# テンソル分解に基づく言語情報表現を用いた言語識別に関する検討*

☆鈴木 颯、齋藤 大輔、峯松 信明 (東大)

## 1 はじめに

言語識別は、入力音声の言語を判定する技術であり、多言語対応の議事録作成支援アプリケーションや国際コールセンター等で用いられている。入力音声から言語を特定するために必要な特徴を抽出することで適切な識別が可能となることが期待されるが、音声は話者やチャネル等の条件によって多様に変化し、これら非言語的特徴の変動による識別性能低下が課題の一つとなっている。近年では、これに着目して提案された i-vector が標準的な言語情報表現の手法になっているが [1]、これは各発話を Gaussian Mixture Model (GMM) でモデル化し、GMM の各分布の平均ベクトルを連結した GMM supervector (GMM-SV) を次元圧縮することによって得られる言語特徴量である。

GMM-SV や i-vector は元来話者識別の分野で提案された特徴量であり [2, 3]、この分野では近年、テンソル分解に基づく話者情報表現が提案されている [4]。これは、行および列がそれぞれ GMM の要素と平均ベクトルに対応するような行列によって一発話を表現し、多数話者分の行列をテンソルとして扱い、テンソル解析を導入することで複数要因からの音響的変動の分離を行なうという手法であり、言語識別にも適用可能であると考えられる。本稿では、テンソル分解に基づく言語表象の効果を実験的に検討することを目的とする。

## 2 先行研究

### 2.1 GMM-SV

GMM-SV は、一発話を GMM でモデル化し、各分布の平均ベクトルを一列に連結した特徴量である [2]。GMM の各分布は音素や単音のような音響的な要素を捉えていることが期待できるため、GMM-SV はそれらの音響的な特徴を発話単位で表現した特徴量と考えることができる。

GMM-SV では一発話の GMM を推定することになるが、これはあらかじめ様々な音声から統計的に話者・言語非依存の Universal Background Model (UBM) を構築しておき、UBM を事前分布として入力発話に対する最大事後確率基準による推定 (MAP 推定) を行なうことで得られる（Fig. 1 参照)。これにより、発話

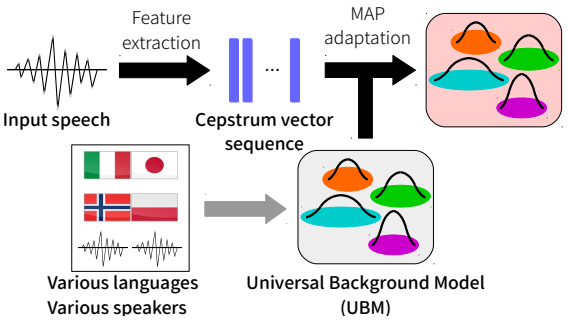


Fig. 1 発話 GMM の推定

GMM の各分布のインデックスと UBM の各分布のインデックスの対応付けが保たれるため、GMM-SV が発話間で比較可能になる。また、一発話という少量のサンプルから安定に GMM を推定できるという利点もある。

### 2.2 i-vector

一発話から抽出された GMM-SV $\boldsymbol{M}$ は、UBM の GMM-SV $\boldsymbol{m}$ と、話者やチャネルの変化による音声のばらつきをモデル化した低次元空間への射影行列 $T$ (Total variability matrix) を用いて

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

と分解される。この $\boldsymbol{w}$ が、i-vector と呼ばれる [3]。GMM-SV には言語識別においてノイズとなりうる話者性・チャネルの成分が含まれたままであるが、言語識別分野における i-vector では低次元空間への射影によってこれらの成分の除去を行なっている。

Total variability matrix による射影は Principal Component Analysis (PCA) に相当する。話者識別分野における i-vector は、GMM-SV に対する PCA という観点から考えると、次に述べる声質変換分野における固有声 (EV) に基づく話者情報表現と基本的に同一視できる。

### 2.3 固有声に基づく話者情報表現

固有声は音声認識におけるモデル適応法として提案された技術であるが [5]、声質変換分野ではこれを適用した固有声変換法 (Eigenvoice Conversion; EVC) が提案されている [6]。GMM に基づく声質変換では、入力話者の特徴量 $\boldsymbol{X}_t$ と出力話者の特徴量 $\boldsymbol{Y}_t$ の結合 GMM を用いて入出力の変換のモデル化を行なうが、EVC ではこの結合 GMM の出力話者側の平均ベ

*A study of language identification using tensor-based language identity representation. by S. Suzuki, D. Saito, and N. Minematsu (The University of Tokyo)

クトルを事前学習用の $S$ 人の話者の特徴を用いて表した EV-GMM $\boldsymbol{\lambda}^{(EV)}$ に基づいて声質変換を行なう。

まず、事前学習用話者毎に GMM を学習し、GMM-SV を抽出する。$S$ 個の GMM-SV に対して特異値分解 (SVD) を行なうことでバイアスベクトルと $K(\leqq S)$ 個の基底ベクトルを求め、これによって話者空間を構築する（Fig. 2 参照）。すなわち、出力話者の GMM-SV $\boldsymbol{M}^{(tar)}$ を以下のようにバイアスベクトルと基底ベクトルの線型結合で表す。

$$\boldsymbol{M}^{(tar)} = \boldsymbol{B}\boldsymbol{w} + \boldsymbol{b} \tag{2}$$

但し $\boldsymbol{B}$ は $K$ 個の基底ベクトルからなる行列である。このように出力話者の GMM-SV が $K$ 次元の重みベクトル $\boldsymbol{w}$ によって制御されるため、$\boldsymbol{w}$ が出力話者を表現していると考えることができる。

重みベクトル $\boldsymbol{w}$ は、出力話者の音声データを用いて最尤基準に基づいて推定することができる。出力話者の特徴量系列を $\boldsymbol{Y}^{(tar)}$ とすると、$\boldsymbol{w}$ は以下のように推定できる。

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \int p(\boldsymbol{X}, \boldsymbol{Y}^{(tar)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) \mathrm{d}\boldsymbol{X} \tag{3}$$

$$= \underset{\boldsymbol{w}}{\operatorname{argmax}}\, p(\boldsymbol{Y}^{(tar)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) \tag{4}$$

EM アルゴリズムを用いて、以下の更新式を得る。

$$\hat{\boldsymbol{w}} = \left\{ \sum_{m=1}^{M} \overline{\gamma}_m^{(tar)} \boldsymbol{B}_m^{\top} \boldsymbol{\Sigma}_m^{(YY)^{-1}} \boldsymbol{B}_m \right\}^{-1} \sum_{m=1}^{M} \boldsymbol{B}_m^{\top} \boldsymbol{\Sigma}_m^{(YY)^{-1}} \overline{\boldsymbol{Y}}_m^{(tar)} \tag{5}$$

$$\overline{\gamma}_m^{(tar)} = \sum_{t=1}^{T} \gamma_{m,t}, \quad \overline{\boldsymbol{Y}}_m^{(tar)} = \sum_{t=1}^{T} \gamma_{m,t} (\boldsymbol{Y}_t^{(tar)} - \boldsymbol{b}_m^{(0)}) \tag{6}$$

$$\gamma_{m,t} = P(m | \boldsymbol{Y}_t^{(tar)}, \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) \tag{7}$$

EVC では事前学習用話者の情報を活用しているため、データが少量でも効率的に出力話者を表現することが可能になっている。しかし、GMM-SV は GMM の各分布の平均ベクトルが単純に一列に並んだ構成になっているため、分布同士の関係性までは考慮されていない。GMM の各分布には相関の高いものも低いものも存在すると考えられるため、それらの関係性を適切に考慮することでより精密な話者情報表現あるいは言語情報表現が可能になることが期待できる。

## 3 テンソル分解に基づく言語情報表現

### 3.1 概要

GMM の各分布の関係性に着目した手法として、テンソル分解に基づく話者情報表現を用いた声質変換 [7] と話者識別 [4] が検討されている。[4, 7] では、GMM の各分布の平均ベクトル群を行列で表し、複

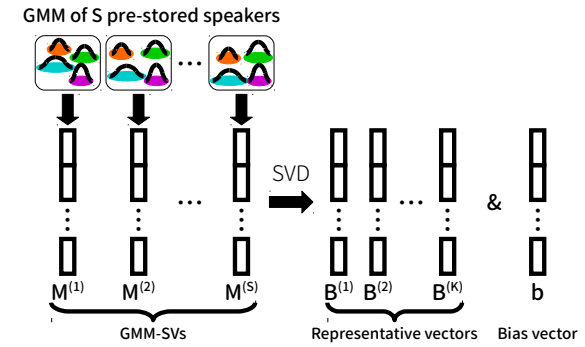


Fig. 2　固有声に基づく話者空間の構築

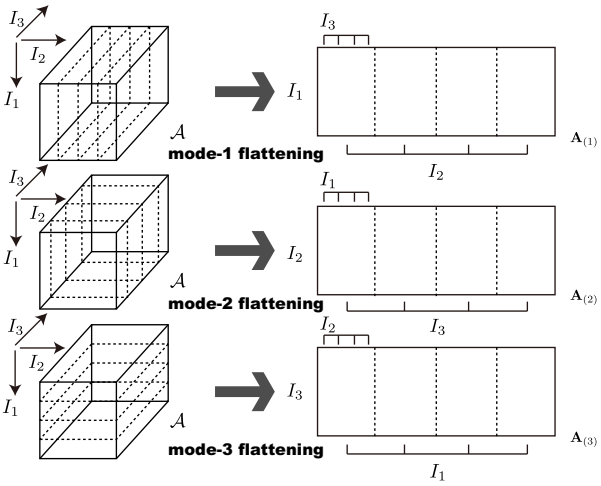


Fig. 3　$I_1 \times I_2 \times I_3$ テンソル $\mathcal{A}$ の平坦化 [7]

数名分の行列をテンソルとして表現する。このテンソルに対して、PCA の射影行列の一つの計算法である SVD を拡張した Tucker 分解と呼ばれるテンソル分解を用いて話者情報を表現している。本稿では、この手法を適用して言語情報表現を行なう。

### 3.2 多重線形解析

テンソルは、行列表現を一般化した多次元配列表現である。テンソルにおける個々のインデックスはモードと呼ばれ、特定のモードに沿ってスライスする平坦化操作によってテンソルを行列の形で表現できる。$\mathcal{A} \in \mathcal{R}^{I_1 \times I_2 \times I_3}$ を 3 階のテンソルとすると、これをそれぞれモード 1, 2, 3 に沿って平坦化した行列 $\mathcal{A}_{(1)}, \mathcal{A}_{(2)}, \mathcal{A}_{(3)}$ は Fig. 3 のようになる。このような平坦化行列を用いて、テンソルと行列間の積が定義できる。テンソル $\mathcal{G} \in \mathcal{R}^{I_1 \times \cdots \times I_N}$ と行列 $\boldsymbol{B} \in \mathcal{R}^{J_n \times I_n}$ のモード $n$ 積 $\mathcal{A} = \mathcal{G} \times_n \boldsymbol{B}$ はモード $n$ の平坦化行列による演算 $\boldsymbol{A}_{(n)} = \boldsymbol{B} \cdot \boldsymbol{G}_{(n)}$ によって定義される。

### 3.3 SVD と Tucker 分解

行列 $\boldsymbol{A} \in \mathcal{R}^{M \times N}$ は、以下のように分解することができる。

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\top} \tag{8}$$

これを SVD と呼ぶ。但し、$\boldsymbol{U} \in \mathcal{R}^{M \times M}, \boldsymbol{S} \in \mathcal{R}^{M \times N}, \boldsymbol{V} \in \mathcal{R}^{N \times N}$ で $\boldsymbol{U}, \boldsymbol{V}$ は直交行列、$\boldsymbol{S}$ は $K$

個の特異値からなる対角行列である。ここで、$K$ は $\boldsymbol{A}$ のランクを表す ($K \leqq \min(M,N)$)。行列 $\boldsymbol{A}$ の SVD を 2 階のテンソルの分解として書き直すと以下のようになる。

$$\boldsymbol{A} = \boldsymbol{S} \times_1 \boldsymbol{U} \times_2 \boldsymbol{V} \tag{9}$$

これを以下のように高階のテンソルに拡張したものを Tucker 分解と呼ぶ [8]。

$$\mathcal{A} = \mathcal{S} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3 \tag{10}$$

但し、$\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3$ が直交行列の場合はコアテンソル $\mathcal{S}$ は密なテンソルとなる。

### 3.4 Tucker 分解による言語情報表現

一発話の GMM の各分布の平均ベクトルを並べて $M \times D$ の行列で表現する。ここで $M$ は混合数、$D$ は平均ベクトルの次元である。始めに全発話の平均を求め、これをバイアス行列 $\boldsymbol{b}$ として各発話の行列から減算しておく。学習データの発話数を $N$ としたとき、$N$ 個の $M \times D$ 行列を 3 階のテンソル $\mathcal{M} \in \mathcal{R}^{M \times D \times N}$ として考えて Tucker 分解を行なうと以下のようになる。

$$\mathcal{M} = \mathcal{G}^{M \times D \times N} \times_1 \boldsymbol{U}^{(M)} \times_2 \boldsymbol{U}^{(D)} \times_3 \boldsymbol{U}^{(N)} \tag{11}$$

但し $\boldsymbol{U}^{(M)} \in \mathcal{R}^{M \times M}, \boldsymbol{U}^{(D)} \in \mathcal{R}^{D \times D}, \boldsymbol{U}^{(N)} \in \mathcal{R}^{N \times N}$ であり、それぞれ GMM 混合数、平均ベクトルの次元、発話インデックスの効果を捉えている。ここで、以下のように $\mathcal{M}$ の第 3 モードを固定することで、特定の発話を表す行列が得られると考えられる。

$$\hat{\boldsymbol{\mu}}^{(n)} = \mathcal{G} \times_1 \boldsymbol{U}^{(M)} \times_2 \boldsymbol{U}^{(D)} \times_3 \boldsymbol{U}^{(N)}(n,:) \tag{12}$$

$\boldsymbol{U}^{(N)}(n,:) \in \mathcal{R}^{1 \times N}$ を重みパラメータ、その他のモード積の項を言語空間の基底と捉えると SVD と等価になる。一方、本稿では GMM の各分布の関係性を考慮するため以下のようなグルーピングを考える。

$$\hat{\boldsymbol{\mu}}^{(n)} = \boldsymbol{U}^{(M)} \left\{ \mathcal{G} \times_2 \boldsymbol{U}^{(D)} \times_3 \boldsymbol{U}^{(N)}(n,:) \right\} \tag{13}$$

$$= \boldsymbol{U}^{(M)} \boldsymbol{W}_n^{\top} \tag{14}$$

$\boldsymbol{U}^{(M)}$ を基底、$\boldsymbol{W}_n \in \mathcal{R}^{D \times M}$ を重み行列とする。次元圧縮の観点から、基底を縮約することで任意の発話は以下のような行列で表される。

$$\boldsymbol{\mu}^{(new)} = \boldsymbol{U}^{(M)} \boldsymbol{W}_{(new)}^{\top} + \boldsymbol{b} \tag{15}$$

$\boldsymbol{U}^{(M)} \in \mathcal{R}^{M \times K} (K \leqq N)$ が表現行列、$\boldsymbol{W}_{(new)} \in \mathcal{R}^{D \times K}$ が重み行列となり、$D \times K$ の重み行列によって発話の言語情報を表現する。

Table 1　音響分析条件

| | |
|---|---|
| サンプリング | 8 bit / 8 kHz |
| 窓 | 25 ms length / 10 ms shift |
| 音響的特徴 | MFCC (7 次元) + SDC (49 次元) |

[4] では式 (15) の最小二乗解として重み行列 $\boldsymbol{W}$ を算出しているが、これも EVC の重みベクトルと同様に、以下のような最尤基準に基づいて $\boldsymbol{W}$ の推定を行なうことができる [7]。

$$\hat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}} \int p(\boldsymbol{X}, \boldsymbol{Y}^{(tar)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{W}) \mathrm{d}\boldsymbol{X} \tag{16}$$

$$= \underset{\boldsymbol{W}}{\operatorname{argmax}}\, p(\boldsymbol{Y}^{(tar)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{W}) \tag{17}$$

EVC と同様に EM アルゴリズムを用いて、以下の更新式を得る。

$$\mathrm{vec}(\boldsymbol{W}) = \left[ \sum_{m=1}^{M} \bar{\gamma}_m^{(tar)} \boldsymbol{U}_m^{\top} \boldsymbol{U}_m \otimes \boldsymbol{\Sigma}_m^{(YY)^{-1}} \right]^{-1} \mathrm{vec}(\boldsymbol{C}) \tag{18}$$

$$\boldsymbol{C} = \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{\Sigma}_m^{(YY)^{-1}} (\boldsymbol{Y}_t^{(tar)} - \boldsymbol{b}_m^{(0)}) \boldsymbol{U}_m \tag{19}$$

$$\boldsymbol{U}_m = \boldsymbol{U}^{(M)}(m,:) \in \mathcal{R}^{1 \times K} \tag{20}$$

$$\boldsymbol{b}_m^{(0)} = \boldsymbol{b}(m,:)^{\top} \in \mathcal{R}^{D \times 1} \tag{21}$$

ここで、vec() は行列を列ベクトルに展開する演算子である。

## 4　言語識別実験

テンソル分解に基づく手法 (Tensor-based) の有効性を検証するため、i-vector と固有声に基づく手法 (EV-based) と合わせて 3 つの手法で言語識別実験を行なった。

### 4.1 実験条件

The National Institute of Standards and Technology Language Recognition Evaluation (NIST LRE) の 2003・2005・2007 年版を用いて、言語識別実験を行なった。このコーパスには電話での会話が継続長 3s, 10s, 30s の 3 つのカテゴリーに分かれて収録されている。NIST LRE 2007 Evaluation Plan に従い、アラビア語・ベンガル語・ペルシア語・ドイツ語・日本語・韓国語・ロシア語・タミル語・タイ語・ベトナム語・中国語・英語・ヒンドゥスタン語の 14 言語を識別対象の言語とした。

各音声データについて Table 1 の条件で音響分析を行ない、Voice Activity Detection (VAD) と Cepstral Mean and Variance Normalization (CMVN), Vocal Tract Length Normalization (VTLN) を施した音響特徴量系列を計算した。UBM は 2,048 混合、対角共分散行列の条件で 24,577 発話から学習した。EV-based, Tensor-based における基底の個数 $K$ はそれぞ

Table 2　14 言語での認識率 [%]

| | WA | | | UA | | |
|---|---|---|---|---|---|---|
| | 3s | 10s | 30s | 3s | 10s | 30s |
| i-vector | **44.40** | **66.36** | **79.19** | **36.90** | **60.95** | **77.00** |
| EV-based (MMSE) | 35.87 | 53.20 | 65.20 | 28.05 | 43.92 | 56.41 |
| EV-based (ML) | 38.83 | 58.80 | 70.11 | 32.10 | 52.05 | 63.42 |
| Tensor-based (MMSE) | 38.23 | 60.56 | 75.67 | 31.42 | 53.56 | 70.64 |
| Tensor-based (ML) | 31.23 | 56.81 | 70.44 | 26.08 | 51.75 | 66.27 |

れ 600, 256 とした。i-vector の次元数は 600 とした。識別には線形サポートベクターマシン (SVM) を用い、23,665 発話で学習、NIST LRE 2007 Evaluation Plan で指定された 6,474 発話（3s, 10s, 30s 各 2,158 発話）でテストを行なった。音響特徴量系列と i-vector の計算には KALDI Toolkit[1]、GMM の学習には Hidden Markov Model Toolkit (HTK)[2]、SVM の実装には LIBLINEAR[3] を用いた。評価として、全テストデータに対する認識率である Weighted Accuracy (WA) と各言語の認識率の平均である Unweighted Accuracy (UA) を計算した。

### 4.2　実験結果

Table 2 に実験結果を示す。EV-based, Tensor-based で重みをそれぞれ式 (2), (15) の最小二乗解として求めた場合を MMSE、重みをそれぞれ式 (4), (17) の最尤基準で求めた場合を ML と表記している。MMSE に関しては WA, UA ともに Tensor-based の認識率が EV-based よりも高くなっており、GMM の各分布の関係性を考慮した Tensor-based の有効性を示せた。一方、MMSE から ML になることで EV-based での認識率は向上したが、Tensor-based での認識率は低下した。そのため ML に関しては、Tensor-based が EV-based の認識率を上回ったのは WA, UA ともに 30s のみとなった。Tensor-based における重み行列は $56 \times 256$、EV-based における重みベクトルは 600 次元となっており、Tensor-based の方が表現力が高いために過学習を生じやすいことが考えられる。

また、3 手法のうち認識率が最高となったのは i-vector であった。i-vector と EV-based は、GMM-SV に対しての PCA という点では共通しているが、i-vector では SVD による決定的な PCA ではなく Probablistic PCA (PPCA) に基づいて確率的なアプローチで言語空間の基底を求めているという違いがある。両者の性能差はこれに起因すると考えられ、Tensor-based にも PPCA を導入することで性能向上が期待できる。

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

## 5　おわりに

本稿では、学習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言語空間を構築し、GMM の各分布の関係性に着目した言語情報表現を用いた言語識別の検討を行なった。同様の検討を話者認識タスクとして行なった [4] と同様に、固有声に基づく言語情報表現と比較して識別性能向上が確認できた。

今後は、テンソル分解に基づく重み行列を推定する際の過学習を防ぐため、最尤推定と事前分布 (UBM) との内挿による MAP 的な重み行列の推定を検討したい。また、PPCA を導入して確率的に言語空間の基底を求めることについても検討したい。

本稿では識別に線形 SVM を用いたが、i-vector の識別でよく用いられている Probabilistic Linear Discriminant Analysis (PLDA) を拡張し、重み行列をそのままの形状で入力できるような行列変量 PLDA を将来的には導入したい。

**謝辞**　本研究の一部は科研費・若手研究 (B) (25730105)、及び基盤研究 (A) (26240022) の助成を受けたものである。

## 参考文献

[1] N. Dehak *et al.*, “Language Recognition via Ivectors and Dimensionality Reduction,” Proc. INTERSPEECH, pp. 857–860, 2011.
[2] W. M. Campbell *et al.*, “Support Vector Machines using GMM Supervectors for Speaker Verification,” IEEE Signal Processing Letters, vol. 13, pp. 308–311, 2006.
[3] N. Dehak *et al.*, “Front-End Factor Analysis for Speaker Verication,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788–798, 2011.
[4] チン・トゥアン・トゥー他, “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,” 日本音響学会春季講演論文集, pp. 217–220, 2015.
[5] R. Kuhn *et al.*, “Rapid Speaker Adaptation in Eigenvoice Space,” IEEE Transactions on Speech and Audio Processing, vol. 8, no. 6, pp. 695–707, 2000.
[6] T. Toda *et al.*, “Eigenvoice Conversion Based on Gaussian Mixture Model,” Proc. INTERSPEECH, pp. 2446–2449, 2006.
[7] D. Saito *et al.*, “One-to-Many Voice Conversion Based on Tensor Representation of Speaker Space,” Proc. INTERSPEECH, pp. 653–656, 2011.
[8] L. R. Tucker, “Some mathematical notes on three-mode factor analysis,” Psychometria, vol. 31, No. 3, pp. 279–311, 1966.