テンソル分解に基づく話者情報表現を用いた話者識別の検討 *

☆チン・トゥアン・トゥー，齋藤大輔，峯松信明，広瀬啓吉 (東大)

1 はじめに

話者識別とは入力音声が事前に登録した話者の内誰のものかを判定する技術であり，近年電話によるバンキング，買い物サービス，情報提供サービスなどで用いられている．話者識別をするためには音声から話者性を表現できる特徴量を抽出する必要があるが，音声には話者性の他に，何を喋ったのか，どのような回線が使われたのかなど，話者とは無関係の情報が含まれる．抽出した話者特徴量にこのような非話者性の情報が混在すると識別性能が低下する場合がある．

話者識別の課題は 1990 年代に提案されたガウス混合モデル (GMM)[1] に基づく方法が基礎となっている．2000 年代に入るとサポートベクタマシン (SVM) という線形分類器が導入され，各話者の GMM のガウス分布の平均ベクトルを連結した GMM スーパーベクトル (GMM-SV) を話者特徴量として用い，SVM で識別する手法が提案された [2]．最近では，GMM-SV を次元圧縮することで得られる i-vector を話者特徴量として用いるの が話者識別における標準的な手法となっている [4]．一方で話者識別のための特徴量としては，テンソルに基づく話者情報表現も可能であり，本稿では，話者識別タスクにおける性能について検討する．

テンソルの表現では複数の軸方向へ情報を分離する特徴を有し，話者性を捉える軸のインデックスを固定することで特定の話者を表現できると期待され，識別精度の向上も見込める．本稿では固有声に基づく声質変換法 (EVC)[3] やテンソル解析の知見から，新しい話者性を表現する特徴量を提案し，提案特徴量によって識別性能が向上することを実験的に示す．

2 先行研究

2.1 GMM-SVM[2]

GMM は確率的なモデルであり複数のガウス分布の混合から構成される．音声信号から抽出した特徴量ベクトルの生成過程がこのモデルに従うと仮定される．d 次元の特徴量ベクトル x を確率変数としたとき確率密度関数 $p(x)$ は以下の式で表される．

$$p(x) = \sum_{k=1}^{K} w_k \mathcal{N}(x; \mu_k, \Sigma_k) \tag{1}$$

但し，K を混合数，w_k を混合の重みとしている．$\mathcal{N}(x; \mu_k, \Sigma_k)$ は k 番目の分布であり，μ_k は平均ベクトル，Σ_k は $d \times d$ の分散共分散行列である．

GMM-SVM では，各入力発話に対して GMM を学習するが，学習データが少ないためモデルが過学習される可能性がある．従って，事前に多人数の音声データを用い，一般的な人間の声を表す Universal Background Model (UBM) を学習し，このモデルを初期モデルとして，最大事後確率学習 (MAP) により各入力発話の GMM を推定する．

2.1.1 最大事後確率 (MAP) 学習

MAP 学習はモデルパラメータを推定法の一つであり，訓練データを用いて事後確率を最大化するようにパラメータを更新する手法である．訓練データ $X = [x_1, x_2, \ldots, x_n]$ があるとき，パラメータ θ を確率変数として扱うと

$$\hat{\theta} = \arg\max_{\theta} p(\theta|X) \tag{2}$$

ベイズ定理により以下のように書き直すことができる．

$$\hat{\theta} = \arg\max_{\theta} p(X|\theta)p(\theta) \tag{3}$$

即ち，尤度 $p(X|\theta)$ と事前確率 $p(\theta)$ の積で表され，対数をとると下記のようになる．

$$\hat{\theta} = \arg\max_{\theta} [\log p(X|\theta) + \log p(\theta)] \tag{4}$$

この式から MAP 推定法は対数尤度に対数事前確率を加えたものを最大にすると分かる．従って，MAP 推定はペナルティー付き最尤推定法とも呼ばれる．

以下に具体的に MAP 学習による GMM の推定手順を示す．あるデータ x_i が観測された時，これが GMM の k 番目のガウス分布から生成される確率 $\gamma_{k,\,i}$ が以下の式で計算できる．

$$\gamma_{k,\,i} = p(k|x_i) = \frac{w_k \mathcal{N}(x_i; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} w_j \mathcal{N}(x_i; \mu_j, \Sigma_j)} \tag{5}$$

従って，k 番目のガウス分布から生成される確率的サンプル数 N_k と確率的平均ベクトル e_k はそれぞれ以

---

*Speaker identification using tensor-based representation of speakers by Trinh Tuan Tu, Daisuke Saito, Nobuaki Minematsu, Keikichi Hirose (The University of Tokyo)

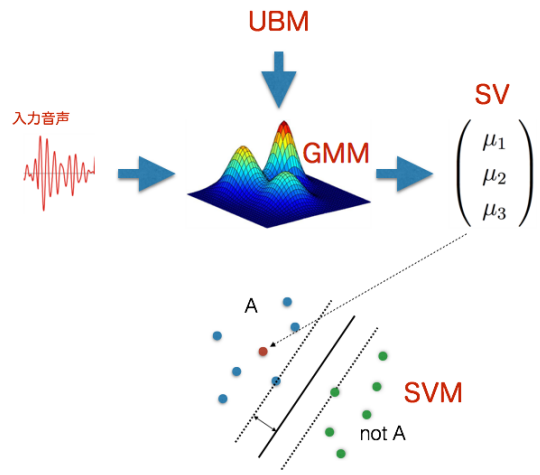

Fig. 1　GMM-SVM による話者識別

下のように表される．

$$\begin{aligned} N_k &= \sum_{i=1}^{N} \gamma_{k,\,i} \\ e_k &= \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{k,\,i} x_i \end{aligned} \tag{6}$$

平均ベクトル μ_k は次のように更新される．

$$\hat{\mu_k} = \frac{\tau}{N_k + \tau} \mu_k + \frac{N_k}{N_k + \tau} e_k \tag{7}$$

理論的に τ は UBM を推定するときの学習データのうち，k 番目のガウス分布から生成されたサンプル数であるが，実際上は固定値を与えることが多い．本稿の範囲では $\tau = 5$ としている．

2.1.2　GMM スーパーベクトルと SVM を用いた識別

GMM-SV とは GMM を構成するガウス分布の平均ベクトルを連結して一つのベクトルにしたものである．GMM の各分布が音素に対応していると考えるならば，スーパーベクトルは，その話者の各音素の音響特徴を連結したベクトルと考えられ，話者の特徴量とする．ある入力音声が話者 A のものであるか否かを図 1 のように SVM で識別する．

2.2　i-vector

i-vector とは GMM スーパーベクトル M を低次元の部分空間へ写像したものであり，以下の式で定義される．

$$M = m + Tw \tag{8}$$

但し，m は UBM から抽出したスーパーベクトル，T は話者とチャネルによる音声のばらつきをモデル化した低次元空間へ射影行列 (基底ベクトル群) である．重みベクトル w が i-vector と呼ばれ，これを話者特徴量として用いる手法が 2011 年に提案された [4]．しかし，GMM-SV を圧縮してから話者特徴量として用いるという観点から考えると，以下で説明する，声質

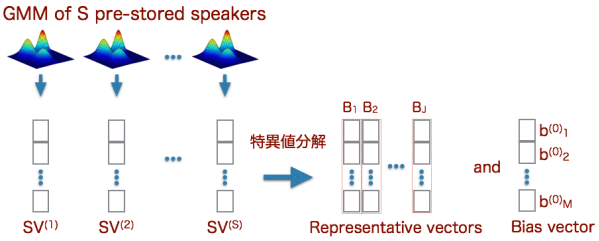

Fig. 2　GMM-SV による話者空間の構築

変換分野における固有声に基づく話者表現と i-vector は基本的に同一視できる．

2.3　固有声 (EV) にに基づく話者表現

この手法では事前学習用 S 人の話者の特徴を用いて任意の目的話者の特徴を表す．事前学習用話者の GMM を学習し GMM-SV を抽出する．図 2 のようにこの S 個の GMM-SV から特異値分解により J 個の基底ベクトル ($J \leq S$) やバイアスベクトルを求め話者空間を構築する．

各目的話者の発話に対して GMM を学習し GMM-SV を抽出して求めた基底やバイアスを用いて以下の式で重みベクトルを導出する．

$$M_s = Bw_s + b \tag{9}$$

但し，b はバイアスベクトル，B は話者空間の J 個の代表ベクトルからなる行列である．

$$B = [B_1, B_2, \ldots, B_J], b = [b_1^{(0)\top}, \ldots, b_M^{(0)\top}]^\top \tag{10}$$

M_s は GMM-SV，w_s は重みベクトルである．この重みベクトルは目的話者が話者空間のどこに位置しているかを意味しており，この手法ではスーパーベクトルの代わりに話者の特徴量として用いられる．

この手法では目的話者に関する情報を補うために事前学習話者の情報が活用されており，目的話者のデータ量が少ない時にも効率的に話者の判別を行うことができる．しかし事前学習用話者の GMM の平均ベクトルを並べて SV にすると各混合の役割が等価になってしまう．M 個の平均ベクトル群には互いに相関の高いベクトル対や低いベクトル対もあり，単純に連結しただけでは，最適化された話者表現であるとは言い難い．しかも，重みベクトルの次元数も事前学習用の話者数で限られている．結果的に必ずしもこのような次元圧縮が話者識別の性能を向上させるとは限らないといえる．

3　提案手法

3.1　多重線形解析

テンソルの定義にはいくつかの方法があるが，古典的な方法ではベクトルや行列を一般化した多次元配列の表現と定義される．テンソルにおける個々のイン

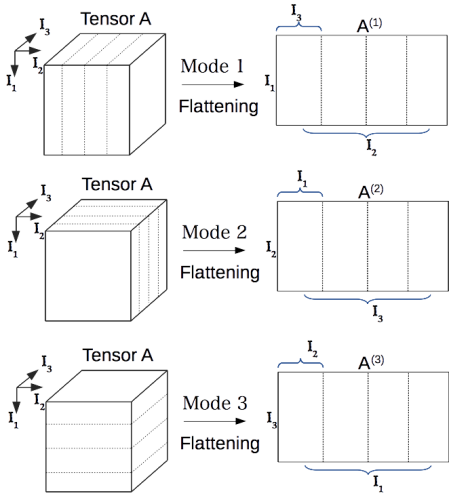

Fig. 3　3 階テンソルの平坦化

デックスがモードとよばれる．あるモードのインデックスに沿うようにスライスし平坦化することでテンソルを行列の形で表現できる．図 3 では $\mathcal{A} \in \mathcal{R}^{I_1 \times I_2 \times I_3}$ の 3 階のテンソルを行列 $A_{(1)}$，$A_{(2)}$，$A_{(3)}$ への平坦化する操作を示している．平坦化操作を用いることで，テンソルと行列間の積を定義できる．テンソル $\mathcal{G} \in \mathcal{R}^{I_1 \times I_2 \times \cdots \times I_N}$ と行列 $B \in \mathcal{R}^{J_n \times I_n}$ の「モード n 積」はテンソル $\mathcal{A} \in R^{I_1 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \times \cdots \times I_N}$ となり，これを $\mathcal{A} = \mathcal{G} \times_n B$ と表現し，平坦化行列を用いて $A_{(n)} = B.G_{(n)}$ のように計算できる．

固有声に基づく話者表現では話者空間の基底を求めるために特異値分解 (SVD) が用いられた．ある行列 A を 2 階のテンソルと見なし，以下の式のように特異値分解することができる．

$$A = USV^\top = S \times_1 U \times_2 V \tag{11}$$

但し，S は対角行列で，U，V はそれぞれ直交行列である．テンソル解析では SVD を一般化した分解がある．例えば，3 階テンソル A は以下のようにコアテンソルと 3 つの行列の積で表現できる．

$$\mathcal{A} = S \times_1 U_1 \times_2 U_2 \times_3 U_3 \tag{12}$$

得られた行列が直交行列である場合，式 (11) が Tucker 分解とよばれる [5]．但し，SVD とは異なり，コアテンソル S は必ずしも対角テンソルとは限らない．

3.2　Tucker 分解による話者空間の構築 [6]

EV に基づく方法と同様，事前学習用 S 人の話者の特徴を用いて話者空間を構築し任意の目的話者の特徴を表現する．事前学習用の各話者の GMM を学習し，スーパーベクトルとは異なり平均ベクトルを行列状に並べ，当該話者を $M \times D$ の行列で表現する．但し，M を GMM の混合数，D を平均ベクトルの次元数とする．次に全話者から得られる話者行列の平均行列を求め，これをバイアス行列 b とする．これを各話者の行列から予め減算しておく．S 人の学習話者に対して求まる S 個の $M \times D$ 行列を 3 階のテンソルとして考え，$\mathcal{M} \in \mathcal{R}^{M \times D \times S}$ で表される．このデータセットを Tucker 分解すると 3 つの基底行列を導出することができ，以下のように表される．

$$\mathcal{M} = \mathcal{G}^{M \times D \times S} \times_1 U^{(M)} \times_2 U^{(D)} \times_3 U^{(S)} \tag{13}$$

但し，$U^{(M)} \in \mathcal{R}^{M \times M}$，$U^{(D)} \in \mathcal{R}^{D \times D}$，$U^{(S)} \in \mathcal{R}^{S \times S}$ である．それぞれの基底行列が GMM 混合要素，平均ベクトルの次元，話者インデックスの効果を捉えている．第三モードを固定することで，特定の話者を表す行列を表現することができる．

$$\hat{\mu}^{(n)} = \mathcal{G} \times_1 U^{(M)} \times_2 U^{(D)} \times_3 U^{(S)}(n, :) \tag{14}$$

グルーピングして書き直すと以下のようになる．

$$\hat{\mu}^{(n)} = U^{(M)} \left\{ \mathcal{G} \times_2 U^{(D)} \times_3 U^{(S)}(n, :) \right\}^\top \tag{15}$$

これは U を基底とし，その他の項を重みパラメータと考えることができ，任意の目的話者の発話は以下のような行列で表される．

$$\mu^{(n)} = U^{(M)} W_n^\top + b \tag{16}$$

但し，$\mu^{(n)}$ は発話 n の GMM の平均ベクトルを並んで得られた行列であり，W_n は発話 n の重み行列である．本研究では重み行列の各行を連結して一つのベクトルとして扱った。これらのベクトルを話者の特徴量とし、SVM で識別を行った．

4　実験

提案手法の有効性を検証するために GMM-SVM，EV に基づく手法，テンソル解析に基づく手法の 3 つの手法で同一の実験条件で話者識別の実験を行い，識別率を比較した．

4.1　実験方法

日本音響学会新聞記事読み上げ音声コーパス JNAS を用いた．このコーパスでは男女各 129 名，合計 258 名の読み上げ音声が収録されている．各話者に対して新聞記事文約 100 文，音声バランス文約 50 文を発声させている．収録環境としてはヘッドセットマイク (HS) と卓上型マイク (DT) で同時録音され，本実験では HS と DT 両方のデータを用いて実験を行った．

音声ファイルからケプストラム分析することによって，ケプストラムベクトル系列に変換し，特徴量ファイルを作成する．分析条件は表 1 で示す．

m001〜m060 の 60 名の男性と f001〜f060 の 60 名の女性を識別対象話者として，他の話者 (138 名) の

Table 1　ケプストラム分析条件

シフト長	10[msec]
窓長	25[msec]
サンプリング	16kHz
特徴ベクトル	MFCC+Energy+Δ

データを用いて UBM を学習する．混合数を 128 にした．ケプストラム分析と GMM，UBM の作成には Hidden Markov Model Toolkit (HTK)[1] を用いた．

EV に基づく手法や提案手法では話者空間の構築に使う事前学習用のデータが必要であるが，本研究では，UBM の学習に用いた話者のデータを活用した．一人あたりに同一内容の発話を HS データから一つ選び，$S = 138$ 発話で話者空間を構築した．

話者特徴量抽出後は，SVM で識別を行う．本研究では，各話者あたりに新聞記事文 20 発話や音声バランス文 10 発話を用いて SVM を学習し，残りの発話をテストに用いた．SVM の学習及びテストにおいて，HS のデータか DT のデータ両方を用いることはせず，どちらか一方を用いて実験を行っている．SVM の実装には LIBLINEAR[2] を用いた．

4.2　実験結果

提案手法では重み行列が $M \times D$ の行列である．但し M は混合数，D は平均ベクトルの次元数であり，$M = 128$，$D = 26$ にしている．しかし実際に求められた全ての重み行列では行 58 以降は，10^{-15} のオーダーの値で埋められていた．これは Tucker 分解することによって情報が圧縮されたと考えられる．よって本実験で重み行列の最初の 57 行だけを用いてベクトル化し，識別実験を行った．

識別結果を表 2 で示す．SVM の学習とテストは HS 同士，DT 同士のミスマッチがない条件，及び HS と DT のミスマッチがある条件の合計 4 種類の条件で行った．SVM の学習とテストに用いたデータタイプが HS-HS，DT-DT，HS-DT，DT-HS の条件においては，EV-based 手法と比べて誤り削減率がそれぞれ 50.6%，70.0%，52.5%，67.7% となっている。どういう条件でも 50% 以上改善しているので，提案手法の優位性を示せた。

また，EV-based の手法による識別性能は GMM-SVM 手法より改善されることが見られなかった．SVM の学習やテストに用いるデータタイプが DT-HS であるときむしろ性能が低くなった．これは事前学習用の話者数が少なすぎるが原因と考えられる．

[1] http://htk.eng.cam.ac.uk
[2] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

Table 2　識別結果

データタイプ		識別率		
SVM 学習	SVM テスト	GMM-SVM	EV-based	Tensor-based
headset	headset	99.523%	99.470%	99.738%
desktop	desktop	99.437%	99.349%	99.805%
headset	desktop	71.572%	75.205%	88.217%
desktop	headset	78.299%	76.640%	92.453%

5　おわりに

本稿では声質変換における知見に基づいて，事前学習用話者のデータセットをテンソルで表現して分解することによって話者空間を構築し，話者識別の検討を行った．目的話者の各発話に対して，この空間のどこに位置するかを重み行列で表現した．この重みを話者の特徴量として SVM で話者識別を行うことによって識別性能向上を確認できた．特に，学習と評価時での環境ミスマッチがある場合に，その効果が大きかった．

今後目的話者のデータ量と事前学習用話者数を変えつつ提案手法と従来手法の性能を比較する予定である．今回 EV-based 手法の性能と比較したが，i-vector の手法も同一の実験条件で再現実験を行い提案手法の性能と比較したい．

今回話者の特徴量を直接に用いて SVM で識別を行った．話者認識に関与しない話者内の変動要因，あるいはチャネルの差異による影響を低減する処理を明示的に入れると識別性能がもっと向上できると考えられる．今後チャネルあるいはセッション変動補正の手法も検討していきたい．

参考文献

[1] D. A. Reynolds *et al.*, “Robust text-independent speaker identification using Gaussian mixture speaker models,” IEEE Trans. Speech Audio Process., vol.3, no.1，pp.72–83, 1995.

[2] W. M. Campbell *et al.*, “Support vector machines using GMM supervectors for speaker verification,” IEEE Signal Processsing Letters, vol.13, pp.308–311, 2006.

[3] 戸田智基 *et al.*, “固有声に基づく性質変換法,” IEICE Technical Report, SP2006–39, 2006.

[4] N. Dehak *et al.*, “Front-end factor analysis for speaker verification,” IEEE Trans. Audio Speech Lang. Process., vol.19, no.4, pp.788–798, 2011.

[5] L. R. Tucker, “Some mathematical notes on three-mode factor analysis,” Psychometrika, vol.31, no.3, pp.279–311, 1966.

[6] D. Saito *et al.*, “One-to-Many Voice Conversion Based on Tensor Representation of Speaker Space,” INTERSPEECH, pp.653-656, 2011.