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の検討*

◎大石康智, 亀岡弘和 (NTT), 小野順貴, 石本祐一 (NII),
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，与えられた時系列音響信号のみから事前知識無しに言語間の類似度を推定し，多言語の分類・類型化を行う手法を確立することである。時系列音響信号の分析に基づいた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。さらに多言語音声翻訳や字幕翻訳などを含むマルチモーダル処理へ展開することで，音声認識・合成といった音声工学分野にとどまらず，言語工学や認知科学などの様々な分野との融合へと発展しうる。一方，言語学的観点からは，文字言語を持たない多数の言語に対して，本手法により，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が期待できる。これによりサイエンスとしての音声言語学の発展とデータ駆動型手法による新たな言語科学の創出に寄与すると考えられる。

これまで，言語の分類・識別は言語学的観点および工学的応用の両面から様々なアプローチがとられてきた。言語学の分野では機能的・地理的比較の手法を用いて，文法，語彙，歴史的・地理的背景などに基づいた言語の分類・類型化が進められてきた [1, 2]。しかしこの手法は研究者個人の観察・内省に基づくものが多く客観性が高いとは言いがたい。一方，工学的な分野では複数言語の大規模音声コーパスを利用して，言語の自動分類が試みられている。これまで，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)，隠れマルコフモデル (HMM)，サポートベクターマシン (SVM)，ニューラルネットワーク (NN) など様々な手法が用いられているが，これらも未だ充分な成果は得られていない [3, 4, 5]。

こうした背景より，我々は音響信号処理分野で発展する非負値行列因子分解 (Nonnegative matrix factorization, NMF)[6, 7] を用いたスパース基底の学習とそれらの時間的な遷移構造をデータのみから学習することにより，大量の多言語コーパスデータから言語の先験知識なしに，各言語の音響的・言語的特徴を抽出する。具体的には，NMF がデータから教師なしで，スパースな基底を学習する能力が高いことに着目し，多言語音声データのメルスペクトログラム（非負データ）から抽出される各言語の特徴的な基底メルスペクトルを，多言語音声分類に応用する。また，近年の関連研究より，言語分類においては音響的な基底スペクトルのみならず，それらの時間遷移の性質を捉えることが重要であることがわかってきた [8]。そこで，NMF のような非負データを対象とする教師なし学習法に，基底の時間的遷移の確率モデルを組み込んだトピック遷移 PLSA(Markovian probabilistic latent semantic analysis) と呼ぶ生成モデルを提案する。そして，基底メルスペクトルと状態遷移確率を学習するためのパラメータ推定アルゴリズムを導出し，その基本動作を検証するための多言語音声識別実験結果を報告する。通常の PLSA と比較して，基底の時間的遷移を表現する提案モデルの有効性を確認した。

2 メルスペクトルのスパース分解表現

音声は，音素という有限個の離散的な要素からなり，その時系列に言語的な情報を載せる情報伝達媒体である。実際に生成されるスペクトルは，音声生成メカニズムにおける力学的な制約があるため，各音素に対応するスペクトル間を階段状に遷移するわけではなく，各音素に対応するスペクトル付近をかすめていくように連続的に時間変化する。すなわち，音声スペクトル時系列は各音素に対応するスペクトルの混ざり具合が連続的に時間変化したものとしてみなすことができる。そこで本研究では，基底スペクトルの和のモデルである NMF を，音韻情報を特徴付けるメルスペクトログラムに適用する。特に NMF の非負値制約により，基底アクティビティがスパースになることから，より言語性を捉えた，言語に支配的な音響的特徴が基底として抽出されることを期待する。

具体的には，NMF に基底の時間的遷移の確率モデルを組み込んだトピック遷移 PLSA を提案する。PLSA[9] は元々テキスト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）に相当する潜在変数を介して，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。PLSA は数学的には，KL 距離に基づく NMF 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。次節では提案モデルの詳細を説明する。

---

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

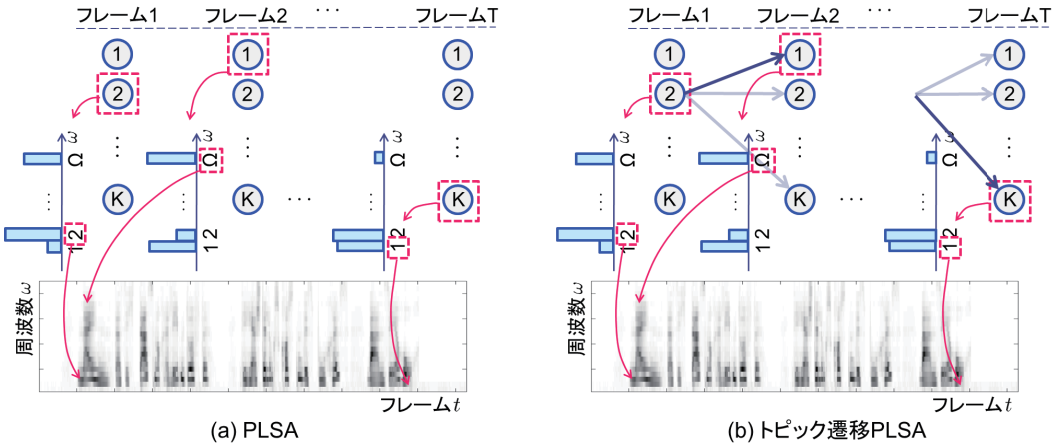

Fig. 1　PLSA およびトピック遷移 PLSA によるメルスペクトログラムの生成モデル

2.1　トピック遷移 PLSA

Fig. 1 (a) は PLSA に基づくメルスペクトログラムの生成過程を表す。まず，言語 l のメルスペクトログラムを $\boldsymbol{Y}^{(l)} = (y^{(l)}_{\omega,t})_{\Omega\times T} \in \mathbb{R}^{\geq 0,\Omega\times T}$ と表現する。ここで，$\omega = 1,\ldots,\Omega$ はメルフィルタのインデックス，$t = 1,\ldots,T$ は分析フレームのインデックスを表す。このスペクトログラムを生成するために，K 個の基底メルスペクトル $\boldsymbol{H}^{(l)} = [\boldsymbol{h}^{(l)}_1,\ldots,\boldsymbol{h}^{(l)}_K]$ を用意する。k 番目の基底メルスペクトルを $\boldsymbol{h}^{(l)}_k = [h^{(l)}_{k,1},\ldots,h^{(l)}_{k,\Omega}]^{\mathrm{T}}$ と表現する。これらの基底は言語 l の音声を作り出すためのアンカーポイント（音素）であり，基底メルスペクトルの各要素は「周波数の出やすさを表す確率」とみなす。フレーム t において，これらの基底メルスペクトルのいずれかが選ばれ，その基底メルスペクトルをパラメータとする多項分布から生成されたものが，時刻 t のメルスペクトルと考える。すなわち，

$$k^{(l)}_t \sim \mathrm{Discrete}(\pi^{(l)}_{1,t},\ldots,\pi^{(l)}_{K,t}) \tag{1}$$

$$\boldsymbol{y}^{(l)}_t | k^{(l)}_t \sim \mathrm{Multinomial}(\boldsymbol{h}^{(l)}_{k^{(l)}_t}) \tag{2}$$

のように書ける。ここで，$\pi^{(l)}_{1,t},\ldots,\pi^{(l)}_{K,t}$ はフレーム t における各基底の生起確率，$\boldsymbol{y}^{(l)}_t = [y^{(l)}_{1,t},\ldots,y^{(l)}_{\Omega,t}]^{\mathrm{T}}$ はフレーム t のメルスペクトルを表す。PLSA はテキストを対象とした自然言語処理の一手法であり，トピックに相当する潜在変数を介して，各文書中に現れる単語の度数データを扱う。PLSA によるメルスペクトログラムの生成モデルでは，基底のインデックス k がトピックに相当する。また，メルスペクトログラムの値 $y_{\omega,t}$ はフレーム（文書）t における周波数（単語）ω の度数と解釈する。本稿では，この基底メルスペクトルが，一つ前の時刻の基底に依存して遷移するモデルへと拡張する（Fig. 1 (b)）。すなわち，

$$k^{(l)}_t | k^{(l)}_{t-1} \sim \mathrm{Discrete}(A^{(l)}_{k^{(l)}_{t-1},1},\ldots,A^{(l)}_{k^{(l)}_{t-1},K}) \tag{3}$$

$$\boldsymbol{y}^{(l)}_t | k^{(l)}_t \sim \mathrm{Multinomial}(\boldsymbol{h}^{(l)}_{k^{(l)}_t}) \tag{4}$$

と書ける。$\boldsymbol{\theta}^{(l)} = \{\boldsymbol{h}^{(l)}_1,\ldots,\boldsymbol{h}^{(l)}_K,\boldsymbol{\pi}^{(l)},\boldsymbol{A}^{(l)}\}$ はパラメータの集合であり，$\boldsymbol{\pi}^{(l)} = \{\pi^{(l)}_1,\ldots,\pi^{(l)}_K\}$ は初期状態確率，$\boldsymbol{A}^{(l)} = (A^{(l)}_{j,k})_{K\times K}$ は状態遷移確率行列を表す（$A^{(l)}_{j,k}$ は基底 j から基底 k への遷移確率を表す）。このように音響的性質と言語的性質がそれぞれ，基底メルスペクトルと状態遷移確率として表現される。

次節では，これらのパラメータの推定アルゴリズムを導出する。便宜上，フレーム t における基底のインデックス k_t の代わりに $\boldsymbol{z}^{(l)}_t = [z^{(l)}_{1,t},\ldots,z^{(l)}_{K,t}]^{\mathrm{T}}$ を表記方法として導入する。これは K 次元の 2 値確率変数であり，どれか 1 つの $z^{(l)}_{k,t}$ だけが 1 で他は 0 とする。

2.2　パラメータ推定アルゴリズム

パラメータ集合 $\boldsymbol{\theta}^{(l)}$ は EM アルゴリズムを利用して効率的に推定できる。ここで Q 関数は

$$\begin{aligned}
&Q(\boldsymbol{\theta}^{(l)},\boldsymbol{\theta}^{(l)\mathrm{old}}) \\
&= \sum_{k=1}^{K}\gamma(z^{(l)}_{1,k})\log\pi^{(l)}_k + \sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K}\xi(z^{(l)}_{t-1,j},z^{(l)}_{t,k})\log A^{(l)}_{j,k} \\
&+ \sum_{t=1}^{T}\sum_{k=1}^{K}\gamma(z^{(l)}_{t,k})\Bigg(\log\Gamma\Big(\sum_{\omega=1}^{\Omega}y^{(l)}_{\omega,t}+1\Big) \\
&\quad - \sum_{\omega=1}^{\Omega}\log\Gamma(y^{(l)}_{\omega,t}+1) + \sum_{\omega=1}^{\Omega}y^{(l)}_{\omega,t}\log h^{(l)}_{k,\omega}\Bigg)
\end{aligned} \tag{5}$$

と書け，$\gamma(\boldsymbol{z}_t)$ は潜在変数 $\boldsymbol{z}_t$ の事後分布，$\xi(\boldsymbol{z}_{t-1},\boldsymbol{z}_t)$ は 2 つの連続した潜在変数の同時事後分布を表す。

E ステップでは，フォワード・バックワードアルゴリズム [10] を利用して，下記の変数の値を求める。

$$\gamma(\boldsymbol{z}^{(l)}_t) = \frac{\alpha(\boldsymbol{z}^{(l)}_t)\beta(\boldsymbol{z}^{(l)}_t)}{p(\boldsymbol{y}^{(l)}_1,\ldots,\boldsymbol{y}^{(l)}_T)}$$

$$\xi(\boldsymbol{z}^{(l)}_{t-1},\boldsymbol{z}^{(l)}_t) = \frac{\alpha(\boldsymbol{z}^{(l)}_{t-1})p(\boldsymbol{y}^{(l)}_t|\boldsymbol{z}^{(l)}_t)p(\boldsymbol{z}^{(l)}_t|\boldsymbol{z}^{(l)}_{t-1})\beta(\boldsymbol{z}^{(l)}_t)}{p(\boldsymbol{y}^{(l)}_1,\ldots,\boldsymbol{y}^{(l)}_T)}$$

これは，$\alpha(\boldsymbol{z}^{(l)}_t)$ と $\beta(\boldsymbol{z}^{(l)}_t)$ を

$$\alpha(\boldsymbol{z}^{(l)}_t) = p(\boldsymbol{y}^{(l)}_t|\boldsymbol{z}^{(l)}_t)\sum_{\boldsymbol{z}^{(l)}_{t-1}}\alpha(\boldsymbol{z}^{(l)}_{t-1})p(\boldsymbol{z}^{(l)}_t|\boldsymbol{z}^{(l)}_{t-1}) \tag{6}$$

$$\beta(\boldsymbol{z}^{(l)}_t) = \sum_{\boldsymbol{z}^{(l)}_{t+1}}\beta(\boldsymbol{z}^{(l)}_{t+1})p(\boldsymbol{y}^{(l)}_{t+1}|\boldsymbol{z}^{(l)}_{t+1})p(\boldsymbol{z}^{(l)}_{t+1}|\boldsymbol{z}^{(l)}_t) \tag{7}$$

に従って逐次的に計算することによって求められる。ここで，尤度は以下のように計算できる。

$$p(\boldsymbol{y}_1^{(l)},\ldots,\boldsymbol{y}_T^{(l)}) = \sum_{\boldsymbol{z}_T^{(l)}} \alpha(\boldsymbol{z}_T^{(l)}) \tag{8}$$

M ステップでは，$\gamma(\boldsymbol{z}_t^{(l)})$ と $\xi(\boldsymbol{z}_{t-1}^{(l)},\boldsymbol{z}_t^{(l)})$ を定数とみなし，パラメータ集合 $\boldsymbol{\theta}^{(l)}$ に関して Q 関数を最大化する。適当なラグランジュ乗数を用いることで，更新式は以下のように導出される。

$$\pi_k^{(l)} = \frac{\gamma(z_{1,k}^{(l)})}{\sum_{j=1}^K \gamma(z_{1,j}^{(l)})},\ h_{\omega,k}^{(l)} = \frac{\sum_{t=1}^T \gamma(z_{t,k}^{(l)}) y_{\omega,t}^{(l)}}{\sum_{t=1}^T \sum_{\omega=1}^\Omega \gamma(z_{t,k}^{(l)}) y_{\omega,t}^{(l)}}$$
$$A_{j,k}^{(l)} = \frac{\sum_{t=2}^T \xi(z_{t-1,j}^{(l)}, z_{t,k}^{(l)})}{\sum_{m=1}^K \sum_{t=2}^T \xi(z_{t-1,j}^{(l)}, z_{t,m}^{(l)})} \tag{9}$$

ここで，$h_{\omega,k}^{(l)}$ が ω に関して正規化されることに注目したい。提案モデル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おいてどういうものが基底スペクトルとして据えられるかということに大きく関係する。更新式から分かるように，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）ほど基底の決定に大きく寄与する。つまり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。

すべての言語 $\{1,\ldots,L\}$ において基底メルスペクトルを共有することもできる。その場合の更新式は下記のように導出される。

$$h_{\omega,k} = \frac{\sum_{l=1}^L \sum_{t=1}^T \gamma(z_{t,k}^{(l)}) y_{\omega,t}^{(l)}}{\sum_{l=1}^L \sum_{t=1}^T \sum_{\omega=1}^\Omega \gamma(z_{t,k}^{(l)}) y_{\omega,t}^{(l)}} \tag{10}$$

3　評価実験

英語，アメリカ英語，アラビア語，中国語，オランダ語，フィンランド語，フランス語，ドイツ語，ギリシャ語，ヒンズー語，ハンガリー語，インドネシア語，イタリア語，日本語，韓国語，ポーランド語，ポルトガル語，ロシア語，スペイン語，スウェーデン語，タイ語からなる合計 21 言語の多言語音声データベースを利用して，提案モデルの基本動作を評価する。このデータベースには，言語ごとにネイティブの話者 8 名（男性 4 名，女性 4 名）による音声が収録されており，各話者が発声した文章の総数は 24 である。文章はすべて異なり，2 つの文章を発声した音声を 1 つのファイルとして収録する。音声信号のサンプリング周波数は 16kHz，量子化ビット数は 16 で収録され，ファイル数は 2016，音声信号の長さの平均は 6.7 秒である。

まず，各ファイルの音声信号はその振幅の最大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 32ms，フレーム長 16ms，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。

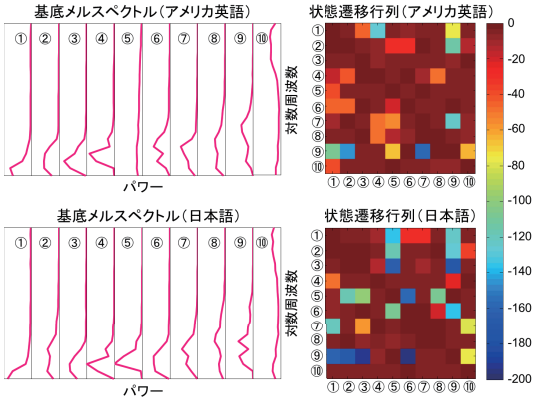

Fig. 2　アメリカ英語と日本語の音声信号から推定された基底と状態遷移行列（$\beta = 0.5$，$K = 10$ とした場合であり，状態遷移確率は対数値である）

$$\boldsymbol{y}_t = [y_{1,t},\ldots,y_{\Omega,t}]^{\mathrm{T}} = [w_{1,t}^\beta,\ldots,w_{\Omega,t}^\beta]^{\mathrm{T}} \tag{11}$$

通常はスペクトル包絡構造を強調するために対数値を用いるが，提案法は非負値制約のため，このような処理を行った。メルフィルタの総数は $\Omega = 22$ とした。

提案アルゴリズムの初期値 $\boldsymbol{\pi}^{(l)}, \boldsymbol{A}^{(l)}$ は乱数を与えて正規化した。また，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基底を，初期値 $\{\boldsymbol{h}_1^{(l)},\ldots,\boldsymbol{h}_K^{(l)}\}$ とした。アルゴリズムの反復回数は 100 回とし，推定結果例を Fig. 2 に示す。アメリカ英語と日本語それぞれ，4 名の話者（男性 2 名，女性 2 名）によるすべての収録音声を学習データとして，式 (6)〜(9) から各パラメータを推定した。最大値となる ω が小さい順序となるように基底を並べ替えて図示した。状態遷移確率はその対数をとって図示した。推定された基底には包絡構造を確認できる。また，英語は日本語に比べて，基底の遷移が多い言語であることがわかった。これは英語の音素数が日本語に比べて多いことにも関連すると考えられる。

次に，β と基底数 K の値を変化させたときのモデルの挙動を検証するために，21 言語の自動識別実験を行った。言語ごとに 4 名の話者（男性 2 名，女性 2 名）を選択し，そのすべての音声信号を学習データとしてパラメータを推定する（言語ごとに学習データのファイル数は合計 48 である）。同時に，これらを評価データとして，音声ごとに計算される事後確率

$$p(l|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_T) \propto p(\boldsymbol{y}_1,\ldots,\boldsymbol{y}_T|l)p(l) \tag{12}$$

に基づいて識別結果を算出した。ここで，事前確率は等確率 $p(l=1) = p(l=2) = \ldots = p(l=L) = 1/L$ と仮定した。学習データと評価データが同じであるため，クローズドな評価実験と言える。Fig. 3 に β と

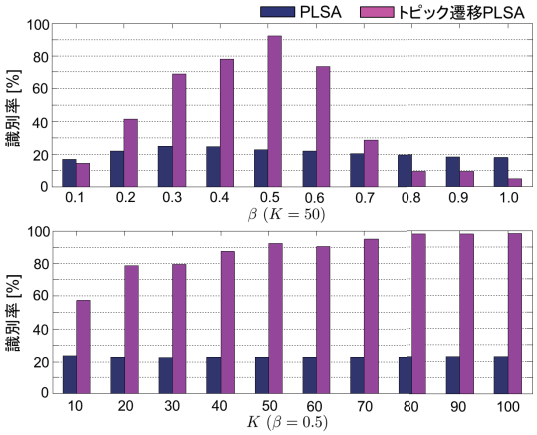

Fig. 3　β と基底数 K の値を検証し，従来の PLSA と比較するための，クローズドデータを利用した評価実験結果

K を変化させたときの識別結果を示す。$K=50$ に固定して β を 0 から 1 までの間で変化させた場合，識別率が最も高かったのは $\beta = 0.5$ のときであった。メルフィルタバンクの出力値をべき乗してスペクトルの包絡構造を強調することの有効性を確認できた。

一方，基底数 K を増やすにつれて，識別性能は向上した。ただし，フォワード・バックワードアルゴリズムの計算コストが大きくなるため，処理時間は増加する。通常の PLSA に比べて提案法の識別性能が高いことも確認できた。基底として表現される各言語の音響的特徴だけでなく，それらの時間遷移も識別に有効な特徴であることがわかった。

最後に，オープンな評価実験結果を Fig. 4 に示す。言語ごとに，残りの 4 名の音声信号を評価データとし，先に推定されたパラメータを用いて計算される，式 (12) の事後確率に基づいて識別した。K を大きくするにつれて，必ずしも識別性能が向上するわけではなく，$K = 60$, 80 のときに，最も高い識別率 18.3%が得られた。これは学習データに対するパラメータの過学習が原因であると考えられる。そもそも学習データ自体，言語ごとに 4 名のみの音声データであるため，全体的に識別性能が著しく低くなったと思われる。文献 [4] で利用される NIST 多言語データベースなどを用いて，話者の多様性を含め，大規模にモデルを学習する必要がある。

4　まとめと今後の課題

事前知識無しに音声信号のみから多言語の分類・類型化を行うために，音響信号処理分野で発展する NMF をベースとしたトピック遷移 PLSA モデルを提案し，EM アルゴリズムによる学習則を導出した。この手法では，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，要素基底と状態遷移確率によって別々に学習される。そして，提案モデルの基本動作を検証するために，21 言語からなる多言語音

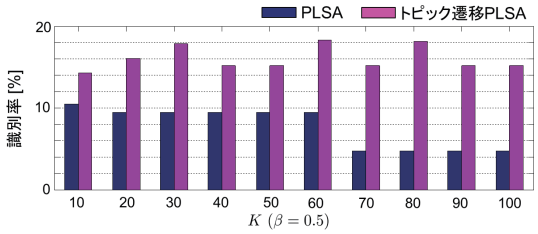

Fig. 4　オープンデータを用いた評価実験結果（$\beta = 0.5$ に固定し，基底数 K を変化させた場合）

声データベースを利用して，言語識別実験を行った。通常の PLSA に比べて，基底の時間的遷移を陽に導入する提案モデルの有効性を確認できたが，オープンデータを用いた識別実験では性能が著しく低下した。今後は大規模データベースを利用して，最適なメルフィルタの数や基底数，分析フレーム長やフレームシフト長，話者性（声道長正規化など）を考慮しながら，提案モデルの学習と評価，従来法との比較実験を行う予定である。

また，提案モデルに識別的アプローチを導入することも今後の課題である。文献 [11] より，識別的アプローチと生成的アプローチの関係は

$$p(l, \theta^{(l)}|y^{(l)}) = \frac{p(y^{(l)}, l|\theta^{(l)})}{\sum_l p(y^{(l)}, l|\theta^{(l)})} p(\theta^{(l)}) \qquad (13)$$

と記述できる。ここで，l は言語ラベル，$y^{(l)}$ は言語 l の観測データを表す。本稿では，右辺の分子を最大化して，言語ごとに $\theta^{(l)}$ を推定する学習則を導出したが，分母を考慮した，各言語ラベルの事後確率を最大化する学習則について検討中である。これにより，言語識別に効果的な（より言語を特徴付ける）メルスペクトルが，基底として学習されることを期待する。

参考文献

[1] 山本秀樹, “世界言語の地理的・系統的語順分布とその変遷,” 渓水社, 2003.
[2] R. G. Gordon *et. al.*, “Ethnologue Languages of the World,” Fifteenth Edition SIL International, 2005.
[3] K. Yeshwant *et. al.*, IEEE Signal Processing Magazine, pp. 33–41, 1994.
[4] C. S. Greenberg *et. al.*, in *Proc. Interspeech 2012*.
[5] P. Matejka *et. al.*, in *Proc. Interspeech 2012*.
[6] T. Virtanen, IEEE TASLP, vol. 15, pp. 1066–1074, 2007.
[7] 緒方, 高木, 音講論 (春), pp. 247-248, 2012.
[8] L. J. Rodriguez-Fuentes *et. al.*, in *Proc. Interspeech 2012*.
[9] T. Hofmann, in *Proc. SIGIR 1999*.
[10] L. R. Rabiner, in *Proc. IEEE*, pp. 257–286, 1989.
[11] J. A. Lasserre *et. al.*, in *Proc. CVPR 2006*.