DNN のための複数変換行列を用いた特徴量空間の適応法*

◎金川裕紀，太刀岡勇気 (三菱電機)，渡部晋治 (MERL)，石井純 (三菱電機)

1 はじめに

特徴量空間の適応処理 (fMLLR) は，求めた変換行列を音響特徴量に乗算するだけでよく，デコード処理とは独立であるとみなせる．このため適応処理とデコード処理にそれぞれ異なる音響モデルを用いることができる．文献 [1] ではこの利点に注目し，精緻なモデル化が可能な一方，適応が困難であった DNN に，fMLLR による変換後の特徴量を入力することの有効性を示している．しかし，fMLLR に使用できる変換行列数は発話/話者毎に 1 つに限られるため，時系列変化する音響特徴量に対し適切に変換行列を割り当てられないといった課題があった．

この課題に対し我々は，既報 [2] にて複数の変換行列を用いた手法を提案し，GMM 音響モデルの認識実験にて，従来の単一の変換行列を用いた場合より優れることを示した．本報では，この提案内容に加え，少量の適応データ使用時の複数の変換行列の過学習に対処するため，変換行列の推定に構造的 MAP 基準を導入する．また DNN への入力に，提案法により変換した音響特徴量を用いた認識実験を行い，単一の変換行列を用いた場合より有効であることを示す．MFCC 特徴量に加え，フィルタバンク特徴量においても有効性を示す．

2 従来の適応手法

2.1 制約付き最尤線形回帰 (CMLLR)

モデル空間の適応である CMLLR[3] 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 $\boldsymbol{\Sigma}_{jm} \in \mathbb{R}^{D\times D}$ 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．

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

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

ここで r は回帰クラスのインデックス，$\boldsymbol{\Theta}_{r(m,j)} \in \mathbb{R}^{D\times D}$ と $\boldsymbol{\varepsilon}_{r(m,j)} \in \mathbb{R}^D$ 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 m と j 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 [4]．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．

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

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

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

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

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

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

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，DNN 音響モデルのスコア計算に適用することはできない．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．

2.2 制約付き事後確率最大線形回帰 (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 [5] 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 $\boldsymbol{O}$ と λ はそれぞれ，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．構造的な事前分布 $P(\mathcal{W})$ を使用する．例えば，CSMAPLR では下記のような事前分布 $P(\boldsymbol{W}_r)$ を使用する．

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

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 [6, 5] と同様，$\boldsymbol{\Omega} = \tau\boldsymbol{I}_D$ と $\boldsymbol{\Psi} = \boldsymbol{I}_{D+1}$ を用いる．τ は，MAP 推定における事前分布の影響をコントロールする正の定数である．$\tau = 0$ のとき，複数の変換行列を CMLLR で推定することと一致する．

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ことで，回帰クラスのインデックス r を無視でき，さらに式 (4) の尤度は次式で書き直すことができる．

$$\mathcal{L}_{jm}(\boldsymbol{o}_t) = |\boldsymbol{A}|\mathcal{N}(\hat{\boldsymbol{o}}_t|\boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}) \tag{10}$$

ここで $\hat{\boldsymbol{o}}$ は変換後の特徴量であり，

$$\hat{\boldsymbol{o}}_t \triangleq \boldsymbol{A}\boldsymbol{o}_t + \boldsymbol{b} \tag{11}$$

---

* Feature-space adaptation using multiple transformation matrices for DNN. by KANAGAWA, Hiroki and TACHIOKA, Yuuki (Mitsubishi Electric Corporation) and WATANABE, Shinji (Mitsubishi Electric Research Labratory) and ISHII, Jun (Mitsubishi Electric Corporation)

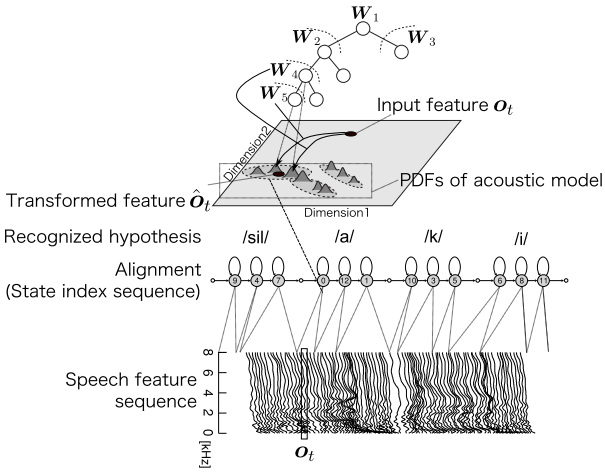

Fig. 1　Outline of the proposed method.

である．したがって適応処理を，デコード処理と独立したフロントエンドの処理として切り分けられるという利点がある．この手法は fMLLR と呼ばれ，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，広く使われている．しかし単一の変換行列しか使えないため，モデル空間の適応よりも性能面で劣るという短所もある．

3　複数の変換行列を用いた特徴量空間の適応法

3.1　複数の変換行列の重み付け

Fig. 1 に提案手法の概要図を示す．この図は「あき」と発話したとき，複数の CMLLR の変換行列を音声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毎に割り当てる．この割り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[1]を用いる．Fig. 1 では，アラインメントを $S = \{s_t | t = 1, \dots T\}$ のように，状態インデックス系列として表現している．ここで T はフレーム数である．s_t を得ることで，対応する GMM の集合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 $\{r(m, s_t)\}_{m \in \mathcal{M}_{s_t}}$ 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 $\left\{\boldsymbol{W}_{r(m,s_t)}\right\}_{m \in \mathcal{M}_{s_t}}$ を対応付けることができる．

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毎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 に従来の適応手法を適用するには，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て実現する必要がある．したがってこれら複数の変換行列の重み付け和をとり，単一の変換行列を推定する．式 (11) 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．

$$\hat{\boldsymbol{o}}_t = \sum\nolimits_{m \in \mathcal{M}_{s_t}} \rho(m, s_t, \boldsymbol{o}_t) \left(\boldsymbol{A}_{r(m,s_t)} \boldsymbol{o}_t + \boldsymbol{b}_{r(m,s_t)}\right) \tag{12}$$

[1]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．これらはあるフレームにおいて，音響特徴量と複数の HMM の状態を代用できるからである．

Algorithm 1 The proposed feature transformation algorithm.

Require: Acoustic feature sequence $\boldsymbol{O} = \{\boldsymbol{o}_t | t = 1, \dots, T\}$ and GMM acoustic model parameters λ
Obtain state sequence $S = \{s_t | t = 1, \dots T\}$ at the first-pass decoding ($S = \mathrm{decode}(\boldsymbol{O})$)
Estimate transformation matrices $\bar{\mathcal{W}}$ by Eq. (8)
for $t = 1, \cdots, T$ **do**
　for $m \in \mathcal{M}_{s_t}$ **do**
$$\hat{\boldsymbol{o}}_t = \sum\nolimits_{m \in \mathcal{M}_{s_t}} \rho(m, s_t, \boldsymbol{o}_t) \left(\boldsymbol{A}_{r(m,s_t)} \boldsymbol{o}_t + \boldsymbol{b}_{r(m,s_t)}\right)$$
$$= \sum\nolimits_{m \in \mathcal{M}_{s_t}} \rho(m, s_t, \boldsymbol{o}_t) \boldsymbol{W}_{r(m,s_t)} \begin{bmatrix} \boldsymbol{o}_t \\ 1 \end{bmatrix}$$
　end for
end for
Second-pass decoding with $\hat{\boldsymbol{O}} = \{\hat{\boldsymbol{o}}_t | t = 1, \dots, T\}$ (GMM/DNN)

ここで $\rho(m, s_t, \boldsymbol{o}_t)$ は，フレーム依存の重みパラメータであり，状態 s_t，GMM の混合 m に対応付けられられている．重みパラメータ $\rho(m, s_t, \boldsymbol{o}_t)$ について，3.2 節で詳細に議論する．この適応は特徴量空間で動作するため，正確な特徴量の変換が GMM 同様 DNN においても実現できる．

アルゴリズム 1 に，提案する fSMAPLR の手順を示す．まず 1 パス目のデコードにより全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．S と $\bar{\mathcal{W}}$ が得られたら，元の音響特徴量 $\boldsymbol{o}_t$ を式 (12) 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．

3.2　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 $\hat{\boldsymbol{o}}_t$ を重み付けされた特徴量変換行列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 つめの重みパラメータ $\rho(m, s_t, \boldsymbol{o}_t)$ として，GMM の混合要素 m に対する事後確率 $\gamma_{m,s_t}(\boldsymbol{o}_t)$ 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，$\gamma_{m,s_t}(\boldsymbol{o}_t)$ は

$$\gamma_{m,s_t}(\boldsymbol{o}_t) = \frac{w_{m,s_t} \mathcal{N}\left(\boldsymbol{o}_t | \boldsymbol{\mu}_{m,s_t}, \boldsymbol{\Sigma}_{m,s_t}\right)}{\sum_{m' \in \mathcal{M}_{s_t}} w_{m',s_t} \mathcal{N}\left(\boldsymbol{o}_t | \boldsymbol{\mu}_{m',s_t}, \boldsymbol{\Sigma}_{m',s_t}\right)} \tag{13}$$

により計算される．ここで，未適応の平均ベクトル $\boldsymbol{\mu}_{m,s_t}$ と対角共分散行列 $\boldsymbol{\Sigma}_{m,s_t}$ を用いる．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 (13) 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 (12) で複数の変換行列に拡張した利点を活かしきれない．

2 つめの重みパラメータとして，GMM の混合重みを用いる．これは式 (13) において，$\mathcal{N}\left(\boldsymbol{o}_t | \boldsymbol{\mu}_{m,s_t}, \boldsymbol{\Sigma}_{m,s_t}\right)$

の項を無視し，下記の近似を行うことに等しい．

$$\gamma_{m,s_t}(\boldsymbol{o}_t) \cong \frac{w_{m,s_t}}{\sum_{m' \in \mathcal{M}_{s_t}} w_{m',s_t}} = w_{m,s_t} \qquad (14)$$

このアプローチを採るのは，$\gamma_{m,s_t}(\boldsymbol{o}_t)$ がスパースとなってしまうことを避けるためである．加えて本アプローチには，式 (13) と異なり尤度計算をする必要がないという利点がある．

4 騒音環境下音声認識実験

4.1 実験条件

騒音下音声認識にて提案手法の有効性を示すため，第 2 回 CHiME チャレンジ [7] の Track 2 を用いる．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採られている．学習セット (si_tr_s) は 83 話者の 7,138 発話 (si84)，評価セット (si_et_05) は 12 話者の 330 発話 (Nov'92)，開発セット (si_dt_05) は 10 名の 409 発話から構成される．GMM と DNN の音響モデルは st_tr_s を用いて学習した．各評価話者の全発話を，適応データおよび評価データとして使用する．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：SNR)−6, −3, 0, 3, 6, 9 dB で重畳し作成した．重畳された騒音は非定常的なものであり，例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 [8] 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，開発セットにて単語誤り率 (word-error rate：WER) が最小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 LDA(linear discriminant analysis)[9] により 40 次元に圧縮する[2]．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)[10] 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 [11] 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 から成る 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対角共分散では精度良く表現できないため，GMM ではフィルタバンク特徴量をモデル化できない [12]．この制約より，フィルタバンクを用いた 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 [13]．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 [12] と同様に適用する．

[2]LDA には動的特徴量は使用していない．

Table 1 WER (%) for the development set with the Track 2 of the second CHiME Challenge when a posterior (Eq. (13)) or a mixture weight (Eq. (14)) is used for the weight in Eq. (12).

the number of transformation matrices	weight	τ (SMAP scale) 0	100	1000
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**

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

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

* significant at the 0.05 level.

音響モデルの学習には文献 [8] 同様，Kaldi ツールキット [14] 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．音響モデルの学習とデコードには Kaldi ツールキット [14] を用い，音響モデルは文献 [8] と同様の手順で学習した．

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，3.2 節で述べた 2 種類の重みパラメータについて検討する．Table 1 に変換行列数が 5, 10 のときの，事後確率 (式 (13)) と GMM の混合重み (式 (14)) 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 0, 100, 1,000 とした．これらの結果から，3.2 節で述べたように，事後確率は各混合間でスパースとなるため，混合重みの方が事後確率より優れることがわかった．なお最適な変換行列数と τ は以降の節にて詳細に議論する．提案する fSMAPLR には本節以降，結果の良かった式 (14) を用いることとする．

4.3 GMM 音響モデルにおける評価

開発セット (si_dt_05) および評価セット (si_et_05) にて性能を評価する．変換行列数および SMAP 係数は，それぞれ開発セット (si_dt_05) での平均 WER が最小であった 5 と 1,000 に固定した．複数の変換行列を用いたモデル空間の適応も有効性も示すため，CSMAPLR[5] についても評価した．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．Table 2 に各 SNR の WER を，また平均 WER を "avg." として示す．

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 に対し全ての SNR で上回り，評価セットの平均 WER で 1.2%（絶

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

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

* significant at the 0.05 level.

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

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

* significant at the 0.05 level.

対値）優れ，5%水準で有意であることがわかった．fSMAPLR は CSMAPLR と同程度の性能であることがわかり，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．

4.4 DNN 音響モデルにおける評価

4.4.1 MFCC 特徴量

本節では，DNN の音響モデルに対する評価を行う．なお変換行列数および SMAP 係数は，それぞれ開発セット (si_dt_05) での平均 WER が最小であった 3 と 1,000 に固定した．開発セットと評価セット (si_et_05) にて提案法の fSMAPLR を，ベースライン (適応なし)，fMLLR と比較し，Table 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．Table 2 と比較すると，DNN は GMM より全てのケースにおいて性能が優れた．

結果より，DNN においても適応が有効であることがわかった．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，5%水準で有意であった．

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) にて提案法の fSMAPLR を，ベースライン (適応なし)，fMLLR と比較し，Table 3 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) 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わかった．

5 おわりに

本報では，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[15, 16] 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 での話者適応学習が挙げられる．

参考文献

[1] T. Yoshioka *et al*, “Investigation of unsupervised adaptation of DNN acoustic models with filter bank input,” Proc. ICASSP, pp.13–16, 2014.

[2] 金川 他, “回帰木に基づく CMLLR 変換行列の特徴量への適用法,” 日本音響学会研究発表会講演論文集 (春季), pp.13–14, 2015.

[3] M. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition.,” Computer Speech and Language, vol.12, pp.75–98, 1998.

[4] M. Gales, “The generation and use of regression class trees for MLLR adaptation,” Technical Report CUED/F-INFENG/TR, vol.263, 1996.

[5] J. Yamagishi *et al*, “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm,” IEEE Trans. Speech and Audio Processing, vol.17, no.1, pp.66–83, 2009.

[6] O.Siohan *et al*, “Structural maximum a posteriori linear regression for fast HMM adaptation,” Computer Speech and Language, vol.16, pp.5–24, 2002.

[7] E. Vincent *et al*, “The second ‘CHiME’ speech separation and recognition challenge: datasets, tasks and baselines,” Proc. ICASSP, pp.126–130, 2013.

[8] Y. Tachioka *et al*, “Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark,” The 2nd International Workshop on Machine Listening in Multisource Environments, 2013.

[9] R. Haeb-Umbach and H. Ney, “Linear discriminant analysis for improved large vocabulary continuous speech recognition,” Proc. ICASSP, pp.13–16, 1992.

[10] M. Gales, “Semi-tied covariance matrices for hidden Markov models,” IEEE Trans. Speech and Audio Processing, vol.3, no.7, pp.272–281, 1999.

[11] T. Anastasakos *et al*, “A compact model for Speaker-Adaptive Training,” Proc. ICSLP, pp.1137–1140, 1996.

[12] T. Sainath *et al*, “Improvements to deep convolutional neural networks for LVCSR,” Proc. ASRU, pp.315–320, 2013.

[13] T. Sainath *et al*, “Deep convolutional neural networks for LVCSR,” Proc. ICASSP, pp.8614–8618, 2013.

[14] D. Povey *et al*, “The Kaldi speech recognition toolkit,” Proc. ASRU, pp.1–4, 2011.

[15] S. Watanabe *et al*, “Bayesian linear regression for hidden Markov model based on optimizing variational bounds,” Proc. MLSP, pp.1–6, 2011.

[16] S.-J.Hahm *et al*, “Feature space variational Bayesian linear regression and its combination with model space VBLR,” Proc. ICASSP, pp.7898–7902, 2013.