社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－ＳＬＰ－69（39）
2007／12／21

頑健なパラメタ推定のための Aggregated EM 法の提案と評価

篠崎 隆宏† Ostendorf, Mari†† 河原 達也†††

† 東京工業大学情報理工学研究科計算工学専攻
〒 152–8552 東京都目黒区大岡山 2-12-1-W8-77
†† Department of Electrical Engineering, University of Washington
Box 352500, Seattle, Washington, U.S.A. 98195-2500
††† 京都大学学術情報メディアセンター
〒 606–8501 京都市左京区吉田二本松町
E-mail: †shinot@furui.cs.titech.ac.jp

あらまし EM アルゴリズムの欠点である過学習の問題を補い、高い汎化能力を持つ学習アルゴリズムとして Aggregated EM 法の提案を行う。従来の EM アルゴリズムが学習ループ中で学習データ全体を用いた単一モデルの推定と尤度の評価を繰り返すのに対して、提案法では学習の各ステージにおいて Bagging と同様に学習セットの部分集合から複数のモデルを推定しそれらの結果を統合することで学習性能の向上を図る。提案手法は区分化した学習セットの各区画に対して求めた十分統計量を活用することで、効率的に動作する。人工的なデータを用いた混合ガウス分布の学習実験により、提案法が従来の EM アルゴリズムと比較して過学習に対して頑健であることを示す。また、中国語放送音声および日本語話し言葉音声を用いた大語彙連続音声認識実験により、提案手法が EM 学習と比較してより多くのパラメタを有効に活用し、単語誤り率の削減に有効であることを示す。

キーワード EM アルゴリズム, 過学習, bagging, 十分統計量

Aggregated EM Algorithm for Robust Parameter Estimation

Takahiro SHINOZAKI†, Mari OSTENDORF††, and Tatsuya KAWAHARA†††

† Department of Computer Science, Tokyo Institute of Technology
2-12-1-W8-77, Ookayama, Meguro-ku, Tokyo, 152–8552 Japan
†† Department of Electrical Engineering, University of Washington
Box 352500, Seattle, Washington, U.S.A. 98195-2500
††† Academic Center for Computing and Media Studies, Kyoto University
Yoshida Nihonmatsu-cho, Sakyo-ku, Kyoto, 606–8501 Japan
E-mail: †shinot@furui.cs.titech.ac.jp

Abstract We propose aggregated EM algorithm that compensates for weaknesses of the EM algorithm by introducing bagging-like approach to estimate likelihood in the training iterations to avoid overtraining. The algorithm consists of E-step and M-step as EM and the multiple models used for the bagging-like operation are efficiently estimated by using a set of sufficient statistics associated with a partitioning of the training data. Analyses using a GMM with artificial data show the proposed algorithm is more robust for overtraining than the conventional EM algorithm. Large vocabulary recognition experiments on Mandarin broadcast news data and Japanese spontaneous speech data show that the method makes better use of more parameters and gives lower recognition error rates than EM training.

Key words EM training, overtraining, bagging, sufficient statistics

1. はじめに

様々な要因で変化する音声を高精度にモデル化するためには、大量のパラメタから構成される複雑なモデルが必要となる。モデルの表現力はモデルのパラメタ数に対して単調に増加するため、精密なモデル化の観点からはパラメタは多ければ多いほどよい。一方、これらパラメタの値は事前に学習データから学習する必要があるが、有限の学習データからのパラメタ推定ではパラメタ数の増大とともにパラメタあたりの学習データが減少するため、推定精度が低下する問題がある。言いかえると、モデルのパラメタが学習データに過度に特化して汎化性が失われてしまい、新しいデータに対応できなくなってしまう過学習の問題が一般的に存在する。

近年一部の音声認識タスクでは数千時間を越えるようなデータが集められており、そのようなタスクでは過学習の問題は比較的小さいと考えられるが、多くのタスクにおいて利用できるデータは数十から数百時間以下にとどまっている。またデータ収集はコストが高く、音声認識技術の普及の観点からも、限られたデータから効率的な学習を行い高い性能のモデルを得られるようにすることは非常に重要である。

このような目的のもと、これまでにクロスバリデーション (CV) を EM [1] の枠組内に導入した CV-EM 手法を提案し音響モデルの学習において有効性を示したが [2]、本研究では CV の代わりに Bagging [3] に似た手法を EM の枠組内に導入することで過学習に対する頑健性を向上させる Aggregated EM (Ag-EM) 法の提案を行い、認識実験によりその効果を示す。Bagging は複数の分類器をサブサンプリングした学習セットから学習し、認識時にはそれらの分類器の結果を統合することで分類精度を向上させる手法である。このため一般には認識時において複数のモデルを並列に走らせることが必要となるが、提案法においては同様の手法が学習アルゴリズム内部に組み込まれるため、最終的に出力されるモデルはこれまでの EM と同じく単一であるという特徴がある。

2. EM アルゴリズムと過学習

EM 学習は繰り返し最適化手法による最尤推定法であり、過学習や局所最適解に対して弱いことが知られている。さらに、混合ガウス分布 (GMM) を利用した隠れマルコフモデル (HMM) では、過学習の極端な場合として、不安定性の問題も存在する。例えば、2 混合の混合ガウス分布を考えると、片方のガウス分布が特定の学習サンプルを非常に小さな分散で覆い、他方のガウス分布が残りのデータをカバーするようにすると、尤度をいくらでも大きくできてしまう。これはモデルの一般性の点で明らかに好ましくないが、このような例は実際に混合ガウス分布の学習においてしばしば観察される。これらの問題を解決できれば、より少ないデータでより精密なモデルを精度よく推定し、認識性能を大きく向上させられると期待される。

EM アルゴリズムの過学習に対する脆弱性は、学習の目的関数として学習セットに対する尤度を用いていることに起因する。EM 学習では、図 1 に示すように、Expectation step (E-step) と Maximization step (M-step) を繰り返しながらモデルパラメタを更新していく。混合ガウス分布 HMM の場合であれば、E-step において状態遷移系列と混合要素の割り当てに関する十分統計量の期待値 (図では SS) が計算され、M-step においてその期待十分統計量に基づいてモデルパラメタ (図では M) が最尤法により更新される。すなわち、E-step ではその学習ステージにおいて仮定されるモデルに基づいた確率推論が行われ、M-step では確率推論の結果をもとに計算された統計量からモデルパラメタの推定が行われる。モデルの推定と尤度の評価が同じ学習データを用いて繰り返し行われることになり、ひとたびモデルが特定のサンプルに過度な尤度を割り当てると次のループではそれがさらに助長される悪循環に陥り、最終的に学習データに特化した一般性のないモデルが出力されることになる。そこで、EM 学習を過学習に対して頑健化するためには、学習内部での尤度評価におけるバイアスを取り除くことが重要と考えられる。

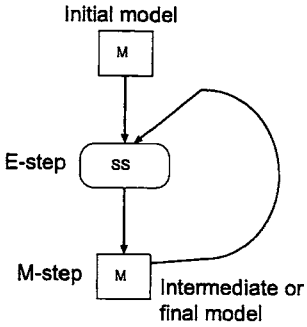

図 1 EM アルゴリズム。SS は十分統計量

3. Aggregated EM アルゴリズム

3.1 アルゴリズムの動作

本研究で提案する Aggregated EM (Ag-EM) アルゴリズムは、学習プロセスとしては並列 EM 学習 [4] の拡張と見ることができる。並列 EM 学習は並列計算機上で効率的に EM 学習を行うための並列化手法であり、図 2 に示すように学習データを K 個の区画に区分化し、各データ区画ごとに独立に期待十分統計量 $SS(k)$ を計算する(注1)。学習データ全体に対する期待十分統計量 SS は、個別に求めた K 個の十分統計量 $SS(k)$ を "足し合わせる" ことで容易に求めることができる。これは、十分統計量が事後確率で重み付けされた定数や観測特徴量に関する和であるためである。モデルパラメタの更新は得られた SS から通常の EM と同様に行われる。

並列 EM では学習データの区分化は単に学習時間の短縮のために行われるが、Ag-EM ではデータの区分化がより本質的な役割を果たす。並列 EM では各 M-step において K 個全ての十分統計量を統合して 1 つのモデルを作成するが、Ag-EM に

(注1)：HTK では、HERest に -p オプションを指定して、.acc ファイルを並列に複数作成することに対応する。

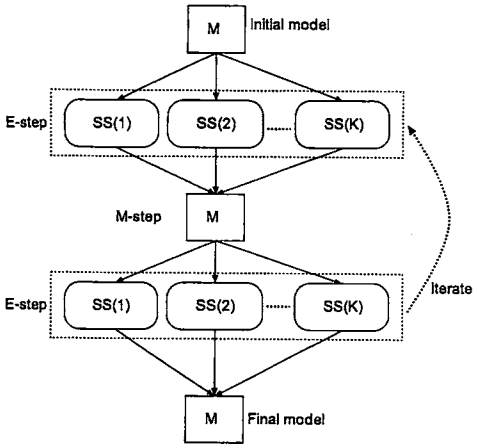

図 2 並列 EM アルゴリズム。SS(k) は k 番目の区画の十分統計量

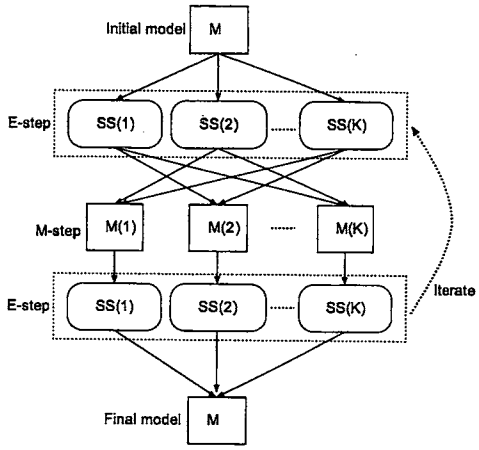

図 3 CV-EM アルゴリズム

おいては図 4 に示すように、K 個中 $K' < K$ 個の区画に対応する十分統計量を足し合わせた、N 個のモデル $M(n)$ を作成する。サブセットの選択において重複を許さないとすると、N の最大値は $C(K, K')$ となる。ただし $C(a, b)$ は a 個の中から b 個の要素を取り出す組合せ数である。

続く E-step では 各データ区画 k に対し、その十分統計量 $SS(k)$ を N 個のモデルを用いて計算した N 個の十分統計量 $SS(n, k)$ の平均 $SS(k) = \frac{1}{N}\sum_{n=1}^{N} SS(n, k)$ として求める。この操作は Bagging における複数分類器の結果統合に対応している。言いかえると、期待十分統計量を N 個の異なるモデルを用いて計算した尤度の平均により求めていることになる。以下では、N をアンサンブルサイズと呼ぶことにする。最終ステージでは全ての十分統計量を統合し、単一のモデルを出力する。

図 3 に示す CV-EM 学習と比較すると、CV-EM 学習では E-step と M-step に用いられる学習データを分離することで尤度評価におけるバイアスを抑制し性能向上を図るのに対し、Ag-EM では E-step と M-step の学習データの間に重複を許す。その代わりに、複数のモデルによる確率推論の結果を統合することで学習時の尤度が特定の学習サンプルに特化することを防ぎ、頑健性の向上を図る。また複数のモデルを用いることから、小さな局所最適解への収束を抑制し、よりよい局所最適解を見つける効果も期待される。

3.2 アルゴリズムの実装

もしも十分統計量を利用しないとすると、Ag-EM の計算量は、学習セットのサイズが T のとき、$O\left(\frac{K'}{K}TN^2\right)$ となる。これは、各モデルが学習セット全体の $\frac{K'}{K}T$ のデータから N 個のモデルを用いて学習され、またその様なモデルが N 個あるためである。しかし図 4 に示す手続きに従い十分統計量を活用することで、各モデルの間で同じデータを繰り返し処理する冗長性を取り除くことができ、計算量を $O\left(\frac{K'}{K}TN\right)$ に抑えることができる。すなわち、アンサンブルサイズ N に比例した計算量とすることができる。

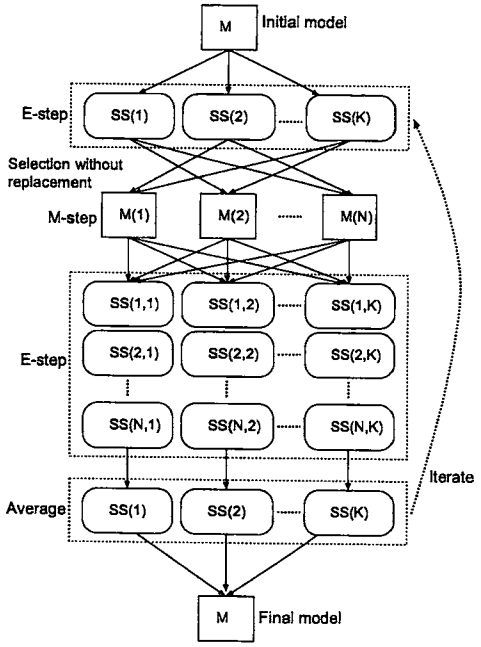

図 4 Aggregated EM (Ag-EM) アルゴリズム。SS(n, k) は k 番目のデータ区画を n 番目のモデル M(n) で処理した十分統計量

学習時に必要となるディスクサイズについては、もし E-step において K 個のデータ区画それぞれに対して N 個の十分統計量を計算し保存したとすると、$O(KN)$ のオーダーの領域が必要とされることになる。しかしながら、同じデータ区画に対する十分統計量は平均としてのみ用いられるため、各区画について十分統計量ファイルを 1 つのみ用意し計算結果を順次積算していくようにすると、$O(K)$ となる。

3.3 学習メタパラメタの設定

Ag-EM における性能向上は複数モデルの統合に基づいている。このため、各学習ステージにおいて用いられるモデルは同

一では意味がなく、適度に異っている必要がある。他方で、それらモデルが完全に独立であり隠れ変数間に対応関係が存在しないと、十分統計量の統合操作が意味をなさない。各ステージにおけるモデル間の相違度は、各モデル間で共有される学習データ量に依存し、$\frac{K'}{K}$ により制御される。以後の実験では事前実験より $\frac{K'}{K} = 0.6$ とした。

Bagging と同様に Ag-EM の学習性能はアンサンブルサイズ N に依存する。基本的にはアンサンブルサイズが大きければ大きいほど高い汎化能力が得られると考えられる。

4. 混合ガウス分布を用いた実験

無作為に設定した 4 次元 8 混合の混合ガウス分布からランダムに抽出した学習用データと評価用データを使った実験を行った。初期モデルとして、学習データ全体の平均と分散を基に初期化した 8 混合のガウス分布を用いた。テストセットのサンプル数は 1000 である。また全ての実験において分散フロア値を 1.0E-5 とした。実験結果から偶然性を取り除くために、各実験条件において独立に抽出したデータを用いた実験を 100 回繰り返し、平均をとった値を最終的な評価値として用いている。

図 5 に 20 サンプルおよび 80 サンプルからなる学習セットを用いて、EM, CV-EM, および Ag-EM により学習したモデルの、テストセットに対する尤度を示す。CV-EM および Ag-EM とも学習セットの分割数は $K = 20$ である。また、Ag-EM においてアンサンブルサイズ N は 8 でり、サブセットの選択数 K' は $\frac{K'}{K} = 0.6$ となるように 12 とした。過学習の影響のため、EM により学習されたモデルのテストセットに対する尤度は学習の繰り返しに対して単調増加とはならない。とくに学習セットが小さい場合には、学習くり返し数が大きくなると、初めのうちは増加した尤度が大幅に低下してしまう現象が見られる。これに対し、CV-EM は EM と比較して過学習に対して頑健であることが分かる。Ag-EM では汎化性能がさらに向上し、学習くり返し数に対しより安定した特性を示し、CV-EM と比較しても大幅に高い尤度が得られることが分かる。この効果はデータ量が少ない場合に特に顕著である。モデルに対するデータ量はパラメタ数に相対的であるため、このことは言いかえると、Ag-EM は EM や CV-EM と比較してより少ないデータからより複雑なモデルを精度よく推定する能力があると言うことができる。

図 6 にアンサンブルサイズ N とモデル性能の関係を示す。Ag-EM 学習において、学習回数は 10、学習セットサイズは 20 であり、$K = 20, K' = 12$ および $K = 10, K' = 6$ の結果を示す。この図でアンサンブルサイズが 8 で $K = 20, K' = 12$ の場合が、図 5 において学習サンプル数が 20 で 10 回 Ag-EM 学習を繰り返した場合に対応する。また、参考のため 20-fold の CV-EM の結果も合わせて示してある。CV-EM は、N に独立である。Ag-EM によるモデルの性能は、N とともにほぼ単調に増加することが分かる。N が小さいと性能も低いが、3 以上で CV-EM の性能を越えることが分かる。また $\frac{K'}{K}$ が一定であれば、K の値が異なってもほぼ同じ性能となることが分かる。ただし、K があまり小さいと、組合せ数で最大値が決まる

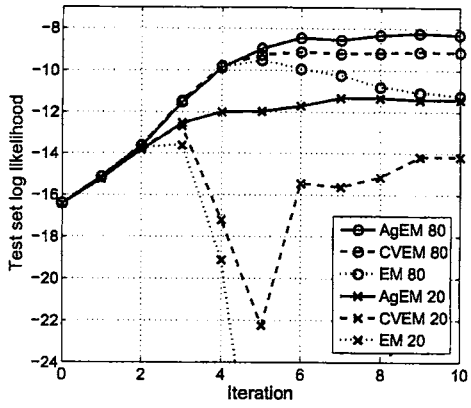

図 5 EM、CV-EM、および Ag-EM により学習した GMM のテストセット尤度。学習セットのサンプル数は 20 および 80 であり、横軸は学習繰り返し数

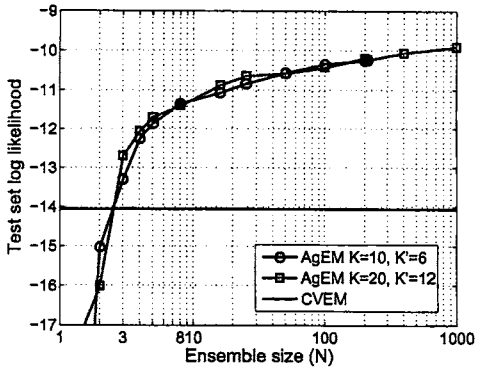

図 6 アンサンブルサイズとモデル性能

N が制約を受ける。

5. 大語彙連続音声認識実験

5.1 中国語放送音声認識タスク

中国語 Hub4 および TDT4 の放送ニュース音声を音響モデルの学習セットとして用いた大語彙連続音声認識実験を行った。学習セットのデータ量は合計で約 97 時間である。認識率の評価は中国語 RT-04 の評価セットを用いて行った。また RT-04 の開発セットをモデル選択に用いた。開発および評価セットは放送音声で、データ量はそれぞれ約 30 分および 1 時間である。認識には Decipher システムを用いている [5]。本実験では言語モデルにトライグラム言語モデル、音響モデルに最尤学習により学習した GENONE モデルを用いている。GENONE モデルは単語内状態共有トライフォンの混合重みを全て展開した構造を持つモデルである。発音辞書の語彙数は 49k である。また、声道長正規化 (VTLN) および MLLR による話者適応を行っている。

実験は $K = 3, K' = 2, N = 3$ と、小さなアンサンブルサイ

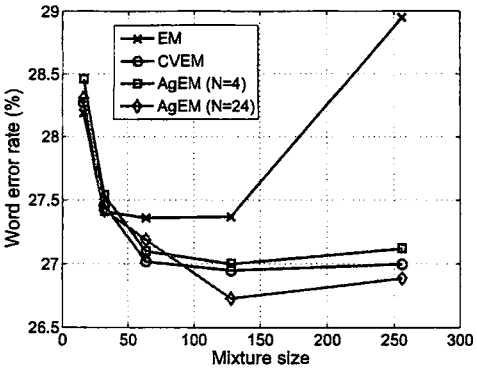

図 7　CSJ30 時間を用いて学習したモデルの単語誤り率

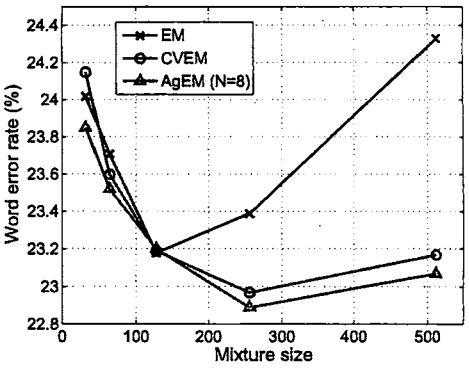

図 8　CSJ254 時間を用いて学習したモデルの単語誤り率

ズを用いた限定的な条件で行ったが、EM を用いたベースラインの文字誤り率 18.9% に対し CV-EM を用いた場合が 18.4%、Ag-EM が 18.5 % と、Ag-EM が CV-EM と同様に認識性能の向上に有効であることが分かる。

5.2　日本語自然発話音声認識タスク

音響モデルは状態共有トライフォンモデルであり、学習セットとして 254 時間の日本語話し言葉コーパス CSJ [6] の学会講演音声を用いた。音響特徴量は 12 次元の MFCC と対数エネルギおよびそれらのデルタとデルタデルタの合計 39 次元である。言語モデルは CSJ 学会講演と模擬講演約 6.8M 形態素から学習したトライグラムである。テストセットは CSJ 学会講演評価セット 10 講演であり、全て男性話者である。HMM の EM 学習には HTK [4]、認識エンジンには Julius [7] を用いた。CV-EM および Ag-EM の学習は、十分統計量の操作をサポートするための変更を加えた HTK を用いて行った。CV-EM において学習セットの分割数は $K = 30$ である。また、Ag-EM では学習セットの分割数 $K = 10$、サブセットの選択数 $K' = 6$ とした。HMM の混合ガウス分布の学習は、EM または提案手法を 5 回繰り返す毎に混合数を倍に増加させることで行った。また、CV-EM および Ag-EM 学習において、学習データのリストは混合数を増加させる毎にランダムに並べ替えを行った。

図 7 に学会講演音声から無作為に抽出した 30 時間のデータを学習セットとして用いた場合のモデルの混合数と単語誤り率の関係を示す。音響モデルの状態数は 1000 である。学習手法ごとの単語誤り率の最小値は EM が 27.4%、CV-EM が 27.0%である。Ag-EM を用いた場合、アンサンブルサイズが $N = 4$ と小さい場合は CV-EM と比較するとやや誤り率が大きいが、$N = 24$ とすると認識性能がさらに向上し、26.7% と最小の単語誤り率が得られた。

図 8 に学会講演音声全体を学習セットとして用いた場合のモデルの混合数と単語誤り率の関係を示す。音響モデルの状態数は 3000 であり、Ag-EM のアンサンブルサイズは $N = 8$ とした。図 7 と同様に、CV-EM および Ag-EM は一定量の学習データに対し EM よりも複雑で高精度なモデルを精度よく推定することができ、低い単語誤り率が得られることが分かる。

6.　考察と今後の課題

本論文では Ag-EM 学習においてモデル推定の分散を減少させ汎化性能を向上させる原理を直感的に説明し、また実験においてその有効性を示したが、Ag-EM アルゴリズム (や CV-EM アルゴリズム) の収束性の証明は今後の課題である。これは、従来の EM アルゴリズムと異なり、Ag-EM や CV-EM においては学習セットに対する尤度の単調増加が保証されず、EM の場合のような単純な収束の議論が成り立たないためである。この問題への直接的な回答ではないが、尤度が単調増加とならない場合の収束性を議論している論文としては、文献 [8] が挙げられる。直観的には、Ag-EM や CV-EM では学習データ量が少なくても E-step における尤度バイアスが少なく、EM において大量のデータを用いた場合の尤度空間を近似していると考えられることから、その近似された尤度空間上で EM を行っていると解釈することができる。この解釈は、モデルのパラメタ数が一定の場合、学習データ量が少ないときには Ag-EM が EM よりも高い性能を示し、データ量の増加とともに次第に同じ性能へと収束していく実験結果とも符合する。

Ag-EM アルゴリズムと同様に過学習をさけ、少量のデータから高い精度のモデルを推定できる手法として変分ベイズ法 [9] が挙げられる。変分ベイズ法とし比較した Ag-EM の特徴としては、全てがデータ駆動で事前分布の設定が必要ないこと、十分統計量を用いた並列化が出来れば目的関数が尤度ではない繰り返し学習法への応用も可能と考えられることなどが挙げられる。

今後の課題としては、尤度推定法のさらなる改良の他、十分統計量を活用する類似のアイデアに基づくモデル構造最適化手法 [10], [11] との組合せ、話者適応や目的関数が尤度でない繰り返し学習法への応用などが挙げられる。

7.　ま　と　め

従来の EM 学習の枠組内に Bagging に似た手法を組み込むことで、過学習に対する頑健性を向上させる手法の提案を行った。Ag-EM は EM や CV-EM と比較して学習計算量がアン

サンプルサイズ N に比例する分増加するが、N の増加とともにより高い汎化性能が期待できる。混合ガウス分布を用いた実験により、Ag-EM を用いることで、EM や CV-EM と比較して学習されたモデルの性能が大幅に向上することを示した。また、大語彙連続音声認識実験においても Ag-EM を学習に用いることで、EM や CV-EM と比較して高い認識性能が得られることを示した。

なお、CV-EM および Ag-EM を実装したサンプルプログラムを Web ページ www.furui.cs.titech.ac.jp/~shinot において公開する予定である。並列化 EM を行う既存のプログラムをこれらのアルゴリズムに対応させることは、基本的には十分統計量に関する操作を追加するだけでよく、一般に容易と考えられる。

謝辞 本研究は第一著者が University of Washington (Seattle, Washington, U.S.A) および International Computer Science Institute (Berkeley, California, USA) に滞在中に行った研究を、京都大学に異動後発展させたものである。本研究は契約番号 HR0011-06-C-0023 により DARPA の支援を得て行われた。配布における制限はない。本論文の見解は著者のものであり、資金を提供した機関の見解を反映するものではない。また、日本語音声を用いた実験は、科研費 (19700167) の助成を受けたものである。

文　献

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. of the Royal Statistical Society*, , no. Series B 39, No. 1, pp. 1–38, 1977.

[2] T. Shinozaki and M. Ostendorf, "Cross-validation and aggregated EM training for robust parameter estimation," *Computer speech and language*, vol. 22, no. 2, pp. 185–195, 2008.

[3] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[4] S. Young *et al.*, *The HTK Book*, Cambridge University Engineering Department, 2005.

[5] M. Y. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin broadcast news speech recognition," in *Proc. ICSLP*, 2006, pp. 1233–1236.

[6] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the Corpus of Spontaneous Japanese," in *Proc. SSPR2003*, 2003, pp. 135–138.

[7] A. Lee, T. Kawahara, and S. Doshita, "An efficient two-pass search algorithm using word trellis index," in *Proc. ICSLP*, 1998, pp. 1831–1834.

[8] A. Gunawardana and W. Byrne, "Convergence theorems for generalized alternating minimization procedures," *Journal of Machine Learning Research*, vol. 6, pp. 2049–2073, 2005.

[9] S. Waterhouse, D. MacKay, and T. Robinson, "Bayesian methods for mixture of experts," in *NIPS8*, 1995, pp. 351–357.

[10] T. Shinozaki, "HMM state clustering based on efficient cross-validation," in *Proc. ICASSP*, Toulouse, 2006, vol. I, pp. 1157–1160.

[11] T. Shinozaki and T. Kawahara, "Gaussian mixture optimization for HMM based on efficient cross-validation," in *Proc. Interspeech*, 2007, pp. 2061–2064.