# INCREMENTAL ADAPTATION BASED ON A MACROSCOPIC TIME EVOLUTION SYSTEM

*Shinji Watanabe and Atsushi Nakamura*

NTT Communication Science Laboratories, NTT Corporation

**ABSTRACT**

In this paper, we propose a new incremental model adaptation approach based on posterior distributions of model parameters. We consider a propagation mechanism of the posterior distributions whereby that the process of posterior refinement is modeled analytically. Then, we derive an incremental estimation algorithm based on a time evolution system, which explicitly includes a discrete stochastic process unlike the conventional Bayesian approaches. This algorithm is viewed as a general solution of the Kalman filter algorithm, where posterior distributions make a transition after every input of an utterance set, and where the evolutions of posterior distributions are represented on a macroscopic time scale.

***Index Terms***— Speech recognition, acoustic model, incremental adaptation, discrete stochastic process, macroscopic time evolution

## 1. INTRODUCTION

In real environments, speech characteristics and their acoustic conditions are constantly changing due to such factors as the variability of speaking style, and the superposition of non-stationary noise. Incremental adaptation techniques for speech recognition are aimed at adjusting acoustic models to track such time-variant conditions whereas batch-type adaptation techniques mainly deal with time-invariant conditions.

The straightforward incremental adaptation approach refines acoustic model parameters, step-by-step, after every partial input of speech, which most typically consists of a small set of utterances. The adaptation starts with a set of initial parameter values as in the case of batch-type adaptation. Then, the refined parameter values are used as initial parameter values in the next refinement step, so that the effect of the refinement propagates in succeeding steps. Since there are unavoidable estimation errors for a refinement that only uses a small amount of data, this adaptation scheme also propagates the errors, and this affects the adaptation stability. On the other hand, an incremental adaptation scheme that estimates the distributions of the parameters instead of the parameters themselves can mitigate the unstable adaptation caused by the influence of estimation errors. Incremental Bayesian approaches, for example, estimate the posterior distributions of model parameters [1, 2] or of transformation parameters [3], and use the results as prior distributions in the next estimation step. Since the effect of refinement is thus propagated via distributions, the influence of the estimation errors is absorbed into the distribution spread, which realizes a stable adaptation.

In this paper, we further enhance this propagation mechanism so that the posterior refinement process is modeled analytically. We derive an incremental estimation algorithm for the posterior distributions of model parameters based on a time evolution system, which explicitly includes a discrete stochastic process unlike the conventional Bayesian approaches. This algorithm is also viewed as a general solution of the Kalman filter algorithm, where posterior distributions make a transition after every input of an utterance set, and where the evolutions of posterior distributions are represented on a macroscopic time scale.

## 2. FORMULATION

### 2.1. Posterior distribution based incremental adaptation

As we start the formulation, we first define the following macroscopic time scale as an adaptation time unit by using a partial time series of adaptation feature vectors based on the utterances.

$$\boldsymbol{O} = \{\underbrace{\boldsymbol{o}_1, ..., \boldsymbol{o}_{N_1}}_{\boldsymbol{O}_1}, ... \underbrace{\boldsymbol{o}_{N_{t-1}+1}, ..., \boldsymbol{o}_{N_{t-1}+N_t}}_{\boldsymbol{O}_t}, ....\} \quad (1)$$

Here, $\boldsymbol{o}_n \in R^D$ denotes a $D$ dimensional feature vector at frame $n$, while $\boldsymbol{O}_t$ denotes a set of feature vectors at macroscopic time $t$.

The straightforward adaptation scheme focuses on updating acoustic model parameters $\theta_t$. Then, an incremental adaptation is realized by constructing a recurrence equation to represent the time evolution from $t-1$ to $t$:

$$\theta_t = T(\theta_{t-1}), \quad (2)$$

where $T(\cdot)$ denotes an arbitrary transformation function, such as the affine transformation used in the MLLR adaptation [4, 5].

On the other hand, this paper focuses on updating a posterior distribution of model parameter $\theta_t$ conditioned on the accumulated data from the beginning of adaptation, i.e., $\boldsymbol{O}^t = \{\boldsymbol{O}_1, \boldsymbol{O}_2, ..., \boldsymbol{O}_t\}$. Namely, the estimation target is, now, the posterior distribution of model parameter $p(\theta_t|\boldsymbol{O}^t)$, instead of model parameter $\theta_t$. Then, we consider the following recurrence equation to represent a time evolution:

$$p(\theta_t|\boldsymbol{O}^t) = \mathcal{T}[p(\theta_{t-1}|\boldsymbol{O}^{t-1})], \quad (3)$$

where $\mathcal{T}[\cdot]$ denotes an arbitrary transformation functional whose argument is posterior distribution $p(\theta_{t-1}|\boldsymbol{O}^{t-1})$. The following sections describe how to realize this incremental adaptation scheme in practice.

### 2.2. Incremental adaptation including a discrete stochastic process

We first introduce a concrete form of functional Eq. (3). By using the probabilistic product formula and Bayes theorem, the following recurrence equation can be derived analytically without using any approximations.

$$p(\theta_t|\boldsymbol{O}^t) \propto p(\boldsymbol{O}_t|\theta_t, \boldsymbol{O}^{t-1}) \int p(\theta_t|\theta_{t-1}, \boldsymbol{O}^{t-1}) p(\theta_{t-1}|\boldsymbol{O}^{t-1}) d\theta_{t-1}. \quad (4)$$

In this paper, instead of dealing with Eq. (4) directly, we introduce the Markov process approximation, i.e., $p(\boldsymbol{O}_t|\theta_t, \boldsymbol{O}^{t-1}) \rightarrow p(\boldsymbol{O}_t|\theta_t)$ and $p(\theta_t|\theta_{t-1}, \boldsymbol{O}^{t-1}) \rightarrow p(\theta_t|\theta_{t-1})$. In practice, we only consider Gaussian mean vector parameter $\boldsymbol{\mu}$ as time-variant, which is assumed to be the dominant parameter in speech recognition, i.e., $\theta \rightarrow \boldsymbol{\mu}$. Then, Eq. (4) can be rewritten as:

$$p(\boldsymbol{\mu}_t|\boldsymbol{O}^t) \propto p(\boldsymbol{O}_t|\boldsymbol{\mu}_t) \int p(\boldsymbol{\mu}_t|\boldsymbol{\mu}_{t-1}) p(\boldsymbol{\mu}_{t-1}|\boldsymbol{O}^{t-1}) d\boldsymbol{\mu}_{t-1}. \quad (5)$$

The right hand side of Eq. (5) consists of three distributions.

1. $p(\boldsymbol{O}_t|\boldsymbol{\mu}_t)$ is an output distribution.
2. $p(\boldsymbol{\mu}_t|\boldsymbol{\mu}_{t-1})$ denotes a discrete stochastic process of $\boldsymbol{\mu}_t$.
3. $p(\boldsymbol{\mu}_{t-1}|\boldsymbol{O}^{t-1})$ is a posterior distribution, which is already estimated in the previous adaptation step $t-1$.

The current posterior distribution $p(\boldsymbol{\mu}_t|\boldsymbol{O}^t)$ is incrementally evolved from the previously estimated posterior distribution $p(\boldsymbol{\mu}_{t-1}|\boldsymbol{O}^{t-1})$, which is dependent on $p(\boldsymbol{O}_t|\boldsymbol{\mu}_t)$ and $p(\boldsymbol{\mu}_t|\boldsymbol{\mu}_{t-1})$. This is an important characteristic of the proposal, which differentiates it from the conventional Bayesian approaches [1–3], by explicitly including the discrete stochastic process. The next section plugs the appropriate concrete forms into each distribution on the right hand side of Eq. (5) to obtain a practical solution for $p(\boldsymbol{\mu}_t|\boldsymbol{O}^t)$, analytically.

### 2.3. Analytic solution based on a linear dynamical system

First, we set the output distribution $p(\boldsymbol{O}_t|\boldsymbol{\mu}_t)$ from a standard acoustic model, which is represented by a Hidden Markov Model (HMM), and a Gaussian Mixture Model (GMM). $p(\boldsymbol{O}_t|\boldsymbol{\mu}_t)$ outputs a partial time series of adaptation feature vectors $\boldsymbol{O}_t = \{\boldsymbol{o}_{N_t+1}, ..., \boldsymbol{o}_{N_t+N_{t+1}}\}$, and is represented by:

$$p(\boldsymbol{O}_t|\boldsymbol{\mu}_t) = \prod_{n=N_t+1}^{N_t+N_{t+1}} \mathcal{N}(\boldsymbol{o}_n|\boldsymbol{\mu}_t, \Sigma), \quad (6)$$

where $\mathcal{N}(\cdot|\boldsymbol{\mu}_t, \Sigma)$ denotes a Gaussian distribution, which has a time-variant mean vector parameter $\boldsymbol{\mu}_t$ and a time-invariant covariance matrix parameter $\Sigma$. $\Sigma$ is obtained from an initial acoustic model. Although Eq. (6) omits state transition and mixture weight parameters, and latent variables included in the HMMs and GMMs, these are considered in the E-step of the EM algorithm.

Second, we introduce an affine transformation to provide a simple representation of the discrete stochastic process of $\boldsymbol{\mu}_t$ as follows:

$$\boldsymbol{\mu}_t = A\boldsymbol{\mu}_{t-1} + \boldsymbol{\nu} + \varepsilon_{t-1}, \quad (7)$$

where $A$ is a $D \times D$ matrix, which denotes a linear transformation consisting of the rotation and scaling of $\boldsymbol{\mu}_{t-1}$, and $\boldsymbol{\nu}$ is a $D$ dimensional vector, which denotes a bias transformation. $\varepsilon_{t-1}$ is called a system noise, which is represented by a Gaussian with a $\mathbf{0}$ mean vector and covariance matrix $U$. This can be viewed as an affine transformation that is fluctuated by noise $\varepsilon_{t-1}$. Then, the discrete stochastic process is represented by the following concrete form:

$$p(\boldsymbol{\mu}_t|\boldsymbol{\mu}_{t-1}) = \mathcal{N}(\boldsymbol{\mu}_t|A\boldsymbol{\mu}_{t-1} + \boldsymbol{\nu}, U). \quad (8)$$

Finally, we assume that $p(\boldsymbol{\mu}_{t-1}|\boldsymbol{O}^{t-1})$ is represented by a Gaussian, which has a mean vector parameter $\hat{\boldsymbol{\mu}}_{t-1}$ and a covariance matrix parameter $\hat{Q}_{t-1}$ as follows:

$$p(\boldsymbol{\mu}_{t-1}|\boldsymbol{O}^{t-1}) = \mathcal{N}(\boldsymbol{\mu}_{t-1}|\hat{\boldsymbol{\mu}}_{t-1}, \hat{Q}_{t-1}). \quad (9)$$

Thus, by substituting Eqs. (6), (8), and (9) into Eq. (5), we can derive the following analytic solution

$$p(\boldsymbol{\mu}_t|\boldsymbol{O}^t) = \mathcal{N}(\boldsymbol{\mu}_t|\hat{\boldsymbol{\mu}}_t, \hat{Q}_t), \quad (10)$$

where

$$\begin{cases} \hat{Q}_t &= ((U + A\hat{Q}_{t-1}A')^{-1} + \zeta_t\Sigma^{-1})^{-1} \\ \hat{K}_t &= \hat{Q}_t\zeta_t\Sigma^{-1} \\ \hat{\boldsymbol{\mu}}_t &= A\hat{\boldsymbol{\mu}}_{t-1} + \boldsymbol{\nu} + \hat{K}_t(\boldsymbol{\mathcal{M}}_t/\zeta_t - A\hat{\boldsymbol{\mu}}_{t-1} - \boldsymbol{\nu}) \end{cases}. \quad (11)$$

$\zeta_t$ is an occupation count and $\boldsymbol{\mathcal{M}}_t$ is first-order statistics, both of which are assigned to the targeted Gaussian. Here $'$ denotes the transpose operation of matrix. Thus, we can update a posterior distribution by updating distribution parameters ($\hat{Q}_t$ and $\hat{\boldsymbol{\mu}}_t$) based on Eq. (11).

The right hand side parameters in Eq. (11) are classified into three types, and are obtained as follows:

- Statistics $\zeta_t$ and $\boldsymbol{\mathcal{M}}_t$ are efficiently computed by the forward-backward or Viterbi algorithm.
- The affine transformation parameters $A$ and $\boldsymbol{\nu}$ can be estimated by the Maximum Likelihood (ML) approach using $\boldsymbol{O}_t$ [4] [1].
- Covariance matrix parameter $U$ of the system noise is assumed to be proportional to that of output distribution $\Sigma$ as $U = (u^0)^{-1}\Sigma$, where $u^0$ is a tuning parameter.

Thus, we can actually implement this incremental adaptation with a single tuning parameter ($u^0$). This corresponds to a solution of a linear dynamical system, where the observable equation corresponds to the output distribution (Eq. (6)) and the state equation corresponds to the discrete stochastic process (Eq. (7)).

[1] Although we can also obtain the transformation parameters using a Bayesian estimation approach or incorporate an incremental adaptation scheme into the transformation parameter estimation [3], we adopt the simple ML estimate form of the transformation parameters to avoid the complicated formulation in this paper.

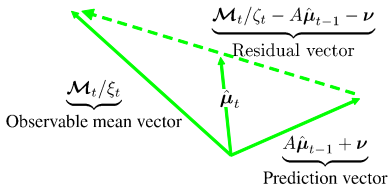


**Fig. 1**. Mean vector parameter $\hat{\boldsymbol{\mu}}_t$ is updated by the prediction and residual vectors.

### 2.4. Macroscopic time evolution system

Eq. (11) is regarded as a general solution of the Kalman filter algorithm in a linear dynamical system. Therefore, according to the standard Kalman filter interpretation, we discuss the meaning of the $\hat{\boldsymbol{\mu}}_t$ update equation in Eq. (11) by rewriting it as follows:

$$\hat{\boldsymbol{\mu}}_t = \underbrace{A\hat{\boldsymbol{\mu}}_{t-1} + \boldsymbol{\nu}}_{\text{prediction}} + \hat{K}_t \underbrace{(\boldsymbol{\mathcal{M}}_t/\zeta_t - A\hat{\boldsymbol{\mu}}_{t-1} - \boldsymbol{\nu})}_{\text{residual}}. \quad (12)$$

From Eq. (12), a state variable ($\hat{\boldsymbol{\mu}}_{t-1}$) is predicted by an affine transformation (parameterized by $A$ and $\boldsymbol{\nu}$). However, the predicted value ($A\hat{\boldsymbol{\mu}}_{t-1} + \boldsymbol{\nu}$) often contains errors resulting from an incorrect estimation, which causes an unstable incremental adaptation due to error propagation. To avoid such unstable estimations, the predicted value is compensated by a residual term, which is obtained as the observable mean vector ($\boldsymbol{\mathcal{M}}_t/\zeta_t$) minus the predicted value. Kalman gain $\hat{K}_t$ controls the degree of this compensation (as shown in Fig. 1). Thus, our approach realizes a stable incremental adaptation, which explicitly includes this prediction (transformation) and error compensation mechanism via the discrete stochastic process, unlike the conventional Bayesian approaches [1–3].

Our solution is represented by macroscopic values, which is unlike the standard Kalman filter solution. For example, the standard Kalman filter is updated frame by frame ($\boldsymbol{o}_{n-1} \rightarrow \boldsymbol{o}_n$) while our solution is updated by a partial time series ($\boldsymbol{O}_{t-1} \rightarrow \boldsymbol{O}_t$). In addition, distribution parameters $\hat{Q}_t$, $\hat{K}_t$, and $\hat{\boldsymbol{\mu}}_t$ in our solution are represented by the statistics $\zeta_t$ and $\boldsymbol{\mathcal{M}}_t$ obtained from $\boldsymbol{O}_t$, while those of the standard Kalman filter are represented by a single frame feature vector $\boldsymbol{o}_n$. Thus, we call our approach a macroscopic time evolution system.

Finally, we comment on parameter $u^0$, which is introduced at the system noise setting in Section 2.3. This parameter plays an important role in connecting the affine transformation and ML approaches as follows:

- Large $u^0$ case ($u^0 \rightarrow \infty$): $\hat{\boldsymbol{\mu}}_t \rightarrow A\hat{\boldsymbol{\mu}}_{t-1} + \boldsymbol{\nu}$ is a mean vector parameter predicted by an affine transformation approach.
- Small $u^0$ case ($u^0 \rightarrow 0$): $\hat{\boldsymbol{\mu}}_t \rightarrow \boldsymbol{\mathcal{M}}_t/\zeta_t$ is an ML estimate obtained by using $\boldsymbol{O}_t$.

Thus, $\hat{\boldsymbol{\mu}}_t$ is represented as an interpolation between the predicted and ML values, and the degree is controlled by $u^0$. This is similar to the methods that combine the transformation and Maximum A Posteriori (MAP) adaptation approaches serially by first estimating the transformation parameters, and then using the MAP adaptation for the transformed model parameters (MLLR-MAP and bias-MAP) [5, 6].

**Table 1**. Experimental conditions for speaker adaptation

| | |
|---|---|
| Sampling rate/quantization | 16 kHz / 16 bit |
| Feature vector | 12 order MFCC with energy |
| (39 dimensions) | +Δ+ΔΔ |
| Window | Hamming |
| Frame size/shift | 25/10 ms |
| Number of temporal HMM states | 3 (Left to right) |
| Number of phoneme categories | 43 |
| Number of context-dependent HMM states | 2,000 |
| Number of mixture components | 16 |
| Initial training data | ASJ read sentences, 10.2 hours (44 males) † |
| Adaptation data | CSJ lectures, 128 × 20 utterances (20 males) ‡ |
| Test data | CSJ lectures, 64,341 words (20 males) ‡ |
| Language model | Standard trigram (made by CSJ transcription) |
| Vocabulary size | 30, 000 |
| Perplexity (OOV rate) | 82.2 (2.1 % ) |

† ASJ (Acoustical Society of Japan) database
‡ CSJ (Corpus of Spontaneous Japanese) database

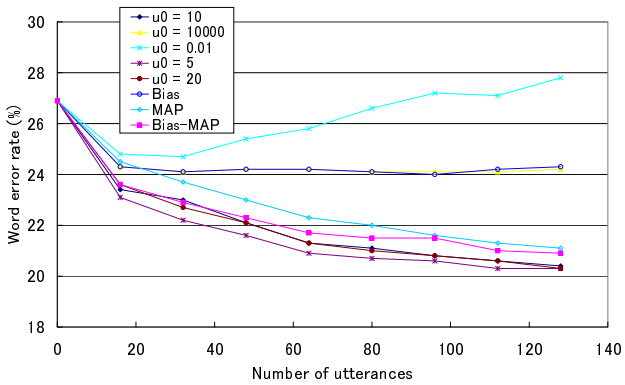


**Fig. 2**. Comparison of the proposed methods with varying $u^0$, conventional bias, MAP, and bias-MAP adaptation within incremental adaptation experiments.

## 3. EXPERIMENTS

We conducted supervised speaker adaptation experiments to examine the basic performance of the proposed incremental adaptation method in terms of (i) stability of adaptation process, (ii) dependence on parameter $u^0$, and (iii) comparison with batch adaptation. Table 1 summarizes the experimental conditions. The initial (prior) acoustic model was constructed from read sentences and we adapted this model using lectures given by 20 males and their transcriptions. All the males delivered more than two lectures, and the latest recorded lectures were used for recognition tests, and the other lectures were used for the adaptation. In this experimental setup, the mismatch between the initial and target conditions is caused not only by the speakers, but also by the difference in speaking styles between read speech and a lecture. The purpose of the adaptation was to eliminate the mismatches stably in an incremental fashion.

The first experiment was aimed at confirming the performance stability for the test data, which were commonly used for examining the performance of adapted model after each incremental step. Each incremental adaptation step used 16 utterances, and eight steps in total were undertaken for each speaker. Figure 2 compares the proposed approach, which involves using different $u^0$ settings, with the MAP and transformation-based adaptations. The transformation parameters were estimated independently of incremental steps, i.e.,

the batch estimation was operated for the transformation parameters in each incremental step. The transformation parameters were shared among several Gaussians based on the familiar Gaussian tree clustering technique [4]. In this experiment for both proposed and conventional approaches, the transformation was the biasing ($A = I$ in Eq. (7)) because it was not feasible to estimate the matrix $A$ properly using only 16 utterances. The MAP-, bias-, and bias-MAP-estimated parameters at each incremental step were used as initial parameters in the next step. We can see that the proposal of the setting of $u^0$ at around 10 and the MAP and bias-MAP performed much more stably than the bias. This is because the proposal, MAP, and bias-MAP are based on the posterior distribution estimation. In addition, the proposal ($u^0 = 5, 10, 20$) performed about 1 % better than MAP, and 0.5 % better than bias-MAP, which serially combining both the MAP and bias adaptation. These results suggest that the proposal fully demonstrated the effect of the prediction and error-compensation mechanism derived from a discrete stochastic process by utilizing both the characteristics of MAP and bias adaptation, as discussed in Section 2.4.

Next, we examined the performance when $u^0$ had extremely small ($0.01$) and extremely large ($10000$) values, which become asymptotically equivalent, as discussed in Section 2.4, to the ML estimation of model parameters and bias estimation, respectively. Because of an insufficient amount of data (16 utterances) for the ML estimation, the setting $u^0 = 0.001$ caused overtraining and increased the errors somewhat as the adaptation proceeded. Although the setting $u^0 = 1000$ did not degrade the performance, the behavior was almost the same as the bias adaptation. By setting $u^0$ around 10, the proposed approach is assumed to make full use of its advantages of the prediction and error-compensation mechanism. These results are consistent with the discussion in Section 2.4. From the above results for various $u^0$ values, we can state that the recognition performance is not very sensitive to the $u^0$ value unless we use an extreme value ($0.01$ or $10000$). This suggests that the proposed adaptation works based on a heavy dependence on $u^0$, and that we do not need to be very careful in choice of $u^0$ value if we choose a value of around 10.

Table 2 shows the results of recognition tests using acoustic models obtained after the final (eighth) step of the proposed incremental adaptation ($u^0 = 10$). The table also includes results for three types of batch adaptation that used all the utterances for the eight steps at one time. When we applied the estimation calculus employed in the proposed incremental method to the batch adaptation, the performance (WER = 19.8 %) was slightly better than that of the incremental method (WER = 20.4 %). However, at the end of the final step, the proposed incremental method performed better than the conventional bias method (WER = 22.1 %), and comparably to the conventional bias-MAP method (WER = 20.2 %). Incremental adaptation, in general, is less advantageous than batch adaptation when both use the same adaptation data since the incremental approach does not guarantee to give optimal estimates for all the adaptation data. The above experimental results prove that the prediction and error-compensation mechanism in the proposed method worked properly and resulted in a quasi-optimum model which, at least, performs comparably to the batch-adapted model. Another experiment with the same purpose was also conducted that employed MLLR instead of bias as the transformation. Since

**Table 2**. Comparison of incremental and batch adaptations by employing bias adaptation.

| | Incremental | Batch | | |
|---|---|---|---|---|
| | Proposal | Proposal | Bias | Bias-MAP |
| WER | 20.4 % | 19.8 % | 22.1 % | 20.2 % |

**Table 3**. Comparison of incremental and batch adaptations by employing MLLR adaptation.

| | Incremental | Batch | | |
|---|---|---|---|---|
| | Proposal | Proposal | MLLR | MLLR-MAP |
| WER | 20.2 % | 19.6 % | 21.5 % | 19.9 % |

16 utterances were insufficient for the MLLR estimation at each incremental step, we made each step use 32 utterances, and reduced the total number of steps to four. The result in Table 3 indicates a very similar tendency to the bias adaptation case, i.e., the proposed incremental adaptation could produce a model that performed comparably to the batch-adapted model.

Thus, we found that the proposed approach provided a stable incremental adaptation process, and the adapted model performed better than those obtained by conventional incremental approaches, and even performed comparably to the batch-adapted model.

## 4. SUMMARY

We proposed a new incremental adaptation method based on a macroscopic time evolution system. Supervised speaker adaptation experiments reveal the stable performance of the proposed method obtained by utilizing the advantages of the prediction and error-compensation mechanism based on the Kalman filter algorithm. We will apply the proposed method to more realistic incremental adaptation tasks such as online and unsupervised tasks [2, 7].

## 5. REFERENCES


[1] G. Zavaliagkos, R. Schwartz, and J. Makhoul, "Batch, incremental and instantaneous adaptation techniques for speech recognition," in *Proc. ICASSP1995*, 1995, vol. 1, pp. 676–679.

[2] 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 SAP*, vol. 5, pp. 161–172, 1997.

[3] K. Yu and M. J. F. Gales, "Incremental adaptation using Bayesian inference," in *Proc. ICASSP 2006*, 2006, vol. 1, pp. 217–220.

[4] 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.

[5] V. Digalakis, D. Ritischev, and L. Neumeyer, "Speaker adaptation using constrained reestimation of Gaussian mixtures," *IEEE Trans. on SAP*, vol. 3, pp. 357–366, 1995.

[6] J. Takahashi and S. Sagayama, "Vector-field-smoothed Bayesian learning for fast and incremental speaker/telephone-channel adaptation," *Computer Speech and Language*, vol. 11, pp. 127–146, 1997.

[7] V. Digalakis, "Online adaptation of hidden Markov models using incremental estimation algorithms," *IEEE Trans. on SAP*, vol. 7, pp. 253–261, 1999.