情報処理学会論文誌 Vol.58 No.9 1555–1564 (Sep. 2017)

音声認識のための回帰木に基づく複数の変換行列の重み付けによる特徴量空間の適応

金川 裕紀[1,a)] 太刀岡 勇気[1,b)] 渡部 晋治[2,c)] 石井 純[1]

受付日 2016年12月9日, 採録日 2017年6月6日

概要：音声認識では適応が重要である．特徴量空間での適応（fMLLR）は，特徴量ベクトル系列に単一の変換行列を乗算することで実現されるため，デコーディング処理とは独立な，特徴量に関する前処理として実装できる．このためガウス混合分布（GMM）と同様にディープ・ニューラルネットワーク（DNN）の音響モデルに対しても適用できる．一方でモデル空間の適応は，回帰木に基づき複数の変換行列を用いることで，単一の変換行列を用いる fMLLR よりも高い精度で適応が可能である．しかしこの手法には 2 つの課題がある．1 つ目は適応とデコードに同じ生成モデル（例：GMM）の音響モデルを共有しなければならず，DNN の音響モデルには適用できないこと，2 つ目は変換行列の数が多くなると，変換行列の推定が過学習しやすいことである．本論文では，1 パスの状態アラインメント情報を用いてフレームごとに対応する複数の変換行列を対応付け，それらを用いて重み付け線形和で表現される変換行列で特徴量変換を行う手法を提案する．さらに 2 つ目の課題に対し，構造的な事前確率の導入により変換行列を MAP 推定する，特徴量空間における構造的事後確率最大線形（fSMAPLR）を提案する．実験より，提案する fSMAPLR は fMLLR の性能を上回った．

キーワード：自動音声認識，適応，特徴量変換，ディープ・ニューラルネットワーク

Feature-space Adaptation with a Weighted Sum of Multiple Transformation Matrices Based on Regression Tree for Automatic Speech Recognition

Hiroki Kanagawa[1,a)] Yuuki Tachioka[1,b)] Shinji Watanabe[2,c)] Jun Ishii[1]

Received: December 9, 2016, Accepted: June 6, 2017

Abstract: In automatic speech recognition, an adaptation is important. Feature-space maximum-likelihood linear regression (fMLLR) transforms acoustic features to adapted ones by a multiplication operation with a single transformation matrix. This property realizes an efficient adaptation performed within a pre-precessing, which is independent of a decoding process, and this type of adaptation can be applied to deep neural network (DNN). On the other hand, model-space adaptations (i.e., CMLLR) improve the performance of fMLLR because it can use multiple transformation matrices based on a regression tree. However, there are two problems in the model-space adaptations: first, these types of adaptation cannot be applied to DNN because adaptation and decoding must share the same generative model, i.e., Gaussian mixture model (GMM). Second, transformation matrices tend to be over-estimated when the number of transformation matrices is large. This paper proposes to use multiple transformation matrices within a feature-space adaptation framework. The proposed method first estimates multiple transformation matrices in the GMM framework according to the first-pass decoding results and the alignments, and then takes a weighted sum of these matrices to obtain a single feature transformation matrix frame-by-frame. In addition, to address the second problem, we propose feature-space structural maximum a posteriori linear regression (fSMAPLR), which introduces hierarchal prior distributions to regularize the MAP estimation. Experimental results show that the proposed fSMAPLR outperformed fMLLR.

Keywords: automatic speech recognition, adaptation, feature-space transformation, deep neural network

1. はじめに

音声認識において適応手法は，音響モデルの学習データと評価データ間に，ミスマッチがある場合において有効である [1], [2]．適応手法は，モデル空間の適応と特徴量空間の適応の 2 手法に分類される．モデル空間の適応の代表的な手法である最尤線形回帰（maximum-likelihood linear regression：MLLR）[3], [4], [5] は，ガウス混合分布（Gaussian mixture model：GMM）に基づく音響モデルの枠組みで提案されてきた．MLLR では回帰木に基づきガウス分布ごとに複数の変換行列を推定し，これらの変換行列をガウス分布の平均ベクトルに乗算することで，単一の変換行列を用いる場合よりも精緻な適応が可能である．しかしながら変換行列数が多くなると過学習しやすくなる傾向がある．この問題を解決するため，回帰木の木構造を利用したベイズ的アプローチを導入する手法が提案されている [6], [7], [8]．構造的事後確率最大線形回帰（structural maximum a posteriori linear regression：SMAPLR）[7] は MLLR の拡張であり，回帰木の木構造に基づく事前分布を導入することで，変換行列を MAP（Maximum a posteriori）推定する．また音響モデルの平均パラメータのみを変換する MLLR を，平均・分散の両パラメータを変換する制約付き MLLR（constrained MLLR：CMLLR [9]）に拡張可能であるのと同様に，SMAPLR も制約付き SMAPLR（constrained SMAPLR：CSMAPLR [10]）に拡張できる．CSMAPLR は CMLLR や SMAPLR よりも，変換行列を安定的に推定できることが報告されている．しかし，複数の変換行列を用いたモデル空間の適応は GMM に特化した手法であることから，これらの適応手法を GMM 以外の音響モデルに適用することは困難である．

一方で特徴量空間の適応は，適応処理をデコード処理から分離できるため，いかなる音響モデルに対しても適用することができる．たとえば特徴量空間の MLLR（feature-space MLLR：fMLLR）は，単一の変換行列を特徴量ベクトルに乗算することにより実現される．他に fMLLR に関連する研究である fMAPLR [11] や fMAPLIN [12] では，適応データ量が少ないときにおいて変換行列の推定の頑健性を向上させることが示されている．またモデル空間の適応と fMLLR を組み合わせた手法 [13] も提案されているが，この方法では特徴量を直接変換することはできない．

このタイプの適応は，精緻なモデル化が可能であるが適応が困難であったディープ・ニューラルネットワーク（deep neural network：DNN）など任意の音響モデルに対しても容易に適用でき，fMLLR による変換後の特徴量を DNN の音響モデルに入力することの有効性が報告されている [14]．さらに，fMLLR のような線形変換層を DNN に組み込む手法として LIN（linear input network）[15], [16], [17] も提案されており，特徴量を変換することを目的とした，非線形活性化関数を用いない線形変換層を DNN の第 1 層として付加する．他にも LIN に関連し，線形変換を行う層を第 1 層以外に挿入する手法が提案されている [18], [19]．これらのアプローチは学習時のように大量データがある場合には DNN を容易に適応できるが，デコード時に線形変換層を頑健に推定するのは難しい．これは DNN のパラメータ数が多いことと，教師なし適応による誤ったアラインメントがパラメータの推定の精度を著しく低下させるからである．

特徴量空間の適応はこのような問題に対しては頑健であるが，使用できる変換行列が単一に限られるためモデル空間の適応と比べ，複雑な音響的な差異を正確に表現することができない．そこでモデル空間，特徴量空間の適応手法の両方の長所を活かすため，複数の変換行列を特徴量空間で用いる手法を提案する．教師あり適応であれば正解トランスクリプションを，教師なし適応であればデコードにより得られた認識結果を用いて，GMM の適応の枠組み（CMLLR）に基づき，複数の変換行列を得る[*1]．そして，これらの変換行列の重み付け和をとり単一の変換行列を推定して特徴量を変換し，再度デコードして最終的な認識結果を得る．この処理は単一の変換行列を乗算するため，fMLLR と比較して計算量はほとんど増加しない．重み付けの重み係数は，教師あり適応であれば正解トランスクリプションに対するアラインメント，教師なし適応であれば 1 パス目のデコードにより得られる HMM の状態アラインメントに基づき，フレームごとに推定される．さらに変換行列推定の過学習を避けるため，構造的事後確率最大化（structual maximum a posterior：SMAP）基準を，特徴量空間の変換行列の推定に導入する．このことより本手法は CSMAPLR の拡張であり，特徴量空間の SMAPLR（fSMAPLR）ととらえることができる．実験結果から提案する fSMAPLR は，fMLLR よりも GMM，DNN 双方の音響モデルで優れることが分かった．

本論文はまず 2 章でモデル空間，特徴量空間における従来の適応手法について述べ，次に 3 章で複数の変換行列を用いた特徴量空間の適応手法を提案する．最後に 4 章で，実験により提案法の有効性を示す．

2. 従来の適応手法

本章では，従来の適応手法について述べる．まず，2.1 節

1 三菱電機株式会社情報技術総合研究所
Information Technology R&D Center, Mitsubishi Electric Corporation, Kamakura, Kanagawa 247–8501, Japan
2 Mitsubishi Electric Research Laboratories, Cambridge, MA 02139–1955, US
a) Kanagawa.Hiroki@ds.MitsubishiElectric.co.jp
b) Tachioka.Yuki@eb.MitsubishiElectric.co.jp
c) watanabe@merl.com

*1 本論文での実験は，教師なし適応とする．

と 2.2 節で述べる 2 つの適応手法はモデル空間の適応である．CMLLR（2.1 節）はモデル空間の適応手法として最も広く用いられている．この手法は，回帰木に基づき複数の変換行列を推定する．CMLLR は適応データに対して過学習する傾向があるため，SMAP 基準を導入した CSMAPLR が提案されている（2.2 節）．最後の適応手法（2.3 節）は単一の変換行列による CMLLR を，特徴量空間に用いたものである．CMLLR で推定する変換行列を単一にする場合，特徴ベクトルを特徴量空間で変換することと等価となることから，fMLLR と呼ばれている．

2.1 制約付き最尤線形回帰（CMLLR）

CMLLR では，ガウス分布における D 次元の平均ベクトル $\boldsymbol{\mu}_{jm} \in \mathbb{R}^D$ と共分散行列 $\boldsymbol{\Sigma}_{jm} \in \mathbb{R}^{D\times D}$ を式 (1), (2) において変換後の平均ベクトル $\hat{\boldsymbol{\mu}}_{jm}$，共分散行列 $\hat{\boldsymbol{\Sigma}}_{jm} \in \mathbb{R}^{D\times D}$ に変換する．j，m はそれぞれ HMM の状態，GMM の混合インデックスである．

$$\hat{\boldsymbol{\mu}}_{jm} = \boldsymbol{\Theta}_{r(j,m)}\boldsymbol{\mu}_{jm} + \boldsymbol{\varepsilon}_{r(j,m)}, \tag{1}$$

$$\hat{\boldsymbol{\Sigma}}_{jm} = \boldsymbol{\Theta}_{r(j,m)}\boldsymbol{\Sigma}_{jm}\boldsymbol{\Theta}^{\top}_{r(j,m)} \tag{2}$$

ここで r は回帰クラスのインデックスを示す．$\boldsymbol{\Theta}_{r(j,m)} \in \mathbb{R}^{D\times D}$ と $\boldsymbol{\varepsilon}_{r(j,m)} \in \mathbb{R}^D$ はそれぞれ変換行列の回転行列，バイアスベクトルである．r は j と m に対してユニークに対応付けられており，この対応付けは回帰木に基づく手法により得られる [5]．もし対角な共分散行列 $\boldsymbol{\Sigma}_{jm}$ が式 (2) によって変換される場合，$\hat{\boldsymbol{\Sigma}}_{jm}$ は全共分散行列となり，尤度計算のコストと音響モデルのサイズが著しく増加してしまう．しかし，t フレーム目の特徴量ベクトル $\boldsymbol{o}_t \in \mathbb{R}^D$ に対する，状態 j，混合 m の全共分散行列のガウス分布の尤度は対角共分散の尤度を用いて以下のように求められる．

$$\mathcal{L}_{jm}(\boldsymbol{o}_t) = \mathcal{N}(\boldsymbol{o}_t|\hat{\boldsymbol{\mu}}_{jm}, \hat{\boldsymbol{\Sigma}}_{jm}) \tag{3}$$

$$= \left|\boldsymbol{A}_{r(j,m)}\right| \mathcal{N}(\hat{\boldsymbol{o}}_{r(j,m),t}|\boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}) \tag{4}$$

ここで $\mathcal{N}$ はガウス分布を示す．回転行列 $\boldsymbol{A}_{r(j,m)}$，バイアスベクトル $\boldsymbol{b}_{r(j,m)}$，変換後の特徴量ベクトル $\hat{\boldsymbol{o}}_{r(j,m),t}$ をそれぞれ次式のように定義する．

$$\boldsymbol{A}_{r(j,m)} \triangleq \boldsymbol{\Theta}^{-1}_{r(j,m)}, \tag{5}$$

$$\boldsymbol{b}_{r(j,m)} \triangleq -\boldsymbol{\Theta}^{-1}_{r(j,m)}\boldsymbol{\varepsilon}_{r(j,m)}, \tag{6}$$

$$\hat{\boldsymbol{o}}_{r(j,m),t} \triangleq \boldsymbol{A}_{r(j,m)}\boldsymbol{o}_t + \boldsymbol{b}_{r(j,m)} = \boldsymbol{W}_{r(j,m)}\begin{bmatrix}\boldsymbol{o}_t \\ 1\end{bmatrix}. \tag{7}$$

したがって式 (3) の代わりに式 (4) を用いることで，全共分散の問題を回避できる．しかし，この手法は GMM の音響モデルの尤度計算に特化しており[*2]，DNN 音響モデルのスコア計算に適用することはできない．**図 1** にモデル空間の適応手法の概略を示す．このタイプの適応手法は，適応処理とデコード処理が組み合わさって実現されるため，同じ音響モデルを両方の処理で共有しなければならない．また CMLLR には，適応データ量が少ない場合に過学習しやすいという問題もある．

[*2] 特徴量ベクトルに対するアフィン変換 $\boldsymbol{W}_{r(j,m)}$ は，状態 j，混合要素 m に強く依存している．

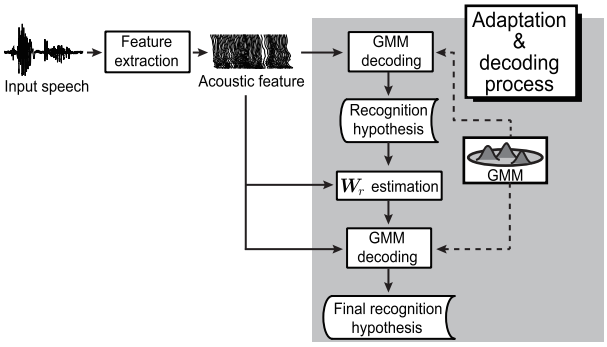


図 1 モデル空間の適応の概要
Fig. 1 Overview of model-space adaptation methods.

2.2 構造的な事前分布を用いた CMLLR（CSMAPLR）

CMLLR の過学習の問題は，ベイズ的アプローチを導入することにより解決できる．CSMAPLR [10] は，変換行列の集合 $\bar{\mathcal{W}} \triangleq \{\bar{\boldsymbol{W}}_r\}_{r=1}^R$ を次式の MAP 基準を用いて推定する．

$$\bar{\mathcal{W}} = \underset{\mathcal{W}}{\operatorname{argmax}}\, P(\mathcal{W})\, P(\boldsymbol{O}|\lambda, \mathcal{W}) \tag{8}$$

ここで R，$\mathcal{W} \triangleq \{\boldsymbol{W}_r\}_{r=1}^R$，$\bar{\boldsymbol{W}}_r$ はそれぞれ回帰クラスの数，最尤推定により求めた変換行列 $\boldsymbol{W}_r$ の集合，MAP 推定により求めた回帰クラス r における変換行列である．また T，$\boldsymbol{O} = \{\boldsymbol{o}_t|t = 1, \ldots, T\}$，$\lambda$ はそれぞれ，フレーム数，特徴量ベクトル系列と GMM のモデルパラメータの集合を示す．本手法では，階層的な事前分布 $P(\mathcal{W})$ を使用する．たとえば，CSMAPLR では下記のような事前分布 $P(\boldsymbol{W}_r)$ を使用する．

$$P(\boldsymbol{W}_r) \propto |\boldsymbol{\Omega}|^{-D/2}\,|\boldsymbol{\Psi}|^{-(D+1)/2} \times \exp\left\{-\frac{1}{2}\mathrm{tr}\big(\boldsymbol{W}_r - \boldsymbol{W}_{\mathsf{pa(r)}}\big)^{\top}\boldsymbol{\Omega}^{-1}\big(\boldsymbol{W}_r - \boldsymbol{W}_{\mathsf{pa(r)}}\big)\boldsymbol{\Psi}^{-1}\right\} \tag{9}$$

ここで $\mathsf{pa(r)}$ は，当該ノード r の親ノードの回帰クラスのインデックスを示す．$\boldsymbol{\Omega} \in \mathbb{R}^{D\times D}$ と $\boldsymbol{\Psi} \in \mathbb{R}^{(D+1)\times(D+1)}$ は事前分布のハイパーパラメータである．本報では，事前分布として先行文献 [7], [10] と同様，$\boldsymbol{\Omega} = \tau\boldsymbol{I}_D$ と $\boldsymbol{\Psi} = \boldsymbol{I}_{D+1}$ を用いる．τ は，事前分布の影響をコントロールするスケーリングパラメータ（SMAP 係数）である．CSMAPLR の回帰クラス r の変換行列 $\bar{\boldsymbol{W}}_r$ の l 行目の列ベクトル $\bar{\boldsymbol{w}}_r^{(l)}$ 推定には，統計量 $\bar{\boldsymbol{y}}_r^{(l)}$，$\bar{\boldsymbol{G}}_r^{(l)}$ が必要である．$\bar{\boldsymbol{y}}_r^{(l)}$，$\bar{\boldsymbol{G}}_r^{(l)}$ は CMLLR の変換行列 $\boldsymbol{W}_r$ の l 行目の要素計算における統計量 $\boldsymbol{y}_r^{(l)}$，$\boldsymbol{G}_r^{(l)}$ に対して親ノードの事前情報の加算により得

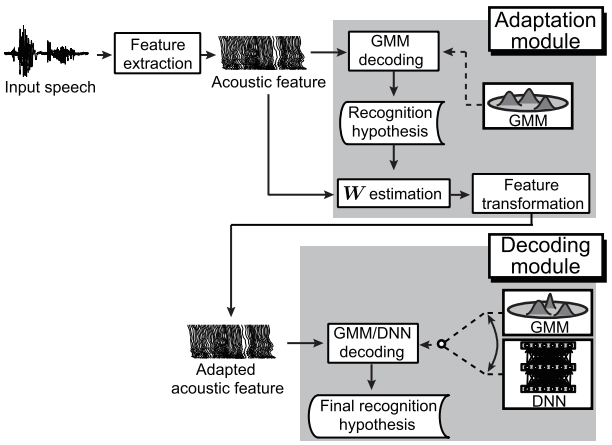


図 2 特徴量空間の適応の概要

Fig. 2 Overview of feature-space adaptation methods.

られる．

$$\bar{\boldsymbol{y}}_r^{(l)} = \boldsymbol{y}_r^{(l)} + \tau^{-1}\boldsymbol{w}_{\mathsf{pa(r)}}^{(l)}, \tag{10}$$

$$\bar{\boldsymbol{G}}_r^{(l)} = \boldsymbol{G}_r^{(l)} + \tau^{-1}\boldsymbol{I}_{D+1} \tag{11}$$

ここで $\boldsymbol{w}_{\mathsf{pa(r)}}^{(l)}$ は親ノードの変換行列 $\boldsymbol{W}_{\mathsf{pa(r)}}$ の l 行目の列ベクトルである．$\tau = \infty$ のときの CSMAPLR は CMLLR と一致するが，これは式 (10)，式 (11) の第二項が 0 となって親ノードの影響がなくなるためである．

2.3 特徴量空間の最尤線形回帰（fMLLR）

2.1 節で述べた CMLLR において，複数の変換行列の代わりに単一の変換行列を用いると，さらに式 (4) の尤度は次式で書き直すことができる．

$$\mathcal{L}_{jm}(\boldsymbol{o}_t) = |\boldsymbol{A}|\mathcal{N}(\hat{\boldsymbol{o}}_t|\boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}) \tag{12}$$

ここで $\hat{\boldsymbol{o}}$ は変換後の特徴量であり，次式で定義される．

$$\hat{\boldsymbol{o}}_t \triangleq \boldsymbol{A}\boldsymbol{o}_t + \boldsymbol{b} \tag{13}$$

したがって適応した特徴量が適応処理により得られ，**図 2** に示すようにデコード処理と分離することができる．このため，fMLLR のような特徴量空間の適応手法は，モデルパラメータの変換が困難な DNN の音響モデルにも適用できることから，広く用いられている．しかし変換行列が単一なため，複数の変換行列を用いた CMLLR よりも性能が劣るという短所もある．

3. 複数の変換行列の重み付け和による特徴量空間の適応法

3.1 複数の変換行列の重み付け法

図 3 に提案手法の概略図を示す．この図は「あき」と発話したとき，5 個の CMLLR 変換行列を音響特徴量に適用する方法を示している．音響特徴量の時系列変化に対処するため，音響特徴量と変換行列をフレームごとに割り当てる．この割当てを実現するため，GMM の音響モデルにより得られる状態アラインメント[*3]を用いる．図 3 では，アラインメントを $S = \{s_t | t = 1, \ldots, T\}$ のように，状態インデックス系列として表現している．s_t を得ることで，対応する GMM の集合 $\mathcal{M}_{s_t}$ を特定することができ，s_t と $\mathcal{M}_{s_t}$ から複数の回帰クラス $\{r(s_t, m)\}_{m\in\mathcal{M}_{s_t}}$ との対応が得られる．したがって，音声特徴量 $\boldsymbol{o}_t$ と複数の変換行列 $\{\boldsymbol{W}_{r(s_t,m)}\}_{m\in\mathcal{M}_{s_t}}$ を対応付けることができる．

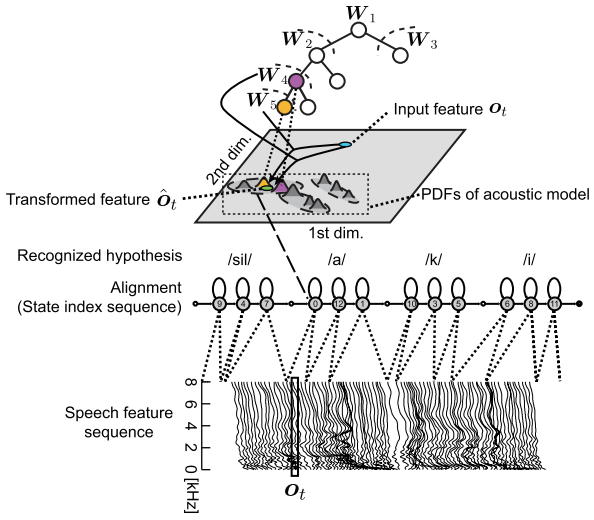


図 3 提案法の概要図

Fig. 3 Outline of the proposed method.

2.1 節で述べたように，モデル空間の適応手法では各ガウス分布に対応する単一の変換行列を用いて HMM の出力確率を計算する．しかし DNN に適用するには GMM 固有の計算を避け，モデル空間ではなく特徴量空間での適応とし，変換行列をモデル空間の $\boldsymbol{W}_r$ から特徴量空間の $\boldsymbol{W}$ に変換する必要がある．したがって，提案法はこれら複数の変換行列 $\boldsymbol{W}_r$ の重み付け和をとり，単一の変換行列を推定する．式 (13) とは異なり，変換後の t フレーム目の特徴量ベクトルを次式で表現する．

$$\begin{aligned}\hat{\boldsymbol{o}}_t &= \sum\nolimits_{m\in\mathcal{M}_{s_t}} \rho(s_t, m, \boldsymbol{o}_t)\left(\boldsymbol{A}_{r(s_t,m)}\boldsymbol{o}_t + \boldsymbol{b}_{r(s_t,m)}\right) \\ &= \sum\nolimits_{m\in\mathcal{M}_{s_t}} \rho(s_t, m, \boldsymbol{o}_t)\boldsymbol{W}_{r(s_t,m)}\begin{bmatrix}\boldsymbol{o}_t \\ 1\end{bmatrix},\end{aligned} \tag{14}$$

ここで $\rho(s_t, m, \boldsymbol{o}_t)$ は，フレーム依存の重みパラメータであり，状態 s_t と GMM の混合要素 m の両方に対応付けられ，具体的には 3.3 節で議論する．**図 4** に提案法の特徴量の変換方法の具体例を示す．状態 s_t の GMM の要素が 5 つのガウス分布（$\mathcal{N}_1, \ldots, \mathcal{N}_5$）から構成されている．重みパラメータ ρ が固定できれば，各ガウス分布に対応する 5 つの変換特徴量の重み付け和により，変換特徴量が得られる．この適応は特徴量空間で動作するため，特徴量の変換が DNN においても GMM 同様に実現できる．

[*3] アラインメントの代わりにラティスや N-best の認識候補を用いることもできる．

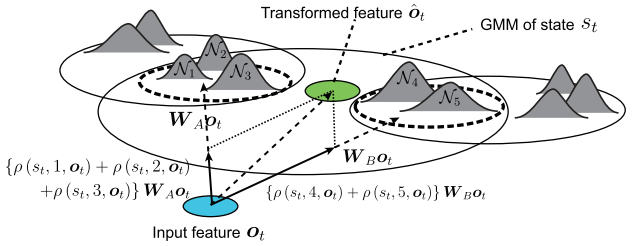


図 4 提案法による特徴量の変換方法の具体例．ここで状態 s_t は 5 つのガウス分布（$\mathcal{N}_{m=\{1,2,3,4,5\}\in s_t} \in \mathcal{M}_{s_t}$）から構成され，変換行列を $\boldsymbol{W}_A$ と $\boldsymbol{W}_B$ で表現する．分布 1，2，3 は $\boldsymbol{W}_A$ を共有し，分布 4，5 は $\boldsymbol{W}_B$ を共有する．また各分布に対する変換行列への重みパラメータは $\rho(s_t,1,\boldsymbol{o}_t),\ldots,\rho(s_t,5,\boldsymbol{o}_t)$ である

Fig. 4 Concrete example of the proposed feature transformation where the component of the state s_t includes five Gaussian distributions ($\mathcal{N}_{m=\{1,2,3,4,5\}\in s_t} \in \mathcal{M}_{s_t}$) and $\boldsymbol{W}_A$ and $\boldsymbol{W}_B$ are transformation matrices. Distributions 1, 2, and 3 share $\boldsymbol{W}_A$ and distributions 4 and 5 share $\boldsymbol{W}_B$. Their weight parameters are $\rho(s_t,1,\boldsymbol{o}_t),\ldots,$ and $\rho(s_t,1,\boldsymbol{o}_t)$.

Algorithm 1 提案する特徴量変換アルゴリズム（The proposed feature transformation algorithm）

Input: Acoustic feature sequence $\boldsymbol{O}=\{\boldsymbol{o}_t|t=1,\ldots,T\}$ and GMM acoustic model parameters λ
Obtain state sequence $S=\{s_t|t=1,\ldots,T\}$ at the first-pass decoding ($S=\mathrm{decode}(\boldsymbol{O})$) (using GMM)
Estimate transformation matrices $\bar{\mathcal{W}}$ by Eq. (8)
for $t=1,\cdots,T$ **do**
  for $m\in\mathcal{M}_{s_t}$ **do**

$$\hat{\boldsymbol{o}}_t=\sum_{m\in\mathcal{M}_{s_t}}\rho(s_t,m,\boldsymbol{o}_t)\left(\boldsymbol{A}_{r(s_t,m)}\boldsymbol{o}_t+\boldsymbol{b}_{r(s_t,m)}\right)$$
$$=\sum_{m\in\mathcal{M}_{s_t}}\rho(s_t,m,\boldsymbol{o}_t)\,\boldsymbol{W}_{r(s_t,m)}\begin{bmatrix}\boldsymbol{o}_t\\1\end{bmatrix}$$

  end for
end for
Second-pass decoding with $\hat{\boldsymbol{O}}=\{\hat{\boldsymbol{o}}_t|t=1,\ldots,T\}$ (using GMM/DNN)

3.2 提案法の適応の手順

アルゴリズム 1 に，提案する fSMAPLR の手順を示す．まず GMM を用いた 1 パス目のデコードによりすべての適応データを用いて，認識候補とコンテキスト依存の状態アラインメント S を得る．次に式 (8) に基づき，複数の変換行列 $\bar{\mathcal{W}}$ を推定する．事前分布の影響を調整するため，CSMAPLR 同様に SMAP 係数 τ を導入する．$\tau=0$ のとき，式 (10)，式 (11) の第一項が無視できるほど第二項が支配的となるため，全ノードの変換行列がルートノードと同じとなる．したがって $\sum_{m=1}^{\mathcal{M}_{s_t}}\rho(s_t,m,\boldsymbol{o}_t)=1$ となるような重み $\rho(s_t,m,\boldsymbol{o}_t)$ を用いるとき，単一の変換行列が使用され，本手法は fMLLR と等価となる．$\bar{\mathcal{W}}$ を推定後，元の音響特徴量 $\boldsymbol{o}_t$ を式 (14) により特徴量 $\hat{\boldsymbol{o}}_t$ に変換する．最後に $\hat{\boldsymbol{o}}_t$ を用いて，GMM もしくは DNN の音響モデルに対し 2 パス目のデコードを行い，最終的な音声認識結果を得る．

3.3 2 種類の重みパラメータについて

3.1 節では，変換後の特徴量ベクトル $\hat{\boldsymbol{o}}_t$ の生成に，重み付けパラメータ $\rho(s_t,m,\boldsymbol{o}_t)$ を用いることを述べた．本節では，2 種類の重みパラメータを使うことを提案する．

まず 1 つ目の重みパラメータ $\rho(s_t,m,\boldsymbol{o}_t)$ として，GMM の混合要素 m に対する事後確率 $\gamma_{s_t,m}(\boldsymbol{o}_t)$ を用いる．状態 s_t はアラインメントから得られているため，$\gamma_{s_t,m}(\boldsymbol{o}_t)$ は次式により計算される．

$$\gamma_{s_t,m}(\boldsymbol{o}_t)=\frac{w_{s_t,m}\mathcal{N}\left(\boldsymbol{o}_t|\boldsymbol{\mu}_{s_t,m},\boldsymbol{\Sigma}_{s_t,m}\right)}{\sum_{m'\in\mathcal{M}_{s_t}}w_{s_t,m'}\mathcal{N}\left(\boldsymbol{o}_t|\boldsymbol{\mu}_{s_t,m'},\boldsymbol{\Sigma}_{s_t,m'}\right)}\tag{15}$$

ここで，未適応の平均ベクトル $\boldsymbol{\mu}_{s_t,m}$ と対角共分散行列 $\boldsymbol{\Sigma}_{s_t,m}$ を用いる[*4]．しかし GMM の混合において，ある特定の混合要素の影響が支配的になり，事後確率が非常にスパースになることがある．すると，式 (15) において単一の変換行列のみを用いることとほぼ等価となり，式 (14) で複数の変換行列に拡張した利点を活かしきれない．

2 つ目の重みパラメータとして，GMM の混合重みを用いる．このアプローチをとるのは，$\gamma_{s_t,m}(\boldsymbol{o}_t)$ がスパースとなってしまうことを避けるためである．GMM の混合重みを用いることは，式 (15) において $\mathcal{N}\left(\boldsymbol{o}_t|\boldsymbol{\mu}_{s_t,m},\boldsymbol{\Sigma}_{s_t,m}\right)$ の項を無視し，下記の近似を行うことに等しい．

$$\rho(s_t,m,\boldsymbol{o}_t)=\gamma(s_t,m,\boldsymbol{o}_t)$$
$$\cong\frac{w(s_t,m)}{\sum_{m'\in\mathcal{M}_{s_t}}w(s_t,m')}=w(s_t,m)\tag{16}$$

ここで s_t はフレーム t に依存するため，$w(s_t,m)$ はフレームごとに異なる．また m も s_t ($m\in\mathcal{M}_{s_t}$) に依存する．式 (16) は式 (15) を用いた場合よりも $\boldsymbol{o}_t$ に依存しないので，外乱の影響を受けにくい．

図 5，**図 6** に 5 つの変換行列の加重和をとる際，それぞれ事後確率，GMM の混合重みを重みパラメータ ρ として用いたとき場合のフレームごとの重みの変化を示す．横軸がフレーム，縦軸は 5 色で示される各変換行列に対する重みを表しており，時間とともに重みがどう変化するかを示す．図 5 に示す事後確率はスパースであるため，各時刻において特定の変換行列の重みが非常に支配的であることが分かる．一方で図 6 に示す GMM の混合重みは，事後確率のようにスパースでないため，特定の変換行列の重みのみに依存しないことが分かる．

[*4] 式 (15) を用いた変換式 (14) は識別的特徴量変換の手法に非常に似ている [20], [21], [22]．しかしこれらの手法は GMM の識別学習の枠組みで動作しており，GMM，DNN の双方での特徴量空間の適応に注力する我々のアプローチとは異なる．

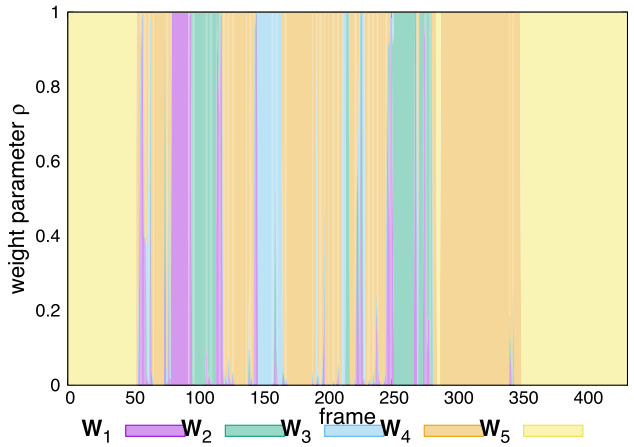


図 5 事後確率（式 (15)）を重みパラメータ $\rho(s_t, m, \boldsymbol{o}_t)$ として用いた場合のフレームごとの変化

Fig. 5 The transitions of weight paramters $\rho(s_t, m, \boldsymbol{o}_t)$ frame by frame. Posteriors (Eq. (15)) are used as weight parameters.

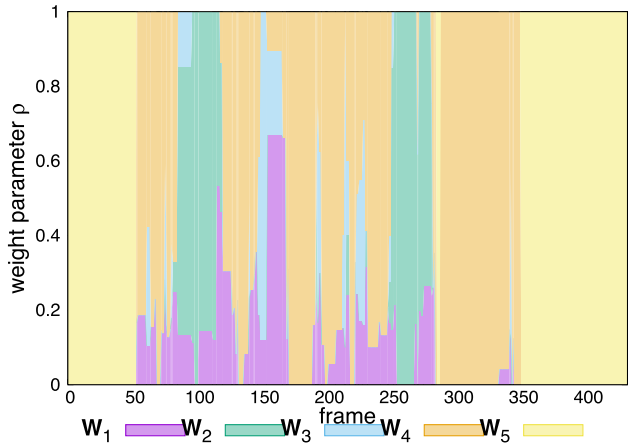


図 6 GMM の混合重み（式 (16)）を重みパラメータ ρ として用いた場合のフレームごとの変化

Fig. 6 The transitions of weight paramters $\rho(s_t, m, \boldsymbol{o}_t)$ frame by frame. GMM mixture weights (Eq. (16)) are used as weight parameters.

4. 第 2 回 CHiME チャレンジによる音声認識実験

4.1 実験条件

騒音下音声認識のタスクである第 2 回 CHiME チャレンジ [23] の Track 2 の孤立（“isolated”）音声[*5]に対して提案手法の有効性を示す実験を行った．Track 2 は中語彙サイズのタスクで，残響かつ騒音環境下で収録されており，ウォール・ストリート・ジャーナルのデータベースから発話がとられている．学習および評価に用いる音声データは，実環境で収録した騒音を，騒音を収録した部屋と同じ部屋で収録した残響音声に対し信号対騒音比（signal-to-noise ratio：SNR）−6，−3，0，3，6，9 dB で重畳し作成した．学習データセット（si_tr_s）は 83 話者の計 7,138 発話（15 [時間]）から構成される．音響モデル（GMM と DNN）は si_tr_s を用いて学習した．性能は，12 話者の計 330 発話（0.67 [時間/SNR] × 6 [SNR]）からなる評価データセット（si_et_05）（Nov’92）と，10 話者の計 409 発話（0.77 [時間/SNR] × 6 [SNR]）からなる開発データセット（si_dt_05）の両方を用いて評価する．各評価話者の全発話（約 4–5 分）を，適応データおよび評価データとして使用する．重畳された騒音は非定常的なものであり，たとえば他話者の発話や，家庭内騒音，音楽が該当する．これらの騒音重畳音声に対し，騒音の影響を低減するために，事前分布に基づくバイナリマスク [24] を前処理に使用した．言語モデルはトライグラムで，サイズは 5k（basic）である．言語重みや変換行列数などのパラメータは，開発セット（si_dt_05）で単語誤り率（word-error rate：WER）が最適となるよう調整した．

実験では 2 種類の音響特徴量を用いる．1 つ目の特徴量は，特徴量変換を用いた MFCC である．0–12 次の静的 MFCC に対し近接する 9 フレームを結合し，生成された計 117 次元の特徴量を線形判別分析（linear discriminant analysis：LDA）[25] により 40 次元に圧縮する[*6]．さらに次元間の相関を低減するため，LDA により変換した特徴量に対し，STC（semi-tied covariance）[26] 行列を適用した．LDA と STC により特徴量を変換した後，話者適応学習 [27] により音響モデルを学習した．2 つ目の特徴量として次元間相関を低減したフィルタバンク特徴量を用いた．0–22 次の静的フィルタバンク特徴量とその Δ および ΔΔ からなる 69 次元のベクトルを使用した．ただしフィルタバンク特徴量は次元間の相関が強く，対角共分散 GMM では次元間相関を精度良く表現できない [28]．このため，フィルタバンクをそのまま用いた fMLLR では音声認識性能を改善できず，適用前に次元間相関を低減しておく必要がある [29]．したがって適応処理では STC 行列 $\boldsymbol{H}$ を次元間相関低減のためフィルタバンク特徴量に適用しておき，デコード処理では fMLLR もしくは fSMAPLR による適応後の特徴量に STC の逆行列 $\boldsymbol{H}^{-1}$ を文献 [28] と同様に適用する．

音響モデルの学習には文献 [24] 同様，Kaldi ツールキット [30] を使用した．トライフォンの GMM 音響モデルは状態数 2,500 であり，ガウス分布の総数は 15,000 である．DNN 音響モデルは 3 つの隠れ層，500,000 個のパラメータを持つ．DNN のクロスエントロピー学習における初期ラーニングレートは 0.02 であり，学習終了時には 0.004 に減少した．ミニバッチサイズは 128 である．音響モデルの学習とデコードには Kaldi ツールキット [30] を用い，音響モデルは文献 [24] と同様の手順で学習した．

[*5] 騒音音声には “isolated” と “embedded” の 2 種類がある．

[*6] LDA には動的特徴量は使用していない．

表 1 CHiME チャレンジトラック 2 開発セット（si_dt_05）での WER [%]．事後確率（式 (15)）と混合重み（式 (16)）は式 (14) の重みに用いた場合のそれぞれの WER を示している

Table 1 WER [%] on the development set of the Track 2 of the second CHiME Challenge when a posterior (Eq. (15)) or a mixture weight (Eq. (16)) is used for the weight ρ in Eq. (14).

the number of $\boldsymbol{W}$	weight ρ	SMAP scale τ ∞	10^{-2}	10^{-3}
5	posterior	39.7	39.6	39.3
	mixture weight	**39.5**	**39.5**	**39.2**
10	posterior	40.8	40.5	39.8
	mixture weight	**40.4**	**40.2**	**39.7**

4.2 変換行列に対する適切な重みパラメータ

提案法を従来法と比較する前に，3.3 節で述べた 2 種類の重みパラメータについて検討する．**表 1** に変換行列数が 5，10 のときの，事後確率（式 (15)）と GMM の混合重み（式 (16)）のそれぞれの平均 WER を示す．fSMAPLR のハイパーパラメータである SMAP 係数 τ は ∞[*7]，10^{-2}，10^{-3} とした．これらの結果から 3.3 節で述べたように，事後確率は各混合間でスパースであるがゆえ，フレームごとの重みの遷移が急峻となるため，混合重みの方が事後確率より優れることが分かった．なお最適な変換行列数と τ は以降の節で詳細に議論する．提案する fSMAPLR には本節以降，結果の良かった式 (16) を用いることとする．

4.3 GMM 音響モデルにおける評価

図 7 に開発セット（si_dt_05）での各 SNR における平均 WER を示す．提案する fSMAPLR は，式 (8) により得られた変換行列を用いて，音響特徴量をフレームごとに式 (14) に基づき変換する．SMAP 係数 τ は ∞，1，10^{-1}，10^{-2}，10^{-3} とした．提案法は変換行列数が 3，5 の場合において，すべての τ で fMLLR より優れた．提案法で変換行列数を 10 より大きくすると性能が劣化するが，これは変換行列がデータ量の少ない子ノードに対して過学習するためである．一方 τ を大きくすることにより，変換行列の増加にともなう過学習を防ぐことができている．このことから階層的な MAP 推定の有効性を確かめられた．これらの結果をもとに，変換行列数と τ をそれぞれ 5，10^{-3} に固定する．

次に開発セット（si_dt_05）および評価セット（si_et_05）で性能を評価する．複数の変換行列を用いた CSMAPLR [10] についても評価し，CSMAPLR の変換行列数と τ は fSMAPLR と同様とした．提案法の fSMAPLR を，ベースライン（適応なし），fMLLR，CSMAPLR と比較した．**表 2** に騒音下音声における各 SNR の WER を，また平均 WER を “avg.” として示す．またクリーン音声に対する WER を “Clean” として示す．

[*7] この場合，CMLLR の変換行列を用いることと等価．

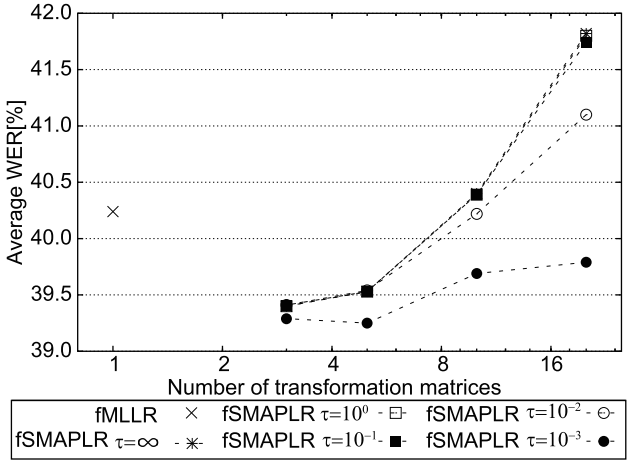


図 7 開発セット（si_dt_05）の孤立（“isolated”）音声に対する GMM 音響モデルを用いた平均 WER [%]．SMAP 係数 τ と変換行列数のパラメータに対する WER の関係性

Fig. 7 Average WER [%] for isolated speech (si_dt_05) with the GMM acoustic model. Parametric study of the SMAP scale τ and the number of transformation matrices.

表 2 孤立（“isolated”）音声の学習セット（si_dt_05）および評価セット（si_et_05）の SNR ごとの WER [%]．GMM 音響モデルを用いた

Table 2 WER [%] for isolated speech (si_{dt,et}_05) with GMM acoustic model in terms of SNR.

	Method	Noisy SNR [dB] −6	−3	0	3	6	9	avg.	Clean
dt	w/o adapt.	67.3	57.6	49.5	43.7	36.9	32.1	47.9	12.3
	fMLLR	61.4	50.2	41.3	34.6	29.1	24.8	40.2	8.9
	fSMAPLR	**61.1**	**49.0***	**40.7**	**33.2***	**28.0***	**23.5***	**39.2***	**8.5***
	CSMAPLR	61.1	50.1	41.0	33.2	27.9	23.9	39.5	8.6
et	w/o adapt.	62.7	54.7	48.0	40.6	35.4	31.8	45.5	13.2
	fMLLR	54.3	45.7	36.9	28.5	23.6	20.1	34.8	8.0
	fSMAPLR	**52.9***	**44.7***	**35.2***	**27.3***	**22.5***	**18.7***	**33.6***	**6.7***
	CSMAPLR	52.7	43.7	35.5	27.4	22.5	19.1	33.5	6.9

* significant at the 5% level.

これらの結果から適応が有効であること，また提案法の fSMAPLR の性能が fMLLR に対しすべての SNR で上回り，評価セットの平均 WER で 1.2%（絶対値）優れ，発話ごとの WER に基づく t 検定により，5%水準で有意であることを確認した．fSMAPLR は CSMAPLR と同程度の性能であることが分かり，このことからモデル空間，特徴量空間の双方において複数の変換行列を使用することの有効性が確かめられた．

また開発セットの結果に注目すると，fMLLR と比べて提案法の改善幅が大きいのは，主に比較的高 SNR 時であったが，これは 1 パス目のデコード結果とアラインメントの推定精度が低 SNR 時より良いためであると考えられる．さらにクリーン音声においても 0.4%（絶対値）優れたこと

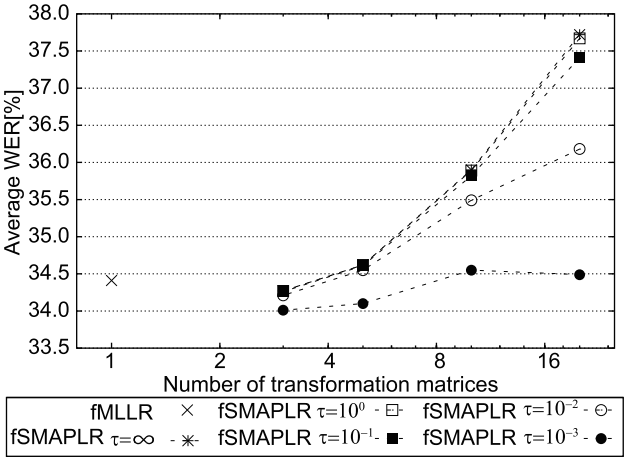


図 8 開発セット（si_dt_05）の孤立（“isolated”）音声に対する DNN 音響モデルを用いた平均 WER [%]

Fig. 8 Average WER [%] for isolated speech (si_dt_05) with the DNN acoustic model.

表 3 孤立（“isolated”）音声の学習セット（si_dt_05）および評価セット（si_et_05）の SNR ごとの WER [%]．MFCC を音響特徴量とし，DNN 音響モデルを用いた

Table 3 WER [%] for isolated speech (si_{dt,et}_05) with DNN acoustic model using MFCC features in terms of SNR.

		Noisy							Clean
		SNR [dB]							
	Method	−6	−3	0	3	6	9	avg.	
dt	w/o adapt.	61.5	51.4	42.9	36.4	32.5	28.1	42.1	10.5
	fMLLR	55.0	**43.1**	35.3	27.9	24.6	20.7	34.4	7.3
	fSMAPLR	**54.7**	**43.1**	**35.0**	**27.3***	**23.7***	**20.2**	**34.0***	**7.2***
et	w/o adapt.	56.3	47.0	39.3	32.7	29.3	26.1	38.5	10.4
	fMLLR	47.0	37.4	29.5	22.0	18.4	15.4	28.3	5.4
	fSMAPLR	**46.6**	**36.4**	**29.2**	**21.6**	**17.2***	**15.0***	**27.6***	**5.2**

* significant at the 5% level.

から，提案法の有効性は騒音環境下に限定されるものではないことも分かった．

4.4 DNN 音響モデルにおける評価

4.4.1 MFCC 特徴量

本項では，DNN の音響モデルに対する評価を行う．**図 8** に，開発セット（si_dt_05）での各 SNR における平均 WER を示す．なお図の表記法は，図 7 と同様である．これらの結果から，変換行列数が 3 の場合，fSMAPLR が fMLLR より優れることが分かった．変換行列数を多く推定し過ぎると，fSMAPLR の性能は GMM での場合と同様，性能が劣化した．本評価結果をもとに，変換行列数と τ をそれぞれ 3，10^{-3} とする．

次に開発セットと評価セット（si_et_05）で提案法の fSMAPLR を，ベースライン（適応なし），fMLLR と比較し，**表 3** に結果を示す．ここで CSMAPLR は 2.1，2.2 節で述べたように，DNN では実現できないことに注意されたい．表 2 と比較すると，DNN は GMM よりすべてのケースにおいて性能が優れた．

表 4 孤立（“isolated”）音声の学習セット（si_dt_05）および評価セット（si_et_05）の SNR ごとの WER [%]．フィルタバンクを音響特徴量とし，DNN 音響モデルを用いた

Table 4 WER [%] for isolated speech (si_{dt,et}_05) with DNN acoustic model using fbank features in terms of SNR.

		Noisy							Clean
		SNR [dB]							
	Method	−6	−3	0	3	6	9	avg.	
dt	w/o adapt.	55.7	44.6	36.4	30.6	25.9	22.5	35.9	8.4
	fMLLR	53.5	**42.8**	34.0	28.7	24.8	19.9	34.0	7.7
	fSMAPLR	**52.7***	43.0	**33.5**	**28.3***	**24.3***	**19.4**	**33.6***	7.7
et	w/o adapt.	47.9	38.7	32.4	24.7	21.4	19.5	30.8	6.7
	fMLLR	**45.2**	35.7	29.1	21.5	18.2	16.6	27.7	5.5
	fSMAPLR	45.3	**35.1**	**28.5**	**21.4**	**18.1**	**16.2***	**27.4***	**5.4**

* significant at the 5% level.

結果より，DNN においても適応が有効であることが分かった．fMLLR と比べると，提案する fSMAPLR の性能は全 SNR において上回り，評価セットの平均 SNR において WER が 0.7%（絶対値）優れ，クリーン音声においても 0.2%（絶対値）優れることが分かった．発話ごとの WER に基づく t 検定により，5%水準で有意であることを確認した．これまでの実験結果から，提案する fSMAPLR は，fMLLR よりも GMM/DNN の双方の音響モデルにおいて性能が優れた．

4.4.2 フィルタバンク特徴量

開発セット（si_dt_05）と評価セット（si_et_05）で提案法の fSMAPLR を，ベースライン（適応なし），fMLLR と比較し，**表 4** にフィルタバンク特徴量を用いた平均 WER を示す．適応なしの性能は，MFCC を用いた場合よりも大幅に改善している．また適応した特徴量に対しては，性能の改善幅は小さいものの，提案法は fMLLR と比較して評価セットの平均 WER で 0.3%の改善（有意差あり），クリーン音声で 0.1%（絶対値）の改善が見られた．

これまでの実験により，提案する fSMAPLR は fMLLR より優れ，MFCC 特徴量とフィルタバンク特徴量の両方で有効であることが分かった．

5. おわりに

本論文では，回帰木に基づく複数の変換行列を用いた特徴量空間の適応法を提案し，さらに変換行列の過学習を防ぐために構造的な MAP 推定を導入した．実験結果から提案法の fSMAPLR は，GMM において fMLLR より優れ，モデル空間の CSMAPLR と同程度の性能を示した．さらに，提案法により変換した特徴量ベクトルを，従来の CSMAPLR では扱えなかった DNN の音響モデルに入力し，fMLLR の性能を上回ることを確認した．今後の課題として，適切な重みパラメータの導出，変換行列の推定における VBLR [8], [13] の導入，また提案法により得られる

変換特徴量を用いたDNNでの話者適応学習 [27] があげられる．

参考文献

[1] Lee, C.-H. and Huo, Q.: On adaptive decision rules and decision parameter adaptation for automatic speech recognition, *Proc. IEEE*, Vol.88, No.8, pp.1241–1269 (2000).

[2] Shinoda, K.: Speaker adaptation techniques for automatic speech recognition, *Proc. APSIPA*, pp.1–8 (2011).

[3] Leggetter, C.J. and Woodland, P.C.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, *Computer Speech and Language*, Vol.9, No.2, pp.171–185 (1995).

[4] Digalakis, V.V., Rtischev, D. and Neumeyer, L.G.: Speaker adaptation using constrained estimation of Gaussian mixtures, *IEEE Trans. Speech and Audio Processing*, Vol.3, No.5, pp.357–366 (1995).

[5] Gales, M.J.: The generation and use of regression class trees for MLLR adaptation, Technical Report CUED/F-INFENG/TR, Vol.263 (1996).

[6] Shinoda, K. and Lee, C.-H.: Structural MAP speaker adaptation using hierarchical priors, *Proc. ASRU*, pp.381–388 (1997).

[7] Siohan, O., Myrvoll, T.A. and Lee, C.-H.: Structural maximum a posteriori linear regression for fast HMM adaptation, *Computer Speech and Language*, Vol.16, No.1, pp.5–24 (2002).

[8] Watanabe, S., Nakamura, A. and Juang, B.-H.: Bayesian linear regression for hidden Markov model based on optimizing variational bounds, *Proc. MLSP*, pp.1–6 (2011).

[9] Gales, M.J.: Maximum likelihood linear transformations for HMM-based speech recognition, *Computer Speech and Language*, Vol.12, No.2, pp.75–98 (1998).

[10] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K. and Isogai, J.: Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.17, No.1, pp.66–83 (2009).

[11] Lei, X., Hamaker, J. and He, X.: Robust feature space adaptation for telephony speech recognition, *Proc. ICSLP*, pp.773–776 (2006).

[12] Huang, Z., Li, J., Siniscalchi, S.M., Chen, I.-F., Weng, C. and Lee, C.-H.: Feature Space Maximum A Posteriori Linear Regression for Adaptation of Deep Neural Networks, *Proc. INTERSPEECH*, pp.2992–2996 (2014).

[13] Hahm, S.-J., Ogawa, A., Delcroix, M., Fujimoto, M., Hori, T. and Nakamura, A.: Feature space variational Bayesian linear regression and its combination with model space VBLR, *Proc. ICASSP*, pp.7898–7902 (2013).

[14] Yoshioka, T., Ragni, A. and Gales, M.J.: Investigation of unsupervised adaptation of DNN acoustic models with filter bank input, *Proc. ICASSP*, pp.13–16 (2014).

[15] Neto, J., Almeida, L., Hochberg, M., Martins, C., Nunes, L., Renals, S. and Robinson, T.: Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system, *Proc. EUROSPEECH*, pp.2171–2174 (1995).

[16] Abrash, V., Franco, H., Sankar, A. and Cohen, M.: Connectionist speaker normalization and adaptation, *Proc. EUROSPEECH*, pp.2183–2186 (1995).

[17] Yao, K., Yu, D., Seide, F., Su, H., Deng, L. and Gong, Y.: Adaptation of context-dependent deep neural networks for automatic speech recognition, *Proc. SLT*, pp.366–369 (2012).

[18] Gemello, R., Mana, F., Scanzio, S., Laface, P. and De Mori, R.: Adaptation of hybrid ANN/HMM models using linear hidden transformations and conservative training, *Proc. ICASSP*, pp.1189–1192 (2006).

[19] Ochiai, T., Matsuda, S., Watanabe, H., Lu, X., Hori, C. and Katagiri, S.: Speaker adaptive training for deep neural networks embedding linear transformation networks, *Proc. ICASSP*, pp.4605–4609 (2015).

[20] Povey, D.: Improvements to fMPE for discriminative training of features, *Proc. INTERSPEECH*, pp.2977–2980 (2005).

[21] Droppo, J. and Acero, A.: Maximum mutual information SPLICE transform for seen and unseen conditions, *Proc. INTERSPEECH*, pp.989–992 (2005).

[22] Zhang, B., Matsoukas, S. and Schwartz, R.M.: Recent progress on the discriminative region-dependent transform for speech feature extraction, *Proc. INTERSPEECH*, pp.1573–1576 (2006).

[23] Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F. and Matassoni, M.: The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines, *Proc. ICASSP*, pp.126–130 (2013).

[24] Tachioka, Y., Watanabe, S., Le Roux, J. and Hershey, J.R.: Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark, *The 2nd International Workshop on Machine Listening in Multisource Environments*, pp.19–24 (2013).

[25] Haeb-Umbach, R. and Ney, H.: Linear discriminant analysis for improved large vocabulary continuous speech recognition, *Proc. ICASSP*, pp.13–16 (1992).

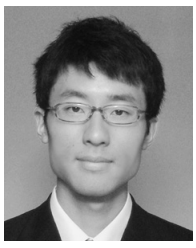
[26] Gales, M.J.: Semi-tied covariance matrices for hidden Markov models, *IEEE Trans. Speech and Audio Processing*, Vol.3, No.7, pp.272–281 (1999).

[27] Anastasakos, T., McDonough, J., Schwartz, R. and Makhoul, J.: A compact model for speaker-adaptive training, *Proc. ICSLP*, pp.1137–1140 (1996).

[28] Sainath, T., Kingsbury, B., Mohamed, A., Dahl, G.E., Saon, G., Soltau, H., Beran, T., Aravkin, A.Y. and Ramabhadran, B.: Improvements to deep convolutional neural networks for LVCSR, *Proc. ASRU*, pp.315–320 (2013).

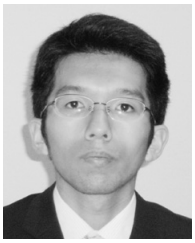
[29] Sainath, T., Mohamed, A., Kingsbury, B. and Ramabhadran, B.: Deep convolutional neural networks for LVCSR, *Proc. ICASSP*, pp.8614–8618 (2013).

[30] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G. and Veselý, K.: The Kaldi speech recognition toolkit, *Proc. ASRU*, pp.1–4 (2011).



金川 裕紀

2011年電気通信大学電気通信学部電子工学科卒業．2013年東京工業大学大学院修士課程修了．同年三菱電機（株）入社．以来，音声認識の研究開発に従事．現在，同社情報技術総合研究所知識情報処理技術部研究員．日本音響学会会員．



太刀岡 勇気 （正会員）

2006 年東京大学工学部建築学科卒業．2008 年同大学大学院修士課程修了．同年三菱電機（株）入社．以来，音声認識の研究開発に従事．現在，同社情報技術総合研究所知識情報処理技術部研究員．2008 年日本建築学会優秀修士論文賞，2014 年日本音響学会粟屋潔学術奨励賞．日本音響学会，計量国語学会各会員．

渡部 晋治

1999 年早稲田大学理工学部物理学科卒業，2001 年同大学大学院修士課程修了．同年 NTT コミュニケーション科学基礎研究所入社．2012 年より Mitsubishi Electric Research Laboratories（MERL）senior principal member．2009 年ジョージア工科大学客員研究員．博士（工学）．音声認識を中心とした音声言語処理の研究に従事．2003 年日本音響学会粟屋潔学術奨励賞，2004 年電子情報通信学会論文賞，2006 年日本音響学会独創研究奨励賞板倉記念，電気通信普及財団テレコムシステム技術賞各受賞．2012 年より IEEE Transaction on Audio, Speech, and Language Processing の Associate Editor，2014 年より IEEE Signal Processing Society, Speech and Language Technical Committee，および APSIPA Speech, Language, and Audio Technical Committee 等を歴任．日本音響学会，電子情報通信学会各会員，IEEE シニア会員．

石井 純

1988 年新潟大学工学部卒業．1990 年同大学大学院修士課程修了．同年三菱電機（株）入社．1995～1997 年 ATR 音声翻訳通信研究所に出向．現在，三菱電機（株）情報技術総合研究所知識情報処理技術部部長．