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

Distant Speech Recognition Using a Microphone Array Network

Alberto Yoshihiro NAKANO[†a)], ***Nonmember***, **Seiichi NAKAGAWA**[†b)], ***Fellow***, ***and*** **Kazumasa YAMAMOTO**[†c)], ***Member***

SUMMARY In this work, spatial information consisting of the position and orientation angle of an acoustic source is estimated by an artificial neural network (ANN). The estimated position of a speaker in an enclosed space is used to refine the estimated time delays for a delay-and-sum beamformer, thus enhancing the output signal. On the other hand, the orientation angle is used to restrict the lexicon used in the recognition phase, assuming that the speaker faces a particular direction while speaking. To compensate the effect of the transmission channel inside a short frame analysis window, a new cepstral mean normalization (CMN) method based on a Gaussian mixture model (GMM) is investigated and shows better performance than the conventional CMN for short utterances. The performance of the proposed method is evaluated through Japanese digit/command recognition experiments.

key words: *distant speech recognition, microphone array network, GMM-based CMN, speaker's position and orientation estimation*

1. Introduction

Microphone arrays [1] have received increased attention in the past few years, particularly for spatial filtering (beamforming) and sound source localization. Compared with a single microphone, a microphone array has a clear advantage in exploiting the spatial characteristics of the sound field. This is the main reason for using a distributed microphone array network to estimate the position and orientation (the facing angle relative to a coordinate system) of a directional acoustic source in an actual enclosed environment for use in distant speech recognition tasks.

Automatic speech recognition (ASR) systems are known to perform well when the speech signal is recorded by a nearby microphone. However, with the increasing demand for hands-free applications, distant microphones must be considered. For distant microphones, the effect of the environment (reverberation and noise) can markedly degrade the speech recognition performance due to a mismatch between the characteristics of the test environment and those of the training environment for the system [2]. Compensating the features used in the recognition system is a way of reducing the mismatch. Cepstral mean normalization (CMN) [3] is a simple method for normalizing the feature space, thereby reducing the channel distortion. The performance of conventional CMN depends mainly on two assumptions: first, the channel effect is within the short frame analysis window used in ASR, and second, the spoken utterance is sufficiently long to model the channel effect reliably. Although the first assumption is rarely true under real conditions, the CMN method provides some improvement even if the reverberation effect is long. Wang et al. [4] used a long window to process the reverberation effect for the stationary part of speech. They confirmed the effectiveness of this method for utterances sufficiently long to model the channel effect; however, for short utterances no significant improvement was observed. We focus on the second assumption in this work: that the speaker moves at every short utterance. When the spoken utterance is short, the normalization factor provided by CMN is highly dependent on the utterance itself and has a negative effect on the normalization procedure; for instance, the cepstral mean of a short utterance composed of a vowel sound distorts the vowel characteristics [5], [6]. To avoid such an effect, we propose a new Gaussian mixture model (GMM)-based CMN based on codeword-dependent cepstral normalization (CDCN) [7]. In our proposed method, the unnormalized training feature space is clustered by a GMM and used to find the nearest mean feature vector to the input testing feature vector. In the next step, the mean feature vector of a training data set normalized by the mean is found. The difference between the input feature vector and the normalized mean feature vector represents the channel effect. Taking the average over the short utterance, the new mean vector for CMN is found. Experiments on a digit recognition task show the efficiency of this method for short-utterance recognition.

Spatial information is defined here as the position (location) and orientation (facing angle) of an acoustic source in an enclosed space. The method proposed in [8], which relies on an artificial neural network (ANN), is employed. In this method, a set of input features describing the position and orientation of the source is used as a direct mapping to the true source position and true source orientation. Then, the spatial information is used to improve the performance of the ASR system. The aim of this research is not to cover all strategies employing spatial information in speech recognition, but to present some practical ideas in this regard, which involves spatial information estimation using ANNs, speech enhancement using delay-and-sum beamformers, transmission-channel-effect compensa-

Manuscript received December 1, 2009.
Manuscript revised February 17, 2010.

[†]The authors are with the Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan.
a) E-mail: alberto@slp.cs.tut.ac.jp
b) E-mail: nakagawa@slp.cs.tut.ac.jp
c) E-mail: kyama@slp.cs.tut.ac.jp

DOI: 10.1587/transinf.E93.D.2451

tion using GMM-based CMN, and lexicon selection using the estimated orientation angles.

We aim to develop an application similar to those developed in the DICIT (Distant-talking Interfaces for Control of Interactive TV) project[†]; however, in our research the goal is to control different devices spatially distributed in a room using speech commands. We use the spatial information estimated by a distributed microphone array network, that is, the estimated speaker's position and facing angle. Using the estimated position information, we enhance the recorded signals by a simple delay-and-sum beamformer, and using the estimated facing angle information we can restrict the recognized vocabulary, assuming that the speaker faces the desired direction before speaking a command.

The outline of this paper is as follows: In Sect. 2, we give a description of the system. In Sect. 3, the conventional CMN and the new GMM-based CMN are described. In Sect. 4, we present the experimental conditions and results, while the last two sections are devoted to a discussion and conclusions.

2. Background

2.1 System Description

Figure 1 illustrates the acquisition process of $\mathbf{x}_{p_m}(t)$ signals comprising a spectral subtraction stage ("SS") to remove stationary noise, resulting in $\tilde{\mathbf{x}}_{p_m}(t)$ signals, a phase alignment stage ("Delay") using time lags estimated by the robust version of the generalized cross-correlation with phase transform (GCC-PHAT) function, and a combination stage ("$\bigoplus$") to obtain a signal $\tilde{\mathbf{x}}_p^{BF}(t)$ for array p in a distributed microphone array network consisting of P arrays. Finally, $\tilde{\mathbf{x}}_p^{BF}(t)$ is used in the automatic recognition system.

2.2 Signal Model

Consider P identical arrays, each one having Q microphones, where each microphone is denoted as m, for $m = 1, \ldots, Q$. Given a signal source $s(t)$, the signal at each microphone can be represented as

$$\mathbf{x}_{p_m}(t) = h_{p_m}(t) * s(t) + n_{p_m}(t), \tag{1}$$

where $p \in \{1, \ldots, P\}$, $m \in \{1, 2, \ldots, Q\}$, "$*$" denotes convolution, $h_{p_m}(t)$ is the reverberation impulse response that

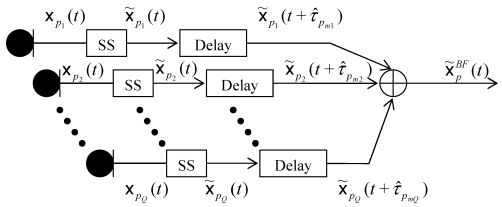

Fig. 1 Illustration of a delay-and-sum beamformer for array p with a prior spectral subtraction (SS) stage applied to each microphone signal $\mathbf{x}_{p_m}(t)$.

describes the propagation path between the source $s(t)$ and the mth microphone of the pth array, and $n_{p_m}(t)$ is the additive background noise. Here, the noise component $n_{p_m}(t)$ is assumed to be different in each microphone.

2.3 Signal Enhancement

2.3.1 Spectral Subtraction

Spectral subtraction (SS) [9] is an efficient method for reducing the spectral effects of acoustically added noise in speech. The method suppresses stationary noise in speech by subtracting the spectral noise bias calculated during nonspeech periods. In practical applications, a voice activity detector (VAD) is necessary to detect speech/nonspeech segments to prevent speech cancellation. In this work, each manually cut test utterance contained a noisy-only segment at the beginning and end of the utterance, which was used to estimate the noise level. We did not use any VAD techniques; this is left as a future problem. As expressed by

$$\mathbf{x}_{p_m}(t) \overset{SS}{\rightarrow} \tilde{\mathbf{x}}_{p_m}(t) \approx h_{p_m}(t) * s(t), \tag{2}$$

the component $n_{p_m}(t)$ in Eq. (1) can be attenuated by spectral subtraction, giving $\tilde{\mathbf{x}}_{p_m}(t)$ as the enhanced version of the signal $\mathbf{x}_{p_m}(t)$.

2.3.2 Delay-and-Sum Beamformer

The purpose of a beamformer [1], [10] is to enhance signals propagated from the desired direction and attenuate interference from other directions. The delay-and-sum beamformer is the simplest example of a beamformer, and is based on the assumption that signals recorded by a microphone array are attenuated and delayed versions of the signal arriving from the desired direction. Using one microphone signal as the reference, appropriate time delays can be estimated between the reference and other signals. Applying these time delays, all signals can be added together in phase and, as a result, phase alignment signals from the desired direction are enhanced whereas signals from other directions are attenuated. That is, for a microphone array p, $\tilde{\mathbf{x}}_p^{BF}(t)$ is the summed signal when appropriate delays are used to phase-align microphone signals, as expressed by

$$\tilde{\mathbf{x}}_p^{BF}(t) = \sum_{n \neq m}^{Q} \tilde{\mathbf{x}}_{p_n}(t + \hat{\tau}_{p_{mn}}), \tag{3}$$

where $\hat{\tau}_{p_{mn}}$ is the time delay between the reference mth microphone and the nth microphone. Although the delay-and-sum method is not a state-of-the-art beamformer technique, it is sufficiently useful to illustrate our proposed distant speech recognition experiments. Here, the delay-and-sum beamformer is applied after a spectral subtraction stage as an additional signal enhancement stage.

[†]http://dicit.fbk.eu/

2.4 Time Delay Estimation

The time delay of arrival (TDOA) is the time lag due to the propagation of a signal to various microphones spatially distributed in a space. In this work, a robust version of the GCC-PHAT function is used for time delay estimation [11], [12]. The proposed method consists of two basic steps; in the first, a binary mask is used to select only high-energy bands in the cross-power spectrum function in the frequency domain (which corresponds to a more prominent peak in the generalized cross-correlation function in the time domain), and in the second step, only frames that yield physically possible delays are combined to generated a robust cross-correlation function. The time delay is estimated by

$$\hat{\tau}_{p_{mn}} = \max_{\tau_{p_{mn}}} \left\{ R(\tau_{p_{mn}}) \right\}, \tag{4}$$

where $\hat{\tau}_{p_{mn}}$ is the time delay estimate (TDE) and $R(\tau_{p_{mn}})$ is the robust GCC-PHAT function. To simplify the notation, the index p in $\hat{\tau}_{p_{mn}}$ will be omitted; thus, the TDE is hereafter expressed by $\hat{\tau}_{mn}$.

2.5 Spatial Information Estimation — Speaker's Position and Orientation Estimation Method

An ANN is used for the position and orientation estimation [8]. ANNs have the ability to learn from and adapt to certain conditions, and can model assumptions about nonlinear/complex physical phenomena, which cannot be solved analytically, responsible for the generation of the given input data of a given process. By employing an ANN, it is expected that combining the parameters estimated by the recorded signals of every array of the network will be sufficient to predict the source orientation and give a more accurate source position than that estimated by a single array.

The estimation of the position and orientation was performed with a different ANN configuration from that used in [8]. The modifications consist of reducing the two-stage ANN in [8] (by first estimating the orientation and then using this information to select a previously trained ANN to estimate the position) to a one-stage ANN (that can estimate the position and orientation simultaneously), and estimating the orientation of the speaker in eight different orientations equally shifted by 45° [north (N), northwest (NW), west (W), southwest (SW), south (S), southeast (SE), east (E), and northeast (NE), relative to a defined coordinate system (see Fig. 5)]. We verify experimentally that the two-stage ANN is slightly better than the one-stage ANN, whereas the latter is easier to implement and is sufficient to demonstrate our method[†]. The one-stage ANN is illustrated in Fig. 2 and consists of a simple fully connected feedforward configuration with one hidden layer. The ANN maps a set of features describing the position of the source to the true source position and the true source orientation. The input layer features are composed of energy-related features consisting of the power values of microphone signals and correlation values between pairs of microphone signals, the TDEs obtained by the robust GCC-PHAT function, and the microphone positions of every array in the array network. The output layer features consist of the true position in three-dimensional (3D) space together with the true speaker orientation, which is one of eight different directions. The power of an array is defined as the highest power value estimated for all microphone signals in the array, while the correlation of an array is defined as the highest correlation value between pairs of microphone signals in the array.

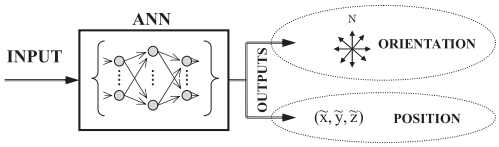

Fig. 2 One-stage ANN. **INPUT**={power + correlation + microphone positions + TDEs}. **OUTPUTS**={one of eight different orientations; position in 3D space $(\tilde{x}, \tilde{y}, \tilde{z})$}.

2.6 Exploring Spatial Information

2.6.1 Reestimation of Time Delay Using the ANN Output (Position in 3D Space)

The TDEs $\hat{\tau}_{mn}$ in Sect. 2.4 can be used in a delay-and-sum beamformer to generate an enhanced speech signal through the process illustrated in Fig. 1. However, the TDE values are degraded by the environment with the result that the signals are not summed correctly in phase. Thus, reliable time delay values, robust to any environmental effect, are necessary.

In this work, the ANN is employed to estimate the position. Once the estimate has been obtained, the time delays are reestimated using the true microphone position. The new values of the time delay $\hat{\tau}'_{mn}$ are then used in the delay-and-sum beamformer.

2.6.2 Resource Management

Processing all the microphone signals of a microphone array network has a high computational cost. Strategies to select a reasonable number of microphone signals are thus necessary. The spatial information of the speaker, in our case the position and orientation, can be used to select the arrays closest to the speaker (assuming that these have a higher signal-to-noise ratio (SNR) than other more distant microphone arrays).

2.6.3 Improving Speech Recognition

Using the estimated orientation, the proposed system explic-

[†]Under the conditions in [12], the two-stage ANN and one-stage ANN respectively yielded 99.5% and 99.4% for the correct orientation ratio, and 23.2 (20.5) and 25.0 (20.5) cm in 3D (2D) space for the average position error.

itly selects a specific lexicon for the application that a person (i.e., a user of the system) is trying to control under the assumption that he or she faces the application when controlling it by speech.

3. Automatic Speech Recognition

Basically, the key to obtaining good recognition performance relies on two factors: (*A*) a large quantity of speech training data to create reliable acoustic and language models, and (*B*) a reduction in the mismatch between training and testing conditions. While (*A*) can be solved with sufficient data, (*B*) caused us the most difficulty. In most cases, training data consist of clean speech data (data recorded with a nearby microphone in a quiet environment), whereas testing data are recorded under adverse conditions (a noisy and reverberant environment with both near and distant microphones). In this section, we present the conventional CMN method and a new CMN method based on GMM to reduce the effects of a reverberant environment.

3.1 Conventional CMN

Conventional CMN is useful in reducing the effect of the early part of reverberation within a frame. For notational simplicity, consider a generic model that illustrates the effect of the environment $h(t)$ on the speech signal $s(t)$ generating the corrupted signal $x(t)$ as

$$x(t) = s(t) * h(t), \tag{5}$$

where "$*$" is the convolution operator. The representation of Eq. (5) in the cepstral domain is obtained by the discrete cosine transform (DCT) of the logarithm of the power spectrum of $x(t)$, under which the convolution operator becomes a simple addition operator,

$$C^x = C^s + C^h, \tag{6}$$

where C^x, C^s, and C^h are the cepstra of $x(t)$, $s(t)$, and $h(t)$, respectively. According to Eq. (6), the effect of $h(t)$ in the cepstral domain is that of an additive bias.

In practice, signals are sampled and acoustic features are estimated within a short-term analysis window assuming quasi-stationarity inside frames. Thus, C^x can be represented by the set of feature vectors $\{C_1, \ldots, C_t, \ldots, C_T\}$, where $t = 1, \ldots, T$ is the time frame index. For a given utterance, the conventional CMN is used to approximately compensate the bias due to the channel distortion as

$$\tilde{C}_t = C_t - \bar{C}, \tag{7}$$

$$\bar{C} = \frac{1}{T}\sum_{t=1}^{T} C_t, \tag{8}$$

where $\tilde{C}_t$, C_t, and $\bar{C}$, are, respectively, the compensated feature vector, the original feature vector, and the average of C_t.

3.2 GMM-Based CMN

3.2.1 Gaussian Mixture Model

A GMM is the weighted sum of M $\mathcal{D}$-dimensional Gaussian densities expressed by

$$P(x;\lambda) = \sum_{i=1}^{M} c_i b_i(x), \tag{9}$$

where x is a $\mathcal{D}$-dimensional random vector; $\lambda = \{c_i, \mu_i, \Sigma_i\}$ denotes the GMM parametric model with mixture weights c_i, mean vector μ_i, and covariance matrix Σ_i; and

$$b_i(x) = \mathcal{N}(x;\mu_i,\Sigma_i) \tag{10}$$

are the densities, for $i = 1, \ldots, M$.

The mixture weight is constrained by

$$\sum_{i=1}^{M} c_i = 1. \tag{11}$$

In our experiment, speaker-independent GMMs were trained by the expectation-maximization (EM) algorithm using the adult male part of the Japanese newspaper article sentence (JNAS) corpus [13] with the HTK toolkit[†].

3.2.2 Compensation Method

An accurate cepstral mean cannot be estimated using Eq. (8), particularly when the utterance is short, because the estimated value is highly dependent on the utterance itself, giving a negative effect on the compensation procedure. For example, the cepstral mean of a short utterance containing a vowel removes the characteristics of the vowel in the normalization procedure. To avoid such a negative effect, the distribution of the non-normalized training data C_{train} is modeled by a GMM of M mixtures, where each mean vector of the GMM is expressed by μ_i, for $i = 1, \ldots, M$. This process divides C_{train} into M different clusters.

The proposed compensation method, illustrated in Fig. 3, consists of finding the nearest mean vector of the GMM to C_t by calculating the likelihood

$$\hat{i}_{(t)} = \arg\max_i \{b_i(C_t)\}, \tag{12}$$

where μ_i is the mean vector of mixture $b_i(\cdot)$ and $\mu_{\hat{i}_{(t)}}$ is the nearest mean vector of the GMM to C_t. On the other hand, the corresponding data set of cluster i is used to determine $\tilde{\mu}_i$, which is the mean vector of the data of cluster i after normalization by the cepstrum mean of cluster i. Using $\tilde{\mu}_i$, the following normalization operation is performed for each frame t:

$$C_t - \tilde{\mu}_{\hat{i}_{(t)}}, \tag{13}$$

$$\text{or} \quad C_t - \mu_{\hat{i}_{(t)}}, \tag{14}$$

[†]http://htk.eng.cam.ac.uk/

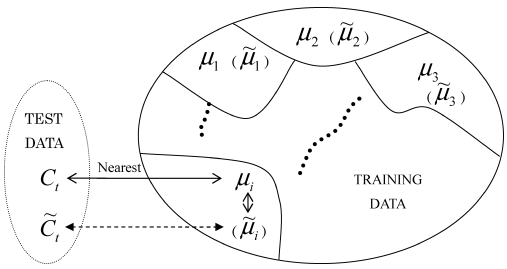

Fig. 3 New GMM-based CMN method.

and taking the expectation of Eqs. (13) and (14) for all input frames T,

$$\Delta C_1 = \frac{1}{T}\sum_{t=1}^{T}\left(C_t - \tilde{\mu}_{\hat{i}_{(t)}}\right), \tag{15}$$

$$\text{or} \quad \Delta C_2 = \frac{1}{T}\sum_{t=1}^{T}\left(C_t - \mu_{\hat{i}_{(t)}}\right), \tag{16}$$

the new channel-effect normalization vector is determined. The normalization is performed as

$$\tilde{C}_t = C_t - \Delta C_1, \tag{17}$$

$$\text{or} \quad \tilde{C}_t = C_t - \Delta C_2 - \bar{C}_{train}, \tag{18}$$

where $\bar{C}_{train}$ is the mean of all training data.

4. Experiments

4.1 Experimental Setup

All experiments were conducted in a 5 m × 6.4 m × 2.65 m room containing eight T-shaped microphone arrays (see Fig. 4), with one array fixed to each wall (arrays A, B, C, and D) and four arrays fixed to the ceiling (arrays E, F, G, and H). To estimate the 3D location it is necessary to use at least four microphones [14], and the T-shaped array is a popular array of four microphones. The computers in the human interaction loop (CHIL)[†] project and [15] also used the T-shaped array. In [12], we compared the position and orientation estimation method based on an ANN with two conventional methods: a TDOA-based position estimation method and the steered response power with phase transform (SRP-PHAT) method, both of which use the T-shaped array. Here we decided to recognize speech using the same microphone array.

Each array was mounted on a structure composed of an acoustic absorber to reduce the effects of reflection near the microphones. The position of the center microphone in each array is fixed and is given in centimeters as follows: A (236.5, 619.0, 206.0), B (497.0, 354.5, 200.0), C (3.0, 354.5, 200.0), D (98.5, 105.0, 200.0), E (130.0, 423.5, 255.0), F (370.0, 423.5, 255.0), G (370.0, 273.5, 255.0), H (130.0, 273.5, 255.0). The distance d between pairs of microphones in each array was set to 20 cm[††]. Five speaking positions (P1 (244, 215.5), P2 (244, 365.5), P3 (144, 415.5), P4 (344, 415.5), P5 (244, 465.5)) were chosen for our experiment.

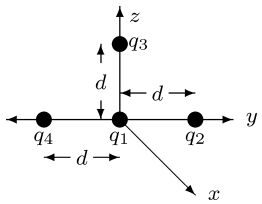

Fig. 4 T-shaped microphone array comprising microphones $\{q_1, q_2, q_3, q_4\}$. d is the distance between adjacent microphones.

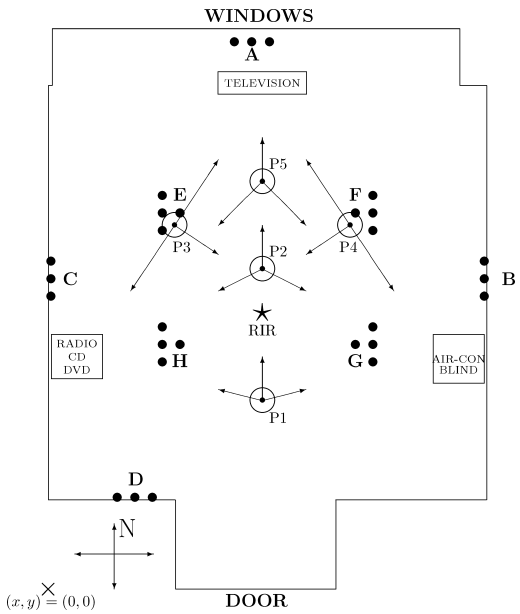

Fig. 5 View of the room from above showing the five speaker positions (P1, P2, P3, P4, P5 – "⊙"), the directions faced (arrows) in the experiment, the microphone array positions (A, B, C, D, E, F, G, H), and the reverberation impulse response measurement position (RIR – "★"). The origin of the coordinate system and the relative orientation are shown at the bottom left.

The array positions and speaker positions are depicted in Fig. 5.

Ten male speakers stood at each position and uttered a list of words in Japanese facing each of the three directions (television, radio/CD/DVD, and air-con/blind). Ta-

[†]http://chil.server.de/servlet/is/101/

[††]Spatial aliasing occurs when the distance between the microphones of an array is greater then half the wavelength of the observed signal in the frequency range of interest (to avoid spatial aliasing at a sampling frequency of 16 kHz considering the entire 8 kHz band, the distance between the microphones must be approximately 2.1 cm; however, the smaller the distance between the microphone pairs, the more difficult it becomes to estimate time delays. The distance of 20 cm was used in [15] and yielded good results in terms of time delay estimation). In practice, spatial aliasing modifies the array beam pattern, such as by reducing the gain in the direction of interest, and creates nulls in the beam pattern where a gain is expected (and vice versa). As a result, the output signal of the beamformer is degraded by the contribution of interference that is not cancelled in the combination process. However, we consider that the main part of the desired signal is still present in the beamformer output, which is used in our experiment.

Table 1 List of words used in our experiment. The left column indicates the type of list (general, television, radio/CD/DVD, air-con/blind), while the right column gives the English translation of the list of words uttered in Japanese.

Purpose	List of words
General (22 words)	one, two, three, four, five, six, seven, eight, nine, zero, yes, no, turn on, turn off, up, down, more/less[a], volume, ok, sleep, open, close
Television (16 words)	next, previous, channel, change, brightness, color, contrast, mute, television, [NHK education, NHK general, Chūkyō TV, Tōkai TV, Nagoya TV, TV Aichi, CBC TV][b]
Radio/CD/DVD (21 words)	station, tuner, music, music name, radio, [NHK 1, NHK 2, Tōkai radio, CBC radio][c], DVD, CD, cancel, start, stop, play, pause, random, track, deck, record music, record movie
Air-con/Blind (8 words)	air conditioner, blind, temperature, heating, cooling, dehumidify, up, down

[a]In Japanese a prefix before a word indicates an increased/decreased effect. But in English a suffix indicates an increased/decreased effect. For example, in English the suffix 'up' in level-up indicates an increased effect.

[b]Listed in brackets are seven Japanese television channels.

[c]Listed in brackets are four Japanese radio stations.

ble 1 gives the list of words translated into English for reference. Each speaker uttered the general list of words plus the specific list related to the direction faced; in other words, if the speaker was facing the television, he would utter words in the general list plus those in the television list; if he was facing the radio/CD/DVD, he would utter words in the general list plus those in the radio/CD/DVD list; and finally, if he was facing the air conditioner/blind, he would utter words in the general list plus those in the air-con/blind list. The procedure was performed twice, resulting in each speaker uttering a total of 222 words per position.

Utterances were recorded at 48 kHz by a 32-channel acquisition system and downsampled to 16 kHz. The acquisition board was manufactured by Tokyo Electron Device Ltd., while the microphones used are the ECM-C10 model produced by Sony Corporation. In the GCC-PHAT analysis, a frame length of 256 samples, a frame shift of 128 samples, and the Hamming window were considered. For each array, a set of three TDOAs $\{\hat{\tau}_{12}, \hat{\tau}_{13}, \hat{\tau}_{14}\}$ was estimated per utterance for pairs $\{1,2\}$, $\{1,3\}$, and $\{1,4\}$, taking microphone 1 as the reference. The measured reverberation time was approximately 330 ms and the background noise level was approximately 35 dBA. The SNR estimated from the recorded signals was approximately 15 dB.

For the speech recognition, a frame size of 25 ms (400 points), a frame shift of 10 ms (160 points), and a Hamming window were used. Then, 27992 utterances read by 175 male speakers (from the JNAS corpus) were used to train 116 Japanese context-independent syllable HMMs including short pauses and silence. Using the context-independent HMMs as base models, 928 context-dependent syllable HMMs with eight left contexts (five vowels, silence, /N/, and the short pause /q/) were obtained [16]. Each continuous-density HMM had five states, with four of them having pdfs of output probability. Each pdf consisted of four Gaussians with full-covariance matrices. The feature space comprised 12 MFCCs, the first and second derivatives of these coefficients plus the first and second derivatives of the power components, generating a total of 38 feature parameters.

For the ANN analysis, a fully connected feedforward ANN was implemented using the Stuttgart Neural Network Simulator (SNNS)[†]. We considered the position estimation in 3D space $(\tilde{x}, \tilde{y}, \tilde{z})$, with eight speaker orientations: N, NW, W, SW, S, SE, E, and NE. The ANN topology used in this study is illustrated in Fig. 2.

For the digit recognition task, the input set comprised the TDE values (three values per array) and microphone positions (three coordinate values per microphone) for all the arrays (a total of 120 input units). Furthermore, in the digit recognition task, the hidden layers had 240 units, and the output layer had 11 units (eight orientations and three values representing the space). In the specific command recognition task, 136 input values (eight power values + eight correlation values + TDEs + microphone position values) and 272 hidden units were used. In the digit recognition task, where the orientation information was not used, the TDEs and microphone positions were sufficient to estimate the position information used to reestimate time delays for the delay-and-sum beamformer. In the command recognition task, the reason for including energy-related features in the estimation method was to obtain better estimates of position and orientation (which were effectively used in the lexicon selection).

The GMMs were trained using male speakers and the JNAS corpus with a sampling frequency of 16 kHz, a frame size of 25 ms, a frame shift of 10 ms, and the Hamming window. The training parameters were 12 MFCCs, with the number of mixtures set to 16, 32, or 64.

4.2 First Task — Digit Recognition

4.2.1 Position and Orientation Estimation by ANN

For the ANN training/testing phase, all recorded data were divided into two sets: nondigits for training (about 73%) and digits for testing (about 27%). As the input for the ANN, the TDE set ($3 \times 8 = 24$ values) and microphone positions ($3 \times 4 \times 8 = 96$ values) were used, while energy-related features were not used. In the training phase, both the correct orientation and the true source position were used as target values. For the true source position, the height of the

[†]http://www.ra.cs.uni-tuebingen.de/SNNS/

mouth of each speaker was used as the z-coordinate. The results in Table 2 are presented in terms of the correct orientation ratio (%) and the average position error (cm), where the correct orientation ratio is the ratio of the total number of correct estimates by the ANN to the total number of input patterns, and the position error is the Euclidean distance between the estimated position and the actual source position. In [8] it was shown that the position estimate obtained by the ANN approach is better than that by a state-of-the-art position localization algorithm such as the SRP-PHAT [1]. The estimated position value in 3D space was used to reestimate $\hat{\tau}_{mn}$, i.e., the TDEs between microphone pairs in each array, and the reestimated values $\hat{\tau}'_{mn}$ were then used in the delay-and-sum beamformer.

4.2.2 Using Conventional CMN

Table 3 gives the digit recognition results for all speakers. In all cases ((a) to (f)), the conventional CMN was used. Experiments without CMN (that is, without cepstral normalization) were not performed because it was shown in [15] that the conventional CMN performed better even for short utterances. In (a), the center microphone in each array was used in the recognition experiment. In (b), the spectral subtraction method was applied to the signal of microphone 1, resulting in a significant improvement in recognition ratio. In (c), the delay-and-sum beamformer was implemented using the time delays estimated by the robust GCC-PHAT function, resulting in a lower recognition ratio than (b). This may be due to the noise level, which degrades the time delay estimation, and thus the signals were not summed in phase. In (d), spectral subtraction was applied to signal (c) resulting in an improvement in the recognition ratio. In (e), spectral subtraction was applied to each signal of the array *a priori*, which were then summed in phase in accordance with the time delay calculated by the robust GCC-PHAT function. The results show that the noise reduction at the beginning of the process has a strong impact on the recognition ratio. In (f), using $\hat{\tau}_{mn}$ in the ANN position and the orientation estimation method, the estimated position was used to reestimate time delays $\hat{\tau}'_{mn}$ between microphone pairs, these delays were used in a delay-and-sum beamformer. Using $\hat{\tau}'_{mn}$, a slight improvement relative to (e) was obtained. The "AVG." column gives the average for all arrays. In "ALL" column, all microphones were used to choose the best result by summing the likelihoods generated in the recognition process of all array signals. The recognized candidate with the highest summed likelihood was chosen as the correct value.

The ceiling arrays (E, F, G, H) tend to yield better results than the wall arrays (A, B, C, D). According to Table 3, array F yields the best results. This is to be expected because the ceiling arrays can perceive signals directly above the speaker and it is more difficult to physically block the ceiling arrays than the wall arrays. Figure 6 depicts the recognition results using only array F for different speaker positions in the six cases, ((a) to (f) in Table 3) and the 32-mixture case (32 Mix. in Table 5). The values in parentheses on the x-axis denote the average distances from the speaker at each position (P1, P2, P3, P4, P5) to array F. It can be seen that position P3 has a lower digit recognition ratio than P1 due to the reflective surfaces (shelves) near P3, whereas there are no obstructions near P1. Signals (a) and (c) have the lowest recognition results for all positions. For P1, signals (e) and (f) yielded recognition ratios greater than 87%. The recognition ratios of signals (b), (e), (f), and (32 Mix.), where spectral subtraction was employed *a priori*, appear to be independent of the distance. Finally, the signal (32 Mix.) yielded the best results for all considered positions.

Table 2 ANN results for the average over all speaker positions and the correct orientation ratio in the case of digits.

Measure	Value
Average position error in 3D space (cm)	34.3
Average position error in 2D space (cm)	29.4
Correct orientation ratio (%)	92.8

Table 3 Digit recognition results by array (A, B, C, D, E, F, G, H) as percentages (%) using the conventional CMN. In "only micro. q_1", only the center microphone of each array was used without any processing; in "SS → only micro. q_1", spectral subtraction is applied to the center microphone of each array; in "beam. $[\hat{\tau}_{mn}]$", beamforming is carried out using time delays calculated by the robust GCC-PHAT function; in "beam. $[\hat{\tau}_{mn}]$ → SS", spectral subtraction is applied to the "beam. $[\hat{\tau}_{mn}]$" signal; in "SS → beam. $[\hat{\tau}_{mn}]$", spectral subtraction is applied to each signal and beamforming is carried out using time delays calculated by the robust GCC-PHAT function; in "SS → beam. $[\hat{\tau}'_{mn}]$", beamforming is carried out using reestimated time delays $\hat{\tau}'_{mn}$. In the "AVG." column, the average over all arrays was calculated; in the "ALL" column, the likelihoods of all arrays were combined.

SIGNAL	ARRAYS									
	A	B	C	D	E	F	G	H	AVG.	ALL
(a) only micro. q_1	66.10	62.53	62.70	64.37	68.43	**71.70**	67.50	66.97	66.29	69.00
(b) SS → only micro. q_1	79.40	77.77	78.40	77.93	81.47	**83.43**	80.20	79.93	79.82	84.96
(c) beam. $[\hat{\tau}_{mn}]$	60.77	61.73	62.20	62.93	66.37	**68.37**	65.80	64.87	64.13	67.06
(d) beam. $[\hat{\tau}_{mn}]$ → SS	76.63	76.07	78.30	77.93	79.00	**80.57**	77.40	76.87	77.85	84.16
(e) SS → beam. $[\hat{\tau}_{mn}]$	79.37	80.07	80.57	79.60	82.90	**84.13**	81.23	80.73	81.08	85.67
(f) SS → beam. $[\hat{\tau}'_{mn}]$	81.40	81.60	82.10	81.33	84.27	**84.63**	82.97	84.20	82.81	86.13

Table 4 Digit recognition results by array (A, B, C, D, E, F, G, H) as percentages (%) using the cepstral mean calculated over all digits in a given position and orientation. "SS → beam. [$\hat{\tau}'_{mn}$]" means beamforming using reestimated delays.

SIGNAL	ARRAYS									
	A	B	C	D	E	F	G	H	AVG.	ALL
SS → beam. [$\hat{\tau}'_{mn}$]	85.53	85.73	86.60	86.13	86.70	**88.17**	86.37	86.77	86.50	89.70

Table 5 Digit recognition results by array (A, B, C, D, E, F, G, H) as percentages (%) using the new GMM-based CMN for the digit recognition task. "SS → beam. [$\hat{\tau}'_{mn}$]" means beamforming using reestimated delays.

SIGNAL	ARRAYS									
SS → beam. [$\hat{\tau}'_{mn}$]	A	B	C	D	E	F	G	H	AVG.	ALL
16 Mix.	82.03	82.10	83.13	81.83	84.83	**86.33**	84.30	84.53	83.64	87.06
32 Mix.	83.30	83.30	84.03	82.80	85.60	**86.93**	85.53	85.43	84.62	87.90
64 Mix.	83.13	83.20	84.33	82.97	86.17	**86.83**	85.27	85.63	84.69	87.86

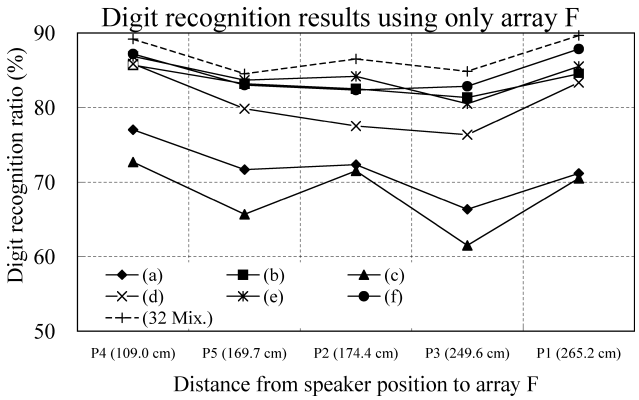

Fig. 6 Digit recognition results considering the distance between the speakers and array F.

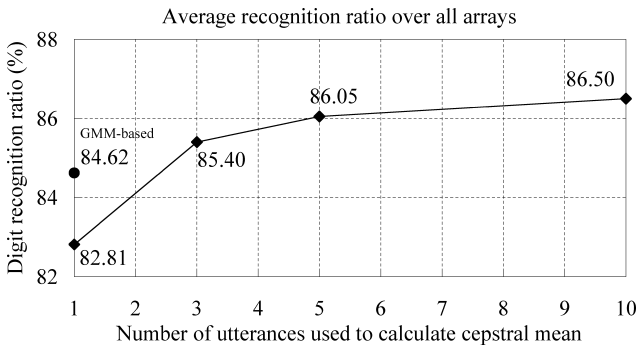

Fig. 7 Number of digit utterances used to calculate the cepstral mean versus the average recognition results over all arrays.

4.2.3 Using the Cepstral Mean Calculated over Many Utterances in CMN

Table 4 gives the digit recognition results for signal (f) when the cepstral mean is calculated over all digits (ten utterances, about four seconds) for each position and orientation obtained using Eq. (8). It can be seen that the longer the signal is, the better the calculated cepstral mean will be, and therefore, the better the modeling of the channel effect will be.

4.2.4 Using the GMM-Based CMN

The experiment in this section utilizes signal (f), the signal giving the best results according to Table 3. Table 5 gives the results using the new GMM-based CMN method for 16, 32, and 64 mixtures using Eq. (17). Initially, improvements were observed compared with the baselines in Table 3. The GMM with 32 mixtures is sufficient to model voice clusters and yields better results than that with 16 or 64 mixtures, which can be explained by the fact that 16 mixtures are too few whereas 64 mixtures are too many to model voice clusters such as phonemes in mismatched environments. Using Eq. (18), the averages "AVG." were 83.32%, 84.37%, and 84.32%, for the GMM with 16, 32, and 64 mixtures, respectively, which are almost the same performances as those obtained by Eq. (17). Although the results are inferior to those in Table 4, where the cepstral mean was calculated using all digits, in Table 5 only the short utterance itself is used in the recognition. In Fig. 7, we present the relationship between the number of digit utterances used to calculated the cepstral mean and the average recognition result over all arrays. We find that three short utterances are sufficient to yield a better performance than the GMM-based CMN.

4.3 Second Task — Specific Command Recognition

The specific commands comprise longer utterances than the digits. For the ANN training/testing phase five different sets were created. All the data from eight speakers were used for training (80%), while all the data from two speakers (20%) were used in the testing phase. Each testing set included utterances from different speakers. There was no overlap between the training and testing sets. As the input for the ANN, the set of TDOAs (24 values), the microphone positions (96 values), the powers (eight values), and the correlation (eight values) were used. In the training phase, both the correct orientation and the true source position were used as target values. For the true source position, the height of the mouth of each speaker was used as the z-coordinate. The

ANN results are presented in Table 6 and only specific commands were used in this experiment.

The estimated position and orientation information was used to improve recognition results. Assuming that the speaker faced the desired direction before speaking a command, a different dictionary was created for each direction. The process was as follows:

1. Estimate the position and orientation.
2. Using the information from Step 1, use a specific dictionary in the recognition task; for instance, if the speaker turns towards the television, use the dictionary that contains only television commands.
3. If the direction faced by the speaker is not directly towards one of the specified directions, use the ANN output to choose the nearest adjacent direction, which has the highest score, to be the estimated orientation.
4. If the strategy in Step 3 does not work, use a dictionary that contains all the specific words.

Table 6 ANN results for the average over all speaker positions and the correct orientation ratio in the case of specific commands.

Measure	Value
Average position error in 3D space (cm)	30.7
Average position error in 2D space (cm)	25.9
Correct orientation ratio (%)	96.0

The strategy in Step 3 does not work when an error occurs in the orientation estimation at positions far from the center of the room. For instance, the speaker at P1 utters a command to the television (N) but the estimated orientation is south (S). The adjacent directions to S are SE or SW. However, SE or SW are not directed to the television (N), Radio/CD/DVD (NW), or air conditioner/blind (NE). Thus, we cannot state the correct orientation with confidence, therefore, we opted to use the entire dictionary in this case.

To evaluate this method, three cases were analyzed:

I. Using a dictionary containing a list of all specific words with the estimated position information.
II. Selecting the list of words depending on the speech direction, assuming a 100% correct orientation estimation, with the estimated position information.
III. Selecting the list of words depending on the speech direction, using the spatial information (the position and orientation) estimated by the ANN.

Tables 7, 8, and 9 give the results for cases I, II, and III, respectively. Table 7 shows the results by array when a unified dictionary was used in the recognition process, resulting in an average recognition ratio of approximately 85%. Table 8 shows the results by array when a specific dictionary was used for each specific direction. Here, it is assumed

Table 7 Command recognition results using a dictionary with all words. In the "AVG." column, the average over all arrays was calculated, in the "CL" column, the results obtained by the closest array to the speaker are given, in the "ALL" column, the likelihoods of all arrays were combined.

	ARRAYS										
SIGNAL	A	B	C	D	E	F	G	H	AVG.	CL	ALL
(a) SS → only micro. q_1	82.18	79.40	82.67	82.09	**84.98**	8378	80.40	83.73	82.40	83.97	91.02
(b) SS → beam. $[\hat{\tau}_{mn}]$	83.62	83.13	85.86	85.53	**86.75**	85.11	82.40	84.77	84.65	85.75	92.37
(c) SS → beam. $[\hat{\tau}'_{mn}]$	84.31	84.08	86.57	85.88	**86.80**	85.02	83.46	85.44	85.20	85.89	92.35
(d) SS → beam. $[\hat{\tau}'_{mn}]_{32Mix.}$	85.33	85.11	87.57	86.40	**88.37**	86.02	85.17	86.80	**86.35**	**87.15**	**92.62**

Table 8 Command recognition results using dictionaries depending on the orientation (assuming 100% correct orientation estimation). In the "AVG." column, the average over all arrays was calculated, in the "CL" column, the results obtained by the closest array to the speaker are given, in the "ALL" column, the likelihoods of all arrays were combined.

	ARRAYS										
SIGNAL	A	B	C	D	E	F	G	H	AVG.	CL	ALL
(a) SS → only micro. q_1	89.06	88.08	89.95	90.11	**91.06**	89.91	88.44	90.93	89.69	90.33	92.15
(b) SS → beam. $[\hat{\tau}_{mn}]$	90.15	90.26	**92.15**	91.91	91.75	91.37	89.53	91.44	91.07	91.22	93.13
(c) SS → beam. $[\hat{\tau}'_{mn}]$	90.66	90.86	92.00	**92.08**	92.06	91.31	90.06	91.88	91.36	91.62	93.06
(d) SS → beam. $[\hat{\tau}'_{mn}]_{32Mix.}$	90.84	91.33	92.15	91.95	**92.51**	91.57	90.84	92.35	**91.69**	**91.86**	**93.22**

Table 9 Command recognition results using dictionaries depending on the orientation (using the orientation estimated by the ANN). In the "AVG." column, the average over all arrays was calculated, in the "CL" column, the results obtained by the closest array to the speaker are given; In the "ALL" column, the likelihoods of all arrays were combined.

	ARRAYS										
SIGNAL	A	B	C	D	E	F	G	H	AVG.	CL	ALL
(a) SS → only micro. q_1	88.95	88.04	89.84	90.04	**91.06**	89.82	88.33	90.77	89.61	90.20	92.13
(b) SS → beam. $[\hat{\tau}_{mn}]$	90.08	90.17	**92.11**	91.82	91.71	91.28	89.37	91.31	90.98	91.15	93.11
(c) SS → beam. $[\hat{\tau}'_{mn}]$	90.55	90.80	**91.97**	**91.97**	**91.97**	91.26	89.95	91.77	91.28	91.55	93.06
(d) SS → beam. $[\hat{\tau}'_{mn}]_{32Mix.}$	90.68	91.24	92.15	91.88	**92.44**	91.48	90.71	92.24	**91.60**	**91.80**	**93.22**

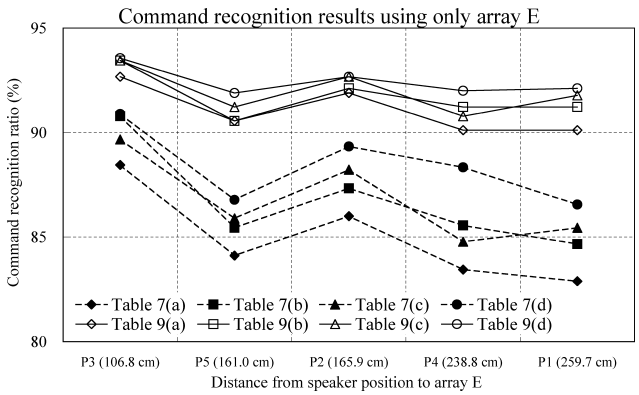

Fig. 8 Command recognition results considering the distance between the speakers and array E.

that the orientation estimation is 100% correct. An average recognition ratio of approximately 91% was obtained. Finally, Table 9 shows the results using the orientation information automatically estimated by the ANN. The average recognition ratio was almost the same as that in Table 8.

Figure 8 illustrates the results of the command recognition task from Tables 7 and 9 using only array E for different speaker positions. Two sets are clearly visible: the upper (Table 9 (a–d)) and lower (Table 7 (a–d)) sets. The upper set, formed by the recognition results using separate dictionaries, is independent of the distance and has a smaller recognition ratio range. The lower set, formed by the recognition results using the dictionary containing all the words, is more dependent on distance and has a larger recognition ratio range. The difference between the upper and lower sets indicates the improvement as a result of restricting the dictionary used on the basis of the orientation information in the recognition process.

5. Discussion

5.1 Comparing the Results of the Digit Recognition Task

Statistical test

To compare the performance of the system for the different approaches adopted in this work, the sign test for matched pairs is used to compare four compensated signals based on the same test data: (b), (e), and (f) in Table 3; and the signal "32 Mix." in Table 5. To compare two methods M_1 and M_2 using the sign test, let n_1 be the number of times that M_1 is true and M_2 is false, and n_2 be the number of times that M_2 is true and M_1 is false. Modeling this process by a binomial distribution with averages $p_1 = p_2 = \frac{1}{2}$, we test the null hypothesis (i.e., no difference between methods M_1 and M_2). Writing $N = n_1 + n_2$, we have

$$Z = \frac{n_2 - N/2}{\sqrt{N/4}}, \tag{19}$$

upon approximating a binomial distribution by the normal distribution $\mathcal{N}(0, 1)$. Comparing (b) $\leftrightarrow$ (e), (e) $\leftrightarrow$ (f), and (f) $\leftrightarrow$ $(f)_{32Mix.}$, we have $|Z|_{e,b} = 6.10$, $|Z|_{f,e} = 10.7$, and $|Z|_{f_{32Mix.},f} = 11.4$, respectively. For a significance level of 1% we have

$$p(\{|Z| \geq 2.65\}) = 0.01, \tag{20}$$

which means that the initial hypothesis can be rejected and the improvements are statistically significant. Thus, $|Z|_{e,b}$ and $|Z|_{f,e}$ validate the improvements shown in Table 3. We can thus conclude that the reestimated $\hat{\tau}'_{mn}$-based delay-and-sum beamformer is superior to the original $\hat{\tau}_{mn}$-based delay-and-sum beamformer.

5.2 GMM-Based CMN

Statistical test

Comparing Tables 3 and 5, the proposed GMM-based CMN shows an improvement over the conventional CMN for short utterances, validated by the sign test ($|Z|_{f_{32Mix.},f} = 11.4$). We can therefore conclude that the GMM-based CMN is superior to the conventional CMN for short utterances (Both methods compensate the distortion due to noise reduction in the same way.). The recognition ratio is, however, inferior to that obtained in Table 4. This may be due to the effect of reverberation in the mapping process ($C_t \leftrightarrow \mu_i$) to the unnormalized GMM mean vector and the influence of the too short utterances. This suggests that a dereverberation method must be applied at the front end before the feature extraction stage.

5.3 Spatial Information

The estimated position and orientation information of the speaker is shown to be effective in improving the performance of the ASR system.

5.3.1 Reestimation of Time Delay

According to Tables 3 and 7, using the reestimated time delays $\hat{\tau}'_{mn}$ derived from the position estimate obtained by the ANN method appropriate for use in a delay-and-sum beamformer to improve the recognition performance.

5.3.2 Decision Methods and Resource Management

In our experiments, we obtained recognition results for each of the eight T-shaped microphone arrays (with a total of 32 microphones). Using all the available results, the best results were obtained by an *integration method* which combined the results of all arrays. It was verified that the "ALL" column in tables of digit and command recognition results has larger recognition ratios than those of individual arrays and those in the "AVG." column.

However, sometimes not all resources are available at the same time owing to the high computational cost of the microphone array network; thus, a *selection method* to select only the necessary number of arrays is suggested. An

example of a *selection method* is given in Tables 7, 8, and 9 for the command recognition task, where the "CL" column represents the results of the closest array to the estimated speaker's position. The values in the "CL" column are still smaller than the "ALL" column results and those of the best array, array "E", showing that different criteria for the *selection method* should be suggested and tested. For instance, an approach selecting/integrating the results of two or three of the closest arrays to the estimated speaker's position.

5.3.3 Orientation Information

Restricting the list of commands depending on the speech direction was shown to be efficient in improving the recognition system performance. Instead of using a dictionary of 45 commands, three different dictionaries with 16 (television), 21 (radio/CD/DVD), and eight (air-con/blind) commands were used. Comparing the results in Table 7 with those in Tables 8 and 9, where the orientation information is used, an improvement of approximately 6% is obtained for the "AVG." values. Comparing Tables 8 and 9, no degradation due to orientation errors is observed. Note that, when the recognized lexicon is restricted by the orientation, the difference between the performances of each array and "ALL" (combining the results of all arrays) is smaller in Tables 8 and 9 than in Table 7. For example, in Table 7 the results of array "E" and "ALL" are 88.37% and 92.62% and in Table 9, those of array "E" and "ALL" are 92.44% and 93.22%, respectively.

6. Conclusions and Future Works

Spatial information estimated by a microphone array network using an ANN was used to improve recognition results in digit and command recognition tasks. In this work, spatial information was used in a simple delay-and-sum beamformer for signal enhancement, in the reestimation of time delays for the delay-and-sum beamformer, and to restrict the recognized vocabulary depending on the estimated source orientation. The results illustrate the potential of using a distributed microphone network in recognition tasks in a real environment, for instance, hands-free and voice command applications. A study on short utterances (digits) was presented, in which a new GMM-based CMN method showed an improvement over the conventional CMN method.

This paper reports an initial study in which a user desires to control a given device using voice commands. To improve the recognition results we combined different methods to deal with the effect of the environment (background noise and the reverberation). In this work, we presented experiments using TV and audio devices, which are typical devices in a living room. However, these devices generate nonstationary disturbances; thus, this task may lead to a misunderstanding but simpler applications such as dimmers, fan heaters, ceiling fans, blinds, and air conditioners in quiet rooms should also be explored. To cope with nonstationary disturbances such as TV and audio devices, more research must be performed such as on an echo cancellation stage that reduces the effect of the nonstationary disturbance.

In our future work, we aim to improve the performance of our system by implementing a dereverberation method and more efficient beamforming techniques [2], [10] such as the minimum variance distortionless response (MVDR) in its adaptive version. We also plan to study echo cancellation methods to cancel nonstationary disturbances at the microphones and to explore the spatial information required for array selection. Resource management will be explored so that only the necessary number of arrays is used aiming at a reduction in power consumption and the amount of data processing.

Acknowledgments

We would like to thank the Global COE program "Frontiers of Intelligent Sensing" and Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT) for supporting our research.

References

[1] M. Brandstein and D. Ward, Microphone Arrays: Signal Processing Techniques and Applications, Springer, New York, 2001.
[2] M. Wolfel and J. Mcdonough, Distant Speech Recognition, Chapter 13, Wiley, Chichester, UK, 2009.
[3] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-29, no.2, pp.254–272, April 1981.
[4] L. Wang, S. Nakagawa, and N. Kitaoka, "Robust speech recognition by combining short-term and long-term spectrum based position-dependent CMN with conventional CMN," IEICE Trans. Inf. & Syst., vol.E91-D, no.3, pp.457–466, March 2008.
[5] A. Vikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," Speech Commun., vol.25, pp.133–147, Aug. 1998.
[6] P. Pujol, D. Macho, and C. Nadeu, "On real-time mean-and-variance normalization of speech recognition features," Proc. ICASSP, pp.773–776, May 2006.
[7] A. Acero and R.M. Stern, "Environmental robustness in automatic speech recogntion," Proc. ICASSP, pp.849–852, April 1990.
[8] A. Nakano, S. Nakagawa, and K. Yamamoto, "Estimating the position and orientation of an acoustic source with a microphone array," Proc. Interspeech, pp.1127–1130, Sept. 2009.
[9] S. Boll, "Supression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process., vol.27, no.2, pp.113–120, April 1979.
[10] E. Hansler and G. Schmidt, Speech and Audio Processing in Adverse Environments, Springer, Berlin, 2008.
[11] A. Nakano, K. Yamamoto, and S. Nakagawa, "Directional acoustic source's position and orientation estimation approach by a microphone array network," IEEE Proc. 13th DSP Workshop & 5th SPE Workshop, pp.606–611, Jan. 2009.
[12] A. Nakano, S. Nakagawa, and K. Yamamoto, "Automatic estimation of position and orientation of an acoustic source by a microphone array network," J. Acoust. Soc. Am., vol.126, no.6, pp.3084–3094, Dec. 2009.
[13] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "Japanese speech corpus for large vocabulary continuous speech recognition research," J. Acoust. Soc. Jpn. (E), vol.20, no.3, pp.199–206, May 1999.
[14] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distance speaker

recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM," Speech Commun., vol.49, pp.501–513, June 2007.
[15] L. Wang, S. Nakagawa, and N. Kitaoka, "Robust distant speech recognition by combining multiple microphone-array processing with position-dependent CMN," EURASIP J. Appl. Signal Process., vol.2006-95491, pp.1–11, Jan. 2006.
[16] J. Zhang, L. Wang, and S. Nakagawa, "LVCSR based on context-dependent syllable acoustic models," Proc. AWSST, pp.81–86, March 2008.

Alberto Yoshihiro Nakano received his B.E. and M.E. degrees from University of São Paulo, Brazil, in 2000 and 2005, respectively. In 2006, he became a research student at Toyohashi University of Technology. Since 2007, he has been enrolled in the doctoral program at Toyohashi University of Technology. His current research interests include signal processing and speech recognition/speech processing. He is a member of ASJ and IEEE.

Seiichi Nakagawa received his Dr. Eng. degree from Kyoto University in 1977. He joined Kyoto University in 1976 as a Research Associate in the Department of Information Sciences. From 1980 to 1983, he was an Assistant Professor, from 1983 to 1990, he was an Associate Professor, and 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 J.C. Bose Memorial Award from the Institution of Electronics and Telecommunication Engineers. His major research interests include automatic speech recognition/speech processing, natural language processing, human interfaces, and artificial intelligence. He is a Fellow of IPSJ.

Kazumasa Yamamoto received his B.E., M.E., and Dr. Eng. degrees in information and computer sciences from Toyohashi University of Technology, Toyohashi, Japan, in 1995, 1997, and 2000, respectively. From 2000 to 2007, he was a Research Associate in the Department of Electrical and Electronic Engineering, Faculty of Engineering, Shinshu University, Nagano, Japan. Since 2007, he has been an Assistant Professor in the Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi, Japan. His current research interests include speech recognition and privacy protection for speech signals. He is a member of ASJ and IPSJ.