社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

話者コードに基づく話者正規化学習を利用したニューラルネット音響モデルの適応

柏木　陽佑†　齋藤　大輔†　峯松　信明†　広瀬　啓吉†

† 東京大学　〒 113–8656 東京都文京区本郷 7–3–1
E-mail: †{kashiwagi,dsk_saito,mine,hirose}@gavo.t.u-tokyo.ac.jp

あらまし　近年，自動音声認識において，その高い認識性能により，deep neural network (DNN) を用いた音響モデルが台頭している．しかし，一般に，DNN 音響モデルは不特定話者のデータで学習されるため，特徴量の分布が実際の特定話者の分布と大きく異なる．したがって，さらなる認識性能の向上のため，DNN 音響モデルの話者適応が注目されている．この内の一つとして，話者コードを用いた DNN の話者適応手法が提案されている．この方法では，話者依存と非依存のネットワークパラメータを別々に学習しており，話者依存／非依存の情報を明確に分離できているとは言えない．一方，話者依存／非依存パラメータの同時推定手法として話者依存層の切り替えによる話者正規化学習も提案されているが，back propagation において話者依存層を切り替える必要があり，学習コストが非常に大きい．そこで，本稿では話者適応の性能向上を目的とした，話者コードをベースとした話者正規化学習と，これを用いた話者適応手法を提案する．話者コードにより話者の情報を制御することで学習時に話者依存の情報と非依存の情報を分け，話者依存／非依存パラメータを同時に学習することにより効果的なネットワークの学習が可能となる．また，話者コードをベースとすることにより，各層のバイアスパラメータを話者コードにより制御することができる．この結果，層のパラメータを切り替える必要がなく，back propagation 時の学習コストの増加を抑えることが可能となる．提案手法の性能を TIMIT データベースを用いた連続音素認識により評価を行い，5.7%の音素認識誤りの削減を実現した．
キーワード　音声認識，音響モデル，話者適応，話者正規化学習，deep neural network

Speaker adaptation using speaker-normalized DNN based on speaker codes

Yosuke KASHIWAGI†, Daisuke SAITO†, Nobuaki MINEMATSU†, and Keikichi HIROSE†

† The University of Tokyo, 7–3–1, Hongo, Bunkyo-ku, Tokyo, 113–8656 Japan
E-mail: †{kashiwagi,dsk_saito,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract　Recently, deep neural network (DNN) becomes one of the main streams of acoustic modeling for automatic speech recognition. Further, speaker adaptation techniques have been tested for DNN-based speech recognition, including one based on a framework of bias adaptation using speaker codes. This paper introduces speaker-normalized training to this framework and experimentally shows its effectiveness. In the conventional method using speaker codes, two kinds of networks of speaker-independent (SI) DNNs and subnetworks for speaker adaptation were trained sequentially. We expect that, by training the SI networks and the subnetworks simultaneously, this method can be tuned so that it can handle both SI information and speaker-dependent (SD) information more adequately. Further, different from the conventional method, the speaker code vector is generated through networks from a 1-of-N speaker representation. This will reduce the training cost of the SI models and the subnetworks and avoid the over-fitting problem. Experimental evaluations using the TIMIT database demonstrate that our proposed training method can reduce the phoneme error rate by 5.7% relative.
Key words　automatic speech recognition, acoustic model, speaker adaptation, speaker normalized training, deep neural network

1. はじめに

自動音声認識における音響モデルの主流は，GMM ベースからニューラルネットベースへと移り変わったと言っても過言ではない．この背景には，deep learning の登場による，deep neural network (DNN) の安定した高い性能の実現の寄与するところが大きい [1]．しかし，DNN ベースの音響モデルは，その高い性能と引き換えに，各パラメータがどのような意味を持つのかを曖昧になった．そのため，GMM ベースの手法が利用できないケースが多い．その一つとして話者適応技術が挙げられる．GMM は生成モデルであり，各混合の分布のパラメータが明示的にではないにしろ，音素などの音響的事象と比較的似た意味を持っていたため，モデルパラメータの制御が容易であった．しかし，DNN は識別モデルであり，かつ非線形変換を多層に渡って積み上げるため，特徴量空間の分割が複雑でありパラメータの制御が困難である．

DNN 音響モデルは特徴量抽出と音素状態識別の両方の機能を内包すると考えられる．これは，従来 GMM ベース音響モデルで採用されていた MFCC などの特徴量よりも，フィルタバンク出力等をそのまま用いた方が高い性能を示すことからもわかる．特徴量抽出は，より認識に適した情報を抽出することであり，音声認識においては話者によるばらつきを排除することに他ならない．そのため，DNN は話者正規化の機能を暗に持つとも考えられる．一方，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とも考えられる [2–10].

話者適応技術の一つとして，Xue 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，あらかじめ，通常の話者非依存の DNN を学習し，各層のバイアス項の制御を，サブネットワークを接続して話者依存の話者コードにより行なう．なお，認識時は，入力話者に対応する話者コードが不明なため，話者依存のパラメータを固定し，back propagation により話者コードを推定することで話者適応を実現する．しかし，この手法は話者非依存の DNN と話者依存のパラメータを別々に学習しており，これにより明確に話者依存／非依存の情報を分離できているとは言えず，話者適応に適した話者正規化が実現できているわけではない．

一方，話者依存／非依存のパラメータの同時推定法として，落合らの話者正規化学習法がある [10]．これは，層ごとに話者依存／非依存を決め，話者依存層を学習データの話者毎に切り替えて学習する．この枠組みを用いて，話者依存／非依存のパラメータの同時推定を行うことでそれぞれの情報を分離することが可能となる．こちらも，認識時は入力話者に対応する話者依存層のパラメータを back propagation により推定し話者適応を行う．しかし，この正規化法は back propagation 時に話者依存層のパラメータを切り替える必要があり，学習コストの削減のためミニバッチ学習を GPU 上で行うことの多い DNN の学習において，このコスト増大は致命的である．

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 の同時推定を行い，かつ，効率的な学習の実現を目的として，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，話者コードはその性質上，学習時は入力特徴量と同様の扱いが可能となる．これにより，各層のパラメータを学習中に切り替える必要がないため，back propagation 中に特別な操作が必要無い．そのため，学習の際のコスト増加を抑えることが可能となる．

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

2.1 話者コードを用いたモデル適応

Xue 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 [7]．概要を図 1 に示す．この手法では，まずあらかじめ通常の DNN の学習を行い，話者非依存のモデルを構築する．その後，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サブネットワークを連結する．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，

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

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．

サブネットワークの学習は，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 $\boldsymbol{S}_c$ は話者毎に切り替えて学習を行う．これにより，サブネットワークのパラメータを通して各層のバイアスパラメータを話者コードにより制御するネットワークモデルを構築することが可能となる．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，

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

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ることができる．なお，入力話者は未知話者であるため，教師あり適応を想定した場合，少量の適応データを用いて back propagation により入力話者の話者コードを推定する．

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．しかし，話者コード $\boldsymbol{S}^{(c)}$ と話者適応サブネットのモデルパラ

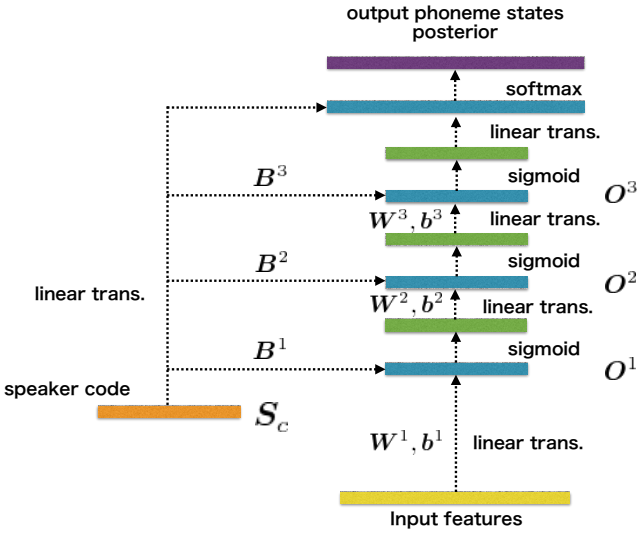

図 1　Direct adaptation of DNNs based on speaker code.

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．

2.2　話者依存層の切り替えによる話者正規化学習

落合らは，話者依存層を切り替えることによる話者正規化学習を提案している．図 2 にモデルの概要を示す．多層ニューラルネットにおいて話者依存／非依存層を設定し，学習時は話者毎に話者依存層のパラメータを切り替える．この枠組みを用いて back propagation により学習することで，話者依存／非依存のパラメータを同時に学習することが可能となる．認識時は，話者非依存層のパラメータを固定し，適応データを用いて back propagation により話者依存層のパラメータを再推定する．これにより話者正規化学習されたニューラルネットを用いた話者適応を実現することができる．なお，この際の話者依存層パラメータの初期値は，話者非依存層のパラメータを固定して全ての学習データで再学習を行ったものを用いている．

しかし，back propagation において話者依存層のパラメータを切り替えることは，非常にコストが大きい．DNN は学習コストの削減のため，学習時にミニバッチを用いて学習することが多い．そのため，各話者のデータ毎にミニバッチを構築することで，話者依存層の切り替えコストを削減することが可能となるが，学習データの偏りの問題が発生してしまう．また，話者依存層を多層に渡って設定した場合，過学習が発生してしまう恐れもある．

3.　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．また，各層ごとに扱うパラメータ数も少ないため，多層に渡ってモデル適応が可能となる．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，

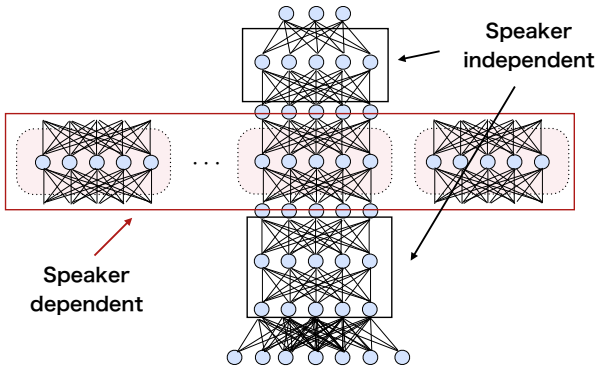

図 2　Speaker-normalized training using switching layers.

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．これは，あらかじめ学習する話者非依存のネットワーク自体がある程度の話者正規化能力を持つためである．

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，話者情報を制御する機能をサブネットワークに集約することができ，より効果的な話者正規化学習を実現することが可能となる．

また，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，back propagation を行う際に必要となる，データの偏りを防ぐための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った．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．また，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とられるが，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の枠組みにより正規化学習を行うことで，モデルパラメータの切り替えが不要となる．これは，話者コードを切り替えることがバイアスパラメータの切り替えに相当するためであり，ミニバッチ中でも各データに対応する話者インデクスさえ用意すれば，データのランダマイズは可能であり，通常の back propagation と同様の枠組みを利用できる．

中間層が 3 層の場合のネットワークの構造を図 3 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

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

とする線形結合層 $\boldsymbol{D}$ の sigmoid 出力を話者コードとして再定義する．以下 $\boldsymbol{D}$ を辞書行列と呼ぶ．sigmoid 出力とすることで，back propagation 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行列である $\boldsymbol{B}^{(l)}$ に集約することを目的としている．

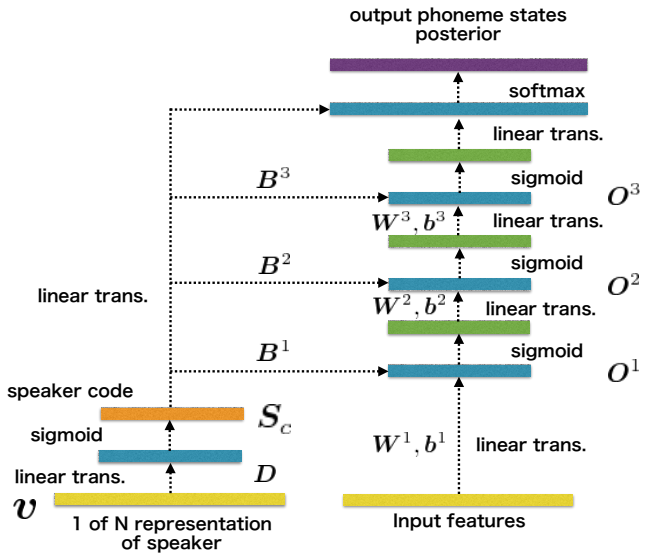

図 3　Direct adaptation of DNNs based on restricted speaker code.

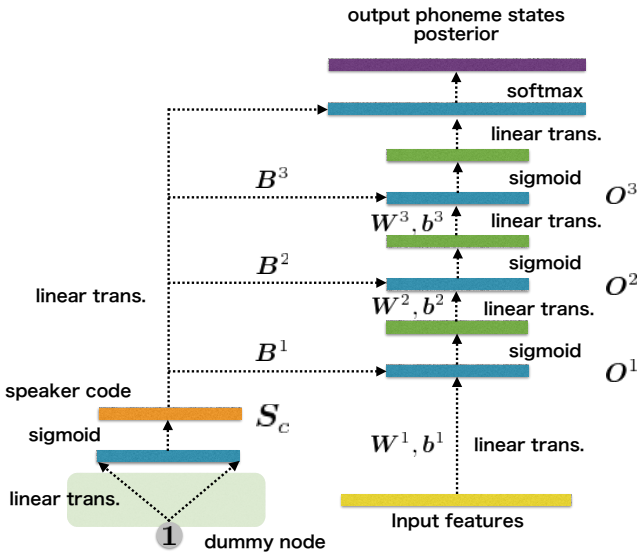

図 4　Estimation of speaker code using dummy node.

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

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

とする．この枠組みを用いて，全てのパラメータを同時に学習することで，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．

3.2　ネットワーク構造

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，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，

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

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．なお，グローバルな話者コードの推定はニューラルネットのパラメータを固定したまま back propagation により推定を行う．これは，図 4 のように話者コードをダミーのスカラーからの線形変換の出力を sigmoid 関数により $[0, 1]$ に正規化したモデルとして考えた場合，線形変換パラメータを推定することに相当する．そのため，back propagation の枠組みで学習することが可能である．なお，話者コードを与えた場合，当然ながら

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

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

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る．したがって，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

認識時に用いるグローバルな話者コードは実際の話者の話者コードとは異なるため，ミスマッチが生じてしまう．そこで，新しく入力された話者に対する話者コードを話者コードに入力することによりモデルの適応を実現する．しかし，学習時と異なり，未知話者に対する話者コードは観測できないため，少量の教師あり適応データを用い，話者コードをモデルパラメータを固定した 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．

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，学習データの内，フレーム単位で 5%を DNN 学習用の開発セットとして用いた．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

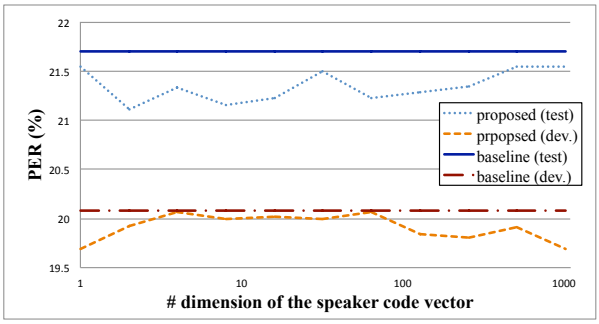

図 5　The phone error rate of DNN acoustic model with restricted speaker code based speaker normalization (phone error rate in %) in TIMIT dataset.

表 1　The phone error rate of DNN acoustic model with restricted speaker code based speaker normalization (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**

DNN の学習の際の結果に対して最良値を採用した．このため，多少ではあるが，提案手法において不利な条件となっている．

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．これは，提案手法においてグローバルな話者コード $\boldsymbol{S}_{global}$ を入力した場合に相当する．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図 5 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．なお，話者コードの次元数が小さい場合に良い性能が得られていることにより，話者コードがボトルネック層として機能していることも期待されるが，学習データ話者の直感的なクラスタリングが実現できている様子はなかった．

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表 1 に示す．これは，提案手法において適応データにより推定した入力話者の話者コード $\boldsymbol{S}_c$ を入力した場合に相当する．適応データには各話者 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．なお，適応時は開発セットの利用が不可能であるため，開発セットを用いずにデコード時の重みを決定した．そのため，図 5 とベースラインの値が異なる．これは，適応時は開発セットの利用が不可能であるためである．また，SC adaptation は先行研究である Xue 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．

表 2　The phone error rate of DNN acoustic model with restricted speaker code based speaker normalization 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**

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表 2 に示す．Monophone HMM と Triphone HMM では MFCC を，それ以外では MFCC+fMLLR を特徴量として用いている．また Proposed + Subspace GMM は提案手法と subspace GMM を system combination したものであり，combination の際の重みは開発セットを用いて決定した．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，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．提案法はサブネットワークを接続し，1-of-N 表現の話者コードを入力としてネットワークの話者性を制御する．この構造を用いて話者依存／非依存のパラメータを同時に学習することで，話者依存／非依存の情報を分離することができ，正規化学習が実現が可能となる．また，話者コードをベースとすることにより，back propagation による学習を行う際に，明示的にモデルパラメータを切り替える必要がないため，学習コストの増加を抑えることができる．

話者のインデックスではなく環境ラベルを用いることで環境正規化学習等も可能となる．さらに，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，話者ラベルとの複合等により，様々な応用が期待される．

文　献

[1] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton, "Acoustic modeling using deep belief networks," in *Audio, Speech, and Language Processing, IEEE Transactions on*. 2012, vol. 20, pp. 14–22, IEEE.

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

[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.," in *SLT*, 2012, pp. 366–369.

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

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

[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," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7893–7897.

[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," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6339–6343.

[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," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6409–6413.

[9] Takuya Yoshioka, Anton Ragni, and Mark JF Gales, "Investigation of unsupervised adaptation of dnn acoustic models with filter bank input," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6394–6398.

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