音声言語情報処理 39－3
(2001. 12. 20)

HMMと音節セグメントの統計量を用いた音節認識

高橋 伸寿† 中川 聖一†

† 豊橋技術科学大学・情報工学系
〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘1-1

E-mail: †{nobutosi,nakagawa}@slp.ics.tut.ac.jp

あらまし 我々はこれまでに、連続する4フレーム分のメルケプストラム係数を1つのベクトルとし、このベクトルに対してKL展開を用いて20次元に圧縮し入力ベクトルとするセグメント単位入力HMMの研究を行なっており、その有効性を示している。しかし、このようなモデルにおいては隣接するフレーム間の相関は考慮されているが、長区間にまたがるフレーム間の相関は考慮されていない。本報告では、より長い区間の特徴ベクトルの相関を表現するために、音節HMMの各状態に割り当てられる特徴量の平均値を連結したベクトルを1つの入力ベクトルとする、音節セグメントの統計量を用いた音声認識手法の提案を行なう。この特徴ベクトルは次元数が大きいため、KL展開を用いて次元圧縮を行なう。音節セグメント統計量はGMMでモデル化する。音節セグメント統計量を用いることにより、より長い区間にわたる特徴ベクトルの相関を表現することが可能となる(例えば、第1状態の特徴ベクトルと第4状態の特徴ベクトルなど)。モデルの学習と評価はベースとなるHMMでのフォースアライメントにより切り出された音節区間を用いて行なった。切り出し区間の音節認識実験ではベースモデルと音節セグメントモデルの両方を併用することにより音節認識率が83.7%から87.7%に向上した。

キーワード 音節認識, セグメントモデル, HMM

Syllable recognition using syllable-segmental statistics and syllable-based HMM

Nobutoshi TAKAHASHI† and Seiichi NAKAGAWA†

† Information and Computer Sciences, Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, 441-8580 Japan

E-mail: †{nobutosi,nakagawa}@slp.ics.tut.ac.jp

Abstract In our previous research, we demonstrated the validity of segmental unit input hidden Markov model (HMM), which regards successive four frame MEL-cepstrum coefficients as a feature vector. The vector is compressed into 20 dimensions using the KL transform. However, the model considers only the correlation between frames in a short section, but not the correlation between the frames over a long section. In this paper, in order to represent the correlation over a long distance, we use the syllable-segmental statistics that are calculated by the concatenation of feature vectors, corresponding to each state in a syllable based HMM. As this concatenated feature vector consists of a high dimension, the dimension is reduced using the K-L transform. The statistics are modeled by a GMM. The use of syllable-segment statistics allows the model to express the correlation between the frames over a long distance (e.g., the correlation between a vector in the first state and a vector in the fourth state in a syllable-based HMM). For modeling and estimating, we conducted a forced Viterbi alignment against continuous speech using a conventional HMM, and then we segmented continuous speech into syllable segments. By combining this approach with a segmental-unit input HMM, the syllable recognition rate was improved to 87.7% from 83.7% for syllables taken from continuous speech, without using a language model.

Key words syllable recognition, segment model, HMM

1. はじめに

隠れマルコフモデル (HMM) は現在の音声認識技術の基本となっている手法の一つである。しかし、基本的なHMMは音声特徴量がフレームごとに独立であるという仮定から、時間的な変動を十分に表現できないという欠点がある。この問題を解決すべく、我々はこれまでに連続する4フレーム分のメルケプストラム係数を1つのベクトルとし、このベクトルに対してKL展開を用いて20次元に圧縮し入力ベクトルとするセグメント単位入力HMMの研究を行なっており、その有効性を示している [1][2][3][4]。しかし、このようなモデルにおいては隣接するフレームの相関は考慮されているが、長区間にまたがるフレーム間の相関は考慮されていない。

長区間にまたがるフレーム間の相関を考慮する方法として、HMMとmulti-layer perceptronを組み合わせる方法がHowellにより提案されている [5]。この手法は、HMMによるViterbi整列により得られるスピーチパターンを時間長正規化し、それにより得られた固定長パターンをmulti-layer perceptronに入力している。Howellは、HMMとperceptronとの尤度の統合による手法は試みていない。一方、HMMとセグメントモデルを統合する研究が行なわれている。この手法では、音声を連続する2つの音素ペアでセグメンテーションし、trajectoryモデルによりモデル化している [6]。古山·小林は、状態遷移が過去の出力確率に依存する部分HMM法を提案し、HMMとの平滑化で高精度な単語認識結果を報告している [7]。

本報告では、より長い区間の特徴ベクトルの相関を表現するために、音節HMMの各状態に割り当てられる特徴量の平均値を連結したベクトルを1つの入力ベクトルとする、音節セグメントの統計量を用いた音声認識手法の検討を行なう。

連続音節中から切り出された音節認識の実験では、HMMと音節セグメントモデルを併用して用いることによりHMM単独での認識率よりも高い認識率が得ることができた。また、特に子音と母音の認識性能に注目し、2種類の音節セグメントモデルとHMMを組み合わせることによりさらなる性能の向上を得ることができた。これらの結果から、連続音節認識への適用による認識率の向上が期待される。

2. 音節セグメントモデル

本報告で提案する音節セグメントモデルでは、各状態での特徴量の平均値を状態数分だけ連結したものを1つの入力ベクトルとする。そのため、ベースとなる音節HMMモデルで各状態に割り当てられている区間の特徴量の平均ベクトルを求め (10次元)、これを状態数分用いて1つの入力ベクトル (40次元) とする分布を新たに推定する。(図1)

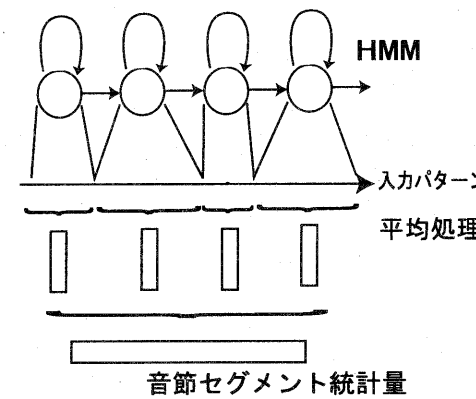

図1 音節セグメントモデルの概念図

これにより、従来手法では状態間や混合分布間の遷移確率では表現できない、例えば1状態目と4状態目といった長い区間での相関を表現することができる。今回音節セグメントモデルを作成するにあたり、まずベースとなる音節HMMモデルを用いて評価用音声データのフォースアライメントを行ない、音節を切り出した。音響モデルの学習文には、日本音響学会データベースATR503文A～Jセットと新聞記事読み上げ音声コーパスの男性話者合計17221文を用いた。HMMを用いたアライメントで得られた時間情報を元に、各音節に割り当てられている区間の特徴ベクトル系列の切り出しを行なった。さらに、それぞれの状態に割り当てられている特徴ベクトルの平均値を求め、それらを連結して一つの入力ベクトルとして各音節毎に混合多次元正規分布でモデル化する。

3. セグメント単位入力HMM

入力シーケンス $y = y_1 y_2 \cdots y_T$ (Tは入力長) と状態シーケンス $x = x_1 x_2 \cdots x_T$ に対して、HMMの出力確率の計算式から以下の導出ができる [3]。

$$P(y_1 \cdots y_T)$$

$$= \sum_x \prod_i P(y_i|y_1 y_2 \cdots y_{i-2} y_{i-1}, x_1 x_2 \cdots x_{i-1} x_i) \times P(x_i|x_1 x_2 \cdots x_{i-1}) \quad (1)$$

$$\simeq \sum_x \prod_i P(y_i|y_{i-3} y_{i-2} y_{i-1}, x_{i-1} x_i) P(x_i|x_{i-1}) \quad (2)$$

$$= \sum_x \prod_i \frac{P(y_{i-3} y_{i-2} y_{i-1} y_i|x_{i-1} x_i)}{P(y_{i-3} y_{i-2} y_{i-1}|x_{i-1} x_i)} P(x_i|x_{i-1}) \quad (3)$$

$$\simeq \sum_x \prod_i \frac{P(y_{i-1} y_i|x_{i-1} x_i)}{P(y_{i-1}|x_{i-1} x_i)} P(x_i|x_{i-1}) \quad (4)$$

$$= \sum_x \prod_i P(y_i|y_{i-1}, x_{i-1} x_i) P(x_i|x_{i-1}) \quad (5)$$

$$\simeq \sum_x \prod_i P(y_{i-1} y_i|x_{i-1} x_i) P(x_i|x_{i-1}) \quad (6)$$

$$\simeq \sum_x \prod_i P(y_i|x_{i-1} x_i) P(x_i|x_{i-1}) \quad (7)$$

式(2)あるいは式(3)が4フレーム幅の条件付HMMである。式(4)あるいは式(5)は2フレーム幅の条件付HMMである。

ここで、式(3)を分子だけを使用するように近似すると、

$$P(y_1 \cdots y_T) \simeq \sum_x \prod_i P(y_{i-3} y_{i-2} y_{i-1} y_i, x_{i-1} x_i) P(x_i|x_{i-1}) \quad (8)$$

となる。式(8)が4フレーム幅のセグメント単位入力HMMを定義する。フレーム単位の入力の代わりにセグメント単位を入力とすることから、これをセグメント単位入力HMMと呼んでいる。

4. 混合要素分布選択の偏り

音節セグメントモデルは、各状態の特徴量の平均を扱うことから、各状態で特徴量が定常であることが望ましい。即ち、もしベースとなるHMMで認識を行なう際にある音節のある状態では同じ混合要素分布を選択しており、更に話者やコンテキスト等の影響によって混合要素分布の選択に特定のパターンが見られるのなら、音節セグメントモデルを用いることによって長い距離の相関を表現するとともに、ある種の要素分布trajectoryモデルのような、本来現れないような要素分布の遷移を抑制する効果を得ることが期待できる。そこで、学習用コーパスでの要素分布選択の偏りを集計した。各音節入力パターンに対して、HMMでのアライメントの結果、音節HMMの各状態で同じ混合要素分布を対応する総フレーム中の5割以上のフレームで選択している音節を特定パターンを持つものとみなし集計した。このような音節は音節によって多少の差があるが平均して全体の音節出現回数の約5割程度であった。また、そのパターン数はほぼ16種類で収束した(例えば、第一状態は第三要素分布、第2状態は第一要素分布、…というようなパターン。可能なパターン総数は、各音節で4状態4混合の場合 $4^4 = 256$)。この結果から、混合分布の選択にはある程度の偏りがあることが予想される。このことから、各状態で選択される混合要素分布の全状態にわたる出現パターンも新しい特徴ベクトルとなりうることがわかる。

5. 認識実験

5.1 実験条件

音声の分析条件は、サンプリング周波数12kHz、分析窓長21.33ms、フレーム周期8msecである。特徴量は10次のLPCメルケプストラムおよびMFCCにΔケプストラム+ΔΔケプストラム+Δパワー+ΔΔパワーを加えた計32次元である。セグメント単位入力の場合は、4フレームの40次をKL展開で20次元に圧縮して用いた。ベースとなるHMMは、連続出力分布型5状態4出力分布型で、各出力分布を全共分散行列4混合分布で表現した音節モデルを用いた[3]。音節数は114である。学習用データの発声者と異なる男性話者10名の新聞読み上げ文(計939文)のベースHMMによる音節認識率を表1に示す。表中のモデルで、frm-HMMはフレーム単位入力HMMを、seg-HMMはセグメント単位入力HMMを用いていることを示している。CO(continous)は、不特定話者連続音節認識実験の認識率である。AL(alignment)はアライメントにより切り出された音節区間の認識実験の認識率である。

表1 ベースHMMの音節認識率[%]

モデル	Acc.	Cor.	Sub.	Ins.	Del.
frm-HMM-CO	67.3	75.7	21.4	8.3	3.0
frm-HMM-AL	81.0	81.0	19.0	0.0	0.0
seg-HMM-CO	70.1	78.3	19.1	8.2	2.7
seg-HMM-AL	81.3	81.3	18.7	0.0	0.0

5.2 音節セグメントモデルのみでの認識実験

音節セグメントモデルは、対角共分散行列と全共分散行列で表現される混合多次元正規分布で作成し

た。混合数は、全共分散行列で1混合と4混合、対角共分散行列で16混合と32混合をそれぞれ作成した。また、音節セグメントの40次元の特徴量をKL展開し、20次元に圧縮した入力ベクトルを用いたモデルもそれぞれ作成した。さらに、同様にしてΔケプストラムを併用したモデルも作成した。Δケプストラムもケプストラムと同様に、ベースHMMで切り出された区間の特徴量の平均を入力としてモデルの作成／スコアの計算をする。それぞれのモデルで、切り出し区間に対する認識実験の音節認識率を表2に示す。

表中のモデル名で、Dは対角共分散行列、Fは全共分散行列、KLは特徴量をKL展開し20次元に圧縮して入力するモデル、数値は混合数を示している。また、CCはベースモデル(frm-HMM)での切り出し区間の認識で正解が得られ、状態セグメントモデルでの認識でも正解が得られた音節数。CSはベースHMMでは正解であったが音節セグメントモデルで誤って認識された音節数。SCはベースHMMで誤って認識されたが音節セグメントモデルで正解が得られた音節数。SSはベースHMMでも音節セグメントモデルでも誤って認識された音節数である。

表2　切り出し区間に対する音節セグメントモデルの認識率[%]及び、ベースHMMとの正解単語傾向の集計

モデル	COR[%]	CC	CS	SC	SS
D16	47.7	16516	14125	1520	5675
KL-D16	55.9	19351	11290	1804	5391
KL-D32	55.5	19198	11443	1813	5382
KL-F1	66.8	23227	7414	2046	5149
KL-F4	68.9	23747	6894	2302	4893
KL-F1+Δ	75.4	25919	4722	2595	4600
KL-F4+Δ	78.4	26512	4129	3141	4054

混合数を増やすことによって音節セグメントモデルの認識性能が向上している。また、直接40次元の入力ベクトルを扱うモデルよりもKL展開して20次元に圧縮した入力ベクトルを用いるモデルの方が認識性能が良い。更に、Δケプストラムを用いることによって認識率が向上し78.4%の正解が得られていることに加え、ベースHMMと音節セグメントモデルの両方で誤って認識された音節数(SS)を4054個(10.7%)にまで減らすことができている。

この結果から、音節セグメントモデルを用いた場合、ベースHMMとは異なる正解の傾向が得られており、この2つを併用することにより認識性能を向上させることが期待できる。

5.3　ベースHMMと音節セグメントモデルとの併用

ベースHMMの音響尤度と音節セグメントモデルの音響尤度を併用して切り出し区間の音節認識実験を行なった。ベースHMMと音節セグメントモデルの尤度の併用により得られる尤度の算出法は以下の通りである。

$$AS = \alpha AS_{HMM} + (1-\alpha) AS_{syllablesegment} * frames, \qquad (9)$$

ここで、AS_{HMM}はHMMにより得られる音響尤度、$AS_{syllablesegment}$は音節セグメントモデルにより得られる音響尤度である。尤度の正規化のために音節セグメントの尤度は切り出し区間のフレーム数倍($frames$)してあり、最終的な尤度はαによる重みづけ和により求める。結果を表3、表4に示す。

表3　ベースモデルと音節モデルの併用による切り出し区間の認識(LPCメルケプストラム)

モデル	認識率[%]
音節セグメントのみ	78.4
frm-HMMのみ	81.0
frm-HMM + 音節セグメントモデル	85.3
seg-HMMのみ	81.3
seg-HMM + 音節セグメントモデル	85.4

表4　ベースモデルと音節モデルの併用による切り出し区間の認識(MFCC)

モデル	認識率[%]
音節セグメントのみ	78.8
frm-HMMのみ	82.9
frm-HMM + 音節セグメントモデル	86.6
seg-HMMのみ	83.7
seg-HMM + 音節セグメントモデル	87.1

ベースHMMのみを用いて認識した場合に比べて、約4%の認識性能の向上が得られている。

αを0から1まで変化させた時の認識率を、図2及び図3に示す。

図中で、左端が$\alpha = 0$すなわち音節セグメントモデルのみでの認識性能であり、右端が$\alpha = 1$すなわちHMMのみでの認識性能である。音節セグメントモデルとHMMを併用することによって、ピークが得られ、双方モデルの併用によって性能が向上していることを示している。

また、この時の子音と母音の認識率をまとめたの

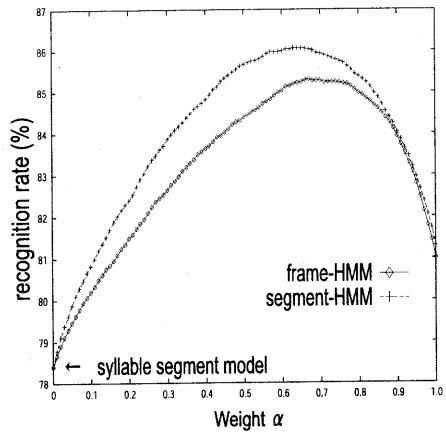

図 2　HMM と音節セグメントモデルの組み合わせによる認識率 (LPC)

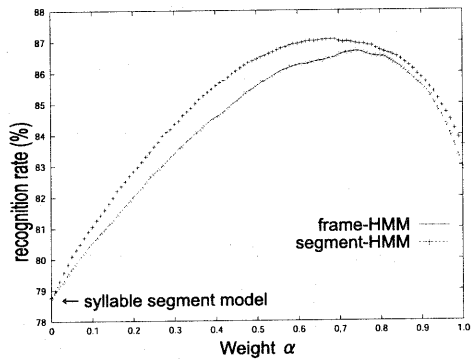

図 3　HMM と音節セグメントモデルの組み合わせによる認識率 (MFCC)

が表5である。ただし、子音の認識率算出時には、単母音は認識対象から除外し、撥音は子音に含めてまとめている。母音の認識率算出時には、撥音を除外している。単独のモデルの使用に比べ、母音、子音ともに向上が見られるが、特に子音の認識率が大きく向上している。

表 5　母音と子音の認識率 [%]

モデル	母音	子音
音節セグメントモデルのみ	92.6	79.5
フレーム入力 HMM のみ	93.4	85.4
セグメント入力 HMM のみ	93.8	86.5
音節セグメントモデル ＋フレーム入力 HMM	94.8	88.1
音節セグメントモデル ＋セグメント入力 HMM	95.1	88.7

5.4　部分的な音節セグメント統計量を用いたモデル

音節セグメントモデルは、40 次元の特徴量を 20 次元に次元圧縮して用いる。すなわち、20 次元で音節全体を表現していることになる。この 40 次元の音節セグメント統計量のうち、前半 2 状態分にあたる 20 次元の統計量を未圧縮でモデル化すれば、子音の情報をより多く表現することができ、子音の認識に特化したモデルを構築することができると考えられる。また逆に、後半 2 状態分 (もしくは 3 状態分) の 20 次元 (もしくは 30 次元) の統計量をモデル化することにより母音の認識に特化したモデルが構築できると考えられる。

そこで、音節セグメント統計量のうち前半 20 次元と 30 次元 (KL 展開により 20 次元に圧縮) を用いたモデルと後半 20 次元を用いたモデルを構築し、母音と子音の認識率を集計した。結果を表 6 に示す。

前半 2 状態分の統計量を用いることにより、全 4 状態分の統計量を用いたモデルに比べて子音の認識率が 79.5%から 86.7%に向上している。しかし、後半 2 状態分および後半 3 状態分の統計量を用いたモデルでは母音認識率が逆に低下している。これは、母音の認識に有効な情報が前半の 1 状態目にも存在しているためと考えられる。

表 6　母音と子音の認識率 [%]

モデル	母音	子音
HMM	93.8	86.5
全 4 状態音節セグメントモデル	92.6	79.5
前半 2 状態音節セグメントモデル	-	86.7
後半 2 状態音節セグメントモデル	91.6	-
後半 3 状態音節セグメントモデル	91.7	-
HMM+全 4 状態音節セグメントモデル	95.1	88.7
HMM+前半 2 状態セグメントモデル ＋全 4 状態音節セグメントモデル	95.2	89.2

5.5　音節セグメントモデル同士の併合

音節セグメント統計量を用いたモデルでも、その統計量の扱いの違いにより子音の認識性能が良いモデルと母音の認識性能が良いモデルを得ることができた。そこで、この 2 種類のモデルを併合することにより更に音節セグメントモデルの音節認識性能を向上させることができるのではないかと考えられる。そこで、同様にして前半 2 状態音節セグメントモデルと全 4 状態音節セグメントモデルの尤度重みづけ和による切り出し区間の音節認識実験を行なった。結果を表 7 に示す。

表 7　4 状態音節セグメントモデルと、2 状態音節セグメントモデルとの組み合わせによる音節認識性能の比較

モデル	音節認識率 [%]
全 4 状態音節セグメントモデル	78.8
前半 2 状態音節セグメントモデル	79.0
全 4 状態モデル＋前半 2 状態モデル	83.7
セグメント単位入力 HMM	83.7

子音の認識性能が良い前半 2 状態音節セグメントモデルと、母音の認識性能が良い全 4 状態音節セグメントモデルを組み合わせることにより、セグメント単位入力 HMM と同等の音節認識性能を得ることができた。

さらに、組み合わせた音節セグメントモデルとセグメント単位入力 HMM とを併合した認識実験を行なった。まず、音節セグメントモデル同士の尤度は表 7 の結果が得られた時の、最も認識率の良かった重みで結合する。これを音節セグメントモデルの尤度とし、式 (9) と同様に HMM との尤度重みづけ和によって認識実験を行なった。結果を表 8 に示す。また、重み α を 0 から 1 まで変化させた時の認識率を図 4 に、上位 N 音節中に正解音節が含まれる割合を表 9 に示す。

表 8　組み合わせた音節セグメントモデルと HMM との併合による認識 [%]

モデル	音節認識率
全 4 状態セグメントモデル＋前半 2 状態モデル	83.7
HMM ＋全 4 状態音節セグメントモデル	87.1
HMM ＋全 4 状態モデル＋前半 2 状態モデル	87.7

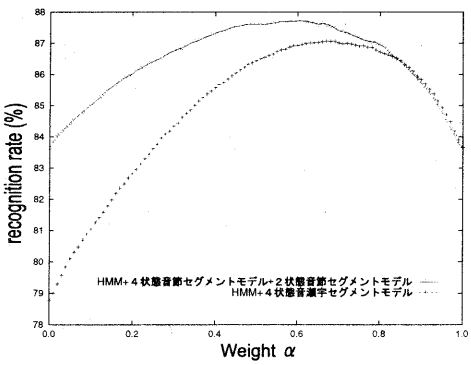

図 4　全 4 状態音節セグメントモデル、前半 2 状態音節セグメントモデル、HMM の組み合わせによる音節認識率

音節セグメントモデルの性能が、全 4 状態音節セグメントモデルのみに比べ向上したことにより、全体的に認識率が向上し、最大で 87.7%の音節認識率が得られている。また、上位 5 音節中にはほぼ正解が含まれていることが見られる。

表 9　各切り出し区間において、上位 N 音節中に正解が含まれる割合 (%))

	N				
モデル	1	2	3	4	5
(LPC) HMM+全 4 状態モデル	86.1	94.5	97.0	98.1	98.7
(MFCC) HMM+全 4 状態モデル	87.1	95.2	97.3	98.2	98.7
(MFCC) 3 モデルの組合せ	87.7	95.4	97.3	98.3	98.7

6. ま と め

HMM の各状態に割り当てられる特徴ベクトルを平均し、全状態のベクトルを連結して 1 つの入力ベクトルとする音節セグメントモデルを作成し評価を行なった。モデルの学習と評価はベースとなる HMM でのフォースアライメントにより切り出された音節区間を用いて行なった。さらに、子音認識性能の高い音節セグメントモデルを組み合わせることにより音節認識率が向上し、切り出し区間の音節認識実験ではベースモデルと音節セグメントモデルを併用することにより認識率が 83.7%から 87.7%に向上した。この結果から、連続音節単語認識実験においても音節セグメントモデルを併用し認識率を向上できることが期待できる。

文　献

[1] 平田好充, 早川勲, 小野義之, 中川聖一: セグメント統計量を用いた HMM による音声認識, 音声技法,SP-96-69(1990)

[2] 中川聖一, 平田好充, 小野義之: 固定長セグメント統計量を用いた HMM による音声認識, 信学論誌,Vol.J75-D-II,No.5,pp.843-851(1992)

[3] 中川聖一, 山本一公：セグメント統計量を用いた隠れマルコフモデルによる音声認識, 信学論誌,Vol.79-D-II,No.12,pp.2032-2038(1996)

[4] 中川聖一, 花井建豪, 山本一公、峯松信明: HMM に基づく音声認識のための音節モデルと triphone モデルの比較, 電子情報通信学会論文誌,VOL.J83-D-II,No.6(2000)

[5] D.N.Howell: The multi-layer perceptron as a discriminating post processor for hidden Markov networds, SPEECH '88,pp.1389-1396(1988)

[6] Hsiao-Wuen Hon, Shankar Kumar, and Kuansan Wang: Unifying HMM and Phone-Pair Segment Models, ICSLP2000,pp.1184-1187(2000)

[7] 古山純子, 小林哲則：部分隠れマルコフモデルによる単語音声認識, 信学論誌,Vol.83-DII,No11,pp.2379-2387(2000)