構造的表象とGMMスーパーベクトル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、広瀬 啓吉 (東大)

1 はじめに

スマートフォン等の音声翻訳アプリケーションや電話受付センターで用いられている技術の1つとして、音声からの言語識別がある。入力音声から言語的特徴を抽出することで言語の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。そのため、非言語的特徴の変動に頑健な言語識別システムの構築が望まれる。本稿では、言語識別において従来用いられてきた混合ガウス分布 (Gaussian Mixture Model, GMM) を基に抽出したGMMスーパーベクトルに加え、GMMを構成するガウス分布の相対関係を捉えた「GMM構造ベクトル」を特徴量とする手法を提案する。GMM構造ベクトルは、非言語的変動に頑健とされる構造的表象の考え方を用いた特徴量であり、GMMスーパーベクトル単体よりも頑健となることが期待される。本稿ではこのGMM構造ベクトルを用いた言語識別実験を行ない、識別性能向上を確認した。

2 先行研究

2.1 音声からの言語識別

言語識別に関する近年の研究では、学習音声から何らかの特徴量を抽出し、これらを用いてSupport vector machine (SVM) 等の識別器を学習させる手法が多く用いられている。[1] では、GMM Universal Background Model (GMM-UBM) に基づいて発話依存GMMを推定し、これから得られるGMMスーパーベクトルを用いている。処理の流れを示す。

1. 様々な言語･話者の音声データから、音響空間の全体的な傾向を表す（言語･話者非依存の）GMM-UBMを推定する。
2. 学習音声の各発声毎に、GMM-UBMをMAP適応し、発話依存GMMを推定する。これを各学習音声、全てについて行なう。
3. 発話依存GMMから、それぞれスーパーベクトルを抽出する。
4. 各発話を表すスーパーベクトルを学習サンプルとし、また各学習サンプルに付属する言語IDをラベルとしてSVMを学習する。

本節では以下、このSV+SVM、及び、「音声の構造的表象」について説明する。

2.2 GMM-UBM

GMM-UBMは、話者・言語非依存のモデル、すなわち人間の声そのものを表すGMMである。大量の話者・言語のデータを用いることで、話者及び言語属性が隠れ変数として扱われ、統計的には話者・言語非依存モデルとしてGMMが推定される。

2.3 MAP適応

MAP適応は、モデルパラメータ推定法の一つである事後確率最大化法（MAP推定）を用いてパラメータを更新する手法である。MAP推定は、入力データ X が得られたとき、パラメータ θ を確率変数として扱い、その事後確率を最大化するように推定する。

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|X) \tag{1}$$

式 (1) をベイズの定理を用いて変形すると

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(X|\theta)p(\theta) \tag{2}$$

となり、尤度 $p(X|\theta)$ と事前確率 $p(\theta)$ の積の形で表せる。最終的にMAP適応による更新後のパラメータは、X に関する最尤推定値と、更新前の事前パラメータとの内挿点に相当する値となる。入力データ X が十分に大きければ、適応後のパラメータは最尤推定値に近づくことになる。

以下に、MAP適応によりGMMの平均ベクトルを更新する計算法を示す。データ $\boldsymbol{x}_i$ が観測されたとき、これがインデックス k のガウス分布 $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k)$ から生成されたとする確率 $(= p(k|\boldsymbol{x}_i))$ を $\gamma_{k,i}$ と置くと、

$$\gamma_{k,i} = \frac{w_k \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k'=1}^{K} w_{k'} \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_{k'}, \Sigma_{k'})} \tag{3}$$

となる。データ $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N$ について、k 番目のガウス分布 $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k)$ から生成されるデータに関する確率的サンプル数 N_k、確率的平均ベクトル $\boldsymbol{e}_k$ は

$$N_k = \sum_{i=1}^{N} \gamma_{k,i} \tag{4}$$

$$\boldsymbol{e}_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{k,i} \boldsymbol{x}_i \tag{5}$$

---

*A study of language identification using structural representation of speech and GMM-based supervector. by S. Suzuki, D. Saito, N. Minematsu, and K. Hirose (The University of Tokyo)

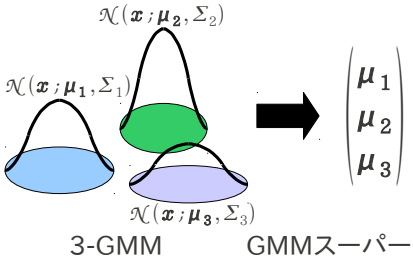

Fig. 1　GMM スーパーベクトル (3 混合)

と表される。このとき、$\boldsymbol{\mu}_k$ を

$$\boldsymbol{\mu}_k^{(X)} = \frac{\tau}{N_k+\tau}\boldsymbol{\mu}_k + \frac{N_k}{N_k+\tau}\boldsymbol{e}_k \qquad (6)$$

で更新する。理論的には τ は、UBM-GMM 推定時の学習サンプルのうち、状態 k から生成される確率的サンプル数であるが、実際には適切な固定値を与えることが多い。式 (6) から $\boldsymbol{\mu}_k^{(X)}$ は、最尤推定パラメータと事前パラメータの内挿点であることが分る。

本稿では、GMM 適応のための入力音声は 3s、10s、30s の場合を検討する。N_k の値はこの入力音声長に比例して変化するため、τ の値も入力音声長に比例して変化させた。その結果、式 (6) における $\boldsymbol{\mu}_k$ と $\boldsymbol{e}_k$ の係数は、どの場合でも凡そ一定となる。

分散についても、確率的な二次十分統計量を用いて同様に適応を行なっている。

2.4　GMM スーパーベクトル

GMM スーパーベクトルは、GMM を構成するガウス分布の平均ベクトル $\boldsymbol{\mu}_k$ を全て連結して 1 つの高次元ベクトルにしたものである [2]。Fig. 1 に 3 混合 GMM の場合を示す。GMM-UBM を適応して構成された GMM は、各ガウス分布に対するインデックス k の割り当てが統一されるため、GMM スーパーベクトル間の比較が意味を持つものとなる。

本稿では、GMM スーパーベクトルをそのまま用いずに、UBM-GMM におけるスーパーベクトルを適応 GMM のスーパーベクトルが持つバイアス項として考え、後者から前者を差し引いた差ベクトルを、スーパーベクトルとして用いる。

2.5　SVM

SVM は 2 クラスの識別を行なう手法である。識別とは、入力特徴ベクトル $\boldsymbol{x}$ に対して、式

$$y = \mathrm{sgn}[f(\boldsymbol{x})] = \begin{cases} 1 & (f(\boldsymbol{x}) \geq 0 \text{ のとき}) \\ -1 & (f(\boldsymbol{x}) < 0 \text{ のとき}) \end{cases} \qquad (7)$$

で表されるように識別関数 $f(\boldsymbol{x})$ の正負によってクラス y を推定することであるが、ここで、$\boldsymbol{x}$ から特徴抽出によって得られた特徴ベクトル $\boldsymbol{\phi}(\boldsymbol{x})$ に対して線

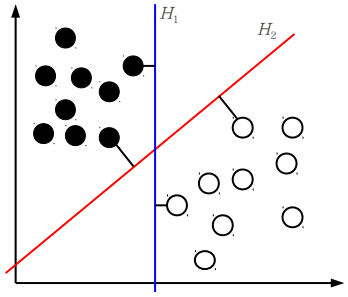

Fig. 2　2 次元データの 2 クラスの分離

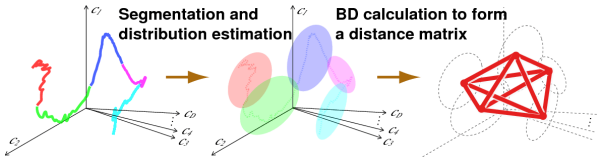

Fig. 3　入力音声の不変構造化

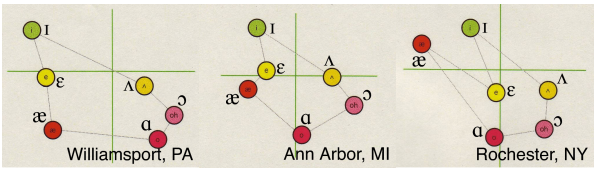

Fig. 4　米語方言における F1/F2 図 [7]

形な識別関数 $f(\boldsymbol{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x})$ を考えると、これは $f(\boldsymbol{x}) = 0$ という超平面で特徴ベクトル空間を 2 つに分割し、一方に 1、もう一方に -1 を割り当てることに対応する。SVM は、このように 2 つのクラスを分離する超平面のうち、最も近くにあるクラス 1 のデータからの距離と最も近くにあるクラス 2 のデータからの距離が等しくなるような、マージン最大の超平面を求める。そのため、クラスの境界付近のデータのみから計算されることになる。Fig. 2 は 2 次元の場合を表している。この場合、直線 H_1、直線 H_2 はいずれも 2 つのクラスを分離できているが、SVM によって求まるマージン最大の境界は直線 H_2 となる。

本稿では SVM を用いて多クラス分類を行なうが、これは「1 つのクラスと残り全てのクラス」の 2 値分類を各クラスについて行ない、識別面から最も遠いクラスを推定結果とする one-versus-rest 法を用いている。

2.6　音声の構造的表象

音声の構造的表象は、音声から非言語的特徴を除去して音声を表象する手法の一つである。非言語的特徴による音響特徴の変動は、特徴量空間の写像として扱えるが、写像不変量のみで音声を表象すれば、それは非言語的特徴に不変な音声特徴となる。

非言語的な空間写像を連続かつ可逆な写像とする。

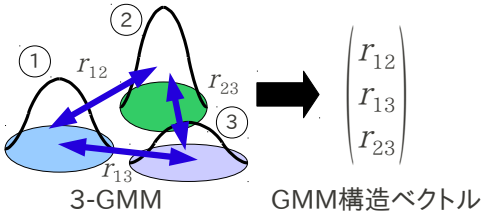

Fig. 5　GMM 構造ベクトル (3 混合)

二分布間の距離尺度の一つであるバタチャリヤ距離（BD 距離）は、任意の連続かつ可逆な写像に対して不変である [3]。写像前後の分布がガウス分布に従うならば、BD 距離は任意の線形変換 $c' = Ac + b$ に対して不変となる [4]。ケプストラムや MFCC を音響特徴量とした場合、声道長の違いは行列の乗算、収録機器の違いはベクトルの加算で近似できるので、BD 距離はこれらの違いに不変な特徴となる。一つの発声（特徴量系列）を分布系列として捉えれば、任意の二分布間 BD 距離からなる距離行列は不変特徴となる（Fig. 3 参照）。これを音声の構造的表象と呼ぶ。

構造的表象は既に音声認識や方言分類でその効果が確認されている [5, 6]。例えば米語における方言は、母音群の幾何学的配置の差として捉えることができる（Fig. 4 参照）。構造的表象は、これら方言間の差異を、性別などに非依存に表象できる [6]。但し [6] で検討された方言分類は、発話内容（読み上げ内容）を固定し、方言差異がより適切に観測され易い環境下での検討である。本稿では発話内容が可変であり、この点が先行研究と大きく異なる。

3　提案手法

GMM スーパーベクトルは元来話者識別のタスクで提案された特徴量である [2]。話者識別、言語識別という独立した識別対象に対して類似した枠組みが使えるのは、SVM 学習における識別超平面推定に負うところが大きい。当然、言語性や方言性を適切に表現でき、かつ GMM スーパーベクトルでは捉えられていない特徴を導入することで、言語識別性能の向上が期待できる。本研究では構造表象を応用し、話者性や雑音等に対して頑健な特徴量として GMM 構造ベクトルを追加する手法を提案する。GMM 構造ベクトルとは、GMM を構成する M 個のガウス分布に対して ${}_M C_2$ 個の BD 距離を求め、これを列ベクトルとして表現したものである。Fig. 5 に 3 混合 GMM の場合の GMM 構造ベクトルを示す。これを GMM スーパーベクトルに追加して言語識別を行なう。GMM スーパーベクトルは M 個の平均ベクトルのみに依存するが、GMM 構造ベクトルは分散項にも依存する。

なお UBM-GMM からも構造ベクトルを抽出することができる。GMM スーパーベクトルと同様に、UBM-GMM 構造ベクトルはバイアス項として考え、適応 GMM から得られる構造ベクトルからこれを引いた差ベクトルを、本稿では構造ベクトルとして使う。

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 (12 次元)
	+ ΔMFCC (12 次元)
	+ Δpower (1 次元)

Table 2　UBM の作成条件

音響的特徴	MFCC (12 次元)
	+ ΔMFCC (12 次元)
	+ Δpower (1 次元)
パラメータ推定	ML 推定
GMM	1024 混合ガウス分布（対角共分散行列）
	128 混合ガウス分布（対角共分散行列）

4　実験

4.1　実験方法

The National Institute of Standards and Technology Language Recognition Evaluation (NIST LRE) の 2003・2005・2007 年のデータベースを用いて言語識別実験を行なった。このデータベースには、電話での会話が継続長 3s・10s・30s という 3 つのカテゴリーに分かれて収録されている。収録言語はアラビア語・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・ヒンドゥー語の 14 言語である。

各データに対して、各フレームのパワー項を用いた自動無音除去を施し、ケプストラム系列に変換した。音響分析条件は Table 1 である。UBM は NIST LRE 2005 の 11,106 発話 (3s・10s・30s 合計) を用いて学習した。その作成条件を Table 2 に示す。以下、GMM スーパーベクトルを「GMM-SV」、GMM 構造ベクトルを「GMM-STR」と略記する。各種特徴量を用いた SVM の学習には、NIST LRE 2003 の 11,118 発話 (3s・10s・30s 合計) と NIST LRE 2007 の追加学習データの 710 発話を用い、テストデータとしては NIST LRE 2007 の 6,474 発話 (3s・10s・30s 合計) を用いた。学習データもテストデータも上に挙げた 14 言語以外の言語を含まない closed-set である。同様に、学習データ・テストデータを中国語・英語・スペイン語の 3 言語のみに絞った実験も行なった。

GMM の作成には Hidden Markov Model Toolkit

Table 3　14 言語での認識率 [%]

	3s	10s	30s
128 GMM-SV (ベースライン)	27.34	40.92	52.73
128 GMM-STR	23.77	35.96	47.45
128 GMM-SV + 128 GMM-STR (提案手法)	27.71	41.71	53.20
1024 GMM-SV (ベースライン)	**31.00**	44.76	57.88
1024 GMM-SV + 128 GMM-STR (提案手法)	30.86	**46.99**	**58.25**

Table 4　3 言語での認識率 [%]

	3s	10s	30s
128 GMM-SV (ベースライン)	**60.14**	71.18	79.27
128 GMM-STR	51.14	62.07	71.98
128 GMM-SV + 128 GMM-STR (提案手法)	55.69	68.11	75.17
1024 GMM-SV (ベースライン)	51.82	60.02	63.44
1024 GMM-SV + 128 GMM-STR (提案手法)	57.97	**72.55**	**80.41**

(HTK)[1]、SVM の実装には LIBLINEAR [2]を用いた。

4.2　実験結果

14 言語での結果をデータの継続長毎に Table 3 に、3 言語での結果も同様、Table 4 に示す。「GMM-SV」・「GMM-STR」の前の数字は GMM の混合数である。なお 1024 混合の GMM-STR は次元数が高くなりすぎるため SVM 学習できなかった。

128 混合について見ると、GMM-SV 単体の性能と比較して、GMM-SV+GMM-STR（提案手法）の認識率は、14 言語の場合では 3s 以外は上回っているが、3 言語の場合では全体的に下回っている。3 言語の場合、出現する単音の種類に対して混合数が多かったために過学習を起こしたことが考えられる。一方、1024 混合では、14 言語・3 言語どちらの場合でも提案手法がベースラインを全体的に上回っている。

また提案手法同士を比較すると、14 言語・3 言語いずれの場合も「1024 GMM-SV + 128 GMM-STR」が「128 GMM-SV + 128 GMM-STR」を上回る認識率となっている。このことより，GMM-STR における GMM 混合数を上げることができれば、更に性能を向上させることができると考えられる。

5　まとめと今後の課題

言語識別において従来用いられてきた特徴量である GMM スーパーベクトルに、構造的表象の考え方を用いた特徴量である GMM 構造ベクトルを付け加えることによって、一定の精度向上を確認できた。

構造ベクトルの場合、GMM 混合数の二乗に比例してベクトル次元数が増加する。今回は 128 混合での検討となったが、PCA などにより次元圧縮するなどしてより高次元の構造ベクトルを検討したい。

また、構造的表象は音声事象群におけるコントラスト特徴（エッジ特徴 [8]）のみを抽出したものであるが、これに対して「Bag of features」の考え方を導入し、抽出されたエッジ群に対するヒストグラムを特徴量とする手法も検討したい。

ベースラインシステムについては、本稿では同じく NIST LRE コーパスを用いた既存システム [1] に性能が及ばなかった。音響分析の前処理や識別器のチューニング等が不十分だと思われるため、これらを改善した上で再評価実験を行ないたい。

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

参考文献

[1] C. H. You *et al.*, “A GMM-supervector approach to language recognition with adaptive relevance factor,” EUSIPCO-2010, pp. 1993–1997, 2010.

[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] 峯松信明他, “構造不変の定理とそれに基づく音声ゲシュタルトの導出,” 電子情報通信学会音声研究会, SP2005, pp. 1–8, 2005.

[4] 峯松信明他, “線形・非線形変換不変の構造的情報表象とそれに基づく音声の音響モデリングに関する理論的考察,” 日本音響学会春季講演論文集, 1-P-12, pp. 147–148, 2007.

[5] 村上隆夫他, “音声の構造的表象に基づく日本語孤立母音系列を対象とした音声認識,” 電子情報通信学会論文誌, J91-A, 2, pp. 181–191, 2008.

[6] X. Ma *et al.*, “Structural analysis of dialects, sub-dialects, and sub-sub-dialects of Chinese,” Proc. INTERSPEECH, pp. 2219-2222, 2009.

[7] W. Labov *et al.*, *Atlas of North American English*, Mouton and Gruyter, 2005.

[8] 齋藤大輔他, “話者不変な相対関係特徴を音響単位とする音響モデリングに関する実験的検討,” 電子情報通信学会音声研究会, SP2009, pp. 7–12, 2009.