話者性の効率的特徴表現を自動抽出可能な生成モデルに基づく話者認識

学籍番号　２３４１３５５６　氏名　服部 貴文
指導教員名　徳田 恵一

1　はじめに

近年の話者認識では，JFA (Joint Factor Analysis) [1, 2] や i-vector [3] と呼ばれる手法が提案され，少量の発話データで高い精度を達成している．これらの手法は，因子分析を利用し，ある入力に対して次元数を削減して話者固有の特徴を抽出するため，少量のデータでも効率的に話者の判別を行うことが可能である．しかし，i-vector などの手法では，各発話の音響特徴量から GMM を作成し，そのモデルパラメータを因子分析の入力として，特徴を抽出している．そのため，抽出される特徴量は音響特徴量ではなく，GMM のモデルパラメータから生成されるため，話者の詳細な特徴を表現できない可能性が存在する．そこで，本研究では音響特徴量を直接表現できる因子分析手法を提案する．提案法では，生成モデルに基づき音響特徴量から直接話者性を抽出することができるため詳細な表現が可能であり，認識精度の改善が期待される．

2　i-vector に基づく話者認識

i-vector に基づく話者認識では，因子分析を利用して，話者に含まれる固有の特徴を抽出し，得られた特徴量を比較することで話者を判別する．この手法では，特徴抽出に因子分析を用いているため，次元を削減して特徴を表現することが可能であり，少量の発話データでも効率良く話者性を表現することができる．

一般的に i-vector に基づく手法では，因子分析の入力として，各発話毎に GMM の平均成分を結合したスーパーベクトルを用いる．今，発話 u から作成された GMM スーパーベクトル $\boldsymbol{M}_u$ は以下で定義される．

$$\boldsymbol{M}_u = \boldsymbol{T}\boldsymbol{w}_u + \boldsymbol{m} \tag{1}$$

ここで $\boldsymbol{M}_u$ は，全話者の発話データから作成される UBM (Universal Background Model) を事前情報として事後確率最大化 (MAP) 法により推定された GMM を用いる．また，$\boldsymbol{m}$ は話者に非依存な GMM スーパーベクトルで UBM から生成され，$\boldsymbol{T}$ は基底ベクトルから構成される固有声行列である．そして，$\boldsymbol{T}$ を用いて推定される $\boldsymbol{w}_u$ は i-vector と呼ばれる因子成分であり，発話の変動を表現している．認識の際には，推定された $\boldsymbol{w}_u$ を用いて，類似度を比較することで，話者の判別を行う．しかし，話者の判別に用いる $\boldsymbol{w}_u$ は，因子分析時に GMM のモデルパラメータを入力としているため，音響特徴量ではなく，GMM のモデルパラメータから生成されている．生成モデルに基づく特徴抽出では，抽出されるモデルパラーメータは話者を表現するために音響特徴量から生成されるべきであり，GMM のスーパーベクトルから生成されるべきではないと考えられる．また，GMM のモデルパラメータを入力としているため，GMM の混合数や事前分布の設定に固有声行列の推定精度が大きく依存する．さらに，直接音響特徴量を表現していないため，話者性を詳細に抽出することができない可能性が存在する．

3　因子分析に基づく音響特徴量の生成モデル

i-vector に基づく手法における問題を解決するため，本研究では話者性の効率的特徴表現を自動抽出可能な生成モデルに基づく話者認識を提案する．提案法では，因子分析を用いて固有声モデルを表現することで，可変長である音響特徴量から直接話者性を抽出する．i-vector に基づく手法と提案法との比較の図を図 1 に示す．

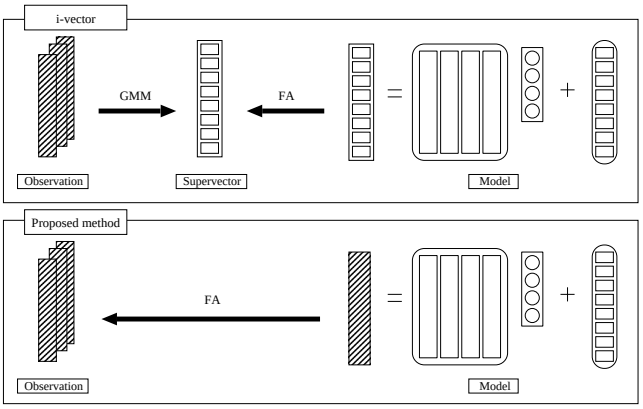

図 1: i-vector と因子分析に基づく音響特徴量の生成モデル

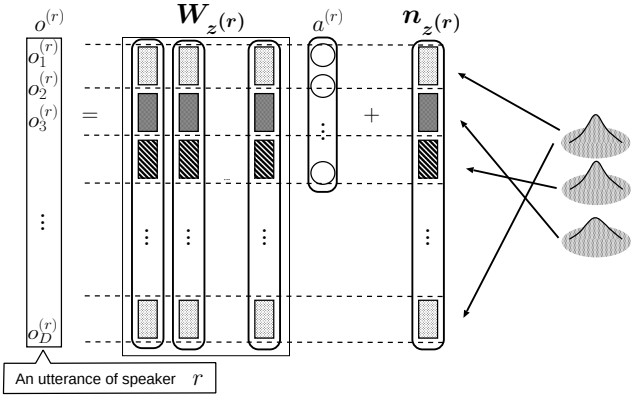

図 2: 因子分析に基づく音響特徴量の生成モデルの構造

提案法では，因子分析における因子負荷行列 $\boldsymbol{W}$ の各列ベクトルを固有声ベクトルとみなし，それらとノイズベクトル $\boldsymbol{n}$ が GMM から出力されると仮定する．因子分析に基づく音響特徴量の生成モデルの構造を図 2 に示す．因子分析に基づく音響特徴量の生成モデルでは，固有声ベクトルが入力データの変動に応じて変形するため，可変である入力データを直接モデル化することが可能である．GMM パラメータは固有声の線形結合として表現されるため，モデルの構造として線形変換による特徴抽出を含むモデルとなる．

話者 r の発話の観測系列を $\boldsymbol{o}^{(r)}$ とすると，このモデルによる観測の生成過程は，次式で表される．ただし以下の表記においては簡単のため用意されるデータは 1 話者につき 1 発話と考える．

$$\boldsymbol{o}^{(r)} = \boldsymbol{W}_{\boldsymbol{z}^{(r)}}\boldsymbol{a}^{(r)} + \boldsymbol{n}_{\boldsymbol{z}^{(r)}} \tag{2}$$

ここで，混合要素系列を $\boldsymbol{z}^{(r)}$，因子を $\boldsymbol{a}^{(r)}$，ノイズベクトルを $\boldsymbol{n}_{\boldsymbol{z}^{(r)}}$ とする．$\boldsymbol{W}_{\boldsymbol{z}^{(r)}}$ と $\boldsymbol{n}_{\boldsymbol{z}^{(r)}}$ は混合要素番号に依存するため，可変長の音響特徴量を直接表現可能である．また，因子ベクトルを各話者ごとに用意しているため因子が話者性を表現するモデルとなっている．全話者の発話データ $\boldsymbol{O} = \{\boldsymbol{o}^{(1)}, \ldots, \boldsymbol{o}^{(R)}\}$

に対する尤度関数は次式で表される．

$$P(\boldsymbol{O}|\boldsymbol{\Lambda}) = \prod_{r} \sum_{\boldsymbol{z}^{(r)}} \int P(\boldsymbol{o}^{(r)}|\, \boldsymbol{a}^{(r)}, \boldsymbol{z}^{(r)}, \boldsymbol{\Lambda}) \times P(\boldsymbol{a}^{(r)}|\boldsymbol{\Lambda}) P(\boldsymbol{z}^{(r)}|\boldsymbol{\Lambda}) d\boldsymbol{a}^{(r)} \quad (3)$$

$$P(\boldsymbol{o}^{(r)}|\boldsymbol{a}^{(r)}, \boldsymbol{z}^{(r)}, \boldsymbol{\Lambda}) = \mathcal{N}(\boldsymbol{o}^{(r)}|\boldsymbol{W}_{\boldsymbol{z}^{(r)}} \boldsymbol{a}^{(r)} + \boldsymbol{\mu}_{\boldsymbol{z}^{(r)}}, \boldsymbol{\Sigma}_{\boldsymbol{z}^{(r)}}) \quad (4)$$

ここで，$\boldsymbol{\Lambda}$ はモデルパラメータを表し，$\boldsymbol{\mu}_{\boldsymbol{z}^{(r)}}$, $\boldsymbol{\Sigma}_{\boldsymbol{z}^{(r)}}$ はそれぞれノイズベクトル $\boldsymbol{n}_{\boldsymbol{z}^{(r)}}$ の平均と分散を表す．因子分析に基づく音響特徴量の生成モデルでは，式 (3) の尤度関数を最大化することで混合要素と固有声ベクトルを同時に最適化する．

4　評価実験

提案法の有効性を確認するために，話者識別実験を行った．実験に用いるデータは，ATR 日本語音声データベース c-set の女性 10 人の音声を用いた．また，学習データは，各話者 10 単語，認識データは，各話者 520 単語を用いた．比較手法として，GMM に基づく話者認識 (GMM-MAP)，i-vector に基づく話者認識 (i-vector)，そして提案法である因子分析に基づく音響特徴量の生成モデルを用いた話者認識 (FA-GMM) を比較する．一般的に i-vector に基づく手法では学習データとして大量の話者から固有声行列を事前に推定し，テスト話者の判別を行うが今回の実験では 10 人分の発話を学習データとする．また，GMM-MAP や i-vector に用いられる GMM の事前分布は UBM を用い，その調整パラメータは事前実験の結果より適切に設定し，基底数や GMM の混合数は認識率が最も高くなるよう調節した．

図 3 に従来法及び提案法の認識結果を示す．また，i-vector と提案法に関する適切な因子数を調べるために，図 4 に混合数 16 の場合のそれぞれの手法について因子数と認識率の関係を示す．認識結果より，従来の GMM-MAP が最も認識率が高い結果となった．一般的には固有声行列の推定に大量のデータを用いて推定を行うため，i-vector に基づく手法の方が GMM-MAP より認識率が高くなるが，今回の実験条件では，テスト話者 10 人のみのデータを用いて推定したことから，固有声行列の推定精度が下がり，i-vector や提案法の認識率が低下したと考えられる．よって，固有声行列推定のデータ量を増加させることにより，GMM-MAP に比べ，i-vector や提案法の認識率が高くなると予想される．また，i-vector と提案法の認識結果を比較すると，ほぼ同等の認識精度となった．i-vector と提案法それぞれ最も認識率が高くなった因子数に注目すると，i-vector では因子が 17 個，提案法では 6 個となり，図 4 を見ても，i-vector の方が話者の特徴を表現するために因子数が必要であることが分かる．このことから，提案法の方が少ないパラメータで効率良く話者の特徴を抽出可能であることが確認できる．また，提案法では音響特徴量から直接因子を推定することができる点で従来の i-vector に比べて詳細に話者性を抽出することができると考えられるが，今回の実験では i-vector と提案法は同等の性能となった．この要因として，提案法では，音響特徴量から因子を推定しているのに対し，i-vector では，全話者のデータを用いた UBM を事前情報として GMM を推定し，そのスーパーベクトルを用いて因子を推定しているため，今回のデータセットではデータ量が少ないことから，その事前情報が効果的に働いたと考えられる．そのため，提案法の推定基準に今回は ML 基準を用いたが，現在 MAP 法や変分ベイズ法など事前分布を利用可能な手法が提案されているため，それらを適用することで提案法の有効性を大きく示すことができると考えられる．よって，今後の課題として，推定基準の改善やデータ量を増加させた場合の認識実験を行う必要がある．

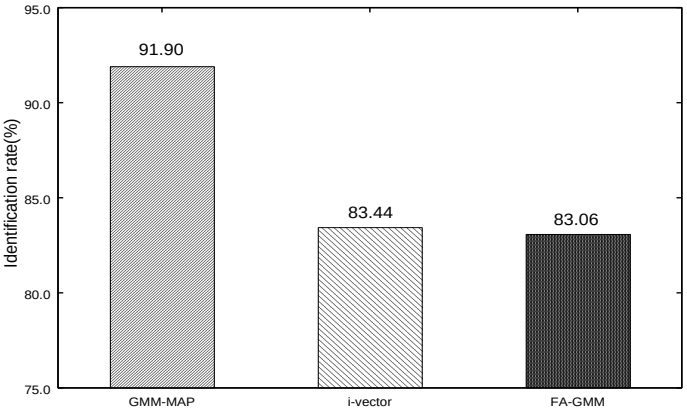

図 3: 認識結果 (対象人数 10 人)

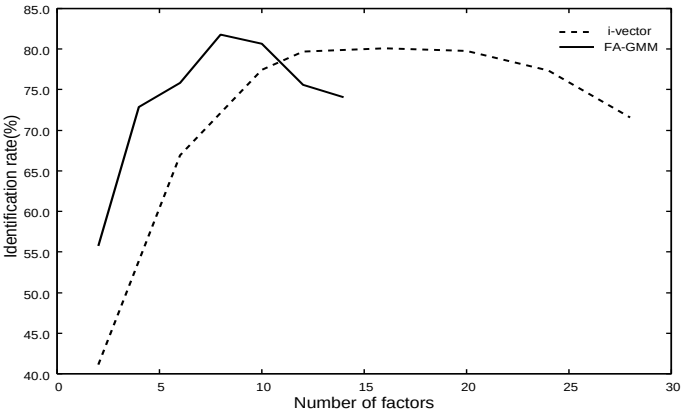

図 4: 因子数と認識率の関係

5　むすび

本研究では近年話者認識の分野で広く用いられている i-vector という手法に対し，因子分析に基づく音響特徴量の生成モデルを用いた手法を提案した．従来の i-vector に基づく手法に比べ，提案法では生成モデルに基づき音響特徴量から直接特徴を抽出することができる．そのため，話者性を詳細に表現することが可能である．話者認識実験の結果，i-vector に基づく手法に比べ，少ないパラメータで効率的に特徴を抽出可能であることが確認できた．しかし，提案法では事前情報を利用していないため，認識精度を比較すると，i-vector に基づく手法と同等の精度となった．そのため，今後の課題として，データ量を増やした場合の認識実験や，提案法に MAP 法や変分ベイズ法などの事前情報を利用可能な推定基準を適用することが挙げられる．

参考文献

[1] P. Kenny, *et al.*, “Speaker and Session Variability in GMM-Based Speaker Verification,” IEEE Trans. on Audio, Speech and Language Processing, vol. 15, no. 4, pp. 1448-1460, 2007.

[2] P. Kenny, *et al.*, “A study of interspeaker variability in speaker verification,” IEEE Trans. on Audio, Speech, and Language Processing,, vol. 16, no. 5, pp. 980-988, 2008.

[3] N. Dehak, *et al.*, “Front-end factor analysis for speaker verification,” IEEE Trans. on Audio, Speech, and Language Processing, no. 99, pp. 1-1, 2010.