Speech Enhancement Based on the Subspace Method

Futoshi Asano, *Member, IEEE*, Satoru Hayamizu, *Member, IEEE*, Takeshi Yamada, *Member, IEEE*, and Satoshi Nakamura, *Member, IEEE*

***Abstract*—A method of speech enhancement using microphone-array signal processing based on the subspace method is proposed and evaluated in this paper. The method consists of the following two stages corresponding to the different types of noise. In the first stage, less-directional ambient noise is reduced by eliminating the noise-dominant subspace. It is realized by weighting the eigenvalues of the spatial correlation matrix. This is based on the fact that the energy of less-directional noise spreads over all eigenvalues while that of directional components is concentrated on a few dominant eigenvalues. In the second stage, the spectrum of the target source is extracted from the mixture of spectra of the multiple directional components remaining in the modified spatial correlation matrix by using a minimum variance beamformer. Finally, the proposed method is evaluated in both a simulated model environment and a real environment.**

***Index Terms*—Automatic speech recognition, beamformer, microphone array, speech enhancement, subspace method.**

NOMENCLATURE

ASR	Automatic speech recognition.
DS	Delay-and-sum.
MV	Minimum variance.
CSS	Coherent subspace.
NSR	Noise-dominant subspace reduction.
GEVD	Generalized eigenvalue decomposition.
D	Number of sources.
M	Number of microphones.
$\mathbf{x}_k$	Input vector.
$\mathbf{a}_{d,k}$	Directional vector.
$\mathbf{n}_k$	Ambient noise vector.
$\mathbf{s}_k$	Directional source spectrum vector.
$\mathbf{P}_k$	Cross-spectrum matrix of directional sources.
$\mathbf{R}_k$	Spatial correlation matrix of $\mathbf{x}_k$.
$\overline{\mathbf{R}}_{k_0}$	Frequency-averaged $\mathbf{R}_k$ at the center frequency f_{k_0}.
$\overline{\mathbf{R}}^{+}_{k_0}$	$\overline{\mathbf{R}}_{k_0}$ after NSR.
$\mathbf{K}_k$	Spatial correlation matrix of $\mathbf{n}_k$.
λ_m	Eigenvalue.
$\mathbf{e}_m$	Eigenvector.
$c_{d,m}$	Normalized contribution.

Manuscript received November 23, 1998; revised December 13, 1999. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kuldip K. Paliwal.

F. Asano and S. Hayamizu are with Electrotechnical Laboratory, Tsukuba 305-8568, Japan (e-mail: asano@etl.go.jp).

T. Yamada is with Tsukuba University, Tsukuba 305-8568, Japan.

S. Nakamura is with ATR Spoken Language Translation Research Laboratories, Nara, Japan.

Publisher Item Identifier S 1063-6676(00)06982-0.

I. INTRODUCTION

WHEN applying automatic speech recognition (ASR) to a real environment, it is indispensable to reduce environmental noise to improve the rate of recognition. Various kinds of speech enhancement/noise reduction techniques have been studied for improving the signal-to-noise ratio (S/N) at the input of ASR. However, since the types of noise varies greatly according to the environment, no one speech enhancement technique is able to cover the whole range of noise.

Fig. 1 shows a rough classification of noise and the corresponding suitable speech enhancement methods. Speech enhancement techniques can be roughly divided into the multi-microphone approach and the single-microphone approach. The multi-microphone approach can further be divided into the spatial inverse type and the acoustic focus type. Fig. 2 shows a typical directivity pattern of these two types.

As depicted in Fig. 2, the spatial inverse approach forms a valley of sensitivity in a certain direction by means of adaptation/learning, and is thus suitable for directional interference. A conventional approach of this type is adaptive beamforming [1]. Currently, blind source separation is being intensively studied [2]. The performance of the reduction of directional interference is considered to be higher than in the other methods, regardless of source characteristics, as long as adaptation/learning is successful and the arrival directions of the interferences are different from that of the target signal.

On the other hand, the acoustic focus method steers a spatial acoustic focus to the target source while reducing the gain in the other directions as depicted in Fig. 2. Since this method has low gain in a wide range of directions, it is suitable for omni-directional or less-directional ambient noise. Delay-and-sum (DS) beamforming is the most widely used conventional method of this type. As compared with the adaptive beamformer, the (deterministically-designed) acoustic focus method usually shows better performance, mainly due to the difficulty in adaptation for the ambient noise in a real environment [3].

For omni- (or less-) directional ambient noise, a single-microphone speech enhancement technique can also be used. As compared with the DS beamformer with a relatively small-sized array (e.g., 50 cm in the largest dimension with 8 microphones as used in the experiment in this paper), conventional single-microphone methods such as the Wiener filter and spectral subtraction show comparable performance as long as the noise is stationary. In the single-microphone method, significant improvement has been made for nonstationary noise [4] (denoted as an advanced single-microphone method in Fig. 1). However, it is still difficult to cover all kinds of nonstationary noise. This is due to the fact that the single-microphone methods utilize *a priori* knowledge of the noise. The acoustic-focus-type method

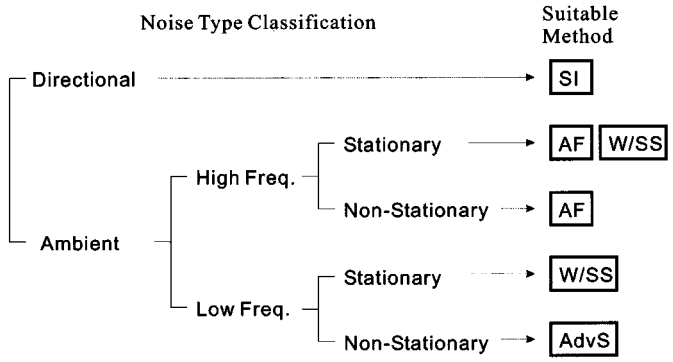

Fig. 1. Classification of noise types and the suitable noise reduction method. SI: spatial inverse; AF: acoustic focus; W/SS: Wiener filter and spectral subtraction; and AdvS: advanced single-microphone method.

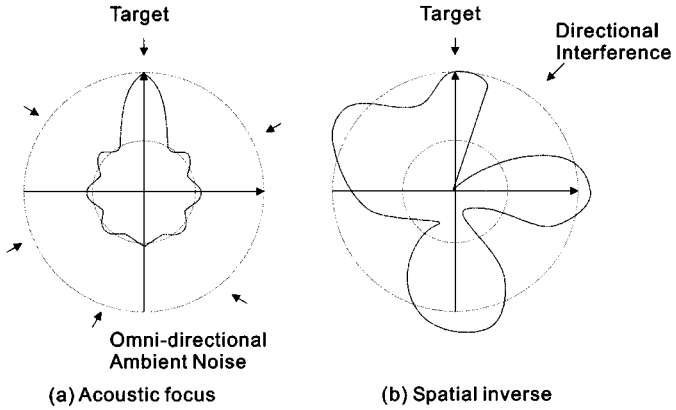

Fig. 2. Directivity pattern of the (a) acoustic-focus-type beamformer and (b) the spatial-inverse-type beamformer.

utilizes only the difference of the spatial characteristics of the signal and noise, and is effective for both stationary and non-stationary ambient noise. In this sense, the acoustic-focus-type method has an advantage over the single-microphone method. However, this advantage for the acoustic focus is limited to the higher frequency range (e.g., over around 1 kHz in the case of the above-mentioned array), due to the phase difference of the input signal being small in the lower frequencies. In the lower frequencies, even if multiple microphones are used, the system is essentially a single-microphone one.

The next step toward covering a wider range of noise types is possibly a combination of the different types of speech enhancement methods. The first attempt of such a combined approach in the field of array processing is a generalized sidelobe canceler (GSC), in which a DS beamformer and an adaptive spatial inverse filter are combined [5].

In this paper, an alternative approach of combining the spatial inverse and the acoustic focus method is proposed. To enhance the performance for omni-directional ambient noise, noise-dominant subspace reduction (NSR) proposed by the authors [6] is employed. In NSR, ambient noise is reduced by weighting eigenvalues of the spatial correlation matrix so that the noise-dominant subspace is reduced. This NSR is then combined with a minimum variance (MV) beamformer, which works as a spatial inverse filter and can extract an arbitrary directional component from the mixed signal. The MV beamformer is modified so that it works with NSR.

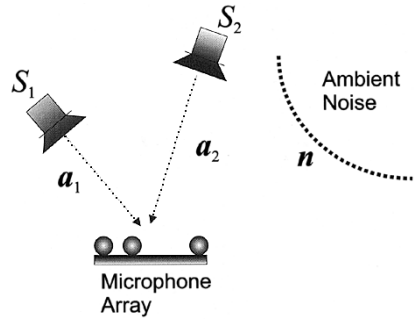

Fig. 3. Model of signal and noise.

The proposed method is based on the subspace method, which has been developed in the field of multi-sensor applications such as radar or sonar as a high resolution DOA (direction of arrival) estimator (e.g., [1]). A similar approach based on the subspace method has also been proposed for single-microphone speech enhancement by Ephraim *et al.* [7]. The proposed multi-microphone subspace method and the single-microphone subspace method utilize the same principle of the subspace method, but work in a different domain. In the multi-microphone method, a subspace corresponds to a certain physical space (spatial frequency region), while a subspace corresponds to a certain frequency region in the single-microphone method.

II. Model of Signal/Noise

Let us consider the acoustic environment as depicted in Fig. 3 where D directional signal/noise and omni-directional (or less-directional) ambient noise coexist. This sound field is observed by a microphone array with M microphones. The direct path from the dth sound source to the mth microphone has a transfer function with the following simple form:

$$A_{m,d}(k) = a_{m,d}(k)e^{-j\omega_k \tau_{m,d}} \tag{1}$$

where $a_{m,d}(k)$ denotes the gain. In an ideal case, $a_{m,d}(k) = 1$ in the far-field condition while $a_{m,d}(k) = 1/r_{m,d}$ for the near-field condition where $r_{m,d}$ is the distance between the dth sound source and the mth microphone. The symbol $\tau_{m,d}$ is the propagation time of sound from the dth sound source to the mth microphone. The symbol k denotes the discrete frequency index. By using the transfer function of the direct path, the input spectrum (Fourier transform of the input signal) observed at the mth microphone, $X_m(k)$, is then expressed as a sum of the D directional components plus ambient noise as

$$X_m(k) = \sum_{d=1}^{D} A_{m,d}(k)S_d(k) + N_m(k). \tag{2}$$

The symbol $S_d(k)$ denotes the spectrum of the dth source. The ambient noise term, $N_m(k)$, represents the sum of all the spectra except those of the direct sounds from the D point-sources. For example, reflection/reverberation of rooms and noise from sources that cannot be represented by the point source such as structural vibration are included in $N_m(k)$. For this kind of noise, the coherence of the input between the microphones is small. Therefore, these noises must be treated in a different way from that of the direct sound in the array processing.

By using the vector notation, (2) can be written as

$$\mathbf{x}_k = \sum_{d=1}^{D} \mathbf{a}_{d,k} S_d(k) + \mathbf{n}_k \tag{3}$$

where $\mathbf{x}_k = [X_1(k), \cdots, X_M(k)]^T$ is termed the input vector. The symbol $\cdot^T$ denotes the transpose. The directional vector is defined as $\mathbf{a}_{d,k} = [A_{1,d}(k), \cdots, A_{M,d}(k)]^T$. The noise vector is defined as $\mathbf{n}_k = [N_1(k), \cdots, N_M(k)]^T$. By using the notations, $\mathbf{A}_k = [\mathbf{a}_{1,k}, \cdots, \mathbf{a}_{D,k}]$ and $\mathbf{s}_k = [S_1(k) \cdots S_D(k)]^T$, (3) is further simplifed as

$$\mathbf{x}_k = \mathbf{A}_k \mathbf{s}_k + \mathbf{n}_k. \tag{4}$$

III. Subspace Method and Beamformer

In this section, signal processing tools used in this paper are briefly reviewed to facilitate understanding the following sections.

A. Subspace Method

Using the input vector $\mathbf{x}_k$, the spatial correlation matrix is defined as

$$\mathbf{R}_k = E[\mathbf{x}_k \mathbf{x}_k^H]. \tag{5}$$

The symbol $\cdot^H$ denotes the Hermitian transpose. Assuming that the directional components and the ambient noise are uncorrelated, $\mathbf{R}_k$ can be written using (4) as

$$\mathbf{R}_k = \mathbf{A}_k \mathbf{P}_k \mathbf{A}_k^H + \mathbf{K}_k. \tag{6}$$

Here, $\mathbf{P}_k$ is the cross-spectrum matrix for the directional sources defined as $\mathbf{P}_k = E[\mathbf{s}_k \mathbf{s}_k^H]$. The matrix $\mathbf{K}_k$ is the spatial correlation matrix of the ambient noise defined as $\mathbf{K}_k = E[\mathbf{n}_k \mathbf{n}_k^H]$.

Next, the generalized eigenvalue decomposition (GEVD, e.g., [8]) is applied to $\mathbf{R}_k$ as

$$\mathbf{R}_k \mathbf{E} = \mathbf{K}_k \mathbf{E} \mathbf{\Lambda}. \tag{7}$$

Here, the eigenvector matrix $\mathbf{E}$ consists of eigenvectors $\{\mathbf{e}_m\}$ as $\mathbf{E} = [\mathbf{e}_1, \cdots, \mathbf{e}_M]$. The eigenvalue matrix $\mathbf{\Lambda}$ has eigenvalues $\{\lambda_m\}$ on the diagonal elements as $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_M)$. Assuming that the power ratio of the directional components to the ambient noise (denoted as direct-ambient ratio) is high, the eigenvalues and eigenvectors have the following properties.

- *Property 1:* The energy of the D directional components is concentrated on the D largest eigenvalues.
- *Property 2:* The energy of the ambient noise equally spreads over all eigenvalues.
- *Property 3:* The eigenvectors $\{\mathbf{e}_1, \cdots, \mathbf{e}_D\}$ corresponding to the D largest eigenvalues become the orthonormal basis of $\Re(\mathbf{A}_k)$, where $\Re(\mathbf{A}_k)$ denotes the column space of $\mathbf{A}_k$.
- *Property 4:* The remaining eigenvectors $\{\mathbf{e}_{D+1}, \cdots, \mathbf{e}_M\}$ become the basis of $\Re(\mathbf{A}_k)^\perp$, where $\Re(\mathbf{A}_k)^\perp$ denotes the orthogonal complement of $\Re(\mathbf{A}_k)$.

The subspaces $\Re(\mathbf{A}_k)$ and $\Re(\mathbf{A}_k)^\perp$ are termed signal subspace and noise subspace, respectively. The reason for using GEVD instead of the standard eigenvalue decomposition $\mathbf{R}_k \mathbf{E} = \mathbf{E} \mathbf{\Lambda}$ is that GEVD diagonalizes $\mathbf{K}_k$ and flattens its eigenvalues (noise whitening) [9]. Therefore, GEVD guarantees Property 2.

B. Coherent Subspace Method

Usually, the spatial correlation matrix is estimated from the microphone array input by replacing the expectation operator $E[\cdot]$ of (5) with time averaging as

$$\mathbf{R}_k = \frac{1}{N} \sum_{n=n_1}^{n_2} \mathbf{R}_{k,n} \tag{8}$$

where $\mathbf{R}_{k,n} = \mathbf{x}_{k,n} \mathbf{x}_{k,n}^H$ and $\mathbf{x}_{k,n}$ denotes the input vector of the nth time frame. The symbols, $[n_1, n_2]$, denote the range of time averaging and $N = n_2 - n_1 + 1$. However, in frame-by-frame processing in ASR, sufficient data for estimating the statistically stable spatial correlation matrix are not available in each frame. Since the estimation of the stable spatial correlation matrix is a key factor in the eigenvalue decomposition in the following sections, the coherent subspace method (CSS) [10] is introduced.

In CSS, time-domain averaging is substituted for by frequency-domain averaging as

$$\overline{\mathbf{R}}_{k_0} = \frac{1}{K} \sum_{k=k_L}^{k_H} \mathbf{T}_k \mathbf{R}_{k,n} \mathbf{T}_k^H. \tag{9}$$

Here, the symbols, $[k_L, k_H]$, denote the range of frequency averaging and $K = k_H - k_L + 1$. The symbol k_o denotes the center frequency. For the sake of simplicity, the index for the time frame n is omitted hereafter. The matrix $\mathbf{T}_k$ is termed "focusing matrix," which executes the following rotation:

$$\mathbf{T}_k \mathbf{A}_k = \mathbf{A}_{k_0}. \tag{10}$$

The function of this rotation can be explained as follows: The directional vectors for the dth source at the different frequencies, $\mathbf{a}_{d,k}$ and $\mathbf{a}_{d,k_0}$, have different directions. The focussing matrix $\mathbf{T}_k$ rotates $\mathbf{a}_{d,k}$ so that $\mathbf{a}_{d,k}$ and $\mathbf{a}_{d,k_0}$ have the same direction for all $d = 1, \cdots, D$. By using this rotation and (6), (9) becomes

$$\overline{\mathbf{R}}_{k_0} = \mathbf{A}_{k_0} \overline{\mathbf{P}}_{k_0} \mathbf{A}_{k_0}^H + \overline{\mathbf{K}}_{k_0} \tag{11}$$

where

$$\overline{\mathbf{P}}_{k_0} = \frac{1}{K} \sum_{k=k_L}^{k_H} \mathbf{P}_k, \tag{12}$$

$$\overline{\mathbf{K}}_{k_0} = \frac{1}{K} \sum_{k=k_L}^{k_H} \mathbf{T}_k \mathbf{K}_k \mathbf{T}_k^H. \tag{13}$$

It can be seen from (11) that the same subspace structure as that in (6) is preserved after frequency averaging. Therefore, Properties 1–4 described in the previous section also hold for $\overline{\mathbf{R}}_{k_0}$. To obtain the focusing matrix $\mathbf{T}_k$, a least square approximation of (10), termed RSS focussing matrix [11], is employed in this paper.

C. Beamformer

The beamformer that extracts the power spectrum of the dth directional component, $P_{d,k}$, from the spatial correlation matrix has the form of

$$P_{d,k} = \mathbf{w}^H \mathbf{R}_k \mathbf{w}. \tag{14}$$

The symbol $\mathbf{w}$ is the coefficient vector. The most widely used beamformer of the acoustic-focus-type is the delay-and-sum (DS) beamformer. The coefficient vector of the DS beamformer which steers the acoustic focus to the dth directional component is given by

$$\mathbf{w}_{\mathrm{DS}} = \frac{\mathbf{a}_{d,k}}{\mathbf{a}_{d,k}^H \mathbf{a}_{d,k}}. \tag{15}$$

On the other hand, the minimum variance (MV) beamformer, which is an adaptive beamformer, is also widely used. The MV beamformer is derived as a result of constrained optimization, in which the all-pass characteristics in the direction of the target signal are the constraint. Under this constraint, the output power (variance) is minimized, resulting in noise reduction. When there are only directional components in the environment, the MV beamformer yields the spatial inverse as depicted in Fig. 2(b). When there is ambient noise together with the directional components, the MV beamformer yields the directivity of a mixture of acoustic focus and the spatial inverse. The coefficients of the MV beamformer are given by

$$\mathbf{w}_{\mathrm{MV}} = \frac{\mathbf{R}_k^{-1} \mathbf{a}_{d,k}}{\mathbf{a}_{d,k}^H \mathbf{R}_k^{-1} \mathbf{a}_{d,k}}. \tag{16}$$

The derivation of these beamformers in detail can be found in textbooks such as [1].

IV. Speech Enhancement Method

A. Reduction of the Noise-Dominant Subspace (NSR)

In this section, a method of reducing ambient noise by manipulating the eigenvalues of the spatial correlation matrix is described. Let us denote the eigenvalue matrix and the eigenvector matrix of $\overline{\mathbf{R}}_{k_0}$ as $\overline{\mathbf{E}} = [\overline{\mathbf{e}}_1, \cdots, \overline{\mathbf{e}}_M]$ and $\overline{\mathbf{\Lambda}} = \mathrm{diag}(\overline{\lambda}_1, \cdots, \overline{\lambda}_M)$, respectively. These are derived from GEVD of $\overline{\mathbf{R}}_{k_0}$ as $\overline{\mathbf{R}}_{k_0}\overline{\mathbf{E}} = \overline{\mathbf{K}}_{k_0}\overline{\mathbf{E}}\,\overline{\mathbf{\Lambda}}$. Reduction of the energy of the ambient noise in $\overline{\mathbf{R}}_{k_0}$ is realized by weighting the eigenvalues so that the noise-dominant eigenvalues are reduced as

$$\overline{\mathbf{\Lambda}}^{+} = \mathrm{diag}\left(g_1\overline{\lambda}_1, \cdots, g_M\overline{\lambda}_M\right) \tag{17}$$

where $\mathbf{g} = [g_1, \cdots, g_M]$ is the weights. Using $\overline{\mathbf{\Lambda}}^{+}$ and $\overline{\mathbf{E}}$, the spatial correlation matrix is then reconstructed as

$$\overline{\mathbf{R}}_{k_0}^{+} = \overline{\mathbf{K}}_{k_0}\overline{\mathbf{E}}\,\overline{\mathbf{\Lambda}}^{+}\overline{\mathbf{E}}^{-1}. \tag{18}$$

Next, how to determine the weights, $\mathbf{g} = [g_1, \cdots, g_M]$, is described. When the direct-ambient ratio is high, the energy of the directional components is concentrated on the D largest eigenvalues while the rest of the eigenvalues contain the energy of the ambient noise only as described in Properties 1 and 2. In this case, the energy of the ambient noise in the noise subspace can be reduced by simply discarding the M–D smallest eigenvalues. This is realized by the following weights

$$\mathbf{g} = \left[\overbrace{1, \cdots, 1}^{D}, \overbrace{0, \cdots, 0}^{M-D}\right] \tag{19}$$

assuming that the eigenvalues are sorted in descending order.

However, Properties 1 and 2 hold only when the direct-ambient ratio is high, i.e., over 0 dB. When the direct-ambient ratio is low, it is no longer guaranteed that the energy of the directional components is concentrated on the D largest eigenvalues, and the energy of the directional components might leak to the other eigenvalues. In this case, "noise-dominant" subspace is identified by using the following projection. The projection of the dth directional vector $\mathbf{a}_{d,k_0}$ onto each eigenvector is

$$\mathbf{p}_{d,m} = [\mathbf{a}_{d,k_0}^H \overline{\mathbf{e}}_m]\overline{\mathbf{e}}_m = C_{d,m}\overline{\mathbf{e}}_m. \tag{20}$$

Here, the coefficient $C_{d,m} = \mathbf{a}_{d,k_0}^H \mathbf{e}_m$ is the contribution of $\mathbf{a}_{d,k_0}$ to the mth subspace. Let us define the normalized contribution, $c_{d,m} = C_{d,m}/\max(C_{d,1}, \cdots, C_{d,M})$, where $\max(\cdots)$ gives the maximum element. If $c_{d,m} \ll 1$, not much of the energy of the dth directional component is contained in the mth subspace. In this case, this mth subspace is expected to be ambient-noise-dominant. Based on this, the weight vector is determined as

$$g_m = 1, \quad \text{when } c_{d,m} \geq c_{thr} \tag{21}$$

for all $1 \leq d \leq D$ and $1 \leq m \leq M$ with the initialization of $\mathbf{g} = [0, \cdots, 0]$. The symbol c_{thr} (≤ 1) is an arbitrary threshold. By using this manipulation, the subspaces where the normalized contribution is smaller than the threshold c_{thr} is discarded. The number of the finally-adopted subspaces, i.e., the number of "1" in $\mathbf{g}$, is denoted as $L(\leq M)$. How to determine the threshold c_{thr} is discussed in a later section based on the simulation.

B. Estimation of the Source Power Spectrum

In this section, an arbitrary directional component is extracted from the ambient-noise-reduced correlation matrix, $\overline{\mathbf{R}}_{k_0}^{+}$, by the beamformer. In the same manner as (14), the band-averaged power spectrum of the dth directional component is estimated from $\overline{\mathbf{R}}_{k_0}^{+}$ by

$$\overline{P}_{d,k_0} = \mathbf{w}^H \overline{\mathbf{R}}_{k_0}^{+} \mathbf{w}. \tag{22}$$

When there is single directional component, i.e., $D = 1$, DS beamformer can be used as $\mathbf{w}_{\mathrm{DS}} = \mathbf{a}_{k_0}[\mathbf{a}_{k_0}^H \mathbf{a}_{k_0}]^{-1}$.

When there are multiple directional components, the dth directional component (target) can be extracted by using the MV beamformer. In the conventional MV beamformer (16), the spatial correlation matrix $\mathbf{R}_k$ is used to derive the coefficient vector $\mathbf{w}_{\mathrm{MV}}$. However, in this paper, the spatial correlation matrix was modified by CSS and NSR in the previous sections, and, therefore, $\mathbf{R}_k$ cannot be used for the proposed method. The next option for the spatial correlation matrix in the MV beamformer is $\overline{\mathbf{R}}_{k_0}^{+}$, which was processed with CSS and NSR. However, the problem of using $\overline{\mathbf{R}}_{k_0}^{+}$ is that the number of averaging is insufficient (see Appendix B for details.) In this paper, therefore, the

following virtual correlation matrix $\mathbf{Q}_{k_0}$, which consists of the estimated directional vector $\mathbf{A}_{k_0}$, is used.

Assuming that NSR is successful, only directional components remain (or are dominant) in $\overline{\mathbf{R}}_{k_0}^{+}$. The spatial correlation matrix of the directional components can be written using the directional vector as

$$\mathbf{Q}_{k_0} = \mathbf{A}_{k_0} \mathbf{\Sigma}_{k_0} \mathbf{A}_{k_0}^{H} \tag{23}$$

where $\mathbf{\Sigma}_{k_0} = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_M^2)$ is the virtual cross-spectrum matrix of the directional sources. For the purpose of deriving the MV beamformer that extracts a certain directional component, this virtual cross-spectrum matrix does not have to reflect the real cross-spectrum matrix, $\overline{\mathbf{P}}_{k_0}$, which is unknown and is to be estimated. Instead, $\mathbf{\Sigma}_{k_0}$ can be arbitrarily chosen according to the desired directivity pattern. A practical choice of $\mathbf{\Sigma}_{k_0}$ is $\mathbf{\Sigma}_{k_0} = \mathbf{I}$ where $\mathbf{I}$ is an identity matrix. By using this virtual correlation matrix, the coefficient vector becomes

$$\mathbf{w}_{\mathrm{MV}} = \frac{\mathbf{Q}_{k_0}^{-1} \mathbf{a}_{d,k_0}}{\mathbf{a}_{d,k_0}^{H} \mathbf{Q}_{k_0}^{-1} \mathbf{a}_{d,k_0}}. \tag{24}$$

Equation (24) yields the spatial inverse filter which passes the dth directional component while reducing the other $D-1$ directional components. This results in directivity similar to that in Fig. 2(b), in which all-pass characteristics are set in the direction of the dth source while nulls are placed in the directions of the other $D-1$ sources. However, the directivity in the directions other than these D sources is indefinite. As long as NSR is successful, this indefinite directivity makes no difference in the output. However, when NSR is imperfect, the residual of the ambient noise remains in $\overline{\mathbf{R}}_{k_0}^{+}$ and might be amplified by the MV beamformer if the directivity except the D directions is indefinite. This can be prevented by adding an omni-directional noise term to the virtual correlation matrix as

$$\mathbf{Q}'_{k_0} = \mathbf{Q}_{k_0} + \gamma^2 \mathbf{I} \tag{25}$$

where $\gamma^2\mathbf{I}$ is the correlation matrix of the virtual omni-directional noise. The parameter γ^2 is the power of the virtual noise and is arbitrarily chosen according to the desired directivity. The value of γ^2 is discussed in the following experiments. Using $\mathbf{Q}'_{k_0}$, the final modified MV beamformer coefficient vector is

$$\mathbf{w}_{\mathrm{MV}} = \frac{\mathbf{Q}'^{-1}_{k_0} \mathbf{a}_{d,k_0}}{\mathbf{a}_{d,k_0}^{H} \mathbf{Q}'^{-1}_{k_0} \mathbf{a}_{d,k_0}}. \tag{26}$$

C. Entire System

Fig. 4 shows a block diagram of the proposed system. In 4(a), the band-averaged spatial correlation matrix, $\overline{\mathbf{R}}_{k_0}$, is estimated using CSS. In 4(b), $\overline{\mathbf{R}}_{k_0}$ is decomposed into subspaces using the generalized eigenvalue decomposition. In 4(c), the directional vectors $\mathbf{A}_{k_0}$ and the number of directional sources D are estimated using a subspace method such as MUSIC [12] and rank analysis [13]. In the later modules, 4(d) and (e), these estimates, $\hat{\mathbf{A}}_{k_0}$ and $\hat{D}$, are used instead of the unknown true $\mathbf{A}_{k_0}$ and D. In 4(d), the energy of the ambient noise is reduced by NSR. Finally, in 4(e), the band-averaged power spectrum of each directional component is estimated by the MV (or DS) beamformer. The estimated band-averaged power spectrum is then transformed to the mel-frequency cepstrum coefficients (MFCC, e.g., [14]) and is used as a feature vector of ASR.

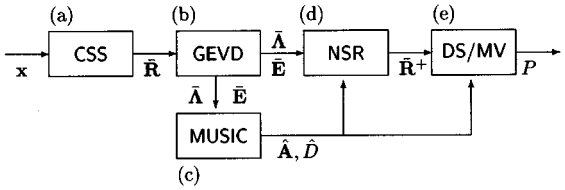

Fig. 4. Block diagram of the proposed method.

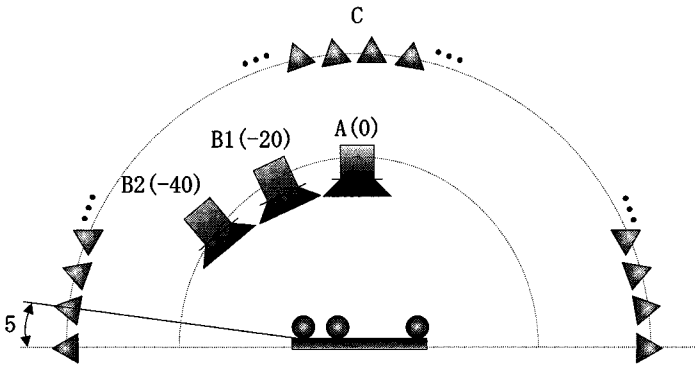

Fig. 5. Configuration of the microphone array and the sound sources in a simulated environment.

V. Experiment I: Model Environment

In this section, the basic characteristics of the proposed method are investigated using a simulated model environment.

A. Conditions

The simulated microphone array was linearly configured with $M = 8$. The interval of the microphones was 6 cm. In the model environment, the directional sources A, B1/B2 and the ambient noise sources C exist as depicted in Fig. 5. As the ambient noise C, a mixture of independent noise coming from $-90°$ to $+90°$ at every 5° was employed to simulate omni-directional noise. As a sound field, far-field condition with no reflection was assumed. The broad-side of the array corresponds to the front (0°). From source A, a speech signal (Japanese words, 1–2 s in duration) was emitted. From source B1/B2, either noise or speech was emitted. As noise, white noise or pink noise (low-frequency dominant, spectral gradient of -6 dB/Oct.) was employed. As a speech signal, the same words as source A but in a different order and by a different speaker were employed. The following two cases were investigated: 1) [directional speech A + ambient noise C]; 2) [directional speech A + directional noise (or speech) B1/B2 + ambient noise C].

The parameters of NSR are summarized in Table I. Since MFCC is employed as a feature vector in ASR, the parameters in CSS, k_0, k_L, and k_H are determined so that the bandwidth of the subbands of CSS is equal in the mel-frequencies. Thus, the number of averages, K, increases with increasing frequency from 10 at the lowest k_0 to 50 at the highest k_0. As the ambient noise correlation matrix, $\overline{\mathbf{K}}_{k_0} = \mathbf{I}$ is employed since ambient

TABLE I
PARAMETERS IN NSR

Parameters	Value
Frame length	512 points (32 ms)
Frame interval	160 points (10 ms)
FFT length	1024 points
Number of subbands	25

TABLE II
PARAMETERS IN ASR

Feature vector	12th order MFCC + 12th order Δ MFCC
ASR	HMM (tied-mixture)
Number of states	3
Number of models	42
Training data	1050 words × 10 subjects
Test data	492 words × 1 subject

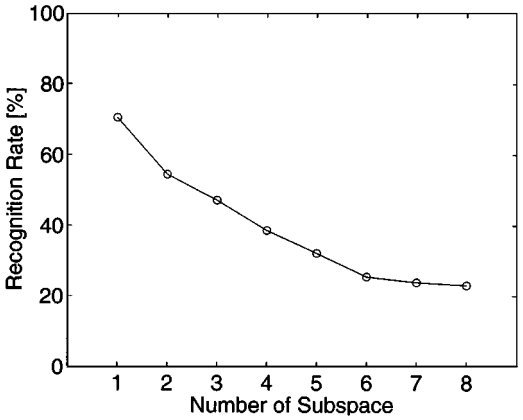

Fig. 6. Recognition rate when the number of the adopted subspaces, L, is varied. The model environment. Case 1) [A + C].

noise C is almost omni-directional and the correlation of the inputs between the microphones is small. In the MV beamformer, $\gamma^2 = ||\mathbf{Q}|| \times 10^{-6}$, where $||\cdot||$ denotes the 2-norm of the matrix. The parameters of ASR are summarized in Table II.

B. Results 1: Single Directional Speech + Ambient Noise

First of all, case 1) [A + C] was investigated. The number of the directional sources was $D = 1$. Fig. 6 shows the recognition rate when the number of the adopted subspaces L was varied. This was a preliminary experiment for determining the value of the threshold c_{thr}. The weight was set to $g_m = 1$ for the L largest normalized contribution c_m. S/N (the power ratio of the target A to the ambient noise C) = 0 dB. S/N was calculated as the ratio of the vowel portion of the speech and the noise. The average power ratio of whole words to noise is roughly 10 dB lower than this S/N. The noise source was white noise. As the spectrum estimator, DS was employed. The combination

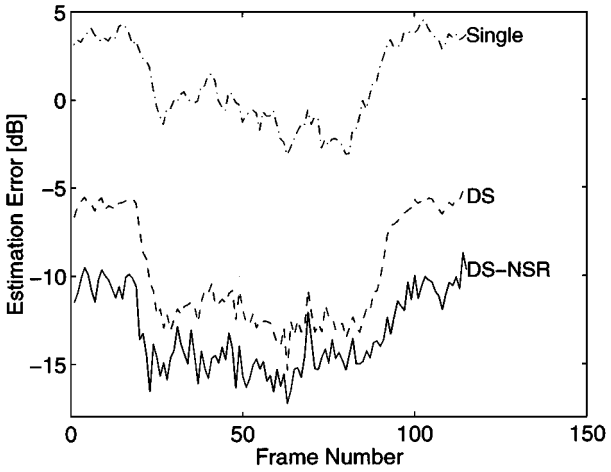

Fig. 7. Estimation error of the band power spectrum of the target speech. Case 1) [A + C]. Model environment. Input S/N = 0 dB.

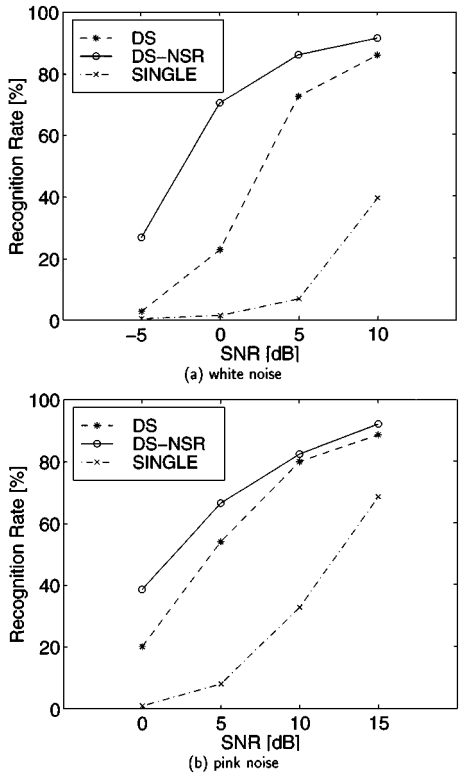

Fig. 8. Recognition rate for different S/N. Case 1) [A + C]. Model environment.

of DS and NSR is referred to as DS-NSR hereafter. When $L = M = 8$, NSR is not conducted. In this case, DS and DS-NSR are equivalent. From this figure, it can be seen that the recognition rate was improved as L decreased. Especially, when $L = 1$, the recognition rate was improved by 47.5%. Based on this result, only the subspace corresponding to the largest contribution was adopted in the succeeding experiments. This is realized by the threshold, $c_{thr} = 1$.

Fig. 8 shows the recognition rate as a function of S/N. For the sake of comparison, the results for DS and the single-microphone are also shown. From Fig. 8(a), which shows the results

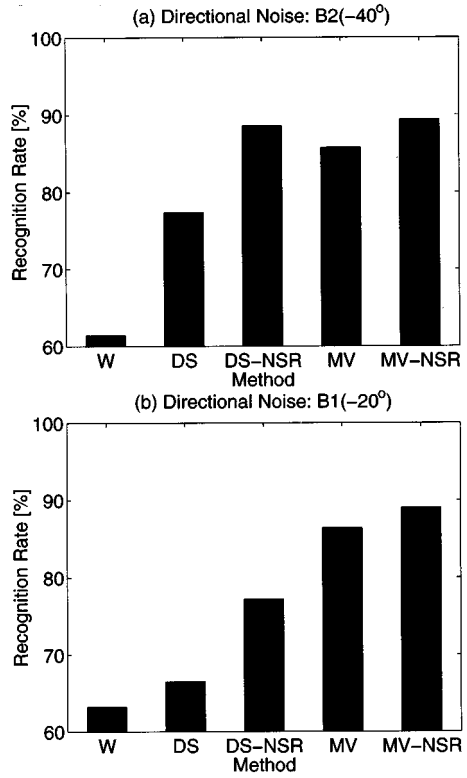

Fig. 9. Recognition rate for case 2) [A + B1/B2 + C]. The directional noise source used is (a) B2 (−40°) and (b) B1 (−20°). Model environment.

TABLE III
RECOGNITION RATE FOR THE SIGNAL SEPARATION EXPERIMENT IN SECTION V-D.

S/N	MV		MV-NSR	
	A	B1	A	B1
5 dB	66.9	71.5	72.3	78.0
10 dB	81.1	86.6	81.5	87.2

for white noise, improvement at S/N = 0 dB is the most significant. Fig. 7 shows the estimation error of the band power spectrum defined as

$$\mathrm{ERR}(n) = \sum_{k_0} |\overline{P}_1(n, k_0) - \overline{P}_1^t(n, k_0)| / E[\overline{P}_1^t(n, k_0)]$$

where $\overline{P}_1(n, k_0)$ and $\overline{P}_1^t(n, k_0)$ are respectively the estimated and the true band power spectrum at the nth frame. The estimation error corresponding to the case of S/N = 0 dB is depicted. From this figure, the estimation error was reduced for DS-NSR by around 5 dB relative to DS.

In the case of the pink noise shown in the Fig. 8(b), the recognition rate for both DS-NSR and DS decreased. This is due to the physical limitation of the acoustic-focus-type method described in the introduction. Comparison of DS-NSR and DS shows an improvement by DS-NSR. However, the rate of the improvement is smaller than that for the white noise.

C. Results 2: Single Directional Speech + Single Directional Noise + Ambient Noise

Next, case 2) [A + B1/B2 + C] was investigated. As a spectrum estimator, both DS and MV were tested. The noise source was white noise for both B1/B2 and C. The power of B1/B2 and C relative to A is 0 dB and −10 dB, respectively. Fig. 9 shows the recognition rate. For the sake of comparison, a single-microphone Wiener filter was also applied (denoted as W in Fig. 9). In the Wiener filter, the spectrum of speech was recovered by using the input of one of the microphones as

$$S_1(k) = \{[X_1^2(k) - E[Z_1^2(k)]]/X_1^2(k)\}^{-1/2} X_1(k)$$

where $Z_1(k)$ is the mixed spectrum of B1/B2 and C at the microphone #1. Since the noise B1/B2 and C is stationary, $E[Z_1^2(k)]$ was calculated in advance.

When employing either B1 and B2, MV-NSR showed the best scores. Thus, the MV-NSR approach is suitable for a mixed environment of directional components and ambient noise. In detail, DS-NSR and MV-NSR show similar scores for B2 (−40°), while MV-NSR is superior to DS-NSR for B1(−20°). The reason for MV-NSR showing better scores for B1 is that the location of the target source A and that of the directional noise source B1 are close, and a part of the energy of the directional noise is within the mainlobe of the DS beamformer, while a spatial null is placed in the direction of B1 for MV. Therefore, MV-NSR is especially useful when the directional noise source is located close to the target source.

D. Results 3: Two Directional Speech Signals + Ambient Noise

In this section, the ability of signal separation using the MV beamformer was tested. The experimental setup was [A + B1 + C]. Different from the previous section, two independent speech signals were emitted from sources A and B1. The two speech signals had equal power. The ambient noise level was −5 dB and −10 dB relative to speech. Two sets of MV beamformer coefficient vectors

$$\mathbf{w}_{\mathrm{MV},1} = \mathbf{Q}'^{-1}_{k_0} \mathbf{a}_{1,k_0} / [\mathbf{a}^H_{1,k_0} \mathbf{Q}'^{-1}_{k_0} \mathbf{a}_{1,k_0}]^{-1}$$

and

$$\mathbf{w}_{\mathrm{MV},2} = \mathbf{Q}'^{-1}_{k_0} \mathbf{a}_{2,k_0} / [\mathbf{a}^H_{2,k_0} \mathbf{Q}'^{-1}_{k_0} \mathbf{a}_{2,k_0}]^{-1}$$

were used to extract the spectra of the two speech signals. Table III shows the recognition rate of each speech signal. From this, it can be seen that the two speech signals were well separated in terms of the ASR rate by the proposed MV-NSR method.

VI. EXPERIMENT II: REAL ENVIRONMENT

In this section, the proposed method is applied to a real environment where diffused noise (room reverberation) exists as ambient noise.

A. Conditions

The room used in the experiment is a meeting room, the size of which is 8.3 m × 7.2 m × 3.2 m. The reverberation time was

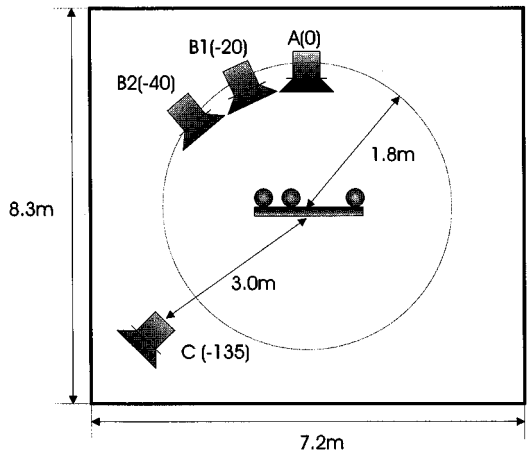

Fig. 10. Configuration of the microphone array and the sound sources in a real environment.

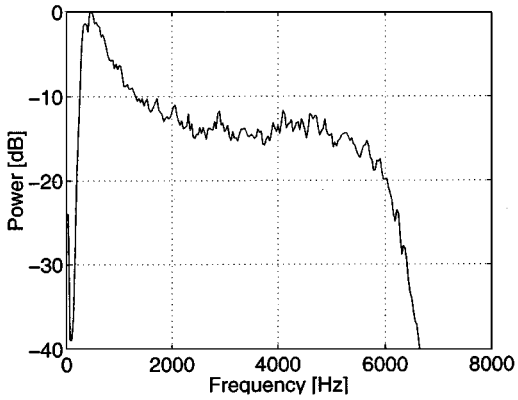

Fig. 11. Spectrum of elevator noise.

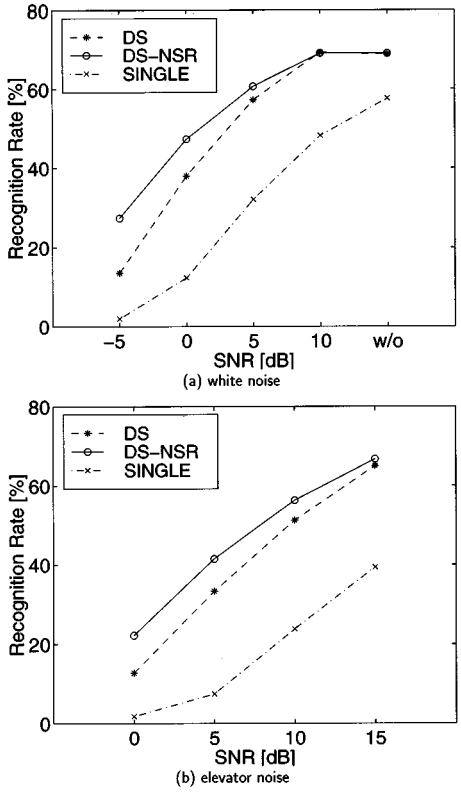

Fig. 12. Recognition rate for different S/N. Case 1) [A + C]. Real environment.

0.42 s. Target source A, directional noise source B1/B2, and ambient noise source C were located as depicted in Fig. 10. To simulate ambient noise, source C was placed facing a corner of the room. Then, the impulse responses from these sources to the microphones were measured. The microphone input was generated by convolving these impulse responses with the source signals. In the impulse responses from the ambient noise source C, direct sound was eliminated to generate diffused noise. The noise sources used were white noise and the noise of an elevator (low-frequency dominant). The spectrum of the elevator noise is shown in Fig. 11. The parameters of NSR and ASR are the same as those of the previous section. The following two cases were investigated: 1) [directional speech A + ambient noise C]; 2) [directional speech A + directional noise B1].

As the ambient noise correlation matrix in the GEVD block, $\overline{\mathbf{K}}_{k_0} = \mathbf{I}$ was employed in the same manner as in the model environment. In real ambient noise, coherence between the microphone inputs exists to some extent, resulting in $\mathbf{K}_k \neq \mathbf{I}$. However, as for the band-averaged correlation matrix $\overline{\mathbf{K}}_{k_0}$ estimated by CSS, the inter-microphone coherence is reduced by the focussing matrix $\mathbf{T}_k$ as long as the arrival direction of the ambient noise is different from the focussing angle, resulting in $\overline{\mathbf{K}}_{k_0} \simeq \mathbf{I}$.

B. Results 1: Single Directional Speech + Ambient Noise

First, case 1) [A + C] was tested. Fig. 12(a) shows the results for the white-noise as an ambient noise source. S/N = w/o corresponds to the case when ambient noise C does not exist. Even in this case, the recognition rate was around 70% for both DS-NSR and DS. This is due to the reverberation of the directional speech. The reason for there being no improvement for DS-NSR as compared with DS is that the ratio of the direct sound to the reverberation is high (over 0 dB) in this case. The reverberation for the directional speech is included in ambient noise that arrives from many directions. Therefore, the eigenvalues for the reverberation have a relatively flat distribution. However, as described in Appendix A, when the direct-ambient ratio is high, the noise subspace reduction is implicitly included in the DS process. Therefore, in this case, DS-NSR and DS are equivalent. As S/N decreased, the recognition rate was improved for DS-NSR. Here, S/N is defined as the ratio of the power of the direct sound of the speech to that of ambient noise C.

In the case where elevator noise was employed as the ambient noise source, the recognition rate was much reduced compared with the case in which white noise was used. This is due to the fact that a large portion of noise energy was concentrated in the low frequencies as shown in Fig. 11. Therefore, in this case, a Wiener filter was further applied to the low frequency range of the output of the array processing to reduce the low frequency component of the noise. The setup of the Wiener filter was the same as that of Section V-C. The range for application of the Wiener filter was the lower 10 mel-frequency bands with a center frequency of 0 to 1388 Hz. This range was determined so that the recognition score was the highest. Fig. 12(b) shows the results for the array signal processing + the Wiener filter. As

shown by this figure, an improvement was found for DS-NSR as compared with DS.

C. Results 2: Single Directional Speech + Single Directional Noise

Next, case 2) [A + B1] was investigated. Though ambient noise source C was not employed, the reverberation for A and B1 existed as natural ambient noise. Therefore, this is a case where directional noise and real ambient noise coexist.

Fig. 13 shows the directivity pattern of the MV beamformer. Fig. 13(a) is the case when $\gamma^2 = \|\mathbf{Q}\| \times 0.03$ while Fig. 13(b) is the case when $\gamma^2 = \|\mathbf{Q}\| \times 0.3$. In 13(a), a deep valley appeared in the direction of B1 while an increase in the gain was found in the lower frequencies in directions other than A and B1. On the other hand, in 13(b), the valley in the direction of B1 is shallower while the increase in the gain in the low frequency is relatively small.

Fig. 14 shows the recognition rate. In this figure, MV1 and MV2 correspond to the case Fig.13(a) $\gamma^2 = \|\mathbf{Q}\| \times 0.03$ and 13(b) $\gamma^2 = \|\mathbf{Q}\| \times 0.3$, respectively. As can be seen from this figure, MV2-NSR achieves the best score. The tendency for MV-NSR to be the best among DS, DS-NSR, MV and MV-NSR is the same as that of the simulation since noise source B1 is close to target source A. As compared with MV1-NSR, MV2-NSR shows a better score. This result shows that, in the case tested in this paper, the directivity with the undesired increase in gain being small is better even when the valley in the directional noise source is shallower.

VII. Conclusion

A method of speech enhancement with noise-dominant subspace reduction and the MV beamformer was proposed and evaluated in this paper. In this method, less-directional ambient noise is eliminated in the subspace domain by reducing the noise-dominant eigenvalues. Then the remaining mixture of the multiple directional components is decomposed into single component corresponding to each sound source by the modified MV beamformer.

From the results of the evaluation experiment with a model environment, it can be seen that the recognition rate was significantly improved by NSR for omni-directional ambient noise. For a mixture of ambient noise and directional interference, the combination of NSR and the MV beamformer (MV-NSR) was effective, especially when the directional interference was located close to the target source. MV-NSR was also able to separate the mixture of multiple speech signals in the presence of ambient noise.

In the experiment in a real environment, an improvement similar to that of the model environment was found. However, the highest performance achieved was around 20% lower than that in the model environment. This is mainly due to the presence of room reverberation. Room reverberation is modeled as $\overline{\mathbf{K}}_{k_0}$ in NSR with the assumption that the ambient noise is independent of the directional components and is stationary. However, the reverberation in the real environment has coherence with the directional components to some extent and dynamically changes its characteristics. This is a problem which should be solved in future work.

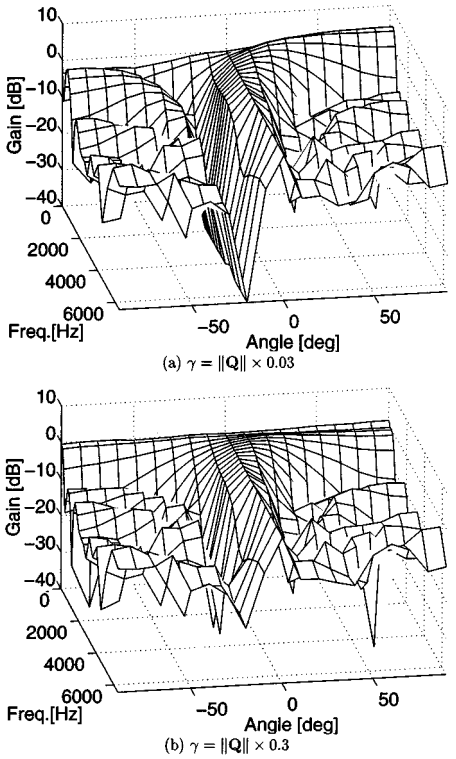

Fig. 13. Directivity pattern of MV beamformer.

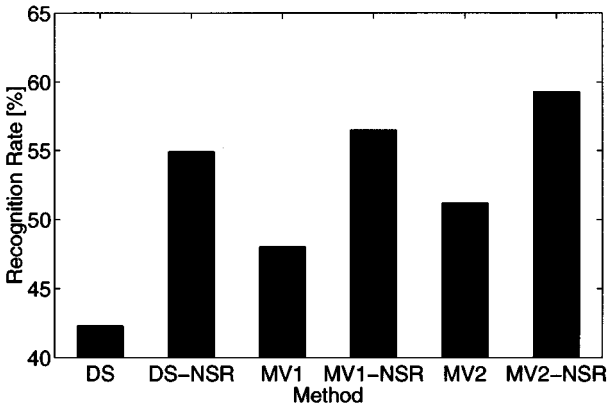

Fig. 14. Recognition rate for case 2) [A + B1]. MV1: $\gamma^2 = \|\mathbf{Q}\| \times 0.03$. MV2: $\gamma^2 = \|\mathbf{Q}\| \times 0.3$. Real environment.

As for the frequency range, NSR was found to be effective for the higher frequency range. This is a physical limitation due to the array configuration used in this paper. For low-frequency-dominant noise, a single-microphone Wiener filter was experimentally employed in the lower frequency region together with the proposed method and was found to be effective. As described in the introduction, each speech enhancement method has its own "territory," and to cover the wide variety of noise, the integration of these methods might be effective. However, the exchange of information between the different speech enhancement modules may pose a problem. For example, to combine array processing with the single-microphone method, the

information of the residual noise after the array processing must be sent to the single-microphone method. In the particular case reported in this paper, information exchange was not necessary since the territory of NSR and the Wiener filter (limited to the low frequency region) did not overlap. However, information exchange will be necessary, especially for the advanced single-microphone method for nonstationary noise, and is considered to be another challenging issue in the field of speech enhancement.

APPENDIX A
RELATION OF NSR AND DS

Let us consider the case of $D = 1$ for the sake of simplicity. When the direct-ambient ratio is high, from Property 4

$$\mathrm{span}(\mathbf{a}_k) = \mathrm{span}(\mathbf{e}_1) = \mathrm{span}(\mathbf{e}_2, \cdots, \mathbf{e}_M)^{\perp}. \tag{27}$$

Using the eigenvalues and eigenvectors of $\mathbf{R}_k$ and assuming that $\mathbf{K}_k = \mathbf{I}$, $\mathbf{R}_k$ can be written as

$$\mathbf{R}_k = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^H + \cdots + \lambda_M \mathbf{e}_M \mathbf{e}_M^H. \tag{28}$$

Using (28) and (15), the estimated spectrum (14) can be written as

$$P_{d,k} = \frac{1}{|\mathbf{a}_k^H \mathbf{a}_k|^2} \{\lambda_1 |\mathbf{a}_k^H \mathbf{e}_1|^2 + \cdots + \lambda_M |\mathbf{a}_k^H \mathbf{e}_M|^2\}. \tag{29}$$

From the orthogonality (27)

$$|\mathbf{a}_k^H \mathbf{e}_m|^2 = 0, \qquad \text{for } m = 2, \cdots, M. \tag{30}$$

This means that the energy in $\mathrm{span}(\mathbf{e}_2, \cdots, \mathbf{e}_M)$ is reduced and, thus, is equivalent to NSR with $L = 1$. Therefore, when the direct-ambient ratio is high, the reduction of noise subspace is implicitly included in the DS beamformer. On the other hand, when the direct-ambient ratio is low and (27) does not hold, the "noise-dominant" subspaces are not perfectly eliminated. In NSR, these noise-dominant subspaces are forcibly eliminated. Since the noise-dominant subspace might include a portion of the target energy, NSR may cause distortion in the estimated target spectrum $P_{d,k}$. Therefore, there is a trade-off between high noise reduction rate and small distortion in estimating $P_{d,k}$. This trade-off should be taken into account depending on the application.

APPENDIX B
USE OF $\mathbf{Q}'_{k_0}$ IN (26)

The correlation matrix processed with CSS and NSR, $\overline{\mathbf{R}}^{+}_{k_0}$, is estimated in every time frame. The number of averaging is $K = 10 - 50$ in this paper as indicated in Section V-A. However, for the purpose of deriving the MV beamformer coefficients, much greater averaging is required, especially when the target signal coexists with the other directional interferences. This is mainly due to the cross terms of the target and the other directional components such as $E[S_1(k)S_2(k)]$ not being zero in $\overline{\mathbf{R}}^{+}_{k_0}$ [15]. These cross terms are theoretically zero if the target and the other directional components are mutually independent. Fig. 15 shows the response of the MV beamformer in the direction of the directional interference derived using $\overline{\mathbf{R}}^{+}_{k_0}$ under the existence of the target (curve A). In comparison with the response using $\overline{\mathbf{R}}^{+}_{k_0}$ without the target (curve B), the reduction in gain of curve A is much smaller. Curve C shows the response with the virtual correlation matrix $\mathbf{Q}'_{k_0}$, which exhibits higher performance.

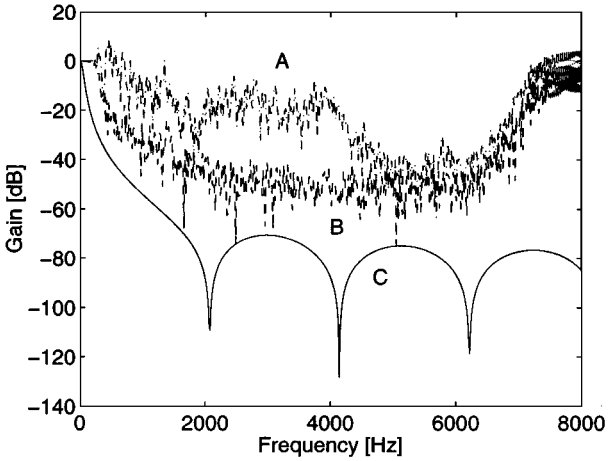

Fig. 15. Frequency response of the MV beamformer in the direction of the directional noise. (a) Designed with the correlation matrix estimated in the presence of both the target speech and the directional noise; (b) designed with the correlation matrix estimated in the presence of the directional noise only; and (c) designed with the virtual correlation matrix $\mathbf{Q}'_{k_0}$.

REFERENCES

[1] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[2] T.-W. Lee, *Independent Component Analysis*. Norwell, MA: Kluwer, 1998.
[3] L. J. Griffiths and K. M. Buckley, "Quiescent pattern control in linearly constrained adaptive arrays," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-35, pp. 917–926, July 1987.
[4] H. Sameti, H. Sheikhzadeh, L. Deng, and R. Brennan, "HMM-based strategies for enhancement of speech signals in nonstationary noise," *IEEE Trans. Speech, Audio Processing*, vol. 6, pp. 445–455, Sept. 1998.
[5] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagat.*, vol. AP-30, pp. 27–34, Jan. 1982.
[6] F. Asano and S. Hayamizu, "Speech enhancement using array signal processing based on the coherent-subspace method," *IEICE Trans. Fundamentals*, vol. E80-A, pp. 2276–2285, Nov. 1997.
[7] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech, Audio Processing*, vol. 3, pp. 251–266, July 1995.
[8] G. Strang, *Linear Algebra and Its Application*. Orlando, FL: Harcourt Brace Jovanovich, 1988.
[9] R. Roy and T. Kailath, "ESPRIT—Estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 37, pp. 984–995, July 1989.
[10] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 823–831, Apr. 1985.
[11] H. Hung and M. Kaveh, "Focussing matrices for coherent signal-subspace processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1272–1281, Aug. 1988.
[12] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. AP-34, pp. 276–280, Mar. 1986.
[13] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 387–392, Apr. 1985.
[14] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan, 1993.

[15] F. Asano, Y. Suzuki, and T. Sone, "Convergence characteristics of the adaptive array using RLS algorithm," *IEICE Trans. Fundamentals*, vol. E80-A, no. 1, pp. 148–158, January 1997.

Futoshi Asano (M'95) received the B.S. degree in electrical engineering and the M.S. and Ph.D. degrees in electrical and communication engineering from Tohoku University, Sendai, Japan, in 1986, 1988, and 1991, respectively.

From 1991 to 1995, he was a Research Associate at RIEC, Tohoku University. From 1993 to 1994, he was a Visiting Researcher at ARL, Pennsylvania State University, University Park. Since 1995, he has been with the Electrotechnical Laboratory, Tsukuba, Japan, where he is currently a Senior Researcher. His research interests include array signal processing, adaptive signal processing, neural network, statistical signal processing, and speech recognition.

Satoru Hayamizu (M'89) received the B.E., M.E., and Dr.E. degrees from Tokyo University, Tokyo, Japan.

Since 1981, he has been working on speech recognition, spoken dialogue, and communication with artifacts at the Electrotechnical Laboratory, Tsukuba, Japan. From 1989 to 1990, he was a Visiting Scholar at Carnegie Mellon University, Pittsburgh, PA, and in 1994 a Visiting Scientist at LIMSI/CNRS.

Dr. Hayamizu is a member of the Institute of Electronics Information and Communication Engineers, the Acoustical Society of Japan, the Japanese Society for Artificial Intelligence, the Association for Natural Language Processing, the Japan Society of Mechanical Engineers, and the International Speech Communication Association.

Takeshi Yamada (M'00) was born in Osaka, Japan, on February 13, 1971. He received the B.Eng. degree from Osaka City University in 1994, and the M.Eng. and Dr.Eng. degrees from Nara Institute of Science and Technology, Nara, Japan, in 1996 and 1999, respectively.

Since 1999, he has been with the Institute of Information Sciences and Electronics, University of Tsukuba, Tsukuba, Japan, where he is a Assistant Professor. His research interest include robust speech recognition, sound scene recognition, microphone array signal processing, and sound field control and reproduction.

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

Satoshi Nakamura (M'90) was born in Japan on August 4, 1958. He received the B.S. degree in electronics engineering from Kyoto Institute of Technology in 1981 and the Ph.D. degree in information science from Kyoto University in 1992.

From 1981 to 1986 and 1990 to 1993, he was with Central Research Laboratory, Sharp Corporation, Nara, Japan, where he was engaged in speech recognition researches. From 1986 to 1989, he was a Researcher with the Speech Processing Department, ATR Interpreting Research Laboratories. From 1994 to 2000, he was an Associate Professor with the Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan. In 1996, he was a Visiting Research Professor with the CAIP Center, Rutgers University, New Brunswick, NJ. He is currently Head of Speech Processing Department, ATR Spoken Language Translation Research Laboratories, Japan. His current research interests include speech recognition, speech translation, spoken dialogue systems, stochastic modeling of speech and microphone array.

Dr. Nakamura received the Awaya Award from the Acoustical Society of Japan in 1992. He is a member of the Acoustical Society of Japan and the Information Processing Society of Japan.