論　文

フレーム間相関を利用した音韻HMMによる音声認識

正員　高橋　敏[†]　　正員　松岡　達雄[†]
正員　南　泰浩[†]　　正員　鹿野　清宏[†]

Speech Recognition Using Phoneme HMMs Constrained by Frame Correlations

Satoshi TAKAHASHI[†], Tatsuo MATSUOKA[†], Yasuhiro MINAMI[†]
and Kiyohiro SHIKANO[†], *Members*

あらまし　現在のHMMの問題点の一つに，出力確率分布が各状態内で常に一定で，音韻特徴量の遷移情報がモデルの仕組みの中に反映されていないという点が挙げられる．しかも，特徴ベクトルの遷移に制約がないので，互いに出力確率が高い特徴ベクトル間の遷移は，学習データ中に観測されなかった遷移でも高い出力確率が与えられている．本論文では，特徴ベクトルの2フレーム間の相関を用いて遷移を制約し，不特定話者用HMMの広がった特徴量分布を，入力話者に適した範囲に制約するBigram制約HMMを提案する．Bigram制約HMMの出力確率は，前時刻の特徴ベクトルの条件付き確率で表現されるので，出力確率分布は各時刻で動的に変化する．また，分布を制約することにより，異なる音韻間の特徴量分布の重なりが減少し，認識率を向上することができる．我々は既に，離散型不特定話者用HMMをもとに，VQコードのBigramを用いて遷移を制約する離散型Bigram制約HMMを提案し，従来のHMMよりも性能が良いことを示した．本論文では，更に高い認識性能を得るために，この手法を半連続型Bigram制約HMM，連続型Bigram制約HMMに拡張した．連続音声中の音韻認識によって評価した結果，入力話者の音声のフレーム間相関情報を用いた場合，半連続型Bigram制約HMMによって平均音韻認識率を65.4%から74.8%に，連続型Bigram制約HMMによって64.8%から74.5%に改善することができた．また，多数話者から抽出した一般的なフレーム間相関情報を用いた場合，連続型Bigram制約HMMによって64.8%から67.5%に改善することができた．

キーワード　不特定話者音声認識，HMM，条件付き出力確率，スペクトル遷移情報

1. まえがき

不特定話者，大語い，連続音声認識という目標に向け，現在，さまざまな研究機関で精力的に研究が進められている．Hidden Markov Model (HMM)を用いた音声認識手法は，この困難なタスクを実現可能なレベルにまで引き上げたキーテクノロジーの一つであると言える．統計モデルであるHMMは，特に，以下の点において利点を有する．

不特定話者：数多くの話者のさまざまな音声データを大量に扱うことができ，しかも，学習アルゴリズムによってモデルの最適化が可能である．

大語い：小さな音声単位（例えば音韻）のモデルを単に連結することによって，あらゆる音韻系列（単語や文章）のモデルを作成できる．

連続音声：調音の揺らぎ，音韻環境の違いによる調音結合の影響などのスペクトルの変動を吸収できる．

しかし，HMMにもいくつかの問題点が指摘されている．本論文で取り上げるHMMの問題点は以下の二つである．

（1）　現在のHMMは，各状態内で出力確率分布が一定で，音韻特徴量の局所的な遷移情報がモデルの仕組みの中に反映されていない．例えば離散型HMMを考えた場合，VQコード系列が“1-1-2-2”でも“1-2-1-2”でも，それらの出力確率が同じ状態内で計算されれば同じ値になってしまう．また，現在のHMMは，特徴量の遷移に関しては何ら制約がないので，互いに高い出力確率をもつ特徴ベクトル間の遷移は，学習データ中に観測されなかった不適切な遷移でも高い確率が与えられてしまう．これでは，音声の重要な要素である特徴量の時系列情報を失っていることになる．この点を改善するために，Δケプストラム[(1)]などパラメー

† NTTヒューマンインタフェース研究所，武蔵野市
NTT Human Interface Laboratories, Musashino-shi, 180 Japan

タに時間情報をもたせる方法，線形予測型 HMM[2]，予測型ニューラルネットと HMM を組み合わせる方法[3] などが提案されている．

（2） 現在の不特定話者用 HMM は，多数話者からなる広がった特徴量分布を用いて認識しているため，異なる音韻間の分布の重なりが多く誤認識を生じやすい．一般に，不特定話者用 HMM は，数多くの話者の音声を使って学習し，どのような入力話者にも対応できるようにする．よって，不特定話者用 HMM の音韻特徴量分布は，複数の話者，あるいは話者クラスタの分布から合成されていると考えられる．学習話者の数を増加させた場合，未学習の特徴量空間が減少する一方で，音韻特徴量分布は次第に広がり，ある話者のある音韻の特徴量分布が，他の話者の異なる音韻の特徴量分布と重なることがしばしば起こる．この分布の重なりが，不特定話者音声認識の性能を劣化させている原因の一つであると考えられる．システムが認識しようとする一人の入力話者に対して必要なのは，広がった多数話者の分布全体ではなく，この入力話者に適した一部分の分布のみである．従って，認識性能を向上させるためには，不特定話者用 HMM が入力話者に適するように自動的に分布を制約し，音韻間の分布の重なりを減少させることが必要である．

本論文では，以上二つの問題を解決するために考案した Bigram 制約 HMM（Bigram-constrained HMM）について述べる[4],[5]．本モデルは，不特定話者用 HMM の多数話者からなる広がった音韻特徴量分布を，2フレーム間の音韻特徴量の相関情報を利用して，入力話者に適した分布に制約するモデルである．不特定話者音声認識においても，ゆう度を計算するフレームの1時刻前の特徴量は観測済みであるから，そこからの遷移を考慮することにより，分布を制約することができる．よって，本モデルは，1時刻前のフレームの音韻特徴量によって，分布が各時刻で動的に変化するのが特徴である．

我々は既に，離散型不特定話者用 HMM をもとに，フレーム間の相関情報として VQ コードの Bigram 情報を用いた離散型 Bigram 制約 HMM を提案し，従来の音韻 HMM よりも性能が良いことを示した[6],[7]．HMM は離散分布型のほかに，連続分布型，半連続分布型があり，これらは離散分布型よりも性能が良いという結果が数多く報告されている[8]~[11]．よって，フレーム間相関を用いた場合でも更に高い性能が期待できる．本論文では Bigram 制約 HMM を半連続型 HMM，連続型 HMM に拡張し，これらのモデルが，従来の HMM や離散型 Bigram 制約 HMM よりも良い性能を有することを示す．

2. フレーム間相関を利用した HMM

Bigram 制約 HMM は音韻特徴量のフレーム間の相関情報を利用し，不特定話者用 HMM の出力確率分布（音韻特徴量分布）を入力話者に適するように制約するモデルである．このモデルの基本的な考え方は以下のようである．ある時刻 t で得られた音韻特徴量に対して確率を計算する場合，時刻 $t-1$ の音韻特徴量は既に観測済みであり，そこから時刻 t に向かって遷移し得る特徴量の範囲はある程度限定できると思われる．よって，特徴量の局所的な遷移情報を不特定話者用 HMM の音韻特徴量分布に反映すれば，広がった分布を入力話者に適するように制約することができる．分布が制約されれば音韻間の分布の重なりが減少し，認識率の向上が期待できる．しかも，Bigram 制約 HMM の出力確率は，時刻 $t-1$ の音韻特徴量の条件付き確率となるので，同じ状態内であっても出力確率分布が各時刻で動的に変化するのが特徴である．

この考えに基づいて離散型分布，半連続型分布，連続型分布，それぞれのタイプの Bigram 制約 HMM を定式化することができる．2フレームの音韻特徴量の相関情報として，離散型，半連続型では VQ コードの Bigram を，連続型では特徴ベクトルの連続分布の条件付き確率を用いる．連続型の場合，VQ コードの Bigram は使用しないが，離散型，半連続型にならい連続型 Bigram 制約 HMM と呼ぶことにする．モデル作成に共通する大まかな処理の流れを以下に，分布制約の概念図を図1に示す．

（1） 多数の学習用話者の音声データを用いて不特定話者用 HMM を作成する（図1（a））．

（2） （1）とは独立に，学習用音声を用いて VQ コードの Bigram（または連続分布の条件付き確率分布）を計算する（図1（b））．

（3） 上記二つのモデルの確率分布を合成し，1時刻前の音韻特徴量を条件とする条件付き出力確率分布を得る（図1（c））．Bigram 制約 HMM を用いた音声認識では，この制約された分布を用いて音韻を認識する．

Wellekens[13] や Brown[14] は，条件付き出力確率分布を最ゆう推定の枠組みで，学習データから直接求める定式化を行っている．しかし，出力確率分布を条件

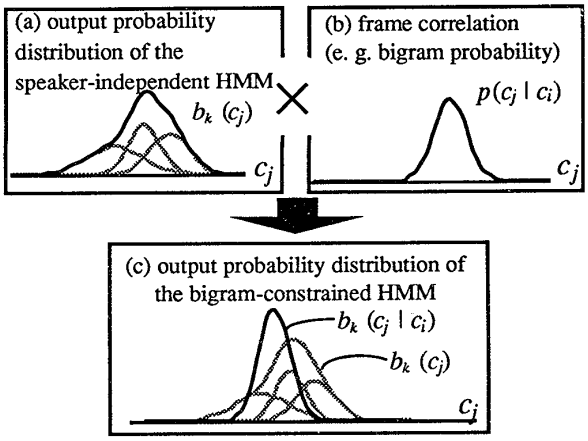


図 1　Bigram 制約 HMM の出力確率分布
Fig. 1　Output probability distribution of the bigram-constrained HMM.

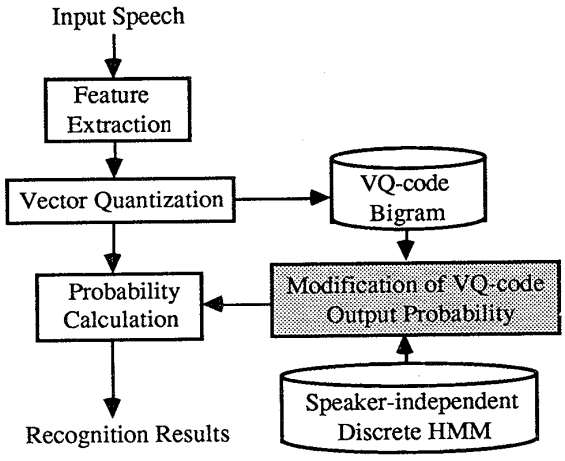


図 2　離散型 Bigram 制約 HMM による認識のブロック図
Fig. 2　Block diagram for recognition using the discrete type bigram-constrained HMM.

付きにすることで推定すべきパラメータ数が膨大になり，実際には，学習データ量の不足から適切なパラメータを求めることは困難であると思われる．ここで提案する方法は，フレーム間相関情報のみを HMM の状態間で「結び」にしたものとみなすこともでき，学習データ量不足の問題を軽減している．

2.1　離散型 Bigram 制約 HMM[5],[6]

離散型 Bigram 制約 HMM では，フレーム間相関情報として VQ コードの Bigram を用いる．手順を以下に示し，ブロック図を図 2 に示す．

（1）　多数の学習話者の音声データを用いて，ユニバーサル VQ コードブックを作成し，これを用いて離散型不特定話者用音韻 HMM を学習する．

（2）　学習用音声をユニバーサル VQ コードブックを用いてベクトル量子化し，Bigram 確率を計算する．

（3）　認識時は，まず，入力音声をベクトル量子化し，得られた VQ コード系列に従い各時刻 t で時刻 $t-1$ の VQ コードとの Bigram 確率を参照する．これを，不特定話者用 HMM の出力シンボル確率に重みづけるようにして，時刻 $t-1$ の VQ コードで条件づけられた出力シンボル確率を求める．

$$b_k(c_j|c_i)=\frac{p(c_j|c_i)b_k(c_j)}{\sum_{m=1}^{M}p(c_m|c_i)b_k(c_m)} \qquad (1)$$

ここで $p(c_j|c_i)$ は，時刻 $t-1$ で i 番目の VQ コード c_i が出現した場合に，時刻 t で VQ コード c_j が出現する Bigram 確率である．$b_k(c_j)$ は，時刻 t で不特定話者用 HMM が VQ コード c_j を出力する確率である．$b_k(c_j|c_i)$ は，Bigram 制約 HMM において，時刻 $t-1$ で VQ コードが c_i であった場合に，時刻 t で VQ コード c_j を出力する確率である．k は状態を，M はコードブックサイズを示す．式(1)の分母は，各状態で，すべての VQ コードに対する確率値の和が 1 になるように正規化する項である．

2.2　半連続型 Bigram 制約 HMM

半連続型 HMM（Semi-continuous HMM または Tied-mixture HMM）[9]~[12] は，連続型と離散型の両方の性質を併せもったモデルと言える．連続型 HMM の観点からは，混合連続分布型 HMM の確率密度関数の平均値，分散を，すべてのモデル，すべての状態で結びとしたモデルである．各モデルのそれぞれの状態における分布形状は，混合分布に対する重み係数を変えることで表現される．一方，離散型 HMM の観点からは，出力シンボル確率を連続分布にしたモデルとみなすことができる．半連続型 HMM において各分布の平均値を推定することは，VQ コードブックの設計を行っていることに等しく，モデルの学習中に VQ コードブックの設計と HMM パラメータの推定を同時に行っている．従って，それらを独立に行っている離散型 HMM の場合よりも，学習データに対してモデルのゆう度が最大になるようにパラメータを決定できる利点がある．

離散型 Bigram 制約 HMM では，VQ コードの Bigram 情報を用いて出力シンボル確率を直接変更したが，半連続型 Bigram 制約 HMM では，各分布の平均値，分散を固定したまま，重み係数のみを Bigram で入力話者に適するように変更する．以下に手順を示し，図 3 にブロック図を示す．

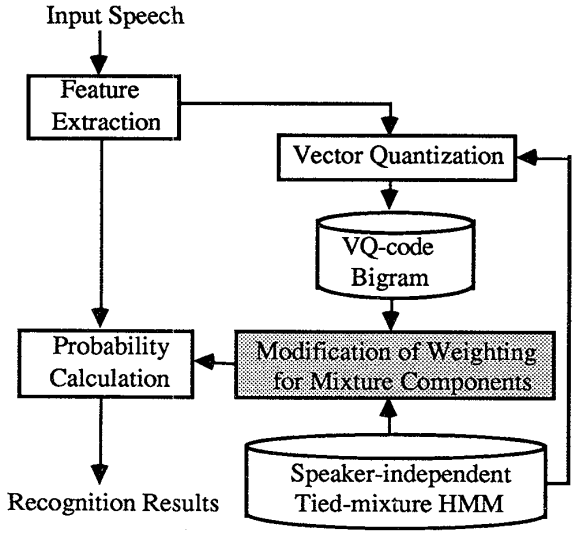


図 3 半連続型 Bigram 制約 HMM による認識のブロック図

Fig. 3 Block diagram for recognition using the tied-mixture type bigram-constrained HMM.

（1） 多数の学習話者の音声データを用いて，半連続型不特定話者用 HMM を学習する．

（2） 半連続型 HMM の各分布の平均値をセントロイドとするユニバーサル VQ コードブックを作成する．

（3） 学習音声をユニバーサル VQ コードブックを用いてベクトル量子化し，Bigram 確率を計算する．

（4） 認識時は入力音声をベクトル量子化し，Bigram 確率を用いて，半連続型 HMM の分布の重み係数を変更する．

$$\hat{\lambda}_{kj}=\frac{p(c_j|c_i)\lambda_{kj}}{\sum_{m=1}^{M}p(c_m|c_i)\lambda_{km}} \tag{2}$$

ここで，λ_{kj} は変更前の，$\hat{\lambda}_{kj}$ は変更後の，j 番目の分布に対する重み係数である．k は状態を示す．c_i は i 番目の分布の平均値をセントロイドにもつ VQ コードである．$p(c_j|c_i)$ は，離散型の場合と同様に Bigram 確率を表す．M は分布の総数（＝コードブックサイズ）である．式(2)の分母は，各状態で，すべての混合分布に対する重み係数の和が 1 になるように正規化する項である．

（5） 特徴ベクトル y に対する確率 $b_k(y)$ を VQ コードの遷移情報によって変更された重み係数を用いて，M 個の連続分布の和によって求め，これを用いて音韻を認識する．

$$b_k(y)=\sum_{m=1}^{M}\hat{\lambda}_{km}b_{km}(y) \tag{3}$$

b_{km} は，状態 k の m 番目の確率密度分布を表す．

2.3 連続型 Bigram 制約 HMM

連続型では，前時刻の特徴ベクトルを条件とする連続分布（正規分布）の条件付き確率分布と，不特定話者用 HMM の確率分布（正規分布）の積によって得られる分布を用いて音韻を認識する．ここでは，不特定話者用 HMM を混合正規分布で，条件付き確率を単一正規分布で近似する．

はじめに，学習データから条件付き確率を計算する[6],[7]．今，ベクトル y が前フレーム y_1，後フレーム y_2 の二つの n 次元の特徴ベクトルからなるとする．

$$y=(y_1, y_2) \tag{4}$$

y は，平均値ベクトル $\mu=(\mu_1, \mu_2)$，共分散行列 Λ の多次元正規分布からのサンプルと仮定する．共分散行列と逆共分散行列をそれぞれ，

$$\Lambda=\begin{bmatrix}\Lambda_{11} & \Lambda_{12}\\ \Lambda_{21} & \Lambda_{22}\end{bmatrix},\quad V=\begin{bmatrix}V_{11} & V_{12}\\ V_{21} & V_{22}\end{bmatrix} \tag{5}$$

とする．但し，サブマトリクスは対角成分のみを考慮する．すなわち，

$$\Lambda_{11}=\begin{bmatrix}\varphi_1^{(1)} & & 0\\ & \ddots & \\ 0 & & \varphi_n^{(1)}\end{bmatrix},\quad \Lambda_{22}=\begin{bmatrix}\varphi_1^{(2)} & & 0\\ & \ddots & \\ 0 & & \varphi_n^{(2)}\end{bmatrix},$$

$$\Lambda_{12}=\Lambda_{21}=\begin{bmatrix}\rho_1 & & 0\\ & \ddots & \\ 0 & & \rho_n\end{bmatrix} \tag{6}$$

ここで，$\varphi_i^{(1)}$ と $\varphi_i^{(2)}$ は，それぞれ前後フレームにおける i 次の特徴パラメータの分散値である．また，ρ_i は $\varphi_i^{(1)}$ と $\varphi_i^{(2)}$ の共分散である．つまり，パラメータのフレーム間の相関は，各次元内でのみ考慮している．$z=y-\mu$，$z_1=y_1-\mu_1$，$z_2=y_2-\mu_2$ とし，$p(y_1, y_2)$ が単一正規分布であると仮定すると，

$$p(y_1, y_2)=\frac{|V|^{1/2}}{(2\pi)^{n/2}}\exp\left(-\frac{1}{2}z^tVz\right) \tag{7}$$

但し，$z^tVz=z_1{}^tV_{11}z_1+2z_1{}^tV_{12}z_2+z_2{}^tV_{22}z_2$ である．ここで右肩の t は転置を表す．条件付き確率の定義から，

$$\begin{aligned}p(y_2|y_1)&=\frac{p(y_1, y_2)}{\int p(y_1, y_2)dy_2}\\ &=\frac{\exp\left(-\frac{1}{2}z^tVz\right)}{\int\exp\left(-\frac{1}{2}z^tVz\right)dy_2}\\ &=\frac{\exp\left\{-\frac{1}{2}(z_2{}^tV_{22}z_2+2z_1{}^tV_{12}z_2)\right\}}{\int\exp\left\{-\frac{1}{2}(z_2{}^tV_{22}z_2+2z_1{}^tV_{12}z_2)\right\}dy_2}\end{aligned} \tag{8}$$

式(8)の分母の指数部を2次形式にした後，y_2 で積分する．

$$z_2{}^tV_{22}z_2+2z_1{}^tV_{12}z_2$$
$$=(z_2+V_{22}{}^{-1}V_{21}z_1)^tV_{22}(z_2+V_{22}{}^{-1}V_{21}z_1)$$
$$-z_1{}^tV_{12}V_{22}{}^{-1}V_{21}z_1 \quad (9)$$

$$\int\exp\left\{-\frac{1}{2}(z_2{}^tV_{22}z_2+2z_1{}^tV_{12}z_2)\right\}dy_2$$
$$=\left(\frac{(2\pi)^{n/2}}{|V_{22}|^{1/2}}\right)\exp\left(\frac{1}{2}z_1{}^tV_{12}V_{22}{}^{-1}V_{21}z_1\right) \quad (10)$$

従って，式(8)は以下のようになる．

$$p(y_2|y_1)$$
$$=\frac{|V_{22}|^{1/2}}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}(z_2{}^tV_{22}z_2+2z_1{}^tV_{12}z_2\right.$$
$$\left.+z_1{}^tV_{12}V_{22}{}^{-1}V_{21}z_1)\right\}$$
$$=\frac{|V_{22}|^{1/2}}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}(z_2+V_{22}{}^{-1}V_{12}z_1)^tV_{22}(z_2\right.$$
$$\left.+V_{22}{}^{-1}V_{12}z_1)\right\} \quad (11)$$

ここで，

$$z_2+V_{22}{}^{-1}V_{12}z_1=y_2-(\mu_2-V_{22}{}^{-1}V_{12}(y_1-\mu_1)) \quad (12)$$

であるから条件付き確率分布は，平均値 $\delta^{(p)}=\mu_2-V_{22}{}^{-1}V_{12}(y_1-\mu_1)$，共分散行列 $\Lambda^{(p)}=V_{22}{}^{-1}$ の単一正規分布になることがわかる，また，

$$\begin{bmatrix}V_{11} & V_{12}\\ V_{21} & V_{22}\end{bmatrix}\begin{bmatrix}\Lambda_{11} & \Lambda_{12}\\ \Lambda_{21} & \Lambda_{22}\end{bmatrix}=I \quad (13)$$

より（I は単位行列），

$$\delta^{(p)}=\mu_2-V_{22}{}^{-1}V_{12}(y_1-\mu_1)$$
$$=\mu_2+\Lambda_{21}\Lambda_{11}{}^{-1}(y_1-\mu_1) \quad (14)$$
$$\Lambda^{(p)}=V_{22}{}^{-1}=\Lambda_{22}-\Lambda_{21}\Lambda_{11}{}^{-1}\Lambda_{12} \quad (15)$$

認識時は図4に示すように，各時刻で式(14)，式(15)に前フレームの特徴ベクトル y_1 を与えて条件付き確率分布を求め，これと HMM の混合正規分布との積を計算し，新たな HMM の条件付き出力確率分布 $b_k(y_2|y_1)$ を得る．

$$b_k(y_2|y_1)=\sum_{i=1}^{M}\hat{\lambda}_{ki}p(y_2|y_1)b_{ki}(y_2) \quad (16)$$

ここで，$b_{ki}(y_2)$ は現時刻の特徴パラメータ y_2 の i 番目の混合分布に対する出力確率である．$\hat{\lambda}_{ki}$ は i 番目の混合分布に対する新しい重み係数である．k は状態を，M は混合分布数を示す．

式(16)において，条件付き確率 $p(y_2|y_1)$ と HMM の状態 k の i 番目の混合正規分布 $b_{ki}(y_2)$ との積は以下のように計算できる．

$$p(y_2|y_1)b_{ki}(y_2)$$
$$=\frac{|V^{(p)}|^{1/2}}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}(y_2-\delta^{(p)})^tV^{(p)}(y_2-\delta^{(p)})\right\}$$
$$\cdot\frac{|V^{(s)ki}|^{1/2}}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}(y_2-\delta^{(s)ki})^tV^{(s)ki}(y_2\right.$$
$$\left.-\delta^{(s)ki})\right\} \quad (17)$$

ここで，$\delta^{(s)ki}$，$V^{(s)ki}$ はそれぞれ，HMM の状態 k の i 番目の混合分布に対する平均値ベクトルおよび逆共分散行列である．

先程と同様に指数部を2次形式にして展開すると，

$$=\frac{|V^{(p)}V^{(s)ki}|^{1/2}}{(2\pi)^n}\exp\left\{-\frac{1}{2}(y_2-\delta)^t\hat{\Lambda}^{-1}(y_2-\delta)\right.$$
$$\left.+\Psi\right\} \quad (18)$$

ここで，

$$\Psi=(\delta^{(p)t}V^{(p)}\delta^{(p)})+(\delta^{(s)ki^t}V^{(s)ki}\delta^{(s)ki})$$
$$-(V^{(p)}\delta^{(p)}+V^{(s)ki}\delta^{(s)ki})^t(V^{(p)}$$
$$+V^{(s)ki})^{-1}(V^{(p)}\delta^{(p)}+V^{(s)ki}\delta^{(s)ki}) \quad (19)$$

平均値ベクトル，

$$\hat{\delta}=(V^{(p)}+V^{(s)ki})^{-1}(V^{(p)}\delta^{(p)}+V^{(s)ki}\delta^{(s)ki})$$
$$=(\Lambda^{(p)-1}+\Lambda^{(s)ki-1})^{-1}(\Lambda^{(p)-1}\delta^{(p)}+\Lambda^{(s)ki-1}\delta^{(s)ki}) \quad (20)$$

共分散行列，

$$\hat{\Lambda}=(V^{(p)}+V^{(s)ki})^{-1} \quad (21)$$

である（これらの導出を付録に示す）．

また，式(16)における新しい重み係数 $\hat{\lambda}_{ki}$ は，不特定話者用 HMM の分布と条件付き確率分布から合成される分布の面積を考慮して決定する．

$$\hat{\lambda}_{ki}=\frac{c_{ki}}{\sum_{m=1}^{M}c_{km}} \quad (22)$$

$$c_{ki}=\int p(y_2|y_1)\lambda_{ki}b_{ki}(y_2)dy_2 \quad (23)$$

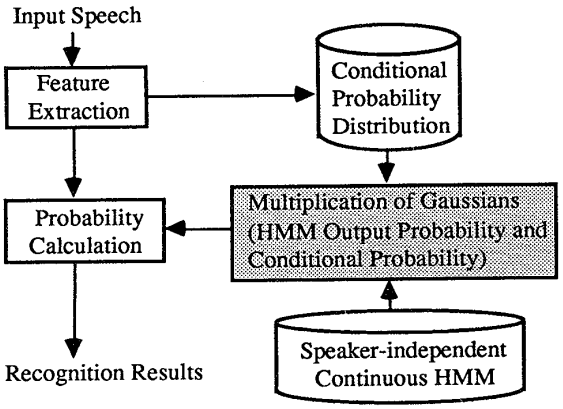


図 4　連続型 Bigram 制約 HMM による認識のブロック図

Fig. 4　Block diagram for recognition using the continuous type bigram-constrained HMM.

ここで，λ_{ki} は変更前の重み係数である．

最後に，Wellekens や Brown らの行った定式化と Bigram 制約 HMM との違いについて述べる．彼らは，式(11)および式(12)で表される条件付き確率分布を HMM の出力確率分布として直接，最ゆう推定した．従って，この分布の形を決定する平均値 $\delta_k^{(p)}(\mu_{k1}, \mu_{k2}$ を含む)，共分散 $\Lambda_k^{(p)}$ を HMM の各状態 k ごとに推定しなければならない．よって，パラメータの数は膨大になり，適切なパラメータを求めるのは困難であると考える．実際，この方式によって，認識性能が改善されたという報告はまだない．一方，Bigram 制約 HMM では，前後フレームの平均値ベクトル μ，共分散行列 Λ は状態に依存せずモデル内で共通とし，しかもこれらを学習データから直接計算する．つまり，隣接する 2 フレームの特徴ベクトルを学習データから取り出して，μ, Λ を計算する．Bigram 制約 HMM では更に，不特定話者用 HMM の分布との合成を行うことによって，より安定的に分布を推定している．

3. 評価実験

3.1 連結学習と音韻ラベルなし評価法

これまでに述べた離散型，半連続型，連続型 Bigram 制約 HMM の性能を，連続音声中の音韻認識によって比較した．ベースとなる不特定話者用モデルはすべて連結学習によって作成し，評価は音韻ラベルなし評価法[15]を用いた．よって，この実験では音韻ラベルを一切使用していない．音韻ラベルなし評価法は，発声テキストをもとに音韻モデルを連結した後，評価する箇所の音韻モデルをすべての音韻モデルに置き換え，それぞれの場合の文章全体のゆう度を計算し，最大ゆう度を与える音韻を認識結果とする方法である．

3.2 実験条件

使用したデータベースを表 1 にまとめる．連結学習に用いたデータは，音響学会の連続音声データベースの音韻バランス文 64 人分である．各話者 150 文ずつ発声している（合計 9,600 文）．評価に用いたデータは，音響学会の連続音声データベースの対話読上げ文 13 人分である．収録機関，発声内容が重ならないように 823 文を選択した．分析条件は，サンプリング周波数 12 kHz，ハミング窓長 32 ms，フレーム周期 8 ms である．連結学習時にはすべての発声テキストを覆うように 43 種類の音韻モデルを作成した．評価は，この中から長母音，2 重母音，拗音等を除き，出現頻度の多い 23 音韻に対して行った．特徴パラメータは，ケプストラム，1 次の線形回帰係数である Δ ケプストラム，Δ パワーを用いた．表 2 に，離散型におけるコードブックサイズ，半連続型における混合正規分布の総分布数，連続型における各状態の混合正規分布の数をパラメータごとに示す．正規分布はすべて対角共分散行列を用いた．HMM の構造はすべて，4 状態 3 ループである．

表 1 実験に使用したデータベース

モデル	用途	使用データベース	文章数
不特定話者用 HMM	学習	音響学会 連続音声データベース 音韻バランス文	9600文 (150文×64人)
	評価	音響学会 連続音声データベース 対話読み上げ文	823文 (計13人)
Bigram	入力話者用	対話読み上げ文 (HMMの評価用データとはテキスト独立)	各入力話者ごとに約50文 (3分間の音声)
	不特定話者用	HMMの学習用データ (音韻バランス文) より抜粋	2000文 (計40人)

表 2 音韻特徴量

音韻特徴量	離散型 (ｺｰﾄﾞﾌﾞｯｸｻｲｽﾞ)	半連続型 (総分布数)	連続型 (分布数／状態)
ケプストラム(16次)	256	256	4
Δ-ケプストラム(16次)	256	256	
Δ-パワー	64	64	

3.3 Bigram の作成

離散型，半連続型用の Bigram，および連続型用の条件付き確率は，それぞれ，入力話者が発声した音声データのみから計算した入力話者用と，多数話者の音声データから計算した不特定話者用の 2 種類を用意した．入力話者用は，入力話者に限定したフレーム間相関情報を含んでおり，不特定話者用は，多数話者に含まれる一般的なフレーム間相関情報を含んでいる．図 5 に示すように，入力話者用 Bigram からは入力話者用 Bigram 制約 HMM を，不特定話者用 Bigram からは不特定話者用 Bigram 制約 HMM を作成することができる．表 1 に示すように，入力話者用は，音響学会の連続音声データベースの対話読上げ文より，各入力話者が発声した平均 50 文（3 分間の音声）を用いて作成した．この文章データは，HMM の評価用とはテキスト独立である．離散型，半連続型用 Bigram を計算する場合，見掛け上のデータ量を増やすために

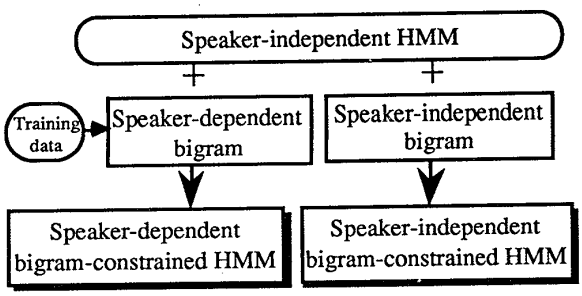


図 5　話者依存/独立 Bigram
Fig. 5　Speaker-dependent/independent bigram.

表 3　音韻平均認識率［%］

HMM Type		離散型	半連続型	連続型
従来の不特定話者用HMM		60.2	65.4	64.8
Bigram-constrained HMM	入力話者用	72.5	74.8	74.5
	不特定話者用	64.8	66.5	67.5

Fuzzy VQ を使用した．不特定話者用は，音韻 HMM の学習に用いた連続音声データベースより，話者，あるいは発声テキストに片寄りがないように 2,000 文を抜き出して作成した．

以前の検討結果から，Bigram を計算する際に，音韻ごとの出現頻度を正規化した方が認識性能が良いことがわかっている[6],[7]．そこで，入力話者用，不特定話者用とも，発声テキストをもとに（教師あり学習）不特定話者用 HMM を連結し，Viterbi セグメンテーションして音韻境界を求めた後，音韻ごとに Bigram（または条件付き確率）を計算した．

3.4 実験結果

各モデルの評価用データに対する 23 音韻平均認識率を表 3 に示す．従来の HMM では，半連続型，連続型の認識率が良く，離散型は若干悪いことがわかる．Bigram 制約 HMM は，いずれの場合も従来の同じ型の HMM の性能を上回っている．今回提案した半連続型，連続型 Bigram 制約 HMM は，以前提案した離散型 Bigram 制約 HMM よりも性能が良くなっており，これらのモデルの有効性が示された．入力話者用 Bigram 制約 HMM は，不特定話者用 Bigram 制約 HMM よりも高い性能が得られているが，これは入力話者に依存したフレーム間相関情報を用いているためで，一種の話者適応になっている．不特定話者用 Bigram 制約 HMM は，話者適応化用の音声データを特に必要とせず，HMM の学習用データからフレーム間相関情報を独立に取り出して利用している．従って，表 3 における従来の不特定話者用 HMM と不特定話者用 Bigram 制約 HMM の性能の差は，従来の HMM にフレーム間相関情報を加えた効果とみることができる．

今後，更に高い性能を得るために，Trigram など，より制約の強い情報を用いることが考えられる．また，入力話者用 Bigram 制約 HMM では，従来の話者適応技術と組み合わせ，適応化後に Bigram を用いる方法も考えられる．

4. むすび

2 フレーム間の特徴パラメータの相関を利用し，不特定話者用 HMM の分布を入力話者に適するように制約する Bigram 制約 HMM を，離散型，半連続型，連続型について定式化し，評価実験を行った．この手法によれば，VQ コードの Bigram（離散型，半連続型の場合），または前フレームの音韻特徴量の条件付き確率（連続分布の場合）を用いることにより，従来の不特定話者用 HMM から，更に性能の良い不特定話者用 Bigram 制約 HMM，あるいは入力話者に適応化した入力話者用 Bigram 制約 HMM を得ることができる．

これらモデルを連続音声中の音韻認識によって評価した．適応化用音声 50 文を用いた入力話者用 Bigram 制約 HMM では，半連続型が平均音韻認識率 74.8% と最も性能が良く，従来の半連続型不特定話者用 HMM よりも 9.4% 性能が改善された．また，連続型においても 74.5% を達成し，従来よりも 9.7% の改善を得た．一方，不特定話者用 Bigram 制約 HMM では連続型が 67.5% と最も良く，従来の連続型 HMM にフレーム間相関情報を加えたことで 2.7% の改善が得られた．本モデルは，従来の HMM に各時刻で出力確率分布が動的に変化する仕組みをもたせたもので，パラメータにケプストラムの回帰係数を用いた場合でも有効であることが実証された．今後，相関情報をより強く，また詳細にすることにより，更に性能の向上が期待できる．

謝辞　日ごろ御指導頂く古井特別研究室古井貞熙室長，有意義な討論をして下さった基礎研究所川端主任研究員，ならびに古井特別研究室の方々に感謝致します．なお実験では，日本音響学会連続音声データベースを使用した．

文　献

（1）　Furui S.: "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE

Trans. Acoust., Speech & Signal Process., **ASSP-34**, 2, pp. 52-59 (1986).
(2) Kenny P., Lenning M. and Mermelstein P.: "A linear predictive HMM for vector-valued observations with applications to speech recognition", IEEE Trans. Acoust., Speech & Signal Process., **ASSP-38**, 2, pp. 220-225 (1990).
(3) 坪香英一："ニューラルネット駆動型 HMM", 信学技報, **SP89**-83 (1989).
(4) 高橋 敏, 南 泰浩, 松岡達雄, 鹿野清宏："フレーム間相関を用いた音韻 HMM", 信学技報, **SP92**-50 (1992).
(5) Takahashi S., Matsuoka T., Minami Y. and Shikano K.: "Phoneme HMMs constrained by frame correlations", Proc. ICASSP-93, **2**, pp. 219-222 (1993).
(6) 高橋 敏, 松岡達雄, 鹿野清宏："VQ コードの Bigram で制約した HMM による音声認識", 信学技報, **SP91**-86 (1991).
(7) Takahashi S., Matsuoka T. and Shikano K.: "Phonemic HMMs constrained by statistical VQ-code transition", Proc. ICASSP-92, pp. 553-556 (1992).
(8) Rabiner L. L., Juang B. -H., Levinson S. E. and Sondhi M. M.: "Recognition of isolated digits using Hidden Markov Models with continuous mixture densities", AT & T Technical Journal, **64**, 6, pp. 1211-1234 (1985).
(9) Bellegarda J. R. and Nahamoo D.: "Tied mixture continuous modeling for speech recognition", IEEE Trans. Acoust., Speech & Signal Process., **ASSP-38**, 12, pp. 2033-2045 (1990).
(10) Paul D. B.: "The Lincoln tied-mixture HMM continuous speech recognizer", Proc. ICASSP-91, pp. 329-332 (1991).
(11) Huang X. D.: "Phoneme classification using semicontinuous Hidden Markov Models", IEEE Trans. Signal Process., **SP-40**, 5, pp. 1062-1067 (1992).
(12) 有木康雄："離散, セミ連続, 連続型 HMM の相互関係", マルコフモデルニューラルネットワークを包含する新しい音声認識手法の総合的研究第 6 回資料 (1990).
(13) Wellekens C. J.: "Explicit time correlation in Hidden Markov Models for speech recognition", Proc. ICASSP-87, pp. 384-386 (1987).
(14) Brown P. F.: "The acoustic-modeling problem in speech recognition", Ph. D dissertation, CMU (1987).
(15) 南 泰浩, 松岡達雄, 鹿野清宏："不特定話者連続音声データベースによる連結学習 HMM の評価", 信学技報, **SP91**-113 (1992).

付　　録

式(16)における条件付き確率 $p(y_2|y_1)$ と HMM の混合正規分布 $b_{ki}(y_2)$ との積の計算を以下に示す.

式(17)から,

$$p(y_2|y_1)b_{ki}(y_2)=\frac{|V^{(p)}V^{(s)ki}|^{1/2}}{(2\pi)^n}\exp\Big[-\frac{1}{2}\{(y_2-\delta^{(p)})^tV^{(p)}(y_2-\delta^{(p)})+(y_2-\delta^{(s)ki})^tV^{(s)ki}(y_2-\delta^{(s)ki})\}\Big]$$

指数部の一部を取り出して 2 次形式にすると,

$$\begin{aligned}&(y_2-\delta^{(p)})^tV^{(p)}(y_2-\delta^{(p)})+(y_2-\delta^{(s)ki})^tV^{(s)ki}(y_2-\delta^{(s)ki})\\&=y_2^t(V^{(p)}+V^{(s)ki})y_2-2y_2^t(V^{(p)}\delta^{(p)}+V^{(s)ki}\delta^{(s)ki})+(\delta^{(p)t}V^{(p)}\delta^{(p)})+(\delta^{(s)kit}V^{(s)ki}\delta^{(s)ki})\\&=\{y_2-(V^{(p)}+V^{(s)ki})^{-1}\cdot(V^{(p)}\delta^{(p)}+V^{(s)ki}\delta^{(s)ki})\}^t(V^{(p)}+V^{(s)ki})\cdot\{y_2-(V^{(p)}+V^{(s)ki})^{-1}(V^{(p)}\delta^{(p)}+V^{(s)ki}\delta^{(s)ki})\}\\&\quad+(\delta^{(p)t}V^{(p)}\delta^{(p)})+(\delta^{(s)kit}V^{(s)ki}\delta^{(s)ki})-(V^{(p)}\delta^{(p)}+V^{(s)ki}\delta^{(s)ki})^t\cdot(V^{(p)}+V^{(s)ki})^{-1}(V^{(p)}\delta^{(p)}+V^{(s)ki}\delta^{(s)ki})\end{aligned}$$

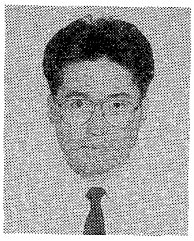
従って, 式(18)の形式と照らし合わせると,

$$p(y_2|y_1)b_{ki}(y_2)=\frac{|V^{(p)}V^{(s)ki}|^{1/2}}{(2\pi)^n}\exp\Big[-\frac{1}{2}\{(y_2-\hat{\delta})^t\hat{\Lambda}^{-1}(y_2-\hat{\delta})+\Psi\}\Big]\qquad(18)$$

から

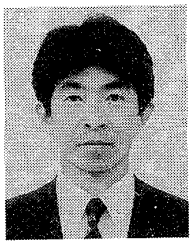
$$\begin{aligned}\hat{\delta}&=(V^{(p)}+V^{(s)ki})^{-1}(V^{(p)}\delta^{(p)}+V^{(s)ki}\delta^{(s)ki})\\\hat{\Lambda}&=(V^{(p)}+V^{(s)ki})^{-1}\\\Psi&=(\delta^{(p)t}V^{(p)}\delta^{(p)})+(\delta^{(s)kit}V^{(s)ki}\delta^{(s)ki})-(V^{(p)}\delta^{(p)}+V^{(s)ki}\delta^{(s)ki})^t(V^{(p)}+V^{(s)ki})^{-1}(V^{(p)}\delta^{(p)}+V^{(s)ki}\delta^{(s)ki})\end{aligned}$$

(平成 5 年 7 月 1 日受付, 9 月 22 日再受付)



高橋　敏

昭 62 早大・理工・電気卒. 平 1 同大大学院修士課程了. 同年日本電信電話(株)入社. NTT ヒューマンインタフェース研究所勤務. 以来, 音声認識の研究に従事. 平 5 日本音響学会粟屋潔学術奨励賞受賞. IEEE, 日本音響学会各会員.



松岡　達雄

昭57早大・理工・電子通信卒．昭59同大大学院修士課程了．同年日本電信電話公社(現NTT)入社．横須賀電気通信研究所においてディジタル電話端末の通話系構成法の研究に従事．昭62より，NTTヒューマンインタフェース研究所においてニューラルネット，HMMによる音声認識の研究に従事．平4～5，AT & T Bell研究所(Murray Hill)において客員研究員として主に話者適応化の研究に従事．現在，NTTヒューマンインタフェース研究所古井特別研究室主任研究員．日本音響学会，IEEE各会員．

南　　泰浩

昭61慶大・理工・電気卒．平3同大大学院博士課程了．同年日本電信電話(株)入社．現在ヒューマンインタフェース研究所研究主任．音声認識の研究に従事．日本音響学会会員．

鹿野　清宏

昭45名大・工・電気卒．昭47同大大学院修士課程了．同年電電公社武蔵野電気通信研究所入所．昭59～61カーネギーメロン大客員研究員．昭61～平2 ATR自動翻訳電話研究所音声情報処理研究室長．現在，NTTヒューマンインタフェース研究所主席研究員．工博．主として音声認識の研究に従事．昭50本会米沢賞，平3 IEEE SP 1990 Senior Award受賞．IEEE，音響学会，情報処理学会各会員．