Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2009, Article ID 512314, 9 pages
doi:10.1155/2009/512314

Research Article

Noise Robust Speech Recognition Applied to Voice-Driven Wheelchair

Akira Sasou and Hiroaki Kojima

Intelligent Media Research Group, Information Technology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Central2,1-1-1, Umezono, Tsukuba, Ibaraki 305-8568, Japan

Correspondence should be addressed to Akira Sasou, a-sasou@aist.go.jp

Received 26 January 2009; Revised 7 May 2009; Accepted 31 July 2009

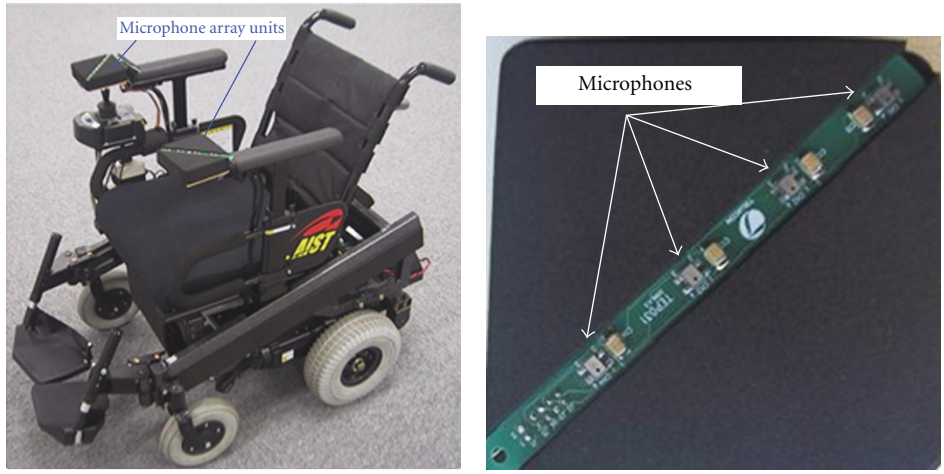
Recommended by Mark Kahrs

Conventional voice-driven wheelchairs usually employ headset microphones that are capable of achieving sufficient recognition accuracy, even in the presence of surrounding noise. However, such interfaces require users to wear sensors such as a headset microphone, which can be an impediment, especially for the hand disabled. Conversely, it is also well known that the speech recognition accuracy drastically degrades when the microphone is placed far from the user. In this paper, we develop a noise robust speech recognition system for a voice-driven wheelchair. This system can achieve almost the same recognition accuracy as the headset microphone without wearing sensors. We verified the effectiveness of our system in experiments in different environments, and confirmed that our system can achieve almost the same recognition accuracy as the headset microphone without wearing sensors.

1. Introduction

Although various voice-driven wheelchairs have already been developed to enable disabled people to move independently, conventional voice-driven wheelchairs still have some associated problems [1–4]. Conventional voice-driven wheelchairs employ a headset microphone that can record the user's voice command in a higher Signal-to-Noise Ratio (SNR), even in the presence of surrounding noise, and can achieve sufficient speech recognition accuracy. However, users need to put on this headset microphone each time they use the wheelchair. In addition, when the headset microphone moves away from the position of the mouth, users need to be able to adjust the position of the headset microphone by themselves. These actions are not always easy, especially for the hand disabled, who are one of the major users of this wheelchair. Since such users need noncontact and nonconstraining interfaces for controlling the wheelchair, we appraised headset microphones as impractical. Conversely, it is also well known that the speech recognition accuracy drastically degrades when the microphone is placed far from the user because surrounding noises can easily interfere with the user's voice.

In this paper, we develop a noise robust speech recognition system for a voice-driven wheelchair [5]. This system can achieve almost the same recognition accuracy as the headset microphone without wearing sensors. To eliminate the need for the user to wear a microphone, we developed a microphone array system that is mounted on a wheelchair. Our proposed microphone array system can easily distinguish the user's utterances from other voices without using a speaker identification technique, and it can achieve precise Voice Activity Detection (VAD). We also adopt a feature compensation technique following to the microphone array system. As a result of combining these two methods, the feature compensation method can utilize the reliable VAD information from the microphone array, which is necessary for correctly compensating the noise-corrupted speech features. And the weak point of the microphone array, which is processing omnidirectional noises, can be compensated for by the feature compensation method. Consequently, our system can be applied to a variety of noise environments.



(a) (b)

Figure 1: A wheelchair with the developed microphone array system. (a) The developed wheelchair. (b)The microphone array unit.

2. Microphone Array System

In a voice-driven wheelchair, headset microphones should be placed as close to the wheelchair user's mouth as possible to overcome the background noise. However, such microphones can be both dangerous and inconvenient for some users, such as those with cerebral palsy, who have involuntary movements. Therefore, the microphone should be positioned sufficiently far enough from the user's mouth so that it does not touch the user's head. However, using this configuration often results in decreased accuracy or functioning by speech recognition systems which are typically sensitive to interference from background noise and other people's voices. To overcome these problems, we employed a microphone array system instead of a headset microphone in the wheelchair we developed (Figure 1(a)). Figure 1(b) shows one of the microphone array units, which consists of two circuit boards. Each circuit board is W130 × D10 × H5 mm in size and has four omnidirectional silicon microphones (Knowles Acoustics, SPM0103ND3-C) soldered in a line at intervals of 3 cm in order to avoid spatial aliasing at frequencies up to 4 kHz. The circuit boards are placed in a diagonal direction on black square sponges on the armrests as shown in Figure 1. Because these black sponges are placed on the edges of the arm rests, the user's head never touches the microphone array system, even during involuntary movements.

2.1. Detection of User Utterances and Noises. A speech recognition system should accept only the user's voice and reject voices coming from other sources. If we adopt a speaker identification technique for this purpose, we need to train the system every time a new user uses the wheelchair. This is not always practical. Instead, with our microphone array system, we can estimate the position or the direction of arrival of the user's voice. That is, the mouth position of the seated user is always in a certain area, which is near the center of the seat at a sitting height. We call this the *user utterance area*. When the position of the voice is estimated to be in the user utterance area, the speech recognition system accepts the voice. However, when a voice is judged to come from outside the user utterance area, the speech recognition system rejects the command. By adopting the microphone array system, we can easily distinguish the user's voice from other voices without any training procedures.

We adopted the MUSIC [6] method for estimating the position and direction of arrival of noises under the assumption that the microphone array system receives a sound source occurring in the user utterance area as a spherical wave. The steering vectors in the user utterance area are defined as follows:

$$\begin{aligned}
\mathbf{P}_q &= [Px_q, Py_q, Pz_q]^T, \quad q = 1, \ldots, 8 \\
R_q &= \left|\mathbf{p}_q - \mathbf{p}_0\right| = \sqrt{\left(Px_q - Px_0\right)^2 + \left(Py_q - Py_0\right)^2 + \left(Pz_q - Pz_0\right)^2} \\
\tau_q &= \frac{R_q}{v}, \qquad g_q = g(\omega, Rq) \\
\mathbf{a}(\omega, \mathbf{P}_0) &= \left[g_1 e^{-j\omega\tau_1}, \ldots, g_8 e^{-j\omega\tau_8}\right]^T,
\end{aligned} \tag{1}$$

where $\mathbf{P}_0$ is the position of the sound source in the user utterance area, $\mathbf{P}_1 \cdots \mathbf{P}_8$ are the positions of the microphones, τ_q is the propagation time, R_q is the distance between the qth microphone and the sound source, v is the sound velocity, $g(\omega, R_q)$ is a distance-gain function, $\mathbf{a}(\omega, \mathbf{P}_0)$ is the steering vector of a user utterance, e is the base of natural logarithms, j is an imaginary number, and T represents the transposition of a vector or matrix. We measured the distance-gain function at several distances and fitted a model function to the measured values. We also assumed that noise sources outside the wheelchair are received as plane waves by the microphone array system. The steering vectors are thus

defined as

$$\mathbf{c}_k = [\cos\phi_k \cos\theta_k, \cos\phi_k \sin\theta_k, \sin\phi_k]$$
$$r_{q,k} = \mathbf{P}_q \cdot \mathbf{c}_k, \qquad T_{q,k} = \frac{r_{q,k}}{v} \tag{2}$$
$$\mathbf{b}(\omega, \theta_k, \phi_k) = [e^{j\omega T_{1,k}}, \ldots, e^{j\omega T_{8,k}}]^T,$$

where $\mathbf{c}_k$ is the normal line vector of the plane wave emitted by the kth outside sound source, θ_k and ϕ_k represent the azimuthal and elevation angles of the kth plane wave, respectively, $r_{q,k}$ and $T_{q,k}$ are the propagation distance and time of the kth plane wave between the qth microphone position and the origin of the coordinate, and $\mathbf{b}(\omega, \theta_k, \phi_k)$ represents the steering vector of the kth plane wave.

The spatial correlation matrix is defined as

$$\mathbf{R}(\omega) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{y}_F(\omega, n)\mathbf{y}_F^H(\omega, n),$$
$$\mathbf{y}_F(\omega, n) = [Y_{F,1}(\omega, n), \ldots, Y_{F,8}(\omega, n)], \tag{3}$$

where $Y_{F,q}(\omega, n)$ represents the FFT of the nth frame received by the qth microphone. The eigenvalue decomposition of $R(\omega)$is given by

$$\mathbf{R}(\omega) = \mathbf{E}(\omega)\mathbf{L}(\omega)\mathbf{E}^{-1}(\omega), \tag{4}$$

where $\mathbf{E}(\omega)$ denotes the eigenvector matrix that consists of the eigenvectors of $\mathbf{R}(\omega)$ as $\mathbf{E}(\omega) = [\mathbf{e}_1(\omega), \ldots, \mathbf{e}_8(\omega)]$, and $\mathbf{L}(\omega)$ is a diagonal matrix whose diagonal elements consist of the eigenvalues $\lambda_1(\omega) \geq \cdots \geq \lambda_8(\omega)$,

$$\mathbf{L}(\omega) = \mathrm{diag}(\lambda_1(\omega), \ldots, \lambda_8(\omega)). \tag{5}$$

The number of sound sources is estimated from the eigenvalues, as follows. First, we evaluate the threshold value, defined as

$$T_{\mathrm{egn}}(\omega) = \lambda_1^{C_{\mathrm{egn}}}(\omega) \times \lambda_8^{(1-C_{\mathrm{egn}})}(\omega), \quad 0 < C_{\mathrm{egn}} < 1, \tag{6}$$

where C_{egn} is a constant that is adjusted experimentally. The number of sound sources $N_{\mathrm{snd}}(\omega)$ is then estimated as the number of eigenvalues larger than the threshold value:

$$\lambda_1(\omega), \ldots, \lambda_{N_{\mathrm{snd}}}(\omega) \geq T_{\mathrm{egn}}(\omega). \tag{7}$$

The eigenvectors corresponding to these eigenvalues form the signal subspace $\mathbf{E}_s(\omega) = [\mathbf{e}_1(\omega), \ldots, \mathbf{e}_{N_{\mathrm{snd}}}(\omega)]$. The remaining eigenvectors $\mathbf{E}_n(\omega) = [\mathbf{e}_{N_{\mathrm{snd}}+1}(\omega), \ldots, \mathbf{e}_8(\omega)]$ are the noise subspace. User utterances are detected according to the following method. First, we search for the position P_0 that absolutely maximizes the following value in the user utterance area

$$\mathcal{Q}(\mathbf{P}) = \frac{1}{\sum_\omega |\mathbf{a}^H(\omega, \mathbf{P})\mathbf{E}_n(\omega)|^2}, \quad \mathbf{P}_0 = \arg\max_{P \in UUA} \mathcal{Q}(\mathbf{P}). \tag{8}$$

If the absolute maximum value $\mathcal{Q}(\mathbf{P}_0)$ exceeds the threshold value T_{usr}, we judge that the user made a sound. The arrival directions of outside sound sources are evaluated as directions that locally maximize the following value:

$$U(\theta, \phi) = \frac{1}{\sum_\omega |\mathbf{b}^H(\omega, \theta, \phi)\mathbf{E}_n(\omega)|^2}. \tag{9}$$

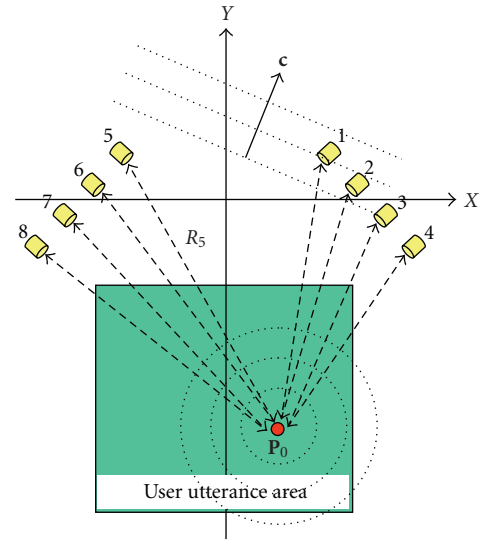


Figure 2: Schematic diagram of wave propagation.

2.2. Enhancement of the User Utterance. When a user utterance and noise occur simultaneously, we need to suppress the noise to recognize the user utterance correctly. For this purpose, we adopted the modified minimum variance beamforming (MVBF) technique [7]. The modified MVBF can generate a spatial inverse filter of high performance with a relatively small amount of data. This capability is suitable for our wheelchair application, because the sound source localization and the spatial inverse filter need to be updated frequently.

In the following, we assume that the estimated position of the user utterance is $\mathbf{P}_0$, the estimated number of noises is K, and the estimated arrival directions of noises are given by $(\theta_k, \phi_k), k = 1, \ldots, K$. Instead of using the estimate of spatial correlation matrix $\mathbf{R}(\omega)$, the modified MVBF uses the following virtual correlation matrix:

$$\mathbf{V}(\omega) = \mathbf{a}(\omega, \mathbf{P}_0) \cdot \mathbf{a}^H(\omega, \mathbf{P}_0) + \sum_{k=1}^{K} \mathbf{b}(\omega, \theta_k, \phi_k) \cdot \mathbf{b}^H(\omega, \theta_k, \phi_k) + \sigma\mathbf{I}. \tag{10}$$

The last term $\sigma\mathbf{I}$ is the correlation matrix of the virtual background noise, and the power of the virtual noise σ can be arbitrarily chosen. By using this virtual correlation matrix, the coefficient vector of the spatial inverse filter becomes

$$\mathbf{w}(\omega) = \frac{\mathbf{V}^{-1}(\omega)\mathbf{a}(\omega, \mathbf{P}_0)}{\mathbf{a}^H(\omega, \mathbf{P}_0)\mathbf{V}^{-1}(\omega)\mathbf{a}(\omega, \mathbf{P}_0)}. \tag{11}$$

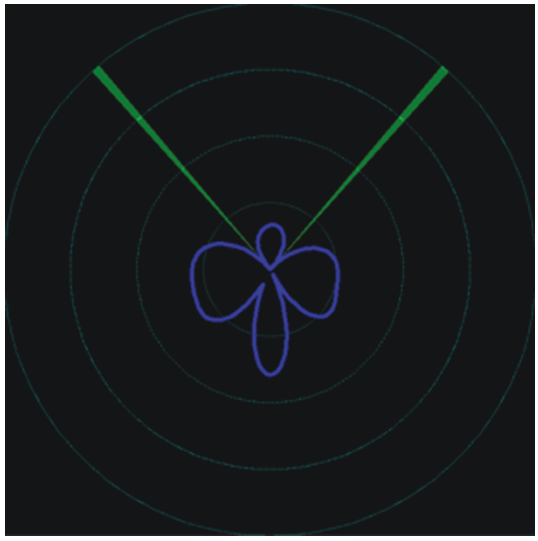


Figure 3: An example of the directional characteristics.

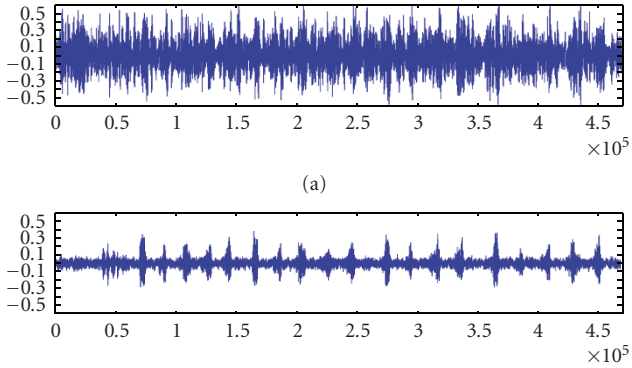


Figure 4: An example of the segregation of a user's voice from surrounding noise sources. (a) Waveform of mixed sounds. (b) Waveform of segregated user voice.

The FFT of the emphasized speech signal is given by

$$\hat{y}_F(\omega, n) = \mathbf{w}^H(\omega) \cdot \mathbf{y}_F(\omega, n). \tag{12}$$

The emphasized speech signal of the time domain is obtained by calculating the inverse FFT of (12).

Figure 3 shows an example of the directional characteristics determined by the modified MVBF. In this example, there are two directional noise sources. The green lines indicate the estimated arrival directions by the MUSIC-based method mentioned above. The blue line indicates the directional characteristics of 1.4 kHz. The gains in the noise directions are set to almost zero. Consequently, the surrounding noises are suppressed well with this beamformer. Figure 4 is an example of the segregation of a user's voice from surrounding noise sources. Two loudspeakers emitting different TV program sounds were placed facing the user, with one to their right and the other to their left. The user uttered voice commands several times. Figure 4(a) shows the waveform of the mixed sounds while Figure 4(b) shows the waveform of the segregated user's voice; the SNR of the user's voice was drastically improved. However, because directional noises cannot be completely eliminated and some omnidirectional noises still exist, the speech recognition accuracy is actually not very high. We therefore apply a feature compensation method after the microphone array processing.

3. HMM-Based Feature Compensation

The microphone array system is very effective at suppressing directional noise sources. However, it tends to be less effective for omnidirectional noises. In order to make the speech recognition more robust in a variety of noise environments, we added hidden Markov models (HMMs), based feature compensation method [8], to the microphone array system.

There is an additional advantage associated with combining the microphone array system and the feature compensation method. The feature compensation method needs precise voice activity detection. Generally, it is not always easy to detect voice activity from a noise-corrupted speech signal of a single channel. Such poor accuracy of voice activity detection degrades the accuracy of feature compensation. In contrast, the microphone array system can detect voice activity even in the presence of surrounding noises. In our system, therefore, the feature compensation method can utilize reliable voice activity detection to guarantee feature compensation accuracy.

The feature compensation method assumes that distortion of the noisy speech feature in the cepstral domain can be divided into stationary and nonstationary distortion components. The temporal trajectory of the nonstationary distortion component is assumed to be zero almost everywhere, although it temporarily changes. The stationary distortion component is absorbed by adding the estimated stationary distortion component to the expectation value of each Gaussian distribution in the output probability density functions (pdfs) of HMMs of the clean speech. The degradation of feature compensation accuracy caused by the non-stationary distortion component is compensated by evaluating each noise-adapted Gaussian distribution's posterior probability multiplied by the forward path probability.

The noisy speech feature $\mathbf{x}_C$ in the cepstral domain can be represented by $\mathbf{x}_C = \mathbf{s}_C + g(\mathbf{s}_C, \mathbf{n}_C)$, where $\mathbf{s}_C$ is the clean speech feature in the cepstral domain, $\mathbf{n}_C$ is the noise feature, and $g(\mathbf{s}_C, \mathbf{n}_C)$ is the distortion component given by

$$g(\mathbf{s}_C, \mathbf{n}_C) = \mathbf{C} \cdot \log[\mathbf{1} + \exp\{\mathbf{C}^{-1} \cdot (\mathbf{n}_C - \mathbf{s}_C)\}], \tag{13}$$

where $\log(\mathbf{a})$ and $\exp(\mathbf{a})$ calculate the logarithm and exponential of each element in a vector $\mathbf{a}$, and $\mathbf{C}$ and $\mathbf{C}^{-1}$denote the DCT matrix and its inverse transformation matrix, respectively.

The feature compensation process consists of the following six steps.

(1) Generate the copied Gaussian distributions of clean speech in an output pdf of each state.

The output pdf of the jth state is represented by

$$b_j(\mathbf{s}_C) = \sum_{m=1}^{M} w_{jm} N\left(\mathbf{s}_C; \mu_{jm}, \mathbf{V}_{jm}\right). \tag{14}$$

(2) Evaluate the stationary distortion component $\mathbf{d}_{jm}$ for each copied Gaussian distribution.

Distortion component $\mathbf{d}_{jm}$ is evaluated using the expectation value of each Gaussian distribution and the noise-only frames prior to each utterance. This can be represented as

$$\mathbf{d}_{jm} = \frac{1}{N_{\mathrm{n}}}\mathbf{C} \cdot \sum_{n=1}^{N_n} \log\left[\mathbf{1} + \exp\left\{\mathbf{C}^{-1} \cdot \left(\mathbf{n}_C(n) - \mu_{jm}\right)\right\}\right], \tag{15}$$

where $\mathbf{n}_C(n)$ represents the noise feature extracted from the noise-only frame, and N_n is the number of noise-only frames.

(3) Adapt each copied Gaussian distribution of clean speech to the noisy speech.

This adaptation can be achieved by adding each evaluated stationary distortion component to the expectation value of each copied Gaussian distribution. In this noise adaptation process, we take into account only the expectation value of each Gaussian distribution. The diagonal covariance matrix in the noise-adapted Gaussian distribution is assumed to be the same as the covariance matrix of the clean speech. The noise-adapted output pdf is given by

$$\hat{b}_j(\mathbf{x}) = \sum_{m=1}^{M} w_{jm} N\left(\mathbf{x}; \mu_{jm} + \mathbf{d}_{jm}, \mathbf{V}_{jm}\right). \tag{16}$$

(4) Evaluate the importance of each noise-adapted Gaussian distribution.

The importance of each noise-adapted Gaussian distribution is evaluated by the posterior probability multiplied by the normalized forward path probability:

$$P_{jm} = \frac{\alpha'(j, n-1) w_{jm} N\left(\mathbf{x}; \mu_{jm} + \mathbf{d}_{jm}, \mathbf{V}_{jm}\right)}{\sum_{s \in \mathrm{AllStates}} \sum_{q=1}^{M} \alpha'(s, n-1) w_{sq} N\left(\mathbf{x}; \mu_{sq} + \mathbf{d}_{sq}, \mathbf{V}_{sq}\right)}, \tag{17}$$

where $\alpha'(j, n)$ denotes the normalized forward path probability, given by

$$\alpha'(j, n-1) = \exp\left\{\frac{\alpha(j, n-1)}{n}\right\}. \tag{18}$$

The forward path probability $\alpha(j, n)$ is obtained from the Viterbi decoding process.

(5) Estimate the average stationary distortion component.

The average stationary distortion component is estimated by averaging the stationary distortion components $\mathbf{d}_{jm}$ weighted by the importance of each noise-adapted Gaussian distribution:

$$\overline{\mathbf{d}} = \sum_{j \in \mathrm{AllStates}} \sum_{m=1}^{M} P_{jm} \mathbf{d}_{jm}. \tag{19}$$

(6) Compensate the noise-corrupted speech feature.

The compensated speech feature $\widetilde{\mathbf{s}}$ is obtained by subtracting the average stationary distortion component from the noise-corrupted speech feature:

$$\widetilde{\mathbf{s}} = \mathbf{x} - \overline{\mathbf{d}}. \tag{20}$$

The original Gaussian distributions of clean speech are used to evaluate the output probability of the compensated speech feature $b_j(\widetilde{\mathbf{s}})$ in the Viterbi decoding process.

4. System Overview

Our system consists of one CPU board of a Pentium-M 2.0 GHz, 8-channel A/D converter, and a DC-DC converter. These devices are placed in an aluminum case of size W30 × H7 × D18 cm, which can be hidden under the seat of the wheelchair. The system devices that can be easily seen are only the microphone array system and the LCD showing the recognition results.

The system embedded in the wheelchair must execute the following five functions: (1) detection of user utterance and noises, (2) enhancement of user utterance, (3) speech feature compensation, (4) speech recognition, and (5) wheelchair control. We developed software that can execute these functions in real time on the CPU board. The sampling rate of the A/D converter is set to 8 kHz due to the limitation of the processing capacity of the CPU board.

A motor controller is connected to the CPU board by an RS232C serial cable. The wheelchair shown in Figure 1 has two motors to drive the left and right wheels independently. The CPU board can dictate the rotation speeds of the motors independently by the controller. So, not only the speed of the wheelchair but also the radius of the wheelchair rotation can be easily controlled from the CPU board.

5. Experiments

The proposed system consists of two noise robust methods: the microphone array and the feature compensation. We first evaluate the relative gains of each method and then compare the performance of the proposed method with the headset microphone.

Table 1: Recognition accuracy of the single microphone.

Single Microphones														
Avg.	Near a kindergarten	A construction Site near a train	Under train rails	In front of an amusement arcade	A building under construction	A public office	Wind noise	Along a big street	A road crossing	A construction site	A shop	In front of a station	In front of a ticket gate	Avg.
Clean	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
SNR20	90.83	98.61	97.08	90.00	97.78	91.67	54.03	99.72	68.33	94.17	94.44	97.78	95.00	89.96
SNR15	73.47	92.08	82.64	58.19	91.11	62.22	34.03	98.47	45.69	73.89	69.58	84.72	69.72	71.99
SNR10	50.69	58.61	52.50	24.72	74.17	34.72	25.69	88.33	27.08	47.22	35.00	49.72	35.42	46.45
SNR5	32.78	33.89	33.89	20.00	49.58	27.64	21.53	59.31	21.53	29.72	21.81	23.75	22.22	30.59
SNR0	25.00	23.47	27.36	20.00	31.11	20.42	21.53	36.81	20.14	21.81	19.86	20.28	20.42	23.71
SNR-5	21.67	20.00	20.28	20.00	23.19	19.86	22.78	29.86	20.00	20.14	19.86	20.00	20.00	21.36
Avg.	49.07	54.44	52.29	38.82	61.16	42.76	29.93	68.75	33.80	47.83	43.43	49.38	43.80	47.34

Table 2: Recognition accuracy of the single microphone followed by feature compensation.

Single Microphone + Feature Compensation														
Avg.	Near a kindergarten	A construction Site near a train	Under train rails	In front of an amusement arcade	A building under construction	A public office	Wind noise	Along a big street	A road crossing	A construction site	A shop	In front of a station	In front of a ticket gate	Avg.
Clean	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
SNR20	97.64	99.72	99.58	99.03	99.72	99.72	97.22	100	99.03	99.58	99.72	99.86	99.58	99.26
SNR15	91.67	99.03	99.31	97.78	98.61	99.72	94.44	100	96.53	99.03	99.17	99.72	98.89	97.99
SNR10	72.78	97.5	98.06	90.28	95.56	97.36	88.19	99.86	88.06	95.56	96.94	98.89	97.22	93.56
SNR5	52.78	92.5	92.78	57.08	91.53	89.17	80.14	98.75	68.19	89.58	91.81	95.56	91.53	83.95
SNR0	40.42	75.42	73.33	25.83	77.92	64.58	69.31	94.31	43.89	68.61	67.78	80.28	67.92	65.35
SNR-5	32.36	45.14	49.44	22.36	53.75	37.36	53.61	85.69	28.47	44.72	36.25	43.61	38.75	43.96
Avg.	64.61	84.89	85.42	65.39	86.18	81.32	80.49	96.44	70.70	82.85	81.95	86.32	82.32	80.68

5.1. Recognition Accuracy Evaluations. To assess the noise robustness and relative gains of each method, we evaluated the recognition accuracies of the following methods:

(i) Method (A): Single microphone,

(ii) Method (B): Single microphone followed by feature compensation,

(iii) Method (C): Microphone array,

(iv) Method (D): Microphone array followed by feature compensation (proposed method).

In the methods using the single microphone, the user's utterances were recorded by the microphone closest to the user on the right-hand microphone array unit. Each voice command was manually segmented to include silence durations and then recognized. The recognition accuracies of the methods using the microphone array were evaluated without any segmentation information except the voice activity detection by the method described in Section 2.1.

We recorded clean speech signals and environmental noises separately and then mixed the digital signals of these together at six different SNR levels (20 dB, 15 dB, 10 dB, 5 dB, 0 dB, −5 dB) to generate noise-corrupted speech signals. We define the SNR of the multichannel signals of the microphone array as follows. Let $S_{T,i}^{MA}(n)$ and $N_{T,i}^{MA}(n)$ represent a clean speech signal and environmental noise, respectively, in the time domain recorded by the ith microphone. The average powers of the clean speech signals and environmental noise signals are given by

$$\overline{S}_{MA} = \frac{1}{8N}\sum_{i=1}^{8}\sum_{n=1}^{N}\left\{S_{T,i}^{MA}(n)\right\}^2,$$
$$\overline{N}_{MA} = \frac{1}{8N}\sum_{i=1}^{8}\sum_{n=1}^{N}\left\{N_{T,i}^{MA}(n)\right\}^2. \tag{21}$$

The SNR of the multichannel signals is given by

$$\mathrm{SNR}_{MA}[\mathrm{dB}] = 10\log_{10}\left(\frac{\overline{S}_{MA}}{\overline{N}_{MA}}\right). \tag{22}$$

To make it possible to compare the results of a single microphone with the results of the microphone array, the SNRs of the noise-corrupted speech signals in the single microphone

Table 3: Recognition accuracy of the microphone array.

Microphone Array														
Avg.	Near a kindergarten	A construction Site near a train	Under train rails	In front of an amusement arcade	A building under construction	A public office	Wind noise	Along a big street	A road crossing	A construction site	A shop	In front of a station	In front of a ticket gate	Avg.
Clean	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
SNR20	99.72	100.00	99.58	96.94	99.86	99.31	89.03	100.00	96.94	99.44	99.58	99.86	99.86	98.47
SNR15	95.97	99.58	96.67	81.81	99.31	95.69	71.94	100.00	85.14	97.78	97.22	98.47	98.19	93.67
SNR10	85.28	93.33	84.86	39.86	96.25	79.17	50.00	99.72	59.03	87.64	81.67	92.08	85.56	79.57
SNR5	62.50	63.06	56.25	22.08	87.92	44.17	38.75	95.00	37.64	58.61	48.06	61.94	50.97	55.92
SNR0	42.92	30.28	36.39	20.00	62.08	29.31	34.44	72.08	24.17	34.03	25.42	28.33	28.06	35.96
SNR-5	26.81	20.69	25.14	20.00	31.39	27.50	32.22	40.97	20.97	25.42	20.42	20.00	21.53	25.62
Avg.	68.87	67.82	66.48	46.78	79.47	62.53	52.73	84.63	53.98	67.15	62.06	66.78	64.03	64.87

Table 4: Recognition accuracy of the microphone array followed by feature compensation.

Microphone Array + Feature Compensation														
	Near a kindergarten	A construction Site near a train	Under train rails	In front of an amusement arcade	A building under construction	A public office	Wind noise	Along a big street	A road crossing	A construction site	A shop	In front of a station	In front of a ticket gate	Avg.
Clean	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
SNR20	99.03	100.00	99.72	98.61	100.00	100.00	99.31	100.00	98.47	99.58	99.86	100.00	99.86	99.57
SNR15	98.06	99.58	99.17	96.39	99.44	99.17	97.64	100.00	92.22	98.47	98.75	99.44	98.19	98.19
SNR10	90.56	98.61	97.08	89.03	96.67	96.39	95.97	99.86	85.00	96.25	94.58	98.47	96.53	95.00
SNR5	78.89	95.14	86.25	78.47	88.19	85.28	90.28	98.61	71.94	83.89	81.53	93.61	89.17	86.25
SNR0	67.08	83.06	65.00	57.22	68.33	62.78	82.36	97.36	55.69	60.69	49.86	78.06	68.47	68.92
SNR-5	54.17	52.78	42.08	41.67	40.69	34.44	67.78	88.19	40.28	37.92	16.53	45.97	39.03	46.27
Avg.	81.30	88.20	81.55	76.90	82.22	79.68	88.89	97.34	73.93	79.47	73.52	85.93	81.88	82.37

experiments were evaluated using all the channel signals of the microphone array in the same manner shown in (21) and (22). The clean speech signals were recorded with the user, sitting in the wheelchair and uttering a command in a silent room. Because the purpose of the experiments was to assess the noise robustness, the users (29 females and 19 males) were able-bodied. The users uttered 13 commands in the user utterance area while looking forward and to the right and left. In this experiment, we used five Japanese commands: *mae* (forward), *migi* (right), *hidari* (left), *ushiro* (backward) and *teishi* (stop). The environmental noises were recorded by actually moving the wheelchair in 13 locations:

(1) a construction site near a train,

(2) a construction site only,

(3) a building under construction,

(4) under train rails,

(5) in front of an amusement arcade,

(6) near a kindergarten,

(7) a public office,

(8) in wind,

(9) along a big street,

(10) a road crossing,

(11) a store,

(12) in front of a train station,

(13) in front of a ticket gate.

The sound source localization and beamforming of the microphone array system were executed every 125 milliseconds. Triphone acoustic models were trained from clean speech data obtained by downsampling the JNAS [9] data to 8 kHz.

Figure 5 shows the average recognition accuracies over all the environmental noises for all the methods. Table 1 shows the evaluated recognition accuracy of the single microphone (Method A). In the table, the average recognition accuracies were calculated using the accuracies ranging from the 20 dB to −5 dB. Table 2 shows the results of the single microphone followed by feature compensation (Method B). The recognition accuracies are drastically improved in comparison with those of the single microphone. Table 3 shows the results of the microphone array (Method C). If

Table 5: Recognition accuracy of a headset microphone.

Headset Microphone														
	near a kinder-garten	a con-struction Site near a train	under train rails	in front of an amuse-ment arcade	a building under con-struction	a public office	wind noise	along a big street	a road crossing	a con-struction site	a shop	in fron of a station	in front of a ticket gate	Avg.
Clean	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
SNR20	98	100	99.2	97.6	98	100	81.2	100	95.2	99.2	98	100	99.6	97.38
SNR15	87.6	100	85.6	79.6	89.2	97.2	66.4	100	81.2	94.4	88	98.8	89.2	89.02
SNR10	70.8	90.8	62.8	36.8	64.8	76.4	43.2	99.6	58.8	81.2	50.4	85.6	59.2	67.72
SNR5	50	58.8	34.8	17.6	40.4	50.8	24.8	94.8	37.2	53.2	28.8	43.2	31.6	43.54
SNR0	32.8	30.4	9.6	20	29.2	39.6	20.8	67.6	29.6	36.4	22	27.2	18.4	29.51
SNR-5	25.6	22	20	20	17.6	30.4	20	41.2	27.6	24.4	22.4	21.6	24.8	24.43
Avg.	60.80	67.00	52.00	45.27	56.53	65.73	42.73	83.87	54.93	64.80	51.60	62.73	53.80	58.60

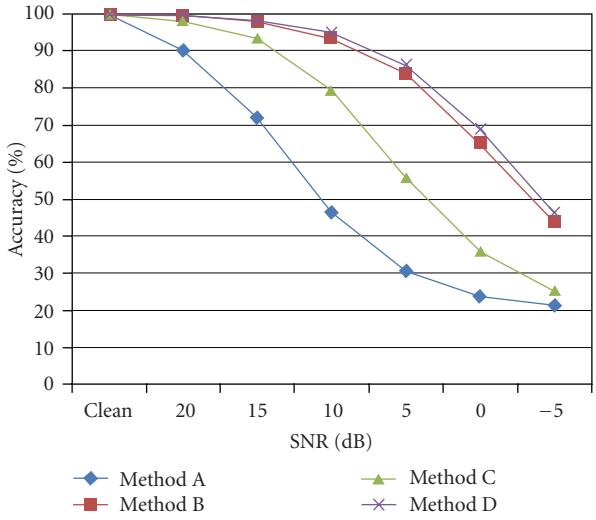


Figure 5: Average accuracies of all the methods.

we compare the results of Method B and Method C, we can say that feature compensation is more effective on these environmental noises than is the microphone array. Table 4 shows the results of the proposed method (Method D). The improvements in the recognition accuracies of the Method D seem to be small in comparison with those of Method B. This is because the environmental noises used in these evaluations tended to be omnidirectional. So, the feature compensation was rather effective than the microphone array. Furthermore, in the evaluations of Method B, each voice command was manually segmented. However, Method D detected each voice command based only on the VAD given by the signal processing of the microphone array. These results imply that the accuracy of VAD based on the microphone array is almost the same as that in manual segmentation. This is a very important benefit of the microphone array in addition to the sound source localization used to distinguish the user's voice from other voices. The microphone array is thus very important for achieving the noise robustness in the proposed method even if the environmental noises are omnidirectional.

5.2. Comparison with Headset Microphone. In this section, we evaluate the recognition accuracy of the conventional headset microphone and compare it with the accuracy of the proposed system.

The clean speech signals of 25 females and 25 males were recorded with the headset microphone (Audio-technica AT810X). The users uttered the same five commands of the previous experiments. We used the same environmental noises of the previous experiments to generate noise-corrupted speech signals by mixing the clean speech signal and the environmental noises at six different SNR levels. Table 5 shows the evaluated recognition accuracy of the headset microphone.

The SNRs of the noise-corrupted speech signals of the headset microphone are evaluated by the following equations:

$$S_{HSM} = \frac{1}{N}\sum_{n=1}^{N}\left\{S_T^{HSM}(n)\right\}^2,$$

$$N_{HSM} = \frac{1}{N}\sum_{n=1}^{N}\left\{N_T^{HSM}(n)\right\}^2, \tag{23}$$

$$\mathrm{SNR}_{HSM}[dB] = 10\log_{10}\left(\frac{S_{HSM}}{N_{HSM}}\right).$$

To compare Table 5 with Table 4, we need to convert the SNRs evaluated by (23) to the SNR defined by (21) and (22), because these SNRs are defined under different conditions. As defined in (2), we assume that the noise source outside the wheelchair is received at each microphone with uniform gains. In addition to this assumption, we assume the headset microphone is omnidirectional. Based on these assumptions, we can also say that $\overline{N}_{MA} \approx N_{HSM}$. Therefore, the headset microphone was placed closer to the user than

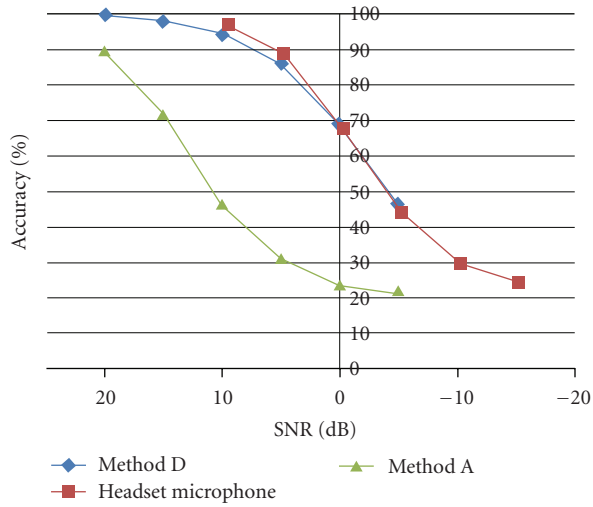


Figure 6: The Performance of the proposed method (Method D) with the headset microphone and the single microphone (Method A) in the microphone array.

the microphone array. We assume the simple relation $\overline{S}_{MA} = S_{HSM}/\alpha, (a > 1)$ between speech signal powers. From these assumptions, we obtain the relation between two SNRs, as follows:

$$\mathrm{SNR}_{MA}[\mathrm{dB}] \approx \mathrm{SNR}_{HSM}[\mathrm{dB}] - 10\log_{10}(\alpha). \quad (24)$$

We actually measured α and obtained the value $10\log_{10}(\alpha) = 10.31[\mathrm{dB}]$.

In Figure 6, the average recognition accuracies over all the environmental noises of proposed method (Method D) are compared with those of the headset microphone. The recognition results of the headset microphone in Table 5 were plotted by shifting the SNR according to (24). The recognition results of the single microphone (Method A) in the microphone array are also plotted. The distance between the headset microphone and the single microphone is approximately 45 cm. The recognition accuracies of the single microphone were drastically degraded. Two microphone array units were also placed at approximately 45 cm from the headset microphone. However, the microphone array was able to achieve almost the same recognition accuracies as those of the headset microphone.

6. Conclusions

We developed a noise robust speech recognition system for a voice-driven wheelchair that combines the microphone array with the feature compensation method. The developed noise robust speech recognition system has the following advantages: (1) the microphone array system can distinguish the user's utterances from other voices without using a speaker identification technique, (2) the accuracy of VAD based on the microphone array is almost the same as that in manual segmentation, (3) the feature compensation method can utilize the reliable information of VAD from the microphone array, and (4) the weak point of the microphone array, which is processing omnidirectional noises, can be compensated by the feature compensation method. Consequently, our system can be applied to various noise environments. We verified the effectiveness of our system in experiments in different environments and confirmed that our system can achieve almost the same recognition accuracy as does the headset microphone without wearing sensors. As a result, we were able to develop a voice-driven wheelchair that does not require the user to wear a headset microphone.

Acknowledgments

This research was conducted as part of "Development of technologies for supporting safe and comfortable lifestyles for persons with disabilities," funded by the Solution-Oriented Research For Science And Technology (SORST) program of the Japan Science and Technology Agency (JST), Ministry of Education, Culture, Sports, Science and Technology (MEXT) of the Japanese Government. This work was also supported by KAKENHI 20700471, funded by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of the Japanese Government.

References

[1] G. E. Miller, T. E. Brown, and W. R. Randolph, "Voice controller for wheelchairs," *Medical & Biological Engineering and Computing*, vol. 23, no. 6, pp. 597–600, 1985.

[2] R. Amori, "VOCOMOTION—an intelligent voice-control system for powered wheelchair," in *Proceedings of the 15th RESNA Annual Conference*, pp. 421–423, Tronto, Canada, 1992.

[3] W. McGuire, "Voice operated wheelchair using digital signal processing technology," in *Proceedings of the 22nd RESNA Annual Conference*, pp. 364–366, 1999.

[4] R. C. Simpson and S. P. Levine, "Voice control of a powered wheelchair," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 10, no. 2, pp. 122–125, 2002.

[5] The demonstration video of the voice-driven wheelchair, http://staff.aist.go.jp/a-sasou/demovideo.html.

[6] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[7] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 497–507, 2000.

[8] A. Sasou, F. Asano, S. Nakamura, and K. Tanaka, "HMM-based noise-robust feature compensation," *Speech Communication*, vol. 48, no. 9, pp. 1100–1111, 2006.

[9] K. Itou, M. Yamamoto, K. Takeda, et al., "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.