Binaural Localization Based on Weighted Wiener Gain Improved by Incremental Source Attenuation

Yoshifumi Nagata, Satoshi Iwasaki, Takahiko Hariyama, Toyota Fujioka, Tomita Obara, Takayuki Wakatake, and Masato Abe, *Member, IEEE*

Abstract—**This paper addresses the problem of direction-of-arrival (DOA) estimation both in azimuthal and elevation angle from binaural sound that is processed with a head-related transfer function (HRTF). Previously, we proposed a weighted Wiener gain (WWG) method for two-dimensional DOA estimation with two-directional microphones. However, for signals processed with HRTFs, peaks in the spatial spectra of WWG indicating true sources can mingle with spurious peaks. To resolve this situation, we propose to apply incremental source attenuation (ISA) in combination with WWG. In fact, ISA reduces spectral components originating from specified sound sources and thereby improves the localization accuracy of the next targeted source in the proposed incremental estimation procedure. We conduct computer simulations using directional microphones and four HRTF sets corresponding to four individuals. The proposed method is compared to two DOA estimation methods that are equivalent to two generalized cross-correlation functions and two high-resolution methods of multiple signal classification (MUSIC) and minimum variance method. For comparison purposes, we introduce binary coherence detection (BCD) to high-resolution methods for emphasizing valid spectral components for localization in multiple source conditions. Evaluation results demonstrate that, although MUSIC with BCD yield comparable performance to that of WWG in conditions where single speech source exists, WWG with ISA surpasses the other methods in conditions including two or three speech sources.**

Index Terms—**Binaural, coherence detection, direction-of-arrival (DOA) estimation, elevation, head-related transfer function (HRTF), incremental source attenuation, minimum variance (MV), multiple signal classification (MUSIC), Wiener gain.**

I. Introduction

DIRECTION-OF-ARRIVAL (DOA) estimation of concurrent multiple sound sources with two-channel input signals is a challenging task. Although many methods have been proposed, such as binaural models and a two-channel version of array signal processing (e.g., [1]–[3]), the two-channel technique remains attractive.

Manuscript received October 24, 2007; revised July 24, 2008. Current version published December 11, 2008. This work was supported by KAKENHI under Grant 19500164. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm Slaney.

Y. Nagata, T. Fujioka, T. Wakatake, and M. Abe are with the Department of Computer and Information Sciences, Iwate University, Morioka 020, Japan (e-mail: nagata@cis.iwate-u.ac.jp).

S. Iwasaki is with Aichi Medical University, Aichi, 480-1195, Japan.

T. Hariyama is with the Department of Biology, Hamamatsu University School of Medicine, Hamamatsu-shi, 431-3192, Japan .

T. Obara is with Panasonic Mobile Communications R&D Lab Company, Ltd., Sendai 981-3206, Japan.

Digital Object Identifier 10.1109/TASL.2008.2006651

Regarding binaural models, several methods are useful in multiple source environments (e.g., [4]–[8]). Among those, the localization possibility of up to five or six speech sources was reported in [6] and [7]. Liu *et al.* [6] used a coincidence model by Jeffress for localization, but the additional "stencil filter" reduced the phase ambiguities in high-frequency components to improve estimation accuracy. Faller *et al.* [7] proposed binary selection of spectral components for extracting binaural cues of interaural time difference (ITD) and interaural level difference (ILD) based on interaural coherence (IC). On the other hand, Braasch introduced subtraction on the cross-correlation function [8] to account for humans' ability to localize two concurrent broadband noises. This idea resembles spectral subtraction on the cross-spectrum [9], which has been developed for speech enhancement. These methods are evaluated mostly in azimuth-only estimation; simultaneous estimation of elevation and azimuth has been described only rarely. Two-dimensional (azimuth-elevation) estimation seems to be challenging and interesting in binaural localization.

Regarding engineering array signal processing, high-resolution methods such as multiple signal classification (MUSIC) [10] and minimum variance (MV) method [11] are popular for use in DOA estimation. For those methods, the number of detectable sources is usually one for two-channel systems. Nevertheless, the limitation is eased when sources radiate speech because of the sparseness of the time-frequency (TF) structure of the speech spectrum. Mohan *et al.* [12] proposed binary coherence detection to use the sparseness of the TF structure actively for selecting spectral components dominated by a single source. Mohan's coherence detection is equivalent to Faller's binary selection. Mohan *et al.* also described an alternative method using the ratio of the eigenvalues for coherence detection. Prior measured head related transfer functions (HRTFs) are useful as steering vectors that are necessary to yield a spatial spectrum over the search area of interest to apply these methods to two-channel observations filtered using HRTFs. Although two-dimensional estimation is considered possible using the HRTF-based steering vector, performance evaluations of such cases are inadequate.

In contrast to the binary selection of valid spectral components described above, we use a method of emphasizing such a component using weighting functions. For this approach, we have proposed introduction of two-channel spectral subtraction (2chSS) [13], followed by noise whitening, into speech enhancement and DOA estimation [14], [15]. In fact, 2chSS is the two-channel version of spectral subtraction [16], but it is applicable to nonstationary noise environments. Two

spectral weighting functions corresponding to 2chSS and noise whitening are combined with a cross-spectrum to yield an accurate estimator of the Wiener gain for the desired signal arriving from the look direction. We designate this estimator as the weighted Wiener gain (WWG). We reported two-dimensional localization exploiting two-directional microphones in [15].

However, our experience has shown that our method, similarly to high-resolution methods, deteriorates when the systems are applied to HRTF-filtered signals in multiple-source environments, mainly because the ILD of the HRTF is small in the frontal vertical plane and in its nearby directions. The small ILD provides low angular resolution in elevation around these directions. Furthermore, WWG-based methods suffer from spurious peaks arising from the complex directional-frequency structure of HRTFs. We propose incremental source attenuation (ISA) for application to WWG in this paper to address these problems. The ISA attenuates spectral components originating from sources located in specified directions. The attenuation is attained using a newly introduced weighting function. During the estimation process, the first-source direction is estimated using the original WWG without ISA. Next, the second source's direction is estimated using WWG from spectra with which the first source contribution is reduced by ISA. The only maximum peak on the spatial spectrum is regarded as the source peak in each estimation step. Third source estimation is facilitated by spectra that have a reduced contribution of the first and the second sources. Similarly, the fourth and greater sources' respective DOAs are estimated using WWG with ISA.

For evaluating the localization performance, we compare the proposed method to the equivalents of two generalized cross correlation functions (GCCs) and two high-resolution methods of MUSIC and MV. Taking into account the fact that the spectral components contributed by multiple sources do not provide proper source directions in these high-resolution methods with two-channel input, we introduce Mohan's binary selection of spectral components. The ratio of the eigenvalues of the input correlation matrix is used for this purpose. We carry out the evaluation with a measured HRTF to simulate cases in which HRTF-filtered observations are assumed. Such a measured HRTF is applicable to, for example, DOA estimation in a robot system, where the incident signals are altered by the robot's body, as reported in [17].

The remainder of this paper is organized as follows. Section II presents a summary of DOA estimation based on WWG. Section III describes the proposed method based on WWG with ISA. Section IV presents an explanation of the experimental setup for evaluation. Section V describes evaluation of the proposed method compared to MUSIC, MV, and two GCCs. Finally, Section VI summarizes the conclusions reached through this study.

II. DOA Estimation Based on WWG

We present a summary of DOA estimation based on WWG, considering HRTF filtered signals. Please see [15] for derivation of the original WWG.

A. Auto Gain Control With Weighted Wiener Gain

We assume that two microphones are placed in a noisy environment. The incident sound from direction d_s is transformed by an HRTF and is received at the microphones. We designate the L-channel and R-channel HRTFs, respectively, corresponding to direction d_s as $H_{x,k}(d_s)$ and $H_{y,k}(d_s)$, where k is the frequency bin number of the discrete Fourier transform (DFT). Let the DFT of the signals received at the microphones be

$$\begin{aligned} X_{O,n,k} &= S_{n,k}H_{x,k}(d_s) + N_{x,n,k} \\ Y_{O,n,k} &= S_{n,k}H_{y,k}(d_s) + N_{y,n,k}, \end{aligned} \tag{1}$$

where $X_{O,n,k}$ and $Y_{O,n,k}$, respectively, denote the DFT of the observed signals at the microphones for the frame n and the kth frequency bin; $S_{n,k}$ denotes the DFT of the source signal, and $N_{x,n,k}$ and $N_{y,n,k}$, respectively, denote those of the noises received at respective microphones.

When d_s is known and the microphone signals are compensated so that the source signal component in each channel is identical, the compensated DFT of each microphone signal is obtained as

$$\begin{aligned} X_{n,k}(d_s) &= S_{n,k} + N_{x,n,k}/H_{x,k}(d_s) \\ Y_{n,k}(d_s) &= S_{n,k} + N_{y,n,k}/H_{y,k}(d_s). \end{aligned} \tag{2}$$

Consider the case in which the average of the compensated signals $Z_{n,k}(d_s) = [X_{n,k}(d_s)+Y_{n,k}(d_s)]/2$ is multiplied using a scalar gain $\alpha_n(d_s)$ for approximating the source signal contained in $Z_{n,k}(d_s)$ as

$$\hat{S}_{n,k} = Z_{n,k}(d_s)\alpha_n(d_s). \tag{3}$$

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

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

where $(^-)$ denotes time-averaging. The weighted version of the Wiener gain is obtained as

$$\alpha_n(d_s) = \frac{\sum_k G_{ss,n,k}\Psi_{n,k}(d_s)}{\sum_k G_{zz,n,k}(d_s)\Psi_{n,k}(d_s)} \tag{5}$$

where $G_{ss,n,k} = \overline{|S_{n,k}|^2}$ and $G_{zz,n,k}(d_s) = \overline{|Z_{n,k}(d_s)|^2}$, respectively, signify the power spectra of the desired signal and the primary signal.

If the background noise is uncorrelated between channels, $G_{ss,n,k}$ in (5) can be replaced by the cross spectrum $G_{xy,n,k}(d_s) = \overline{X^*_{n,k}(d_s)Y_{n,k}(d_s)}$, in which $(^*)$ signifies operation of the complex conjugate. Thereby, we introduce a positive constant β to control the effect of the weighting function and rewrite (5) as

$$\alpha_n(d_s) = \frac{\sum_k \mathrm{Re}\left[G_{xy,n,k}(d_s)\right]\Psi^{\beta}_{n,k}(d_s)}{\sum_k G_{zz,n,k}(d_s)\Psi^{\beta}_{n,k}(d_s)}. \tag{6}$$

Parameters β were determined empirically. We assume that $\Psi_{n,k}(d_s)$ has a real value. The imaginary part of $G_{xy,n,k}(d_s)$ can be ignored because the signals that come from the direction d_s are assumed to be identical among channels by spectral compensation. We intended for the weighting function $\Psi_{n,k}(d_s)$ to whiten the noise components to reduce them with the averaging along frequency bin k. To that end, we chose $\Psi_{n,k}(d_s)$ to approximate the inverse of the noise spectrum as

$$\Psi_{n,k}(d_s) = 1/G_{dd,n,k}(d_s) \tag{7}$$

$$G_{dd,n,k}(d_s) = \overline{|X_{n,k}(d_s) - Y_{n,k}(d_s)|^2} \tag{8}$$

because it is difficult to estimate the noise spectrum directly from observations.

B. 2chSS-Based Weighting Function

In addition to averaging with whitening, we introduced a 2chSS-based weighting function for reducing correlated noise components in $G_{xy,n,k}(d_s)$. In fact, 2chSS is a modification of the Griffiths–Jim generalized sidelobe canceller (GSC) [18]. 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. The reference signal is assumed to contain only the noise component that is usually obtained as the differenced signal between the two input signals.

In our arrangement, the primary power spectrum corresponds to $|G_{xy,n,k}(d_s)|$ and the reference power spectrum corresponds to $G_{dd,n,k}(d_s)$. To allow these correspondences, we modified 2chSS, as

$$\begin{aligned}\tilde{G}_{ss,n,k} &= \max\left[|G_{xy,n,k}(d_s)| - \gamma G_{dd,n,k}(d_s)/\nu_{n,k}(d_s), 0\right] \\ &= |G_{xy,n,k}(d_s)|\,\Phi_{n,k,\gamma}(d_s)\end{aligned} \tag{9}$$

$$\Phi_{n,k,\gamma}(d_s) = \max\left[1 - \frac{\gamma G_{dd,n,k}(d_s)/\nu_{n,k}(d_s)}{|G_{xy,n,k}(d_s)|}, 0\right]. \tag{10}$$

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, $\nu_{n,k}(d_s)$ is the compensation coefficient, and $\Phi_{n,k,\gamma}(d_s)$ is the resultant 2chSS-based weighting function. The value of this function is restricted to be varied from 0 to 1 using the "max" operation. We use the compensation coefficient $\nu_{n,k}(d_s) = 1$ for all k and d_s in DOA estimation because we cannot expect precise estimation of this value in multiple source conditions or in the case where the noise duration is short for estimation. For details related to calculation of the compensation coefficient for speech enhancement, see [13] and [14].

The WWG corresponding to direction d_s is the total gain that is obtained by multiplying (6) by (10) as

$$\rho_n(d_s) = \frac{\sum_k \mathrm{Re}\left[G_{xy,n,k}(d_s)\right] \Psi^{\beta}_{n,k}(d_s)\Phi_{n,k,\gamma}(d_s)}{\sum_k G_{zz,n,k}\Psi^{\beta}_{n,k}(d_s)}. \tag{11}$$

Parameters β and γ were determined empirically. In fact, WWG is the estimate of broadband signal-to-signal+noise ratio. Therefore, WWG can be regarded as a degree of existence of the signal arriving from the look direction. Speech enhancement based on auto gain control (AGC) using WWG is performed using

$$\hat{S}_{n,k} = Z_{n,k}(d_s)\rho_n(d_s). \tag{12}$$

We note that the inversion of HRTF in (2) can impart an adverse effect on estimating WWG. Deep notches that are generally observable in the high-frequency region of HRTFs are of particular concern. However, we use a limited frequency range for investigation, as described later; this range does not contain such deep notches in the HRTFs that we used for the examination. Therefore, we do not take particular processing to avoid this effect here.

C. DOA Estimation Based on WWG

Next, consider the case in which d_s is unknown and the input signals are compensated as if an imaginary source is presented in the look direction d. We designate the directional pattern obtained by $\rho_n(d)$ by changing the look direction d as the WWG spatial spectrum. Actually, $\rho_n(d)$ takes a large value when the compensated signals are identical between channels. Therefore, $\rho_n(d)$ usually has a peak at $d = d_s$. Consequently, we can estimate source directions using peak picking. The time-averaged spatial spectrum based on WWG is obtainable as $\overline{\rho_n(d)}$. However, we instead use

$$S_{\mathrm{WWG}}(d) = \frac{\overline{\sum_k \mathrm{Re}\left[G_{xy,n,k}(d)\right] \Psi^{\beta}_{n,k}(d)\Phi_{n,k,\gamma}(d)}}{\overline{\sum_k G_{zz,n,k}\Psi^{\beta}_{n,k}(d)}} \tag{13}$$

to emphasize the time frames that have larger power. We use (13) in the evaluation. We note that a WWG calculation with reduced computation is possible [15].

III. WWG With Incremental Source Attenuation

In this section, we propose a new method to improve the localization performance of WWG in the presence of multiple sources. For simplicity, in this section, we describe equations omitting the frame number n.

First, we let the direction located by the maximum peak on $S_{\mathrm{WWG}}(d)$ be

$$d_1 = \arg\left[\max_d S_{\mathrm{WWG}}(d)\right]. \tag{14}$$

We regard the source indicated by this direction as a first found source because of the observation that the maximum peak represents a correct source direction with high certainty, even in a multiple-source environment, as shown in the experiments.

Next, we reduce the spectral components in the input spectra that are contributed by the first source. We expect that this reduction raises the second source's peak, mingled among spurious peaks generated by the first source. Considering the fact that a

2chSS-based weighting function $\Phi_{k,\gamma}(d_1)$ approximately represents the signal-to-signal+noise ratio of signal arriving from direction d_1 and that $\Phi_{k,\gamma}(d_1)$ varies from 0 to 1, then

$$1-\Phi_{k,\gamma}(d_1)=\gamma\frac{G_{dd,k}(d_1)}{|G_{xy,k}(d_1)|} \tag{15}$$

is considered to be suitable as a weighting function for reducing the peak at d_1 in place of noise. Because the above constant γ has no effect when (15) is used as a weighting function, we omit γ in (15) and modify it as

$$\Upsilon_k(d_1)=\min\left[\frac{G_{dd,k}(d_1)}{|G_{xy,k}(d_1)|},1\right]. \tag{16}$$

We use the "min" operator to restrict the function to work as an attenuator.

Next, we propose a new version of the WWG spectrum by integrating the above weighting function as

$$S_{\mathrm{WWG-ISA}}(d,d_1)=\frac{\overline{\sum_k \mathrm{Re}\left[G_{xy,k}(d)\right]\Psi_k^{\beta}(d)\Phi_{k,\gamma}(d)\Upsilon_k^{\mu}(d_1)}}{\overline{\sum_k G_{zz,n,k}\Psi_k^{\beta}(d)\Upsilon_k^{\mu}(d_1)}} \tag{17}$$

where μ is the constant to control the attenuation. We let the direction located by the maximum peak on the new spatial spectrum $S_{\mathrm{WWG-ISA}}(d,d_1)$ be

$$d_2=\arg\left[\max_d S_{\mathrm{WWG-ISA}}(d,d_1)\right]. \tag{18}$$

We regard this direction as the second found source. We can locate the third and greater sources by applying this process until we reach the assumed number of sources. The spatial spectrum for estimating Nth source ($N\geq 2$) and the direction of the Nth source are expressed, respectively, as

$$S_{\mathrm{WWG-ISA}}(d,d_1,d_2,\ldots,d_{N-1}) =\frac{\overline{\sum_k \mathrm{Re}\left[G_{xy,k}(d)\right]\Psi_k^{\beta}(d)\Phi_{k,\gamma}(d)\prod_{i=1}^{N-1}\Upsilon_k^{\mu}(d_i)}}{\overline{\sum_k G_{zz,n,k}\Psi_k^{\beta}(d)\prod_{i=1}^{N-1}\Upsilon_k^{\mu}(d_i)}} \tag{19}$$

$$d_N=\arg\left[\max_d \mathrm{S}_{\mathrm{WWG-ISA}}(d,d_1,\ldots,d_{N-1})\right]. \tag{20}$$

IV. Experimental Setup

A. Methods for Comparison

For comparative evaluation among methods, we describe two kinds of GCC [19] and two high-resolution methods: MUSIC and MV. We assign "GCC-PHAT" to an equivalent to GCC with the amplitude spectrum whitened by $|G_{xy,n,k}(d)|$. The time-averaged spatial spectrum of GCC-PHAT is calculated as

$$S_{\mathrm{GCC-PHAT}}(d)=\overline{\sum_k \mathrm{Re}\left[G_{xy,n,k}(d)\right]/|G_{xy,n,k}(d)|}. \tag{21}$$

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

$$\Gamma_{n,k}(d)=\frac{\zeta_{n,k}^2}{\left(1-\zeta_{n,k}^2\right)|G_{xy,n,k}(d)|} \tag{22}$$

$$\zeta_{n,k}^2=|W_{xy,n,k}|^2/(W_{xx,n,k}W_{yy,n,k}) \tag{23}$$

where $\zeta_{n,k}^2$ is the squared coherence function between L-channel and R-channel. The weighting function $\Gamma_{n,k}(d)$ is intended to minimize the uncorrelated noise power between channels [19]. We calculate the time-averaged spatial spectrum of this method as the following:

$$S_{\mathrm{GCC-ML}}(d)=\overline{\sum_k \mathrm{Re}\left[G_{xy,n,k}(d)\right]\Gamma_{n,k}(d)}. \tag{24}$$

For MUSIC and MV, we introduce the binary selection of spectral components based on coherence detection (CD), as described in Section I. We average the spatial spectra both on the frequency axis and over time as

$$S_{\mathrm{MUSIC-CD}}(d)=\overline{\sum_k r_{n,k}\log\left(\frac{1}{U_k(d)^* E_{n,k}E_{n,k}^* U_k(d)}\right)} \tag{25}$$

$$S_{\mathrm{MV-CD}}(d)=\overline{\sum_k \frac{r_{n,k}}{U_k(d)^* R_{n,k}U_k(d)}}^{-1} \tag{26}$$

$$R_{n,k}=\begin{bmatrix}W_{xx,n,k} & W_{xy,n,k}^*\\ W_{xy,n,k} & W_{yy,n,k}\end{bmatrix} \doteq\begin{bmatrix}\overline{X_{O,n,k}X_{O,n,k}^*} & \overline{X_{O,n,k}Y_{O,n,k}^*}\\ \overline{Y_{O,n,k}X_{O,n,k}^*} & \overline{Y_{O,n,k}Y_{O,n,k}^*}\end{bmatrix} \tag{27}$$

$$U_k(d)=\{H_{x,n,k},H_{y,n,k}\} \tag{28}$$

$$r_{n,k}=\begin{cases}1, & \text{if } \frac{e_{L,n,k}}{e_{S,n,k}}>r_{th}\\ 0, & \text{else}\end{cases} \tag{29}$$

where $R_{n,k}$ denotes the covariance matrix of the input signal, $E_{n,k}$ denotes the eigenvector corresponding to the smaller eigenvalue of the eigendecomposition of $R_{n,k}$, $e_{L,n,k}$ and $e_{S,n,k}$, respectively, represent the larger and smaller eigenvalues, and r_{th} is a constant to determine whether the spectral component is coherent or not.

It might be possible to replace the observed spectra $W_{xx,n,k}$, $W_{yy,n,k}$, and $W_{xy,n,k}$ that are contained in $R_{n,k}$ with the compensated versions of $G_{xx,n,k}(d)$, $G_{yy,n,k}(d)$, and $G_{xy,n,k}(d)$, respectively. The performance of this method is unknown, but it is unattractive because it requires eigendecomposition of the compensated covariance matrix for every look direction d. Therefore, we do not describe this method here.

As an alternative to CD, emphasis of the coherent spectral components based on the ratio of the larger and smaller eigenvalues, i.e.,

$$r_{n,k}=\frac{e_{L,n,k}}{e_{S,n,k}} \tag{30}$$

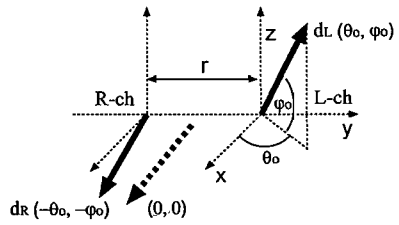

Fig. 1. Arrangement of directional microphones in rotational symmetry.

TABLE I
DIRECTIONS MEASURED FOR HRTF

elevation ($-60° \leq \varphi \leq 90°$)	azimuth step
$+90°$	$360°$
$+85°, +80°$	$30°$
$+75°, +70°$	$15°$
$+65°, \pm 60°$	$10°$
$-55° - - + 55°$	$5°$

is worth mentioning. However, because this method and binary selection (29) provided similar results in preliminary experiments, we do not describe the method in this paper.

B. Simulation Conditions

For an evaluation of the DOA estimation performance, we assume the following two conditions for sound acquisition.

- Condition D) Two directional microphones positioned in a free sound field receive direct sound from sources. We assume no reflection and no diffraction.
- Condition H) Two omni-directional microphones at the eardrum acquire incident sound altered by human HRTFs. The human head is assumed to be set in a free sound field.

Condition D is included to investigate the applicability of the proposed method not only to an HRTF case but also to a case using popular directional microphones. Effects of room reverberation and reflection are not incorporated in this study. Evaluation in such adverse conditions is interesting, but it is beyond the scope of this study.

1) Condition D: In this condition, we assume that each source sound arrives at the microphones as a plane wave. Furthermore, two uni-cardioid microphones are assumed for use. Therefore, the time delay between source signals received at the L-channel and R-channel and amplitude response of each source sound at each microphone are incorporated to calculate the microphone signals.

Fig. 1 shows the arrangement of the directional microphones. We set the front directions of the microphones to have rotational symmetry of 180°; we set an inter-microphone distance to 15 cm, the L-channel microphone direction (θ_o, φ_o) to (60°, 60°), and that of the R-channel to $(-60°, -60°)$.

2) Condition H: We use four sets of individual human HRTFs for evaluation. We previously measured those HRTFs in an anechoic chamber. For the HRTF measurement, we made four precise head figures corresponding to two females and two males. The head figures were generated from both three-dimensional data of the ear bone structure obtained using X-ray computer tomography and data of the head surface shape measured using a laser scanner. We mounted a binaural microphone (4101; B&K) without a supporting wire at the position of the eardrum of each head figure. We radiated a time-stretched pulse (TSP) [20] using a loudspeaker (TD712z; Eclipse) as a test signal to measure their HRTFs. The distance between the head figure and the speaker was 1.5 m. The shoulder reflections do not exist in the measured head-related impulse response (HRIR) because the head figures have no body under the neck.

Table I shows that the angular range of elevation is $-60° \leq \varphi \leq 90°$ with 5° step and that of the azimuth is $-180° \leq \theta \leq 180°$, with steps varied according to the elevation. We denote the two male HRTFs as (M1) and (M2), and the two female ones as (F1) and (F2). We have not compensated the speaker and microphone characteristics because they are flat in the frequency range used for simulations. Furthermore, we consider that their effect on localization accuracy is negligible. In the computer simulation for HRTF cases, the source directions for generating input signals are chosen randomly from among the measured directions described above. We used linear interpolation of HRTFs to address unmeasured directions that were necessary to calculate the spatial spectra.

C. Measures for Evaluation

1) Source Detection Rate (SDR): We use the SDR as an evaluation measure to examine performance through various source directions and through various quantities of sources

$$SDR(M_s, e_p) = K_{\text{success}}(M_s, e_p)/K_{\text{total}}(M_s). \quad (31)$$

In that equation, M_s is the assumed number of sources, e_p is the permissible error for judging detections as either successful or failed, $K_{\text{total}}(M_s)$ is the number of trials of DOA detection, and $K_{\text{success}}(M_s, e_p)$ is the number of successful trials. For SDR measurements, the spatial spectra are first calculated using each DOA method; then the M_s peaks in the spatial spectra are detected assuming that M_s is known *a priori*. Directions of the detected peaks are compared to the actual ones. Then, the detection is judged according to that comparison. The permissible error defines a permissible margin of the angular distance between the true direction and the detected peak direction. We use $e_p = 5°$ to exhibit SDR results because the performance difference among methods is clearly observable around this value, as presented in [15].

Many sets of source directions are generated randomly to distribute uniformly in a specified angular range to avoid bias introduced using a specific source direction. Thereafter, the number of sources M_s is varied and the detection rates are obtained for each DOA method and in each M_s. The number of sources M_s is varied from one to three; 1000 sets of source directions are generated randomly for each M_s. They are commonly used for each head model in Condition H. Thereby, $K_{\text{total}}(M_s) = 4000$ for Condition H, and $K_{\text{total}}(M_s) = 1000$ for Condition D are used.

2) Maximum Peak Correct Rate: The WWG-ISA procedure regards the direction of the maximum peak on the WWG spatial spectrum as a source direction. Consequently, whether the maximum peak corresponds to the true source or not is an important consideration, particularly at the stage of estimating the first source direction. Therefore, we calculate the maximum peak

TABLE II
ANGULAR RANGES FOR SPATIAL SPECTRUM CALCULATION

		Condition-D	Condition-H
azimuth-only case	azimuth	$-90° \leq \theta \leq 90°$	$-180° \leq \theta \leq 180°$
	elevation	$\varphi = 0°$	$\varphi = 0°$
	step	1°	1°
two-dimensional case	azimuth	$-90° \leq \theta \leq 90°$	$-180° \leq \theta \leq 180°$
	elevation	$-90° \leq \varphi \leq 90°$	$-60° \leq \varphi \leq 90°$
	step	2°	5°

TABLE III
ANGULAR RANGES FOR SOURCE DIRECTION SETS

		Condition-D	Condition-H
azimuth-only case	azimuth	$-85° \leq \theta \leq 85°$	$-180° \leq \theta \leq 180°$
	elevation	$\varphi = 0°$	$\varphi = 0°$
	min. dist.	10°	10°
two-dimensional case	azimuth	$-85° \leq \theta \leq 85°$	$-180° \leq \theta \leq 180°$
	elevation	$-85° \leq \varphi \leq 85°$	$-55° \leq \varphi \leq 90°$
	min. dist.	10°	10°

correct rate (MCR) as an additional measure to confirm the validity of the ISA procedure. This measure represents the ability to estimate the direction of a dominant source in a multiple-source environment. Therefore, we believe it is useful to reveal the robustness of the DOA estimation methods.

3) Root Mean Squared Error (RMSE): We additionally calculate the RMSE of the angular distance between the peak directions and their true directions to investigate the accuracy of the peak position. The same spatial spectra obtained for SDR measurements are used for this evaluation. Only the pairs of the true peak direction and its corresponding peaks that resulted from successful detection are considered. The calculated results are shown as a function of the permissible error.

D. Analysis Condition

For evaluation, we commonly use a sampling frequency of 48 kHz, FFT of 2048-point length with a Hanning window, a frame shift of 1024 points, and a frequency band including 260 Hz to 4 kHz for calculating the spatial spectra. The number of averaging iterations to estimate spectra $W_{xx,k}$, $W_{yy,k}$, and $W_{xy,k}$ to 3(85 ms) was determined through preliminary experiments.

Additionally, we set the index of the weighting function β to 1, the strength of the 2chSS γ to 1, and the parameter for WWG-ISA μ to 1.5 for Condition H. For Condition D, we set β to 0.75, γ to 3 and μ to 1.2. As the threshold of the binary selection (29), we used $r_{th} = 300$. These values were also determined through preliminary experiments.

For the performance evaluation, we investigated a one-dimensional localization (azimuth only) in addition to the two-dimensional (azimuth and elevation) one. We calculated the spatial spectrum in an angular range depending on the conditions. They are presented in Table II. The "step" rows in Table II indicates the angular step for calculating the spatial spectra. In addition, Table III summarizes the angular ranges for source directions to be generated randomly in the evaluation. We restricted every distance between sources to be larger than 10°, as described in the "min. dist." rows in Table III.

In Condition D, directions in the half-sphere of the backside ($-180° \leq \theta \leq -90°$ and $90° \leq \theta \leq 180°$) are excluded from the evaluation because amplitudes of the incident sounds from the back side are attenuated by the uni-cardioid directivity. Degraded localization of those sounds spoils the overall performance. Other types of directivity are useful to avoid this problem. However, the aim of examining Condition D is not comparison to the performance in Condition H, but to confirm the validity of the proposed method also in a case where popular directional microphones are used. Although performance differences depending on the microphone directivity are interesting to us, they are beyond the scope of this paper.

E. Speech and Noise Data

We used speech samples uttered by ten males from a Japanese speech database "ETL-WD I & II," which was distributed by the National Institute of Advanced Industrial Science and Technology (AIST). Because these signals were sampled at 16 kHz, we converted the sampling frequency to 48 kHz. We concatenated the samples of 492 words by each talker and excluded silent segments based on the short-term signal power from the concatenated speech signals. Thereby, we obtained ten speech signals of about 230 s without silence pauses. As background noise, computer fan noise that radiated simultaneously from many computers in the experiment room was recorded using binaural microphones mounted at the position of the eardrum of the head model (F2). The speech and noise samples described above are filtered using a 301-point linear phase FIR filter with a passband of 260 Hz to 4 kHz for ease of adjusting the signal-to-noise ratio (SNR) in the experiments.

For simulations of localization, the M_s speech signals were chosen from the above ten signals and were assigned to the assumed M_s sound sources. The assignment of the speech signals to the sources was changed by each trial of localization. In each trial of the localization, we used speech and background noise segments with respective durations of 1 s. They were chosen randomly from the above ten band-passed speech and noise samples. In case of multiple source conditions, the amplitude of each speech segment corresponding to each source is adjusted to have equal power, i.e., we set all the sources to have equal power in each simulation. The SNRs in the experiments were specified as the ratio of this equal power to the background noise power.

V. EVALUATION

A. Spatial Spectra

Before SDR evaluation of the DOA estimation methods, we present the spatial spectra obtained using WWG and WWG-ISA to emphasize the change of spectra shape according to the ISA steps. We assumed two sources (A and B) of equal power: the direction of source A is (0°, 0°); that of source B is (20°, 50°).

First, we present Fig. 2 to depict the WWG and WWG-ISA spectra obtained in Condition D at SNR = 10 dB. The displayed spectra are normalized by the maximum value in each spectrum. The true directions are indicated by arrows in the figure. Fig. 2(1) portrays that the WWG spectrum exhibits two clear peaks at the true source directions. Source A has the maximum height in the spectrum. Next, Fig. 2(2) depicts the WWG-ISA spectrum, which has attenuated contribution of source A. We can observe that the peak of source A is almost reduced and that

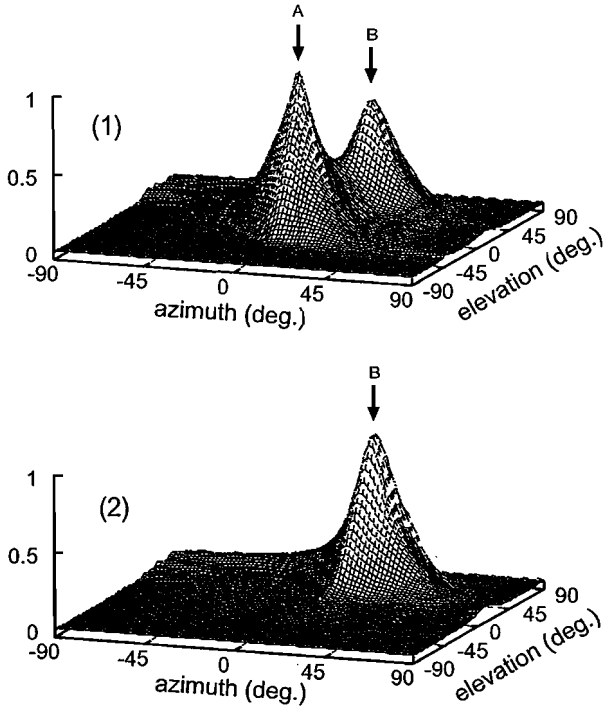

Fig. 2. WWG and WWG-ISA spatial spectra obtained in Condition D [(1) WWG spectrum, (2) source A is attenuated by WWG-ISA].

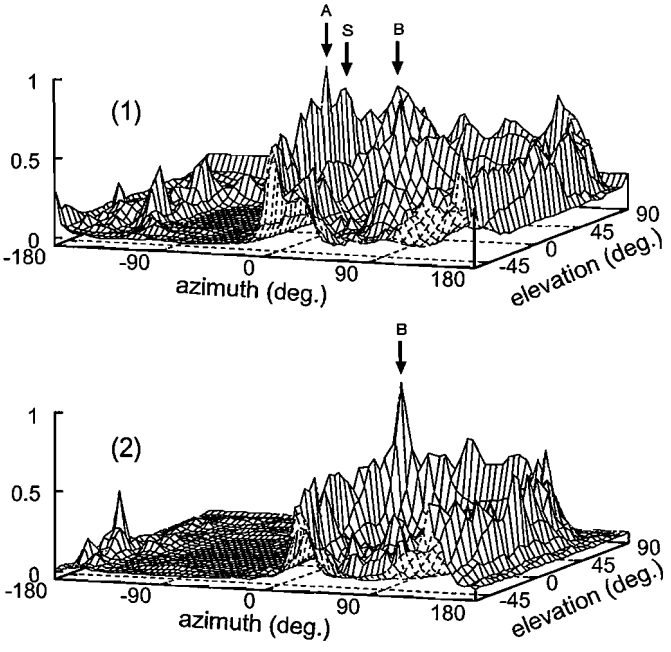

Fig. 3. WWG and WWG-ISA spectra obtained in Condition H [(1) WWG spectrum, (2) source A is attenuated by WWG-ISA].

the peak of source B in this figure also indicates the correct direction of source B. We can readily distinguish two source directions from the ordinal WWG spectrum, as observed in Fig. 2(1). Therefore, the ISA step is not necessary in this condition.

Subsequently, we show the results obtained in Condition H at SNR = 10 dB in Fig. 3; HRTF (M1) was used for the calculation. The true directions are indicated by arrow A and arrow B in the figure, as in Fig. 2. Fig. 3(1) shows that the WWG spectrum contains many peaks, in contrast to that obtained in Condition D. The second highest peak is a spurious peak indicated by arrow S; and the third peak indicates source B. Nevertheless, the maximum peak indicates the true direction of source A. Fig. 3(2) depicts the WWG-ISA spectrum with the attenuated contribution of source A. It is apparent that the peak of source A and accompanying peaks are reduced and that the maximum peak in this spectrum indicates the correct direction of source B.

B. Localization Performance

Now, SDR, MCR, and RMSE are examined to evaluate localization performance quantitatively. In addition to the two-dimensional (azimuth and elevation) case, a one-dimensional case (azimuth only) is examined here. However, we do not show MCR and RMSE results for the one-dimensional case because they are similar to those of the two-dimensional case.

For the WWG-ISA procedure, the maximum peak is selected from the spectrum obtained at each step of WWG-ISA using (19) and (20), and the accumulated M_s peak directions are regarded as the DOA estimates. For methods other than WWG-ISA, the highest M_s peaks are selected at once from the spectrum obtained using each method. The detected M_s peak directions are compared to the true M_s source directions. The detected peak that is nearest to each true direction is marked as the corresponding detected peak of the true source. Correspondence is regarded as valid when the difference between the true source direction and the corresponding detected peak is smaller than the permissible error. We regard the detection as successful when all true sources have valid corresponding peaks in a one-to-one relation.

1) One-Dimensional Case (Condition D): Here, we conducted an evaluation for the azimuth-only case.

First, the SDR results obtained in Condition D are depicted in Fig. 4. Fig. 4(1) shows that SDRs $\geq$ 90% is attained when M_s = 1 at SNR $\geq$ 10 dB by all the methods examined. The WWG is superior to the other methods, particularly at SNR $\leq$ 5 dB. Two GCCs deteriorate greatly, although the other methods do so moderately and WWG-ISA exhibits the highest performance when M_s = 2. The SDRs of MV-CD, MUSIC-CD, WWG, and WWG-ISA at the SNR = 10 dB are about 85%, 86%, 92%, and 94%, respectively. The superiority of WWG-ISA is more remarkable for M_s = 3. The respective SDRs of MV-CD, MUSIC-CD, WWG, and WWG-ISA at the SNR = 10 dB when M_s = 3 are 52%, 49%, 73%, and 78%.

2) One-Dimensional Case (Condition H): Next, Fig. 5 shows the SDR results obtained in Condition H. This figure shows that SDRs $\geq$ 95% are attained when M_s = 1 by WWG and MUSIC-CD at SNR $\geq$ 5 dB; also, MUSIC-CD is slightly better than WWG at SNR $\geq$ 0 dB. When M_s = 2, the two GCCs and MV-CD deteriorate by less than 35% at SNR = 10 dB, whereas MUSIC-CD, WWG, and WWG-ISA attain SDRs of 53%, 71%, and 92%, respectively. The performance degradation for increased M_s is greater than that in Condition D. Nevertheless, WWG-ISA exhibits the highest performance. The superiority of WWG-ISA is more remarkable when M_s = 3. The

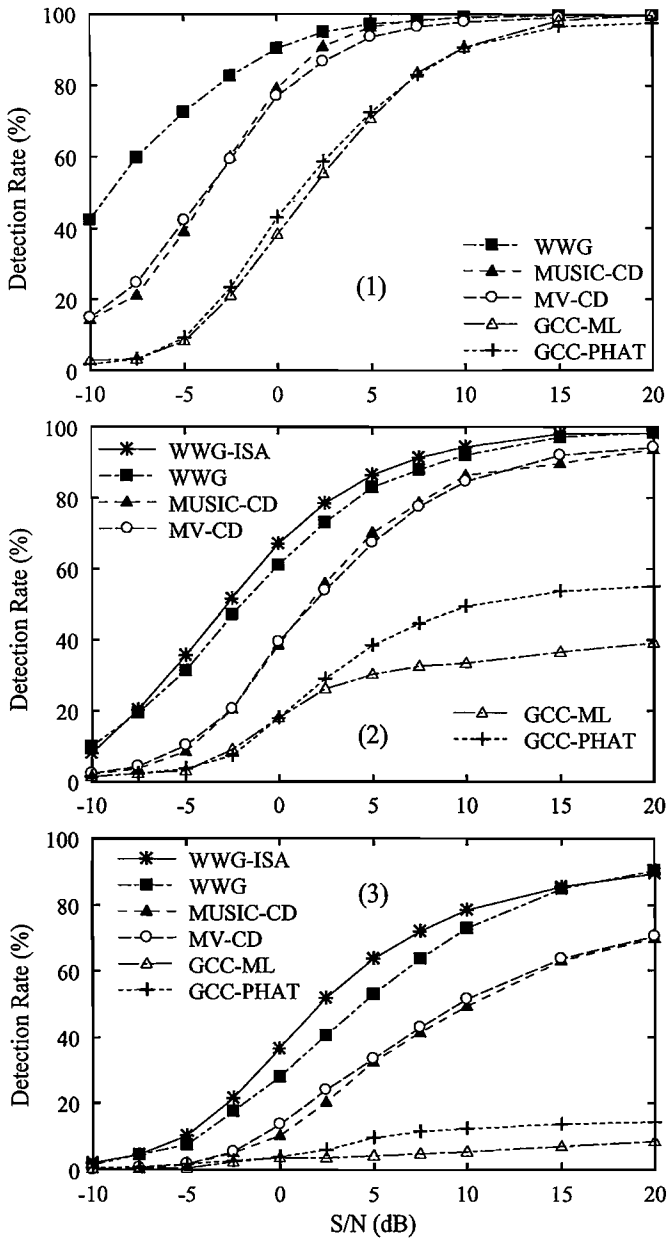

Fig. 4. SDR versus SNR in one-dimensional case in Condition D (Directional Microphones) at permissible error (e_p) = 5° [(1) single source, (2) two sources, (3) three sources].

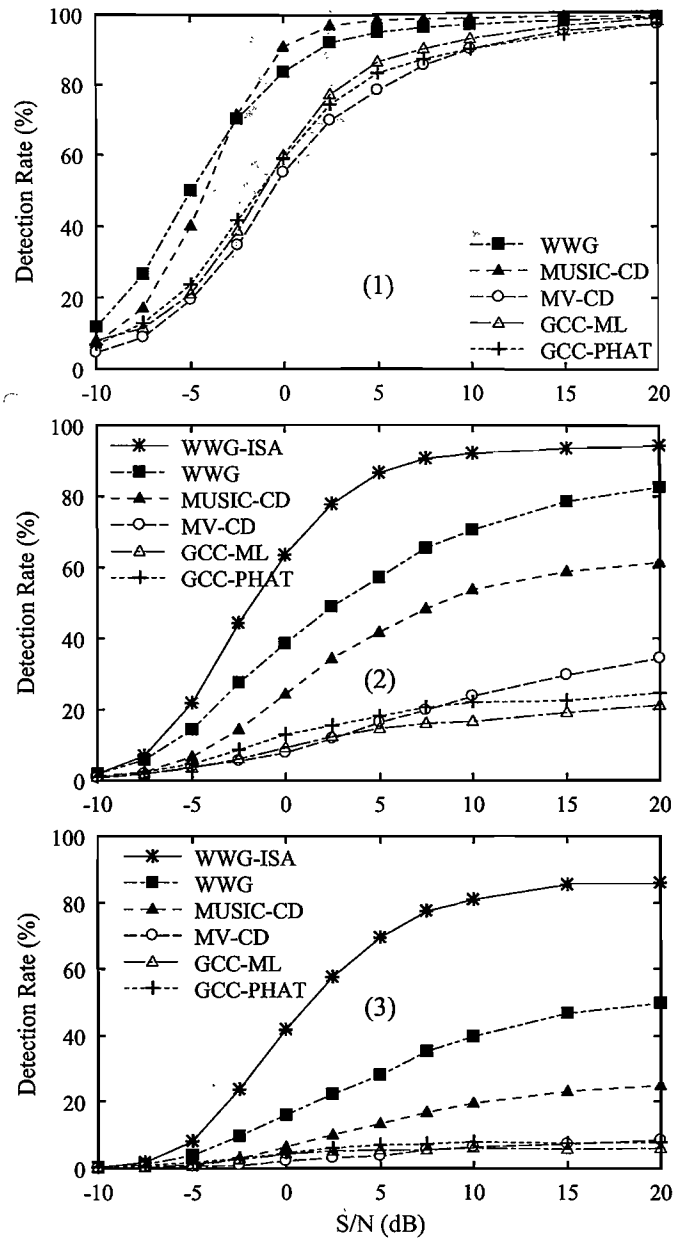

Fig. 5. SDR versus SNR in a one-dimensional case in Condition H (HRTF) at permissible error (e_p) = 5° [(1) single source, (2) two sources, (3) three sources].

respective SDRs of MUSIC-CD, WWG, and WWG-ISA at the SNR = 10 dB when $M_s = 3$ are 20%, 40%, and 81%.

3) Two-Dimensional Case (Condition D): Next, the SDR, MCR, and RMSE in a two-dimensional case are examined. The obtained results in Condition D are depicted, respectively, in Figs. 6–8. The SDR and MCR results were obtained at $e_p = 5°$ as functions of SNR; the RMSE results were obtained at SNR = 10 dB as functions of e_p.

Fig. 6(1) shows that SDRs larger than 90% are attained by MUSIC-CD, MV-CD, and WWG when $M_s = 1$ at SNR $\geq$ 7.5 dB, although the SDRs of two GCCs are very low. The lowest performance, that of GCC-PHAT, is attributable to the fact that GCC-PHAT has no angular resolution in elevation because the amplitude difference between the two channels is omitted through normalization by $|G_{xy,n,k}(d)|$, as expressed in (21). It is readily apparent that WWG-ISA is superior to the other methods, but the SDR improvement compared to WWG is not so great when $M_s = 2$ [Fig. 6(2)]. The respective SDRs of MV-CD, MUSIC-CD, WWG, and WWG-ISA at SNR = 10 dB when $M_s = 2$ are 75%, 67%, 79%, and 85%. The superiority of WWG-ISA is greater at high SNRs when $M_s = 3$ [Fig. 6(3)].

Next, the MCR results are portrayed in Fig. 7. The MCR difference between WWG and MUSIC-CD is small; WWG attains its highest MCR at SNR $\leq$ 5 dB and is comparable to MUSIC-CD at SNR > 7.5 dB.

The RMSE results are shown in Fig. 8. We display the result only when the number of successful detections is greater than 50 samples. Thereby, we avoid inclusion of less reliable data of

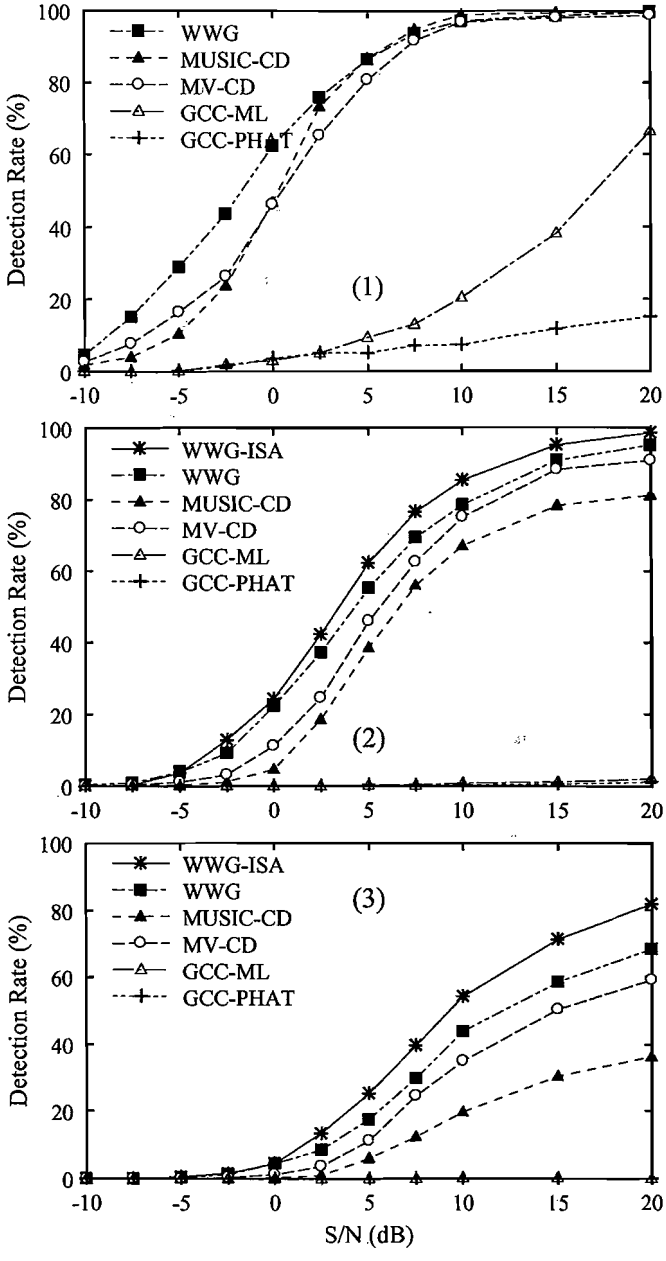

Fig. 6. SDR versus SNR in a two-dimensional case in Condition D (Directional Microphones) at permissible error $(e_p) = 5°$ [(1) single source, (2) two sources, (3) three sources].

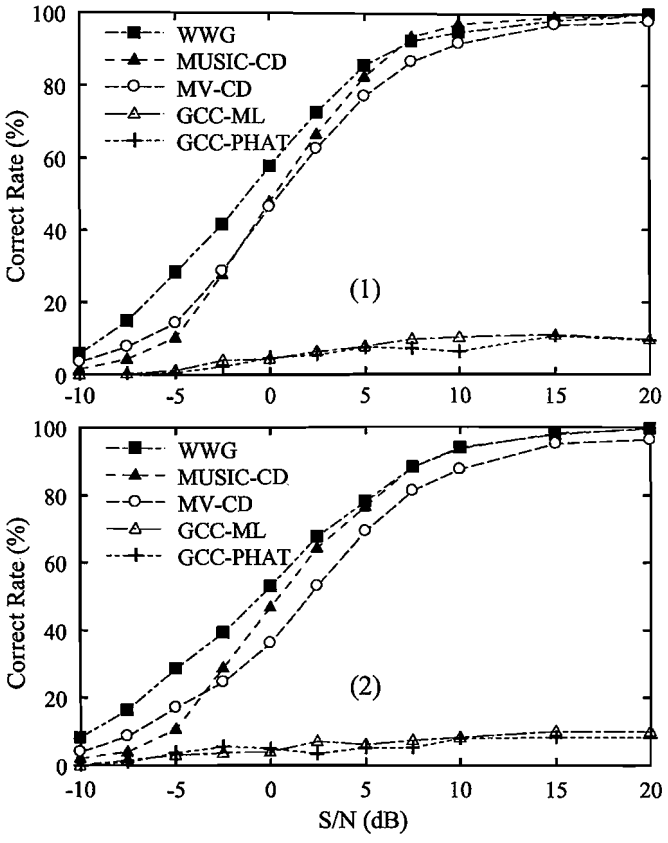

Fig. 7. MCR versus SNR in a two-dimensional case in Condition D (Directional Microphones) at permissible error $(e_p) = 5°$ [(1) two sources, (2) three sources].

the small samples used for calculating the RMSE. This figure shows that the RMSEs of GCC-PHAT and GCC-ML are larger than those obtained using the other methods. These values increase mostly in relation to the permissible error. These methods have insufficient successful detections to display all permissive errors when $M_s = 3$. By contrast, MUSIC-CD, WWG, and WWG-ISA have small RMSEs. They are almost constant when $e_p > 10°$. We can observe that these methods degrade slightly as M_s increases. Their performances in terms of RMSE are almost equivalent.

In consideration of the results of SDR, MCR, and RMSE obtained in Condition D, we can confirm that WWG-ISA achieves the highest detection rate among the methods that were compared in multiple source environments while retaining detection accuracy equivalent to high-resolution methods.

4) Two-Dimensional Case (Condition H): We present evaluation results for the two-dimensional case in Condition H. Figs. 9 and 10, respectively, depict SDR and MCR results. Fig. 9(1) shows that MUSIC-CD and WWG attain SDRs larger than 90% when $M_s = 1$ at the SNR $\geq$ 5 dB, whereas the other methods yield less than 50% at SNR = 5 dB. We can observe that WWG-ISA is far superior to the other methods when $M_s \geq 2$ [Fig. 9(2)]. The respective SDRs of MUSIC-CD, WWG, and WWG-ISA at SNR = 10 dB when $M_s = 2$ are 27%, 48%, and 90%. Regarding the MCR performance, similar results to those in Condition D are obtained for WWG and MUSIC-CD, but the MV-CD performance is degraded compared to that in Condition D.

Subsequently, the RMSE results are displayed in Fig. 11. This figure shows that the RMSEs of GCC-PHAT, GCC-ML, and MV-CD are larger than those obtained using the other methods and degrade greatly through an increase of M_s. In contrast to those methods, the RMSEs of MUSIC-CD, WWG and WWG-ISA are small and degradation through the increase of M_s is moderate. We can observe that WWG-ISA has the smallest RMSEs among the methods when $M_s = 2$ at $e_p \geq 15°$ and when $M_s = 3$ at $e_p \geq 10°$.

The results obtained for the two-dimensional case in Condition H demonstrate that the proposed WWG-ISA method has

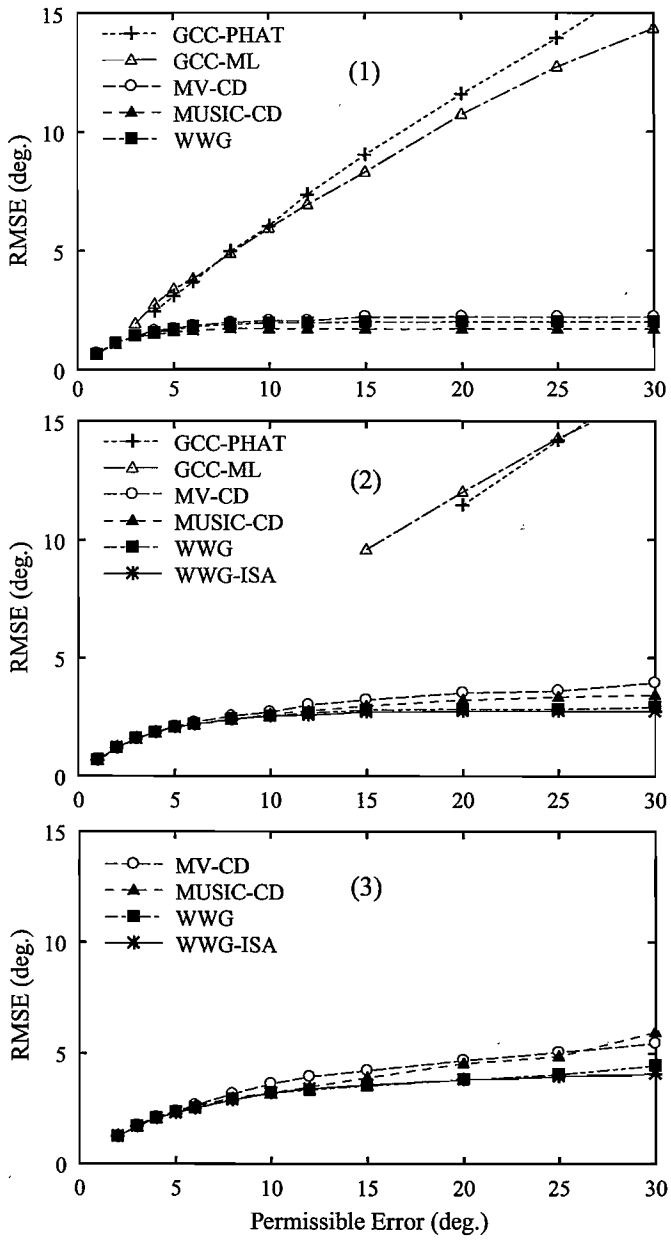

Fig. 8. RMSE versus permissible error in a two-dimensional case in Condition D (Directional Microphones) at SNR = 10 dB [(1) single source, (2) two sources, (3) three sources].

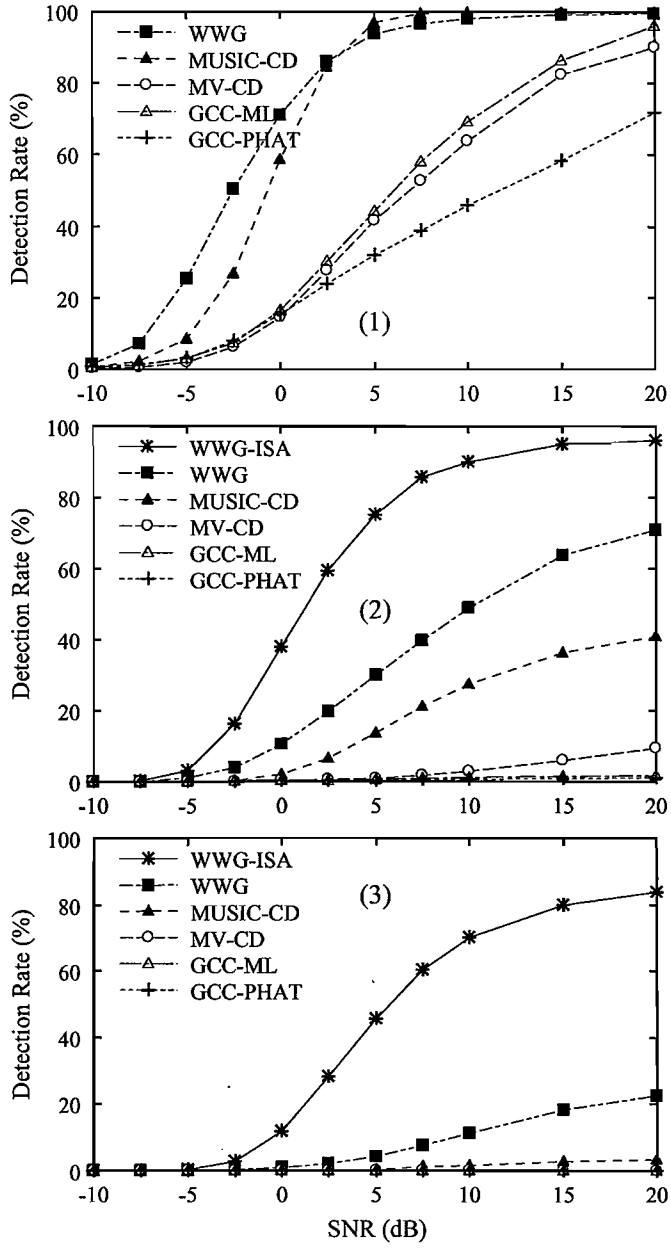

Fig. 9. SDR versus SNR in a two-dimensional case in Condition H (HRTF) at permissible error (e_p) = 5° [(1) single source, (2) two sources, (3) three sources].

higher accuracy and much more robustness to the increase of sound sources than the other compared methods do. We confirmed that WWG-ISA can attain high performance in two-dimensional localization of multiple sound sources for the case in which the incident signals are altered by HRTFs.

C. Detailed Investigation for Condition H

In this section, we further investigate the SDR performance in Condition H. For detailed investigation, we also examine the dependency of SDR on the source separation and their absolute directions when two concurrent sources are present. We omit presentation of the results of the two GCC methods because no result worth noting was obtained.

1) Random Source Direction in Elevation-Only: First, we examine the SDR performance of the elevation-only case, in which source directions are generated randomly in the frontal vertical plane. The angular range for source elevation is from $-55°$ to 90°; and that for estimating spatial spectra is from $-60°$ to 90° with step of 1°. The frontal directions ($\theta = 0°$) and backward directions ($\theta = \pm 180°$) were unified and treated as one region. Therefore, the results include front-back confusions. The other conditions for simulation are the same as those of the azimuth-only case.

Fig. 12 shows the results obtained. As displayed in this figure, the overall tendency resembles that of the azimuth-only case described in Section V-B.2 (Fig. 5). The SDRs of WWG and WWG-ISA in multiple source conditions are lower than those in

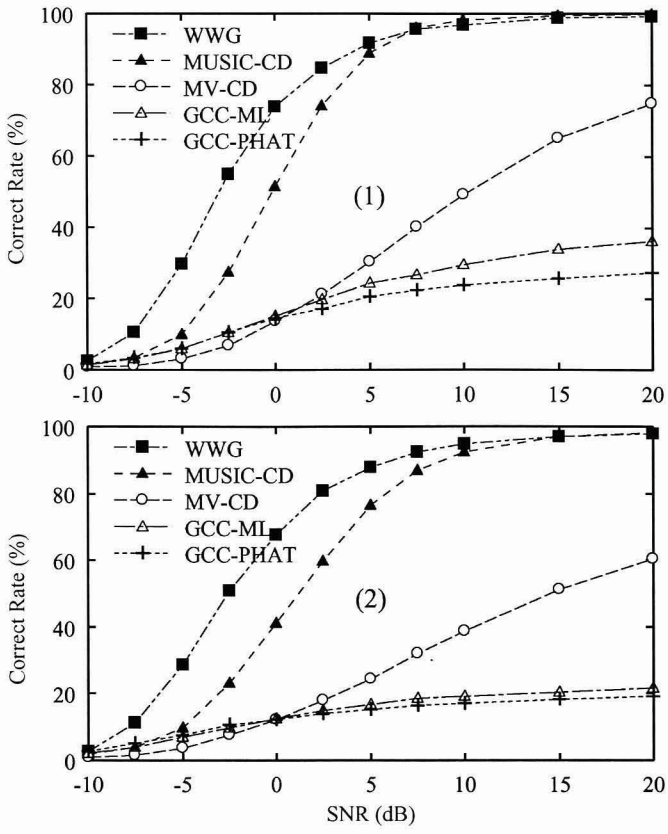

Fig. 10. MCR versus SNR in a two-dimensional case in Condition H (HRTF) at permissible error $(e_p) = 5°$ [(1) two sources, (2) three sources].

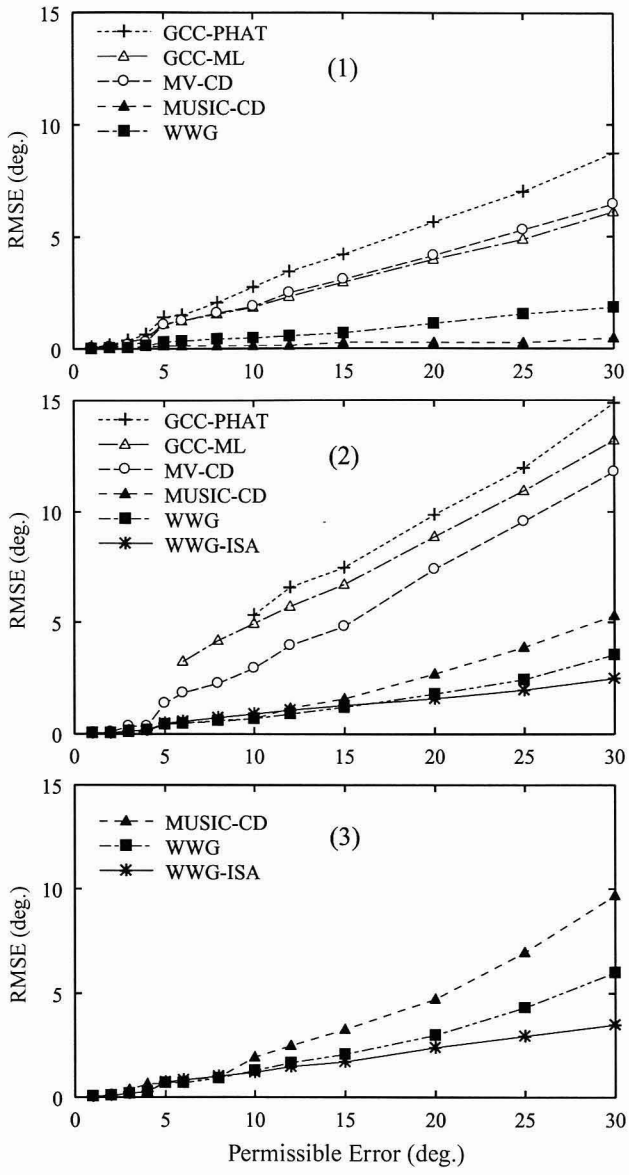

Fig. 11. RMSE versus permissible error in a two-dimensional case in Condition H (HRTF) at SNR = 10 dB [(1) single source, (2) two sources, (3) three sources].

the azimuth-only case. In this case also, WWG-ISA outperforms the other methods.

2) Dependency on Source Separation and Absolute Location: Next, we examine the dependence of SDR on the source separation and their absolute directions. We calculate SDRs when two sources (source A and source B) are present at two different directions, maintaining constant angular separation of ϕ. Here, SDR measurements of two kinds are performed. One is the azimuth-only case, in which source A is set at (θ_A, φ_A) and source B is set at $(\theta_A + \phi, \varphi_A)$ varying θ_A with fixed φ_A and ϕ. The other is the elevation-only case, in which source A is set at (θ_A, φ_A) and source B is set at $(\theta_A, \varphi_A + \phi)$ varying φ_A with fixed θ_A and ϕ. The total number of trials for each varied direction of θ_A or φ_A is 1200. Only the signal content is changed from trial to trial. We note that the resultant SDR curves were almost unchanged by varying the number of trials from 600 to 1200 in the preliminary experiment.

Fig. 13 shows the results obtained in the azimuth-only case where θ_A was varied from $-180°$ to $180°$ with a step of $5°$ with $\varphi_A = 0°$ and source separation $\phi = 5°$. The SNR was set to 10 dB. The results are shown as a function of θ_A. We observe that SDRs of all the method degrade at azimuth around $\theta = \pm 90°$. This degradation corresponds to the fact that ITD is almost constant around this azimuth and that the resolution along azimuth becomes low. We note that the results obtained when the source separations were $\phi = 10°$ and $20°$ also have such degradation around $\theta = \pm 90°$, but the degradation is less than those at $\phi = 5°$.

Finally, Figs. 14 and 15 show the results obtained in the elevation-only case where elevation φ_A was varied from $-55°$ to $90°$ with step of $5°$ with source separation $\phi = 5°$. The SNR was set to 10 dB. The results are shown as a function of φ_A. Fig. 14 corresponds to the case in which the azimuth θ_A is fixed to $0°$ and Fig. 15 corresponds to $90°$.

We observe that WWG, MUSIC-CD, and MV-CD tend to exhibit higher performances as φ_A gets closer to $0°$ when $\theta_A = 0°$ (Fig. 14), whereas those when $\theta_A = 90°$ (Fig. 15) exhibit lower performances as the elevation approaches $0°$. We note that this tendency is weaker when the source separations were $\phi = 10°$ and $20°$ in both cases of $\theta_A = 0°$ and $\theta_A = 90°$. These results indicate that the angular resolutions along the elevation of these three methods are higher around elevation $= 0°$ in the frontal vertical plane, but are lower when $\theta_A = 90°$.

On the other hand, WWG-ISA exhibits high SDRs in most elevations. This result demonstrates that WWG-ISA can reduce

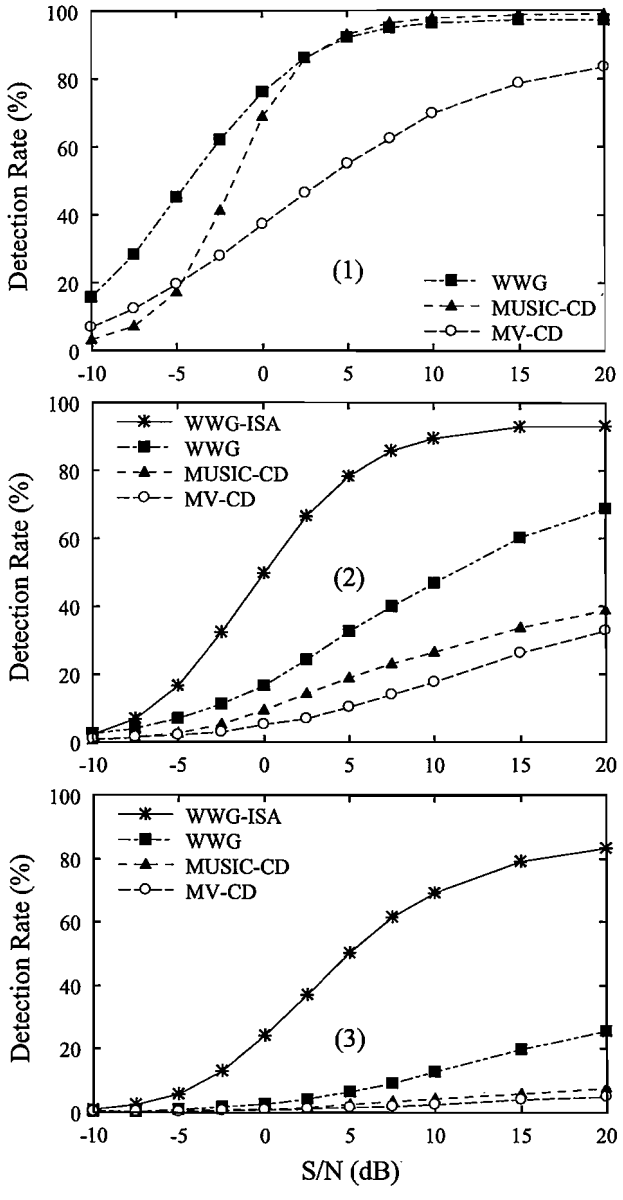

Fig. 12. SDR versus SNR in a one-dimensional case (elevation only) in Condition H (HRTF) at permissible error $(e_p) = 5°$ [(1) single source, (2) two sources, (3) three sources].

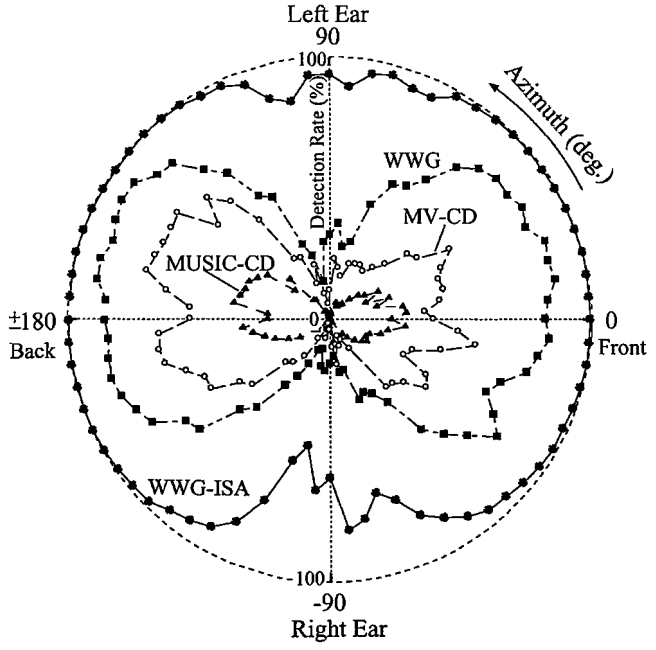

Fig. 13. SDR versus azimuth (θ_A) in a one-dimensional case (azimuth-only) in Condition H (HRTF) at permissible error $(e_p) = 5°$, SNR = 10 dB, angular separation in azimuth $\phi = 5°$, elevation $\varphi_A = 0°$.

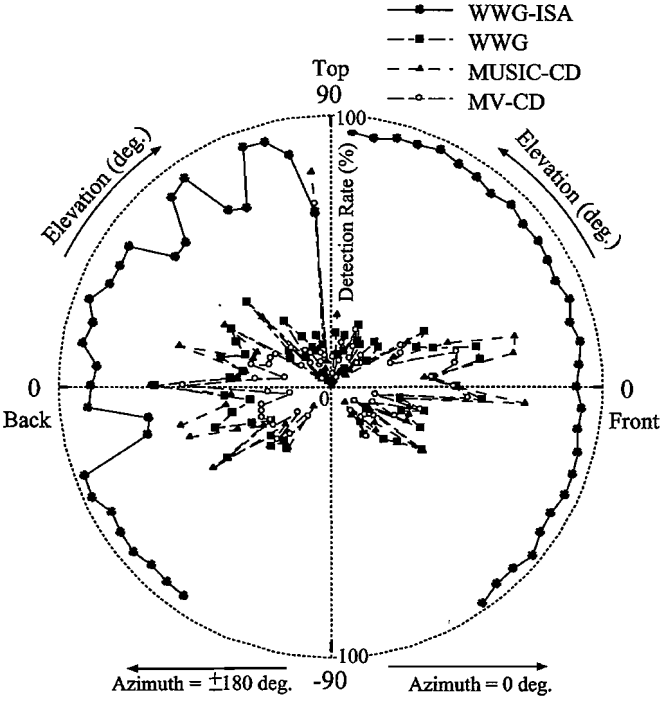

Fig. 14. SDR versus elevation (φ_A) in a one-dimensional case (elevation-only) in Condition H (HRTF) at permissible error $(e_p) = 5°$, SNR = 10 dB, angular separation $\phi = 5°$, azimuth $\theta_A = 0°, \pm 180°$.

degradation arising from the limited resolution of WWG. We note that degradations of WWG-ISA down to about 70% in some elevations are observable when $\theta_A = \pm 180°$ (Fig. 14, left half). Those degradations are mainly attributable to the result obtained using the head model (M2) which has many front-back errors in these elevations when $\theta_A = \pm 180°$. It is interesting to examine a smaller separation than 5°, but HRTFs measured using smaller angular steps are necessary for evaluation.

VI. Conclusion

This paper presented a new method of two-dimensional DOA estimation for an incident signal affected by HRTFs. The proposed method is based on the procedure designated as incremental source attenuation applied to the weighted Wiener gain. We compared the performance in terms of source detection rates, maximum peak correct rates, and root mean squared error with the two equivalents to the generalized cross correlation function and to the popular high-resolution methods of MUSIC and MV with coherence detection. Results show that, although the respective performances of both the generalized cross correlation functions and the high-resolution methods degrade remarkably with more than two sources in the condition when filtering by HRTFs alters the spectral content of the incident

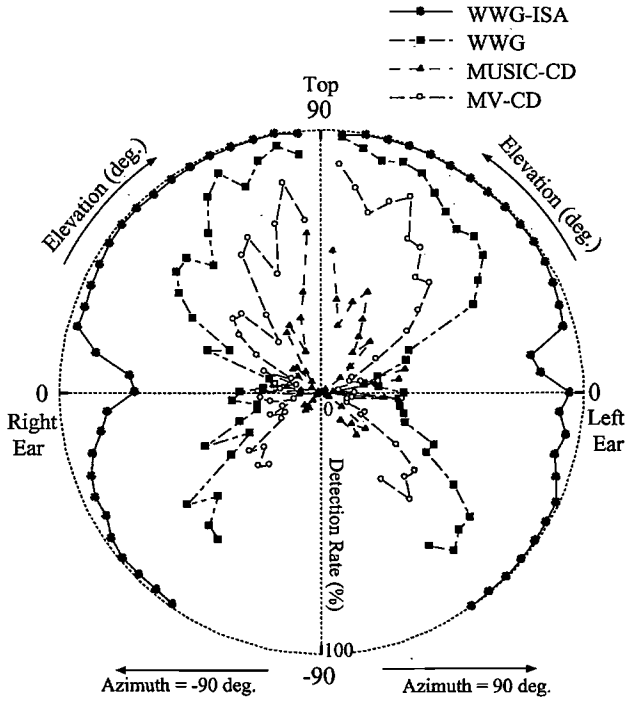

Fig. 15. SDR versus elevation (φ_A) in a one-dimensional case (elevation-only) in Condition H (HRTF) at permissible error (e_p) = 5°, SNR = 10 dB, angular separation ϕ = 5°, azimuth $\theta_A = \pm 90°$.

signal, the performance degradation of the proposed method is small and detection rates are 92% in the azimuth-only case, 90% in the elevation-only case, and 90% in the two-dimensional case at the permissible error = 5°, even in cases where the actual background noise with SNR = 10 dB and two sources are present. Furthermore, the proposed method attained the highest detection accuracy among the methods examined, as indicated by the results of the root mean squared error. These results demonstrate the superiority of the proposed method over other methods, particularly in adverse conditions with multiple sound sources. Although the number of detectable sources is increased by the proposed method, the determination of the number of sources remains as an important problem. We are considering the use of the maximum peak's height obtained at each step of the proposed incremental procedure.

ACKNOWLEDGMENT

The authors would like to thank Dr. H. Isoda, Department of Radiology, Hamamatsu University School of Medicine, for technical supports of positron emission tomography.

REFERENCES

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

[2] , D. Wang and G. J. Brown, Eds., *Computational Audio Scene Analysis*. New York: Wiley, 2006.

[3] , J. Blauert, Ed., *Communication Acoustics*. New York: Springer-Verlag, 2005.

[4] M. Bodden, "Modeling human sound-source localization and the cocktail-party-effect," *Acta Acoust.*, pp. 43–55, 1993.

[5] W. Linderman, "Extension of a binaural cross-correlation model by contralateral inhibition. i. Simulation of lateralization for stationary signals," *J. Acoust. Soc. Amer.*, vol. 80, pp. 1608–1622, 1986.

[6] C. Liu, B. C. Wheeler, W. D. O'Brien, Jr., R. C. Bilger, C. R. Lansing, and A. S. Feng, "Location of multiple sound sources with two microphones," *J. Acoust. Soc. Amer.*, vol. 108, pp. 1888–1905, Oct. 2000.

[7] C. Faller and J. Merimaa, "Sound localization in complex listening situations," *J. Acoust. Soc. Amer.*, vol. 116, pp. 3075–3089, Nov. 2004.

[8] J. Braasch, "Localization in the presence of a distracter and reverberation in the frontal horizontal plane: Ii. Model," *Acta Acoust. United With Acoust.*, vol. 88, pp. 956–969, Nov. 2002.

[9] R. Le Bouquin-Jeannès, A. A. Azirani, and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 484–487, Sep. 1997.

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

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

[12] S. Mohan, M. L. Kramer, B. C. Wheeler, and D. L. Jones, "Localization of nonstationary sources using a coherence test," in *Proc. IEEE Workshop Statistical Signal Process.*, 2003, pp. 470–473.

[13] 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. Fundamentals*, vol. E79-A, pp. 2151–2158, Dec. 1996.

[14] 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.

[15] Y. Nagata, F. Fujioka, and M. Abe, "Two-dimensional doa estimation of sound sources based on weighted Wiener gain exploiting two directional microphones," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 416–429, Feb. 2007.

[16] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[17] F. Asano, "Sound localization for robots (in Japanese)," *J. Acoust. Soc. Japan*, vol. 63, pp. 41–46, Jan. 2007.

[18] 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.

[19] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[20] F. Asano [Online]. Available: http://www.tosa.mri.co.jp/sounddb/tsp/index.html.

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 Research and Development Center, Toshiba Corporation, where he has been 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.

Prof. Nagata is a member of Acoustical Society of Japan, Information Processing Society of Japan, and the Institute of Electronics, Information, and Communication Engineers.

Satoshi Iwasaki graduated from Medical Department, Mie University, Tsu, Japan, in 1986 and received the Ph.D. degree from the Otolarygology Department, Hamamatsu University School of Medicine, Hamamatsu-shi, Japan, in 1995.

He was an Associate Professor at Hamamatsu University School of Medicine in 2000 and a Professor of Aichi Medical University School of Medicine in 2007. His research interest includes cochlear implant and hearing disturbance.

Prof. Iwaski is a member of the American Association for Research in Otolaryngology, the American Academy of Otolaryngology-Head and Neck Surgery, the Japan Otolaryngology Society, and the Japan Audiological Society.

Takahiko Hariyama studied biology at Yokohama City University, Okayama University, and Tohoku University, and received the Ph.D. degree from Kyushu University.

His main research has been focused of invertebrate photoreceptors: structures, physiology, and photoreceptor optics. In 2004, he was appointed Professor of Biology at Hamamatsu University School of Medicine, Hamamatsu-shi, Japan. Recently, his research interests have expanded to "Umwelt"; the world in which animal live using the relationship between the behavior and the sensory processing, including human echo-location.

Toyota Fujioka 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.

Tomita Obara received the B.E. and M.E. degrees from the Department of Computer and Information Science, Faculty of Engineering, Iwate University, Morioka, Japan, in 2005 and 2007, respectively.

He is currently with Panasonic Mobile Communications R&D Lab. His research interests include speech signal processing.

Takayuki Wakatake received the B.E. degree in computer and information science from Iwate University, Morioka, Japan, in 2007. He is currently pursuing the M.S. degree in the Graduate School of Engineering, Iwate University.

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, Acoustical Society of Japan, the Institute of Noise Control Engineering of Japan, 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.