# 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

10.21437/Interspeech.2011-411

## 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-

Copyright © 2011 ISCA

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 $\Theta_T$, 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 $\Lambda_T$, :

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

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.

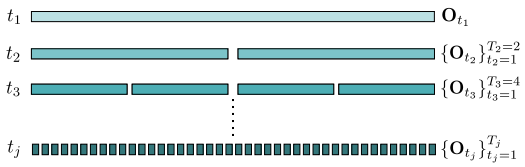


Figure 1: Example multiscale representation of feature sequences.

## 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 $k$th 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 $\phi^r$ 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 $\alpha_T$ 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:

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

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:

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

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) | $+\Delta+\Delta\Delta$ (CMS) |
| Window | Hamming |
| Frame size/shift | 25/10 ms |
| Num. of temporal HMM states | 3 (left to right) |
| Num. of phoneme categories | 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 (39 dimensions) | 12 order MFCC with energy $+\Delta+\Delta\Delta$ (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** |