Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 95491, Pages 1–11
DOI 10.1155/ASP/2006/95491

Robust Distant Speech Recognition by Combining Multiple Microphone-Array Processing with Position-Dependent CMN

Longbiao Wang, Norihide Kitaoka, and Seiichi Nakagawa

Department of Information and Computer Sciences, Toyohashi University of Technology, Toyahashi-shi 441-8580, Japan

Received 29 December 2005; Revised 20 May 2006; Accepted 11 June 2006

We propose robust distant speech recognition by combining multiple microphone-array processing with position-dependent cepstral mean normalization (CMN). In the recognition stage, the system estimates the speaker position and adopts compensation parameters estimated a priori corresponding to the estimated position. Then the system applies CMN to the speech (i.e., position-dependent CMN) and performs speech recognition for each channel. The features obtained from the multiple channels are integrated with the following two types of processings. The first method is to use the maximum vote or the maximum summation likelihood of recognition results from multiple channels to obtain the final result, which is called *multiple-decoder processing*. The second method is to calculate the output probability of each input at frame level, and a single decoder using these output probabilities is used to perform speech recognition. This is called *single-decoder processing*, resulting in lower computational cost. We combine the delay-and-sum beamforming with *multiple-decoder processing* or *single-decoder processing*, which is termed multiple microphone-array processing. We conducted the experiments of our proposed method using a limited vocabulary (100 words) distant isolated word recognition in a real environment. The proposed multiple microphone-array processing using multiple decoders with position-dependent CMN achieved a 3.2% improvement (50% relative error reduction rate) over the delay-and-sum beamforming with conventional CMN (i.e., the conventional method). The multiple microphone-array processing using a single decoder needs about one-third the computational time of that using multiple decoders without degrading speech recognition performance.

1. INTRODUCTION

Automatic speech recognition (ASR) systems are known to perform reasonably well when the speech signals are captured using a close-talking microphone. However, there are many environments where the use of a close-talking microphone is undesirable for reasons of safety or convenience. Hands-free speech communication [1–5] has been more and more popular in some special environments such as an office or the cabin of a car. Unfortunately, in a distant environment, channel distortion may drastically degrade speech recognition performance. This is mostly caused by the mismatch between the practical environment and the training environment.

Compensating an input feature is the main way to reduce a mismatch. Cepstral mean normalization (CMN) has been used to reduce channel distortion as a simple and effective way of normalizing the feature space [6, 7]. CMN reduces errors caused by the mismatch between test and training conditions, and it is also very simple to implement. Thus, it has been adopted in many current systems. However, the system should wait until the end of speech to activate the recognition procedure when adopting a conventional CMN [6]. The other problem is that the accurate cepstral mean cannot be estimated especially when the utterance is short. However, the recognition of short utterances such as commands, city names is very important in many applications. In [8], the CMN was modified to estimate compensation parameters from a few past utterances for real-time recognition. But in a distant environment, the transmission characteristics from different speaker positions are very different. This means that the method in [8] cannot track the rapid change of the transmission characteristics caused by change in the speaker position, and thus cannot compensate for the mismatch in the context of hands-free speech recognition.

In this paper, we propose a robust speech recognition method using a new real-time CMN based on speaker position, which we call position-dependent CMN. We measured the transmission characteristics (the compensation parameters for position-dependent CMN) from some grid points in the room a priori. Four microphones were arranged in a T-shape on a plane, and the sound source position was estimated by time delay of arrival (TDOA) among the microphones [9–11]. The system then adopts the compensation parameter corresponding to the estimated position and applies a channel distortion compensation method to the speech (i.e., position-dependent CMN) and performs speech recognition. Speech recognition uses the input features compensated by proposed position-dependent CMN. In our method, cepstral means have to be estimated a priori from utterances spoken in each area, but this is costly. The simple solution is to use utterances emitted from a loudspeaker to estimate them. But they cannot be used to compensate for real utterances spoken by a human, because of the effects of recording and playing equipment. We also solve this problem by compensating the mismatch between voices from human and loudspeaker using compensation parameters estimated by a low-cost method.

In a distant environment, the speech signal received by a microphone is affected by the microphone position and the distance from the sound source to the microphone. If an utterance suffers fatal degradation by such effects, the system cannot recognize it correctly. Fortunately, the transmission characteristics from the sound source to every microphone should be different, and the effect of channel distortion for every microphone (it may contain estimation errors) should also be different. Therefore, complementary use of multiple microphones may achieve robust recognition. In this paper, the maximum vote (i.e., *voting method* (VM)) or the maximum summation likelihood (i.e., *maximum-summation-likelihood method* (MSLM)) of all channels is used to obtain the final result [12], which is called *multiple-decoder processing*. This should obtain robust performance in a distant environment. However, the computational complexity of *multiple-decoder processing* is K (the number of input streams) times that of a single input. To reduce the computational cost, the output probability of each input is calculated at frame level, and a single decoder using these output probabilities is used to perform speech recognition, which is called *single-decoder processing*.

Even when using multiple channels, each channel obtained from a single microphone is not stable because it does not utilize the spatial information. On the other hand, beamforming is one of the simplest and the most robust means of spatial filtering, which can discriminate between signals based on the physical locations of the signal sources [13]. Therefore beamforming cannot only separate multiple sound sources but also suppress reverberation for the speech source of interest. Many microphone-array-based speech recognition systems have successfully used delay-and-sum processing to improve recognition performance because of its simplicity, and it remains the method of choice for many array-based speech recognition systems [2, 3, 5, 14]. Nevertheless, beams with a different property would be formed depending on the array structure, sensor spacing, and sensor quality [15]. Using a different sensor array, more robust spatial filtering would be obtained in a real environment. In this paper, a delay-and-sum beamforming combined with *multiple-decoder processing* or *single-decoder processing* is proposed. This is called multiple microphone-array processing. Furthermore, position-dependent CMN (PDCMN) is also integrated with the multiple microphone-array processing.

Section 2 describes the 3D space speaker position estimation based on the time delay of arrival (TDOA). An environmentally robust real-time effective channel compensation method called position-dependent CMN is described in Section 3. A multiple microphone-array processing using multiple decoders or single decoder is proposed in Section 4, while Section 5 describes the experimental results of distant speech recognition in a real environment. Finally, Section 6 summarizes the paper and describes future directions.

2. SPEAKER POSITION ESTIMATION

Speaker localization based on time delay of arrival (TDOA) between distinct microphone pairs has been shown to be effectively implementable and to provide good performance even in a moderately reverberant environment and in noisy conditions [9, 11, 16–18]. Speaker localization in an acoustical environment involves two steps. The first step is estimation of time delays between pairs of microphones. The next step is to use these delays to estimate the speaker location.

The performance of TDOA estimation is very important to the speaker localization accuracy. The prevalent technique for TDOA estimation is based upon generalized cross-correlation (GCC) in which the delay estimation is obtained as the time lag which maximizes the cross correlation between filtered versions of the received signals [10]. In [9, 18, 19], some more effective TDOA estimation methods in noisy and reverberant acoustic environments were proposed.

It should be recalled, however, that it is necessary to find the speaker position using estimated delays. The maximum likelihood (ML) location estimate is one of the common methods because of its proven asymptotic consistency. It does not have a closed-form solution for the speaker position because of the nonlinearity of the hyperbolic equations. The Newton-Raphson iterative method [20], Gauss-Newton method [21], and least-mean-squares (LMS) algorithm are among possible choices to find the solution. However, for these iterative approaches, selecting a good initial guess to avoid a local minimum is difficult, the convergence consumes much computation time, and the optimal solution cannot be guaranteed. Therefore, it is our opinion that an ML location estimate is not suitable for real-time implementation of a speaker localization system.

We earlier proposed a method to estimate the speaker position using a closed-form solution [22]. Using this method, the speaker position can be estimated in real time using TDOAs. This method involves relatively low computational cost, and there is no position estimation error if the TDOA estimation is correct because no assumption is needed for the relative position between the microphones and the sound source. Of course, this approach leads to an estimation error caused by the measuring error of TDOA. If there are more than 4 microphones, we can also estimate the location by using the other combinations of 4 microphones. Thus, we can estimate the location by the average of estimated locations at only a small computational cost.

As will be mentioned in Section 5.1, we did not use position estimation for experiments but assumed that we could estimate accurate position because various previous works revealed the sufficient accuracy of the methods based on TDOA for our purpose.

3. POSITION-DEPENDENT CMN

3.1. Conventional CMN and real-time CMN

A simple and effective way of channel normalization is to subtract the mean of each cepstrum coefficient (CMN) [6, 7], which will remove time-invariant distortions caused by the transmission channel and the recording device.

When speech s is corrupted by convolutional noise h and additive noise n, the observed speech x becomes

$$x = h \otimes s + n. \tag{1}$$

Spectral subtraction, and so forth, can be used to compensate for the additive noise, and then the channel noise can be compensated by the CMN. In this paper, we propose methods to compensate for the effect of channel distortion dependent on speaker position. For the sake of simplicity, we assumed that the additive noises were negligible or well reduced by other methods. So the effect of additive noise was ignored in this paper. We did, in fact, conduct our experiment in a silent seminar room. So (1) is modified as $x = h \otimes s$.

Cepstrum is obtained by DCT transforming a logarithm of a power spectrum of the signal (i.e., $C^x = \mathrm{DCT}(\log|\mathrm{DFT}(x)|^2)$), and thus (1) becomes

$$C^x = C^h + C^s, \tag{2}$$

where C^x, C^h, and C^s express the cepstrums of observed speech x, transmission characteristics h, and clean speech s, respectively.

Based on this, the convolutional noise is considered as additive bias in the cepstral domain, so the noise (transmission characteristics or channel distortion) can be compensated by CMN in the cepstral domain as

$$\widetilde{C}_t = C_t - \Delta C \quad (t = 0, \ldots, T), \tag{3}$$

where $\widetilde{C}_t$ and C_t are compensated and original cepstrums at time frame t, respectively.

In conventional CMN, the compensation parameter ΔC is approximated by

$$\Delta C \approx \overline{C_t} - \overline{C}_{\text{train}}, \tag{4}$$

where $\overline{C}_t$ and $\overline{C}_{\text{train}}$ are cepstral means of utterances to be recognized and those to be used to train the speaker-independent acoustical model, respectively. Thus, when using conventional CMN, the compensation parameter ΔC can be calculated at the end of input speech. This prevents real-time processing of speech recognition. The other problem of conventional CMN is that accurate cepstral means cannot be estimated especially when the utterance is short.

We solve these problems under the assumption that the channel distortion does not change drastically. In our method, the compensation parameter is calculated from utterances recorded a priori. The new compensation parameter is defined by

$$\Delta C = \overline{C}_{\text{environment}} - \overline{C}_{\text{train}}, \tag{5}$$

where $\overline{C}_{\text{enviornment}}$ is the cepstral mean of utterances recorded in a practical environment a priori. Using this method, the compensation parameter can be applied from the beginning of recognition of current utterance. Moreover, as the compensation parameter is estimated from a sufficient number of cepstral coefficients of utterances, so it can compensate for the distortion better than the conventional CMN. We call this method *real-time CMN*. In our early work [8], the compensation parameter is calculated from past recognized utterances. Thus, the calculation of the compensation parameter for the nth utterance is

$$\Delta C^{(n)} = (1 - \alpha)\Delta C^{(n-1)} - \alpha \times \left(\overline{C}_{\text{train}} - \overline{C^{(n-1)}}\right), \tag{6}$$

where $\Delta C^{(n)}$ and $\Delta C^{(n-1)}$ are the compensation parameters for the nth and $(n-1)$th utterances, respectively, and $\overline{C^{(n-1)}}$ is the mean of cepstrums of the $(n-1)$th utterance. Using this method, the compensation parameter can be calculated before recognition of the nth utterance. This method can indeed track the slow changes in transmission characteristics, but the characteristic changes caused by the change in speaker position or speaker are beyond the tracking ability of this method.

3.2. Incorporate speaker position information into real-time CMN

In a real distant environment, the transmission characteristics of different speaker positions are very different because of the distance between the speaker and the microphone, and the reverberation of the room. Hence, the performance of a speech recognition system based on real-time CMN will be drastically degraded because of the great change of channel distortion.

In this paper, we incorporate speaker position information into real-time CMN [23]. We call this method *position-dependent* CMN. The new compensation parameter for position-dependent CMN is defined by

$$\Delta C = \overline{C}_{\text{position}} - \overline{C}_{\text{train}}, \tag{7}$$

where $\overline{C}_{\text{position}}$ is the cepstral mean of utterances affected by the transmission characteristics between a certain position and the microphone. In our experiments in Section 5, we divide the room into 12 areas as in Figure 1 and measure the $\overline{C}_{\text{position}}$ corresponding to each area.

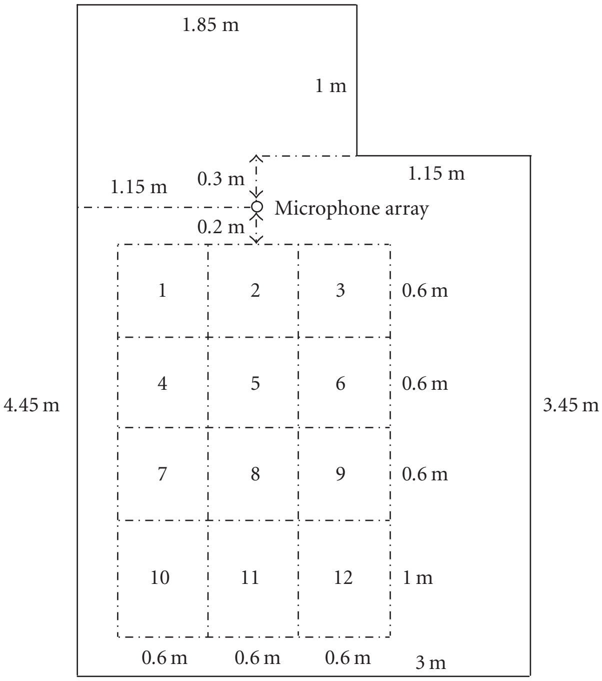

Figure 1: Room configuration (room size: (W) 3 m × (L) 3.45 m × (H) 2.6 m).

3.3. Problem and solution

In position-dependent CMN, the compensation parameters should be calculated a priori depending on the area, but it is not realistic to record a sufficient amount of utterances spoken in each area by a sufficient number of humans because that would take too much time. Thus, in our experiment, the utterances were emitted from a loudspeaker in each area. However, because the cepstral means were estimated by using utterances distorted by the transmission characteristics of the channel including the loudspeaker, they cannot be used to compensate for real utterances spoken by human.

In this paper, we solve this problem by compensating the mismatch between voices from humans and loudspeaker. An observed cepstrum of a distant human's utterance is as follows:

$$C^x_{\text{human}} = C^s_{\text{human}} + C^h_{\text{environment}}, \tag{8}$$

where C^x_{human}, C^s_{human}, and $C^h_{\text{environment}}$ are the cepstrums of observed human utterance, emitted human utterance, and transmission characteristics from human's mouth to the microphone, respectively. However, an observed cepstrum of a distant loudspeaker's utterance is as follows:

$$\begin{aligned} C^x_{\text{loudspeaker}} &= C^s_{\text{loudspeaker}} + C^h_{\text{environment}} \\ &= C^s_{\text{human}} + C^h_{\text{loudspeaker}} + C^h_{\text{environment}}, \end{aligned} \tag{9}$$

where $C^x_{\text{loudspeaker}}$, $C^s_{\text{loudspeaker}}$, and $C^h_{\text{loudspeaker}}$ are the cepstrums of observed speech emitted by the loudspeaker, human utterances emitted by the loudspeaker, and transmission characteristics of the loudspeaker, respectively. That is, the speech emitted by the loudspeaker is human speech corrupted by the transmission characteristics of the loudspeaker. The difference between (8) and (9) is $C^h_{\text{loudspeaker}}$, and this is independent of the other environment such as the speaker position.

Thus, the compensation parameter ΔC in (7) is modified as

$$\Delta C = \{\overline{C}_{\text{position}} - \overline{C}_{\text{train}}\} - \{\overline{C}_{\text{loudspeaker}} - \overline{C}_{\text{human}}\}, \tag{10}$$

where $\overline{C}_{\text{human}}$ and $\overline{C}_{\text{loudspeaker}}$ are cepstral means of close-talking human utterances and those of utterances from a close-loudspeaker. We used far fewer human utterances to estimate $\overline{C}_{\text{human}}$ than to estimate position-dependent cepstral means. In addition, we need only close-talking utterances, which are easier to record than distant-talking utterances.

A detailed illustration is shown in Figure 2.

4. MULTIPLE MICROPHONE SPEECH PROCESSING

The *voting method* (VM) and *maximum-summation-likelihood method* (MSLM) using multiple decoders (i.e., *multiple-decoder processing*) are proposed in Section 4.1. To reduce the computational cost of the methods described in Section 4.1, a multiple-microphone processing using a single decoder (i.e., *single-decoder processing*) is proposed in Section 4.2. In Section 4.3, we combine *multiple-decoder processing* or *single-decoder processing* with the delay-and-sum beamforming.

4.1. Multiple-decoder processing

In this section, we proposed a novel multiple-microphone processing using multiple decoders, which is called *multiple-decoder processing*. The procedure of multiple-microphone processing using multiple decoders is shown in Figure 3, in which all results obtained by different decoders are inputted to a so-called VM or MSLM decision method to obtain the final result.

4.1.1. Voting method

Because of the subtle differences in the features between input streams, different channels may lead to different results for a certain utterance. To achieve robust speech recognition for the multiple channels, a good decision method for the final result from the results obtained from these channels is important. The signal received by each channel is recognized independently, and the system votes for a word according to the recognition result. Then the word which obtained the maximum number of votes is selected as the final recognition result, which is called *voting method* (VM). The *voting method* is defined as

$$\begin{aligned} \widehat{W} &= \arg\max_{W_R} \sum_{i=1}^{\#\text{channel}} I(W_i, W_R), \\ I(W_i, W_R) &= \begin{cases} 1 & \text{if } (W_i = W_R), \\ 0 & \text{otherwise}, \end{cases} \end{aligned} \tag{11}$$

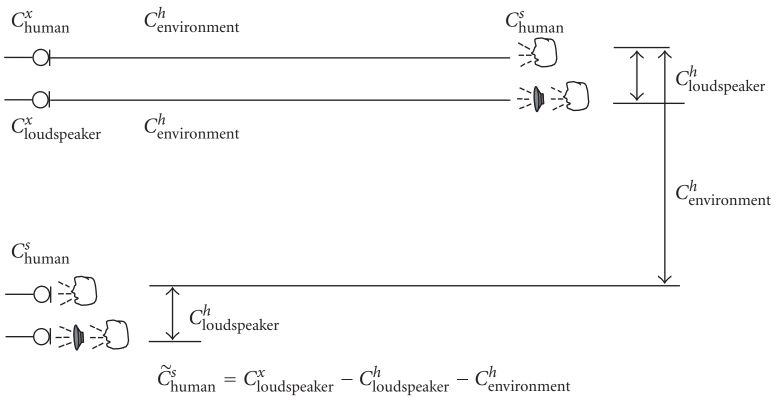
C^x_{human}
$C^h_{\text{environment}}$
C^s_{human}
$C^h_{\text{loudspeaker}}$
$C^x_{\text{loudspeaker}}$
$C^h_{\text{environment}}$
$C^h_{\text{environment}}$
C^s_{human}
$C^h_{\text{loudspeaker}}$
$\tilde{C}^s_{\text{human}} = C^x_{\text{loudspeaker}} - C^h_{\text{loudspeaker}} - C^h_{\text{environment}}$

Figure 2: Illustration of compensation of transmission characteristics between human and loudspeaker (same microphone).

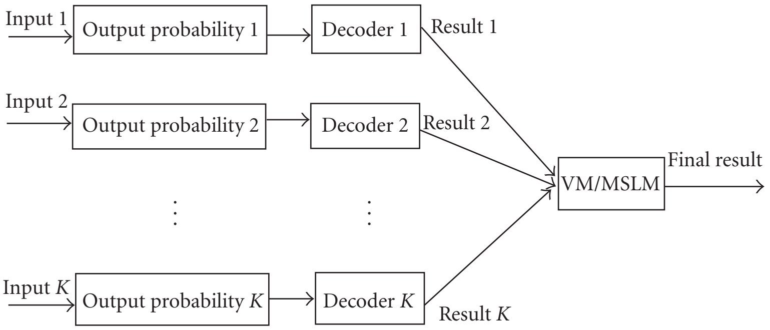

Figure 3: Illustration of multiple-microphone processing using multiple decoders (utterance level).

where W_i is the recognition result of the ith channel, and $I(W_i, W_R)$ denotes an indicator. If there are more than two results that obtain the same number of votes, the result of the microphone which is nearest to the sound source is selected as the final result. In our proposed position-dependent CMN method, speaker position is estimated a priori, so it is possible to calculate the distance from the microphone to the speaker.

4.1.2. Maximum-summation-likelihood method

The likelihood of each microphone can be seen as a potential confidence score, so combining the likelihood of all channels should obtain a robust recognition result. In this paper, the maximal summation likelihood is defined as

$$\widehat{W} = \arg\max_{W_R} \sum_{i=1}^{\#\text{channel}} L_{W_R}(i), \qquad (12)$$

where $L_{W_R}(i)$ indicates the log likelihood of W_R obtained from ith channel. We call this the *maximum-summation-likelihood method* (MSLM). In other words, it is a maximum production rule of probabilities.

4.2. Single-decoder processing

The multiple-microphone processing using multiple decoders may be more robust than a single channel. However, the computational complexity of multiple-microphone processing using multiple decoders is K (the number of input channels) times that of a single input. To reduce the computational cost, instead of obtaining multiple hypotheses or likelihoods at the utterance level using multiple decoders, the output probability of each input is calculated at frame level, and a single decoder using these output probabilities is used to perform speech recognition. We call this method *single-decoder processing*, and Figure 4 shows its processing procedure.

In a multiple-decoder method, a conventional Viterbi algorithm [24] is used in each decoder, and the probability $\alpha(t, j, k)$ of the most likely state sequence at time t which has generated the observation sequence $O_k(1) \cdots O_k(t)$ (until time t) of kth input ($1 \le k \le K$) and ends in state j is defined by

$$\alpha(t, j, k) = \max_{1 \le i \le S} \left\{ \alpha(t-1, i, k) a_{ij} \sum_m \lambda_{mj} b_{mj}\left(O_k(t)\right) \right\}, \qquad (13)$$

where $a_{ij} = P(s_t = j \mid s_{t-1} = i)$ is the transition probability from state i to state j, $1 \le i, j \le S$, $2 \le t \le T$; $b_{mj}(O_k(t))$ is the output probability of mth Gaussian mixture ($1 \le m \le M$) for an observation sequence $O_k(t)$ at state j; and λ_{mj} is the mixture weights. In the multiple-decoder method shown as Figure 3, the Viterbi algorithm is performed by each decoder independently, so K (the number of input streams) times computational complexity is required. Thus, both the calculation of output probability and the rest of the processing cost such as finding a best path (state sequence), and so forth, are K times that of a single input.

In order to use a single decoder for multiple inputs shown in Figure 4, we modify the Viterbi algorithm as follows:

$$\alpha(t, j) = \max_{1 \le i \le S} \left\{ \alpha(t-1, i) a_{ij} \max_k \sum_m \lambda_{mj} b_{mj}(O_k(t)) \right\}. \tag{14}$$

In (14), the maximum output probability of all K inputs at time t and state j is used. So only one best state sequence for all K inputs using the maximum output probability of all K inputs is obtained. This means that extra $K-1$ times only the calculation of the output probability is required compared to that of a single input.

Here, we investigate further reduction of the computational cost. We assume that the output probabilities of K features at time t from each Gaussian component are similar to each other. Hence, if we obtained the maximum output probability of the 1st input from the $\hat{m}$th component among those in state j, it is highly likely that the maximum output probability of kth input will also be obtained from $\hat{m}$th component. Thus, we modify (14) as follows:

$$\begin{gathered} \alpha(t, j) = \max_{1 \le i \le S} \left\{ \alpha(t-1, i) a_{ij} \max_k b_{\hat{m}j}(O_k(t)) \right\}, \\ \hat{m} = \arg\max_m \lambda_{mj} b_{mj}(O_1(t)). \end{gathered} \tag{15}$$

In (15), only extra $(M+K-1)/M - 1 = (K-1)/M$ times calculation of output probability is required compared to that of a single input. The methods defined by (14) and (15) both involve multiple-microphone processing using the single decoder shown in Figure 4. To distinguish these two methods, the method given by (14) is called the *full-mixture single-decoder method*, while the method given by (15) is called the *single-mixture single-decoder method*.

4.3. Multiple microphone-array processing

Many microphone-array-based speech recognition systems have successfully used delay-and-sum processing to improve recognition performance because of its spatial filtering ability and simplicity, so it remains the method of choice for many array-based speech recognition systems [3, 4, 13]. Beamforming can suppress reverberation for the speech source of interest. Beams with different properties would be formed by the array structure, sensor spacing, and sensor quality [15]. As described in Sections 4.1 and 4.2, the multiple-microphone-array processing using multiple decoders or a single decoder should obtain a more robust performance than a single channel or a single-microphone array, because only microphone-array processing may yield estimation error. We integrated a set of the delay-and-sum beamforming with *multiple- or single-decoder processings*.

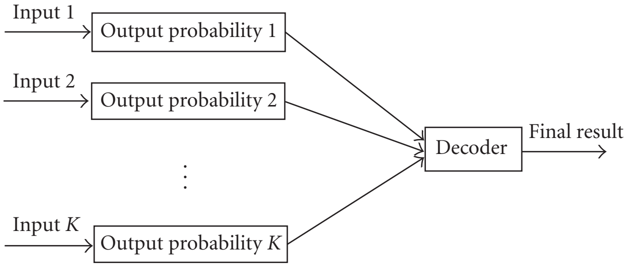

Figure 4: Illustration of multiple microphone processing using single decoder (frame level).

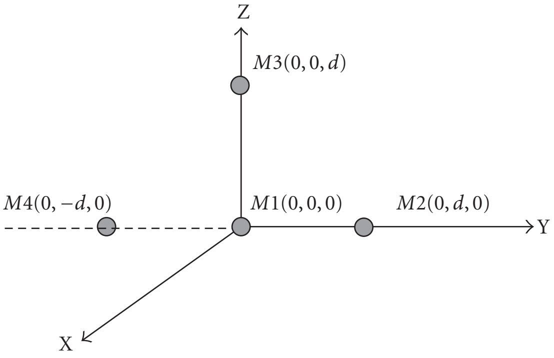

Figure 5: Microphones' setup (d = 20 cm).

In this paper, the 4 T-shaped microphones are set as shown in Figure 5. Array 1 (microphones 1, 2, 3), array 2 (microphones 1, 2, 4), array 3 (microphones 1, 3, 4), array 4 (microphones 2, 3, 4), and array 5 (microphones 1, 2, 3, 4) are used as individual arrays, and thus we can obtain 5 channel input streams using delay-and-sum beamforming. These streams are used as inputs of the *multiple- or single-decoder processings* to obtain the final result. We call this method multiple microphone-array processing. These streams can also be compensated by the proposed position-dependent CMN, and so forth, before they are inputted into *multiple-decoder processing* or *single-decoder processing*.

5. EXPERIMENTS

5.1. Experimental setup

We performed the experiment in the room shown in Figure 6 measuring 3.45 m × 3 m × 2.6 m without additive noise. The room was divided into the $12(3 \times 4)$ rectangular areas shown in Figure 1, where the area size is 60 cm×60 cm. We measured the transmission characteristics (i.e., the mean cepstrums of utterances recorded a priori) from the center of each area. In our experiments, the room was set up as the seminar room shown in Figure 6 with a whiteboard beside the left wall, one table and some chairs in the center of the room, one TV and some other tables, and so forth.

In our method, the estimated speaker position should be used to determine the area (60 cm × 60 cm) in which the speaker should be. It has been shown in [25] that an average location error of less than 10 cm could be achieved using only 4 microphones in a room measuring 6 m×10 m×3 m, in which source positions are uniformly distributed in 6 m×6 m area. In our past study [22], we also revealed that the speaker position could be estimated with estimation errors of 20–25 cm by the 4 T-shaped microphone system as shown in Figure 5 without interpolation between consecutive samples. In the present study, therefore, we assumed that the position area was accurately estimated, and we purely evaluated only our proposed speech recognition methods.

Twenty male speakers uttered 200 isolated words, each with a close microphone. The average time of all utterances was about 0.6 second. For the utterances of each speaker, the first 100 words were used as test data and the rest for estimation of cepstral mean $\overline{C}_{\text{position}}$ in (7) and (10). All the utterances were emitted from a loudspeaker located in the center of each area and recorded for test and estimation of $\overline{C}_{\text{position}}$ to simulate the utterances spoken at various positions. The sampling frequency was 12 kHz. The frame length was 21.3 ms, and the frame shift was 8 ms with a 256-point Hamming window. Then, 116 Japanese speaker-independent syllable-based HMMs (strictly speaking, mora-unit HMMs [26]) were trained using 27992 utterances read by 175 male speakers (JNAS corpus). Each continuous-density HMM had 5 states, 4 with pdfs of output probability. Each pdf consisted of 4 Gaussians with full-covariance matrices. The feature space was comprised of 10 MFCCs. First- and second-order derivatives of the cepstrums plus first and second derivatives of the power component were also included.

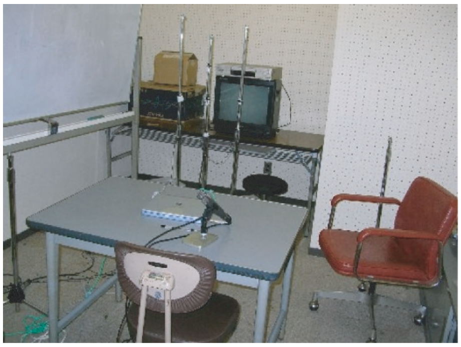

Figure 6: Experimental environment.

Table 1: Recognition results emitted by a loudspeaker (average of results obtained by 4 independent microphones: %).

Area	W/O CMN	Conv. CMN	CM of area 5	PICMN	PDCMN
1	86.3	92.8	94.2	95.0	95.7
2	95.4	95.7	97.4	97.7	97.4
3	94.3	94.6	96.8	97.1	96.8
4	87.4	92.9	93.1	94.6	95.6
5	92.1	93.8	96.3	96.0	96.3
6	90.9	93.2	95.2	96.1	95.9
7	89.0	92.4	94.3	94.8	95.7
8	91.4	91.4	93.8	94.1	94.7
9	92.3	93.1	95.9	96.4	96.0
10	84.9	90.0	90.5	91.8	93.5
11	86.9	90.9	91.7	93.2	94.1
12	85.9	89.8	90.9	93.3	93.3
Average	89.7	92.5	94.2	94.9	95.4

5.2. *Recognition experiment by single microphone*

5.2.1. *Recognition experiment for speech emitted by a loudspeaker*

We conducted the speech recognition experiment of isolated words emitted by a loudspeaker using a single microphone in a distant environment.

The recognition results are shown in Table 1. The proposed method is referred to as PDCMN (position-dependent CMN). In Table 1, the average results obtained by the 4 independent microphones shown in Figure 5 are indicated. In Table 1, PDCMN is compared with the baseline (recognition without CMN), conventional CMN, "CM of area 5," and PICMN (position-independent CMN). Area 5 is in the center of all 12 areas, and "CM of area 5" means that a fixed cepstral mean (CM) in the central area was used to compensate for the input features of all 12 areas. PICMN means the method by which the averaged compensation parameters over 12 areas were used. Without CMN, the recognition rate was drastically degraded according to the distance between the sound source and the microphone. Conventional CMN could not obtain enough improvement because the average duration of all utterances was too short (about 0.6 second). By compensating the transmission characteristics using the compensation parameters measured a priori, all CM of area 5, PICMN, and PDCMN effectively improved the performance of speech recognition from without CMN and conventional CMN.

In a distant environment, the reflection may be very strong and may be very different depending on the given areas, so the difference of transmission characteristics in each area should be very large. In other words, obstacles caused complex reflection patterns depending on the speaker positions. The proposed PDCMN could also achieve more effective improvement than "CM of area 5" and PICMN. The PDCMN achieved a relative error reduction rate of 55.3% from without CMN, 38.7% from conventional CMN, 20.7% from CM of area 5, and 9.8% from PICMN, respectively. The experimental result also shows that the greater the distance between the sound source and the microphone, the greater the improvement.

The differences of the performance between the PDCMN and PICMN/CM of area 5 were significant, but not too large. When assuming larger area, the performance difference must

Table 2: Recognition results of human utterances (results obtained by microphone 1 shown in Figure 5 (%)).

Area	W/O CMN	Conv. CMN	CMN by human utterances	CMN by utterances from a loudspeaker	Proposed method
5	95.8	94.6	96.6	96.0	96.8
9	93.4	90.6	94.4	91.2	94.2
10	84.8	83.8	89.8	83.0	90.0
Average	91.3	89.7	93.6	90.1	93.7

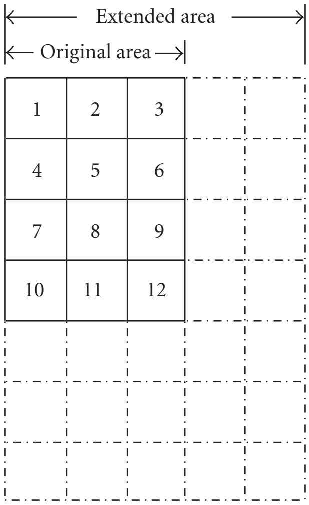

Figure 7: Extended area.

be much larger. So, we assume the extended area described in Figure 7 and then the area 12 of the original area correspond to the center of the extended area. We used "CM of area 12" to compensate the utterances emitted from area 1 to simulate the extended area. The result degraded from 94.2% (CM of area 5) to 92.9%. This was much inferior to that of PDCMN (95.7%). These results indicated that the proposed method works much better in the larger area. This degradation means a larger variation of the transmission characteristics, and this variation must cause the degradation of the performance of PICMN.

5.3. Recognition experiment of speech uttered by humans

We also conducted experiments with real utterances spoken by humans using a single microphone (i.e., microphone 1 in Figure 5 in this case).

The utterances were directly spoken by 5 male speakers instead of those emitted from a loudspeaker in the first experiment. The experimental results are shown in Table 2, in which "CMN by human utterances" means the result of CMN with the cepstral means of real utterances recorded along with the test set (i.e., the ideal case). "CMN by utterances from a loudspeaker" means the result of CMN with the cepstral means of utterances emitted by a loudspeaker. The "proposed method" is the result of the proposed CMN given by (10) which compensated for the mismatch between human (real) and loudspeaker (simulator). In the cases of CMN by human utterances and proposed method, we estimated the compensation parameters for a certain speaker from the utterances by the other 4 persons. We also conducted recognition experiments without CMN and with conventional CMN. Since the utterances were too short (about 0.6 s) to estimate the accurate cepstral mean, conventional CMN was not robust in this case. In Table 1, the utterances were emitted by a loudspeaker whose distortion is relatively large. Hence, the gain of compensating these transmission characteristics is greater than the loss caused by the inaccurate cepstral mean estimated by short utterances. Conventional CMN worked better than without CMN. On the contrary, in Table 2, the utterances were spoken by humans, so the transmission characteristics were much smaller than those in Table 1. Then the degradation caused by the inaccurately estimated cepstral mean became dominant, and the conventional CMN worked even worse than without CMN. The results show that the proposed method could approximate the CMN with the human cepstral mean and was better than the CMN with the loudspeaker cepstral mean.

5.4. Experimental results for multiple-microphone speech processing

The experiments in Section 5.3 showed that the proposed method given by (10) could well compensate for the mismatch between voices from humans and the loudspeaker. For convenience's sake, we used utterances emitted from a loudspeaker to evaluate the multiple-microphone speech processing methods.

The recognition results of a single microphone and multiple microphones are compared in Table 3. The multiple-microphone processing methods described in Section 4.1 which use multiple decoders were conducted. Both *voting method* (VM) and *maximum-summation-likelihood method* (MSLM) are more robust than single-microphone processing. The MSLM achieved a relative error reduction rate of 21.6% from single-microphone processing. The VM and MSLM could achieve a similar result to the conventional delay-and-sum beamforming. By combining the MSLM with beamforming based on position-dependent CMN, an 11.1% relative error reduction rate was achieved from beamforming based on position-dependent CMN, and a 50% relative error reduction rate was achieved from beamforming with conventional CMN (i.e., a conventional method). The MSLM

Table 3: Comparison of recognition accuracy of single microphone with multiple microphones using multiple decoders (%).

	Single	Multiple microphones								
	micro-	VM	MSLM	Beamforming					VM +	MSLM +
	phone			Array 1	Array 2	Array 3	Array 4	Array 5	beamforming	beamforming
W/O CMN	89.7	91.3	91.6	90.9	91.3	91.3	91.0	91.4	91.9	91.9
Conv. CMN	92.5	94.2	94.5	93.5	93.3	93.3	93.4	93.6	94.1	94.2
PICMN	94.9	96.0	96.2	95.7	96.1	95.8	96.0	96.1	96.3	96.4
PDCMN	95.4	96.4	96.6	96.2	96.3	96.2	96.1	96.4	96.7	96.8

Table 4: Comparison of recognition accuracy of multiple microphone-array processing using single decoder with that using multiple decoders (%).

		Multiple decoders (see Table 3)		Single decoder	
		VM + beamforming	MSLM + beamforming	*Full-mixture* + beamforming	*Single-mixture* + beamforming
Recognition rate	W/O CMN	91.9	91.9	93.0	92.0
	Conv. CMN	94.1	94.2	93.9	92.9
	PICMN	96.3	96.4	96.5	96.1
	PDCMN	96.7	96.8	96.9	96.6
Computation ratio		5	5	3.58	1.77

proved more robust than the VM in almost all cases because the summation of the likelihoods can be seen as the potential confidence of all channels. The proposed PDCMN achieved more efficient improvement than PICMN by using multiple microphones. In the case of MSLM combining with beamforming, PDCMN achieved a relative error reduction rate of 11.1% from PICMN. Both PDCMN and PICMN could improve speech recognition performance significantly more than without CMN and conventional CMN. It is not necessary for PICMN to estimate the speaker postion. Therefore, PICMN may also be a good choice because it simplifies system implementation.

As described in Section 4.2, the computational cost of multiple-microphone processing using multiple decoders given by (13) is 5 (the number of microphone arrays) times that of a single channel. Experiments were also conducted on a *full-mixture single-decoder processing* given by (14) and *single-mixture single-decoder processing* given by (15). The computational costs of *full-mixture single-decoder processing* and *single-mixture single-decoder processing* are 3.58 times and 1.77 times that of a single channel, respectively. The recognition results of the multiple microphone-array processing using the multiple decoders and single decoder are shown in Table 4. Since the multiple microphone-array processing using the *full-mixture single decoder* selected a maximum likelihood of each input sequence at every frame, it achieved slightly more improvement than the multiple microphone-array processing using the multiple decoders. The multiple microphone-array processing using the *single-mixture single decoder* reduced computational cost about 50% more than that using the *full-mixture single decoder*. In theory, the improvement of computational complexity between the *single-mixture single-decoder processing* and the multiple-microphone processing using the multiple decoders is determined by the number of inputs K and the number of Guassian mixtures M, as decribed in Section 4.2. The larger the number of Gaussian mixtures was, the greater the reduction of computational cost became. In our experiments, the number of Gaussian mixtures was 4. Comparing the results in Tables 3 and 4, the delay-and-sum beamforming using the *single-mixture single decoder* based on position-dependent CMN achieved a 3.0% improvement (46.9% relative error reduction rate) over the delay-and-sum beamforming based on conventional CMN with 1.77 times the computational cost.

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed a robust distant speech recognition system based on position-dependent CMN using multiple microphones. At first, the 3D space speaker position could be quickly estimated, and then a channel distortion compensation method based on position-dependent CMN was adopted to compensate for the transmission characteristics. The proposed method improved the speech recognition performance more than not only conventional CMN but also position-independent CMN. If the utterance contained more than 3 words (about 2), the recognition rate of the conventional CMN could approximate that of PDCMN in this experimental situation. However, it is unavailable in many short utterance recognition systems. We also compensated for the mismatch between the cepstral means of utterances spoken by humans and those emitted from a loudspeaker. Our experiments showed that the proposed method could also well compensate for the mismatch between voices from humans and the loudspeaker. Multimicrophone speech processing technology such as the *Vot-*

ing method and the *Maximum-summation-likelihood method* was used to obtain robust distant speech recognition. To reduce the computational cost, the output probability of each input was calculated at frame level, and a single decoder using these output probabilities was used to perform speech recognition. Furthermore, we combined delay-and-sum beamforming with *multiple-decoder processing* or *single-decoder processing*. The proposed multiple microphone-array using the single decoder achieved a significant improvement over the single-microphone array. Combining the multiple microphone-array using the single decoder with position-dependent CMN, a 3.0% improvement (46.9% relative error reduction rate) over the delay-and-sum beamforming with conventional CMN was achieved in a real environment at 1.77 times the computational cost.

In future work, we will integrate the speaker position estimation with our speech recognition methods. Furthermore, we will also attempt to track a moving speaker and expand our speech recognition method to accommodate an adverse acoustic environment.

REFERENCES

[1] B. H. Juang and F. K. Soong, "Hands-free telecommunications," in *Proceedings of the International Workshop on Hands-Free Speech Communication (HSC '01)*, pp. 5–10, Kyoto, Japan, April 2001.

[2] M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, "Experiments of hands-free connected digit recognition using a microphone array," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 490–497, Santa Barbara, Calif, USA, December 1997.

[3] T. B. Hughes, H.-S. Kim, J. H. DiBiase, and H. F. Silverman, "Performance of an HMM speech recognizer using a real-time tracking microphone array as input," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 346–349, 1999.

[4] T. Takiguchi, S. Nakamura, and K. Shikano, "HMM-separation-based speech recognition for a distant moving speaker," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 127–140, 2001.

[5] M. L. Seltzer, B. Raj, and R. M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 489–498, 2004.

[6] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[7] F. Liu, R. M. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition," in *Proceedings of the ARPA Speech and Natural Language Workshop*, pp. 69–74, Princeton, NJ, USA, March 1993.

[8] N. Kitaoka, I. Akahori, and S. Nakagawa, "Speech recognition under noisy environments using spectral subtraction with smoothing of time direction and real-time cepstral mean normalization," in *Proceedings of the International Workshop on Hands-Free Speech Communication (HSC '01)*, pp. 159–162, Kyoto, Japan, April 2001.

[9] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1110–1124, 2003.

[10] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[11] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using CSP analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, vol. 2, pp. 921–924, Atlanta, Ga, USA, May 1996.

[12] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speech recognition based on position dependent CMN using a novel multiple microphone processing technique," in *Proceedings of the 9th European Conference on Speech Communication and Technology (EUROSPEECH '05)*, pp. 2661–2664, Lisbon, Portugal, September 2005.

[13] B. Van Veen and K. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[14] T. Yamada, S. Nakamura, and K. Shikano, "Distant-talking speech recognition based on a 3-D Viterbi search using a microphone array," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 48–56, 2002.

[15] J. Flanagan, J. Johnston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *The Journal of the Acoustical Society of America*, vol. 78, no. 5, pp. 1508–1518, 1985.

[16] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, "Real-time passive source localization: a practical linear-correction least-squares approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[17] M. Brandstein, *A framework for speech source localization using sensor arrays*, Ph.D. thesis, Brown University, Providence, RI, USA, 1995.

[18] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, chapter 8, pp. 157–180, Springer, Berlin, Germany, 2001.

[19] V. Raykar, B. Yegnanarayana, S. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 751–760, 2005.

[20] Y. Bard, *Nonlinear Parameter Estimation*, Academic Press, New York, NY, USA, 1974.

[21] W. Foy, "Position-location solutions by Taylor-series estimation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 12, no. 2, pp. 187–194, 1976.

[22] L. Wang, N. Kitaoka, and S. Nakagawa, "Distant speech recognition based on position dependent cepstral mean normalization," in *Proceedings of the 6th IASTED International Conference on Signal and Image Processing (SIP '04)*, pp. 249–254, Honolulu, Hawaii, USA, August 2004.

[23] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speech recognition based on position dependent CMN," in *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP '04)*, pp. 2409–2052, Jeju Island, Korea, October 2004.

[24] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[25] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, 1997.

[26] S. Nakagawa, K. Hanai, K. Yamamoto, and N. Minematsu, "Comparison of syllable-based HMMs and triphone-based HMMs in Japanese speech recognition," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 393–396, Keystone, Colo, USA, December 1999.

Longbiao Wang received his B.E. degree from Fuzhou University, China, in 2000 and an M.E. degree from Toyohashi University of Technology, Japan, in 2005. He is now a Ph.D. student at Toyohashi University of Technology, Japan. From July 2000 to August 2002, he had been working at the China Construction Bank. His research interests include robust speech recognition, speaker recognition, and source localization. He is a Member of the Institute of Electronics, Information and Communication Engineers (IEICE), and the Acoustical Society of Japan (ASJ).

Norihide Kitaoka received his B.E. and M.E. degrees from Kyoto University in 1992 and 1994, respectively, and a Dr. Engineering degree from Toyohashi University of Technology in 2000. He joined Denso Corporation, Japan, in 1994. He then joined the Department of Information and Computer Sciences at Toyohashi University of Technology as a Research Associate in 2001 and has been a Lecturer since 2003. His research interests include speech processing, speech recognition, and spoken dialog. He is a Member of the IEICE, the Information Processing Society of Japan (IPSJ), the ASJ, and the Japan Society for Artificial Intelligence (JSAI).

Seiichi Nakagawa received his B.E. and M.E. degrees from the Kyoto Institute of Technology, in 1971 and 1973, respectively, and Dr. of Engineering degree from Kyoto University in 1977. He joined the Faculty of Kyoto University, in 1976, as a Research Associate in the Department of Information Sciences. From 1980 to 1983, he had been an Assistant Professor, and from 1983 to 1990 he had been an Associate Professor. Since 1990 he has been a Professor in the Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi. From 1985 to 1986, he had been a Visiting Scientist in the Department of Computer Science, Carnegie-Mellon University, Pittsburgh, USA. He received the 1997/2001 Paper Award from the IEICE and the 1988 JC Bose Memorial Award from the Institution of Electronic Telecommunication Engineers His major interests in research include automatic speech recognition/speech processing, natural language processing, human interface, and artificial intelligence.