論文 / 著書情報
Article / Book Information

論題(和文)	話者適応(サーベイ)
Title(English)	
著者(和文)	篠田浩一
Authors(English)	Koichi Shinoda
出典(和文)	第3回音声言語シンポジウム講演論文集, Vol. , No. , pp.
Citation(English)	, Vol. , No. , pp.
発行日 / Pub. date	2001, 12
権利情報 / Copyright	

話者適応
(サーベイ)

篠田 浩一 (東京大学)

2001年12月20日

はじめに

- 近日中にホームページに掲載
  http://hil.t.u-tokyo.ac.jp/~k-shino/index-j.shtml
- 今までのサーベイ
  - [Woodland 99]
  - [Lee & Huo 2000]
  - [Sagayama, Shinoda, Nakai, and Shimodaira 2001]
- 英文文献のみ
- 独断と偏見

話者適応とは？

音声認識において、使用時の少量の発声を用いて、認識性能を改善させる技術

- 少量データで性能が改善
- データ量が増加するに従い、特定話者の性能に近付く

「事前知識を利用した Rapid Training」

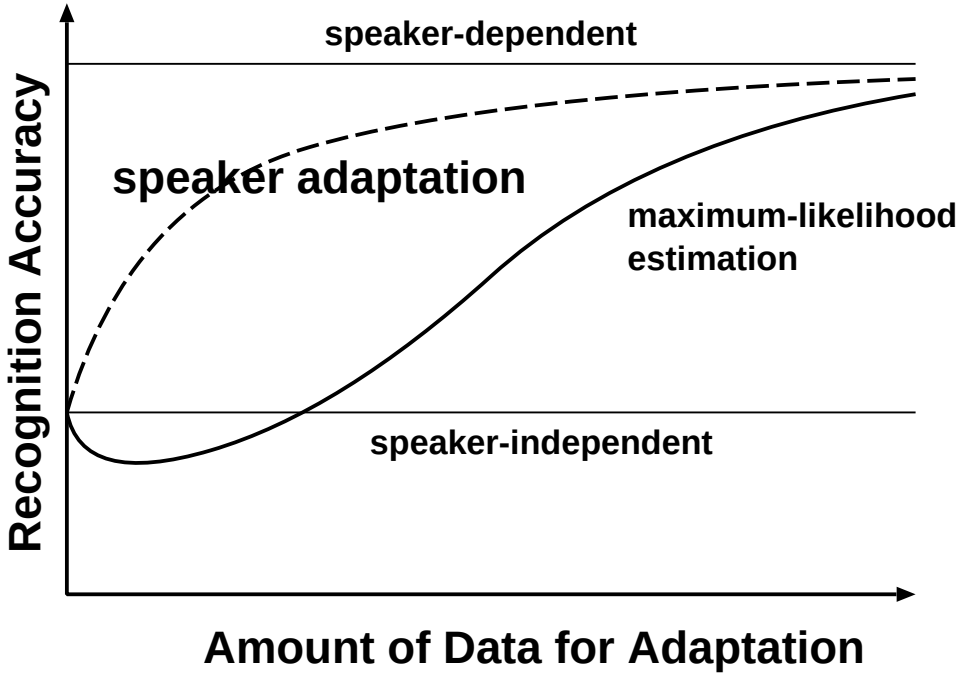

図 1. 話者適応に対する要件

話者適応の基本

- 対象：混合ガウス分布連続HMMにおけるガウス分布の平均ベクトル (と共分散)
- 方法：事前知識の活用
  - パラメータの事前分布 ··· MAP
  - 音響特徴量空間における写像関数 ··· シフト、MLLR
  - 構成因子の変動の観測データ ··· Jacobian
  - 話者毎のパラメータ分布 ··· EigenVoice
  - 特徴量空間の構造 ··· Structural Approach(SMAP など)
- 教師あり手法と教師なし手法
  教師なし手法：多くの場合、認識＋教師あり
- バッチ手法とオンライン手法
  オンライン手法：メモリレス、不安定
- 話者適応を前提とした学習 (Speaker Adaptive Training;SAT)

事後確率最大化(Maximum A Posteriori; MAP)

- MAP とは？
  - パラメータの事前分布を仮定し事後確率を最大にするパラメータを推定
  - 事前分布：自己共役分布 (事後分布が事前分布と同じ形になる) が便利

例：正規分布の平均ベクトル μ が未知、分散 σ^2 は既知
事前分布が正規分布 $\mathcal{N}(x|\mu_0, \sigma_0)$、観測データ $x_1, \ldots, x_N$ のとき

$$\mu = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{1}{\sigma^2}\Sigma_i^N x_i}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}}$$

- 特徴
  - 最尤推定量への収束が保証
  - データに出現しないパラメータは更新されない→改善速度が遅い

MAP を用いた手法

- 混合正規分布 [Lee et al. 91, Gauvain & Lee 94]
  事前分布：normal-Wishart 分布
- EMAP(Extended MAP) [Stern & Lasry 87, Zavaliagkos et al. 95]
  事前分布：2 つのパラメータの同時確率分布
- オンライン [Huo & Lee 97, Huo & Lee 98]

写像関数を用いる手法(1) – シフト

- SBR(Signam Bias Removal) [Rahim & Juang 96]
- スペクトル内挿 [Shinoda et al. 91]
- VFS(Vector Field Smoothing) [Ohkura et al. 92]
- シフトの分布を考慮 [Sankar & Lee 96]

特徴

- MLLR に内包される
- より少ないデータ量で十分

写像関数を用いる手法(2) – MLLR

Maximum Likelihood Linear Regression

- アフィン変換 [Leggetter & Woodland 95]

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

- Constrainet MLLR [Digalalkis & Neumeyer 96]
  特徴量「空間」の変形

特徴

- 比較的少量のデータで安定して動作
- 特徴量空間を区分しなくてもある程度の性能
- データ量が少ないとき：推定パラメータを対角成分やその近傍に限定

※注意　　真の写像を「線形近似」

Jacobian

[Sagayama et al. 97, Yamaguchi et al. 97]

$$\Delta \mathbf{y} = A\Delta \mathbf{a} + B\Delta \mathbf{b} + C\Delta \mathbf{c} + \ldots$$

- **分析的手法：全体の変化を因子毎の変化に分解**
- **ある因子の変化を観測 → その因子に起因する全体の変化を推測**

特徴

- **雑音、声道長など個々の因子の変化が観測可能な条件下で有効**

EigenVoice

[Kuhn et al. 98, Kuhn et al. 2000, Kuhn et al. 2001]

- 話者と音韻の直積空間における主成分分析
  話者特徴量空間を低次元部分空間に射影

特徴

- 多くの話者の多量の発声データが必要
- 極めて少ない発声で安定動作
- 発声数が多くなると性能向上度合が減少

※ 話者クラスタリング
[Kosaka et al. 94, Padmanabhan et al. 98, Yoshizawa et al. 2001]

直積空間における領域を選択

Structural Approach

- 推定パラメータ数とデータ量との関係に焦点
- 特徴量空間をクラスタリング
- データ量に応じてクラスタ数を変化させる

クラスタの階層構造を用いた自律的制御

- [Furui 89]
- [Shinoda & Watanabe 95]
- [Shinoda & Watanabe 96]
- [Kannan & Khudanpur 99]

構造的手法の例

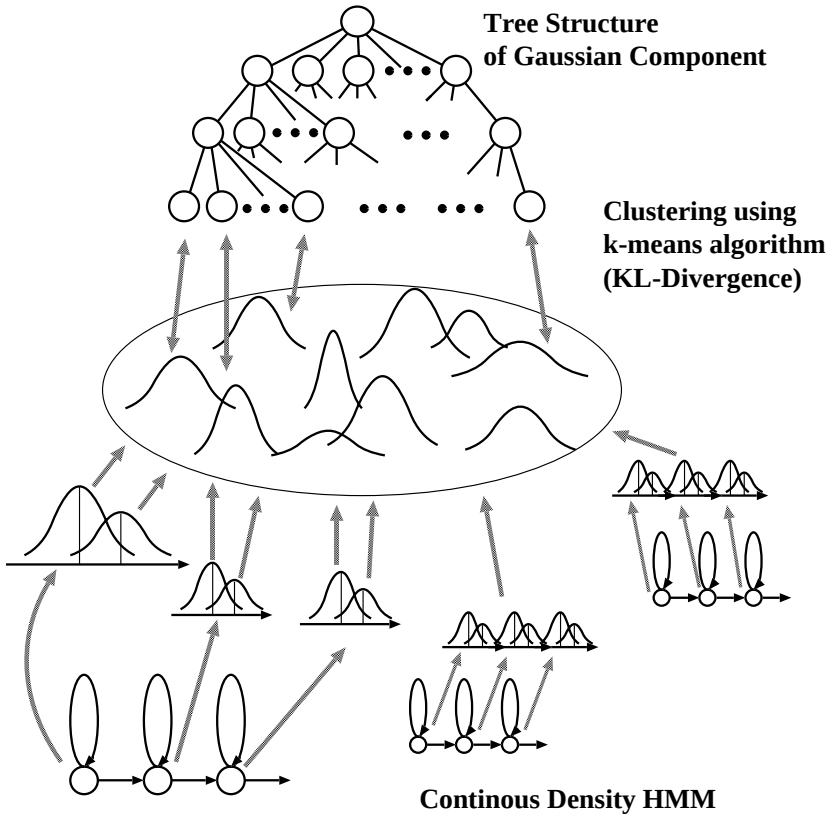

図 2. 特徴量空間における構造の例：ガウス分布木構造

MAP と写像関数手法と構造的手法

排他的ではなく、それぞれを組み合わせて使用可能

- **MAP とシフト**
  MAP-VFS [Takahashi & Sagayama 95, Tonomura et al. 95]
- **MAP とアフィン変換**
  MAPLR(MAP+MLLR)[Siohan et al. 99, Chesta et al. 99]
- **MAP と構造的手法**
  SMAP [Shinoda & Lee 97, Shinoda & Lee 2001]
- **MAP とアフィン変換と構造的手法**
  SMAPLR [Siohan et al. 2000, Myrvall et al. 2000]
- **MLLR と EigenVoice**
  [Chen & Wang 2001, Wang et al. 2001]

話者適応マップ

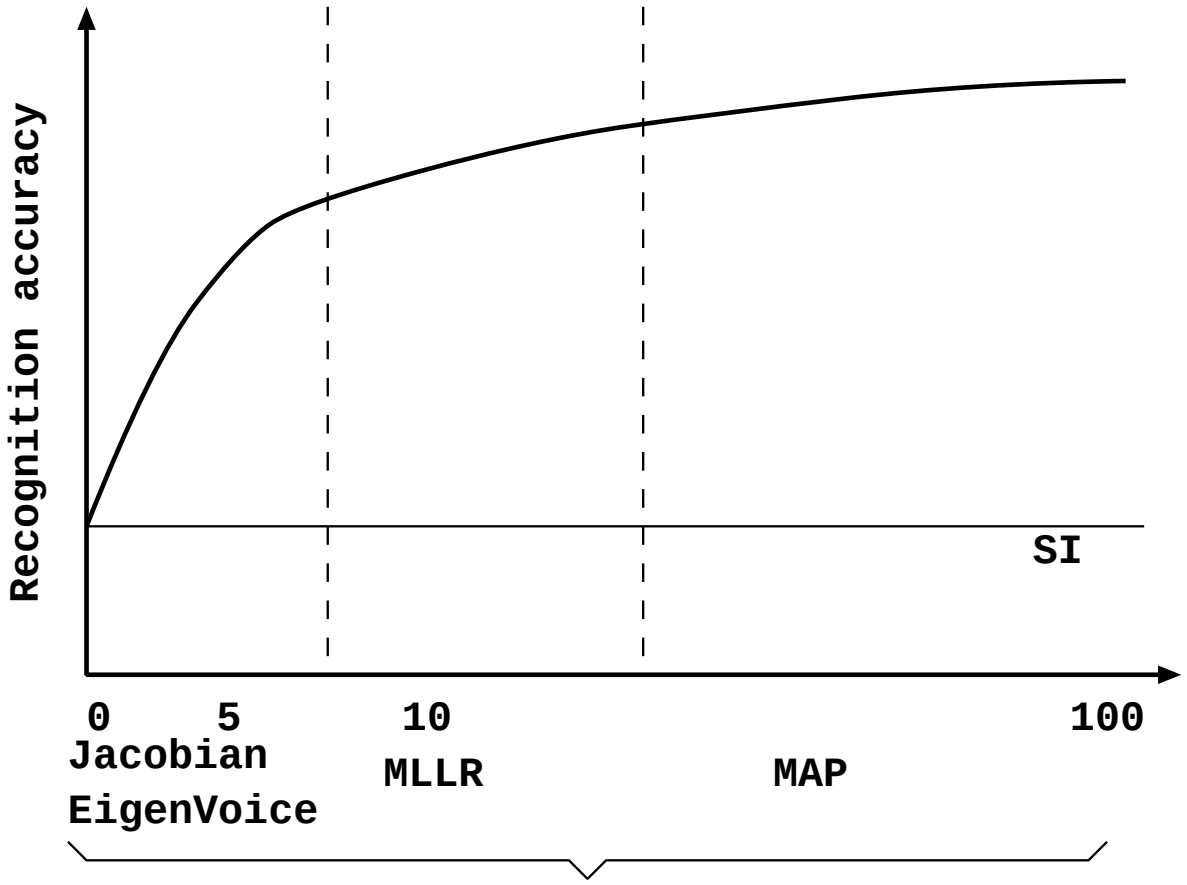

図 3. 話者適応手法の適用範囲

話者適応を前提とした学習

Speaker Adaptive Training (SAT)
[Anastasakos et al. 96, Pye & Woodland 97, Jin et al. 98, Welling et al. 98, Gales 98]

1. ある標準的な話者 (正準話者) を想定
2. 各々の学習話者について以下を実行
   (a) 特徴量空間において、話者から正準話者への写像を推定
   (b) 推定した写像を用いて話者のデータを変換
3. 全学習話者の写像後のデータで学習

特徴

- 少ないデータで推定可能な写像の選択が重要
- 以下の手法は SAT に内包

  CMN(Cepstrum Mean Normalization) **[Atal 74]**

  VTLN(Vocal Tract Length Normalization)
  [Lee & Rose 96, Zhan & Westohal 97, Emori & Shinoda 2001, Pitz et al. 2001]

今後の方向性

- 話者データの増加
  - 話者データの使用方法 [Yoshizawa et al. 2001]
  - EigenVoice 的手法の進展 (FisherVoice など)
  - 話者性と音韻性の分離 [Nishida & Ariki 2001]
  - 話者と音響を統一的に扱う枠組 [Kenny et al. 2001]
- 状態クラスタリングとの組み合わせ [Huo & Ma 99]
- 特徴分析段階における話者性 [Saon et al. 2001]
- 分析的な手法 (Jacobian Approach)
- 非線形変換を用いる手法 [Surendran et al. 99]
- データ量に依存しないシームレスな手法

参考文献

References

[Anastasakos et al. 96] T. Anastasakos, J. McDonough, R. Schwartz, and John Makhoul, "A compact model for speaker-adaptive training," in Proc. ICSLP96, vol. 2, FrP2L1.3, 1996.

[Atal 74] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Amer., vol. 55, pp. 1304-1312, 1974.

[Chen & Wang 2001] K. Chen and H. Wang, "Eigenspace-based maximum a posteriori linear regression for rapid speaker adaptation," *in Proc. ICASSP-2001*, P.3.2, 2001.

[Chesta et al. 99] C. Chesta, O. Siohan, and C.-H. Lee, "Maximum a posteriori linear regression for hidden Markov model adaptation," *In Proc. EuroSpeech99*, pp. 211-214, 1999.

[Digalalkis & Neumeyer 96] V.V. Digalakis and L.G. Neumeyer, "Speaker Adaptation Using Combined Transformation and Bayesian Methods," IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 4, pp. 294-300, 1996.

[Emori & Shinoda 2001] T. Emori and K. Shinoda, "Rapid Vocal Tract Length Normalization using Maximum Likelihood Estimation," *in Proc. EuroSpeech2001*, pp. 1649-1652, 2001.

[Furui 89] S. Furui, "Unsupervised Speaker Adaptation Method Based on Hierarchical Spectral Clustering," *Proc. ICASSP-89*, pp. 286-289, Glasgow, 1989.

[Gales 98] M. J. F. Gales, "Cluster Adaptive Training for Speech Recognition," *Proc. ICSLP-98*, pp. 1783-1786, Sydney, 1998.

[Gauvain & Lee 94] 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*, pp. 291-298, Vol. 2, No. 2, April 1994.

[Huo & Lee 97] Q. Huo and C.-H. Lee, "On-line Adaptive Learning of the Continuous Density Hidden Markov Model Based on Approximate Recursive Bayes Estimate," *IEEE Trans. on Audio and Speech Processing*, Vol. 5, No. 2, pp. 161-172, March 1997.

[Huo & Lee 98] Q. Huo and C.-H. Lee, "On-Line Adaptive Learning of the Correlated Continuous-Density Hidden Markov Model for Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 4, pp. 386-397, 1998.

[Huo & Ma 99] Q. Huo and B. Ma, "Irrelevant variability normalization in learning HMM state tying from data based on phonetic decition tree," *in Proc. ICASSP-99*, pp. 577-580, 1999.

[Jin et al. 98] H. Jin, S. Matsoukas, R. Schwartz, and F. Kubaka, "Fast robust inverse transform speaker adapted training using diagonal transformations," in Proc. ICASSP98, vol. 2, pp. 785-788, 1997.

[Kannan & Khudanpur 99] A. Kannan and S. P. Khudanpur, "Tree-Structured Models of Parameter Dependence for Rapid Adaptation in Large Vocabulary Conversational Speech Recognition, " *Proc. ICASSP-99*, pp. 769-772, Phoenix, May 1999.

[Kenny et al. 2001] P. Kenny, G. Boulianne, and P. Dumouchel, "Inter-speaker correlations, intra-speaker correlations and Bayesian adaptation,", in Proc. Isca ITR-Workshop2001, Sophia-Antipolis, 2001.

[Kosaka et al. 94] T. Kosaka, S. Matsunaga and S. Sagayama, "Tree-structured speaker clustering for speaker-independent continuous speech recognition," *in Proc. ICSLP-94*, pp.1375-1378, Yokohama, 1994.

[Kuhn et al. 98] R. Kuhn, P. Nguyen, J.-C. Janqua, L. Goldwasser, N. Niedzielski, S. Finke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," *in Proc. ICSLP-98*, pp. 1771-1774, 1998.

[Kuhn et al. 2000] R. Kuhn, J.-C. Janqua, P. Nguyen, and N. Niedzielski, "Rapid Speaker Adaptation in Eigenvoice Space Robust Speech Recognition," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 6, pp. 695-707, 2000.

[Kuhn et al. 2001] R. Kuhn, E. Perronnin, P. Nguyen, J.-C. Janqua and L. Rigazio, "Very fast adaptation with a compact context-dependent Eigenvoice model," *in Proc. ICASSP-2001*, L.5.6, 2001.

[Lee et al. 91] C.-H. Lee, C.-H. Lin and B.-H. Juang, "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," in *IEEE Trans. Acoustic, Speech and Signal Proc.*, Vol. ASSP-39, No. 4, pp. 806-814, April 1991.

[Lee & Rose 96] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in Proc. ICASSP96, vol. 1, pp. 353-356, 1996.

[Lee & Huo 2000] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *in Proc. IEEE*, vol. 88, no. 8, 2000.

[Leggetter & Woodland 95] C.J. 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.

[Myrvall et al. 2000] T.-A. Myrvoll, O. Siohan, C.-H. Lee, and W. Chou, "Structural maximum a posteriori linear regression for unsupervised speaker adaptation," *In Proc. ICSLP2000*, 2000.

[Nishida & Ariki 2001] N. Nishida and Y. Ariki, "Speaker recogntion by seperating phonetic space and speaker space," *in Proc. Eurospeech2001*, 2001.

[Ohkura et al. 92] K. Ohkura, M. Sugiyama and S. Sagayama, "Speaker Adaptation Based on Transfer-Vector-Field Smoothing with Continuous Mixture Density HMMs", *Proc. ICSLP-92*, pp.369-372, Alberta, 1992.

[Padmanabhan et al. 98] M. Padmanabhan, L. R. Bahl, D. Nahamoo and M. A. Picheny, "Speaker clustering and transformation for speaker adaptation in speech recognition systems,", *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 71-77, 1998.

[Pitz et al. 2001] M. Pitz, S. Molau, R. Schluter, H. Ney, "Vocal tract normalization equals linear transformation in cepstrum space," *in Proc. Eurospeech-2001*, pp. 2653-2656, 2001.

[Pye & Woodland 97] D. Pye and P .C. Woodland, "Experiments in speaker normalization and adaptation for large vocabulary speech recognition," in Proc. ICASSP97, vol. 2, pp. 1047-1050, 1997.

[Rahim & Juang 96] M. Rahim and B.-H. Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition," IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 1, pp.19-30, 1996.

[Sagayama et al. 97] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, "Jacobian Approach to Fast Acoustic Model Adaptation," *in Proc. ICASSP-97*, pp. 835-838, 1997

[Sagayama, Shinoda, Nakai, and Shimodaira 2001] S. Sagayama, K. Shinoda, M. Nakai and H. Shimodaira "Analytic Methods for Acoustic Model Adapation: A Review," in Proc. Isca ITR-Workshop2001, pp. 67-76, Sophia-Antipolis, 2001.

[Sankar & Lee 96] A. Sankar and C-.H. Lee, "A Maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition," IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 3, pp. 190-202, 1996.

[Saon et al. 2001] G. Saon, G. Zweig, and M. Padmanablan, "Lenear feature space projections for speaker adaptation," *in Proc. ICASSP2001*, 2001.

[Siohan et al. 99] O. Siohan, C. Chesta and C.-H. Lee, "Hidden Markov Model Adaptation Using Maximum A Posteriori Linear Regression," Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions, pp.147-150, Tampere, Finland, May 1999.

[Siohan et al. 2000] O. Siohan, T.-A. Myrvoll, and C.-H. Lee, " Structural maximum a Posteriori linear regression for fast HMM adaptation," *In Workshop on ISCA ITRW ASR2000*, 2000.

[Shinoda et al. 91] K. Shinoda, K. Iso, and T. Watanabe, "Speaker Adaptation for Demi-Syllable-Based Continuous-Density HMM," *Proc. ICASSP-91,* pp. 857-860, Toronto, 1991.

[Shinoda & Watanabe 95] K. Shinoda and T. Watanabe, "Speaker Adaptation with Autonomous Control Using Tree Structure," *Proc. EuroSpeech-95*, pp. 1143-1146, 1995.

[Shinoda & Watanabe 96] K. Shinoda and T. Watanabe, "Speaker Adaptation with Autonomous Model Complexity Control by MDL Principle," *Proc. ICASSP-96*, pp.717-720, 1996.

[Shinoda & Lee 97] K. Shinoda and C.-H. Lee, "Structural MAP Speaker Adaptation Using Hierarchical Priors," *Proc. of IEEE Workshop on Speech Recognition and Understanding*, 1997.

[Shinoda & Lee 2001] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 3, pp. 276-287, 2001.

[Stern & Lasry 87] R.M. Stern and M.J. Lasry, "Dynamic Speaker Adaptation for Feature-Based Isolated Word Recognition," *IEEE Trans. on Audio and Speech Processing*, Vol. 35, No. 6, pp. 751-763, 1987.

[Surendran et al. 99] A.C. Surendran, C.H. Lee, and M. Rahim, "Non-Linear Compensation for Stochastic Matching," *IEEE Trans. on Audio and Speech Processing*, pp. 643-655, Nov. 1999.

[Takahashi & Sagayama 95] J. Takahashi and S. Sagayama, "Vector-field-smoothed Bayesian Learning for Incremental Speaker Adaptation," *Proc. ICASSP-95*, pp. 688-691, Detroit, May. 1995.

[Tonomura et al. 95] M. Tonomura, T. Kosaka ,and S. Matsunaga, "Speaker Adaptation Based on Transfer-Vector-Field-Smoothing Using Maximum A Posteriori Probability Estimation," *Proc. ICASSP-95*, pp. 688-691, Detroit, May. 1995.

[Wang et al. 2001] N. J.-C. Wang, S. S.-M. Lee, F. Seide, and L.-S. Lee, "Rapid speaker adaptation using a priori knowledge by Eigenspace analysis of MLLR parameters," *in Proc. ICASSP-2001*, P.3.2, 2001.

[Welling et al. 98] L. Welling, R. Haeb-Umbach, X. Aubert, and N. Harberland, "A study on speaker normalization using vocal tract normalization and speaker adaptive training," in Proc. ICASSP98, vol. 2 , pp. 797-800, 1998.

[Woodland 99] P. C. Woodland, "Speaker adaptation: techniques and challenges," *in Proc. 1999 IEEE Workshop Automatic Speech Recognition and Understanding*, Keystone, 1999.

[Yamaguchi et al. 97] Y. Yamaguchi, S. Takahashi, and S. Sagayama, "Fast Adaptation of Acoustic Models to Environmental Noise Using Jacobian Adaptation Algorithm," *in Proc. Eurospeech-97*, pp. 2051-2054, 1997

[Yoshizawa et al. 2001] S. Yoshizawa, A. Baba, K. Matsunami, Y. Mera, M. Yamada, and K. Shikano, "Uusupervised speaker adaptation based on the sufficient HMM statistics of selected speakers,", *Proc. ICASSP2001*, 2001.

[Zavaliagkos et al. 95] G. Zavaliagkos, R. Schwartz, and J. McDonough, "Maximum A Posteriori Adaptation for Large-Scale HMM Recognizers," *Proc. ICASSP-95*, pp. 725-728, Detroit, May. 1995.

[Zhan & Westohal 97] P. Zhan, M. Westohal, "Speaker normalization based on frequency warping," in Proc. ICASSP97, pp.1039-1042, 1997.