# 回帰木に基づく CMLLR 変換行列の特徴量への適用法*

◎金川裕紀，太刀岡勇気，石井純 (三菱電機)

## 1 はじめに

騒音下や，未知の話者に対する音声認識では，話者適応が有効である．話者適応には大きく分けて 2 つの手法があり，1 つは音響モデルを入力特徴量にマッチさせるよう変換するモデル空間での適応 (モデル適応) 手法，もう 1 つは入力特徴量を音響モデルにマッチさせるよう変換する特徴量空間での適応 (特徴量適応) 手法である．

MLLR（Maximum Likelihood Linear Regression）[1] に代表されるモデル適応では，音素などのコンテキスト情報とモデルパラメータが関連付けられていることから，コンテキストに応じて異なる変換行列を使い分けることが可能である．この性質を利用した適応性能高度化の試みとして，文献 [2] では複数の変換行列を回帰木で共有化し，コンテキストに応じた変換行列を用いてモデルパラメータを適切に変換する手法を提案している．

特徴量適応 (fMLLR : feature-space MLLR) では，あらかじめ特徴量に対し変換行列を適用しておけばよく，CMLLR(Constrained MLLR)[3] 等で生成した単一の変換行列を全フレームの特徴量に適用することで，適応を音声認識のフロントエンドの処理として切り分けることができる．この切り分けにより，GMM とは異なる基準の音響モデルを使用することが容易になる．例えば単一の fMLLR 変換行列による適応をフロントエンドに用い，変換行列適用後の特徴量を DNN(Deep Neural Network) に入力する試みがなされている [4]．特徴量適応にも回帰木を用いることでさらなる性能の向上が期待できるが，モデル適応のように変換行列と変換の対象が一意に対応づけられている必要があるため，これまでは回帰木について求めた変換行列を直接特徴量に適用できず，CMLLR 変換行列を用いた適応はモデル適応に限定されていた．

そこで本稿ではモデル適応と特徴量適応の利点の両立を目的として，回帰木について求めた複数の CMLLR 変換行列を特徴量に適用する手法を提案する．

## 2 CMLLR による話者適応

CMLLR では，特徴量を音響モデルにマッチするような変換を行う変換行列を求める．次元 $D$，時刻 $t$ の音響特徴量ベクトル $\boldsymbol{o}_t \in \mathbb{R}^{D\times 1}$ に対し，アフィン変換 $\boldsymbol{W}_{r(m,j)} = \left[\boldsymbol{A}_{r(m,j)}\, \boldsymbol{b}_{r(m,j)}\right] \in \mathbb{R}^{D\times(D+1)}$ を用いて次式の変換を行う．

$$\hat{\boldsymbol{o}}_{r(m,j),t} = \boldsymbol{A}_{r(m,j)}\boldsymbol{o}_t + \boldsymbol{b}_{r(m,j)} \qquad (1)$$

ここで $r\,(m,j)$ は回帰クラスのインデックスであり，GMM(Gaussian Mixture Model) の分布番号 $m$，HMM(Hidden Markov Model) の状態番号 $j$ から一意に対応づけられている．単一の変換行列を用いる場合，変換行列が回帰クラス $r$ に依存しないため全フレーム $(1 \le t \le T)$ の特徴量に対して適用しておけばよく，特徴量適応となる．CMLLR を回帰木に基づき求める場合は，$\boldsymbol{o}_t$ と $r\,(m,j)$ が一意に対応付けられていないためモデル適応となり，学習器や認識器内部で尤度計算時に式 (1) の変換を行うこととなる．

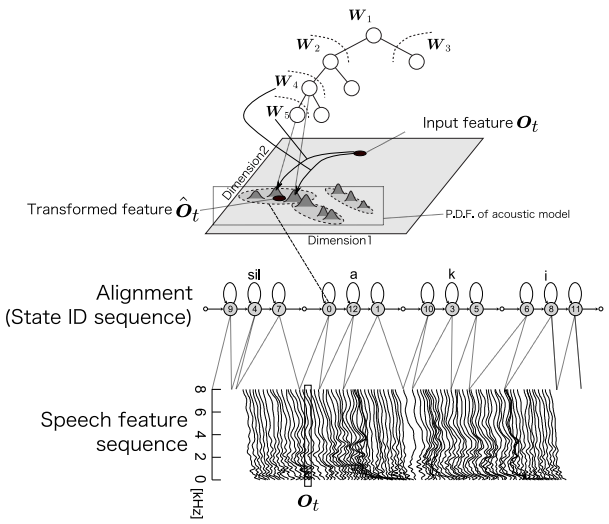


Fig. 1 An outline of the proposed method.

## 3 回帰木に基づく CMLLR 変換行列の特徴量への適用法

Fig.1 に提案法の概略を示す．図では「あき」と発話した時，木構造に基づき求めた CMLLR 変換行列を特徴量に適用する場合の動作を示している．

本手法では，フレーム毎に変化する音響特徴量に対して変換行列を切り替えて使用するために，音響特徴量と変換行列との対応をとる．この対応関係を得るため，アラインメントを利用する．図にはアラインメントとして HMM の状態番号系列を示しており，状態番号 $j$ から GMM を取得することができる．したがって状態番号 $j$ と GMM の各分布番号 $m$ から回帰木の変換行列 $\boldsymbol{W}_{r(m,j)}$ を定めることができ，音響特徴量と変換行列の対応をとることができる．

対応付けによって音響特徴量 $\boldsymbol{o}_t$ には，GMM の混合数 $M$ 個の変換行列が割り当てられる．ここで，これらの変換行列を用いて $\boldsymbol{o}_t$ を変換するため次式の変換を行う．

$$\hat{\boldsymbol{o}}_t = \sum_{m=1}^{M} w_m \left(\boldsymbol{A}_{r(m,j)}\boldsymbol{o}_t + \boldsymbol{b}_{r(m,j)}\right) \qquad (2)$$

ここで $w_m$ は混合インデックス $m$ の分布に対する重みである．

---

* A feature-space speaker adaptation technique by applying regression tree-based CMLLR transformation matrices to speech features. by KANAGAWA, Hiroki, TACHIOKA, Yuuki and ISHII, Jun (Mitsubishi Electric Corp.)

デコード時には，式 (2) により得た変換後の音響特徴量を認識器に入力するだけでよく，モデル適応のように尤度計算の毎に変換行列を適用する必要がないという利点がある．

手順を以下にまとめる．

1. 音声特徴量を認識器に入力し，認識結果とアラインメントを得る．
2. 音響モデルから状態番号と分布を対応付けた回帰木を生成する．
3. 認識結果と音声特徴量と回帰木を用いて変換行列を推定する．
4. アラインメントと音声特徴量と回帰木，変換行列を用いて式 (2) により，変換特徴量 $\hat{\boldsymbol{o}}_t$ を得る．
5. $\hat{\boldsymbol{o}}_t$ を用いてデコードし，最終的な認識結果を得る．

## 4 実験

### 4.1 実験条件

提案手法の有効性を評価するため，第 2 回 CHiME チャレンジ [5] の Track2 における騒音重畳データ (isolated) により評価した．なお使用する騒音重畳データに対し，事前分布に基づくバイナリマスクの騒音抑圧処理 [6] をフロントエンドで適用する．

Track2 の評価セット (si_et_05) には 12 話者の 330 発話が含まれており，発話は Wall Street Journal データベース (WSJ0) から取られている．評価および話者適応における変換行列生成には，各評価話者に対する全発話を使用した．重畳されている騒音は他の話者の発話や，家庭内の騒音等の非定常性のものである．評価においては，これらの騒音を信号対雑音比 (SNR) が 0，3，6，9dB になるように重畳したデータを使用する．音響モデルは状態数 2,500，ガウス分布の全体数が 15,000 のトライフォンモデルとし，Track2 の学習セット (si_tr_s) に含まれる 83 話者の 7,138 発話で学習した．音響特徴量には，13 次元の MFCC とその Δ および ΔΔ から成る 39 次元のベクトルを使用した．言語モデルにはサイズが 5k のものを使用した．言語モデル重みは，Track2 の開発セット (si_dt_05) を用いて調整した．

### 4.2 実験結果

図 2 に各手法における，各 SNR の単語誤り率 (WER) の平均値を示す．縦軸は単語誤り率を示す．縦軸の w/o adaptation, fMLLR Global, CMLLR Tree, fMLLR Tree はそれぞれ話者適応なし，全フレームの音響特徴量に対して単一の変換行列を用いて変換する手法，回帰木を用いてモデル適応する手法，回帰木を用いて特徴量適応する手法を意味する．また凡例の w/o DT，w/ DT はそれぞれ音響モデルの学習に識別学習を用いたか否かを示す．

図より話者適応の有無を比較すると，話者適応の有無で WER が識別学習なしで 4〜6% 程度，識別学習ありで 7〜8%程度改善しており，話者適応の有効性が確認できる．また fMLLR Global と CMLLR Tree, fMLLR Tree を比較すると回帰木を使用する手法の WER が低く，複数の変換行列を使用することが有効

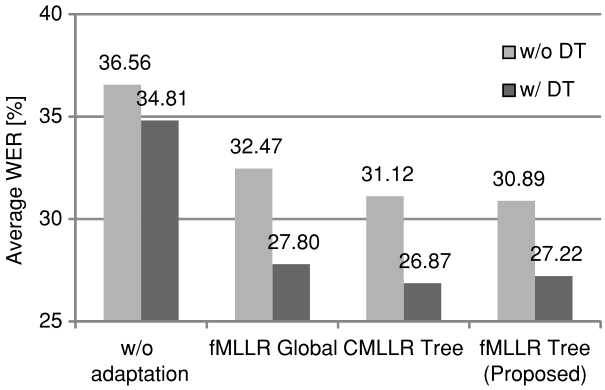


Fig. 2 Average WER[%] for isolated speech (si_et_05) with noise suppression by prior-based binary masking.

であることがわかる．CMLLR Tree, fMLLR Tree を比較すると同程度の性能を示しており，後者はモデル適応の利点である回帰木が利用可能でかつ，適応処理の切り分けが容易な特徴量適応であることから， 提案法は両適応手法の利点を併せ持つといえる．

## 5 おわりに

回帰木に基づき求めた CMLLR 変換行列を特徴量に適用する，特徴量空間での話者適応手法を提案した．実験結果から提案法が，単一の変換行列を用いた適応手法よりも優れ，かつ複数の変換行列を用いたモデル適応手法と同程度の性能であること確認した．今後は，本手法により得られる特徴量を DNN に適用し，評価する予定である．

## 参考文献

[1] C. Leggetter *et al.*, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, vol.9, pp.171–185, 1995.

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

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

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

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

[6] 太刀岡勇気 他, “騒音環境下音声認識に対する識別的アプローチの有効性 第 2 回 CHiME チャレンジ,” 音講論 (秋), pp.1–4, 2013.