MULTI-MODEL NOISE SUPPRESSION USING PARTICLE FILTERING

Takatoshi Jitsuhiro, Tomoji Toriyama, and Kiyoshi Kogure

ATR Knowledge Science Laboratories
2-2-2 Hikaridai, "Keihanna Science City," Kyoto, 619-0288, Japan
{takatoshi.jitsuhiro, toriyama, kogure}@atr.jp

ABSTRACT

We propose a noise suppression method based on multi-model compositions using particle filtering. In real environments, input speech for speech recognition includes many kinds of noise signals. For such noisy speech, we previously proposed Multi-model Noise Suppression (MM-NS) that uses many kinds of noise models and their compositions obtained from training data. However, since MM-NS only uses the static property of noise models, handling unknown noise distributions is difficult. We introduce a particle filter into MM-NS. The distributions of noise models are used as prior distributions of particle filtering to increase the accuracy of the estimation of noise signals for input data. We evaluated this method using the E-Nightingale task, which contains voice memoranda spoken by nurses during actual work at hospitals. The proposed method outperformed the original MM-NS.

Index Terms— speech recognition, noise suppression, model composition, particle filter, E-Nightingale project

1. INTRODUCTION

We have been working on the E-Nightingale Project to establish the fundamental technology for a knowledge sharing system based on understanding everyday activities and situations[1]. The project focuses on the prevention and reduction of medical malpractice in medical care domains. We have been collecting and analyzing voice memoranda recorded by nurses about their services and tasks while working. Recently, we started to evaluate the speech recognition performance of these voice memoranda. However, this recognition task is difficult because they are very noisy spontaneous speech that includes many kinds of noise signals and other voices. Now, we are working on noise suppression for speech recognition.

Many noise suppression methods have been proposed to improve the speech recognition performance of noisy speech. For stationary noise signals, Spectral Subtraction[2] and Parallel Model Combination[3] have been proposed. The Gaussian Mixture Model (GMM) based Minimum Mean-Squared Error (MMSE) method[4] assumes that input noise is stationary but fluctuating. Recently, noise suppression research has focused on non-stationary noise, including a sequential EM approach[5], a particle filtering approach[6], and so on. Since these methods usually assume that only one kind of noise signal exists, applying them to noisy speech that includes many kinds of noise signals is difficult. In general, not only stationary noise signals but also accidental noise signals occur in real environments. Furthermore, obtaining the actual noise signals from input signals is very difficult.

To solve this problem, we previously proposed Multi-model Noise Suppression (MM-NS), which includes many kinds of noise models, uses multi-pass search to find noise label sequences, and suppresses noise signals by a GMM-based MMSE method extended to multi-model compositions[7]. Although this method needs training data for noise models, it obtained better performance than the conventional method, Single-Model Noise Suppression (SM-NS). However, it is difficult for MM-NS to obtain good performance for unknown noise distributions because it only uses static models obtained from training data. MM-NS can handle unknown distributions by a large number of mixture components and weighting using posterior probabilities in the GMM-based MMSE scheme. However, this scheme is insufficient, and requires a huge number of distributions for composite models. When the number of mixture components for each noise model increases, the total number of distributions grows exponentially.

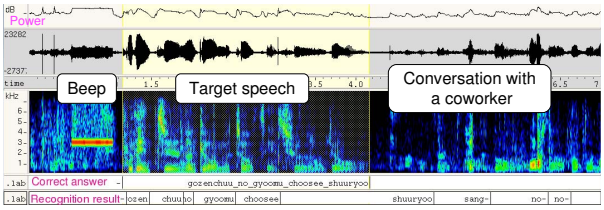

Fig. 1. Wave sample including target speech

Therefore, MM-NS needs a dynamic estimation scheme for noise signals. Using particle filtering, we propose a new noise suppression method, MM-NS, to realize non-linear noise estimation. First, MM-NS obtains the best label sequence from an input speech by multi-pass search using multi-label noise models, multi-label n-gram models, and a multi-label lexicon. Second, using labels assigned to each frame, a GMM-based MMSE extended to multi-model compositions is performed. Into this new method, a particle filter with Markov Chain Monte Carlo (MCMC) is integrated that can dynamically estimate noise signals. When new noise is detected by label recognition, particles are sampled from the model of the detected noise as prior distributions. Therefore, a particle filter can estimate the current noise distribution more precisely. Even if a noise signal is unknown, this method can estimate approximate noise distribution from preceding frames by particle filtering. Therefore, this new method includes the dynamic estimation of noise signals to solve previous problems.

The rest of our paper is organized as follows. First, in Section 2, we briefly explain our motivation, the E-Nightingale project, and its recognition task. Next, our proposed method is described in Section 3. In Section 4, we perform experiments and report results and conclude this paper in Section 5.

2. E-NIGHTINGALE PROJECT

One purpose of the E-Nightingale project is establishing technology using wearable computers and sensor networks to support nursing services[1]. To analyze daily nursing activities, we recorded and collected voice memoranda in real environments while nurses were working. We asked them to record short sentences about each nursing event using IC recorders with small microphones attached to

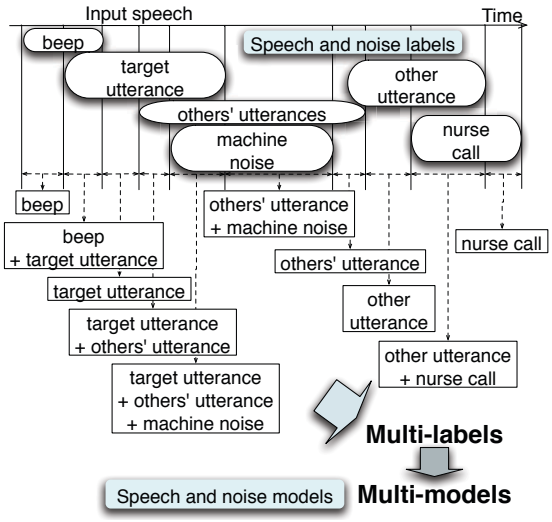

Fig. 2. Example of multi-layered labels and multi-labels

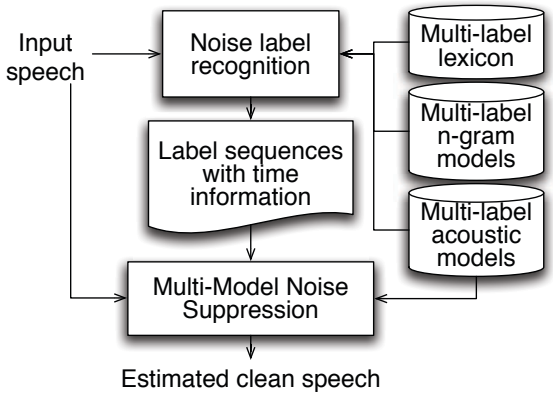

Fig. 3. Overview of Multi-model Noise Suppression

their chests. Figure 1 shows a sample of recorded speech where a nurse said, "the service adjustment meeting is finished." This sample includes a beep, a target utterance needed for analysis, conversations with a coworker, and other persons' speech as background noise. Recognizing such utterances is very difficult because many kinds of non-stationary noise signals are included, and the utterances are not so long, but they include many kinds of spontaneous speech, e.g., small and ambiguous voices with local accents. These data also include many general and essential speech recognition problems.

3. MULTI-MODEL NOISE SUPPRESSION USING PARTICLE FILTERING

3.1. Overview of MM-NS

In this section, we describe an overview of our proposed method. One key point is that it uses model compositions to represent overlapped noise signals. First, we define "multi-label" to represent the labels of composite models. To consider overlapped noise signals, we first made each speech or noise model and then combinations among them. These acoustic models are called multi-label acoustic models or "multi-models." Figure 2 shows an example of multi-layered noise labels and multi-labels.

In this paper, we used GMMs to represent them. We also made a multi-label lexicon and multi-label n-gram models from multi-label training data.

Figure 3 shows the flow of our method. Using the above models

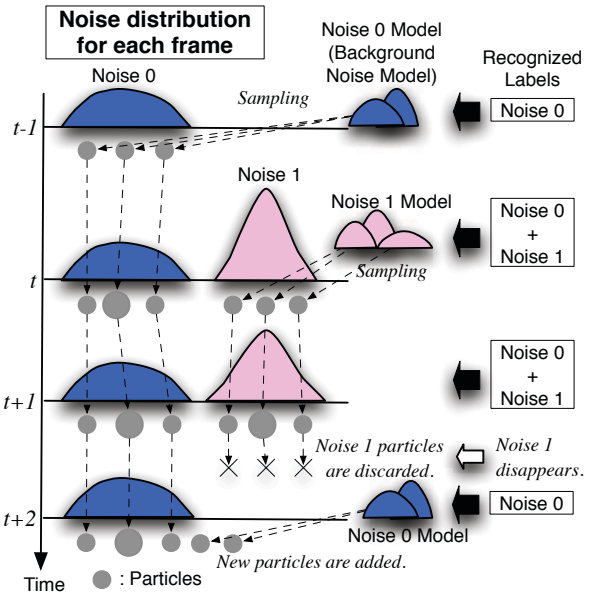

Fig. 4. Overview of particle filtering for MM-NS. This figure shows noise distributions for each frame. To represent the "Noise 0" distribution as a background noise at the first frame, particles are sampled from the "Noise 0" model. At each frame, a noise distribution are estimated by particle filtering. When "Noise 1" is detected by label recognition, particles from the "Noise 1" model are added to estimate the current "Noise 1" distribution. When "Noise 1" disappears, its particles are discarded, and new particles of "Noise 0" are added to reset particles.

and the lexicon, the best multi-label sequence is obtained by multi-pass search. After that, model-based frame-wise noise suppression is performed.

3.2. Noise suppression procedure

In the log Mel-spectral domain, when $\mathbf{x}_t, \mathbf{s}_t$, and $\mathbf{n}_t(n)$ are vector representations at the t-th frame of observed noisy speech, clean speech, and the n-th noise, respectively, the observation process of $\mathbf{x}_t$ can be written as:

$$\begin{aligned}\mathbf{x}_t &= \mathbf{s}_t + \log\left[\mathbf{I} + \exp\left\{\log\left(\sum_{n=1}^{N}\exp\left(\mathbf{n}_t(n)\right)\right) - \mathbf{s}_t\right\}\right] \\ &\quad + \mathbf{v}_t \\ &= \mathbf{s}_t + \log\left\{\mathbf{I} + \exp(\mathbf{n}_t - \mathbf{s}_t)\right\} + \mathbf{v}_t \\ &= \mathbf{s}_t + g(\mathbf{s}_t, \mathbf{n}_t) + \mathbf{v}_t \quad (1) \\ &= f(\mathbf{s}_t, \mathbf{n}_t) + \mathbf{v}_t, \quad (2) \\ \mathbf{v}_t &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\mathbf{x}), \quad (3)\end{aligned}$$

where $\mathbf{n}_t$ is composite noise including all noise at time t and $g(\mathbf{s}_t, \mathbf{n}_t)$ is a mismatch factor between clean speech $\mathbf{s}_t$ and noisy observation $\mathbf{x}_t$. $\mathbf{v}_t$ denotes error signal. $\mathbf{\Sigma}_\mathbf{x}$ is the covariance matrix of noisy speech model.

Since this approach can detect voice activity intervals, noise suppression can be done separately for each interval. We define mismatch factor $g(\mathbf{s}_t, \mathbf{n}_t)$ for each mixture component as:

$$g(\mathbf{s}_{t,l}, \mathbf{n}_{t,l}) = \begin{cases} \mu_{\mathbf{x},l} - \mu_{\mathbf{s},l} & \text{for target utterances,} \\ \mu_{\mathbf{x},l} - \varepsilon & \text{for the others,} \end{cases} \quad (4)$$

where $\mu_{\mathbf{x},l}$ is the l-th mean vector of the noisy speech model and is made by combining the l-th mean vector $\mu_{\mathbf{s},l}$ of the clean speech

model and an estimated noise model for the target utterance. ε is a small positive number that can control the power of residual signals after noise suppression.

Furthermore, we assume that the state transition process of the l-th mixture component $\mathbf{n}_{t,l}(n)$ in noise $\mathbf{n}_t(n)$ can be modeled by the following random walk process:

$$\mathbf{n}_{t+1,l}(n) = \mathbf{n}_{t,l}(n) + \mathbf{w}_t, \tag{5}$$

$$\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{n}(n),l}), \tag{6}$$

where $\mathbf{w}_t$ is the driving noise and $\mathbf{\Sigma}_{\mathbf{n}(n),l}$ is the covariance matrix of noise $\mathbf{n}_{t,l}(n)$. We define a dynamic system using Eqs. (2) and (5), and define a particle filter similar to [6].

Our proposed method used an extended Kalman particle filter with residual sampling and MCMC as did [6]. To introduce it to MM-NS, the distributions of noise models are used as priors for particles. Figure 4 shows an overview of particle filtering for MM-NS. Initially, particles for noise distributions are drawn from a background noise model "Noise 0." When a new noise, "Noise 1," is detected at time t by the noise label recognition, new particles are drawn from the "Noise 1" model. In the next frame, "Noise 1" particles are estimated by the extended Kalman filter. When "Noise 1" disappears, its particles are also discarded. At this time, new particles are sampled from "Noise 0" to reset particles.

The proposed method can especially handle accidental noise signals. Furthermore, in this method, particle pairs between clean speech and noise models must be made in particle filtering at each frame. Therefore, particles must be made from combinations of particles of clean speech and noise.

Here, we show the algorithm of the noise suppression procedure:

Multi-model Noise Suppression based on Particle Filtering:

1. Initialization
   (a) Frame index: $t = 0$
   (b) For $i = 1, \ldots, I$
   draw particles $\mathbf{n}_0^{(i)}$ from the background noise model. i is the particle's index.
2. For $t = 1, 2, \ldots, T$
   (a) Importance sampling step (particle filtering):
   i. If this frame is a speech interval,
   for $i = 1, \ldots, I$, draw speech particles $\mathbf{s}_t^{(i)}$ from the clean speech model

$$\mathbf{s}_t^{(i)} \sim \sum_{l=1}^{L_{\mathbf{s}}} w_{\mathbf{s},l} \mathcal{N}(\mu_{\mathbf{s},l}, \mathbf{\Sigma}_{\mathbf{s},l}),$$

   where $\mu_{\mathbf{s},l}$ and $\mathbf{\Sigma}_{\mathbf{s},l}$ denote the l-th mean vector and covariance matrix of the speech model, respectively, and $L_{\mathbf{s}}$ is the number of mixture components.
   ii. If new noise $\mathbf{n}(n)$ is detected,
   for $j = 1, \ldots, J$, draw the noise particles,

$$\mathbf{n}_{t-1}^{(j)} \sim \sum_{l=1}^{L_{\mathbf{n}(n)}} w_{\mathbf{n}(n),l} \mathcal{N}(\mu_{\mathbf{n}(n),l}, \mathbf{\Sigma}_{\mathbf{n}(n),l}),$$

   where j is the index of noise particle, $\mu_{\mathbf{n}(n),l}$ and $\mathbf{\Sigma}_{\mathbf{n}(n),l}$ denote the l-th mean vector and covariance matrix of the n-th noise model, respectively, and $L_{\mathbf{n}(n)}$ is the number of mixture components.
   iii. Remake particle sets for noisy speech combined between $\mathbf{s}_t^{(i)}$ and $\mathbf{n}_{t-1}^{(j)}$. Total number of particles: $K = I \times (J + 1)$
   iv. For $k = 1, \ldots, K$
   update the particles with the extended Kalman filter. Estimate $\hat{\mathbf{n}}_t^{(k)}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{n}_t}^{(k)}$.
   v. For $k = 1, \ldots, K$, $w_t^{(k)} \propto w_{t-1}^{(k)} p(\mathbf{x}_t | \mathbf{n}_t^{(k)})$.
   vi. For $k = 1, \ldots, K$,
   obtain normalized weights, $\hat{w}_t^{(k)} = w_t^{(k)} / \sum_{k=1}^{K} w_t^{(k)}$.
   (b) Residual sampling step:
   Multiply/suppress particles with high/low importance weights, respectively.
   (c) MCMC step:
   Apply Metropolis-Hastings sampling.
   (d) Estimation of a noise posterior distribution:
   Obtain a posterior distribution from particles.

$$p(\mathbf{n}_{0:t} | \mathbf{x}_{0:t}) \simeq \sum_{k=1}^{K} \hat{w}_t^{(k)} p(\mathbf{n}_{0:t}^{(k)} | \mathbf{x}_{0:t}) = \mathcal{N}(\mu_{\hat{\mathbf{n}}_t}, \mathbf{\Sigma}_{\hat{\mathbf{n}}_t}),$$

   where $\mu_{\hat{\mathbf{n}}_t}$ and $\mathbf{\Sigma}_{\hat{\mathbf{n}}_t}$ are the estimated mean vector and covariance matrix of the noise model, respectively.
   (e) GMM-based MMSE estimation of clean speech: Obtain estimated clean speech using the mismatch factor.

$$\hat{\mathbf{s}}_t = \mathbf{x}_t - \sum_{l=1}^{L_{\mathbf{s}}} P(l | \mathbf{x}_t) g(\mathbf{s}_{t,l}, \hat{\mathbf{n}}_{t,l}).$$

   $P(l|\mathbf{x}_t)$ is given by

$$P(l | \mathbf{x}_t) = \frac{w_{\mathbf{s},l} \mathcal{N}(\mathbf{x}_t; \mu_{\mathbf{x},l}, \mathbf{\Sigma}_{\mathbf{x},l})}{\sum_m^{L_{\mathbf{s}}} w_{\mathbf{s},m} \mathcal{N}(\mathbf{x}_t; \mu_{\mathbf{x},m}, \mathbf{\Sigma}_{\mathbf{x},m})},$$

   where $\mu_{\mathbf{x},l}$ and $\mathbf{\Sigma}_{\mathbf{x},l}$ denote the mean vector and the covariance matrix of the noisy speech model, respectively. They are estimated from the clean speech model and the estimated noise model $\mathcal{N}(\mu_{\hat{\mathbf{n}}_t}, \mathbf{\Sigma}_{\hat{\mathbf{n}}_t})$ by applying the first order Taylor series expansion[8].

This method can estimate the current noise distribution: that is, only a single distribution is obtained. Although the original MM-NS required the calculation of mixture components for noise models, this proposed method does not. Therefore, its GMM-based MMSE estimation of clean speech is easier than the original MM-NS.

4. EXPERIMENTS

4.1. Experimental setup

Our experimental conditions are identical as [7], except for the conditions of the particle filters. The E-Nightingale data were recorded in a Japanese hospital. The data collected on the first day were used for evaluation. The length of each file was 10 sec including one target utterance. The data from the second day were used as training data to adapt the acoustic models to speakers and to create noise GMMs for noise suppression. In this paper, diagonal covariance matrices were used for all distributions. Table 1 shows the details of the experimental conditions. Test data included 208 utterances with 1,051 words spoken by eight speakers who were selected as ordinary speakers and included both in the test and adaptation data.

For noise suppression, an HTK version 3.3 was used to extract feature parameters and train GMMs. 24-order outputs of log Mel-filter bank "FBANK" were used as feature parameters. We used MFCC models converted from FBANK models for noise label recognition. Speaker-adapted GMMs were used as clean speech models. The remaining speech and noise models were generated as GMMs with four, eight, and 12 mixture components. In this training data, 32 kinds of noise models were obtained including a target speech model. The total number of models including the composite models was 194. In this data, the number of models combined into one model was two, or three. Furthermore, estimated background noise for each input speech was combined into all models when noise suppression was performed.

Table 1. Experimental conditions

Common conditions	
Analysis conditions	16kHz sampling rate, 16 bit Frame shift: 10 ms, frame length: 20 ms
Test data	8 females (208 utterances, 1,051 words) Average SNR: 8.25 dB
Noise label recognition & noise suppression	
Tools	HTK Ver. 3.3 (for GMM training) ATRASR Ver. 3.6 (for decoding)
Feature parameters	24 Mel-filter bank (for noise suppression) 12 MFCC and 0th MFCC (for search)
Acoustic models	32 basic GMMs (speech and noise) Training data: about one hour 162 composite models
	Clean speech: Speaker-adapted GMM about 200 mixture components
Language models	Multi-label bigram, multi-label trigram Training data: 354 utterances
Lexicon	194 multi-labels
Particles	Original PF: 300 particles PF-MM-NS: 110 particles
Speech recognition	
Tools	ATRASR Ver.3.6
Feature parameters	12 MFCC, 12 ΔMFCC, Δlog power Cepstral Mean Subtraction (CMS)
AMs	2,086 states with five mixture components Speaker-adapted models
AM Training DB	Topology training data: 37 hours Re-training DB: 21 hours (female only)
LMs	Word bigram, word trigram Out of Vocabulary (OOV) rate: 2.36%
LM Training DB	E-Nightingale data: Nine days, 9,936 utterances
Lexicon	2,636 words

For particle filters, 300 particles were used for the original particle filter method[6], and 110 particles were used for our proposed method. Both Real Time Factors of noise suppression processing of these methods were about 2.5 by Intel Pentium-D 3.2 GHz.

As a speech recognizer and a training tool, we used the ATRASR large-vocabulary speech recognition system version 3.6 developed by the ATR Spoken Language Communication Labs. An acoustic model for word recognition was generated by MDL-SSS[9]. MAP-VFS[10] was used as the speaker-adaptation method.

4.2. Experimental results

Figure 5 shows word accuracy rates for several methods. "SM-NS" is a Single-Model Noise Suppression method using one distribution for a noise model. "PF" denotes the original particle filter-based method[6]. This method could not obtain better performance in this task than the baseline because it made many insertion errors, and tracking accidental noise is difficult. "PF-MM-NS (4 mix.)," "PF-MM-NS (8 mix.)," and "PF-MM-NS (12 mix.)" are particle filtering methods based on MM-NS using noise models with four, eight, and 12 mixture components, respectively. These patterns obtained better performance than the MM-NS with identical noise models. These results show that our proposed method can handle accidental noise signals using noise models and particle filtering.

5. CONCLUSION

We proposed Multi-model Noise Suppression (MM-NS) based on particle filtering. This method can handle many kinds of noise signals including speech data recorded in real environments. Furthermore, it can estimate unknown noise distribution using a particle filter with Markov Chain Monte Carlo. It obtained better performance than the original MM-NS with identical noise models.

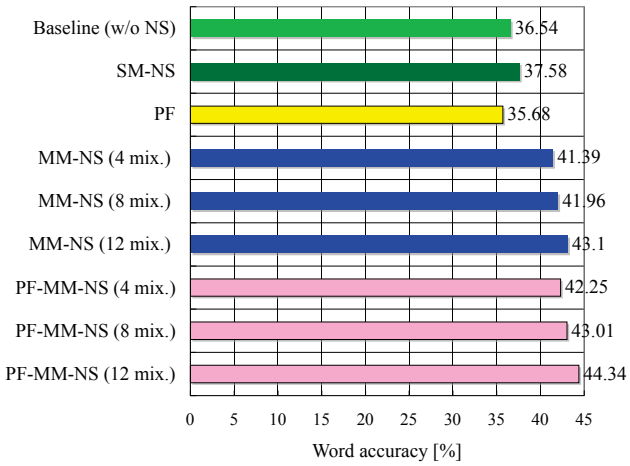

Fig. 5. Word accuracy

6. ACKNOWLEDGEMENTS

This research was supported by the National Institute of Information and Communications Technology of Japan. We'd like to thank all the nurses for their cooperation, the ATR-SLC members for their tools, and the ATR-KSL members for their collaboration.

7. REFERENCES

[1] K. Kogure, "Toward a knowledge sharing system based on understanding everyday activities and situations – Introduction to the E-Nightingale project –," in *Proc. the Workshop on Knowledge Sharing for Everyday Life 2006 (KSEL2006)*, 2006, pp. 1–8.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 27, pp. 113–120, 1979.

[3] M. F. J. Gales, *Model-based techniques for noise robust speech recognition*, Ph.D. thesis, University of Cambridge, 1995.

[4] J. C. Segura, A. de la Torre, M. C. Benitez, and A. M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II database and tasks," in *Proc. EUROSPEECH2001*, 2001, vol. 1, pp. 221–224.

[5] K. Yao, K. K. Paliwal, and S. Nakamura, "Noise adaptive speech recognition based on sequential noise parameter estimation," *Speech Communication*, vol. 42, no. 1, pp. 5–23, 2004.

[6] M. Fujimoto and S. Nakamura, "A non-stationary noise suppression method based on particle filtering and Polyak averaging," *IEICE Trans. Inf. & Syst.*, vol. E89-D, no. 3, pp. 922–930, 2006.

[7] T. Jitsuhiro, T. Toriyama, and K. Kogure, "Robust speech recognition using noise suppression based on multiple composite models and multi-pass search," in *Proc. ASRU2007*, 2007, to apear.

[8] P. J. Moreno, *Speech recognition in noisy environments*, Ph.D. thesis, Carnegie Mellon University, 1996.

[9] T. Jitsuhiro, T. Matsui, and S. Nakamura, "Automatic generation of non-uniform HMM topologies based on the MDL criterion," *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 8, pp. 2121–2129, 2004.

[10] M. Tonomura, T. Kosaka, and S. Matsunaga, "Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation," *Computer Speech and Language*, vol. 10, pp. 117–132, 1996.