LG-002

移動ベクトルのコース/ファイン学習法にもとづく音響モデル適応
Acoustic model adaptation based on coarse/fine training of transfer vectors

渡部 晋治 † 中村 篤 †
Shinji Watanabe Atsushi Nakamura

1. まえがき

音声認識技術はヒューマンインターフェースにおいて大きく期待される技術である．しかし，音声認識は現状において学習データとして事前に収集できない未知環境要因の存在により，実環境では著しく性能が劣化することが知られている．そのため，少量の未知環境データをもとに音声認識用音響モデルを素早くその環境に適応させるモデル適応技術が大変重要となっており，盛んに研究されている (例えば [1–4]).

代表的な音響モデル適応は大きく 2 つに分類される．一つは事前知識を利用してガウス分布統計量をベイズ推定により求めるベイズアプローチ [2]，もう一つはガウス分布統計量の変換写像を推定する変換写像アプローチである [1, 3]．一般に，変換写像アプローチは共有化ガウス分布クラス (コースクラス) を用いるため，推定するパラメータが少なく過学習が緩和され，少量適応データにおいてベイズアプローチを上回る．逆に，適応データが適度に多い場合，一般に，ガウス分布クラス (ファインクラス) で推定を行うベイズアプローチは，その精密な推定ゆえに，変換写像アプローチを上回る．このように，従来手法は適応データ量に応じて一長一短であり，データ量によらず常に十分な性能を示す音響モデル適応技術が求められている．

本稿では，コース/ファイン両クラスにおけるパラメータ推定を内包する音響モデル適応を提案する．まず初めに，初期モデルから適応モデルへのガウス分布平均パラメータの移動ベクトルに着目する．このとき，移動ベクトルを方向ベクトルとスケーリングファクターに分解し，それぞれをコースクラス，ファインクラスの推定により個別に求める．これらは最尤法・事後確率最大化法・変分ベイズ法 (Variational Bayes: VB [5]) を用いて解析的且つ効率良く求めることができる．VB においてはモデル構造に対する VB 事後確率値を導出することにより，ガウス分布共有クラスを学習データ量に応じて自動的に決定できる．本稿では移動ベクトルのコース/ファイン学習における VB 解を示す．

2. 移動ベクトルのコース/ファイン学習法

未知環境データ (適応データ) を用いて初期音響モデルから適応音響モデルを再構築する場合，推定されるガウス分布 k の適応化平均ベクトル $\boldsymbol{\mu}_k^{new}$ は初期モデルの平均ベクトル $\boldsymbol{\mu}_k^{ini}$ 及び 適応データの平均ベクトル $\boldsymbol{m}_k$ 間の内挿ベクトルで表現される（図 1 参照）．ここで，図 1 における移動ベクトル $\boldsymbol{\mu}_k^{new} - \boldsymbol{\mu}_k^{ini}$ に着目すると，移動ベクトルは次のように方向ベクトル $\boldsymbol{\delta}_k$ およびスケーリングファクター g_k に分解することができる．

$$\boldsymbol{\mu}_k^{new} = \boldsymbol{\mu}_k^{ini} + g_k \boldsymbol{\delta}_k.$$

†日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

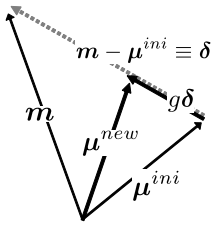

図 1: 方向ベクトル $\boldsymbol{\delta}$ 及び スケーリングファクター g.

本研究の要点は**方向ベクトル $\boldsymbol{\delta}$ 及び スケーリングファクター g をガウス分布クラス k を含む異なる上位クラス i_k 及び j_k で推定する**所にある．つまり，

$$\boldsymbol{\mu}_k^{new} = \boldsymbol{\mu}_k^{ini} + \underline{g_k \boldsymbol{\delta}_k} \longrightarrow \boldsymbol{\mu}_k^{new} = \boldsymbol{\mu}_k^{ini} + \underline{g_{j_k} \boldsymbol{\delta}_{i_k}}.$$

このとき，スケーリングファクター g のパラメータ数は 1 であり，特徴量次元分ある方向ベクトル $\boldsymbol{\delta}$ のパラメータ数に比べて非常に小さい．そこで，移動ベクトルの推定に際し，(i) 方向ベクトルの推定においては共有化ガウス分布クラスに割り当てられた十分多い適応データを用い (コースクラス推定)，(ii) スケーリングファクターの推定においてはガウス分布に割り当てられた少量データを用いる (ファインクラス推定) ことにより，コース/ファイン両クラスにおけるパラメータ推定を内包することができる (図 2 参照)．これを移動ベクトルのコース/ファイン学習と呼ぶ (Coarse/Fine Training of transfer vectors: CFT).

3. 変分ベイズ法による解析解

VB は事後分布の実用的近似解を提供する強力な手法であり [5]，音声認識においてもその効果が十分示されている [6]．適応データの D 次元特徴量ベクトル集合を $\boldsymbol{O} = \{\boldsymbol{o}^t \in \mathcal{R}^D : t = 1, ..., T\}$ としたとき，方向ベクトル $\boldsymbol{\delta}_{i_k}$ の VB 事後分布 $\widetilde{q}(\boldsymbol{\delta}_{i_k}|\boldsymbol{O})$ はガウス分布 $\mathcal{N}(\boldsymbol{\delta}_{i_k}|\widetilde{\boldsymbol{\alpha}}_{i_k}, \widetilde{\Omega}_{i_k})$ で表現され，それらのハイパーパラメータ $\widetilde{\boldsymbol{\alpha}}_{i_k}$ および $\widetilde{\Omega}_{i_k}$ は初期モデルの共分散行列 Σ_k を用いて次のように与えられる．

$$\begin{cases} \widetilde{\boldsymbol{\alpha}}_{i_k} \equiv \widetilde{\Omega}_{i_k}\big((\Omega_{i_k}^0)^{-1}\boldsymbol{\alpha}_{i_k}^0 + \sum_{k\in i} \zeta_k \widetilde{u}_{j_k} (\Sigma_k)^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k^{ini})\big) \\ \widetilde{\Omega}_{i_k} \equiv \big((\Omega_{i_k}^0)^{-1} + \sum_{k\in i} \zeta_k (\widetilde{v}_{j_k} + (\widetilde{u}_{j_k})^2)(\Sigma_k)^{-1}\big)^{-1}. \end{cases}$$

ここで $\boldsymbol{\alpha}_{i_k}^0$ 及び $\Omega_{i_k}^0$ は事前分布 $\mathcal{N}(\boldsymbol{\delta}_{i_k}|\boldsymbol{\alpha}_{i_k}^0, \Omega_{i_k}^0)$ のハイパーパラメータであり，ζ_k 及び $\widehat{\boldsymbol{\mu}}_k \equiv \sum_t \zeta_k^t \boldsymbol{o}^t / \zeta_k$ は

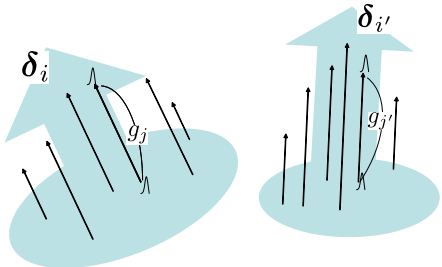

図 2: $\boldsymbol{\delta}_i$ and g_j により推定される平均の移動ベクトル.

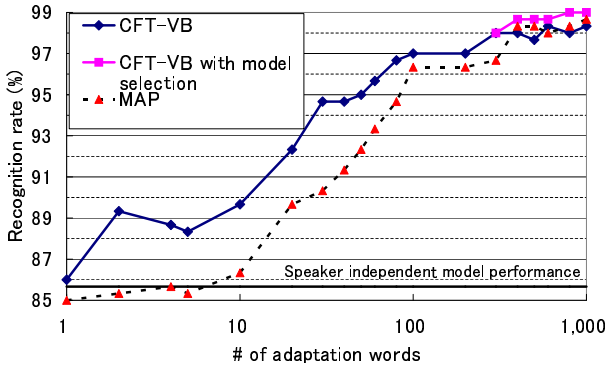

図 3: 適応データ量に応じた認識率.

ガウス分布 k におけるフレーム数及び適応データの平均ベクトルを表す.

同様にスケーリングファクター g_{j_k} の VB 事後分布 $\widetilde{q}(g_{j_k}|\boldsymbol{O})$ はガウス分布 $\mathcal{N}(g_{j_k}|\widetilde{u}_{j_k},\widetilde{v}_{j_k})$ で表現され，それらのハイパーパラメータ $\widetilde{u}_{j_k}$ 及び $\widetilde{v}_{j_k}$ は次のように与えられる.

$$\begin{cases} \widetilde{u}_{j_k} \equiv \left((v^0_{j_k})^{-1}u^0_{j_k} + \sum_{k\in j}\zeta_k\widetilde{\boldsymbol{\alpha}}'_{i_k}(\Sigma_k)^{-1}(\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}^{ini}_k)\right)\widetilde{v}_{j_k} \\ \widetilde{v}_{j_k} \equiv \left((v^0_{j_k})^{-1} + \sum_{k\in j}\zeta_k \mathrm{tr}((\widetilde{\boldsymbol{\alpha}}_{i_k}\widetilde{\boldsymbol{\alpha}}'_{i_k} + \widetilde{\Omega}_{i_k})(\Sigma_k)^{-1})\right)^{-1}. \end{cases}$$

ここで $u^0_{i_k}$ および $v^0_{i_k}$ は事前分布 $\mathcal{N}(\boldsymbol{\delta}_{i_k}|u^0_{i_k},v^0_{i_k})$ のハイパーパラメータであり，$'$ および tr は行列の転置及びトレースを表す．これらの計算は，VB 版の期待値最大化法を元に効率よく求めることができる．また，モデル構造に対する VB 事後確率値を計算することにより，学習データ量に応じたクラスの自動制御が可能となる．これを CFT-VB と呼ぶ.

4. 話者適応実験

教師あり話者適応実験を通じて，提案法である CFT-VB 適応の，適応データ量の増加に応じた認識性能を調べた．このとき，比較対照として，ファインクラスを用いたベイズアプローチの代表的手法である MAP 適応 [2] を用いた．CFT-VB において方向ベクトル $\boldsymbol{\delta}_{i_k}$ のクラスは，トライフォン HMM の共有状態中のガウス分布が共有化されたクラスを用い，スケーリングファクター g_{j_k} のクラスは個々のガウス分布クラスを用いた（つまり $g_{j_k} \rightarrow g_k$）．実験条件は表 1 及び 2 のとおりであり，CFT-VB, MAP 両実験を通じて共分散行列は一定とした．初期モデルは電子協都市名データのべ 4800 単語 (男性話者 50 名) を用い，適応・評価データには ATR 孤立単語データを用い男性話者一人の 1000 単語を適応データに 300 単語を評価データに用いた．適応用データからランダムに単語を抜き出し，データ量の異なる 19 種類の適応データセットを用いて適応実験を行った.

図 3 は適応データ学習量に応じた CFT-VB と MAP の，不特定話者音響モデル (85.7 %) からの性能改善比較

表 1: Acoustic conditions

Sampling rate	16 kHz (quantization 16 bit)
Feature vector	12 - order MFCC with Δ MFCC
Window	Hamming
Frame size/shift	25/10 ms

表 2: Acoustic model structure

# of states	3 (left to right)
# of phoneme categories	27
# of clustered states	324
Output distribution	8 components GMM

を示したグラフである．特に 100 単語以下の適応実験において CFT-VB は MAP に比べて最大で 4 % 程性能が上回っている．また，100 単語以上になると，CFT-VB は MAP と同等の認識性能を示している．これらの結果は，コース/ファイン両クラスの学習を行うことにより任意の学習データにおいて認識性能の改善が得られる CFT 適応の効果を示している.

次に，VB のモデル選択機能を利用した適応実験結果を示す．今回の実験では，方向ベクトル $\boldsymbol{\delta}_{i_k}$ のクラスに対して先程の実験で用いた共有化状態クラス (ファインクラス) 及び，ガウス分布クラス (コースクラス) を用意し，VB を用いて学習データに応じてクラスの自動選択を行った．VB のモデル選択により，適応データ 400 単語のとき，方向ベクトル $\boldsymbol{\delta}_{i_k}$ のクラスはコースクラスからファインクラスに移った（図 3 参照）．この結果は，400 単語付近でパラメータあたりの適応データが十分多くなったため，より精密なファインクラスが選ばれたことを示している．またこのとき，CFT-VB は MAP に比べて学習データが十分多い領域において 1 % 程性能が上回っている．理由として，この領域においては，MAP 適応と CFT-VB は双方ともファインクラス推定を行っているため，スケーリングファクターの分だけ，CFT-VB のモデルがより精密になったためと考えられる.

このように，CFT-VB はコース/ファイン学習及び VB のモデル選択により任意の学習データにおいて従来の MAP 適応を上回る結果を示した.

5. むすび

本稿では移動ベクトルのコース/ファイン学習法 (CFT) にもとづく斬新な音響モデル適応手法を提案した．CFT は教師あり話者適応実験において従来の MAP 適応を任意の学習データにおいて上回った．CFT は効率的で且つ強力な適応技術であり，今後は教師なし適応やオンライン適応等でその効果を示していきたい.

参考文献

[1] 大倉他．混合連続分布 HMM 移動ベクトル場平滑化話者適応方式．信学論 D-II, Vol. J76 - D-II, pp. 2469–2476, 1993.

[2] J-L. Gauvain and C-H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. on SAP*, Vol. 2, pp. 291–298, 1994.

[3] C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, Vol. 9, pp. 171–185, 1995.

[4] 篠田浩一．話者適応 (サーベイ)．第 3 回 音声言語シンポジウム, 2001.

[5] H. Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. UAI 15*, 1999.

[6] S. Watanabe et. al. *Application of variational Bayesian approach to speech recognition.* NIPS 2002, MIT Press, 2003.