A UNIFIED INTERPRETATION OF ADAPTATION APPROACHES BASED ON A MACROSCOPIC TIME EVOLUTION SYSTEM AND INDIRECT/DIRECT ADAPTATION APPROACHES

Shinji Watanabe and Atsushi Nakamura

NTT Communication Science Laboratories, NTT Corporation, Japan

ABSTRACT

Incremental adaptation techniques for speech recognition are aimed at adjusting acoustic models quickly and stably to time-variant acoustic characteristics due to temporal changes of speaker, speaking style, noise source, etc. 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. In this paper, we provide a unified interpretation of the proposal and the two major conventional approaches of indirect adaptation via transformation parameters (e.g. Maximum Likelihood Linear Regression (MLLR)) and direct adaptation of classifier parameters (e.g. Maximum A Posteriori (MAP)). We reveal analytically and experimentally that the proposed incremental adaptation involves both the conventional and their combinatorial approaches, and simultaneously possesses their quick and stable adaptation characteristics.

Index Terms— speech recognition, acoustic model, incremental adaptation, macroscopic time evolution, indirect/direct adaptation

1. INTRODUCTION

In real environments, there inevitably exist time-variant and time-invariant mismatches between the acoustic characteristics of training and unseen data that depend on the speaker, speaking style, and noise varieties and their temporal changes. Acoustic model adaptation techniques aim to compensate for such mismatches in a batch or incremental manner, and are roughly classified into two standard approaches, i.e., indirect and direct adaptation [1].

The indirect adaptation approach, as typified by Maximum Likelihood Linear Regression (MLLR) adaptation, does not estimate the target model directly, but estimates mapping or transformation from the initial to target models *indirectly* [2–4]. Model parameters are usually grouped into classes in advance, and we estimate a set of transformation parameters for each class, so that a reasonable amount of data is available for estimating each transformation. The direct adaptation approach, as typified by Maximum A Posteriori (MAP) adaptation, *directly* estimates individual parameters in the target model, taking account of both data and prior distributions [5–8]. An advantage of indirect adaptation over direct adaptation is the *quick* effect of adaptation for a small amount of data. This is because there are fewer free parameters to be estimated owing to the use of parameter classes where model parameters in the same class are commonly transformed. On the other hand, an advantage of direct adaptation over indirect adaptation is its *stable* property, where the performance of an adapted model steadily approaches that of a condition-dependent model based on the Bayesian theory. However, there are unavoidable estimation errors in both indirect and direct adaptation approaches for a refinement that only uses a small amount of data. Therefore, especially in the incremental adaptation case, the both approaches may also propagate the errors, and this affects the adaptation performance.

In [9], we focused on the influence of the propagation of the estimation errors appeared in incremental adaptation, and proposed a novel incremental adaptation framework based on a macroscopic time evolution system. In the proposed framework, the dynamics of acoustic model parameters are tracked by incremental update of posterior distributions. The use of the posterior distributions can well mitigate the error propagation. The proposed algorithm of our incremental update involves a prediction and correction step in accordance with the Kalman filter theory [10], and this achieves the quickness and stability in adaptation, which have been in trade-off relationship in the conventional indirect/direct approaches.

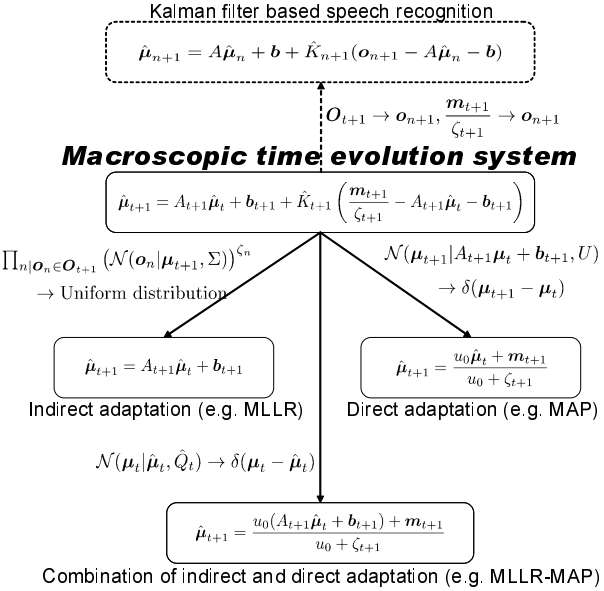


Fig. 1. Correlation diagram of each adaptation method.

In this paper, we further investigate the mechanism of the proposed adaptation algorithm to provide the quickness and stability, analytically and experimentally. Firstly, we provide a unified interpretation of adaptation techniques, which involves the conventional indirect/direct approaches, based on the macroscopic time evolution system. We also prove that the above interpretation can be extended so as to involve cascade combinations of indirect and direct approaches (Bias-MAP, MLLR-MAP [11, 12]). Figure 1 shows the relationship of the proposed and conventional adaptation methods based on the obtained interpretation. Finally, we verify the appropriateness of our interpretation by examining the quickness and stability of these adaptation methods in unsupervised incremental adaptation experiments.

2. MACROSCOPIC TIME EVOLUTION SYSTEM

This section briefly introduces an incremental adaptation framework based on a macroscopic time evolution system [9]. In the macroscopic time evolution system, we assume that acoustic features are changed based on a chunk unit, which consists of several utterances. Then, the accumulated feature vectors ($\boldsymbol{O}^t$), which is a frame-based

sequence, can be regarded as a chunk-based sequence:

$$\boldsymbol{O}^t = \{\underbrace{\boldsymbol{o}_1, ..., \boldsymbol{o}_{n_1}}_{\boldsymbol{O}_1}, \underbrace{\boldsymbol{o}_{n_1+1}, ..., \boldsymbol{o}_{n_1+n_2}}_{\boldsymbol{O}_2}, ..., \underbrace{\boldsymbol{o}_{n_{t-1}+1}, ..., \boldsymbol{o}_{n_{t-1}+n_t}}_{\boldsymbol{O}_t}\}.$$

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 chunk t (ex. 16 utterances). 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. 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 $p(\boldsymbol{\mu}_{t+1}|\boldsymbol{O}^{t+1})$ [9] as:

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

The right hand side of time evolution equation (1) consists of three distributions.

(A) $p(\boldsymbol{O}_{t+1}|\boldsymbol{\mu}_{t+1})$ is the *output distribution*.

(B) $p(\boldsymbol{\mu}_{t+1}|\boldsymbol{\mu}_t)$ is the *discrete stochastic process* of $\boldsymbol{\mu}_t$.

(C) $p(\boldsymbol{\mu}_t|\boldsymbol{O}^t)$ is the *current posterior distribution*, which is already estimated in the current adaptation step t.

In [9], we provide concrete Gaussian forms with these three distributions.

(A) Output distribution $\Leftarrow$ Continuous density HMM [1]

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

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

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

where A_{t+1} and $\boldsymbol{b}_{t+1}$ are affine transformation parameters, which are shared by several Gaussians, and can be estimated by the standard MLLR algorithm by using $\boldsymbol{O}_{t+1}$ [3, 4]. U is the covariance matrix of the system noise, and is assumed to be proportional to Σ as $U \triangleq (u_0)^{-1}\Sigma$, where u_0 is a tuning parameter.

(C) Current posterior distribution $\Leftarrow$ Conjugate distribution

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

where we adopt a conjugate distribution of $\boldsymbol{\mu}_t$ [5], 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 resultant 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

$$\begin{aligned}
\hat{Q}_{t+1} &\triangleq (((u_0)^{-1}\Sigma + A_{t+1}\hat{Q}_t A'_{t+1})^{-1} + \zeta_{t+1}(\Sigma)^{-1})^{-1} \\
\hat{\boldsymbol{\mu}}_{t+1} &\triangleq \underbrace{A_{t+1}\hat{\boldsymbol{\mu}}_t + \boldsymbol{b}_{t+1}}_{\text{Prediction}} + \underbrace{\hat{K}_{t+1}}_{\text{Kalman gain}} \underbrace{\left(\frac{\boldsymbol{m}_{t+1}}{\zeta_{t+1}} - A_{t+1}\hat{\boldsymbol{\mu}}_t - \boldsymbol{b}_{t+1}\right)}_{\text{Innovation}}.
\end{aligned} \tag{5}$$

Here $'$ denotes the transpose operation of the matrix. $\hat{K}_{t+1}$ is a Kalman gain defined as $\hat{K}_{t+1} \triangleq \hat{Q}_{t+1}\zeta_{t+1}(\Sigma)^{-1}$. 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 at chunk $t+1$, i.e.,

$$\begin{cases} \zeta_{t+1} & \triangleq \sum_{n|\boldsymbol{o}_n \in \boldsymbol{O}_{t+1}} \zeta_n \\ \boldsymbol{m}_{t+1} & \triangleq \sum_{n|\boldsymbol{o}_n \in \boldsymbol{O}_{t+1}} \zeta_n \boldsymbol{o}_n \end{cases}.$$

Thus, we can update the posterior distribution given the succeeding speech chunk $\boldsymbol{O}_{t+1}$.

In [9], we provide the solution (Eq. (5)) with the Kalman filter interpretation. The first two terms on the right hand side of $\hat{\boldsymbol{\mu}}_{t+1}$ in Eq. (5) are known as the "prediction term" with respect to the Kalman filtering, which has a quick adaptation property. However, the prediction often contains errors because the parameters are estimated using only a limited amount of data. The errors might propagate expansively in the process of successive updating, and this causes the incremental adaptation to lose stability. To avoid this problem, the prediction result is corrected with an innovator, which is obtained as the expectation vector of the observation ($\boldsymbol{m}_{t+1}/\zeta_{t+1}$) minus the prediction vector. The Kalman gain $\hat{K}_{t+1}$ controls the degree of correction. This scheme forms the core of "predictor-corrector algorithm," which is known as the most powerful advantage of Kalman filtering as regards incremental adaptation issues [10], which often employ a frame-by-frame type formulation. Since the proposed adaptation employs a chunk-by-chunk type formulation, and is driven by chunk statistics ζ_{t+1}, and $\boldsymbol{m}_{t+1}$, we call it incremental adaptation based on a macroscopic time evolution system. Thus, our framework includes the predictor-corrector algorithm explicitly, and therefore we can expect a quick and stable incremental adaptation for speech recognition.

3. UNIFIED INTERPRETATION OF PROPOSED AND CONVENTIONAL APPROACHES

This section theoretically discusses the relationships among proposed framework and conventional adaptation approaches, and provides a unified interpretation of them. In particular, we show, by considering the limit of three Gaussian distributions (Eqs. (2), (3), and (4)) in Eq. (1), that Eq. (5) can be simplified so as to be equivalent to each of the conventional approaches.

3.1. Connection to indirect/direct adaptation

An indirect adaptation approach can be derived by disregarding the effect from the output distribution in the proposal. Namely, we consider a limit of $p(\boldsymbol{O}_{t+1}|\boldsymbol{\mu}_{t+1})$ as its variance approaches infinity[2]:

$$\begin{aligned}
&p(\boldsymbol{O}_{t+1}|\boldsymbol{\mu}_{t+1}) : \\
&\prod\nolimits_{n|\boldsymbol{o}_n \in \boldsymbol{O}_{t+1}} \left(\mathcal{N}(\boldsymbol{o}_n|\boldsymbol{\mu}_{t+1}, \Sigma)\right)^{\zeta_n} \rightarrow \text{Uniform distribution.}
\end{aligned} \tag{6}$$

At this limit, the estimation of $\boldsymbol{\mu}$ becomes independent of the output distribution. By applying the replacement expressed in Eq. (6) to Eq. (2), and by solving Eq. (1), we obtain a simplified solution for $\hat{\boldsymbol{\mu}}_{t+1}$ instead of Eq. (5) as follows:

$$\hat{\boldsymbol{\mu}}_{t+1} \rightarrow A_{t+1}\hat{\boldsymbol{\mu}}_t + \boldsymbol{b}_{t+1}. \tag{7}$$

[1] We consider the auxiliary function form instead of the output distribution in continuous density HMMs owing to the existence of latent variables, and omit state transition and mixture weight parameters.

[2] To make the output distribution non-informative, we virtually assumed a multivariate uniform distribution on infinite ranges. The effect of the uniform distribution is absorbed into a normalization factor in the calculation.

This is equivalent to a type of indirect adaptation, the MLLR transformation of the mean vectors [3, 4].

Also, a direct adaptation approach can be derived by discarding the effect from the discrete stochastic process of model parameters in the proposal. We now consider a limit of $p(\boldsymbol{\mu}_{t+1}|\boldsymbol{\mu}_t)$ as its variance approaches zero:

$$p(\boldsymbol{\mu}_{t+1}|\boldsymbol{\mu}_t) : \mathcal{N}(\boldsymbol{\mu}_{t+1}|A_{t+1}\boldsymbol{\mu}_t + \boldsymbol{b}_{t+1}, U) \rightarrow \delta(\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t), \quad (8)$$

where $\delta(\boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t)$ is a Dirac δ function. At this limit, the stochastic process of model parameters does not work at all. By solving Eq. (1) with Eq. (8) instead of Eq. (3), we obtain another simplified the solution for $\hat{\boldsymbol{\mu}}_{t+1}$ as follows:

$$\hat{\boldsymbol{\mu}}_{t+1} \rightarrow \frac{u_0\hat{\boldsymbol{\mu}}_t + \boldsymbol{m}_{t+1}}{u_0 + \zeta_{t+1}}. \quad (9)$$

This is equivalent to a type of direct adaptation, the MAP adaptation of the mean vectors [5]. Note that the role of system noise parameter u_0 in the proposal is the same as that of the hyper-parameter used in the MAP adaptation, which controls the balance between the statistics from data and prior knowledge.

These two relationships (Eqs. (7) and (9)) prove the fact that the macroscopic time evolution system theoretically involves conventional indirect and direct adaptation approaches. Accordingly, it is reasonable that the proposal can possess both quickness and stability when it does not discard either effect from the output distribution or discrete stochastic process (Eq. (5)).

3.2. Connection to combinatorial adaptation

An alternative method that can involve indirect and direct adaptation approaches is a cascade type combination of the two approaches, i.e., first estimating the transformation parameters, and then applying the Bayesian adaptation for the transformed model parameters (Bias-MAP, MLLR-MAP) [11, 12]. Our discussion now moves on to the relationship between our framework and the combinatorial adaptation methods. We consider a limit of mean posterior distribution $p(\boldsymbol{\mu}_t|\boldsymbol{O}^t)$ as its variance approaches zero:

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

In other words, we here point-estimate $\boldsymbol{\mu}$. By applying the replacement expressed in Eq. (10) to Eq. (4), and by solving Eq. (1), we obtain

$$\hat{\boldsymbol{\mu}}_{t+1} \rightarrow \frac{u_0(A_{t+1}\hat{\boldsymbol{\mu}}_t + \boldsymbol{b}_{t+1}) + \boldsymbol{m}_{t+1}}{u_0 + \zeta_{t+1}}. \quad (11)$$

The solution thus becomes equivalent to the combinatorial methods [11, 12]. Note that the difference between our approach (Eq. (5)) and the combinatorial methods is whether or not covariance matrix $\hat{Q}_t$ is considered. Because of this effect of the covariance matrix, our approach is expected to be more robust than the combinatorial methods.

Throughout the above discussions, it has been proved that our framework provides a unified view of the conventional indirect, direct, and their combinatorial adaptation methods. It stands to reason that the advantages of conventional methods (quickness and stability) are inherited to and robustly enhanced in our framework. Figure 1 depicts a correlation of adaptation methods that have been discussed in this section. The Kalman filter based speech recognition, which is not in the context of this discussion, is also included in the figure, We have already discussed the relationship between this and our framework in [9].

4. EXPERIMENTS

We conducted a series of experiments for verifying whether the relationship that we discussed in the previous section was properly reflected in adaptation performance. Figure 2 shows design of unsupervised incremental adaptation experiments. Here, it was assumed

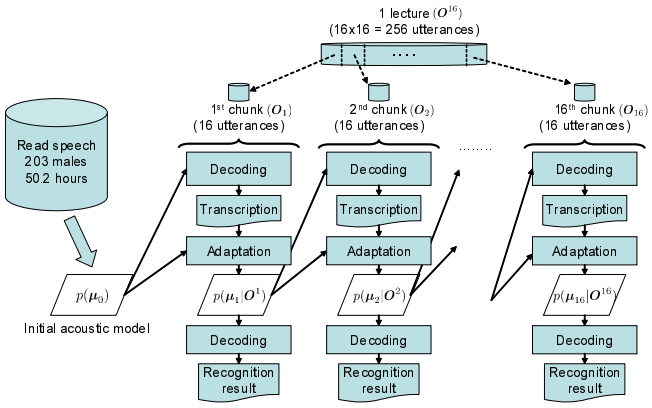


Fig. 2. Experimental flow of unsupervised incremental adaptation from read speech to lectures.

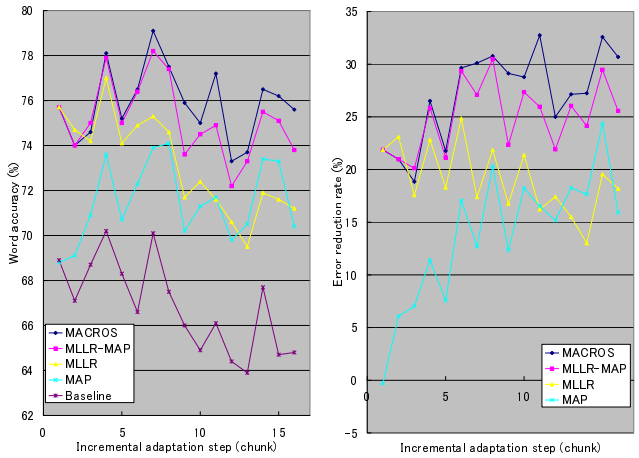
that a speaker-independent read speech model was adapted to lecture speech. We prepared 10 lectures from the Corpus of Spontaneous Japanese (CSJ [13]), which contained more than 256 utterances, and divided each lecture into 16 chunks, each of which consisted of 16 utterances. The incremental adaptation proceeded by the chunk as a basic unit. Then, the following three operations were performed for a set of utterances in each chunk 1) transcribing each utterance by automatic speech recognition using the previously obtained set of models, 2) applying adaptation to the previously obtained set of models by using the transcriptions, and 3) again recognizing the utterances using the adapted set of models. The adaptation performance (word accuracy and error reduction rate) to be hereinafter used were calculated by averaging the above results for the 10 lectures. Acoustic and language model conditions are shown in Table 1. In the indirect adaptation approach, we adopted the MLLR adaptation, where a shared MLLR parameter structure was obtained by using a common Gaussian tree construction, which was controlled by an occupancy threshold (= 5000) [3, 4]. Then, the obtained MLLR parameters (A_{t+1} and $\boldsymbol{b}_{t+1}$) and sufficient statistics (ζ_{t+1} and $\boldsymbol{m}_{t+1}$) in chunk $t+1$ were also used in the combinatorial approach (Eq. (11)) and proposals (Eq. (5)). In the direct adaptation approach, we adopted the MAP adaptation [5]. We set $u_0 = 10$ with reference to the result in [5, 9], which is used for the system noise parameter of the proposal (Eq. (5)) and the hyper-parameter of the MAP (Eq. (9)) and MLLR-MAP adaptation (Eq. (11)) [3].

Figure 3 compares adaptation based on a macroscopic time evolution system (MACROS) with indirect adaptation (MLLR), direct adaptation (MAP), and combinatorial adaptation (MLLR-MAP) in terms of adaptation performance. MLLR quickly improved the models even by adaptation that only used the first chunk, and MAP did not quickly but stably improved the performance, as incremental steps of adaptation proceeded. These characteristics are easily observable from the results of error reduction rates, which were calculated from the non-adapted (baseline) word error rates (Figure 3 (b)). On the other hand, MACROS and MLLR-MAP performed well for almost all chunks, differently from MLLR and MAP. Fig-

Table 1. Acoustic and language model conditions

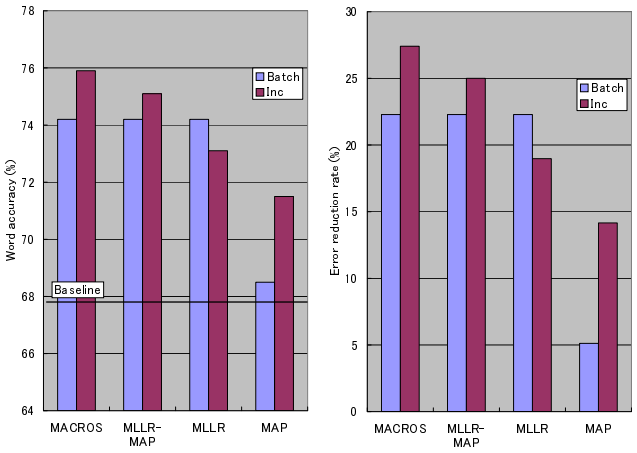
Sampling rate/quantization	16 kHz / 16 bit
Feature vector (39 dimensions)	12 order MFCC with energy $+\Delta+\Delta\Delta$
Window	Hamming
Frame size/shift	25/10 ms
Number of temporal HMM states	3 (left to right)
Number of phoneme categories	43
Number of context-dependent HMM states	2,000
Number of mixture components	16
Language model	Standard trigram (made by CSJ transcription)
Vocabulary size	30, 000
Perplexity (OOV rate)	82.2 (2.1 %)

[3] When MACROS and MLLR-MAP were performed in the 1st chunk and batch adaptation, we set u_0 to a large initial value (10,000)



(a) Word accuracy (b) Error reduction rate

Fig. 3. Comparison of macroscopic time evolution system (MACROS) with indirect adaptation (MLLR), direct adaptation (MAP), and combinatorial adaptation (MLLR-MAP). Error reduction rates in (b) were calculated from the non-adapted (baseline) word error rates.



(a) Word accuracy (b) Error reduction rate

Fig. 4. Total word accuracies (a) and error reduction rates (b) of incremental adaptation results for all the chunks and batch adaptation results that uses all the chunks simultaneously (non-adapted (baseline) performance = 66.8 %).

ure 4 shows the average word accuracies of all the chunks. We can see that MACROS (75.9%) was better than MLLR-MAP (75.1%) by 0.8 points. It would appear that MARCOS worked more robustly than MLLR-MAP owing to the effect of the covariance ($\hat{Q}$) considered in the posterior distributions. Thus, the proposed approach outperformed the conventional indirect and direct adaptation and their combinatorial method by robustly utilizing their practical advantages of quickness and stability. These results are consistent with the findings of theoretical discussions in Section 3.

Figure 4 also shows a comparison between the MACROS and batch adaptation which used all of 16 chunks at once. We see that MACROS performed better than the batch adaptation results. As we know from the results for a non-adapted model (baseline) in Figure 3 (a), the word accuracies in their original conditions were much different by chunks. This suggests that the acoustic conditions of speech can change during a period of lecture. Since the batch adaptation only utilizes temporally averaged statistics from data, it could not deal with the temporal change of conditions. In contrast, the proposed adaptation based on a macroscopic time evolution system (MACROS) could appropriately track the temporal change of conditions in a long lecture-type speech by incremental update of posterior distributions.

5. SUMMARY

This paper analytically and experimentally revealed that incremental adaptation based on a macroscopic time evolution system involves both indirect and direct adaptation approaches and simultaneously possesses their advantages of quickness and stability. The proposed incremental adaptation framework is based on a macroscopic time evolution system where posterior distributions are updated chunk-by-chunk. At this time, we require 16 utterances for a chunk to appropriately estimate the transformation parameters. In order to follow acoustic the temporal change of acoustic conditions more flexibly, our framework should work with fewer utterances, for example, by estimating transformation parameters based on Bayesian approaches (e.g. [14]).

6. REFERENCES

[1] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," in *Proceedings of the IEEE*, 2000, vol. 88, pp. 1241–1269.

[2] K. Shinoda and T. Watanabe, "Speaker adaptation with autonomous control using tree structure," in *Proc. EUROSPEECH95*, 1995, pp. 1143–1146.

[3] 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.

[4] 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.

[5] 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.

[6] T. Matsuoka and C.-H. Lee, "A study of on-line Bayesian adaptation for HMM-based speech recognition," in *Proc. EUROSPEECH'93*, 1993, pp. 815–818.

[7] 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.

[8] 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.

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

[10] G. Welch and G. Bishop, "An introduction to the Kalman filter," Tech. Rep. TR95-041, University of North Carolina at Chapel Hill, 1995.

[11] 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.

[12] V. Digalakis and L. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. on SAP*, vol. 4, pp. 294–300, 1996.

[13] S. Furui, K. Maekawa, and M H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," in *Proc. ASR2000*, 2000, pp. 244–248.

[14] K. Yu and M. J. F. Gales, "Bayesian adaptive inference and adaptive training," *IEEE Trans. on ASLP*, vol. 15, pp. 1932–1943, 2007.