制約付き話者コードの同時推定による ニューラルネット音響モデルの話者正規化学習*

◎柏木陽佑, 齋藤大輔, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

近年，Deep Learning 技術の出現により非線形な識別モデルである多層ニューラルネットを用いた音響モデルが高い認識性能を示し，注目されている [1]．一般に，Deep Neural Network (DNN) 音響モデルにおいて，DNN は各時間におけるセグメント特徴量を入力として音素状態を識別する．DNN がその多層の構造により特徴量抽出と識別に相当する機能を併せ持つと考えた場合，DNN は潜在的に話者正規化の機能を保持していると考えられる．しかし，近年，DNN の話者適応の研究が盛んに行われており，話者適応により認識精度の向上が経験的に得られることから，DNN の持つ話者正規化機能が不十分であることは明白である [2–10]．

この原因として，入力情報の欠如が考えられる．一般に DNN の入力として用いる数十フレームのセグメント特徴量のみからでは，話者の同定は困難である．通常，話者の情報は長時間に表れるものであり，一般に話者識別に用いられる i-vector などではある程度の長さの音声により UBM を適応する．つまり，数十フレームのセグメント特徴量のみからでは話者の違いによる変動と音素の違いによる変動を back propagation 時に明確に分離することができないことを示している．従って，一般的な音響モデルは不特定話者の学習データからフレーム単位で特徴量を抽出し学習するため，各入力特徴量が同一話者から発話されたという情報が十分に反映されていない．しかし，入力フレーム数が増大することは過学習を引き起こし，畳み込み構造やリカレント構造を導入する場合は実装コスト等の増大が懸念される．

さて，話者適応技術として Ossama らの提案した話者コードを用いた話者適応がある [7]．これは，各層のバイアス項を話者毎に学習した話者コードの線形変換により計算する．[10] と同様に DNN の学習時に各セグメントがどの話者に所属するかの情報を与えることで学習効率を上げていると解釈することができる．

本提案手法では，話者を表現する特徴量と音素状態識別を行う DNN の同時推定を行うことを目的として，話者コードベースの話者正規化学習を提案する．

話者コードは少数のパラメータにより各層のバイアスを制御することが可能である．DNN が多層の構造により段階的な特徴量変換を行っていると考えると，話者コードにより異なる特徴量ドメインにおいて話者依存の成分を分離することが期待できる．また，各話者毎の学習データ数はそれほど多くないことが十分考えられるため，話者コードをボトルネック層として設計することで制約を導入する．これにより完全に話者毎により学習されるパラメータ群，話者クラス毎に学習されるパラメータ群，全話者データにより学習されるパラメータ群に分離することで，話者属性の階層的モデリングが可能になると考えられる。

本稿の構成を示す．まず，2 節で先行研究について触れる．3 節で提案手法の定式化を行い，4 節で音素認識実験によりその性能を示す．最後に 5 節でまとめる．

2 関連研究

Ossama らは話者コード (Speaker code) を用いた DNN の適応手法を提案している．概要を Fig. 1 に示す．予め学習された話者非依存の DNN に対して，各層と話者コードと呼ばれる話者の特徴を表現するベクトルを連結する．これにより，層 l の出力 $\boldsymbol{O}^{(l)}$ は，$\boldsymbol{S}_c$ を話者 c の話者コードベクトルとすると，

$$\boldsymbol{O}^{(l)} = \boldsymbol{\sigma}(\boldsymbol{W}^{(l)}\boldsymbol{O}^{(l-1)} + \boldsymbol{b}^{(l)} + \boldsymbol{B}^{(l)}\boldsymbol{S}_c) \qquad (1)$$

として，計算する．なお，ここで，$\boldsymbol{W}^{(l)}$ は l と $l-1$ 層間の線形変換パラメータであり，$\boldsymbol{B}^{(l)}$ は l 層への話者適応の重みパラメータである．また，$\boldsymbol{\sigma}$ は vector sigmoid 関数である．学習は，予め学習した DNN のパラメータ $\boldsymbol{W}^{(l)}$ を固定した back propagation によって行うが，話者コードと各層間の連結行列 $\boldsymbol{B}^{(l)}$ は全ての話者の学習データを用いて学習するのに対し，話者コード $\boldsymbol{S}_c$ は話者毎に切り替えて学習を行う．なお，話者コードの back propagation はクロスエントロピー最小化などの目的関数を E とした場合，

$$\frac{\partial E}{\partial \boldsymbol{S}_{c,k}} = \frac{1}{L}\sum_l \sum_j \frac{\partial E}{\partial \boldsymbol{O}_j^{(l)}}(1-\boldsymbol{O}_j^{(l)})\boldsymbol{O}_j^{(l)}\boldsymbol{B}_{kj}^{(l)} \qquad (2)$$

として計算することができる．これにより，話者依存の情報が話者コードに集約され，話者非依存の変換が線形変換行列 $\boldsymbol{B}^{(l)}$ に集約される．

モデル適応の際は，対応する話者の話者コードを入力することで効率的にモデルパラメータを制御す

*Speaker-normalized training of DNN-based acoustic models by simultaneous estimation of weight parameters and speaker codes, by Y.Kashiwagi D.Saito, N.Minematsu, and K.Hirose (The University of Tokyo)

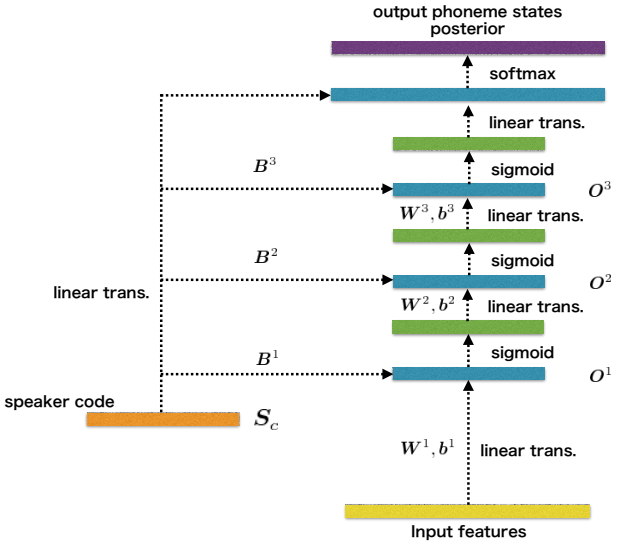

Fig. 1　Direct adaptation of DNNs based on peaker code.

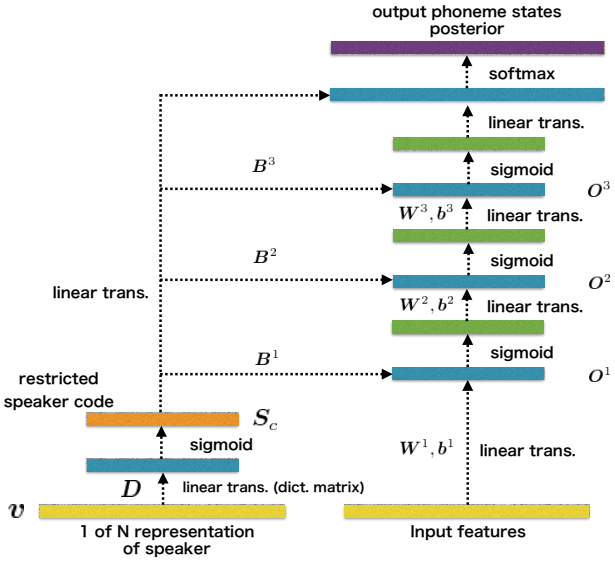

Fig. 2　Direct adaptation of DNNs based on restricted peaker code.

ることができる．しかし，ここで問題となるのが，音声認識時には，適応する話者の話者コードが観測できない点である．これは，教師あり適応を想定すれば，適応データを用いて back propagation により学習時と同様の方法で適応話者の話者コードを推定することができる．

このモデルは，ベースの話者非依存 DNN のモデルパラメータ $\boldsymbol{W}, \boldsymbol{b}$，話者適応サブネットのモデルパラメータ $\boldsymbol{B}$，そして話者コード $\boldsymbol{S}_c$ の 3 種類のパラメータを学習する必要がある．話者コード $\boldsymbol{S}^{(c)}$ と話者適応サブネットのモデルパラメータ $\boldsymbol{B}^{(l)}$ は同時に学習するが，話者非依存の DNN のモデルパラメータはバイアスを除いて固定するため，各層間の線形変換行列の学習時に話者依存の情報の分離が不十分となる恐れがある．そのため，効率的に話者依存の情報を話者適応サブネットの学習に利用できているとは言えない．

3　制約付き話者コードを用いた話者正規化学習

3.1　話者依存/非依存パラメータの同時推定

話者コードを用いたモデル適応はバイアス成分に限定して話者の変動を表現することで，各層毎に話者適応を行うことができる．これはニューラルネットの各層が異なるドメインの特徴量における変換を行っていると考えると，特徴量の異なるドメインにおいて適応を行うことに相当し，複雑な表現が可能となる．しかし話者コードを用いた適応は、ベースとなるニューラルネットを予め学習するため，サブネットの学習時に話者に依存する情報が、効果的に伝搬しない．そこで，本提案手法は話者コード型の適応手法をベースとし，話者非依存のネットワークと話者依存ネットワークの同時推定学習を行う．これにより，話者非依存の情報で話者非依存サブネットを適切に学習する話者正規化学習を実現することが可能となる．

また、本提案手法に付随した利点として学習時にモデル切り替えが不要である点がある．話者毎に特定層のパラメータを全て切り替えるタイプの正規化学習は，入力セグメントのランダマイズが困難であるという問題がある．ランダムに並び替えた各セグメント毎にネットワーク構造を変更しなければならず，これは計算コストが非常に高い．高速化のためミニバッチに同一話者のデータを集約する等の手段が考えられるが，逐次更新を行う場合学習データの偏りは防ぐことができない．これは話者非依存部分のサブネットの学習にとって望ましいとは言えない．それに対して，本提案手法は話者コードをベースとした各層におけるバイアスの制御にのみ着目することで，効果的に話者依存と非依存のサブネットを学習することが可能となる．

3.2　ネットワーク構造

中間層が 3 層の場合のネットワークの構造を Fig. 2 に示す．学習データ中に存在する話者数が N 人の場合，各ビンを各話者に対応させた 1 of N ベクトル $\boldsymbol{v} = [v_1, \ldots, v_n]^\top$ を考えると

$$\boldsymbol{S}_c \quad = \quad \boldsymbol{\sigma}(\boldsymbol{D}\boldsymbol{v}) \begin{cases} v_n = 1 \ (n = c) \\ v_n = 0 \ (n \neq c) \end{cases} \qquad (3)$$

とする線形結合層 $\boldsymbol{D}$ の sigmoid 出力を話者コードとして定義する．以下 $\boldsymbol{D}$ を辞書行列と呼ぶ．ノード数を減らし，かつ sigmoid 出力とすることで，話者コードをボトルネック層として自然に解釈することが可能

となる．これはつまり，BP 時の話者コードが閉区間 $[0,1]$ の値のみを取るという制約を設けることに相当する．話者コードが全ての層との連結を持つことを考慮すると，それぞれの層が持つ意味が異なるため，スケーリングの効果は各層との連結ベクトルである $\boldsymbol{B}^{(l)}$ に集約することを目的としている．

そして，各中間層の出力 $\boldsymbol{O}_{1,2,\ldots,l}$ を

$$\boldsymbol{O}^{(l)} = \sigma(\boldsymbol{A}^{(l)}\boldsymbol{O}^{(l-1)} + \boldsymbol{b}^{(l)} + \boldsymbol{B}^{(l)}\boldsymbol{S}_c) \quad (4)$$

とする．これによりバイアス成分をグローバルな成分と話者に依存する成分に分離してモデル化することが可能となる．

学習時は，入力特徴に加え各話者を 1 of N 表現で表したベクトルを入力とし，back propagation により学習する．これは，学習話者毎に話者コードを切り替えて学習していることに相当する．ただし，辞書行列によりバイアスを制御するため，先行研究と異なり，明示的にモデルパラメータや話者コードを入れ替える必要がない．そのため，このモデル構造を通常の BP と同様のアルゴリズムで学習することにより，各セグメント単位でモデルパラメータの切り替えに相当する効果が得られるため，学習データの偏りを回避することが可能となる．

認識時は，認識する対象話者の話者コードを観測することができない．そこで，全話者のデータを用いてグローバルな話者コードを擬似的に与えて認識に用いる．これは，

$$\boldsymbol{O}^{(l)} = \boldsymbol{\sigma}(\boldsymbol{W}^{(l)}\boldsymbol{O}^{(l-1)} + \boldsymbol{b}^{(l)} + \boldsymbol{B}^{(l)}\boldsymbol{S}_{global}) \quad (5)$$

として各隠れ層の出力を計算すれば良いことに相当する．これは，当然ながら

$$\hat{\boldsymbol{b}}^{(l)} = \boldsymbol{b}^{(l)} + \boldsymbol{B}^{(l)}\boldsymbol{S}_{global} \quad (6)$$

$$\boldsymbol{O}^{(l)} = \boldsymbol{\sigma}(\boldsymbol{A}^{(l)}\boldsymbol{O}^{(l-1)} + \hat{\boldsymbol{b}}^{(l)}) \quad (7)$$

としてバイアス項を纏めることにより，通常のニューラルネットの形に変形することができ，認識時は既存の枠組みを流用することが可能となる．

3.3 話者適応

話者適応は，新しく入ってきた話者に対して，話者コードを推定することに他ならない．適応データを用いて話者コードを back propagation により推定する．この際，話者コードの初期値には正規化学習を行った際に推定したグローバルな話者コードを用いる．

4 実験

4.1 実験条件

提案手法により学習を行った DNN 音響モデルの評価を TIMIT データベースを用いた連続音素認識実験により行った．音響モデルの学習には各話者 8 発話，全 462 話者の計 3696 発話を用い，評価セットは TIMIT のコアテストセットである 24 話者計 192 発話を用いた．また，デコード時の音響モデルの重みの調整として 50 話者 400 発話の開発セットを用いる．なお，学習セットと評価セット，開発セットで話者の重複はない．

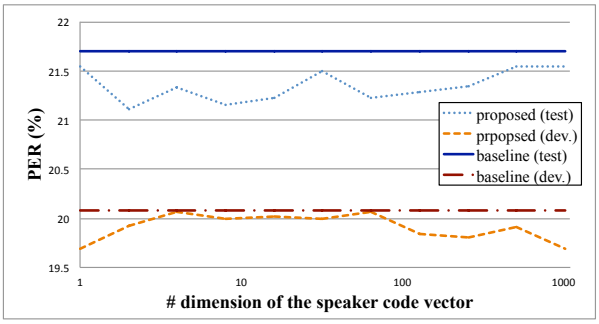

Fig. 3 The phone erorr rate of restricted speaker code based speaker normalized DNN acoustic model (phone error rate in %) in TIMIT dataset.

DNN モデルの学習に用いる音素状態ラベルのアライメントには fMLLR を行ったトライフォンモデルを用い，音素状態は計 1951 状態である．DNN は隠れ層 6 層，各層 2048 ノード，活性化関数には sigmoid 関数を用いた．入力特徴量として MFCC に対して fMLLR を行ったものを用いている．なお，DNN の学習の際は開発セットと学習データで話者の重複が必要なため，学習データをフレーム単位で 5%を DNN 学習用の開発セットとして用いている．事前学習には stacked denoising autoencoder を用いており，また，dropout は行っていない．さらに，学習の際のハイパーパラメータのチューニングに起因する影響を軽減するため，ラーニングレート，エポック数等の設定は通常の DNN の学習の際の結果に対して最良値を採用した．

4.2 話者正規化 DNN の性能評価

提案手法により話者正規化学習を行った DNN の性能を比較する．話者コードの次元数と音素認識誤り率の対応をプロットした実験結果を Fig. 3 に示す．なお，baseline は，通常の DNN 音響モデルである．開発セットと評価セット共に特に話者コードの次元数が小さい場合，通常の DNN 音響モデルと比較して提案手法の方が良い結果が得られている．これにより，提案手法においては，話者に由来する各層のバイアスの変動が話者コード側のネットワークによって適切に吸収されて学習されていると考えられる．

4.3 話者正規化 DNN を用いた話者適応性能の評価

本提案手法により話者正規化を行った DNN をベースとした話者適応の結果を Table 1 に示す．適応デー

Table 1　The phone erorr rate of restricted speaker code based speaker normalized DNN acoustic model (phone error rate in %) in TIMIT dataset.

	PER (%)
baseline	21.50
+ BP adaptation	21.42
+ SC adaptation	21.21
SC-based normalized DNN	21.11
+ adaptation	20.98

Table 2　The phone erorr rate of restricted speaker code based speaker normalized DNN acoustic model compared with other methods (phone error rate in %) in TIMIT dataset.

	PER (%)
Monophone HMM	34.30
Triphone HMM	30.42
Triphone HMM (SAT)	25.47
Subsupeace GMM	23.06
Basic DNN/HMM	21.50
Proposed	21.11
Proposed + Subspace GMM	20.64

タには各話者 2 発話を用いた．なお，この 2 発話は TIMIT 中の “sa”ラベルが振ってあるものであり，各話者について共通の文章となっている．適応の際のラーニングレート，バッチサイズ，エポック数等のハイパーパラメータ群は，各手法において最良値を採用した．baseline は通常の DNN での認識結果であり，BP adaptation は適応データを用いて back propagation により追加学習を行ったものである．また，SC adaptation は先行研究である Ossama らの手法により適応を行ったものである．なお，先行研究，提案手法共に話者コードの次元数は 2 とした．

提案手法である SC-based normalized DNN は話者コードにグローバルな値を入れた場合でも既に他手法の適応後と同等の誤り率を得ることが出来ている．さらに，話者適応を行うことで，僅かではあるが誤り率を削減することが可能となる．なお，全体的にモデル適応の効果が低いが，これは入力特徴量に対して fMLLR を行っているため，あらかじめ話者によるばらつきが低減されているためと考えられる．

4.4　他手法との比較

提案手法により正規化学習を行った音響モデルと様々な音響モデルとの性能比較を Table 2 に示す．最終的に subspace GMM と提案手法の system combination により 20.64 %の音素認識誤り率を得ることができた．

5　おわりに

本稿では，制約付き話者コードを用いた DNN 音響モデルの話者正規化学習を提案した．提案法は 1-of-N 表現のコードを入力として用いているため，例えば雑音環境下音声認識における雑音環境のラベルの利用や，さらに今回の話者ラベルとの複合等により，様々な応用が期待される．そして，直感的に音素情報と直交すると考えられる，話者性や環境の違いなどのラベルを用いて DNN の学習をより効果的に行うことが可能であることを示唆している．

参考文献

[1] Abdel-rahman Mohamed, George E. Dahl, and Geoffrey Hinton: “Acoustic modeling using deep belief networks,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 1, pp. 14–22, 2012.

[2] Frank Seide, Gang Li, Xie Chen, and Dong Yu: “Feature engineering in context-dependent deep neural networks for conversational speech transcription,” Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, pp. 24–29, 2011.

[3] Kaisheng Yao, Dong Yu, Frank Seide, Hang Su, Li Deng, and Yifan Gong: “Adaptation of Context-Dependent Deep Neural Networks for Automatic Speech Recognition,” ACL Workshop on Spoken Language Technology, pp. 366–369, 2012.

[4] Hank Liao: “Speaker adaptation of context dependent deep neural networks” Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 7947–7951, 2013.

[5] George Saon, Hagen Soltau, David Nahamoo and Michael Picheny: “Speaker adaptation of neural network acoustic models using i-vectors” Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, pp. 55–59, 2013.

[6] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and Frank Seide: “KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition” Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 7893–7897, 2013.

[7] Shaofei Xue, Ossama Abdel-Hamid, Hui Jiang, and Lirong Dai: “Direct adaptation of hybrid DNN/HMM model for fast speaker adaptation in LVCSR based on speaker code” Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp.6389–6393, 2014.

[8] Jian Xue, Jinyu Li, Dong Yu, Mike Seltzer, and Yifan Gong: “Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network” Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 6409–6413, 2014.

[9] Takuya Yoshioka, Anton Ragni, Mark J. F. Gales: “Investigation of unsupervised adaptation of DNN acoustic models with filter bank input,” Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 6394–6398, 2014.

[10] Tsubasa Ochiai, Shigeki Matsuda, Xugang Lu, Chiori Hor, and Shigeru Katagiri: “Speaker adaptive training using deep neural networks.” Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 6349–6353, 2014.