情報処理学会論文誌 Vol.59 No.4 1295–1308 (Apr. 2018)

推薦論文

可変長時系列パターン分類のための大幾何マージン最小分類誤り学習法の提案とその実験的評価

松廣 達也[1] 橋本 哲也[1] 北岡 見生代[1] ア デイビッド[1] 落合 翼[1] 渡辺 秀行[2] 片桐 滋[1,a)]
大崎 美穂[1]

受付日 2017年5月31日, 採録日 2018年1月15日

概要：最近，識別学習法の 1 つ，最小分類誤り（MCE: Minimum Classification Error）学習法が，大幾何マージン最小分類誤り（LGM-MCE: Large Geometric Margin Minimum Classification Error）学習法に拡張された．LGM-MCE 法は，利用可能な学習用標本上の分類誤り数損失の最小化と，標本空間における幾何マージンの最大化とを同時に行うものである．学習用標本に対する過学習を抑制することで，この新しい LGM-MCE 法は，理想的な最小分類誤り確率状態に対応する分類器状態を近似する．その有効性は，プロトタイプ型分類器用に実装され，様々な固定次元ベクトルパターンの分類課題において実証されている．本稿は，可変長パターンに対しても過学習が抑制された望ましい分類が実現されることを目指し，状態遷移モデル型分類器を用いて可変長パターンのための幾何マージンを新たに導出し，それを用いた LGM-MCE 学習法を構築し，その有効性を 3 種の音声認識課題において評価した結果を報告するものである．実験では，LGM-MCE 学習法と，その基盤となった MCE 学習法とを比較する．実験の結果，いずれの課題においても，LGM-MCE 学習法が，過学習を抑制し，未知の試験用標本上で高い分類精度を達成することが示される．比較実験に用いる LGM-MCE 学習法の実装は，適応型損失最小化手段である確率的降下法を用いている．本稿では追加的に，高速な並列処理に適した RPROP 法に基づく実装も行い，その動作確認結果も報告する．

キーワード：最小分類誤り学習法，確率的降下法，RPROP 法

Development and Experimental Ealuation of Large Geometric Margin Minimum Classification Error Training for Classification of Variable-length Time Series Patterns

Tatsuya Matsuhiro[1] Tetsuya Hashimoto[1] Mikiyo Kitaoka[1] David Ha[1] Tsubasa Ochiai[1]
Hideyuki Watanabe[2] Shigeru Katagiri[1,a)] Miho Ohsaki[1]

Received: May 31, 2017, Accepted: January 15, 2018

Abstract: Recently, one of the latest discriminative training methods, Minimum Classification Error (MCE) training, was extended to Large Geometric Margin MCE (LGM-MCE) training. This new version of MCE training, LGM-MCE training, simultaneously maximizes the geometric margin in the sample space as well as minimizes the classification error count loss over the training samples in hand. It is expected to effectively approximate an optimal classifier status, which corresponds to the desirable, minimum classification error probability condition, by suppressing overlearning. So far, the LGM-MCE training method has been implemented for prototype-based classifiers, and its effectiveness was demonstrated in various fixed-dimension pattern classification tasks. In this paper, to achieve the ideal overlearning-suppressed classification for variable-length patterns, we newly define a new LGM-MCE training procedure for the state-transition-model-based classifier, and evaluate its effect in three speech pattern classification tasks. In experiments, we compare LGM-MCE training and its baseline, MCE training, and show that LGM-MCE training successfully suppresses overlearning and surpasses MCE training in classification accuracy over unknown testing patterns. The LGM-MCE training implementation used in experiments adopts an adaptive Probabilistic-Descent-method-based loss minimization procedure. In the paper, we also introduce an alternative RPROP-method-based implementation, which will be useful for fast parallel run of loss minimization.

Keywords: minimum classification error training, probabilistic descent method, RPROP method

1. はじめに

統計的パターン認識における理想は，最小分類誤り確率状態，いわゆるベイズリスク状態に対応する分類[*1]を行うことである [1]．しかし，無限個のパターン標本によって定義されるベイズリスクは本質的に観測が難しい．したがって実際には，利用可能な有限のパターン標本に基づいて，最小分類誤り確率状態を近似するクラス境界，あるいはそれを導く分類器の実現が目指されている [1], [2]．

そうした実現を効率的に目指す手法として，パーセプトロン学習の時代から識別学習法が幅広く研究されてきた [3]．識別学習は，最小分類誤り確率状態を，学習の評価基準，すなわち損失として近似する定義法に応じて，二乗誤差損失最小化法やクロスエントロピー最小化法，ヒンジ損失最小化法など，実に様々な様式で定形化されてきた [1], [4], [5]．

理想状況が最小分類誤り確率状態であるとき，分類誤り数を損失とすることは自然である．この理解に基づき，関数マージン最小分類誤り（FM-MCE: Functional Margin Minimum Classification Error）学習法[*2]が提案され，特に音声認識器のための識別学習法として広く用いられてきた [4], [6]．しかし，識別学習法の効率の良さは，翻って，学習用標本に対する過剰適応，すなわち過学習を引き起こしやすくなる．この弱点に関し，FM-MCE 学習も例外ではなく，正則化や損失の改良などに基づく様々な改良が行われてきた [7], [8]．

一般に，過学習は最小分類誤り確率の過小推定を招く．過学習された認識器は，学習用標本に対して高い認識精度を達成する一方で，未知の試験用標本に対しては低い精度の達成にとどまりやすい．したがって，過学習の抑制は，必ずしも保証されるものではないものの，未知標本に対する認識性能の向上を期待させる．実際，識別学習だけでなく広範な学習法において用いられている正則化の概念は，この期待に基づいている [2]．また，サポートベクタマシン（SVM: Support Vector Machine）の登場によって注目を集める幾何マージン最大化の概念も同様である [9], [10]．

SVM は，完璧な認識が困難な現実の認識課題の多くにおいては，非線形カーネル写像をともなう線形識別関数の学習法として，ヒンジ損失の最小化と幾何マージンの最大化（過学習の抑制）を目指す．実際，多くの評価実験においてその有用性が認められ，有力な認識器実現法として定着しつつある．しかし，SVM は，基本的に固定次元ベクトルパターンを入力とし，可変長時系列パターンの認識への直接的な適用は必ずしも容易ではない．また，学習手続き自体あるいは学習後の認識器規模が膨大になるスケーラビリティの問題もかかえている．したがって，様々な認識課題において高い（未知標本に対する）認識性能の達成が実証されてはいるものの，こうした問題の解決が望まれてきた [11]．

こうした状況の中で，FM-MCE 学習法の過学習を抑制し，また同時に，SVM が持つ可変長パターンへの適用の困難さを解決する識別学習法として，大幾何マージン最小分類誤り（LGM-MCE: Large Geometric Margin Minimum Classification Error）学習法が提案された [12]．この新しい学習法の定形化においては，まず，基本的に識別関数の種類に制約されない一般的な幾何マージンが導入された．そして，幾何マージンとの親和性が高い距離型識別関数を用いるプロトタイプ型分類器に実装され，様々な固定次元ベクトルパターン分類課題においてその有用性が明らかにされた．しかしその一方で，元々 FM-MCE 学習法の長所でもあった，音声パターンのような可変長時系列パターンの分類のための実装とその評価は，まだ必ずしも十分ではなかった．実際，著者らが知る限り，可変長時系列パターンにおける幾何マージンの定式化さえも行われてこなかった．また，たとえば動的計画法に基づくパターンの非線形伸縮，すなわち動的時間軸伸縮（DTW: Dynamic Time Warping）をともなって定義される可変長パターン間の距離は，固定次元空間における幾何学的距離とまったく同様に扱うことは難しく，固定次元の距離空間において認められた幾何マージン増大による過学習抑制効果が，そうした可変長パターンの距離空間においても認められるとは限らない．

本稿は，上記の不十分さや未解明部分の解消を目指し，状態遷移モデル型識別関数を用いて，可変長パターンのための幾何マージンを新たに定式化し，その識別関数を用いる可変長時系列パターン分類のための LGM-MCE 学習法を提案するものである．提案手法の評価は，複数の音声認識課題における，LGM-MCE 学習法とその基盤でもある FM-MCE 学習法との比較実験によって行う．固定次元パターン分類において LGM-MCE 学習法が評価された際には，FM-MCE 学習法だけでなく，過学習抑制を狙った一般化学習量子化（GLVQ: Generalized Learning Vector Quantization）法との比較も行われた [12]．しかしそこで

[1] 同志社大学
Doshisha University, Kyotanabe, Kyoto 610–0394, Japan
[2] 株式会社国際電気通信基礎技術研究所
Advanced Telecommunications Research Institute International, Sagara-gun, Kyoto 619–0288, Japan
[a)] skatagiri@mail.doshisha.ac.jp

本論文の内容は 2016 年 9 月の情報処理学会関西支部支部大会にて報告され，同支部長により情報処理学会論文誌ジャーナルへの掲載が推薦された論文である．

[*1] 本稿では，「特徴抽出（Feature extraction）」と「分類（Classification）」からなる過程を「認識（Recognition）」と呼び，「識別」という用語は，識別関数（Discriminant function）や識別学習（Discriminative training）を指す際に用いている．

[*2] 元々 MCE 学習法と呼ばれたものであるが，後述する改良型の MCE 学習法である LGM-MCE 学習法との区別を明確にするため，本稿では，当初の MCE 学習法を FM-MCE 学習法と呼ぶこととする．

は，GLVQ 法の過学習抑制機構は必ずしも幾何マージンを増加させるものでないことが示され，かつその抑制力は FM-MCE 学習法のそれとほぼ同程度であることも示されている [12]．こうした結果を受け，本稿における LGM-MCE 学習法の比較対象は，FM-MCE 学習法によって代表させている．

MCE 学習における損失の最小化には，最急降下法や確率的降下（PD: Probabilistic Descent）法 [13] *3が用いられることが多い．そうした中で，本稿では特に，適応学習型の PD 法を採用する．なお，最急降下法においても PD 法においても，そこで用いられる学習係数の設定は，実験や経験に頼らざるをえないやっかいな問題であり，本稿におけるような実験を行う際にもその改善が望まれる．そこで，本稿における研究では，最急降下法における学習係数の設定が不要となる RPROP 法 [14], [15] を FM-MCE 法と LGM-MCE 法との双方に実装し，その動作の検証も行った．得られた結果は，LGM-MCE 学習法の有用性を検証する本稿の主たる文脈とやや離れるため，付録において紹介する．

2. 準備

2.1 分類課題および分類規則

可変長時系列パターン標本 $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T]$ を J 個のクラス $\{C_j \ (j = 1, \ldots, J)\}$ のうちの 1 つに分類する課題を考える．課題は，以下のように規則化することができる．

$$C(\boldsymbol{X}) = C_i \quad \text{iff } i = \arg\min_j \ g_j(\boldsymbol{X}; \boldsymbol{\Lambda}). \tag{1}$$

なおここで，$\boldsymbol{x}_t$ は系列中のフレーム時刻指標 t における δ 次元の（音声）音響特徴ベクトルであり，T は音響特徴ベクトルの数（フレーム数）で表す，$\boldsymbol{X}$ の長さである．また，$\boldsymbol{\Lambda}$ は分類器パラメータ（クラスモデルパラメータ）の集合であり，$C(\,)$ は分類オペレータである．説明の具体化のため，課題は孤立単語音声認識とし，分類すべきクラスは単語クラスとする．

2.2 状態遷移モデル型識別関数を用いる分類器

上記の課題に対処するため，音声認識で広く用いられている混合ガウス分布・状態遷移モデル型，いわゆる隠れマルコフモデル（HMM: Hidden Markov Model）型識別関数（例：文献 [4]）を採用することは一案である．しかし本稿では，幾何マージンの定義の正確さを優先し，マルチプロトタイプ距離・状態遷移モデル型識別関数 [16] を採用する．可変長パターン間の類似性を測るとき，HMM 型であってもマルチプロトタイプ距離・状態遷移モデル型であっても，

*3 PD 法は，最近では確率的勾配降下（SGD: Stochastic Gradient Desent）法と呼ばれることも多い．しかし PD 法という呼称は，60 年代から用いられてきた長い歴史を持つ．

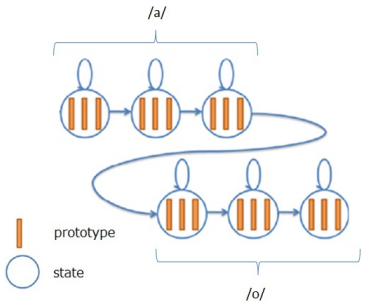


図 1 単語 “あお” の状態遷移型クラスモデルの概念図
Fig. 1 Conceptual diagram of state transition model of word “/a/ /o/”.

その尺度にパターン長の伸縮にともなう非線形性が加わることは基本的に避けにくい．一方，固定次元の音響ベクトル空間においては，ガウス分布のような確率型識別関数に基づく幾何マージンの定義に近似を避けることが難しいのに対し，距離型識別関数に基づく幾何マージンは近似をともなわず厳密に定義することができる [12]．したがって，距離型識別関数を採用することでより正確に定義される幾何マージンを用いることが可能となり，本稿ではこの点を優先した．

図 1 に，各クラスの識別関数を構成するマルチプロトタイプ距離・状態遷移モデルを図解する．単語に相当するクラスモデルは，状態遷移構造を持つ音素モデルを連結して構成する．図では，/a/と/o/との 2 つの音素に対応する音素モデル（3 状態かつ 3 プロトタイプ（/状態）の例）を連結して単語 “あお” に対するクラスモデルを構成している．

後続する学習法の定義においては，音素モデルは，S 個の状態を持ち，また各状態において I 個のプロトタイプを持つものとし，たとえば，h 番目の音素の s 番目の状態の i 番目のプロトタイプは $\boldsymbol{r}_i^{h,s}$ と表す．FM-MCE 学習法および LGM-MCE 学習法における学習対象はこのプロトタイプであり，式 (1) 中の $\boldsymbol{\Lambda}$ は $\boldsymbol{\Lambda} = \{\boldsymbol{r}_i^{h,s}\}_{h=1,s=1,i=1}^{H,\ S,\ I_{h,s}}$ と表されることになる．なおここで，H と S，$I_{h,s}$ は，それぞれ，音素クラスの数と，音素クラスに用いられる状態遷移モデルの状態数，その第 h 音素クラスモデルの第 s 状態に配置されるプロトタイプ数を示している．

固定次元ベクトルパターンの次元数と異なり，音声のような可変長時系列パターンにおけるフレーム数はパターンごとに異なる．したがって，そのようなパターンと（状態遷移型）モデルとの間の距離を測る際には工夫が必要である．音声認識分野で広く行われてきたように，本稿で扱う分類器も，DTW による累積距離計算を採用する．このとき，各単語クラスに対する識別関数を，DTW による最小累積距離として次のように定義する．

$$g_j(\boldsymbol{X}; \boldsymbol{\Lambda}) = \frac{1}{T} \sum_{t=1}^{T} \| \boldsymbol{x}_t - \boldsymbol{r}_{i(\varphi_{j,t}, \theta_{j,t}, t)}^{\varphi_{j,t}\ \theta_{j,t}} \|^2. \tag{2}$$

ここで，最適な経路 $\{(\varphi_{j,1}\ \theta_{j,1}); (\varphi_{j,2}\ \theta_{j,2}); \ldots; (\varphi_{j,T}\ \theta_{j,T})\}$

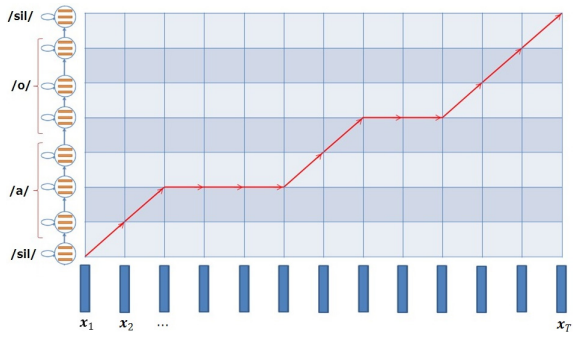


図 2 DTW による識別関数値の計算の概念
Fig. 2 Concept of DTW-based distance calculation.

は，DTW によって自動的に見出される．また，$\varphi_{j,t}$ は第 j クラスにおける時刻 t での音素であり，$\theta_{j,t}$ は第 j クラスにおける時刻 t における状態の指標である．さらに，$i\left(\varphi_{j,t},\theta_{j,t},t\right) = \arg\min_{i=1}^{I} \| \boldsymbol{x}_t - \boldsymbol{r}_i^{\varphi_{j,t}\ \theta_{j,t}} \|^2$ である．図 2 は，この DTW に基づく識別関数の計算過程を図解している．図は，横軸に入力パターンの音声特徴ベクトルを置き，縦軸に音声区間の始めと終わりに無音区間（“sil” で示す）を持つ，単語クラス “あお” の状態遷移モデルを置き，両者の間で行われる DTW の手続きを示している．

2.3 関数マージン最小分類誤り学習法

LGM-MCE 学習法の新しい実装などを議論するとき，その基盤となる FM-MCE 学習法の定式化の利用は必須である．準備の一環として，本節において FM-MCE 学習法の概要をふりかえる．

2.3.1 関数マージン型誤分類尺度

FM-MCE 学習は，まず，学習用標本 $\boldsymbol{X}$（$\in C_y$）に対する，分類判断の正誤およびその程度を表す関数マージン型誤分類尺度を式 (2) の識別関数を用いて

$$d_y\left(\boldsymbol{X};\boldsymbol{\Lambda}\right) = g_y\left(\boldsymbol{X};\boldsymbol{\Lambda}\right) - \left[\frac{1}{J-1}\sum_{j,j\neq y} g_j\left(\boldsymbol{X};\boldsymbol{\Lambda}\right)^{-\psi}\right]^{-\frac{1}{\psi}} \tag{3}$$

と定義する[*4]．ここで ψ は正の定数であり，$\psi \to \infty$ としたとき，式 (3) の誤分類尺度は次のように簡単化される．

$$\begin{aligned} &d_y\left(\boldsymbol{X};\boldsymbol{\Lambda}\right) = g_y\left(\boldsymbol{X};\boldsymbol{\Lambda}\right) - g_{y^*}\left(\boldsymbol{X};\boldsymbol{\Lambda}\right), \\ &y^* = \arg\min_{j,j\neq y} g_j\left(\boldsymbol{X};\boldsymbol{\Lambda}\right). \end{aligned} \tag{4}$$

ここで，式 (4) が，優れて規則 (1) における操作 min の表現になっていることが分かる．以下では，この一貫性の維持と処理の簡単化とのため，誤分類尺度には式 (4) を用いるものとする．以降では，便利のため，C_{y^*} をベスト・インコレクト・クラスと呼ぶ．

[*4] 元々 FM-MCE 学習法においては，式 (3) は単に誤分類尺度と呼ばれていたが，LGM-MCE 学習法が用いる尺度との対比のため，関数マージン型誤分類尺度と呼ぶ．

2.3.2 平滑な分類誤り数損失

式 (4) が示すように，FM-MCE 学習法は，誤分類尺度の符号を用いて分類の正誤を表す．すなわち，正値が誤分類を，負値が正分類を表す．したがって，誤分類尺度の関数として

$$\ell_y\left(\boldsymbol{X};\boldsymbol{\Lambda}\right) = \begin{cases} 1 & \text{if} \quad d_y\left(\boldsymbol{X};\boldsymbol{\Lambda}\right) > 0 \\ 0 & \text{if} \quad d_y\left(\boldsymbol{X};\boldsymbol{\Lambda}\right) < 0 \end{cases} \tag{5}$$

の 0-1 損失関数を用いることによって，学習用標本に対する分類誤りの個数を求めることが可能となる．しかし，式 (5) は学習対象である $\boldsymbol{\Lambda}$ に関して微分不可能であり，さらにほとんど常にその勾配値はゼロであり，勾配法を用いて分類器の学習をすることは難しい．そこで，式 (5) を，

$$\ell_y\left(\boldsymbol{X};\boldsymbol{\Lambda}\right) = \frac{1}{1+\exp\left(-\alpha d_y\left(\boldsymbol{X};\boldsymbol{\Lambda}\right)\right)} \tag{6}$$

の平滑な分類誤り数損失で置き換える．ここで α は正の定数であり，損失関数の平滑度を制御する．

最小分類誤り確率状態の近似を目指す FM-MCE 学習は，式 (6) の損失を多数の学習用標本のそれぞれに適用し，そうして得られる以下の経験的平均損失

$$L\left(\boldsymbol{\Lambda}\right) = \frac{1}{N}\sum_{n=1}^{N} \ell_{y_n}\left(\boldsymbol{X}_n;\boldsymbol{\Lambda}\right) \tag{7}$$

の最小化を目指す．ここで $\boldsymbol{X}_n$ は n 番目の学習用標本を表し，y_n はその学習用標本が所属するクラスを表している．なお，この学習用標本を表す指標に合わせて，$\boldsymbol{X}_n$ の要素に関しては $\boldsymbol{X}_n = [\boldsymbol{x}_1^n,\ldots,\boldsymbol{x}_t^n,\ldots,\boldsymbol{x}_{T_n}^n]$ と記述することとする．

2.3.3 確率的降下法による経験的平均損失の最小化

1 つの学習用標本ごとに状態遷移モデルのプロトタイプを更新し，学習用標本を取り替えながらその更新を繰り返す PD 法の m 回目（$m = 1,2,3,\ldots$）の更新において，プロトタイプ $\boldsymbol{r}_i^{h,s}$ に関する更新は

$$\boldsymbol{r}_i^{h,s(m+1)} = \boldsymbol{r}_i^{h,s(m)} - \epsilon_m \nabla_{\boldsymbol{r}_i^{h,s}} \ell_y(\boldsymbol{X};\boldsymbol{\Lambda}^{(m)}) \tag{8}$$

となる．ここで，$\boldsymbol{r}_i^{h,s(m)}$ と $\boldsymbol{r}_i^{h,s(m+1)}$ は，それぞれプロトタイプ $\boldsymbol{r}_i^{h,s}$ の繰返し指標 m における更新前と更新後の状態（値）であり，さらに $\boldsymbol{\Lambda}^{(m)}$ と $\boldsymbol{X}$，y，ϵ_m とは，それぞれ m における，全プロトタイプの更新前の状態と，学習用標本，学習用標本の正解クラスの指標，学習係数（$\epsilon_m > 0$）である．また，$\nabla_{\boldsymbol{r}_i^{h,s}} \ell_y(\boldsymbol{X};\boldsymbol{\Lambda})$ は，微分の鎖則により，以下のように展開される．

$$\begin{aligned} &\nabla_{\boldsymbol{r}_i^{h,s}} \ell_y(\boldsymbol{X};\boldsymbol{\Lambda}) \\ &= \frac{\partial \ell_y(\boldsymbol{X};\boldsymbol{\Lambda})}{\partial d_y(\boldsymbol{X};\boldsymbol{\Lambda})} \cdot \left(\frac{\partial d_y(\boldsymbol{X};\boldsymbol{\Lambda})}{\partial g_y(\boldsymbol{X};\boldsymbol{\Lambda})} \cdot \nabla_{\boldsymbol{r}_i^{h,s}} g_y(\boldsymbol{X};\boldsymbol{\Lambda}) \right. \\ &\quad \left. + \frac{\partial d_y(\boldsymbol{X};\boldsymbol{\Lambda})}{\partial g_{y^*}(\boldsymbol{X};\boldsymbol{\Lambda})} \cdot \nabla_{\boldsymbol{r}_i^{h,s}} g_{y^*}(\boldsymbol{X};\boldsymbol{\Lambda}) \right). \end{aligned} \tag{9}$$

3. 状態遷移モデル型識別関数を用いる分類器のための大幾何マージン最小分類誤り学習法 [17], [18], [19]

3.1 可変長パターンのための幾何マージン

幾何マージンは，固定次元ベクトルパターンの空間において，クラス境界とその最近傍パターンの間のユークリッド距離として定義される [10]．また，固定次元ベクトルパターン分類のための LGM-MCE 学習法，特に距離型識別関数を用いるプロトタイプ型分類器のための適用においては，その幾何マージンは，符号を反転した誤分類尺度値に厳密に等しいことが示されている [12]．

本来，可変長パターンのための幾何マージンは，可変長パターンの距離空間を定義し，その空間において導出されるべきである．しかし，そうした可変長パターンの距離空間は必ずしも十分に解明されていない．一方，DTW に基づく距離が，音声などの時系列パターン間の距離として広く普及している．こうした状況を考えたうえで，可変長パターン分類における LGM-MCE 学習法を用いた分類器の実装を可能にするため，DTW に基づく距離の定義の下で，固定次元ベクトルパターンの場合 [12] と同じようにして幾何マージンを導出する．

まず，可変長パターン間の距離を，DTW に基づく最小累積距離として定義する．各クラスの識別関数値は，それぞれのクラスにおける（DTW によって選択可能な）プロトタイプ系列の中で，この DTW 距離の意味で入力パターン標本に最近傍のプロトタイプ系列と入力パターンとの間の距離となる．したがって，クラス境界は，この識別関数値の意味で，正解クラスとベスト・インコレクト・クラスとの識別関数値が等しい点の集合となり，以下のように表される．

$$\beta_y(\boldsymbol{\Lambda}) = \{\boldsymbol{X} \in \mathcal{X} | d_y(\boldsymbol{X}; \boldsymbol{\Lambda}) = 0\}. \tag{10}$$

ここで $\mathcal{X}$ は標本空間を示す．幾何マージン ρ は，クラス境界に最近傍の（正しく分類された）入力パターン標本 $\boldsymbol{X}^\dagger$ と，それに対応するクラス境界上の最近傍点 $\boldsymbol{X}^\ddagger$ との間の DTW 距離と定義する．なお，$\boldsymbol{X}^\dagger$ と $\boldsymbol{X}^\ddagger$ はそれぞれ以下のように表される．

$$\boldsymbol{X}^\dagger = \left[\boldsymbol{x}^\dagger{}_1, \cdots, \boldsymbol{x}^\dagger{}_t, \cdots, \boldsymbol{x}^\dagger{}_{T^\dagger}\right]. \tag{11}$$

$$\boldsymbol{X}^\ddagger = \left[\boldsymbol{x}^\ddagger{}_1, \cdots, \boldsymbol{x}^\ddagger{}_t, \cdots, \boldsymbol{x}^\ddagger{}_{T^\ddagger}\right]. \tag{12}$$

したがって，幾何マージン ρ は次のように定義できる．

$$\rho = \sqrt{\frac{1}{T^\ddagger}\sum_{t=1}^{T^\ddagger} \| \boldsymbol{x}^\ddagger{}_t - \boldsymbol{x}^\dagger{}_{\pi(t)} \|^2}. \tag{13}$$

ここで $\pi(t)$ は $\boldsymbol{X}^\ddagger$ との最小累積距離を計算する際の最適経路であり，簡単化のため $T^\ddagger \leq T^\dagger$ および $T^\dagger = \pi\left(T^\ddagger\right)$ とする．定義より，$\boldsymbol{X}^\ddagger$ は以下の制約条件付き最小化問題の解として与えられる．

$$\begin{aligned} &\min_{\boldsymbol{X}} \frac{1}{T}\sum_{t=1}^{T} \| \boldsymbol{x}_t - \boldsymbol{x}^\dagger{}_{\pi(t)} \|^2 \\ &\quad \text{subject to: } \boldsymbol{X} \in \beta_y(\boldsymbol{\Lambda}). \end{aligned} \tag{14}$$

次に，ラグランジュ未定乗数 λ を導入し，次式の評価関数を考える．

$$V(\boldsymbol{X}, \lambda) = \frac{1}{T}\sum_{t=1}^{T} \| \boldsymbol{x}_t - \boldsymbol{x}^\dagger{}_{\pi(t)} \|^2 + \lambda d_y(\boldsymbol{X}; \boldsymbol{\Lambda}). \tag{15}$$

式 (15)，式 (4)，式 (2) より，$\boldsymbol{X}^\ddagger$ は次式を満たさなければならない．

$$\begin{aligned} &\frac{2}{T^\ddagger}(\boldsymbol{x}^\ddagger_t - \boldsymbol{x}^\dagger{}_{\pi(t)}) + \lambda\{\frac{2}{T^\ddagger}(\boldsymbol{x}^\ddagger_t - \boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)}) \\ &\quad -\frac{2}{T^\ddagger}(\boldsymbol{x}^\ddagger_t - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)})\} = 0 \\ &\quad (t = 1, \cdots, T^\ddagger), \end{aligned} \tag{16}$$

$$\begin{aligned} &\sum_{t=1}^{T^\ddagger} \| \boldsymbol{x}^\ddagger_t - \boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} \|^2 \\ &\quad = \sum_{t=1}^{T^\ddagger} \| \boldsymbol{x}^\ddagger_t - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)} \|^2. \end{aligned} \tag{17}$$

さらに，式 (16) より，

$$\begin{aligned} &\boldsymbol{x}^\ddagger_t - \boldsymbol{x}^\dagger{}_{\pi(t)} = \lambda\left(\boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)}\right) \\ &(t = 1, \cdots, T^\ddagger). \end{aligned} \tag{18}$$

式 (13) と式 (18) から幾何マージン ρ は次のように書き換えられる．

$$\rho = \frac{|\lambda|}{\sqrt{T^\ddagger}} \sqrt{\sum_{t=1}^{T^\ddagger} \| \boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)} \|^2}. \tag{19}$$

また，式 (17) を変形すると以下のようになる．

$$\begin{aligned} &\sum_{t=1}^{T^\ddagger} \boldsymbol{x}^{\ddagger\mathrm{T}}_t \left(\boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)}\right) \\ &= \frac{1}{2}\sum_{t=1}^{T^\ddagger} \| \boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} \|^2 - \frac{1}{2}\sum_{t=1}^{T^\ddagger} \| \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)} \|^2, \end{aligned} \tag{20}$$

ここで T はベクトルの転置である．さらにこの式 (20) に対して両辺に $\sum_{t=1}^{T^\ddagger} \boldsymbol{x}^\dagger_{\pi(t)}{}^{\mathrm{T}} \left(\boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)}\right)$ を減算することで次式を得る．

$$\begin{aligned} &\sum_{t=1}^{T^\ddagger} \left(\boldsymbol{x}^\ddagger_t - \boldsymbol{x}^\dagger_{\pi(t)}\right)^{\mathrm{T}} \left(\boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)}\right) \\ &\quad = \frac{1}{2}\left(\sum_{t=1}^{T^\ddagger} \| \boldsymbol{x}^\dagger_{\pi(t)} - \boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} \|^2 \right. \\ &\qquad \left. - \sum_{t=1}^{T^\ddagger} \| \boldsymbol{x}^\dagger_{\pi(t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)} \|^2 \right). \end{aligned} \tag{21}$$

式 (18) を式 (21) に代入し，変形すると以下のようになる．

$$\lambda = \frac{1}{2\sum_{t=1}^{T^\ddagger} \| \boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)} \|^2} \times \left(\sum_{t=1}^{T^\ddagger} \| \boldsymbol{x}^\dagger_{\pi(t)} - \boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} \|^2 - \sum_{t=1}^{T^\ddagger} \| \boldsymbol{x}^\dagger_{\pi(t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)} \|^2 \right). \tag{22}$$

さらに，この式 (22) を式 (19) に代入することにより，幾何マージン ρ は次のようになる．

$$\rho = \frac{1}{\sqrt{T^\ddagger}} \times \frac{1}{2\sqrt{\sum_{t=1}^{T^\ddagger} \| \boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)} \|^2}} \times \left| \sum_{t=1}^{T^\ddagger} \| \boldsymbol{x}^\dagger_{\pi(t)} - \boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} \|^2 - \sum_{t=1}^{T^\ddagger} \| \boldsymbol{x}^\dagger_{\pi(t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)} \|^2 \right|. \tag{23}$$

しかしながら，導出した幾何マージン ρ は算出不可能な $\boldsymbol{X}^\ddagger$ との最適経路などが含まれており，求めることが不可能である．そこで，$\boldsymbol{X}^\ddagger$ と $\boldsymbol{X}^\dagger$ の最小累積距離を計算するときに ${\boldsymbol{x}^\ddagger}_t$ と ${\boldsymbol{x}^\dagger}_{\pi(t)}$ が対応し，さらに ${\boldsymbol{x}^\ddagger}_t$ が $(\varphi_{y,t}, \theta_{y,t})$ と対応することをふまえて，$(\varphi_{y,t}, \theta_{y,t})$ を ${\boldsymbol{x}^\dagger}_{\pi(t)}$ に対する最適な状態対であると考える．そこで 2 つの仮定を行う．

(仮定 1)　$T^\ddagger$ と $\pi(T^\ddagger)$ は大きく変わらない．

(仮定 2)　$\boldsymbol{X}^\dagger$ はクラス境界に十分近い．

(仮定 1) より式 (23) の右辺の分子に存在する $[{\boldsymbol{x}^\dagger}_{\pi(1)}, \cdots, {\boldsymbol{x}^\dagger}_{\pi(t)}, \cdots, {\boldsymbol{x}^\dagger}_{\pi(T^\ddagger)}]$ を $[{\boldsymbol{x}^\dagger}_1, \cdots, {\boldsymbol{x}^\dagger}_t, \cdots, {\boldsymbol{x}^\dagger}_{T^\dagger}]$ に近似的に置き換えることができる．また，(仮定 2) より $\boldsymbol{X}^\ddagger$ に関する連結モデルの最適経路であった $(\varphi_{y,t}\ \ \theta_{y,t})$ を $\boldsymbol{X}^\dagger$ に関するものと近似的に見なすことが可能となる．さらに，$\boldsymbol{X}^\dagger$ が正分類であるため，式 (23) は以下のように近似的に表すことができる．

$$\rho \approx (-1) \times \frac{\sqrt{T^\dagger} d_y(\boldsymbol{X}^\dagger; \boldsymbol{\Lambda})}{2\sqrt{\sum_{t=1}^{T^\dagger} \| \boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)} \|^2}} \tag{24}$$

3.2 誤分類尺度と平滑な分類誤り数損失

式 (24) から，符号を反転した幾何マージンは，モデルベクトル系列間の DTW に基づく “近さ” によって正規化された誤分類尺度 $d_y(\boldsymbol{X}; \boldsymbol{\Lambda})$ にほかならないことが分かる．そこで，この符号を反転した幾何マージンを，

$$D_y(\boldsymbol{X}; \boldsymbol{\Lambda}) = -\rho \tag{25}$$

のように新たに誤分類尺度 $D_y(\boldsymbol{X}; \boldsymbol{\Lambda})$ として定義し，それを用いた MCE 学習を行う．幾何マージンを誤分類尺度とする MCE 学習は，分類判断の信頼度をより向上させうるはずである．すなわち，学習の進展とともに，分類判断結果は，誤分類尺度の負の領域において，その絶対値がより大きな領域に写像されることになる．結果的に，符号を反転した幾何マージンを誤分類尺度とする MCE 学習は，平滑な分類誤り数損失の最小化がそのまま幾何マージンの増大化を導き，この同時最適化を通して理想的な分類器パラメータの状態を追及する．

基本的に，FM-MCE 学習法と LGM-MCE 学習法の違いは，上記のような誤分類尺度の違いのみである．この違いに基づき，FM-MCE 学習法における式 (6) と式 (7) における $d_y(\boldsymbol{X}; \boldsymbol{\Lambda})$ を $D_y(\boldsymbol{X}; \boldsymbol{\Lambda})$ によって置き換え，LGM-MCE 学習法のための平滑な分類誤り数損失

$$\ell_y(\boldsymbol{X}; \boldsymbol{\Lambda}) = \frac{1}{1 + \exp(-\alpha D_y(\boldsymbol{X}; \boldsymbol{\Lambda}))} \tag{26}$$

と経験的平均損失

$$L(\boldsymbol{\Lambda}) = \frac{1}{N} \sum_{n=1}^{N} \ell_{y_n}(\boldsymbol{X}_n; \boldsymbol{\Lambda}) \tag{27}$$

を得る．なお，LGM-MCE 学習法における $\ell_y(\boldsymbol{X}; \boldsymbol{\Lambda})$ は，直接的には $d_y(\boldsymbol{X}; \boldsymbol{\Lambda})$ ではなく $D_y(\boldsymbol{X}; \boldsymbol{\Lambda})$ の関数であることに注意が必要である．

式 (24) の定義から，誤分類尺度 $D_y(\boldsymbol{X}; \boldsymbol{\Lambda})$ の負の領域における絶対値が，可変長パターン $\boldsymbol{X}$ とクラス境界との距離，すなわち幾何マージンであった．したがって，LGM-MCE 学習による式 (27) 中の $L(\boldsymbol{\Lambda})$ の最小化は，同時に，$\boldsymbol{\Lambda}$ に対応する分類判断が，負の領域における $D_y(\boldsymbol{X}; \boldsymbol{\Lambda})$ の値の絶対値がより大きくなるように $\boldsymbol{\Lambda}$ を更新する．こうして，LGM-MCE 学習法は，平滑な分類誤り数損失からなる経験的平均損失の最小化と幾何マージンの最大化とを同時に進めることになる．

3.3 確率的降下法による経験的平均損失の最小化

LGM-MCE 学習法における経験的平均損失の最小化（同時に幾何マージンの最大化も含む）の手続きも，基本的には FM-MCE 学習法におけるものと同様に導出でき，PD 法による更新式は，誤分類尺度の定義の違いに基づき更新量を決定する微分の結果が異なる（後述）ものの，形式的には式 (8) と同一となる．便利のため，更新式を再掲する．

$$\boldsymbol{r}_i^{h,s(m+1)} = \boldsymbol{r}_i^{h,s(m)} - \epsilon_m \nabla_{\boldsymbol{r}_i^{h,s}} \ell_y(\boldsymbol{X}; \boldsymbol{\Lambda}^{(m)}). \tag{28}$$

式 (28) による LGM-MCE 学習は，式中の第 2 項によって特徴づけられる．すなわち，$\nabla_{\boldsymbol{r}_i^{h,s}} \ell_y(\boldsymbol{X}; \boldsymbol{\Lambda})$ は

$$\nabla_{\boldsymbol{r}_i^{h,s}}\ell_y(\boldsymbol{X};\boldsymbol{\Lambda}) = \frac{\partial\ell_y(\boldsymbol{X};\boldsymbol{\Lambda})}{\partial D_y(\boldsymbol{X};\boldsymbol{\Lambda})}\cdot\left\{\frac{\sqrt{T}}{N(\boldsymbol{\Lambda})}\cdot\left(\frac{\partial d_y(\boldsymbol{X};\boldsymbol{\Lambda})}{\partial g_y(\boldsymbol{X};\boldsymbol{\Lambda})}\cdot\nabla_{\boldsymbol{r}_i^{h,s}}g_y(\boldsymbol{X};\boldsymbol{\Lambda}) + \frac{\partial d_y(\boldsymbol{X};\boldsymbol{\Lambda})}{\partial g_{y^*}(\boldsymbol{X};\boldsymbol{\Lambda})}\cdot\nabla_{\boldsymbol{r}_i^{h,s}}g_{y^*}(\boldsymbol{X};\boldsymbol{\Lambda})\right) - \frac{\sqrt{T}d_y(\boldsymbol{x}_1^T;\boldsymbol{\Lambda})}{N(\boldsymbol{\Lambda})^2}\cdot\nabla_{\boldsymbol{r}_i^{h,s}}N(\boldsymbol{\Lambda})\right\} \tag{29}$$

となる．ここで

$$N(\boldsymbol{\Lambda}) = 2\sqrt{\sum_{t=1}^{T}\left\|\boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)}\right\|^2} \tag{30}$$

であり，さらに

$$\frac{\partial\ell_y(\boldsymbol{X};\boldsymbol{\Lambda})}{\partial D_y(\boldsymbol{X};\boldsymbol{\Lambda})} = \alpha\ell_y(\boldsymbol{X};\boldsymbol{\Lambda})\{1-\ell_y(\boldsymbol{X};\boldsymbol{\Lambda})\}, \tag{31}$$

$$\nabla_{\boldsymbol{r}_i^{h,s}}N(\boldsymbol{\Lambda}) = \frac{2\sum_{t=1}^{T}(Q)}{\sqrt{\sum_{t=1}^{T}\left\|\boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)}\right\|^2}}, \tag{32}$$

$$Q = \{\delta(\varphi_{y,t}-h)\delta(\theta_{y,t}-s)\delta(i(\varphi_{y,t},\theta_{y,t},t)-i) - \delta(\varphi_{y^*,t}-h)\delta(\theta_{y^*,t}-s)\delta(i(\varphi_{y^*,t},\theta_{y^*,t},t)-i)\} \times\left(\boldsymbol{r}^{\varphi_{y,t}\ \theta_{y,t}}_{i(\varphi_{y,t},\theta_{y,t},t)} - \boldsymbol{r}^{\varphi_{y^*,t}\ \theta_{y^*,t}}_{i(\varphi_{y^*,t},\theta_{y^*,t},t)}\right) \tag{33}$$

である．

結果的に，LGM-MCE 学習法における更新式 (28) は，FM-MCE 学習における式 (9) と異なる式 (29) の $\nabla_{\boldsymbol{r}_i^{h,s}}\ell_y(\boldsymbol{X};\boldsymbol{\Lambda})$ をプロトタイプの更新量の決定に用いることで，FM-MCE 学習法における更新式 (8) とは異なるプロトタイプの状態をもたらすことになる．

4. 評価実験

4.1 音声データ

評価は，ETL-WD-I&Ⅱ データベースと東北大–松下単語（TMW: Tohoku University Matsushita Word）音声データベースを用いる 2 種の孤立単語音声認識課題と，TIMIT データベースを用いる 1 種の音素認識課題とを用いて行った．各データベースの主な仕様は以下のとおりである．

ETL-WD-I&Ⅱ データベース[*5]は，20 名の発話者（男女各 10 名）による 1,542 語の読み上げ単語音声で構成されている．ただし，一部の音声が収録されておらず，実験には，男声と女声がともに存在する 984 単語の音声データを用いた．したがって，課題は，各クラスのパターン標本数が 20 の，984 クラスの単語分類課題（標本総数は 19,680）とした．

TMW データベース[*6]は，60 名の発話者（男女各 30 名）による，212 語の音韻バランス単語の読み上げ音声で構成されている．ここでも一部の音声が収録されておらず，課題は，各クラスのパターン標本数をおよそ 60 とする，212 クラスの単語分類課題（標本総数は 12,649）とした．

TIMIT データベース[*7]は，486 名の英語発話者（男女それぞれ 342 名と 144 名）による連続文音声データからなる．付属の，音素単位の書き起こしと音声波形の音素境界情報とを用いて，54 クラスの音素認識用のデータベースを作成した．このデータベースの標本総数は 157,699 となった．

なお，学習時に適切に選定されるべきハイパーパラメータ（4.4 節参照）の不適切な設定は，過学習に対して，副作用的で解析が困難な影響を及ぼす．本稿における研究のように過学習現象の分析を目指す場合，そうした副作用的な影響は避ける必要がある．また，そうしたハイパーパラメータの設定には，多数回の学習実験の繰返しを必要とする．こうした点を考慮し，ハイパーパラメータの設定を十分に吟味できるように，1 回の学習実験に要する時間があまり膨大にならない程度の規模のデータベースを選択した．

4.2 音響特徴表現とクラスモデル

各時間窓位置（前出のフレームに対応）における音響特徴量には，12 次元のメル周波数ケプストラム係数（MFCC: Mel-Frequency Cepstrum Coefficient）とその窓内の音声信号のパワーとからなる 13 次元の変数ベクトルと，さらにそれらの各変数の 5 フレーム内の変動を近似する線形回帰直線の傾きからなる 13 次元ベクトルを合わせた，合計 26 次元ベクトルのベクトルを用いた．したがって，入力音声パターンは，この 26 次元の音響特徴ベクトル系列として表した．

本実験に用いた各音素の状態遷移モデルは，それぞれ 3 状態からなり（$S=3$），各状態に 3 つのプロトタイプを配置した（$I_{h,s}=3$）．ただし，孤立単語音声認識において無音区間を表す “sil” については，1 状態モデルを用い，その状態におけるプロトタイプ数も 1 とした．各状態遷移モデルのプロトタイプの初期化にはセグメンタル K 平均法を用いた [20]．

4.3 学習と試験のためのデータ分割法

過学習にかかわる学習法の特性を調べるため，各課題における音声データは，学習用標本群と試験用標本群とに分割して，それぞれを学習とその結果の試験（評価）に供した（Hold-Out 法）．**表 1** に，各データベースにおける学習用と試験用との標本分割の詳細を示す．いずれのデータベースにおいても，学習用標本群と試験用標本群は，互いに独立である．なお一般に，学習された認識器が取り扱う試験

[*5] http://research.nii.ac.jp/src/ETL-WD.html
[*6] http://research.nii.ac.jp/src/TMW.html
[*7] https://catalog.ldc.upenn.edu/ldc93s1.html

表 1 Hold-Out 法におけるデータ分割
Table 1 Data splitting for hold-out evaluation scheme.

データベース名	学習用標本数	試験用標本数
ETL-WD-I&Ⅱ	5,904	13,776
TMW	4,233	8,416
TIMIT	149,967	7,732

用標本の数に比べれば，学習に利用可能な学習用標本の数は小さい．この点を考慮し，ETL-WD-I&Ⅱは学習用標本に話者 6 名分，試験用標本に話者 14 名分を設定し，TMW は学習用標本に話者 20 名分，試験用標本に話者 40 名分を設定することで，学習用標本数よりも試験用標本数を大きくなるようにしている．一方，TIMIT データベースに関しては，広く定着している学習用標本群と Core Test と呼ばれる試験用標本群との分割をそのまま採用した．

便利のため，以下では，学習用標本集合を用いた評価を Closed Test と，未知の試験用標本集合を用いた評価のことを Open Test と呼ぶこととする．

4.4 ハイパーパラメータの設定

多くの学習法と同様に，FM-MCE 学習法も LGM-MCE 学習法も，学習対象である $\mathbf{\Lambda}$ のほかに学習の進行を制御するハイパーパラメータを持つ．一般に，ハイパーパラメータは，学習用標本群とも試験用標本群とも独立な検証用標本群を用いて，その検証用標本群に対する分類精度が高くなるように最適設定される．しかし，そのようなデータの 3 分割は，各標本群の数を減らし，学習あるいは試験の信頼度を低下させうる [21]．実際，分割の仕方によって，それぞれの標本群に対する分類精度は大きく変動し [21]，この 3 分割法は必ずしも合理的とは思えない．この点を考慮し，本稿の実験では，検証用標本群は準備せず，学習用標本群に対して最も高い分類精度が出るようハイパーパラメータを設定し，そうして得られた分類器の試験用標本上の分類精度を通して過学習特性を調査した．なお，3 分割法を用いた場合であっても，LGM-MCE 学習法が FM-MCE 学習法よりも高い過学習抑制力を持つことは，固定次元パターンの実験において明らかにされている [12]．

両 MCE 学習法におけるハイパーパラメータは，仮想的に学習用標本を増やす効果を持つ，平滑分類誤り数損失の損失平滑度と，経験的平均損失の最小化における収束を制御する，学習係数と最大エポック数とである．ここでエポックとは，PD 法において，全学習用標本を（通常，無作為順で）1 度ずつパラメータ更新に供する過程を指す．したがって，学習用標本は，最大エポック数分，繰返し学習に利用されることになる．

まず，損失平滑度と学習係数に関しては，**表 2** に示すように，データベース別に一定の探索範囲の中を試行的に設定し，最良の（学習用標本群に対して最も高い分類精度を出す）値を探索した．なお，それぞれの探索範囲と探索の粒度（表中の括弧内の数字）は，予備実験によって設定した．

表 2 ハイパーパラメータ設定の概要
Table 2 Overview of hyper-parameter settings.

データベース名	パラメータ	FM-MCE	LGM-MCE
ETL-WD-I&Ⅱ	損失平滑度	0.5-8.0 (0.5)	5-50 (5)
	学習係数	0.5-4.0 (0.5)	0.4-2.0 (0.2)
TMW	損失平滑度	0.5-5.0 (0.5)	5-50 (5)
	学習係数	0.1-1.5 (0.2)	0.1-1.5 (0.2)
TIMIT	損失平滑度	1.0-2.0 (0.5)	10-20 (5)
	学習係数	0.05, 0.1	0.05, 0.1

一方，最大エポック数に関しては，予備実験を通して選択した，いずれのデータベースに関する学習においても経験的平均損失値が十分に収束できる値，すなわち，ETL-WD-I&Ⅱおよび TMW では 100 とし，TIMIT では 500 とした．

本稿では前述したように試行錯誤的に損失平滑度と学習係数の設定を行ったが，この設定方法は多くの時間がかかるという問題がある．その解決を目指し，損失平滑度に関しては，誤分類尺度空間における確率分布の最尤推定を通して，それを自動的に選定する手法が提案されている [22]．しかし，学習係数に関しては，その設定の理論的基盤が必ずしも明確でなく，その自動化は容易でないように考えられる．こうした状況を受け，経験的ではあるものの種々の実験で有効性が示されている，学習係数を自動的に調整する機構を持つ損失最小化法，RPROP 法を FM-MCE 学習法と LGM-MCE 学習法とに適用し，その効果を調査した．この結果は，1 章にも述べたように付録において紹介する．

4.5 結果と考察

学習が，過学習を避け，最小分類誤り確率状態の優れた推定を行った場合，得られた分類器は，学習用標本群に対する分類精度（Closed Test の結果）を過剰に高めることなく，むしろ，未知なる試験用標本群に対する分類精度（Open Test の結果）を高めることが期待される．したがって，上述のように，学習用標本群に対する分類精度が上がるようにハイパーパラメータの設定を行った学習を通して得られた Open Test の結果を観測することで，学習手続きの最小分類誤り確率状態の推定能力を評価することができる．

以下に 3 種の認識課題ごとに，結果を表にまとめる．いずれの表においても，各学習法に対する分類精度は，学習用標本群に対して最も高い分類精度を達成するように学習された分類器を用いて得たものである．

ETL-WD-I&Ⅱに対する実験結果を**表 3** に示す．表 3 は表 2 で示すハイパーパラメータの組合せである 128 回の実験（FM-MCE 学習法の場合）と 90 回の実験（LGM-MCE

表 3 ETL-WD-I&Ⅱにおける分類率（%）
Table 3 Classification rates of ETL-WD-I&II.

	FM-MCE	LGM-MCE
Closed Test	99.98	99.98
Open Test	90.53	**92.00**

表 4 TMWにおける分類率（%）
Table 4 Classification rates of TMW.

	FM-MCE	LGM-MCE
Closed Test	100.00	100.00
Open Test (Max)	96.49	**97.37**
Open Test (Ave)	96.25	**97.04**

表 5 TIMITにおける分類率（%）
Table 5 Classification rates of TIMIT.

	FM-MCE	LGM-MCE
Closed Test	**73.89**	73.09
Open Test	66.59	**67.76**

学習法の場合）の中でClosed Testの結果が一番高いものを表にしている．表3から，Closed Testにおける分類率の差はないものの，Open Testの分類率において，LGM-MCE学習法がFM-MCE学習法を上回っていることが分かる．目指す最小分類誤り確率状態の推定に関し，LGM-MCE学習法の方がFM-MCE学習法よりも勝っていることを読み取ることができる．

次に，TMWに対する実験結果を**表 4**に示す．この実験では，各学習法に対してハイパーパラメータの組合せ数分の80回ずつの実験を行った．しかしながら，ETL-WD-I&Ⅱの実験とは異なり，FM-MCE学習法で12個の，LGM-MCE学習法で8個の組合せにおいて，Closed Testの分類精度が100%に達成する結果が存在した．そこで，表中のOpen Test (Max)には，それらの複数の結果の中で，試験用標本群に対する最も高い分類精度を示し，またOpen Test (Ave)には，それらの複数の結果のそれぞれに対応する試験用標本群に対する分類精度の平均を示した．

TMWデータベースは212クラス問題であるため，ETL-WD-I&Ⅱデータベースの984クラス問題と比べて容易なタスクと推測された．このことは，FM-MCE学習法とLGM-MCE学習法の両手法において，分類精度が100%を達成するハイパーパラメータの組合せが複数あったことからもうかがい知ることができる．表から，Open Test (Max)とOpen Test (Ave)のいずれにおいても，LGM-MCE学習法の優位性を読み取ることができる．また特にOpen Test (Ave)における優位性は，LGM-MCE学習法の効果の安定性を示唆するものと考えられる．

続いて，TIMITに対する実験結果を**表 5**に示す．この実験では，エポック数を500とした関係で，先の2つの実験と比べると少ない6回ずつの実験であるが，ETL-WD-I&Ⅱと同様にその中でClosed Testの結果が一番高いものを表にしている．表のClosed Testの結果が73%程度であり，先の2つの課題と比べて難しいことがうかがえる．またこの課題では，先行する2つの課題と比べて圧倒的に多数の学習用標本を用いている．このような，分類が難しく，かつ学習の規模が大きな課題においてもまた，表のOpen Testの結果はLGM-MCE学習法の優位性を示している．

表 6 3種のデータベースにおける3つの上位分類率の平均と標準偏差（%）
Table 6 Averages and standard devisions of classification rates for 3 databases.

	FM-MCE	LGM-MCE
ETL-WD (Closed Test)	**99.98** (0.01)	99.97 (0.01)
ETL-WD (Open Test)	91.11 (0.99)	**92.90** (0.68)
TMW (Closed Test)	100.00 (0.00)	100.00 (0.00)
TMW (Open Test)	96.49 (0.00)	**97.26** (0.10)
TIMIT (Closed Test)	**73.80** (0.08)	72.85 (0.28)
TIMIT (Open Test)	66.96 (0.26)	**67.55** (0.25)

表3から表5までの結果は，学習用標本群に対して最も高い分類精度を出した（TMWデータベースに関する一部を除いて）単独の学習結果に基づくものであった．観測の信頼性を高めるため，**表 6**に，それぞれのデータベースに関する実験において，FM-MCE学習法あるいはLGM-MCE学習法が学習用標本群に対してもたらした上位3つの分類精度に着目し，それらの平均と標準偏差を，それらの学習結果に対応する試験用標本群に対する分類率の平均と標準偏差とともにとりまとめた．表から，この複数の学習結果においても，LGM-MCE学習法はFM-MCE学習法よりも高い試験用標本分類率を達成し，安定的に過学習抑制の効果を発揮していることを読み取ることができる．

本研究で採用したPD法などのパラメータ更新の繰返しをともなう学習においては，その繰返しの数，すなわちエポック数の不足が，副作用的に過学習抑制効果をもたらすことがある[*8]．本実験が，そうした副作用に影響されることなく，FM-MCE学習法とLGM-MCE学習法それぞれの本来の学習性能を検証していることを確認するため，TMWデータベースに関する表4中の“Open Test (Max)”欄，すなわち学習用標本群に対して100%の分類率を達成した場合の中で試験用標本群に対して最も高い分類率をもたらした場合に注目し，その学習が，学習エポックの増加とともにもたらした分類率の変化を観察した（**図 3**）．図中，横軸はエポック数であり，縦軸は分類誤り率である．また，橙色実線と赤色実線は，それぞれLGM-MCE学習法の学習用標本群と試験用標本群に対する分類誤り率を示し，水色実線と青色実線は，それぞれFM-MCE学習法の学習用標本群と試験用標本群に対する分類誤り率を示している．

[*8] 広く用いられている正則化は，この副作用的効果を積極的に利用するものである．

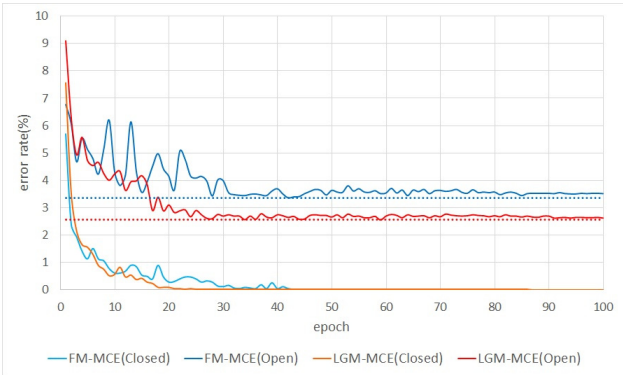


図 3 学習の進捗にともなう分類精度の推移

Fig. 3 Changes in classification error rates along training progress.

さらに，赤色点線（直線）は LGM-MCE 学習法によって試験用標本群に対して達成された最小の誤り率を，青色点線（直線）は FM-MCE 学習法によって試験用標本群に対して達成された最小の誤り率を示している．図から，いずれの学習法も，学習用標本群に対して 40 エポック前後で 0%の誤り率を達成する一方で，試験用標本群に対しては，LGM-MCE 学習法が到達できた最小の分類誤り率付近に安定的に収束しているのに対し，FM-MCE 学習法は，比較的早いエポック（40 エポック周辺）においていったん最小分類誤り率に達した後，わずかながら誤り率が増加してそのまま収束するという，過学習にしばしば見られる現象が現れていることが分かる．これらの結果から，いずれの学習法も十分に学習が繰り返され，その結果には不測の副作用的な影響は入っていないことも，FM-MCE 学習法の収束には過学習現象が現れているのに対し，LGM-MCE 学習法にはそれが見られなかったことも理解できる．

パラメータ更新の不足が副作用的に過学習抑制効果を生み出しうることを述べたが，分類器のクラスモデルの表現力の不足がやはり副作用的に過学習を抑えることがある．本研究では，この点も考慮したうえで比較的標準的なクラスモデル（3 状態状態遷移モデル）を採用した．実際，表 3 から表 6 の結果に見られるように，学習用標本群と試験用標本群との分類率には大きな乖離があり，用いた 3 種のデータのいずれの実験においても過学習が発生していたものと考えられる．なお，LGM-MCE 学習法の過学習抑制は，分類誤り数損失の平滑性にともなって生み出される，学習標本周辺における仮想標本生成効果によることが分かっている [23]．したがって，過少な学習標本を用いる学習においては，標本分布がまばらになりすぎ，その抑制効果が現れにくくなるおそれがある．こうした問題の有無を確認するため，TMW データを用いた追加の調査を行った．表 4 の実験では 20 名分のデータを学習用（学習用標本数：4,233）に，また 40 名分を試験用（試験標本数：8,416）に用いたのに対し，ここでは 10 名分を学習用（学習用標本数：2,114）に，50 名分を試験用（試験用標本数：10,535）に用いて実験を行った．ハイパーパラメータ設定に関しては，FM-MCE 学習法においては損失平滑度を 1.0～4.0（0.5 刻み）の範囲で，学習係数を 0.5～1.1（0.2 刻み）の範囲で調査し，LGM-MCE 学習法においては損失平滑度を 5～30（5 刻み）の範囲で，学習係数を 0.7～1.3（0.2 刻み）の範囲で調査した．こうして得られた 28 個の FM-MCE 学習法の結果と 24 個の LGM-MCE 学習法の結果の中から，学習用標本群に対して最も高い分類率（100%であった）を達成した分類器を用いて得た分類率と，学習用標本群に対して 100%を達成した分類器のそれぞれがもたらした試験用標本群に対する分類率の平均を**表 7** に示す．表から分かるように，この表 1 とは異なる（より厳しい過学習状況の発生が予想される）データ分割においても，LGM-MCE 学習法は，FM-MCE 学習法より安定的に高い試験用標本分類率を達成し，過学習抑制効果を発揮していることを読み取ることができる．

表 7 TMW における分類率（学習用標本数：2,114，試験標本数：10,535）（%）．

Table 7 Classification rates of TMW(Training samples: 2,114, Testing samples: 10,535).

	FM-MCE	LGM-MCE
Closed Test	100.00	100.00
Open Test (Max)	95.06	**95.84**
Open Test (Ave)	94.61	**95.51**

5. おわりに

可変長時系列パターンの分類においては必ずしも十分に評価されていなかった LGM-MCE 学習法について，状態遷移モデル型分類器を用いた実装を行い，3 種の音声認識課題においてその実験的な調査を行った．実験では，利用可能な標本を学習用標本群と試験用標本群とに 2 分割し，学習用標本群を用いて分類器クラスモデルパラメータとハイパーパラメータの双方を最適化する学習を行い，そうして得られた分類器の試験用標本群に対する分類精度によって学習法の評価を行った．その結果，従来型の FM-MCE 学習法と比べ，提案した LGM-MCE 学習法が，未知標本に対して安定的に高い分類精度を達成しうる，すなわち，学習用標本に対する過学習が抑制された（最小分類誤り確率状態により近い）分類器クラスモデルパラメータの正確な推定を行いうることを確認することができた．

最近の音声認識の研究においては，本稿で用いた（HMM も含む）状態遷移モデル型クラスモデルの前段に（一般にクラス共通の）深層ニューラルネットワーク（DNN: Deep Neural Network）を備えた，ハイブリッド DNN-HMM 型音声認識器の利用が注目を集めている [24]．この新しく強力なハイブリッド型音声認識器を用いて LGM-MCE 学習法の評価を行うべきとも考えられる．多くの場合，そうし

たハイブリッド型音声認識器は，その規模の大きさゆえに，異なる学習基準を用いた部分的な学習の組合せによって学習される．すなわち，まず後段部分の HMM クラスモデルが，たとえば単語クラスのような認識器出力のレベルで学習され，その学習結果を利用して，遡って前段部分の DNN 特徴変換部が，単語よりも小さなセノンのようなクラスのレベルで学習される．しかし，様々な要因が複雑に影響し，そのような異なる基準を用いた学習の組合せは，必ずしも最終段階の性能向上につながらない．これは，認識器全体の統合的な最適化を目指す End-to-End 学習が継続的に研究されているゆえんでもある（文献 [25], [26] など）．また，こうした強力な音声認識器でさえも話者や発話環境の多様性を必ずしも十分には獲得できず，ハイブリッド DNN-HMM 型音声認識器のための話者適応技術などもさかんに研究されている（文献 [27], [28] など）．基本的に，ハイブリッド DNN-HMM 型音声認識器は多数の学習対象モデルパラメータを持つ一方で，適応学習用標本の量は限られる．したがって，適応学習における過学習の抑制はいっそう重要となる．本稿で提案した可変長パターン分類用の LGM-MCE 学習法を，こうした統合的学習や適応学習の文脈において研究することは，次の重要な課題と思われる．

なお，1 章でも触れ，付録で実験結果を紹介してもいるように，LGM-MCE 学習法が採用している PD 法における学習係数の最適化は，必ずしも容易ではない[*9]．その結果として，実験に多くの時間を要するばかりか，学習法が本来達成しうる性能を十分に引き出すこと自体も難しくなっている．したがって，この学習係数を合理的に設定する手法の開発が望まれる．また，本稿で用いた可変長パターンのための DTW 型距離が最良である保証はない．より適切な距離空間における幾何マージンを増大させることによって，過学習抑制効果をいっそう高めうる可能性がある．これらの点も，重要な課題であると思われる．

謝辞 本研究の一部は，科研費（JP26280063）および文部科学省平成 26 年度私立大学戦略的研究基盤形成支援事業・進化適応型自動車運転支援システム「ドライバ・イン・ザ・ループ」研究拠点形成の支援を受けて行われた．また，ETL-WD-I&Ⅱ データベースと東北大–松下単語音声データベースは，国立情報学研究所音声資源コンソーシアムからの提供を受けた．著者一同，ご支援に感謝いたします．

参考文献

[1] Duda, R.O., Hart, P.E., Stork, D.G.（著），尾上守夫（監訳）：パターン識別，新技術コミュニケーションズ，東京 (2001).
[2] Bishop, C.M.（著），元田 浩，栗田多喜夫，樋口知之，松本裕治，村田 昇（監訳）：パターン認識と機械学習，シュプリンガー・ジャパン，東京 (2007).
[3] Nilsson, N.: *The Mathmatical Foundations of Learning Machines*, Morgan Kaufmann, San Mateo (1990).
[4] Jiang, H.: Discriminative Training of HMMs for Automatic Speech Recognition: A Survey, *Comput. Speech Lang.*, Vol.24, No.4, pp.589–608 (2010).
[5] Schluter, R., Macherey, W., Muller, B. and Ney, H.: Comparison of Discriminative Training Criteria and Optimization Methods for Speech Recognition, *Speech Commun.*, Vol.34, pp.287–310 (June 2001).
[6] Juang, B.-H. and Katagiri, S.: Discriminative Learning for Minimum Error Classification, *IEEE Trans. Signal Process.*, Vol.40, No.12, pp.3043–3054 (1992).
[7] Liu, C., Jiang, H. and Rigazio, L.: Recent Improvement on Maximum Relative Margin Estimation of HMMs for Speech Recognition, *Proc. ICASSP*, Vol.1, pp.269–272 (May 2006).
[8] Yu, D., Deng, L., He, X. and Acero, A.: Large-Margin Minimum Classification Error Training: A Theoretical Risk Minimization Perspective, *Comput. Speech Lang.*, Vol.22, pp.415–429 (Oct. 2008).
[9] Vapnik, V.N.: *The Nature of Statistical Learning Theory,* Springer-Verlag, New York (1995).
[10] Cristianini, N.，Shawe-Taylor, J.（著），大北 剛（訳）：サポートベクターマシン入門，共立出版，東京 (2005).
[11] He, T. and Huo, Q.: A Study of A New Misclassification Measure for Minimum Classification Error Training of Prototype-Based Pattern Classifiers, *Proc. ICPR* (Dec. 2008).
[12] 渡辺秀行, 片桐 滋, 山田幸太, マクダーモット・エリック, 中村 篤，渡部普治，大崎美穂：幾何マージンに基づく誤分類尺度を用いた最小分類誤り学習法，電子情報通信学会論文誌 D，Vol.J94-D, No.10, pp.1664–1675 (2011).
[13] Amari, S.: A Theory of Adaptive Pattern Classifiers, *IEEE Trans. Electron. Comput.*, Vol.EC-16, pp.299–307 (Mar. 1967).
[14] Riedmiller, M. and Braun, H.: A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm, *Proc. ICNN*, pp.586–591 (1993).
[15] Igel, C. and Husken, M.: Improving the Rprop Learning Algorithm, *Proc. NC2000*, pp.115–121 (2000).
[16] McDermott, E. and Katagiri, S.: Prototype-based Dsicriminative Training for Various Speech Units, *Proc. ICASSP*, Vol.1, pp.417–420 (Mar. 1992).
[17] 橋本哲也, 北岡見生代, 渡辺秀行, 片桐 滋, ル・シュガン, 堀 智織，大崎美穂：大幾何マージン最小分類誤り学習法を用いた音声パターン認識，日本音響学会講演論文集，1-P-25, pp.165–168 (Mar. 2015).
[18] Kitaoka, M., Hashimoto, T., Ochiai, T., Katagiri, S., Ohsaki, M., Watanabe, H., Lu, X. and Kawai, H.: Speech Pattern Classification Using Large Geometric Margin Minimum Classification Error Training, *Proc. TENCON* (Nov. 2015)
[19] 松廣達也，北岡見生代，ア・デイビッド，渡辺秀行，片桐滋，大崎美穂：大幾何マージン最小分類誤り学習法を用いた音声認識に関する実験的評価，情報処理学会関西支部大会，G-08 (Sep. 2016).
[20] Juang, B.-H. and Rabiner, L.R.: The segmental k-means algorithm for estimating parameters of hidden markov models, *IEEE Trans. Acoustics, Speech and Signal Process*, Vol.38, No.9, pp.1639–1641 (1990).
[21] 白石裕之，渡辺秀行，片桐 滋，ル シュガン，堀 智織，大崎美穂：大幾何マージン最小分類誤り学習法におけるデータ分割法と未知標本耐性の関係について，電子情報通信学会信学技報，Vol.114, No.409, PRMU2014-101,

[*9] バッチ型の最急降下法でも基本的には同様であり，DNN の学習に用いられている学習法においてもこの問題は共通している．

pp.177–182 (2015).
[22] Watanabe, H., Tokuno, J., Ohashi, T., Katagiri, S., Ohsaki, M., Matsuda, S. and Kashioka, H.: Minimum Classification Error Training Incorporating Automatic Loss Smoothness Determination, *J. Signal Process. Syst.*, Vol.74, No.3, pp.311–322 (2014).
[23] Watanabe, H., Ohashi, T., Katagiri, S., Ohsaki, M., Matsuda, S. and Kashioka, H.: Robust and Efficient Pattern Classification using Large Geometric Margin Minimum Classification Error Training, *J. Signal Process. Syst.*, Vol.74, No.3, pp.297–310 (2014).
[24] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. and Kingsbury, B.: Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups, *IEEE Signal Process. Mag.*, Vol.29, No.6, pp.82–97 (2012).
[25] Biem, A., Katagiri, S. and Juang, B.-H.: Pattern Recognition Using Discriminative Feature Extraction, *IEEE Trans. Signal Process.*, Vol.45, No.2, pp.500–504 (1997).
[26] Chorowski, J., Bahdanau, D., Kyunghyun, C. and Bengio, Y.: Attention-based Models for Speech Recognition, *Proc. NIPS*, pp.577–585 (Dec. 2015).
[27] Gemello, R., Mana, F., Scanzio, S., Laface, P. and De Mori, R.: Linear Hidden Transformations for Adaptation of Hybrid ANN/HMM Models, *Speech Commun.*, Vol.49, No.10, pp.827–835 (2007).
[28] Ochiai, T., Matsuda, S., Watanabe, H., Lu, X., Hori, C., Kawai, H. and Katagiri, S.: Speaker Adaptive Training Localizing Speaker Modules in DNN for Hybrid DNN-HMM Speech Recognizers, *IEICE Trans. Inf. & Syst.*, Vol.E99-D, No.10, pp.2431–2443 (2016).

付　　録

A.1 MCE学習におけるRPROP法による損失最小化 [19]

PD法や最急降下法における学習係数の値は，実験による試行錯誤などを通して経験的に設定せざるをえない．その設定には多くの時間を要し，なんらかの改善法の開発が期待される．本稿では，学習係数の値を陽に設定する必要のないRPROP法 [14], [15] に着目し，それを用いたLGM-MCE学習法の（および，比較のためにFM-MCE学習法の）学習手続きを導出する．

A.1.1 概要

RPROP法は，以下のように勾配の符号のみを用いて，あらかじめ設定しておいたパラメータ更新量を最適な値に向けて変化させつつ，損失の超曲面を下るようにしてその局所的最小状態の発見を目指すものである．また，PD法とは異なり，RPROP法ではパラメータベクトルの要素ごとに更新量が変化する．

$$r_i^{h,s}{}_{(\mathrm{dim})}^{(m+1)} = r_i^{h,s}{}_{(\mathrm{dim})}^{(m)} - \mathrm{sign}\left(\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m)}\right) \times \triangle_i^{h,s}{}_{(\mathrm{dim})}^{(m)}, \tag{A.1}$$

Algorithm 1 RPROP$^+$ algorithm

for each $r_i^{h,s}{}_{(\mathrm{dim})}$ **do**
  if $\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m-1)} \times \frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m)} > 0$ **then**
    $\triangle_i^{h,s}{}_{(\mathrm{dim})}^{(m)} = \min\left(\triangle_i^{h,s}{}_{(\mathrm{dim})}^{(m-1)} \times \eta^+, \triangle_{max}\right)$
    $\triangle r_i^{h,s}{}_{(\mathrm{dim})}^{(m)} = -\mathrm{sign}\left(\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m)}\right) \times \triangle_i^{h,s}{}_{(\mathrm{dim})}^{(m)}$
    $r_i^{h,s}{}_{(\mathrm{dim})}^{(m+1)} = r_i^{h,s}{}_{(\mathrm{dim})}^{(m)} + \triangle r_i^{h,s}{}_{(\mathrm{dim})}^{(m)}$
  else if $\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m-1)} \times \frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m)} < 0$ **then**
    $\triangle_i^{h,s}{}_{(\mathrm{dim})}^{(m)} = \max\left(\triangle_i^{h,s}{}_{(\mathrm{dim})}^{(m-1)} \times \eta^-, \triangle_{min}\right)$
    $r_i^{h,s}{}_{(\mathrm{dim})}^{(m+1)} = r_i^{h,s}{}_{(\mathrm{dim})}^{(m)} - \triangle r_i^{h,s}{}_{(\mathrm{dim})}^{(m-1)}$
    $\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m)} = 0$
  else
    $\triangle_i^{h,s}{}_{(\mathrm{dim})}^{(m)} = \triangle_i^{h,s}{}_{(\mathrm{dim})}^{(m-1)}$
    $\triangle r_i^{h,s}{}_{(\mathrm{dim})}^{(m)} = -\mathrm{sign}\left(\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m)}\right) \times \triangle_i^{h,s}{}_{(\mathrm{dim})}^{(m)}$
    $r_i^{h,s}{}_{(\mathrm{dim})}^{(m+1)} = r_i^{h,s}{}_{(\mathrm{dim})}^{(m)} + \triangle r_i^{h,s}{}_{(\mathrm{dim})}^{(m)}$
  end if
end for

ここで $\triangle_i^{h,s}{}_{(\mathrm{dim})}^{(m)}$ は，更新繰返しの m ステップにおける h 番目の音素の s 番目の状態の i 番目のプロトタイプの dim 次元のパラメータに対する更新量を表している．この更新量は，たとえば現在 m ステップの更新であるとした場合，それに先行する $(m-1)$ ステップの勾配の符合と m ステップの勾配の符合が等しいときはさらに大きく曲面を下ることができると見なして更新量を増やし，符号が異なる場合は局小解を飛び越えたと見なして $(m-1)$ ステップのパラメータに戻し，さらに更新量も減らす処理を行うことで更新量を調整する．

RPROP法には複数の版があり，特にRPROP$^+$ と呼ばれるRPROP法を用いた [14], [15]．RPROP$^+$ の具体的な更新手続きを，式 (A.1) に基づく疑似コードとしてAlgorithm 1 に示す．疑似コード中の $\triangle_{max}$ は更新量の最大値であり，$\triangle_{min}$ は更新量の最小値である．また，$0 < \eta^- < 1 < \eta^+$ である．$\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}^{(m)}$ は m ステップにおける経験的平均損失を $r_i^{h,s}{}_{(\mathrm{dim})}$ で微分した値を表す．

A.1.2 パラメータ更新量

疑似コードの手順に従い，分類器パラメータの更新を行う．疑似コード中の $\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}}$ は鎖則によって以下のように表すことができる．

$$\frac{\partial L(\boldsymbol{\Lambda})}{\partial r_i^{h,s}{}_{(\mathrm{dim})}} = \frac{1}{N}\sum_{n=1}^{N} \nabla_{r_i^{h,s}{}_{(\mathrm{dim})}} \ell_{y_n}(\boldsymbol{X}_n; \boldsymbol{\Lambda}), \tag{A.2}$$

ここでパラメータ更新量 $\nabla_{r_i^{h,s}{}_{(\mathrm{dim})}} \ell_{y_n}(\boldsymbol{X}_n; \boldsymbol{\Lambda})$ は，式 (9) や式 (29) において，パラメータがベクトル $\boldsymbol{r}_i^{h,s}$ であると

きの dim 次元目の要素に関するものである．FM-MCE 学習法と LGM-MCE 学習法に対し，パラメータ更新量の具体的な定義はそれぞれ次のよう与えられる．

・FM-MCE 学習法の場合

$$
\begin{aligned}
&\nabla_{r_i^{h,s}{}_{(\mathrm{dim})}} \ell_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda}) \\
&= \alpha \ell_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda})\{1-\ell_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda})\} \\
&\times \Bigg(1 \times \Bigg\{ -\frac{2}{T}\sum_{t=1}^{T} \delta(\varphi_{y,t}-h)\delta(\theta_{y,t}-s) \\
&\times \delta(i(\varphi_{y,t},\theta_{y,t},t)-i)\left(x_t^n{}_{(\mathrm{dim})} - r_i^{h,s}{}_{(\mathrm{dim})}\right) \Bigg\} \\
&+ (-1) \times \Bigg\{ -\frac{2}{T}\sum_{t=1}^{T} \delta(\varphi_{y^*,t}-h)\delta(\theta_{y^*,t}-s) \\
&\times \delta(i(\varphi_{y^*,t},\theta_{y^*,t},t)-i)\left(x_t^n{}_{(\mathrm{dim})} - r_i^{h,s}{}_{(\mathrm{dim})}\right) \Bigg\} \Bigg).
\end{aligned}
\tag{A.3}
$$

・LGM-MCE 学習法の場合

$$
\begin{aligned}
&\nabla_{r_i^{h,s}{}_{(\mathrm{dim})}} \ell_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda}) \\
&= \alpha \ell_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda})\{1-\ell_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda})\} \\
&\times \Bigg\{ \frac{\sqrt{T}}{N(\boldsymbol{\Lambda})} \times \Bigg(1 \times \Bigg\{ -\frac{2}{T}\sum_{t=1}^{T} \delta(\varphi_{y_n,t}-h)\delta(\theta_{y_n,t}-s) \\
&\times \delta(i(\varphi_{y_n,t},\theta_{y_n,t},t)-i)\left(x_t^n{}_{(\mathrm{dim})} - r_i^{h,s}{}_{(\mathrm{dim})}\right) \Bigg\} \\
&+ (-1) \times \Bigg\{ -\frac{2}{T}\sum_{t=1}^{T} \delta(\varphi_{y_n{}^*,t}-h)\delta(\theta_{y_n{}^*,t}-s) \\
&\times \delta(i(\varphi_{y_n{}^*,t},\theta_{y_n{}^*,t},t)-i)\left(x_t^n{}_{(\mathrm{dim})} - r_i^{h,s}{}_{(\mathrm{dim})}\right) \Bigg\} \Bigg) \\
&- \frac{\sqrt{T} d_{y_n}(\boldsymbol{X}_n;\boldsymbol{\Lambda})}{N(\boldsymbol{\Lambda})^2} \\
&\times \frac{2\sum_{t=1}^{T}(Q')}{\sqrt{\sum_{t=1}^{T} \| \boldsymbol{r}^{\varphi_{y_n,t}\ \theta_{y_n,t}}_{i(\varphi_{y_n,t},\theta_{y_n,t},t)} - \boldsymbol{r}^{\varphi_{y_n{}^*,t}\ \theta_{y_n{}^*,t}}_{i(\varphi_{y_n{}^*,t},\theta_{y_n{}^*,t},t)} \|^2}} \Bigg\},
\end{aligned}
\tag{A.4}
$$

ここで，Q' は次式で表すものとする．

$$
\begin{aligned}
Q' &= \{\delta(\varphi_{y,t}-h)\delta(\theta_{y,t}-s)\delta(i(\varphi_{y,t},\theta_{y,t},t)-i) \\
&-\delta(\varphi_{y^*,t}-h)\delta(\theta_{y^*,t}-s)\delta(i(\varphi_{y^*,t},\theta_{y^*,t},t)-i)\} \\
&\times \left(r^{\varphi_{y_n,t}\ \theta_{y_n,t}}_{i(\varphi_{y_n,t},\theta_{y_n,t},t)_{(\mathrm{dim})}} - r^{\varphi_{y_n{}^*,t}\ \theta_{y_n{}^*,t}}_{i(\varphi_{y_n{}^*,t},\theta_{y_n{}^*,t},t)_{(\mathrm{dim})}} \right).
\end{aligned}
\tag{A.5}
$$

A.1.3 動作確認および評価

ETL-WD-I&Ⅱ データベースを用いて，RPROP 法による FM-MCE 学習法と LGM-MCE 学習法との動作確認を行った．

なお，RPROP 法では，学習係数の設定はなくなったものの，更新量 $\triangle_i^{h,s(m)}{}_{(\mathrm{dim})}$ の初期値の設定を行う必要がある．しかし，これまでの報告 [14], [15] から，学習結果に対するこの設定の感度は比較的低いとされている．この点も含めて，**表 A·1** に示す，損失平滑度と更新量の初期値とをハイパーパラメータとした学習実験を行った．表から分かるように，PD 法における学習係数の設定と比べ，更新量の初期値の設定はかなり粗い粒度で行った．最大エポック数は，PD 法と同じ 100 とした．

表 A·1 RPROP のためのハイパーパラメータ設定
Table A·1 Hyper-parameter settings for RPROP.

パラメータ	FM-MCE	LGM-MCE
損失平滑度	0.5-8.0 (0.5)	5-50 (5)
更新量の初期値	0.2-0.5 (0.1)	0.2-0.5 (0.1)

表 A·2 RPROP 法による分類率（%）
Table A·2 Classification rates by RPROP method.

	FM-MCE (PD)	LGM-MCE (PD)	FM-MCE (RPROP)	LGM-MCE (RPROP)
Closed Test	99.98	99.98	99.92	99.93
Open Test	90.53	92.00	92.61	92.15

表 A·2 に，PD 法の結果と比較する形で，RPROP 法による分類精度を示す．表から，FM-MCE 学習法と LGM-MCE 学習法のいずれにおいても，RPROP 法が PD 法とほぼ同等の Closed Test 結果を出していることが分かる．結果は，RPROP 法による学習が期待どおりに動作していることを示唆しているように思われる．なお，FM-MCE 学習法と LGM-MCE 学習法のいずれにおいても，RPROP 法の Closed Test の精度は PD 法のそれより若干低い．これは，RPROP 法における近似によるもののように考えられる．一方，RPROP 法による FM-MCE 学習法の Open Test の精度は，PD 法のそれを上回り，さらに LGM-MCE 学習法の RPROP 法の精度をも上回っていた．これは，学習における近似が，その副作用として過学習を抑制した結果であるように推察される．しかし，こうした副作用は，元々陽に制御することが難しく，RPROP 法を用いる MCE 学習法は，理想的な（最小分類誤り確率状態に対応する）分類器の実現を分析的に探求するための手法としては最良とはいい難い．

一方，パラメータの更新を全学習用標本に対してまとめて行う RPROP 法は，手続きの並列化による学習の高速化に明らかに適している．表 A·2 の結果から，RPROP 法を用いる MCE 学習が，そうした目的に十分に対応しうることが期待できる．

推薦文

関西支部では支部大会において優れた内容の論文に対し推薦論文を選定することとした．そこで，支部大会で発表

された論文のうち 6 ページに満たないものを除く 24 件を対象とし，各セッションの座長および実行委員から広く推薦を集めて候補論文を選出した．各論文に対し事後評価者 2 名の評価を加え，実行委員会による審議を経て 2 件の推薦論文候補を決定した．本論文では，可変長の音声パターン分類においては必ずしも十分に評価されていなかった大幾何マージン最小分類誤り（LGM-MCE: Large Geometric Margin-MCE）学習法について，単語音声パターンの分類実験を通して，実験的な評価を行っている．関数マージン最小分類誤り（FM-MCE: Functional Margin-MCE）学習法との比較を行うことで可変長の音声パターン分類における LGM-MCE 学習法の有用性を明らかにしており，評価に値する．

（情報処理学会関西支部支部長　安本慶一）

松廣 達也　（学生会員）

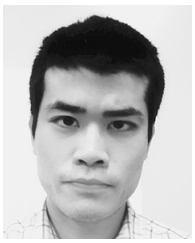
2016 年同志社大学理工学部情報システムデザイン学科卒業．現在，同大学院理工学研究科博士前期課程在学．電子情報通信学会学生会員．

橋本 哲也

2013 年同志社大学理工学部情報システムデザイン学科卒業．2015 年同大学院理工学研究科博士前期課程修了．現在，ダイキン工業株式会社に勤務．

北岡 見生代

2014 年同志社大学理工学部情報システムデザイン学科卒業．2016 年同大学院理工学研究科博士前期課程修了．現在，株式会社ラクスに勤務．



ア デイビッド　（正会員）

2016 年 Ecole Centrale de Lille 修士課程修了．2016 年同志社大学大学院理工学研究科博士前期課程修了．現在，同大学院理工学研究科博士後期課程在学．

落合 翼

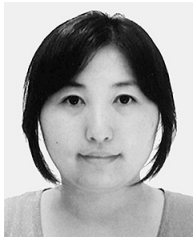
2013 年同志社大学理工学部情報システムデザイン学科卒業．2015 年同大学院理工学研究科博士前期課程修了．現在，同大学院理工学研究科博士後期課程在学．IEEE，日本音響学会，電子情報通信学会各会員．

渡辺 秀行

1988 年北海道大学工学部電子工学科卒業．1990 年同大学院修士課程修了．1993 年同博士課程修了．博士（工学）．1993 年国際電気通信基礎技術研究所（ATR）入社．2009 年独立行政法人情報通信研究機構（NICT）に移籍．2016 年より ATR 脳情報通信総合研究所連携研究員．パターン認識，音声情報処理，信号処理に関する研究に従事．電子情報通信学会，日本音響学会，IEEE 各会員．

片桐 滋　（正会員）

1977 年東北大学工学部電気工学科卒業．1979 年同大学院工学研究科修士課程修了．1982 年同大学院工学研究科博士課程修了．日本電信電話公社（現，日本電信電話株式会社），（株）国際電気通信基礎技術研究所を経て，現在，同志社大学理工学部教授．工学博士．IEEE，日本音響学会，電子情報通信学会各会員．



大崎 美穂　（正会員）

1994 年九州芸術工科大学（現，九州大学）芸術工学部音響設計学科卒業．1996 年同大学院芸術工学研究科博士前期課程修了．1999 年博士後期課程修了．静岡大学情報学部助手，同志社大学工学部専任講師，准教授を経て，現在，同志社大学理工学部教授．博士（工学）．IEEE，人工知能学会，知能情報ファジィ学会，音響学会各会員．