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

2003－ＳＬＰ－49　（９）
2003／12／18

# モデル統合に基づく高速 EM 学習法

芳澤 伸一† 　鹿野 清宏‡

†松下電器産業株式会社　先端技術研究所　〒619-0237 京都府相楽郡精華町光台 3-4
‡奈良先端科学技術大学院大学　〒630-0101 奈良県生駒市高山町 8916-5
E-mail: †yoshizawa.shinichi@jp.panasonic.com　‡shikano@is.aist-nara.ac.jp

**あらまし**　学習済みの音響モデルを利用して，利用条件に適した音響モデルを短時間に EM 学習する方法を提案する．ここでは，話者適応による認識実験により提案法の有効性を示す．

**キーワード**　EM アルゴリズム，最尤推定，モデル統合，音響モデル

# Rapid EM Training Technique Based on Model-Integration

Shinichi YOSHIZAWA† and Kiyohiro SHIKANO‡

†Matsushita Electric Industrial Co., Ltd. 3-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
‡Nara Institute of Science and Technology 8916-5 Takayama-cho, Ikoma-shi, Nara 630-0192 Japan
E-mail: †yoshizawa.shinichi@jp.panasonic.com　‡shikano@is.aist-nara.ac.jp

**Abstract**　We propose a rapid EM training technique using pre -trained acoustic models and evaluate it successfully.

**Keyword**　Expectation-Maximization algorithm, Maximum -likelihood estimation, Model –integration , Acoustic model

## 1. はじめに

音響モデルを構築する場合，一般に，バウム・ウェルチによる学習法[1]が用いられる．この方法では，最尤基準によりモデルを学習するため高い認識性能を獲得することができる．しかし，学習に大量の音声データを用いるため学習時間が膨大になるという課題がある．

そこで本論文では，事前に構築された様々な音響モデルを利用して，利用条件に適した音響モデルを短時間に構築する方法を提案する[2][3]．提案法は，大量の音声データの代わりに，事前に学習された音響モデルの少数の統計パラメータを用いて学習するため，短時間に音響モデルを獲得することができる．また，最尤基準に基づいた学習であるため，バウム・ウェルチの学習法と同様に高い認識性能を獲得することができる．本論文では，話者適応による認識実験により提案法の有効性を示す．

## 2. モデル統合に基づく高速ＥＭ学習法

提案法の概念図を図１に示す．提案法は３つのステップにより構成される．第１ステップでは，構築したい音響モデルの利用条件に適した学習済みの音響モデル（以下，学習モデルと呼ぶ）を選択する．話者適応の場合は，利用者の１文章発声により，利用者に音響的に近い話者の音声で学習された音響モデル（学習モデル）を選択する方法がある[6][7]．第２ステップでは，構築する音響モデルの統計パラメータ（混合重み，平均，分散）の初期値を設定する．第３ステップでは，選択した学習モデルを用いて，利用条件に適した音響モデルを EM 学習により獲得する．

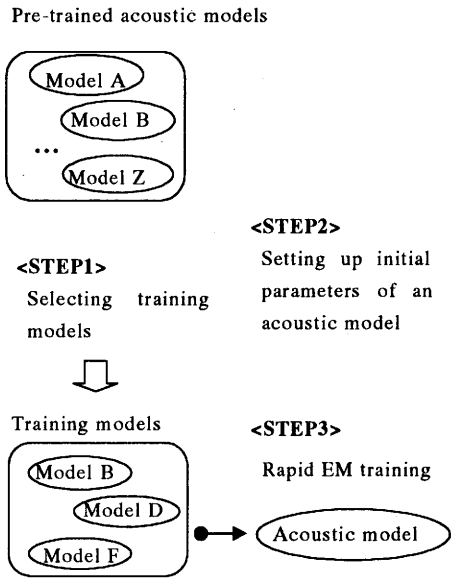


図１　Rapid EM training

## 2.1. 学習モデルの選択

第１ステップでは，事前に学習された音響モデルの中から，学習モデルとして利用条件に適した複数の音響モデルを選択する．例えば，話者適応の場合は，利用者に音響的に近い話者の音声で学習された音響モデル（学習モデル）を選択する．また，騒音適応の場合は,利用環境に近い騒音下で学習された音響モデル(学習モデル）を選択する．

ここで話者適応の場合について具体的に述べる．初めに，様々な話者の音声で学習した複数の音響モデル（特定話者モデル）を準備する．同時に，音響モデルを選択するための選択モデルを準備する．ここでは選択モデルとして,発声内容を区別せず 64 混合の混合ガウス分布で構築したものを準備する．そして，利用者の１文章発声を選択モデルに入力し，尤度が高い複数の話者の音響モデルを選択する．ここで選択された音響モデルは，利用者に音響的に近い音声データで学習されたモデルである[6][7]．

## 2.2. 音響モデルの初期値の設定

第２ステップでは，構築する音響モデルの統計パラメータの初期値を設定する．本論文では，不特定話者モデルを事前に作成しておき音響モデルの初期値として利用する．他の方法として，学習モデルを利用して初期値を設定する方法がある．

また，バウム・ウェルチによる学習法と同様に，認識処理時間，認識精度などを考慮して，構築する音響モデルのガウス分布の混合分布数を設定することができる．例えば，短時間に認識結果を獲得したい場合などは，混合分布数を少なく設定してモデルを単純化することにより，認識時における処理時間を削減することができる．

## 2.3. 高速ＥＭ学習

第３ステップでは，第２ステップで設定した初期値を用いて，第１ステップで選択した学習モデルに対する尤度（対数尤度）を最大化するように音響モデルを学習する．

学習モデルは，構築したい音響モデルの利用条件に適した音声データの特徴を反映したモデルである．そのため，学習モデルに対する尤度を最大化することにより適切な音響モデルを学習することができる．

また，バウム・ウェルチの学習法とは異なり，音声データそのものではなく，その音声データで学習された学習モデルの少数の統計パラメータを用いて学習するため,短時間に音響モデルを獲得することができる．

### 2.3.1. 最適化関数（学習モデルに対する尤度）

提案法における最適化関数を数式１に示す．最適化関数は，バウム・ウェルチの学習法と同じ尤度関数である．異なるのは，音声データ１つずつに対する尤度ではなく，学習モデルに対する尤度である点である．

**数式 1**

$$\log P = \sum_{i=1}^{N_g} \int_{-\infty}^{\infty} \{\log[\sum_{m=1}^{M_f} \omega_{f(m)} f(x;\mu_{f(m)},\sigma^2_{f(m)})] \sum_{l=1}^{L_{g(i)}} \omega_{g(l,i)} g(x;\mu_{g(l,i)},\sigma^2_{g(l,i)})\} dx$$

ここで $N_g$ は学習モデル数であり，$f(.)$，$g(.)$はそれぞれ，構築する音響モデル，学習モデルのガウス分布であり，$M_f$，$L_{g(i)}$ は混合分布数，$\omega_{f(m)}$，$\omega_{g(l,i)}$ は混合重み，$\mu_{f(m)}$，$\mu_{g(l,i)}$ は平均，$\sigma^2_{f(m)}$，$\sigma^2_{g(l,i)}$ は分散である．

### 2.3.2. 統計パラメータの更新

提案法では，最適化関数（数式１）を最大化することにより，音響モデルの統計パラメータを学習する．

音響モデルの統計パラメータの更新式を，数式２，数式３，数式４に示す．

**数式 2**

$$\omega_{f(m)[t+1]} = \frac{\sum_{i=1}^{N_g} A(m,i)_{[t]}}{\sum_{k=1}^{M_f}\sum_{i=1}^{N_g} A(k,i)_{[t]}} \qquad (m = 1,2,...,M_f)$$

**数式 3**

$$\mu_{f(m,j)[t+1]} = \frac{\sum_{i=1}^{N_g} B(m,i,j)_{[t]}}{\sum_{i=1}^{N_g} A(m,i)_{[t]}} \qquad (m = 1,2,...,M_f, j = 1,2,...,J)$$

**数式 4**

$$\sigma^2_{f(m,j)[t+1]} = \frac{\sum_{i=1}^{N_g} C(m,i,j)_{[t]}}{\sum_{i=1}^{N_g} A(m,i)_{[t]}}$$

$$(m = 1,2,...,M_f)$$

ここでtは学習回数であり，jはxの要素（次元）のインデックスを示す．ここで，

**数式 5**

$$A(m,i)_{[t]} = \int_{-\infty}^{\infty} \sum_{l=1}^{L_{g(i)}} \gamma(x;m)_{[t]} \omega_{g(l,i)} g(x;\mu_{g(l,i)},\sigma^2_{g(l,i)}) dx$$

**数式 6**

$$B(m,i,j)_{[t]} = \int_{-\infty}^{\infty} x_j \sum_{l=1}^{L_{g(i)}} \gamma(x;m)_{[t]} \omega_{g(l,i)} g(x;\mu_{g(l,i)},\sigma^2_{g(l,i)}) dx$$

**数式 7**

$$C(m,i,j)_{[t]} = \int_{-\infty}^{\infty} (x_j - \mu_{f(m,j)})^2 \times \sum_{l=1}^{L_{g(i)}} \gamma(x,m)_{[t]} \omega_{g(l,i)} g(x;\mu_{g(l,i)},\sigma^2_{g(l,i)}) dx$$

であり，

**数式 8**

$$\gamma(x;m)_{[t]} = \frac{\omega_{f(m)[t]} f(x;\mu_{f(m)[t]},\sigma^2_{f(m)[t]})}{\sum_{k=1}^{M_f} \omega_{f(k)[t]} f(x;\mu_{f(k)[t]},\sigma^2_{f(k)[t]})}$$

である．$\gamma(x;m)_{[t]}$は，t回目の学習により獲得した$f(\cdot)_{[t]}(=\omega_{f(\cdot)[t]} f(x;\mu_{f(\cdot)[t]},\sigma^2_{f(\cdot)[t]}))$に依存した値になる．

また，音響モデルの状態遷移確率の更新式を数式9に示す．

**数式 9**

$$T_f[i][j] = \frac{\sum_{k=1}^{N_g} T_{g(k)}[i][j]}{\sum_{j=1}^{N_{st}} \sum_{k=1}^{N_g} T_{g(k)}[i][j]}$$

ここで$T_f[i][j]$，$T_{g(k)}[i][j]$はそれぞれ，構築する音響モデル，学習モデルのi番目の状態からj番目の状態への状態遷移確率であり，$N_{st}$は状態数である．

### 2.3.3. $\gamma(x,m)_{[t]}$の近似

ここで，$\gamma(x;m)_{[t]}$（数式8）について考察する．$\gamma(x;m)_{[t]}$は，特徴量xにおける，構築する音響モデルの全てのガウス分布$f(k)_{[t]}(k=1,2,...,M_f)$に対する，更新を行うm番目のガウス分布$f(m)_{[t]}$の寄与度である．図3は，$f(m)_{[t]}$と$\gamma(x;m)_{[t]}$の関係を表す概念図である．$f(m)_{[t]}$に近い領域は$f(m)_{[t]}$の寄与度が大きいため$\gamma(x;m)_{[t]}=1$，それ以外の領域は$\gamma(x;m)_{[t]}=0$となる．

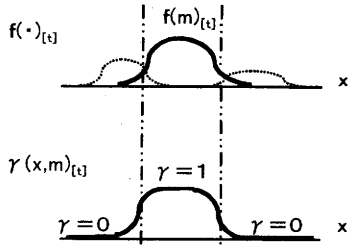


図3 $\gamma(x,m)_{[t]}$

このことから，学習モデルのガウス分布に対する$\gamma(x;m)_{[t]}$の値は，以下に示す数式10で近似できる．すなわち，図4に示すように，$f(m)_{[t]}$との分布間距離が近い学習モデルのガウス分布に対して$\gamma(x;m)_{[t]} \approx 1$，それ以外のガウス分布に対して$\gamma(x;m)_{[t]} \approx 0$と近似で

きる．

**数式 10**

$$\gamma(x;m)_{[t]} \doteqdot \begin{cases} 1 & g(\cdot)\ \ nearby\ \ f(m)_{[t]} \\ 0 & otherwise \end{cases}$$

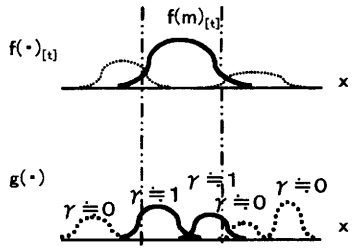


**図4 γ of Gaussians of training models**

ここで，バウム・ウェルチの学習法で学習した音響モデル（ここでは，不特定話者モデル，モノフォン16混合）におけるガウス分布の位置関係についてみてみる．図5は，最も近いガウス分布間のマハラノビス距離の平均を模式的に表したものであり，距離の平均は3.3であった．これより，構築したい音響モデルのガウス分布の位置関係は図3，図4のように表現できることがわかる．

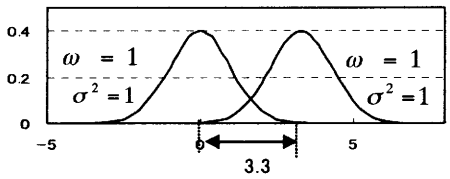


**図5 Distance between nearest Gaussians**

最終的には，$\gamma(x;m)_{[t]}$の近似（数式10）を利用することで，統計パラメータの更新式（数式2，数式3，数式4）は，数式11，数式12，数式13となる．

**数式 11**

$$\omega_{f(m)[t+1]} \approx \frac{\sum_{i=1}^{N_g}\sum_{l=1}^{L_{g(i)}}\gamma(x;m)_{[t]}\omega_{g(l,i)}}{\sum_{k=1}^{M_f}\sum_{i=1}^{N_g}\sum_{l=1}^{L_{g(i)}}\gamma(x;m)_{[t]}\omega_{g(l,i)}}$$

$$(m = 1,2,...,M_f)$$

**数式 12**

$$\mu_{f(m,j)[t+1]} \approx \frac{\sum_{i=1}^{N_g}\sum_{l=1}^{L_{g(i)}}\gamma(x;m)_{[t]}\omega_{g(l,i)}\mu_{g(l,i,j)}}{\sum_{i=1}^{N_g}\sum_{l=1}^{L_{g(i)}}\gamma(x;m)_{[t]}\omega_{g(l,i)}}$$

$$(m = 1,2,...,M_f, j = 1,2,...,J)$$

**数式 13**

$$\sigma^2_{f(m,j)[t+1]} \approx \frac{\sum_{i=1}^{N_g}\sum_{l=1}^{L_{g(i)}}\gamma(x,m)_{[t]}\omega_{g(l,i)}\{\sigma^2_{g(l,i,j)} + (\mu_{g(l,i,j)} - \mu_{f(m,j)})^2\}}{\sum_{i=1}^{N_g}\sum_{l=1}^{L_{g(i)}}\gamma(x;m)_{[t]}\omega_{g(l,i)}}$$

$$(m = 1,2,...,M_f, j = 1,2,...,J)$$

ただし，$\gamma(x;m)_{[t]}$は数式10により決定する．

上式を用いることで，音響モデルを，学習モデルの統計パラメータを用いて短時間に学習することができる．

## 3. 認識実験

ここで，話者適応による認識実験により提案法の有効性を示す．

データベースとして，日本音響学会によるJNASデータベース[4]を利用する．JNASは306人の話者の音声

データにより構成されており，各話者は約 150 文章の発声データをもつ．ここでは静かな環境で収録した音声データを用いる．サンプリング周波数は 16kHz，量子化ビットレートは 16bit である．特徴量として，窓シフト長 10ms で分析した 12 次元の MFCC（Mel-Frequency Cepstrum Coefficient）と 12 次元のデルタケプストラム，デルタパワーを用いる．特徴量抽出において CMN（Cepstrum Mean Normalization）処理が施されている．20k の新聞記事により構築した言語モデルとデコーダとして Julius[5]を用いる．音響モデルは，43 音素の，モノフォン 16 混合もしくはモノフォン 64 混合を用いる．

事前に準備する音響モデル（学習モデルの候補）として，評価話者を除いた 260 人の話者のモデル（260 個）を用いる．そして，利用者の 1 文章発声を用いて，準備した音響モデルの中から尤度の高い上位 $N_g$ 個の音響モデルを学習モデルとして選択する．ここでは，モノフォン 16 混合の音響モデルを作成する場合は $N_g=10$ とし，モノフォン 64 混合の音響モデルを作成する場合は $N_g=20$ とした．この値は，十分統計量による話者適応[6][7]を参考にした．評価用データとして，上記音声データに含まれない 46 人の発声データを用いる．

比較として，バウム・ウェルチの学習法における認識結果も合わせて示す．バウム・ウェルチの学習法では，学習モデルを学習した全ての音声データ（$N_g$ × 約 150 文章発声）を用いて利用者の音響モデルを学習する．

### 3.1. 学習時間と認識率

ここでは学習時間と認識率について考察する．

表 1 と表 2 に高速 EM 学習法（提案法）とバウム・ウェルチの学習法における認識結果を示す．表には，1 回の学習における学習時間と認識率(単語正解精度)が示されている．学習時間は 1.3GHz の CPU パワーをもつコンピュータで学習したときの結果である．

表 1 には，学習モデル($N_g=10$)，構築する音響モデルともにモノフォン 16 混合を用い，音響モデルの初期値として不特定話者（260 人分）で作成したモノフォン 16 混合を用いた場合の結果が示されており，表 2 には，学習モデル($N_g=20$)，構築する音響モデルともにモノフォン 64 混合を用い，音響モデルの初期値として不特定話者（260 人分）で作成したモノフォン 64 混合を用いた場合の結果が示されている．

表 1，表 2 の結果より，提案法は，バウム・ウェルチの学習法と比較し，短時間に高い認識性能を獲得できることがわかる．

表 1 学習時間と認識率 （モノフォン 16 混合）

| | 学習時間 | 単語正解精度 | |
|---|---|---|---|
| | | 学習 1 回 | 学習前 |
| 高速 EM 学習法（提案法） | 12 秒 | 85.7% | 82.2% |
| バウム・ウェルチ | 1510 秒 | 85.0% | |

表 2 学習時間と認識率 （モノフォン 64 混合）

| | 学習時間 | 単語正解精度 | |
|---|---|---|---|
| | | 学習 1 回 | 学習前 |
| 高速 EM 学習法（提案法） | 365 秒 | 89.3% | 86.3% |
| バウム・ウェルチ | 9400 秒 | 89.2% | |

### 3.2. 学習回数と認識率

ここで学習回数と認識率の関係について考察する．

図 6，図 7 に高速 EM 学習法（提案法）とバウム・ウェルチの学習法における認識結果を示す．

図 6 は，学習モデル($N_g=10$)，構築する音響モデルともにモノフォン 16 混合を用い，音響モデルの初期値として不特定話者（260 人分）で作成したモノフォン 16 混合を用いた場合の結果であり，図 7 は，学習モデル($N_g=20$)，構築する音響モデルともにモノフォン 64 混合を用い，音響モデルの初期値として不特定話者（260 人分）で作成したモノフォン 64 混合を用いた場合の結果である．

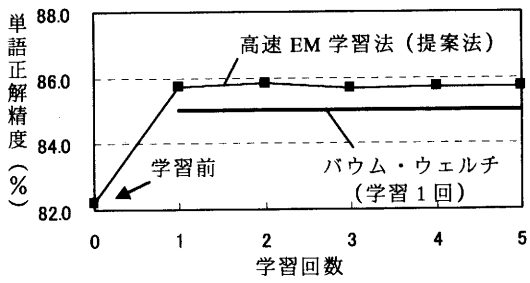


図 6 学習回数と認識率 （モノフォン 16 混合）

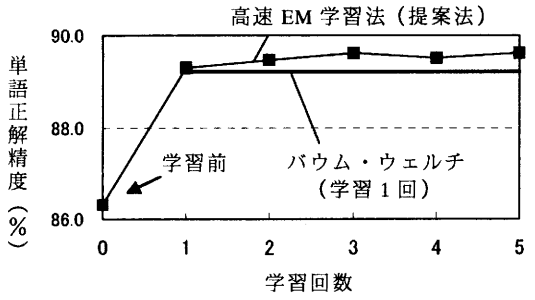


図7 学習回数と認識率 （モノフォン 64 混合）

図6，図7の結果より，モノフォン 16 混合の場合では，1回の学習で高い認識性能が獲得でき，モノフォン 64 混合の場合では，学習を繰り返すごとに認識性能が改善されることがわかる．

### 3.3. 混合分布数の設定

混合分布数を変更した場合について考察する．ここでは，学習モデル($N_g=10$)としてモノフォン 64 混合を準備し，モノフォン 16 混合の音響モデルを作成した．

図8に認識結果を示す．ここでは，音響モデルの初期値として不特定話者（260 人分）で作成したモノフォン 16 混合を用いた場合（■印）と，不特定話者（260 人分）で作成したモノフォン 64 混合の中から任意の 16 混合を用いた場合（▲印）における結果を示す．

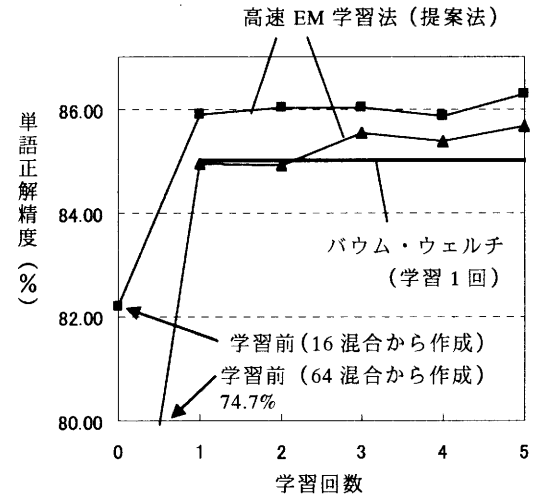


図8 学習回数と認識率
（モノフォン 64 混合→モノフォン 16 混合）

図8の結果より，提案法は，バウム・ウェルチの学習法と比較して，高い認識率が獲得できていることがわかる．また，学習を繰り返すごとに認識性能が改善されることがわかる．また，音響モデルの初期値として適切な値を設定することで，より高い認識性能が獲得できることがわかる．

## 4. まとめ

学習済みの音響モデルを利用して，利用条件に適した音響モデルを短時間に EM 学習する方法を提案した．話者適応による認識実験により提案法の有効性を示した．

今後は，$\gamma$ の近似方法や，音響モデルの初期値の設定方法，EM 学習における局所解問題について考察する予定である．また，騒音環境下での評価もあわせて行っていきたいと考えている．

## 文 献

[1] Lawrence Rabiner, Biing-Hwang Juang 共著，古井貞熙 監訳，音声認識の基礎（下）第6章，NTT アドバンステクノロジ株式会社，1995.

[2] S.Yoshizawa and K.Shikano, "Model-Integration Rapid Training based on Maximum Likelihood for Speech Recognition," Proc. of Eurospeech2003, pp.2621-2624, Geneva, Switzerland, Sept.2003.

[3] 芳澤伸一，鹿野清宏，"最尤推定に基づくモデル統合学習法，" 日本音響学会 2003 年秋季研究発表会，分冊 I，pp.105-106，Sept.2003.

[4] K.Ito, M.Yamamoto, K.Takeda, T.Takezawa, T.Matsuoka, T.Kobayashi, K.Shikano and S.Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," The Journal of the Acoustical Society of Japan (E), Vol.20, pp.199-206, 1999.

[5] T.Kawahara, A.Lee, T.Kobayashi, K.Takeda, N.Minematsu, S.Sagayama, T.Utsuro,and K.Shikano, "Free software toolkit for Japanese large vocabulary continuous speech recognition," Proc. of ICSLP, pp.476-479, 2000.

[6] S.Yoshizawa, A.Baba, K.Matsunami, Y.Mera, M.Yamada and K.Shikano, "Unsupervised Speaker Adaptation Based on Sufficient HMM Statistics of Selected Speakers," Proc. of ICASSP, pp.1269-1272, 2000.

[7] 芳澤伸一，馬場朗，松浪加奈子，米良祐一郎，山田実一，李晃伸，鹿野清宏，"十分統計量と話者距離を用いた音韻モデルの教師なし学習法，" 信学論 DII，Vol.J85-D-II, No.3, pp.382-389, 2002.