Vol.2021-NL-251 No.5
Vol.2021-SLP-139 No.5
2021/12/1

# GMM適応に基づく話者認識における混合要素整合の検討

野田 脩介[1]　川端 豪[1]

**概要**：話者認識のための定番手法である混合ガウスモデル（GMM）において，混合要素の順番についてなんら制限が設けられていないため，GMM 適応における EM アルゴリズムの進行につれて，混合要素の平均値の大小関係が入れ替わることがある．これは GMM スーパーベクトルの不整合を引き起こし，話者認識性能の劣化につながる．本報告はまず，GMM 適応における EM アルゴリズムに最大事後確率（MAP）基準を用いた，パラメータ推定の安定化効果を確認する．次に，適応後の GMM スーパーベクトルの要素を特徴量ごとに整列（Sort）することによって混合要素の不整合を防止する．話者認識実験の結果，両者を併用（MAP&Sort）した場合に最良の結果が得られた．

# A Study of Mean Order Alignments for the GMM Adaptation-based Speaker Recognition

***Abstract:*** The GMM (Gaussian Mixture Model) adaptation is a promising technology for speaker recognition systems. Because the GMM element order is arbitrary, GMM mean values can shift in order during EM (Expectation Maximization) iterations. This phenomenon leads to the miss alignments of GMM supervectors and degrades the speaker recognition accuracy. This paper, first, describes the MAP (Maximum A posteriori Probability) approach for stabilizing the GMM parameter estimation. Second, we introduce the "Sort" approach for re-arranging the mean values sorted in order in each speech feature. Experiment results show that the combination of MAP and Sort approaches well prevent miss alignments problem.

## 1. はじめに

話者照合・話者認識研究の歴史は長く，音声のあらゆる分野と関連を持っている．米国で 1993 年頃，クレジットカードに話者照合機能を付加するフィールドテスト等が行われるなど，実用システムの波及効果も大きい [1].

近年，音声による個人認証の視点から，話者照合の研究が著しく発展している．現在主流になっている i-vector 法では，ある話者の音響的な特徴がどのような特徴空間に分布しているかが重要になり，その第一段階として混合ガウスモデル（GMM, Gaussian Mixture Model）による分布近似が用いられる [2].

高見らは GMM 適応による話者モデルの構成過程において，GMM の混合要素である正規分布の平均値が適応の途中で交差する現象について報告している．またこれは話者認識精度に悪影響を及ぼしている [3].

GMM 適応を用いる話者認識においては，多数の話者の音声データから作成した共通の特徴分布モデルを初期値として，各話者の音声データを用いて GMM をその話者に適応していく．ある話者のある特徴量に対する分布が，共通分布と大きく異なっていれば，適応データの影響を強く受けて一部混合要素の平均値が入れ替わっても不思議ではない．Reynolds らは，この GMM 適応において最大事後確率（MAP, Maximum Aposteriori Probability）基準を導入することによってパラメータの推定が安定することを示した [4].

本報告は，話者認識のための GMM 適応過程で発生する平均値交差問題への対策として，次のような検討を行う．

- GMM 適応におけるパラメーター（平均値，分散，混合係数）の推定に，ベイズの定理に基づき事後確率を最大化することによって推定の安定性を高める MAP 基準の導入を試みる（MAP）.
- GMM の混合要素となる正規分布の平均値をソートすることによって，各話者の GMM スーパーベクトルと入力音声の GMM スーパーベクトルの間で要素の順番の不適合を解消する（Sort）.

[1] 関西学院大学 理工学部
Kwansei Gakuin University，School of Science and Technology

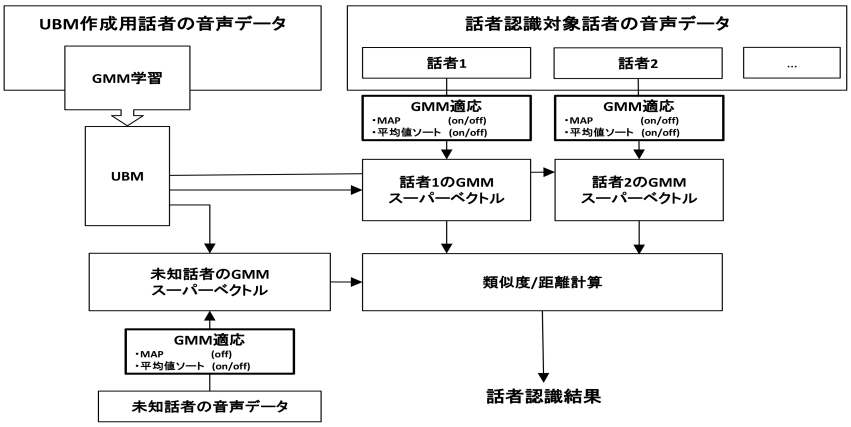


図 1　GMM スーパーベクトルに基づく話者認識

- 上記 2 種類の対処法を併用する（MAP&Sort）.

また，話者照合・話者認識においては学習データと評価データの発話長が大きく異なると精度が低下するといわれている [5].

本報告では，適応データをごく短い区間（4 秒）から順次増加させ，発話長が短い場合の性能特性についても観察する.

## 2. GMM スーパーベクトルに基づく話者認識

### 2.1 話者認識の枠組み

ある 1 つの特徴量に注目し，その確率分布を混合ガウスモデル (GMM, Gaussian Mixture Model) によって表現する．これは複数の正規分布の重み付き和によって任意の確率分布を表現する手法である．任意の確率分布 $p(x)$ を $K$ 個の正規分布 $N(x|\mu_k, \sigma_k^2)$ の和として表現する.

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \sigma_k^2) \qquad (1)$$

ここで $\mu_k$，$\sigma_k^2$，$\pi_k$ は $k$ 番目の正規分布の平均値，分散，混合係数である.

図 1 に，GMM スーパーベクトルに基づく話者認識処理の流れを示す．図の左上にある UBM 作成用話者の音声データに対して GMM 学習を適用し，共通の特徴分布モデル（UBM, Universal Background Model）を作成する.

図の右上にある話者認識対象の音声データに対して GMM 適応を行うことによって，各話者の GMM を作成する．このとき，前述の UBM を適応の初期値として用いる．図中の GMM 適応のブロックには「MAP」および「平均値ソート」の記述があり，本研究ではこれらの機能の「on/off」の組み合わせを検討する．これらの機能については後の節で詳述する．各特徴量に対し 1 次元の GMM を適用し，混合要素である正規分布の平均値をすべての特徴量について集めたものを GMM スーパーベクトルと呼ぶ.

一方，図の左下から未知話者の音声データが入力されると，このデータに対して GMM 適応を行い，またその平均値を集めて GMM スーパーベクトルを作成する．この未知話者のスーパーベクトルと，あらかじめ求められている各認識対象話者のスーパーベクトルとの類似度，あるいは距離を計算することによって話者の判定が行われる.

### 2.2 平均値交差による混合要素の不整合

式 (1) において正規分布の要素番号と平均値の値の間には関係がないので，混合要素が特に平均値の順に並ぶことはない．そのために GMM の適応の過程において平均値の大小が入れ替わることがある.

図 2, 図 3 に平均値交差の実例を示す．ある話者の GMM を求める過程において，あるスペクトル特徴量に注目し，その分布を混合数 8 の GMM で表現する．UBM を初期値とし，EM アルゴリズムの繰り返し計算を用いてパラメータを求めていく．図の横軸は EM アルゴリズムの繰り返し回数（EM 回数）を表している．縦軸は特徴値であり，各混合要素の平均値が EM 回数によって変化していく.

図 2 は特徴量（MFCC1 次）について，EM 回数の増加に伴って各混合要素の平均値がどのように変化しているかを示したものである．いくつかの混合要素の平均値はもともと近い値に固まっているが，EM 回数 15 付近で順番が入れ替わっている現象が観察される.

図 3 は特徴量（MFCC3 次）について，EM 回数と混合要素の平均値の変化を示したものである．前の例では比較的早い時期に平均値交差が起きていたが，こちらの場合は学習がだいぶ進んだ EM 回数 80 回付近で平均値が交差している．学習の進行状態に関わらずいろいろな EM 回数で

Vol.2021-NL-251 No.5
Vol.2021-SLP-139 No.5
2021/12/1

平均値の交差が観察された．

このような平均値交差が起こることによって，GMM の混合要素の順番と平均値の順番が一致しないことが起こりうる．

さて，図 1 に示したように，話者の判定は入力音声の GMM スーパーベクトルと認識対象話者の GMM スーパーベクトルの類似度や距離を用いて行われる．このとき，GMM の混合要素の順番が入れ替わってしまうと適切な類似度が計算できなくなる．

図 4 を用いて，ベクトル要素の交差が類似度計算に及ぼす悪影響について説明する．2 次元のベクトルを想定し，あるカテゴリーの平均ベクトルが $\mu = (0.2, 0.8)$，入力ベクトルが $x = (0.3, 0.5)$ とする．両者のコサイン類似度は角度 $\theta$ で表される．ベクトル $\mu$ において要素の交差が起きた場合，ベクトル $\mu' = (0.8, 0.2)$ と入力ベクトルのコサイン類似度は角度 $\theta'$ で計算され，正しい角度との間に違いが生じてしまう．

次節において，この GMM 適応過程における平均値交差の問題を回避するために，MAP 基準の利用およびスーパーベクトル要素の平均値ソートについて検討する．

## 3. 混合要素の整合

前節で述べた GMM 適応の進行に伴う平均値交差に対処する手法として，GMM 適応における MAP 基準の利用，および適応の後処理として平均値について検討する．以下詳述する．

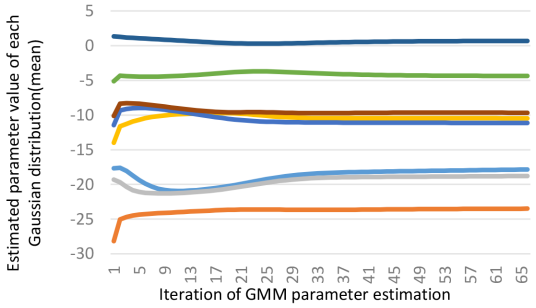


**図 2** 平均値交差の実例（MFCC1 次）

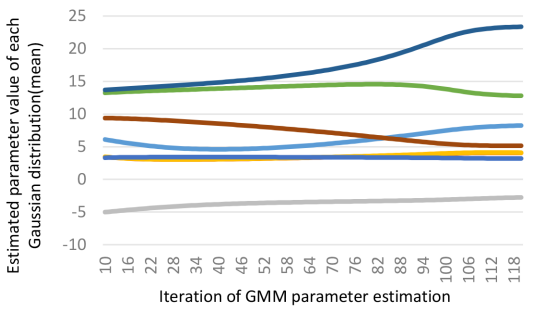


**図 3** 平均値交差の実例（MFCC3 次）

### 3.1 MAP 推定

MAP（Maximum Aposteriori Probability）基準とは，統計的パラメータ推定において，事後確率を最大にするようにパラメータを決定する方法である．GMM 適応におけるパラメータとは，各混合要素に対する平均値，分散，混合係数であり，これらのパラメータを観測データに基づいて逐次的に推定していく必要がある．この時，観測データのみを用いて条件付き確率（尤度）が最大になるようにパラメータを決定するのが最尤推定法であり，観測データとパラメータの事前確率分布から計算される事後確率（Aposteriori Probability）が最大になるようにパラメータを推定するのが MAP 推定法である．事前確率分布というのは 1 観測ステップ前の事後確率分布なので，過去のパラメーター推定の結果が現在のパラメータ推定に反映されるのでパラメータが急に大きく変化しなくなり推定の安定性が増すと考えられている．

このように MAP 基準を用いたパラメータ推定では，事前分布と適応データのバランスを取りながら推定を進めるので，適応データの分布の偏りに大きく影響されず，適応の進行に伴ってパラメータが急に変化することを避けることができる．適応データの偏りに対する急なパラメータ調整は，前節で述べた平均値交差の原因のひとつと考えられるので，GMM 適応において MAP 基準を利用すればこの問題の解決となる可能性がある．

Reynolds らは GMM 適応に基づく話者照合技術において，UBM から話者モデルを導出するための MAP 推定に基づく定式化を行った [4]．本研究では平均値交差の問題を解決するひとつの手法として GMM 適応における MAP 基準の導入を検討する．

以下のような手順（EM アルゴリズム）で適応計算を行う．

1. 混合要素 $k$ $(k = 1, \ldots, K)$ に対する，平均 $\mu_k$，分散 $\sigma_k^2$，混合係数 $\pi_k$ を UBM の値で初期化する，対数尤度の初期値を計算する．

2. 「E ステップ」では，音声データのある区間（フレー

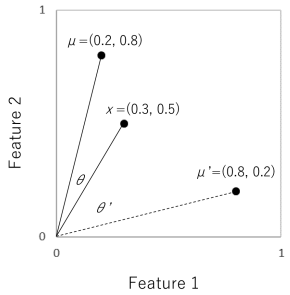


**図 4** ベクトル要素交差の類似度計算への悪影響

ム）$x_i$ が，混合要素 $k$ にどのくらい関わっているのかを計算する．

$$\gamma_{ik} = \frac{\pi_k N(x_i|\mu_k, \sigma_k^2)}{\sum_{j=1}^{K} \pi_j N(x_i|\mu_j, \sigma_j^2)} \tag{2}$$

ただし，

$$N(x_n|\mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{(x_n-\mu_k)^2}{2\sigma_k^2}\right\} \tag{3}$$

音声データの全区間（フレーム）の特徴量 $x_i(i = 1, \ldots, n)$ に対し，平均値 $E_k(x)$ および二乗平均値 $E_k(x^2)$ を計算する．

$$E_k(x) = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik}\ x_i \tag{4}$$

$$E_k(x^2) = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik}\ x_i^2 \tag{5}$$

ただし，データの総フレーム数を $N$ として，$N_k$ は以下のように計算される．

$$N_k = \sum_{i=1}^{N} \gamma_{ik} \tag{6}$$

3. 「M ステップ」では，各混合要素について関わっている音声データを集めて正規分布の平均値，分散，混合係数を再計算する．

$$\mu_k^{new} = \alpha_k E_k(x) + (1-\alpha_k)\mu_k \tag{7}$$

$$\sigma_k^{2\ new} = \alpha_k E_k(x^2) + (1-\alpha_k)(\sigma_k^2 + \mu_k^2) - \mu_k^{2\ new} \tag{8}$$

$$\pi_k^{new} = \frac{\alpha_k N_k/N + (1-\alpha_k)\pi_k}{\sum_{k=1}^{K}\{\alpha_k N_k/N + (1-\alpha_k)\pi_k\}} \tag{9}$$

ただし，

$$\alpha_k = \frac{N_k}{N_k + r} \tag{10}$$

4. 対数尤度

$$\sum_{n=1}^{N} ln\left\{\sum_{k=1}^{K} \pi_k^{new} N(x_n|\mu_k^{new}, \sigma_k^{2\ new})\right\} \tag{11}$$

を計算し，対数尤度の変化が閾値 $\epsilon$ より大きければステップ 2 に戻る．

なお，閾値 $\epsilon$ は（$1.0 \times 10^{-6}$）とする．

（10）式に現れる $\alpha_k$ は，MAP 基準によるパラメータ推定において事前分布と新たな入力データのバランスを指定する「関連係数」$r$ を用いて計算される．この関連係数は文献 [4] によれば $r = 8, \ldots, 20$ の範囲で話者認識精度にあまり影響しない．本研究においては同文献にならい $r = 16$ と設定した．また，MAP 基準を用いない場合は，$r = 0$ と設定すればよい．

### 3.2 平均値ソート

GMM スーパーベクトルは，特徴数×混合数の要素数を持ち，特徴量に対応するブロックから構成される．GMM 適応の結果から平均値を集めてスーパーベクトルを作成する際に，各ブロックの内部でその要素となる平均値を特徴値の小さなものから大きなものの順にソートする．このソート操作を未知話者のスーパーベクトル作成，および各認識対象話者のスーパーベクトル作成時に行うことで，その特徴量ブロックの内部では値の順番にベクトル要素が対応付けられることにより，適応の過程で平均値交差が起きていても適切な要素同士を整合できることが期待される．

前項で述べた MAP 推定によって，平均値交差問題が完全に解決されていれば，この平均値ソートはなんの効果ももたらさない．逆に言えば MAP 推定と平均値ソートを同時に行うことによって話者認識性能が向上するようであれば，MAP 基準によるパラメータ推定の安定化は，平均値交差の防止の観点からは完全でないことがわかる．

MAP 推定よりも平均値ソート手法の方が少ない計算量で実現できるので，平均値ソートのみを用いた手法が MAP 推定と同等の精度を実現できるのであれば，計算量の観点から有利である．

以上のような観点から次節においては，平均値交差への対処として，下記の 3 つの手法を用いた話者認識実験を行い，それらの効果を考察する．

(a) MAP 推定

(b) 平均値ソート

(c) MAP 推定と平均値ソートの併用

## 4. 話者認識実験

### 4.1 音声データベース

話者認識技術の応用領域として音声対話システムのフロントエンドが考えられる．対話の初期段階で発話者を判定し，適切な応答を返すことはシステムに対する親近感を向上させる効果がある．このような，音声対話セッションの

初期段階における相手判定のための話者認識の研究に利用できる音声データベースとして，日本音響学会研究用連続音声データベース（ASJ-JIPDEC）がある [6].

音声の収録条件は，標本化周波数 16kHz，ビット長 16bit，モノラルである．その中にある音韻バランス文 A セットを話者認識のために利用する．

### 4.2 GMM 適応による話者モデルの学習と評価

ASJ-JIPDEC に含まれる 64 名の話者 (男性 30，女性 34) のうち 32 名を UBM 作成用，残り 32 名を話者認識評価用に用いる．認識対象話者によって発声された A セット（50 文）の音韻バランス文のうち，奇数番号を学習用，偶数番号を適応・評価用に用いる．25ms の分析窓で 10ms フレームごとに分析し，12 次元の MFCC に変換する [7]．これに対数パワーの値を加え，13 次元の音響特徴量を計算しておく．表 1 に音声の分析条件をまとめておく．

まず，UBM 作成用話者 32 名の音声データを用いて，話者に共通の特徴分布モデル (UBM) を作成する．EM アルゴリズムによる GMM 学習を行い，各特徴量について混合数 8 の平均値・分散・混合係数を学習する．

次に，この UBM を初期値として，各認識対応話者の学習データ (奇数番号の文) を用いて，GMM 適応を行うことにより，その話者の GMM を求める．この GMM 適応に際し，前節に述べた MAP 推定の on/off，平均値ソートの on/off を組み合わせ，それぞれの効果を検討する．

話者認識実験への入力は，各話者について評価用データ（偶数番号の文）をすべて接続した約 90 秒の音声を，4 秒ずつ増加させて，対話セッションの開始間もない少量の音声データで話者認識しなければいけない状況をシミュレートする．評価データの長さのバリエーションは 4 秒から 4 秒ずつ増加させ実験を行う．これによって，適応データが十分な長さある場合の話者認識精度だけでなく，適応データが短い場合の認識精度も評価できる．

ある長さの未知話者の音声から求めたスーパーベクトルと，あらかじめ計算した各認識対象話者のスーパーベクトルの類似度，あるいは距離を計算し，話者認識精度を求める．認識対象となる話者の数は 32 である．話者判定のための類似度・距離尺度として，本報告ではコサイン類似度およびユークリッド距離を用いる．

**表 1** 音声の分析条件

| | |
|---|---|
| サンプリング周波数 | 11kHz |
| 特徴量 | MFCC12 個と対数パワー 1 個（13 次元ベクトル） |
| フレーム周期 | 10ms |
| 分析窓長 | 25ms |
| 分析窓 | Hamming |
| 高域協調の係数 | 0.97 |
| フィルタバンクの数 | 24 |
| MFCC の次元 | 12 |

## 5. 結果，考察

未知話者の発声データを 4 秒ずつ増加させて，話者認識精度の変化を観察した．認識対象話者数は 32 である．

GMM スーパーベクトル間の類似度・距離尺度として，コサイン類似度を用いた場合の話者認識精度の変化を，図 5 に示す．横軸は GMM 適応に用いられる未知話者の発声データの積算時間を表しており，最短 4 秒から 4 秒ずつ増加させる．縦軸は話者認識精度（%）である．図中に表示する 4 つの折れ線は，平均値交差対策なし（実線，base），MAP 推定（破線，MAP），平均値ソート（点線，Sort），MAP 推定と平均値ソートの併用（一点鎖線，MAP&Sort）を表している．

どの方式においても，30 秒程度の適応データが利用できる場合は，ほぼ 100%の話者認識精度が達成できている．しかし，より適応データが短い条件では，方式による差が表れてきた．

なにがしかの平均値交差対策を行ったものは，何も対策をしていないものと比べて折れ線の上下変動が少なく，適応データ量の変化に対しての安定度が増していることがわかる．

MAP と base を比較すると，適応データがごく短い（〜8 秒）うちは base のほうが良い精度を示しているが，適応データが長くなるにしたがって安定して MAP のほうが精度が高くなる．

Sort と base を比較すると，12 秒以上の領域で精度が良くなる点は MAP と似ているが，20 秒越えたあたりで若干精度が悪くなっている．このように MAP と Sort は適応データが 30 秒よりも短い条件においていずれも base の性能を改善しているが，その挙動は異なっており「平均値交差の防止」にとどまらない要素が潜在している．よって両者を併用すればさらに精度が改善される可能性がある．

MAP&Sort とそれ以外の方法を比較するといずれの適応データ量に対しても最良の結果が得られており，両者を併用する効果が確認できた．

GMM スーパーベクトル間の類似度・距離尺度として，ユークリッド距離を用いた場合の話者認識精度の変化を，図 6 に示す．縦軸，横軸は図 5 と同じである．

この場合も，30 秒程度の適応データがあればどの方式もほぼ 100%の話者認識率を達成できる．

MAP と base を比較すると，MAP は適応データが長くなると安定して精度が高くなっている．

Sort と base を比較すると，コサイン類似度の場合と同じ適応データ量で同じように精度が良くなっている．

MAP&Sort とそれ以外の方法を比較すると 12 秒以上の領域で最良の結果が得られており，ユークリッド距離を用

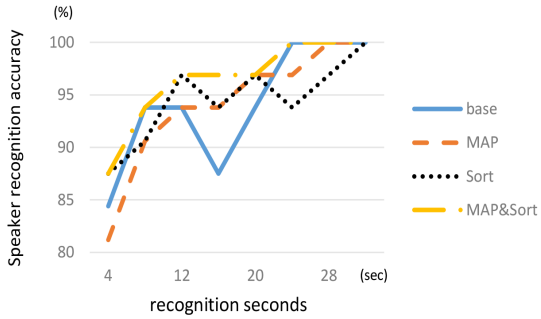


図 5　適応データの増加に対する話者認識精度の変化（コサイン類似度）

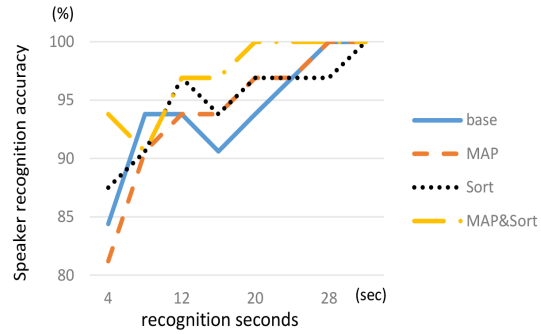


図 6　適応データの増加に対する話者認識精度の変化（ユークリッド距離）

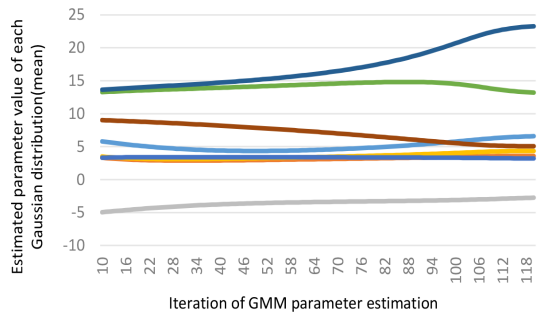


図 7　平均値交差の実例（MFCC3 次，MAP 推定）

いる場合でも両者を併用する効果が確認できた．

## 6. おわりに

GMM スーパーベクトルに基づく話者認識において，GMM 適応の過程で発生する平均値交差問題に対する緩和策を検討した．具体的には，GMM 適応への MAP 推定の利用，GMM 混合要素の平均値のソート，および両者の併用を試み，未知話者の適応データが短い条件下での話者認識精度の変化を観察した．

ASJ-JIPDEC の 32 名の話者データから UBM を学習し，この UBM を初期値として別の 32 名の話者に対する GMM 適応を行った．話者認識実験への入力は，各認識対応話者について評価用データ（偶数番号の文）をすべて接続した約 90 秒の音声を，4 秒ずつ増加させて，対話セッションの開始間もない少量の音声データで話者認識しなければいけない状況をシミュレートした．

平均値交差対策なし，MAP 推定，平均値ソート，MAP 推定と平均値ソートの併用の 4 つの手法を比較評価した．

どの方式においても，30 秒程度の適応データが利用できる場合は，ほぼ 100%の話者認識精度が達成できている．より適応データが短い条件では，方式による差が表れ，なにがしかの平均値交差対策を行ったものは，何も対策をしていないものと比べて認識精度の上下変動が少なく，適応データ量の変化に対しての安定度が増していることがわかった．

MAP 推定と平均値ソートはそれぞれ認識精度に対して良い効果を及ぼすが，その変化特性は異なっており「平均値交差の防止」にとどまらない要素が潜在していることが想像できる．

3.2 節において，MAP 推定と平均値ソートを同時に行うことによって話者認識性能が向上するようであれば，MAP 基準によるパラメータ推定の安定化は，平均値交差の防止の観点からは完全でないと述べたが，この実験結果はそれを裏付けるものである．

図 3 に示したように，GMM 適応の進行に応じて平均値の交差が起きないかどうかを視察によって観察しているが，GMM 適応に MAP 推定を利用することによって，確かにその頻度は減少している．しかし，完全に平均値交差を回避したわけではなく時々それを見つけることがあった．図 7 に MFCC3 次特徴量について MAP 推定を用いて GMM 適応を行っている際に発見した平均値交差の実例を示す．

本報告では GMM 適応における平均値交差の回避という観点から，MAP 推定の利用と平均値ソートの導入を検討したが，実験結果から見て両者は異なる特性を持っており，両者の併用によって最良の性能が得られた．

## 参考文献

[1] 古井貞熙：声の個人性の話, 音響誌, 51, 11, 876-881(1995).
[2] 小川哲司, 塩田さやか：i-vector を用いた話者認識, 音響誌, 70, 6, 332-339(2014).
[3] 高見順子, 川端豪：GMM 適応速度と到達精度に基づく音声対話システムフロントエンドのための話者認識性能の評価法, 信学技報, 118, 426, SP2018-59, 35-40 (2019).
[4] Reynolds et al.：Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing, 10, 19-41(2000).
[5] 塩田さやか：音声を用いた生体認証技術, 話者照合ソフトウェア入門, システム/制御/情報, 62, 2, 63-68(2018).
[6] 小林哲則, 板橋秀一, 速水悟, 竹沢寿幸：日本音響学会研究用連続音声データベース, 音響誌, 48, 12, 888-893(1992).
[7] HTK book, http://speech.ee.ntu.edu.tw/homework/DSP_HW2-1/htkbook.pdf.