トピック遷移 PLSA に基づくメルスペクトログラム生成モデルを用いた多言語音声分類手法の評価*

◎大石康智, 亀岡弘和 (NTT), 小野順貴 (NII),
石本祐一 (国語研), 松井知子 (統数研), 板橋秀一 (産総研)

1　はじめに

本研究の目的は，与えられた音声信号のみから事前知識無しに言語間の類似度を推定し，多言語を分類する手法を確立することである。多言語音声の分類は，直接的には言語識別技術の基盤となり，多言語音声認識の前処理としての応用が期待される。一方，言語学的観点からは，文字を持たない多数の言語に対して，音素に近い要素の抽出が可能となり，それらの言語の記述および言語系統の解明を期待できる。

言語の分類・識別はこれまで，言語学的観点および工学的応用の両面から研究が進められている。特に工学的な分野では，大規模な多言語音声コーパスを利用した言語の自動識別が試みられている [1, 2]。音響特徴量として MFCC や I-vector，音素認識結果に基づく N-gram などを用い，識別器としてガウス混合モデル (GMM) や隠れマルコフモデル (HMM) などが用いられるが，充分な成果は得られていない [3–6]。

我々は非負値行列因子分解 (NMF)[7, 8] を基礎とし，その基底の時間遷移の確率モデルを導入したトピック遷移 PLSA を提案した [9]。PLSA[10] は文書を対象とする自然言語処理の一手法であり，トピック（話題）に相当する隠れ変数を介して，各文書中に現れる単語の度数データを扱う確率モデルである。数学的には，I-divergence 型 NMF[7] と等価であるが，NMF の定式化では導入が困難であった基底の時間遷移のモデリングが，PLSA の場合には隠れ変数の遷移確率として自然に導入できる。提案法を音声のメルスペクトログラムに適用し，先験知識なしに，各言語の音響的特徴と言語的特徴を抽出して言語識別に応用した結果，通常の PLSA と比較して，基底の時間遷移を表現する提案法の有効性を確認した。本稿では，より大規模な多言語音声コーパスを利用し，提案法の言語識別への応用可能性を検討する。また，提案法の統計モデルとしての性質について議論する。

2　トピック遷移 PLSA

メルスペクトログラムを $\boldsymbol{Y} = (y_{\omega,t})_{\Omega\times T} \in \mathbb{R}^{\geq 0, \Omega\times T}$ と表現する。ω はメルフィルタのインデックス，t はフレームのインデックスを表す。ここで，K 個の基底スペクトル $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K]$ と遷移確率行列 $\boldsymbol{A} = (A_{j,k})_{K\times K}$（$A_{j,k}$ は基底 j から基底 k への遷移確率

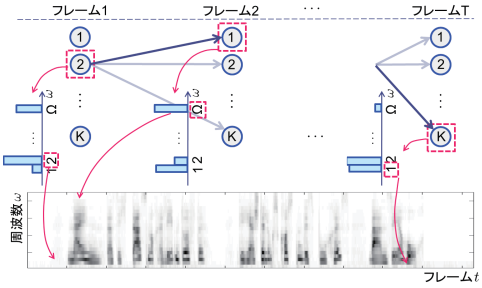

Fig. 1　メルスペクトログラムの生成モデル

を表す）を用意する。また k 番目の基底スペクトルを $\boldsymbol{h}_k = [h_{1,k}, \ldots, h_{\Omega,k}]^{\mathrm{T}}$ と表す。このベクトルの要素は「周波数の出やすさを表す確率」と解釈される。フレーム t の基底 k_t は，一つ前のフレームの基底 k_{t-1} に依存して選択され，その基底をパラメータとする多項分布から生成されたものが，フレーム t のメルスペクトルと考える (Fig. 1)。

言語の音響的および言語的特徴が，基底スペクトルと遷移確率として表現される提案法は基本的には出力分布が多項分布である HMM と解釈できるが，この多項分布仮定がパラメータ学習にどのような影響を与えるか議論する。文献 [9] より，基底の更新式は，

$$h_{\omega,k} = \frac{\sum_{t=1}^{T}\gamma(z_{t,k})y_{\omega,t}}{\sum_{t=1}^{T}\gamma(z_{t,k})\sum_{\omega=1}^{\Omega}y_{\omega,t}} = \frac{\sum_{t=1}^{T}\gamma(z_{t,k})\sum_{\omega=1}^{\Omega}y_{\omega,t}\cdot\frac{y_{\omega,t}}{\sum_{\omega=1}^{\Omega}y_{\omega,t}}}{\sum_{t=1}^{T}\gamma(z_{t,k})\sum_{\omega=1}^{\Omega}y_{\omega,t}} \quad (1)$$

と書ける。$\gamma(z_{t,k})$ は Forward-Backward アルゴリズムによって得られる値であり，フレーム t の観測スペクトルを基底 k に分配する割合を表す。式 (1) より，観測スペクトル系列の中でスケールが大きいもの（大きな声で発声されているスペクトル）ほど基底の更新に大きく寄与する。これは言語に支配的な音声の基底を得たいとする本研究の動機にまさに合致した性質と言える。一方，出力分布をガウス分布と仮定すると，その平均の更新式は

$$\mu_{\omega,k} = \frac{\sum_{t=1}^{T}\gamma(z_{t,k})\cdot y_{\omega,t}}{\sum_{t=1}^{T}\gamma(z_{t,k})} \quad (2)$$

と書ける。これは声の大きさに関係なく，観測スペクトル系列を均等な重み付けによって，平均が推定されることを意味する。このような統計モデルとしての性質の違いが，言語識別にどのように影響を与えるか，次節の評価実験にて検証する。

*Evaluation of Language Classification using Generative Model of Mel-scale Spectrogram based on Markovian PLSA. by OHISHI, Yasunori, KAMEOKA, Hirokazu (NTT), ONO, Nobutaka (NII), ISHIMOTO, Yuichi (NINJAL), MATSUI, Tomoko (ISM), ITAHASHI, Shuichi (AIST)

3　評価実験

GLOBALPHONE コーパス [11] における，アラビア語 (AR)，ブルガリア語 (BL)，クロアチア語 (CR)，チェコ語 (CZ)，フランス語 (FR)，ドイツ語 (GE)，日本語 (JA)，韓国語 (KO)，中国語 (MA)，ポルトガル語 (PO)，ポーランド語 (PL)，ロシア語 (RU)，スペイン語 (SP)，スウェーデン語 (SW)，タイ語 (TH)，トルコ語 (TU)，ベトナム語 (VN) の計 17 言語の音声データを利用して，言語識別の観点から提案法の有効性を評価する。このコーパスには，母国語話者によって読み上げられた新聞記事の音声がサンプリング周波数 16 kHz，量子化数 16 ビットで収録されている。

まず，発話毎に音声信号はその振幅の絶対値の平均値によって除算され，音量が正規化される。そして，フレームシフト長 32 ms，フレーム長 16 ms，ハニング窓を用いてフレームに分割され，短時間フーリエ変換によってパワースペクトログラムに変換される。最後に，各フレームのパワースペクトルをメルフィルタバンク処理し，その出力値 $\{w_{1,t}, w_{2,t}, \ldots, w_{\Omega,t}\}$ を下記のように β 乗したものをメルスペクトルとする。

$$\boldsymbol{y}_t = [y_{1,t}, \ldots, y_{\Omega,t}]^{\mathrm{T}} = [w_{1,t}^{\beta}, \ldots, w_{\Omega,t}^{\beta}]^{\mathrm{T}} \quad (3)$$

提案法は非負値制約のため，スペクトル包絡構造を強調するためにこのような処理を行った。文献 [9] の結果を踏まえて，$\Omega = 22, \beta = 0.5$ とした。また，多項分布は離散型確率分布であるため，度数の最小単位を 0.1 とし，メルスペクトログラムの各要素を数え上げて整数値で表現した。遷移確率の初期値は乱数によって与えられ，基底スペクトルの初期値はあらかじめメルスペクトログラムに NMF を適用して得られた結果を利用した。学習則の反復回数は 100 回とした。

GLOBALPHONE コーパスは様々な録音環境下で多数の話者による音声が収録されているため，言語毎にランダムに選択された 2 時間分の音声を基底スペクトルと遷移確率行列の学習に利用する。評価データとして，学習データとオープンになるように話者を選択し，30 秒の音声を言語毎に 100 サンプル用意した。サンプルごとに計算される，言語に対する事後確率に基づいて，偽陽性率と偽陰性率の相加平均 C_{avg}[1] を評価尺度として利用する。前節で議論した，HMM の出力分布をガウス分布とした手法 (Gaussian HMM と呼ぶ) と性能を比較する。

識別性能を Fig. 2 に示す。基底数（HMM の状態数）を $K = 80$ に増やすことによって，全体的に性能は向上するものの，Gaussian HMM に比べて性能は低下した。性能で Gaussian HMM を上回ることはなかったものの，提案モデルの学習則が正しく動いていることが改めて確認された。提案法において，学習データに対するモデルの尤度が Gaussian HMM に比べて 10 倍程度小さかったため，まずは基底数を増やして調整することが今後の課題として挙げられる。

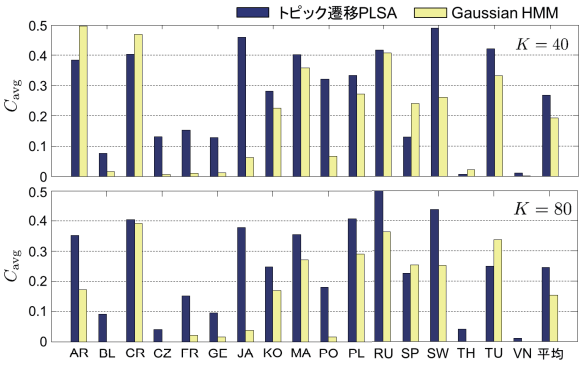

Fig. 2　各言語の識別性能の比較: トピック遷移 PLSA と Gaussian HMM を比較する。C_{avg} が小さい値ほど性能が高いことを示す。

4　おわりに

GLOBALPHONE コーパスを利用して，トピック遷移 PLSA モデルの言語識別への応用可能性を評価した。提案モデルでは，言語が持つ音響的な性質と音素遷移を含む言語的な性質がそれぞれ，基底と遷移確率によって別々に学習される。性能で Gaussian HMM を上回ることはなかったものの，提案モデルの学習則の動作を改めて確認した。

今後の課題は，NIST(National Institute of Standards and Technology) の言語識別評価で使われるデータベースを利用して最先端の手法と比較すること，基底数を調整することである。文献 [13] のように，多項分布によるスペクトルの生成過程を考える際のスケール（度数の最小単位）を検討することも必要である。また，NMF を含め，基底の状態遷移をモデル化する研究が精力的に取り組まれているため [12]，モデルの構成を見直すことや新たな応用先を探すことも課題として挙げられる。

参考文献

[1] Greenberg *et. al.*, in *Proc. Interspeech 2012*.
[2] Rodriguez-Fuentes *et. al.*, in *Proc. Interspeech 2012*.
[3] Yeshwant *et. al.*, IEEE Signal Processing Magazine, pp. 33–41, 1994.
[4] Li *et. al.*, in *Proc. IEEE 2013*.
[5] Huang *et. al.*, in *Proc. ICASSP 2013*.
[6] Lawson *et. al.*, in *Proc. Interspeech 2013*.
[7] Virtanen, IEEE TASLP, Vol. 15, pp. 1066–1074, 2007.
[8] 石井他, 音講論 (秋), pp. 245-248, 2013.
[9] 大石他, 音講論 (春), pp. 445–448, 2013.
[10] T. Hofmann, in *Proc. SIGIR 1999*.
[11] Schultz, in *Proc. ICSLP 2002*.
[12] ルルー他, 音講論 (春), pp. 807–808, 2013.
[13] Hoffman, in *Proc. ICASSP 2012*.