事前情報を利用した基底 fMLLR のための重み係数推定法*

◎金川裕紀，太刀岡勇気，石井純 (三菱電機)

1 はじめに

騒音、残響下などでは、話者適応が音声認識性能の向上に有効である。中でも特徴量最尤線形回帰 (feature-space maximum likelihood linear regression : fMLLR)[1, 2] が最もよく用いられている。fMLLR は特徴量ベクトルに対して行列を乗算するだけでよく、音響モデルを変化させなくてよいため、DNN-HMM の音響モデルに入力することができ、汎用性に優れている。しかし fMLLR は、1 発話などの極めて少量の適応データからでは適応が過学習してしまい、適切な変換行列が求められないことが知られている。この問題を解決する方法の 1 つとして基底 fMLLR[3] が提案されている。基底 fMLLR は事前に学習した複数の基底行列を用いて、適応時は基底行列への重みを求める。変換行列の全要素を適応データのみから求める fMLLR とは異なり、基底 fMLLR が求めるのは重みだけでよいので推定パラメータ数が比較的少なく過学習に頑健である。

本報は、従来適応データのみから基底行列の重みを求めていたところに、学習データから得られる事前情報を反映させる手法を提案する。実験により提案法の有効性を示す。

2 fMLLR

fMLLR は、GMM-HMM の音響モデルの平均ベクトルと共分散行列を共通の行列で変換する制約付き MLLR[1, 2] の行列数を単一にした場合の手法であり、次式にて時刻 t の特徴量ベクトルを $\boldsymbol{o}_t$ を $\hat{\boldsymbol{o}}_t$ にアフィン変換する。

$$\hat{\boldsymbol{o}}_t = \boldsymbol{A}\boldsymbol{o}_t + \boldsymbol{b} = \boldsymbol{W} \begin{bmatrix} \boldsymbol{o}_t^\top & 1 \end{bmatrix}^\top \tag{1}$$

ここで $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b} \end{bmatrix}$、$\top$ はそれぞれ時刻 t の変換行列、転置を示す。fMLLR はこの変換行列 $\boldsymbol{W}$ を適応データのみから求める。パラメータ数は行列の要素数に等しく、特徴量を D 次元のベクトルとすると $D(D+1)$ 個となり、これを適応データのみから求めるため、極少量データの場合は過学習してしまう。

3 基底 fMLLR

少量データに頑健でない fMLLR に対し、基底 fMLLR[3] は推定すべきパラメータ数を少なくするため、直接変換行列 $\boldsymbol{W}$ を推定するのではなく、N 個の基底行列 $\boldsymbol{W}_{1:N_{\max}}$ の重み付けにより表現する。ここで $N_{\max} = D(D+1)$ である。基底 fMLLR は大きく分けて 2 つのステップにより実現される。ここで学習データを効率的に正規化できる $N_{\max}$ 個の基底を求め、その特異値の降順にインデックス n が $1 \leq n \leq N_{\max}$ で振られる。1 つ目は学習ステップであり、学習データから複数の基底行列 $\boldsymbol{W}_{1:N_{\max}}$ を求める。

基底行列 $\boldsymbol{W}_{1:N_{\max}}$ は、学習データの話者 s に対してそれぞれ求めた $\mathcal{Q}$ 関数の勾配より導かれる $N_{\max}$ 次元のベクトル $\boldsymbol{p}_{(s)}^{\text{train}}$ の自己相関行列 $\boldsymbol{M}_{(s)} = \boldsymbol{p}_{(s)}^{\text{train}} \boldsymbol{p}_{(s)}^{\text{train}\top}$ を全話者について足して得られた行列 $\boldsymbol{M}$ を特異値分解することにより得られる。特異値分解により、行列 $\boldsymbol{M}$ を構成するうえで寄与度が高い基底行列 $\boldsymbol{W}_n$ を求めることができる。

2 つ目は適応ステップであり、適応データを用いて式 (2) により変換行列 $\boldsymbol{W}$ を反復法により推定する。

$$\boldsymbol{W} := \boldsymbol{W}_{\text{prev}} + \sum_{n=1}^{N} d_n \boldsymbol{W}_n \tag{2}$$

ここで $\boldsymbol{W}_{\text{prev}}$、$N$ はそれぞれ 1 反復前の変換行列、使用する基底行列数であり、基底行列 $\boldsymbol{W}_n$ への重み d_n は式 (3) で求められる。

$$d_n = \text{tr}\left(\boldsymbol{W}_n^\top \boldsymbol{P}^{\text{adapt}}\right) \tag{3}$$

ここで $\boldsymbol{P}^{\text{adapt}} \in \mathbb{R}^{D\times(D+1)}$ は、適応話者に対する $\mathcal{Q}$ 関数の勾配より導かれる行列であり、$\boldsymbol{W}$ と適応データから求められる。$\boldsymbol{W}_{\text{prev}}$ の初期値は単位行列 $\boldsymbol{I} \in \mathbb{R}^{D\times(D+1)}$ である。また文献 [3] では過学習を防ぐため、適応データ量の状態占有確率のフレーム和 β に応じて使用する基底行列数を式 (4) で示される N に制限しており、本報告もそれに倣った。

$$N = \min(\eta\beta, N_{\max}) \tag{4}$$

ここで $\eta = 0.2$ である。反復により求められる尤度の上がり幅が閾値より小さくなったとき、また反復が規定の回数に達したときに反復を打ち切る。

4 基底 fMLLR への事前情報の導入

3 節では、基底行列 $\boldsymbol{W}_{1:N_{\max}}$ と変換行列 $\boldsymbol{W}$ の推定法について述べた。基底行列は学習データへの寄与の降順にインデックスが付与されているが、学習データに対する基底行列の寄与度の高低に依らず、適応時

*Estimation method of weight parameters for basis fMLLR by using prior information. by KANAGAWA, Hiroki and TACHIOKA, Yuuki and ISHII, Jun (Mitsubishi Electric Corporation)

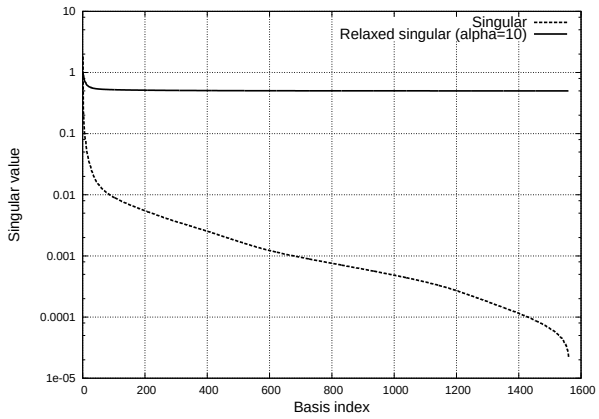

Fig. 1　Singular values and relaxed ones with sigmoid function (Eq. (5)).

には式 (3) のように同じ重みで扱っているため、パラメータ推定がうまくいかない場合がある。そこで適応時に、基底行列学習時に得た特異値 $\sigma_{1:N_{\max}}$ を事前情報として用いることにする。ただし先頭の数個の基底行列を除き、ほとんどの特異値は非常に小さく、基底値が大きい基底行列のみに依存してしまうので次式のシグモイド関数により特異値の影響を調整する。

$$\rho_n = \left(1 + e^{-\alpha\sigma_n}\right)^{-1} \tag{5}$$

図 1 の破線と実線に、それぞれ特異値 $\sigma_{1:N_{\max}}$ と、式 (5) で $\alpha = 10$ としたときの $\rho_{1:N_{\max}}$ を示す。横軸は基底行列のインデックス n である。n の増大に伴い、破線に示す特異値 $\sigma_{1:N_{\max}}$ は急激に小さい値を取るが、実線に示す $\rho_{1:N_{\max}}$ は緩やかに減少し 0.5 に収束する。

提案法は式 (5) により得た $\rho_{1:N_{\max}}$ を用いて、式 (6) により変換行列 $\boldsymbol{W}$ および重み d_n を更新する。

$$\boldsymbol{W} := \boldsymbol{W}_{\text{prev}} + \sum_{n=1}^{N} \rho_n d_n \boldsymbol{W}_n \tag{6}$$

5　実験

5.1　音声認識タスク

発話内容は Wall Street Journal で (WSJCAM0)、以下の 2 種があり、本報では比較的定常的な騒音が存在する室の実測データである “RealData” を使用する [4]。学習セット、開発セット、評価セットが提供され、音響モデルは学習セットにより学習し、言語モデル重みと基底行列 $\boldsymbol{W}_n$ は、開発セットの単語誤り率 (WER) で調整し、式 (5) における最適な α は 10 であった。語彙は 5 k で、tri-gram 言語モデルを使った。適応は発話単位で行われ、1 発話 (5〜6 秒) の音声のみから変換行列が推定される。なお音声は 1ch で、信号処理には [5] を使用した。音響特徴量は、13 次元の MFCC とその動的特徴量 (Δ, $\Delta\Delta$) である。

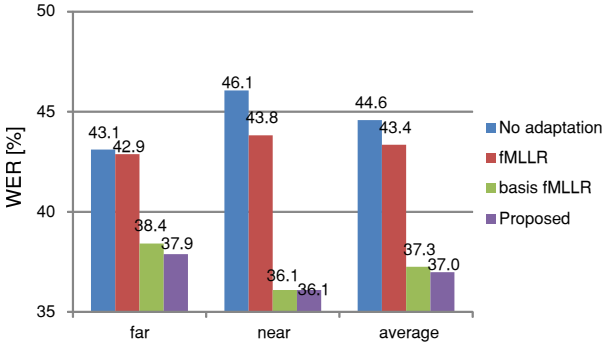

Fig. 2　Average WER[%] on the REVERB Challenge (evaluation set).

5.2　音声認識実験

評価セットにおいて提案法 (Proposed) を、適応なし (No adaptation)、fMLLR、基底 fMLLR(basis fMLLR) と比較する。図 2 に遠距離音声 (far)、近距離音声 (near) に対する WER、またその平均 WER(average) を示す。適応なしと比較し fMLLR が優れたが、そのゲインは小さかった。これは適応データが非常に少なく適応が過学習し、十分な性能が得られていないためである。一方、基底 fMLLR は fMLLR と比較して大幅に WER を改善し、少量データでの適応で有効であった。さらに提案法は、近距離 (near) では基底 fMLLR と WER が同じで、遠距離 (far) では基底 fMLLR より絶対値で 0.5% 優れており、事前情報の使用の有効性を確認した。

6　おわりに

基底行列生成時の特異値を、適応時に事前情報として導入した。残響下において fMLLR、基底 fMLLR よりも提案法が有効であることを示した。今後の課題は、パラメータ α の適切な設定法である。

参考文献

[1] M.J.F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition.,” Computer Speech and Language, **12**, 75–98 (1998).

[2] V. Digalakis, D. Ritischev, and L. Neumeyer, “Speaker adaptation using constrained estimation of Gaussian mixtures,” IEEE Trans. Speech and Audio Processing, **3**, 357–366 (1995).

[3] D. Povey and K. Yao, “A basis representation of constrained MLLR transforms for robust adaptation,” Computer Speech and Language, **26**, 35–51 (2012).

[4] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, “The REVERB Challenge: A common evaluation framework for dereverberation and recognition of reverberant speech,” Proc. WASPAA, 1–4 (2013).

[5] 太刀岡勇気, 成田知宏, 渡部晋治, “残響除去手法とシステム統合手法の種々の残響環境に対する有効性: REVERB チャレンジ,” 情報処理学会研究報告, SLP-105(6), 1–6 (2015).