PAPER

Speech Enhancement by Spectral Subtraction Based on Subspace Decomposition

Takahiro MURAKAMI[†a)], *Student Member*, **Tetsuya HOYA**[††b)], *Nonmember*, *and* **Yoshihisa ISHIDA**[†c)], *Member*

SUMMARY This paper presents a novel algorithm for spectral subtraction (SS). The method is derived from a relation between the spectrum obtained by the discrete Fourier transform (DFT) and that by a subspace decomposition method. By using the relation, it is shown that a noise reduction algorithm based on subspace decomposition is led to an SS method in which noise components in an observed signal are eliminated by subtracting variance of noise process in the frequency domain. Moreover, it is shown that the method can significantly reduce computational complexity in comparison with the method based on the standard subspace decomposition. In a similar manner to the conventional SS methods, our method also exploits the variance of noise process estimated from a preceding segment where speech is absent, whereas the noise is present. In order to more reliably detect such non-speech segments, a novel robust voice activity detector (VAD) is then proposed. The VAD utilizes the spread of eigenvalues of an autocorrelation matrix corresponding to the observed signal. Simulation results show that the proposed method yields an improved enhancement quality in comparison with the conventional SS based schemes.

key words: *speech enhancement, spectral subtraction, subspace decomposition, MUSIC algorithm*

1. Introduction

In general speech applications such as automatic speech recognizers, hands-free mobile telephony, or hearing aids, noise reduction is necessary in order to provide better utilities. Spectral subtraction (SS) based methods are well-known for such purpose in speech signal processing [1]–[4]. The SS carries out the noise reduction by subtracting an estimate of the noise spectrum from the noisy signal. In the conventional SS methods, the estimate of the noise spectrum is obtained from the preceding segments where speech is absent, under the assumption that the statistics of the noise process do not vary rapidly in time. Therefore, the SS generally requires a voice activity detector (VAD) in order to detect the non-speech segments and it is well-known that the performance of the SS is dependent upon the VAD. Especially in the noisy environment, a robust VAD is inevitable for the SS.

Martin proposed the nonlinear spectral subtraction (NSS) [2], [3] which does not require any VAD. In the NSS, the noise spectrum in the observed speech is estimated by using the minimum statistics obtained from several subsequent frames. Despite that NSS does not require the VAD, the performance of NSS is quite dependent upon the choice of many parameters, for instance, spectral floor constant, over subtraction factor, and smoothing constant. In practice, to find the reasonable choice of the parameters is very hard.

Recently, a number of methods for speech enhancement based on subspace decomposition have been developed [5]–[14]. In the subspace decomposition methods, the observed signals are expanded with orthonormal bases and such bases are partitioned into two disjoint subsets, i.e., the bases spanning the signal subspace and those spanning the noise. Then, noise reduction is achieved by exploiting the subspace estimates, e.g., by orthonormally projecting the observed signal onto the estimated signal part. In general, the subspace decomposition is carried out by employing the singular value decomposition (SVD) or the eigendecomposition (ED). However, since the algebraic complexity of both the SVD and ED is proportional to the length of analysis frame, the subspace decomposition is computationally heavy when a long analysis frame is used. Therefore, in order to alleviate the complexity due to the subspace decomposition, a large number of adaptive tracking algorithms have been proposed so far [11], [13], [15]–[20].

The proposed method in this paper is essentially based on the subspace decomposition. In the method, we exploit the multiple signal classification (MUSIC) algorithm [4], [21]. The MUSIC algorithm is a subspace decomposition method to estimate the frequencies of sinusoids of the signal contaminated with additive white noise. Generally, in the MUSIC algorithm, the noise subspace estimated by the ED of autocorrelation matrix is used. The frequencies estimated by MUSIC algorithm are then utilized for noise reduction, which is based on the maximum likelihood method [23]. In contrast, within this paper, by approximating the orthonormal bases spanning both the signal and noise subspace to the Fourier bases, a relation between the discrete Fourier transform (DFT) and MUSIC spectra is firstly derived. Then, in terms of the orthonormal bases so estimated, it is shown that the noise reduction method based on the MUSIC algorithm combined with the maximum likelihood estimate can lead to an SS based method in which noise reduction is performed by subtracting the estimated variance of noise process from the observed signal in the frequency domain. Since the method does not involve any heavy alge-

Manuscript received April 27, 2004.
Manuscript revised September 2, 2004.
Final manuscript received December 2, 2004.

[†] The authors are with the Department of Electronics and Communications, Meiji University, Kawasaki-shi, 214-8571 Japan.

[††] The author is with the Laboratory for Advanced Brain Signal Processing, Wako-shi, 351-0198 Japan.

a) E-mail: tmura@isc.meiji.ac.jp
b) E-mail: hoya@brain.riken.jp
c) E-mail: ishida@isc.meiji.ac.jp

DOI: 10.1093/ietfec/e88–a.3.690

braic computation such as the ED, the computational complexity in the proposed method is greatly alleviated in comparison with the standard MUSIC algorithm combined with the maximum likelihood estimate. Second, for the application to speech signals, a novel VAD for reliably estimating the variance of noise process is proposed. The VAD is developed under the assumption that the eigenvalues of the autocorrelation matrix associated with the noise are approximated to the variance of the noise, whereas those associated with the noisy speech are not approximated to a unique value but spread within a certain range. Later, it will be confirmed that this assumption can analytically be validated.

2. Review of the MUSIC Algorithm for Noise Reduction

Let an N-sample observed signal vector $\boldsymbol{y} = [y(0), y(1), \cdots, y(N-1)]^T$ (T: a vector or matrix transpose) be

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{n}$ are respectively the target and noise signal vectors and $\boldsymbol{x}$ is composed of P ($< N$) sinusoids as follows:

$$\boldsymbol{x} = \sum_{k=0}^{P-1} X(f_k)\boldsymbol{s}(f_k) \tag{2}$$

$$\boldsymbol{s}(f_k) = [1, e^{j2\pi f_k}, \cdots, e^{j2\pi f_k(N-1)}]^T \tag{3}$$

where $\boldsymbol{s}(f_k)$ and $X(f_k)$ ($k = 0, 1, \cdots, P-1$) are respectively the sinusoidal signal vectors and the complex amplitudes at the unknown frequencies f_k. This expression is referred to as *complex sinusoid model*. Note that f_k in this model is any frequency, while in the discrete Fourier transform (DFT), the frequency is given by the fixed value $f_k = l/N$ ($l \in \{0, 1, \cdots, N-1\}$). Then, the noise is often modeled as Gaussian random process due to the central limit theorem [22]. In this paper, by taking this general principle into account, $\boldsymbol{n}$ is assumed to be zero-mean Gaussian white noise with variance σ_n^2 and uncorrelated with $\boldsymbol{x}$.

The autocorrelation matrix of $\boldsymbol{y}$ is defined as

$$\boldsymbol{R}_{yy} = E[\boldsymbol{y}\boldsymbol{y}^H] \tag{4}$$

where $E[\cdot]$ and H denote the expectation operation and the Hermitian transpose of a vector (or matrix), respectively. Since $\boldsymbol{x}$ and $\boldsymbol{n}$ are uncorrelated with each other, (4) can be rewritten by

$$\begin{aligned}\boldsymbol{R}_{yy} &= E[\boldsymbol{x}\boldsymbol{x}^H] + E[\boldsymbol{n}\boldsymbol{n}^H]\\ &= \boldsymbol{R}_{xx} + \boldsymbol{R}_{nn}\\ &= \boldsymbol{R}_{xx} + \sigma_n^2\boldsymbol{I}\end{aligned} \tag{5}$$

where $\boldsymbol{R}_{xx}$ and $\boldsymbol{R}_{nn} = \sigma_n^2\boldsymbol{I}$ are respectively the autocorrelation matrices of $\boldsymbol{x}$ and $\boldsymbol{n}$. Then, the eigen-decomposition (ED) of $\boldsymbol{R}_{yy}$ is expressed in the form

$$\boldsymbol{R}_{yy} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1} \tag{6}$$

where the diagonal elements in $\boldsymbol{D} = diag(\lambda_0, \lambda_1, \cdots, \lambda_{N-1})$ and the columns of $\boldsymbol{V} = [\boldsymbol{v}_0, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_{N-1}]$ are the eigenvalues and corresponding eigenvectors of $\boldsymbol{R}_{yy}$, respectively. From (5), λ_k are given by

$$\lambda_k = \mu_k + \sigma_n^2, \ (k = 0, 1, \cdots, N-1) \tag{7}$$

where μ_k ($k = 0, 1, \cdots, N-1$) are the eigenvalues of $\boldsymbol{R}_{xx}$. Since the target signal $\boldsymbol{x}$ consists of P sinusoids, μ_k are obtained as P positive eigenvalues and $N-P$ zeros. Therefore, λ_k satisfy the relation

$$\begin{cases}\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_{P-1} > \sigma_n^2\\ \lambda_P = \lambda_{P+1} = \cdots = \lambda_{N-1} = \sigma_n^2\end{cases}. \tag{8}$$

The relation (8) indicates that $\{\boldsymbol{v}_0, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_{N-1}\}$ can be partitioned into two disjoint subsets. Namely, the first set $\{\boldsymbol{v}_0, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_{P-1}\}$ associated with the P largest eigenvalues spans the *signal subspace*, whereas the second $\{\boldsymbol{v}_P, \boldsymbol{v}_{P+1}, \cdots, \boldsymbol{v}_{N-1}\}$ associated with the $N-P$ smallest eigenvalues (i.e., corresponding to σ_n^2) spans the *noise subspace*. Since the signal and noise subspace are mutually orthogonal, the sinusoidal signal vectors given by (3) are accordingly orthogonal to the noise subspace:

$$\begin{aligned}&\boldsymbol{s}^H(f_k)\boldsymbol{v}_l = 0,\\ &(k=0, 1, \cdots, P-1; l=P, P+1, \cdots, N-1)\end{aligned} \tag{9}$$

Then, the MUSIC spectrum of $\boldsymbol{y}$ is defined as

$$Y_{MUSIC}(f) = \frac{1}{\sum_{l=P}^{N-1} |\boldsymbol{s}^H(f)\boldsymbol{v}_l|^2} \tag{10}$$

where f is an arbitrary frequency. From (9), $Y_{MUSIC}(f)$ is sharply peaked at $f = f_k$ ($k = 0, 1, \cdots, P-1$). Therefore, the estimated frequencies $\hat{f}_k$ ($k = 0, 1, \cdots, P-1$) corresponding to $\boldsymbol{x}$ can be obtained by simply taking the P peaks on the MUSIC spectrum. Finally, the estimated frequencies $\hat{f}_k$ are utilized for eliminating the noise in $\boldsymbol{y}$. Noise reduction in $\boldsymbol{y}$ is implemented based on the maximum likelihood estimate [23]:

$$\hat{\boldsymbol{x}} = \boldsymbol{S}(\boldsymbol{S}^H\boldsymbol{S})^{-1}\boldsymbol{S}^H\boldsymbol{y} \tag{11}$$

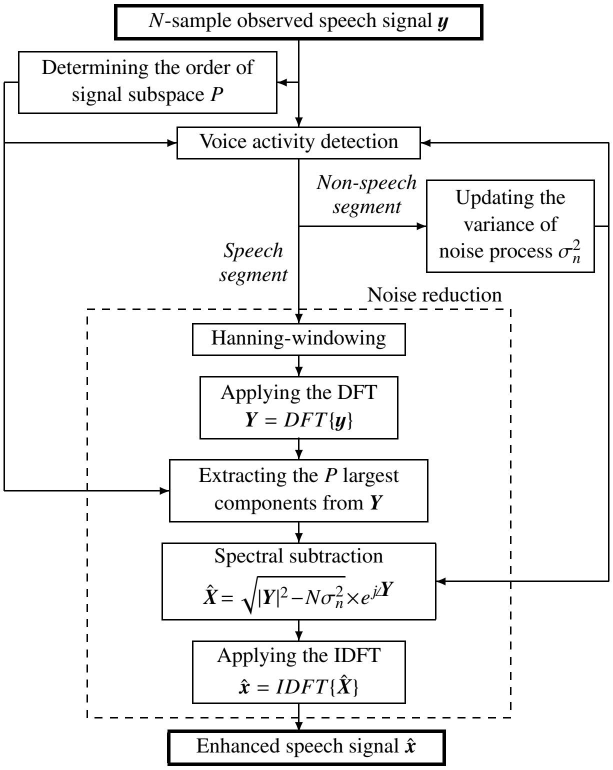
$$\boldsymbol{S} = [\boldsymbol{s}(\hat{f}_0), \boldsymbol{s}(\hat{f}_1), \cdots, \boldsymbol{s}(\hat{f}_{P-1})] \tag{12}$$

where $\hat{\boldsymbol{x}}$ is an estimate of the target signal $\boldsymbol{x}$.

In general, the MUSIC algorithm combined with the maximum likelihood estimate is commonly used for noise reduction. However, both the MUSIC algorithm and the maximum likelihood estimate involve rather heavy computation, for instance, the ED and matrix inversion. Therefore, in practice, it is necessary to alleviate such computational load.

3. Proposed Method

Figure 1 summarizes the procedure for speech enhancement proposed in this paper. In the figure, $\boldsymbol{y}$ is an N-sample observed signal vector, $\boldsymbol{Y}$ is the DFT spectrum of $\boldsymbol{y}$, $\hat{\boldsymbol{X}}$ is the



N-sample observed speech signal $\boldsymbol{y}$
Determining the order of signal subspace P
Voice activity detection
Non-speech segment
Updating the variance of noise process σ_n^2
Speech segment
Noise reduction
Hanning-windowing
Applying the DFT
$\boldsymbol{Y} = DFT\{\boldsymbol{y}\}$
Extracting the P largest components from $\boldsymbol{Y}$
Spectral subtraction
$\hat{\boldsymbol{X}} = \sqrt{|\boldsymbol{Y}|^2 - N\sigma_n^2} \times e^{j\angle\boldsymbol{Y}}$
Applying the IDFT
$\hat{\boldsymbol{x}} = IDFT\{\hat{\boldsymbol{X}}\}$
Enhanced speech signal $\hat{\boldsymbol{x}}$

Fig. 1 Summary of the procedure for speech enhancement.

spectrum obtained by the newly proposed spectral subtraction (SS), $\hat{\boldsymbol{x}}$ is the enhanced speech, and σ_n^2 is the variance of noise process. As in Fig. 1, the method carries out noise reduction without involving algebraic complex calculation such as the ED and matrix inversion. Therefore, it is considered that the method is well suited for real-time implementations.

As shown, the method is similar to the combination of the classical threshold technique and the conventional spectral subtraction (SS). In both the threshold technique and SS, however, the performance is generally dependent upon the choice of parameters, especially the threshold value in the threshold technique and the subtraction factor in the SS, and thus to find the optimal choice of such parameters is normally very hard. In contrast, the proposed method gives a reasonable choice of parameters, since the method is based on the subspace decomposition method. As shown in Fig. 1, the method extracts the P largest frequency components from the DFT spectrum of $\boldsymbol{y}$ and then subtracts $N\sigma_n^2$ from these extracted frequency components without using the subtraction factor. Moreover, both the parameters, P and $N\sigma_n^2$, are explicitly given by the MUSIC algorithm combined with the maximum likelihood estimation as described in the following sections.

In this section, it is firstly shown that the ED of the autocorrelation matrix is approximated to Fourier bases expansion. This approximation yields a relation between the DFT and MUSIC spectra. Then, by using this relation, it is shown that the noise reduction algorithm based on the combination of the MUSIC algorithm and the maximum likelihood estimate results in a simple algorithm which does not involve heavy computation. Moreover, in order to perform noise reduction further, a novel SS method is derived by exploiting the property in (11).

3.1 Approximating the Eigen-Decomposition of the Autocorrelation Matrix

In general, the autocorrelation matrix $\boldsymbol{R}_{yy}$ is estimated by an ensemble average as

$$\boldsymbol{R}_{yy} = \frac{1}{M}\sum_{m=0}^{M-1} \boldsymbol{y}(m)\boldsymbol{y}^H(m) \tag{13}$$

where $\boldsymbol{y}(m) = [y(m), y(m+1), \cdots, y(m+N-1)]^T$ $(m = 0, 1, \cdots, M-1)$ is the observed signal vector in the m-th analysis frame and M is the number of analysis frames. In this paper, in order to alleviate the computational complexity in the ED, we consider the general assumption that $\boldsymbol{y}(m)$ has an implicit periodicity with period N as in the DFT theorem, i.e.,

$$y(m+N) = y(m). \tag{14}$$

Under this assumption, the ED of $\boldsymbol{R}_{yy}$ can be approximated to Fourier bases expansion. In (13), by using the Fourier bases, $\boldsymbol{y}(m)$ is expressed in the form

$$\boldsymbol{y}(m) = \boldsymbol{W}\boldsymbol{a}(m) \tag{15}$$

$$\boldsymbol{W} = [\boldsymbol{w}_0, \boldsymbol{w}_1, \cdots, \boldsymbol{w}_{N-1}] \tag{16}$$

$$\boldsymbol{w}_k = [1, e^{j2\pi k/N}, \cdots, e^{j2\pi k(N-1)/N}]^T \tag{17}$$

$$\boldsymbol{a}(m) = \frac{1}{N}[Y(0;m), Y(1;m), \cdots, Y(N-1;m)]^T \tag{18}$$

where $\boldsymbol{w}_k$ and $Y(k;m)$ $(k = 0, 1, \cdots, N-1)$ are the Fourier basis vector and the DFT spectrum of $\boldsymbol{y}(m)$ at the k-th frequency bin, respectively. Then, under the assumption (14), the eigenvalues and eigenvectors of $\boldsymbol{R}_{yy}$ are respectively approximated to

$$\lambda_l = \frac{|Y(k;0)|^2}{N} \tag{19}$$

$$\boldsymbol{v}_l = \boldsymbol{w}_k, \ (k, l = 0, 1, \cdots, N-1) \tag{20}$$

(see Appendix). Note that in general $k \neq l$, since k and l denote the indices in order of the frequency and amplitude, respectively.

3.2 Relation between the DFT and MUSIC Spectra

It has been shown that, under the assumption (14), the eigenvectors of $\boldsymbol{R}_{yy}$ are approximated to the Fourier bases as in (20). This approximation implies that the inner product of the sinusoidal signal vector and the eigenvector is equivalent to that of the sinusoidal signal vectors, since the Fourier basis is also given as the sinusoidal signal vector at the frequency k/N $(k = 0, 1, \cdots, N-1)$. Therefore, it is considered that the MUSIC spectrum defined by (10) yields a simple form as follows:

The inner product of the sinusoidal signal vector and the eigenvector is given by

$$\begin{aligned} \boldsymbol{s}^H(f)\boldsymbol{v}_l &= \boldsymbol{s}^H(f)\boldsymbol{w}_k \\ &= \sum_{n=0}^{N-1} e^{-j2\pi fn} e^{j2\pi kn/N} \\ &= \begin{cases} N, \left(f = \dfrac{k}{N}\right) \\ 0, \left(f \in \left\{0, \dfrac{1}{N}, \cdots, \dfrac{N-1}{N}\right\} \cap f \neq \dfrac{k}{N}\right) \\ c_f, \left(f \notin \left\{0, \dfrac{1}{N}, \cdots, \dfrac{N-1}{N}\right\}\right) \end{cases}, \\ &(k, l = 0, 1, \cdots, N-1) \end{aligned} \tag{21}$$

where $c_f(\neq 0)$ is any complex value. The relation (21) indicates that the denominator in (10) is obtained as follows:

- If f is equal to one of the frequencies associated with the P largest components in the DFT spectrum of $\boldsymbol{y}(0)$,

$$\sum_{l=P}^{N-1} |\boldsymbol{s}^H(f)\boldsymbol{v}_l|^2 = 0. \tag{22}$$

- Else if f is equal to one of the frequencies associated with the $N-P$ smallest components in the DFT spectrum of $\boldsymbol{y}(0)$,

$$\sum_{l=P}^{N-1} |\boldsymbol{s}^H(f)\boldsymbol{v}_l|^2 = N^2. \tag{23}$$

- Otherwise, if f is not equal to $\dfrac{k}{N}$ $(k = 0, 1, \cdots, N-1)$,

$$\sum_{l=P}^{N-1} |\boldsymbol{s}^H(f)\boldsymbol{v}_l|^2 > 0. \tag{24}$$

It is then evident that, from the relations (22)–(24), the MUSIC spectrum has poles only in the case of (22). From this, it is also said that the MUSIC spectrum is closely related with the DFT spectrum, i.e., (10) has P poles at the frequencies which are identical to those of the P largest components in the DFT spectrum of $\boldsymbol{y}(0)$.

3.3 Spectral Subtraction Based on the Subspace Decomposition

In the noise reduction algorithm based on the maximum likelihood estimate, as in (11), the matrix $\boldsymbol{S}$ is comprised of P sinusoidal signal vectors $\boldsymbol{s}(\hat{f}_k)$ $(k = 0, 1, \cdots, P-1)$ whose frequencies $\hat{f}_k$ are estimated from the MUSIC spectrum. On the other hand, as shown in Sect. 3.2, the estimated frequencies $\hat{f}_k$ obtained from the MUSIC spectrum are equivalent to the frequencies of the P largest components in the DFT spectrum, i.e., $\hat{f}_k$ satisfy the relation

$$\hat{f}_k \in \left\{0, \frac{1}{N}, \cdots, \frac{N-1}{N}\right\}, \ (k = 0, 1, \cdots, P-1). \tag{25}$$

This relation implies that the computation of (11) can be simplified due to the orthogonal property of $\boldsymbol{s}(\hat{f}_k)$.

Substituting (15) in (11), $\hat{\boldsymbol{x}}$ is rewritten as

$$\hat{\boldsymbol{x}} = \boldsymbol{S}\boldsymbol{b} \tag{26}$$

$$\boldsymbol{b} = (\boldsymbol{S}^H\boldsymbol{S})^{-1}\boldsymbol{S}^H\boldsymbol{W}\boldsymbol{a}(m). \tag{27}$$

In (26) and (27), since the frequencies $\hat{f}_k$ are expressed by (25), the columns of $\boldsymbol{S}$, $\boldsymbol{s}(\hat{f}_k)$, are mutually orthogonal:

$$\boldsymbol{s}^H(\hat{f}_k)\boldsymbol{s}(\hat{f}_l) = \begin{cases} N, \ (\hat{f}_k = \hat{f}_l) \\ 0, \ (\hat{f}_k \neq \hat{f}_l) \end{cases}, \quad (k, l = 0, 1, \cdots, P-1). \tag{28}$$

Then, from (28), the matrix inversion $(\boldsymbol{S}^H\boldsymbol{S})^{-1}$ is expressed in the form

$$(\boldsymbol{S}^H\boldsymbol{S})^{-1} = \frac{1}{N}\boldsymbol{I}. \tag{29}$$

In addition, since $\hat{f}_k$ is given by the relation (25), $\boldsymbol{s}(\hat{f}_k)$ and the columns of $\boldsymbol{W}$, $\boldsymbol{w}_l$, mutually exhibit the orthogonal property as

$$\boldsymbol{s}^H(\hat{f}_k)\boldsymbol{w}_l = \begin{cases} N, \left(\hat{f}_k = \dfrac{l}{N}\right) \\ 0, \left(\hat{f}_k \neq \dfrac{l}{N}\right) \end{cases}, \quad (k = 0, 1, \cdots, P-1; l = 0, 1, \cdots, N-1). \tag{30}$$

Therefore, it is now clear that the vector $\boldsymbol{b}$ given by (27) is composed of the P largest elements of $\boldsymbol{a}(m)$ by the relations (29) and (30). This indicates that the noise reduction algorithm based on a combination of the MUSIC algorithm and the maximum likelihood estimate is similar to the classical threshold technique in which the noise reduction is performed by extracting the relatively large components from the DFT spectrum. In other words, it is said that the classical threshold technique can be derived within the context of the subspace decomposition. Moreover, from the relations (26), (27), (29) and (30), the number of the frequency components which are extracted for reconstructing the target signal is equal to P (i.e., the order of the signal subspace), while such number is determined empirically in the conventional method. The method for estimating the order of the signal subspace P is described later.

In this way, extraction of the P largest components from the DFT spectrum leads to noise reduction. However, the frequency components so extracted contain the noise, since the noise components are spread over all the frequencies. Then, in order to eliminate the noise in the frequency components, we here propose a novel SS method.

By the analogy to (19), the eigenvalues of $\boldsymbol{R}_{yy}$ is given by using the elements of $\boldsymbol{b}$:

$$\lambda_k = \frac{|b_k|^2}{N}, \ (k = 0, 1, \cdots, P-1) \tag{31}$$

where b_k $(k = 0, 1, \cdots, P-1)$ is the k-th row elements of $\boldsymbol{b}$. From (31), b_k can be expressed in terms of λ_k as

$$\begin{aligned} b_k &= |b_k| \times e^{j\angle b_k} \\ &= \sqrt{N\lambda_k} \times e^{j\angle b_k},\ (k = 0, 1, \cdots, P-1). \end{aligned} \tag{32}$$

In (31) and (32), λ_k contains the noise components as in (7). Then, the noise components in λ_k are eliminated by subtracting σ_n^2:

$$\begin{aligned} \lambda'_k &= \lambda_k - \sigma_n^2 \\ &= \frac{|b_k|^2}{N} - \sigma_n^2,\ (k = 0, 1, \cdots, P-1). \end{aligned} \tag{33}$$

Thus, the relation

$$\sqrt{N\lambda'_k} = \sqrt{|b_k|^2 - N\sigma_n^2},\ (k = 0, 1, \cdots, P-1) \tag{34}$$

yields the elimination of the noise components in $|b_k|$. Hence the noise reduction in b_k is performed by

$$b'_k = \sqrt{|b_k|^2 - N\sigma_n^2} \times e^{j\angle b_k},\ (k = 0, 1, \cdots, P-1). \tag{35}$$

Finally, the estimated signal $\hat{\boldsymbol{x}}$ is obtained as

$$\hat{\boldsymbol{x}} = \boldsymbol{S}\boldsymbol{b}' \tag{36}$$

$$\boldsymbol{b}' = [b'_0, b'_1, \cdots, b'_{P-1}]^T. \tag{37}$$

In (35), the proposed method can also be seen as one of the SS methods. Generally, in the conventional SS methods, the statistics of noise process (e.g., the amplitude spectrum of noise) are multiplied by the subtraction factor and then subtracted from the spectrum of the observed signal. However, the optimal choice of the subtraction factor is normally quite hard. Instead of employing the subtraction factor, the proposed method carries out noise reduction by only subtracting $N\sigma_n^2$ from the power spectrum of $\boldsymbol{y}$. As described above, $N\sigma_n^2$ is derived by approximating the MUSIC algorithm. Therefore, the method is efficient not only to alleviate the computational complexity in the MUSIC algorithm combined with the maximum likelihood estimate but also to eliminate the noise components without using the subtraction factor.

3.4 Attenuation of the Processing Distortion

It has been described that the method derived above is classified into the power spectral subtraction due to (35). However, it is well-known that the SS based methods suffer from the self-producing noise, that is, "musical noise." This undesirable artifact greatly deteriorates the intelligibility in the enhanced speech. Therefore, in the SS based methods, the attenuation of musical noise is a key to improve the enhancement quality.

It is known that the musical noise is caused by the random variations in the noise spectrum. This indicates that suppression of such random variations leads to the attenuation of musical noise. Then, in the proposed method, the observed signal is Hanning-windowed to obtain the DFT spectra $\boldsymbol{a}(m)$. This is based on the fact that Hanning-windowing in the time domain is equivalent to the convolution in the frequency domain. Since the convolution by Hanning-windowing in the frequency domain operates to weight a few consecutive frequency bins and then add to the original bins, the noise spectrum is slightly smoothed. Therefore, the variations in the noise spectrum are attenuated by Hanning-windowing. In various SS based methods, Hanning-windowing combined with the overlap-add operation is generally employed to avoid the discontinuity between the adjacent frames. However, in this paper, it is justified that Hanning-windowing combined with the overlap-add operation is effective in order not only to avoid discontinuity but also to attenuate the musical noise.

Moreover, the characteristic differences between musical noise and speech can also be utilized for reducing such undesirable noise. One of the most important characteristics of musical noise is that the majority of the frequency components consisting musical noise have the duration shorter than about 20 [msec], whereas the duration of the speech components is considerably long [4]. Therefore, the frequency components which last no more than 20 [msec] are identified as the musical noise components and eliminated in the enhanced speech.

4. Determining the Order of the Signal Subspace

In the conventional subspace-oriented methods, one of the key issues is to determine the order of the signal subspace P. For determining P, one of the well-known methods is to minimize both the Akaike's information criterion (AIC) [24] and the minimum description length (MDL) [25], [26]. However, in general, the method requires a relatively large number of frames of $\boldsymbol{y}$ in order to obtain the better estimate of P.

Another approach is to determine P from the spread of eigenvalues of $\boldsymbol{R}_{yy}$. As in (8), P can be determined from the number of the eigenvalues which are greater than σ_n^2. In practice, however, the resulting eigenvalues are not approximated to a unique value for a finite analysis frame length. Thus, it seems rather difficult to determine P directly from the spread of eigenvalues.

On the other hand, the spread of eigenvalues of $\boldsymbol{R}_{nn}$ differs from those of $\boldsymbol{R}_{xx}$. To illustrate this, Fig. 2 shows an example of the eigenvalues of $\boldsymbol{R}_{nn}$ and $\boldsymbol{R}_{xx}$. In the figure, the solid and broken lines are respectively the eigenvalues of $\boldsymbol{R}_{nn}$ obtained by using noise (which is assumed to be Gaussian) and those of $\boldsymbol{R}_{xx}$ by using clean speech (vowel /a/ uttered by a Japanese female). The dotted line is the variance (both noise and speech are normalized to variance unity). As shown in the figure, the eigenvalues of $\boldsymbol{R}_{nn}$ relatively concentrate around the variance, while the eigenvalues of $\boldsymbol{R}_{xx}$ are spread from the value 0 to greater than 10 (in this example, the maximum value was about 43). As in Fig. 2, it is seen that the eigenvalues of $\boldsymbol{R}_{nn}$ are close to σ_n^2. Then, from this observation and the hypothesis that the eigenvalues of $\boldsymbol{R}_{yy}$ are obtained from (7), the estimation of P is relatively straightforward, by regarding σ_n^2 as the threshold value for separating the eigenvalues into those associated with the signal and noise subspace. Hence, in this paper, P is deter-

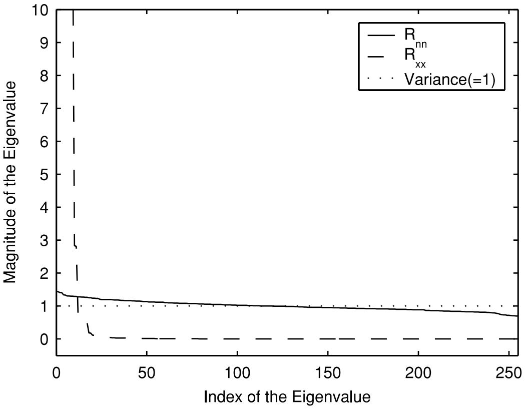


Fig. 2 Eigenvalues of autocorrelation matrices.

mined from the number of the eigenvalues of $\boldsymbol{R}_{yy}$ which are greater than σ_n^2.

In the proposed method, $\boldsymbol{n}$ is assumed to be Gaussian, which is considered to be sufficient to describe general situations. As mentioned in Sect. 2, the central limit theorem implies that the distribution of noise can eventually be approximated to Gaussian when there are multiple noise sources, for example, an air-conditioner, a vehicle, and a factory. Therefore, it is considered that the proposed method is effective to a certain extent in the real environment. In Sect. 6, we will investigate the performance of the proposed method in both the cases of computer generated Gaussian and real recorded noise.

5. Estimating the Variance of Noise Process

In the proposed method (as described in both Sects.3 and 4), the variance of noise process, σ_n^2, is exploited for both noise reduction and subspace decomposition. However, in practice, the true value of σ_n^2 cannot be obtained, since, as mentioned earlier, the length of analysis frame is finite.

Consider that $\boldsymbol{y}$ is composed of only noise, i.e., $\boldsymbol{y} = \boldsymbol{n}$. Then, for a finite frame length, the variance of $\boldsymbol{y}$ cannot be obtained. However, if the analysis frame length of $\boldsymbol{y}$ is sufficiently long, it is satisfactory to use the instantaneous estimate of the variance in each frame of $\boldsymbol{y}$ within the proposed noise reduction method. Therefore, an instantaneous estimate of the variance σ_n^2 is used within the proposed method.

In realistic situations, the noise process is usually nonstationary, while the speech utterance normally consists of separated sentences with multiple of silent periods. Therefore, under the assumption that the variance of the noise process does not vary rapidly in time, σ_n^2 in the speech segments can be regarded as nearly the same as the estimated value in the last segment where the speech is absent but noise is present.

In order to estimate the variance of noise process, the proposed method requires a voice activity detector (VAD) which detects the non-speech segments from speech signals. For the purpose of VAD, the following parameters are commonly utilized: zero-crossing rate, signal energy, or one sample delay correlation coefficient [11], [13], [27], [28]. In general, speech segments are detected by simply comparing the parameters so obtained with the threshold values, which are chosen heuristically or obtained using a certain number of previous frames. However, the threshold value must be varied according to, e.g., the instantaneous SNR or amplitude of $\boldsymbol{y}$. Thus, we propose a simple VAD which does not require the adjustment of such threshold value.

As in (8), the eigenvalues of $\boldsymbol{R}_{yy}$ in the speech segments are given by

$$\lambda_k > \lambda_l, \; (k=0,1,\cdots,P-1; l=P,P+1,\cdots,N-1), \tag{38}$$

while all the eigenvalues are considered to be identical, when $\boldsymbol{y}$ is composed of only noise, namely

$$\lambda_k = \sigma_n^2, \; (k = 0,1,\cdots,N-1). \tag{39}$$

In practice, as in the example in Fig. 2, it is considered that the eigenvalues in the non-speech segments are not approximated to a single value. As mentioned in Sect. 4, however, the eigenvalues of $\boldsymbol{R}_{nn}$ are considered to be nearly constant in comparison with those of $\boldsymbol{R}_{xx}$. Therefore, the difference between the speech activity and silence appears in terms of the spread of eigenvalues. Then, we define the VAD:

$$D_{VAD} = 10\log_{10}\left(\frac{(N-P)\displaystyle\sum_{k=0}^{P-1}|\lambda_k - \hat{\sigma}_n^2|^2}{P\displaystyle\sum_{l=P}^{N-1}|\lambda_l - \hat{\sigma}_n^2|^2} \right) \tag{40}$$

where D_{VAD} indicates that the spread of eigenvalues, $\hat{\sigma}_n^2$ is the estimated variance of noise process obtained from the previous non-speech segment and P is the number of eigenvalues which are greater than $\hat{\sigma}_n^2$. Over several subsequent non-speech segments, D_{VAD} is expected to be nearly a constant value, whereas, in the case of speech activity, D_{VAD} is large. Hence, $\hat{\sigma}_n^2$ is updated as

$$\hat{\sigma}_n^2(m+1) = \begin{cases} \sigma_y^2(m), & \left(D_{VAD}(m) \le D_{threshold}\right) \\ \hat{\sigma}_n^2(m), & \left(D_{VAD}(m) > D_{threshold}\right) \end{cases}, \quad (m = 0,1,\cdots) \tag{41}$$

where $D_{VAD}(m)$, $\hat{\sigma}_n^2(m)$, σ_y^2 $(m = 0,1,\cdots)$, and $D_{threshold}$ are the spread of eigenvalues, the estimated variance of noise process, the variance of the observed signal in the m-th frame, and the threshold value for the VAD, respectively. Since the proposed VAD is based on the spread of eigenvalues of the autocorrelation matrix, it is not necessary to vary $D_{threshold}$ according to the instantaneous SNR or amplitude of $\boldsymbol{y}$. In the method, under the assumption that the observed signal does not begin immediately with speech, $D_{threshold}$ is determined by averaging $D_{VAD}(m)$ in the first few frames of the observed signal. In addition, in this paper, the variance of the observed signal in the first frame is used for giving the initial value $\hat{\sigma}_n^2(0)$.

6. Simulation Study

6.1 Parameter Settings

In the simulation study, the performance obtained by the proposed method was compared with the conventional SS method (SS), NSS, and the MUSIC algorithm combined with the maximum likelihood estimate (MUSIC+MLE). In SS, the VAD proposed in Sect. 5 was used for examining the performance of the proposed VAD. In the case of NSS, in order to see how the performance varies, two different parameter settings shown in Table 1 were attempted: the parameters were optimized to eliminate the residual noise (NSS1) and attenuate the distortion of speech (NSS2). These parameters were set by the separate simulation study. In addition, the over subtraction factor in NSS was adjusted as the function of SNR at each frequency (see, e.g., [4]). In MUSIC+MLE, the order of the signal subspace P was determined by using the method in which the AIC was minimized [24], [26].

For the speech signals $\boldsymbol{x}$, the utterances by three male and two female speakers were used. Each utterance was the speech "Sakura ga saita" in Japanese, sampled originally at 44.1 [kHz], and then down-sampled to 11.025 [kHz].

In order to validate the proposed method, we investigated the following two cases for the noise signal $\boldsymbol{n}$: the noise components are 1) the random variables generated from Gaussian distribution and 2) real fan noise signals. Both the noise signals were assumed to be zero-mean and variance unity. Then, the amplitude of both the noise signals was adjusted according to the input SNR from −5 dB to 15 dB.

The observed signal was divided into a multiple number of frames by applying an overlap-add window. The length of each frame was $N = 512$ and the adjacent frames were overlapped at every $N/2 = 256$ samples for giving a good trade-off in terms of the performance and the computational complexity. For determining $D_{threshold}$, the first four frames in the observed signal were used.

6.2 Performance Measurements

For the evaluation of the enhancement quality, the commonly used objective measurements in terms of both the segmental SNR [29] and the averaged cepstral distance [30] were considered:

Table 1 Parameters used for NSS.

Setting Name	NSS1	NSS2
DFT Length	512	1024
Decimation Ratio	128	256
Window Length for Minimum Search	4	5
Spectral Floor Constant	0.02	0.05
Smoothing Constant for Signal Power Estimate	0.75	0.85
Smoothing Constant for Noise Power Estimate	0.89	0.63
Overestimation factor	1.5	2

The segmental SNR is defined as

$$SNR_{seg} = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\frac{\sum_{k=0}^{N_{SNR}} x_m^2(k)}{\sum_{k=0}^{N_{SNR}} (x_m(k) - \hat{x}_m(k))^2} \tag{42}$$

where $x_m(k)$ and $\hat{x}_m(k)$ are respectively the original and the estimated speech at the m-th frame, N_{SNR} is the length of analysis frames (set to 256), and M is the number of frames where speech is present. The determination of speech presence was achieved by manual inspection of the clean speech.

The averaged cepstral distance is given by

$$d_{cep} = \frac{1}{M}\sum_{m=0}^{M-1}\sum_{k=0}^{2q-1}\left(c_m(k) - \hat{c}_m(k)\right)^2 \tag{43}$$

where $c_m(k)$ and $\hat{c}_m(k)$ are respectively the cepstral coefficients corresponding to the original speech and the enhanced speech, and q is the order of the model (chosen to 8).

6.3 Simulation Results

6.3.1 Gaussian Noise Case

Figure 3 shows the simulation results using the speech sample uttered by a female speaker. As shown in Figs. 3(d) and 3(e), it is observed that the enhanced speech by NSS1 is greatly attenuated and the waveform is greatly distorted in comparison with the original speech shown in Fig. 3(a), while the residual noise is noticeably observed from the enhanced speech by NSS2. From these results, it is shown that NSS cannot suppress the distortion of speech and the residual noise simultaneously. On the other hand, as in Fig. 3(f), although MUSIC+MLE attenuates both the distortion of speech and the noise components in comparison with NSS, the residual noise is still observed. In contrast, in Fig. 3(g), the proposed method substantially reduces the noise components while preserving the overall shape of the speech.

In Figs. 4, (a) and (b) respectively show the performance comparison in terms of the segmental SNR and the averaged cepstral distance. In the figures, both the segmental SNR and the averaged cepstral distance were averaged over the five speech samples. As shown in the figures, both the performances obtained by the combination of the conventional SS and the proposed VAD (SS) are better than those by NSS1 and NSS2. From this, it is considered that the VAD proposed in Sect. 5 gives better performance of estimating the statistics of the noise process as compared with NSS even in the low SNR environments. In addition, as in Fig. 4, though the same VAD was employed, the performance obtained by the proposed method is superior to that by SS. Moreover, the proposed method improves the segmental SNR and the cepstral distance around 2 dB and 0.2,

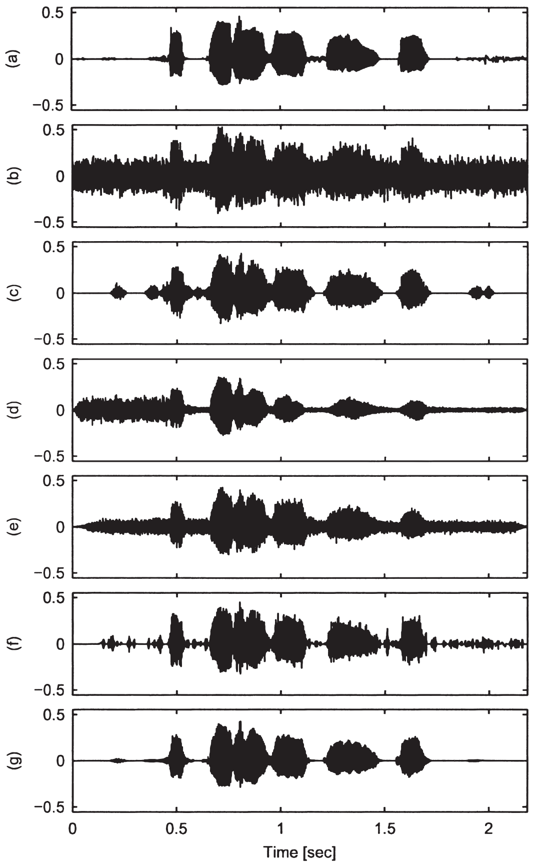
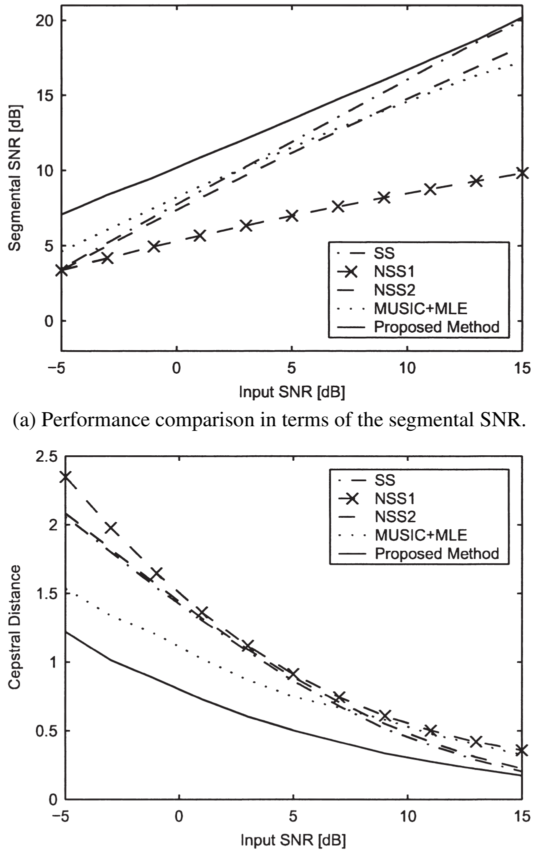


Fig. 3 Simulation results of the speech sample uttered by a female speaker (Gaussian noise case): (a) Original speech, (b) Noisy speech (Input SNR=0 dB), (c) Enhanced speech by SS, (d) Enhanced speech by NSS1, (e) Enhanced speech by NSS2, (f) Enhanced speech by MUSIC+MLE, and (g) Enhanced speech by the proposed method.

respectively, as compared with MUSIC+MLE. These indicate that both the speech and noise signals are appropriately modeled by the method proposed in Sect. 3. As a result, the proposed method gives the best performance of the five methods. Especially, at input SNRs<0 dB in Fig. 4, the proposed method improves the segmental SNR more than 3 dB and the averaged cepstral distance obtained by the proposed method is more than 40% shorter in comparison with the conventional SS based methods.

Table 2 shows the performance comparison in terms of the ratio of computation. In the table, the ratio of computation was normalized to SS=1. As in the table, it is clearly seen that the proposed method reduced the ratio of computation to around 0.1% as compared with MUSIC+MLE. In addition, in comparison with both the NSS1 and NSS2, the computational load is alleviated by the proposed method. Therefore, it is shown that the proposed method can be well suited to the real-time application.

In the informal listening tests, it was confirmed that in both the cases of SS and MUSIC+MLE, the residual noise is observed and it is noticeable. By the residual noise, the intelligibility of speech was considerably degraded, particularly in very low SNR environments as input SNRs<0 dB. On the other hand, the listening tests showed that NSS1 deteriorates the intelligibility of speech, since NSS1 reduces not only the noise but also the speech components, while with the NSS2, the residual noise is still heard. Both the degradation of intelligibility by NSS1 and the residual noise by NSS2 were noticeable especially at the input SNRs<5 dB. In contrast, in the listening tests, it was confirmed that the proposed method eliminates only the noise components with much less distortion. At the input SNRs<5 dB, in the enhanced speech by the proposed method, less amount of musical noise was observed as compared with that by the conventional SS based methods.



(a) Performance comparison in terms of the segmental SNR.

(b) Performance comparison in terms of the averaged cepstral distance.

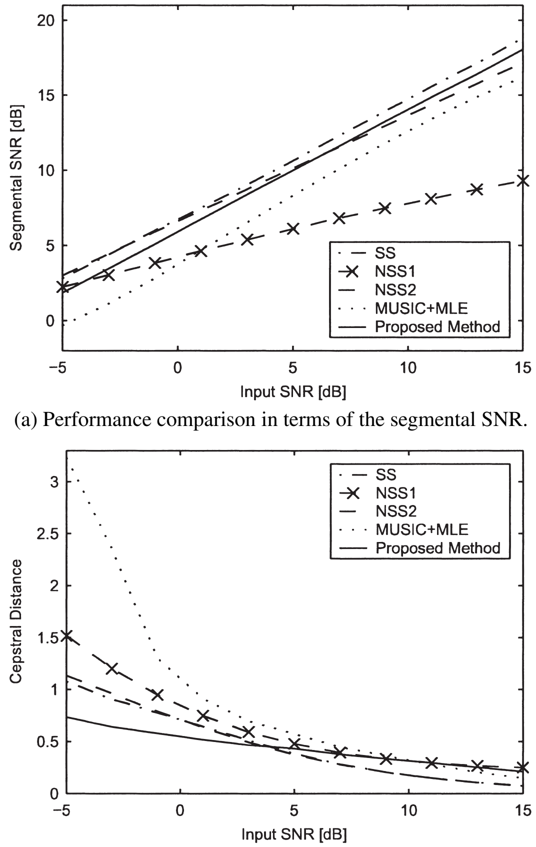
Fig. 4 Performance comparison in the case of Gaussian noise.

Table 2 Performance comparison in terms of the ratio of computation.

	Ratio of Computation (Normalized to SS=1)
SS	1
NSS1	2.2
NSS2	2.2
MUSIC+MLE	1101.8
Proposed Method	1.1

6.3.2 Real Fan Noise Case

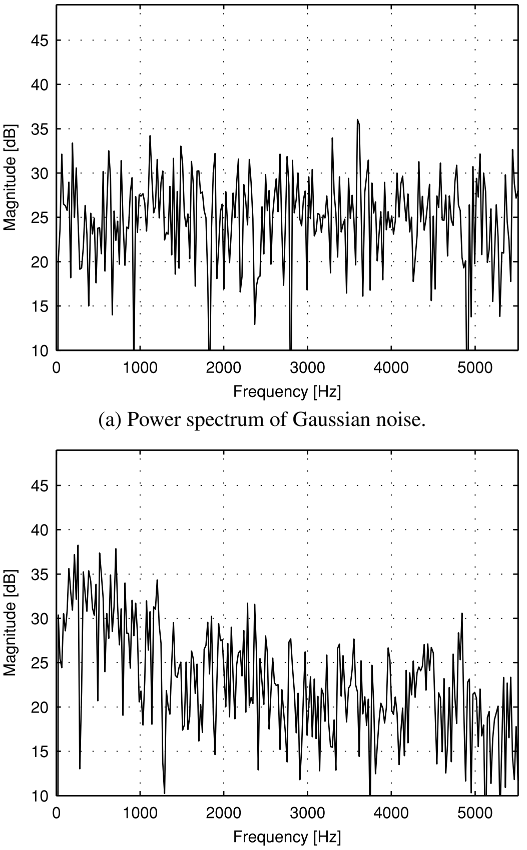
Figure 5 shows the performance comparison in the case of using the real fan noise. In the figure, as in Fig. 4, the re-



(a) Performance comparison in terms of the segmental SNR.

(b) Performance comparison in terms of the averaged cepstral distance.

Fig. 5 Performance comparison in the case of real fan noise.



(a) Power spectrum of Gaussian noise.

(b) Power spectrum of real fan noise.

Fig. 6 Comparison of the noise spectra.

sults shown are averaged over the five speech samples. As shown in Figs. 5(a) and 5(b), both the performances with SS are better than those with NSS1 and NSS2. Therefore, it was confirmed that the proposed VAD is still effective even in the real environment. In contrast, as in Fig. 5(a), at input SNRs>2 dB, SS gives the best performance of the five method, and at input SNRs<5 dB, the segmental SNR obtained by the proposed method is slightly degraded in comparison with that by both SS and NSS2. On the other hand, in Fig. 5(b), at the input SNRs<4 dB, the performance in terms of the averaged cepstral distance obtained by the proposed method is the best of all of five methods. However, at higher input SNRs, the averaged cepstral distance obtained by both SS and NSS2 is the best of performance in terms of the five methods.

In the informal listening tests, it was seen that in SS, NSS1, NSS2, and MUSIC+MLE, the results are similar to the case of the Gaussian noise, i.e., SS, NSS2, and MUSIC+MLE) the residual noise was observed, and NSS1) the intelligibility of speech was degraded. In contrast, the listening tests showed that the enhanced speech obtained by the proposed method somewhat sounds like a mixture of the original speech and low-pass filtered noise. It is considered that the main cause of the residual noise in the enhanced speech is due to the fact that the power of the fan noise used in the simulation study rather concentrates in a relatively lower frequency range. To illustrate this, Fig. 6 shows the comparison between the Gaussian noise and the real fan noise spectra. As shown in the figure, it is observed that the real fan noise is composed of the large lower frequency components and, in contrast, the small higher frequency components, whereas the power of the Gaussian noise spreads uniformly over all the frequencies. Since the proposed method subtracts the variance of the noise from the power spectrum of the noisy speech uniformly by (35), it is considered that the lower frequency components of the noise are left in the enhanced speech. It is also considered that deterioration in the enhancement quality by the proposed method as in Fig. 5 is caused by the residual noise in the enhanced speech. Therefore, it is inevitable that the model of noise described in this paper is expanded into the case of non-Gaussian distribution in order to obtain a further improvement.

7. Conclusion

A novel method of noise reduction for speech signals using the SS based on the subspace decomposition has been proposed. In this paper, by approximating the orthonormal bases, which span both the signal and noise subspace, to the Fourier bases, the relation between the DFT and MUSIC spectra has been derived. Then, we have shown that the noise reduction algorithm using the MUSIC algorithm combined with the maximum likelihood estimate results in a novel power spectral subtraction method by exploiting the relation between the DFT and MUSIC spectra. Moreover, for estimating the variance of noise process, we have also proposed the robust VAD based on the spread of eigenvalues of the autocorrelation matrix.

In the simulation, it has been observed that the enhancement quality obtained by our method is superior to the quality obtained by the NSS in the case of the Gaussian noise, while in the case of the real fan noise, the proposed method is effective to a certain extent. In addition, the simulation results have shown that the proposed method is well suited to the real-time implementation. In the informal listening tests, it has been confirmed that the proposed method effectively reduces the noise components with much less distortion in the enhanced speech as compared with the NSS. Future work includes a thorough investigation in order to obtain a further enhancement in the case of the non-Gaussian noise.

Acknowledgement

The authors would like to thank Dr. Toshihisa Tanaka, the Department of Electrical and Electronic Engineering, Tokyo University of Agriculture and Technology (TUAT), for the fruitful discussions to improve the draft of the original manuscript. The authors are also grateful to the anonymous reviewers for their helpful suggestions in improving the quality of this paper. This research is supported by a Grand-in-Aid for JSPS Fellows and JSPS Research Fellowships for Young Scientists.

References

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

[2] R. Martin, “Spectral subtraction based on minimum statistics,” Proc. EUSPICO-94, pp.1182–1185, Edinburgh, 1994.

[3] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” IEEE Trans. Speech Audio Process., vol.9, no.5, pp.484–487, July 2001.

[4] S.V. Vaseghi, Advanced Signal Processing and Digital Noise Reduction, John Wiley & Sons, New York, 2000.

[5] 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, July 1995.

[6] E. Grivel, M. Gabrea, and M. Najim, “Speech enhancement as a realisation issue,” Signal Process., vol.82, pp.1963–1978, 1997.

[7] 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, Sept. 2000.

[8] M. Dendrinos, S. Bakamidis, and G. Carayannis, “Speech enhancement from noise: A regenerative approach,” Speech Commun., vol.10, pp.45–47, Feb. 1991.

[9] J. Huang and Y. Zhao, “An energy-constrained signal subspace method for speech enhancement and recognition in white and colored noises,” Speech Commun., vol.26, pp.165–181, Nov. 1998.

[10] S. Doclo and M. Moonen, “SVD-based optimal filtering with application to noise reduction in speech signals,” IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp.143–146, New Plats, New York, Oct. 1999.

[11] A. Rezayee and S. Gazor, “An adaptive KLT approach for speech enhancement,” IEEE Trans. Speech Audio Process., vol.9, no.2, pp.87–95, Feb. 2001.

[12] U. Mittal and N. Phamdo, “Signal/noise KLT based approach for enhancing speech degraded by colored noise,” IEEE Trans. Speech Audio Process., vol.8, no.2, pp.159–167, March 2000.

[13] S. Affers and Y. Grenier, “A signal subspace tracking algorithm for microphone array processing of speech,” IEEE Trans. Speech Audio Process., vol.5, no.5, pp.425–437, Sept. 1997.

[14] F. Asano and S. Hayamizu, “Speech enhancement using array signal processing based on the coherent-subspace method,” IEICE Trans. Fundamentals, vol.E80-A, no.11, pp.2276–2285, Nov. 1997.

[15] D.-Z. Feng, Z. Bao, and X.-D. Zhang, “A bi-iteration instrumental variable noise-subspace tracking algorithm,” Signal Process., vol.81, pp.2215–2221, 2001.

[16] C.E. Davila, “Efficient, high performance, subspace tracking for time-domain data,” IEEE Trans. Signal Process., vol.48, no.12, pp.3307–3315, Dec. 2000.

[17] J.-L. Yu, “A novel subspace tracking using correlation based projection approximation,” Signal Process., vol.80, pp.2517–2525, 2000.

[18] K.V.S. Babu, Y. Yoganandam, and V.U. Reddy, “Adaptive estimation of eigensubspace and tracking the directions of arrival,” Signal Process., vol.68, pp.317–339, 1998.

[19] M.A. Hasan, “DOA and frequency estimation using fast subspace algorithm,” Signal Process., vol.77, pp.49–62, 1999.

[20] P. Strobach, “Square Hankel SVD subspace tracking algorithms,” Signal Process., vol.57, pp.1–18, 1997.

[21] R.O. Schmidt, “Multiple emitter location and signal parameter estimation,” IEEE Trans. Antennas Propag., vol.AP-34, no.3, pp.276–280, March 1986.

[22] A. Papoulis and S. Unnikrishna Pillai, Probability, Random Variables and Stochastic Processes, McGraw-Hill, New York, 2002.

[23] N. Kikuma, Adaptive Signal Processing with Array Antenna, Science and Technology Publishing Company, Tokyo, 1998.

[24] H. Akaike, “A new look at the statistical model identification,” IEEE Trans. Autom. Control, vol.AC-19, no.6, pp.716–723, Dec. 1974.

[25] J. Rissanen, “Modeling by shortest data description,” Automatica, vol.14, pp.465–471, 1978.

[26] M. Wax and T. Kailath, “Detection of signals by information theoretic criteria,” IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-33, no.2, pp.387–392, April 1985.

[27] R. Le Bouquin-Jeannes and G. Faucon, “Study of a voice activity detector and its influence on a noise reduction system,” Speech Commun., vol.16, no.3, pp.245–254, 1995.

[28] M.H. Savoji, “A robust algorithm for accurate end-pointing of speech signals,” Speech Commun., vol.8, no.1, pp.45–60, 1989.

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

[30] R. Le Bouquin-Jennes, A. Akbari Azirani, and G. Faucon, “Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator,” IEEE Trans. Speech Audio Process., vol.5, no.5, pp.484–487, Sept. 1997.

Appendix: Approximation of the ED

The relations (19) and (20) are derived under the assumption (14) as follows:

By substituting (15) in (13), $\boldsymbol{R}_{yy}$ is rewritten by

$$\boldsymbol{R}_{yy} = \frac{1}{M} \sum_{m=0}^{M-1} \Big(\boldsymbol{W}\boldsymbol{a}(m)\Big)\Big(\boldsymbol{W}\boldsymbol{a}(m)\Big)^{H}$$

$$
\begin{aligned}
&= \boldsymbol{W}\left(\frac{1}{M}\sum_{m=0}^{M-1}\boldsymbol{a}(m)\boldsymbol{a}^H(m)\right)\boldsymbol{W}^H \\
&= \boldsymbol{W}\boldsymbol{R}_{aa}\boldsymbol{W}^H \qquad (\mathrm{A}\cdot 1)
\end{aligned}
$$

where

$$\boldsymbol{R}_{aa} = \frac{1}{M}\sum_{m=0}^{M-1}\boldsymbol{a}(m)\boldsymbol{a}^H(m) \qquad (\mathrm{A}\cdot 2)$$

is the autocorrelation matrix of $\boldsymbol{a}(m)$. In (A· 1), the columns of $\boldsymbol{W}$ (i.e., the Fourier basis vectors $\boldsymbol{w}_k$) are mutually orthogonal as

$$\boldsymbol{w}_k^H\boldsymbol{w}_l = \begin{cases} N, & (k = l) \\ 0, & (k \neq l) \end{cases}, \ (k, l = 0, 1, \cdots, N-1), \qquad (\mathrm{A}\cdot 3)$$

since the elements of $\boldsymbol{w}_k$ are equally spaced in angle around the unit circle in the complex plane. This orthogonal property indicates that

$$\boldsymbol{W}^H\boldsymbol{W} = N\boldsymbol{I}. \qquad (\mathrm{A}\cdot 4)$$

Therefore, $\boldsymbol{W}^H$ is given by

$$\boldsymbol{W}^H = N\boldsymbol{W}^{-1}. \qquad (\mathrm{A}\cdot 5)$$

Substituting (A· 5) in (A· 1), we have

$$
\begin{aligned}
\boldsymbol{R}_{yy} &= \boldsymbol{W}\boldsymbol{R}_{aa}(N\boldsymbol{W}^{-1}) \\
&= \boldsymbol{W}(N\boldsymbol{R}_{aa})\boldsymbol{W}^{-1}. \qquad (\mathrm{A}\cdot 6)
\end{aligned}
$$

The relation (A· 6) has a similar structure to (6). However, in (A· 6), $N\boldsymbol{R}_{aa}$ is not a diagonal matrix in general, since, by substituting (18) in (A· 2), the elements of $\boldsymbol{R}_{aa}$ are expressed as

$$
\begin{aligned}
r_{kl} &= \frac{1}{M}\frac{1}{N^2}\sum_{m=0}^{M-1} Y(k;m)\overline{Y(l;m)}, \\
&(k, l = 0, 1, \cdots, N-1) \qquad (\mathrm{A}\cdot 7)
\end{aligned}
$$

where r_{kl} is the k-th row and l-th column element of $\boldsymbol{R}_{aa}$ and $\overline{Y(l;m)}$ is the complex conjugate of $Y(l;m)$. In contrast, under the assumption (14), $\boldsymbol{R}_{aa}$ can be diagonalized. From (14), $Y(k;m)$ is given by

$$
\begin{aligned}
&Y(k;m) = Y(k;0)e^{j2\pi km/N}, \\
&(k = 0, 1, \cdots, N-1; m = 0, 1, \cdots, M-1). \qquad (\mathrm{A}\cdot 8)
\end{aligned}
$$

Substituting (A· 8) in (A· 7), r_{kl} is rewritten as

$$
\begin{aligned}
r_{kl} &= \frac{1}{M}\frac{1}{N^2}\sum_{m=0}^{M-1} Y(k;0)\overline{Y(l;0)}e^{j2\pi(k-l)m/N}, \\
&(k, l = 0, 1, \cdots, N-1; m = 0, 1, \cdots, M-1). \qquad (\mathrm{A}\cdot 9)
\end{aligned}
$$

Since $e^{j2\pi(k-l)m/N}$ $(m = 0, 1, \cdots, M-1)$ are equally spaced in angle around the unit circle in the complex plane, $\sum_{m=0}^{M-1} Y(k;0)\overline{Y(l;0)}e^{j2\pi(k-l)m/N}$ in (A· 9) can be represented by

$$
\begin{aligned}
&\sum_{m=0}^{M-1} Y(k;0)\overline{Y(l;0)}e^{j2\pi(k-l)m/N} \\
&= \begin{cases} M|Y(k;0)|^2, & (k = l) \\ \displaystyle\sum_{m=0}^{Q-1} Y(k;0)\overline{Y(l;0)}e^{j2\pi(k-l)m/N}, & (k \neq l) \end{cases}, \\
&(k, l = 0, 1, \cdots, N-1) \qquad (\mathrm{A}\cdot 10)
\end{aligned}
$$

where $Q = M \bmod N$. Hence we have

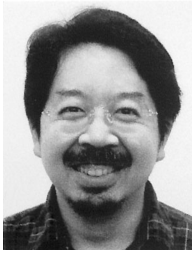
$$
\begin{aligned}
&\lim_{M\to\infty} r_{kl} = \begin{cases} \dfrac{|Y(k;0)|^2}{N^2}, & (k = l) \\ 0, & (k \neq l) \end{cases}, \\
&(k, l = 0, 1, \cdots, N-1). \qquad (\mathrm{A}\cdot 11)
\end{aligned}
$$

From the relation (A· 11), $\boldsymbol{R}_{aa}$ is given in terms of the diagonal matrix:

$$\boldsymbol{R}_{aa} = diag\left(\frac{|Y(0;0)|^2}{N^2}, \frac{|Y(1;0)|^2}{N^2}, \cdots, \frac{|Y(N-1;0)|^2}{N^2}\right). \qquad (\mathrm{A}\cdot 12)$$

Therefore, under the assumption (14), the relation (A· 6) represents the ED of $\boldsymbol{R}_{yy}$. Thus, the eigenvalues and eigenvectors of $\boldsymbol{R}_{yy}$ are respectively approximated to the relations (19) and (20).

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.



Tetsuya Hoya 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, in electrical engineering. He received the Ph.D. degree from Imperial College of Science, Technology and Medicine, University of 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, from 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 so far and is the author of the book *Artificial Mind System—Kernel Memory Approach* (to appear from Springer-Verlag). He is a member of IEEE and was a committee member of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA-2003).

Yoshihisa Ishida received the B.Sc., M.Sc., and Ph.D. degrees in Electrical Engineering from Meiji University, Kawasaki, Japan, in 1970, 1972 and 1978, respectively. In 1975 he joined the Department of Electrical Engineering, Meiji University, as a Research assistant. He then became a Lecturer and an Associate Professor in 1978 and 1981, respectively. He is currently a Professor at the Department of Electronics and Communications, Meiji University. His current research interests are in the area of digital signal processing, speech analysis. He is a member of IEEE, ASJ and an editing committee member of Japan Fluid Power System Society.