ON-LINE ADAPTATION AND BAYESIAN DETECTION OF ENVIRONMENTAL CHANGES BASED ON A MACROSCOPIC TIME EVOLUTION SYSTEM

Shinji Watanabe and Atsushi Nakamura

NTT Communication Science Laboratories, NTT Corporation, Japan

ABSTRACT

Acoustic characteristics are often changed over time as a result of various factors including changes of speakers, speaking styles, and noise sources. Incremental adaptation techniques for speech recognition are aimed at adjusting acoustic models quickly and stably to such time-variant acoustic characteristics. Recently we proposed a novel incremental adaptation framework based on a macroscopic time evolution system, which models the time-variant characteristics by successively updating posterior distributions of acoustic model parameters. This paper proposes fast incremental adaptation based on a macroscopic time evolution system that realizes an utterance-by-utterance update by approximating the posterior distributions. This adaptation was used to perform on-line adaptation of Japanese broadcast news for very large vocabulary continuous speech recognition (700k vocabulary size) in real time. The word accuracy was improved from 73.9% to 85.1%. In addition, by incorporating a Bayesian model selection approach, we realized the simultaneous on-line adaptation and detection of environmental changes.

Index Terms— Speech recognition, on-line adaptation, acoustic model, macroscopic time evolution system, model selection

1. INTRODUCTION

Speech information has a hierarchical structure on various time scales, as shown in Figure 1. In conventional speech recognition, the dynamics of speech on each time scale are processed or modeled by an appropriate technique (e.g. speech in a frame unit is processed by signal processing and feature extraction. Hidden Markov Model (HMM) state and phoneme units are represented by acoustic models. Transitions from a phoneme unit to a word unit are represented by lexical models. Word and word sequence units are modeled by n-gram language models). However, speech includes information on a larger scale than the conventional time scale (at most n-gram) used in speech recognition. For example, changes of topics, speakers, noise environments, and speaking rhythms often occur on an utterance time scale in usual conversations, and this information should be also modeled by incorporating conventional speech recognition modeling.

We used the above picture of the speech information hierarchy as a basis for our work on modeling acoustic characteristics, which are often changed as a result of various factors over a long-time scale. In real environments, these temporal variations often cause time-variant mismatches between the acoustic characteristics of training data and unseen speech input. Incremental adaptation techniques are aimed to compensate for such mismatches. In [1], we propose a novel incremental adaptation framework based on a macroscopic time evolution system. Here, the term "macroscopic time" scale means that the posterior distributions are updated at rather long intervals. The proposed incremental update involves a predictor-corrector algorithm in accordance with the Kalman filter theory. We also provide a unified interpretation of the proposal and the two main conventional approaches

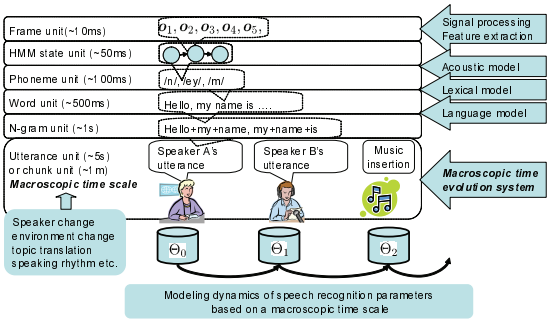

Fig. 1. Speech information hierarchy.

of indirect adaptation via transformation parameters (e.g. Maximum Likelihood Linear Regression (MLLR) [2]) and direct adaptation of classifier parameters (e.g. Maximum A Posteriori (MAP) [3]) in [4].

A part of our previous implementation often adopts the MLLR algorithm as a discrete stochastic process. Generally, MLLR requires more than about 10 utterances ($\approx$ 1 minute) to sufficiently estimate the transformation parameters. In addition, the posterior update equation incurs high computational costs that result from matrix calculations. These properties are not a serious disadvantage as regards off-line (batch) speech recognition tasks. However, on-line speech recognition tasks require quick recognition results when users finish speaking. Moreover, conversations in real environments often include large changes utterance by utterance. Therefore, to track such environmental changes appropriately and achieve a quick response, this paper proposes fast incremental adaptation based on a macroscopic time evolution system with an utterance unit update ($\approx$ a few seconds). The proposed method approximates the matrix representation to the scalar and vector representations by restricting the number of free parameters in the posterior distributions and by adopting bias adaptation [5, 6] instead of MLLR. This realizes the fast, quick response, and low memory implementation of the macroscopic time evolution system while maintaining high recognition performance.

This adaptation is used to recognize Japanese broadcast news, which includes various environmental changes (speaker changes, music insertions, and hookups). Namely, the proposed method is used to perform on-line adaptation for very large vocabulary continuous speech recognition (700k vocabulary size) in real time. Note that our approach temporally varies a single model to track any environmental change while approaches using a multiple speaker model basically try to track specific speakers among reference speakers (or clusters of speakers) [7, 8]. Note also that most adaptation approaches are based on (feature-space) MLLR and a sufficient amount of statistics is accumulated from the previously estimated statistics. However, the proposed adaptation based on a macroscopic time evolution system does not have such a time dependency as regards the statistics accumulation, and the time dependency is simply modeled in the Kalman filter part of the macroscopic time evolution sys-

tem. Finally, by incorporating a Bayesian model selection approach, we realized the simultaneous on-line adaptation and detection of environmental changes.

2. MACROSCOPIC TIME EVOLUTION SYSTEM

This section briefly introduces an incremental adaptation framework based on a macroscopic time evolution system [1]. In the macroscopic time evolution system, we assume that acoustic features are changed based on an utterance or a chunk (several utterances) unit. Then, the accumulated set of feature vectors ($\boldsymbol{o}^t$), which is a frame-based sequence, can be regarded as an utterance-based (or chunk-based) sequence:

$$\boldsymbol{o}^t = \{\underbrace{\boldsymbol{o}_{n=1}, ..., \boldsymbol{o}_{n_1}}_{\boldsymbol{o}_{t=1}}, \underbrace{\boldsymbol{o}_{n_1+1}, ..., \boldsymbol{o}_{n_1+n_2}}_{\boldsymbol{o}_{t=2}}, ..., \underbrace{\boldsymbol{o}_{n_{t-1}+1}, ..., \boldsymbol{o}_{n_{t-1}+n_t}}_{\boldsymbol{o}_t}\}.$$

Here, $\boldsymbol{o}_n \in R^D$ denotes a D dimensional feature vector at frame n (ex. 10 ms), while $\boldsymbol{o}_t$ denotes a set of feature vectors at t. Then, posterior distributions of acoustic model parameters, such as the mean vectors ($\boldsymbol{\mu}$) of Gaussians in continuous density HMMs, are incrementally updated on this macroscopic time scale (t). Here, we target an arbitrary Gaussian mean vector parameter in an acoustic model, and omit the Gaussian index from the parameter. By using the Markov assumption and probabilistic formulae, we analytically derive a time evolution equation from $p(\boldsymbol{\mu}_t|\boldsymbol{o}^t)$ to $p(\boldsymbol{\mu}_{t+1}|\boldsymbol{o}^{t+1})$ [1] as:

$$p(\boldsymbol{\mu}_{t+1}|\boldsymbol{o}^{t+1}) \propto \underbrace{p(\boldsymbol{o}_{t+1}|\boldsymbol{\mu}_{t+1})}_{\text{(A)}} \int \underbrace{p(\boldsymbol{\mu}_{t+1}|\boldsymbol{\mu}_t)}_{\text{(B)}} \underbrace{p(\boldsymbol{\mu}_t|\boldsymbol{o}^t)}_{\text{(C)}} d\boldsymbol{\mu}_t, \quad (1)$$

where $\propto$ denotes a proportional relation. The right hand side of time evolution equation (1) consists of three distributions. In [1], we provide concrete Gaussian forms with these three distributions.

(A) Output distribution (Auxiliary function) $\Leftarrow$ Continuous density HMM:

$$p(\boldsymbol{o}_{t+1}|\boldsymbol{\mu}_{t+1}) = \prod\nolimits_{n|\boldsymbol{o}_n \in \boldsymbol{o}_{t+1}} \left(\mathcal{N}(\boldsymbol{o}_n|\boldsymbol{\mu}_{t+1}, \Sigma)\right)^{\zeta_n}, \quad (2)$$

where Σ is the covariance matrix of a targeted Gaussian, and ζ_n is the occupation probability assigned to the targeted Gaussian at frame n, which is obtained by the E-step of the EM algorithm.

(B) Discrete stochastic process $\Leftarrow$ Linear dynamical system:

$$p(\boldsymbol{\mu}_{t+1}|\boldsymbol{\mu}_t) = \mathcal{N}(\boldsymbol{\mu}_{t+1}|A_{t+1}\boldsymbol{\mu}_t + \boldsymbol{b}_{t+1}, U), \quad (3)$$

where A_{t+1} and $\boldsymbol{b}_{t+1}$ are affine transformation parameters, which are shared by several Gaussians, and can be estimated with the standard MLLR algorithm by using $\boldsymbol{o}_{t+1}$ [2]. U is the covariance matrix of the system noise.

(C) Succeeding posterior distribution $\Leftarrow$ Conjugate distribution:

$$p(\boldsymbol{\mu}_t|\boldsymbol{o}^t) = \mathcal{N}(\boldsymbol{\mu}_t|\hat{\boldsymbol{\mu}}_t, \hat{Q}_t), \quad (4)$$

where we adopt a conjugate distribution of $\boldsymbol{\mu}_t$ [3], i.e. $\boldsymbol{\mu}_t$ is distributed by a Gaussian of $\hat{\boldsymbol{\mu}}_t$ and $\hat{Q}_t$.

Then, the succeeding posterior distribution can be derived analytically by substituting the above three Gaussians (Eqs. (2), (3), and (4)) into Eq. (1). The resulting posterior also becomes a Gaussian distribution:

$$p(\boldsymbol{\mu}_{t+1}|\boldsymbol{o}^{t+1}) = \mathcal{N}(\boldsymbol{\mu}_{t+1}|\hat{\boldsymbol{\mu}}_{t+1}, \hat{Q}_{t+1}),$$

where

$$\begin{aligned} \hat{Q}_{t+1} &\triangleq ((U + A_{t+1}\hat{Q}_t A'_{t+1})^{-1} + \zeta_{t+1}(\Sigma)^{-1})^{-1} \\ \hat{\boldsymbol{\mu}}_{t+1} &\triangleq A_{t+1}\hat{\boldsymbol{\mu}}_t + \boldsymbol{b}_{t+1} \\ &\quad + \hat{Q}_{t+1}\zeta_{t+1}(\Sigma)^{-1}\left(\frac{\boldsymbol{m}_{t+1}}{\zeta_{t+1}} - A_{t+1}\hat{\boldsymbol{\mu}}_t - \boldsymbol{b}_{t+1}\right) \end{aligned}. \quad (5)$$

Here $'$ denotes the transpose operation of the matrix. ζ_{t+1} is the accumulated occupation count and $\boldsymbol{m}_{t+1}$ is the accumulated first-order statistics, both of which are assigned to a targeted Gaussian in $t+1$, i.e., $\zeta_{t+1} \triangleq \sum_{n|\boldsymbol{o}_n \in \boldsymbol{o}_{t+1}} \zeta_n$ and $\boldsymbol{m}_{t+1} \triangleq \sum_{n|\boldsymbol{o}_n \in \boldsymbol{o}_{t+1}} \zeta_n \boldsymbol{o}_n$. Thus, we can update the posterior distribution given the succeeding speech utterance (or chunk) $\boldsymbol{o}_{t+1}$.

In [1], we provide the solution (Eq. (5)) with the Kalman filter interpretation. The proposed algorithm of our incremental update involves a prediction and correction step in accordance with the Kalman filter theory, and this achieves a *balanced adaptation between quickness and stability*. In [4], we also investigate the proposed framework in terms of the adaptation techniques in speech recognition. We provide a unified interpretation of adaptation techniques, which involves the conventional MLLR, MAP [2, 3], and their combination approaches, based on the macroscopic time evolution system.

However, to perform update Eq. (5) quickly with a small amount of data (a few seconds), there are certain problems that originate from the matrix computation:

- Eq. (5) requires matrix operations (add, product, inverse) for all Gaussians. The computation cost of these operations is very high.
- The inverse operation of the matrix makes the computation unstable for anomalous data, which rarely occurs when the amount of data is very small.
- In addition to the usual acoustic model parameters, we require the distribution parameter $\hat{Q}_t$ for all Gaussians. $\hat{Q}_t$ is a symmetric full matrix and requires a large memory (e.g. 10 times more than the usual acoustic model).

In addition to the above problems caused by the matrix representation, we require a sufficient amount of adaptation data (about 10 utterances) to estimate A_{t+1} and $\boldsymbol{b}_{t+1}$ used in Eq. (5) [2]. If we estimate A_{t+1} and $\boldsymbol{b}_{t+1}$ with one utterance, over-training problems occur. The next section considers these problems to realize the one utterance update, and hereafter we use t as an utterance index.

3. FAST INCREMENTAL ADAPTATION BY USING BIAS TRANSFORMATION AND VARIANCE SCALING FACTOR

First, to realize fast incremental adaptation based on a macroscopic time evolution system, we adopt bias adaptation approaches, which regard A_{t+1} as the unit matrix [5, 6], i.e., Eq. (3) is represented as:

$$p(\boldsymbol{\mu}_{t+1}|\boldsymbol{\mu}_t) = \mathcal{N}(\boldsymbol{\mu}_{t+1}|\boldsymbol{\mu}_t + \boldsymbol{b}_{t+1}, (u_0)^{-1}\Sigma). \quad (6)$$

where U is assumed to be proportional to Σ as $U \triangleq (u_0)^{-1}\Sigma$. The number of free parameters is reduced from $D \times (D+1)$ to D. Therefore, bias adaptation does not require as much adaptation data as MLLR does, which leads to a quick response in on-line adaptation. We also restrict the number of free parameters in the posterior distribution $p(\boldsymbol{\mu}_t|\boldsymbol{o}^t)$ by introducing the inverse of scaling factor $\hat{r}_t$, which is proportional to the original covariance matrix Σ in the targeted Gaussian of an acoustic model. Then, Eq. (4) is represented as:

$$p(\boldsymbol{\mu}_t|\boldsymbol{o}^t) = \mathcal{N}(\boldsymbol{\mu}_t|\hat{\boldsymbol{\mu}}_t, (\hat{r}_t)^{-1}\Sigma). \quad (7)$$

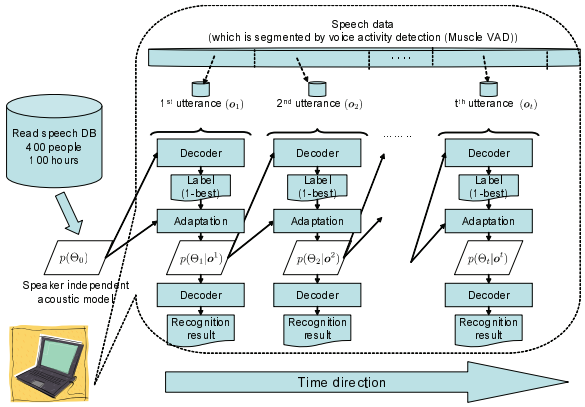

Fig. 2. Speech recognition with on-line adaptation based on the proposed macroscopic time evolution system.

These constraints are summarized as follows:

$$\begin{cases} A_{t+1} & = I \\ U & = (u_0)^{-1}\Sigma \\ \hat{Q}_t & = (\hat{r}_t)^{-1}\Sigma \end{cases}. \tag{8}$$

By substituting Eq. (8) into update Eq. (5), the update equation is represented as follows:

$$\begin{cases} \hat{r}_{t+1} & = ((u_0)^{-1} + (\hat{r}_t)^{-1})^{-1} + \zeta_{t+1} \\ \hat{\boldsymbol{\mu}}_{t+1} & = \hat{\boldsymbol{\mu}}_t + \boldsymbol{b}_{t+1} + \frac{\boldsymbol{m}_{t+1} - \zeta_{t+1}(\hat{\boldsymbol{\mu}}_t + \boldsymbol{b}_{t+1})}{\hat{r}_{t+1}} \end{cases}, \tag{9}$$

and the resulting posterior distribution is

$$p(\boldsymbol{\mu}_{t+1}|\boldsymbol{o}^{t+1}) = \mathcal{N}(\boldsymbol{\mu}_{t+1}|\hat{\boldsymbol{\mu}}_{t+1}, (\hat{r}_{t+1})^{-1}\Sigma). \tag{10}$$

By comparison with Eq. (5), Eq. (9) does not include any matrix representation (Note that we use capital letters as matrix variables in this paper, and these have been eliminated from Eq. (9)). Therefore, the computational problems that originate from the matrix representation, as discussed in Section 2, are also avoided. In addition, we can reduce the memory size by memorizing the scalar value $\hat{r}_{t+1}$ instead of the matrix value $\hat{Q}_{t+1}$. Consequently, the macroscopic time evolution system using Eq. (9) achieves fast, stable, and low memory properties in addition to a quick response by adopting the bias adaptation approach.

Thus, a macroscopic time evolution system is realized even when the amount of adaptation data in the update is small (one utterance). Section 4 applies this utterance-by-utterance macroscopic time evolution system to on-line acoustic model adaptation in broadcast news. In addition, section 5 uses it to detect environmental changes by incorporating a Bayesian model selection approach.

4. ON-LINE ADAPTATION BASED ON A MACROSCOPIC TIME EVOLUTION SYSTEM

This section explains an implementation and experimental results of speech recognition with on-line adaptation based on the proposed macroscopic time evolution system.

4.1. Implementation

Figure 2 shows an implementation scheme for on-line adaptation. A speaker-independent read speech model was adapted to three sets of Japanese broadcast news (the total size of the test sets is 4,235 words). The broadcast news was provided by NTT Cyber Space Laboratories. The read speech acoustic model was constructed by using 100 hours of Japanese speech spoken by 400 people. The acoustic and language model conditions are shown in Table 1. Note that we undertook all our experiments with very large vocabulary continuous speech recognition (700k vocabulary size). From speech data, an utterance at $t+1$ was extracted by voice activity detection based on a switching Kalman filter and speech periodic to aperiodic component ratio (Muscle VAD) [9], then, 1) transcribing the utterance by automatic speech recognition using the previously obtained model Θ_t, 2) applying the adaptation (Eq. (9)) by using the transcription (1-best hypothesis), and 3) again recognizing the utterance using the adapted model Θ_{t+1}. In this algorithm, step 2) was started when the VAD detected an endpoint. Therefore, each output of the recognition results was delayed by the time of steps 2) and 3) (less than half of the utterance time in this experiment). The bias ($\boldsymbol{b}$) vectors were estimated independently for each utterance. $\boldsymbol{b}$ was shared among several Gaussians by using a Gaussian tree structure technique to avoid over-training [5], where we set the occupancy threshold at 60 empirically. In MLLR (A and $\boldsymbol{b}$), we set the occupancy threshold at 5,000. We also set the system noise parameter u_0 at $5,000$ in Eq. (9). All of these on-line adaptation and recognition procedures were implemented on a laptop computer (Lenovo ThinkPad x60, Intel Core 2Duo Processor T7200(1.66GHz), 3GB memory size).

4.2. Experimental results

Table 2 summarizes the recognition results. The baseline word accuracy was 73.9%. The proposed approach (On-line MACROS) improves the word accuracy to 85.1%, which is 0.3 % better than the straightforward implementation of on-line bias adaptation (On-line bias). On-line bias successively updated the parameters by using the previously estimated parameters as initial parameters in estimation. Therefore, the superiority of On-line MACROS is owing to the update based on the macroscopic time evolution system (Eq. (9)). The proposed method also significantly reduced the memory size from 152 MB (Eq. (5)) to 0.19 MB (Eq. (9)) for the posterior distribution parameters by comparison with the previous implementation [1] and achieved real time decoding ($\times$ 0.77). Thus, the proposed method achieved high performance on-line adaptation for very large vocabulary continuous speech recognition in real time. We also compare On-line MACROS with On-line MLLRs, which were implemented similarly to On-line bias although their incremental update sizes were 2 and 16 utterances respectively, not one utterance, to avoid overtraining. MACROS was better than On-line MLLRs (semi-batch). In addition, we also show that the on-line results were better than the off-line (batch) results (bias and MLLR) in this experiment. These suggest that this task includes so many temporal variations utterance by utterance that the batch and semi-batch adaptation could not mitigate the temporal mismatches.

Table 1. Acoustic and language model conditions

Sampling rate/quantization	16 kHz / 16 bit
Feature vector (39 dimensions)	12 order MFCC with energy +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	3,042
Number of mixture components	16
Language model	Trigram (with Kneser-Ney smoothing)
Corpus	Mainichi newspaper articles (14 years)
Vocabulary size	700,000

5. DETECTION OF ENVIRONMENTAL CHANGES BY USING BAYESIAN MODEL SELECTION

This section introduces a model selection function in the utterance-by-utterance macroscopic time evolution system. The model selection is derived from a Posterior distribution-based Bayesian Criterion (PBC) [10]. The model selection function is used to detect the acoustic environment changes.

5.1. Posterior distribution-based Bayesian Criterion (PBC)

Since we obtain posterior distributions of model parameters in Eq. (9) ($p(\boldsymbol{\mu}_{t+1}|\boldsymbol{o}^{t+1})$) in on-line adaptation, we can obtain a posterior-based model selection criterion based on a Bayesian approach [10]. The criterion is based on the logarithm of the posterior distribution of model complexity $p(m|\boldsymbol{o}^{t+1})$, which is proportional to the sum of the negative free energy $\mathcal{F}(\boldsymbol{o}_{t+1})$ and the logarithm of the prior distribution of model complexity $p(m)$, i.e.,

$$\log p(m|\boldsymbol{o}^{t+1}) \propto \mathcal{F}(\boldsymbol{o}^{t+1}) + \log p(m) \triangleq \mathcal{L}(\boldsymbol{o}^{t+1}).$$

The negative free energy $\mathcal{F}(\boldsymbol{o}_{t+1})$ is calculated from Eqs. (2), (6), (7), and (10), as follows:

$$\begin{aligned}
&\mathcal{F}(\boldsymbol{o}_{t+1}) \\
&= \sum_k \left\langle \log \frac{p(\boldsymbol{o}_{t+1}|\boldsymbol{\mu}^k_{t+1}) \int p(\boldsymbol{\mu}^k_{t+1}|\boldsymbol{\mu}^k_t) p(\boldsymbol{\mu}^k_t|\boldsymbol{o}^t) d\boldsymbol{\mu}^k_t}{p(\boldsymbol{\mu}^k_{t+1}|\boldsymbol{o}^{t+1})} \right\rangle_{p(\boldsymbol{\mu}^k_{t+1}|\boldsymbol{o}^{t+1})} \\
&= -\frac{1}{2} \sum_k \Bigg(\zeta^k_{t+1} \log |\Sigma^k| - \log \bar{r}^k_{t+1} + \log \hat{r}^k_{t+1} + \sum_{d=1}^{D} (\Sigma^k_{dd})^{-1} \\
&\quad \left(V^k_{t+1,dd} + \bar{r}^k_{t+1} (\hat{\boldsymbol{\mu}}^k_{t,d} + \boldsymbol{b}^k_{t+1,d})^2 - \hat{r}^k_{t+1} (\hat{\boldsymbol{\mu}}^k_{t+1,d})^2 \right) \Bigg),
\end{aligned}$$

where

$$\begin{cases} \bar{r}^k_{t+1} & \triangleq ((u_0)^{-1} + (\hat{r}^k_t)^{-1})^{-1} \\ V^k_{t+1} & \triangleq \sum_{n|\boldsymbol{o}_n \in \boldsymbol{o}_{t+1}} \zeta^k_n \boldsymbol{o}_n \boldsymbol{o}'_n \end{cases}.$$

k is a Gaussian index, which has been omitted in the previous sections for simplicity. $<>_p$ means the expectation with respect to p. V^k_{t+1} is the accumulated second-order statistics. We use the score of Bayesian Information Criterion (BIC) as values of $\log p(m)$. BIC is often used to segment speech [11] by modeling a fraction of feature vectors as Gaussians. Namely, scores of the BIC detection are used as prior information in calculating $p(m|\boldsymbol{o}^{t+1})$. In addition to the BIC information, this criterion includes information of acoustic model adaptation, which can provide more appropriate model selection than BIC.

5.2. On-line speech segmentation by using Bayesian model selection

Based on PBC, we can judge whether utterances t and $t+1$ are segmented or not by considering the difference between the corresponding objective functions ($\Delta\mathcal{L}(\boldsymbol{o}_t, \boldsymbol{o}_{t+1}) \triangleq \mathcal{L}(\boldsymbol{o}_{t+1}) + \mathcal{L}(\boldsymbol{o}_t) - \mathcal{L}(\boldsymbol{o}_t, \boldsymbol{o}_{t+1})$). Namely, if $\Delta\mathcal{L}(\boldsymbol{o}_t, \boldsymbol{o}_{t+1}) \geq 0$, we set a boundary between t and $t+1$ since the posterior probability of setting the boundary is larger than that of concatenating t and $t+1$. As a preliminary experiment, we also used the broadcast news for the evaluation task where there were 44 changes in acoustic environments resulting from speaker changes, music insertions, and hookups. We summarized the segmentation results in Table 3 by using BIC and the proposed PBC segmentation. PBC slightly improved BIC by improving the precision score, which shows the effectiveness of PBC. However, this result is preliminary, and we will further investigate the effectiveness of PBC by demonstrating a larger evaluation task.

Table 2. Word accuracies for on-line and off-line adaptation approaches

Baseline	73.9 %
On-line MACROS	85.1 % (1 utterance update)
On-line bias	84.8 % (1 utterance update)
On-line MLLR	82.6 % (2 utterance update)
On-line MLLR	84.5 % (16 utterance update)
Off-line bias	80.8 %
Off-line MLLR	81.5 %

Table 3. Segmentation results for broadcast news

	BIC	PBC
Precision	0.55	0.57
Recall	0.91	0.91
F-measure	0.68	0.70

6. SUMMARY

This paper described the realization of on-line adaptation for very large vocabulary continuous speech recognition (700k vocabulary size) in real time based on a macroscopic time evolution system. In addition, by incorporating a Bayesian model selection approach, we can simultaneously realize on-line adaptation and the detection of the environmental changes. Future work will compare our approach with other approaches, where feature-space MLLR and multiple speaker model approaches were adopted [7, 8], in detail experimentally. We will also consider providing incremental adaptation approaches of language models with the proposed approach.

7. REFERENCES

[1] S. Watanabe and A. Nakamura, "Incremental adaptation based on a macroscopic time evolution system," in *Proc. ICASSP 2007*, 2007, vol. 4, pp. 769–772.

[2] 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.

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

[4] S. Watanabe and A. Nakamura, "A unified interpretation of adaptation techniques based on a macroscopic time evolution system with indirect/direct approaches," in *ICASSP 2008*, 2008, pp. 4285–4288.

[5] K. Shinoda and T. Watanabe, "Speaker adaptation with autonomous control using tree structure," in *Proc. EUROSPEECH1995*, 1995, pp. 1143–1146.

[6] M. Rahim and B-H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. on SAP*, vol. 4, pp. 19–30, 1996.

[7] D. Liu, D. Kiecza, A. Srivastava, and F. Kubala, "Online speaker adaptation and tracking for real-time speech recognition," in *Proc. Interspeech '2005*, 2005, pp. 281–284.

[8] U. Remes, J. Pylkkonen, and M. Kurimo, "Segregation of speakers for speaker adaptation in TV news audio," in *Proc. ICASSP 2007*, 2007, vol. 4, pp. 481–484.

[9] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme," in *ICASSP 2008*, 2008, pp. 4441–4444.

[10] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Constructing shared-state hidden Markov models based on a Bayesian approach," in *Proc. ICSLP2002*, 2002, vol. 4, pp. 2669–2672.

[11] S. Chen and R. Gopinath, "Model selection in acoustic modeling," in *Proc. Eurospeech1999*, 1999, vol. 3, pp. 1087–1090.