Deep Neural Network 混合モデルを用いた環境・話者適応の検討*

☆柏木陽佑 (東大), 久保陽太郎, 中村篤 (NTT 研究所), 峯松信明, 広瀬啓吉 (東大)

1 はじめに

音声認識において，モデルと観測特徴量間のミスマッチは認識率を下げる大きな要因の 1 つである．そのため，このミスマッチを軽減するために，音声特徴量の正規化 [1] や，モデルの適応 [2] など様々な手法が検討されて来た．近年，ミスマッチに頑健な音響モデルの枠組みとして Deep Neural Network (DNN) が非常に注目されている [3]．従来の GMM を用いた音響モデルと比較して，DNN 音響モデルは高い認識率を示すことが知られている．

DNN は従来の GMM を用いた音響モデルよりミスマッチに対して頑健であると言える．しかし，ミスマッチの影響を完全に取り除くことができるわけではなく，DNN 音響モデル対する適応の検討がなされている [4,5]．例えば，[5] は DNN を特徴量抽出部分と識別部分の組み合わせであると考え，DNN の最も出力側に近い隠れ層への入力を MLLR (Maximum Likelihood Linear Regression) と同等の変換によって適応することを行っている．しかし，MLLR の枠組みによる適応は適応データの蓄積が必要であり，特に DNN はパラメータ数が多いために適応データの数が多くなると考えられる．

それに対し，VTLN (Vocal Tract Length Normalization) [6] のような発話単位での高速な適応は適応データを蓄積できない場合でも利用可能である．VTLN は声道長正規化を行うもので，このパラメータを発話単位で変化させることで高速な適応を可能としている．VTLN と同様，音声認識器の適応を複数パラメータの選択問題に帰着できれば，パラメータ数の多いモデルであっても高速に適応できるということが期待される．

そこで，本研究では，DNN の混合モデルを定義することにより，混合モデル内の要素選択によって話者・環境適応することを検討する．また，そのために話者・環境のクラスタごとに構成した DNN を組み合わせて混合モデル化し，学習データから同時に異なる性質の DNN を得る枠組みを提案する．そして，認識時に，適切に DNN を選択することによって，DNN を話者や環境に対して適応することを可能とする手法も提案する．

提案法では，VTLN 同様，話者 (環境) による変動は発話内で一定であることを仮定し，発話単位とフレーム単位の二つの潜在変数を持つ階層モデルとして表現するアプローチ [7,8] を導入する．これによって，話者や環境の情報の適切な利用と，フレーム単位での識別精度の両立を可能とする．

この枠組みにより学習することで，同時に複数の異なる性質の DNN を構築することが可能となる．学習の結果，環境ラベルが，話者性，環境などによって分かれることによって，より適切に音素 HMM 状態と特徴量の関係をモデル化することが期待され，またその結果，認識時に適切にモデルを選択することで話者・環境適応が可能となることが期待される．

本稿の構成を以下に示す．第 2 節では従来の DNN 音響モデルについて説明を行う．第 3 節で提案手法について説明を行い，第 4 節で実験によって提案手法の有効性を示す．その後，第 5 節でまとめと今後の展望について述べる．

2 Deep Neural Network 音響モデル

DNN は隠れ層を 2 層以上持つ多層パーセプトロンであり，非常に強力な識別器として近年注目されている．従来よりも多数の層それぞれが特徴量抽出に相当する機能を獲得することで非常に複雑な空間での識別をも可能とする．しかし，そのような多層パーセプトロンにはモデルパラメータの最適化が非常に難しいという問題があった．これに対して Hinton らは Restricted Boltzmann Machine (RBM) として学習したパラメータを各層の初期モデルとして用いてバックプロパゲーションにより学習することで解決する手法を提案した [9]．このような背景により，近年音声認識の音響モデルとして DNN が非常に注目され利用されている．

音声認識は式 (1) のように観測特徴量系列 $\boldsymbol{x}$ が得られた時，正解テキスト l の推定値 $\hat{l}$ を求めるタスクである．

$$\hat{l} = \underset{l}{\operatorname{argmax}} \log p(l|\boldsymbol{x}) \tag{1}$$

式 (1) を音響モデルから算出される確率 $p(\boldsymbol{x}|l)$ と言語モデルから算出される確率 $p(l)$ の積で表すと

$$\hat{l} = \underset{l}{\operatorname{argmax}} \log p(\boldsymbol{x}|l)p(l) \tag{2}$$

となり，この音響モデル部分 ($p(\boldsymbol{x}|l)$) を DNN でモデル化することを考える．音声信号の時間方向の伸縮構造の表現には，従来通り隠れマルコフモデル (HMM) を用いることを考え，音響モデル確率を HMM 状態系列で周辺化した形で記述すると

$$p(\boldsymbol{x}|l) \overset{\text{def}}{=} \sum_{\boldsymbol{s}} \left\{ \prod_t q(s_t|\boldsymbol{x}_t, \boldsymbol{\Lambda}) \right\} p(\boldsymbol{x}) p(\boldsymbol{s}|l) \tag{3}$$

$$q(s_t|\boldsymbol{x}_t, \boldsymbol{\Lambda}) = \frac{p(s_t|\boldsymbol{x}_t, \boldsymbol{\Lambda})}{p(s_t)}$$

となる．ここで $\boldsymbol{\Lambda}$ は DNN のパラメータ，$\boldsymbol{s}$ は HMM 状態系列である．$q(s_t|\boldsymbol{x}_t, \boldsymbol{\Lambda})$ は $p(\boldsymbol{x}_t|s_t, \boldsymbol{\Lambda})$ に比例する項であり，$p(s_t)$ は一様分布を仮定する，あるいはコーパス中の出現回数をカウントし，最尤推定により求める．このようにして，識別モデルである DNN を音響モデルとして利用することが可能となっている．

*A study on environmental and speaker adaptation using deep neural network mixture models by Y.Kashiwagi (The University of Tokyo), Y.Kubo, A.Nakamura (NTT), N.Minematsu, and K.Hirose (The University of Tokyo)

3 提案手法

DNN を用いた音響モデルを話者・環境適応するに当たり，まず，特徴量の分布が話者性・環境などに相当する隠れ状態インデクス（環境ラベル）を持つものとし，その上で，ある環境ラベルにおける音素 HMM 状態と特徴量の関係を DNN でモデル化することを考える．しかし，環境ラベルを直接観測することは困難であるため，周辺分布を考慮することでより適切に音響モデルを構築することを考える．まず，提案手法を定式化し，学習方法と認識時の応用方法について述べる．

3.1 定式化

文 l が与えられた時の，環境ラベル k と入力特徴量系列 $\boldsymbol{x}$ との結合確率を $p(\boldsymbol{x}, k|l)$ とし，これを最大化するものを音声認識結果とするモデルを考える．

$$\hat{l} = \operatorname*{argmax}_{l} p(\boldsymbol{x}, k|l) \tag{4}$$

しかし，k は観測することができないため，式 (4) を直接最大化することは困難である．そこで，代わりに k についての周辺分布を最大化する．HMM 状態系列 s を考慮すると，

$$\begin{aligned}\hat{l} &= \operatorname*{argmax}_{l} \sum_{k} p(\boldsymbol{x}, k|l) \qquad (5)\\ &\approx \operatorname*{argmax}_{l} \sum_{\boldsymbol{s}} \sum_{k} p(\boldsymbol{s}|\boldsymbol{x}, k) p(\boldsymbol{x}|k) p(k) p(\boldsymbol{s}|l)\end{aligned}$$

ここで，$p(k) = \pi_k$ とすることにより

$$\hat{l} = \operatorname*{argmax}_{l} \sum_{\boldsymbol{s}} \sum_{k} \pi_k \left\{ \prod_{t} p(s_t|\boldsymbol{x}_t, \boldsymbol{\Lambda}_k) \right\} \times p(\boldsymbol{x}|\boldsymbol{\Theta}_k) p(\boldsymbol{s}|l) \tag{6}$$

とし，$p(s_t|\boldsymbol{x}_t, \boldsymbol{\Lambda}_k)$ を DNN，$p(\boldsymbol{x}|\boldsymbol{\Theta}_k)$ を正規分布もしくは GMM でモデル化する．ここで，$\boldsymbol{\Lambda}_k$ と $\boldsymbol{\Theta}_k$ は，それぞれある隠れ状態 k に対応するモデルパラメータである．このようにモデル化を行うことで，フレーム単位での音素 HMM 状態識別性能の高い DNN を $p(s_t|\boldsymbol{x}_t, \boldsymbol{\Lambda}_k)$ でモデル化しているのに対し，話者性や環境の影響を大きく受ける発話全体の特徴量系列の分布を $p(\boldsymbol{x}|\boldsymbol{\Theta}_k)$ で考慮することが可能となる．

3.2 学習

学習時は，入力特徴量と，あらかじめ GMM/HMM によりアライメントされた正解出力音素 HMM 状態系列の組の学習データを利用してモデルパラメータ $\boldsymbol{\Gamma} = \{\boldsymbol{\Lambda}_k, \boldsymbol{\Theta}_k, \pi_k | \forall k\}$ の最尤推定を行う．このモデルは隠れ状態 k を持つために，EM-algorithm を用いて学習する必要がある．学習データのインデックスを $j = \{1, 2, \ldots, J\}$ とすると，学習データは音素 HMM 状態系列 $\boldsymbol{s}^j$ と入力特徴量系列 $\boldsymbol{x}^j$ の集合として表すことが出来る．学習時の目的関数は，

$$\operatorname*{argmax}_{\boldsymbol{\Gamma}} \sum_{j} \sum_{\boldsymbol{s}} \sum_{k} \pi_k \left\{ \prod_{t} p(s_t|\boldsymbol{x}_t, \boldsymbol{\Lambda}_k) \right\} \times p(\boldsymbol{x}|\boldsymbol{\Theta}_k) p(\boldsymbol{s}|l) \tag{7}$$

となり，これを直接最大化することは困難であるため，EM-algorithm を用いて最適化することを考える．まず，完全対数尤度の下界として Q 関数を以下のようにおく．

$$Q(\boldsymbol{\Gamma}|\boldsymbol{\Gamma}^{(n)}) \stackrel{\text{def}}{=} \sum_{j} \sum_{k} p(k|\boldsymbol{s}, \boldsymbol{x}, \boldsymbol{\Gamma}^{(n)}) \times \log \pi_k \prod_{t} \{p(s_t|\boldsymbol{x}_t, \boldsymbol{\Lambda}_k)\} p(\boldsymbol{x}|\boldsymbol{\Theta}_k) \tag{8}$$

ここで，$\boldsymbol{\Gamma}^{(n)}$ は EM-algorithm の n 番目のイテレーションにおけるパラメータの推定値である．EM-algorithm の E-step ではこの Q 関数中の寄与率項 $(p(k|s, x, \boldsymbol{\Gamma}^{(n)}))$ を以下のように計算する．

$$\begin{aligned}&p(k|\boldsymbol{x}^j, \boldsymbol{s}^j, \boldsymbol{\Gamma})\\ &= \frac{\pi_k \prod_t \left\{ p(s_t^j|\boldsymbol{x}_t^j, \boldsymbol{\Lambda}_k) \right\} p(\boldsymbol{x}^j|\boldsymbol{\Theta}_k)}{\sum_{k'} \pi_{k'} \prod_t \left\{ p(s_t^j|\boldsymbol{x}_t^j, \boldsymbol{\Lambda}_{k'}) \right\} p(\boldsymbol{x}^j|\boldsymbol{\Theta}_{k'})}\end{aligned} \tag{9}$$

EM-algorithm の M-step では，E-step で求めたそれぞれの隠れ状態インデックスへの寄与率を元に，Q 関数 (式 (8)) を最大化するようにモデルパラメータを更新する．

$$\boldsymbol{\Gamma}^{(n+1)} = \operatorname*{argmax}_{\boldsymbol{\Gamma}} Q(\boldsymbol{\Gamma}|\boldsymbol{\Gamma}^{(n)}) \tag{10}$$

ここで，$\boldsymbol{\Gamma}$ の中の DNN パラメータに相当する部分 $\boldsymbol{\Lambda}_k(\forall k)$ は解析的に求めることができないため，ニューラルネットに対する反復解法である Back Propagation を M-step の中で用いる必要がある．学習時のアルゴリズムをまとめると以下のようになる．

repeat
  // E-step
  estimate γ_k^j ***from*** $\boldsymbol{\Gamma}^{(n)}$
  Viterbi approximation $\gamma_k^j = 1$ or 0
  // M-step
  ML estimation $\boldsymbol{\Theta}^{(n+1)}$ and $\pi^{(n+1)}$
  repeat
    compute $\delta\boldsymbol{\Lambda}_k(\forall k)$ update $\boldsymbol{\Lambda}_k(\forall k)$
    update $\boldsymbol{\Lambda}$
  until $\boldsymbol{\Lambda}^{(n+1)}$ converge
until the parameters converge

最適化のアルゴリズムは上のように，EM-algorithm によるループの部分と，バックプロパゲーションによるループの部分から構成される，二重の最適化となっている

3.3 認識

認識時は学習によって得られたパラメータを利用し，

$$\hat{l} = \operatorname*{argmax}_{l} \log \sum_{\boldsymbol{s}} \sum_{k} \pi_k p(\boldsymbol{s}|\boldsymbol{x}, \boldsymbol{\Lambda}_k) \times p(\boldsymbol{x}|\boldsymbol{\Theta}_k) p(\boldsymbol{x}|\boldsymbol{s}) p(\boldsymbol{s}|l) p(l) \tag{11}$$

として $\hat{l}$ を求める．この際，各環境ラベルへの寄与率に対して Viterbi 近似を行う．この近似によるモデル選択は，認識時の π_k を認識データを用いて適応することに相当する．しかし，認識時の場合 1 を観測でき

ないため，寄与率 (式 (9)) を直接求めることは難しい．そこで，未知のデータに対して，複数の環境ラベルに対応するモデルから適切に用いるモデルを選択する近似手法を 4 つ提案する．

モデル尤度基準　学習時と同様の基準でコンポーネントを選択する．それぞれのコンポーネントの出力尤度を計算し，最も高い尤度を出力するクラス k についての Viterbi 近似を行うことによってモデル選択を実現している．

$$\begin{aligned}\hat{k} &\approx \operatorname*{argmax}_k \log \pi_k p(\boldsymbol{x}|\boldsymbol{\Theta}_k) p(\boldsymbol{s}'|\boldsymbol{x}, \boldsymbol{\Lambda}_k) \\ &\quad \times p(\boldsymbol{x}|\boldsymbol{s}') p(\boldsymbol{s}'|l) p(l) \quad (12) \\ \hat{l} &\approx \operatorname*{argmax}_l \log p(\boldsymbol{s}'|\boldsymbol{x}, \boldsymbol{\Lambda}_{\hat{k}}) p(\boldsymbol{s}'|l) p(l)\end{aligned}$$

ここで $\boldsymbol{s}'$ と k' はそれぞれ Viterbi 近似後の $\boldsymbol{s}$ と k である．この基準は EM 学習時の寄与率の計算と同様の計算に基づくため，最も学習則に近い．

音素識別率基準　式 (13) のように発話単位の特徴量分布を考慮せずにコンポーネントを選択する．これは，発話単位の特徴量分布形として正規分布もしくは GMM を仮定しているため，認識の際に DNN の出力する尤度と比較して信頼できない可能性があるためである．

$$\begin{aligned}\hat{k} &\approx \operatorname*{argmax}_k \log p(\boldsymbol{s}'|\boldsymbol{x}, \boldsymbol{\Lambda}_k) p(\boldsymbol{s}'|l) p(l) \\ \hat{l} &\approx \operatorname*{argmax}_l \log p(\boldsymbol{s}'|\boldsymbol{x}, \boldsymbol{\Lambda}_{\hat{k}}) p(\boldsymbol{s}'|l) p(l) \quad (13)\end{aligned}$$

この基準は，音素列の確率のみを考慮し利用するモデルを選択していくことから，システムコンビネーションに近いと言える．

特徴量分布基準　式 (14) のように話者や環境に対する尤度でコンポーネントを選択する場合も考える．この場合は，話者・環境適応という考えに最もマッチしていると考えられるが，実際の分布がモデルから逸脱してしまう特徴量に対しての精度が信頼できない．

$$\begin{aligned}\hat{k} &\approx \operatorname*{argmax}_k \log \pi_k p(\boldsymbol{x}|\boldsymbol{\Theta}_k) \\ \hat{l} &\approx \operatorname*{argmax}_l \log p(\boldsymbol{s}'|\boldsymbol{x}, \boldsymbol{\Lambda}_{\hat{k}}) p(\boldsymbol{s}'|l) p(l) \quad (14)\end{aligned}$$

特徴量の分布が近ければ同じ話者や環境であるとの仮定を重視しているため，発話単位での特徴量分布ではあるが，SPLICE [10] と非常に似た基準となっている．

フレーム単位統合　また，発話単位での正規分布でのコンポーネント選択の効果を調べるために，式 (15) のようにフレーム単位で正規分布もしくは GMM からの尤度を計算して DNN の尤度と統合する場合も考慮する．これは，上記の 3 つの基準とは異なり統合を行っており，モデル選択ではない．

$$\begin{aligned}\hat{l} &\approx \operatorname*{argmax}_l \log \left\{ \prod_t \sum_k \pi_k p(s'_t|x_t, \boldsymbol{\Lambda}_k) p(x_t|\boldsymbol{\Theta}_k) \right\} \\ &\quad \times p(\boldsymbol{s}'|l) p(l) \quad (15)\end{aligned}$$

4　実験

提案手法の有効性と，認識時のモデル選択基準の評価を音声認識実験によって比較した．実験に用いるコーパスは TIMIT を採用した．学習セットと評価セットの発話数はそれぞれ 3,696 発話と 192 発話である．特徴量は MFCC ＋対数パワー＋Δ＋ΔΔの 39 次元を用いた．

初期検討として今回はモデルの混合数を全ての場合で 2 とした．DNN はデベロップメントセットで最も良い結果を得ることができた中間層が 5 層のものを採用した．DNN の各中間層のノード数は 2048 と 1024 の場合を試した．特徴量分布としてガウス分布と GMM の場合を考え，それぞれの分散行列は，ガウス分布の場合は全共分散行列を，GMM の場合は対角共分散行列を仮定した．また，GMM の場合，その混合数は 128 である．EM-algorithm の初期値は隠れ変数をランダムで与えている．ただし，計算量削減のために，式 (16) のようにもっとも寄与率の大きなクラスとそれ以外に 2 値で割り振った．

$$\begin{aligned}\hat{k}_j &= \operatorname*{argmax}_k p(k|\boldsymbol{x}^j, \boldsymbol{s}^j, \boldsymbol{\Gamma}) \\ \gamma_k^j &= \begin{cases} 1 & k_j = \hat{k}_j \\ 0 & k_j \neq \hat{k}_j \end{cases} \quad (16)\end{aligned}$$

Fig. 1 に音声認識結果の音素誤り率（Phoneme Error Rate ; PER）を示す．それぞれ中間層のノード数が 1024 の場合と 2048 の場合の結果であり，EM-algorithm の各イテレーションの中で最も認識率が良かった時の結果である．特徴量分布に正規分布と GMM を利用した場合それぞれにおける，認識時のモデルコンポーネント選択方法 3 通りとフレーム単位でコンポーネントの尤度を算出する場合の比較を行った．また，ベースラインとして，混合モデル化をしていない従来の DNN 音響モデルを用いた際の認識結果も併せて記載した．

本提案手法により，ベースラインである従来の DNN 音響モデルと同程度もしくは低い PER を得ることができた．これにより，本稿で提案した手法による話者・環境適応が DNN 音響モデルの高精度化に寄与する可能性が示された．

最も良い性能を得られた実験条件は，中間層のノード数を 2048，特徴量の分布として正規分布を仮定し，音素識別基準に基づいたモデル選択を行なうといった条件であった．これは学習の際も安定して収束することが観測された条件であった．ノード数 1024，特徴量の分布を GMM で認識時に特徴量の分布の尤度でコンポーネント選択する場合も同程度に結果が良いが，こちらは学習の収束が不安定であることが観測された．

特徴量の分布を GMM でモデル化した場合，特徴量分布基準でモデル選択を行う場合が最も PER が小さくなる．これは特徴量の分布は正規分布よりも GMM の方が精密に表現することができるためだと考えられる．

また，DNN の中間層のノード数が 1024 の場合であっても，提案手法により適応を行なうことで，2048 の場合と同等以上の認識結果を得られることがわかった．これは，話者性や環境に関する普遍性を得る上

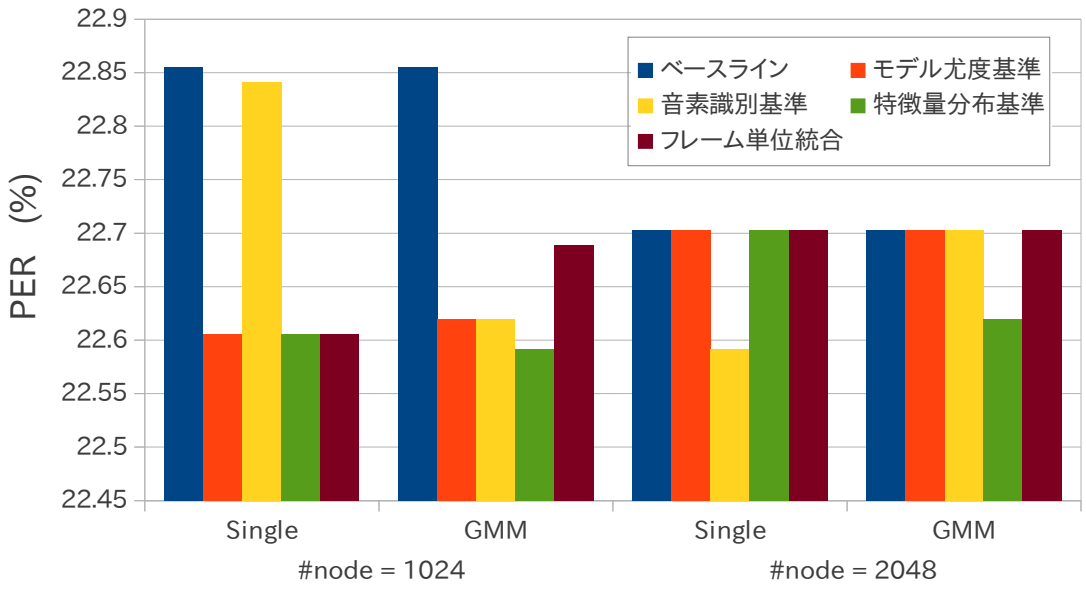

Fig. 1　Phoneme error rates (number of hidden node = 1024 or 2048).

で，DNN のノード数を増加させるよりも提案手法のように混合モデル化を行うことの優位性を示唆するものと考えられる．

逆に，混合モデル化を行う場合はノード数が 2048 の場合，特徴量分布の精度に依存して最適な基準が大きく変わる．学習時，環境ラベル毎に学習データを 0/1 の寄与率で割り当てているため，実質の学習データが半減していることによる過学習が原因だと考えられる．そのため，本提案手法では，学習データ数との兼ね合いによりモデル混合数に応じて適切に中間層のノード数を変化させる必要があると考えられる．本実験で行なったノード数の調整に基づくパラメータ数の調整の他に，層数の調整などに基づくパラメータ数の調整も考えられるが，これは今後の課題としたい．

5　まとめ

本稿では，特徴量空間を確率モデルによって分割し，その分割された部分空間一つ一つに DNN による識別器を持つ混合 DNN モデルを提案し，混合重みを発話毎によって切り換えるとことで話者・環境適応を行なう音声認識器を構築した．実験により，この枠組みを用いた DNN 音響モデルに対する適応が効果的であることがわかった．

今後，コンポーネントの数を増やして更なる検討を行うとともに，それに伴う計算量の増大に対処するための高速化を検討する．具体的には DNN のノード数や層数の調整を行なうことによる高速化や，より洗練された近似の導入による高速化を検討したい．また，今回は評価セットの認識率に基づいて繰り返し計算数を決定したため特に GMM の場合の過学習を考慮できていない．今後，適切に過学習を回避するために，EM-algorithm を何らかの基準により適切に終了するなどの方法の導入の検討を行う．

参考文献

[1] Viikki, O. and Laurila, K. , “Cepstral domain segmental feature vector normalization for noise robust speech recognition,” Speech Communication, vol. 25, no. 1, pp. 133–147, 1998.

[2] Gales, M. J. F. and Woodland, P. C. , “Mean and covariance adaptation within MLLR framework,” Computer Speech and Language, vol. 10, no.4, pp. 249–264, 1996.

[3] Mohamed, A. and Dahl, G.E. and Hinton, G. , “Acoustic modeling using deep belief networks,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 1, pp.14–22, 2012.

[4] Trmal, J. and Zelinka, J. and Müller, L. , “On Speaker Adaptive Training of Artificial Neural Networks,” Eleventh Annual Conference of the International Speech Communication Association, 2010.

[5] Yao, K. and Yu, D. and Seide, F. and Su, H. and Deng, L. and Gong, Y. , “Adaptation of Context-Dependent Deep Neural Networks for Automatic Speech Recognition,” ACL Workshop on Spoken Language Technology, pp. 366–369, 2012.

[6] Eide, E. and Gish, H. , “A parametric approach to vocal tract length normalization,” ICASSP96, vol. 1, pp. 346–348, 1996.

[7] Seong-Jun, H. and Ohkawa, Y. and Ito, M. and SUZUKI, M. and Ito A. and MAKINO, S. , “Speech Recognition under Multiple Noise Environment Based on Multi-Mixture HMM and Weight Optimization by the Aspect Model,” IEICE transactions on information and systems, vol. 93, no. 9, pp. 2407–2416, 2010.

[8] Watanabe, S. and Mochihashi, D. and Hori, T. and Nakamura, A. , “Gibbs Sampling Based Multi-Scale Mixture Model for Speaker Clustering,” Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp. 4524–4527, 2011.

[9] Hinton, G. E., Osindero, S. and Teh, Y., “A fast learning algorithm for deep belief nets,” Neural Computation, vol. 18, no. 7, pp. 1527–1554, 2006.

[10] Droppo, J. and Deng, L. and Acero, A. “Evaluation of the SPLICE algorithm on the Aurora2 database,” Proc. Eurospeech, vol. 1, pp. 217–220, 2001.