TR-IT-0214

話者適応・話者正規化を用いた不特定話者音声認識

Speaker-Independent Speech Recognition Using Speaker Adaptation and Speaker Normalization techniques

石井 純
Jun Ishii

1997.3.31

話者適応技術に関しての先行研究の概要，及び筆者が1995年4月1日から1997年3月31日まで，ＡＴＲ音声翻訳通信研究所で行なった，話者適応，及び話者正規化技術を用いた不特定話者音声認識の研究に関する研究発表論文リストを載せた報告書である．研究の内容は「音響モデルの生成過程を利用した話者適応における共有構造の決定法」，「最大事後確率推定方式を用いた重回帰写像モデルによる話者適応方式」及び「重回帰写像モデルに基づく話者正規化と話者適応方式」である．

もくじ

第 1 章　話者適応技術の概要

1　はじめに

話者適応技術は不特定話者音声認識システムの性能向上のために，少量の特定話者音声データを利用して特定話者モデルのパラメータを推定する技術である．話者適応は，学習データ不十分な場合のモデルの学習であるので，

- パラメータの推定精度の向上
- 未学習パラメータの推定

が研究の中心となっている．ここでは連続混合分布型 HMM (CDHMM) を基本とした音声認識システムに関する話者適応技術を紹介する．

2　話者適応の分類

話者適応の学習という観点からは以下のように分類される．

1. 適応学習データの内容の既知 / 未知
    - 教師あり (Supervised Training)
    - 教師なし (Unsupervised Training)
2. 適応学習データの一括 / 逐次処理
    - 蓄積型 (Batch)
    - 逐次型 (Online, Incremental)

音声認識のアプリケーションを想定すれば教師なし / 逐次型学習の話者適応が望まれる．

また，適応モデルパラメータの推定法に着目した場合，以下の 3 つに大別されると考えられる．

1. モデルパラメータの統計的推定法
2. 写像法
3. 話者選択法

以下にモデルパラメータの統計的推定法，写像法，話者選択法についての先行研究を紹介する．

3　モデルパラメータの統計的推定法

適応データ量を考慮したパラメータ推定方法.

- 最大事後確率 (MAP) 推定法
  最大事後確率 (maximum a posteriori:MAP) 推定 [1] は，新たに得られたデータからモデルパラメータを推定するときに，事前に得られている知識 (事前知識) を効果的に利用する方式である．学習データ $\mathbf{o}$ が与えられた時，MAP 推定による推定パラメータ λ_{MAP} は下式で与えられる．

$$\lambda_{MAP} = \arg\max_{\lambda}\{P(\lambda|\mathbf{o})\} = arg\max_{\lambda}\{P(\mathbf{o}|\lambda)P_0(\lambda)\}$$

  ここで $P_0(\lambda)$ は事前分布である．平均ベクトルの事前分布が $N(\rho,\tau^2)$ のとき，MAP 推定による推定値 μ_{MAP} は

$$\mu_{MAP} = \frac{n\tau^2}{\sigma^2+n\tau^2}\overline{\mathbf{o}} + \frac{\sigma^2}{\sigma^2+n\tau^2}\rho$$

  となる．ここで n は学習サンプル数，$\overline{\mathbf{o}}$ は学習データの平均値，σ^2 は各ガウス分布の分散である．統計的手法に関しては多くの研究機関において検討されいる．以下に代表的なものを列挙する．

- 出力分布の分散，状態遷移確率，混合分布の場合の重み係数の MAP 推定 [2][3]
- 逐次型の MAP 推定 [4]
- 拡張 MAP 推定法 [5]
  事前分布の平均ベクトル間の相関を考慮した MAP 推定方式
- 再帰ベイズ則による逐次型話者適応 [6]

4　写像法

適応データでは学習できないパラメータの推定に関する研究である．初期モデルのパラメータと適応話者モデルのパラメータを写像関係として，推定を行なう．

4.1　補間・平滑化法

適応学習データを用い，ML 再推定によって得られるガウス分布の平均ベクトルと，初期モデルの差である移動ベクトルを求め，未学習の平均ベクトルを事後的に補間する．

- スペクトル内挿法 [7]
  ガウス分布の未学習平均ベクトルを近傍の移動ベクトルから補間推定.

- 移動ベクトル場平滑化方式 (VFS) [8]
  ガウス分布の未学習平均ベクトルを近傍の移動ベクトルから補間推定し，更に，平滑化を行なう．

- 補間・平滑化方式と MAP 推定との統合
  - MAP/VFS [9][10]
    VFS の移動ベクトル推定を MAP 推定を用いて行なう．
  - MAP/VFS-MCE [11]
    MAP/VFS へ更に識別学習 (minimum classification error training: MCE) の手法を加えた．

4.2 重回帰写像モデル

適応前のガウス分布の平均ベクトル μ(次元数 n) を下の式で適応後の平均ベクトル $\hat{\mu}$ へ変換する．

$$\hat{\mu} = A\mu + b$$

$$A \quad : \quad n \times n \text{ の行列}$$
$$b \quad : \quad n \text{ 次元の定数ベクトル}$$

演算量が少なく，また実験的に定める係数がない (少ない) ので扱い易い．以下に重回帰写像モデルに基づく先行研究を列挙する．

- Maximum likelihood linear regression(MLLR)[12]
  最尤推定によって回帰係数を推定
- MLLR による教師なし話者適応 [13]
- 行列 A の対角成分，副対角成分が同一の値をとる行列に拘束 [14]
- 行列 A を対角行列とし，EM アルゴリズムによってガウス平均と分散の両方を求める [15]
- 重回帰写像モデルと MAP 推定の統合 [16][17][18]

4.3 パラメータ共有化法

一般に話者適応データが少量であるのに対し，HMM のパラメータは多数である．従って，HMM のパラメータを共有化 (sharing, tying) し，パラメータ数を減らして推定精度を向上される方法．

- 木構造化されたパラメータによる共有化法
  - 状態の木構造共有構造 [19]
  - 基底分布の木構造共有構造
    クラスタリングの基準
    * kullback divergence の距離によるクラスタリング [20]
    * 話者適応を行なった際の移動ベクトルの相関 [21]
- パラメータ共有化法と MAP 推定の統合 [22][23]
- パラメータ共有化法と重回帰写像モデルとの統合 [12]

5　話者選択法

予め複数の話者のモデルを用意しておき，その中から特定話者モデルを選択し，認識を行なう．より短い学習音声から話者適応を行なうことができる．また選択後に更に話者適応を行なう場合においても，選択された音響モデルは良い初期モデルであるとから正確な話者適応が実現できる．

- 木構造話者クラスタによる話者選択 [24].
  適応学習データを用い尤度によって選択する．さらに
  - 木構造話者クラスタによる選択後に MAP-VFS [10]
    により，認識性能の向上が得られている．
- 不特定話者モデルからの移動ベクトルを基準に話者を選択 [25]
- 話者クラスタからの話者選択後に重回帰写像モデルによる適応 [26]

6　その他

6.1　教師なし話者適応

- subword モデルと言語情報を利用した方式 [27]
  全ての音素 HMM の音素 bigram 確率値を遷移確率として結合したエルゴディック HMM と VFS を統合した．
- N-best 認識結果から教師信号を得る方法
  - N-best 認識結果を用いて話者適応を行ない，適応後の尤度が最も高い結果を教師信号とする [28].

– N-best 認識結果において 1 位の認識結果の尤度と 2 位以下の認識結果の尤度差が閾値より大きければ教師信号として採用する [29].

6.2 話者適応の初期モデルの検討

一般に不特定話者モデルを初期モデルとして話者適応を行なうが，不特定話者モデルは複数の話者の性質を含んでいる．従って，ある特定の話者へ適応する場合には，適応に不都合な話者の性質が正確な話者適応の妨げになっていると考えられる．そこで，話者適応にとって有効な初期モデルの作成に関する研究が行なわれている．

- MLLR を用いた話者正規化による初期モデルの作成 [30][18]
- 移動ベクトル用いた話者正規化による初期モデルの作成 [31]

7 まとめ

話者適応技術の先行研究をモデルパラメータの統計的推定法，写像法，話者選択法に分類して紹介した．現状では上記の 3 の方法を統合した方式が最も良い結果を得ている．

今後は，教師なし，逐次型の話者適応の検討，良い初期モデルの獲得，環境適応や話者正規化を含めた研究を行なうべきであろう．

参考文献

[1] C.-H. Lee, C.-H. Lin and B.-H.Juang: "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," IEEE Trans. on Signal Processing, Vol. 39, No. 4, pp. 806-814 (1991)

[2] J.-L. Gauvain and C.-H. Lee: "Speaker adaptation based on map estimation of hmm parameters," Proc. of ICASSP 93, pp. 558-561 (1993)

[3] J.-L. Gauvain and C.-H. Lee: "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp. 291-298 (1994)

[4] 松岡達雄, C.-H. Lee: "最大事後確率推定法 (MAP 推定法) によるオンライン話者適応化," 信学技報, **SP93**-133 (1994.1)

[5] R. M. Stern *IEEE trans. ASSP*, 35, pp. 751, (1987)

[6] Q. Hou and C. -H. Lee: "A Study on On-Line Quasi-Bayes Adaptation for CDHMM-Based Speech Recognition," *Proc. of ICASSP96*, pp. 705-708, (1996)

[7] 篠田, 磯, 渡辺: "音声認識のためのスペクトル内挿を用いた話者適応化," 信学会論文誌, Vol. J77-A, No. 2, pp. 120-127, (1994)

[8] K. Ohkura, M. Sugiyama and S. Sagayama: "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs," *Proc. of ICSLP92*, pp. 369-372, (1992)

[9] J. Takahashi and S. Sagayama: "Telephone Line Characteristic Adaptation Using Vector Field Smoothing Technique," *Proc. of ICSLP94*, pp. 991-993, (1994)

[10] M. Tonomura, T. Kosaka and S. Matsunaga: "Speaker Adaptation Based on Transfer Vector Field Smoothing Using Maximum a Posteriori Probability Estimation," *Proc. of ICASSP95*, pp. 688-691, (1995)

[11] J. Takahashi and S. Sagayama: "Minimum Classification Error Training for a Small Amount of Data Enhanced by Vector-Field-Smoothing Bayesian Learning," *Proc. of ICASSP96*, pp. 597-600, (1996)

[12] C. L. Leggetter and P. C. Woodland: "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, Vol. 9, pp. 171-185 (1995)

[13] P. C. Woodland, D. Pye, M. J. F. Gales: "Iterative Unsupervised Adaptation Using Maximum Likelihood Linear Regression," *Proc. of ICSLP96*, pp. 1133-1136, (1996)

[14] S. J. Cox and J. S. Bridle: "Unsupervised speaker adaptation by probabilistic spectrum fitting," *Proc. of ICASSP 89*, pp. 294-297 (1989)

[15] V. V. Digalakis, D. Rtischev, L. G. Neumeyer: "Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures," *IEEE Trans. on speech and audio processing*, Vol. 3, No. 5, (1995)

[16] V. Digalakis and L. Neumeyer: "Speaker Adaptation Using Combined Transformation and Bayesian Method," *Proc. of ICASSP 95*, pp. 680-683 (1995)

[17] G. Zavaliagkos, R. Schwartz and J. McDonough: "Maximum a Posteriori Adaptation for Large Scale HMM Recognizers," *Proc. of ICASSP 96*, pp. 725-728 (1996)

[18] J. Ishii, M. Tonomura: "Speaker Normalization and Adaptation Based on Linear Transformation," Proc. of ICASSP97, (1997) (to appear)

[19] S. J. Young and P. C. Woodland: "State Clustering in Hidden Markov Model-Based Continuous Speech Recognition," *Computer Speech and Language*, vol. 8, pp. 369-383, (1994)

[20] K. Shinoda and T. Watanabe: "Speaker Adaptation with Autonomous Control Using Tree Structure," *Proc. of EUROSPEECH95*, pp. 1143-1146, (1995)

[21] 高橋, 嵯峨山: "学習移動ベクトルの相関により tying した音響モデルの共有構造," 音講論集, **2-2-13**, (1995.9)

[22] 柴田, 松本: "差分ベクトルの木構造の結びに成分分布の信頼度を考慮した話者適応," *SP96-90*, pp. 21-28, (1997)

[23] J. Ishii, M. Tonomura, and S. Matsunaga: "Speaker Adaptation Using Tree Structured Shared-State HMMs," *Proc. of ICSLP 96*, pp. 1149-1152, (1996)

[24] 小坂, 松永, 嵯峨山: "話者適応のための木構造話者クラスタリング," 信学技法, SP93-110, (1993.12)

[25] 大倉, 大西, 飯田: "複数代表話者の話者空間移動ベクトルに基づく不特定話者 HMM の話者適応化," 信学会論文誌, Vol. J79-D-II, No. 5, pp. 667-674, (1996.5)

[26] M. Padmanabhan, L. R. Bahl, D.Nahamoo, M. A. Picheny: "Speaker Clustering and Transformation for Speaker Adaptation in Large-Vocabulary Speech Recognition," *Proc. of ICASSP96*, pp.701-704, (1996)

[27] 宮沢，大倉，嵯峨山: "全音素エルゴディック HMM を用いた教師なし話者適応," 信学会論文誌, A Vol. J77-A, No. 2, pp. 112-119, (1994.2)

[28] T. Matsui, S. Furui: "N-best-based Instantaneous Speaker Adaptation Method for Speech Recognition," *Proc. of ICSLP96*, pp. 973-976, (1996)

[29] S. Homma, J. Takahashi, S. Sagayama: "Iterative Unsupervised Speaker Adaptation for Batch Dictation," *Proc. of ICSLP96*, pp. 1141-1144, (1996)

[30] T. Anastasakos, J. McDonough, R. Schwartz, J. Makhoul: "Compact Model for Speaker-Adaptive Training," *Proc. of ICSLP 96*, pp. 1137-1140 (1996)

[31] N. Iwahashi: "Novel Training Method for Classifiers used in Speaker Adaptation," *Proc. of ICSLP 96*, pp. 2119-2122 (1996)

第 2 章　研究内容

1　研究の概要

筆者が話者適応、話者正規化技術に関して発表した文献を載せる。筆者は不特定話者音声認識の性能向上のために、話者適応方式、及び話者正規化方式を検討した。

- 音響モデルの生成過程を利用した話者適応における共有構造の決定法

  一般に話者適応においては適応学習データの数が少ないために共有化によってパラメータ数を減らし適応処理を行なうことが有効である。我々は音響モデルが持つ音素環境の類似性による共有化を検討し、音響モデルを生成する逐次状態分割法の分割過程を考慮した話者適応方式を提案した。本稿ではこの音素環境を考慮した共有化構造を、適応データ量に応じた適応手法である最大事後確率推定法 (MAP 推定法) を導入した MAP/VFS へ適用することと、適応データ数によって VFS の補間、平滑化処理に用いる近傍ベクトル数の制御に利用することを検討し、音素認識実験で評価した。(参考文献: 付録 A1, ICSLP96)

- 最大事後確率推定方式を用いた重回帰写像モデルによる話者適応方式

  連続分布型 HMM を用いた話者適応の一つとして重回帰写像モデルに基づく方式があり、その有効性が報告されている。ここでは、最尤を基準に変換行列を推定する Maximum Likelihood Linear Regression (MLLR) について検討した。この方式において変換行列が Full Matrix である場合、少量の適応データに対しては認識率が低下する問題がある。この問題に対処するために、MLLR に最大事後確率 (MAP) 推定法を導入することを検討。(参考文献: 付録 A 2, 平成 8 年秋季音響学会)

- 重回帰写像モデルに基づく話者正規化と話者適応方式

  重回帰写像モデルを用いた話者正規化モデルの生成と話者適応方式を提案する。話者正規化の目的は、識別性能が高く、また話者適応の初期モデルとしても有効な音響モデルの獲得である。話者正規化モデルの作成は、重回帰写像モデルを用いて変換係数を求めた後、話者性を表していると考えられる定数項ベクトルの値を学習ベクトルから引き去ることで行なう。また、この話者正規化モデルを初期モデルとした話者適応として、重回帰写像モデルに最大事後確率 (MAP) 推定法を組み合わせた方式を提案する。話者正規化を行なった音響モデルは識別性能が高いことを示す共に、話者適応の初期モデルとしても有効であることを示した。(参考文献: 付録 A3, ICASSP97 論文)

2　発表文献リスト

1. 石井純, 外村政啓, 松永昭一:
"逐次状態分割法の分割過程を考慮した移動ベクトル場平滑化の検討,"
日本音響学会平成 7 年秋季研究発表会講演論文集, **3-2-14**, (1995.9)

2. 石井純, 外村政啓, 松永昭一:
"逐次状態分割法の分割過程を考慮した話者適応方式の検討,"
日本音響学会平成 8 年春季研究発表会講演論文集, **1-5-23**, (1996.3)

3. J. Ishii, M. Tonomura, S. Matsunaga:
"Speaker Adaptation Using Tree Structured Shared-State HMMs,"
Proc. of ICSLP96, pp. 1149-1152, (1996.10)

4. 石井純, 外村政啓:
"重回帰モデルに基づく話者適応方式の検討,"
日本音響学会平成 8 年秋季研究発表会講演論文集, **3-3-17**, (1996.9)

5. 石井純, 外村政啓:
"重回帰写像モデルを用いた話者正規化と話者適応化方式,"
電子情報通信学会技術研究報告 [音声], **SP96-91**, (1997.1)

6. 石井純, 外村政啓:
"重回帰写像モデルを用いた話者適応のための話者正規化方式,"
日本音響学会平成 9 年春季研究発表会講演論文集, (1997.3)

7. J. Ishii, M. Tonomura:
"Speaker normalization and adaptation based on linear transformation,"
ICASSP97, (1997.4)

注：文献略語

- ICSLP : International Conference on Spoken Language Processing
- ICASSP: International Conference on Acoustics, Speech, and Signal Processing

3　付録

これらの方法実験結果を発表した, ICSLP96 論文, 平成 8 年秋季音響学会, ICASSP97 論文を付録 A1, A2, A3に示す.