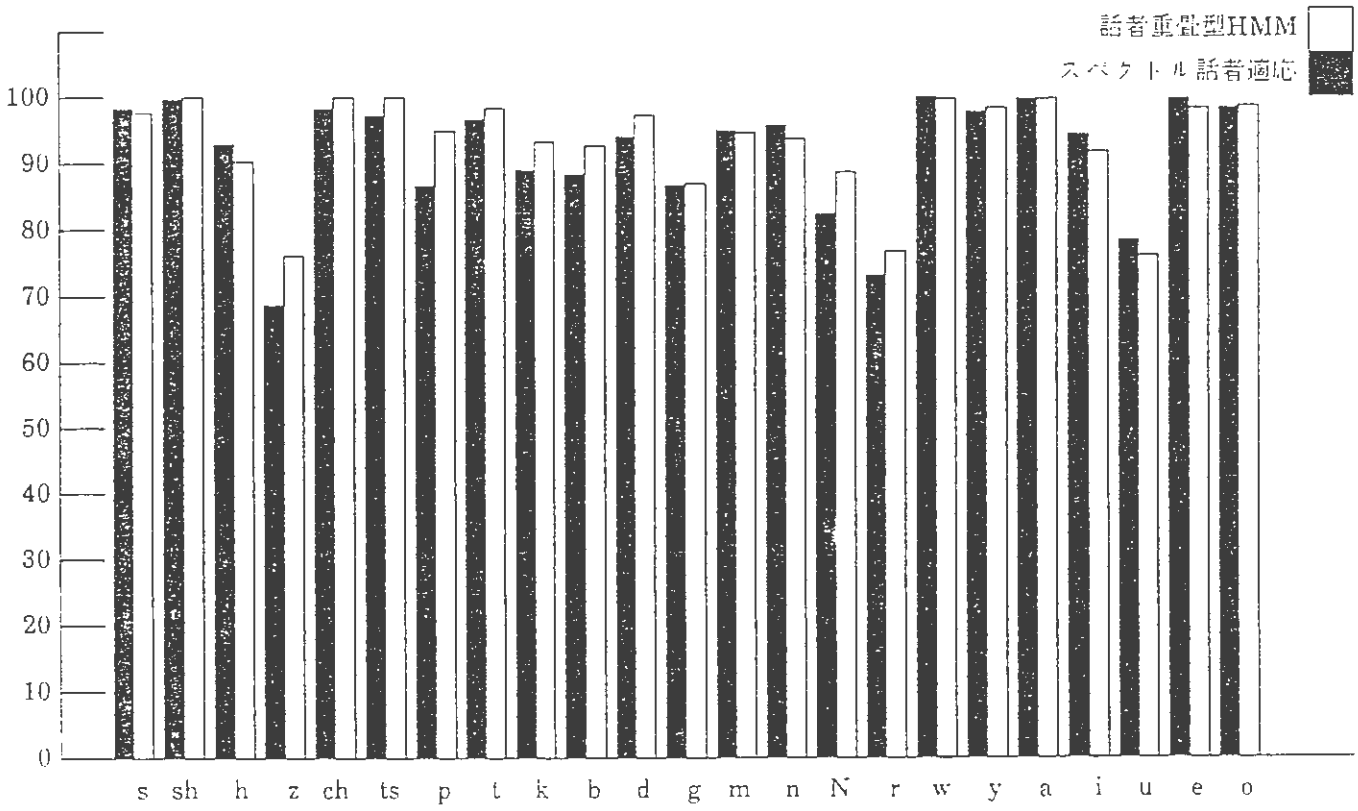


図5　第三候補までの累積音韻認識率(%)

表2　文法の複雑性

異なり語彙数	1035
タスクエントロピー	17.0
音韻パープレキシティ	5.9

表3　文節累積認識率　(%)

	スペクトル適応化					話者重畳型HMM					特定話者		
	fsu→mht	mau→mht	fsu→mnm	mau→mnm	平均	fsu→mht	mau→mht	fsu→mnm	mau→mnm	平均	fsu	mau	平均
1	83.1	75.3	73.7	64.5	74.2	80.9	68.5	84.2	67.0	75.2	89.2	81.7	85.5
~2	92.1	87.1	87.4	82.1	87.2	91.4	85.7	89.9	84.9	88.0	96.8	93.9	95.4
~5	97.1	95.3	96.8	91.4	95.2	97.8	93.5	96.8	92.8	95.2	98.9	98.9	98.9

表4　文節累積認識率　(%)

	話者重畳型HMM＋スペクトル適応化				
	fsu ↓ mht	mau ↓ mht	fsu ↓ mnm	mau ↓ mnm	平均
1	83.5	74.2	76.3	65.6	74.9
~2	90.6	87.8	88.5	84.2	87.8
~5	98.6	95.0	96.4	94.6	96.2