Internal Use Only（非公開）

ＴＲ－ＳＬＴ－００３０

調音運動を考慮した音声認識法

A speech recognition method considering articulatory movements

飯塚　陽介　　マルコフ　コンスタンティン
Yosuke Iiduka　　Konstantin Markov
中村　哲
Satoshi Nakamura

2002 年 12 月 27 日

現在、音声認識の技術は理想的な環境、発話に対して実用に近い状態になっているが実環境や自然な発話に対してはまだ改良の余地がある。その問題点の一つとして調音結合について十分な対処ができていないという点がある。音声認識で行われている連続的な音声波形から離散的な音素系列へのマッピングは音声生成過程の逆過程の一種であると考えられるが、従来の音声認識では音声生成のメカニズムを取り入れていない。調音結合などの問題を解決してより高い認識率を得るために音声生成のメカニズムを音声認識に取り入れる必要があると考えられる。

本研究ではBayesianNetworkとHMMを組み合わせたHybrid HMM/BNモデルに基づいて実測した調音パラメータを用い、音声と調音位置の依存関係を表現するモデルを作成した。音素認識実験を行ったところ、音声データのみを用いた従来のHMMよりも、HMM/BNモデルの方が得られた認識率は、約2%高くなっていた。この結果から、音声認識に音声生成過程を取り入れることが有効であることを明らかにした。

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目２番地２ TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai “Keihanna Science City” 619-0288,Japan
Telephone:+81-774-95-1301
Fax　　:+81-774-95-1308

©2002（株）国際電気通信基礎技術研究所
©2002 Advanced Telecommunication Research Institute International

目 次

1 はじめに

従来の音声認識で行われている連続的な音声波形から離散的な音素系列へのマッピングは音声生成過程の逆過程の一種であると考えられるが、実際は音声生成のメカニズムを取り入れている研究はまだ少ない。音声認識における問題の一つとして、調音結合について十分な対処ができていないという点がある。調音結合は調音器官の運動学的特性に起因した現象であるので、調音結合による影響を解決し、音声認識率を向上させるためには音声生成のメカニズムを音声認識に取り入れる必要があると考えられる。本研究では音声生成のメカニズムを考慮した音声認識法を提案する。ここでは実際に観測した調音運動データに基づいて音声生成メカニズムを認識に取り込むことによりその有効性について考察する。

1.1 本研究の背景

音声認識の技術は、HMM を用いることで理想的な環境では実用に近い状態になっているが、実環境ではまだ改良が必要である。問題点の一つとして会話音声や連続音声の場合、発話過程において、隣接する音素が影響しあって変化する調音結合などの影響により従来法では音響モデルの標本の分布が重なり合うオーバーラップが生じてしまい、認識率に影響が現れる。音素環境などの要因によって音素の特徴が変化するのは、音声の生成が顎、唇、舌などから構成される力学系の運動特性に拘束されるため、隣接した音素の特徴が調音器官の運動パターン上で相互に重なり合い、時間的に広がりを持った変動として現れるためである。したがって、音声情報処理における調音結合の問題をより本質的に解決するには、調音運動を用いて音声の情報表現をおこなうことが必要であると考えられる。人間の音声生成メカニズムを考慮した研究として、Gao らはカルマンフィルタや多層パーセプトロンを用いた調音結合モデルを構築し、認識に応用した [7]。Hodgen らは音声と調音位置とのマッピングを行う MO-MALCOM を構築した。仮想調音空間において調音運動の連続性を導入することによりスペクトル上の不連続性を補整している [8]。Lee らはカルマンフィルタを用いて調音器官の相互関係を考慮して調音運動の推定を行っている [9]。

1.2 本研究の目的

本研究では音声生成過程のうち音声波形の一つ前のステップである調音運動について着目した。発話される音声の特徴は調音位置によって大きく左右されると考えられるので、音声と調音位運動の情報には依存関係があると考えられる。また、調音運動は実際に観測することが出来るので、本研究では実際に測定された調音運動を音声認識に取り入れることを提案する。その手法として、認識時に常に調音運動を観測することは容易ではないので観測されている調音運動と音声との関連について分析し、従来法では音声波形から音素系列へ直接マッピングしていたものを、調音運動を音声と音素の間の一つの拘束条件として考えてモデルに取り入れることを考える。その手法として BayesianNetwork と HMM を組み合わせた Hybrid HMM/BN モデルを用いる。実際に収録した調音パラメータを取り入れて、音声と調音位置の依存関係を表現したモデルを作成して認識実験を行い、音声認識

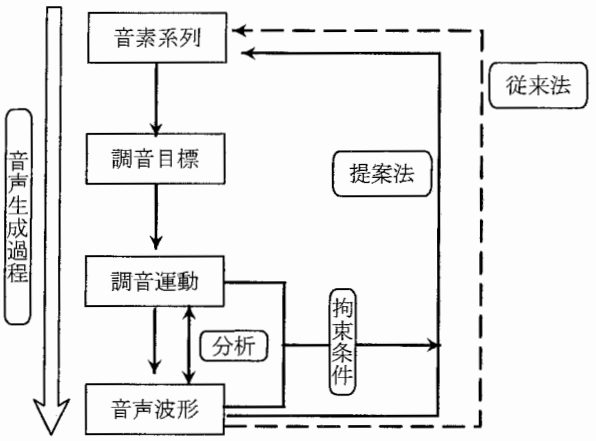


Fig. 1: 音声生成過程と認識手法

に音声生成のメカニズムを取り入れることの有効性について検討する。

2 音声生成過程と調音運動

2.1 発声器官の構造

人間が音声を生成するプロセスの第一段階は、相手に伝えたい内容を表現する言語を選択し、それを文法に合う言語的形式に変えることである。次に、これに従って脳から発声器官に運動神経指令が出され、発声器官の種々の筋肉が動いて、空気振動としての音声が作られる。この音声波は空気中や伝達系を伝わって相手の耳に届き、その聞き手の聴音器官を働かせて、神経パルスとして聴覚神経系を経て脳に伝えられる。こうして話し手が伝えようとした言語情報は、相手に理解されることとなる。生成された音声は話して自身の耳にも伝わり、話し手はこのフィードバックを聞きながら、絶えず発声器官の制御を行っている。この音声の生成と受容の密接な結びつきは、言葉の鎖 (speech chain) と呼ばれている。

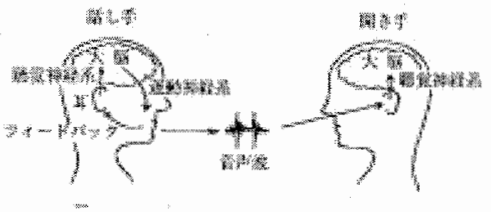


Fig. 2: 音声の生成と受容の密接な結びつき 言葉の鎖 (speech chain)

人間の発声器官の構造は図 3[2] に示す通りで、全体として一つの連続した管を成している。腹筋が横隔膜を押し上げることによって、肺から押し出された空気は、気管を通った後、喉頭の声門 (glottis) すなわち左右の声帯 (vocal cords) の間を通る。通常の呼吸の時は声門は大きく開いているが、声を出そうとすると声帯が接近する。この間を肺からの空気が通り抜けようとするために、空気流と声帯との相互作用により、声門が周期的に開閉し、ほぼ規則的な空気の断続が生ずる。これは非対称三角波で近似でき、これが音声の音源となる。これを喉頭原音または声帯音源 (glottal source) と呼ぶ。声帯の緊張が大きく、かつ肺からの空気圧が高いと、声帯の振動周期（基本周期,fundamental period）が短くなって、音源の音の高さが高くなり、逆の時は低くなる。基本周期の逆数を基本周波数（fundamental frequency;F0）と呼び、声の高さ（ピッチ,pitch）に対応する。　　喉頭より上の部分は声道 (vocal tract) と呼ばれ、成人では約 15-17cm の長さがあり、顎、舌、口唇などを動かすことによって、音源波に音色が付与され [a][o] の音声となる。鼻腔は、軟口蓋（口蓋帆）を持ち上げることにより、声道の形を調整することを調音 (articulation) と呼び、声道の各部分の動きを調音運動と呼ぶ。調音に用いられる声道の各部分を調音器官 (articulatory organ) と呼び、その中で、舌、口唇、口蓋帆のように自由に動けるものを特に調音器 (articulator)、調音によって生ずる声道の狭めの位置を調音点または調音位置 (place of articulation) と呼ぶ。調音によって種々の音色が付与されるのは、伝達特性が変わり、共鳴作用によって周波数的にエネルギーの強弱が生ずるためである。

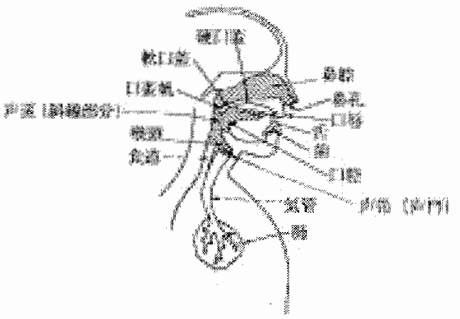


Fig. 3: 人間の発声器官の構造

図 4[1] に、世界の言語に含まれる代表的な母音を示す。横の位置は、左ほど声道の狭めが前（口唇に近い方）にあることを示す。縦の位置は、下ほど下顎が下がって開いていることを示す。図中、一対の母音が左右に密接しているのは、口唇の調音のみ異なる一組の母音で、左側が口唇の狭めのないもの、右側が口唇の突き出しを伴うものを示す。顎の開きの広いものでは口唇の狭めが起こりにくい。

日本語子音の分類を図 5[3] に示す。肺から空気流を音声に変換する機構は声帯振動の他に二つの方法があり、これによって摩擦音と破裂音が生成される。摩擦音は、舌または喉頭によって声道のある場所に狭めを作り、そこを空気が通り抜けるときに乱流が生じて雑音的な音を生ずるものである。破裂音は、舌や口唇で声道を遮断することによって空気流が

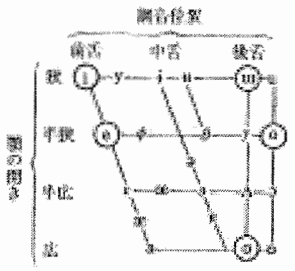


Fig. 4: 世界の言語に含まれる代表的な母音（◯は標準的な日本語で用いられる母音）

を一時的に止め、空気がその後方にたまって圧力が高まったところで、急に開放してインパルス的な音を生ずるものである。破擦音は、破擦後に同じ場所での狭めを保って摩擦音に移行することにより生成するもので、二つの子音が連結したものである。これらの音の生成は、声帯の振動の有無とは独立で、声帯の振動を伴うものを有声子音 (voiced consonant)、伴わないものを無声子音 (unvoiced consonant) と呼ぶ。摩擦音、破裂音などの様式を、調音様式 (manner of articulation) と呼ぶ。

調音様式＼調音位置	口唇		歯，歯茎		口蓋		声門
	有声	無声	有声	無声	有声	無声	無声
摩擦音		ɸ(*)	z	s	ʒ	ʃ	h(*)
破擦音			dz	ts	dʒ	tʃ	
破裂音	b	p	d	t	g	k	
半母音	w		ɾ(**)		j		
鼻音(***)	m		n		ŋ		

*) 日本語のハ行音は特殊な構造を持つ．
**) 日本語のラ行音は弾音として分類される．
***) 日本語のンの音は撥音と呼ばれ，環境によって種々に変化する．

Fig. 5: 日本語子音の分類

そのほかの子音に、半母音と鼻音がある。半母音は母音と同じ機構で生成され、乱流やインパルス的な音を伴うことはないが、母音のように持続せず過渡的である。鼻音は、軟口蓋が下がるとともに、口腔のいずれかの位置で気流を遮断することによって、鼻腔にも空気が供給され、声道に分岐が生じた形で生成される。なお、声帯を振動させず、やや開いて、その部分で乱流を生じさせると、気音やささやき声を作ることができる。　日本語のン（/N/）は撥音と呼ばれ、前後の音素、特に後続音によって [m],[n] あるいは種々の鼻音化母音として現れる。鼻音化母音とは、母音発声時に口蓋帆が下がって、鼻腔が口腔

とともに声道の一部を構成することによる母音の音色の変化である。日本語のラ音は舌先を歯茎の後部に当てて不完全な短い閉鎖をつくることによって生成され、弾音と呼ばれる。日本語にはないが [l] の音は側音と呼ばれ、[l] と [r] の音をまとめて流音と呼ぶことがある。日本語のハ音は、後続母音によってさまざまな調音特性を示し、音源の位置は声門とも調音点のいずれとも特定できない。　母音、子音にかかわりなく、声帯振動を伴う音を有声音 (voiced sound)、伴わない音を無声音 (unvoiced sound) と呼ぶ。　音声は以上のように、音源の性質と声道の共鳴特性によって特徴付けられる。

2.2　調音結合

音素を特徴付ける優勢な周波数成分は、声道の共振周波数に対応し、フォルマント (formant) と呼ばれる。有声音には通常 3 個程度の特徴的なフォルマントがあり、周波数の低い方から、第 1、第 2、第 3 フォルマントと呼ばれる。音韻性の点から特に重要なのは、第 1 と第 2 フォルマントである。フォルマントのピークの周波数をフォルマント周波数と呼ぶ。フォルマント周波数は、同じ音素でも、発話者によって異なり、声道の形が急には変化できないために、前後の音素の影響を受けて更に変化してしまう。　人間の自然な発話では、音素環境や韻律によって音素の特徴が変化する。このような現象を調音結合 (coarticulation) と言う。調音結合とは隣接した音素の特徴が調音器官上で相互にオーバーラップし、時間的に広がりを持った変動性として現れる現象である。そのため隣りの音素だけでなく、更に先の音素にも及ぶことがある。　連続音声の特徴である調音結合には、協調動作、先行性調音、キャリーオーバー、なまけなどの現象が知られている。協調動作とは、個々の音素に固有の声道形状を形成する上で、顎、唇、舌の運動が相補的、補完的な動きをすることである。先行調音とは、音素の部分的な調音特徴が時間的に先行して達成される現象のことである。キャリーオーバーとは、先行音素の特徴が後続音の調音に影響を与える現象を示す。また、なまけとは、発話速度に依存して調音運動が中立化し、孤立発声時の調音とは異なったものとなる現象のことである。これらの現象の背景には、顎、唇、舌などの調音器官が個々に運動学上の自由度を持つと同時に、全体として多自由度の力学系を構成していることや、調音器官の運動がその力学的構造が持つ動特性に拘束されることなどが関係している。そのため、観測される音響的現象もかなり複雑になる。　図 6[4] は、第 1 フォルマント (F_1) を横軸に第 2 フォルマント (F_2) を縦軸にとった $F_1 - F_2$ 平面に、日本語 5 母音の分布を示したものである。男性、女性各約 30 名が発声した単独 5 母音の分布、平均値、および標準偏差を示している。連続音声中の母音は、調音結合のために大きな広がりを持つので、この図よりも母音間の分布の重なりは大きくなる。　一方、子音は波形の周期性の有無（有声、無声）やスペクトルのほか、継続時間長、スペクトルと波形の時間的変化などに、それぞれの音素の特徴がある。子音は定常的な区間を持たないことが多いので母音との調音結合の影響を受けて、特徴が大きく変化する。

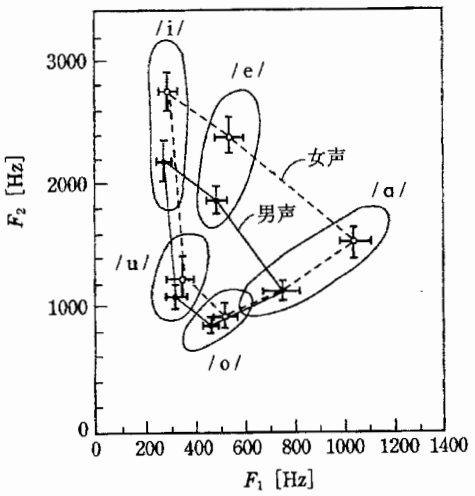


Fig. 6: 日本語母音の第 1、第 2 フォルマンと周波数の分布

2.3 音声認識における調音結合の問題点

一般に音声認識は多くのデータを分析し、そこから得られた特徴空間内における特徴の分布をもとしてに幾つかの正規分布を決定し、これを認識の際に用いる。多くの音声認識システムでは、音素の定常部に着目している。勿論、定常部以外は着目していないわけでなく、triphone というモデルのように前後の音素も考慮に入れてモデル化をする場合もある。調音結合は前後の音素だけでなく、さらに先の音素の影響もあると考えられているのでその点も考慮に入れたモデルを作成しなければならない。ただ、この場合は考えてみれば分かるように、音素の組合せの数は多くなってしまい現実的にモデル化することは困難である。

音声認識は観測された音響的特徴（スペクトル等）から発声された音素系列を推定する逆問題である。この推定の問題は観測されたものでは完全に元のものを表現できない不良設定問題であると考えられる。

観測されたものが何であるかを解くために様々な拘束条件を与えなければならない。

triphone も前後の音素による音素の特徴の変化を拘束条件として取り入れているが、この場合、信号源は何かわからないので、観測されたものから分布を表現するので、複雑な分布でしか表現できない。

調音器官の位置を表す調音パラメータの利点として以下のような点があげられる。

- 調音器官の生理的、物理的性質を直接反映することができる。
- 時間的に比較的ゆっくりと変化するため、効率的な符号化に適する。
- 調音パラメータの時間的補間は物理的に意味を持つ、つまり調音パラメータの補間値で表現される声道形状は現実に存在する。調音モデルを用いた音声合成では、パラ

メータの補間が容易である。また、このような性質は他のパラメータにおいては必ずしも存在するものではない。

音声は調音位置によって各音素の特徴が大きく左右されると考えられ、調音運動を拘束条件として考えた場合、音声の信号源（調音位置）が分かるので音声の出力確率分単を、簡単で、音声生成の理論に基づいているような分布で表現できると考えられる。

3 調音運動データを音声認識に取り入れることの有効性についての検討

調音運動データが音声認識に取り入れるのに有効なデータであるか調べるために音素識別実験を行う。ここでは、音声データ、調音運動データ、音声と調音運動を合わせたデータの３つを用いて、音素の推定実験を行う。

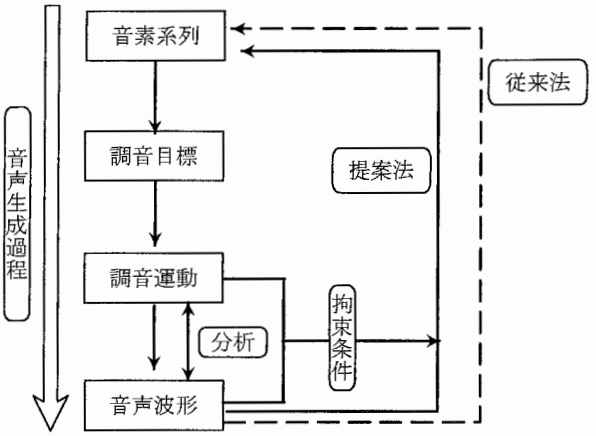


Fig. 7: ３つのデータを用いた音素認識実験

3.1 実験データ

本実験では日本人男性話者３名が発話したデータに関して実験を行った。各話者に関してモデルを作成して、認識実験を行う。学習データ、テストデータは表 1に示す通りである。テストデータには学習データに含まれない文章を用いた。

Table 1: ３話者の実験データ

話者	発話データ	学習データ	テストデータ
MH	354 文章	304 文章	50 文章
TM	375 文章	325 文章	50 文章
TO	363 文章	313 文章	50 文章

ここで３つのデータセットについて説明する。

3.1.1 音声データ

音声データを以下に示す条件で音響分析を行った。

- サンプリング周波数：16kHz
- プリエンファシス：0.97
- 分析窓：Hamming 窓
- 分析窓長：20ms
- 窓間隔：8ms (音響パラメータの周期)
- 特徴パラメータ :48 次 MFCC
  (MFCC 16,Δ MFCC 16,ΔΔMFCC 16)
- 周波数分析 :等メル間隔フィルタバンク
- フィルタバンク :32 チャンネル

音声認識では、通常音声データに対して特徴パラメータ抽出を行い、スペクトルパラメータに変換したものを扱う。特徴パラメータ抽出を行う方法として実用的なものにフィルタバンク分析 (filter bank analysis) と線形予測符合化 (linear predictive cording) がある。本実験では、ハードウェアによる実時間分析の実現が容易であることからフィルタバンク分析を用いて特徴パラメータ抽出を行った。

特徴パラメータは通常、ケプストラム、メルケプストラムなどが用いられている。人の聴覚は音の高さに関してメル (mel) 尺度と呼ばれる対数に近い非線形の特性を示し、低い周波数では細かく、高い周波数では荒い周波数分解能を持つ。このため、音声認識の分野においても特徴パラメータにメルケプストラムが広く用いられている。

フィルタバンク分析を用いた特徴パラメータ抽出の基本的な方法を以下に示す。

1. FFT によるスペクトルを元に、メルスケール上に等間隔に配置された帯域フィルタバンクの出力を抽出する。この様子を図に示す。
2. この出力を対数変換する。
3. 逆フーリエ変換することによりケプストラム係数に変換したパラメータをメル周波数ケプストラム係数 (MFCC;Mel Frequency Cepstrum Coefficient) という。

本研究ではこのようにして抽出された MFCC と、ΔMFCC (MFCC の一次差分)、ΔΔMFCC (MFCC の二次差分) を合わせて、48 次の MFCC を音響パラメータとして実験に用いる。

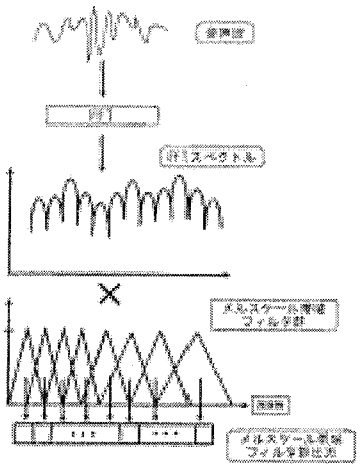


Fig. 8: FFT に基づくメルスケール帯域フィルタ群分析の手順

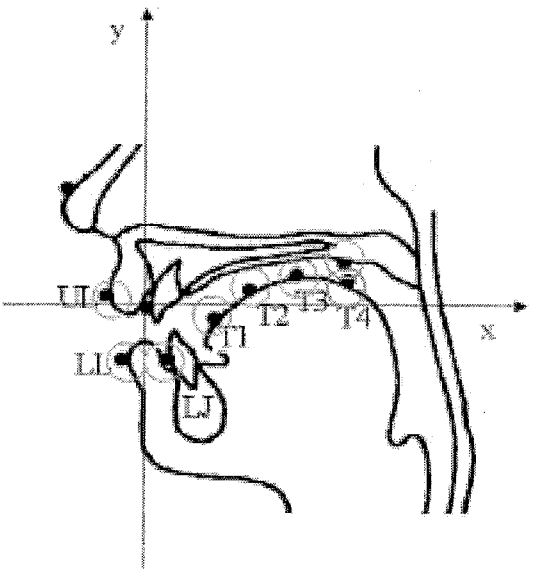


Fig. 9: 観測点

3.1.2 調音運動データ

本研究で用いた調音運動データは NTT の磁気センサシステム EMA(electro-magnetic articukometer) を用いて測定されたものを用いた。調音運動データは正中矢状断面上における上唇、下唇、軟口蓋、下顎にそれぞれ 1 点、舌尖部から舌背部までの区間で等間隔に 4 点、舌 1、舌 2、舌 3、舌 4 の計 8 点がある。上顎に取りつけたコイルの位置を原点として、前部から後部への水平位置を x 軸、下部から上部への垂直方向を y 軸として、合計 16 点の位置データがサンプリング周波数 250Hz(4ms:調音パラメータの周期) で収録されている。

上唇、下唇、下顎で唇の突出し、開口などを、舌 4 点で舌の位置、狭めを、軟口蓋は結合度により鼻腔の影響を表現することができると考えられる。また、各調音器官や声道は 3 次元構造を持つが、声道形状に関する情報をもっとも多く含む正中矢状断面で考える。

本実験では、16 個の位置データに対してそれぞれ移動距離に対して単位時間内の変化率 (速度)、さらにその速度の単位時間内の変化率 (加速度) を求め、48 次の特徴パラメータを作成して実験に用いた。

3.1.3 調音運動と音声を合わせたデータ

調音運動と音声を合わせたデータの作成法として、画像音声統合システムで行われている初期統合の手法で行い、音声の特徴と調音運動の特徴を合わせて一つのパラメータとして用いる。ここでは、音声に関して MFCC(16 次),ΔMFCC(16 次) の 32 次の音響パラメータを用い、調音運動に関しては位置データ (16 次) を調音パラメータとして用いて、それらすべてを合わせて、48 次のパラメータとして用いた。これは、音声、調音運動によるそれぞれの音素認識実験結果から、音声を用いた方が調音運動を用いた場合より高い認識結果が得られたため、音声の情報を多く持たせるためにこのようなパラメータを用いた。

3 つのパラメータに関してそれぞれ HMM を作成する。今回、実験に用いた HMM は共分散行列の非対角要素をすべて 0 に設定した対角共分散行列 HMM を用い、図 2 に示すような自己遷移する状態が 3 の left-to-right monophone モデルを用いた。音素 HMM は 29 種類作成し、混合数を 1,2,3,4,5,6,8,12,16 と変えて実験を行った。

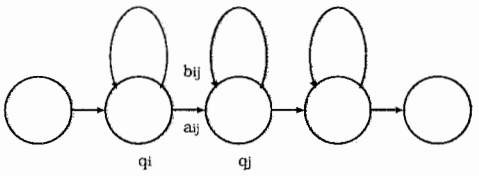


Fig. 10: 状態数 3 の HMM

3.2 実験結果

以下に、各話者ごとの音素認識実験の結果を示す。

認識率の評価の方法として2種類ある。まず、エラーとして

- S:(substitution errors; 置換エラー)
- D:(deletion errors; 脱落エラー)
- I:(insertion errors; 挿入エラー)

の3つがある。置換エラー (S) は異なる音素に置換された誤りの数、脱落エラー (D) は正しい音素が脱落した誤りの数、挿入エラー (I) は逆に異なった音素が挿入されたエラーの数である。

認識の評価法は正解率 (Percent Correct) と正解精度 (Percent Accuracy) の2つがあり、以下のようにして計算される。ここで N は入力音声中の総音素数である。

$$\text{正解率 } (PercentCorrect) = \frac{N - D - S}{N} * 100 \tag{1}$$

$$\text{正解精度 } (PercentAccuracy) = \frac{N - D - S - I}{N} * 100 \tag{2}$$

正解率は挿入エラーを考慮に入れていないが、正解精度では考慮に入れるので、正解精度の方が厳しい評価になる。

本研究での音素認識結果の認識率として、正解精度 (Percent Accuracy) を用いる事とし、以下の認識結果はすべて正解精度を示す。

Table 2: 話者 MH の初期統合モデルによる音素認識実験

混合数	調音運動 (%)	音声 (%)	調音+音声 (%)
1	71.10	78.66	78.82
2	74.88	81.90	84.28
3	75.85	83.52	85.95
4	76.34	84.60	86.62
5	76.93	85.20	87.30
8	77.74	85.90	87.74
12	80.44	86.55	89.09
16	76.99	84.93	86.71

調音運動のみを用いた場合と音声のみを用いた場合を比較すると、音声の方が高い認識率を得る事が分かった。次に、音声と調音運動を合わせたデータを用いた場合と音声のみを用いた結果を比較すると、音声と調音運動を合わせた方が良い認識率を得ることができた。これは、調音データは音声には無い情報を持っていると言うことができる。

Table 3: 話者 TM の初期統合モデルによる音素認識実験

混合数	調音運動 (%)	音声 (%)	調音+音声 (%)
1	64.83	76.39	79.85
2	73.10	82.17	82.98
3	76.39	84.98	85.79
4	76.23	84.55	85.04
5	77.74	84.93	85.63
8	78.07	85.58	87.47
12	79.52	87.03	88.60
16	78.61	85.25	86.17

Table 4: 話者 TO の初期統合モデルによる音素認識実験

混合数	調音運動 (%)	音声 (%)	調音+音声 (%)
1	61.70	69.64	73.47
2	70.50	71.64	78.71
3	71.85	73.80	80.01
4	72.45	74.88	80.71
5	72.72	75.31	81.36
8	73.04	76.23	81.90
12	75.31	78.55	82.77
16	73.74	75.63	81.20

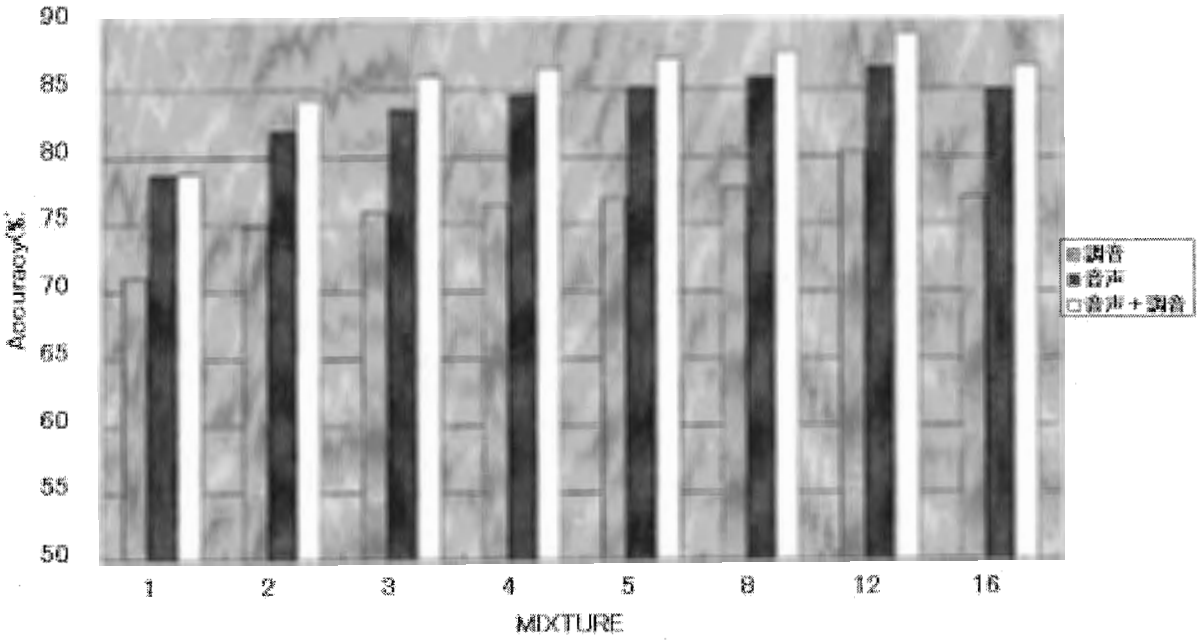


Fig. 11: MH 話者の初期統合モデルによる音素認識実験

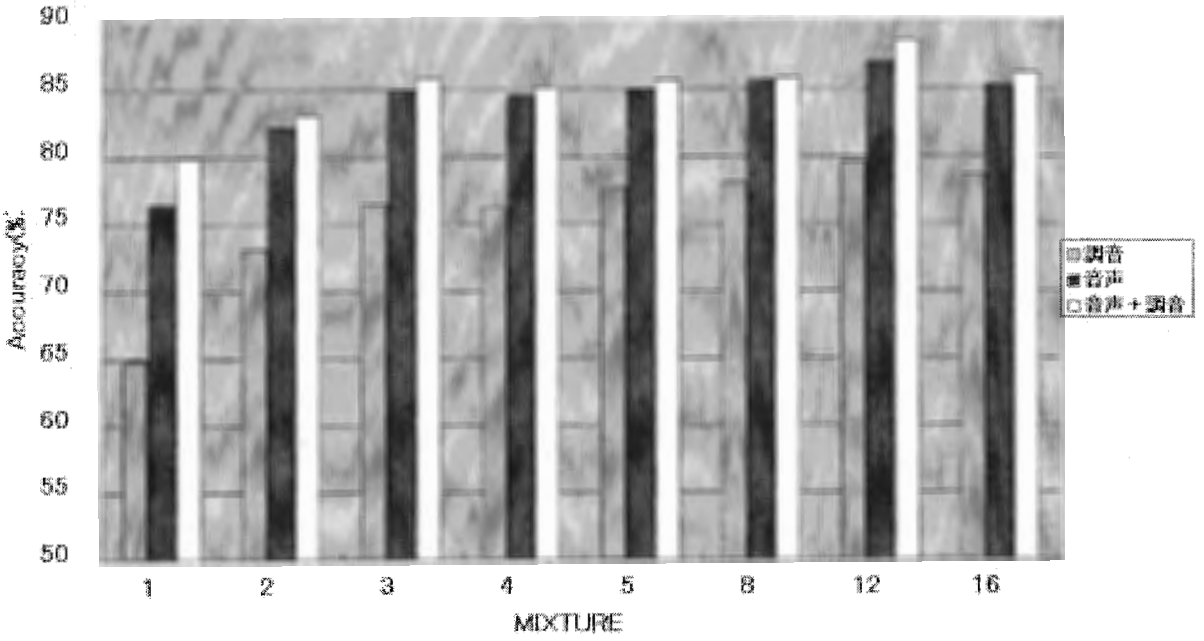


Fig. 12: TM 話者の初期統合モデルによる音素認識実験

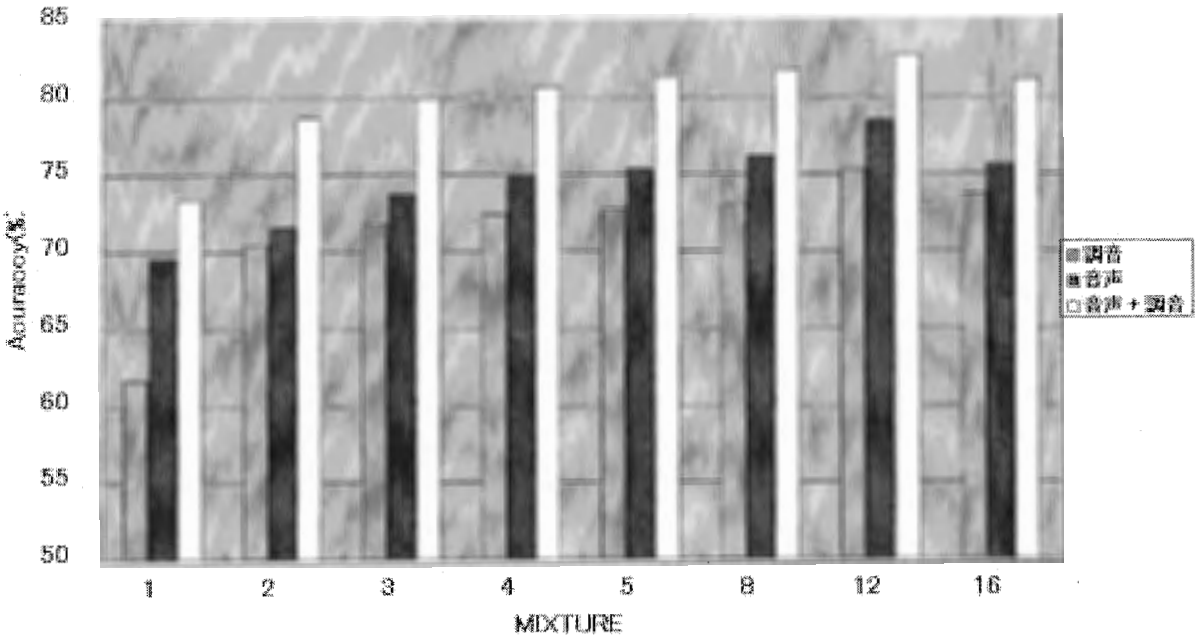


Fig. 13: 話者 TO の初期統合モデルによる音素認識実験

音声と同様に調音運動を情報として音声認識に取り入れる事が有効であると考えられる。今回、調音パラメータと音響パラメータを一つのパラメータとして扱い、モデルを作成したが、これは、音声と同期した調音運動データがある場合にのみ用いることができる。しかし、音声に同期した調音運動データを常に観測することは困難であるので、調音運動と音声との間の関係を拘束条件として音声認識のモデルに取り入れて、音声のみのデータを認識する方法を考えなければならない。その手法として、BayesianNetwork と HMM を組み合わせた Hybrid HMM/BN モデルを用いて、音響パラメータと調音パラメータとの間に関係付けて利用することを試みる。

4 音声認識とベイジアンネットワーク

4.1 従来の音声認識法（HMM）

音声認識の基本的な問題は、発話された音声に一致するような単語列を書き出すことができるかということである。統計的なアプローチとしてこの問題は与えられた音声データから尤もらしい単語列を見つけるということである。つまり $P(W|X)$ の確率が最大になる単語列 W を求めるということである。(X は音声データから得られた特徴ベクトル、W は単語列) この式はベイズの定理より以下のように表すことができる

$$P(W|X) = \frac{p(X|W)P(W)}{p(X)} \tag{3}$$

これは音響モデルと言語モデルの２つの部分に分けて推定を行う。音響モデルはデータによって推定され、言語モデルは単語列の事前確率として推定される。これは言語モデルと音響モデルを独立にも扱うことができることを意味している。音響モデルが HMM に基づく場合、単語列は HMM の系列によって表される。少ない語彙の音声認識において一つの単語に対して一つの HMM が用いられる。しかし、大語彙のサブワード単位の HMM に対しては、HMM は多くの学習データセットが必要になる。この場合、サブワードは音素などが用いられる。

図 14は音素を表すときに用いられる、状態数３の left-to-right のモデルである。確率分布 $P(x|q)$ は混合ガウス分布で表現される。

4.2 ベイジアンネットワーク

ユーザの意図のように確定値を得ることが難しいものがある。これらを体系的に取り扱うために確率変数を用いることができる。以降では簡便のため確率変数が離散の場合を中心に説明する。例えば複雑な要因やノイズの影響などによって不確定さを含む対象を確率変数として大文字 X で表し、その変数がとりえる具体値は小文字 $x_1, x_2, \cdots, x_n$ のように表す。ここで変数の確率分布を考えることができ，そのおおまかな傾向はその統計量 (例えば平均や分散、エントロピーなど) によって特徴づけることができる。

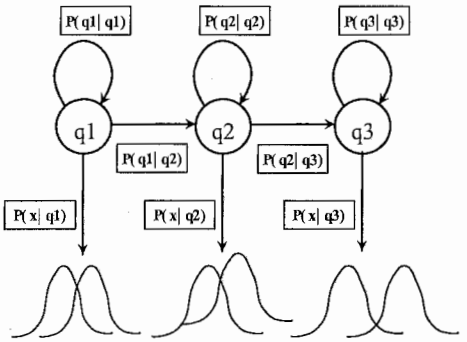


Fig. 14: 状態 3 の left-to-rightHMM モデル

次に変数間の依存関係を考える。例えば変数 X_i と X_j があり「if $X_i = a$ then $X_j = b$」というルールが成立しているとき、X_j が X_i に依存しているという。現実に起きている様々な事象を考えるとこうした変数間の依存関係は複雑であり、「if $X_1 = a_1, \cdots, X_i = a_i, \cdots$,then $X_j = b$ 」のように明示的に全ての関係を列挙することは現実的でない．また予期できない要因によってルールが常に成立するとは限らない場合もある．こうした不確実性を吸収し，また依存関係の程度を定量的に表すため，「$X_i = a_i$ あるとき $X_j = b$ である確率は $P(X_j = b|X_i = a_i)$」という確率的な言明を与える．二つの量 x, y の間の一意的な依存関係は，例えば関数 $y = f(x)$ によって表せる．同様に，確率変数 X_i,X_j の依存関係は条件付き確率分布 $P(X_j|X_i)$ によって表すことができる．これは X_i のとる値に応じて，X_j の分布が影響をうけ，その依存関係の定量的関係が条件付き確率分布 $P(X_j|X_i)$ で定められることを示している．

確率変数をノードで表し，因果関係や相関関係といった依存する関係を持つ変数の間にリンクを張ったグラフ構造による確率モデルが確率ネットワーク (あるいはグラフィカルモデル [Whittaker 90, 宮川 98]) と呼ばれ，その中でとくにリンクが因果関係の方向に向きを持ち，このリンクをたどったパスが循環しない，非循環有向グラフで表されるモデルがベイジアンネットワークである．ベイジアンネットは確率変数間の定性的な依存関係をグラフ構造によって表し，変数間の定量的な依存関係はその変数の間に定義される条件付き確率によって表すことで問題領域をモデル化する．先に述べた X_i, X_j の間の条件付き依存性をベイジアンネットワークでは向きのついたリンクによって $X_i \rightarrow X_j$ と表し，X_i を親ノード，X_j は子ノードと呼ぶ．親ノードが複数あるとき子ノード X_j の親ノードの集合を $\pi(X_j) = \{X_1, \cdots, X_i\}$ と書くことにする．この場合の変数 X_j に関する依存関係は条件付き確率,

$$P(X_j|\pi(X_j)) \tag{4}$$

で定義され，これは X_j を子ノード, $\pi(X_j)$ を親ノード群とする木構造になる．さらに n 個の確率変数 $X_1, \cdots, X_n$ があるとき，全ての確率変数の同時確率分布は式 (2) のようにな

り，各子ノードとその親ノード群からなる局所木を組み合わせたグラフ構造で表せる図15.

$$P(X_1, \cdots, X_n) = \prod_j P(X_j|\pi(X_j)) \tag{5}$$

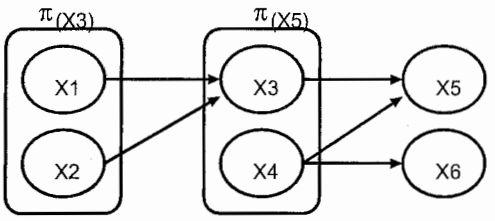


Fig. 15: ベイジアンネットワーク

つまり，式5の左辺の同時確率分布は局所的な木構造に分割した右辺の各項の積として計算される.

このようなグラフ構造と、各ノードに割り当てた条件付き確率集合により、ベイジアンネットワークが構成される。離散変数の場合、子ノードの親ノードに関する条件付き確率はすべての状態における条件付き確率を並べた表、CPT(Conditional Probability Table) によって表現される。

4.3 ベイジアンネットワークの変形-Dynamic Bayesian Network-

以上で説明したベイジアンネットワークを実際に適用する場合，変数のとり方によって様々なバリエーションを与えることができる．例えばノイズを含むアナログセンサからの観測値などは連続確率変数になるので、離散確率変数のかわりに連続確率変数 でベイジアンネットワークを構成する。この場合には条件付き確率は連続的な関数によって表す必要があり、適当なパラメータを導入して特定の関数族を仮定することで表現する。とくにガウス分布が用いられることが多く、このように構成したベイジアンネットワークはガウシアンネットワーク [Geiger 94] あるいは Conditional Gaussian[Lauritzen 89] と呼ばれる．例えば親ノード $X = x$, 子ノード $Y = y$ とすると，条件付き確率は,

$$P(y|x) = g(x)e^{-f(x,y)} \tag{6}$$

のように書ける (f, g は適当な関数として与える)

また時間遅れ (状態遷移) を持つ系をモデル化する場合のベイジアンネットワークとして Dynamic Bayesian Network(DBN) と呼ばれるものがある [Dean 89]。時系列を扱う確率モデルとして工学的に重要なマルコフチェイン，あるいはその発展形であり，音声認識や遺伝子情報処理で使われている隠れマルコフモデル （Hidden Markov Model:HMM） 4[Rabiner 93] があるが，このような時間的な遷移をモデル化するために， DBN はベイジアンネット

ワークの一部の有向リンクに状態遷移と時間順序の意味を持たせ，時間軸に沿って必要な回数だけ繰り返し適用することで，時刻パラメータを持つ有限個の（独立かつ同一分布に従う）確率変数 $\{X_1(t), \cdots, X_N(t)\}$ の分布を表現する．隠れマルコフモデルは状態遷移と出力として確率を割り当てるのに対し，DBN は全てのノードに対する条件付き確率として割り当てる点が異なり，ある時刻の状態は任意の複数のノードの組で表せる．

5 Hybrid HMM/BN model

[10]

5.1 概要

HMM を用いた音声認識では、観測される状態分布を混合ガウス分布で表現してきた。多くの場合、観測されたものからスペクトルを特徴抽出する。音声認識において高い認識精度を得るにはこれらの特徴だけでは不十分である。多くの研究者たちは付加的な特徴を HMM システムに取り入れることを試みている。Baysian network(BN) は多くの研究者から HMM に取って代わるものとして注目を浴びている。BN は人工知能（AI）の研究分野ではよく知られているが音声認識では新しいトピックである。BN は多くの異なったランダムな変数の同時確率分布を適切な構造でモデル化することができ、特に音声の時間上の特徴をモデル化するのに (DBN) が適切である。孤立単語認識の単語モデルとして DBN を用いた研究が実際に行われている [11][12]。これらの研究では DBN は HMM を一般化したものとみなされていて、音声のスペクトル情報は簡単に追加的な知識、調音器官の特徴や発話スタイルなどを加えることができる。

BN の有利な点として、追加された特徴は認識するときに観測できなくても良いということがあげられる。このような有利な点があるにも関わらず BN を用いた音声認識においては少数の孤立単語認識に制限されていた。その理由は BN のパラメータの学習と推定のアルゴリズムは連続音声認識、特に大語彙連続音声認識には適切ではなかったためである。

Hybrid HMM/BN モデル [10] は、HMM と BN を組み合わせることで上記の欠点を解決することができる。本研究では、この Hybrid HMM/BN モデルを調音パラメータと音響パラメータの 関係づけに利用することを試みる。このモデルは音声信号の特徴を HMM 状態遷移でモデル化して、BN は HMM 状態分布をモデル化するのに用いられる。これらは階層的に表され、BN は下層で HMM は上層として表現される。このモデルの有利な点は従来の HMM として振る舞うので変更せずに既存の認識アルゴリズムを用いることが可能で、大語彙認識システムに不可欠な subword や word の両方について用いることができるということである。

5.2 モデルの構造

音声のような時系列過程は DBN によって表される。一般的な HMM を DBN として表現したものを図 15に示す。Q_t、Y_t はそれぞれ時刻 t の状態変数、観測値変数を表している。

各状態間の矢印は状態遷移確率を表し、状態と観測値の間の矢印はHMMの状態に関する条件付き確率分布を表している。また、丸は連続値、四角は離散値を表し、影がついているものは観測できる変数、ついてないものは観測できない（隠れている）変数を表している

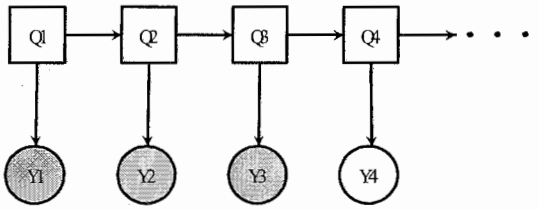


Fig. 16: HMMをDBNで表した場合

hybridHMM/BNモデルについていくつかのステップで説明する。まず、図16のDBNについて考え、ノード間の弧を消して考える。そうすると図17のように時刻tに対応する複数の独立したBNが得られる。次に、時間遷移を従来のHMMにより決めて、これらのBNを適当なHMMの状態に割り当てるとすると、時間を表すインデックス（番号）が消えて、すべてのBNが同じ構造を持ち図18のようなひとつのBNとして表すことができ、変数Qは音響モデルにおけるすべてのHMMの状態番号の値を取り、状態確率分布$P(Y|Q=q_{ij})$はアーク(弧)によって表される。

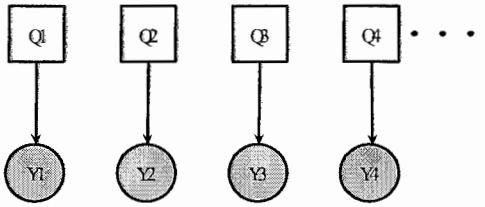


Fig. 17: 各時刻に関するBN

Fig. 18: 状態BN

言い換えると、従来の HMM で用いられているガウス分布の代わりに状態分布モデルとして BN を取り入れるということである。

このようなモデル hybrid HMM/BN モデルと呼ぶ。HMM と BN をこのようにして組み合わせることにより HMM/BN モデルを図 19のように、BN が下層、HMM が上層として階層的に表すことができる。状態変数 Q は BN においては観測可能であるが、上の HMM の層では隠れていることになる。

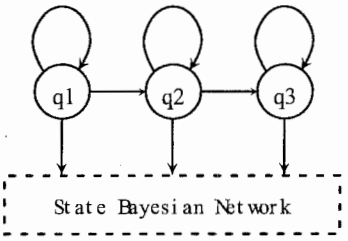


Fig. 19: HMM/BN モデルの構造

状態 BN は簡単に付加的な知識を表す他の任意の変数を含むように拡張が可能である。拡張された BN の図式的な構造は、データから学習されるよりも変数間の関係の知識によって課すことができる。例として、状態 BN の可能な構造を図 20で示す。例えば、この図の変数 X は環境ノイズの種類を表すことができ、他の W や Z は話者 ID（識別番号）やその人の母国語を表すことが可能である。観測される確率分布 Y はノイズにより変化し、因果関係があると考えられ、話者や言語によっても音声の特徴が変化し、話者とその人の母国語にも因果関係があると考えられるので BN によってさまざまな因果関係を表現することが可能になる。

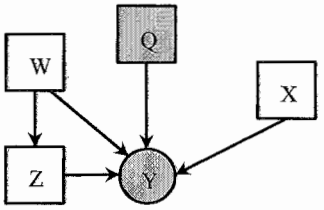


Fig. 20: BN の構造

BN を持ちいる利点は、連続変数と離散変数を簡単に結合できという点である。HMM/BN は一般的な HMM と同様に扱うことができ、HMM/BN モデルはすべてのシステムにおいて HMM の代わりに用いられることが可能である。

5.3 モデルの学習と認識

HMM/BN モデルの学習は、Viterbi アルゴリズムに基づいている。以下にそのアルゴリズムを示す。

1. 音素 HMM を用いて Viterbi アライメントを行い各フレームをある音素 HMM の状態に割り当てる。
2. 各音素の各状態に対応する音声データと BN の変数（X、W、Z など）を集め、モデル作成する。
3. 作成したモデルを学習データによって状態遷移確率を再推定する。
4. 再推定されたモデルを用い、1 に戻って繰り返す。

次章において、Hybrid HMM/BN の学習をより詳しく説明する。

HMM/BN モデルの認識には従来の HMM の場合と同じように Viterbi アルゴリズムが用いられる。ここで、すべての状態 Q に対する $P(y|Q)$ を計算する必要がある。この値を一般的な推定アルゴリズムを用いて、BN 確率モデルから推定することが可能である。

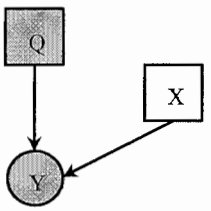


Fig. 21: BN の構造

この図 20の BN に対する同時確率モデルは以下に示す連鎖法則によって説明することができる。

$$P(Y,X,Q) = P(Y|X,Q) * P(X|Q) * P(Q) \tag{7}$$

X と Q は独立変数なので（それぞれの間に弧が無く、繋っていない）上に示す式はこのように書き換えられる

$$P(Y,X,Q) = P(Y|X,Q) * P(X) * P(Q) \tag{8}$$

そして、求めたい状態 Q における Y の出力確率分布は $P(Y|Q)$ は以下のように計算出来る。

$$\begin{aligned} P(Y|Q) &= \frac{P(Y,Q)}{P(Q)} = \frac{\sum_x P(Y,X=x,Q)}{P(Q)} \\ &= \frac{\sum_x P(Y|X=x,Q) * P(X=x) * P(Q)}{P(Q)} \\ &= \sum_x P(Y|X=x,Q) * P(X=x) \end{aligned} \tag{9}$$

多くの実験において、$P(X)$ はすべての $X = x$ に対して同じであると仮定することができ、上式は以下のようになる

$$P(Y|Q) = \frac{1}{N(x)} \sum_{x} P(Y|X = x, Q) \tag{10}$$

N(x) は X が取ることができる値の数である。

6 調音パラメータを用いた Hybrid HMM/BN モデル

音声の特徴は調音位置によって決まるので調音パラメータと音声には因果関係があると考えられ、hybrid HMM/BN モデル取り入れることで、その関係を表現することができる。この場合、調音パラメータは前章の図 21における付加的な変数 X として用いる。図 21における X のような付加的な変数は観測できないパラメータとして扱えるので、音声認識の時、調音パラメータが観測できなくても用いることができるので調音パラメータをこのモデルに取り入れることが可能であると考えられる。

付加的な変数は hybrid HMM/BN モデルで用いる場合には離散値として用いる。本研究で用いている調音パラメータは 16 次の位置データである。このパラメータをそのまま用いずに分析を行って抽出した離散値を用いる。調音パラメータを離散値として扱わなければならないのでまず、主成分分析を行い次元数を落とし、ベクトル量子化を行い離散値として扱うことを考える。

以下に、本研究における調音運動の分析として用いた主成分分析（PCA）とベクトル量子化（VQ）について説明する。

6.1 調音パラメータの分析方法

6.1.1 主成分分析 (PCA)

主成分分析 (Principal Component Analysis;PCA) は、多変量解析の手法の一つで、複数の変数がもつ情報をひとまとめにして要約するということになる。つまり、いくつかの変数から、主成分と呼ばれる合成変数を作るということである。この主成分は、互いに無相関となるように算出される。主成分分析の概要を理解するために、簡単な例をとりあげて説明する。一番簡単な例は２次元のデータを１次元に要約するという問題である。

以下に示す表のように、それぞれの学生ごとに、２科目の試験の成績が記録されている。すなわち、２次元のデータである。この試験の成績を使って、総合順位を算出したい。一番簡単な方法は、単純に合計点を計算することである。ただし、試験によっては難易度も違うだろうし、科目ごとに重みをつけたいということもあるはずである。たとえば、英語のできる人よりも数学のできる人 を優先させたい、ということなら、数学の点数の重みを大きくすればいい

そこで、このような「総合順位」を求めるという問題は、一般的には、以下のような線形式の係数を決定する、という問題に置きかえることができる。

$$z = \beta_1 x + \beta_2 y \tag{11}$$

Table 5: 2科目の試験成績

学生	数学	英語
1	87	85
2	79	75
3	97	75
4	66	97
5	83	66
6	67	54
7	84	65
8	54	86
9	56	69
10	67	89

この時、係数β_1やβ_2をどのように定めれば良いか、ということ自体はその時その時で様々な考え方がある。単純な合計点だけ求めれば良いのなら、どちらも1にすれば良い。たとえば数学の重みを英語の倍にしたいなら、$\beta_1 = 2$、$\beta_2 = 1$にすれば良い。

この時、特にどの科目を優先させたいということはないが、なるべく「情報の損失」がないようにしたい、と考えたとする。元々2次元のデータを1次元にしてしまうのだから、必ず情報が損失する。式が与えられたとき、2次元のデータから1次元のデータを計算することは常に可能だが、逆に1次元のデータから2次元のデータを復元することは不可能であ。すなわち情報が損失している。ここで、情報の損失がなるべく少なくなるようにということを考える。情報量が多いということはその変数の分散が大きいという関係があると言える。合成変数の分散がなるべく大きくなるようにすることを考える。

合成変数の分散を計算すると

$$\sigma_z^2 = \beta_1^2\sigma_x^2 + 2\beta_1\beta_2\sigma_{xy} + \beta_2^2\sigma_y^2 \tag{12}$$

この式からわかるように、係数を大きくすればするほど分散が大きくなってしまう。そこで、係数ベクトルの長さを1とするいう制約条件をつけて解くことにする。つまり

$$\beta_1^2 + \beta_2^2 = 1 \tag{13}$$

ここで、ラグランジェの未定係数法を用いると、

$$L = \beta_1^2\sigma_x^2 + 2\beta_1\beta_2\sigma_{xy} + \beta_2^2\sigma_y^2 - \lambda(\beta_1^2 + \beta_2^2 - 1) \tag{14}$$

となり、それぞれの係数ごとに偏微分した式を0とおいた、

$$\frac{\partial L}{\partial \beta_1} = 2\beta_1\sigma_x^2 + 2\beta_2\sigma_{xy} - 2\lambda\beta_1 = 0 \tag{15}$$

$$\frac{\partial L}{\partial \beta_2} = 2\beta_2\sigma_y^2 + 2\beta_1\sigma_{xy} - 2\lambda\beta_2 = 0 \tag{16}$$

を連立方程式として解けば良い、ということになる。ここで、分散・共分散行列、および係数ベクトルを用いて書き直すと、

$$\begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \lambda \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \tag{17}$$

となる。すなわち、求める係数ベクトルは、分散・共分散行列の固有ベクトルとなることがわかる (λ は固有値)。

ここまでの議論は、データが２次元だけでなく、多次元のデータであっても同じように成立する。すなわち、多次元で表現されているデータの分散・共分散行列の最大固有値に対応する固有ベクトルを求めるという問題と、その多次元変数の線形結合で表わされる合成変数の分散を最大化させる係数ベクトルを求めるという問題とは等価である。例としてｎ次の変数ベクトル $x = (x_1, x_2, \cdots, x_n)$ を考える。x について大きさＮの標本 $x_1, \cdots, x_N$ が与えられているとするとき、この平均ベクトル、分散共分散行列をそれぞれ μ,Σ とおくと

$$\mu = \frac{1}{N}\sum_{j=1}^{N} x_j \tag{18}$$

$$\Sigma = \frac{1}{N}\sum_{j=1}^{N} (x_j - \mu)(x_j - \mu)^t \tag{19}$$

と表すことができる。

Σ の固有値 $(\lambda \geq, \cdots, \geq \lambda_n)$ を対角要素とする行列と固有ベクトル行列をそれぞれ

$$\Lambda = (\lambda_1, \cdots, \lambda_n), \quad \Gamma = (\gamma_1, \cdots, \gamma_n) \tag{20}$$

とおけば

$$\Sigma = \Gamma\Lambda\Gamma^t, \quad \Lambda^t\Lambda = I \tag{21}$$

となる。

よって主成分 y はベクトル γ と x の一次結合として

$$y = \gamma^t(x - \mu), \gamma^t\gamma = 1 \tag{22}$$

と表される。ここで、$y_i = \gamma_i^t(x - \mu)$ を x の第 i 主成分、γ_i を第 i 主成分ベクトルと呼ぶ。

このように、分散・共分散行列の固有値を求め、それらを大きさの順にならべたとき、最大固有値に対応する固有ベクトルを係数ベクトルとする線形式を、そのデータの「第一主成分」と呼ぶ。同じように２番目に大きい固有値に対応する固有ベクトルを係数ベクトルとする線形式を「第二主成分」と呼ぶ。以下同じように、元々のデータがｎ次元で記述されているなら、「第ｎ主成分」まで存在する。

また

$$\lambda_j / \sum_{i=1}^{n} \lambda_i, \quad (j = 1, \cdots, n) \tag{23}$$

を第 i 主成分の寄与率といい、y_i の分散の全変動に対する割合を表している。

$$\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{n} \lambda_i \quad (k \leq n) \tag{24}$$

を第 k 主成分までの累積寄与率という。この値でどれくらい元のデータの情報を失わずに表現できるかを判断する。この値を見て主成分の数を決める。

6.1.2 ベクトル量子化 (VQ)

[1] 音声波形の瞬時値のディジタル化はスカラ量子化と呼ばれる。それに対して、複数の値の組（ベクトル）をまとめて一つの符号で表現する量子化方法をベクトル量子化 (Vector Quantization;VQ) という。これは波形の時系列や音声の特徴パラメータ（LPC などによるベクトル表現）に適応され、ベクトルの低ビット表現が可能であることから、主にデー多量の削減化方法として用いられてきた。また、ベクトル時系列を一括して量子化する方法もあり、マトリクス量子化と呼ばれている。ここで例として、ベクトル量子化手法による LPC パラメータの高性能符号化法について考えてみることにする。音声波形は LPC 分析により短時間区間ごとに LPC パラメータに変換され、ベクトルの時系列 $x_1, x_2, \cdots, x_I$ として表現される。各ベクトル x_i はあらかじめ求められている典型的なベクトル集合 $\{y_1, y_2, \cdots, y_N\}$（コードブック）の一つの要素であるベクトル $y_{\hat{n}}$（コードベクトル）に置換される。ここで、

$$\hat{n} = \arg\min_{n} d(x_i, y_n) \tag{25}$$

もし、コードブックが $N = 2^k$ であれば、各ベクトル x_i は k ビットで表現されることになる。当然、N を大きくすれば量子化誤差は小さくなるが、N を固定してもコードベクトルの選び方によって量子化誤差は変化する。母集団の分布の偏りを利用して、ベクトル量子化による歪みがサンプル全体として最小になるコードブックを求めることが必要になる。

多数の訓練サンプルからこれらのサンプル集合を最も良く近似する N 個のコードベクトルを求めるには、クラスタリング手法によって行われる。音声認識にマルチテンプレート作成のためのクラスタリング技法としては $k-means$ 法が良く知られている。本研究では、代表的なコードブック作成アルゴリズムとして LBG アルゴリズムと 2 分割繰り返しアルゴリズムを用い、符号長探索、学習の距離尺度にはユークリッド距離を用いた。以下にそのアルゴリズムを示す。[6]

ここで、セントロイドベクトル $\hat{x}$ を次式で定義する

$$\hat{x} = C(x_0, x_1, \cdots, x_n) = \arg\min_{x} \frac{1}{n} \sum_{i=0}^{n-1} d(x, x_i) \tag{26}$$

LBG アルゴリズム

1. 初期化

   訓練サンプル集合　　：$\{x_j,; j=0,\cdots,n-1\}$
   コードブックサイズ　　：N
   コードブックの初期集合：$A_0^{(N)}=y_0^{(0)},\cdots,y_{N-1}^{(0)}$
   ステップ数　　：$m=0$
   平均歪み　　：$D_{-1}=\infty$
   歪み閾値　　：ϵ

2. x_j を $A_m^{(N)}$ によって N 個の部分集合 $\{S_i; i=0,\cdots,N-1\}$ に分割する。すなわち
   もしすべての t について、$d(x_j,y_i^{(m)})<d(x_j,y_t^{(m)})$ なら

$$x_j\in S_i$$

3. 平均歪みを計算

$$D_t=\frac{1}{n}\sum_{i=0}^{N-1}\sum_{j\in S_i}^{n-1}\left\{d(x_j,y_i^{(m)})|x_j\in S_i\right\}$$

   もし、$(D_{m-1}-D_m)/D_m<\epsilon$ ならば終了し、$A_m^{(N)}$ をコードブックとする。それ以外なら 4 へ。

4. $A_{m+1}^{(N)}=y_0^{(m+1)},\cdots,y_{N-1}^{(m+1)}$ を求める。ここで
   $y_i^{m+1}=C(\{S_i\}), m=m+1$ として 2 へ

2 分割繰り返しアルゴリズム

1. 初期化

   Δ :大きさの小さい適当なベクトル
   $A_{0,1}=C(x_0,x_1,\cdots,x_{n-1})$:全サンプルのセントロイド
   $M=1$

2. $A_{0,M}=\{y_0,y_1,\cdots,y_{M-1}\}$ に対して、各 y_i に近接した 二つのベクトル

$$y_i+\Delta, y_i-\Delta$$

   に分ける

$$\{y_0+\Delta,y_0-\Delta,y_1+\Delta,y_1-\Delta,\cdots,y_{M-1}+\Delta,y_{M-1}-\Delta\}$$

   を改めて $A_{0,2M}=y_0,y_1,\cdots,y_{2M-1}$ とする。

3. $A_{0,2M}$ を初期値として上記のLBGアルゴリズムによって準最適な $A_{0,2M}=y_i; i=0,1,\cdots,2M-1$ を求める。M=N なら終了それ以外なら M=2M として 2 へ

6.2 調音パラメータを用いた Hybrid HMM/BN model

分析し抽出された調音パラメータを x、観測された音声パラメータを y、HMM の状態を Q とすると、HMM/BN モデルは図のように書くことができる、破線で囲った部分は BN を表している。また、丸は連続値、四角は離散値を表し、影がついているものは観測できる変数、ついてないものは観測できない（隠れている）変数を表している

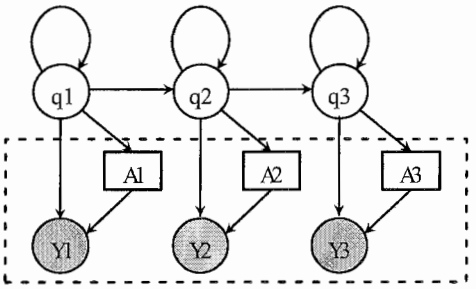


Fig. 22: 調音パラメータを用いた HMM/BN モデル

ここで、状態遷移の弧を考えずに BN の一つの状態を取り出して見ると図のようになる。

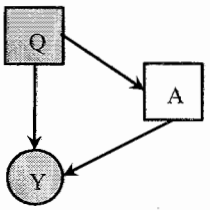


Fig. 23: 調音パラメータを用いた BN

このモデルの同時確率は

$$P(Y, A, Q) = P(Y|A, Q) * P(A|Q) * P(Q) \tag{27}$$

A は Q に依存している。

そして、求めたい確率 $P(Y|Q)$ は

$$P(Y|Q) = \frac{P(Y, Q)}{P(Q)} = \frac{\sum_a P(Y, A = a, Q)}{P(Q)}$$

$$= \frac{\sum_a P(Y|A = a, Q) * P(A = a|Q) * P(Q)}{P(Q)}$$

$$= \sum_{a} P(Y|A=a,Q) * P(A=a|Q) \tag{28}$$

と表すことができる。従来のHMMと同様に、確率分布$P(Y|A=a,Q)$はガウス分布でモデル化でき、$P(A=a|Q)$はその混合重みとして考えることができる。

ここで本研究で用いた、調音パラメータを取り入れたHMM/BNモデルの学習のアルゴリズムについて説明する。

1. 音素HMMを用いてViterbiアライメントを行い各フレームをある音素HMMの状態に割り当てる 。
2. 各音素HMMの各状態に対応する音響パラメータと調音パラメータから得られた同じラベルを集め、ラベルごとに音響パラメータを分け、モデルパラメータ（平均、分散）を求める。混合重み係数は

$$P(A=a|Q) = \frac{\text{クラス } a \text{ のサンプル数}}{\text{全てのクラスのサンプル数}}$$

として求める。このとき、ある閾値を設定しその数以下のサンプル数を持つクラスは混合重みを0として、用いない。
3. 作成したモデルを学習データによって状態遷移確率を再推定する。本研究ではHTKを用いて推定を行った。
4. 再推定されたモデルを用い、1に戻って繰り返す。

このようにして状態遷移確率を従来のHMM法と同様にして求め、出力確率分布はBNを用いて求めてモデルを作成する。

7 HMM/BN モデルを用いた音素認識実験

次に、実際に HMM/BN モデルを用いて実験を行った結果を示す。実験条件は、3 章の音素認識実験と同じ条件で特徴抽出を行い 16 次の MFCC、ΔMFCC、ΔΔMFCC、48 次の音響パラメータを用いた。調音パラメータは位置データ、16 次を用いて分析を行う。

HMM は 3 章の実験と同様に状態数 3 の対角共分散 HMM、音素 HMM29 種類として実験を行った。

7.1 各話者モデルでの実験

まず、本実験に用いたデータは日本人男性話者 3 名が発話したデータは以下に示す通りで、それぞれの話者に関してモデルを作成し、音素認識実験を行った。認識には HTK を用いて行った。

Table 6: 各話者の実験データ

話者	発話データ	学習データ	テストデータ
TO	363 文章	313 文章	50 文章
TM	375 文章	325 文章	50 文章
MH	354 文章	304 文章	50 文章

16 次の調音パラメータに対し、主成分分析を行った。その結果、各話者データともに 4 番目までの累積寄与率は MH に関して 89.8%、TM に関して 91.2%、TO に関して 88.2% となった。これにより 4 つの成分までで、十分に元のデータを表現できると判断できるので、本実験では、第 4 主成分までを用いることとした。求めた変換行列を用いて 16 次の調音パラメータを 4 次のパラメータに変換し、コードブックサイズは 4 と 8 の 2 種類についてベクトル量子化を行った。それらのデータを用いてそれぞれ HMM/BN モデルを作成し音素認識実験を行った。　ここでコードブックサイズを 4 とした場合の HMM/BN モデルと音声のみのデータを用いて作成した従来の HMM による音素認識実験を行った結果を示す。結果は正解精度（Accuracy）で表している。

Table 7: コードブックサイズ 4 の各話者の HMM/BN モデルによる音素認識結果

	HMM (MIX2)	HMM (MIX3)	HMM/BN(VQ4)
MH	81.90%	83.52%	84.12% (2.47)
TM	82.17%	84.98%	83.58% (2.54)
TO	71.64%	73.80%	75.09% (2.67)

コードブックサイズがHMM/BNモデルにおけるガウス分布の混合数と考えることができるが、実際にはすべての音素の各状態にラベルがすべて現れるというこはなく、また数が少ない場合は混合重み0として無いものとして扱っているので、実際にコードブックサイズがモデルの混合数になるということは無い。そこで、HMM/BNモデルのすべての混合数を本実験では音素HMM29個、状態数3としてモデルを作成しているので実際にはモデルは87個の状態を持っていることになるので、HMM/BNモデルに現れるガウス分布の数を87で割り、一状態あたりの混合数を求めて表のHMM/BNモデルの認識値の後の括弧内に示す。この値を元に比較する音声のみを使って作成されたHMMの混合数を決める。ここではHMM/BNモデルの混合数が2から3の間の値となったので、HMMの混合数が2,3の場合について比較を行った。

話者TMに関してはHMM/BNモデルのほうがHMMに比べて悪くなったが話者MH、TOに関しては約1%向上した。

次にコードブックサイズ8とした場合の認識結果を示す。

Table 8: コードブックサイズ8の各話者のHMM/BNモデルによる音素認識結果

	HMM (MIX3)	HMM (MIX4)	HMM/BN (VQ8)
MH	83.52%	84.60%	84.76%(3.70)
TM	84.98%	84.55%	84.76%(4.08)
TO	73.80%	74.88%	75.53%(3.89)

HMM/BNモデルの1状態当りの混合数が3から4であるので、従来のHMMの混合数3と4のモデルと比較を行った。話者TMに関しては混合数4のHMMと比較すると向上したが、混合数3のHMMと比較すると認識率が下がっている。話者MH、TOに関してはコードブックサイズ4の場合と同様に認識率の向上が見られた。

同様にコードブックサイズを16、32、64としてモデルを作成した場合の結果を以下に示す。

Table 9: コードブックサイズ16の各話者のHMM/BNモデルによる音素認識結果

	HMM (MIX4)	HMM (MIX5)	HMM/BN (VQ16)
MH	84.60%	85.20%	85.58%(4.73)
TM	84.55%	84.93%	85.31%(4.74)
TO	74.88%	75.31%	77.57%(4.95)

コードブックサイズを増やしていくとHMMとHMM/BNの差が大きくなった。音声のみを用いて作成されたHMMよりも少ない混合数で高い認識率を得ることができた。

次に、調音パラメータの分析について別の手法を考える。

Table 10: コードブックサイズ 32 の各話者の HMM/BN モデルによる音素認識結果

	HMM (MIX8)	HMM/BN(VQ32)
MH	85.90%	86.76% (7.57)
TM	85.58%	87.14% (7.95)
TO	76.23%	77.90% (7.79)

Table 11: コードブックサイズ 64 の各話者の HMM/BN モデルによる音素認識結果

	HMM (MIX12)	HMM/BN(VQ64)
MH	86.55%	86.44% (12.3)
TM	87.03%	87.47% (12.1)
TO	78.55%	78.95% (12.0)

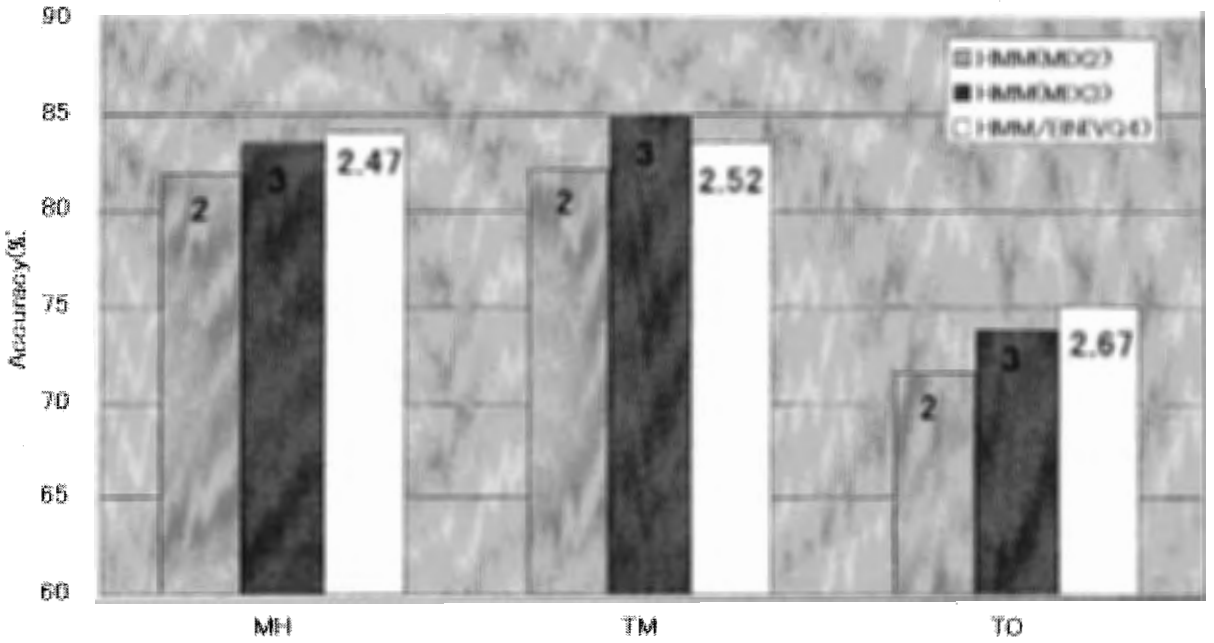


Fig. 24: コードブックサイズ 4 の各話者の HMM/BN モデルによる音素認識結果

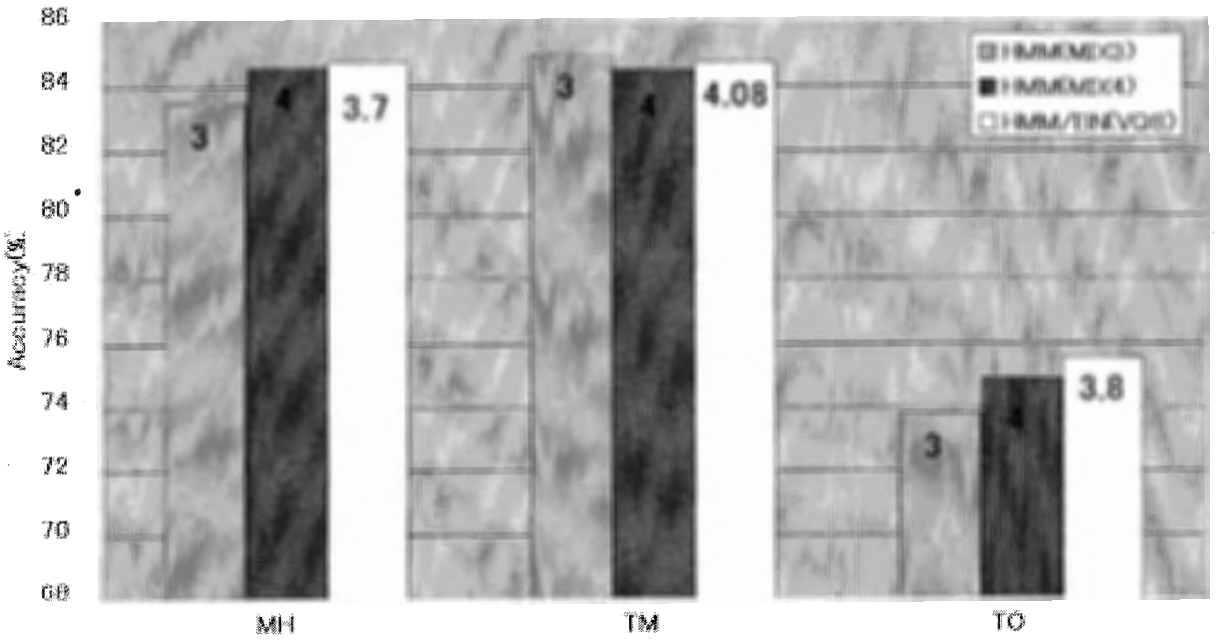


Fig. 25: コードブックサイズ 8 の各話者の HMM/BN モデルによる音素認識結果

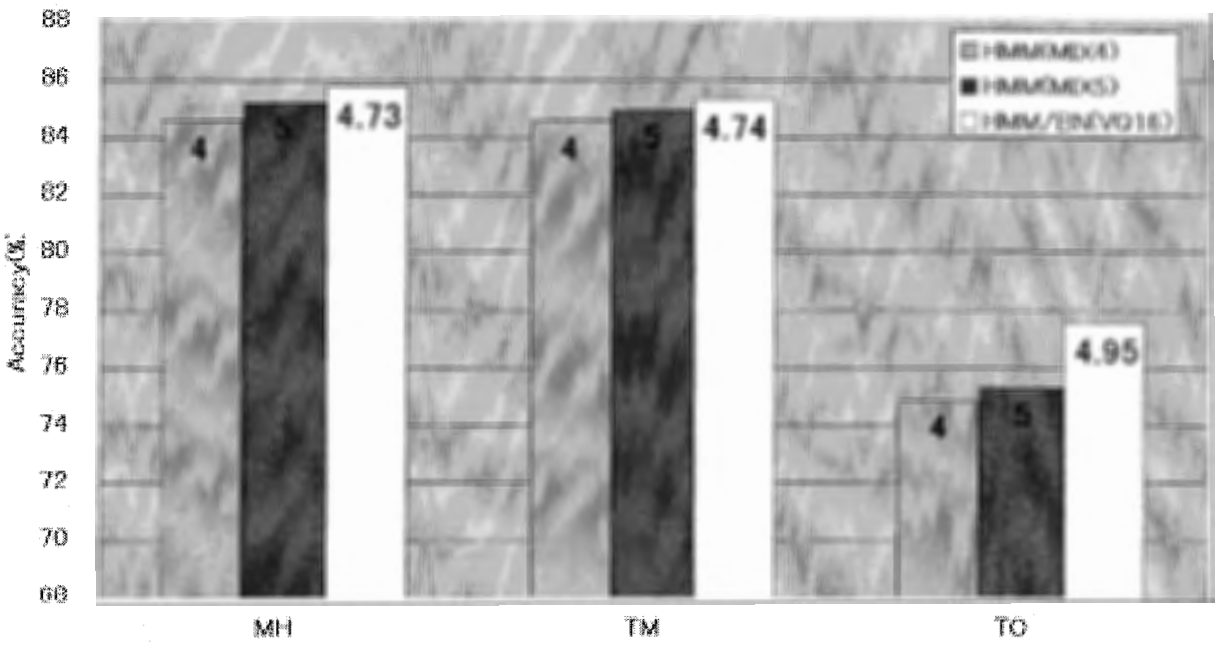


Fig. 26: コードブックサイズ 16 の各話者の HMM/BN モデルによる音素認識結果

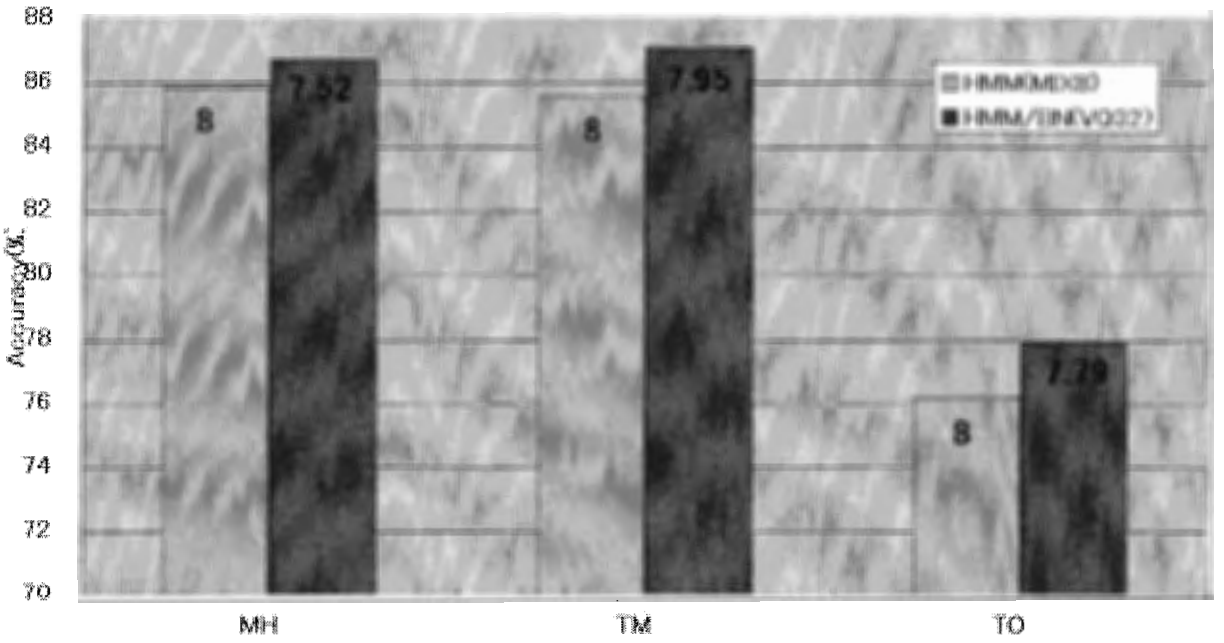


Fig. 27: コードブックサイズ 32 の各話者の HMM/BN モデルによる音素認識結果

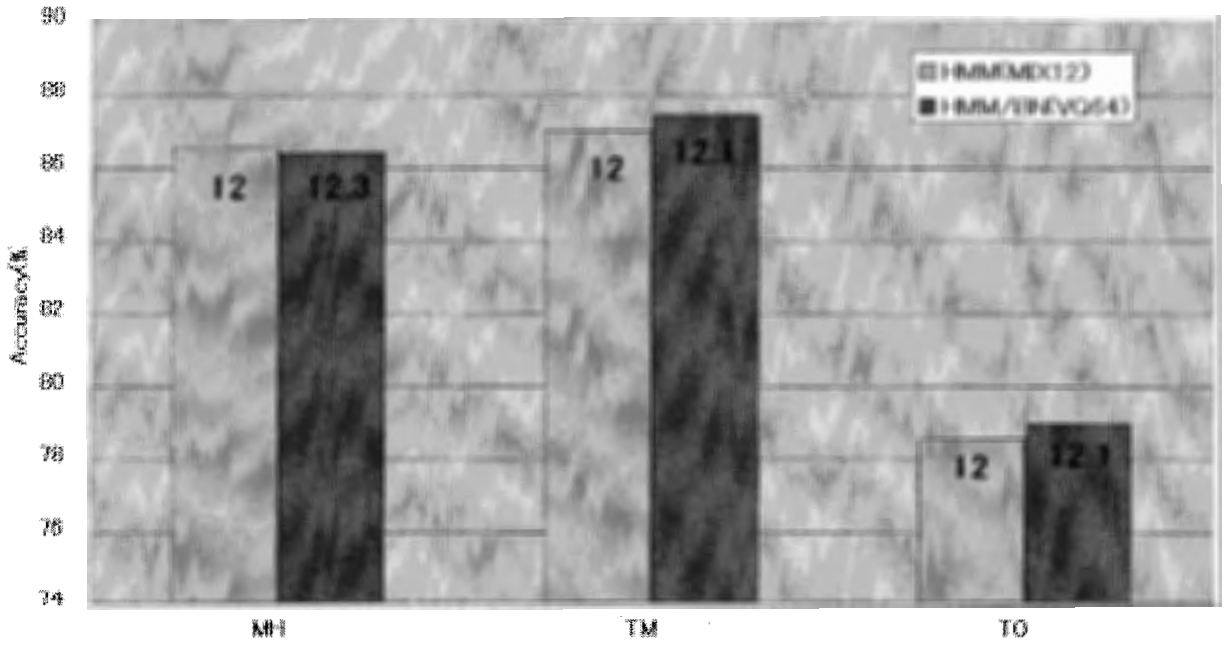
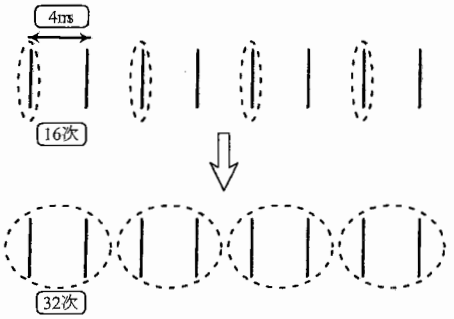


Fig. 28: コードブックサイズ 64 の各話者の HMM/BN モデルによる音素認識結果

本研究で用いている調音パラメータは4ms間隔で収録されていて、先の実験では音声の分析間隔にあわせて8ms間隔でデータを用いている。ここで図29のように用いていない4ms後のパラメータも合わせて32次のパラメータとして用いることを考える。16次の場合と同様にPCAを行ってベクトル量子化を行いモデルに取り入れる。



- Fig. 29: 16次と32次の調音パラメータ

各話者についてPCAを行った結果、第8次主成分までの累積寄与率が話者MHで97.6%、話者TMで98.3%, 話者TOで97.6%であったので、第8主成分までを用いて32次のパラメータを8次に変換して、コードブックサイズを8としてベクトル量子化を行いモデルを作成した。コードブックサイズ8のモデルで16次のパラメータを用いた場合と32次のパラメータを用いた場合の結果を以下に示す。括弧内は1状態当りの平均混合数を表す。

Table 12: 16次と32次の調音パラメータを用いた場合の比較（コードブックサイズ8の各話者のHMM/BNモデル）

	16次	32次
MH	84.76% (3.70)	85.20% (3.78)
TM	84.76% (4.08)	84.12% (4.16)
TO	75.53% (3.89)	75.57% (3.75)

話者MH、TOに関して32次のパラメータを用いることで認識率の向上が見られた。調音運動の時間的な変化も考慮にいれることで認識率が向上したと考えられる。

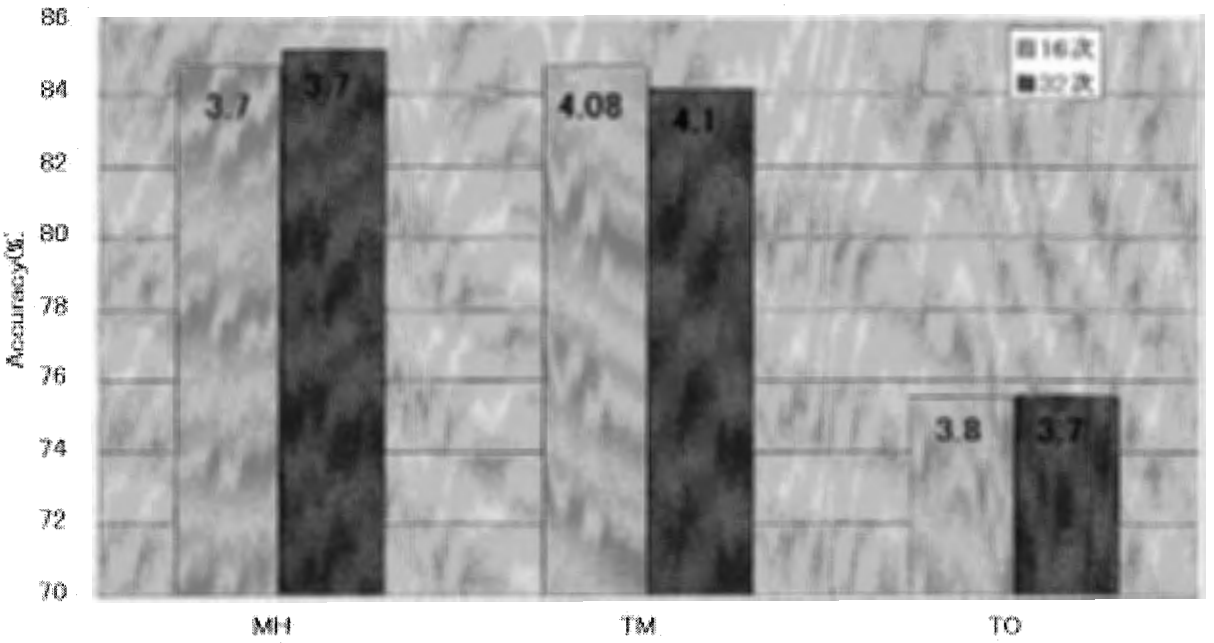


Fig. 30: 16 次と 32 次の調音パラメータを用いた場合の比較（コードブックサイズ 8 の各話者の HMM/BN モデル）

7.2 全話者モデルでの実験

次に、3人分の学習データを用いて作成したモデルでの認識実験を行う。実験データを以下に示す。各話者のモデルを作成した時と同様にPCAを行った結果、第4主成分までで、

Table 13: 全話者の実験データ

話者	学習データ	テストデータ
TO	$300*3=900$ 文章	50 文章
TM		50 文章
MH		50 文章

90.4%の累積寄与率を得ることができたので、16次の調音パラメータを4次へ変換し、ベクトル量子化を行い、モデルを作成した。

コードブックサイズを8としたときのモデルの認識結果を示す。

Table 14: コードブックサイズ8の全話者HMM/BNモデルでの音素認識結果

	HMM (MIX5)	HMM (MIX6)	HMM/BN (VQ8,MIX 5.60)
MH	81.42%	82.33%	81.85%
TM	81.42%	81.96%	81.36%
TO	72.50%	73.74%	74.28%

HMM/BNモデルの1状態当りの平均混合数が5.60となったので、混合数5のHMMと混合数6のHMMとの比較を行った。話者MH、TMに関しては認識率は上がらなかったが、TOに関しては認識率が向上した。

コードブックサイズ8の場合、話者毎にクラスが別れる傾向があった。話者MHのデータは2つのクラスに、話者TMのデータも別の2つのクラスに別れ、話者TOはまた別の4つのクラスに別れて、合計8個のクラスになるという結果になった。このようなクラスタリングが行われたため、TOの情報を多く含むようなモデルになってしまったため、TOのみ認識率が上がったと考えられる。

そこで、12のコードブックをつくる際、各話者のデータを4つのクラスに分かれるようにコードブックを作成してベクトル量子化を行い、モデルを作成した。そのモデルで認識実験を行った結果を示す。

同じ混合数7のモデルでHMMよりも調音パラメータを用いたHMM/BNモデルの方が2から3%認識率が上がった。コードブックサイズ8のHMM/BNと比べると話者TOに関しては若干下がったが、MH、TMに関しては認識率が3%上昇した。12のクラスに分けたため、話者ごとの特性を表現できるモデルになったと考えられ、従来の音声のみを用

Table 15: コードブックサイズ 12 の全話者 HMM/BN モデルでの音素認識結果

	HMM (MIX7)	HMM/BN (VQ12,MIX 7.1)
MH	81.47%	84.33%
TM	81.42%	84.39%
TO	71.37%	73.53%

いた HMM よりも認識率が良くなった。以下にコードブックサイズを 24、48 としてモデルを作成し、実験を行った結果を示す。

Table 16: コードブックサイズ 24 の全話者 HMM/BN モデルでの音素認識結果

	HMM (MIX9)	HMM (MIX10)	HMM/BN (VQ24,MIX 9.71)
MH	81.47%	81.96%	84.17%
TM	82.39%	83.09%	85.62%
TO	72.23%	73.91%	75.63%

Table 17: コードブックサイズ 48 の全話者 HMM/BN モデルでの音素認識結果

	HMM (MIX16)	HMM/BN (VQ48,MIX 15.1)
MH	82.06%	85.63%
TM	82.98%	86.44%
TO	73.10%	76.50%

コードブックサイズを増やすことでさらに HMM/BN モデルのほうが認識率が向上した。これは音声と調音位置の依存関係を HMM/BN によって表現出来たのと、各話者の特徴を表現出来るモデルであったためであると考えられる。

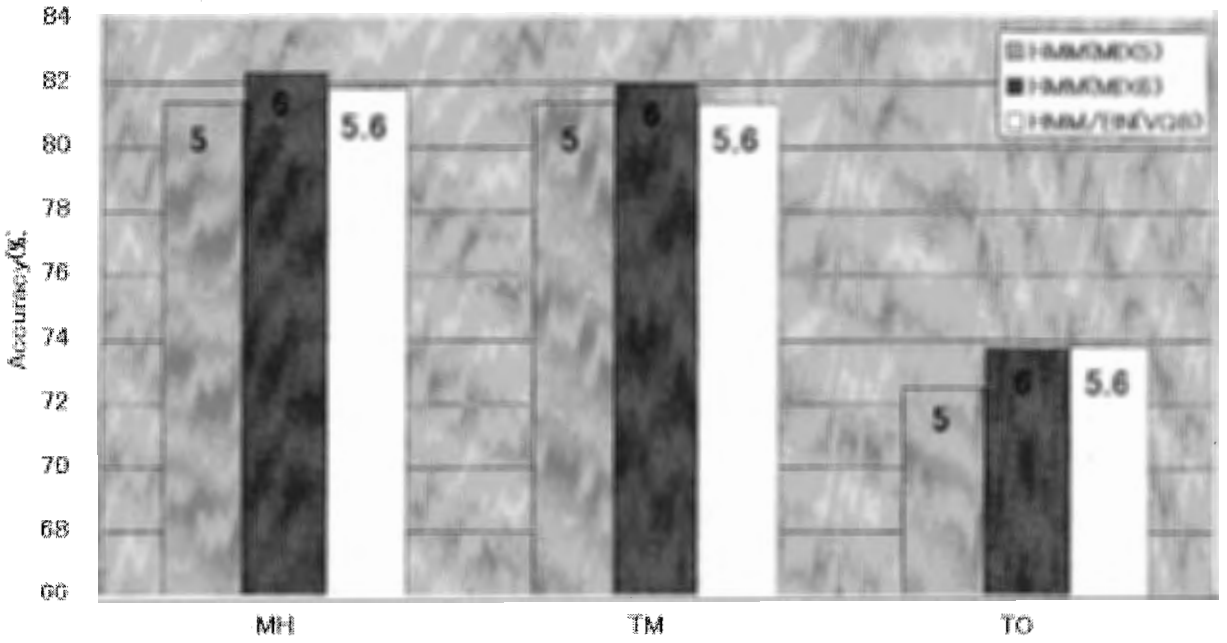


Fig. 31: コードブックサイズ 8 の全話者 HMM/BN モデルでの音素認識結果

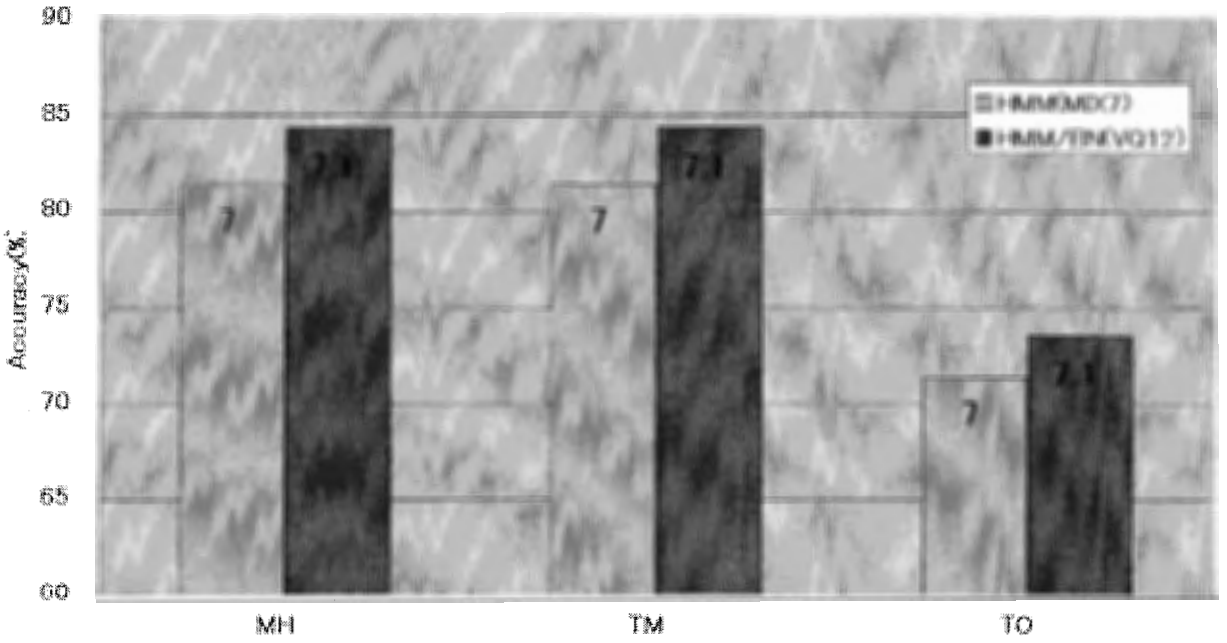


Fig. 32: コードブックサイズ 12 の全話者 HMM/BN モデルでの音素認識結果

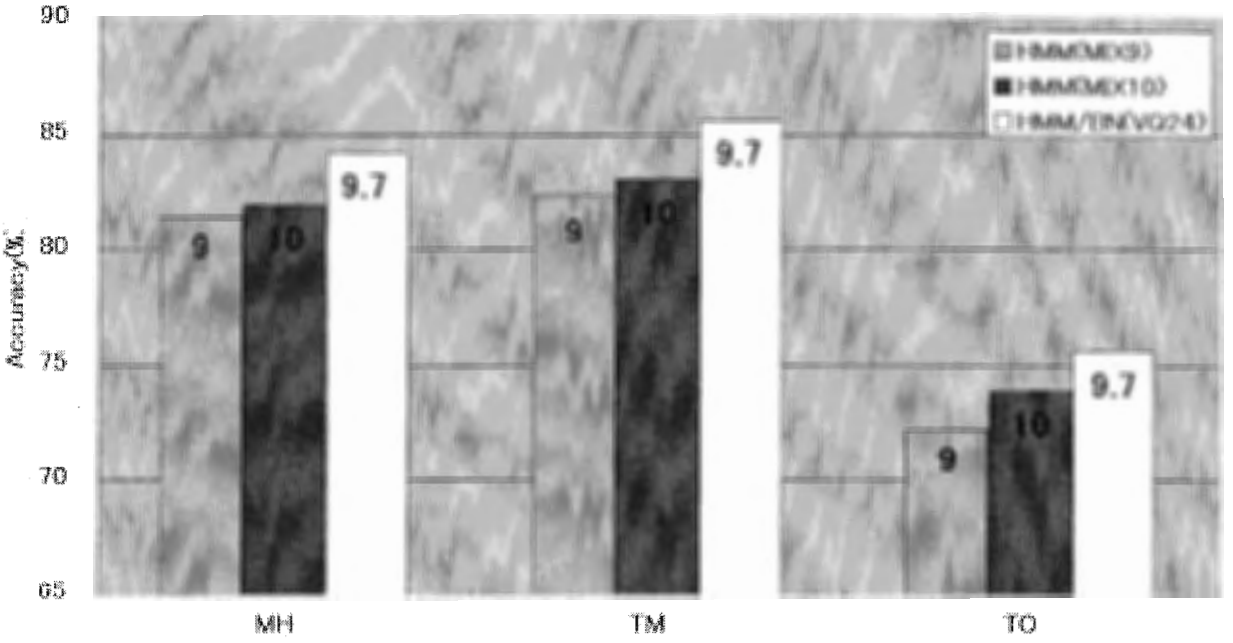


Fig. 33: コードブックサイズ 24 の全話者 HMM/BN モデルでの音素認識結果

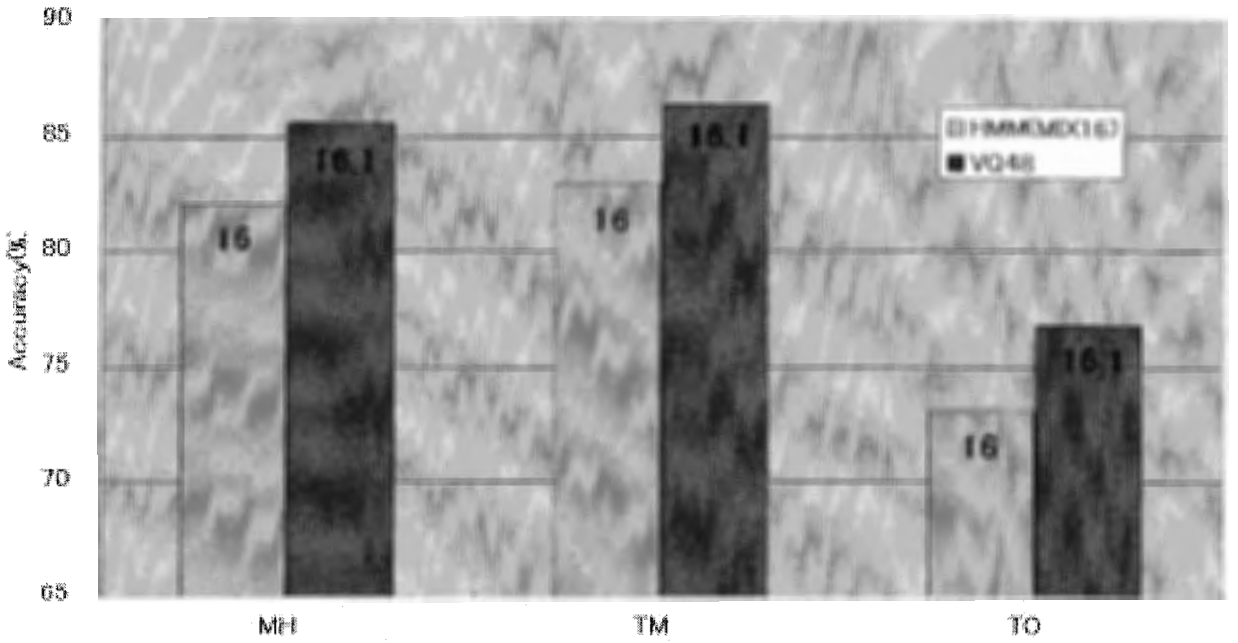


Fig. 34: コードブックサイズ 48 の全話者 HMM/BN モデルでの音素認識結果

次に、各話者モデルと同様に、調音パラメータを 32 次にして、コードブックサイズ 8、12 でベクトル量子化してモデルを作成した結果を示す。32 次のパラメータを PCA したときの第 8 主成分までの累積寄与率は 98.4% であった。

Table 18: 16 次と 32 次の調音パラメータの比較（コードブックサイズ 8 の全話者 HMM/BN モデル）

	16 次 (MIX 5.6)	32 次 (MIX 5.59)
MH	81.85%	81.52%
TM	81.36%	81.47%
TO	73.85%	74.28%

話者 MH に関しては認識率は上がらなかったが、TM、TO に関しては認識率が向上した。

Table 19: 16 次と 32 次の調音パラメータの比較（コードブックサイズ 12 の全話者 HMM/BN モデル）

	16 次 (MIX 7.1)	32 次 (MIX 7.15)
MH	84.33%	84.49%
TM	84.39%	84.01%
TO	73.53%	74.88%

パラメータを増やすことにより、話者 MH、TO に関して認識率が向上した。4ms 後の調音パラメータも取り入れることで、その相関も考慮にいれることができ、制約条件として働いたと考えることができる。

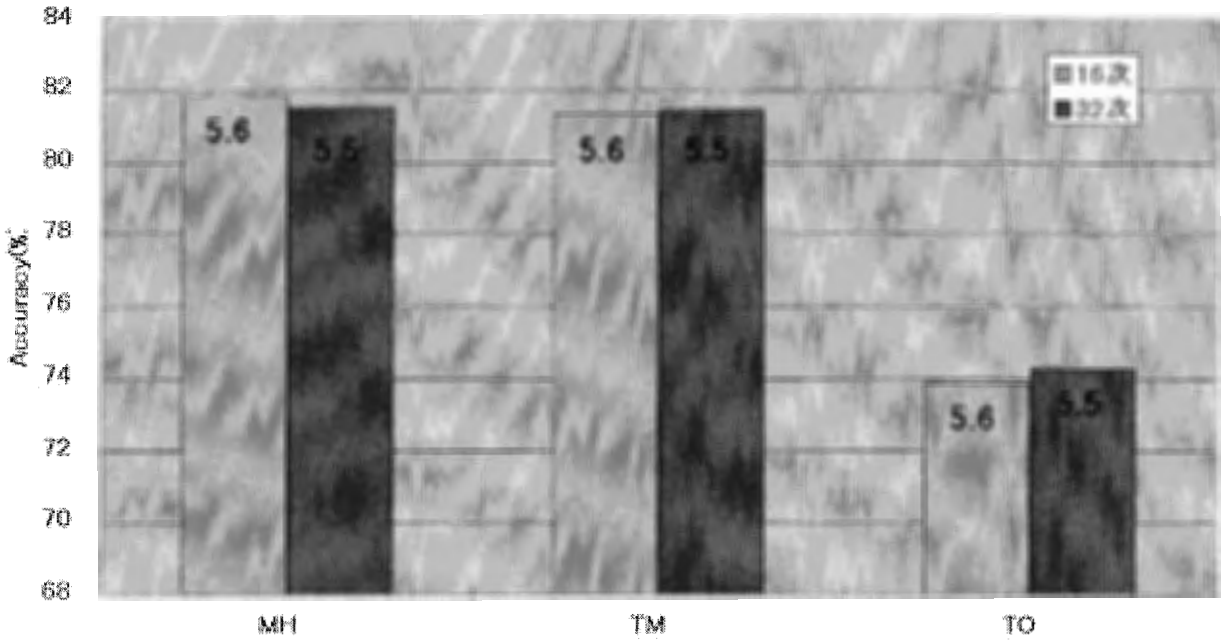


Fig. 35: 16 次と 32 次の調音パラメータの比較（コードブックサイズ 8 の全話者 HMM/BN モデル）

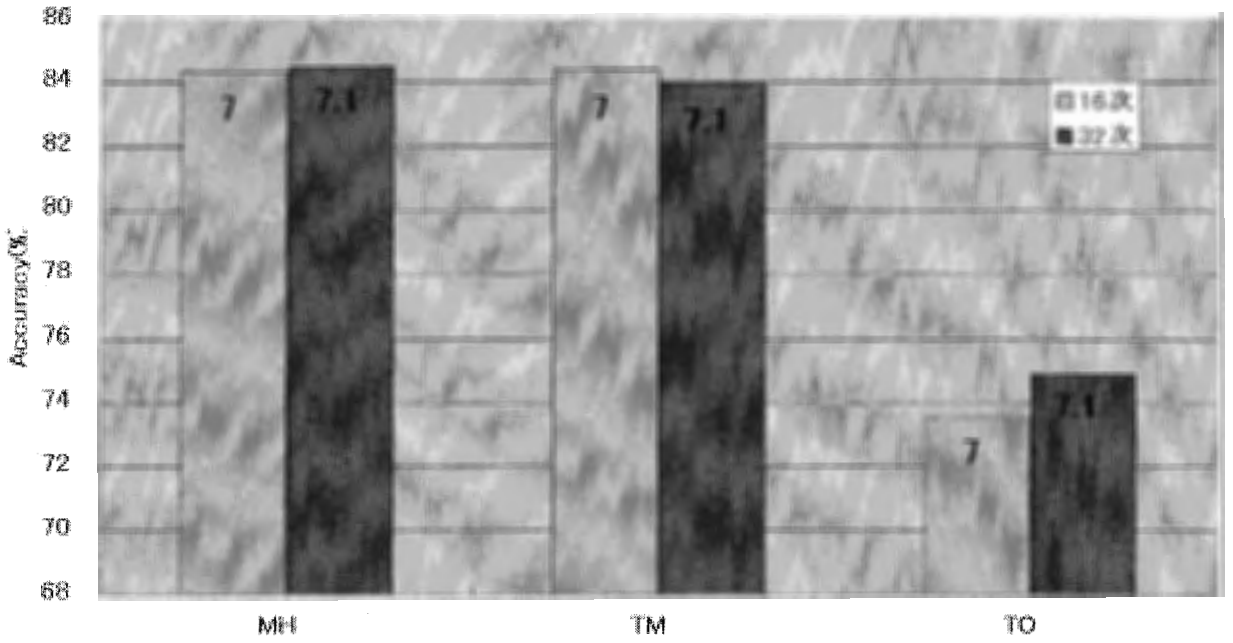


Fig. 36: 16 次と 32 次の調音パラメータの比較（コードブックサイズ 12 の全話者 HMM/BN モデル）

8 考察

8.1 各話者ごとのHMM/BNモデルについて

HMM/BNモデルに調音パラメータを取り入れてモデルを作成した。HMM/BNにおいては、調音パラメータをベクトル量子化して得られたラベル毎に、対応する音響パラメータをクラスタリングを行い、それぞれの出力確率分布をガウス分布と見なおすことにより、音声と調音位置の依存関係を表現することが出来た。同程度の混合数を持つHMM（音声のみ）と比較したところ、HMM/BNモデルのほうが認識率が高くなっていることが確認できた。これは、調音器官の調音位置に関する情報を用いたことにより音声データのみにある曖昧さをある程度を抑えたと考えられる。これは音声には含まれていない有効な情報を取り入れるこ とが出来たと考えられる。

話者TMにおいてVQ4とVQ8（混合数3、4）ではHMMの方が高い認識率になっている。それは、少ない混合数でデータの分布を表現する場合、表現の正確さはデータの分布に左右されるので、不安定な要素があると考えられる。混合数を増やすことによって表現の正確さがデータの分布に依存しなくなるので、安定な認識率が得られる。そのため、不特定話者の場合、混合数の少ないモデルより、混合数の多いモデルで議論をなされている。また、混合数を単純に増やすことで認識率が良くなるというわけではなく、多くしても逆に誤差を生じる可能性があるので、データによって適切な混合数があると考えられる。

各話者モデルにおいてHMMとHMM/BNの3人の結果の平均を求めたグラフを図37に示す。混合数3の場合、HMMの方が高いが混合数4以降でHMM/BNの方が高くなっている。この場合、混合数を4から12の間で調音パラメータを取り入れたHMM/BNモデルが最適であると考えられる。

調音位置のパラメータは16次となっている。時間の情報を取り入れるため、時間上（4ms間隔）で隣接する16次をあわせて32次にして用いた。32次を16次の場合に比較すると、若干、認識率が向上することが分かったが、その違いはそれほどなかった。これは、4ms後の情報を取り入れることで時間的な変化も考慮に入れることができるという利点と空間的に不連続で対応関係がはっきりしないものを一つのパラメータとして扱ってしまっているという欠点があり、その両方が打ち消しあってあまり認識率に効果が現れなかったと考えられる。

8.2 全話者HMM/BNモデルについて

全話者のデータにより作成したHMM/BNモデルを用いて実験を行なったところ、コードブックサイズが8の場合、音声のみのHMMと比較すると、TOだけ認識率が高くなったが、他のMH、TMに関しては認識率が低くなった。この場合のコードブックはMHに関して2つ、TMに関して2つ、TOに関して4つのクラスに分かれるようなセントロイドを持つコードブックであるので、作成されたHMM/BNモデルではTOの情報がより多くなり、TOだけ認識率が高くなったと考えられる。この問題を解決するため、MH、TMのセントロイドを2分割して、合計12個のセントロイドを持つようなコードブックを全話者の

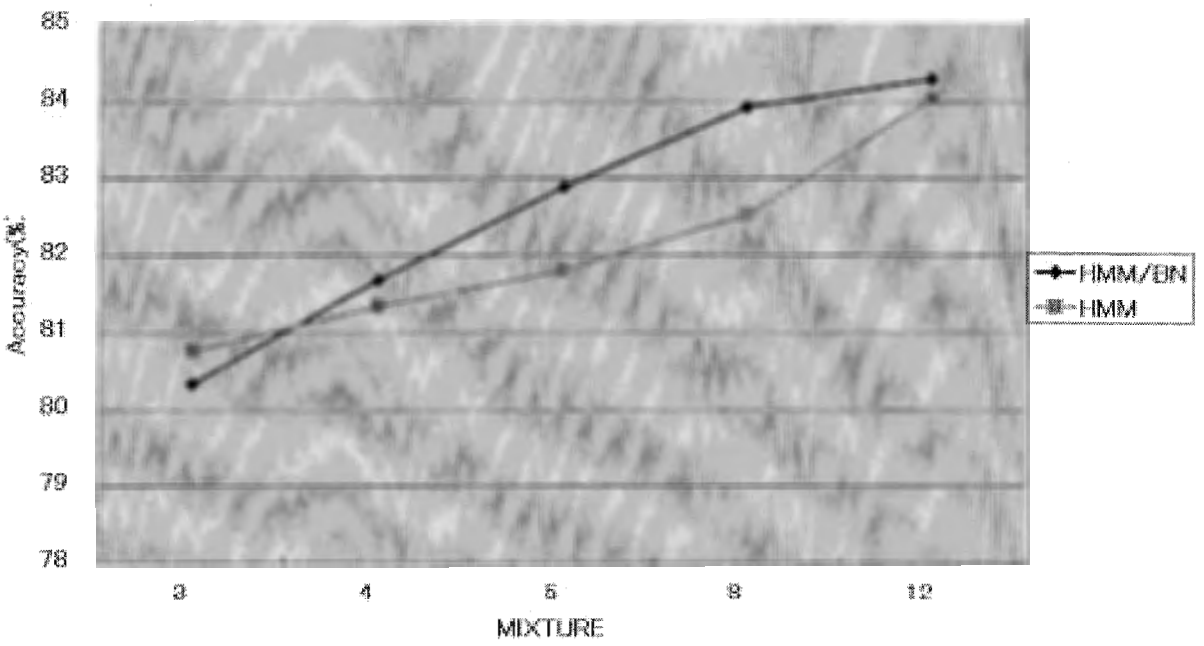


Fig. 37: 各話者モデルの HMM と HMM/BN の比較

データから作成して話者ごとにそれぞれ4つのクラスを持つようなコードブックを作成した。このコードブックを用いてモデルを作成し、各話者で音素認識実験を行った結果、同程度の混合数を持つHMMよりもHMM/BNの方が高くなった。これは全話者のデータで一つのモデルを作成する時、各話者の情報を均等に持つことが重要であることを示唆している。その後、24、48とコードブックサイズを増やした場合もHMM/BNモデルのほうが高い認識率を得ることができた。これは、音声と調音位置の依存関係をHMM/BNによって表現できたのと、各話者の特徴を表現出来るモデルであったためであると考えられる。

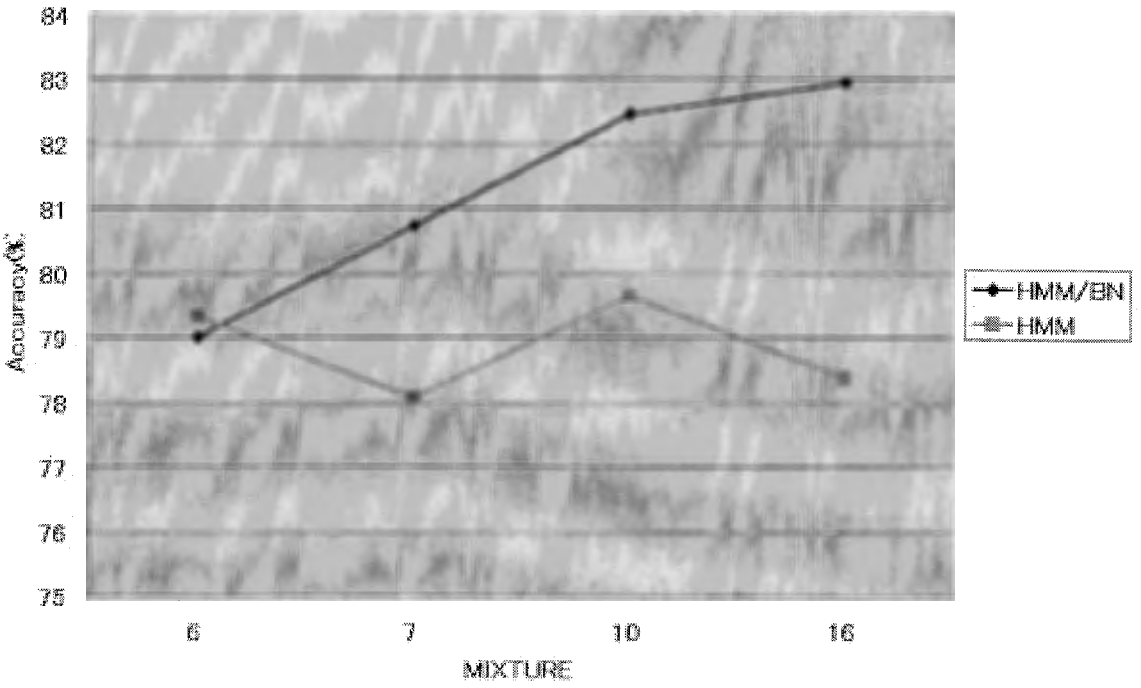


Fig. 38: 全話者モデルのHMMとHMM/BNの比較

8.3 初期統合モデルとの比較

音響パラメータと調音パラメータを一つのパラメータに統合し、モデルを作成し、音素認識実験を3章で行った。その結果と音声のみを用いたHMM、調音パラメータを取り入れたHMM/BNモデルを用いて音素認識実験を行った結果を示す。

初期統合モデルとHMM/BNモデルを比較すると混合数8のときHMM/BNの認識率が初期統合の結果に最も近くなった。初期統合の場合、音声に同期した調音運動のデータのみ用いることができるがHMM/BNモデルの場合、認識データの調音運動が観測できない、音声のみの場合でも調音運動を知っている場合に近い認識率を得ることができた。し

かし、初期統合の結果と差があるのは、HMM/BN モデルでは調音運動にある有用な情報がまだ十分に利用されていないことを示唆している。この結果より、調音運動の速度（Δ）成分などを用いることによってさらに調音運動の情報を有効に利用するモデルの作成が期待できると考えられる。

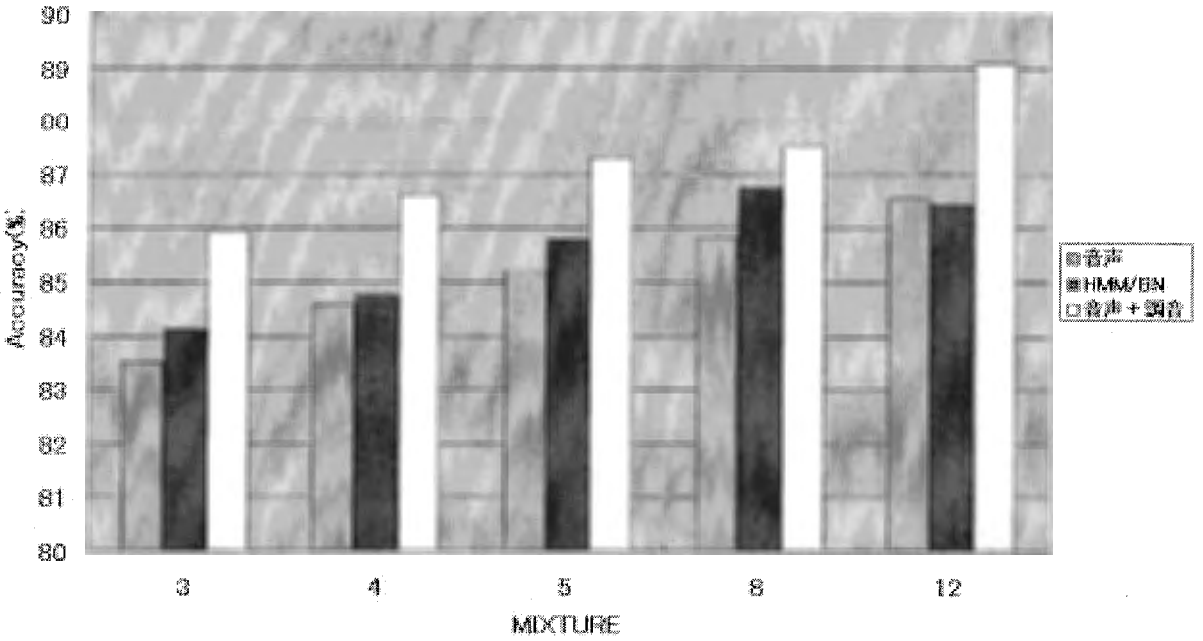


Fig. 39: 話者 MH に関する音素認識結果の比較

9 コードブックサイズと認識率

コードブックサイズを大きくすることで認識率が上がることがわかったが、今回の実験ではサイズが各話者モデルで 64、全話者モデルで 48 までの実験を行った。図 42に各話者の HMM/BN モデルを用いた認識結果ついて、図 43に全話者 HMM/BN モデルを用いた認識結果について、横軸コードブックサイズ、縦軸認識率（Accuracy）としたグラフを示す。コードブックサイズと認識率の関係を示す。各話者モデルでは話者 MH が 64 で若干認識率が下がったが、他の 2 人に関してはまだ認識率が上がっていて、3 話者の平均（mean）を見ても上がっているのが分かる。全話者モデルでも、コードブックサイズを大きくするほど認識率が上がっている。コードブックサイズを大きくすると認識率が向上している。これは調音位置の特徴をより細かく分けることで、より詳しく調音位置の情報を表現できていると言うことが分かる。そこでコードブックをさらに増やし、モデルを作成して実験を行い、適切なコードブックサイズについて検討を行う必要がある。

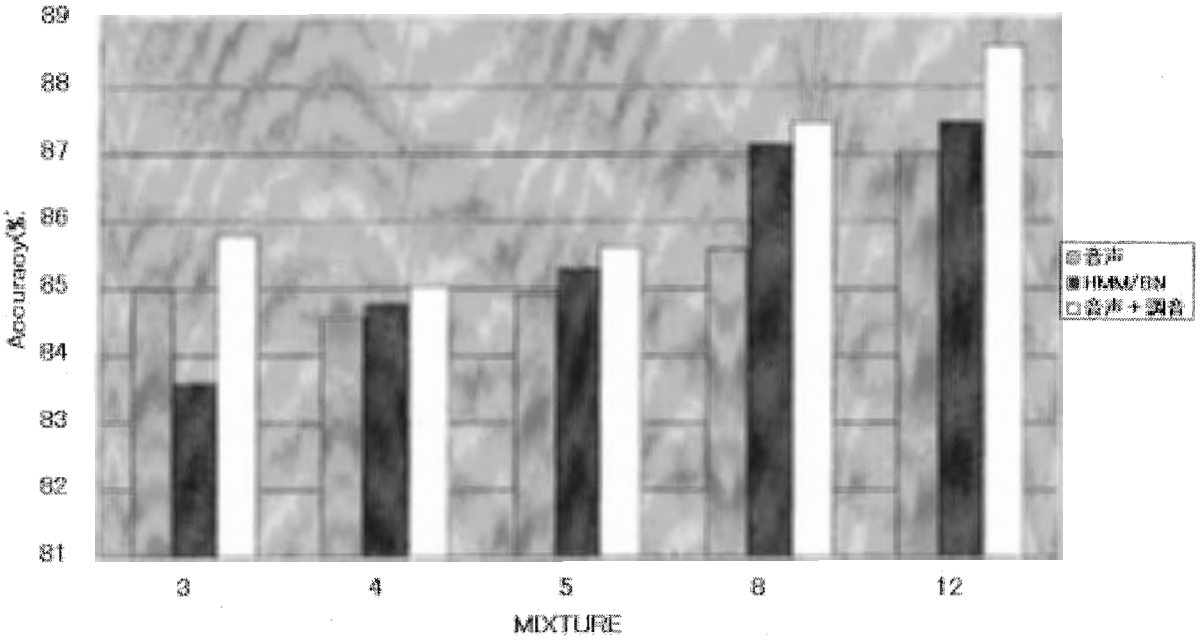


Fig. 40: 話者 TM に関する音素認識結果の比較

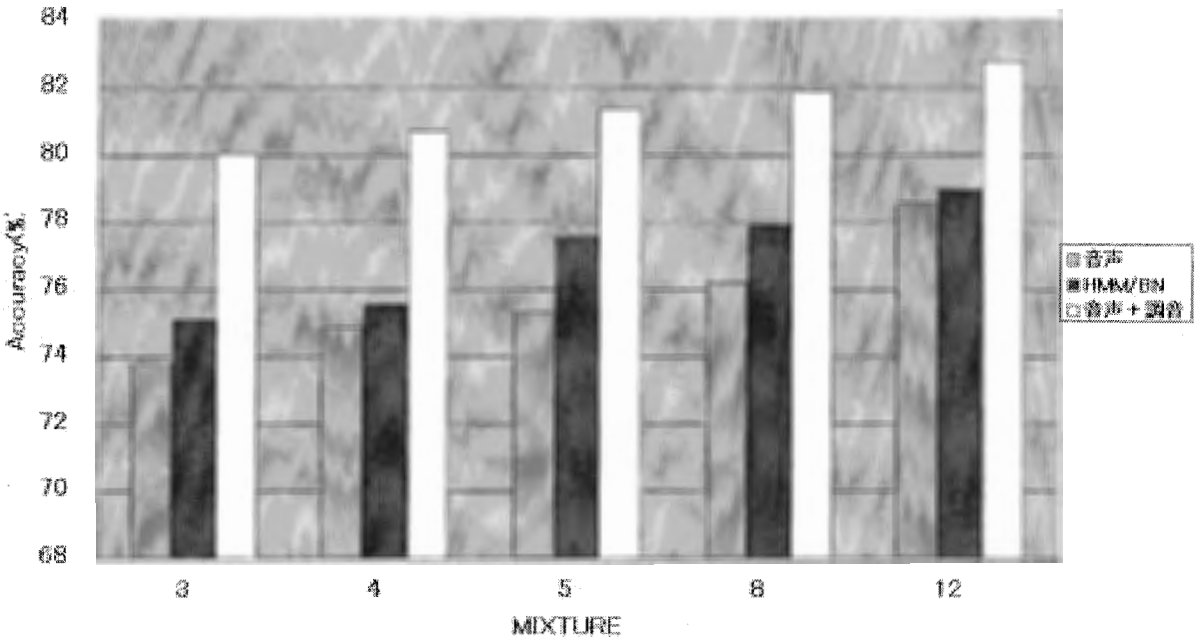


Fig. 41: 話者 TO に関する音素認識結果の比較

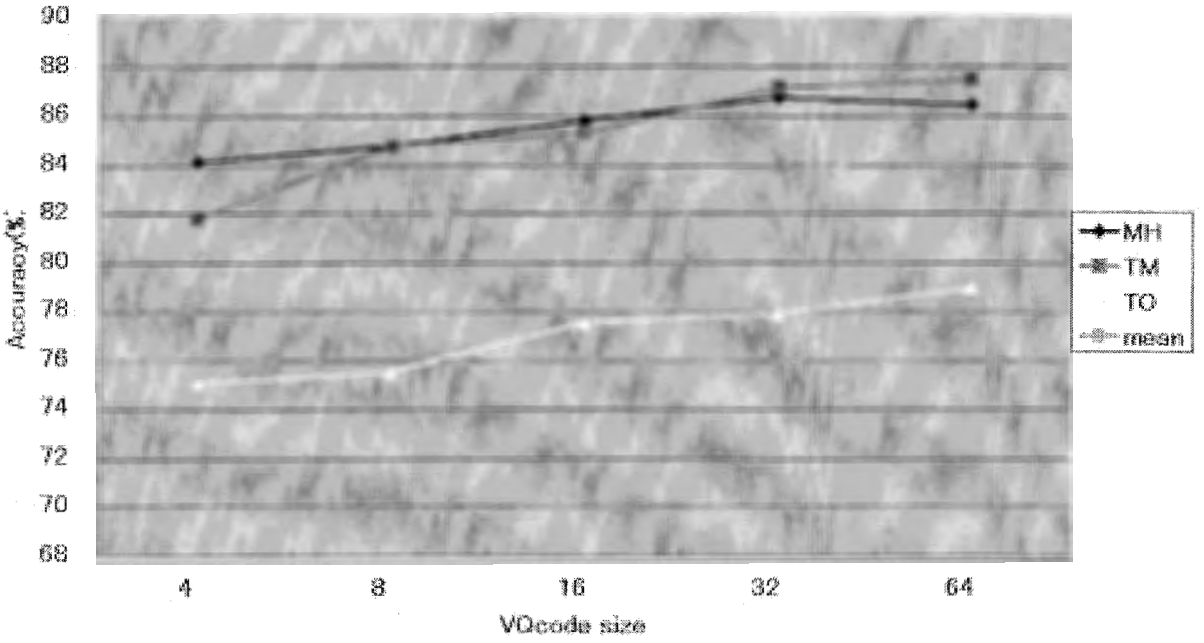


Fig. 42: 各話者モデルにおけるコードブックサイズと認識率の変化

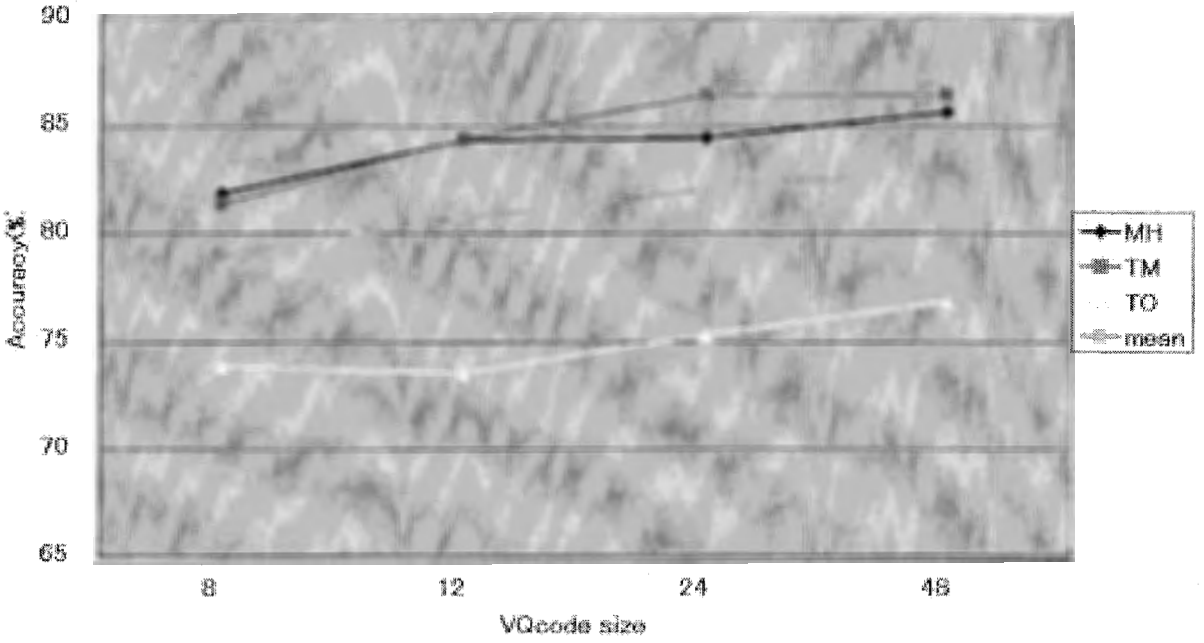


Fig. 43: 全話者モデルにおけるコードブックサイズと認識率の変化

10 今後の課題

- 調音パラメータを速度データ（Δ:1 次差分）のパラメータを用いてモデルを作成し実験を行う。例えば、/asa/と/aka/の最後の/a/は同じ/a/でも前についている子音によって特徴が変化する。音声生成で考えると、/s/と/k/の調音は大きく異なり、調音パラメータの変化も HMM/BN モデルに取り入れることでその違いを表現できると考えられる。
- 各話者、全話者 HMM/BN モデルともに、コードブックサイズをさらに大きくしてモデルを作成し、実験を行い適切なコードブックサイズについて検討を行う必要があると考えられる。
- 今後の課題として、今回の結果から調音運動を音声認識に取り入れることが有効であることが分かった。しかし、今回用いた HMM/BN モデルはグラフィカルモデル（依存関係をグラフを用いて統計モデルで表現したもの）であり、正確な意味で「音声生成メカニズムを取り入れた音声認識」であるとはこの段階では言えない。今後は音声生成モデルを用いた音声認識に発展させる必要があると考えられる。

参考文献

[1] 古井貞熙著 “音声情報処理”, 森北出版

[2] 古井貞熙著 “ディジタル音声処理”, 東海大学出版会、1985

[3] 服部四郎 “音声学”、岩波全書 131、岩波書店、1992

[4] 佐藤大和 “男女声の声質情報を決める要素”、研究実用化報告（NTT）、24,5,pp.977-993,1975

[5] 音声認識セミナー資料、情報処理学会 音声言語情報処理研究会 1998 8 月 23 日-27 日

[6] 電子情報通信学会 “確率モデルによる音声認識”, コロナ社 1988

[7] Yuqing Gao,Raimo Bakis,Jing Huang,Bing Xiang “Multistage coarticulation model combining articulatory,formant and cepstral features” ,SPS5 Seeon 2000 ,Page241

[8] John Hodgen , Patrick Valdez “Bridging the gap between speech production and speech recognition”

[9] Li DENG,Leo J.LEE & Paul FIEGUTH “A functional articulatory dynamic model for speech production”

[10] Konstantin MARKOV,Satoshi NAKAMURA “A Hybrid HMM/BN Acoustic Model for Autimatic Speech Recognition” IEICE TRANS.INF.& SYST.,VOL.E86-D,NO.1 JANUARY 2003

[11] Geoffrey Zweig and Stuard Russell "Probabilistic modeling with Bayesian Networks for automatic speech recognition" ICSLP,pp.3010-3013,1988

[12] K.Daoudi,D.Fohr and C.Antoine,"A new approach for multi-band speech recognition based on probabilistic graphical models" ICSLP,vol.1,pp.329-332,2000