Vol.2016-SLP-112 No.1
2016/7/28

識別的推定法に基づく音声の構造的表象を制約として用いたニューラルネットワーク音響モデルの話者適応

柏木 陽佑[1,†1,a)]　齋藤 大輔[1,b)]　峯松 信明[1,c)]

概要：本稿では，音声の構造的表象を制約として用いたニューラルネット音響モデルの話者適応を提案する．分布間距離をニューラルネットワークにより推定する手法を用いることで，ニューラルネット音響モデルが想定する特徴量分布の構成する構造的表象を容易に計算することができる．話者適応の前後において音声の構造 的表象が大きく変化しないという仮定に基づき，話者適応の際の正則化項として音声の構造的表象を導入することで，11.2%のエラー削減が可能となった．

キーワード：自動音声認識，音響モデル，話者適応，音声の構造的表象，分布間距離

Speaker adaptation for deep neural network acoustic models based on discriminative estimation of structural constraints

Kashiwagi Yosuke[1,†1,a)]　Saito Daisuke[1,b)]　Minematsu Nobuaki[1,c)]

1. はじめに

自動音声認識において，入力話者の違いは認識率を低下させる大きな要因となり，これに対処するために従来より話者適応等の話者の違いに対して頑健性を確保する手法が多く研究されてきた．この多くはガウス分布をベースとした生成モデルに対するアプローチ，もしくは特徴量ドメインにおける正規化手法であった [1, 2]．一方近年，音声認識はニューラルネットワーク，特に Deep Neural Network (DNN) の発展に伴い大きな変革が生じた．音声認識における音響モデルも DNN をベースとしたものが主流となり，より高い認識性能の実現が可能となった．しかし，DNN はそのモデル構造の複雑さからパラメータと実際の物理的意味との対応が非常に取りづらいという特徴がある．これは，DNN の高い識別性を実現する利点でもあるが，モデルの制御が困難という欠点でもある．そのため，DNN 音響モデルに対する話者適応技術は未だ発展途上であり，重要な課題であると言える．

DNN 音響モデルに対する話者適応技術として，適応データを用いたモデルの再学習がある．これは，大量の不特定話者のデータにより学習した不特定話者音響モデルのパラメータを初期値として，適応データを用いて再学習をするという非常に単純なアプローチであるが，多くの研究で認識性能の向上が報告されており効果的であると言える [3]．

しかし，ガウス混合モデルをベースとするモデルに対しての話者適応と異なり，DNN 音響モデルは分布の形状を陽に仮定しないため，その話者適応は容易に過学習してしまう．そのため，再学習時に事前分布や Kullback–Leibler divergence (KL 距離) を用いた正則化手法が提案されている [3,4]．しかし，これらはあくまで不特定話者音響モデルのパラメータから大きく変化しないという制約であり，その背後にある音声学的な知見にまで言及しているとは言い

[1] 東京大学
The Uniersity of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan
[†1] 現在，ソニー株式会社
Presently with SONY
[a)] kashiwagi@gavo.t.u-tokyo.ac.jp
[b)] dsk_saito@gavo.t.u-tokyo.ac.jp
[c)] mine@gavo.t.u-tokyo.ac.jp

難い．これは，話者によって変わる音響量と，話者によっても変わらない音響量との区別が不十分なために生じる．

そこで，本研究では音声の構造的表象を制約として用いたDNN音響モデルの話者適応について提案し，教師なし適応による実験によりその有効性を検討する．話者の性別や年齢の違いは，ケプストラム空間上においてアフィン変換として表現可能であることが知られている [5]．音声の構造的表象は各音響イベント間の分布間距離を利用した特徴量であるが，f-divergenceに代表される分布間距離はアフィン変換に対して不変である．そのため，音声の構造的表象は話者性の違いに対して頑健であると言える [6]．この性質を利用して，音声の構造的表象を制約としてモデルパラメータの更新を行う手法も提案されている [7, 8]．本研究では，音響モデルの再学習時に音声の構造的表象を同時に計算するには，不特定話者モデルの持つ音声の構造的表象と，適応後の話者依存モデルの持つ音声の構造的表象が大きく変化しないという制約を正則化として導入する．ただし，音響モデルの学習時に音声の構造的表象を計算することは，各音響イベントのガウス分布のパラメータを計算する必要がある．そのため，DNNを用いてガウス分布のパラメータを経由せずに識別的に分布間距離を計算する手法 [9] を導入することで，効率的にパラメータの更新を行う．

2. DNN音響モデルと話者適応

2.1 DNN音響モデル

音声は時系列データであるため，時系列情報は隠れマルコフモデル（Hidden Markov Model; HMM）により，セグメント単位ではDNNなどのニューラルネットワークによってモデル化することが一般的である [10]．音声認識におけるDNNの出力は音素状態事後確率ラベルであるため，DNNの学習はクロスエントロピー最小化基準のもと，誤差逆伝搬法により学習する．不特定話者のデータ集合を $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T, y_1, y_2, \ldots, y_T\}$ とすると，

$$\hat{\Theta}_g = \underset{\Theta}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} D(\boldsymbol{x}_t) \tag{1}$$

$$D(\boldsymbol{x}_t) = \left\{ -\sum_{q}^{Q} \tilde{p}(y_t = q|\boldsymbol{x}_t) \ln p(y_t = q|\boldsymbol{x}_t, \Theta) \right\} \tag{2}$$

ここで，$\boldsymbol{x}_t$ はデータインデクス t における入力特徴量ベクトル，y_t はインデクス t における出力ラベルであり，$p(y_t = q|\boldsymbol{x}_t)$ は音素状態インデクス q の事後確率に相当する．また，$\tilde{p}(y_t = q|\boldsymbol{x}_t)$ は正解ラベル（事後確率）である．

2.2 再学習によるDNN音響モデルの話者適応

DNNは多層の複雑な構造を持つために，各パラメータの意味づけが直感的でない．そのため，ガウス混合モデルのように話者性に対応するパラメータの制御が困難である．そこで，あらかじめ大量の不特定話者のデータにより学習された不特定話者モデルのパラメータを初期値として，適応データを利用した再学習により全体のモデルパラメータを更新する手法が用いられることが多い．

未知話者の適応データ $\{\boldsymbol{x}'_1, \boldsymbol{x}'_2, \ldots, \boldsymbol{x}'_{T'}\}$ が得られた際に，この適応データを用いてDNNを再学習することができる．まず，適応データを不特定話者モデルを用いてデコードすることにより，フレーム毎の音素状態事後確率ラベル $\{p_d(y'_1 = q|\boldsymbol{x}'_1), p_d(y'_2 = q|\boldsymbol{x}'_2), \ldots, p_d(y'_{T'} = q|\boldsymbol{x}'_{T'})\}$ を用意する．その後，不特定話者モデルのパラメータ Θ_g を初期値として，再学習によりパラメータを更新する．再学習時の目的関数 (4) は不特定話者モデルの学習時と同じであるが，誤差逆伝搬法の際に学習時のエポック数や学習係数等の制御を行うことで過適応を抑える．

$$\hat{\Theta}_{adapt} = \underset{\Theta}{\operatorname{argmin}} \frac{1}{T'} \sum_{t'=1}^{T'} D_{adapt}(\boldsymbol{x}_{t'}) \tag{3}$$

$$D_{adapt}(\boldsymbol{x}_{t'}) = \left\{ -\sum_{q}^{Q} p_d(y'_{t'} = q|\boldsymbol{x}'_{t'}, \Theta_g) \ln p(y_{t'} = q|\boldsymbol{x}'_{t'}, \Theta) \right\} \tag{4}$$

なお，適応データの音素状態事後確率ラベルを用意する際に，正解テキストを用いて強制アライメントを行うことで教師あり適応，正解テキストを用いずにデコード結果を用いることで教師なし適応となる．

2.3 分布間距離を制約として用いた話者適応

再学習法の学習時の目的関数を変更する手法としてKL距離を用いた正則化手法が提案されている [4]．再学習によるDNNの話者適応は一般にエポック数や学習係数の制御によってモデルの過適応を抑えるが，適応データが少量の場合，もしくは教師なし適応のような適応データの音素状態事後確率ラベルが不安定である場合，学習係数等の制御では不十分である．これはDNNの持つ高い表現力により，適応データの偏りなどによって音素状態を識別するモデルの性質が簡単に失われてしまうためである．そこで，KL距離を利用した正則化をかけることで，少量の適応データでも初期モデルからモデルパラメータが大きく外れないような制約をかけて適応を行う．

再学習は，あらかじめ学習された不特定話者モデルのパラメータ Θ_g を初期値として，目的関数 (6) を用いてパラメータの更新を行う．

$$\hat{\Theta}_{adapt} = \underset{\Theta}{\operatorname{argmin}} \frac{1}{T'} \sum_{t'=1}^{T'} D_{kl}(\boldsymbol{x}_{t'}) \tag{5}$$

$$D_{kl}(\boldsymbol{x}_{t'}) = \left\{ -\sum_{i}^{Q} \tilde{p}(y_{t'} = q_i|\boldsymbol{x}_{t'}) \ln p(y_{t'} = q_i|\boldsymbol{x}_{t'}, \Theta) \right\} \tag{6}$$

$$\tilde{p}(y_{t'} = q_i|\boldsymbol{x}_{t'}) = \rho\, p(y_{t'} = q_i|\boldsymbol{x}_{t'}, \Theta_g) + (1-\rho)\, p_d(y_{t'} = q_i|\boldsymbol{x}_{t'}, \Theta_g) \tag{7}$$

ここで，$p(y_{t'} = q_i|\boldsymbol{x}_{t'}, \Theta_g)$ は話者非依存モデルから計算される事後確率であり，ρ はその重みである．つまり，話者非依存モデルパラメータを事前分布とした正則化をかけつつ再学習を行うことに相当する．

3. 音声の構造的表象を制約として用いたニューラルネットワーク音響モデルの適応

3.1 音声の構造的表象

音声の構造的表象は，各音響イベント間の分布間距離を要素としてもつ特徴量表現である．音響イベントを N 個の音素状態とすると，音声の構造的表象は $N \times (N-1)/2$ の次元数を持つ [11].

音声の構造的表象を構成する分布間距離としては f-divergence の一種であるバタチャリヤ距離を用いることが一般的である．

$$BD(a,b) = -\ln \int \sqrt{p(\boldsymbol{x}|y=a)p(\boldsymbol{x}|y=b)}d\boldsymbol{x} \quad (8)$$

となる．ここで，a と b は音響イベントのラベル，本研究では音素状態ラベルである．各音響イベントの特徴量分布をガウス分布と仮定した場合のバタチャリヤ距離は，

$$BD(a,b) = \frac{1}{8}(\boldsymbol{\mu}^{(a)} - \boldsymbol{\mu}^{(b)})^\top \Sigma^{-1}(\boldsymbol{\mu}^{(a)} - \boldsymbol{\mu}^{(b)}) + \frac{1}{2}\ln\left(\frac{\det\Sigma}{\sqrt{\det\Sigma^{(a)}\det\Sigma^{(b)}}}\right) \quad (9)$$

$$\Sigma = \frac{\Sigma^{(a)} + \Sigma^{(b)}}{2} \quad (10)$$

となり，$\boldsymbol{\mu}^{()}$，$\Sigma^{()}$ は各音響イベントのガウス分布の平均と分散である．

各音素状態の分布をガウス分布と仮定した場合，f-divergence はあらゆるアフィン変換に対して不変である．ケプストラム空間において，話者の性別や年齢の違いはアフィン変換によってよく表現できることが知られている [5]．そのため，音声の構造的表象は話者の性別や年齢の違いに対して頑健な特徴量であると言える [6]．この性質を利用して，音声の構造的表象は音響モデルを学習する際の制約として用いられている [7]．しかし，真の特徴量分布はガウス分布に近いものの，複雑な形をしていると考えられるため，ガウス分布では十分とは言えない．

3.2 音声の構造的表象の識別的計算法

仮定した分布と真の分布との間でミスマッチがある場合，その分布を用いて計算された分布間距離も真に得たい分布間距離との間でミスマッチが生じてしまう．音声の構造的表象の識別的計算法は，この問題を回避する方法の一つである．本稿では，識別モデルとして DNN を想定する．入力特徴量から該当するフレームの音素状態ラベルを識別する DNN 音響モデルが学習されている場合，特徴量が与えられた場合の音素状態ラベルに対する事後確率 $p(y=a|\boldsymbol{x})$，$p(y=b|\boldsymbol{x})$ が DNN により直接計算することができる．

式 (8) にベイズ則を導入することによって，バタチャリヤ距離は，

$$BD(a,b) = -\ln \int \sqrt{p(\boldsymbol{x}|y=a)p(\boldsymbol{x}|y=b)}d\boldsymbol{x} \quad (11)$$

$$= -\ln \int \sqrt{\frac{p(y=a|\boldsymbol{x})p(\boldsymbol{x})}{p(y=a)}\frac{p(y=b|\boldsymbol{x})p(\boldsymbol{x})}{p(y=b)}}d\boldsymbol{x} \quad (12)$$

$$= -\ln \int p(\boldsymbol{x})\sqrt{p(y=a|\boldsymbol{x})p(y=b|\boldsymbol{x})}d\boldsymbol{x} + \tfrac{1}{2}\ln p(y=a) + \tfrac{1}{2}\ln p(y=b) \quad (13)$$

として計算することができる．これにより，生成モデルのパラメータを用いずに，分布間距離を計算することが可能となる．

3.3 目的関数への音声の構造的表象の導入

音声の構造的表象の考えを基にした場合，音響イベント間の分布間距離は話者によらず不変であることが望ましい．しかし，過適応が生じる際は，適応データの偏り等に起因する場合が多いため，適応後のモデルの想定する特徴量空間における音声の構造的表象が変化してしまうことが想定される．そこで，本提案手法では，音声の構造的表象を基にした正則化を導入することで，再学習による DNN 音響モデルの話者適応の際に過適応を抑える．再学習時の目的関数を，あらかじめ学習された不特定話者モデルのパラメータ Θ_g を初期値として

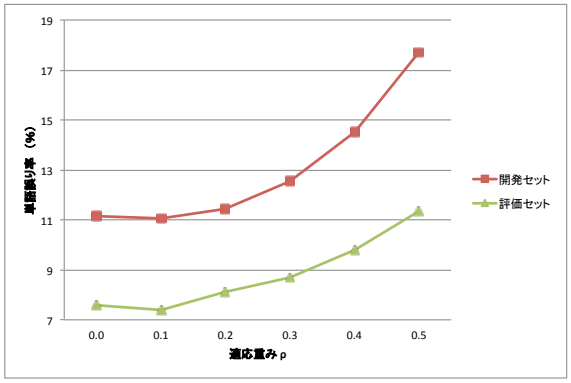
$$\hat{\Theta}_{adapt} = \underset{\Theta}{\operatorname{argmin}}\, \rho D_{st} + (1-\rho)\frac{1}{T'}\sum_{t'=1}^{T'} D_{adapt}(\boldsymbol{x}_t) \quad (14)$$

$$D_{adapt}(\boldsymbol{x}_{t'}) = \left\{ -\sum_i^Q p_d(y_{t'} = q_i|\boldsymbol{x}_{t'}, \Theta_g)\ln p(y_t = q_i|\boldsymbol{x}_{t'}) \right\} \quad (15)$$

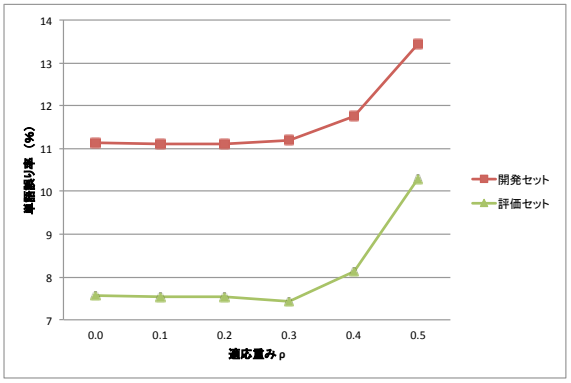
$$D_{st} = \frac{1}{\hat{Q}\times\hat{Q}}\left| \sum_i^{\hat{Q}}\sum_j^{\hat{Q}} BD^{SD}(q_i,q_j) - \sum_i^{\hat{Q}}\sum_j^{\hat{Q}} BD^{SI}(q_i,q_j) \right| \quad (16)$$

とする．ここで，$BD^{SD}(q_i,q_j)$ は適応データと適応後のモデルから得られるバタチャリヤ距離である．また，$BD^{SI}(q_i,q_j)$ は適応前の不特定話者モデルと，不特定話者のデータにより得られる，不特定話者モデルの持つ音響イベント間のバタチャリヤ距離である．$\hat{Q}$ は，音素状態を共有した後の集合であり，これは DNN 音響モデルの出力であるトライフォン音素状態は次元数が大きいため，モノフォンなどの数まで削減することを視野に入れている．さらに，無音や子音などは話者によって不変である，もしくは大差がないことが考えられるため，それらを除外したような集合等も考えられる．この目的関数を用いて誤差逆伝搬法による再学習によってパラメータの更新を行う．

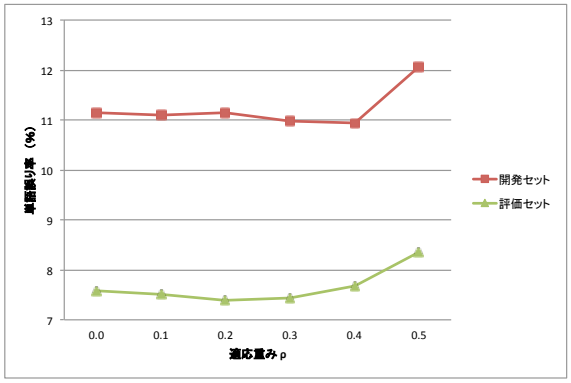
なお，バタチャリヤ距離の計算には，各音響イベントに対する確率密度関数が必要となるが，ニューラルネット



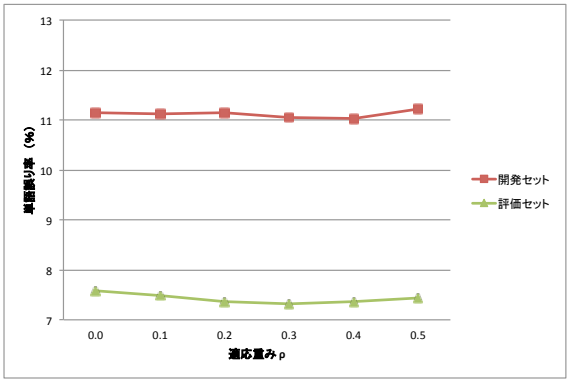
(a) 共有を行わず 1552 のトライフォン状態のまま計算した場合. (Prop. 1552)



(b) モノフォンラベルである 42 音素に共有した場合. (Prop. 42)



(c) モノフォンラベルから無音ラベルに相当するものを除いた 39 のラベルに共有した場合. (Prop. 39)



(d) 母音のみの 15 音素のモノフォンラベルに共有した場合. (Prop. 15)

図 1 適応重みを変化させた時の単語誤り率.

ワークの学習時に各音響イベントの確率密度関数を計算することは非効率である．そこで，分布間距離をニューラルネットワークを用いて識別的に計算する [9]．これにより，誤差逆伝搬法のミニバッチ単位で容易に音声の構造的表象を計算することが可能となる．それぞれのバタチャリヤ距離はニューラルネットワークを用いて

$$BD^{SD}(q_i,q_j) \approx -\ln\frac{1}{T'}\sum_{t'}\sqrt{p(y_{t'}=q_i|\boldsymbol{x}_{t'},\Theta)p(y_{t'}=q_j|\boldsymbol{x}_{t'},\Theta)} \quad (17)$$
$$+\frac{1}{2}\ln p(y=q_i)$$
$$+\frac{1}{2}\ln p(y=q_j)$$
$$=-\ln\frac{1}{T'}\sum_{t'}\sqrt{p(y_{t'}=q_i|\boldsymbol{x}_{t'},\Theta)p(y_{t'}=q_j|\boldsymbol{x}_{t'},\Theta)} \quad (18)$$
$$+const.$$

$$BD^{SI}(q_i,q_j) \approx -\ln\frac{1}{T'}\sum_{t'}\sqrt{p(y_{t'}=q_i|\boldsymbol{x}_{t'},\Theta_g)p(y_{t'}=q_j|\boldsymbol{x}_{t'},\Theta_g)} \quad (19)$$
$$+\frac{1}{2}\ln p(y=q_i)$$
$$+\frac{1}{2}\ln p(y=q_j)$$
$$=-\ln\frac{1}{T'}\sum_{t}\sqrt{p(y_{t'}=q_i|\boldsymbol{x}_{t'},\Theta_g)p(y_{t'}=q_j|\boldsymbol{x}_{t'},\Theta_g)} \quad (20)$$
$$+const.$$

として計算する．なお，それぞれの音響イベントの事前確率に対応する $p(y=q_i)$ は，適応前と後で変化しないと仮定することで定数項として扱う．

4. 実験

評価に用いるデータベースとして英語の大語彙音声認識用のデータベースである WSJ (01, 02) を用いた．学習データセットは 37,416 発話であり，評価データセットは 333 発話である．また，デコード時の音響モデルスコアと言語モデルスコアの重み決定のため，開発データセット 503 発話が用意されている．また，評価セットは各話者約 40 発話，開発セットは各話者 50 発話程度ある．

認識システムは KALDI に付属の WSJ データベースの認識スクリプトをベースとして用いた [12]．入力特徴量はメルフィルタバンク出力の 40 次元であり，音響モデルは DNN/HMM である．音響モデルのネットワークの入力は当該フレームとその前後 5 フレーム，計 11 フレームを入力とした．ネットワークの構造は中間層が 6 層であり，隠れ層は各層 1024 ノードである．また，活性化関数は maxout を採用した．出力は 1552 音素状態であり，学習データのラベルはあらかじめ学習した GMM/HMM のトライフォン音響モデルを用いた強制アライメントによってラベルを

表 1 認識実験結果：各状態共有の条件，適応重み ρ における単語誤り率．（開発セット/評価セット）

	適応なし	$\rho = 0.0$	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$
KL 正則化			11.20/7.51	11.20/7.53	11.25/7.55	11.32/7.57	11.34/7.57
Prop. 1552			11.06/7.39	11.42/8.12	12.55/8.67	14.50/9.80	17.69/11.34
Prop. 42	11.79/8.24	11.14/7.57	11.10/7.53	11.11/7.53	11.19/7.43	11.76/8.12	13.43/10.28
Prop. 39			11.09/7.50	11.15/7.39	10.98/7.43	10.94/7.69	12.06/8.36
Prop. 15			11.11/7.48	11.15/7.37	**11.05/7.32**	11.03/7.37	11.23/7.44

付与した．このトライフォン音響モデルは，特徴量として MFCC とその 1 次，2 次微分の 39 次元に平均分散正規化を行ったものを用いてモデル化した．なお，言語モデルは CMU の言語モデルを用いた．

話者適応は教師なし適応を想定する．認識するデータに対して不特定話者モデルにより認識を行った結果を用いて擬似正解ラベルを作成する．これを適応データとして再学習によりモデル適応を行った．なお， 適応時における音響モデルスコアと言語モデルスコアの重みも開発セットを用いて決定している．適応の際の音素状態数の共有条件として，1) 共有を行わず 1552 のトライフォン状態のまま計算した場合，2) モノフォンラベルである 42 音素に共有した場合，3) モノフォンラベルから無音ラベルに相当するものを除いた 39 のラベルに共有した場合，4) 母音のみの 15 音素のモノフォンラベルに共有した場合，の 4 パターンを検証した．それぞれの条件における単語誤り率を Fig. 1(a),1(b),1(c),1(d) に示す．また，それぞれの値の詳細は Table 1 に示す．$\rho = 0.0$ は正則化を行わずに再学習により適応した場合に相当し，Table 1 中の「適応なし」は適応を行っていない不特定話者モデルを用いて認識を行った結果である．また，Table 1 中の KL 正則化は KL 距離を制約として用いて再学習を行う先行研究 [4] である．

音素状態を共有せずに用いた場合（Fig. 1(a), Prop. 1552）では，適応重みが $\rho = 0.1$ と小さい場合には単語誤り率が減少するが，それ以降は急速に誤り率が増加してしまう．また，モノフォンラベルである 42 音素に共有した場合（Fig. 1(b), Prop. 42）では，適応重みの増加による誤り率の減少が抑えられており，音素状態を共有することの効果の有効性がわかる．さらに，モノフォンラベルから無音ラベルに相当するものを除いた 39 のラベルに共有した場合（Fig. 1(c), Prop. 39）では，僅かではあるが誤り率が減少しており，話者によって共通であると考えられる無音に相当するラベルを取り除いたことによって構造特徴が安定したことが寄与したと考えられる．母音のみの 15 音素のモノフォンラベルに共有した場合（Fig. 1(d), Prop. 15）では，適応重みを増加させた場合の単語誤り率の増加も抑えられており，適応重みが $\rho = 0.3$ の場合に最も良い単語誤り率が得られ，3.3%のエラー削減率が得られた．破擦音などの話者毎によって差が生じづらい子音の影響が取り除かれるため，さらに構造特徴が安定したと考えられる．

また，どの条件でも適応重みをより大きくした際に適応性能が程度の差はあれ低下することは共通して見られた．これは，構造的表象が時系列情報を吸収するため，適応重みを大きくした場合に逆にセグメント単位の音素状態識別性能が低下してしまうためだと考えられる．なお，先行研究である KL 距離を用いた正則化は本実験では適応性能の向上に寄与しなかった．これは，教師なし適応ではあるが比較的適応データの多い実験条件であるため，MAP 適応のように適応前と適応後のパラメータとの内挿を行う KL 距離の制約の効果が働かなかったためだと考えられる．

5. まとめ

本稿では，ニューラルネットワーク音響モデルの再学習を利用した話者適応の際に，音声の構造的表象を制約として用いる手法を提案した．ケプストラム空間上における各音響イベントの分布をガウス分布として仮定した場合，音声の構造的表象は話者の違いに対して頑健な特徴量であると言える．本研究では，音声の構造的表象のこの性質を音響モデルの話者適応に利用した．その際に，ニューラルネットワークの学習時に音声の構造的表象を効率的に計算するために，ニューラルネットワークを用いて識別的に分布間距離を計算する手法を導入した．また，音声の構造的表象を構築する際の音響イベントの単位を，もともとのトライフォン音素状態から共有することで，無音や子音などの影響を取り除くことを行った．これにより，適応の際の正則化として音声の構造的表象が有効に機能することが実験によって明らかになり，無音や子音を除外することで識別的計算法においても音声の構造的表象の安定性が上昇することがわかった．最終的に，正則化を行わない再学習による話者適応と比較して，母音のみの 15 音素のモノフォンラベルに共有した場合で 3.3%のエラー削減率が得られた．

謝辞

本研究の一部は科研費・基盤研究 (A) (26240022) 及び，特別研究員奨励費 (269167) の助成を受けた．

参考文献

[1] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[2] F. Seide, G. Li, X. Chen, and D. 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, pp. 24–29, 2011.

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

[4] D. Yu, K. Yao, H. Su, G. Li, and F. 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, pp. 7893–7897, 2013.

[5] M. Pitz, S. Molau, R. Schlüter, and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space." in *INTERSPEECH*, pp. 2653–2656, 2001.

[6] N. Minematsu, S. Asakawa, M. Suzuki, and Y. Qiao, "Speech structure and its application to robust speech processing," *New Generation Computing*, vol. 28, no. 3, pp. 299–319, 2010.

[7] Y. Qiao, M. Suzuki, N. Minematsu, and K. Hirose, "Structure-constrained distribution matching using quadratic programming and its application to pronunciation evaluation," in *Pattern Recognition (ACPR), 2011 First Asian Conference on*. IEEE, pp. 350–354, 2011.

[8] 内田秀継, 齋藤大輔, 峯松信明, "音声の構造的表象を用いた未観測調音運動の推定法の検討," **電子情報通信学会信学技報**, pp. 7–12, 2016.

[9] 柏木陽佑, 張聡穎, 齋藤大輔, 峯松信明, "識別的アプローチによる分布間距離推定の検討とその言語識別への応用," **日本音響学会秋期講演論文集**, pp. 31–34, 2015.

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

[11] N. Minematsu, "Yet another acoustic representation of speech sounds," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1, 2004, pp. 585–588.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, iEEE, 2011.