基底 fMLLR のための効率的な基底行列選択*

◎金川裕紀、太刀岡勇気、成田知宏 (三菱電機)

1 はじめに

騒音や残響が存在する実環境下では適応が音声認識性能の向上に有効であり、中でも特徴量空間最尤線形回帰 (feature-space maximum likelihood linear regression : fMLLR)[1, 2] がよく用いられている。fMLLR は入力音声の特徴量ベクトルに対して変換行列を用いて適応処理を行うものであり、音響モデルを適応する必要がない。このため高い認識性能が得られる深層ニューラルネットワーク（deep neural network : DNN）音響モデル [3] に対しても利用可能であり、汎用性に優れている [4, 5]。しかし fMLLR は、推定すべきパラメータが多いことから 1 発話などの極めて少量の適応データでは過学習してしまい、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[6] 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。基底 fMLLR をより高度化するため、以前筆者らは推定ステップに、学習データから得られる事前情報を加味して重み係数を算出することで性能が向上することを確認した [7]。今回は、類似する基底行列同士をクラスタリングにより集約することで、従来の基底 fMLLR では過学習への対策として使用を制限されていた基底行列を活用する手法を提案し、実験により提案法の有効性を確認する。

2 従来の適応手法

本章では従来の適応手法について説明する。従来法として特徴量空間の最尤線形回帰 (fMLLR) および基底 fMLLR の概要を、それぞれ 2.1 節および 2.2 節にて述べる。

2.1 特徴量空間の最尤線形回帰 (fMLLR)

fMLLR は、混合ガウス分布（Gaussian mixture model : GMM）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 $D \times (D+1)$ 個となり、これを適応データから求めるため、極少量データの場合は過学習してしまう。

2.2 基底 fMLLR

少量データに頑健でない fMLLR に対し、基底 fMLLR[6] 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、次式のように N 個の基底行列 $\boldsymbol{W}_n \in \mathbb{R}^{D\times(D+1)}$ $(1 \le n \le N_{\max})$ の重み付けにより表現する。

$$\boldsymbol{W} = \boldsymbol{I} + \sum_{n=1}^{N} d_n \boldsymbol{W}_n \tag{2}$$

ここで $N_{\max} = D\,(D+1)$ である。

基底 fMLLR は大きく分けて 3 つのステップにより実現される。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 $\boldsymbol{W}_n$ を求める。基底行列 $\boldsymbol{W}_n$ 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ベクトル $\boldsymbol{p}_{(s)}^{\text{train}} \in \mathbb{R}^{N_{\max}\times 1}$ の自己相関行列 $\boldsymbol{M}_{(s)} = \boldsymbol{p}_{(s)}^{\text{train}} \boldsymbol{p}_{(s)}^{\text{train}^\top} \in \mathbb{R}^{N_{\max}\times N_{\max}}$ を全話者について足して得られた行列 $\boldsymbol{M} \in \mathbb{R}^{N_{\max}\times N_{\max}}$ 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 (3) により変換行列 $\boldsymbol{W}$ を反復法により推定する。

$$\boldsymbol{W} \leftarrow \boldsymbol{W} + \sum_{n=1}^{N} d_n \boldsymbol{W}_n \tag{3}$$

ここで記号 $\leftarrow$、N はそれぞれ更新式の記号、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 (4) で求められる。

$$d_n = \mathrm{tr}\left(\boldsymbol{W}_n^\top \boldsymbol{P}^{\text{adapt}}\right) \tag{4}$$

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 [6] では過学習を防ぐため、適応データ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 (5) で示される N に制限しており、本報告もそれに倣った。

$$N = \min\left(\eta\beta, N_{\max}\right) \tag{5}$$

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、もしくは反復

*Effective basis matrices selection for basis fMLLR. by KANAGAWA, Hiroki and TACHIOKA, Yuuki and NARITA, Tomohiro (Mitsubishi Electric Corporation)

が規定の回数に達したときに反復を打ち切る。3 つ目は認識ステップであり、fMLLR と同じく適応後の特徴量 $\hat{\boldsymbol{o}}_t$ と音響モデルを用いてデコードし、最終認識結果を得る。

3 効率的な基底行列の選択

3.1 全基底行列から選択する場合

2.2 節では、D 次元の特徴ベクトルに対して基底行列 $\boldsymbol{W}_n$ が $N_{\max}$（$= D(D+1)$）個求まることを述べた。しかし、実際に適応データが入力されるとき、これらの基底行列への重みの個数は入力フレーム数に応じて式 (5) の N 個に限定される。このため特異値の降順としたときのインデックス N 以降の基底行列を無視することで、変換行列の表現力が低下する恐れがある。

本報ではこの問題を低減するために、N 以降のインデックスを持つ基底行列も適応時に考慮されるよう、$N_{\max}$ 個の基底行列を任意の N_{target} 個にクラスタリングする。基底行列をクラスタリングするために、まず次式の距離行列 $\boldsymbol{Z} \in \mathbb{R}^{N_{\max} \times N_{\max}}$ を定義する。

$$\boldsymbol{Z} = \begin{bmatrix} z_{11} & \cdots & z_{1N_{\max}} \\ \vdots & \ddots & \vdots \\ z_{N_{\max}1} & \cdots & z_{N_{\max}N_{\max}} \end{bmatrix} \tag{6}$$

ここで要素 z_{mn} は基底行列 $\boldsymbol{W}_m$、$\boldsymbol{W}_n$ 間の距離を示す。距離尺度[1]には最大列和ノルム (1-ノルム)[2]

$$z_{mn} = \|\boldsymbol{W}_m - \boldsymbol{W}_n\|_1 \tag{7}$$

もしくは最大特異値ノルム (スペクトルノルム)[3]

$$z_{mn} = \|\boldsymbol{W}_m - \boldsymbol{W}_n\|_2 \tag{8}$$

を用いる。

Fig. 1 に全基底行列から選択する場合の概略図を示す。式 (6) の距離行列 $\boldsymbol{Z}$ を生成後、階層的クラスター分析により、N_{target} 個のクラスに分類する。分類後、そのクラスの中で最も特異値 σ_n が大きい基底行列を選択[4]してこれをクラスタリングの結果とし、N_{target} 個の基底行列を再構成する。

3.2 部分的に基底行列を選択する場合

3.1 節とは異なり、特異値の大きいの基底行列を信頼しつつ特異値の小さい基底行列を効率的に利用する方法として、基底行列すべてをクラスタリングするのではなく、部分的にクラスタリングする方法も考えられる。Fig. 2 に部分的に基底行列を選択する場合の概略図を示す。式 (6) の距離行列 $\boldsymbol{Z}$ を $N_{\max} - M$

[1]距離尺度として、ほかにもノロベニウスノルムなど他基準についても検討したが、予備検討の結果この 2 つに限定した。

[2]$\boldsymbol{A} \in \mathbb{R}^{I \times J}$ に対し、$\|\boldsymbol{A}\|_1 = \max_{1 \le j \le J} \sum_{i=1}^{I} |a_{ij}|$

[3]$\boldsymbol{A} \in \mathbb{R}^{I \times J}$ に対し、$\|\boldsymbol{A}\|_2 = \sigma_{\max}(\boldsymbol{A})$（ただし $\sigma_{\max}(\boldsymbol{A})$ は行列 $\boldsymbol{A}$ の最大特異値を表す。）

[4]予備実験にてクラスに属する基底行列の平均値をとったケースと比較したところ、こちらのほうがよかった。

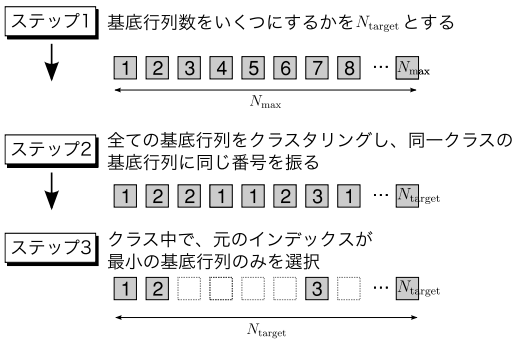

Fig. 1 An example of aggregating basis matrices from all basis matrices.

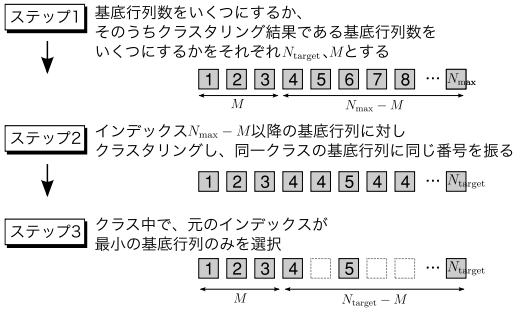

Fig. 2 An example of partial aggregating basis matrices.

以降のインデックスの基底行列について生成後、階層的クラスター分析により 3.1 節と同様に N_{target} 個のクラスに分類する。そのクラスの中でインデックスが最小の基底行列を選択してこれをクラスタリングの結果とし、N_{target} 個の基底行列を再構成する。具体的には使用する N 個の基底行列のうち、前半の M 個はオリジナルのものを用い、残りの $N - M$ 個の基底行列は、前半の M 個の基底行列を除く $N_{\max} - M$ 個の行列をクラスタリングすることで $N - M$ 個に集約することが考えられる。なお 3.1 節の方法は、本節において $M = 0$ の場合と等価であると見なすことができる。

4 実験

4.1 音声認識タスク

REVERB チャレンジ [8] のデータセットを用いた音声認識実験を行う。本データセットは残響下での音声認識タスクで構成され、発話内容は Wall Street Journal で (WSJCAM0) であり、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 “REAL-DATA” を使用する。学習セット、開発セット (dt)、評価セット (et) が提供され、音響モデルは学習セットにより学習し、言語モデル重みは開発セットの単語誤り率 (WER) で調整した。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 [9] 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,

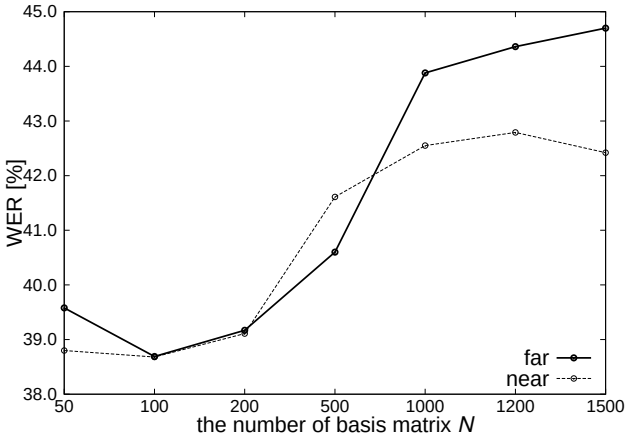

Fig. 3　WER [%] for the development set of the REVERB Challenge. Parametric study of the number of basis matrices.

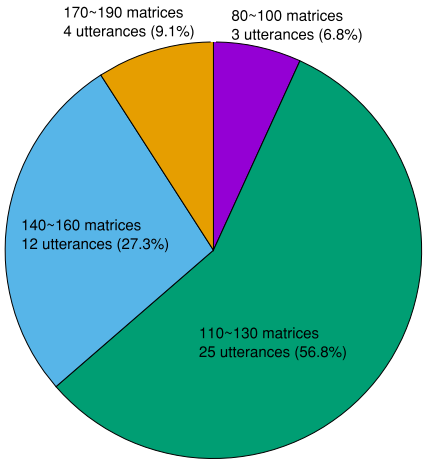

Fig. 4　Detail of the number of used basis matrices for development set.

ΔΔ) である。

4.2　基底行列の効率的な選択の効果

4.2.1　基底行列数と単語誤り率の関係

まず適切な基底行列数を決定する。基底 fMLLR と提案する基底行列の選択法と比較する前に、使用する基底行列数が性能に及ぼす影響を調査した。基底 fMLLR では通常、式 (5) により入力音声に応じて使用する基底行列数 N を制限するが、ここでは N を全発話で共通とした。Fig. 3 の横軸に使用した基底行列数、縦軸に開発セット (dt) における単語誤り率を示す。図中に示す far と near はそれぞれ遠距離音声、近距離音声を表す。

まず $N = 50$ と基底行列数が少ないとき、遠距離音声 (far) では単語誤り率が N =100、200 のときと比べて大きい。これは加重和での表現力が少ない基底のため限定されているためである。また $N \geq 200$ の結果を見ると基底行列数が大きくなるに従って単語誤り率が増加しているが、これは推定すべきパラメータの増加による過学習によるものと考えられ、基底行列数を適切に決定することが重要であることを確認した。最終的に開発セットにおいて遠距離、近距離音声の平均単語誤り率が小さかったのは $N = 100$、$N = 200$ の順であった。Fig. 4 に開発セットにおいて、基底 fMLLR で使われた基底行列と、そのときの発話数と、およびそれが発話全体に占める割合を円グラフに示す。今回のタスクでは Fig. 4 から発話が短く、すべての発話において基底行列を 200 個未満しか使用していないことから、200 個以上の基底行列に対して式 (4) の重み係数 d_n を求めることは有効でないことがわかった。

4.2.2　基底行列の効率的な選択

次に 4.2.1 節の知見をもとに、基底行列数が 100 個または 200 個となるよう 3.1 節の方法により集約し、

Table 1　Average WER [%] for isolated speech (dt) with respect to distance measures and the number of basis matrices N_{target}.

距離尺度	基底行列数	平均 WER	
	N_{target}	far	near
最大列和	100	39.5	39.4
ノルム（式 (7)）	200	39.1	39.8
最大特異値	100	38.8	39.0
ノルム（式 (8)）	200	**38.7**	**38.7**

距離尺度と基底行列数の関係を調査した。基底行列間の距離尺度には最大列和ノルムと最大特異値ノルムを、クラスタリング法には ward 法 [10, 11] を用いた。集約後の基底行列数を N_{target}、クラスタリングされた基底行列数を M とするときの、基底行列数 N_{target} に占めるクラスタリング基底行列の割合=(M/N_{target}) が $N = 100$ で 0, 30, 50, 80, 100%、$N = 200$ で 0, 25, 40, 50, 75, 100% となるような条件の平均単語誤り率を開発セットについて求めた。距離尺度、基底行列数 N_{target} ごとに求めた単語誤り率を Table 1 に示す。

結果から距離尺度が最大特異値ノルムで、基底行列数 N_{target} が 200 のとき、遠距離音声、近距離音声ともに単語誤り率が最小であった。また結果から、最大列和ノルムより最大特異値ノルムのほうが誤り率が小さいことがわかった。先の検討の基底行列数と単語誤り率の比較 (4.2.1 節参照) では基底行列数が 100 のときに単語誤り率最小であったが、基底行列を選択した後では基底行列数が 200 のときが優れた。これは従来利用できなかった後半インデックスの基底行列が基底選択により利用可能となったことにより加重和の表現力向上が、推定すべき重みパラメータ数の増加による過学習の影響を上回った結果と考えられる。

次に評価セット (et) にて、開発セットにて単語誤

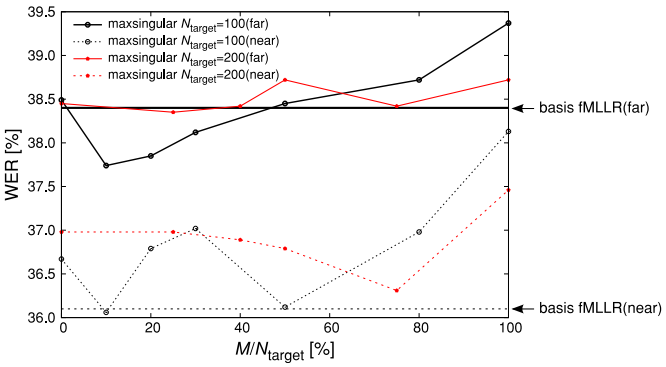

Fig. 5　WER [%] for the evaluation set of the REVERB Challenge. Parametric study of the ratio of the number of clustered basis matrices M to the number of basis matrices N_{target}.

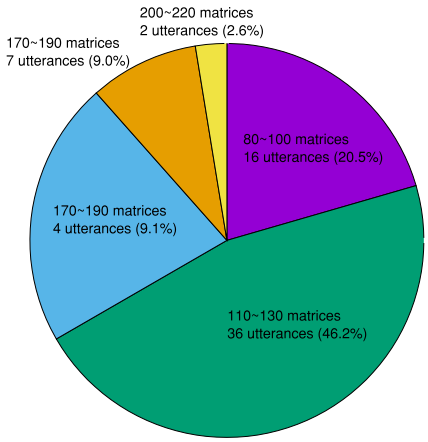

Fig. 6　Detail of the number of used basis matrices for evaluation set.

り率が小さかった最大特異値ノルムを距離尺度として、基底行列数が 100、200 の条件で評価した。Fig. 5 に基底行列数 N_{target} と、基底行列数 N_{target} に占めるクラスタリング基底行列の割合の関係を示す。横軸に基底行列数に占めるクラスタリング基底行列の割合 ($=M/N_{\text{target}}$) を、縦軸に単語誤り率を示す。凡例には遠距離音声 (far) か近距離音声 (near)、N_{target} が 100 か 200 かの各条件において、ベースラインの基底 fMLLR と基底選択を用いた方法を表した。結果から $M/N_{\text{target}} = 10$ で $N_{\text{target}} = 100$ のとき、ベースラインの基底 fMLLR を遠距離、近距離音声でともに優れた。またクラスタリングを実施しない場合が $M/N_{\text{target}} = 0$ の結果に相当する。M/N_{target} が 0 より大きい箇所で単語誤り率が小さい箇所があることから、クラスタリングされた基底行列数 M を適切に調整することでクラスタリングによる効果が得られることがわかった。基底行列数に注目すると $N_{\text{target}} = 100$ の場合は、M/N_{target} が 10〜30 で遠距離において優れたが、$N_{\text{target}} = 200$ の多くの場合 $N_{\text{target}} = 100$ より単語誤り率が大きかった。なお適応なし、fMLLR の単語誤り率はそれぞれ遠距離音声で 43.1%, 41.9%, 近距離音声で 43.4%、42.7%であり、基底 fMLLR および提案法はそれと比較して大きく単語誤りを削減したことがわかった。

Fig. 4 と同じく、評価セットに対しても使用される基底行列数の内訳を調査した結果を Fig. 6 に示す。図から、評価セットは開発セットと比較して基底行列数の内訳に差異が見られ、特に基底行列を 80 から 100 個使用する発話が多いことがわかった。Fig. 5 において N_{target} が 200 より 100 のほうが結果が良いことから、発話長に応じて式 (5) 同様、N_{target} を適切に設定することが有効であることが考えられる。

5　おわりに

基底行列間の距離に基づくクラスタリングにより基底行列を選択し、従来無視されていた基底行列を活用する方法を提案した。提案法はクラスタリングによる基底行列数を適切に設定することで、残響下音声認識において適応なし、fMLLR、基底 fMLLR よりも有効であった。今後は、以前に提案した事前情報を用いた基底 fMLLR[7] と基底行列の選択を併用した場合の評価、および選択する基底行列数の適切な設定法を提案する。

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] G. Hinton, L. Deng, D. Yu, G. Dahl, A. rahmanMohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” IEEE Signal Processing Magazine, **28**, 82–97 (2012).

[4] T. Yoshioka, A. Ragni, and M.J.F. Gales, “Investigation of unsupervised adaptation of DNN acoustic models with filter bank input,” Proc. ICASSP, 13–16 (2014).

[5] H. Kanagawa, Y. Tachioka, S. Watanabe, and J. Ishii, “Feature-space structural MAPLR with regression tree-based multiple transformation matrices for DNN,” Proc. APSIPA, 86–92 (2015).

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

[7] 金川裕紀，太刀岡勇気，石井純，“事前情報を利用した基底 fMLLR のための重み係数推定法，” 日本音響学会研究発表会講演論文集 (春季)，41–42 (2016).

[8] 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).

[9] 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).

[10] J.H. Ward, “Hierarchical groupings to optimize an objective function,” Journal of the American Statistical Association, **58**, 234–244 (1963).

[11] S.D. Kamvar, D. Klein, and C.D. Manning, “Interpreting and extending classical agglomerative clustering algorithms using a model-based approach,” Proc. of The 19th Int’l Conf. on Machine Learning, 283–290 (2002).