Stereophonic Noise Reduction Using a Combined Sliding Subspace Projection and Adaptive Signal Enhancement

Tetsuya Hoya, *Member, IEEE*, Toshihisa Tanaka, *Member, IEEE*, Andrzej Cichocki, *Member, IEEE*, Takahiro Murakami, Gen Hori, and Jonathon A. Chambers, *Senior Member, IEEE*

***Abstract*—A novel stereophonic noise reduction method is proposed. This method is based upon a combination of a subspace approach realized in a sliding window operation and two-channel adaptive signal enhancing. The signal obtained from the signal subspace is used as the input signal to the adaptive signal enhancer for each channel, instead of noise, as in the ordinary adaptive noise canceling scheme. Simulation results based upon real stereophonic speech contaminated by noise components show that the proposed method gives improved enhancement quality in terms of both segmental gain and cepstral distance performance indices in comparison with conventional nonlinear spectral subtraction approaches.**

***Index Terms*—Sliding subspace projection, speech enhancement, stereophonic noise reduction.**

I. Introduction

IN THE LAST few decades, noise reduction has been a topic of great interest in speech enhancement. One of the classical and most commonly used methods is based upon nonlinear spectral subtraction (NSS) [1]–[5]. In NSS methods, both the speech and noise spectra of the noisy speech data are independently estimated by using sample statistics obtained over some number of frames, and then noise reduction is performed by subtracting the spectrum of the noise from that of the observed data.

Due to the block processing based approach, however, it is well known that such methods introduce annoying artifacts, which are often referred to as undesirable "musical tone", in the enhanced speech. Moreover, in many cases, such methods also remove some speech components in the spectra which are fundamental to the intelligibility of the speech. This is a particular problem at lower SNR's. The performance is also quite dependent on the choice of many parameters, such as, spectral subtraction floor, over-subtraction factors, or over-subtraction corner frequency parameters. To find the optimal choice of these parameters in practice is therefore very difficult.

Recently, in the study of blind signal processing, one of the most active potential application areas has been speech separation [6] and a number of methods for blind separation/deconvolution of speech have been developed [7]–[10]. These methods work quite well when each sensor is located close to each source. However, separation of the speech from noise is still difficult when all the sensors are located close to one dominant source but far from the others, as in cocktail party situations. This sensor configuration is typically employed in practice, for example, as in stereo conferencing systems; two microphones being placed in front of the speaker at a reasonable distance. Moreover, the existing blind separation/deconvolution methods quite often fail to work where there are more sources than sensors.

In contrast, in the study of biomedical engineering, it has been reported that the utility of the subspace method implemented using the singular value decomposition (SVD) is to successfully enhance nonaveraged data (e.g., [11], [12]). In the technique, the space of the observed data is partitioned into signal and noise subspaces. Elimination of the noise is thereby achieved by orthonormal projection of the observed signal onto the signal subspace, with the assumption that the signal and noise subspaces are orthogonal.

In recent study, a number of SVD based methods have also been developed for speech enhancement [13]–[19]. For instance, a Toeplitz (or Hankel) structured data matrix representation is employed within the subspace decomposition operation, and thereby the data matrix is decomposed into signal-plus-noise subspace and a noise subspace rather than signal and noise subspaces (see [14], [15], and [19]). However, little attention has generally been paid to the extension to multichannel outputs.

In this paper, we propose a novel multichannel signal enhancement scheme using a combination of a subspace estimation method and a multichannel adaptive signal enhancement (ASE) approach in order to tackle the aforementioned problems. The objective of the approach is to estimate the received signal at the sensors after the removal of noise components from each, instead of recovering/extracting the original source signals. In the special case where the number of sensors is exactly two, it is then considered that the problem is to recover the stereophonic signal from the two channel noisy observations.

In the proposed method, rather than requiring additional microphones to provide separate noise references, a sliding subspace projection (SSP) is used, which operates as a sliding-win-

Manuscript received February 13, 2003; revised February 19, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Futoshi Asano.

T. Hoya, T. Tanaka, and G. Hori are with the Laboratory for Advanced Brain Signal Processing, Saitama 351-0198, Japan.

A. Cichocki is with the Laboratory for Advanced Brain Signal Processing, Saitama 351-0198, Japan, on leave from the Warsaw University of Technology, Warsaw, Poland.

T. Murakami is with the Department of Electronics and Communication Engineering, Meiji University, Kawasaki, Japan.

J. A. Chambers is with the Cardiff School of Engineering, Cardiff University, Cardiff CF24 OYF, U.K.

Digital Object Identifier 10.1109/TSA.2005.845792

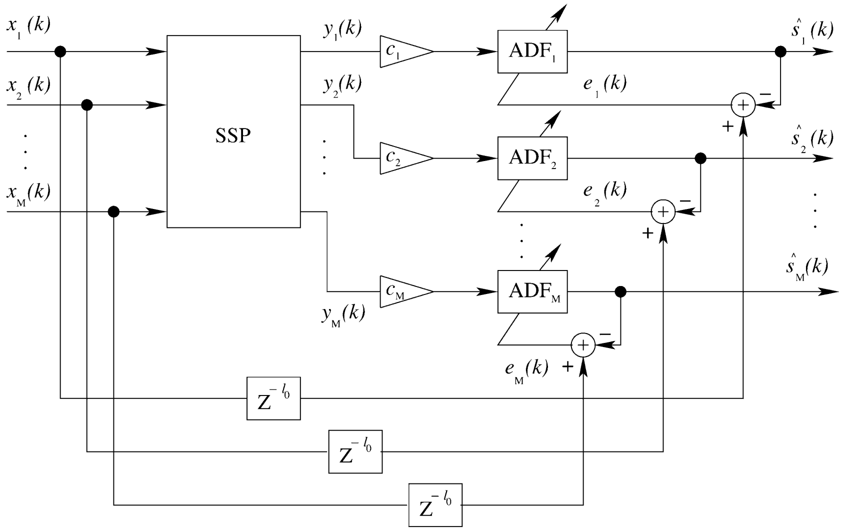


Fig. 1. Block diagram of the proposed multichannel noise reduction system.

dowed subspace noise reduction processor, in order to extract the source signals for a bank of adaptive signal enhancers. In stereophonic situations, the role of the SSP is to extract the (monaural) source signal.

For the actual signal enhancement, a bank of modified adaptive signal (line) enhancers is used. For each channel, the enhanced signal obtained from the SSP is given to the adaptive filter as the source signal for the compensation of the stereophonic image. The philosophy of this approach is that the quality of the outputs of the SSP will be improved by the adaptive filters.

In [21], a similar approach by integrating Wiener filtering and a single stage SVD was proposed and applied to the recovery of evoked potentials. In that method, a Wiener filtering approach is firstly implemented to extract the overall shape of the evoked potentials and then SVD is used to enhance the filtered version of the raw data. The method, however, requires a relatively large number of sensors and will not work efficiently when the number of the sensors is very small (e.g., only two microphones are available). In speech enhancement, this limits its utility in practical situations, e.g., stereophonic noise reduction. By virtue of the SSP as preprocessing, the proposed method is thus advantageous in this respect.

II. Multichannel Signal Enhancement by a Combination of an SSP and ASE

In the general case of an array of sensors, the M-channel observed sensor signals $x_i(k)$ $(i = 1, 2, \dots, M)$ can be written in the form

$$x_i(k) = s_i(k) + n_i(k), \quad (i = 1, 2, \dots, M) \tag{1}$$

where $s_i(k)$ and $n_i(k)$ are respectively the target and noise components within the observation $x_i(k)$.

The block diagram of the proposed multichannel noise reduction system is illustrated in Fig. 1. In the figure, $y_i(k)$ denotes the i-th signal obtained from the SSP, and $\hat{s}_i(k)$ is the i-th enhanced version of the target signal $s_i(k)$.

In this paper, we assume that the target signals $s_i(k)$ are speech signals arriving at the respective sensors and that the noise process is zero-mean, additive, and uncorrelated with the speech signals. Thus, under the assumption that $s_i(k)$ are generated from one single speaker, it can be considered that the speech signals $s_i(k)$ are strongly correlated to each other and thus that we can exploit the property of the strong correlation for noise reduction by a subspace method.

In other words, we can reduce the additive noise by projecting the observed signal onto the subspace of which the energy of the signal is mostly concentrated. The problem here, however, is that, since speech signals are usually nonstationary processes, the correlation matrix can be time-variant. Moreover, it is considered that the subspace projection reduces the dimensionality of the signal space, e.g., a stereophonic signal pair can be reduced to a monaural signal.

To solve these problems, we thus propose to use a combined subspace projection operated within a sliding-window and signal enhancers realized by adaptive filters. The former technique can estimate the correlation matrices adaptively, whereas the latter expands the reduced space into the original whole signal space again.

A. The Subspace Projection for Noise Reduction

The subspace projection of a given signal data matrix contains information about the signal energy, the noise level, and the number of sources. By using a subspace projection, it is thus possible to divide approximately the observed noisy data into the subspaces of the signal of interest and the noise [21]–[23]. A summary of the noise reduction technique using the subspace projection is given as follows:

Let $\boldsymbol{X}$ be the available data in the form of an $L \times M$ matrix

$$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_M] \tag{2}$$

where the column vector $\boldsymbol{x}_i$ $(i = 1, 2, \dots, M)$ is written as

$$\boldsymbol{x}_i = [x_i(0), x_i(1), \dots, x_i(L-1)]^T \quad (T : \text{transpose}). \tag{3}$$

Then, the eigenvalue decomposition (EVD) of the autocorrelation matrix of $\boldsymbol{X}$ (for $M < L$) is given by

$$\boldsymbol{X}^T \boldsymbol{X} = \boldsymbol{V} \boldsymbol{\Sigma} \boldsymbol{V}^T \tag{4}$$

where the matrix $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_M] \in \Re^{M \times M}$ is orthogonal such that $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_M$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_M) \in \Re^{M \times M}$, with eigenvalues $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_M \geq 0$. The columns in $\boldsymbol{V}$ are the eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$. The eigenvalues in $\boldsymbol{\Sigma}$ contain some information about the number of signals, signal energy, and the noise level. It is well known that if the signal-to-noise ratio (SNR) is sufficiently high (e.g., see [12]), the eigenvalues can be ordered in such a manner as

$$\sigma_1 > \sigma_2 > \cdots > \sigma_s \gg \sigma_{s+1} > \sigma_{s+2} \cdots > \sigma_M \tag{5}$$

and the autocorrelation matrix $\boldsymbol{X}^T\boldsymbol{X}$ can be decomposed as

$$\boldsymbol{X}^T\boldsymbol{X} = [\boldsymbol{V}_s \quad \boldsymbol{V}_n] \begin{bmatrix} \boldsymbol{\Sigma}_s & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Sigma}_n \end{bmatrix} [\boldsymbol{V}_s \quad \boldsymbol{V}_n]^T \tag{6}$$

where $\boldsymbol{\Sigma}_s$ contains the s largest eigenvalues associated with s signals with the highest energy (i.e., $\sigma_1, \sigma_2, \ldots, \sigma_s$) and $\boldsymbol{\Sigma}_n$ contains $(M - s)$ eigenvalues $(\sigma_{s+1}, \sigma_{s+2}, \ldots, \sigma_M)$. It is then considered that $\boldsymbol{V_s}$ contains s eigenvectors associated with the signal part, whereas $\boldsymbol{V}_n$ contains $(M - s)$ eigenvectors associated with the noise. The subspace spanned by the columns of $\boldsymbol{V}_s$ is thus referred to as the signal subspace, whereas that spanned by the columns of $\boldsymbol{V}_n$ corresponds to the noise subspace.

Then, the signal and noise subspace are mutually orthogonal and orthonormally projecting the observed noisy data onto the signal subspace leads to noise reduction. The data matrix after the noise reduction $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_M]$, where $\boldsymbol{y}_i = [y_i(0), y_i(1), \ldots, y_i(L-1)]^T$, is given by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{V}_s\boldsymbol{V}_s^T \tag{7}$$

which describes the orthonormal projection onto the signal space.

This approach is quite beneficial to practical situations, since we do not need to assume/know in advance the locations of the noise sources.

For instance, in stereophonic situations, since both the speech components $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ are strongly correlated with each other, even if the rank is reduced to one for the noise reduction purpose (i.e., by taking only the eigenvector corresponding to the eigenvalue with the highest energy σ_1), it is still possible to recover $\boldsymbol{s}_i$ from $\boldsymbol{y}_i$ by using adaptive filters as the post-processors to be described in Section II-C.

B. Sliding Subspace Projection

As in Fig. 1, the SSP acts as a sliding-window noise reduction block. To illustrate the difference between the SSP and the conventional frame-based operation (e.g., see [21], [22]), Fig. 2 is given. In the figure, $\boldsymbol{x}$ is a row vector of the autocorrelation matrix in (2), i.e., $\boldsymbol{x}(k) = [x_1(k), x_2(k), \ldots, x_M(k)]$ $(k = 0, 1, 2, \ldots)$.

Then, given the previous L past samples for each channel at time instance k $(\geq L)$ and using (7), the new input matrix to the SSP $\boldsymbol{X}(k)$ $(L \times M)$ can be written

$$\boldsymbol{X}(k) = \begin{bmatrix} \boldsymbol{P}\boldsymbol{X}(k-1)\boldsymbol{V}_s(k-1)\boldsymbol{V}_s(k-1)^T \\ \boldsymbol{x}(k) \end{bmatrix}, \quad \boldsymbol{P} = [\boldsymbol{0}_{(L-1)\times 1}\boldsymbol{I}_{L-1}], \quad (L-1 \times L) \tag{8}$$

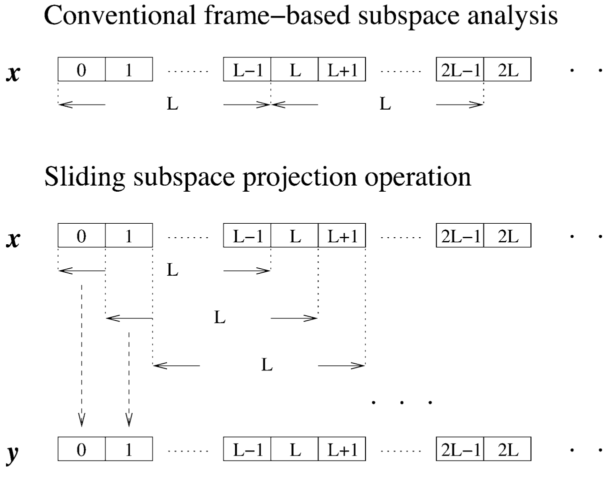

Conventional frame–based subspace analysis
x
0
1
L–1
L
L+1
2L–1
2L
L
L
Sliding subspace projection operation
x
0
1
L–1
L
L+1
2L–1
2L
L
L
L
y
0
1
L–1
L
L+1
2L–1
2L

Fig. 2. Illustration of an SSP operation.

TABLE I
VALUES FOR p_1, $p_{2,1}$, AND $p_{2,2}$

Speech Sample	p_1	$p_{2,1}$	$p_{2,2}$
No.1	116	308	322
No.2	134	346	345
No.3	130	317	330

TABLE II
PARAMETERS USED FOR NSS

Setting Name	NSS1	NSS2	NSS3
Smoothing Time Constant for Signal Power Estimate	0.04	1	0.01
Subtraction Floor	0.02	0.2	0.01
Oversubtraction Corner Frequency	800	1000	600
Oversubtraction Scale Factor	4	5	3
Smoothing Time Constant for Signal Power-Estimate in Noise Estimation	0.1	0.1	0.1
FFT Window Length	0.032	0.032	0.032
Length of Minimum Filter	1.5	1.5	1.5
Time Constant for Oversubtraction Factor	0.04	0.04	0.04
Num. of Minimization Buffers	4	4	4
Oversampling Constant	4	4	4
Noise Estimate Compensation	1.5	1.5	1.5

where $\boldsymbol{V}_s(k)$ denotes the signal subspace matrix obtained at time instance k and

$$\boldsymbol{X}(0) = \begin{bmatrix} \boldsymbol{0}_{(L-1)\times M} \\ \boldsymbol{x}(0) \end{bmatrix}. \tag{9}$$

Then, the first row of the new input matrix $\boldsymbol{X}(k)$ given in (8) corresponds to the M-channel signals after the SSP operation $\boldsymbol{y}(k) = [y_1(k), y_2(k), \ldots, y_M(k)]^T$

$$\boldsymbol{y}(k) = \boldsymbol{X}(k)^T\boldsymbol{q}, \quad \boldsymbol{q} = [1, 0, 0, \ldots, 0]^T \quad (L \times 1). \tag{10}$$

Note that in (8) the first $(L-1)$ rows of $\boldsymbol{X}(k)$ are obtained from the previous SSP operation, whereas the last row is taken

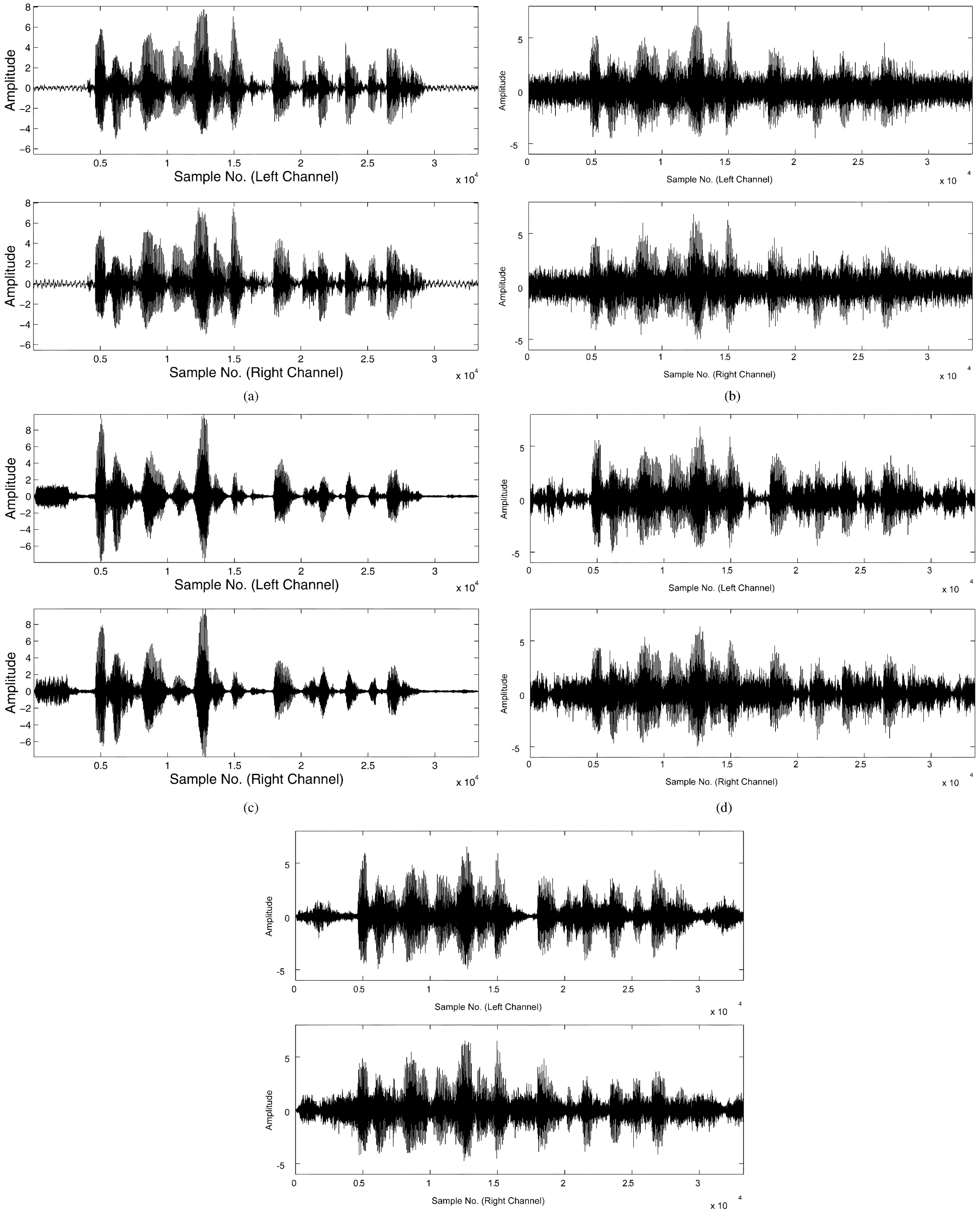
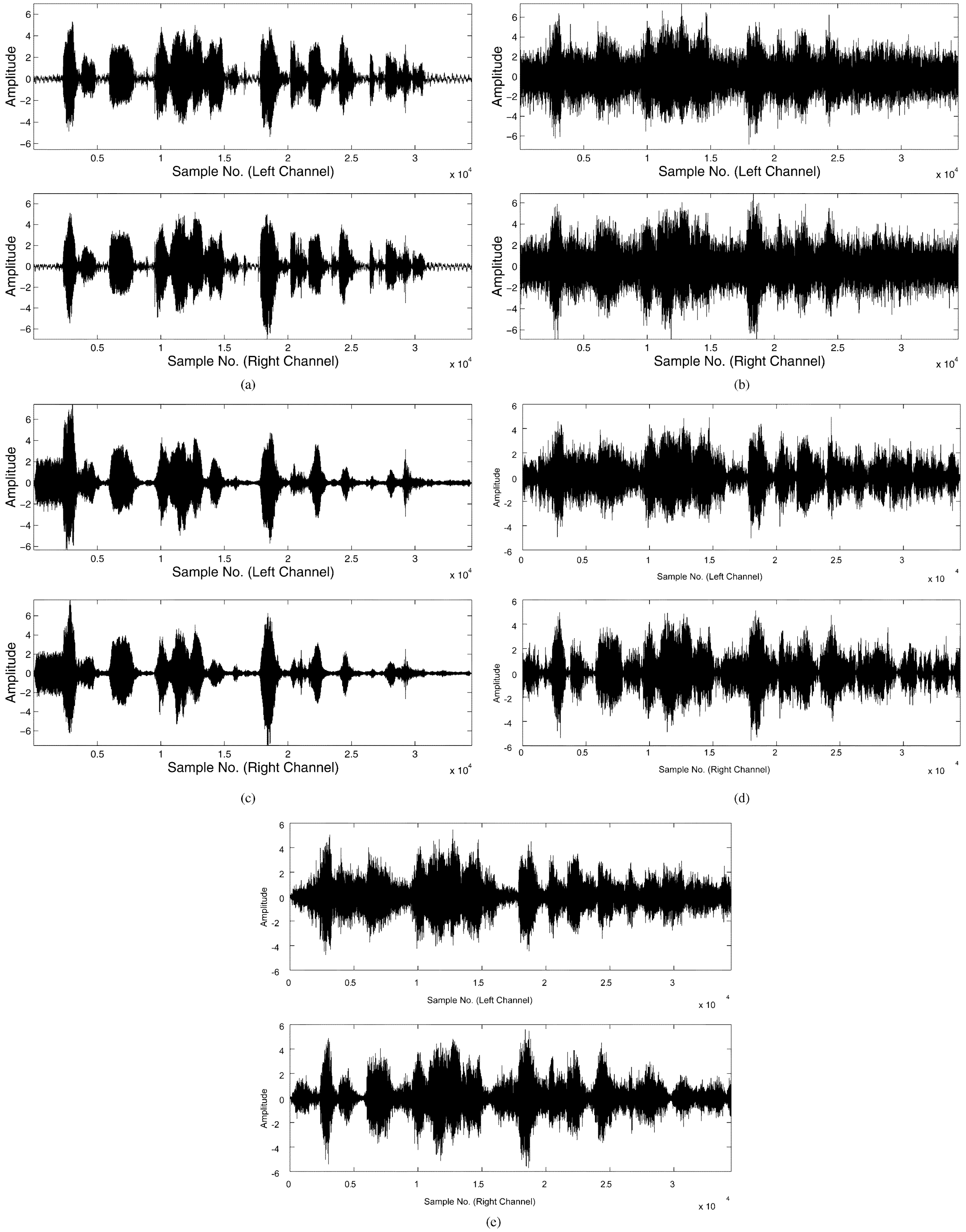


Fig. 3. Simulation results—using speech sample no. 1 (with two additive i.i.d. noise case). (a) Clean speech. (b) Noisy data $(\mathrm{SNR} = 3\ \mathrm{dB})$. (c) Enhanced speech by NSS. (d) Enhanced speech by SSP. (e) Enhanced speech by $\mathrm{SSP} + \mathrm{DASE}$.

from the data obtained from the original observation. Then, at this point, as in Fig. 2, the new data $\boldsymbol{x}(k)$ remains intact and the rest $(L-1)$ data vectors, i.e., those obtained by the product $\boldsymbol{PX}(k)$, will be replaced by the subsequent subspace projection operations. It is thus considered that this recursive operation is similar to the concept of data-reusing [24] or fixed point iteration [25] in which the input data at the same data point is repeatedly used for, i.e., improving the convergence rate in adaptive algorithms. Related to the subspace based noise reduction as a sliding window operation, it has been shown that a truncated SVD operation is identical to an array of analysis-synthesis finite impulse response (FIR) filter pairs connected in parallel



(a)

(b)

(c)

(d)

(e)

Fig. 4. Simulation results—using speech sample no. 2 (with two additive i.i.d. noise case). (a) Clean speech. (b) Noisy data $(\mathrm{SNR} = 0\ \mathrm{dB})$. (c) Enhanced speech by NSS. (d) Enhanced speech by SSP. (e) Enhanced speech by $\mathrm{SSP} + \mathrm{DASE}$.

[26]. It is then expected that this approach still works when the number of the sensors M is small, as in ordinary stereophonic situations.

In addition, we can intuitively justify the effectiveness of using the SSP as follows: for large noise and very limited numbers of samples (this choice must, of course, relate to the stationarity of the noise), a single SSP (sliding window) operation may perform only rough or approximate decomposition to both the signal and noise subspace. In other words, we are not able to ideally decompose the noisy sensor vector space into a signal subspace and its noise counterpart within a distinct frame. In one single frame, we rather perform decomposition into a signal-plus-noise subspace and a noise subspace [14].

C. Multichannel Adaptive Signal Enhancement

After the extraction of each signal, a multichannel adaptive signal enhancer (ASE) is used to enhance the observed signal.

Since the respective input signals to the signal enhancer are strongly correlated with the corresponding signals of interest, the i-th adaptive filter functions to recover the original signal in each channel from the signal $y_i(k)$ using the delayed version of the reference signal $x_i(k - l_0)$. In the diagram in Fig. 1, the delay factor l_0 is given by

$$l_0 = \frac{L_f - 1}{2} \tag{11}$$

where L_f is the length of each adaptive filter. The insertion of a delay factor is necessary in order to shift the center lag of the reference signals in not only the positive but also the negative time direction by the adaptive filters.

This scheme is then somewhat related to direction of arrival (DOA) estimation using adaptive filters [28] and similar to adaptive line enhancers (ALE, see e.g., [20]). However, unlike an ordinary ALE, the reference signal in each channel is not taken from the original input but the observation $x_i(k)$ and the input signal to the adaptive filter is the delayed version of the original input signal, as in Fig. 1. Moreover, as we elucidate in the context of stereophonic noise reduction described in Section II-D, the role of the adaptive filters is different from the DOA; it is considered that the adaptive filters are always adjusting the essential elements with respect to the recovery of the stereophonic image, e.g., both the delay and amplitude in one channel against the other.

In addition, in Fig. 1, c_i are appropriately chosen constants and used to adjust the scaling of the corresponding input signals to the adaptive filters. These scaling factors are normally necessary since the choice will affect the initial tracking ability of the adaptive algorithms in terms of stereophonic compensation and may be determined *a priori* with keeping a good-trade off between the initial tracking performance and the signal distortion.

Eventually, as in Fig. 1, the enhanced signal $\hat{s}_i(k)$ is obtained simply from the filter output.

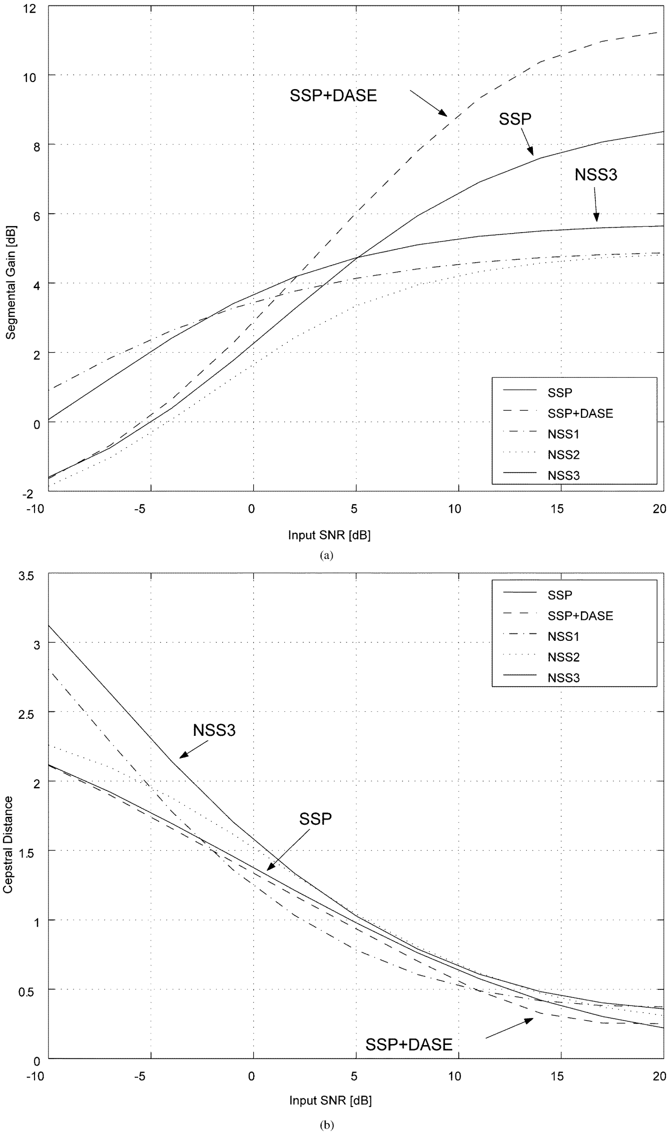
D. Stereophonic Noise Reduction

In this paper, the following model is considered as the two-channel observation $x_i(k)$ $(i = 1, 2)$:

$$\begin{aligned} x_1(k) &= a \cdot s_1(k) + n_1(k), \\ x_2(k) &= a \cdot s_2(k) + n_2(k) \end{aligned} \tag{12}$$

where $s_1(k)$ and $s_2(k)$ respectively correspond to the left and right channel speech signal, $n_1(k)$ and $n_2(k)$ are the noise components, and the constant 'a' controls the input SNR.

In the above, the number of the sources can be seen to be four; two stereophonic speech components and the other two for the noise sources, while the number of the sensors is assumed to be two $(M = 2)$, as in stereophonic representative of many teleconferencing systems.



(a)

(b)

Fig. 5. Performance comparison using noise components modeled by two i.i.d. normally distributed random signals. (a) Comparison of the segmental gain (b) Comparison of the cepstral distance.

Hence, this seems to be really problematic since "There are more sources than sensors." However, in stereophonic noise reduction, the components $s_i(k)$ can be approximated by

$$\begin{aligned} s_1(k) &= \boldsymbol{h}_1^T(k)\boldsymbol{s}(k), \\ s_2(k) &= \boldsymbol{h}_2^T(k)\boldsymbol{s}(k) \end{aligned} \tag{13}$$

where $\boldsymbol{h}_i(k) = [h_i(0), h_i(1), \ldots, h_i(L_s - 1)]^T$ $(i = 1, 2)$ are the impulse response vectors of the acoustic transfer functions between the signal (speech) source and the microphones with length L_s, and $\boldsymbol{s}(k) = [s(k), s(k-1), \ldots, s(k - L_s + 1)]^T$ is the speech source signal vector.

Therefore, it is considered that the respective stereophonic speech components $s_i(k)$ $(i = 1, 2)$ are generated from one speech source using two (sufficiently long) filters $\boldsymbol{h}_i$ and, in reality, the stereophonic speech components $s_i(k)$ are strongly correlated with each other.

In the SSP described earlier, the orthonormal projection of each observation $x_i(k)$ onto the estimated signal subspace by

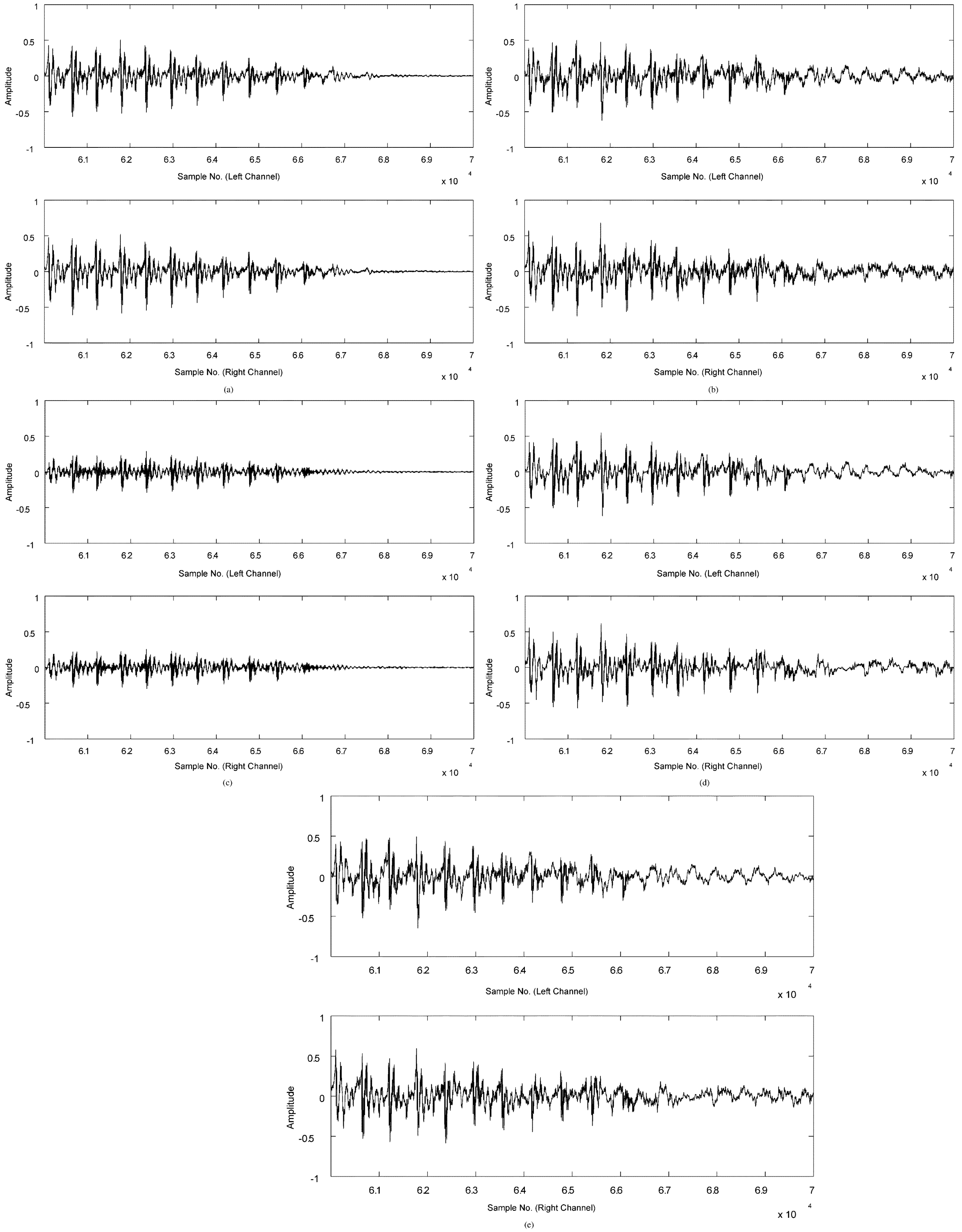


Fig. 6. Simulation results—using speech sample no. 4 (sampled at 48 (kHz), with the real stereophonic fan noise components). (a) Clean speech. (b) Noisy data $(\mathrm{SNR} = 2\ \mathrm{dB})$. (c) Enhanced speech by NSS. (d) Enhanced speech by only SSP. (e) Enhanced speech by $\mathrm{SSP} + \mathrm{DASE}$.

the SSP leads to reduction of the noise in each channel. However, since the projection is essentially performed using only a single orthonormal vector which corresponds to the speech source, this may cause distortion of the stereophonic image in the extracted speech signals $y_1(k)$ and $y_2(k)$. In other words, the SSP is performed to recover a single speech source from the two observations $x_i(k)$.

In the proposed method, the adaptive signal enhancers are thus employed in order to compensate for the stereophonic image. Since, as in the block diagram in Fig. 1, the error signals $e_i(k)$ $(i = 1, 2)$ contain the information about the stereophonic image (because the observations $x_i(k)$ contain true stereophonic signals), the adaptive filters (with sufficient filter lengths) essentially adjust the delay and the amplitude of the signal in each channel, both of which are of fundamental to recover the stereophonic sound, and therefore are considered to compensate for the stereophonic image in each channel.

III. Simulation Study

A. Parameter Settings

For the speech components $s_i(k)$, four stereophonically recorded speech data were used. For the first three, the sentence was "Pleasant zoos are rarely reached by efficient transportation" in English. Each utterance was recorded by one female and two male speakers in a nonreverberant room, sampled originally at 48 (kHz) and down-sampled to 8 (kHz). Each untrained speaker was asked not to move their head from the center of the two microphones (the distance between the two mics. is 50 (cm)). For the fourth data, the speech utterance was recorded by a male Japanese native speaker in an ordinary room (the shape is rectangular and its size is 200 (cm) wide, 350 (cm) long, and 230 (cm) tall) of a house near the kitchen system, without any sound shielding equipped, and, the sentence was "Hajime-mashite, Kon-nichiwa" in Japanese ("How do you do, hello"). The speech data were then normalized to have unity variance.

In order to validate the proposed scheme, we tested two cases for the noise components $n_i(k)$ $(i = 1, 2)$: the two noise components are 1) synthetically generated i.i.d. sequences, and 2) the real stereophonic fan noise data recorded in the same room and condition as those for the fourth speech data.

The two i.i.d. noise components assumed were the signals generated from 1) uniform distribution (using MATLAB function, *rand()*) shifted to lie within the range from -0.5 to 0.5, and 2) Normal distribution (using MATLAB function, *randn()*).

For the SSP, the length of the analysis matrix is fixed to 32. In a separate simulation study, we confirmed that this is a reasonable choice for giving a good trade-off in terms of the performance and the computational complexity, since the SSP (i.e., the EVD) operation is the most computationally demanding part within the proposed scheme (e.g., for the actual computation, applying the Cholesky's decomposition requires $O(L^3/3)$).

For the ASE, the standard normalized-LMS algorithm (e.g., see [20]) was used to adjust the filter coefficients in the dual adaptive signal enhancer (DASE, i.e., the case where $M = 2$ in Fig. 1). For each adaptive filter, the learning constant was chosen as 0.5. The filter length was fixed to 51, which allows approximately 3 (ms) of delay in left/right channel, and, within this range, neither precedence effect (or, alternatively, Haas effect) nor echo effect will occur [29]. Moreover, the scalar constants c_i were empirically fixed to 0.1 for both the left and right channels, which was empirically found to moderately suppress the distortion and satisfied a good trade-off between a reasonable stereophonic image compensation and signal distortion.

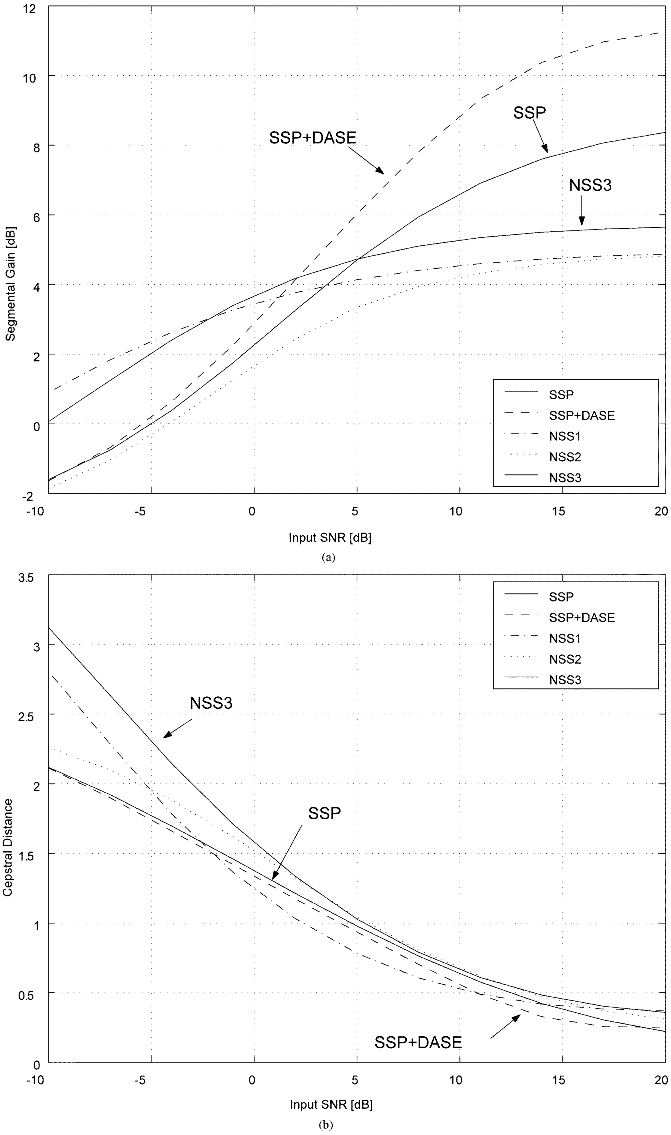


Fig. 7. Performance comparison using the real stereophonic fan noise components. (a) Comparison of the segmental gain and (b) Comparison of the cepstral distance.

B. Performance Measurements

For the evaluation of the enhancement quality, the objective measurement in terms of both the segmental gain in SNR and averaged cepstral distance was considered. In this paper, the measurement in terms of the segmental gain in SNR is employed, instead of the ordinary segmental SNR (see, e.g., [30] and [31]), in order to clarify how much gain in the context of noise reduction is obtained at each frame, rather than checking merely the signal-to-noise ratio.[1]

[1] Imagine the situation where both the input and output SNRs are high (at 5 dB and 22 dB for the input and output SNR, respectively, say). Then, the conventional segmental SNR cannot fully explain how much amount of noise reduction we gain, if the input SNR varies greatly (from 5 dB to 20 dB, say). Hence, we consider the segmental gain in SNR as a measurement for noise reduction in this paper.

Then, the segmental gain in SNR (dB) is defined as

$$\begin{aligned} G(dB) &= \text{Segmental SNR (Output)} \\ &\quad - \text{Segmental SNR (Input)} \\ &= \frac{1}{Mp_1}\sum_{i=1}^{M}\sum_{j=1}^{p_1}\left\{10\log_{10}\frac{\|\boldsymbol{s}_i\|_2^2}{\|\boldsymbol{s}_i-\hat{\boldsymbol{s}}_i\|_2^2}\right. \\ &\qquad\qquad \left. -10\log_{10}\frac{\|\boldsymbol{s}_i\|_2^2}{\|\boldsymbol{n}_i\|_2^2}\right\}, \\ &= \frac{1}{Mp_1}\sum_{i=1}^{M}\sum_{j=1}^{p_1}10\log_{10}\frac{\|\boldsymbol{n}_i\|_2^2}{\|\boldsymbol{s}_i-\hat{\boldsymbol{s}}_i\|_2^2} \end{aligned} \quad (14)$$

where $M = 2$ (stereophonic), $\boldsymbol{s}_i = [s_i(k), s_i(k+1), \ldots, s_i(k+N_f-1)]^T$, $\hat{\boldsymbol{s}}_i = [\hat{s}_i(k), \hat{s}_i(k+1), \ldots, \hat{s}_i(k+N_f-1)]^T$, $\boldsymbol{n}_i = [n_i(k), n_i(k+1), \ldots, n_i(k+N_f-1)]^T$, $(k = (j-1)N_f, (j-1)N_f+1, \ldots, jN_f-1, j = 1, 2, \ldots, p_1)$ are respectively the clean speech, enhanced speech, and the noise signal vector, and where N_f is the number of the samples in each frame ($N_f = 256$, in this paper) and p_1 is the number of the frames.

The averaged cepstral distance is given by

$$d_{cep} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{p_{2,i}}\sum_{j=1}^{p_2}\sum_{k=1}^{2q}\left(c_{i,k}(j) - c'_{i,k}(j)\right)^2 \quad (15)$$

where $c_{i,k}(j)$ and $c'_{i,k}(j)$ are the cepstral coefficients corresponding to the clean and the enhanced signal at left/right channel, respectively. The parameter q is the order of the model (chosen as 8), and $p_{2,i}$ ($i = 1, 2$, in this paper) is the number of frames where speech is present[2]. The determination of speech presence was achieved by manual inspection of the clean speech signals. (Note that normally the numbers of the frames $p_1 \neq p_{2,i}$.)

C. Simulation Results

For both the two additive i.i.d. and real stereophonic fan noise cases, performance comparisons are made using 1) NSS, 2) only the SSP, and 3) SSP + DASE. In the case of NSS, three different parameter settings were attempted (i.e., indicated as NSS1, NSS2, and NSS3 in Figs. 5 and 7) in order to see how the performance varies. As shown in Table II, the first four parameters of the NSS method, which were empirically found to affect the performance mostly, were varied arbitrarily, with keeping as much as a reasonable trade-off between the noise reduction performance and the amount of musical noise introduced or the distortion within the shape of the enhanced speech, while the other parameters were remained the same for all the three settings.

1) Two Additive i.i.d. Noise Case: Due to the constraints on space, we present only the simulation results using two speech samples in this sub-section.

In Figs. 3 and 4, (a) shows the clean speech data (both the left and right channel signals), (b) the noisy speech (assuming the input $\text{SNR} = 3$ (dB) for speech sample no. 1, while $\text{SNR} = 0$ (dB) for speech sample no. 2), (c) the enhanced speech by dual-mono nonlinear spectral subtraction (NSS) algorithm (with the parameter setting as NSS1), (d) the enhanced speech by only SSP (SSP), and (e) the enhanced speech by SSP + DASE (i.e., a combination of the SSP and DASE) method, respectively. For the simulation using speech sample no. 1, the two i.i.d. noise components used were the random signals generated from the Normal distribution, while the random signals generated from the uniform distribution were used for speech sample no. 2.

Since the performance in terms of both the segmental gain and cepstral distance was similar to each other for the two distributions, we show only the case of the Normal distribution. Fig. 5(a) shows a comparison of the segmental gain [given by (14)] versus input SNR, using the two-channel observations with the additive noise components generated from two i.i.d. Normal distribution. The results shown are those averaged over the three speech samples. Table I shows the actual values for p_1, $p_{2,1}$, and $p_{2,2}$ used to compute (14) and (15), respectively. In the figure, the performance of the three different noise reduction algorithms, i.e., 1) SSP (using only an SSP), 2) SSP + DASE (a combination of an SSP and DASE), and 3) NSS algorithm, is compared.

In the figure, at lower SNRs, the performance with NSS is better than the other two, while at higher SNRs the SSP+DASE algorithm is the best. However, at lower SNRs, as in Fig. 5(a), the performance in terms of cepstral distance with NSS (for all the three parameter settings) is poorest amongst the three. As in Fig. 5(a), at around $\text{SNR} > 7$ (dB), it is clearly seen that the combination of the SSP and DASE yields performance improvement of more than 3 (dB) over the case using only the SSP.

2) Real Stereophonic Fan Noise Case: For the real stereophonic fan noise case, we used the data originally sampled at 48 (kHz) and performed the simulations.

Fig. 6 shows the simulation results using the real fan noise data. Fig. 6(a) shows the part of the clean speech data (sampled at 48 (kHz), using Speech Sample no. 4, note that the sample number for display is limited from sample no. 60 001 to 70 000 for a clear presentation of the results.), (b) the noisy speech (assuming the input $\text{SNR} = 2$ (dB)), (c) the enhanced speech by dual-mono nonlinear spectral subtraction (NSS) algorithm (with the parameter setting NSS1), (d) the enhanced speech by only the SSP, and (e) the enhanced speech by the SSP + DASE, respectively.

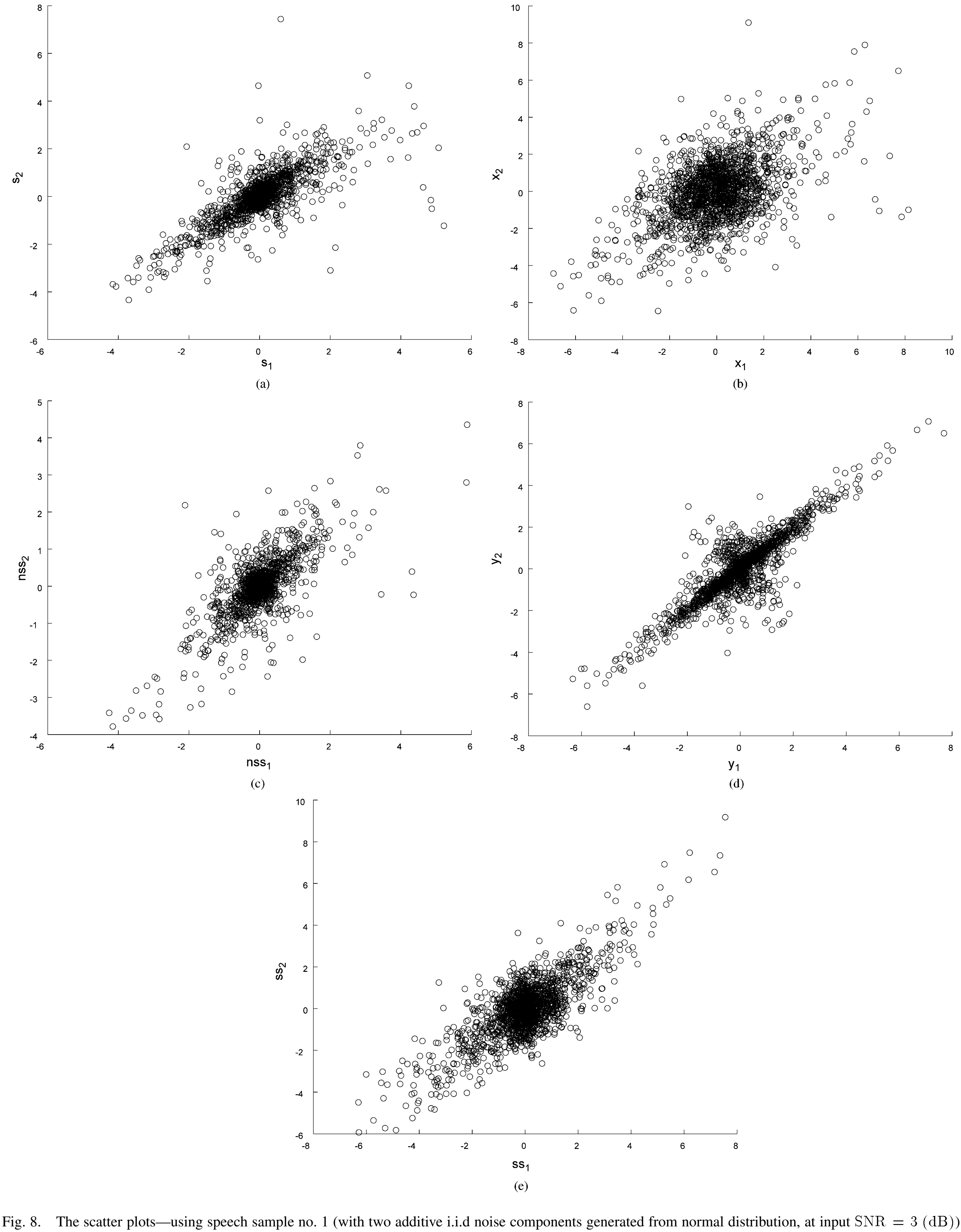
Fig. 7 shows a comparison of the segmental gain and the cepstral distance, respectively. In the figure, the performance of the three different noise reduction methods, i.e., 1) NSS, 2) SSP, and 3) SSP + DASE, is compared, as for the i.i.d. noise case.

In Fig. 7(a), similar to the two i.i.d. noise case, the performance with NSS is better than the other two at lower SNRs, In contrast, as in Fig. 7(b), note that at lower SNRs, the case with SSP + DASE is, however, best among the three methods. This also coincided with the informal listening tests.

D. Discussion

These simulation results indicate that the NSS method removes not only the noise but some parts of the speech. Moreover, as in Figs. 3, 4, and 6(c), it is clearly seen that some voiced speech parts are eliminated or greatly changed in shape. This was also confirmed by informal listening tests, in which the enhanced speech obtained from the NSS sounds 'hollow' besides the additive musical tonal noise. In contrast, in the listening tests, it was also observed that the enhanced speech by the other two methods does not have such artifacts or distortion

[2] For the cepstral distance measurement, the number of the frames may vary for left/right channel, since in this paper a single threshold value according to the left channel signal was used to determine whether speech is present or not.



(a)

(b)

(c)

(d)

(e)

Fig. 8. The scatter plots—using speech sample no. 1 (with two additive i.i.d noise components generated from normal distribution, at input $\mathrm{SNR} = 3\ (\mathrm{dB})$). (a) Clean speech. (b) Noisy data $(\mathrm{SNR} = 3\ \mathrm{dB})$. (c) Enhanced speech by NSS. (d) Enhanced speech by SSP. (e) Enhanced speech by $\mathrm{SSP} + \mathrm{DASE}$.

and the noise level is certainly decreased preserving the components that are of fundamental importance to the stereophonic image.

In the listening tests, it was also confirmed that the speech obtained from the SSP sounds rather dual-mono or that the spatial image is gone, but the stereophonic image is, to a great extent, recovered in the enhanced speech obtained after the post-processing by the adaptive FIR filters.

It can be said that these empirical facts agree with the hypothesis in Section II-D in which the adaptive filters can compensate for the stereophonic image from the signals obtained by the SSP using the information contained in the true stereophonic observations $x_i(k)$.

As the performance improvement of SSP together with the DASE approach observed in Figs. 5 and 7(a) and (b) compared to that of only using an SSP, the enhanced signal obtained after the DASE is much closer to the original stereophonic speech signal than that after the SSP. Moreover, to see intuitively how the stereophonic image in the enhanced signals can be recovered, the scatter plots are shown in Fig. 8, where the parameter settings are all the same as those for Fig. 3 (i.e., using speech sample no. 1, the input $\mathrm{SNR} = 3\ (\mathrm{dB})$, and the two additive i.i.d noise components generated from Normal distribution). In Fig. 8(e) (in the figure the labels 'ss_1' and 'ss_2' correspond to $\hat{s}_1$ and $\hat{s}_2$, respectively, whereas those 'nss_1' and 'nss_2' correspond respectively to the enhanced signals obtained by the NSS method), it is observed that the pattern of the scatter plot for the enhanced speech after the $\mathrm{SSP} + \mathrm{DASE}$ somewhat approaches that of the original stereophonic speech as in Fig. 8(a), in comparison with that for the speech after only the SSP shown in Fig. 8(d) is considered as rather monaural (which also agreed with the informal listening tests), since the distribution of the data points are more concentrated around the line $s_1 = s_2$ than the case of $\mathrm{SSP} + \mathrm{DASE}$.

In Fig. 8(c), it is also observed that some data points in the original signals are missing (especially at lower-left corner) and that the shape of the cluster is somewhat changed for the enhanced signal by the NSS. This coincides with the empirical fact that the enhanced speech by the NSS can greatly be changed in shape.

From these observations, it can be empirically justified that the adaptive filters work to compensate for the stereophonic image.

IV. Conclusion

In this paper, a novel multichannel noise reduction method has been proposed by a combination of SSP and multichannel adaptive signal enhancement technique. In the proposed method, the SSP is used for the extraction of the signal of interest to the adaptive filter in each channel, and actual signal enhancement is performed by the adaptive approach. The proposed methods have been applied to stereophonic noise reduction, where the number of the sensors is just two. In the simulation study, it has been shown that the performance with the proposed methods is superior to the conventional NSS approach. Our simulation study also shows that the adaptive filters can compensate for the stereophonic image.

Acknowledgment

The authors would like to thank the anonymous reviewers for their encouraging/insightful comments to improve the draft of the manuscript. They also would like to thank Dr. R. R. Gharieb for the fruitful discussions.

References

[1] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. EUSIPCO-94*, Edinburgh, U.K., 1994, pp. 1182–1185.

[2] F. Xie and D. Van Compernolle, "Speech enhancement by spectral magnitude estimation—A unifying approach," *Speech Commun.*, vol. 19, no. 2, pp. 89–104, Aug. 1996.

[3] H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction using dual microphones," in *Int. Workshop on Acoustic Echo and Noise Control*, Sep. 1999, pp. 60–63.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[5] Gustafsson *et al.*, "System and Method for Dual Microphone Signal Noise Reduction Using Spectral Subtraction," U.S. Patent 6 549 586, Apr. 2003.

[6] S. Haykin, *Unsupervised Adaptive Filtering*. New York: Wiley, 2000, vol. I & II.

[7] S. Amari and A. Cichocki, "Adaptive blind signal processing—Neural network approaches," *Proc. IEEE*, vol. 86, pp. 2026–2048, Oct. 1998.

[8] K. Torkkola, "Blind separation of delayed sources based on information maximization," in *Proc. ICASSP-96*, 1996, pp. 3509–3512.

[9] H. L. N. Thi and C. Jutten, "Blind source separation for convolutive mixtures," *Signal Process.*, vol. 45, no. 2, pp. 209–229, 1995.

[10] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Process.*, vol. 24, no. 1, pp. 1–10, 1991.

[11] P. A. Karjalainen, J. P. Kaipio, A. S. Koistinen, and M. Vuhkonen, "Subspace regularization method for the single-trial estimation of evoked potentials," *IEEE Trans. Biomed. Eng.*, vol. 40, pp. 849–860, Jul. 1999.

[12] T. Kobayashi and S. Kuriki, "Principle component elimination method for the improvement of S/N in evoked neuromagnetic field measurements," *IEEE Trans. Biomed. Eng.*, vol. 46, pp. 951–958, Aug. 1999.

[13] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Commun.*, vol. 10, pp. 45–57, Feb. 1991.

[14] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[15] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 439–448, Nov. 1995.

[16] P. S. K. Hansen, "Signal Subspace Methods for Speech Enhancement," Ph.D. dissertation, Technical Univ. Denmark, Lyngby, 1997.

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

[18] S. Doclo and M. Moonen, "Multi-microphone noise reduction using GSVD-based optimal filtering with ANC postprocessing stage," in *Proc. 9th IEEE Digital Sig. Proc. Workshop*, Hunt, TX, USA, Oct. 2000.

[19] ——, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Processing*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[20] S. Haykin, *Adaptive Filter Theory*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1991.

[21] A. Cichocki, R. R. Gharieb, and T. Hoya, "Efficient extraction of evoked potentials by combination of Wiener filtering and subspace methods," in *Proc. ICASSP-2001*, vol. 5, Salt Lake City, UT, May 2001, pp. 3117–3120.

[22] P. K. Sadasivan and D. N. Dutt, "SVD based technique for noise reduction in electroencephalographic signals," *Signal Process.*, vol. 55, no. 2, pp. 179–189, 1996.

[23] A. Cichocki and S. Amari, *Adaptive Blind Signal And Image Processing*. New York: Wiley, 2002.

[24] J. A. Apolinario, M. L. R. de Campos, and P. S. R. Diniz, "Convergence analysis of the binormalized data-reusing LMS algorithm," in *Proc. European Conference on Circuit Theory and Design*, Budapest, Hungary, Sep. 1997, pp. 972–977.

[25] N. Forsyth, J. A. Chambers, and P. A. Naylor, "A noise robust alternating fixed-point algorithm for stereophonic acoustic echo cancellation," *Electron. Lett.*, vol. 35, no. 21, pp. 1812–3, Oct. 1999.

[26] P. C. Hansen and S. H. Jensen, "FIR filter representation of reduced-rank noise reduction," *IEEE Trans. Signal Process.*, vol. 46, no. 6, pp. 1737–1741, Jun. 1998.

[27] D. Callaerts, J. Vandewalle, W. Sansen, and M. Moonen, "On-Line algorithm for signal separation based on SVD," in *SVD and Signal Processing—Algorithms, Applications and Architectures*. New York: Elsevier, 1988, vol. 1, pp. 269–275.

[28] C. C. Ko and C. S. Siddharth, "Rejection and tracking of an unknown broadband source in a two-element array through least square approximation of inter-element delay," *IEEE Signal Processing Lett.*, vol. 6, no. 5, pp. 122–125, May 1999.

[29] C. Hugonnet and P. Walder, *Stereophonic Sound Recording—Theory and Practice*. New York: Wiley, 1998.

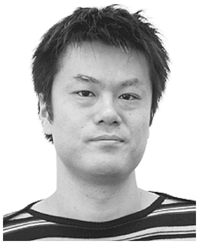
[30] J. R. Deller, Jr, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: Macmillan, 1993.

[31] R. Le Bouquin-Jennes, A. A. Akbari, 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.

Tetsuya Hoya (M'01) was born in Tokyo, Japan, on September 15, 1969. He received the B.Sc. and M.Sc. degrees both from Meiji University, Japan, in 1992 and 1994, respectively, both in electrical engineering. He received the Ph.D. degree from Imperial College of Science, Technology and Medicine, University of London, London, U.K., in 1998.

From April 1994 to September 1994, he was a Research Assistant at Department of Electronics and Communication, Graduate School of Meiji University, Japan. He was then a student at Department of Electrical and Electronics Engineering, Imperial College of Science, Technology and Medicine, from October 1994 to December 1997. He was a Postdoctoral Research Associate at Department of Electrical and Electronics Engineering, Imperial College, London, from September 1997 to August 2000. Since October 2000, he has been a Research Scientist within the Brain Science Institute, RIKEN (The Institute of Physical and Chemical Research), Japan, and a Visiting Lecturer at Saitama Institute Technology, Japan, since April 2003. His research interest focuses on a wide spectrum of brain science: artificial intelligence, cognitive neuroscience, combinatoric optimization, computational linguistics, consciousness studies, electroencephalography, neural networks (connectionism), philosophy, psychology, robotics, and signal processing. He has published more than 30 technical papers and is the author of the book *Artificial Mind System—Kernel Memory Approach* (Berlin, Germany: Springer-Verlag, 2005).

Dr. Hoya was a Committee Member of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA-2003).



Toshihisa Tanaka (S'98–M'02) received the B.E., M.E., and Ph.D. (Dr. Eng.) degrees from Tokyo Institute of Technology, Tokyo, Japan, in 1997, 2000, and 2002, respectively. From 1997 to 1998, he was a visiting student at Korea University under the study-abroad program of the Monbusho.

From 2000 to 2002, he was with Tokyo Institute of Technology as a JSPS Research Fellow. Since 2002, he has been a Research Scientist of the Laboratory for Advanced Brain Signal Processing at Brain Science Institute, RIKEN. In 2004, he joined the Department of Electrical and Electronic Engineering, Tokyo University of Agriculture and Technology (TUAT), where he is currently a Lecturer. His research interests include image and signal processing, multirate systems, blind signal separation, and adaptive signal processing.

Dr. Tanaka was a local organizing committee member of ICA 2003 and is an international program committee member of ICA 2004. He is a member of IEICE and EURASIP. He received the 15th TAF Telecom System Technical Student Award in 2001 and the Tejima Memorial Award in 2003.

Andrzej Cichocki (M'96) was born in Poland. He received the M.Sc. (with honors), Ph.D., and Habilitate Doctorate (Doctor of Science) degrees, all in electrical engineering, from Warsaw University of Technology, Warsaw, Poland, in 1972, 1975, and 1982, respectively.

Since 1972, he has been with the Institute of Theory of Electrical Engineering, Measurements and Information Systems at the Warsaw University of Technology, where he became a Full Professor in 1991. He is the co-author of three books: *MOS Switched-Capacitor and Continuous-Time Integrated Circuits and Systems* (Berlin, Germany: Springer-Verlag, 1989), *Neural Networks for Optimization and Signal Processing* (New York: Teubner-Wiley, 1993/1994), and *Adaptive Blind Signal and Image Processing* (New York: Wiley, 2003) and more than 150 research journal papers. He spent a few years at University Erlangen-Nuernberg (Germany) as Alexander Humboldt Research Fellow and Guest Professor. Since 1995 he has been working in the Brain Science Institute RIKEN (Japan), as a team leader of the Laboratory for Open Information Systems and currently as a head of laboratory for Advanced Brain Signal Processing. His current research interests include biomedical signal and image processing, especially analysis and processing of multi-sensory, multi-modal data.

Dr. Cichocki is a member of the IEEE Signal Processing Technical Committee for Machine Learning for Signal Processing and IEEE Circuits and Systems Technical Committee for Blind Signal Processing.

Takahiro Murakami received the B.Sc. and M.Sc. degrees in electrical engineering from Meiji University, Kawasaki, Japan, in 2000 and 2002, respectively. He is currently working toward the Ph.D. at Graduate School of Electrical Engineering, Meiji University.

His research interests include speech signal processing and digital signal processing.

Mr. Murakami is a student member of the Institute of Electronics, Information, and Communication Engineers (IEICE).

Gen Hori was born in Tokyo, Japan. He received the B.Sc. and M.Sc. degrees in mathematical engineering and the Ph.D. degree in information engineering in 1991, 1993 and 1996 respectively, all from the University of Tokyo.

From April 1996 to September 1998, he was a Research Fellow of the Japan Society for the Promotion of Science. Since 1998, he has been a researcher with Brain Science Institute, RIKEN, Japan. His research interests include independent component analysis (ICA) and matrix dynamical systems with application to signal processing.

Jonathon A. Chambers (S'85–M'85–SM'98) was born in Peterborough, U.K., in March 1960.

He holds a Cardiff Professorial Research Fellowship in Digital Signal Processing at Cardiff University, U.K. His research interests are in the areas of adaptive, blind, and statistical signal processing with applications in wireless communications and intelligent sensors. He is the co-author of the research monograph *Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability* (New York: Wiley, 2002).

Dr. Chambers has served as an Associate Editor for IEEE Transactions on Signal Processing and the IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing. He is currently an Associate Editor for the IEEE Signal Processing Letters and the *EURASIP Journal on Wireless Communications and Networking*. He is the co-recipient of two Institute of Electrical Engineers (IEE) premium awards. He is a Member of the European Signal Processing Society ADCOM and has served as the Chairman of the U.K. IEE Professional Group on Signal Processing.