PAPER Special Section on Processing Natural Speech Variability for Improved Verbal Human-Computer Interaction

Speech Recognition under Multiple Noise Environment Based on Multi-Mixture HMM and Weight Optimization by the Aspect Model

Seong-Jun HAHM[†a)], ***Student Member***, **Yuichi OHKAWA**[††], **Masashi ITO**[†], **Motoyuki SUZUKI**[†††], **Akinori ITO**[†], ***Members***, ***and*** **Shozo MAKINO**[†], ***Fellow***

SUMMARY In this paper, we propose an acoustic model that is robust to multiple noise environments, as well as a method for adapting the acoustic model to an environment to improve the model. The model is called "the multi-mixture model," which is based on a mixture of different HMMs each of which is trained using speech under different noise conditions. Speech recognition experiments showed that the proposed model performs better than the conventional multi-condition model. The method for adaptation is based on the aspect model, which is a "mixture-of-mixture" model. To realize adaptation using extremely small amount of adaptation data (i.e., a few seconds), we train a small number of mixture models, which can be interpreted as models for "clusters" of noise environments. Then, the models are mixed using weights, which are determined according to the adaptation data. The experimental results showed that the adaptation based on the aspect model improved the word accuracy in a heavy noise environment and showed no performance deterioration for all noise conditions, while the conventional methods either did not improve the performance or showed both improvement and degradation of recognition performance according to noise conditions.

key words: *multi-mixture HMM, noise-independent acoustic model, aspect model, speech recognition in noisy environment*

1. Introduction

Background noise has always posed a serious problem in speech recognition systems. Many methods have been proposed for solving background noise problems [1]; these methods can be classified into three types: recording-based [2]–[4], analysis-based [5]–[8], and model-based methods [9]–[11]. The recording-based method utilizes multiple input signals, such as signals from a microphone array, to emphasize speech signals. These emphasized signals are used as the input for the hidden Markov model (HMM). The analysis-based method utilizes a special representation of speech signals that is robust against additive noises. The spectral subtraction method [5] and the perceptual linear predictive analysis [6] are the widely used algorithms belonging to this category. The model-based method utilizes the HMM that is not trained with clean speech, but is directly trained with a mixture of speech and environmental noise signals. The recording-, analysis-, and model-based methods for noise-robust speech recognition are complementary; these three methods can be combined together for improving the robustness in a noisy environment [12]–[14].

In this study, we focus on the model-based method for noise-robust speech recognition in various noise environments. A multi-condition HMM (MC-HMM) can be used in the model-based approach [11]. This method trains an HMM using training data that contains speech signals corrupted by various environmental noises. The trained HMM is known to be robust against various types of noises used in the training [11]. However, the recognition accuracy of the MC-HMM is not as high as that of an HMM trained using speech signals in the matched noise environment.

A noise adaptation technique, which tunes an HMM to recognize speech in a specific noise environment using small amount of data obtained in the target environment, is effective in solving this problem. The maximum likelihood linear regression (MLLR) [16], a famous algorithm used for speaker adaptation, can be easily applied for noise adaptation. Although a combination of the MC-HMM and the MLLR is effective for speech recognition in noisy environments, it requires a large amount of adaptation data (more than 10 sentences) for achieving sufficient performance [18]. Another noise adaptation approach is based on tree-structured clustering method [21]. In this method, tree-structured clustering [17] is performed on various noise and signal-to-noise ratio (SNR) conditions. Then, based on the ML criterion, the HMM that best matches the input speech was selected by tracing the tree from top to bottom. Furthermore, MLLR adaptation is performed using the selected HMM to reduce mismatches with the input speech.

In this paper, we propose a new acoustic model called a multi-mixture HMM (MM-HMM), and its adaptation technique, termed as an aspect model, for speech recognition in noisy environments. These models have high recognition accuracy even when a very small amount (around 1 s) of observed data are used.

This paper is organized as follows. In Sect. 2, we introduce the MM-HMM, which improves the performance of MC-HMM. In Sect. 3, we provide an overview of the aspect model approach and discuss the potential of the technique

Manuscript received December 2, 2009.
Manuscript revised February 23, 2010.

[†]The authors are with the Graduate School of Engineering, Tohoku University, Sendai-shi, 980–8579 Japan.
[††]The author is with the Graduate School of Educational Informatics, Tohoku University, Sendai-shi, 980–8579 Japan.
[†††]The author is with Institute of Technology and Science, The University of Tokushima, Tokushima-shi, 770–8501 Japan
a) E-mail: branden65@makino.ecei.tohoku.ac.jp

DOI: 10.1587/transinf.E93.D.2407

using the aspect model.

2. MM-HMM for Multi-Condition Training

In this section, we propose an MM-HMM, which is an improved version of the MC-HMM. The problem with the MC-HMM is that it is difficult to train a large number of parameters for Gaussian mixtures of HMM states using the EM algorithm. Using the proposed method, we can use a large number of Gaussian mixture components without suffering from parameter estimation.

2.1 MC-HMM

Lipmann et al. proposed a method for training an HMM that is robust to various noise environments [11]. In this method, various types of noises are used for training. First, corrupted speech data are obtained for each background noise, and then, all the data are used for training. The resulting model contains all variations of speakers and environments; therefore, it is expected to be robust to variations in both speaker and environment.

This type of trained HMM is called MC-HMM. While the MC-HMM is quite simple, it is known to be robust against various noises. Therefore, this method is regarded as a "standard" for the noise-robust acoustic model [15].

The disadvantage of the MC-HMM is that it is difficult to train models with a large number of parameters. Since the variation in environmental noises is considerably wider than that in speakers, a model for various speakers and noise environments should have considerably larger number of parameters than an HMM in a single environment. However, it is difficult to train a model with a large number of parameters because the solution of the EM algorithm tends to converge a local maximum when the number of parameters is large.

2.2 Meaning of Gaussian Mixture Distribution

In most of the continuous density HMMs, a Gaussian mixture distribution is employed as a probability density function of a state. A Gaussian mixture distribution is expressed as follows:

$$p(\boldsymbol{x}) = \sum_{k=1}^{M} \eta_k N(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{1}$$

where

$$\sum_{k=1}^{M} \eta_k = 1. \tag{2}$$

Here, $\boldsymbol{x}$ is a feature vector, M is the number of mixture, $N(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate Gaussian density function, $\boldsymbol{\mu}_k$ is a mean vector of the k-th distribution, and $\boldsymbol{\Sigma}_k$ is a covariance matrix of the k-th distribution.

In HMMs used for speech recognition in a certain noise environment (including the case of "clean" environment), using Gaussian mixture distributions is regarded as a method of approximating the distributions of feature vectors. Since the shape of the true distribution of feature vectors is unknown, we employ a Gaussian mixture distribution that can express various types of distributions and adjust the parameters of the distribution using the EM algorithm so that the mixture distribution approximates the true distribution of the feature vectors. We call this type of Gaussian mixture distribution as the "mixture for approximation."

The other interpretation of a mixture distribution is that a feature vector is generated from a different information source. Suppose a feature vector belongs to a distinct noise environment, while we do not know which environment the vector belongs to. If a distribution of vectors in a certain environment is approximated by a Gaussian distribution, the distribution of the observed vectors can be expressed as a mixture of distributions of all environments.

$$p(\boldsymbol{x}) = \sum_{j=1}^{N} \gamma_j N(\boldsymbol{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j). \tag{3}$$

In this case, N is the number of environments, $N(\boldsymbol{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ is a distribution of the j-th environment, and γ_j is a prior probability of the j-th environment. We call this type of mixture distribution as the "mixture for alternatives." The difference between the "mixture for alternatives" and the "mixture for approximation" is that each component of the mixture distribution for alternatives has a different meaning (i.e., a specific environment). In a mixture distribution for alternatives, if we know the environment j_0 to which the vectors belong, we can adjust γ_j as

$$\gamma_j = \begin{cases} 1 & \text{if} \quad j = j_0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

for maximizing expectations of the probability $p(\boldsymbol{x})$, but this kind of adjustment does not make sense for a mixture for approximation because a Gaussian component of the mixture distribution does not have any specific meaning.

2.3 MM-HMM

In an MC-HMM, all variations in feature vectors attributed to phoneme environments, speaker variations and noise environments are jointly expressed as a Gaussian mixture distribution. Here, we assume that different noise environments are "alternatives" explained above. Then, we can decompose a Gaussian mixture distribution into "mixture for alternatives" and "mixture for approximation," as follows:

$$p(\boldsymbol{x}) = \sum_{j=1}^{N} \gamma_j \psi_j(\boldsymbol{x}) \tag{5}$$

$$= \sum_{j=1}^{N} \gamma_j \sum_{k=1}^{M} \eta_{j,k} N(\boldsymbol{x}; \boldsymbol{\mu}_{j,k}, \boldsymbol{\Sigma}_{j,k}), \tag{6}$$

where γ_j is a prior probability of the j-th environment and

$\psi_j(\boldsymbol{x})$ is a probability density of feature vectors in the j-th environment, which is expressed as a mixture of M Gaussian components.

If we have no prior knowledge of the noise environment, the distribution of γ_j is assumed to be uniform. Thus, the distribution becomes

$$p(\boldsymbol{x}) = \sum_{j=1}^{N} \frac{1}{N} \psi_j(\boldsymbol{x}) \tag{7}$$

$$= \sum_{j=1}^{N} \frac{1}{N} \sum_{k=1}^{M} \eta_{j,k} N(\boldsymbol{x}; \boldsymbol{\mu}_{j,k}, \boldsymbol{\Sigma}_{j,k}). \tag{8}$$

The MM-HMM employs the above-mentioned mixture distribution as probability density functions of all states. The resulting model is quite simple: just train HMMs for all noise environments individually, and then, combine all HMMs by mixing all distribution functions state by state.

The construction procedure of the MM-HMM is as follows:

1. *Training of each single-noise HMM for each noise*:
   For each noise used for training $(\varphi_1, \varphi_2, \cdots \varphi_N)$, a single-noise HMM $\mathcal{M}^{(\varphi_j)}$ is trained using the speech data corrupted by the noise φ_j. All single-noise HMMs have the same topology (left-to-right HMM) with the same number of states. Each state has a mixture Gaussian distribution with the same number of Gaussian components.
2. *Combining all Gaussian components of every single-noise HMM into unified HMM*:
   The HMM $\mathcal{M}^{(\Phi)}$ is constructed by combining all Gaussian components of single-noise HMMs $\mathcal{M}^{(\varphi_j)}$, where Φ denotes a set of noises ($\Phi = \{\varphi_1, \varphi_2, \cdots, \varphi_N\}$).
   The output probability density distribution $p_s(\boldsymbol{x})$ at the s-th state in $\mathcal{M}^{(\Phi)}$ is given by Eq. (9).

$$p_s(\boldsymbol{x}) = \frac{1}{N} \sum_{j=1}^{N} \psi_{j,s}(\boldsymbol{x}), \tag{9}$$

   where $\psi_{j,s}(\boldsymbol{x})$ denotes an output probability density distribution at the s-th state in $\mathcal{M}^{(\varphi_j)}$.

The MM-HMM is expected to be robust to noise environments included in the training data. Moreover, it is also expected to be robust to an "unknown" environment that is not included in the training data, as long as the distribution of feature vectors in the unknown environment is approximated by a linear combination of distributions of the environments used for training. Of course, it is not always true, but we can expect this assumption to hold when many noisy environments are used for the training.

2.4 Experiments

In order to confirm the effectiveness of the MM-HMMs, several speech recognition experiments were carried out. The recognition performance of MM-HMMs, clean HMMs and MC-HMMs was compared.

Table 1 Noise data used in the experiments.

Purpose	Tape No.	Noise Types
Training	(4)	1. Exhibition hall
	(5)	2. Station (concourse)
	(7)	3. Plant Machinery
	(7)	4. Metal factory
	(8)	5. Sorting field
	(9)	6. Highway
	(9)	7. Highway intersection
	(10)	8. Crowded street
	(12)	9. Train (local trains)
	(13)	10. Computer Room (medium)
	(16)	11. Noise of the fan coil air conditioning
	(17)	12. Hall elevator (department stores)
Testing & Adaptation	(3)	13. Exhibition hall
	(8)	14. Coating Room
	(10)	15. Crowded street
	(14)	16. Computer Room (workstation)
	(17)	17. Hall elevator (office)

Seventeen types of background noises shown in Table 1 were used for the experiments. Twelve noises were used as training data, and the other five noises were used as test data. Note that the noise "Exhibition hall" and "Crowded street" are involved in both of the training and test data; however, these noise signals were recorded at different places. Therefore, we regard these noise signals as belonging to different environments.

The training signals were generated by combining clean speech signals and noise signals with four SNR, 5, 10, 15 and 20 dB. The test signals were generated similarly, with SNR conditions of 0, 5, 10, 15, 20, 25 and 30 dB. The clean speech signals were also used as test data.

All single-noise HMMs and MC-HMM consisted of tied-state triphones. The structure of tied states was automatically determined by the decision tree method, and the structures of all HMMs were determined separately. A state in the MM-HMM is constructed by combining the states in the single-noise HMMs. When combining the Gaussian mixture, a uniform distribution is used (i.e., $\gamma_j = 1/N$). 13-dimensional Mel-frequency cepstral coefficient (MFCCs) feature vectors excluding the frame log power were extracted from the pre-emphasized speech signal every 10 ms using a 25 ms Hamming window. The MFCCs and ΔMFCCs were concatenated to form 25-dimensional feature vectors. Cepstral mean normalization was used. The performance was evaluated using word accuracy. The other experimental conditions are shown in Table 2.

The speech recognition experiments for various SNR conditions were carried out. The MM-HMM was constructed from 48 single-noise HMMs (12 noise variations $\times$ 4 SNR variations), with each state having 768 ($16 \times 12 \times 4$) mixtures. The MC-HMMs were trained using all available training samples. In the test data, SNR was set to 0, 5, 10, 15, 20, 25, 30, and ∞ dB.

Table 2 Experimental conditions.

Speech database	ATR speech dialogue database for spontaneous speech recognition [24]
Noise database	JEIDA noise database [25]
Training data	1,500 sentences uttered by 1,070 speakers
Test data	100 sentences uttered by 20 speakers (5 for each speaker, 10 males and 10 females)
No. of components in a state	
clean HMM	16, 32, 64
MM-HMM	768 ($16 \times 12 \times 4$)
MC-HMM	16, 32, 64, 128
Language model	Trigram trained by transcription of speech database except test sentences
Vocabulary size	7,000
Decoder	Julius [26]

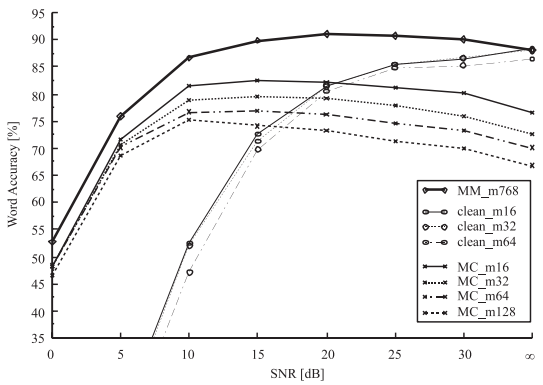


Fig. 1 Word accuracy of MC model, clean speech model, and MM-HMM.

Figure 1 shows the experimental results. In the figures, "MM" denotes "MM-HMM"; "clean", "clean HMM"; "MC", "MC-HMM", and the number after "m", the number of mixtures. These results show that the MM-HMM has the highest accuracy for all SNRs. Among the MC-HMMs, the best result was obtained when the number of mixture was 16, showing that it was difficult to train all variations from noisy environments, speakers, and other fluctuations altogether using the EM algorithm.

3. Adjustment of Prior Probabilities Using an Aspect Model

3.1 Overview of Adjustment of Prior Probabilities

In Eq. (8), we assumed that we have no knowledge of the noise environment. If we observe noise or noise-added speech in the target environment, we can adjust the prior probability γ_j so that the overall probability increases. This approach is similar to noise environment adaptation [18], [19]. The basic idea is to change γ_j using the EM algorithm so that the likelihood of the adaptation data becomes maximum. A straightforward method of adjusting the prior probabilities is to adjust γ_i directly. However, if the amount of adaptation data is extremely small (i.e., a few seconds), the estimation of parameters become unstable.

A conventional approach of reducing the number of parameters to be adapted is to use the clustering technique for noisy speech [20], [21]. In this type of approach, the noise environments are classified (either strictly by using ordinary clustering or softly using a fuzzy clustering) into a few clusters, and the adaptation is performed on the basis of the model for the nearest cluster.

A disadvantage of the conventional approaches is that the adaptation result is not guaranteed to be optimum from the maximum likelihood point of view, because clustering and adaptation are performed independently. To obtain the optimum result for both clustering and adaptation, we employ an aspect model [22] for reducing the number of parameters to be adjusted.

The probability distribution function for the sample $\boldsymbol{x}$ is as follows. First, we consider the adaptation of a distribution of a specific state. In this case, the probability distribution is expressed as

$$p(\boldsymbol{x}|\Xi, \Lambda) = \sum_{z=1}^{Z} \xi_z p(\boldsymbol{x}|\Lambda_z) \tag{10}$$

$$= \sum_{z=1}^{Z} \xi_z \sum_{n=1}^{N} \lambda_{n,z} \psi_n(\boldsymbol{x}) \tag{11}$$

where

$$\Xi = \{\xi_1, \ldots, \xi_Z\}, \tag{12}$$

$$\Lambda_z = \{\lambda_{1,z}, \ldots, \lambda_{N,z}\}. \tag{13}$$

$$\Lambda = \{\Lambda_1, \ldots, \Lambda_Z\} \tag{14}$$

$$\psi_n(\boldsymbol{x}) = \sum_{k=0}^{M} \eta_k N(\boldsymbol{x}; \boldsymbol{\mu}_{n,k}, \boldsymbol{\Sigma}_{n,k}). \tag{15}$$

The variable $\lambda_{n,z}$ is the first-level weighting and ξ_z is the second-level weighting of the z-th cluster.

This model is interpreted as follows. First, all noise environments are "softly" classified into a few clusters (i.e., *aspect model*). Here, $p(\boldsymbol{x}|\Lambda_z)$ denotes a probability density function of samples in the z-th aspect model. Then, the probabilities from all aspect models are combined using the weight ξ_z. Under this interpretation, $\lambda_{n,z}$ denotes the degree of belongingness of the n-th environment belongs to the z-th cluster, and ξ_z denotes the probability that the current environment belongs to the z-th cluster.

Comparing Eq. (10) with Eq. (6), we obtain

$$\gamma_j = \sum_{z} \xi_z \lambda_{j,z}. \tag{16}$$

When adjusting the probability γ_j, we adjust only ξ_z instead of adjusting γ_j, because $\lambda_{j,z}$ (which denotes the probability of the j-th environment belonging to the z-th aspect model) is independent of the current environment. If $Z < N$, the estimation of parameters from a small amount of adaptation data becomes easier than simply adjusting all of γ_j.

3.2 Training Aspect Model

To estimate Ξ and Λ, we have to make some kind of assumption regarding ξ_z and $\lambda_{j,z}$ so that each $p(\boldsymbol{x}|\Lambda_z)$ becomes a "basis distribution" for expressing distributions of various noise environments. The basic idea of training of the aspect models is to estimate Λ so that mixture of $p(\boldsymbol{x}|\Lambda_z)$ can be used to approximate each of the single-noise models $\psi_j(\boldsymbol{x})$. Since the number of the aspect models Z is smaller than that of the single-noise models N, we can expect that the trained aspect models are trained to express any distribution. The optimization of basis distributions is based on the maximum likelihood criterion, which is the advantage of the proposed method.

We define the probability distribution function for the noise j and sample $\boldsymbol{x}$ as

$$p(\boldsymbol{x}|\Xi_j, \Lambda) = \sum_{z=1}^{Z} \xi_{z,j} \sum_{n=1}^{N} \lambda_{n,z} \psi_n(\boldsymbol{x}), \tag{17}$$

where

$$\Xi_j = \left\{\xi_{1,j}, \ldots, \xi_{Z,j}\right\}. \tag{18}$$

Let $\boldsymbol{X}_j$ be a set of samples that belong to the j-th noise environment:

$$\boldsymbol{X}_j = \{\boldsymbol{x}_{i,j}\} \tag{19}$$

Then, the objective function of the training of the aspect models is the total likelihood of the samples, given by

$$L(\Xi_1, \ldots, \Xi_N, \Lambda) = \sum_{j=1}^{N} \sum_{i} \log p(\boldsymbol{x}_{i,j}|\Xi_j, \Lambda). \tag{20}$$

The EM algorithm for maximizing Eq. (20) is as follows. First, we randomly initialize $\xi_{z,j}$ and $\lambda_{n,z}$. Next, we define $\alpha_{i,j,z}$ and $\beta_{i,j,n}$ as

$$\alpha_{i,j,z} = \frac{\xi_{z,j} \sum_n \lambda_{n,z} \psi_n(\boldsymbol{x}_{i,j})}{\sum_z \xi_{z,j} \sum_n \lambda_{n,z} \psi_n(\boldsymbol{x}_{i,j})} \tag{21}$$

$$\beta_{i,j,n} = \frac{\sum_z \xi_{z,j} \lambda_{n,z} \psi_n(\boldsymbol{x}_{i,j})}{\sum_n \sum_z \xi_{z,j} \lambda_{n,z} \psi_n(\boldsymbol{x}_{i,j})}. \tag{22}$$

Then, we re-estimate $\lambda_{n,z}$ and $\xi_{z,j}$ as

$$\lambda'_{n,z} = \frac{\sum_j \sum_i \alpha_{i,j,z} \beta_{i,j,n}}{\sum_n \sum_j \sum_i \alpha_{i,j,z} \beta_{i,j,n}}, \tag{23}$$

$$\xi'_{z,j} = \frac{\sum_i \alpha_{i,j,z}}{\sum_z \sum_i \alpha_{i,j,z}}. \tag{24}$$

After training $\lambda_{n,z}$ and $\xi_{z,j}$, only $\lambda_{n,z}$ are remaining for the calculation of the aspect models. $\xi_{z,j}$ are not used for the adaptation. The average of $\xi_{z,j}$ over j are used as initial values of the adaptation.

$$\bar{\xi}_z^{(0)} = \frac{1}{N} \sum_{j=1}^{N} \xi_{z,j}. \tag{25}$$

When using the aspect model for the HMM, we have to apply the above-mentioned method to distributions in many states. In this case, we use state-dependent $\lambda_{n,z}$ (i.e., $\lambda_{n,z,s}$) and state-independent $\xi_{z,j}$. The reason behind using state-dependent $\lambda_{n,z,s}$ is that effect of environmental noise on speech differs from phoneme to phoneme. For example, when the noise level is not high, vowels are not affected by the noise strongly because they have large power, while plosive consonants are affected by the noise more strongly. The state-dependent $\lambda_{n,z,s}$ can express this kind of difference. Note that the state-dependency of $\lambda_{n,z,s}$ is independent of the parameters to be adapted for environment adaptation, because we adjust only ξ_z for the adaptation. To estimate these parameters, $\alpha_{i,j,z}$ and $\beta_{i,j,n}$ are also made state dependent (i.e., $\alpha_{i,j,z,s}$ and $\beta_{i,j,n,s}$). Therefore, Eq. (11) is rewritten as follows:

$$p(\boldsymbol{x}|\Xi, \Lambda_s) = \sum_z \xi_z \sum_n \lambda_{n,z,s} \psi_{n,s}(\boldsymbol{x}), \tag{26}$$

where

$$\Lambda_s = \{\Lambda_{1,s}, \ldots, \Lambda_{Z,s}\} \tag{27}$$

$$\Lambda_{z,s} = \{\lambda_{1,z,s}, \ldots, \lambda_{N,z,s}\} \tag{28}$$

$$\psi_{n,s}(\boldsymbol{x}) = \sum_{k=0}^{M} \eta_{s,k} N(\boldsymbol{x}; \boldsymbol{\mu}_{n,s,k}, \boldsymbol{\Sigma}_{n,s,k}) \tag{29}$$

In this case, $\lambda_{n,z,s}$ is a state-dependent weight from n to the aspect model z at the state s, whereas ξ_z is still independent of states. The mixture weights $\lambda_{n,z,s}$ from the training samples are trained using the EM algorithm.

3.3 Online Adaptation Using Aspect Model

For the adaptation of the aspect model, the EM algorithm is applied for estimating $\bar{\xi}_z$, which is the updated ξ_z. When the adaptation data $y_1, y_2, \ldots, y_n$ are given, $\bar{\xi}_z$ is calculated as

$$\bar{\xi}_z^{(k+1)} = \frac{\sum_s \sum_i \bar{\xi}_z^{(k)} \sum_n \lambda_{n,z,s} \psi_{n,s}(y_i^{(s)})}{\sum_z \sum_s \sum_i \bar{\xi}_z^{(k)} \sum_n \lambda_{n,z,s} \psi_{n,s}(y_i^{(s)})}, \tag{30}$$

where k is the number of iterations and

$$\psi_{n,s}(y_i^{(s)}) = \begin{cases} \psi_{n,s}(y_i) & \text{if } y_i \text{ belongs to state } s \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

After estimating $\bar{\xi}_z$, we obtain a mixture model adapted to the data as

$$p(x|\bar{\Xi}, \Lambda_s) = \sum_{z=1}^{Z} \bar{\xi}_z \sum_{n=1}^{N} \lambda_{n,z,s} \psi_{n,s}(x). \tag{32}$$

Figure 2 shows a block diagram of the noise-adaptive speech recognition system based on the aspect model. In the training phase, single-noise HMMs are trained using the training data. MM-HMMs can be obtained by combining all single-noise HMMs. Using the MM-HMMs and training data, the aspect models are computed. In the adaptation

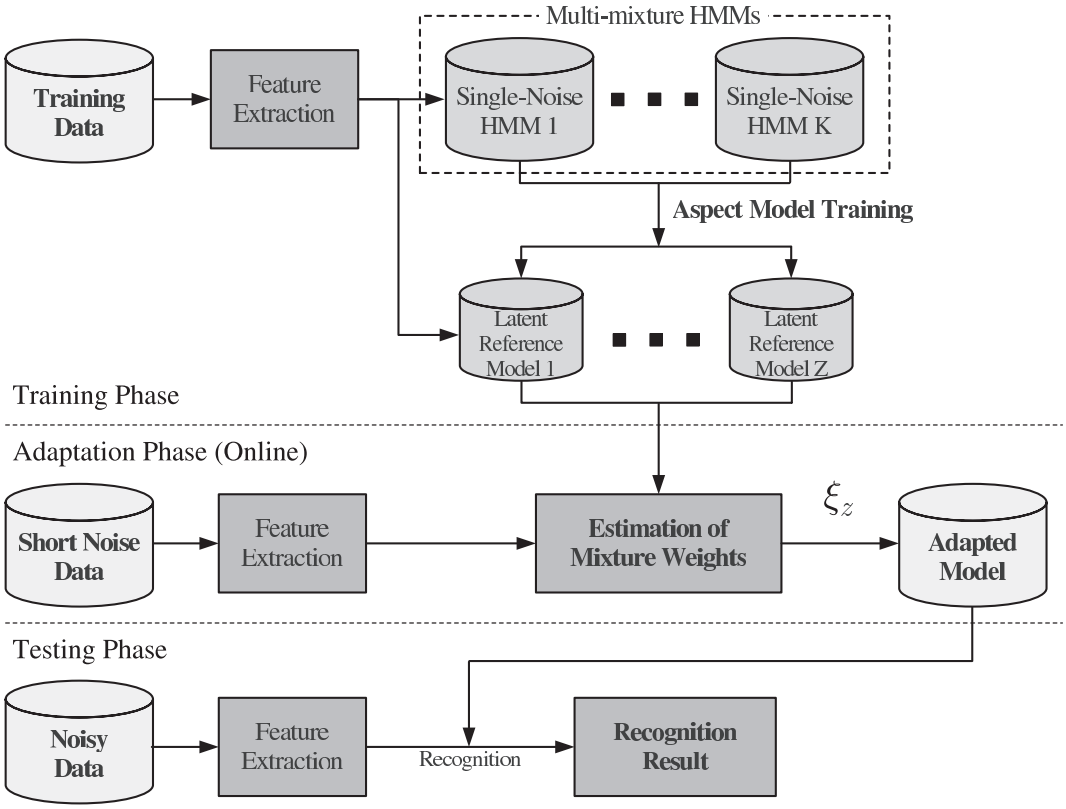


Fig. 2 Environment adaptation system using an aspect model.

phase, the original aspect model is adjusted using the adaptation data. Here, the second-level weightings of each latent reference model, ξ_z, are the unit for adaptation instead of the single-noise model set.

Since we adjust only Z parameters in the adaptation phase, we can estimate the parameters without suffering from the complexity of the original model. Therefore, we can employ HMMs with a large number of mixture components without degrading the performance of adaptation.

3.4 Experiments

3.4.1 Experimental Conditions

Speech recognition experiments were conducted for investigating the performance of environment adaptation using the aspect model. In these experiments, the SNR of the test data was set from 0 to 20 dB. The speech for adaptation were generated by connecting sentences spoken by the same speaker of the test data. Then, noise signals of the same type as the test data were added to the adaptation speech signals with the same SNR as the test data. The speech and noise signals of the adaptation data were not included in either the training or the test data. One adaptation speech signal was used for a speaker and an environment for recognizing five sentences uttered by the same speaker in the same environment. Adaptation experiments were conducted for 20 speakers (10 males, 10 females) in five environments, yielding 500 sentences in total. The other experimental conditions were same as the experiments described in Sect. 2.4.

3.4.2 MM and Conventional Noise Adaptation Methods

First, we confirmed that the plain MM-model was better than a simple noise-reduction method such as SS to compare the proposed method with methods that use information regarding the noise signals. In addition, we tried to apply MLLR adaptation to confirm whether MLLR adaptation works when a few seconds of adaptation data is used.

Among various extended versions of SS, we used multi-band SS (MBSS) [27] in this study. The spectral floor parameter was set to $\beta = 0.03$, which gave the best results in the preliminary experiment. For estimating the noise spectrum, 0.1 s of noise data was used. For the MLLR, a global transformation matrix was used for adaptation because the length of the speech for adaptation was not quite long (5 s).

Figure 3 shows the experimental results. The performance of the MLLR degraded as compared to the original MM models. This could be because the adaptation data are not sufficient for MLLR adaptation. Further, the word accuracy of MBSS did not improve and was not as effective as that of the MM-HMMs.

3.4.3 MM and the Proposed Noise Adaptation Method

Next, we carried out experiments to investigate the effect of noise adaptation (i.e., weight optimization) of the MM-model using the aspect model. In this experiment, we com-

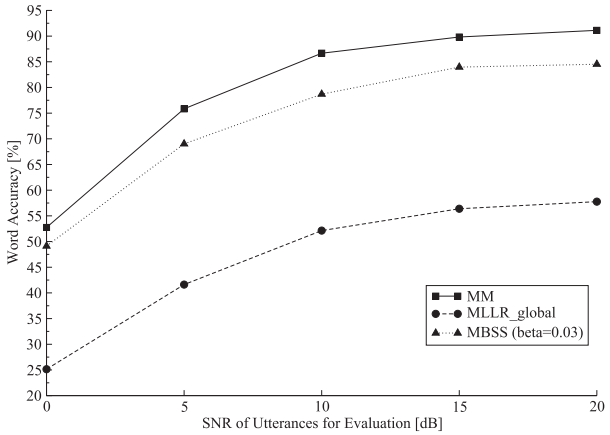


Fig. 3 Word accuracy of SS, MLLR, and MM-HMM.

pared the word accuracy of the model for SNR of 0, 10, and 20 dB. We used five types of adaptation data: 0.1 s and 0.5 s signals that contained only noise, 1 s, 2 s, and 5 s signals that contained speech with noise. When the noise-only signals were used, we used only silence models of the HMM for adaptation.

In this experiment, we used 2, 5 and 10 aspect models. As the number of noise environments of the training data was 48 (12 noises $\times$ 4 SNRs), the aspect model reduced the number of adaptation parameters to 4%, 10% and 21%, respectively.

The results for different numbers of aspect models are shown in Fig. 4. First, as compared to the MM-HMM, the weight optimization did improve word accuracy in most of the conditions, but the degree of improvement was different for different conditions. When the environmental noise was heavy (i.e., 0 dB), the improvement was at the most 4.1 points using 10 aspect models and 5 s of adaptation speech. Using noisy speech was effective in the 0 dB condition as compared to using only noise signals, but the length of the adaptation speech had no significant effect. In the better noise cases (i.e., 10 and 20 dB), the effect of adaptation decreased (0.52 point for 10 dB, and 0.62 point for 20 dB).

Figure 5 shows the result for 0, 10, and 20 dB SNR conditions when 1 s of adaptation data and 10 aspect models were used. We can see that the improvement was the highest for the 0 dB condition. The reason why there was no significant improvement may be that the effect of noise on the variation of speech signals was smaller than the effect of other factors such as the speaker.

3.4.4 Comparison of Parameter Initialization

As shown in Eq. (25), the initial values of $\bar{\xi}_z$ are calculated by averaging $\xi_{z,j}$ with respect to j. We compared the initialization of parameters with simply using the uniform value (i.e., $\bar{\xi}_z^{(0)} = 1/Z$). Figure 6 shows the experimental result when number of aspect models were 10 and SNR were 0 dB. As this result indicates, parameter initialization using $\xi_{z,j}$ is slightly better than just using the uniform value, but the dif-

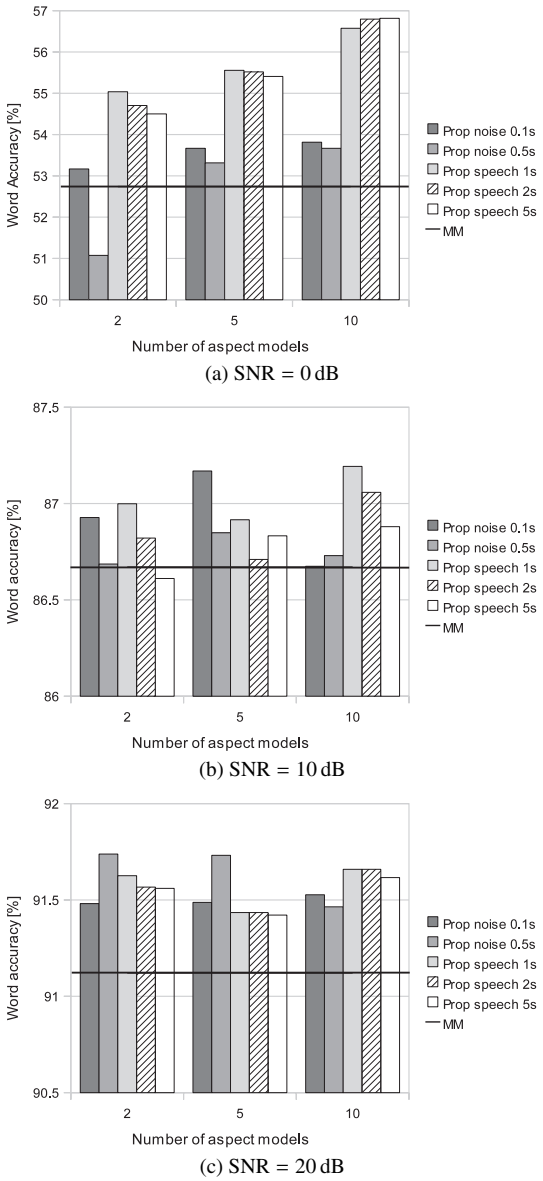


Fig. 4 Effect of noise adaptation for different number of aspect models.

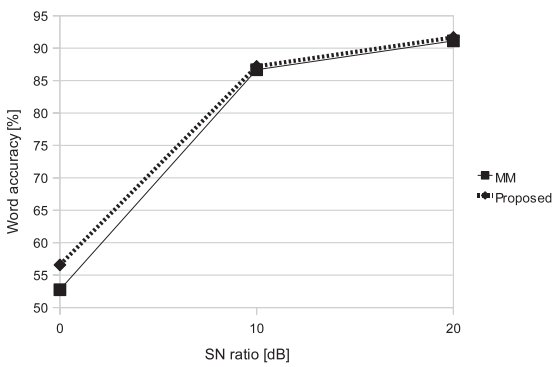


Fig. 5 Effect of noise adaptation for various SNR (1 s adaptation data, 10 aspect models).

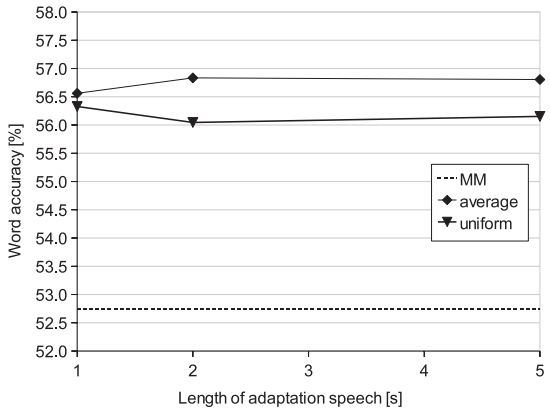


Fig. 6 Effect of initial value of $\bar{\xi}_z$ (0 dB SNR, 10 aspect models).

ference is not large.

3.4.5 Comparison with Other Noise Adaptation Methods

We conducted experiments for comparing the proposed method with two adaptation methods. First, we considered a method that optimizes γ_j directly based on EM algorithm. On optimizing γ_j, we used uniform values as initial values of γ_j.

Next, we carried out experiments to compare the performance between the cluster-based noise adaptation and proposed methods [21]. In this experiment, we used a top-down clustering method based on Bhattacharyya distance [17], [23]. Of the constructed tree-structure, the root node is identical to MC-HMMs and the leaf node is the same as single-noise HMMs. The depth of levels is 8 and the total number of node is 94. After constructing tree-structure, each node model is expressed as a tied-state left-to-right 16-mixture context-dependent triphone model.

As for the computational complexity of the selection method, it requires calculation of likelihood of the adaptation data for almost half of clusters when tree-based pruning is employed. In our experiment, as we used 94 clusters, we need to calculate likelihood nearly 50 times (the actual number of calculation depends on the situation). On the other hand, the proposed method need to calculate Eq. (30) for adaptation. The computational complexity of calculation of likelihood in Eq. (30) is in proportion to number of environments for training. Note that re-calculation of likelihood is not needed for each iteration because only $\bar{\xi}_z$ are changed in the iteration. For we used 48 noises, number of likelihood calculation of our method is comparable to the selection method.

Figure 7 shows the recognition results for the aspect models, EM training and model selection methods. The EM-based optimization could not improve the recognition performance, which seems to be caused by the number of adaptation parameters (48), which was too many to estimate from only 5 s of adaptation data. The method based on the model selection gave the best results when SNR was 0 dB and length of the adaptation speech was 1 s or 5 s. However,

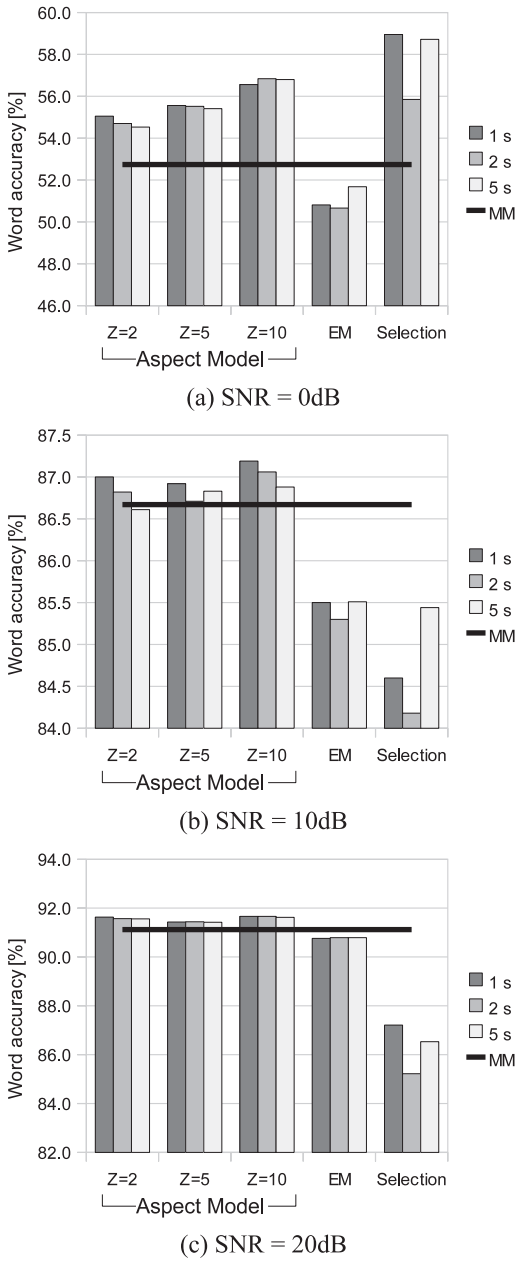


Fig. 7 Comparison between aspect models and other optimization methods.

the model selection method could not improve the word accuracy in 10 dB and 20 dB case, showing that the generated clusters were mainly determined by the speech with heavy noises. Conversely, the proposed method showed stable improvement under all SNR conditions. The stable improvement of the proposed method is caused by the training procedure of the aspect models; since the aspect models are trained so that the aspect models reproduce a probability distribution of any specific noise and SNR in the training data, the aspect models can express speech under any noise environment regardless of its SNR condition.

4. Conclusions

In this paper, we proposed an acoustic model that is robust to various environmental noises. The MM-HMM was obtained by combining HMMs that are trained using corrupted speech data each containing different types of background noises. Experimental results showed that the MM-HMM exhibited the best recognition performance for any type of noise and any variation in SNRs. On the basis of the MM-HMM, we investigated the noise-robust speech recognition method using an adaptation approach. The aspect model is a two-level mixture model, which can reduce the number of free parameters for adaptation.

We evaluated the performance of the proposed method through an experiment. Although we used noisy speech data of very short length, the performance of the proposed method was higher than that of MM-HMMs under heavy noise condition. In addition, we compared the proposed method with the EM-based weight optimization and the model selection method based on a tree-structured model clusters. As a result, the proposed method outperformed the existing adaptation methods except the model selection method at 0 dB SNR condition. Moreover, the proposed method improved the recognition performance constantly regardless of the SNR conditions.

References

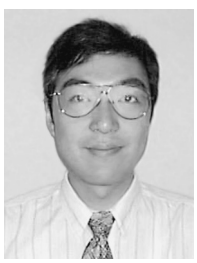
[1] S.V. Vaseghi and B.P. Milner, "Noise compensation methods for hidden Markov model speech recognition in adverse environments," IEEE Trans. Speech Audio Process., vol.5, no.1, pp.11–21, 1997.
[2] J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," J. Acoust. Soc. Am., vol.78, no.5, pp.1508–1518, 1985.
[3] M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, "Microphone array based speech recognition with different talker-array positions," Proc. ICASSP'97, pp.227–230, 1997.
[4] J. Adcock, Y. Gotoh, D. Mashao, and H. Silverman, "Microphone-array speech recognition via incremental MAP training," Proc. ICASSP'96, pp.897–900, 1996.
[5] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-27, no.2, pp.113–120, 1979.
[6] H. Hermansky, A. Bayya, N. Morgan, and P. Kohn, "Compensation for the effect of the communication channel in perceptual linear predictive (PLP) analysis of speech," Proc. EUROSPEECH'91, pp.1367–1370, 1991.
[7] H. Matsumoto, Y. Nakatoh, and Y. Furuhata, "An efficient Mel-LPC analysis method for speech recognition," Proc. ICSLP'98, pp.1051–1054, 1998.
[8] S. Ikabal, H. Misra, and H. Bourlard, "Phase autocorrelation (PAC) derived robust speech features," Proc. ICASSP, pp.133–136, 2003.
[9] M.J.F. Gales and S.J. Young, "HMM recognition in noise using parallel model combination," Proc. EUROSPEECH, pp.837–840, 1993.
[10] F. Martin, K. Shikano, and Y. Minami, "Recognition of noisy speech by composition of hidden Markov models," Proc. EUROSPEECH, pp.1031–1034, 1993.
[11] R.P. Lippmann, E.A. Martin, and D.B. Paul, "Multi-style training for robust isolated-word speech recognition," Proc. ICASSP'87, pp.705–708, 1987.
[12] R.M. Stern, F.-H. Liu, Y. Ohshima, T.M. Sullivan, and A. Acero, "Multiple approaches to robust speech recognition," Proc. ICSLP, pp.274–277, 1992.
[13] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," Proc. ICSLP, 2000.
[14] E. Visser, M. Otsuka, and T.-W. Lee, "A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments," Speech Commun., vol.41, no.2–3, pp.393–407, 2003.
[15] D. Pierce and H.G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," Proc. Interspeech, 2000.
[16] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," Comput. Speech Lang., vol.9, no.2, pp.171–185, 1995.
[17] T. Kosaka and S. Sagayama, "Tree-structured speaker clustering for fast speaker adaptation," Proc. ICASSP, pp.245–248, 1994.
[18] M. Gales, D. Pye, and P. Woodland, "Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation," Proc. ICSLP, 1996.
[19] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, "Jacobian approach to fast acoustic model adaptation," Proc. ICASSP, pp.835–838, 1997.
[20] H.M. Cung and Y. Normandin, "Noise adaptation algorithms for robust speech recognition," Speech Commun., vol.12, no.3, pp.267–276, 1993.
[21] Z. Zhang, T. Sugimura, and S. Furui, "Tree-structured clustering methods for piecewise linear-transformation-based noise adaptation," IEICE Trans. Inf. & Syst., vol.E88-D, no.9, pp.2168–2176, Sept. 2005.
[22] S.-J. Hahm, A. Ito, S. Makino, and M. Suzuki, "A fast speaker adaptation method using aspect model," Proc. Interspeech, pp.1221–1224, 2008.
[23] M. Suzuki, T. Abe, H. Mori, S. Makino, and H. Aso, "Speaker adaptation using phoneme-dependent tree-structured speaker clustering," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J83-D, no.11, pp.981–989, Nov. 1999.
[24] T. Matsui, M. Naito, H. Singer, A. Nakamura, and Y. Sagisaka, "A large-scale Japanese speech database with wide regional and age distribution," Proc. ASJ Fall Meeting, pp.169–170, 1999.
[25] S. Itahashi, "Creating speech corpora for speech science and technology," IEICE Trans., vol.E74, no.7, pp.1906–1910, July 1991.
[26] A. Lee, T. Kawahara, and K. Shikano, "Julius — An open source real-time large vocabulary recognition engine," Proc. EUROSPEECH, pp.1691–1694, 2001.
[27] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," Proc. ICASSP 2002, pp.4164–4164, 2002.

Seong-Jun Hahm was born in Seoul, Korea in 1976. He received B.S. degrees from Kookmin University, Seoul, Korea in 2003 and M.S. degree from Yeungnam University, Gyeongsan, Korea in 2006, respectively. He is currently pursuing the Ph. D. degree in Tohoku University, Sendai, Japan. His research interests include rapid speaker and environment adaptation and speaker recognition. He is a member of the Acoustical Society of Japan.

Yuichi Ohkawa was born in Osaka, Japan in 1975. He received B.E., M.E. and Ph.D. degrees from Tohoku University, Sendai, Japan in 1998, 2000 and 2006, respectively. Since 2003, he has worked with the Graduate School of Engineering, Tohoku University as a research associate. He is now an Assistant Professor of the Graduate School of Educational Informatics, Tohoku University. He has been engaged in spoken language processing and educational technology. He is a member of the Acoustical Society of Japan.

Masashi Ito was born in Fukushima, Japan in 1970. He received B.E., M.E., and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1994, 1996, and 2006, respectively. From 1996 to 2004 he worked with Wako Research Center, Honda R & D Co., Ltd., Japan. He is now an Assistant Professor of the Graduate School of Engineering, Tohoku University. His interests include speech perception, processing, and recognition. He is a member of the Acoustical Society of Japan.

Motoyuki Suzuki was born in Chiba, Japan in 1970. He received B.E., M.E. and Ph.D. degrees from Tohoku University, Sendai, Japan, in 1993, 1995 and 2004, respectively. Since 1996, he has worked with the Computer Center and the Information Synergy Center, Tohoku University as a research associate. From 2006 to 2007 he worked with the Centre for Speech Technology Research, University of Edinburgh, UK, as a visiting researcher. He is now an Associate Professor of the Institute of Technology and Science, The University of Tokushima. His interests include spoken language processing, music information retrieval and pattern recognition using statistical modeling. He is a member of the Acoustical Society of Japan and the Information Processing Society of Japan.



Akinori Ito was born in Yamagata, Japan in 1963. He received B.E., M.E. and Ph.D. degrees from Tohoku University, Sendai, Japan in 1984, 1986 and 1992, respectively. Since 1992, he has worked with the Research Center for Information Sciences and Education Center for Information Processing, Tohoku University. He was with the Faculty of Engineering, Yamagata University from 1995 to 2002. From 1998 to 1999 he worked with the College of Engineering, Boston University, MA, USA as a visiting scholar. He is now an Associate Professor of the Graduate School of Engineering, Tohoku University. He has engaged in spoken language processing, statistical text processing and audio signal processing. He is a member of the Acoustical Society of Japan, the Information Processing Society of Japan and the IEEE.

Shozo Makino was born in Osaka, Japan on January 3, 1947. He received B.E., M.E. and Dr. Eng. degrees from Tohoku University, Sendai, Japan in 1969, 1971 and 1974, respectively. Since 1974, he has worked with the Research Institute of Electrical Communication, Research Center for Applied Information Sciences, Graduate School of Information Science, Computer Center and Information Synergy Center, as a Research Associate, an Associate Professor and a Professor. He is now a Professor of the Graduate School of Engineering, Tohoku University. He has been engaged in spoken language processing, CALL systems, autonomous robot systems, speech corpora, music information processing, image recognition and understanding, natural language processing, semantic web searches and digital signal processing.