情報処理学会論文誌 Vol.54 No.2 436–442 (Feb. 2013)

識別学習を用いた離散混合分布HMMによる音声認識

小坂 哲夫[1,a)] 加藤 正治[1,b)]

受付日 2012年5月22日, 採録日 2012年11月2日

概要：これまで我々は離散混合分布 HMM による音声認識の検討を行ってきた．離散混合分布 HMM では混合分布やサブベクトル量子化，尤度補償などを行うことにより一般的な離散分布 HMM と比較し高い性能を得ることが可能となっている．認識実験の結果雑音音声認識に対して有効であること，および講演音声認識において連続分布 HMM を用いたシステムとほぼ同等の性能が得られることを示してきた．本研究では，さらなる性能向上を目指し，離散混合分布の識別学習の有効性を検討する．識別学習としては最大相互情報量（MMI）に基づく手法を用い，日本語話し言葉コーパス（CSJ）で評価を行った．その結果従来法である最尤推定法（ML）や最大事後確率推定法（MAP）と比較し高い性能が得られることが分かった．

キーワード：音声認識，隠れマルコフモデル，識別学習，離散確率分布

Speech Recognition by Using Discrete-Mixture HMMs Based on Discriminative Training

Tetsuo Kosaka[1,a)] Masaharu Kato[1,b)]

Received: May 22, 2012, Accepted: November 2, 2012

Abstract: Previously, we had investigated speech recognition by using discrete-mixture hidden Markov models (DMHMMs). The DMHMM yields a higher recognition performance than conventional discrete HMM because it uses mixture-density, subvector quantization, likelihood compensation, etc. From the results obtained through speech recognition experiments, the DMHMM-based system showed better performance under noise conditions. Moreover, it showed similar performance to the continuous-mixture HMM-based system in lecture speech recognition. In this paper, we investigate a discriminative learning approach for DMHMMs to further improve their recognition performance. We use the maximum mutual information (MMI) estimation criterion as discriminative learning. The proposed method was evaluated on a large-scale spontaneous speech database, "Corpus of Spontaneous Japanese". From the results, the MMI estimation showed better performance than conventional approaches such as maximum likelihood estimation or maximum *a posteriori* estimation.

Keywords: speech recognition, hidden Markov model, discriminative training, discrete probability distribution

1. はじめに

近年の音声コーパスの整備により，数百時間から数千時間規模の大量の音声データを音響モデルの学習に使用することが可能となった [1]．日本においても日本語話し言葉コーパス（Corpus of Spontaneous Japanese: CSJ）の整備により数百時間規模のデータが使用可能となっている [2]．音声認識はこの大量の音声データを利用した統計的な音響モデルに基づき行われている．近年は連続分布型隠れマルコフモデル（Hidden Markov Model: HMM）によるモデリングが主流となっている．

学習データの量とモデルの構造やパラメータは密接な関係がある．一般的には，学習データが少ない場合は，パラメータ数の少ないモデルが，多い場合はパラメータ数の多いモデルが適している．このため，1980 年代に統計的モデルが音声認識に積極的に使われるようになってから，学習

[1] 山形大学大学院理工学研究科
Graduate School of Science and Engineering, Yamagata University, Yonezawa, Yamagata 992–8510, Japan
[a)] tkosaka@yz.yamagata-u.ac.jp
[b)] katoh@yz.yamagata-u.ac.jp

データ量の規模の増大とともに，よりパラメータ数の多い複雑なモデルが使用されるようになってきている．たとえば従来効果が少ないと考えられてきた連続分布 HMM の全共分散パラメータや quinphone などについても，近年効果が得られることが分かっている（たとえば文献 [3], [4]）．またデータ量の増大は学習の方法とも関連がある．HMM の学習は最尤推定や識別学習が用いられるが，以前はデータ量の不足により識別学習は十分な成果が得られていなかった [5]．音声認識において識別学習が効果を発揮するためには 100 時間超の音声データが必要といわれているが [6]，近年ではそれだけのデータ量も揃い識別学習も利用が可能となってきた．

以上のように学習データの増大により従来十分には性能が得られなかった手法の見直しが進んでいる．離散分布 HMM についても同様に連続分布 HMM と比較して性能が低いと考えられてきたが，これまでの検討の結果，離散混合分布 HMM（DMHMM）[7] が雑音音声認識に対して有効であること [8]，および講演音声認識において混合連続分布 HMM（CMHMM）を用いたシステムとほぼ同等の性能が得られること [9] が分かった．

文献 [9] では，CMHMM および DMHMM それぞれについて学習には最尤（ML）推定および最大事後確率（MAP）推定を用いている．しかし近年 CMHMM においては種々の識別学習が検討され性能向上が得られている．特に Minimum Classification Error（MCE）および Maximum Mutual Information（MMI）に基づく方法が有効であることが確認されている [10], [11]．一方 DMHMM に関しては ML 推定以外では MAP 推定については検討を行っているが [8]，識別学習については未検討であった．

そこで本研究では DMHMM のさらなる性能向上を目指し，DMHMM の識別学習を提案しその有効性を検討する．CMHMM においては MCE や MMI の派生形も種々提案され成功を収めているが（たとえば MPE/MWE [11], Boosted MMI [12] など），本研究ではまずベースとなる MMI 推定について，DMHMM のパラメータ推定法として有効であるかの検討を行った．識別学習は大規模なデータを用いた場合効果が発揮されるため，提案法の有効性検証のための認識実験は CSJ で行い，従来法である ML 推定および MAP 推定との比較を行った．

2. 離散混合分布 HMM の識別学習

2.1 離散混合分布 HMM

CMHMM では正規分布を仮定することにより，効率良くパラメータの推定を行うことができる．このため現在音声認識のモデルとしては主流の方法となっている．一方離散分布型 HMM の場合，量子化サイズを小さくすると量子化歪みが大きくなり，逆にサイズを大きくすると学習データが不足し，十分にパラメータ推定ができないという問題がある．一方 DMHMM では，スカラ量子化 [7] あるいはサブベクトル量子化 [13] の利用により量子化サイズを削減しパラメータ推定の問題に対処している．これまでの検討結果からサブベクトル量子化のほうが性能が高いことが分かっているため [13]，本研究では両者のうちサブベクトル量子化を利用する．以下にその手法を示す．

(1) 入力ベクトル $\mathbf{o}_t$ を S 個のサブベクトルに分割する．このとき，特徴量の隣接するものをひとまとめにする．

$$\mathbf{o}_t = [\mathbf{o}_{1t}, \cdots, \mathbf{o}_{st}, \cdots, \mathbf{o}_{St}] \tag{1}$$

(2) コードブック作成用データを用いて，各サブベクトルごとのコードブックを作成する．サブベクトル s において，入力 $\mathbf{o}$ は以下のように量子化される．

$$q(\mathbf{o}_t) = [q_1(\mathbf{o}_{1t}), \cdots, q_s(\mathbf{o}_{st}), \cdots, q_S(\mathbf{o}_{St})] \tag{2}$$

このとき，DMHMM の出力 $b_i(\mathbf{o}_t)$ は次のように求められる．

$$b_i(\mathbf{o}_t) = \sum_m w_{im} \prod_s \theta_{sim}(q_s(\mathbf{o}_{st})) \tag{3}$$

ここで θ_{sim} はサブベクトル s，状態 i，混合要素 m における離散確率，w_{im} は混合分布の重み係数である．上式は，混合内で異なるサブベクトル間の離散確率は互いに独立だが，状態内の異なるサブベクトル間の従属性は，混合要素でモデル化されるという仮定に基づく．よって状態内でサブベクトル間の相関がないならば混合要素に分ける必要がなくなるが，相関があれば，混合要素に分けた場合認識性能の向上が期待できる．

2.2 離散分布パラメータの推定

本研究ではパラメータ推定として DMHMM の最大相互情報量基準（MMI）による推定法を提案する．本節ではこの MMI 推定，および比較として用いる最尤推定（ML）と最大事後確率推定（MAP）について説明する．

2.2.1 最大相互情報量基準による推定

離散混合分布 HMM のサブベクトル s，状態 i，混合 m における k 番目の離散確率値 $\theta_{sim}^{MMI}(k)$ は以下のように求められる．

$$\theta_{sim}^{MMI}(k) = \frac{\theta_{sim}(k)(\frac{\partial \mathcal{F}}{\partial \theta_{sim}(k)} + E)}{\sum_{k'} \theta_{sim}(k')(\frac{\partial \mathcal{F}}{\partial \theta_{sim}(k')} + E)} \tag{4}$$

ここで

$$\frac{\partial \mathcal{F}}{\partial \theta_{sim}(k)} = \frac{1}{\theta_{sim}(k)}(\gamma_{simk} - \gamma_{simk}^{gen}) \tag{5}$$

$$\gamma_{simk} = \sum_{t=1}^{T} \gamma_{im}(t)\delta(q_s(\mathbf{o}_{st}), k) \tag{6}$$

$$\delta(q_s(\mathbf{o}_{st}), k) = \begin{cases} 1 & q_s(\mathbf{o}_{st}) = k \\ 0 & \text{otherwise} \end{cases}$$

$\gamma_{im}(t)$ は時刻 t で状態 i，混合要素 m に存在する確率である．識別学習では正解だけでなく誤認識結果を対立候補として使用する．γ_{simk} は正解データより求めた EM カウントである．また γ^{gen}_{simk} は一般モデルから得られるカウントである．これはすべての起こりうる音素列に対応するモデル $\mathcal{M}_{gen}$ を使って計算するが，実際は認識に使用するモデルを $\mathcal{M}_{gen}$ の近似として用いる．E は収束の速さを制御する定数であり，これを小さくすると収束は速くなるが学習は不安定となる．逆に E を大きくすると学習が進まない．式 (5) は小さい値のパラメータにきわめて敏感に反応する．そこで Merialdo の提案した近似法 [5] に従い，以下のように求める．

$$\frac{\partial \mathcal{F}}{\partial \theta_{sim}(k)} \approx \frac{\gamma_{simk}}{\sum_{k'} \gamma_{simk'}} - \frac{\gamma^{gen}_{simk}}{\sum_{k'} \gamma^{gen}_{simk'}} \tag{7}$$

式 (4) において E の値が小さくなるほど収束は速くなるが，収束が保証されなくなり不安定となる．この値の設定に関しては Gopalakrishnan らが提案した以下の方法 [14] を用いる．

$$E = \max_{\theta_{sim}(k)} \left\{ -\frac{\partial \mathcal{F}}{\partial \theta_{sim}(k)}, 0 \right\} + \epsilon_e \tag{8}$$

ここで ϵ_e は小さな正の定数であり，本研究では様々な値で検討を行う．E については，すべてのモデルで共通の値を用いる方法のほかに，モデルや状態，離散分布ごとに別の値を設定するという方法も考えられる．たとえば音素 p ごとに設定する場合，以下のように音素ごとの最大値を計算する．

$$E(p) = \max_{\theta_{sim}(k) \in p} \left\{ -\frac{\partial \mathcal{F}}{\partial \theta_{sim}(k)}, 0 \right\} + \epsilon_e \tag{9}$$

本研究では，すべてのモデルで共通に設定した場合と，モデルごとに設定する場合の 2 種類の検討を行う．

混合重みの推定も，式 (4) と同様に求めることができる．状態 i，混合 m の混合重み w_{im} の再推定式は以下で与えられる．

$$w^{MMI}_{im} = \frac{w_{im} \left\{ \frac{\partial \mathcal{F}}{\partial w_{im}} + C \right\}}{\sum_{m'} w_{im'} \left\{ \frac{\partial \mathcal{F}}{\partial w_{im'}} + C \right\}} \tag{10}$$

また定数 C については式 (8) と同様以下で与えられる．

$$C = \max_{w_{im}} \left\{ -\frac{\partial \mathcal{F}}{\partial w_{im}}, 0 \right\} + \epsilon_c \tag{11}$$

この定数についても，すべての音素で共通な場合と音素ごと別に設定する場合の 2 種類について検討を行う．

2.2.2 最尤推定と MAP 推定

本項では DMHMM の最尤推定（ML）と最大事後確率推定（MAP）について説明する．通常の最尤推定では事前分布の影響を無視するが，MAP 推定では事前分布も考慮に入れたパラメータ推定を行う．離散出力確率の ML 推定値は以下のように求められる．

$$\theta^{ML}_{sim}(k) = \frac{\gamma_{simk}}{\sum_{k'} \gamma_{simk'}} \tag{12}$$

また離散出力確率の MAP 推定値 $\theta^{MAP}_{sim}(k)$ は事前分布をディレクレ分布とした場合，

$$\theta^{MAP}_{sim}(k) = \frac{(\nu_{simk} - 1) + n_{im} \cdot \theta^{ML}_{sim}(k)}{\sum_{k'} (\nu_{simk'} - 1) + n_{im}} \tag{13}$$

$$n_{im} = \sum_{k'} \gamma_{simk'} \tag{14}$$

で求められる．ここで ν_{simk} は事前分布パラメータである．

$$\nu_{simk} = \tau_M \cdot \theta_{sim}(k)^0 + 1 \tag{15}$$

と仮定すると，

$$\theta^{MAP}_{sim}(k) = \frac{\tau_M \cdot \theta_{sim}(k)^0 + n_{im} \cdot \theta^{ML}_{sim}(k)}{\tau_M + n_{im}} \tag{16}$$

となる．ここで τ_M は事前知識の確からしさに関係する係数であり，今回は開発セットを用いて実験的に求めた．DMHMM に関しては出力分布の平均値だけでなく，混合係数，状態遷移確率についても MAP 推定が可能であるが，これらのパラメータについては ML 推定により求めた．

2.3 事前分布

事前分布をどのように定めるかは MAP 推定における重要な点となるが，本研究では文献 [8] と同様に，まず CMHMM のパラメータ推定を行いそれを DMHMM のパラメータに変換することにより与えた．CMHMM におけるサブベクトル s，状態 i の m 番目の混合分布は以下のとおり与えられる．

$$\begin{aligned} b'_{sim}(\boldsymbol{o}_{st}) = & \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}_{sim}|^{\frac{1}{2}}} \\ & \cdot \exp\left[-\frac{1}{2} (\boldsymbol{o}_{st} - \boldsymbol{\mu}_{sim})^t \boldsymbol{\Sigma}^{-1}_{sim} (\boldsymbol{o}_{st} - \boldsymbol{\mu}_{sim}) \right] \end{aligned} \tag{17}$$

ここで $\boldsymbol{\mu}_{sim}$ は平均ベクトル，$\boldsymbol{\Sigma}_{sim}$ は共分散行列，d はサブベクトル s の次元数である．このとき事前分布を以下の式で求める．

$$\theta^0_{sim}(k) = \frac{b'_{sim}(\boldsymbol{v}_s(k))}{\sum_{k'} b'_{sim}(\boldsymbol{v}_s(k'))} \tag{18}$$

ここで $\boldsymbol{v}_s(k)$ はサブベクトル s のセントロイドを示す．この $\theta^0_{sim}(k)$ は正規分布の制約下で求められるが式 (12) の $\theta^{ML}_{sim}(k)$ や式 (16) の $\theta^{MAP}_{sim}(k)$ はそのような制約を受けない．このため離散分布パラメータの学習により，より複雑な分布形状が表現できる可能性がある．以上求めた分布は MAP 推定の事前分布のほか，DMHMM の各種パラメータ推定の初期モデルとしても使用する．

2.4 音素グラフを利用した識別学習

MMI により学習を行う場合対立候補をどのように与えるかが問題となるが，計算量削減のため，Valtchev らが提案した音素グラフによる方法を用いる [15]．これは学習データを用いて認識を行い音素グラフを生成し，グラフ内の誤りを含む音素系列を利用する方法である．生成された音素グラフに含まれる誤りに対立候補を限定し，それをグラフ表現することで，処理量の削減が可能となる．音素グラフを用いた MMI 学習アルゴリズムの概要は以下のとおりである．

(1) 学習データの各発話に対して認識を行い音素グラフを生成する．
(2) 上記で生成された音素グラフは正解が含まれていることは保障されないため，正解音素列を付加する．
(3) 上記発話ごとの音素グラフを用い，グラフの各エッジにおいて時刻 t における存在確率を求める．
(4) 上記存在確率を正解および誤りについてそれぞれ，すべてのエッジおよび学習データで加算することにより，γ_{simk} および γ_{simk}^{gen} を求める．

2.5 DMHMM の尤度補償

DMHMM において，式 (3) の $\theta_{sim}(q_s(\mathbf{o}_{st}))$ のいずれかのサブベクトルの確率が 0 またはそれに近い値になると，出力確率も 0 にきわめて近い値になる．この場合対数計算を行うと，わずかの入力値の違いが大きな尤度差となって表れる．これは認識性能に悪影響を与えるため，離散確率に一律に閾値を設け，確率が閾値を下回った場合，閾値に置き換えるフロアリング処理が効果的である．この方法は特に雑音下の音声認識ではきわめて効果的であり，雑音の種類により多少の違いはあるが，出力確率の 7 割程度をフロアリングすることにより性能向上が得られることが分かっている [8]．またクリーン音声でも多少の効果があることが分かっているため，本実験でもこの手法を用いた．方法としては以下の式のとおり，すべての出力確率に対し一定の閾値 dth を与える．

$$\theta'_{sim}(q_s(\mathbf{o}_{st})) = \begin{cases} \theta_{sim}(q_s(\mathbf{o}_{st})) & \text{if } \theta_{sim}(q_s(\mathbf{o}_{st})) \geq dth \\ dth & \text{else} \end{cases} \tag{19}$$

2.6 I-smoothing

CMHMM に関して ML 推定と識別学習の補間の目的で I-smoothing [11] と呼ばれる手法が提案されている．本研究では，この手法と同様，両者のスムージングを目的として以下の手法を検討する．スムージングは式 (4) を以下のように変形して行う．

$$\theta_{sim}^{Ism}(k) = \frac{\theta_{sim}(k)(\frac{\partial \mathcal{F}}{\partial \theta_{sim}(k)} + E) + \tau_I \cdot \theta_{sim}^{ML}(k)}{\sum_{k'} \theta_{sim}(k')(\frac{\partial \mathcal{F}}{\partial \theta_{sim}(k')} + E) + \tau_I} \tag{20}$$

τ_I はスムージングの度合いを決める係数であり，この値が 0 の場合識別学習のみを行った場合と等しく，$\tau_I \to \infty$ の場合最尤推定と等しくなる．

3. 認識実験

3.1 概要

本研究で提案した DMHMM の識別学習の有効性を，講演音声認識により検討した．検討にあたっては MMI 推定，ML 推定, MAP 推定の 3 種類の比較を行った．また MMI 推定に関して，学習におけるパラメータ C や E を音素モデル共通で設定する方法と音素ごとに設定する方法について比較検討した．さらに I-smoothing の効果について検討した．

3.2 実験条件

データベースとして日本語話し言葉コーパス（CSJ）を用いる．音声分析のフレーム長/周期は 25 ms/8 ms，特徴ベクトルは 1〜12 次の MFCC と対数パワー，およびその 1 次と 2 次の回帰係数の計 39 次元とし，発話ごとのケプストラム平均正規化を行った．音響モデルについては，3,000 状態 16 混合の CMHMM を 5 回 ML 推定で学習した後，式 (18) により DMHMM に変換し，さらに MAP 推定で 2 回学習したモデルを今回の実験の初期モデルとした．CMHMM における ML 推定は正規分布の拘束下での学習となるが，DMHMM 変換後はそのような拘束がない学習となる．DMHMM のサブベクトル量子化におけるコードブックサイズは，対数パワーについて 64，また MFCC の 1 次，2 次の 2 次元サブベクトルに対し 64，以下同様にして 2 次元ごとのサブベクトルに対し 64 と設定した．さらに 1 次，2 次の回帰係数についても同様に設定した．このため全体としてはサイズが 64 のコードブックを 21 個設定した．学習データとしては CSJ の男性話者と女性話者の学会講演，計 963 講演 203 時間を用いた．この初期モデルを基に MMI 推定・MAP 推定・ML 推定の 3 種類の方法によりさらに複数回学習を行い．学習回数ごとに認識実験を行って性能を比較した．

MMI 推定における対立候補用の音素グラフは初期モデルを用い学習セットから得られたものを使用する．本来は 1 回の学習ごとに音素誤り傾向は若干異なるので，音素グラフを作り直す必要があるが，予備検討の結果数回程度の学習の場合その効果は少ないことが分かったため，今回は初期モデルで作成したものをその後の繰返し学習でもそのまま使用する．

MMI 推定における尤度閾値 dth はこれまでの検討結果より，1.0×10^{-5} と定めた．また MMI 推定における ϵ_c や ϵ_e および MAP 推定における τ のパラメータの設定には開発セットを用いた．開発セットとしては，学習セットおよび評価セットとは異なる学会講演男性話者 5 講演を用いる．また評価には CSJ のテストセット 1，学会講演男性話

者 10 講演およびテストセット 3，模擬講演 10 講演（男女各 5 講演）を用いる．

認識用デコーダは第 1 パスで bigram・triphone，第 2 パスで trigram を用いる 2 パスデコーダを用いる [16], [17]．言語モデルの学習テキストは，CSJ の男性＋女性話者の学会講演＋模擬講演の 2,668 講演である．総単語数約 686.3 万のテキストから作成した．語彙エントリ数は学会講演 2 回以上，模擬講演 2 回以上の計 47,099 語である．

3.3 実験結果

まず MMI の ϵ_c や ϵ_e の最適な値についての検討を開発セットを用いて行った．まず上記初期モデルを用い，ϵ_c および ϵ_e について 40 種の組合せについて 1 回学習を行う．これにより得られた 40 個のモデルを開発セットを用いて評価し，最高の性能を示すモデルを選択し 2 回目の学習を行う．これを繰り返すことにより学習を進める．

開発セットにおける学習回数ごとの単語誤り率（WER）および得られたパラメータの値を**表 1**，**表 2** に示す．表 1 は C や E をモデル共通とした場合，表 2 は音素ごとに設定した場合の結果である．また表中の ϵ_c，ϵ_e は，認識結果より選択された最適なパラメータの値を示す．結果より学習回数の増加とともに，性能が向上することが分かる．ただし開発セットの場合，認識結果が最良となるようパラメータを選択しているので，別途評価セットで評価する必要がある．パラメータの設定に関してモデル共通と音素モデルごとを比較すると，モデル共通でより良い結果が得られた．MAP 推定の τ についても最適な値を求めるために開発セットを用いて認識実験を行った．5 種類の値で比較した結果 $\tau = 200$ で学習回数 3 回のとき最良値 WER 19.31%を得た．

次に，開発セットで選択された最適なパラメータを用いて評価セットで評価を行う．**表 3** にテストセット 1 および 3 の結果を，また**図 1** にテストセット 1 の結果を示す．図および表では MMI 推定と MAP 推定，ML 推定を比較する．base として示されているモデルは，CMHMM を ML 推定で 5 回，さらに DMHMM に変換後 MAP 推定で 2 回学習したものである．MAP や ML 推定での性能向上が少ないのは，すでにある程度学習されているためであると考えられる．MAP，ML とも 1，2 回の学習でわずかに性能が向上するが，それ以降は過学習のため性能が低下した．

表 1 開発セットにおける学習回数ごとの WER（%）．C および E を音素モデル共通で設定

Table 1 WER (%) for development set. Values of C and E are shared by all phonemes.

baseline	1 回目	2 回目	3 回目	4 回目	5 回目
ϵ_c	0.01	0.001	0.001	0.001	0.001
ϵ_e	2.5	0.01	0.001	0.001	1.5
19.59	19.52	19.41	19.26	19.18	19.14

表 2 開発セットにおける学習回数ごとの WER（%）．C および E を音素モデルごと設定

Table 2 WER (%) for development set. Values of C and E are set separately for phonemes.

baseline	1 回目	2 回目	3 回目	4 回目	5 回目
ϵ_c	0.01	0.01	0.01	0.01	0.01
ϵ_e	2.0	0.1	0.01	0.01	2.0
19.59	19.48	19.47	19.40	19.38	19.62

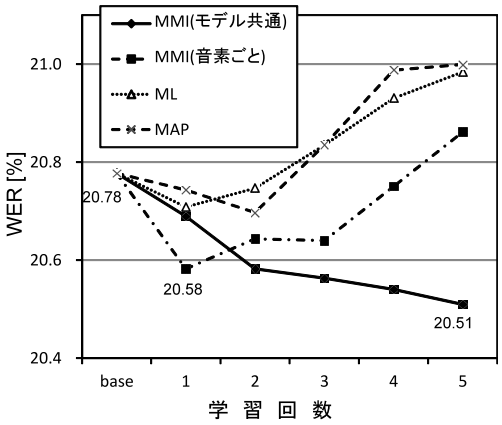

図 1 評価セットでの学習回数に対する WER の変化（%）（テストセット 1）

Fig. 1 WER (%) for evaluation set (test-set 1).

表 3 評価セットにおける学習回数ごとの WER（%）

Table 3 WER (%) for evaluation set.

学習回数	base	1 回目	2 回目	3 回目	4 回目	5 回目
テストセット 1（単語数：26,139）						
MMI（モデル共通）	20.78	20.69	20.58	20.56	20.54	**20.51**
MMI（音素ごと）		**20.58**	20.64	20.64	20.75	20.86
ML		**20.71**	20.75	20.83	20.93	20.98
MAP		20.74	**20.70**	20.83	20.99	21.00
テストセット 1+3（単語数：43,283）						
MMI（モデル共通）	22.09	21.99	21.96	21.81	21.72	**21.67**
MMI（音素ごと）		21.89	21.81	21.81	**21.81**	21.92
ML		22.09	**22.05**	22.14	22.20	22.21
MAP		22.08	**22.06**	22.13	22.24	22.28

表 4 I-smoothing を用いた認識実験の WER（%）（テストセット 1）
Table 4 WER (%) by using I-smoothing method (test-set 1).

τ_I	0.0	0.1	0.2	0.3	0.4	0.5	1.0
開発セット	19.14	**19.13**	19.15	19.15	19.14	19.14	19.24
評価セット	20.51	20.48	**20.47**	20.49	20.49	20.50	20.51

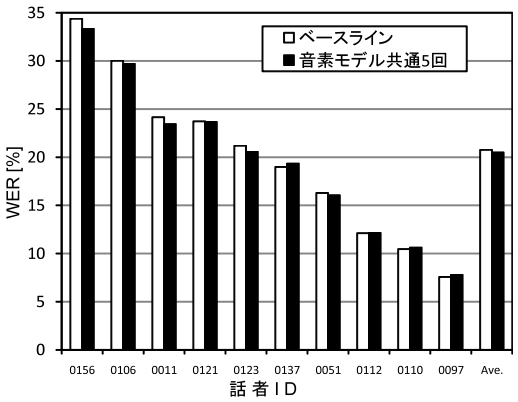

図 2 評価セットの話者ごとの WER（%）（テストセット 1）
Fig. 2 WER (%) of each speaker for evaluation set (test-set 1).

一方 MMI ではいずれの方法でも従来法の ML や MAP と比較して性能向上が得られる．テストセット 1+3 の評価において，従来法の最良値は ML 推定 2 回学習の 22.05%であるのに対し，MMI で音素ごとにパラメータを設定した場合は 21.81%（学習回数 4 回），音素モデル共通でパラメータ設定を行った場合 21.67%（学習回数 5 回）が得られた．符号検定法 [18] により検定を実施したところ前者は危険率 5%（$p = 0.0232$），後者は 1%（$p = 0.00002$）で有意という結果が得られた．学習回数 6 回ではモデル共通でも性能が低下した．パラメータの設定に関しては，開発セットでの結果と同様モデル共通でより良い結果が得られた．音素ごとにパラメータを設定する場合，式 (9) の偏微分の項を計算するためのデータ量が減少し，安定して求めることができないためと考えられる．以上より，DMHMM における MMI 推定の有効性が示された．

比較のために CMHMM についても MMI 推定を行った．3.2 節の実験条件で述べた DMHMM 変換前の 3,000 状態 16 混合，5 回 ML 推定の CMHMM をベースとし，DMHMM と同様に開発セットを用いて各種パラメータを定めた．最良の条件は MMI 4 回学習，モデル共通のパラメータ設定であり，このときテストセット 1 で評価するとベースの WER が 21.34%に対し MMI 学習後は 20.79%が得られた．この場合の改善率は 2.59%であり，DMHMM における改善率 1.30%（20.78%から 20.51%）を上回っている．DMHMM はパラメータ数が確率密度 1 分布あたり CMHMM の約 17.2 倍であり，データ量に対するパラメータ数の多さが影響している可能性がある．

テストセット 1 における話者ごとの評価結果を**図 2** に示す．MMI 推定の影響を見るため，ベースラインと MMI の音素モデル共通でパラメータを設定し学習したものを比較した．この場合，特に認識性能の低い話者で MMI の効果が高いのに対し，元々認識性能の高い話者では，性能向上が少ないか逆に悪化する．文献 [15] においても話者ごとの認識性能の分析を行っているが，必ずしもこのような傾向は見られない．このためこの傾向は識別学習特有の傾向とはいえないと考えられるが，CMHMM と DMHMM の違いなど種々の実験条件の違いがあるため，今後検討が必要である．

次にテストセット 1 における I-smoothing の結果を**表 4** に示す．図 1 の実験でモデル共通にパラメータを設定した場合の MMI の 5 回学習で最良の結果を得ているため，この条件での比較を行った．学習 4 回目のモデルを初期モデルとし，スムージング係数 τ_I を変えながら 5 回目の学習を行った．認識実験は開発セットおよび評価セットの両方で行った．開発セットでは $\tau_I = 0.1$ のとき最良の WER，19.13%を得た．また評価セットでは $\tau_I = 0.2$ のとき最良の WER，20.47%を得た．以上のように最良の τ_I は開発セットと評価セットで類似した値となった．これまでの実験と同様に開発セットでの最良のパラメータを評価実験で用いるとすると，20.48%が評価セットでの結果となる．$\tau_I = 0.0$ の 20.51%が通常の MMI 推定の結果である．また $\tau_I = \infty$ に相当する ML 推定の結果は 20.98%である．通常の MMI と比較すると I-smoothing 使用において WER は減少したが，その差は少なく有意な結果とはいえなかった．これはスムージングする相手の ML 推定の性能が低く，スムージングの効果が十分には表れなかったためと考えられる．

4. まとめ

本研究では離散混合分布 HMM における識別学習法を提案し，その有効性を講演音声認識により検討した．識別学習としては最大相互情報量基準を用いた．また，識別学習におけるパラメータの設定に関して C や E をモデル共通で設定する場合と，音素モデルごとに設定する場合の比較を行った．結論として，ML 推定や MAP 推定と比較し，本提案法は有効であることが分かった．パラメータ C，E の設定に関しては音素ごとよりモデル共通のほうで高い性能が得られた．また，I-smoothing と同様の手法の検討も行ったが性能向上は若干にとどまった．これはスムージングする相手である ML 推定の性能が低かったためと考えられる．

本研究では，DMHMM における識別学習の有効性を示すため，基本となる MMI を用いたが，CMHMM においては音素誤り基準を導入した MPE が成功を収めてい

る [11]．さらに fMPE により特徴量空間への拡張が行われている [19]．国内における CSJ を用いた識別学習の例を見ると，文献 [20] では MCE を用いて 1〜3 ポイント程度の向上，また文献 [21] において MPE で 1.75 ポイントの向上が得られている．よってこれらの方法が DMHMM においても有効であるか今後検討が必要である．また DMHMM は雑音下では CMHMM に比べ有効であるが，クリーンな環境では従来の CMHMM と比較すると性能差が少ないため [22] 利用するメリットが少ないと考えられる．しかし，CMHMM とは誤認識の傾向が違うため，DMHMM の認識結果と CMHMM の認識結果を統合することによりクリーンな環境での性能向上も見込まれる [22]．今後は，このシステム統合の方法についても種々検討しクリーンな環境におけるさらなる性能向上を目指す．

謝辞 本研究の一部は科学研究費（課題番号 22500144）によった．また 3.3 節の実験にご協力いただいた佐藤元宣君に感謝する．

参考文献

[1] Chen, S.F. et al.: Advance in speech transcription at IBM under the DARPA EARS program, *IEEE Trans. Audio, Speech, and Language Process.*, Vol.14, No.5, pp.1596–1608 (2006).
[2] Furui, S., Nakamura, M., Ichiba, T. and Iwano, K.: Analysis and recognition of spontaneous speech using corpus of spontaneous Japanese, *Speech Communication*, Vol.47, pp.208–219 (2005).
[3] 中村 篤ほか：音声認識システム SOLON の日本語話し言葉コーパスによる評価（2006 年度版），信学技報，SP2006-127, pp.73–78 (2006).
[4] 加藤正治，小坂哲夫，伊藤彰則，牧野正三：Quinphone HM-Net に基づく講演音声認識，音講論秋，1-9-7, pp.21–24 (2010).
[5] Merialdo, B.: Phonetic recognition using hidden Markov models and maximum mutual information training, *Proc. ICASSP88*, pp.111–114 (1988).
[6] 渡部晋治：小特集—自動音声認識研究の動向と展望，音声認識における音響モデル，音響誌，Vol.66, No.1, pp.18–22 (2010).
[7] Takahashi, S., Aikawa, K. and Sagayama, S.: Discrete mixture HMM, *Proc. ICASSP97*, pp.971–974 (1997).
[8] Kosaka, T., Katoh, M. and Kohda, M.: Robust speech recognition using discrete-mixture HMMs, *IEICE Trans. Inf. & Syst.*, Vol.E88-D, No.12, pp.2811–2818 (2005).
[9] Kosaka, T., Yamamoto, A., Kumakura, T., Kato, M. and Kohda, M.: Lecture speech recognition using discrete-mixture HMMs, *IEEJ Trans. Electrical and Electronic Engineering*, Vol.6, No.1, pp.23–29 (2011).
[10] McDermott, E. et al.: Discriminative training for large-vocabulary speech recognition using minimum classification error, *IEEE Trans. Audio, Speech, and Language Process.*, Vol.15, No.1, pp.203–223 (2007).
[11] Povey, D. and Woodland, P.C.: Minimum Phone Error and I-SMOOTHING for Improved Discriminative Training, *Proc. ICASSP2002*, pp.105–108 (2002).
[12] Povey, D., Kanevsky, D., Kingshury, B., Ramabhadran, B., Saon, G. and Visweswariah, K.: Boosted MMI for Model and Feature-Space Discriminative Training, *Proc. ICASSP2008*, pp.4057–4060 (2008).
[13] Tsakalidis, S., Digalakis, V. and Newmeyer, L.: Efficient speech recognition using subvector quantization and discrete-mixture HMMs, *Proc. ICASSP99*, pp.569–572 (1999).
[14] Gopalakrishnan, P.S. et al.: A Generalization of the Baum algorithm to rational objective functions, *Proc. ICASSP89*, pp.631–634 (1989).
[15] Valtchev, V., Odell, J.J., Woodland, P.C. and Young, S.J.: MMIE training of large vocabulary recognition systems, *Speech Communication*, Vol.22, No.4, pp.303–314 (1997).
[16] 堀 貴明，岡 直生，加藤正治，伊藤彰則，好田正紀：大語彙連続音声認識のための音素グラフに基づく仮説制限法の検討，情報処理学会論文誌，Vol.40, No.4, pp.1365–1373 (1999).
[17] 堤 怜介，加藤正治，小坂哲夫，好田正紀：発音変形依存モデルを用いた講演音声認識，信学論，Vol.J89-D, No.2, pp.305–313 (2006).
[18] 中川聖一，高木英行：パターン認識における有意差検定と音声認識システムの評価法，音響誌，Vol.50, No.10, pp.849–854 (1994).
[19] Povey, D. et al.: fMPE: Discriminatively trained features for speech recognition, *Proc. ICASSP2005*, pp.961–964 (2005).
[20] 中村 篤ほか：音声認識システム SOLON の日本語話し言葉コーパス（公開版 Ver.1.0）による評価，信学技報，SP2005-106, pp.7–12 (2005).
[21] 篠崎隆宏，久保田雄，ディクソン・ポール，古井貞熙：識別学習モデルと教師なし CV 適応を用いた CSJ 講演音声認識，音講論春，1-6-14, pp.37–38 (2010).
[22] Kosaka, T., Goto, K., Ito, T. and Kato, M.: Lecture speech recognition by combining word graphs of various acoustic models, *Proc. Interspeech2010*, pp.2978–2981 (2010).

小坂 哲夫 （正会員）

1984 年東北大学工学部電気工学科卒業．1986 年東北大学大学院博士前期課程修了．同年キヤノン（株）に入社．1991 年（株）ATR 自動翻訳電話研究所に出向．2002 年山形大学工学部情報科学科助教授．現在，山形大学大学院理工学研究科教授．博士（情報科学）．音声情報処理の研究に従事．1995 年電子情報通信学会論文賞．IEEE Senior Member．日本音響学会，電子情報通信学会各会員．

加藤 正治

1991 年山形大学工学部情報工学科卒業．1993 年山形大学大学院修士課程修了．同年山形大学工学部助手．現在，山形大学大学院理工学研究科助教．博士（工学）．音声認識等，音声情報処理の研究に従事．日本音響学会会員．