Analytic Methods for Acoustic Model Adaptation: A Review

Shigeki Sagayama[1,2], *Koichi Shinoda*[3], *Mitsuru Nakai*[2] *and Hiroshi Shimodaira*[2]

[1] *The University of Tokyo*, Bunkyo-ku, Tokyo 113-8656 Japan
[2] *Japan Advanced Institute of Science and Technology*, Tatsu-no-kuchi, Ishikawa 923-1292 Japan
[3] *NEC Laboratories, NEC Corporation*, Miyazaki, Miyamae-ku, 216-8555 Japan
E-mail: sagayama@hil.t.u-tokyo.ac.jp, {sagayama,mit,sim}@jaist.ac.jp, k-shinoda@bc.jp.nec.com

Abstract

This paper discusses analytic methods of acoustic model adaptation for automatic speech recognition and reviews other major methods. The main purpose of this paper is to demonstrate the potential of analytic approach for model adaptation. As an example of analytic methods, Jacobian Adaptation (JA) is intensively discussed and its potential of applicability to speech recognition problems is revealed. Vector Field Smoothing (VFS) is introduced as an extension of a special case of JA. Other method reviewed in this paper include Maximum A Posteriori (MAP) estimation, transformation-based approaches including Maximum Likelihood Linear Regression (MLLR), structural approaches, model selection including Eigenvoice, and feature compensation including Speaker Adaptive Training (SAT).

1 Introduction

Automatic speech recognition systems using continuous density hidden Markov models (HMMs) have been recently used in various applications, and speaker-independent (SI) systems have been constructed using speech samples collected from many speakers. It has been reported, however, that the performance of SI HMMs is often degraded when there is a mismatch between the training and testing environments. For example, when the acoustic characteristics of a new speaker are very different from those of the speakers in the training data, the recognition accuracy for the new speaker might be far below the average accuracy. Other major differences causing mismatches are those due to different microphones, channels, and noise environments.

Many techniques compensating the degradation caused by mismatches have been developed. They are roughly grouped into two categories, namely: (1) *feature compensation* (e.g. [30]), in which the process of feature extraction is modified; and (2) *model adaptation* (e.g. [16, 32]), in which the parameters of recognition models are adjusted. Combining these two techniques has been shown effective (e.g. [42]). Since it is almost impossible to cover all these techniques, the readers are also referred to two recent reviews [62, 31].

This paper mainly focuses on acoustic model adaptation. The fairly common framework in this domain is as follows. A sufficiently good acoustic model for an environment, named A, is given. When a small amount of data samples from the target environment, named B, is given, find a good way to adjust the initial model (originally for A) so that the adapted model performs well in B. If the mismatch between the acoustic model and the real environment is too large for any adaptation method to compensate, model selection from multiple models seems to be the best way to improve the performance. Therefore, model adaptation treats relatively small model mismatches.

In speaker adaptation case, adaptation is done from the initial speaker, A (or speaker-independent), to the target speaker, B. Similarly, adaptation to environmental noise and transmission channel can be considered as acoustic model adaptation from condition A to condition B.

"Analytic methods" refers to methods that utilizes analytic functions and their handling such as differentiation in this paper. One of typical approaches in this category is Jacobian approach which is contrasted with Bayesian approach. First, we discuss Jacobian adaptation for the case that the mapping function from the cause to the results are known and the difference between conditions A and B is given. Second, we discuss Vector Field Smoothing as a case where the mapping function is unknown. These are followed by a review of other major methods.

2 Jacobian Adaptation (JA)

2.1 Mathematical formulation

Jacobian approach (JA) [39, 63], proposed in 1997[1], is one of analytic approaches to adapting initial acoustic models under an initial condition, A, to a target condition, B, assuming that the difference between the two conditions are relatively small. It has been shown that this method provides a computationally efficient and effective algorithm for adaptation of acoustic models to a given noisy condition [45, 5].

This method provides a wide applicability to analytic relationships between causes and results. In general, if the results $\boldsymbol{Y}$ is an analytic function of causes $\boldsymbol{X}$, namely,

$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{X}), \tag{1}$$

$\Delta\boldsymbol{X}$, i.e., a small change in vector $\boldsymbol{X}$, causes $\Delta\boldsymbol{Y}$, i.e., another small change in vector $\boldsymbol{Y}$, which is well

[1] This idea was first conceived in 1992, experimented and presented in Japanese in 1996 and presented in English in 1997.

approximated by

$$\Delta \boldsymbol{Y} = \frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{X}} \Delta \boldsymbol{X} \tag{2}$$

where $\frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{X}} = \frac{\partial}{\partial \boldsymbol{X}} \boldsymbol{f}(\boldsymbol{X})$ represents a Jacobian matrix whose (i,j)-component is $\frac{\partial Y_j}{\partial X_i}$.

This simple math suggests a lot about model adaptation. In speech recognition, mismatch between an acoustic model and the target speaker and environment often causes a serious degradation of the performance. If the mismatch is so large that any adaptation method can not compensate it, model selection from multiple models seems to be the best way to improve the performance. Therefore, in most cases, model adaptation focuses on relatively small model mismatches. If there is a small mismatch $\Delta \boldsymbol{X}$, the following formula means that a small change in $\boldsymbol{X}$ causes another small change in $\boldsymbol{f}(\boldsymbol{X})$

$$\boldsymbol{f}(\boldsymbol{X}+\Delta \boldsymbol{X}) = \boldsymbol{f}(\boldsymbol{X}) + \frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{X}} \Delta \boldsymbol{X}. \tag{3}$$

For example, suppose $\boldsymbol{X}$ is a feature vector representing the noise spectrum in condition A, and $f(\boldsymbol{X})$ is the mean of a state output distribution of feature vectors in the initial noisy speech model. If the noise feature vector observed in condition B is slightly different from $\boldsymbol{X}$ and if the difference is $\Delta \boldsymbol{X}$, the mean vector of the noisy speech model should be adjusted to be $f(\boldsymbol{X}+\Delta \boldsymbol{X})$ which can be well approximated and easily calculated by the right-hand side of Eq.(3). Being linear, this calculation is generally fast and requires a small amount of computation.

Note that function $\boldsymbol{f}(\boldsymbol{X})$ can be any analytic function and can be non-linear. Also, this relation can be extended to multilayer relationships between the causes and results. If $\boldsymbol{U}$ is a variable related to $\boldsymbol{X}$, we can extend Eq.(2) to

$$\Delta \boldsymbol{Y} = \frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{X}} \Delta \boldsymbol{X} = \frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{X}} \frac{\partial \boldsymbol{X}}{\partial \boldsymbol{U}} \Delta \boldsymbol{U}. \tag{4}$$

Thus, we can even think of "Jacobian Network" which relates multiple nodes representing different factors by Jacobians just like the concept of "Bayesian Network".

2.2 Higher-order approximation

To make this linear approximation more accurate [63], one can introduce higher orders in the Taylor series of function $f(\boldsymbol{x})$ of vector $\boldsymbol{x}$ given by:

$$\boldsymbol{f}(\boldsymbol{X}+\Delta \boldsymbol{X}) = \sum_{k=0}^{\infty} \frac{1}{k!} (\Delta \boldsymbol{X}^T \boldsymbol{\nabla})^k \boldsymbol{f}(\boldsymbol{X}) \tag{5}$$

where

$$\boldsymbol{\nabla} = \left(\frac{\partial}{\partial X_1}, \frac{\partial}{\partial X_2}, \cdots, \frac{\partial}{\partial X_n}\right)^T \tag{6}$$

denotes the Laplacian operator. Note that the first-order term of the Taylor series is equivalent to Jacobian.

As for the use of Taylor series, Moreno et al. [34] defined a specific function in the spectral domain (not including cepstral features and other general cases):

$$z = x + q + \log(1 + e^{n-x-q}) \tag{7}$$

and used its Taylor series for adaptation to noise, where z, x and n represent log-spectra of degraded speech, clean speech and noise, and q is an unknown parameter that represents the effect of linear filtering in the log-spectrum domain. This paper is related to the Jacobian approach in the sense that first-order term of the Taylor series is Jacobian.

2.3 Jacobian adaptation to noisy cepstra

One of most common special cases is that the feature vector is cepstrum and the environmental factor is additive noise.

In the linear-scale power spectrum domain, noisy speech is represented by $\boldsymbol{Y} = \boldsymbol{S} + \boldsymbol{N}$ where n-dimensional vectors $\boldsymbol{S}$, $\boldsymbol{N}$, and $\boldsymbol{Y}$ denote the clean speech spectrum, additive noise spectrum, and the resulted composite speech spectrum, respectively. Here, vector operations are applied component by component[2]. The signal-to-noise ratio is not treated separately; the noise power is included in $\boldsymbol{N}$. In the cepstral domain, the composite speech cepstrum, $\boldsymbol{C}_Y$, is a non-linear function of the speech cepstrum, $\boldsymbol{C}_S$, and the noise cepstrum, $\boldsymbol{C}_N$, as follows:

$$\boldsymbol{C}_Y = \boldsymbol{F}^* \Big[\log \big\{ \exp(\boldsymbol{F}\boldsymbol{C}_S) + \exp(\boldsymbol{F}\boldsymbol{C}_N) \big\} \Big] \tag{8}$$

since these cepstra are related to the respective spectra by

$$\log \boldsymbol{Y} = \boldsymbol{F}\boldsymbol{C}_Y, \ \log \boldsymbol{S} = \boldsymbol{F}\boldsymbol{C}_S, \ \log \boldsymbol{N} = \boldsymbol{F}\boldsymbol{C}_N \tag{9}$$

where $\boldsymbol{F}$ is the $(m \times n)$ Fourier transform matrix (m and n represent the resolutions of frequency and quefrency axes) and $\boldsymbol{F}^*$ is the transposed complex conjugate of $\boldsymbol{F}$ that $\boldsymbol{F}^*\boldsymbol{F} = \boldsymbol{1}$.

The Jacobian matrix of the above non-linear function of $\boldsymbol{C}_S$ and $\boldsymbol{C}_N$ is given as follows:

$$\begin{aligned} \boldsymbol{J}_C &\equiv \frac{\partial \boldsymbol{C}_Y}{\partial \boldsymbol{C}_N} = \frac{\partial \boldsymbol{C}_Y}{\partial \log \boldsymbol{Y}} \frac{\partial \log \boldsymbol{Y}}{\partial \boldsymbol{Y}} \frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{N}} \frac{\partial \boldsymbol{N}}{\partial \log \boldsymbol{N}} \frac{\partial \log \boldsymbol{N}}{\partial \boldsymbol{C}_N} \\ &= \boldsymbol{F}^* \frac{\boldsymbol{1}}{\boldsymbol{H}(\boldsymbol{S}+\boldsymbol{N})} \boldsymbol{H}\boldsymbol{N}\boldsymbol{F} = \boldsymbol{F}^* \frac{\boldsymbol{N}}{\boldsymbol{S}+\boldsymbol{N}} \boldsymbol{F} \end{aligned} \tag{10}$$

This gives a practical calculation of Jacobian components:

$$\left[\frac{\partial \boldsymbol{C}_Y}{\partial \boldsymbol{C}_N}\right]_{ij} = \sum_k F_{ik}^{-1} \frac{N_k}{S_k + N_k} F_{kj} \tag{11}$$

Thus, if differences between the initial and observed conditions, A and B, is found in the cepstrum domain, i.e., $\Delta \boldsymbol{C}_N$, the composite cepstrum is approximately computed by Eq. (10).

[2] For vectors $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{ab} = (a_1 b_1, \cdots, a_n b_n)^T$, $\mathbf{a/b} = (a_1/b_1, \cdots, a_n/b_n)^T$, $\exp \mathbf{a} = (\exp a_1, \cdots, \exp a_n)^T$, and $\log \mathbf{a} = (\log a_1, \cdots, \log a_n)^T$. One can regard vectors as diagonal matrices for consistency with matrix arithmetic.

2.4 Jacobian adaptation of time derivatives

To demonstrate that the analytic approach can handle the time derivatives, we consider that the cepstrum is a continuous function of time from which we usually observe a sample sequence for discrete time points. Denote by $\dot{\boldsymbol{C}}$ the time derivative of $\boldsymbol{C}$. Because the time derivative of spectrum $\dot{\boldsymbol{S}}$ is related to the time-derivative of cepstrum $\dot{\boldsymbol{C}}$ by

$$\dot{\boldsymbol{S}}=\frac{\partial}{\partial t}\{\exp(\log \boldsymbol{S})\}=\exp(\log \boldsymbol{S})\frac{\partial}{\partial t}(\log \boldsymbol{S})=\boldsymbol{S}\boldsymbol{F}\dot{\boldsymbol{C}}_S \quad (12)$$

we obtain the Jacobian matrix of the time derivative of the composite cepstrum from Eq.(11) by using a relation between the time derivative of the linear spectrum and the cepstrum:

$$\begin{aligned}\frac{\partial \dot{\boldsymbol{C}}_Y}{\partial \boldsymbol{C}_N} &= \frac{\partial}{\partial t}\frac{\partial \boldsymbol{C}_Y}{\partial \boldsymbol{C}_N} = \boldsymbol{F}^*\frac{\partial}{\partial t}\left(\frac{\boldsymbol{N}}{\boldsymbol{S}+\boldsymbol{N}}\right)\boldsymbol{F} \\ &= \boldsymbol{F}^*\left(\frac{\dot{\boldsymbol{N}}\boldsymbol{S}-\boldsymbol{N}\dot{\boldsymbol{S}}}{(\boldsymbol{S}+\boldsymbol{N})^2}\right)\boldsymbol{F} \end{aligned} \quad (13)$$

which leads to the practical calculation:

$$\left[\frac{\partial \dot{\boldsymbol{C}}_{S+N}}{\partial \boldsymbol{C}_N}\right]_{ij} = \sum_k F^{-1}_{ik}\frac{\dot{N}_k S_k - N_k \dot{S}_k}{(S_k+N_k)^2}F_{kj} \quad (14)$$

though the mean of the delta cepstrum of the noise signal can be assumed to be close to 0 and the above formula is not used in practice.

2.5 Jacobian Adaptation of Means and Variances

Another demonstration of analytic approach is adaptation of mean vector and covariance matrix of a distribution. Assuming that the variance of the distribution of $\boldsymbol{C}_Y$ is sufficiently small and stays within the effective range of linear (Jacobian) approximation, we can extend the above relationship of point-to-point mapping to mean vectors of statistical distributions. In other words, if the Jacobian matrix can be regarded as a constant within the distribution range, small changes of mean values of $\boldsymbol{C}_Y$, and $\boldsymbol{C}_N$ are related to each other by ($\overline{\boldsymbol{X}}$ denotes the mean of $\boldsymbol{X}$.):

$$\Delta\overline{\boldsymbol{C}_Y} = \boldsymbol{J}_C\Delta\overline{\boldsymbol{C}_N} \quad (15)$$

$$\Delta Cov[\boldsymbol{C}_Y] = \boldsymbol{J}_C\ \Delta Cov[\boldsymbol{C}_N]\ \boldsymbol{J}_C^T \quad (16)$$

if $\boldsymbol{C}_Y$, $\boldsymbol{C}_N$, and $\Delta\boldsymbol{C}_N$ are statistically independent. In practical cases, however, the latter formula is not fully utilized due to lack of sufficient data to accurately obtain $\Delta Cov[\boldsymbol{C}_N]$.

2.6 Practical issues

In practical applications of speech recognition where environmental conditions often vary from time to time (e.g., mobile applications) or with each usage (e.g., telephone applications), acoustic model mismatch results in a serious degradation of performance. Only if a sufficient amount of training data is available, retraining acoustic models using the actual noise environment is feasible, but not in most cases.

Acoustic model composition from clean speech model and noise model such as PMC [14] (or NOVO [33]) requires too much computation to follow in real-time the instantaneous changes in noise spectrum and level. Moreover, these methods tend to require a considerable amount of training noise sample data.

Jacobian adaptation is advantageous in this point of view. The practical procedure is as follows:

Training Phase: In the training phase,

Step 1 - Train the model under the initial condition A**.** Assume an initial noise condition as the reference (Condition A). Train the initial speech model (CM-HMMs) under the initial condition with real or simulated (e.g., noise-added) speech data, or use PMC [14] (or NOVO [33]) to compose HMMs from clean speech and noise models. Also from the initial condition, obtain initial means of noise, $\overline{\boldsymbol{C}_N}$.

Step 2 - Calculate Jacobian matrices. For each mean vector of all the mixture components in the CM-HMMs, calculate the corresponding linear spectrum $\boldsymbol{S}$ [with Eq. (9)] and Jacobian matrices $\boldsymbol{J}_C$ for the cepstrum [Eq. (11)].

In the recognition phase,

Recognition Phase:

Step 3 - Observe noise and channel under the target condition B**.** Obtain the noise and channel cepstra and find the differences of the mean vectors, $\Delta\overline{\boldsymbol{C}_N}$, between the initial and target conditions (i.e., Conditions A and B).

Step 4 - Update mean vectors and variance matrices. Update all cepstrum and delta-cepstrum mean vectors and variances of mixture components in the HMMs by Eqs.(16) and (**??**), namely,

$$\overline{\boldsymbol{C}_Y} \leftarrow \overline{\boldsymbol{C}_Y} + \boldsymbol{J}_C\Delta\overline{\boldsymbol{C}_N} \quad (17)$$

It should be noted that the Fourier transform is essentially the cosine transform in real symmetric spectrum cases, namely, $F_{ij} = \cos\frac{2ij\pi}{n}$ and the computation amount can be reduced to one quarter by handling the positive frequency only. One implementation is

$$F_{ik} = \cos\frac{i(k+0.5)\pi}{N} \quad (18)$$

where N stands for the number of frequency points. Cepstrum can be simply replaced by MFCC in the above formulation where the corresponding spectrum is replaced by mel-frequency warped spectrum.

Cerisara et al. [5] and Sarikaya et al. [43] proposed to emphasize the noise spectrum as to improve linear approximation to cover a wider range and yielded even better results compared with PMC [14]. They also suggest to cluster Jacobian matrices into a small number to save the memory space without no significant degradation of performance.

2.7 Adaptation to noise and channel

The Jacobian adaptation algorithm for additive noise and multiplicative channel can be easily derived by replacing Eq.(8) with

$$\boldsymbol{C}_Y = \boldsymbol{F}^* \Big[\log \big\{ \exp(\boldsymbol{F}\boldsymbol{C}_S) + \exp(\boldsymbol{F}\boldsymbol{C}_N) \big\} \Big] + \boldsymbol{C}_H \quad (19)$$

and by following the same derivation as in the noise-only case. to obtain an almost same algorithm except for an additional term representing the channel difference as follows:

$$\overline{\boldsymbol{C}_Y} \leftarrow \overline{\boldsymbol{C}_Y} + \boldsymbol{J}_C \Delta\overline{\boldsymbol{C}_N} + \Delta\overline{\boldsymbol{C}_H} \quad (20)$$

which is useful if noise and channel cepstra can be observed separately.

2.8 Joint adaptation to noise and channel

Noise and channel differences, however, can not be observed separately in most cases such as in telephone speech recognition. This problem of unknown noise and channels involves a joint estimation problem for noise and channel differences between conditions A and B. Noting that Eq. (10) holds at mean vectors of all distributions aligned to the input speech, we have

$$\begin{cases} \Delta\overline{\boldsymbol{C}_Y^{(1)}} = \boldsymbol{J}_C^{(1)} \Delta\overline{\boldsymbol{C}_N} + \Delta\overline{\boldsymbol{C}_H} + \boldsymbol{\epsilon}^{(1)} \\ \Delta\overline{\boldsymbol{C}_Y^{(2)}} = \boldsymbol{J}_C^{(2)} \Delta\overline{\boldsymbol{C}_N} + \Delta\overline{\boldsymbol{C}_H} + \boldsymbol{\epsilon}^{(2)} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \Delta\overline{\boldsymbol{C}_Y^{(M)}} = \boldsymbol{J}_C^{(M)} \Delta\overline{\boldsymbol{C}_N} + \Delta\overline{\boldsymbol{C}_H} + \boldsymbol{\epsilon}^{(M)} \end{cases} \quad (21)$$

where M is the number of distributions contained in the acoustic model Viterbi aligned to the input speech (that can be the total number of mixture components) and $\epsilon^{(i)}$ is the i-th error term. Thus, if a small amount of input speech with unknown noise and channel is given with its phonetical transcription, we can obtain the joint estimate of $\Delta\overline{\boldsymbol{C}_N}$ and $\Delta\overline{\boldsymbol{C}_H}$ by simple least squares estimation. Once the linear decomposition into $\Delta\overline{\boldsymbol{C}_N}$ and $\Delta\overline{\boldsymbol{C}_H}$ are estimated, we can apply Eq. (10) to estimate all mean vectors of the model.

It should be noted that, even if some equations in the above simultaneous linear equations are missing, the common coefficients, $\Delta\overline{\boldsymbol{C}_N}$ and $\Delta\overline{\boldsymbol{C}_H}$, are estimated through least squares fit and applied to missing equations. This gives a solution to the "unseen context" problem. One typical result of experimental performance evaluation gave a 10% error reduction rate for 8 word utterances for supervised adaptation.

2.9 Adaptation to VTL

The Jacobian approach can be extended to include some aspects of speaker differences [40]. If the vocal tract length becomes λ-times longer, the corresponding speech spectrum changes from $f(\omega)$ into $f(\lambda\omega)$. The Jacobian matrix $\boldsymbol{J}_V$ of resulted cepstrum in respect to λ matrix has only one column and appears like a vector.

Combining noise, channel, and vocal tract length factors, we can express the small changes in the composite cepstrum as follows using small changes in noise cepstrum, channel cepstrum, and the vocal length stretch coefficient:

$$\Delta\boldsymbol{C}_Y = \boldsymbol{J}_N \Delta\boldsymbol{C}_N + \Delta\boldsymbol{C}_H + \boldsymbol{J}_V \Delta\lambda \quad (22)$$

In Eqs.(9), if we replace the Fourier Transform $\boldsymbol{F}$ with a λ-stretched Fourier transform $\boldsymbol{F}^\lambda$, or, if we use

$$F_{ik}^\lambda = \cos\frac{\lambda i(k+0.5)\pi}{N} \quad (23)$$

instead of Eq.(18), the λ-stretched speech spectrum $\tilde{\boldsymbol{S}}$ is given by

$$\log\tilde{\boldsymbol{S}} = \boldsymbol{F}^\lambda \boldsymbol{C}_S. \quad (24)$$

The λ-stretched cepstrum $\tilde{\boldsymbol{C}}_S$ is thus expressed as

$$\tilde{\boldsymbol{C}}_S = \boldsymbol{F}^{-1}\boldsymbol{F}^\lambda \boldsymbol{C}_S \quad (25)$$

whose i-th component is given by

$$\tilde{C_{Si}} = \sum_{j=1}^{N} F_{ij}^{-1} \sum_{k=1}^{p} F_{jk}^\lambda C_{Sk} \quad (26)$$

from which the i-th component of its Jacobian in respect to λ is derived by differentiating it by λ as

$$J_{\lambda i} = \sum_{j=1}^{N} F_{ij}^{-1} \sum_{k=1}^{p} \frac{-j(k+0.5)\pi}{N} G_{jk} C_{Sk} \quad (27)$$

where matrix G represents the sine transform.

The joint estimation of noise, channel, and vocal tract length can be formulated in the same way as described for the noise and channel case [41].

3 Vector Field Smoothing (VFS)

Vector Field Smoothing (VFS) [35, 17], proposed in 1992, is an effective and easy-to-use method for speaker and channel adaptation and actually is frequently used. It can be understood as a non-parametric version of Jacobian adaptation. Assume that a set of good initial acoustic models is given but only a limited amount of training data is available for adaptation to the target speaker B. The initial model has been well trained by a large amount of data from speaker A (or speaker-independent). VFS assumes that the mapping function from speaker A to speaker B is unknown but known to be smooth, i.e., does not change suddenly in the feature vector space.

VFS speaker adaptation is performed as follows. Given the speaker B's arbitrary utterance with its phonetical transcription, mean vectors of constituent distributions in the model are retrained through embedded training using the initial models of speaker A. The difference $\Delta\boldsymbol{C}_Y$ between mean vectors before and after retraining is found. These vectors are regarded as samples of underlying true $\Delta\boldsymbol{C}_Y$ containing statistical

fluctuations and estimation errors. If the contained errors are supposed to be random, they can be reduced by spatial smoothing such as:

$$\Delta\tilde{\boldsymbol{C}}_Y{}^{(i)} = \sum_{j\in K_i} w_j \Delta\boldsymbol{C}_Y^{(j)} \tag{28}$$

where w_j means the spatial smoothing filter weights and K_i is the k-nearest neighbor of the distribution i. The weight w_i can be any that satisfies good characteristics of smoothing. In the original paper, Ohkura et al. [35] used a fuzzy membership function, while some others used Gaussian smoothing windows later. The smoothing window spreadth is empirically determined so as to make a best compromise between spatial resolution and statistical error reduction.

Missing vectors between corresponding states (or Gaussians) of speakers A and B are automatically estimated and recovered from the vector field through smoothing, just as in joint Jacobian adaptation to noise and channel. This feature is one of outstanding advantages of this method and inherits the advantage of spatial interpolation from a coding paper [44].

This algorithm is usually performed in the cepstral ot MFCC multidimensional space. The difference between linear microphone characteristics can be normalized by parallel shifting the acoustic models in the cepstrum multidimensional space since a linear system represented by convolution in the time domain is transformed into multiplication in the spectral domain, and addition in the logarithmic spectral domain and the cepstral domains. The spectral change from noiseless to noisy environments is also supposed to be continuous and smooth in the cepstral space. Thus, VFS adapts the system to the new speaker, new microphone, new noise environment, and some other factors affecting the speech spectra simultaneously.

VFS can be interpreted as position-dependent channel bias vectors. In the cepstral domain, linear channel affects cepstrum vectors $\Delta\boldsymbol{C}$ as a constant bias vector $\Delta\boldsymbol{C}$ that is added to all cepstral vectors in an acoustic model. This is interpreted as a uniform vecter field in a cepstral vector space in which vector points are carried along the vector field just like floating objects carried by a flow. If the cepstral difference between speakers could be simply modeled by such a uniform vector field, speaker adaptation would be a simple addition of $\Delta\boldsymbol{C}$ to all cepstral points. The bias vector, $\Delta\boldsymbol{C}$, is, however, not uniform throughout the vector space but dependent on the location. Thus, $\Delta\boldsymbol{C}$ is a function of $\boldsymbol{C}$ not known in the analytic sense.

4 Maximum A Posteriori (MAP) Estimation

Maximum *A Posteriori* (MAP) estimation [9, 16] has been widely used approach for model adaptation. We consider the case where the parametric form of the probabilistic density function (pdf) $p(x)$, where x is a k-component vector-valued random variable, is the multivariate Gaussian pdf,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{k}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right], \tag{29}$$

while neither the mean vector $\boldsymbol{\mu}$ nor the variance $\boldsymbol{\Sigma}$ are known. Let $\mathcal{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$ be a set of observed samples, which are assumed to be independent and identically distributed (i.i.d.). Our goal is to estimate the parameter set $\theta = \{\boldsymbol{\mu},\boldsymbol{\Sigma}\}$ by using the observation samples $\mathcal{X}$.

Maximum likelihood (ML) estimation is often used for this purpose. In the ML estimation, the parameter set which maximizes the following likelihood function is chosen,

$$f(\mathcal{X}|\theta) = \prod_{n=1}^{N} p(\mathbf{x}_n|\theta). \tag{30}$$

The resulting maximum likelihood estimate, $\tilde{\theta} = \{\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}\}$, is calculated as follows,

$$\tilde{\boldsymbol{\mu}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n, \tag{31}$$

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\tilde{\boldsymbol{\mu}})(\mathbf{x}_n-\tilde{\boldsymbol{\mu}})^T. \tag{32}$$

In the maximum *a posteriori* (MAP) estimation [9, 16], it is assumed that the parameter set θ is a random vector in the parameter space and it has a prior distribution $p(\theta)$. Let $p(\theta|\mathcal{X})$ be the posterior pdf that is obtained after the observation of $\mathcal{X}$. Then, using Bayes' rule,

$$\begin{aligned} p(\theta|\mathcal{X}) &= \frac{p(\mathcal{X}|\theta)p(\theta)}{\int p(\mathcal{X}|\theta)p(\theta)d\theta} \\ &= C\prod_{n=1}^{N} p(\mathbf{x}_n|\theta)p(\theta), \end{aligned} \tag{33}$$

where C is a scale factor that depends on $\mathcal{X}$ but is independent of θ. The MAP estimate $\hat{\theta}$ is defined as the mode of the posterior pdf,

$$\begin{aligned} \hat{\theta} &= \operatorname*{argmax}_{\theta} p(\theta|\mathcal{X}) \\ &= \operatorname*{argmax}_{\theta} \prod_{n=1}^{N} p(\mathbf{x}_n|\theta)p(\theta). \end{aligned} \tag{34}$$

The choice of the prior pdf is a key issue in MAP estimation. Mainly from the viewpoint of tractability, the conjugate prior pdf is often used; when using it, the resulting posterior pdf is in the same family as the one that the prior pdf belongs to. One such pdf for the multivariate Gaussian pdf $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the normal-Wishart density of the form,

$$\begin{aligned} g(\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0,\alpha,\tau) \propto \\ |\boldsymbol{\Sigma}|^{-\frac{\alpha-k}{2}}\exp\left[\frac{\tau}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right] \\ \times\exp\left[-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}^{-1})\right], \end{aligned} \tag{35}$$

where $(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0,\alpha,\tau)$ are the prior density parameters such that $\alpha > k-1$, $\tau > 0$, $\boldsymbol{\mu}_0$ is a vector of dimension k, and $\boldsymbol{\Sigma}_0$ is a $k\times k$ positive definite matrix.

Then, the MAP estimate $\hat{\theta} = \{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\}$ is the one that maximizes the following function,

$$g(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{X}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) g(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \quad (36)$$

After simple calculations, we get

$$\hat{\boldsymbol{\mu}} = \frac{\tau\boldsymbol{\mu}_0 + \sum_{n=1}^{N} \mathbf{x}_n}{\tau + N}, \quad (37)$$

$$\hat{\boldsymbol{\Sigma}} = \frac{\boldsymbol{\Sigma}_0 + \sum_{n=1}^{N} (\mathbf{x}_n - \hat{\boldsymbol{\mu}})(\mathbf{x}_n - \hat{\boldsymbol{\mu}})^T + \tau(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}})^T}{(\alpha - k) + N}. \quad (38)$$

The improvement obtained with MAP estimation is significantly larger than that obtained with ML estimation, especially when the amount of adaptation data is small. It should be noted that as the number of samples, N, increases, the MAP estimate $\hat{\theta} = \{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\}$ approaches the ML estimate $\tilde{\theta} = \{\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}\}$.

A quasi-Bayes approach [18] has also been adopted and extended the MAP framework to *on-line* MAP adaptation.

5 Transformation-based Approach

5.1 overview

Another category of adaptation techniques, which do not use the MAP framework, are often referred to as *transformation-based* approaches, such as *cepstrum mean normalization* (CMN) [1], *signal bias removal* (SBR) [38], *maximum likelihood linear regression*(ML LR)[32], *spectral interpolation* [46], *vector field smoothing* (VFS) [35], *stochastic matching* (SM) [42], *nonlinear stochastic matching* [54] and *predictive adaptation* [55]. This family of techniques limits the number of free parameters by tying the HMM parameters or by applying some constraints on the parameters in order to improve recognition accuracies with a small amount of data.

5.2 Maximum Likelihood Linear Regression (MLLR)

Maximum Likelihood Linear Regression (MLLR) [32] is one of most famous transformation-based approach. In MLLR, the mean vectors of Gaussian distributions in HMMs are modified using an affine transformation,

$$\hat{\mu} = \mathbf{A}\mu + b, \quad (39)$$

where A is an $n \times n$ matrix, and b is a vector of dimension n, in which n is the dimension of the observation vector. This equation is rewritten as follows:

$$\hat{\mu} = \mathbf{W}\xi, \quad (40)$$

where $\mathbf{W}$ is an $n \times (n+1)$ matrix and $\xi^T = (1, \mu_1, \ldots, \mu_n)$. The matrix W is estimated by E-M algorithm using the adaptation data.

In this approach, the Gaussian distributions in HMMs are clustered into some groups and one transformation matrix is shared in the distributions in each group. Usually, one global matrix is share among all the distributions or one matrix is provided for each phone.

It should be noted that it is assumed here that the mapping of mean vectors can be efficiently approximated by an affine transformation. It is clear, however, that this assumption is not true when one global matrix is used for all the distributions. Therefore, the key issue in this approach is how to provide the clusters of distributions. The number of clusters should be controlled according to the amount of data to avoid the data insufficiency problems.

6 Structural approach

It is desirable that adaptation improves speech recognition accuracies even when little adaptation data is given and more importantly it yields performance equal to or better than that obtained using *maximum likelihood* (ML) estimation when enough data is available. Few methods, however, achieve both objectives.

In MAP estimation methods, HMM parameters of different speech units are often assumed to be independent. Therefore, each model can be adapted only if the corresponding speech unit has been observed in the current set of adaptation data. The improvement is consequently rather small when the amount of adaptation data is extremely limited.

In transformation-based approach, when the amount of adaptation data exceeds a certain value, the recognition accuracy often becomes inferior to that obtained with ML estimation of the model parameters. This is because a model with a small number of free parameters could not fully utilize the potential information embedded in the large amount of data.

6.1 Adaptation using fixed structure

Because the MAP approaches and the transformation based methods are not capable of either improving recognition accuracy when little data are available or exploiting the information in a large amount of data, several algorithms supplementing those techniques have been developed. The *Extended MAP* (EMAP) method [53, 65], and the *quasi-Bayes* technique with *correlated* mean vectors [19] are extensions of the MAP approaches. They increase the recognition rates obtained with a small amount of data by taking into account the *a priori* knowledge in the correlation between the parameters modeling different speech units. For example, the pair-wise correlation between the mean vectors could be used to enhance estimation of the mean parameters of some speech units even if they are not directly observed in the adaptation data and therefore the recognition rates are significantly improved [19]. Although these methods are in theory quite general,

they need to impose some approximation in practice because it is difficult to estimate such correlations precisely when the amount of training data is small. In [53, 65], for example, the model parameter space was divided into several subspaces, the ideal number might depend on the amount of adaptation data available.

It is also possible to extend the known ML techniques, such as MLLR to incorporate the MAP estimation criterion. The recently proposed *maximum a posteriori linear regression* (MAPLR) [52] algorithm improves MLLR in a way similar to MAP enhancement over ML for HMM parameter estimation. Combinations of MAP and transformation-based approaches have also been studied intensively ([10, 8, 56, 57]). Notable examples were in combining MLLR and MAP [10] and combining MAP and VFS [56, 57]. Chien *et al.* [8] reported that significantly better recognition accuracy can be obtained by combining MAP and SM with an iterative process.

It should be noted that Furui [13] has already developed an unsupervised adaptation method that utilizes a hierarchical structure for vector quantization.

6.2 Adaptation using flexible structure

The shortcoming of these combined methods is the use of fixed *structures*, i.e. fixed ways of parameter tying, in the acoustic space. Therefore they have only been shown useful with adaptation data sizes within a narrow range. To alleviate this problem, a *tree structure* has been used in adjusting the number of layers in a tree and the degree of parameter tying according to the amount of available data (e.g. [47, 48]).

Recently, Shinoda and Lee proposed *structural Bayes adaptation* framework that achieves the two desired objectives mentioned earlier. In this approach, we take advantage of the nice asymptotic property of MAP estimation for large size adaptation and the flexible parameter tying strategy in a tree for small sample adaptation.

For example, in *structural maximum a posteriori* (SMAP) algorithm [49, 50, 51], the prior knowledge in a tree node is used to construct prior density needed for MAP estimation of all the parameters in the successive child nodes. 20Three key steps are required in formulating the proposed SMAP approach. They are described in the next three Sections. First, a tree with a uniform structure is constructed to characterize the acoustic space represented by the HMM parameters. Next, given all the density clusters used to characterize nodes in a tree, a Gaussian density is estimated, which summarize all the Gaussian components in the cluster so that the likelihood of a sequence of observation vectors representing the adaptation data can be evaluated at the node level and therefore the MAP estimate at any node in the tree can be computed. For the third step, the prior density at each tree node needs to be defined. In this step, *hierarchical prior evolution approximation* is used, in which it is assumed that the hyperparameters charactering the prior density at each node are evaluated based on the knowledge embedded in the prior density of its parent node.

Recently this structural Bayes approach has been also combined with MAPLR approach (SMAPLR [52]) and proved to be effective.

7 Model Selection

In model selection scheme, a number of models prepared beforehand, and the optimal model for a new speaker is selected. Speaker clustering (e.g., [24, 36]) has been mostly employed in this scheme, while some novel techniques has been recently developed.

7.1 Eigenvoice

Recently Eigenvoice approach [25, 26] has been proposed for speaker adaptation. In this approach, first the speaker-dependent models from many speakers are provided, and then, the principle component analysis (PCA) is carried out for model parameters of each speaker model, and the lower order eigen-vectors are selected as *eiganvoices*. For a new speaker, the weight for each eigenvoice are estimated in a maximum-likelihood estimation using a small amount of adaptation data.

It should be noted that, since the model parameters for each speaker models are usually very large, it is almost impossible to use all of them. The parameters used for PCA has to be carefully selected from the model parameters. The Gaussian mean vectors of single-Gaussian monophone models are often used.

This method is proved to be significantly useful especially when the adaptation data available is extremely small [4, 27]. It should be noted that the method in this category are also effective as a bootstrap for the model adaptation methods such as MAP or MLLR (e.g. [7, 22, 58]).

8 Feature compensation

8.1 CMN and VTLN

In feature compensation, features that are dependent on individual speakers or channels are subtracted from observed features. Extensive studies have been conducted on cepstrum mean normalization (CMN) [1] and vocal tract length normalization (VTLN) [12].

In CMN, the long-term average of the cepstrum is subtracted from the cepstrum of each data frame. This helps eliminate changes created not only by differences among individual speakers, but also by environmental noise and channel changes, for which changes are much slower than the changing phonetic features of speech itself.

Since vocal tract length differs from speaker to speaker, so do the formant frequencies in the power spectrum for each speaker. In VTLN, vocal tract lengths are estimated using each speaker's spectrum, but since it is difficult to precisely estimate a vocal tract length from a spectrum, some studies have used a maximum-likelihood VTLN (ML-VTLN) selection method [29, 59, 61, 64]. With this method, a number of param-

eters, each of which represents a different vocal tract length, are prepared beforehand, and the parameter that maximizes the likelihood of the data is selected.

8.2 Speaker Adaptive Training (SAT)

A method called speaker adaptive training (SAT) has recently come into frequent use [2, 21, 37, 60]. Here, so long as speaker adaptation will always be carried out for each speaker, a standard-speaker-dependent model (i.e., a speaker-dependent model based on the speech of a *standard* speaker) will be more appropriate for use as the initial model than a speaker-independent model (i.e., a model representing variations in the utterances of a large number of speakers). In SAT, the parameters for a standard-speaker-dependent model are estimated in the following process. First, a mapping from the parameters of the model created for each individual speaker to those of an initial model is estimated. Second, this estimated mapping is used to map the utterance data for each speaker. Third, this mapped data is used to train the standard-speaker-dependent model. This process is iterated until convergence. While an affine transformation is often used for the mapping, since the number of parameters to be estimated is relatively large, it is difficult to precisely estimate its parameters when the number of utterances is small.

Recently, cluster adaptive training (CAT) [15] has been proposed. In this method, more than one models constructed by speaker clustering are employed for SAT, and the models themselves and the weights between them are simultaneously estimated.

Both feature compensation and SAT remove the variations in input speech data caused by differences in individual speaker characteristics. The effectiveness of these methods depends mainly on the method used to conduct the mapping, for which precise estimation needs to be achieved on the basis of only a small amount of data from each speaker.

9 Conclusion

In this paper, analytic methods for model adaptation such as Jacobian Adaptation and Vector Field Smoothing has been mainly discussed and followed by a review of other major methods. Analytic approaches provides simple, easy-to-use, and efficient alogorithms for model adaptation to multiple factors.

References

[1] 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.

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

[3] L.R. Bahl, P.V. de Sousa, P.S. Gopalakrishnam, D. Nahamoo, and M.A. Picheny, "Decision Trees for Phonological Rules in Continuous Speech", *Proc. ICASSP91*, Toronto, pp.185-188, 1991.

[4] H. Botterweck, "Very fast adaptation for large vocabulary continuous speech recognition," *in Proc. ICSLP-2000*, pp.179-182, Beijing, 2000.

[5] C. Cerisara, L. Rigazio, R. Boman, and J.-C. Junqua, "Transformation of Jacobian Matrices for Noisy Speech Recognition," *Proc. ICSLP2000 (Beijing)*, pp. 369-372, 2000.

[6] C. Cerisara, L. Rigazio, R. Boman and J.-C. Junqua, "Environmental Adaptation Based on First Order Approximation," *Proc. ICASSP2001 (Salt Lake City)*, in CD-ROM, 2001.

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

[8] J.-T. Chien, C.-H. Lee, and H.-C. Wang, "Improved Bayesian Learning of Hidden Markov Models for Speaker Adaptation," *Proc. ICASSP-97*, pp. 1027-1039, 1997.

[9] M.H. DeGroot, *Statistical Decision Theory and Bayesian Analysis*, McGraw-Hill, 1970.

[10] 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.

[11] V. Digalakis, S. Berkowitz, E. L. Bocchieri, C. Boulis, W. J Byrne, H. Collier, A. Corduneanu, A. Kannan, S. P. Khudanpur, and A. Sankar, "Rapid Speech Recognizer Adaptation to New Speakers," *Proc. ICASSP-99*, pp. 765-768, Phoenix, May 1999.

[12] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP96*, vol. 1, pp. 346-3483, 1996.

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

[14] M. J. F. Gales and S. J. Young, "An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise," *Proc. ICASSP92*, pp. 233–236, 1992.

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

[16] 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.

[17] H. Hattori and S. Sagayama, "Vector Field Smotthing Principle for Speaker Adaptation," *Proc. ICSLP92*, pp. 381–384, Banff, October, 1992.

[18] Q. Huo and C.-H. Lee, "On-line Adaptive Learning of the Continuous Density Hidden Markov Model Based on Approximate Recursive Bayes Es-

timate," *IEEE Trans. on Audio and Speech Processing*, Vol. 5, No. 2, pp. 161-172, March 1997.

[19] 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.

[20] H. Jiang, K. Hirose, and Qiang Huo, "Robust Speech Recognition Based on Viterbi Bayesian Predictive Classification," *Proc. ICASSP-97*, pp. 1551-1554, Berlin, April, 1997.

[21] 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.

[22] E. Jon, D. K. Kim, and N. S. Kim, "EMAP-based speaker adaptation with robust correlation estimation," *in Proc. ICASSP-2001*, P.3.3, 2001.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] C.-H. Lee, "On Stochastic Feature and Model Compensation Approaches to Robust Speech Recognition," *Speech Communication*, Vol. 25, pp. 29-47, 1998.

[31] 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.

[32] 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.

[33] F. Martin, et al., "Recognition of Noisy Speech by Using the Composition of Hidden Markov Models," *Proc. 1992 Fall ASJ Conf.,* 1-7-10, 1992.

[34] P. J. Moreno, B. Raj, and R. Stern, "A Vector Taylor Series Approach for Environment-Independent Speech Recognition," *Proc. ICASSP96*, pp. 733–736, 1996.

[35] 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, Banff, 1992.

[36] 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.

[37] 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.

[38] 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.

[39] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, "Jacobian Approach to Fast Acoustic Model Adaptation," *Proc. ICASSP97*, Vol. 2, pp. 835–838, 1997.

[40] Shigeki Sagayama, Yutaka Kato, Mitsuru Nakai and Hiroshi Shimodaira, "Jacobian Approach to Joint Adaptation to Noise, Channel and Vocal Tract Length," *Proc. Int. Workshop on Adaptation Methods (Sophia-Antipolis)*, this issue, 2001.

[41] Nobuyoshi Sakai, Mitsuru Nakai and Hiroshi Shimodaira and Shigeki Sagayama, "Simultaneous Adaptation to Noise and Channel Distortion adn Speaker Using Jacobian Adaptation," *Proc. 2001 ASJ Autumn Conference (Oita)*, to appear.

[42] 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.

[43] R. Sarikaya and J. H. L. Hansen, "Improved Jacobian Adaptation for Fast Acoustic Adaptation in Noisy Speech Recognition," *Proc. ICSLP2000 (Beijing)*, pp. 702–705, 2000.

[44] Shiraki, Y., and Honda, M., "LPC Speech Coding Based on Variable-Length Segment Quantization," IEEE Trans. ASSP, Vol.36, No. 9, pp. 1437–1444, Sep 1988.

[45] Hiroshi Shimodaira, Toshihiko Akae, Mitsuru Nakai, Shigeki Sagayama, "Jacobian Adaptation

of HMM with Initial Model Selection for Noisy Speech Recognition," *Proc. ICSLP2000 (Beijing)*, pp.1003-1006, 2000.

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

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

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

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

[50] K. Shinoda and C.-H. Lee, "Unsupervised Adaptation Using Structural Bayes Approach," *Proc. ICASSP98*, pp. II 793-796, 1998.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] A. Surendran, and C.-H. Lee, "Bayesian Predictive Approach to Adaptation of HMMs," *Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, pp.155-158, Tampere, Finland, May 1999.

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

[57] 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.

[58] 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.

[59] S. Wegmann, D. Macllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc. ICASSP96*, vol. 1, pp. 339-341, 1996.

[60] 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.

[61] L. Welling, S. Kanthak, and H. Ney, "Improved methods for vocal tract normalization," in *Proc. ICASSP99*, no. 1436, 1999.

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

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

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

[65] 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.