Adaptive & Discriminative Speech Modeling to Cope with Temporal Changes of Environments

Georgia Institute of Technology
Professor Biing-Hwang (Fred) Juang
In collaboration with Shinji Watanabe

30th September 2011

Contents

1. Summary

The collaboration between Gatech and NTT in the current period (9/2010-8/2011) focuses on two main ideas: multiple time-scale evolution modeling of speech and a variational Bayes approach to discriminative training. The research work has produced very positive results, which are published in two papers, one at Interspeech (Florence, Italy, August 2011) and the other IEEE Workshop on Machine Learning for Signal Processing (Beijing, China, September 2011).

The current research focus is motivated by the general issue of robust acoustic modeling of speech for achieving superior and reliable performance across various application environments. As indicated in the previous report, in real environments, characteristics of a speech signal vary considerably, due to changes of contents, speakers, and ambience. There is thus an acute need in the methodology for the construction of robust and high-accuracy speech models that adaptively respond to these changes of the environment.

In particular, the approach that is being taken in this collaboration is based on discriminative modeling. We aim to apply our studies and new techniques to the task of speaker diarization and acoustic event detection in addition to speech recognition. In order to realize this application, we attempt to extend the incremental adaptation, as follows:

- Combination of multi-scale time evolution models to cope with various temporal characteristic changes in order to achieve robust incremental adaptation;
- A fully Bayesian treatment of the macroscopic time evolution system, in the spirit of regularized optimization, to enhance the generalizability of the model.

2. Multiple Time-Scale Evolution Model of Speech

Incremental statistical model adaptation based on a macroscopic time evolution model was proposed in 2006; it aims at tracking the temporal changes in the signal characteristics. The original algorithm, nevertheless, makes an assumption on the rate of temporal changes, which translates into a prescribed size of the temporal window over which speech sentences are used to facilitate the adaptation. However, the change in speech characteristics may be originated from various factors, at various (temporal) rates; for example, the ambient noise in a room may not change drastically over time but in a meeting environment the speaker change may be rather abrupt. These temporal changes have their own dynamics and therefore, it is necessary to extend the original single time evolution model to a multi-scale time evolution model, which has the potential of greatly increasing the model's robustness as it optimizes the parameters to match the nature of the characteristic change. Our accomplishment has been specifically on the **Integration of multiple time evolution streams in incremental adaptation systems.**

Our research result is reported in the attached paper, entitled "Model Adaptation for Automatic Speech Recognition Based on Multiple Time Scale Evolution," which is published at **Interspeech** 2011, Florence, Italy, August 2011. See Attachment A.

3. Variational Bayes method for regularized adaptive modeling with affine feature transformation

Model adaptation based on linear transformation on the feature vector has been reported to be effective in enhancing the system performance when the testing signal properties deviate from those in the signal used for system training. In the prevalent method of MLLR (maximum likelihood linear regression), a clustered set of regression matrices over a hierarchy of the modes of the collective acoustic model set is obtained. A further attempt to enhance the adaptation method with discriminative modeling objectives was proposed (Shin, Kim and Juang, ICASSP 2010). These regression or transformation models, while showing encouraging performance

improvements, are obtained through an unregularized procedure, causing concerns over its ultimate generalizability. This issue is also interesting when we incorporate the discriminative approach into an incremental adaptation framework with explicit time evolution model constraints. We have developed a Bayesian treatment of a linear regression (affine transformation) model of the HMM parameters. The work is summarized in a paper published at the IEEE Workshop on Machine Learning for Signal Processing (Beijing, China, September 2011), entitled "Bayesian Linear Regression for Hidden Markov Model Based on Optimizing Variational Bounds," which is included in this report as Attachment B.

Attachment A

Model Adaptation for Automatic Speech Recognition Based on Multiple Time Scale Evolution

Shinji Watanabe[1], *Atsushi Nakamura*[1], *and Biing-Hwang (Fred) Juang*[2]

[1]NTT Communication Science Laboratories, NTT Corporation
[2]Center for Signal and Image Processing, Georgia Institute of Technology
{watanabe.shinji,nakamura.atsushi}@lab.ntt.co.jp, juang@ece.gatech.edu

Abstract

The change in speech characteristics is originated from various factors, at various (temporal) rates in a real world conversation. These temporal changes have their own dynamics and therefore, we propose to extend the single (time-) incremental adaptations to a multiscale adaptation, which has the potential of greatly increasing the model's robustness as it will include adaptation mechanism to approximate the nature of the characteristic change. The formulation of the incremental adaptation assumes a time evolution system of the model, where the posterior distributions, used in the decision process, are successively updated based on a macroscopic time scale in accordance with the Kalman filter theory. In this paper, we extend the original incremental adaptation scheme, based on a single time scale, to multiple time scales, and apply the method to the adaptation of both the acoustic model and the language model. We further investigate methods to integrate the multi-scale adaptation scheme to realize the robust speech recognition performance. Large vocabulary continuous speech recognition experiments for English and Japanese lectures revealed the importance of modeling multiscale properties in speech recognition.

Index Terms: speech recognition, incremental adaptation, multiscale, time evolution system

1. Introduction

Recently work on automatic speech recognition has been shifting from laboratory simulations to more challenging real-world applications (e.g., broadcast news, meeting, and lecture recognition [1, 2]). In this situation, we are faced with various speech characteristics that speech recognition research has not thoroughly addressed yet. For example, in a real world conversation, speech characteristics could vary over a set of utterances due to change in speaker, speaking style, emotion, ambient noise, and topic. Conventional on-line incremental adaptation of acoustic and language models aims to model these changes in the speech characteristics usually at rather long, i.e., macroscopic time scales [3–5].

Nevertheless, it is important to note that these macroscopic changes are originated from various factors, at various (temporal) rates; for example, in a lecture recognition task, the ambient noise in a room may not change drastically over time while the speaking style and topic may change rather abruptly, as the lecture continues. These temporal changes have their own dynamics and therefore, we propose to extend the single incremental adaptations [4, 5] to one with multiple time scales (*multiscale*) taken into account, which has the potential of greatly increasing the model's robustness as it will include adaptation mechanism to approximate the nature of the characteristic change. There have been previous works that deal with temporal changes in speech characteristics at multiple time scales on feature or segmental (i.e. microscopic) units basis [6, 7] in speech recognition. Unlike these previous works, our approach focuses on relatively macroscopic time periods. Namely, whereas the motivation of [6, 7] is, for example, to model various speech dynamics governed by articulatory and noise factors observed with short time scales, the proposed approach deals with various speech dynamics governed by conversational factors, which manifest at a time scale much beyond the feature or articulatory level. The conversational factors will impact upon the acoustical and linguistic characteristics, and this paper formulates an incremental speech recognition process for both acoustic and language models.

The formulation of the incremental adaptation is based on the time evolution systems of acoustic and language models, where the posterior distributions are successively updated based on a macroscopic time scale in accordance with the Kalman filter theory. Then, we realize a multiscale adaptation by integrating the multiple single time evolution models, which are updated based on various time scales, to a multiscale time evolution model. The integration is performed in an ensemble classification, and we use a frame-based system combination approach [8]. Our experiments involve two lecture recognition tasks (Corpus of Spontaneous Japanese (CSJ [1]) and MIT-OpenCourseWare (MIT-OCW [2])) and the results show the effectiveness of the proposed approach.

2. Time evolution system perspective of automatic speech recognition

We first formulate a single-scale incremental speech recognition process for acoustic and language models from the perspective of time evolution systems within a probabilistic framework. Let $\{\mathbf{o}_n \in \mathbb{R}^D | n = 1, \cdots N\}$ be a D-dimensional feature vector sequence, and $\{w_m \in \mathbb{V} | m = 1, \cdots M\}$ be the corresponding word sequence with vocabulary size $|\mathbb{V}|$. In this paper, we consider that the feature and word sequences ($\{\mathbf{o}_n\}_{n=1}^N$ and $\{w_m\}_{m=1}^M$) are segmented (manually or automatically) as follows:

$$\{\mathbf{O}_t\}_{t=1}^T = \{\underbrace{\mathbf{o}_1, \cdots \mathbf{o}_{N_1}}_{\mathbf{o}_{t=1}}, \cdots, \underbrace{\mathbf{o}_{N_{T-1}+1}, \cdots, \mathbf{o}_N}_{\mathbf{o}_T}\},$$
$$\{\mathbf{W}_t\}_{t=1}^T = \{\underbrace{w_1, \cdots w_{M_1}}_{\mathbf{W}_{t=1}}, \cdots, \underbrace{w_{M_{T-1}+1}, \cdots, w_M}_{\mathbf{W}_T}\},$$

where t denotes a *macroscopic* time unit (e.g., an utterance or a set of utterances). Let $\mathbf{O}$ be an unknown feature vector sequence. Then, an incremental automatic speech recognizer, given previous data $\{\mathbf{O}_t, \mathbf{W}_t\}_{t=1}^T$, outputs a word sequence $\tilde{\mathbf{W}}$ based on the well-known Maximum A Posteriori (MAP) classi-

fication as follows:

$$\tilde{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}}\, p(\mathbf{W}|\mathbf{O}, \{\mathbf{O}_t, \mathbf{W}_t\}_{t=1}^{T}). \quad (1)$$

Similar to the standard decomposition of acoustic and language models, the posterior $p(\mathbf{W}|\mathbf{O}, \{\mathbf{O}_t, \mathbf{W}_t\}_{t=1}^{T})$ is decomposed into two models as follows:

$$p(\mathbf{W}|\mathbf{O}, \{\mathbf{O}_t, \mathbf{W}_t\}_{t=1}^{T}) \propto \underbrace{p(\mathbf{O}|\{\mathbf{O}_t, \mathbf{W}_t\}_{t=1}^{T})}_{\text{Incremental AM}} \underbrace{p(\mathbf{W}|\{\mathbf{W}_t\}_{t=1}^{T})}_{\text{Incremental LM}}. \quad (2)$$

Here, we assume the conditional independence of feature vector and word sequences, i.e., $p(\mathbf{W}|\{\mathbf{O}_t, \mathbf{W}_t\}_{t=1}^{T}) \approx p(\mathbf{W}|\{\mathbf{W}_t\}_{t=1}^{T})$, as usual. Since the decomposed pdfs respectively predict $\mathbf{O}$ and $\mathbf{W}$ given previous data, these pdfs can be interpreted as predictive distributions based on incrementally adapted acoustic and language models. The following subsections introduce time evolution system perspectives to these distributions.

2.1. Macroscopic time evolution system of acoustic models

Now we focus on the predictive distribution of the incremental acoustic model adaptation in Eq. (2). By introducing a set of current acoustic model (i.e., HMM) parameters Θ_T, the inference of acoustic models can be rewritten as follows:

$$p(\mathbf{O}|\{\mathbf{O}_t, \mathbf{W}_t\}_{t=1}^{T}) = \int p(\mathbf{O}|\Theta_T) p(\Theta_T|\{\mathbf{O}_t, \mathbf{W}_t\}_{t=1}^{T}) d\Theta_T. \quad (3)$$

Usually, instead of integrating out the posterior distribution of acoustic model parameters, we first point-estimate the model parameter values ($\bar{\Theta}_T$) based on the ML or MAP criterion, plug $\bar{\Theta}_T$ into the output distribution $p(\mathbf{O}|\Theta_T)$, and use it as a predictive distribution.

Thus, to obtain the predictive distribution (Eq. (3)), we require the posterior distribution of the acoustic model parameters $p(\Theta_T|\{\mathbf{O}_t, \mathbf{W}_t\}_{t=1}^{T})$. The posterior distribution can be recursively obtained from the previously estimated posterior distribution as follows:

$$p(\Theta_T|\{\mathbf{O}_t, \mathbf{W}_t\}_{t=1}^{T}) = p(\mathbf{O}_T|\Theta_T) \times \int p(\Theta_T|\Theta_{T-1}) p(\Theta_{T-1}|\{\mathbf{O}_t, \mathbf{W}_t\}_{t=1}^{T-1}) d\Theta_{t-1}, \quad (4)$$

where $p(\mathbf{O}_T|\Theta_T)$ is an output distribution (namely a likelihood function of an HMM), and $p(\Theta_T|\Theta_{T-1})$ corresponds to the dynamics of the HMM parameters. By using linear (MLLR type) dynamics[1] for $p(\Theta_T|\Theta_{T-1})$, Eq. (4) can be analytically solved as a time evolution system of the HMM parameters in accordance with the Kalman filter theory [4]. This incremental adaptation can achieve robustness based on the predictor-corrector algorithm of the Kalman filter theory, which theoretically involves the conventional MAP and MLLR and their combinatorial adaptation approaches.

2.2. Topic tracking language models

Similar to the predictive distribution of the incremental acoustic model adaptation, the inference of language models in Eq. (2) can be rewritten as follows by introducing a set of current language model (i.e., a topic model and an n-gram if we use a topic-based n-gram language model) parameters Λ_T, :

$$p(\mathbf{W}|\{\mathbf{W}_t\}_{t=1}^{T}) = \int p(\mathbf{W}|\Lambda_T) p(\Lambda_T|\{\mathbf{W}_t\}_{t=1}^{T}) d\Lambda_T. \quad (5)$$

The plug-in approach is also usually used instead of Eq. (5) to obtain a predictive distribution. The posterior distribution for language model parameters can be recursively obtained from the previously estimated posterior distribution as follows:

$$p(\Lambda_T|\{\mathbf{W}_t\}_{t=1}^{T}) = p(\mathbf{W}_T|\Lambda_T) \times \int p(\Lambda_T|\Lambda_{T-1}) p(\Lambda_{T-1}|\{\mathbf{W}_t\}_{t=1}^{T-1}) d\Lambda_{T-1}, \quad (6)$$

where $p(\mathbf{W}_T|\Lambda_T)$ is an output distribution (namely a multinomial distribution for an n-gram), and $p(\Lambda_T|\Lambda_{T-1})$ corresponds to the dynamics of the n-gram and topic model parameters. [5] uses a Latent Dirichlet Allocation (LDA)-based topic model, and employs topic model dynamics represented by a Dirichlet distribution[2]. Then, Eq. (6) can also be analytically solved as a time evolution system of the language model parameters in accordance with a discrete version of the Kalman filter theory. This approach can track topics as adaptation continues, and is called the topic tracking language model. The inference part of this adaptation is performed by collapsed Gibbs sampling.

Thus, we reveal that an incremental speech recognition process can be viewed as time evolution systems of acoustic and language models within a probabilistic framework. The next section extends the single time evolution system to a multiscale time evolution system.

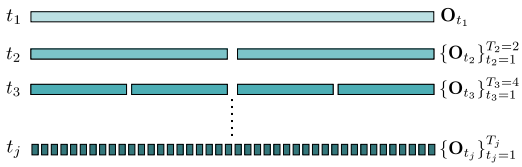

Figure 1: Example multiscale representation of feature sequences.

3. Multiscale time evolution system

This paper considers a multiscale adaptation that has the various step sizes of incremental adaptations. For example, one adaptation size consisted of one utterance to track the abrupt changes in utterance level in speech (e.g., speakers and speaking styles), and another adaptation size consisted of some dozens of utterances to track the long-term changes in speech (e.g., room acoustics and topics), as shown in Figure 1. Each model can be obtained by incrementally updating acoustic and language models with its adaptation size. Then, the problem is how to integrate these models to perform a multiscale adaptation.

This is similar to the problem of multistream speech recognition, and there are several ways to realize integration. For example, multistream adaptation [3] integrates several acoustic model adaptation streams by linearly interpolating Gaussian mean vectors of the streams in HMMs. [6] focuses on the bias (shift) vectors of Gaussian mean vectors, and these are estimated by linearly interpolating various-scale bias (shift) vectors. Topic-based multiscale language models are realized by linearly interpolating various-scale n-gram probabilities [9].

[1] The stochastic dynamics of kth Gaussian mean vector in an HMM can be represented as a Gaussian distribution (i.e., $\mathcal{N}(\boldsymbol{\mu}_T^k|\mathbf{A}_T^k\boldsymbol{\mu}_{T-1}^k + \mathbf{b}_T^k)$ where $(\mathbf{A}_T^k, \mathbf{b}_T^k)$ is an affine transformation matrix).

[2] The stochastic dynamics of topic proportion probability ϕ^r for latent topic r can be represented as a Dirichlet distribution (i.e., $\mathcal{D}(\{\phi_T^r\}_r|\{\alpha_T\boldsymbol{\phi}_{T-1}^r\}_r)$ where α_T is a precision).

The above parameter-level integration can tightly integrate multiple models into one multiscale model. However, the integration process often becomes complex, and it limits integration with the same model structures (i.e., the same model topology of HMMs or n-grams). To avoid these problems, this paper proposes the use of hypothesis-level ensemble classification integration as a simple realization of multiscale integration.

Let j be an adaptation scale index, and J be the number of scales. We consider j as a random variable, and represent the multiscale MAP classification from Eq. (1), as follows:

$$\begin{aligned}\tilde{\mathbf{W}} &= \operatorname*{argmax}_{\mathbf{W}} p(\mathbf{W}|\mathbf{O}, \{\mathbf{o}_n\}_{n=1}^{N}, \{w_m\}_{m=1}^{M}) \\ &= \operatorname*{argmax}_{\mathbf{W}} \sum_{j=1}^{J} p(\mathbf{W}|\mathbf{O}, \{\mathbf{o}_n\}_{n=1}^{N}, \{w_m\}_{m=1}^{M}, j) \\ &\qquad p(j|\mathbf{O}, \{\mathbf{o}_n\}_{n=1}^{N}, \{w_m\}_{m=1}^{M}) \\ &\approx \operatorname*{argmax}_{\mathbf{W}} \sum_{j=1}^{J} p(\mathbf{W}|\mathbf{O}, \{\mathbf{O}_{t_j}, \mathbf{W}_{t_j}\}_{t_j=1}^{T_j}),\end{aligned} \tag{7}$$

where we assume that $p(j|\mathbf{O}, \{\mathbf{o}_n\}_{n=1}^{N}, \{w_m\}_{m=1}^{M})$ is a uniform distribution. This fusion corresponds to using a hypothesis-level ensemble classification of the multiscale integration. If we introduce the predictive distributions of the acoustic and language models with each adaptation scale j, Eq. (7) can be represented as follows:

$$\begin{aligned}\tilde{\mathbf{W}} = \operatorname*{argmax}_{\mathbf{W}} &\sum_{j=1}^{J} p(\mathbf{O}|\{\mathbf{O}_{t_j}, \mathbf{W}_{t_j}\}_{t_j=1}^{T_j}) \\ &p(\mathbf{W}|\{\mathbf{W}_{t_j}\}_{t_j=1}^{T_j}).\end{aligned} \tag{8}$$

Thus, we derive a multiscale time evolution system of the acoustic and language models as an ensemble classification problem. This paper uses frame-level hypothesis integration [8] for this problem, which is based on the definition of a time frame-wise word error cost function in a minimum Bayes risk framework.

4. Experiments

We show the effectiveness of the multiscale time evolution system by performing large vocabulary continuous speech recognition experiments using English and Japanese lectures. We used an MIT OpenCourseWare (OCW) task [2] and a CSJ task [1]. The experimental conditions for the MIT-OCW task are summarized in Table 1. The initial acoustic model was constructed by using variational Bayesian triphone clustering [10] and differenced Maximum Mutual Information (dMMI) training [11]. The development set consisted of 2 lectures (3,460 utterances, 23,720 words, and 2.1 hours) and the evaluation set consisted of 8 lectures (6,989 utterances, 72,159 words, and 7.8 hours). The development set was used to tune the adaptation parameters (e.g., the occupancy threshold in the MLLR adaptation, system noise parameter, language model weights). We used a one-pass WFST-based decoder that employs a pair of WFSTs for composition during decoding by a fast on-the-fly composition technique [12].

In the incremental adaptation process, the following three operations were performed in each adaptation unit: 1) obtaining lattice-based hypotheses of utterances by automatic speech recognition using a previously obtained set of models, 2) applying the adaptation to the previously obtained set of models by using the lattices, and 3) again recognizing the utterances by using the lattices and the adapted set of models.

Table 1: Experimental conditions for an MIT-OCW task

Sampling rate/quantization	16 kHz / 16 bit
Observation vector	12 order MFCC with energy
(39 dimensions)	+Δ+$\Delta\Delta$ (CMS)
Window	Hamming
Frame size/shift	25/10 ms
Num. of temporal HMM states	3 (left to right)
Num. of phoneme categories	52
Num. of clustered HMM states	2,565
Num. of mixture components / state	32
Language model	3-gram (KN discounting)
Vocabulary size	44K

Table 2: Word error rate (%) of multi-scale time evolution system of *acoustic* models in an MIT-OCW task.

Method	Dev.	Eval.
Baseline	25.5	28.3
Single scale (4)	26.9	23.4
Single scale (8)	25.5	23.1
Single scale (16)	25.0	23.5
Single scale (32)	24.6	24.2
Single scale (64)	24.9	25.2
Multi scale	**24.3**	**22.4**

Table 2 describes the results of single-scale time evolution systems based on 4, 8, 16, 32, and 64 utterances as a time unit and the multiscale time evolution system. In the MIT-OCW task, we only used acoustic model adaptation based on Section 2.1. Within the results of single-scale time evolution systems, Single scale (32) achieves the best score in the development set, while Single scale (8) performs best in the evaluation set. This result shows that each lecture has its own appropriate-size dynamics (e.g., 32 utterances for the development set, and 8 utterances for the evaluation set), and suffers the limitation of the single-scale time evolution system with one time scale. However, the multiscale time evolution system increased the model's robustness because it would include adaptation mechanism to approximate the nature of the change in characteristics, and improved the recognition performance for both the development and evaluation sets. Thus, we showed the effectiveness of the multiscale time evolution system of acoustic models.

In the CSJ task, we further examined the multiscale time evolution systems of both acoustic and language models. The experimental conditions for the CSJ task are summarized in Table 3. The initial acoustic and language models were trained by discriminative approaches [11, 13]. We also used the on-the-fly WFST-based decoder [12]. We used CSJ testset 2 as a development set (10 lectures, 794 utterances, 26,798 words, and 2.2 hours) and CSJ testset 1 as an evaluation set (10 lectures, 977 utterances, 26,329 words, and 2.0 hours). The development set was used to tune the adaptation parameters of the acoustic and language models similar to the MIT-OCW task. In this experiment, the utterances were automatically segmented from the lectures using non-linear Kalman filtering based VAD [14].

Tables 4 and 5 show the results for the time evolution systems of the acoustic and language models, respectively. The language model adaptation described in Section 2.2 dealt with uni-gram language models, and was applied to n-gram language models by using rescaling techniques. In both experiments, the multiscale time evolution system achieved the best scores, which shows the effectiveness of the consideration of the multiscale dynamics among the acoustical and linguistic characteristics. Finally, in Table 6, we show how we realized the simultaneous incremental adaptation of acoustic and language models by successively adapting these models in a cascade manner.

This result also shows the effectiveness of the multiscale time evolution system, and finally improved WERs by 5.0 % and 3.8 % in the development and evaluation sets, respectively.

Thus, from a series of experimental results, we revealed the importance of modeling multiscale properties in speech recognition.

5. Conclusion

This paper proposes a multiscale time evolution system for acoustic and language models, and lecture recognition tasks show the effectiveness of the proposed approach experimentally. We believe that the consideration of multiscale acoustic and linguistic characteristics in speech is an essential problem that must be overcome if we are to appropriately model speech dynamics in real world applications, and this direction is supported by our experimental results to some extent. Future work will examine the integration aspect of the multiscale time evolution system to achieve more robust integration. We will also apply the multiscale time evolution system to meeting and broadcast news tasks where speaker and topic changes occur frequently.

6. Acknowledgements

We thank the MIT Spoken Language Systems Group for helping us to perform speech recognition experiments based on MIT-OCW [2]. We also thank Dr. Hoffmeister at RWTH Aachen University (currently at Yap Inc.) for helping us to use a frame-based system combination technique.

7. References

[1] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proceedings of LREC2000*, 2000, vol. 2, pp. 947–952.

[2] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent progress in the MIT Spoken Lecture Processing Project," in *Proc. Interspeech'07*, 2007, pp. 2553–2556.

[3] Q. Huo and B. Ma, "Online adaptive learning of continuous-density hidden Markov models based on multiple-stream prior evolution and posterior pooling," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 388–398, 2001.

[4] S. Watanabe and A. Nakamura, "Predictor–corrector adaptation by using time evolution system with macroscopic time scale," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 395–406, 2010.

[5] S. Watanabe, T. Iwata, T. Hori, A. Sako, and Y. Ariki, "Topic tracking language model for speech recognition," *Computer Speech and Language*, vol. 25, no. 2, pp. 440–461, 2011.

[6] A. Kannan and M. Ostendorf, "Modeling dependency in adaptation of acoustic models using multiscale tree processes," in *Eurospeech'97*, 1997, vol. 4, pp. 1863–1866.

Table 3: Experimental conditions for a CSJ task

Sampling rate/quantization	16 kHz / 16 bit
Observation vector	12 order MFCC with energy
(39 dimensions)	$+\Delta+\Delta\Delta$ (CMS)
Window	Hamming
Frame size/shift	25/10 ms
Num. of temporal HMM states	3 (left to right)
Num. of phoneme categories	43
Num. of clustered HMM states	5,000
Num. of mixture components / state	32
Language model	3-gram (Good Turing) + discriminative LM [13]
Vocabulary size	100K

Table 4: Word error rate (%) of multiscale time evolution system of *acoustic* models in a CSJ task.

Method	Dev.	Eval.
Baseline	17.9	21.0
Single scale (4)	15.4	19.4
Single scale (8)	14.4	18.9
Single scale (16)	14.5	19.1
Single scale (32)	14.7	19.3
Single scale (64)	14.5	18.9
Multiscale	**13.8**	**18.3**

Table 5: Word error rate (%) of multiscale time evolution system of *language* models in a CSJ task.

Method	Dev.	Eval.
Baseline	17.9	21.0
Single scale (1)	16.2	19.3
Single scale (2)	16.0	19.6
Single scale (4)	16.0	19.5
Single scale (8)	16.0	19.5
Single scale (16)	15.8	19.3
Single scale (32)	16.0	19.3
Single scale (64)	16.0	19.5
Multiscale	**15.2**	**18.8**

[7] N. Morgan et al., "Pushing the envelope-aside: Beyond the spectral envelope as the fundamental representation for speech recognition," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 81–88, 2005.

[8] B. Hoffmeister, T. Klein, R. Schlüter, and H. Ney, "Frame based system combination and a comparison with weighted ROVER and CNC," in *Proc. Interspeech'06*, 2006.

[9] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda, "Online multiscale dynamic topic models," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 663–672.

[10] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 365–381, 2004.

[11] E. McDermott, S. Watanabe, and A. Nakamura, "Discrminative training based on an integrated view of MPE and MMI in margin and error space," in *Proc. ICASSP'10*, 2010, pp. 4894–4897.

[12] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1352–1365, 2007.

[13] T. Oba, T. Hori, and A. Nakamura, "A study of efficient discriminative word sequences for reranking of recognition results based on n-gram counts," in *Proc. Interspeech'07*, 2007, pp. 1753–1756.

[14] M. Fujimoto, K. Ishizuka, and H. Kato, "Noise robust voice activity detection based on statistical model and parallel non-linear Kalman filtering," in *Proc. ICASSP'07*, 2007, vol. 4, pp. 797–800.

Table 6: Word error rate (%) of multiscale time evolution system of *acoustic* and *language* model adaptation in a CSJ task.

Method	Dev.	Eval.
Baseline	17.9	21.0
Single scale (4)	15.9	18.6
Single scale (8)	13.5	18.0
Single scale (16)	13.9	18.0
Single scale (32)	13.7	17.9
Single scale (64)	13.3	17.9
Multiscale	**12.9**	**17.2**

Attachment B

BAYESIAN LINEAR REGRESSION FOR HIDDEN MARKOV MODEL BASED ON OPTIMIZING VARIATIONAL BOUNDS

Shinji Watanabe, Atsushi Nakamura

NTT Communication Science Laboratories
NTT Corporation
{watanabe.shinji,nakamura.atsushi}@lab.ntt.co.jp

Biing-Hwang (Fred) Juang

Center for Signal and Image Processing
Georgia Institute of Technology
juang@ece.gatech.edu

ABSTRACT

Linear regression for Hidden Markov Model (HMM) parameters is widely used for the adaptive training of time series pattern analysis especially for speech processing. This paper realizes a fully Bayesian treatment of linear regression for HMMs by using variational techniques. This paper analytically derives the variational lower bound of the marginalized log-likelihood of the linear regression. By using the variational lower bound as an objective function, we can optimize the model topology and hyper-parameters of the linear regression without controlling them as tuning parameters; thus, we realize linear regression for HMM parameters in a non-parametric Bayes manner. Experiments on large vocabulary continuous speech recognition confirm the generalization effect of the proposed approach, especially for small quantities of adaptation data.

1. INTRODUCTION

Hidden Markov Models (HMM) have been widely used for time series analysis (e.g., speech, text, and image processing). HMM parameters can be trained by using a large amount of data based on statistical approaches. However, since we cannot collect data for all environments in advance, the adaptive training of HMM parameters to a specific environment with a small amount of data is an important issue. In speech recognition research, linear regression for Hidden Markov Model (HMM) parameters has been developed for the adaptive training of HMMs [1, 2].

The linear regression approach for HMM parameters estimates the linear (affine) transformation parameters from the initial to target HMM parameters, instead of the HMM parameters themselves. Then, by sharing a linear transformation parameter among many Gaussians in the HMMs, the number of free parameters becomes small, and therefore the linear transformation parameters can be trained with a small amount of adaptation data. The parameters are usually estimated by using the maximum likelihood EM algorithm (called Maximum Likelihood Linear Regression (MLLR)). In addition, to achieve more generalization abilities than MLLR, some approximated Bayesian approaches are applied to the estimation of the linear regression parameter (e.g., Maximum A Posteriori Linear Regression (MAPLR) [3] and structural MAPLR (SMAPLR) [4]).

Recently, a fully Bayesian treatment of latent models has been developed in the machine learning field based on a *variational technique* [5–7]. This variational Bayes is successfully applied to HMM training in speech recognition [8–15], and the estimation of the transformation parameters of HMMs [16, 17]. In this paper, we provide an analytical solution for the fully Bayesian linear regression using variational Bayes by deriving the variational lower bound (evidence) of the marginalized log-likelihood. By using this lower bound as an objective function, we can optimize the model topology and hyper-parameters in the linear regression without controlling them as tuning parameters; thus, we realize linear regression for HMM parameters in a non-parametric Bayes manner. We performed unsupervised speaker adaptation experiments for a large vocabulary continuous speech recognition task, and confirmed the effectiveness of the proposed approach.

2. LINEAR REGRESSION FOR HIDDEN MARKOV MODELS BASED ON VARIANCE NORMALIZED REPRESENTATION

This section briefly explains a solution for the linear regression parameters for hidden Markov models within a maximum likelihood EM algorithm framework. This paper uses a solution based on a variance normalized representation of Gaussian mean vectors to make the solution simple[1]. In this paper, we only focus on the transformation of Gaussian mean vectors in HMMs.

2.1. Maximum likelihood EM algorithm for hidden Markov models

First, we explain the EM algorithm of the conventional HMM parameter estimation. Let $\mathbf{O} \triangleq \{\mathbf{o}_t \in \mathbb{R}^D | t = 1, \cdots, T\}$ be a set of D dimensional feature vectors for T speech frames, and $\mathbf{V}$ be the corresponding set of latent variables. The latent variables in a continuous density HMM are composed of HMM states and mixture components of Gaussian Mixture Models (GMMs). The EM algorithm deals with the following auxiliary function as an optimization function instead of directly using the data likelihood:

$$Q(\mathbf{\Theta}) \triangleq \langle p(\mathbf{O}, \mathbf{V}|\mathbf{\Theta}) \rangle_{p(\mathbf{V}|\mathbf{O})}, \tag{1}$$

where $\mathbf{\Theta}$ is a set of HMM parameters. $p(\mathbf{V}|\mathbf{O})$ is a posterior distribution of latent variables with the previously estimated HMM parameters. Eq (1) is an expected value, and is efficiently computed by using the forward-backward algorithm as the E-step of the EM algorithm.

The M-step of the EM algorithm estimates HMM parameters, as follows:

$$\bar{\mathbf{\Theta}} = \underset{\mathbf{\Theta}}{\operatorname{argmax}}\, Q(\mathbf{\Theta}). \tag{2}$$

[1] This is first described in [18] as normalized domain MLLR. The structural Bayes approach [19] for bias vector estimation in HMM adaptation also uses this normalized representation.

The E-step and M-step are performed iteratively with a convergence judgment, and finally we obtain the HMM parameters.

2.2. Maximum likelihood solution based on EM algorithm and variance normalized representation

This section focuses on the linear transformation parameters within the EM algorithm. Similar to Eq. (1), the auxiliary function with respect to a set of transformation parameters $\mathbf{W}$ can be represented as follows:

$$Q(\mathbf{W}) = \sum_{t,k} \zeta_{k,t} \log \mathcal{N}(\mathbf{o}_t | \boldsymbol{\mu}_k^{ad}, \boldsymbol{\Sigma}_k), \tag{3}$$

where $\zeta_{k,t}$ is the posterior probability of mixture component k at t. $\boldsymbol{\mu}_k^{ad}$ is a transformed mean vector with $\mathbf{W}$, and the concrete form of this vector is discussed in the next paragraph. In the Q function, we disregard the parameters of the state transition probabilities and the mixture weights since they do not depend on the optimization with respect to $\mathbf{W}$. $\mathcal{N}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean parameter $\boldsymbol{\mu}$ and covariance matrix parameter $\boldsymbol{\Sigma}$, and is defined as follows:

$$\begin{aligned}&\mathcal{N}(\mathbf{o}_t | \boldsymbol{\mu}_k^{ad}, \boldsymbol{\Sigma}_k) \\ &\triangleq g(\boldsymbol{\Sigma}_k) \exp\left(-\frac{1}{2}\mathrm{tr}\left[(\boldsymbol{\Sigma}_k)^{-1}(\mathbf{o}_t - \boldsymbol{\mu}_k^{ad})(\mathbf{o}_t - \boldsymbol{\mu}_k^{ad})'\right]\right),\end{aligned} \tag{4}$$

where tr[·] and $'$ mean the trace and transposition operations of a matrix, respectively. $g(\boldsymbol{\Sigma}_k)$ is a normalization constant, and is defined as follows:

$$g(\boldsymbol{\Sigma}_k) \triangleq (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}}. \tag{5}$$

In the following paragraphs, we derive Eq. (3) as a function of $\mathbf{W}$ to optimize $\mathbf{W}$, similar to Eq. (2).

We consider the concrete form of the vector transformed mean vector based on the variance normalized representation. We first define Cholesky decomposition matrix $\mathbf{C}_k$ as follows:

$$\boldsymbol{\Sigma}_k \triangleq \mathbf{C}_k (\mathbf{C}_k)'. \tag{6}$$

$\mathbf{C}_k$ is a $D \times D$ triangular matrix. Then, the affine transformation of a Gaussian mean vector in a covariance normalized space is represented as follows:

$$\boldsymbol{\mu}_k^{ad} = \mathbf{C}_k \mathbf{W}_j \begin{pmatrix} 1 \\ (\mathbf{C}_k)^{-1} \boldsymbol{\mu}_k^{ini} \end{pmatrix}. \triangleq \mathbf{C}_k \mathbf{W}_j \boldsymbol{\xi}_k. \tag{7}$$

$\boldsymbol{\xi}_k$ is an augmented vector of an initial (non-adapted) Gaussian mean vector $\boldsymbol{\mu}_k^{ini}$. $\mathbf{W}_j$ is a $D \times (D+1)$ affine transformation matrix. Here j is a cluster index where each cluster holds a set of Gaussians. Namely, transformation parameter $\mathbf{W}_j$ is shared among a set of Gaussians c_j. The clustered structure of the Gaussians is usually represented as a binary tree where a set of Gaussians belongs to each node.

The Q function of $\mathbf{W} = \{\mathbf{W}_j\}_j$ is represented by substituting Eqs. (7) and (4) into Eq. (3) as follows:

$$\begin{aligned}&\sum_{k \in c_j, t} \zeta_{k,t} \log \mathcal{N}(\mathbf{o}_t | \mathbf{C}_k \mathbf{W}_j \boldsymbol{\xi}_k, \boldsymbol{\Sigma}_k) \\ &= \sum_{k \in c_j} \zeta_k \log g(\boldsymbol{\Sigma}_k) \\ &\quad - \frac{1}{2}\mathrm{tr}\left[\mathbf{W}_j' \mathbf{W}_j \boldsymbol{\Xi}_j - 2\mathbf{W}_j' \mathbf{Z}_j + \sum_{k \in c_j} \boldsymbol{\Sigma}_k^{-1} \mathbf{S}_k\right],\end{aligned} \tag{8}$$

where $\boldsymbol{\Xi}_j$ and $\mathbf{Z}_j$ are 0th and 1st order statistics of linear regression parameters defined as:

$$\begin{cases} \boldsymbol{\Xi}_j \triangleq \sum_{k \in c_j} \boldsymbol{\xi}_k (\boldsymbol{\xi}_k)' \sum_t \zeta_{k,t}. \\ \mathbf{Z}_j \triangleq \sum_{k \in c_j} \zeta_k (\mathbf{C}_k)^{-1} \left(\sum_t \zeta_{k,t} \mathbf{o}_t\right) (\boldsymbol{\xi}_k)'. \end{cases} \tag{9}$$

Here $\mathbf{Z}_j$ is a $D \times (D+1)$ matrix and $\boldsymbol{\Xi}_j$ is a $(D+1) \times (D+1)$ symmetric matrix. ζ_k and $\mathbf{S}_k$ are defined as follows:

$$\begin{cases} \zeta_k = \sum_t \zeta_{k,t} \\ \mathbf{S}_k = \sum_t \zeta_{k,t} \mathbf{o}_t \mathbf{o}_t' \end{cases} \tag{10}$$

These are the 0th and 2nd order sufficient statistics of Gaussians in HMMs, respectively.

Since Eq. (8) is represented as a function of $\mathbf{W}$, we can obtain the optimal $\bar{\mathbf{W}}$, similar to Eq. (2).. By differentiating the Q function with respect to $\mathbf{W}_j$, we can derive the following equation:

$$\frac{\partial}{\partial \mathbf{W}_j} Q(\mathbf{W}) = 0. \Rightarrow \mathbf{Z}_j - \bar{\mathbf{W}}_j \boldsymbol{\Xi}_j = 0. \tag{11}$$

Thus, we can obtain the following analytical solution unless we assume a diagonal covariance matrix:

$$\bar{\mathbf{W}}_j = \mathbf{Z}_j \boldsymbol{\Xi}_j^{-1}. \tag{12}$$

Therefore, the optimized mean vector parameter is represented as:

$$\boldsymbol{\mu}_k^{ad} = \mathbf{C}_k \mathbf{Z}_j \boldsymbol{\Xi}_j^{-1} \boldsymbol{\xi}_k. \tag{13}$$

This solution means that the affine transformation occurs in the covariance-normalized space. This solution corresponds to the M-step of the EM algorithm, and the E-step is similar to that of HMMs.

3. BAYESIAN LINEAR REGRESSION

This section provides an analytical solution for Bayesian linear regression by using a variational lower bound.

3.1. Variational lower bound

With regard to the variational Bayesian approaches, we first focus on the lower bound of the marginalized likelihood with a set of hyper-parameters $\boldsymbol{\Psi}$ and a model structure m[2], as follows:

$$\begin{aligned}&\log p(\mathbf{O}|m, \boldsymbol{\Psi}) \\ &= \log\left(\int \sum_{\mathbf{V}} p(\mathbf{O}, \mathbf{V}|\mathbf{W}) p(\mathbf{W}|\mathbf{O}; m, \boldsymbol{\Psi}) d\mathbf{W}\right) \\ &\geq \underbrace{\left\langle \log \frac{p(\mathbf{O}, \mathbf{V}|\mathbf{W}) p(\mathbf{W}; m, \boldsymbol{\Psi})}{q(\mathbf{W}|\mathbf{O}; m, \boldsymbol{\Psi}) q(\mathbf{V}|\mathbf{O}; m, \boldsymbol{\Psi})} \right\rangle_{q(\mathbf{W}|\mathbf{O}; m, \boldsymbol{\Psi}) q(\mathbf{V}|\mathbf{O}; m, \boldsymbol{\Psi})}}_{\triangleq \mathcal{F}(\boldsymbol{\Phi}, m)},\end{aligned} \tag{14}$$

[2] $\boldsymbol{\Psi}$ and m can also be marginalized by setting their distributions. This paper point-estimates $\boldsymbol{\Psi}$ and m by a MAP approach.

where $p(\mathbf{W}|\mathbf{O}; m, \mathbf{\Psi})$ is a prior distribution of $\mathbf{W}$, and $q(\mathbf{W}|\mathbf{O}; m, \mathbf{\Psi})$ and $q(\mathbf{V}|\mathbf{O}; m, \mathbf{\Psi})$ are arbitrary distributions. For simplicity, we omit m, $\mathbf{\Phi}$, and $\mathbf{O}$ from these distributions. The variational Bayes regards variational lower bound $\mathcal{F}(\mathbf{\Phi}, m)$ as an objective function

The variational lower bound is a better approximation of the marginalized log likelihood than the auxiliary functions of maximum likelihood EM and maximum a posteriori EM algorithms that point-estimate model parameters, especially for small amount of training data [5–7]. Therefore, the variational Bayes can mitigate the sparse data problem that the conventional approaches face on Although it contains the complicated integration and expectation, it can be simplified analytically by using conjugate distributions as prior distributions, and assuming the conditional independence on the posterior distributions.

From variational calculus, we obtain the optimal posterior distributions $q(\mathbf{W})$ and $q(\mathbf{V})$ (we call them VB posteriors), as follows:

$$\begin{aligned} \tilde{q}(\mathbf{W}) &= \operatorname*{argmax}_{q(\mathbf{W})} \mathcal{F}(\mathbf{\Phi}, m) \\ \tilde{q}(\mathbf{V}) &= \operatorname*{argmax}_{q(\mathbf{V})} \mathcal{F}(\mathbf{\Phi}, m) \end{aligned} \tag{15}$$

This optimization steps are performed alternately, and finally obtain local optimum solutions, similar to the EM algorithm. To solve the VB posteriors, we first set a prior distribution $p(\mathbf{W})$ in the next section.

3.2. Conjugate distribution setting

With a Bayesian approach, we need to set a prior distribution of $\mathbf{W}$. A conjugate distribution is preferable as regards obtaining an analytical solution, and we set a matrix normal distribution similar to Maximum A Posteriori Linear Regression (MAPLR [3]). A matrix normal distribution is defined as follows:

$$\begin{aligned} p(\mathbf{W}) &= \prod_j p(\mathbf{W}_j) = \prod_j \mathcal{N}(\mathbf{W}_j|\mathbf{M}_j, \mathbf{\Phi}_j, \mathbf{\Omega}_j) \\ &\triangleq \prod_j \frac{\exp\left(-\frac{1}{2}\mathrm{tr}\left[(\mathbf{W}_j - \mathbf{M}_j)'\mathbf{\Phi}_j^{-1}(\mathbf{W}_j - \mathbf{M}_j)\mathbf{\Omega}_j^{-1}\right]\right)}{(2\pi)^{D(D+1)/2}|\mathbf{\Omega}_j|^{D/2}|\mathbf{\Phi}_j|^{(D+1)/2}}, \end{aligned} \tag{16}$$

where $\mathbf{M}_j$ is a $D \times (D+1)$ matrix, $\mathbf{\Omega}_j$ is a $(D+1) \times (D+1)$ matrix, and $\mathbf{\Phi}_j$ is a $D \times D$ matrix. These are hyper-parameters of the matrix normal distribution. There are many hyper-parameters to be set, and this makes the implementation complicated. In this paper, we try to find another conjugate distribution with fewer hyper-parameters than Eq. (16). To obtain a simple solution for the final analytical results, we set the following constraints on $\mathbf{\Omega}_j$ and $\mathbf{\Phi}_j$:

$$\begin{aligned} \mathbf{\Phi}_j &\approx \mathbf{I}_D, \\ \mathbf{\Omega}_j &\approx \rho_j^{-1}\mathbf{I}_{D+1}, \end{aligned} \tag{17}$$

where $\mathbf{I}_D$ is the $D \times D$ identity matrix. ρ_j indicates a precision parameter. Then, Eq. (16) can be rewritten as follows:

$$\begin{aligned} &\mathcal{N}(\mathbf{W}_j|\mathbf{M}_j, \mathbf{I}_D, \rho_j^{-1}\mathbf{I}_{D+1}) \\ &= g(\rho_j^{-1}\mathbf{I}_{D+1}) \exp\left(-\frac{1}{2}\mathrm{tr}\left[\rho_j(\mathbf{W}_j - \mathbf{M}_j)'(\mathbf{W}_j - \mathbf{M}_j)\right]\right), \end{aligned} \tag{18}$$

where $g(\rho_j^{-1}\mathbf{I}_{D+1})$ is a normalization constant, and defined as

$$g(\rho_j^{-1}\mathbf{I}_{D+1}) \triangleq \left(\frac{\rho_j}{2\pi}\right)^{\frac{D(D+1)}{2}}. \tag{19}$$

These approximations can derive simple solutions for Bayesian linear regression.

3.3. Posterior distribution of transformation matrix

From the variational calculation for $\mathcal{F}(\mathbf{\Phi}, m)$ with respect to $q(\mathbf{W}_j$, we obtain the following posterior distribution:

$$\tilde{q}(\mathbf{W}_j) \propto p(\mathbf{W}_j) \exp\left(\langle p(\mathbf{O}, \mathbf{V}|\mathbf{W}_j)\rangle_{q(\mathbf{V})}\right). \tag{20}$$

Here, we assume the following conditional independence of the posterior distributions for each $\mathbf{W}_j$, i.e.,

$$q(\mathbf{W}) = \prod_j q(\mathbf{W}_j). \tag{21}$$

After expectation with respect to $q(\mathbf{V})$, we can obtain the following expression:

$$\tilde{q}(\mathbf{W}_j) \propto p(\mathbf{W}_j) \exp\left(\sum_{t,k\in c_j} \zeta_{k,t} \log \mathcal{N}(\mathbf{o}_t|\mathbf{C}_k\mathbf{W}_j\boldsymbol{\xi}_k, \mathbf{\Sigma}_k)\right). \tag{22}$$

By substituting Eqs. (8) and (18) into Eq. (22), we can finally derive the quadratic form of $\mathbf{W}_j$ as follows:

$$\begin{aligned} &\log(\tilde{q}(\mathbf{W}_j)) \\ &\propto -\frac{1}{2}\mathrm{tr}\left[\rho_j\mathbf{W}_j'\mathbf{W}_j + \mathbf{W}_j'\mathbf{W}_j\mathbf{\Xi}_j - 2\rho_j\mathbf{W}_j'\mathbf{M}_j - 2\mathbf{W}_j'\mathbf{Z}_j\right] \\ &= -\frac{1}{2}\mathrm{tr}\left[\mathbf{W}_j'\mathbf{W}_j(\rho_j\mathbf{I}_{D+1} + \mathbf{\Xi}_j) - 2\mathbf{W}_j'(\rho_j\mathbf{M}_j + \mathbf{Z}_j)\right], \end{aligned} \tag{23}$$

where we disregard the terms that do not depend on $\mathbf{W}_j$. Thus, by defining the following matrix variables

$$\begin{aligned} \tilde{\mathbf{\Omega}}_j &= (\rho_j\mathbf{I}_{D+1} + \mathbf{\Xi}_j)^{-1} \\ \tilde{\mathbf{M}}_j &= (\rho_j\mathbf{M}_j + \mathbf{Z}_j)\,\tilde{\mathbf{\Omega}} \end{aligned} \tag{24}$$

we can finally obtain the posterior distribution of $\mathbf{W}_j$, as follows:

$$\begin{aligned} \tilde{q}(\mathbf{W}_j) &= \mathcal{N}(\mathbf{W}_j|\tilde{\mathbf{M}}_j, \mathbf{I}_D, \tilde{\mathbf{\Omega}}_j) \\ &= g(\tilde{\mathbf{\Omega}}_j) \exp\left(-\frac{1}{2}\mathrm{tr}\left[(\mathbf{W}_j - \tilde{\mathbf{M}}_j)'(\mathbf{W}_j - \tilde{\mathbf{M}}_j)\tilde{\mathbf{\Omega}}_j^{-1}\right]\right) \end{aligned} \tag{25}$$

where

$$g(\tilde{\mathbf{\Omega}}_j) \triangleq (2\pi)^{-\frac{D(D+1)}{2}}|\tilde{\mathbf{\Omega}}_j|^{-\frac{D}{2}} \tag{26}$$

The posterior distribution also becomes a matrix normal distribution since we use a conjugate prior distribution for $\mathbf{W}_j$. From Eq (24), $\tilde{\mathbf{M}}_j$ are linearly interpolated by hyper-parameter $\mathbf{M}_j$ and the 1st order statistics of the linear regression matrix $\mathbf{Z}_j$. ρ_j controls the balance between the effects of the prior distribution and adaptation data.

3.4. Posterior distribution of latent variables

From the variational calculation of $\mathcal{F}(\mathbf{\Phi}, m)$ with respect to $q(\mathbf{V})$, we also obtain the following posterior distribution:

$$\tilde{q}(\mathbf{V}) \propto \exp\left(\langle p(\mathbf{O}, \mathbf{V}|\mathbf{W})\rangle_{q(\mathbf{W})}\right). \tag{27}$$

To obtain the above VB posteriors of latent variables, we have to consider the following integral.

$$\int q(\mathbf{W}_j) \log \mathcal{N}(\mathbf{o}_t|\mathbf{C}_k\mathbf{W}_j\boldsymbol{\xi}_k, \boldsymbol{\Sigma}_k) d\mathbf{W}_j \tag{28}$$

By substituting Eqs. (25) and (4) into Eq. (28), the equation is represented as:

$$\begin{aligned}
&\int q(\mathbf{W}_j) \log \mathcal{N}(\mathbf{o}_t|\mathbf{C}_k\mathbf{W}_j\boldsymbol{\xi}_k, \boldsymbol{\Sigma}_k) d\mathbf{W}_j \\
&= -\frac{D}{2}(2\pi)^{\frac{D(D+1)}{2}} |\tilde{\boldsymbol{\Omega}}_j|^{\frac{D}{2}} \mathrm{tr}\left[\boldsymbol{\xi}_k\boldsymbol{\xi}_k'\right] \mathrm{tr}\left[\tilde{\boldsymbol{\Omega}}_j\right] \\
&\quad - \frac{D}{2}\log(2\pi|\boldsymbol{\Sigma}_k|) - \frac{1}{2}(\mathbf{o}_t - \tilde{\boldsymbol{\mu}}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{o}_t - \tilde{\boldsymbol{\mu}}_k)
\end{aligned} \tag{29}$$

The frame-dependent term (3rd term) is equivalent to the E-step of conventional MLLR, which means that the computation time is almost the same as that of the conventional MLLR E-step.

3.5. Variational lower bound

The variational lower bound defined in Eq. (14) is decomposed as follows:

$$\begin{aligned}
&\mathcal{F}(m, \mathbf{\Psi}) \\
&= \underbrace{\left\langle \log \frac{p(\mathbf{O}, \mathbf{V}|\mathbf{W})p(\mathbf{W})}{q(\mathbf{W})} \right\rangle_{\substack{q(\mathbf{W}) \\ q(\mathbf{V})}}}_{\triangleq \mathcal{L}(m, \mathbf{\Psi})} - \langle \log q(\mathbf{V})\rangle_{q(\mathbf{V})}.
\end{aligned} \tag{30}$$

The second term, which consists of $q(\mathbf{V})$, is an entropy value and is calculated at the E-step in the VB EM algorithm. The first term ($\mathcal{L}(m, \mathbf{\Psi})$) is a logarithmic evidence term for m and $\mathbf{\Psi}$ and we can obtain an analytical solution of $\mathcal{L}(m, \mathbf{\Psi})$. Because of the conditional independence assumption in Eq. (21), $\mathcal{L}(m, \mathbf{\Psi})$ can be represented as the summation over cluster j, as follows:

$$\mathcal{L}(m, \mathbf{\Psi}) = \sum_j \left\langle \log \frac{p(\mathbf{O}, \mathbf{V}|\mathbf{W}_j)p(\mathbf{W}_j)}{q(\mathbf{W}_j)} \right\rangle_{\substack{q(\mathbf{W}_j) \\ q(\mathbf{V})}} \tag{31}$$

To derive an analytical solution, we first consider the expectation with respect to only $q(\mathbf{V})$ for cluster j. By substituting Eqs. (8), (18), and (25) into $\mathcal{L}(m, \mathbf{\Psi})$, and by using Eq (24), the expectation can be rewritten, as follows:

$$\begin{aligned}
&\left\langle \log \frac{p(\mathbf{O}, \mathbf{V}|\mathbf{W}_j)p(\mathbf{W}_j)}{q(\mathbf{W}_j)} \right\rangle_{q(\mathbf{V})} \\
&= \sum_{k\in c_j, t} \log \frac{(\mathcal{N}(\mathbf{o}_t|\mathbf{C}_k\mathbf{W}_j\boldsymbol{\xi}_k))^{\zeta_{t,k}} p(\mathbf{W}_j)}{q(\mathbf{W}_j)} \\
&= \sum_{k\in c_j} \zeta_k \log g(\boldsymbol{\Sigma}_k) + \log \frac{g(\rho_j^{-1}\mathbf{I}_{D+1})}{g(\tilde{\boldsymbol{\Omega}}_j)} \\
&\quad - \frac{1}{2}\mathrm{tr}\left[\rho_j\mathbf{M}_j'\mathbf{M}_j - \tilde{\mathbf{M}}_j'\tilde{\mathbf{M}}_j\tilde{\boldsymbol{\Omega}}_j^{-1} + \sum_{k\in c_j}\boldsymbol{\Sigma}_k^{-1}\mathbf{S}_k\right]
\end{aligned} \tag{32}$$

The obtained result does not depend on $\mathbf{W}_j$. Therefore, the expectation with respect to $q(\mathbf{W}_j)$ can be disregarded in $\mathcal{L}(m, \mathbf{\Psi})$. Consequently, we can obtain the following analytical result for the lower bound:

$$\begin{aligned}
&\mathcal{L}(m, \mathbf{\Psi}) \\
&= \sum_j \left(-\frac{D}{2}\log(2\pi)\sum_{k\in c_j}\zeta_k - \frac{1}{2}\sum_{k\in c_j}\zeta_k \log|\boldsymbol{\Sigma}_k| \right. \\
&\quad + \frac{D(D+1)}{2}\log\rho_j + \frac{D}{2}\log|\tilde{\boldsymbol{\Omega}}_j| \\
&\quad \left. - \frac{1}{2}\mathrm{tr}\left[\rho_j\mathbf{M}_j'\mathbf{M}_j - \tilde{\mathbf{M}}_j'\tilde{\mathbf{M}}_j\tilde{\boldsymbol{\Omega}}_j^{-1} + \sum_{k\in c_j}\boldsymbol{\Sigma}_k^{-1}\mathbf{S}_k\right]\right)
\end{aligned} \tag{33}$$

The first line of the obtained result corresponds to the likelihood value given the amounts of data and the covariance matrices of the Gaussians. The other terms consider the effect of the prior and posterior distributions of the model parameters. This is used as an optimization criterion with respect to model structure m and hyper-parameters $\mathbf{\Psi}$.

Note that the objective function can be represented as the summation over cluster j because of the conditional independence assumption in Eq. (21). This representation property is used for our model structure optimization in Section 4.2 within a binary tree structure representing a set of Gaussians used in the conventional MLLR.

4. OPTIMIZATION OF HYPER-PARAMETERS AND MODEL STRUCTURE

In this section, we describe how to optimize hyper-parameters $\mathbf{\Psi}$ and model structure m by using the variational lower bound as an objective function.

4.1. Structural prior setting and hyper-parameter optimization

Even though we set approximations in Eq. (17), we still have many hyper-parameters for setting ($\{\mathbf{M}_j\}_j$ and $\{\rho_j\}_j$). Therefore, we employ an efficient prior setting for $\{\mathbf{M}_j\}_j$ based on structural Bayesian approaches [4, 19]. The structural Bayesian approaches utilize a binary tree structure representing a set of Gaussians in MLLR. For example, if we focus on node (cluster) j, we use the posterior distribution parameter $\tilde{\mathbf{M}}_{(p)}$ at the parent node (p) of j as the prior distribution parameter of j, as follows:

$$\tilde{\mathbf{M}}_j = \left(\rho_j\tilde{\mathbf{M}}_{(p)} + \mathbf{Z}_j\right)\tilde{\boldsymbol{\Omega}}_j \tag{34}$$

If j is a root node, we set $(\mathbf{0}, \mathbf{I}_D)$ as the prior distribution parameters. Since the number of Gaussians included in a parent node is larger than that in a child node, the amount of adaptation data assigned to the parent node is greater than that assigned to the child node. Therefore, $\tilde{\mathbf{M}}_j$ is estimated by using reliably estimated $\tilde{\mathbf{M}}_{(p)}$ as the prior parameter.

In addition, we also optimize the other hyper-parameter ρ_j by using $\mathcal{L}(m, \mathbf{\Psi})$ as follows:

$$\tilde{\rho}_j = \underset{\rho_j}{\mathrm{argmax}}\, \mathcal{L}(m, \mathbf{\Psi}) \tag{35}$$

Thus, we remove the manual tuning of the hyper-parameters with the proposed approach. This is an advantage of the proposed approach as regards SMAPLR [4], since SMAPLR has to tune its hyper-parameters corresponding to $\{\rho_j\}_j$.

4.2. Model selection

The remaining tuning parameter in the proposed approach is how many clusters we prepare. This is a model selection problem, and we can also automatically obtain the number of clusters by optimizing the variational lower bound. In the binary tree structure, we use $j(p)$ as a parent node, and $j(c1)$ and $j(c2)$ as child nodes. Then, the objective function at node j can be defined from Eq. (33) as follows:

$$\begin{aligned}\mathcal{L}_j \triangleq & -\frac{D}{2}\log(2\pi)\sum_{k\in c_j}\zeta_k - \frac{1}{2}\sum_{k\in c_j}\zeta_k \log|\boldsymbol{\Sigma}_k| \\ & + \frac{D(D+1)}{2}\log\rho_j + \frac{D}{2}\log|\tilde{\boldsymbol{\Omega}}_j| \\ & - \frac{1}{2}\mathrm{tr}\left[\rho_j \mathbf{M}_j'\mathbf{M}_j - \tilde{\mathbf{M}}_j'\tilde{\mathbf{M}}_j\tilde{\boldsymbol{\Omega}}_j^{-1} + \sum_{k\in c_j}\boldsymbol{\Sigma}_k^{-1}\mathbf{S}_k\right].\end{aligned} \tag{36}$$

This difference function is used for a stopping criterion in the top-down clustering strategy.

We focus on node $j(p)$ in the tree, and if the node is not a leaf node, we compute the following difference between the logarithmic evidence of the parent and child nodes

$$\Delta\mathcal{L} \triangleq \mathcal{L}_{j(p)} - \mathcal{L}_{j(c1)} - \mathcal{L}_{j(c2)}, \tag{37}$$

Then, if the sign of $\Delta\mathcal{L}$ is positive, we continue the clustering for $j(c1)$ and $j(c2)$, and if it is negative, we stop the clustering at $j(p)$. By checking the signs of $\Delta\mathcal{L}$ for all possible nodes, and only using the nodes that have positive signs, we can obtain the appropriate structure of clusters that maximizes the variational lower bound. This optimization is efficiently accomplished by using a depth-first search. This approach is similar to the tree-based triphone clustering based on VB [10].

Thus, by optimizing the hyper-parameters and model structure, we can avoid setting any tuning parameters.

5. EXPERIMENTS

This section shows the effectiveness of the proposed approach by employing large vocabulary continuous speech recognition. We used a Corpus of Spontaneous Japanese (CSJ) task [20].

5.1. Experimental condition

The training data for constructing the initial (non-adapted) acoustic model consisted of 961 talks from the CSJ conference presentations (234 hours of speech data), and the training data for the language model construction consisted of 2,672 talks from the complete CSJ speech data (6.8M word transcriptions). The test set consisted of 10 talks (2.4 hours, 26,798 words). Table 1 provides acoustic and language model information [21]. We used a state-of-the-art acoustic model, which is a context-dependent model with a continuous density HMM. The HMM parameters were estimated based on a discriminative training (Minimum Classification Error: MCE) approach [22]. Lexical and language models were also obtained by employing all the CSJ speech data. We used a 3-gram model with a Good-Turing smoothing technique. The OOV rates were 2.3 % and the test set perplexities were 82.4. The acoustic model construction, LVCSR decoding, and the following acoustic model adaptation procedures were performed with the NTT speech recognition platform SOLON [23].

Table 1. Experimental setup for CSJ

Sampling rate	16 kHz
Feature type	MFCC + Energy $+\Delta + \Delta\Delta$ (39 dim.)
Frame length	25 ms
Frame shift	10 ms
Window type	Hamming
# of categories	43 phonemes
Context-dependent	5,000 HMM states (3-state left to right)
HMM topology	32 GMM components
Training method	Discriminative training (MCE)
Language model	3-gram (Good-Turing smoothing)
Vocabulary size	100,808
Perplexity	82.4
OOV rate	2.3 %

5.2. Experimental result

To check whether the proposed approach steadily increase the variational lower bound for each optimization in Section 4, Table 2 examines the values of the variational lower bound for each condition. Namely, we compare the proposed approach that optimized both model structure and hyper-parameters, as discussed in Section 4 with those did not optimize each or none of them, in terms of the $\mathcal{L}(m, \boldsymbol{\Psi})$ value. Table 2 shows that the proposed approach steadily increased the $\mathcal{L}(m, \boldsymbol{\Psi})$ value. Therefore, this result indicates that the optimization worked well by obtaining appropriate hyper-parameters and model structure.

Table 2. Variational lower bound for each optimization.

Optimization	Variational lower bound
No optimization ($\approx$ SMAPLR)	2.50E+05
Hyper-parameter	2.54E+05
Model structure	2.62E+05
Hyper-parameter and model structure	2.67E+05

Next, Figure 1 compares the proposed approach with MLLR based on the maximum likelihood estimation, and SMAPLR based on the approximate Bayesian estimation, as regards the Word Error Rate (WER) for various amounts of adaptation data. With a small amount of adaptation data, the proposed approach outperformed the conventional approaches by at most 1.0 %, while with a large amount of adaptation data, the accuracies of all approaches were comparable. This property is theoretically reasonable since the variational lower bound would be tighter than the EM-based objective function in a small amount of data, while would approach to it in a large amount of data asymptotically. Therefore, we conclude that this improvement came from the optimization of the hyper-parameters and model structure of the proposed approach, in addition to the mitigation of sparse data problem based on the Bayesian approach.

Thus, from the values of the lower bound and recognition result, we show the effectiveness of the proposed approach.

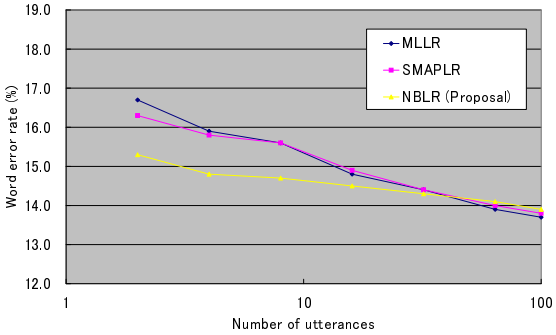

Fig. 1. Word error rate of conventional MLLR, SMAPLR, and the proposed Bayesian Linear Regression (NBLR).

6. SUMMARY

This paper realized the fully Bayesian treatment of linear regression for HMMs by using variational techniques. The derived lower bound of the marginalized log-likelihood can be used for optimizing the hyper-parameters and model structure, which was confirmed by speech recognition experiments. Future work will focus on combining modern machine learning techniques with the proposed approach (e.g., variational inference for Dirichlet process mixtures [24]).

7. REFERENCES

[1] 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.

[2] V. Digalakis, D. Ritischev, and L. Neumeyer, "Speaker adaptation using constrained reestimation of Gaussian mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 357–366, 1995.

[3] C. Chesta, O. Siohan, and C.-H. Lee, "Maximum a posteriori linear regression for hidden Markov model adaptation," in *Proc. Eurospeech1999*, 1999, vol. 1, pp. 211–214.

[4] O. Siohan, T.A. Myrvoll, and C.H. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer Speech & Language*, vol. 16, no. 1, pp. 5–24, 2002.

[5] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.

[6] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proc. Uncertainty in Artificial Intelligence (UAI) 15*, 1999.

[7] N. Ueda and Z. Ghahramani, "Bayesian model search for mixture models based on optimizing variational bounds," *Neural Networks*, vol. 15, pp. 1223–1241, 2002.

[8] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, *Application of variational Bayesian approach to speech recognition*, NIPS 2002, MIT Press, 2002.

[9] F. Valente and C. Wellekens, "Variational Bayesian GMM for speech recognition," in *Proc. Eurospeech2003*, 2003, pp. 441–444.

[10] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 365–381, 2004.

[11] P. Somervuo, "Comparison of ML, MAP, and VB based acoustic models in large vocabulary speech recognition," in *Proc. ICSLP2004*, 2004, vol. 1, pp. 830–833.

[12] T. Jitsuhiro and S. Nakamura, "Automatic generation of non-uniform HMM structures based on variational Bayesian approach," in *Proc. ICASSP2004*, 2004, vol. 1, pp. 805–808.

[13] K. Hashimoto, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "Bayesian context clustering using cross valid prior distribution for hmm-based speech recognition," in *Proc. Interspeech'08*, 2008.

[14] A. Ogawa and S. Takahashi, "Weighted distance measures for efficient reduction of gaussian mixture components in hmm-based acoustic model," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4173–4176.

[15] N. Ding and Z. Ou, "Variational nonparametric bayesian hidden markov model," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2098–2101.

[16] S. Watanabe and A. Nakamura, "Acoustic model adaptation based on coarse/fine training of transfer vectors and its application to a speaker adaptation task," in *Proc. ISLP*. Citeseer, 2004, pp. 2933–2936.

[17] K. Yu and M. J. F. Gales, "Incremental adaptation using Bayesian inference," in *Proceedings of IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP 2006)*, 2006, vol. 1, pp. 217–220.

[18] M. J. F. Gales and P. C. Woodland, "Variance compensation within the MLLR framework," Tech. Rep. 242, Cambridge University Engineering Department, 1996.

[19] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 276–287, 2001.

[20] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proceedings of LREC2000*, 2000, vol. 2, pp. 947–952.

[21] A. Nakamura, T. Oba, S. Watanabe, K. Ishizuka, M. Fujimoto, T. Hori, E. McDermott, and Y. Minami, "Evaluation of the SOLON speech recognition system : 2006 benchmark using the Corpus of Spontaneous Japanese," *IPSJ SIG Notes*, vol. 2006, no. 136, pp. 251–256, 2006, (in Japanese).

[22] E. McDermott, T.J. Hazen, J. Le Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 203–223, 2007.

[23] T. Hori, "NTT Speech recognizer with OutLook On the Next generation: SOLON," in *Proc. NTT Workshop on Communication Scene Analysis*, 2004, vol. 1, SP-6.

[24] D.M. Blei and M.I. Jordan, "Variational inference for dirichlet process mixtures," *Bayesian Analysis*, vol. 1, no. 1, pp. 121–144, 2006.