Two-Dimensional DOA Estimation of Sound Sources Based on Weighted Wiener Gain Exploiting Two-Directional Microphones

Yoshifumi Nagata, Toyota Fujioka, and Masato Abe, *Member, IEEE*

***Abstract*—We propose a new method for estimating directions of arrival (DOAs) of sound sources, both in azimuthal and elevation angle, using two directional microphones. This method adopts weighted Wiener gain (WWG) for DOA estimation. WWG is an estimate of the Wiener gain that we proposed for use in automatic gain control to enhance speech that is degraded by additive noise. Angular resolution of WWG arises from spectral subtraction (SS)-based noise reduction involved in the WWG calculation, which enhances the signal from the look direction while suppressing signals from other directions. Because WWG involves two-channel SS, which can deal with instantaneous noise, noise sources need not to be stationary, as they must be with ordinary single-channel SS. We further propose the exploitation of a pair of directional microphones whose front directions are arranged in rotational symmetry. The time difference and amplitude difference between the two-channel signal provided by the microphones are utilized to yield a two-dimensional resolution of DOA. We evaluated the proposed method through computer simulations and compared it to three DOA estimation methods that are based on a cross-correlation function and two popular high-resolution methods of multiple signal classification and minimum variance method. Evaluation results of the source detection rate and estimation accuracy demonstrate the remarkable superiority of our method compared to the other methods in conditions where multiple speech sources exist.**

***Index Terms*—Directional microphone, direction of arrival (DOA) estimation, elevation angle, two-channel, multiple signal classification (MUSIC), minimum variance (MV), Wiener gain.**

I. Introduction

The ARRIVAL angle of a sound signal is considered to be important information for many applications: noise reduction, speech dialog systems, robot audition, etc. Among the many direction of arrival (DOA) estimation methods that have been proposed for array signal processing (e.g., [1]) with various numbers of sensors, the two-channel technique is particularly attractive because of its hardware costs and processing costs. However, performance limitations of the two-channel technique in a multiple source environment are important because popular high-resolution methods such as multiple signal classification (MUSIC) [2] and minimum variance (MV) method [3] are ineffective for conditions in which the sound sources are more numerous or equal in number to microphones. Consequently, a two-channel system permits the use of only one source. The detectable number of sources in a DOA estimation is an important factor, particularly for systems with a small number of microphones, because the number of sound sources can change frequently in a practical condition. Sound sources can easily overwhelm the theoretical limit of a method. Even in such cases, stable performance is desirable to retain estimation ability, particularly for a slight increase in the number of sources. Therefore, we are interested in the behavior of the DOA estimation system in cases where the sources are more numerous than the theoretically detectable number.

The theoretically detectable number of sources in DOA estimation depends on the principle for reducing the contribution of the source signals arriving from outside the look direction. Most high-resolution methods steer nulls on the array directivity to the source directions to reduce the contribution of source signals. This operation provides sharp peaks at source directions on the "spatial spectrum," which is a plot of the array response, as a function of the imaginary look direction. The number of detectable sources in the high-resolution method is known to be equal to the number of nulls that can be generated on the directivity: It is usually $M-1$, where M is the number of microphones. This limitation seems to have been moderated by the method using a nonlinear complementary beamformer [4]; it was raised to $2(M-1)$, i.e., two, for a two-channel system. However, this method involves nonlinear searching, which can be computationally intensive for broadband two-dimensional (2-D) DOA estimation because the nonlinear search must be repeated for every imaginary look direction and for every spectral component that is necessary to obtain a spatial spectrum. The method based on classical time delay estimation using the generalized cross correlation function [5] does not have that theoretical limitation, but its detection performance is generally lower than that of the high-resolution methods because noise reduction involved in this method is merely the result of averaging.

We consider the DOA estimation problem in a multiple sound source environment. To simplify the problem while retaining its realism, we restrict the condition to cases in which every sound source radiates a nonstationary broadband signal, such as speech. In this condition, we infer that all source spectra mutually differ. Furthermore, each source spectrum changes independently over time. This nonstationarity implies that some spectral components originate mainly from one source at a specified frequency or within a specified time frame. If those spectral components were known, we could select them and average the spatial spectrum obtained from each of them using, e.g., MUSIC.

Manuscript received December 17, 2004; revised March 24, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bayya Yegnanarayana.

The authors are with the Department of Computer and Information Science, Iwate University, Morioka 020-8551, Japan (e-mail: nagata@cis.iwate-u.ac.jp; toy@cis.iwate-u.ac.jp).

Digital Object Identifier 10.1109/TASL.2006.881699

1558-7916/$25.00 © 2006 IEEE

Thereby, we obtain the spatial spectrum that reflects the DOA of all sources. Consequently, an increased number of detectable sources is possible, but it is usually difficult to know such components before the DOAs are known.

To perform "blind" averaging of the spatial spectrum that requires no relation between spectral components and their original sources, we propose the use of weighted Wiener gain (WWG) [6] for DOA estimation. The WWG is an estimate of the Wiener gain we have proposed to utilize for speech enhancement: WWG is based on automatic gain control (AGC). Two weighting functions that are applied for the gain estimation yield an accurate estimate of the Wiener gain, even in conditions where impulsive noises exist. One weighting function is related to the two-channel version [7] of spectral subtraction [8] (2chSS); the other is a function to whiten the noise spectrum for improving noise reduction in the step of averaging along the frequency axis in WWG calculation. The 2chSS-related function uses the spectral difference between the desired signal arriving from the look direction and other signals. Noise reduction in 2chSS process is not based on null steering, but on SS. For that reason, a theoretical limitation of the detectable number does not exist. We consider that this characteristic of WWG is useful to estimate DOA in a two-channel system.

Estimation of the source's elevation angle poses an important problem. Two-channel DOA estimation systems usually exploit time differences of signals between channels. Therefore, only the azimuthal resolution of angle is available when microphones are arranged in a horizontal plane. In this case, directional microphone is useful to obtain resolution in the elevation angle. The difference in amplitude arises between the channels if the front directions of the directional microphones are arranged either not to coincide or not to belong to the horizontal plane. This information can bring resolution in the elevation angle. Therefore, we further propose to exploit signals that are acquired by a pair of directional microphones that are arranged in rotational symmetry. This arrangement can clarify both the time delay and the amplitude difference between channels. For those reasons, it is expected to yield 2-D resolution of DOA estimation.

The remainder of this paper is organized as follows. Section II describes a brief summary of speech enhancement based on WWG. Section III describes the proposed method of DOA estimation based on WWG. Section IV describes the experimental setup for evaluation. Section V describes evaluation of the proposed method compared with MUSIC, MV and three cross-correlation-based methods. Finally, Section VI summarizes the conclusion.

II. Speech Enhancement Based on Automatic Gain Control With Weighted Wiener Gain

A. *Automatic Gain Control With Weighted Wiener Gain*

We assume that two directional microphones are placed in a noisy environment to receive identical desired signals, as shown in Fig. 1. Let the discrete time samples of the signals received at the microphones be

$$\begin{aligned} x(i) &= s(i) + n_x(i) \\ y(i) &= s(i) + n_y(i) \end{aligned} \tag{1}$$

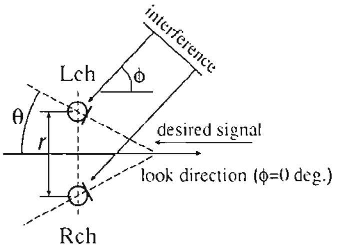

Fig. 1. Microphone arrangement for speech enhancement based on WWG.

where $x(i)$ and $y(i)$, respectively, denote the L-channel and R-channel microphone signals, $s(i)$ is the desired signal, and $n_x(i)$ and $n_y(i)$ are the noises received at respective microphones. Subjecting the above samples to short-time discrete Fourier transform (DFT), we obtain

$$\begin{aligned} X_{n,k} &= S_{n,k} + N_{x,n,k} \\ Y_{n,k} &= S_{n,k} + N_{y,n,k} \end{aligned} \tag{2}$$

where $X_{n,k}$ and $Y_{n,k}$ denote the DFT of the $x(i)$ and $y(i)$ for the frame n and the kth frequency bin; $S_{n,k}$ denotes that of $s(i)$, and $N_{x,n,k}$ and $N_{y,n,k}$, respectively, denote those of $n_x(i)$ and $n_y(i)$.

Consider the case in which the average of the received signals $Z_{n,k} = (X_{n,k} + Y_{n,k})/2$ is multiplied by a scalar gain ρ_n for approximating the desired signal contained in $Z_{n,k}$ as

$$\hat{S}_{n,k} = Z_{n,k}\rho_n. \tag{3}$$

Gain ρ_n is obtainable as a weighted least-squares solution to minimize the following cost function assuming that gain ρ_n and weighting function $\Psi_{n,k}$ are constant within the period of time-averaging, as

$$J(\rho_n) = \sum_k \overline{|Z_{n,k}\rho_n - S_{n,k}|^2 \Psi_{n,k}} \tag{4}$$

$$= \frac{1}{2L+1} \sum_k \sum_{j=n-L}^{j<=n-L} |Z_{j,k}\rho_n - S_{j,k}|^2 \Psi_{n,k} \tag{5}$$

where $(^{-})$ denotes time averaging and $2L+1$ denotes the number of frames for time averaging. Because $J(\rho_n)$ is a quadratic form of ρ_n, the optimum value of ρ_n is obtained by letting $dJ(\rho_n)/d\rho_n = 0$. Then, we obtain

$$\sum_k (2\overline{|Z_{n,k}|^2}\rho_n - 2\mathrm{Re}[\overline{S_{n,k}Z^*_{n,k}}])\Psi_{n,k} = 0 \tag{6}$$

where $\mathrm{Re}[]$ denotes the operation that takes a real part of the complex number. Because the desired signal and the noises are assumed to be uncorrelated, $\overline{S_{n,k}Z^*_{n,k}} = \overline{|S_{n,k}|^2}$. Therefore, the weighted version of the Wiener gain is obtained as

$$\rho_n = \frac{\sum_k G_{ss,n,k}\Psi_{n,k}}{\sum_k G_{zz,n,k}\Psi_{n,k}} \tag{7}$$

where

$$G_{ss,n,k} = \overline{|S_{n,k}|^2} = \frac{1}{2L+1} \sum_{j=n-L}^{j<=n-L} |S_{j,k}|^2 \tag{8}$$

and

$$G_{zz,n,k} = \overline{|Z_{n,k}|^2} = \frac{1}{2L+1}\sum_{j=n-L}^{j<=n+L}|Z_{j,k}|^2 \qquad (9)$$

respectively denote the power spectra of the desired signal and the primary signal. The power spectra of $X_{n,k}$ and $Y_{n,k}$ are obtained similarly as

$$G_{xx,n,k} = \overline{|X_{n,k}|^2} = \frac{1}{2L+1}\sum_{j=n-L}^{j<=n+L}|X_{j,k}|^2 \qquad (10)$$

$$G_{yy,n,k} = \overline{|Y_{n,k}|^2} = \frac{1}{2L+1}\sum_{j=n-L}^{j<=n+L}|Y_{j,k}|^2. \qquad (11)$$

If no interference is present and the background noise is uncorrelated between channels, $G_{ss,n,k}$ in (7) can be replaced by the cross spectrum, as

$$G_{xy,n,k} = \overline{X^*_{n,k}Y_{n,k}} = \frac{1}{2L+1}\sum_{j=n-L}^{j<=n+L}X^*_{j,k}Y_{j,k} \qquad (12)$$

where (*) signifies operation of the complex conjugate. Thereby, (7) becomes

$$\rho_n = \frac{\sum_k \mathrm{Re}(G_{xy,n,k})\Psi_{n,k}}{\sum_k G_{zz,n,k}\Psi_{n,k}}. \qquad (13)$$

We assume that $\Psi_{n,k}$ has a real value. The imaginary part of $G_{xy,n,k}$ can be ignored because the signals that come from the look direction are assumed to be identical among channels.

Next, we assume that the received noise signals contain broadband interference arriving from angle ϕ and uncorrelated background noise. Taking into account that the received interference signals differ in amplitude and phase in the microphone arrangement described above, (2) becomes

$$\begin{aligned} X_{n,k} &= S_{n,k} + V_{n,k} + B_{x,n,k} \\ Y_{n,k} &= S_{n,k} + \alpha_{\phi,k}V_{n,k}e^{-j2\pi f_s\tau k/K} + B_{y,n,k} \end{aligned} \qquad (14)$$

where $B_{x,n,k}$ and $B_{y,n,k}$ are the DFTs of uncorrelated background noise, $V_{n,k}$ is the DFT of the interference, $\alpha_{\phi,k}$ is the relative amplitude of the interference normalized by that contained in the L-channel signal, f_s is the sampling frequency, τ is the time delay of the interference between channels, and K represents the point length of the DFT.

In this case, the numerator of (13) is expressed as the following by substituting (14) into (13)

$$\begin{aligned} &\sum_k \mathrm{Re}(G_{xy,n,k})\Psi_{n,k} \\ &\quad = \sum_k \mathrm{Re}(\overline{|S_{n,k}|^2} + \alpha_{\phi,k}\overline{|V_{n,k}|^2}e^{-j2\pi f_s\tau k/K})\Psi_{n,k}. \end{aligned} \qquad (15)$$

If the weighting function $\Psi_{n,k}$ functions to whiten the second term in (15) and the phase $2\pi f_s\tau k/K$ is distributed uniformly within the range of $-\pi$ to π, the summation along frequency bin k reduces the summed power of the interference to a lower level than that of the desired signal. We can closely simulate this condition in most cases by taking a sufficient inter-microphone distance. Because it is difficult to estimate the interference spectrum directly from observations, we have chosen $\Psi_{n,k}$ to approximate the inverse of the noise spectrum as

$$\Psi_{n,k} = 1/G_{dd,n,k}, \qquad (16)$$

$$G_{dd,n,k} = \overline{|X_{n,k} - Y_{n,k}|^2} \qquad (17)$$

$$= \frac{1}{2L+1}\sum_{j=n-L}^{j<=n+L}|X_{j,k} - Y_{j,k}|^2 \qquad (18)$$

where $G_{dd,n,k}$ is the power spectrum of the differenced signal $X_{n,k} - Y_{n,k}$. The desired signal is reduced by the differencing operation for use as an approximation of the noise signal. In addition, $G_{dd,n,k}$ is used by the 2chSS-based noise reduction, as described later.

Components of the interference signal that is contained in the differenced spectrum are expressed as the following:

$$V'_{n,k} = V_{n,k}(1 - \alpha_{\phi,k}e^{-j2\pi f_s\tau k/K}). \qquad (19)$$

If $\alpha_{\phi,k} = 1$, this operation can produce zero values on the interference spectrum. Noise spectrum distortion that is attributable to the differencing operation can not be disregarded for 2chSS-based speech enhancement. It is compensated in the process of 2chSS, as described in the next section.

On the other hand, the spectral zeros that are attributable to differencing are avoidable using the microphone arrangement, as shown in Fig. 1, because $\alpha_{\phi,k} \neq 1$. In addition, the distortion does not directly affect the broad band gain estimation because the averaging over speech frequency band can moderate that effect. For that reason, we use $\Psi_{n,k}$ without compensation as the whitening function. We note that bare differencing $X_{n,k} - Y_{n,k}$ provides zeros in the source spectrum for all frequencies in the case where the sound source is immediately in front of the array. However, in the DOA estimation process using WWG as described later in Section III, the microphone signals are to be modified before differencing according to the imaginary look direction d to obtain a spatial spectrum. Because each modification corresponding to each channel is different from every other when d does not coincide with the front direction, a spectral zero is avoidable. Otherwise, when d coincides with the source direction, the sound source is no longer regarded as a noise source, but as a desired source. The gain (13) is optimally obtained with the function Ψ, which reflects the other noise sources. Therefore, we do not need to regard this case as an exception.

B. 2chSS-Based Weighting Function

In addition to the above averaging with whitening, we have introduced 2chSS-based noise reduction for reducing correlated noise components in $G_{xy,n,k}$. 2chSS is a modification of the Griffiths–Jim generalized sidelobe canceler (GSC) [9], which performs noise cancellation using the differenced signal between the channels as a reference signal and the averaged signal of the input channels as the primary signal. Whereas GSC estimates a transfer function between the reference and the primary signal, 2chSS estimates the imaginary transfer function between the power spectra of the two signals. This imaginary transfer function is a set of real-valued coefficients called compensation coefficients. In our setup, the primary power spectrum corresponds to $G_{xy,n,k}$ and the reference power spectrum corresponds to $G_{dd,n,k}$. To allow these correspondences, we modified 2chSS as

$$\tilde{G}_{ss,n,k} = |G_{xy,n,k}| - \gamma G_{dd,n,k}/\nu_{n,k} \tag{20}$$

$$= |G_{xy,n,k}|\Phi_{n,k,\gamma}, \tag{21}$$

$$\Phi_{n,k,\gamma} = \frac{|G_{xy,n,k}| - \gamma G_{dd,n,k}/\nu_{n,k}}{|G_{xy,n,k}|}. \tag{22}$$

In those equations, $\tilde{G}_{ss,n,k}$ is an estimate of the desired power spectrum, γ is a positive constant to control the strength of the subtraction, $\nu_{n,k}$ is the compensation coefficient, and $\Phi_{n,k,\gamma}$ is the resultant 2chSS-based weighting function.

We calculated the compensation coefficient $\nu_{n,k}$ as

$$\nu_{n,k} = \frac{D_{n,k}}{|Q_{xy,n,k}|}, \tag{23}$$

$$D_{n,k} = \begin{cases} |X_{n,k} - Y_{n,k}|^2\lambda + D_{n-1,k}(1-\lambda) \\ \qquad\qquad\qquad\qquad \text{(noise period)} \\ D_{n-1,k} \qquad\qquad\quad \text{(speech period)} \end{cases} \tag{24}$$

$$Q_{xy,n,k} = \begin{cases} X^*_{n,k}Y_{n,k}\lambda + Q_{xy,n-1,k}(1-\lambda) \\ \qquad\qquad\qquad\qquad \text{(noise period)} \\ Q_{xy,n-1,k} \qquad\qquad\quad \text{(speech period)} \end{cases} \tag{25}$$

where $D_{n,k}$ is the averaged differenced spectrum in the noise period, $Q_{xy,n,k}$ is the cross spectrum in the noise period, and λ represents the learning factor. The noise period is determined according to the criterion described later in this section.

The WWG is the total gain that is obtained by combining (13) with (22) as

$$\rho'_n(\beta,\gamma) = \frac{\sum_k \mathrm{Re}(G_{xy,n,k})\Psi^{\beta}_{n,k}\Phi_{n,k,\gamma}}{\sum_k G_{zz,n,k}\Psi^{\beta}_{n,k}} \tag{26}$$

where β is a positive constant that is introduced to control the strength of the whitening. Because the noise periods are very short, it is difficult to obtain accurate compensation coefficients $\nu_{n,k}$ in (22) in cases where impulsive disturbances arise. However, in such cases, whitening combined with noise reduction realized by the above gain has been demonstrated to be particularly effective. Parameters β and γ were determined empirically because this is an ad hoc combination. For estimation of $\Phi_{n,k,\gamma}$, we determined the noise period based on the following criterion:

$$C_n = \rho'_n(\beta,\gamma)|_{\nu_{n,k}=1.0}. \tag{27}$$

Because a fixed value $\nu_{n,k} = 1.0$ is used as the compensation coefficient, this criterion requires no detection of the noise period.

Finally, speech enhancement based on AGC with WWG is performed as

$$\hat{S}_{n,k} = Z_{n,k}\rho'_n(\beta,\gamma). \tag{28}$$

Because WWG is the estimate of broadband signal-to-signal + noise ratio, WWG can be regarded as a degree of existence of the signal arriving from the look direction. For that reason, we consider that WWG is a useful parameter for speech enhancement, for speech detection, and for DOA estimation.

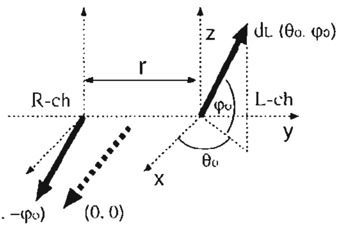

Fig. 2. Arrangement of directional microphones in rotational symmetry.

We can extend WWG of the two-channel system to a multichannel version using a simple averaging of WWG obtained from different combinations of channel pairs, as

$$\rho'(\beta,\gamma) = \frac{\sum_{\mathrm{P}}\sum_k \mathrm{Re}(G_{pq,k})\Psi^{\beta}_{pq,k}\Phi_{pq,k,\gamma}}{\sum_{\mathrm{P}}\sum_k G_{zz_{pq},k}\Psi^{\beta}_{pq,k}} \tag{29}$$

where P is a set of the combinations of all different pairs of channels and $p, q (p \neq q)$ are the two channel numbers from set P.

III. DOA Estimation Based on WWG

A. *Two-Channel Microphone System for 2-D DOA Estimation*

Next, we describe the proposed method of DOA estimation based on WWG. We assume that two directional microphones are placed as depicted in Fig. 2. Two thick solid arrows show the front directions of the microphones. The microphones are placed at the respective roots of the arrows. These directions of L-channel and R-channel microphones are expressed in polar coordinates; they are represented, respectively, as $d_L = (\theta_o, \varphi_o)$ and $d_R = (-\theta_o, -\varphi_o)$. Both microphones are assumed to have identical directivity. This arrangement represents the rotational symmetry of 180°. The sensitivities of the two microphones at directions (0°, 0°) are identical; those at the other directions are different in most cases, excluding direction $(-180°, -180°)$, which is the inverse of (0°, 0°). We ignore the effect of the signal from the inverse direction because sensitivities at the inverse are lower than that of (0°, 0°) when usual directional microphones, e.g., uni-cardioid microphones, are used. Consequently, we assume that the only signal that arrives from (0°, 0°) provides identical signals between the channels.

WWG takes a large value when signals are identical between channels. The magnitude of WWG is between 0 and 1 in most cases. Negative values of WWG rarely occur because spectral components that provide negative correlation usually accompany those having positive correlation in cases of broadband signal processing. If negative gain is obtained, we can use zero instead, as $\max(\rho'_n(\beta,\gamma), 0)$.

B. *WWG Spatial Spectrum*

The spatial spectrum of WWG for DOA estimation requires the value of WWG that corresponds to an arbitrary direction.

This value is available by simulating a case in which the look direction of the system is steered to that direction. To steer the look direction to a specified direction, the input spectrum $X_{n,k}, Y_{n,k}$ is modified both in phase and amplitude as

$$X'_{n,k}(d) = X_{n,k}A_{x,k}(d) \tag{30}$$
$$Y'_{n,k}(d) = Y_{n,k}A_{y,k}(d) \tag{31}$$
$$A_{x,k}(d) = 1/D_{x,k}(d) \tag{32}$$
$$A_{y,k}(d) = \exp(j2\pi f_s \tau_d k/K)/D_{y,k}(d) \tag{33}$$

where $d = (\theta, \varphi)$ is the look direction, $A_{x,k}(d)$ and $A_{y,k}(d)$, respectively, identify the compensation spectra for $X_{n,k}$ and $Y_{n,k}$, τ_d is the inter-channel time delay of the signal arriving from direction d, and $D_{x,k}(d), D_{y,k}(d)$ are the relative directivity normalized by the response of each microphone's front direction. We assume that $D_{x,k}(d)$ and $D_{y,k}(d)$ are known, by a preliminary measurement if necessary. The compensation described above creates a coincidence between the channels in both amplitude and phase of the signal components originating from the source at the look direction d. For simplicity, in this section, we describe the equations omitting the frame number n.

Because ρ' is dependent on G_{xy}, G_{zz}, and G_{dd}, we modify them to yield G'_{xy},G'_{zz} and G'_{dd} using (30)–(33) and obtain the steered version of WWG corresponding to the look direction d as

$$\rho'(\beta, \gamma, d) = \frac{\sum_k \mathrm{Re}[G'_{xy,k}(d)]\Psi'^{\beta}_k(d)\Phi'_{k,\gamma}(d)}{\sum_k G'_{zz,k}(d)\Psi'^{\beta}_k(d)} \tag{34}$$

$$G'_{xy,k}(d) = \overline{X'^*_k(d)Y'_k(d)} = G_{xy,k}A^*_{x,k}(d)A_{y,k}(d)$$
$$G'_{zz,k}(d) = \overline{|(X'_k(d) + Y'_k(d))/2|^2} \tag{35}$$
$$= G_{xx,k}|A_{x,k}(d)|^2/4 + G_{yy,k}|A_{y,k}(d)|^2/4$$
$$+ \mathrm{Re}[G_{xy,k}A^*_{x,k}(d)A_{y,k}(d)]/2 \tag{36}$$
$$G'_{dd,k}(d) = \overline{|X'_k(d) - Y'_k(d)|^2}$$
$$= G_{xx,k}|A_{x,k}(d)|^2 + G_{yy,k}|A_{y,k}(d)|^2$$
$$- 2\mathrm{Re}[G_{xy,k}A^*_{x,k}(d)A_{y,k}(d)] \tag{37}$$

where

$$\Psi'_k(d) = 1/G'_{dd,k}(d). \tag{38}$$
$$\Phi'_{k,\gamma}(d) = \frac{|G'_{xy,k}(d)| - \gamma G'_{dd,k}(d)/\nu'_k(d)}{|G'_{xy,k}(d)|} \tag{39}$$

are the modified weighting functions. Also, $\nu'_k(d)$ is the modified version of the compensation coefficient ν_k; $\nu'_k(d)$ depends on the results of desired signal detection shown in (24) and (25). Then, estimation of $\nu'_k(d)$ requires signal detection in every look direction d led by the steering. We omit this calculation and use $\nu'_k(d) = 1$ for all k and d because this estimation is computationally intensive, particularly when numerous look directions are required, as in the 2-D DOA estimation. Consequently, from (34), we can calculate steered WWG using the observed spectra G_{xx}, G_{yy} and G_{xy} without re-averaging for obtaining the steered spectra $G'_{xy}(d)$,$G'_{zz}(d)$ and $G'_{dd}(d)$ to make it correspond to every look direction. Thus, we can obtain the WWG spatial spectrum with reduced computation and can estimate DOA from the spectrum using peak picking, as described in a subsequent evaluation.

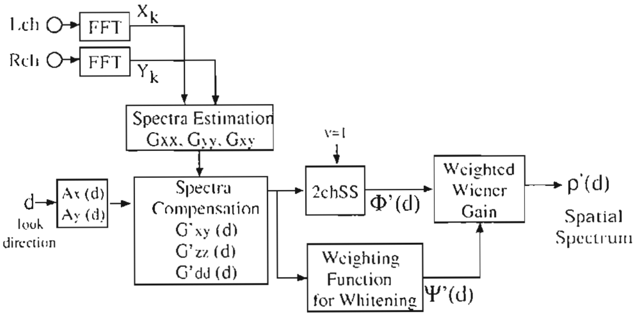

Fig. 3. Block diagram of the proposed DOA estimation.

C. Processing System

A block diagram of the proposed DOA estimation system based on WWG is depicted in Fig. 3.

First, the DFTs of the received signals are obtained via fast Fourier transform (FFT). The power and cross spectra G_{xx}, G_{yy}, and G_{xy} are estimated from the FFT spectra. Next, the modified spectra $G'_{xy,k}(d), G'_{zz,k}(d)$, and $G'_{dd,k}(d)$ are calculated from the power and cross spectra using expressions from (35) to (37) according to the look direction d. The spatial spectrum is obtained by changing d in the specified ranges of azimuth and elevation angle.

In this system, we use FFT of 512-point length with a Hanning window, a frame shift of 256 points, a frequency band for calculating the spatial spectra between 260 Hz and 4 kHz, and an inter-microphone distance r of 15 cm. The number of averaging iterations to estimate spectra $G_{xx,k}, G_{yy,k}$, and $G_{xy,k}$ to 3 ($L = 1.93$ ms) was determined through preliminary experiments. In addition, as described in the evaluation section, we set the index of the weighting function β to 0.75, the strength of the 2chSS γ to 3, the L-channel microphone direction d_L to (60°, 60°), and that of R-channel d_R to (−60°, −60°).

IV. Experimental Setup

A. Measures for Evaluation

1) Source Detection Rate: Performance of DOA estimation has been evaluated in the literature in terms of the shape of the spatial spectrum. Aspects that have been emphasized in those evaluations include angular resolution, peak sharpness, and accuracy of peak direction. Those aspects are important for quantitative evaluation, but we consider that they do not represent overall efficiency when used in a practical system. We consider that stabilizing factors through use of various source directions and through various numbers of sources are also important because both can vary unexpectedly in practical situations. The DOAs of a few dominant sources are expected to be estimated even in such conditions. In addition, abrupt performance degradation caused by a small increase in the number of sources is not desirable.

For the purposes described above, we propose the use of the source detection rate (SDR) as an evaluation measure

$$\mathrm{SDR}(N_s) = K_{\mathrm{success}}(N_s)/K_{\mathrm{total}}(N_s). \tag{40}$$

In that equation, N_s is the assumed number of sources, $K_{\text{total}}(N_s)$ is the number of trials of DOA detection, and $K_{\text{success}}(N_s)$ is the number of successful trials. For SDR measurements, the spatial spectra are first calculated using each DOA method; then the peaks in the spatial spectra are detected assuming that N_s is known *a priori*. Directions of the detected peaks are compared to the actual ones. Then, the detection is judged as either successful or failed according to that comparison. Many sets of source directions are generated randomly to avoid the bias introduced by a specific source direction. Then, each set is used to simulate the sound field for calculating spatial spectra. Thereafter, the number of sources N_s is varied and the detection rates are obtained for each DOA method and in each N_s.

The SDR measure is sensitive to criteria that determine the detection as successful or failed. Therefore, we shall later introduce "permissible error," which defines a permissible margin of the angular distance between the true source direction and the peak direction. To evaluate the performance, the experimental results of this study are shown as a function of the permissible error.

2) Accuracy of DOA: The SDR measure has estimation accuracy because SDR is shown as a function of the permissible error. That accuracy is higher when a higher SDR is achieved with smaller permissible error. Nevertheless, to exhibit the accuracy of DOA directly, we calculate the root mean squared error of the angular distance between the true source direction and detected peak direction.

B. Simulation Conditions

For this simulation, we assume a free sound field in which each source sound arrives at the microphones as a plane wave. Therefore, the time delay and amplitude response of each source sound at each microphone are mentioned to calculate the microphone signals. We assume that the amplitude response depends on the arrival angle of sound, as determined solely by microphone directivity. It is noteworthy that, even in the case in which noise is assumed as uncorrelated between channels, source signals function as correlated noise in multiple source conditions.

Evaluation using acquired data in real conditions or evaluation using simulated or measured impulse responses is desirable because reflection and reverberation are important considerations for practical use. However, measurement of microphone directivity is necessary in these cases for accurate evaluation because the proposed method of DOA estimation assumes that microphone directivity is known. Such investigation is beyond the scope of this paper because investigation in a reverberant condition requires additional intensive treatment.

C. Speech and Noise Data

1) Speech Data: For the preliminary investigation in V-A and V-B, we recorded speech samples of three sentences that were uttered by one male and one female. Moreover, we use speech samples of another 50 sentences from a Japanese speech corpus [10] uttered by one male for evaluation of SDR, as described in V-C, and for DOA accuracy in V-D. The male utterer of the former three sentences and that in the speech database are different. We denote the data set of the former six speech samples of the three sentences as "Data (A)," and the remaining 50 samples as "Data (B)."

TABLE I
CHOICE OF SPEECH SAMPLES FOR PRELIMINARY INVESTIGATIONS

Sentence	Data ID (utter)	Number of sources				
		1	2	3	4	5
a	a-1 (female)	o	o	o	o	o
	a-2 (male)	-	o	o	o	o
b	b-1 (female)	-	-	o	o	o
	b-2 (male)	-	-	-	o	o
c	c-1 (female)	-	-	-	-	o

For the preliminary investigation, we choose some samples in Data (A) depending on the number of sound sources assumed in the simulation condition, as described in Table I. The number of sources is from one to five; the mark "o" in Table I indicates that the speech sample is used for calculation. On the other hand, we take speech samples sequentially from 50 samples in Data (B); the samples are rotated after the last sample is taken. The beginning of all utterances is adjusted to coincide approximately in time.

2) Noise Data: This study uses background noise for evaluation. It had been recorded previously in a computer room that had computer fan noise.

D. Methods for Comparison

For comparison, we use three methods based on the cross correlation function that are often used for DOA or time delay estimation. We also use two popular high-resolution methods: MUSIC and MV. These two high-resolution methods are generally applicable to conditions in which the number of microphones is larger than that of sources ($M > N_s$). We are interested in cases where this condition does not hold, i.e., $M \leq N_s$. We consider that these methods are applicable to such conditions in cases where the source signals are broadband and nonstationary for the reason stated in Section I.

Regarding the cross-correlation-based method, we mention the ordinary cross-correlation function and the two types of generalized cross-correlation functions (GCCs) that have different weighting functions. We respectively designate these three methods as "OCC," "GCC-1," and "GCC-2." We assign OCC as the method using the ordinary cross spectrum, which is equivalent to the ordinary cross correlation function. The time averaged spatial spectrum of OCC is calculated as

$$S_{\text{OCC}}(d) = \overline{\sum_k \mathrm{Re}[G'_{xy,n,k}(d)]}. \tag{41}$$

We assign GCC-1 to an equivalent to GCC with the amplitude spectrum whitened by the inverse of the cross spectrum $G_{xy,n,k}$. The GCC-1 method has been referred to as "PHAT" in the literature [5]. We calculate the time averaged spatial spectrum of this method as the following:

$$S_{\text{GCC-1}}(d) = \overline{\sum_k \mathrm{Re}[G'_{xy,n,k}(d)]/|G'_{xy,n,k}(d)|}. \tag{42}$$

Here, GCC-2 is assigned to an equivalent to GCC with the amplitude spectrum whitened by the weighting function $\Gamma_k(d)$ shown below, as

$$\Gamma_{n,k}(d) = \zeta_{n,k}^2 \Big/ \left[(1-\zeta_{n,k}^2)|G'_{xy,n,k}(d)|\right] \tag{43}$$
$$\zeta_{n,k}^2 = |G_{xy,n,k}|^2/(G_{xx,n,k}G_{yy,n,k}) \tag{44}$$

where $\zeta_{n,k}^2$ is the squared coherence function between the channels. The weighting function $\Gamma_{n,k}(d)$ is intended to minimize the uncorrelated noise power between channels [5]. We used the fact that $\zeta_{n,k}^2$ does not change depending on the steering direction d in (43). We calculate the time averaged spatial spectrum of this method as the following:

$$S_{\mathrm{GCC-2}}(d) = \overline{\sum_k \mathrm{Re}[G'_{xy,n,k}(d)]\Gamma_{n,k}(d)}. \tag{45}$$

For the two high-resolution methods, we average the spatial spectra both on the frequency axis and over time as

$$S_{\mathrm{MUSIC}}(d) = 10^{\overline{\sum_k \log\left(\frac{1}{a_k(d)^* e_{n,k} e_{n,k}^T a_k(d)}\right)}} \tag{46}$$
$$S_{\mathrm{MV}}(d) = \overline{\sum_k 1 \Big/ \left(a_k(d)^* R_{n,k}^{-1} a_k(d)\right)} \tag{47}$$

where $R_{n,k}$ denotes the covariance matrix of the input signal

$$\begin{aligned} R_{n,k} &= \begin{bmatrix} \overline{X_{n,k}X_{n,k}^*} & \overline{X_{n,k}Y_{n,k}^*} \\ \overline{Y_{n,k}X_{n,k}^*} & \overline{Y_{n,k}Y_{n,k}^*} \end{bmatrix} \\ &= \begin{bmatrix} G_{xx,n,k} & G_{xy,n,k}^* \\ G_{xy,n,k} & G_{yy,n,k} \end{bmatrix} \end{aligned} \tag{48}$$

and $e_{n,k}$ denotes the eigenvector corresponding to the smaller eigenvalue of the eigendecomposition of $R_{n,k}$. The mode vector $a_k(d)$ is expressed as

$$a_k(d) = \{D_{x,k}(d), D_{y,k}(d)\exp(-j2\pi f_s \tau_d k/K)\}^T \tag{49}$$

where τ_d is the inter-channel time delay of a signal arriving from direction d, f_s is the sampling frequency, and K is the FFT point length. Diagonal elements of the covariance matrix $R_{n,k}$ are increased by 0.1% of $(G_{xx,n,k} + G_{yy,n,k})/2$ to stabilize the estimation of its inverse matrix and eigendecomposition numerically. They are respectively required for MV and MUSIC.

In calculating the MUSIC spectrum, the eigenvector that represents the noise space is generally required. If we were able to determine the spectral components that originated from one source signal, we could then use only those components and thereby obtain the eigenvector as that which corresponds to the smaller eigenvalue. Unfortunately, it is difficult to select those components in a multiple source condition. For that reason, we use the eigenvector that corresponds to the smaller eigenvalue for all components as an approximation. Spectral components that are contributed by multiple sources provide no correct peaks of sources, but they do provide gradual spatial spectra in most cases. These spatial spectra only raise the background level in their averaged spectrum. Therefore, correct peaks that are provided by the spectral components that originate from one source are maintained. Consequently, DOA estimation of multiple sources is possible using MUSIC.

For averaging the MUSIC spectrum, we average the log spectrum both in time and in frequency, as expressed in (46). Other methods of averaging, e.g., taking a logarithm after linear averaging, are possible. However, averaging by (46) showed a stable and superior result in preliminary experiments compared to those of other averaging methods. Averaging in the frequency axis is possible in the step of estimating the covariance matrix, as adopted in the method using a focusing matrix (e.g., [11]). However, averaging of the multiple frequency bin increases the chance for contribution by signals from multiple sources to the covariance matrix. This contribution renders the 2 × 2 covariance matrix as having no noise space. Therefore, we use no averaging along the frequency axis in the covariance matrix estimation.

Spatial spectra are typically displayed in decibels in the literature, but this study uses a linear scale to display them because the dynamic range of the spatial spectrum is sufficiently small to use a log scale in the case of $M \leq N_s$, which is not the proper condition for MUSIC and MV.

V. Evaluation

A. Directivity of WWG

Preliminary to performance comparison, we investigated the directivity of WWG to ensure the suitability of parameters (β, γ) of the two weighting functions and the relation between the microphone arrangement and the angular resolution. We also calculated the directivity as a system response that corresponds to the arrival direction of a signal where only one source is assumed to be present. The system response is WWG here. For calculation of directivity, we varied both the azimuth and elevation angle of the one assumed source from $-90°$ to $90°$ with a $2°$ step. Speech signals in Data (A) were concatenated and used as the source signal. The signal duration was about 15 s; WWG was averaged over that duration.

First, we present results corresponding to different values of parameters β and γ, as shown in Fig. 4. Here, the microphone direction is set to $d_L = (60°, 60°), d_R = (-60°, -60°)$. Calculated results are shown as a bird's-eye view diagram. The look direction $d = (0°, 0°)$ is at the center of each figure. Figs. 4(A)–(F), respectively, correspond to results when $(\beta, \gamma) = (0, 0), (0, 1.5), (0, 3), (0.4, 0), (0.75, 0)$, and $(0.75, 3)$. The results show that the directivity of WWG provides low angular resolution when no weighting functions are enabled ($(\beta, \gamma) = (0, 0)$), as shown in Fig. 4(A). Because WWG with $\beta = \gamma = 0$ is equivalent to the ordinary cross spectrum, we can infer that the ordinary cross correlation function also has insufficient resolution to detect multiple sources in this experimental setup. Different from that case, Fig. 4(B) and (C) show that a discontinuity between the region of the mainlobe around the look direction and the region surrounding the mainlobe emerges when the 2chSS weighting function is enabled ($\gamma > 0$). The level of the surrounding region is sufficiently low and the mainlobe width decreases as γ increases. On the other hand, the mainlobe is conical when the whitening function is enabled ($\beta > 0$) and the level of the surrounding region is also raised, as shown in Fig. 4(D) and (E). Then, Fig. 4(F), which is obtained when both weighting functions are

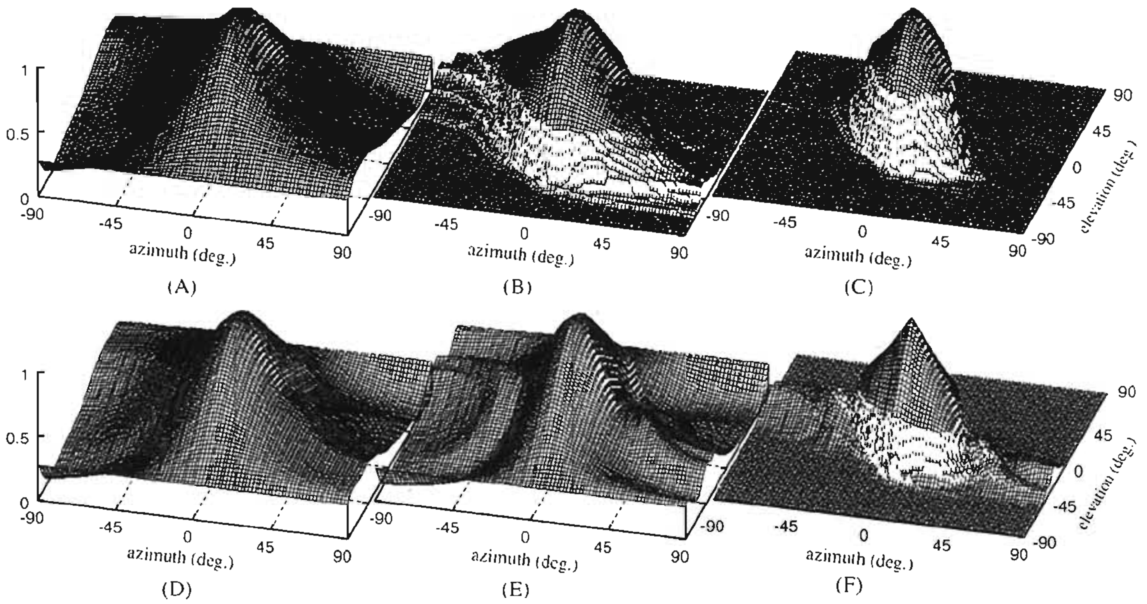

Fig. 4. Directivity of WWG versus parameters β, γ. ($\theta_o = \varphi_o = 60°$). (A) $(\beta, \gamma) = (0, 0)$, (B) $(0, 1.5)$, (C) $(0, 3)$, (D) $(0.4, 0)$, (E) $(0.75, 0)$, (F) $(0.75, 3)$.

enabled ($(\beta, \gamma) = (0.75, 3)$), shows the directivity of the sharp mainlobe with a low surrounding region. These results confirm that 2-D angular resolution can be realized by WWG with a circular symmetric two-channel microphone and that both β and γ serve to increase the angular resolution, at least in the case where a single source exists.

Next, we investigated the relation between the microphone arrangement and the angular resolution. For simplicity, we set the parameters of the microphone direction as $\theta_o = \varphi_o$: $d_L = (\theta_o, \theta_o)$ and $d_R = (-\theta_o, -\theta_o)$. Calculated results using different values of θ_o are shown in Fig. 5. $(\beta, \gamma) = (0.75, 3)$ was used here. Fig. 5(A)–(D), respectively, correspond to results when $\theta_o = 0°, 30°, 60°$, and $90°$. As the figure illustrates, the mainlobe is broader along the elevation angle when $\theta_o = 0°$ and $30°$; it narrows as the tilt angle θ_o increases. The resolution in elevation angle seems to be at a maximum when $\theta_o = 90°$, but we consider that $\theta_o = 60°$ is better for practical use because of the fewer bad effects that are attributable to the signal from the inverse direction $(-180°, -180°)$.

We assumed a single source condition to yield the results described above. Therefore, we cannot directly confirm from them that the spatial spectrum in a multiple source condition also has peaks whose shapes exactly reflect the above directivity. Nevertheless, we infer that the resolution can be higher as the directivity becomes sharper and that the opposite cannot be true. For that reason, we use $\theta_o = 60°$ with $(\beta, \gamma) = (0.75, 3)$ for subsequent evaluation.

B. Spatial Spectra

We calculate the spatial spectra of the three methods WWG, MUSIC, and MV by changing the number of sources N_s to emphasize the difference of the peak shapes depending on the methods. We omit displaying those of OCC, GCC-1, and GCC-2 because they exhibit no clear peak corresponding to the sound sources when multiple sources exist. Speech signals in Data (A) were used as the source signal and the first one-second of the utterances were used. Calculations are performed for azimuths of $-90° \leq \theta \leq 90°$ with 2° step and elevation angles $-90° \leq \varphi \leq 90°$ with 2° step.

Fig. 6(1)–(3) respectively depict the results obtained in cases where $N_s = 1, 2$, and 3. The source arrangement is shown in Table II. As shown in Fig. 6(1), all spectra obtained using these three methods exhibit one clear peak at the true source direction in the case of $N_s = 1$. The peak of WWG is broader than that obtained using the other methods. Therefore, the angular resolution of WWG seems to be lower than those of the other methods. In the case of $N_s = 2$, we observe that the MUSIC spectrum has only one peak, which seems to result from the fusion of the two peaks of sources 1 and 2, as shown in Fig. 6(2). Two peaks corresponding to the two sources emerged in the MV spectrum, but the peak height of source 1 is much smaller than that of source 2. Consequently, detection of the two sources appears to be difficult. In contrast, the WWG spectrum exhibits two clear peaks in the true source directions.

Directions of source 1 and 2 provide the same time difference between channels. For that reason, only the amplitude difference is valid to distinguish the two sources. We can observe that resolution of the elevation angle in MUSIC and MV is lower than that of WWG. In the case of $N_s = 3$, a similar result to the case of $N_s = 2$ is obtained. Three clear peaks are visible in the WWG spectrum, as depicted in Fig. 6(3), whereas only two peaks are visible in MUSIC and MV spectra.

The MV spectrum peak appears sharper than that of MUSIC when $N_s = 1$ because of the few averaging iterations of the spectra to estimate the covariance matrix (three times). The peak sharpness in the MUSIC spectrum increases as the averaging times of the spectra increase when $N_s = 1$. Nevertheless, the peaks broaden as the averaging iterations increase when $N_s > 1$ because of the contribution of multiple sources to the covariance matrix. The use of three averaging iterations is inferred to be

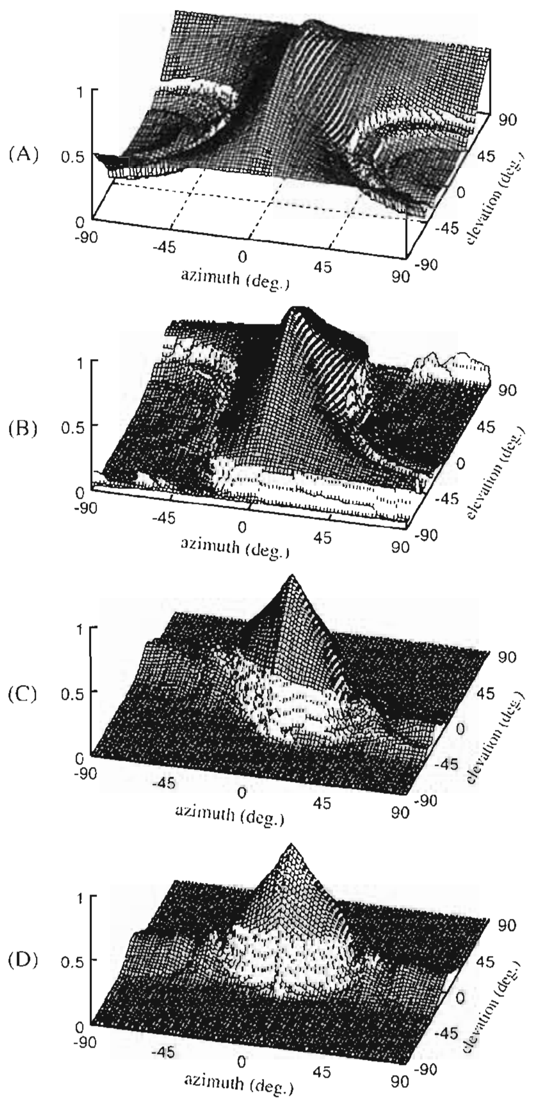

Fig. 5. Directivity of WWG versus microphone tilt angles. ($(\beta, \gamma) = (0.75, 3)$). (A) $(\theta_0, \varphi_0) = (0°, 0°)$, (B) $(\theta_0, \varphi_0) = (30°, 30°)$, (C) $(\theta_0, \varphi_0) = (60°, 60°)$, (D) $(\theta_0, \varphi_0) = (90°, 90°)$.

best for MUSIC in this experimental setting. Similar results are obtained for MV.

As shown in Fig. 6(2), the WWG spectrum obtained in cases with multiple sources tends to exhibit bothersome ripple-like peaks. These peaks can increase false detection of sources, thereby rendering an inaccurate estimation of the number of sources. Estimation of the number of sources is an important problem. However, no good method exists to solve that problem when $N_s \geq M$. Investigation of this problem is continuing, but estimation of the number of sources is beyond the scope of this paper.

C. Source Detection Rate

Next, the SDR described in Section IV-A1 is examined to evaluate DOA estimation performance quantitatively. In addition to the two-dimensional (azimuth and elevation) case, a one-dimensional case (azimuth only) is examined here. The number of sources N_s is set from one to five; 1000 sets of source directions are used for each N_s. Speech data set Data (B) is used for the source signals and the required number of samples is taken sequentially from the data set. The samples are rotated after the last sample is taken.

Spatial spectra used for DOA detection are calculated in the specified angular range corresponding to a one-dimensional case or two-dimensional case. For detecting the DOAs, we assume that the number of sources N_s is known, as stated in Section IV-A1 and that the highest N_s peaks are selected from the spectra. The detected N_s peak directions are compared to the true N_s source directions. The detected peak that is nearest to each true direction is marked as the corresponding detected peak of the true source. We use the permissible error to validate the correspondence. Correspondence is regarded as valid when the difference between the true source direction and the corresponding detected peak is smaller than the permissible error. Then, we regard the detection as being successful when all true sources have valid corresponding peaks in a one-to-one relation. We regard it as being failed if one or more true sources have no valid corresponding peak. The resultant SDR is obtained as a function of the permissible error.

1) One-Dimensional Case (SDR in Azimuth Only): First, we conducted an SDR evaluation for the azimuth-only case. Directions of each source set are generated randomly to distribute them uniformly in the range of $-85° \leq \theta \leq 85°$. The elevation angles of all sources are set as $\varphi = 0°$, where each angular distance between sources is restricted to be larger than 10°. Spatial spectra used for DOA detection are calculated in the angular range of $-90° \leq \theta \leq 90°$ with 1° step at $\varphi = 0°$.

The obtained result is shown in Fig. 7. That figure illustrates a case in which SNR = 10 dB. Regarding the WWG performance, we calculated the SDR with $(\beta, \gamma) = (0.75, 3)$. This figure comprises five panels that correspond to the number of sound sources N_s.

As shown in Fig. 7, an SDR of almost 100% is attained when $N_s = 1$ by GCC-1, GCC-2, and WWG at the permissible error = 10°. However, the SDRs of MV and MUSIC when $N_s = 1$ are degraded by about 95% because MV and MUSIC spatial spectra exhibit broad peaks in cases where the true source is located far from the front direction, e.g., at greater than 80° or at less than −80°. The peak maximum does not exist within the scanning range in those cases. For the increase of N_s, the performance of all methods degrades. When $N_s = 2$, the degradation of the three cross-correlation-based methods is large and that of OCC is particularly large. On the other hand, performance degradations of MV and MUSIC at the permissible error = 10° are about 22%, and that of WWG is less than 4% in the case of $N_s = 2$. When $N_s \geq 3$, the performance difference between WWG and the other methods becomes larger and WWG remains to exhibit the highest performance. When $N_s = 3$, the respective SDR of WWG, MV and MUSIC at the permissible error = 10° are 84%, 46% and 45%. The performance superiority of WWG is more remarkable when $N_s \geq 4$. These results indicate that WWG attains much better performance than the other methods, despite the increase of sound sources.

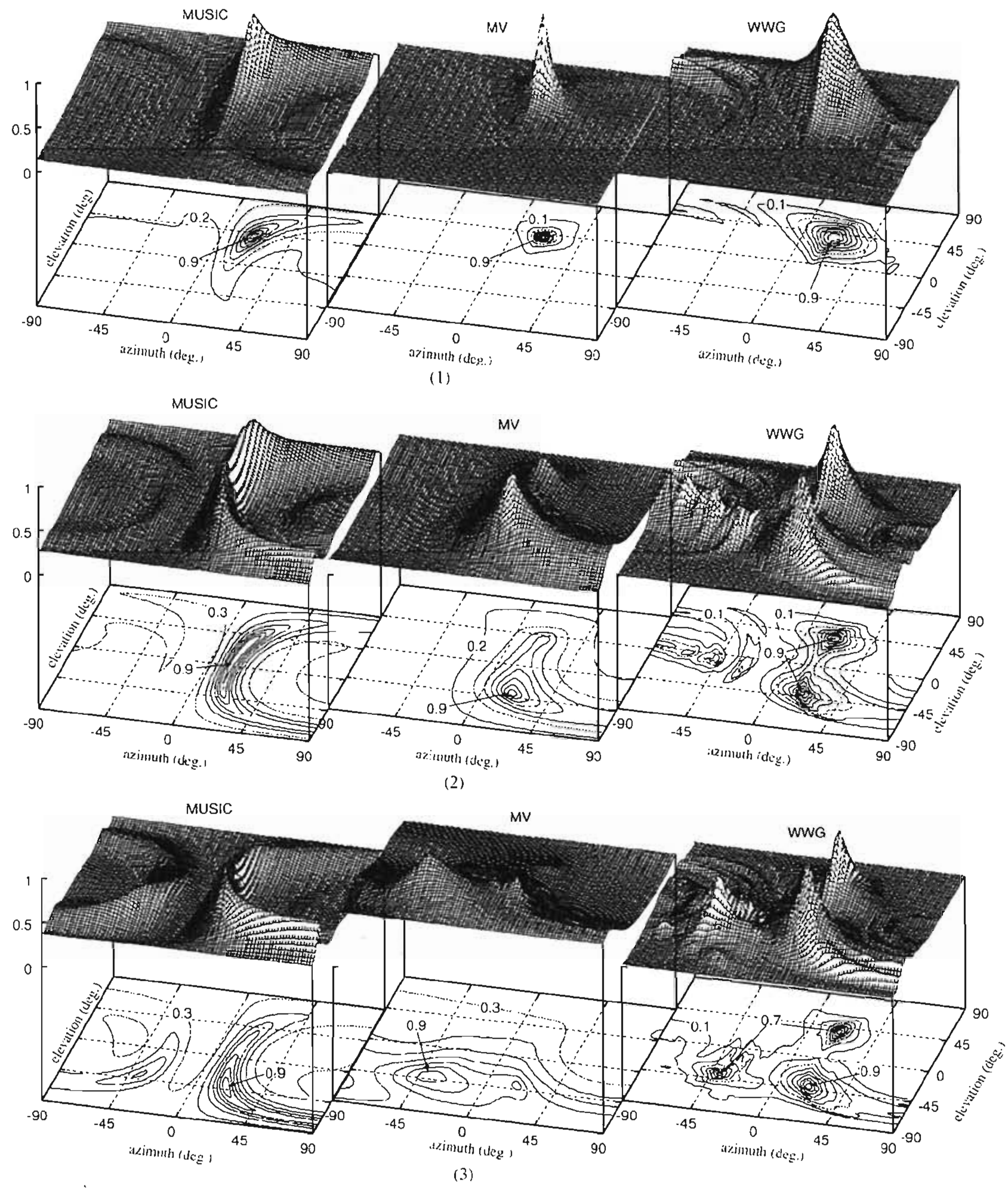

Fig. 6. Spatial spectra obtained using the three methods. (1) One sound source. (2) Two sound sources. (3) Three sound sources.

TABLE II
DIRECTION OF SOUND SOURCES

N_s	Direction (azimuth, elevation)		
	Source 1 (a-1)	Source 2 (a-2)	Source 3 (b-1)
1	(20°,40°)	–	–
2	(20°,40°)	(20°,-40°)	–
3	(20°,40°)	(20°,-40°)	(-40°,-40°)

2) Two-Dimensional Case: Next, the SDR in the 2-D case is examined. The angular range of the source directions that were randomly generated is set as $-85° \leq \theta \leq 85°$ and $-85° \leq \varphi \leq 85°$, where every angular distance between sources is restricted to be larger than 10°. Spatial spectra used for DOA detection are calculated in the angular range of $-90° \leq \theta \leq 90°$ and $-90° \leq \varphi \leq 90°$ with 2° step. The result at SNR = 10 dB is shown in Fig. 8. As in the one-dimensional case, $(\beta, \gamma) = (0.75, 3)$ is used to calculate WWG spectra.

As shown in this figure, SDRs larger than 93% are attained by MUSIC, MV, and WWG when $N_s = 1$ at the permissible error= 10°. However, the SDRs of the three cross-correlation-based methods are degraded and that of GCC-1 is particularly low even in the case where $N_s = 1$. The SDR degradation of GCC-1 arises from the fact that the amplitude difference between the two microphones is omitted through normalization

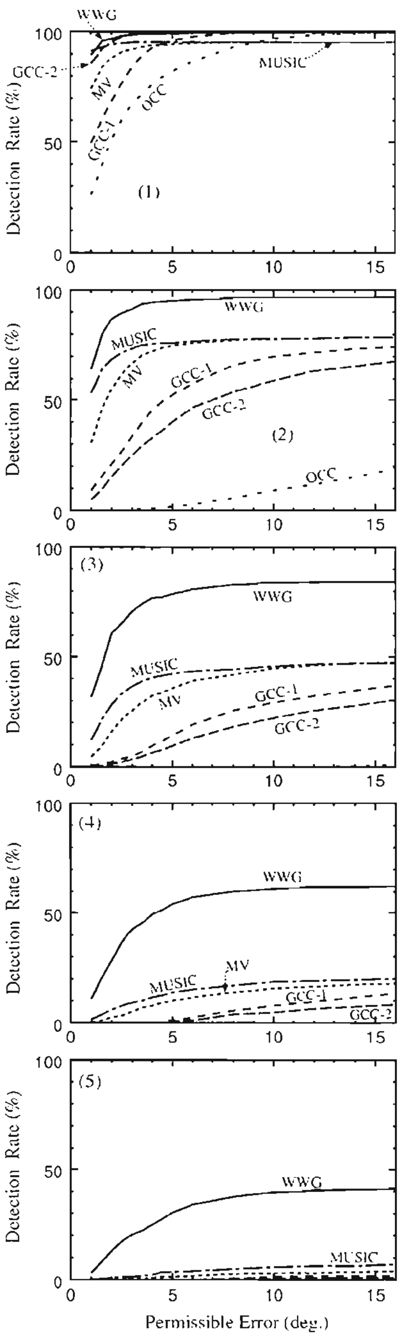

Fig. 7. Source Detection Rate (SDR) versus permissible error in the one-dimensional case for SNR = 10 dB. (1) One source. (2) Two sources. (3) Three sources. (4) Four sources. (5) Five sources.

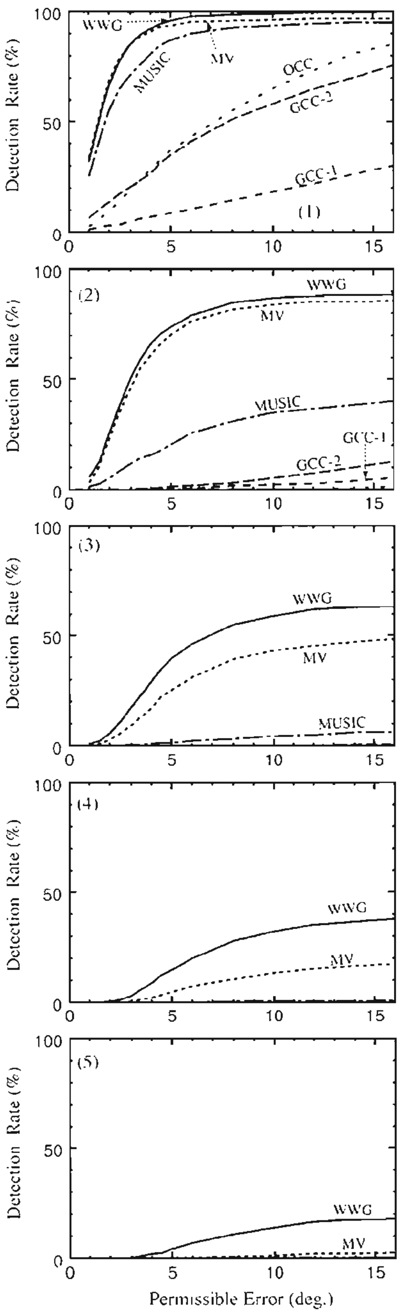

Fig. 8. Source detection rate (SDR) versus permissible error in the two-dimensional case at SNR = 10 dB. (1) One source. (2) Two sources. (3) Three sources. (4) Four sources. (5) Five sources.

by $|G'_{xy,n,k}(d)|$, as shown in (42). Resolution in the elevation angle is lost by this normalization. For the increase of N_s, the performance of all methods degrades as N_s increases. When $N_s = 2$, the SDR of MUSIC is decreased to about 35% when the permissible error= 10°. Those of OCC, GCC-1, and GCC-2 become less than 5%. Nevertheless, the performance degradations of WWG and MV in the case of $N_s = 2$ are small. When $N_s \geq 3$, the SDR of MUSIC decreases rapidly as N_s increases, in contrast to the result of WWG. The result of MV is intermediate between MUSIC and WWG. When $N_s = 3$, the respective SDR of WWG, MV, and MUSIC at the permissible error = 10° are 60%, 43%, and 4%.

Similarly to the one-dimensional case described in the previous Section V-C1, the results described above for the two-dimensional case again indicate that the proposed method is much more robust to the increase of sound sources than are other methods with which it was compared here.

D. DOA Accuracy

Finally, to examine accuracy of the peak position, we calculate the root mean squared error (RMSE) of the angular distance between the peak direction and its true direction. The same spatial spectra obtained in the previous Section V-C2 of SDR measurement with the two-dimensional case are used for this evaluation. Only the pairs of the true peak direction and its corresponding peaks that resulted from successful detection are taken into account. The calculated results are shown as a function of the permissible error, as was done previously for the SDR evaluation.

The obtained results at SNR = 10 dB are shown in Figs. 9 and 10. These figures, respectively, illustrate the RMSE versus the permissible error in azimuth and in the elevation angle. We display the result only when the number of successful detections is greater than 5% (50 samples) of the number of all trials. Thereby, we avoid inclusion of less reliable data of the small samples used for calculating the RMSE.

Regarding the estimation of the azimuth's angle, Fig. 9 shows that the RMSEs of GCC-1 and GCC-2 are lower than those obtained by the other methods when $N_s = 1$ at the permissible error $< 12^\circ$ [Fig. 9(1)]. These values increase mostly in proportion to the permissible error. In contrast, the RMSEs of WWG, MV and MUSIC are greater than those of GCC-1 and GCC-2 when the permissible error $< 12^\circ$, but they remain lower than 4° even when the permissive error $\geq 12^\circ$.

In the case of $N_s = 2$, as shown in Fig. 9(2), the results of OCC and GCC-1 are not displayed because OCC and GCC-1 had lower SDR than 5% for all permissible errors and GCC-2 did so when the permissible error $< 10^\circ$. In this case, we observe that the performance of MUSIC and GCC-2 degrades, whereas those of WWG and MV remain almost unchanged compared to the case of $N_s = 1$. When $N_s = 3$ and 4, as shown in Fig. 9(3) and (4), only WWG and MV have sufficient successive detections for display. We observe that RMSE of WWG remains lower than that of MV and is mostly unchanged through increase of N_s.

In the case of estimating the elevation angle, we can observe a similar tendency to those of results of azimuth estimation, as shown in Fig. 10. Both WWG and MV have similar RMSEs that are lower than those of azimuth estimation. In consideration of the results of SDR measurement and RMSE measurement, we can confirm that WWG achieves the highest SDR and accuracy simultaneously among the methods that were compared in multiple source environments.

E. Angular Resolution

In this analysis, we investigate the angular resolution of WWG. We assume that the two sources that have equal power are present near the direction of (0°, 0°); we further assume that the source signals are taken from data set Data (A). The SNR was set to 20 dB. We calculate the spatial spectra in the case in which one (source-a) of the two sources is fixed at $d_a = (0^\circ, 0^\circ)$ and the other source (source-b) changes its direction near (0°, 0°) with a 2° step to find the minimum angular resolution. We denote the direction of source-b as $d_b = (\theta_b, \varphi_b)$.

First, we show results obtained when the direction of source-b is changed along the azimuthal axis while its elevation angle is fixed as 0°. The spatial spectrum of GCC-2, WWG, MUSIC, and MV along the azimuthal axis (elevation = 0°) is shown in Fig. 11. Four panels at the left side of Fig. 11 correspond to this case. The displayed spatial spectra are shifted along the vertical axis as the azimuth of source-b shifts. The arrows indicate the source directions. As shown in that figure, the WWG spectra peaks corresponding to the two sources are distinguishable when the angular distance between the sources is greater than 4°. Therefore, we infer that the minimum angular resolution of WWG along the azimuthal axis is about 3°. Similarly, that of MV is inferred from the measurement to be about 5° in this setup.

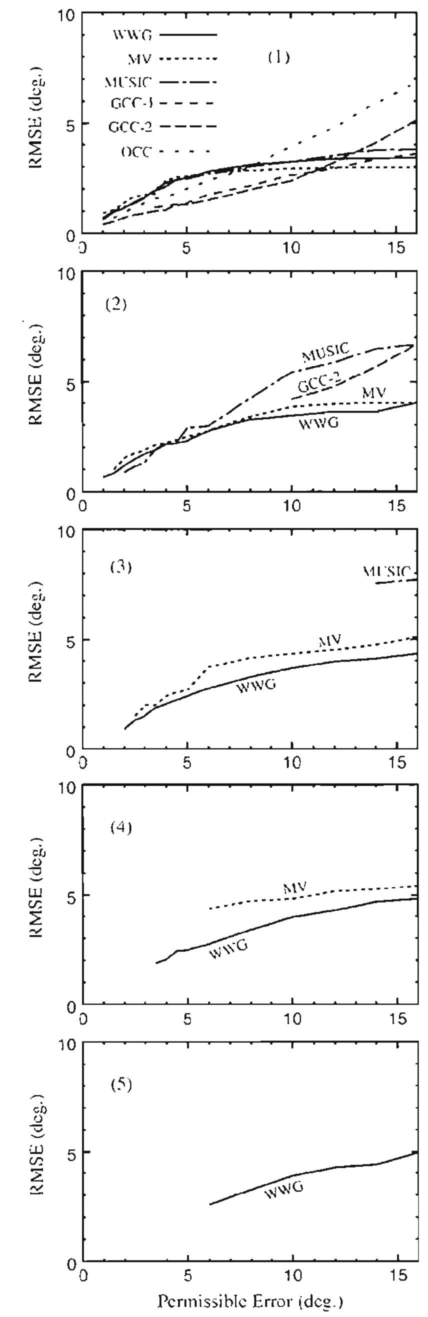

Fig. 9. RMSE in azimuth angle versus permissible error. (1) One source. (2) Two sources. (3) Three sources. (4) Four sources. (5) Five sources.

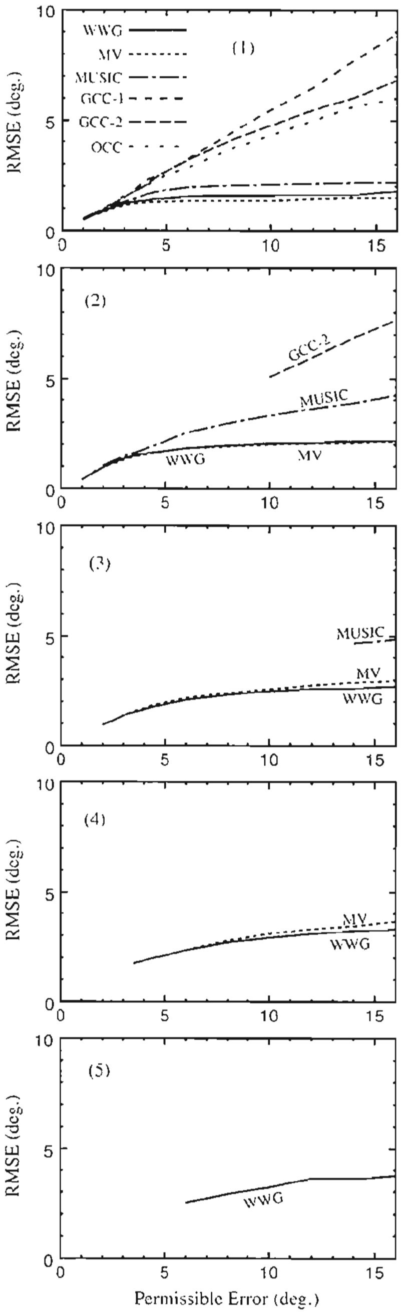

Fig. 10. RMSE in elevation angle versus permissible error. (1) One source. (2) Two sources. (3) Three sources. (4) Four sources. (5) Five sources.

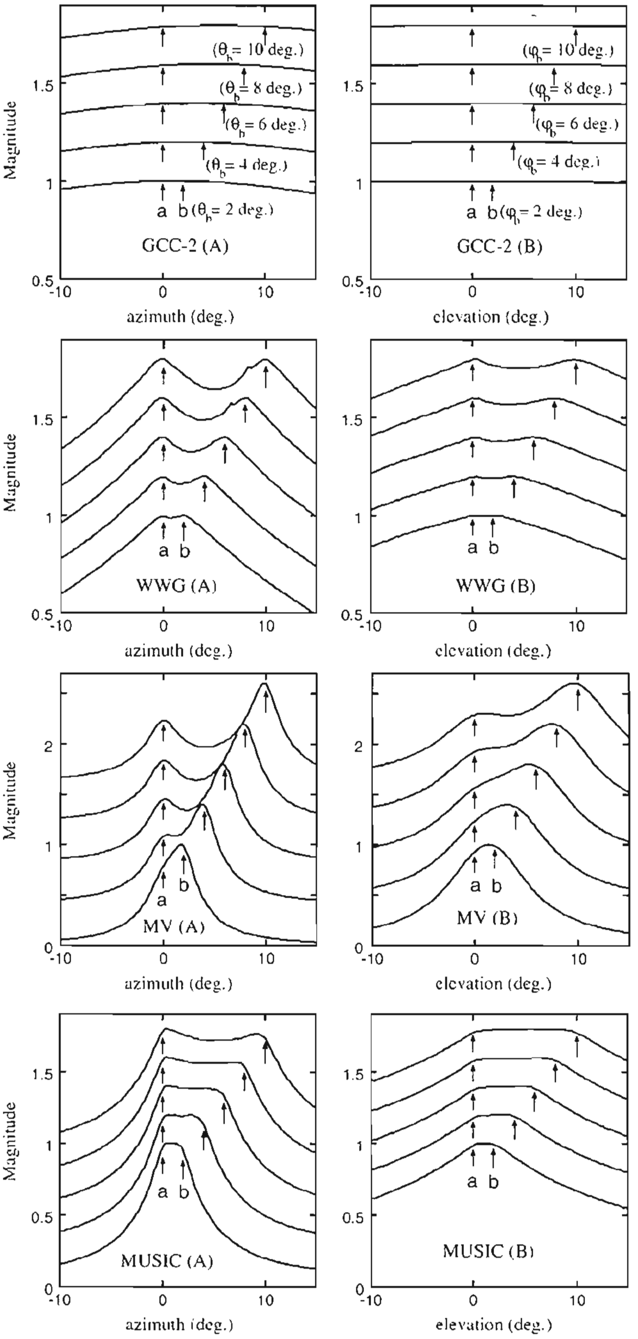

Fig. 11. Angular resolution in azimuth and elevation angle.

Next, we show the results obtained when the direction of source-b changes along the elevation axis, while its azimuth is fixed at 0°. The spatial spectrum along the elevation axis (azimuth $= 0^\circ$) is shown in the four panels at the right side of Fig. 11. As shown in that figure, the minimum angular resolution of WWG along the elevation axis is about 8°; that of MV is about 10°. We observe that GCC-2 has no resolution in the range of this examination and that the resolution of MUSIC is also low, particularly in the elevation angle.

We note that bias error is observed in MV and MUSIC spectra, whereas that in WWG spectra appears to be very small. Moreover, we observe that MV spectra exhibit peaks of the two sources with different height, although the two sources have equal power. This phenomenon coincides to the 2-D spatial spectra of MV, as shown in Fig. 6(2) and Fig. 6(3); it seems to degrade the MV performance.

The results described above verify that the minimum angular resolution of the azimuthal axis in this setup is higher than that of the elevation axis for all methods and that WWG has the highest resolution in both axes.

VI. Conclusion

This study proposed a new method of DOA estimation using two directional microphones based on the weighted Wiener gain. A circular symmetry arrangement using two directional microphones was also proposed to enable estimation not only of azimuth, but of elevation angle. We compared the performance in terms of source detection rates and detection accuracy with the three equivalents to the methods based on the cross-correlation function and the popular high-resolution methods of MUSIC and MV. Results show that, although the respective performances of both the cross-correlation based methods and the high-resolution methods degrade remarkably with more than two sources, the performance degradation of the proposed method is moderate and a detection rate of 84% in the azimuth-only case and 60% in the two-dimensional case were attained at the permissible error $= 10^\circ$, even in cases where actual background noise with SNR $= 10$ dB and three sources are present. These results demonstrate the superiority of the proposed method over other methods, particularly in adverse conditions where sound sources are more numerous than microphones. Although the number of detectable sources is increased by the proposed method, determination of the number of sources remains as an important problem. We are continuing our examination of this problem.

References

[1] S. U. Pillai, Ed., *Array Signal Processing*. New York: Springer-Verlag, 1989.

[2] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.

[3] J. Capon, "High-resolution frequency-wavenumber specirum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[4] H. Kamiyanagida, H. Saruwatari, K. Takeda, F. Itakura, and K. Shikano, "Direction of arrival estimation using nonlinear microphone array," *IEICE Trans. Fund.*, vol. E84-A, pp. 999–1009, Apr. 2001.

[5] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 8, pp. 320–327, Aug. 1976.

[6] Y. Nagata, T. Fujioka, and M. Abe, "Speech enhancement based on auto gain control," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 177–190, Jan. 2006.

[7] H. Y. Kim, F. Asano, Y. Suzuki, and T. Sone, "Speech enhancement based on short-time spectral amplitude estimation with two-channel beamformer," *IEICE Trans. Fund.*, vol. E79-A, pp. 2151–2158, Dec. 1996.

[8] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 113–120, Apr. 1979.

[9] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagat.*, vol. AP-30, no. 1, pp. 27–34, Jan. 1982.

[10] S. Itabashi, Ed., *Continuous Speech Corpus for Research Vol. 1–7*. Tokyo, Japan: AI and Fuzzy Promotion Center. Japan Information Processing Development Corp., 1991.

[11] J. W. Choi and Y. H. Kim, "Estimation of locations and strengths of broadband planar and spherical noise sources using coherent signal subspace," *J. Acoust. Soc. Amer.*, vol. 98, pp. 2082–2093, Oct. 1995.

Yoshifumi Nagata received the B.E. degree in electronics in 1984 and the M.E. and Dr.Eng. degrees in information science in 1987 and 1990, respectively, all from Tohoku University, Sendai, Japan.

In 1990, he joined the Research and Development Center, Toshiba Corporation, where he was engaged in research and development of speech processing systems. Since 1997, he has been an Associate Professor at Iwate University, Morioka, Japan. His interests include multimedia human interface and speech signal processing.

Dr. Nagata is a member of the Acoustical Society of Japan and the Information Processing Society of Japan.

Toyota Fujioka was born in Akita, Japan, on August 21, 1969. He received the B.E. and M.E. degrees in electrical and electronic engineering from the Mining Collage, Akita University, Akita, Japan, in 1992 and 1994, respectively, and the Ph.D. degree in electrical and communication engineering from Tohoku University, Sendai, Japan, in 1997.

He is currently a Research Associate in the Department of Computer and Information Science Faculty of Engineering, Iwate University, Morioka, Japan. His research interests include parallel computer and data compression.

Dr. Fujioka is a member of the Information Processing Society of Japan.

Masato Abe (M'85) received the B.E., M.E., and Ph.D. degrees in electrical engineering from Tohoku University, Sendai, Japan, in 1976, 1978, and 1981, respectively.

From 1981 to 1989, he was a Research Associate with the Research Center for Applied Information Sciences, Tohoku University. From 1989 to 1996, he was an Associate Professor in the Department of Information Science, Iwate University, Morioka, Japan. His research interests include digital signal processing for acoustics and computer architecture.

Dr. Abe is a member of the Acoustical Society of America, the Acoustical Society of Japan, the Institute of Noise Control Engineering of Japan, the Information Processing Society of Japan, the Institute of Electronics, Information and Communication Engineers, the Association for Computing Machinery, and the Japan Society of Mechanical Engineers.