PAPER Special Section on Robust Speech Processing in Realistic Environments

Robust Speech Recognition by Combining Short-Term and Long-Term Spectrum Based Position-Dependent CMN with Conventional CMN

Longbiao WANG[†a)], ***Student Member***, **Seiichi NAKAGAWA**[†], ***Fellow***, ***and*** **Norihide KITAOKA**[††], ***Member***

SUMMARY In a distant-talking environment, the length of channel impulse response is longer than the short-term spectral analysis window. Conventional short-term spectrum based Cepstral Mean Normalization (CMN) is therefore, not effective under these conditions. In this paper, we propose a robust speech recognition method by combining a short-term spectrum based CMN with a long-term one. We assume that a static speech segment (such as a vowel, for example) affected by reverberation, can be modeled by a long-term cepstral analysis. Thus, the effect of long reverberation on a static speech segment may be compensated by the long-term spectrum based CMN. The cepstral distance of neighboring frames is used to discriminate the static speech segment (long-term spectrum) and the non-static speech segment (short-term spectrum). The cepstra of the static and non-static speech segments are normalized by the corresponding cepstral means. In a previous study, we proposed an environmentally robust speech recognition method based on Position-Dependent CMN (PDCMN) to compensate for channel distortion depending on speaker position, and which is more efficient than conventional CMN. In this paper, the concept of combining short-term and long-term spectrum based CMN is extended to PDCMN. We call this Variable Term spectrum based PDCMN (VT-PDCMN). Since PDCMN/VT-PDCMN cannot normalize speaker variations because a position-dependent cepstral mean contains the average speaker characteristics over all speakers, we also combine PDCMN/VT-PDCMN with conventional CMN in this study. We conducted the experiments based on our proposed method using limited vocabulary (100 words) distant-talking isolated word recognition in a real environment. The proposed method achieved a relative error reduction rate of 60.9% over the conventional short-term spectrum based CMN and 30.6% over the short-term spectrum based PDCMN.

key words: *robust speech recognition, distant-talking environment, CMN, long-term spectrum*

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 such microphones is undesirable for reasons of safety or convenience. Hands-free speech communication [1]–[3] has become more popular in special environments such as the office or a car. Unfortunately, in a distant-talking environment, channel distortion can drastically degrade the speech recognition performance. This is predominantly caused by the mismatch between the real and the training environments.

Compensating an input feature is the main way to reduce the mismatch. Cepstral Mean Normalization (CMN) is a simple and effective way of normalizing the feature space and thereby reducing channel distortion [4]–[7]. CMN reduces the errors caused by the mismatch between test and training conditions, and it is also very simple to implement. It has, therefore, been adopted in many current systems. In order to be effective for CMN, the length of the channel impulse response needs to be shorter than the short-term spectral analysis window which is usually 16 ms - 25 ms. However, the duration of the impulse response of reverberation usually has a much longer tail in a distant-talking environment. Therefore, conventional CMN, in which cepstral means are estimated from the entire current utterance using the short-term analysis window, is not effective under these conditions. Several studies have focused on decreasing the above problem. Raut et al. [8], [9] use preceding states as units of preceding speech segments, and by estimating their contributions to the current state using a maximum likelihood function, they adapt the models accordingly. In this paper, we address the effect of long reverberation by feature-based compensation method that is easier to be implemented. In [10], [11] a multiresolution channel normalization based speech recognition front end has been implemented by subtracting the mean of the log magnitude spectrum using a long-term spectral analysis window. At first, they used a long time window (high frequency resolution; 2 seconds) analysis and applied channel normalization. Then, they transformed the long-time representation to a short-time representation. Finally, cepstral domain features were computed for speech recognition. In this paper, we directly normalized the cepstral domain feature based on long-term spectrum corresponding to static speech signal and short-term spectrum corresponding to non-static speech signal for speech recognition in one step.

In this paper, we propose robust speech recognition by combining a short-term spectrum based CMN with a long-term spectrum based CMN, which we call *Variable-Term spectrum based CMN (VT-CMN)*. We assume that static speech segments (such as vowels, for example) affected by reverberation can be modeled by a long-term cepstral analysis. Thus, the effect of long reverberation on a static speech segment may be compensated by the long-term spectrum based CMN. For speech recognition, short-term and long-

Manuscript received July 2, 2007.
Manuscript revised September 6, 2007.

[†]The authors are with Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan.
[††]The author is with Nagoya University, Nagoya-shi, 464–8603 Japan.
a) E-mail: wang@slp.ics.tut.ac.jp

DOI: 10.1093/ietisy/e91–d.3.457

term cepstral coefficients are extracted *a priori*. The cepstral distance of neighboring frames is used to discriminate the static and non-static speech segments. A speech segment with a smaller variance between neighboring frames is detected as a static speech segment. The cepstra of static and non-static speech segments are normalized by the corresponding cepstral means.

In conventional CMN, the cepstral mean is previously estimated by averaging along the entire current utterance and is kept constant during the normalization. However, this off-line estimation involves a long delay that is likely to be unacceptable when the utterance is long. If the utterance is short, an accurate cepstral mean cannot be estimated. Various window CMN methods have been used to normalize the feature vectors in an on-line version [6], [7]. However, a tradeoff exists between delay and recognition error [7]. Thus, the usual CMN cannot achieve good recognition performance with a short delay. In our previous study [12], [13], we proposed a robust speech recognition method using a new real-time CMN based on speaker position, which we call Position-Dependent CMN (PDCMN). In this method, we measure the transmission characteristics (the compensation parameters for position-dependent CMN) from certain grid points in a room *a priori*. The system then adopts the compensation parameter corresponding to the estimated position, applies a channel distortion compensation method to the speech (that is, position-dependent CMN) and performs speech recognition. It is shown in [13] that PDCMN is more efficient for speech recognition in a distant-talking environment than conventional CMN. In this paper, position-dependent cepstral means are estimated from short-term cepstra using non-static speech segments and from long-term cepstra using static speech segments. The cepstra of the static and non-static speech segments are then subtracted from the corresponding cepstral means depending on the speaker position. We call this method *Variable-Term spectrum based PDCMN (VT-PDCMN)*.

PDCMN[†] or VT-PDCMN can indeed compensate efficiently for the channel transmission characteristics depending on speaker position, but cannot normalize the speaker variation because a position-dependent cepstral mean does not contain speaker characteristics. On the contrary, the conventional CMN can compensate for both the transmission and the speaker variations, but cannot achieve good recognition performance for short utterances because the sufficient phonemics balance cannot be obtained. Both variations perform additional operations in the cepstral domain. Thus, the combination of position-dependent cepstral mean and conventional cepstral mean may simultaneously compensate for the channel distortion and speaker variation effectively. In this paper, the sum of weights of position-dependent cepstral and conventional cepstral mean is set to 1 because the transmission characteristics should not be over normalized. In other words, since both the position-dependent cepstral mean and the conventional cepstral mean contain the channel transmission characteristics, the channel distortion would be normalized twice if each weight is set to 1. Indeed, we also conducted experiments using the various weights (the sum of weight was not equal 1) of two kinds of cepstral mean, the results became worse because the amplitude of weight-sum of two kinds of cepstral mean mismatched the real value.

In this paper, we propose a robust distant speech recognition by combining PDCMN/VT-PDCMN with conventional CMN to address the above problems. The *a priori* estimated position-dependent cepstral mean is linearly combined with an utterance-wise cepstral mean using the following two combination methods. The first method uses a fixed weighting coefficient over the whole test data to obtain the combinational CMN, and this is called *fixed-weight combinational CMN*. However, the optimal weight seems to depend on the speaker position and the length of the utterance to be recognized. Thus, a fixed weighting coefficient does not obtain the optimal result. A variable weighting coefficient may produce better performance. A single input feature compensated by the combinational cepstral means with different weighting coefficients generates multiple input features. Thus, the problem becomes how to obtain the optimal performance for the given multiple input features. Voting on the different hypotheses generated from the multiple input features has been studied in [12], [14]. In [15], a new algorithm to select a suitable channel for speech recognition using the output of the speech recognizer has been proposed. All the methods discussed above use the output hypotheses generated by multiple decoders to estimate the final result. In our previous study [13], we proposed the combination of multiple input streams at frame level using a single decoder. In this paper, we extend this method to the combination of PDCMN/VT-PDCMN and conventional CMN. The second method for obtaining the combinational CMN involves calculating the output probability of each input feature at frame level, and a single decoder using these output probabilities is used to perform speech recognition. This is called *variable-weight combinational CMN* and is very easy to implement in both isolated word recognition systems and continuous speech recognition systems.

Section 2 describes the combination of short-term and long-term spectrum based CMN. An environmentally robust real-time Position-Dependent CMN (PDCMN) and *Variable-Term spectrum based PDCMN (VT-PDCMN)* are described in Sect. 3. The combination of PDCMN/VT-PDCMN and conventional CMN is proposed in Sect. 4. Section 5 describes the experimental results of distant-talking speech recognition in a real environment. Finally, Sect. 6 summarizes the paper and describes future work.

2. Variable-Term Spectrum Based CMN

2.1 Conventional Short-Term Spectrum Based CMN

A simple and effective way of channel normalization is to

[†]For the sake of convenience, CMN refers to short-term spectrum based CMN, and PDCMN refers to short-term spectrum based PDCMN in this paper.

subtract the mean of each cepstral coefficient (CMN) [4], [5], [16], which removes time-invariant distortions caused by the transmission channel and the recording device.

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

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

We, however, conducted our experiments in a silent seminar room, with the result that the effect of noise is ignored in this paper. So Eq. (1) becomes $x[l] = h[l] \otimes s[l]$.

CMN has been used to compensate for the convolution distortion. In order for CMN to be effective, the length of the impulse response has to be shorter than the short-term spectral analysis window. However, in a distant-talking environment, the length of impulse response is longer than the short-term spectral analysis window, and therefore the late effect of impulse response cannot be compensated.

To analyze the effect of impulse response, the impulse response $h[l]$ can be separated into two parts $h_1[l]$ and $h_2[l]$ as

$$h_1[l] = \begin{cases} h[l] & 1 < \mathrm{L} \\ 0 & \text{otherwise} \end{cases}, h_2[l] = \begin{cases} h[l+L] & 1 \geq 0 \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where L is the length of the spectral analysis window, and $h[l] = h_1[l] + \delta(l-L) \otimes h_2[l]$. $\delta()$ is a dirac delta function (that is, unit impulse function). The formula (1) can be rewritten as

$$x[l] = s[l] \otimes h_1[l] + s[l-L] \otimes h_2[l], \tag{3}$$

where the early effect is within a frame (analysis window), and the late effect is over multiple frames.

In [17], the early term of Eq. (3) was compensated by conventional CMN, whereas the late term of Eq. (3) was treated as additive noise, and a noise reduction technique based on spectrum subtraction was applied. In this paper, we focus on increasing the length of the analysis window L, which reduces the early effect of the impulse response (that is, the first term of Eq. (3)) as much as possible.

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

$$C^x = C^h + C^s, \tag{4}$$

where C^x, C^h and C^s express the cepstra 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:

$$\tilde{C}_t = C_t^x - \Delta C, \qquad (t = 0, \ldots, T), \tag{5}$$

$$\Delta C \approx \bar{C}^x - \bar{C}^{train}, \tag{6}$$

where $\tilde{C}_t$ and C_t^x are the compensated and original cepstra at time frame t, and $\bar{C}^x$ and $\bar{C}^{train}$ are the cepstral means of utterances to be recognized and those to be used to train the speaker-independent acoustical model, respectively.

2.2 Combination of Short-Term and Long-Term Spectrum Based CMN

In the traditional method, a short-term cepstral analysis is used. However, the duration of impulse response of reverberation usually has a much longer tail in a distant-talking environment. Therefore, conventional CMN is not effective under these conditions.

For the static part of speech signals, the spectrum can be extracted by the long-term analysis window because the speech signal is stationary. We assume that a static speech segment affected by long reverberation can be modeled by the long-term spectrum based CMN. Thus, the effect of long reverberation on a static speech signal may be compensated by the long-term spectrum based CMN. On the other hand, for the non-static part of speech signals, the Fourier transform cannot be applied to a long-term analysis window because the long-term speech signal is not stationary. This result in long-term analysis window based spectrum yields too low time resolution for transient speech. Thus, the long-term CMN cannot be applied to non-static part of speech signals because long-term cepstral mean is not available too. In the case of a non-static speech segment, the traditional short-term spectrum based CMN is used. Thus, the combination of short-term and long-term spectrum based CMN is defined as [18]:

$$\begin{aligned} \tilde{C}_t = C_t^x - \Delta C &= \begin{cases} C_t^{x\_short} - \Delta C^{short} \\ C_t^{x\_long} - \Delta C^{long} \end{cases} \\ &= \begin{cases} C_t^{x\_short} - (\bar{C}^{x\_short} - \bar{C}^{train\_short}) \\ C_t^{x\_long} - (\bar{C}^{x\_long} - \bar{C}^{train\_long}) \end{cases} \\ &\quad \begin{cases} \text{if } t\text{-}th \text{ speech segment is non-static} \\ \text{if } t\text{-}th \text{ speech segment is static} \end{cases}, \end{aligned} \tag{7}$$

where $C_t^{x\_short}$ and $C_t^{x\_long}$ are the original short-term and long-term cepstra at time frame t, $\bar{C}^{x\_short}$ and $\bar{C}^{x\_long}$ are short-term and long-term cepstral means of utterances to be recognized, and $\bar{C}^{train\_short}$ and $\bar{C}^{train\_long}$ are short-term and long-term cepstral means of utterances to be used to train the speaker-independent acoustical model, respectively.

2.3 Static and Non-static Speech Segment Detection

Test and training utterances include static and non-static speech. In order to estimate the cepstral means of static and non-static speech segments and to normalize the corresponding cepstral features, static speech segment detection and non-static speech segment detection are necessary and important for speech recognition using the proposed method. It is well known that a static speech segment has

smaller variance between neighboring frames than a non-static speech segment. To discriminate the static and non-static speech segments, the cepstral distance of neighboring frames is defined as:

$$D(C_t, C_{t+1}) = \sum_{m=1}^{M} |C_t^m - C_{t+1}^m|, \tag{8}$$

where C_t^m is the m-th cepstrum of the *t-th* frame. A speech segment is more likely to be a static speech segment when the cepstral distance between neighboring frames is small. For the sake of simplicity, a certain percentage of speech segments with smaller cepstral distances is identified as static speech segments in this paper.

3. Variable-Term Spectrum Based Position-Dependent CMN

3.1 Position-Dependent CMN

In a real distant-talking environment, the transmission characteristics of different speaker positions differ because of the distance between the speaker and the microphone, and the reverberation of the room [19]. In [12], [13], we proposed an environmentally robust speech recognition method based on Position-Dependent CMN (PDCMN). For PDCMN, the compensation parameter in Eq. (6) is defined by:

$$\Delta C = \bar{C}^{position} - \bar{C}^{train}, \tag{9}$$

where $\bar{C}^{position}$ is the cepstral mean of utterances affected by the transmission characteristics between a certain position and the microphone. Both $\bar{C}^{position}$ and $\bar{C}^{train}$ are estimated from short-term cepstra. In our experiments in Sect. 5, we divide the room into 12 areas as shown in Fig. 1 and measure the $\bar{C}^{position}$ corresponding to each area.

We measure the transmission characteristics (the compensation parameters for position-dependent CMN) from some grid points in the room *a priori*. The system estimates the speaker position in a 3-D space based on microphone arrays [19]. Four microphones are arranged in a T-shape on a plane, and the sound source position is estimated by Time Delay of Arrival (TDOA) among the microphones [20]–[22]. The system then adopts the compensation parameter corresponding to the estimated position, applies a channel distortion compensation method to the speech (that is, position-dependent CMN), and performs speech recognition.

3.2 Combination of Short-Term and Long-Term Spectrum Based PDCMN

We extend the concept of combining short-term and long-term spectrum based CMN to PDCMN, which we call Variable-Term spectrum based PDCMN (VT-PDCMN). $\bar{C}^{position}$ and $\bar{C}^{train}$ are both estimated by averaging short-term cepstra obtained from non-static speech segments and long-term cepstra obtained from static speech segments.

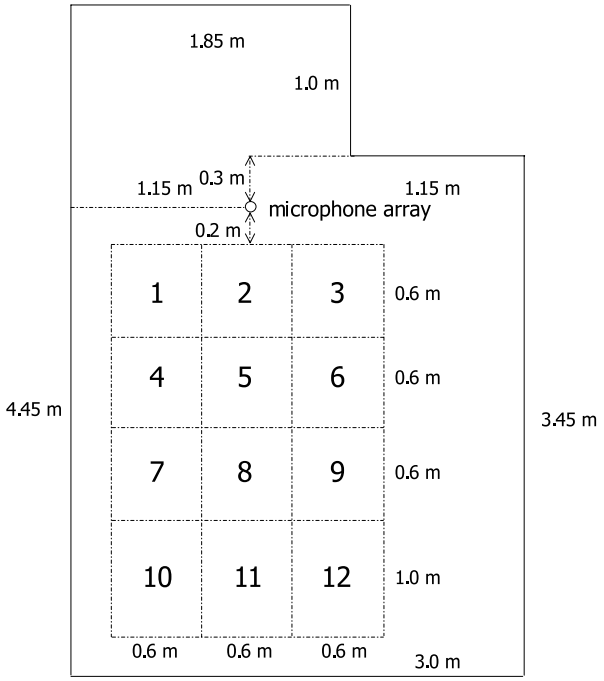


Fig. 1 Room configuration. (room size: (W) 3 m × (L) 3.45 m × (H) 2.6 m)

The cepstrum of the *t-th* speech segment C_t^x is compensated by $\Delta C = \bar{C}^{position} - \bar{C}^{train}$, while the corresponding $\bar{C}^{position}$ and $\bar{C}^{train}$ are selected as:

$$\bar{C}^{position} = \begin{cases} \bar{C}^{position\_short} & \text{if } t\text{-}th \text{ segment is non-static} \\ \bar{C}^{position\_long} & \text{if } t\text{-}th \text{ segment is static} \end{cases}, \tag{10}$$

$$\bar{C}^{train} = \begin{cases} \bar{C}^{train\_short} & \text{if } t\text{-}th \text{ segment is non-static} \\ \bar{C}^{train\_long} & \text{if } t\text{-}th \text{ segment is static} \end{cases}, \tag{11}$$

where $\bar{C}^{position\_short}$, $\bar{C}^{position\_long}$ are short-term and long-term cepstral means of utterances emitted from a certain position, and $\bar{C}^{train\_short}$ and $\bar{C}^{train\_long}$ are short-term and long-term cepstral means of utterances to be used to train the speaker-independent acoustical model, respectively.

4. Combination of PDCMN/VT-PDCMN and Conventional CMN

4.1 Fixed-Weight Combinational CMN

To compensate the channel distortion and speaker characteristics simultaneously, a short-term or variable-term spectrum based position-dependent cepstral mean is combined linearly with the conventional cepstral mean [23]. The new compensation parameter ΔC for combinational CMN is defined by:

$$\begin{aligned} \Delta C &= \lambda(\bar{C}^{position} - \bar{C}^{train}) + (1-\lambda)(\bar{C}_t^x - \bar{C}^{train}) \\ &= \lambda \bar{C}^{position} + (1-\lambda)\bar{C}_t^x - \bar{C}^{train}, \end{aligned} \tag{12}$$

where λ denotes a weighting coefficient. In Sect. 4, $\bar{C}_{position}$ and $\bar{C}_{train}$ are estimated by averaging short-term cepstra for PDCMN given by Eq. (9) and variable-term cepstra for VT-PDCMN given by Eqs. (10) and (11). When a fixed λ is used for the entire test data, the method is known as *fixed-weight combinational CMN*.

4.2 Variable-Weight Combinational CMN

In Sect. 4.1, a fixed weighting coefficient λ is used to combine PDCMN/VT-PDCMN with the conventional CMN. The effect of the channel distortion (that is, position-dependent cepstral mean) depends on speaker position and the confidence of the estimated speaker characteristics (that is, the conventional cepstral mean) depends on the length of the utterance. Therefore, the weighting coefficient λ should be adjusted according to the speaker position and the length of the utterance. A single input feature compensated by the combinational cepstral means with different weighting coefficients generates multiple input features. Thus, the problem becomes how to obtain the optimal performance given the multiple input features.

Given a set of variable weights λ_k, an automatic decision algorithm for the optimal weighting coefficient λ is required. In a previous study [13], we proposed an optimal input decision algorithm, which calculates the output probability of each input stream at frame level and selects the input with maximum probability as the optimal input. We extend and modify this algorithm to the so-called *variable-weight combinational CMN*. Indeed, the proposed *variable-weight combinational CMN* can automatically select the optimal weight coefficient at frame level from within the range of given weight coefficients.

For multiple inputs, a conventional Viterbi algorithm [24] is used for each input stream, k. 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 the *k-th* input ($1 \le k \le K$) and ends in state j is defined by:

$$\alpha(t, j, k) = \max_{1 \le i \le S} \{\alpha(t-1, i, k) a_{ij} b_j(O_k(t))\}, \tag{13}$$

$$O_k(t) = \tilde{C}_t - (\lambda_k \bar{C}^{position} + (1-\lambda_k)\bar{C}_t^x - \bar{C}^{train}).$$

where $a_{ij} = P(s_t = j | s_{t-1} = i)$ is the transition probability from state i to state j, $1 \le i, j \le S$, $2 \le t \le T$; and $b_j(O_k(t))$ is the output probability for an observation sequence $O_k(t)$ at state j. λ_k is the *k-th* weighting coefficient.

In this conventional multiple-decoder method, the Viterbi algorithm is performed for each input stream independently, resulting in a computational complexity of K (the number of input streams). Thus, both the calculation of output probability and the rest of the processing costs 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, we modify Eq. (13) as follows:

$$\alpha(t, j) = \max_{1 \le i \le S} \{\alpha(t-1, i) a_{ij} \max_k b_j(O_k(t))\}. \tag{14}$$

This method is called *single decoder processing*. In Eq. (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 an extra $K-1$ times the calculation of only the output probability is required compared to that of a single input. Furthermore, the derivatives of the K input cepstrums (Δcepstrum) compensated by different combinational cepstral means have the same values. Thus, the calculation depending only on the derivatives can be shared by the input streams.

5. Experiments

5.1 Experimental Setup

We performed the experiment in a room, measuring 3.45 m× 3 m × 2.6 m, without additive noise, as shown in Fig. 2. The room was divided into the 12 (3×4) rectangular areas shown in Fig. 1, where each area is 60 cm×60 cm. We measured the transmission characteristics (that is, the mean cepstrums of utterances recorded *a priori*) from the center of each area. For our experiments, the room was set up as the seminar room shown in Fig. 2 with a whiteboard beside the left wall, a table and a few chairs in the center of the room, a TV and some other tables, etc. The reverberation time of this room was about 150 ms.

4 microphones in a T-shape as shown in Fig. 3 were used. The first microphone (M1) was regarded as the reference and was placed at the origin of the coordinate system. The distance between a pair of microphones was 20 cm. In this paper, the first microphone (M1) was used for single microphone processing, and 4 microphones in a T-shape were used for microphone array processing (delay-and-sum beamforming [25], [26]). Delay-and-sum 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. Therefore beamforming can not only separate multiple sound sources but

Fig. 2 Experimental environment.

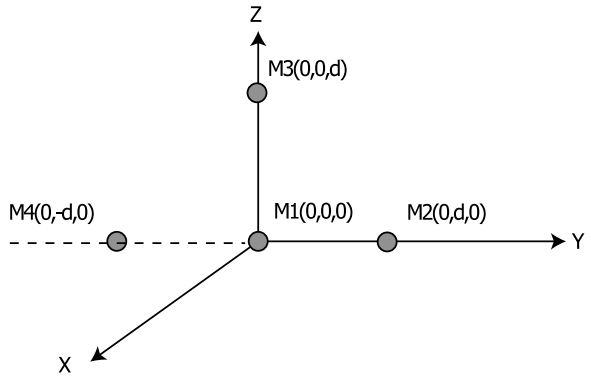


Fig. 3 Setup of microphones. (d = 20 cm)

also suppress reverberation for the speech source of interest.

In our method, the estimated speaker position is used to determine the area (60 cm × 60 cm) in which the speaker should be. In a past study [19], we revealed that speaker positions 1, 5 and 9 shown in Fig. 1 could be estimated with estimation errors of less than 10 cm by the T-shaped microphone system shown as Fig. 3, and that 99.8% positions would be within the correct area. In the present study, therefore, we assumed that the position area was accurately estimated, and we evaluated only our proposed speech recognition methods.

Twenty male speakers each with a close-microphone, uttered 200 isolated words. The 200 isolated words are phonetic balance common isolated words selected from Tohoku University and Panasonic isolated spoken word database [27]. The average time of all utterances was about 0.6 seconds. For each speaker, the first 100 words were used as test data and the remainder for estimating the cepstral mean $\bar{C}^{position}$ in the short-term spectrum based on Eq. (9) and the variable-term spectrum based on Eq. (10)†. To simulate the utterances spoken at various positions, all the utterances were emitted by a loudspeaker located in the center of each area and recorded for test purposes and the estimation of $\bar{C}^{position}$. The sampling frequency was 12 kHz. The frame length was 21.3 ms (256-point) for a short-term cepstrum, and 37.3 ms (448-point) for a long-term cepstrum. To compensate for the effect of long reverberation, it seems that longer analysis window is more effective. However, there exists a tradeoff between temporal resolution and frequency resolution. The longer the analysis window was, the worse the temporal resolution was. Furthermore, the speech segment (even the vowel etc.) should no longer be a stationary signal if the analysis window it too long, the recognition performance would became worse in that case. In this paper, the length of long-term analysis window was empirically determined. The result based on 448-point window obtained the best performance. By the way, the result based on 448-point window was almost the same as that based on 512-point window and was significantly better than that based on other length of windows. A frame shift of 8 ms (96-point) was used for both short-term and long-term cepstra. The number of filters in a filter bank was 24, and the length of cepstral liftering was 22. Then, 116 Japanese speaker-independent syllable-based HMMs (strictly speaking, mora-unit HMMs [28]) 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 comprised 10 MFCCs. First- and second-order derivatives of the cepstra plus first and second derivatives of the power component were also included.

Table 1 The individual speech recognition results for short-term and long-term cepstra. Cepstral means were estimated from 100 isolated words for each speaker. (single microphone: %)

Recognition method	short-term spectrum based CMN	long-term spectrum based CMN	
Acoustic model	short-term cepstrum	short-term cepstrum	long-term cepstrum
Area 10	95.0	94.3	94.1
Area 11	95.1	94.8	94.4
Area 12	95.2	93.6	94.0
Ave.	95.1	94.2	94.2

When using the *variable-weight combinational CMN*, the optimal weighting coefficient was not empirically determined for the entire test data or development data as in *fixed-weight combinational CMN*, but was automatically selected at frame level from within the range of given weight coefficients. In this paper, the number of weight coefficient K was set as 3. For the single microphone, λ_1, λ_2 and λ_3 were set as 0.6, 0.7 and 0.8, respectively. For the microphone array, λ_1, λ_2 and λ_3 were set as 0.4, 0.5 and 0.6, respectively.

5.2 Preliminary Experimental Results Based on the Combination of Short-Term and Long-Term Spectrum Based CMN

We conducted the preliminary speech recognition experiment using a single microphone and combining short-term and long-term spectrum based CMN as proposed in Sect. 2. The utterances emitted by a loudspeaker located in areas 10, 11 and 12 as shown in Fig. 1 were used as test data.

The individual results for the single microphone based on short-term cepstrum and long-term cepstrum are compared in Table 1. Cepstral means were estimated from 100 isolated words for each speaker, and then CMN was performed. The results based on long-term cepstrum were worse than those based on short-term cepstrum because numerous speech segments of test data were not static signals and could not be analyzed by the long-term window (≈ 37.3 ms). For static signals analyzed by the long-term window, both the short-term HMMs and long-term HMMs were used as acoustic models. Since a considerable number of the speech segments of training data was not static and could not be analyzed by the long-term window, parameters of the long-term HMMs could not be estimated ac-

†For the speech recognition method combining short-term and long-term spectrum based CMN, estimation of a position-dependent cepstral mean $\bar{C}^{position}$ is not necessary.

curately. Thus, the results based on long-term HMMs were slightly worse than those based on short-term HMMs. In the following part of this paper, the same short-term syllable-based HMMs were used as acoustic models for both the short-term spectrum based CMN and the long-term spectrum based CMN.

The results of combining short-term and long-term spectrum based CMN are shown in Table 2. Cepstral means were estimated from 1 word, 10 words and 100 words for each speaker. 30% of speech segments with a smaller cepstral distance were identified as static speech segments, and this was empirically determined. For CMN with 1 word, since the average duration of the static speech signal of all utterances was too short (about 0.6 second $\times$ 30% = 0.18 second), accurate cepstral means could not be estimated for the static speech signal (long-term spectrum). Thus, the combination of short-term and long-term spectrum based CMN did not improve recognition performance for short utterances. For CMN with 10 words or 100 words, the proposed combination method effectively improved recognition performance. The experimental results also show that the longer the speech data which is used to estimate the cepstral mean, the greater the improvement. The proposed combination of the short-term and long-term spectrum based CMN using 100 words for cepstral mean estimation achieved a 14.3% relative error reduction rate over the conventional short-term spectrum based CMN.

5.3 Experimental Results Based on the Combination of PDCMN/VT-PDCMN and Conventional CMN

The variable-term spectrum based CMN (that is, the combination of short-term and long-term spectrum based CMN) improved the recognition rate when the length of the utterance to be recognized was long enough. However, this precludes real-time processing of speech recognition. Furthermore, the variable-term spectrum based CMN degraded the recognition rate when the length of utterance to be recognized was too short. In this section, we conducted the experiments based on a combination of environmentally robust real-time PDCMN/VT-PDCMN and conventional CMN. Both the single microphone and the T-shape 4 microphone array were used. Short-term cepstral means were estimated from one isolated word (about 0.6 seconds) for conventional CMN.

The average results of all 12 areas based on a combination of PDCMN/VT-PDCMN and conventional CMN for both the single microphone and the microphone array are summarized in Table 3. The detailed experimental results for every area are shown in Table 4 for the single microphone and Table 5 for the microphone array. By compensating the transmission characteristics using the compensation parameters measured *a priori* from sufficient utterances for each area, the short-term spectrum based PDCMN given by Eq. (9) effectively improved the speech recognition performance in all 12 areas for both the single microphone and the microphone array, compared to the conventional CMN. For the microphone array, the conventional CMN and the proposed PDCMN were applied after the delay-and-sum beamforming. The proposed method outperformed the conventional CMN (that is, a typical channel normalization method for dereverberation), microphone array processing (that is, a spatial filtering for dereverberation) and the combination method of conventional CMN and microphone array processing. Furthermore, CMN based dereververation methods are easy to be combined with many other dereverberation methods such as representations RelAtive SpecTrA (RASTA) filtering, low-pass AutoRegressive Moving Average (ARMA) filtering [29]–[31], etc. RASTA applies a band-pass filter to the energy in each frequency subband in order to smooth over noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel [29], [32]. The band-pass nature of the RASTA filter and mean subtraction of CMN both result in a feature vector stream with mean of zero. In many cases, the performance based on CMN was similar to that based on RASTA under convolutional noise [33], [34]. In [33], RASTA filtering and CMN were examined as methods for normalization. Experiments showed that RASTA filtering results in slightly better performance on

Table 3 Speech recognition results for the combination of PDCMN/VT-PDCMN and conventional CMN (%).

	Single microphone	Microphone array
W/O CMN	90.1	91.4
Conv. CMN	92.9	93.6
PDCMN	**95.8**	**96.4**
VT-PDCMN	96.2	96.7
PDCMN + Conv. CMN (fixed-weight)	96.3	96.9
VT-PDCMN + Conv. CMN (fixed-weight)	96.6	97.1
VT-PDCMN + Conv. CMN (variable-weight)	**97.0**	**97.5**

Table 2 Speech recognition results for the combination of short-term and long-term spectrum based CMN. Cepstral means were estimated from 1 word, 10 words and 100 words for each speaker. (single microphone: %)

Area	1 word		10 words		100 words	
	Conv. CMN	proposed	Conv. CMN	proposed	Conv. CMN	proposed
10	90.4	88.6	94.4	94.8	95.0	95.7
11	90.9	88.5	94.2	94.9	95.1	96.1
12	89.9	88.7	94.8	94.7	95.2	95.6
Ave.	90.4	88.6	94.5	94.8	95.1	95.8

Table 4 Speech recognition results for the combination of PDCMN/VT-PDCMN and conventional CMN using a single microphone (%).

Area	Conv. CMN	PD-CMN	VT-PDCMN	PDCMN + Conv. CMN (fixed-weight)	VT-PDCMN + Conv. CMN (fixed-weight)	VT-PDCMN + Conv. CMN (variable-weight)
1	93.1	96.6	97.0	96.9	97.2	97.8
2	96.1	97.8	97.7	97.9	98.1	98.6
3	94.9	96.6	97.2	97.1	97.6	98.1
4	93.6	95.9	96.2	96.2	96.5	97.0
5	94.4	96.6	97.1	97.2	97.6	97.8
6	93.7	96.0	96.4	97.0	97.1	97.5
7	92.4	95.8	95.9	96.4	96.6	97.0
8	91.1	95.1	95.2	95.2	95.7	96.1
9	93.8	96.9	96.9	97.4	97.3	97.7
10	90.4	94.4	94.6	94.4	95.1	95.5
11	90.9	94.0	95.3	94.9	95.4	95.6
12	89.9	93.6	94.6	94.7	95.2	95.5
Ave.	92.9	95.8	96.2	96.3	96.6	97.0

Table 5 Speech recognition results for the combination of PDCMN/VT-PDCMN and conventional CMN using a microphone array (%).

Area	Conv. CMN	PD-CMN	VT-PDCMN	PDCMN + Conv. CMN (fixed-weight)	VT-PDCMN + Conv. CMN (fixed-weight)	VT-PDCMN + Conv. CMN (variable-weight)
1	94.8	97.2	97.5	97.7	97.8	98.2
2	96.0	98.1	98.2	97.9	98.3	98.6
3	94.7	96.9	97.3	97.7	98.0	98.3
4	93.1	95.9	96.6	96.7	97.3	97.6
5	94.5	97.0	97.3	98.0	97.8	98.4
6	94.6	97.3	97.1	97.7	97.9	98.2
7	94.1	96.8	97.0	96.9	97.0	97.5
8	93.0	95.9	96.4	96.2	96.8	96.9
9	93.8	96.5	97.0	97.2	97.1	97.7
10	91.5	94.8	95.7	95.4	95.6	96.1
11	91.9	95.5	95.7	95.8	95.9	96.7
12	91.4	94.4	94.7	95.2	95.4	95.8
Ave.	93.6	96.4	96.7	96.9	97.1	97.5

the unconstrained monophone task than CMN. In [34], the classical RASTA filtering resulted in decreased recognition performance when compared to CMN. Phase-corrected RASTA reached the same performance level as obtained for CMN for a medium and large vocabulary continuous speech recognition task. The phase-corrected RASTA is a technique that consists of classical RASTA filtering followed by a phase correction operation. In some cases, the combination of CMN and RASTA can give better results than either of the techniques alone [30], [32]. Therefore, our proposed method is effective than some typical dereververation techniques such as the conventional CMN and the delay-and-sum beamforming. Moreover, the proposed method is easy to be combined with many other dereverberation methods such as beamforming, RASTA filtering, ARMA filtering, etc., and a furthermore improvement should be obtained. Thus, in this paper, we did not compare our proposed method with other dereverberation methods such as RASTA filtering, ARMA filtering, etc.

Since the effect of long reverberation on a static speech segment could be compensated by the long-term spectrum based CMN, the combination of short-term spectrum based PDCMN and long-term spectrum based PDCMN (that is, Variable-Term spectrum based PDCMN (VT-PDCMN)) further improved the speech recognition performance. VT-PDCMN achieved a relative error reduction rate of 9.5% over PDCMN for the single microphone and 8.3% over PDCMN for the microphone array. 40% of speech segments for the single microphone and 30% of speech segments for the microphone array with smaller cepstral distances were identified as static speech segments, and this was empirically determined.

The combination of short-term spectrum based PDCMN and conventional CMN with fixed-weight compensated the channel distortion and speaker characteristics simultaneously, so an 11.9% relative error reduction rate was achieved over PDCMN for the single microphone and a 13.9% relative error reduction rate was achieved over PDCMN for the microphone array. When VT-PDCMN was combined with conventional CMN using fixed-weight, it achieved a relative error reduction rate of 19.0% over PDCMN for the single microphone and 19.4% over PDCMN for the microphone array. The best average performance was obtained with the weight coefficient $\lambda = 0.7$ for the single microphone and $\lambda = 0.5$ for the microphone array.

Finally, the combination of VT-PDCMN and conventional CMN with variable-weight achieved the best recognition performance of all the methods because the optimal weighting coefficients were selected at each frame in an utterance. In other words, when using the *variable-weight combinational CMN*, the optimal weighting coefficient was not empirically determined for the entire test data or development data as in *fixed-weight combinational CMN*, but was automatically selected at frame level from within the range of given weight coefficients. For the single microphone, a 4.1% improvement (57.7% relative error reduction rate) over conventional CMN, and a 1.2% (28.6% relative error reduction rate) over PDCMN were achieved. For the microphone array, a 3.9% improvement (60.9% relative error reduction rate) over conventional CMN, and a 1.1% (30.6% relative error reduction rate) over PDCMN were achieved. The computational cost of the *variable-weight combinational CMN* was only 1.26 times that of the other methods even when 3 input streams were used.

6. Conclusion and Future Work

In a distant-talking environment, the length of channel impulse response is longer than the short-term spectral analysis window which is usually 16 ms - 25 ms. Therefore, conventional short-term spectrum based CMN is not effective in these conditions. We have proposed a robust distant-talking speech recognition method by combining a short-term spectrum based CMN with a long-term spectrum based CMN. We have assumed that a static speech segment affected by reverberation can be modeled by a long-term cepstral analysis. Thus, the effect of long reverberation on a static speech segment may be compensated by the long-term spectrum based CMN. The cepstral distance of neighboring frames is used to discriminate the static speech segment and non-static speech segment. The cepstra of static and non-static speech segments are normalized by the corresponding cepstral means. In this paper, the concept of variable-term spectrum based CMN has been extended to a robust speech recognition method based on Position-Dependent CMN (PDCMN) to compensate for channel distortion depending on speaker position. We call this method Variable-Term spectrum based PDCMN (VT-PDCMN). Since PDCMN/VT-PDCMN cannot normalize speaker variation, we have further combined PDCMN/VT-PDCMN with conventional CMN to compensate simultaneously for the channel distortion and speaker characteristics. The short-term spectrum or variable spectrum based position-dependent cepstral mean is combined linearly with a conventional cepstral mean using the following two types of processing. The first method uses a fixed weighting coefficient over the whole test data to obtain the combinational CMN, and this is called *fixed-weight combinational CMN*. The second method calculates the output probability of multiple features compensated by a variable weighting coefficient at each frame, and a single decoder using these output probabilities is used to perform speech recognition. This is called *variable-weight combinational CMN*. We conducted the experiments of our proposed method using limited vocabulary (100 words) distant-talking isolated word recognition in a real environment. The combination of VT-PDCMN and conventional CMN with variable-weight achieved a relative error reduction rate of 60.9% over the conventional short-term spectrum based CMN and 30.6% over the short-term spectrum based PDCMN using a T-shape 4 microphone array.

In our future work, we aim to subtract the late term of Eq. (3) based on spectrum subtraction and to normalize the early term of Eq. (3) by combining variable-term spectrum based PDCMN with conventional CMN as proposed in this paper.

Acknowledgments

This work was supported by The Global COE Program "Frontiers of Intelligent Human Sensing", from the ministry of Education, Culture, Sports, Science and Technology.

References

[1] 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 Trans. Speech Audio Process., vol.7, no.3, pp.346–349, May 1999.

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

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

[4] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust. Speech Signal Process., vol.29, no.2, pp.254–272, 1981.

[5] F. Liu, R. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition," Proc. ARPA Speech and Nat. Language Workshop, pp.69–74, 1993.

[6] A. Vikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," Speech Commun., vol.25, no.1-3, pp.133–147, 1998.

[7] P. Pujol, D. Macho, and C. Nadeu, "On real-time mean-and-variance normalization of speech recognition features," Proc. ICASSP-2006, pp.773–776, 2006.

[8] C. Raut, T. Nishimoto, and S. Sagayama, "Model adaptation by splitting of HMM for long reverberation," Proc. INTERSPEECH-2005, pp.277–280, 2005.

[9] C. Raut, T. Nishimoto, and S. Sagayama, "Adaptation for long convolutional distortion by maximum likelihood based state filtering approach," Proc. ICASSP-2006, vol.1, pp.1133–1136, 2006.

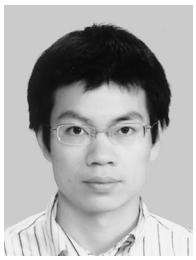
[10] C. Avendano, Temporal processing of speech in a time feature space, Ph.D. Thesis, Oregon Graduate Institute of Science & Technology, April 1997.

[11] C. Avendano, S. Tibrewala, and H. Hermansky, "Multiresolution channel normalization for ASR in reverberation environments," Proc. EUROSPEECH-1997, pp.1107–1110, 1997.

[12] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speech recognition based on position dependent CMN using a novel multiple microphone processing technique," Proc. EUROSPEECH-2005, pp.2661–2664, 2005.

[13] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speech recognition by combining multiple microphone-array processing

with position-dependent CMN," EURASIP J. Appl. Signal Process., vol.2006, Article ID 95491, pp.1–11, 2006.
[14] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," IEEE ASRU Workshop, pp.347–352, 1997.
[15] Y. Obuchi, "Mixture weight optimization for dual-microphone MFCC combination," IEEE ASRU Workshop, pp.325–330, 2005.
[16] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Am., vol.55, pp.1304–1312, 1974.
[17] Q. Jin, Y. Pan, and T. Schultz, "Far-field speaker recognition," Proc. ICASSP-2006, vol.1, pp.937–940, 2006.
[18] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust speech recognition by combining short-term spectrum based CMN with long-term spectrum based CMN," The Japan-China Joint Conference on Acoustics (JCA2007), P–2–13, June 2007.
[19] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM," Speech Commun., vol.49, no.6, pp.501–513, June 2007.
[20] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-24, no.4, pp.320–327, Aug. 1976.
[21] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," IEEE Trans. Speech Audio Process., vol.5, no.3, pp.288–292, 1997.
[22] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localisation in noisy and reverberant acoustic environments," EURASIP J. Applied Signal Processing, vol.2003, no.11, pp.1110–1124, Oct. 2003.
[23] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speech recognition by combining position-dependent CMN with Conventional CMN," Proc. ICASSP-2007, vol.4, pp.817–820, 2007.
[24] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," IEEE Trans. Inf. Theory, vol.13, no.2, pp.260–269, 1967.
[25] B. Van Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," IEEE Acoust. Speech Signal Process. Mag., vol.5, no.2, pp.4–24, April 1988.
[26] J. Flanagan, J. Johnston, R. Zahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," J. Acoust. Soc. Am., vol.78, pp.1508–1518, June 1985.
[27] S. Makino, K. Niyada, Y. Mafune, and K. Kido, "Tohoku University and Panasonic isolated spoken word database," J. Acoust. Soc. Jpn., vol.48, no.12, pp.899–905, Dec. 1992.
[28] S. Nakagawa, K. Hanai, K. Yamamoto, and N. Minematsu, "Comparison of syllable-based HMMs and triphone-based HMMs in Japanese speech recognition," Proc. International Workshop on Automatic Speech Recognition and Understanding, pp.393–396, 1999.
[29] H. Hermansky and N. Morgan, "RASTA processing of speech," IEEE Trans. Speech Audio Process., vol.2, no.4, pp.578–589, Oct. 1994.
[30] X., Xiao, E. Chng, and H. Li, "Normalizing the speech modulation spectrum for robust speech recognition," Proc. ICASSP-2007, vol.4, pp.1021–1024, 2007.
[31] C. Chen, J. Blimes, and K. Kirchhoff, "Low-resource noise-robust feature post-processing on AURORA 2.0," Proc. ICSLP-2004, pp.2445–2448, 2004.
[32] H. Hermansky and N. Morgan, "RASTA processing of speech," Proc. 1993 IEEE Speech Recogn. Workshop, Snowbird, UT, Dec. 1993.
[33] B. Milner, "A comparason of front-end configurations for robust speech recognition," Proc. ICASSP-2002, vol.1, pp.797–800, 2002.
[34] J. Veth and L. Boves, "On the efficiency of classical RASTA filtering for continuous speech recognition: Keeping the balance between acoustic pre-processing and acoustic modelling," Speech Commun., vol.39, no.3-4, pp.269–286, Feb. 2003.



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 worked at the China Construction Bank. His research interests include robust speech recognition, speaker recognition and source localization. He is a member of the Acoustical Society of Japan (ASJ).

Seiichi Nakagawa received a Dr. of Eng. 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. He moved to Toyohashi University of Technology in 1980. From 1980 to 1983 he was an Assistant Professor, and from 1983 to 1990 he was 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 was 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 Electro. Telecomm. Engrs. His major interests in research include automatic speech recognition/speech processing, natural language processing, human interface, and artificial intelligence. He is a fellow of the Information Processing Society of Japan (IPSJ).

Norihide Kitaoka received his B.E. and M.E. degrees from Kyoto University in 1992 and 1994, respectively, and a Dr. Eng. 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 from 2003 to 2006 he was a Lecturer. Since 2006 he has been an Associate Professor in the Department of Media Science, Nagoya University, Nagoya. His research interests include speech processing, speech recognition and spoken dialog. He is a member of the IPSJ, the ASJ and the Japan Society for Artificial Intelligence (JSAI).