Vol.2011-SLP-85 No.13
2011/2/5

Hidden Conditional Neural Fields を用いた音声認識における目的関数と階層的音素事後確率特徴量の検討

藤井康寿[†1]　山本一公[†1]　中川聖一[†1]

我々は，Hidden Conditional Neural Fields(HCNF) を用いた音声認識手法について検討を進めている．本稿では，HCNF を学習するための目的関数として，正解状態系列が一意に定まらない場合においても正解状態系列に対するエラー数を考慮した学習が可能となる Hidden Boosted MMI(HB-MMI) を提案する．HB-MMI を用いることで，過学習が起こりにくい状況では認識率を改善できることがわかった．本稿では，HCNF が出力する音素事後確率を次段の HCNF の特徴量とする階層的音素事後確率特徴量を用いた音声認識手法についても述べる．階層的音素事後確率特徴量を単独で用いる場合でも認識率の改善を得ることができたが，HB-MMI 学習を組み合わせることで，さらなる改善を得ることができた．

Investigations of Objective Functions and Hierarchical Phoneme Posterior Feature for Hidden Conditional Neural Fields based Automatic Speech Recognition

Yasuhisa Fujii,[†1] Kazumasa Yamamoto[†1]
and Seiichi Nakagawa [†1]

We have investigated automatic speech recognition using Hidden Conditional Neural Fields(HCNF). In this paper, we propose a new objective function, Hidden Boosted MMI(HB-MMI), which can consider the number of errors in training data even if the correct state sequence is not known for training HCNF. The experimental results show that HB-MMI can improve recognition accuracy when overfitting does not occur. In this paper, we also present an automatic speech recognition method using hierarchical phoneme posterior feature where the output of the first HCNF is used for the input of the second HCNF. The experimental results show that the feature can improve the recognition accuracy. By using both of the proposed methods, we obtained further improvement.

1. はじめに

HMM は系列モデリングのための優れたモデルであるため音声認識において広く用いられてきたが，HMM が抱える本質的な欠点もいくつか指摘されている[1)]．特に，状態が与えられた上でのフレーム間の特徴量の独立性を仮定しているために，数フレームにまたがる特徴を十分に考慮できないこと，また，本質的に生成モデルであるがゆえに識別能力に欠けることの 2 点は特に解決すべき問題である．これまで，HMM を用いた枠組みのなかでこれらの欠点を克服するための様々な研究が行われてきた[2)–5)]．

我々は，数フレームにまたがる特徴を考慮でき，かつ，識別能力が高いモデルを用いた方法として，Hidden Conditional Neural Fields(HCNF) を用いた音声認識手法を提案している[6)]．HCNF は，数フレームにまたがる非線形な特徴量間の関係をモデル化でき，かつ，識別モデルであるが故に生得的に識別能力が高いため，前述の 2 つの問題点を解決することができる手法である．実際に，TIMIT コーパス上での音素認識実験によって，HCNF の有効性が示されている[6)]．

文献 6) では，HCNF のパラメータを事後確率最大化の枠組み (MMI) で学習する手法を提案したが，事後確率の上昇は必ずしも認識率の上昇に直接関係しないため，学習基準としては学習データに対するエラーを直接的に考慮できることが望ましい．そのため，本稿では，学習データに対するエラーを直接的に考慮できる目的関数として，HMM における Boosted MMI[7)] を HCNF に適用した Hidden Boosted MMI(HB-MMI) を提案する．HCNF では，正解状態系列が観測不可能である (隠れている) ために，正解状態系列に対するエラー数を直接定義できないが，HB-MMI 学習では，各フレームにおける各状態の期待値を計算することでエラー数の期待値を算出する．

本稿では，近年注目を集める Tandem[8)] や CRANDEM[9)] などの音素事後確率特徴を用いた手法[10)] を参考に，HCNF が出力する各状態に対する事後確率を後段の HCNF への入力とする階層的音素事後確率特徴量を用いた音声認識手法についても述べる．音素事後確率特徴量は，MFCC や PLP などの音響特徴量と比べて，コンテキストや話者性，雑音などに頑健であることや，長時間の音素事後確率特徴の変化を見ることで誤りの生じやすい音韻的なパターンを検出できることなどから注目を集めており，音素事後確率特徴量を使用することで認識率の改善が期待できる[10)]．

本稿の構成は以下の通りである．まず，2 節において HCNF を用いた音声認識について述べる．3 節では，HCNF において正解状態系列が観測不可能である場合でも学習データに

†1 豊橋技術科学大学情報・知能工学系
Department of Computer Science and Engineering, Toyohashi University of Technology

対するエラー数を考慮可能な目的関数である HB-MMI について述べる．4 節では，HCNF を用いた階層的音素事後確率特徴量による音声認識手法について述べる．5 節で評価実験を行い提案法を評価する．最後に 6 節でまとめを述べる．

2. Hidden Conditional Neural Fields を用いた音声認識

2.1 定 式 化

HCNF を用いた音声認識では，音響特徴量の系列 $X = (x_1, x_2, \ldots, x_T)$ が与えられた上で，対応するラベル列が $Y = (y_1, y_2, \ldots, y_T)$ である確率を以下のように計算する[6]．

$$P(Y|X) = \frac{1}{Z(X)} \sum_{S} \exp(\kappa(\Phi_n(X, Y, S) + \Psi_n(X, Y, S))) \tag{1}$$

ここで，$Z(X)$ は正規化項であり以下の様に定義される．

$$Z(X) = \sum_{Y'} \sum_{S} \exp(\kappa(\Phi_n(X, Y', S) + \Psi_n(X, Y', S))) \tag{2}$$

S は隠れ状態系列，κ は state-flattening 係数である[11]．$\Phi_n(X, Y, S)$ および $\Psi_n(X, Y, S)$ はそれぞれ観測特徴，遷移特徴と呼ばれ，以下のように定義される．

$$\Phi_n(X, Y, S) = \sum_{t} \sum_{g}^{K} w_{y_t, s_t, g} h(\theta^T_{y_t, s_t, g} \phi(X, Y, S, t)) \tag{3}$$

$$\Psi_n(X, Y, S) = \sum_{j} u_j \sum_{t} \psi_j(X, Y, S, t, t-1) \tag{4}$$

ここで，$\phi(X, Y, S, t)$ はフレーム t から得られる特徴量を並べたベクトルであり，$\theta_{y,s,g}$ はこれに対応する重みベクトルである．$w_{y,s,g}$ はゲート関数 $h(x)$ の出力に対応する重みである．$\psi_j(X, Y, S, t, t-1)$ はフレーム $t-1$ および t から得られる特徴量であり，u_j は対応する重みである．本稿において，$h(x)$ は以下のように定義される．

$$h(x) = \frac{c}{1 + \exp(-\alpha(x - \beta))} - b \tag{5}$$

b および c はゲート関数の値域を変更するための項であり，α および β はゲート関数の形を制御するための項である．

2.2 学 習

HCNF の学習は，以下の目的関数を最小化する $\lambda = \{w_{y,s,g}, \theta_{y,s,g}, u_j\}$ を発見する問題として定式化できる．

$$f(\lambda; D) = \ell(\lambda; D) + r(\lambda) \tag{6}$$

$\ell(\lambda; D)$ は学習データ $D = \{X^i, Y^i\}, i = 0, \ldots, N$ に基づいて定義される損失関数であり，$r(\lambda)$ は正則化項である．$r(\lambda)$ としては L1 正則化や L2 正則化などを用いることができる．

文献 6) では，事後確率最大化に基づいた損失関数として $\ell(\lambda; D)$ を以下のように定義した．

$$\begin{aligned} \ell_{MMI}(\lambda; D) &= -\sum_i \log P(Y^i|X^i) \\ &= -\sum_i \log \frac{\sum_S \exp(\kappa(\Phi_n(X, Y, S) + \Psi_n(X, Y, S)))}{\sum_{Y'} \sum_S \exp(\kappa(\Phi_n(X, Y', S) + \Psi_n(X, Y', S)))} \end{aligned} \tag{7}$$

$\ell_{MMI}(\lambda; D)$ の $\lambda = \{w_{y,s,g}, \theta_{y,s,g}, u_j\}$ による偏微分が計算可能であれば，勾配に基づく各種最適化手法を適用可能である．$\ell_{MMI}(\lambda; D)$ の $w_{y,s,g}$，$\theta_{y,s,g}$，u_j による偏微分は以下のように計算できる．

$$\begin{aligned} \frac{\partial \ell_{MMI}(\lambda; D)}{\partial w_{y,s,g}} =& -\kappa \sum_i E\left[\sum_t h(\theta^T_{y,s,g} \phi(X^i, Y^i, S, t))\right]_{S|X^i, Y^i} \\ &+ \kappa \sum_i E\left[\sum_t h(\theta^T_{y,s,g} \phi(X^i, Y, S, t))\right]_{Y,S|X^i} \end{aligned} \tag{8}$$

$$\begin{aligned} \frac{\partial \ell_{MMI}(\lambda; D)}{\partial \theta_{y,s,g}} =& -\kappa \sum_i E\left[\sum_t w_{y,s,g} \frac{\partial h(\theta^T_{y,s,g} \phi(X^i, Y^i, S, t))}{\partial \theta_{y,s,g}}\right]_{S|X^i, Y^i} \\ &+ \kappa \sum_i E\left[\sum_t w_{y,s,g} \frac{\partial h(\theta^T_{y,s,g} \phi(X^i, Y, S, t))}{\partial \theta_{y,s,g}}\right]_{Y,S|X^i} \end{aligned} \tag{9}$$

$$\begin{aligned} \frac{\partial \ell_{MMI}(\lambda; D)}{\partial u_j} =& -\kappa \sum_i E\left[\sum_t \psi_j(X^i, Y^i, S, t, t-1)\right]_{S|Y^i, X^i} \\ &+ \kappa \sum_i E\left[\sum_t \psi_j(X^i, Y, S, t, t-1)\right]_{Y,S|X^i} \end{aligned} \tag{10}$$

式 (5) の微分は以下のように計算できる．

$$\frac{dh(x)}{dx} = \frac{\alpha}{c}(b + h(x))(c - b - h(x)) \tag{11}$$

2.3 推 論

文献 12) では，HCRF の推論において，上位 N ベストを保持しながら探索を行うことで隠れ状態 S を周辺化しながら最尤の Y を求めるアルゴリズムを採用している．HCNF においても同様のアルゴリズムを使用することが可能であるが，本稿では，既存のデコーダとの親和性の高さから，HCNF の推論は Viterbi アルゴリズムによって行う．すなわち，隠れ状態 S を周辺化することなく最尤の系列 S を求めることで Y を推論する．

3. 目的関数の検討

3.1 Boosted MMI[7)]

HMM ベースの音声認識システムにおける Boosted MMI(B-MMI) 学習の目的関数は以下のように設定される[7)].

$$\ell_{BMMI-HMM}(\lambda) = \sum_{i=1}^{R} \log \frac{p_\lambda(X_i|Y_i)^\kappa P(Y_i)}{\sum_Y p_\lambda(X_i|Y)^\kappa P(Y) \exp(-bA(Y,Y_i))} \tag{12}$$

ここで，$p_\lambda(X|Y)$ は HMM による尤度，$P(Y)$ は言語モデルによる確率を表し，$A(Y,Y_i)$ は正解ラベル列 Y_i に対する Y の正解数を示す．B-MMI では，分母側が表す競合仮説において正解精度の高い仮説のスコアを相対的に弱める働きを持つ項 $\exp(-bA(Y,Y_i))$ を導入することで，競合仮説との間のマージンを強要するように目的関数を設定している．これは，対数線形モデルの場合には，SVM で使用される Hinge Loss を滑らかに近似した関数となる[13)].

3.2 Hidden Boosted MMI

式 (12) のように，HCNF の MMI 学習に相当する式 (7) に式 (12) の $\exp(-bA(Y,Y_i))$ に相当する項を加えることで，学習データに対するエラーを考慮した目的関数を設定することができると考えられる．文献 13) では，状態レベルのエラー関数を用いて，HCRF に対して同様の考え方を導入した手法を提案している．本稿においては，ラベル列 (単語列や音素列) に対するエラーではなく，状態レベルでのエラーを考える．文献 13) では，正解状態系列が既知であり，アライメントが固定されている場合の方法であったが，本稿では，正解状態系列が既知でない場合でも使用できる方法を提案する．正解状態系列が既知でない場合でも使用できることから，以後本稿で提案する目的関数を Hidden Boosted MMI(HB-MMI) と呼ぶ．HB-MMI の目的関数は，以下のように定義される．

$$\ell_{HB-MMI}(\lambda; D) = -\sum_i \log \frac{\sum_S \exp(\kappa(\Phi_n(X,Y,S) + \Psi_n(X,Y,S)))}{\sum_{Y'} \sum_S \exp(\kappa(\Phi_n(X,Y',S) + \Psi_n(X,Y',S))) \exp(-b\mathrm{Acc}(S,D^i))} \tag{13}$$

ここで，$\mathrm{Acc}(S,D^i)$ は学習データ D^i(音響特徴量の系列 X^i と対応する正解ラベル列 Y^i) が与えられたときの状態系列 S の期待正解状態数である．まず，状態系列 S' を正解とした時の S の正解数を S' と S の一致フレーム数として以下のように定義する．

$$\mathrm{Acc}(S,S') = \sum_t \delta(s_t, s'_t) \tag{14}$$

これを用いて，$\mathrm{Acc}(S,D^i)$ を以下のように定義する．

$$\begin{aligned}
\mathrm{Acc}(S,D^i) &= \sum_{S'} P(S'|Y^i,X^i)\mathrm{Acc}(S,S') \\
&= \sum_{S'} P(S'|Y^i,X^i) \sum_t \delta(s_t,s'_t) \\
&= \sum_t \sum_{S'} P(S'|Y^i,X^i)\delta(s_t,s'_t) \\
&= \sum_t \gamma^i(s_t,t)
\end{aligned} \tag{15}$$

$\gamma^i(s,t)$ は正解データ D^i が与えられたときのフレーム t における状態 s の事後確率である．

$$\gamma^i(s,t) = \sum_{S'} P(S'|Y^i,X^i)\delta(s,s'_t) \tag{16}$$

すなわち，$\mathrm{Acc}(S,D^i)$ は，正解データ D^i が与えられた上での，状態系列 S の各フレーム t における状態 s_t の期待正解率の和として定義される．$\mathrm{Acc}(S,D^i)$ を定数項として扱えば，式 (13) の偏微分は式 (8)-(10) と同一となる．式 (15) は，フレーム t に局所的な値の和として表現されるため，偏微分の計算に必要な期待値計算は，フレーム t における各状態のスコアに $-b\gamma^i(s_t,t)$ を加えるだけで，従来のアルゴリズムに変更を加えることなく行える．

4. 階層的音素事後確率特徴量の検討

1 節で述べた HMM の欠点を補うために，長時間の特徴の変化を考慮することができ，かつ識別的能力が高いモデルを HMM と組み合わせる試みが存在する．例えば，Tandem システムは，MLP を用いて予め音響特徴量系列を音素および弁別特徴の事後確率に変換し，これを HMM の入力とするモデルである[8)]．特徴抽出に MLP ではなく Conditional Random Fields(CRF) を使用する試みもある[9)]．文献 10) では，MLP が出力する音素事後確率特徴量を HMM ではなく，さらにもう一度別の MLP の入力とする方法を提案している．音素事後確率特徴量は，MFCC や PLP などの音響特徴量と比べて，コンテキストや話者性，雑音などに頑健であることや，長時間の音素事後確率特徴の変化を見ることで誤りの生じやすい音韻的なパターンを検出できることなどから注目を集めている[10)].

本稿では，文献 10) で用いられた MLP のかわりに提案法である HCNF を用いて，1 段目の HCNF の出力を 2 段目の HCNF への入力とする手法を提案する．本手法では，2 段目の HCNF の入力として，1 段目の HCNF による各状態の各フレームにおける事後確率を用いる．本稿ではこれを階層的音素事後確率特徴量と呼ぶ．ある入力系列 X に対する状態 s のフレーム t における階層的音素事後確率特徴量 $\gamma(s,t|X)$ は以下のように定義される．

$$\gamma(s,t|X) = \sum_{Y'} \sum_{S'} P(Y',S'|X)\delta(s,s'_t) \tag{17}$$

階層的音素事後確率特徴量を用いた音声認識では，式 (17) に基づいて各フレーム毎に各

状態の階層的音素事後確率特徴量を並べたベクトルを作成し，2 段目の HCNF への入力とすることで最終的な認識結果を得る．

5. 実　　験

5.1 実験条件

TIMIT コーパスおよび ASJ+JNAS コーパスを用いた連続音素認識タスクによって提案法の評価を行う． TIMIT コーパスおよび ASJ+JNAS コーパスの文数と話者数をそれぞれ表 1，表 2 に示す．TIMIT コーパスを用いた実験では，文献[14] に従い，TIMIT コーパスで定義される 61 音素から，学習時には 48 音素，評価時にはさらにそこから 39 音素にマッピングして評価を行った．ASJ+JNAS コーパスを用いた実験では，学習，評価時ともに 43 音素を用いた．HCNF は，各音素について 3 状態を持つ left-to-right 型の monophone モデルを使用した．音響特徴量として MFCC13 次元を抽出し[⋆1]その Δ と ΔΔ を計算した (ASJ+JNAS の実験では MFCC の 0 次の項の代わりにパワーを使用した)．観測特徴としては，以下に示す特徴量を使用した．

$$\phi_s^{M1}(X,y,s) = \sum_{t=1}^{T} \delta(s_t = s) x_t \tag{18}$$

$$\phi_s^{M2}(X,y,s) = \sum_{t=1}^{T} \delta(s_t = s) x_t^2 \tag{19}$$

$$\phi_s^{Occ}(X,y,s) = \sum_{t=1}^{T} \delta(s_t = s) \tag{20}$$

$$\phi_y^{Uni}(X,y,s) = \sum_{t=1}^{T} \delta(y_t = y) \tag{21}$$

Δ特徴量を使用する場合，M1 および M2 の次元数は各状態あたり 39 次元となり，Δ特徴量を使用しない場合は 13 次元となる．遷移特徴としては，以下に示す現在の状態と直前の状態とのペアで活性化する状態のバイグラム特徴 (Tr) およびラベルのバイグラム特徴 (Bi) を使用した[⋆2]．

$$\psi_{ss'}^{Tr}(X,y,y',s,s') = \sum_{t=1}^{T} \delta(s_t = s)\delta(s_{t-1} = s') \tag{22}$$

$$\psi_{yy'}^{Bi}(X,y,y',s,s') = \sum_{t=1}^{T} \delta(y_t = y)\delta(y_{t-1} = y') \tag{23}$$

M1 と M2 の特徴量は，学習データ全体でフレーム毎の平均が 0，分散が 1 となるように正規化し，該当フレームに加えて前後 4 フレーム (計 9 フレーム) 結合して使用した．階層的音素事後確率特徴量を使用する場合には，前後 4 フレーム (計 9 フレーム) に加えて，前後 11(計 23 フレーム) 結合する場合の実験も行った．HCNF のパラメータは-0.5 から 0.5 の間でランダムに初期化した．Δ特徴量を使用する場合のゲート数は $K = 4$，使用しない場合のゲート数は $K = 16$ とした．階層的音素事後確率特徴量を用いる場合，2 段目の HCNF のゲート数は全て $K = 4$ とした．state-flattening 係数 κ は全て 0.1 を使用した．式 (5) のパラメータは $\alpha = 0.1$，$\beta = 0.0$，$b = 3.0$，$c = 6.0$ を使用した．階層的音素事後確率特徴量を用いる場合は $\alpha = 1.0$ とした．学習は全て SGD で行い，FOBOS[15] を用いて L2 正則化を行った．正則化パラメータ C[6] は全て 1.0 を使用した．TIMIT では 30 回繰り返し学習し，ASJ+JNAS では 15 回繰り返し学習した．階層的音素事後確率特徴を用いる場合は，2 段目の HCNF は 10 回繰り返し学習を行った．

表 1 TIMIT コーパスの文数と話者数

データ	文数	話者数	音素数
学習	3696	462	140225
テスト	192	24	7215

表 2 ASJ+JNAS コーパスの文数と話者数

データ	文数	話者数	音素数
学習	20337	133	1269999
テスト (IPA100 文)	100	23	6021

5.2 目的関数の検討の実験結果

5.2.1 TIMIT コーパスでの実験結果

TIMIT コーパス上で，Δ特徴量を使用して HCNF を HB-MMI 学習する場合に，式 (13) における b を変化させた場合の音素認識結果を表 3 に示す．b はどの程度マージンを強要するかを示す項であり，大きいほど学習エラーに対して厳しくなる．表 3 で $b = 0$ の場合はマージンを強要しない，すなわち，式 (7) の目的関数を使用する場合であり，事後確率最大化学習 (MMI) に対応する．表 3 から，Δ特徴量を使用する場合には HB-MMI を使用しても認識率の向上は得られていないことがわかる．表の “学習” は，学習データに対するエラー率を表す (学習エラー集計時の音素の種類数は 48 である)．b を大きくすることで学習エラーが減少していることから，HB-MMI が学習エラーを考慮した目的関数となっていることがわかる．評価データに対するエラーは減少していないため，いわゆる過学習が起きていると考えられる．これらの結果は，ほぼ同条件で HMM を MPE 学習した場合と遜色ない認識結果である[6]．

表 4 にΔ特徴量を使用せずに HB-MMI 学習した場合の認識結果を示す．Δ特徴量を使用

⋆1 サンプリング周波数=16kHz，プリエンファシス=0.97，分析窓長=25ms，フレームシフト=10ms
⋆2 音素認識の場合には，Bi は音素バイグラム相当の特徴量である

表 3 TIMIT コーパスでΔ特徴量を使用して HB-MMI 学習した場合の音素認識結果 [%]

b	Del	Ins	Subs	PER	学習
0.0	7.7	2.1	18.2	28.0	16.1
1.0	7.1	2.5	18.9	28.5	13.8
3.0	6.6	3.1	19.5	29.2	11.9
5.0	6.0	3.5	19.7	29.1	11.4
HMM(MPE,diag,32mix)[6]				28.4	18.4

表 4 TIMIT コーパスでΔ特徴量を使用せずに HB-MMI 学習した場合の音素認識結果 [%]

b	Del	Ins	Subs	PER	学習
0.0	11.9	1.2	17.7	30.9	28.7
1.0	10.3	1.5	17.6	29.4	26.8
3.0	8.7	1.9	17.3	27.9	24.8
5.0	8.3	2.5	18.1	28.8	23.9
10.0	7.6	2.9	18.5	29.1	23.9

する場合とは異なって，HB-MMI 学習を行うことで認識率の改善が得られた．通常の MMI 学習を行った場合の認識率が PER=30.9%であったのに対し，認識率が最大となった $b = 3.0$ の場合には絶対値で 3.0%改善し 27.9%となった．この値は表 3 における $b = 0.0$ の値とほぼ同等であり，Δ特徴量を使用しない場合でもΔ特徴量を使用する場合と同等の認識率が得られている．これより，HCNF のゲート関数によって音素認識においてΔ特徴量相当の特徴を捉えることができたといえる．表 4 の $b = 3.0$ の場合と表 3 の $b = 0.0$ の場合を比較すると，表 4 の $b = 3.0$ の場合の方が約 1.5 倍学習エラーが高いことがわかる．学習エラーが高いにもかかわらず評価データに対する認識率はほぼ同等であることから，Δ特徴量を使用しない方がより一般的で頑健なモデルを学習できている可能性が高い．Δ特徴を使用する場合のパラメータ数は 409968，使用しない場合のパラメータ数は 546480 であり，Δ特徴量を使用しない場合の方がパラメータ数が多い．このことから，Δ特徴量を使用しない場合には，パラメータ数が同等以上の場合でも，通常の MMI 学習では HCNF で識別的な特徴を捉えることが困難であったと考えられる． HB-MMI によって学習エラーが減少するように，すなわち識別的な特徴が HCNF によって捉えられるように学習を行うことで，Δ特徴量を使用しない場合に認識率が向上したと考えられる．

5.2.2 ASJ+JNAS コーパスでの実験結果

TIMIT コーパスを用いた認識実験では，Δ特徴量を使用する場合に HB-MMI による認識率の向上が見られなかったが，学習データに対するエラー率は減少していたことから，いわゆる過学習が生じた考えられる．学習データサイズと HB-MMI 学習との関係を調べるため，TIMIT コーパスよりも学習データが多い ASJ+JNAS コーパスを用いた認識実験を行った．表 5 に ASJ+JNAS コーパスで HB-MMI 学習を行った場合の認識結果を示す．表 5 より，学習データの増えた ASJ+JNAS コーパスでは，HB-MMI 学習を行うことで認識率が改善したことがわかる．このことから，学習データが多い場合には HB-MMI は効果的であり，今後大規模化を行っていく上で有効であるといえる．

比較のために，HMM を MLE，MMI，MPE 基準で学習した場合の認識結果を表 5 下部に示す．HMM は文献 6) と同様の条件で学習した．HB-MMI を用いることで，HCNF が MMI および MPE 学習した HMM を上回っている．表 3 と表 5 を比較すると，学習データが増えることで HB-MMI 学習が効果的となり，HCNF の認識結果が優位に HMM を上回るようになると推察できる．

表 5 ASJ+JNAS コーパスで目的関数に HB-MMI 学習を行った場合の実験結果 [%]

b	Del	Ins	Subs	PER	学習
0.0	6.2	0.8	8.7	15.7	14.3
1.0	5.0	1.0	8.6	14.6	13.0
3.0	4.5	1.3	8.1	13.9	12.8
HMM(MLE,diag,32mix)	6.4	1.2	11.4	19.0	18.7
HMM(MMI,diag,32mix)	5.0	1.2	9.5	15.8	15.6
HMM(MPE,diag,32mix)	5.5	0.9	8.6	15.0	13.2

5.3 階層的音素事後確率特徴量の検討の実験結果

5.2.1 節の TIMIT コーパス上での認識結果を元に，階層的音素事後確率特徴量を使用した認識実験を行った．1 段目の HCNF として，Δを用いる表 3 の $b = 0.0$ および，Δを用いない表 4 の $b = 3.0$ のモデルを使用した．2 段目の HCNF は，文献 10) を参考に，窓幅 9 と 23 の場合を検討した．1 段目にΔを用いる場合の認識結果を表 6，Δを用いない場合の認識結果を表 7 に示す．表 6，7 より，1 段目の HCNF でΔを使用するかどうかにかかわらず，階層的音素事後確率特徴量を使用することで認識率が改善していることがわかる．また，窓長 9 と 23 の場合を比べると，窓長 23 の場合の方が認識率が高い．この結果は文献 10) の結果と一致している．1 段目にΔを使用した場合と使用しない場合の結果を比べると，注目すべきことに，Δを使用しない場合の方が総じて認識率が高い．特に，1 段目にΔを使用した場合には，HB-MMI を使用しても認識率の向上が見られなかったが，1 段目にΔを使用しない場合には，HB-MMI を使用することでさらなる認識率の向上が得られた．1 段目にΔを使用せず，窓長 23，$b = 5.0$ の場合に最高の認識率を達成し，PER=25.3%を得た．これは，階層的音素事後確率特徴量を使用しない場合である PER=27.9%と比べて絶対値で 2.6%良い値である．これらの結果から，階層的音素事後確率特徴量は認識率を改善するうえで有効な特徴量であるといえ，1 段目の HCNF としてはできるだけ一般的で頑健なモデルを使用するのが良いと考えられる．

6. おわりに

本稿では，正解状態系列が未知の場合でも正解状態系列に対するエラー数を考慮した学習を行うことができる HB-MMI を用いた HCNF の学習手法および，音素事後確率特徴量を使用した音声認識手法について述べた．本稿で提案した HB-MMI 学習を用いること

表 6 TIMIT コーパスで音素事後確率特徴量を用いた場合の認識結果 (1 段目にデルタを使用) [%]

窓長	b	Del	Ins	Subs	PER	学習
9(1 段目)	0.0	7.7	2.1	18.2	28.0	16.1
9(2 段目)	0.0	6.6	2.5	17.4	26.6	13.4
9(2 段目)	3.0	5.7	2.9	18.2	26.9	12.6
23(2 段目)	0.0	6.4	2.5	17.3	26.2	10.9
23(2 段目)	3.0	5.6	2.9	18.4	26.9	9.4

表 7 TIMIT コーパスで音素事後確率特徴量を用いた場合の認識結果 (1 段目にデルタを不使用) [%]

窓長	b	Del	Ins	Subs	PER	学習
9(1 段目)	3.0	8.7	1.9	17.3	27.9	24.8
9(2 段目)	0.0	8.8	1.7	15.9	26.4	22.5
9(2 段目)	3.0	7.9	2.0	15.8	25.7	21.8
23(2 段目)	0.0	8.3	1.8	16.0	26.1	18.6
23(2 段目)	3.0	7.3	2.2	16.0	25.5	17.0
23(2 段目)	5.0	7.0	2.3	16.0	25.3	16.6
23(2 段目)	10.0	6.9	2.4	16.6	25.9	16.6

で，TIMIT コーパス上でΔ特徴量を使用しない場合および，TIMIT コーパスよりも大きな ASJ+JNAS コーパスを用いた場合には，認識率の改善を得ることができた．HB-MMI は，大規模なコーパスを用いる場合に効果が高いことが予想されるため，今後大規模化を行っていく上で効果的であると考えられる．音素事後確率特徴量を用いた音声認識手法については，音素事後確率特徴量を用いることで認識率の改善を得ることができ，音素事後確率特徴量が認識率を改善するために有効であることがわかった．音素事後確率特徴量を用いる場合に，HB-MMI 学習を行った場合の認識率が最も良く，HB-MMI 学習は音素事後確率特徴量を使用する場合にも効果的であった．

今後の課題としては，CSJ などのより大規模なコーパスでの実験，コンテキスト依存性の導入などがあげられる．また，今後大規模化を行っていく上では，計算処理の並列化が必要になると考えられる．

謝　　辞

本研究は文部科学省グローバル COE プログラム「インテリジェントセンシングのフロンティア」の支援を受けた。

参 考 文 献

1) Bilmes, J.A.: What HMMs Can Do, *IEICE TRANS. INF. & SYST.*, Vol.E89-D, No.3, pp.1–24 (2006).
2) Furui, S.: Speaker-independent isolated word recognition using dynamic features of speech spectrum, *IEEE Transactions of Acoustics Speech and Signal Processing*, Vol.34, No.1, pp.52 – 59 (1986).
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，電子情報通信学会論文誌， Vol.J79-D-II, No.12, pp.2032–2038 (1996).
4) Kanedera, N., Arai, T., Hermansky, H. and Pavel, M.: On the Relative Importance of Various Components of the Modulation Spectrum for Automatic Speech Recognition, *Speech Communication*, Vol.28, pp.43–55 (1999).
5) Povey, D.: Discriminative Training for Large Vocabulary Speech Recognition, PhD Thesis, Cambridge University Engineering Dept (2003).
6) 藤井， 山本， 中川：Hidden Conditional Neural Fields を用いた音声認識の検討，Proc. 情報処理学会研究報告，SLP-83-1 (2010).
7) Povey, D., Kanevsky, D. and Kingsbury, B.: Boosted MMI for Model and Feature-Space Discriminative Training, *Proc. ICASSP*, pp.4058 – 4061 (2008).
8) Hermansky, H., Ellis, D. and Sharma, S.: Tandem connectionist feature stream extraction for conventional HMM systems, *Proc. ICASSP* (2000).
9) Fosler.-L., E. and Morris, J.: Crandem systems: Conditional Random Field Acoustic Models for Hidden Markov Models, *Proc. ICASSP*, pp.4049–4052 (2008).
10) Pinto, J., Garimella, S., M.-Doss, M., Hermansky, H. and Bourland, H.: Analysis of MLP-Based Hierarchical Phoneme Posterior Probability Estimatior, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.19, No.2, pp.225–241 (2011).
11) Mahajan, M., Gunawardana, A. and Acero, A.: Training algorithms for hidden conditional random fields, *Proc. ICASSP*, pp.I–273–I–276 (2006).
12) Sung, Y.-H. and Jurafsky, D.: Hidden Conditional Random Fields for Phone Recognition, *Proc. ASRU*, pp.107 – 112 (2009).
13) Heigold, G., Deselaers, T., Schlüter, R. and Ney, H.: Modified MMI/MPE: A Direct Evaluation of the Margin in Speech Recognition, *Proc. ICML* (2008).
14) Lee, K.-F. and HON, H.-W.: Speaker-Independent Phone Recognition Using Hidden Markov Models, *IEEE Transactions of Acoustics Speech and Signal Processing*, Vol.37, No.11, pp.1641 – 1648 (1989).
15) Duchi, J. and Singer, Y.: Efficient Learning using Forward-Backward Splitting, *Proc. NIPS* (2009).