SPEECH ENHANCEMENT USING CSS-BASED ARRAY PROCESSING

Futoshi Asano *Satoru Hayamizu*

Electrotechnical Laboratory
1-1-4, Umezono, Tsukuba, 305, Japan
asano(/hayamizu)@etl.go.jp

ABSTRACT

A method for recovering the LPC spectrum from a microphone array input signal corrupted by ambient noise is proposed. This method is based on the CSS (coherent subspace) method, which is designed for DOA (direction of arrival) estimation of broadband array input signals. The noise energy is reduced in the subspace domain by the maximum likelihood method. To enhance the performance of noise reduction, elimination of noise-dominant subspace using projection is further employed, which is effective when the SNR is low and classification of noise and signals in the subspace domain is difficult. The results of the simulation show that some small formants, which cannot be estimated by the conventional delay-and-sum beamformer, were well estimated by the proposed method.

1. INTRODUCTION

When using speech recognition systems in noisy environments, speech enhancement/noise reduction techniques are indispensable. Adaptive beamforming techniques are quite effective for reducing directional noise. For less-directional noise such as ambient noise, however, the performance of the system is similar to or less than that of the classical delay-and-sum beamformer and is not satisfactory [1]. In this paper, a method for recovering the LPC spectrum of speech from microphone-array inputs, which are corrupted by less-directional ambient noise, is proposed. This method is based on a combination of the coherent subspace (CSS) method [2] and the maximum likelihood method. The CSS method, which was developed in the field of DOA (direction of arrival) estimation, classifies signals and noise in the subspace domain. In this paper, this classification is applied to speech enhancement. However, the classification is effective only when SNR is relatively high. Therefore, in this paper, a method using projection for classifying signals and noise in the subspace domain even when SNR is low is also proposed. Using this, the noise-dominated subspace can be effectively eliminated from the correlation matrix.

2. MODEL OF SIGNAL

Let us assume that there are D directional sources and omni-directional ambient noise in the environment. A mixture of these signals and noise is observed by M microphones. The input vector, which consists of the DFT of the input signal at each microphone $X_m(\omega_j)$, is expressed as

$$\mathbf{x}(\omega_j) = [X_1(\omega_j), \cdots, X_M(\omega_j)]^T \quad (1)$$
$$= \mathbf{A}(\omega_j)\mathbf{s}(\omega_j) + \mathbf{n}(\omega_j) \quad (2)$$

Matrix $\mathbf{A}(\omega_j)$ and vector $\mathbf{s}(\omega_j)$ are defined as

$$\mathbf{A}(\omega_j) = [\mathbf{a}_1(\omega_j) \cdots \mathbf{a}_D(\omega_j)], \quad (3)$$
$$\mathbf{s}(\omega_j) = [S_1(\omega_j) \cdots S_D(\omega_j)]^T, \quad (4)$$

where $S_i(\omega_j)$ denotes the short time DFT of the ith directional signal. The directional vector, $\mathbf{a}_i(\omega)$, consists of time-delays for the ith signal at microphones as

$$\mathbf{a}_i(\omega_j) = [e^{-j\omega_j\tau_1(\theta_i)} \cdots e^{-j\omega_j\tau_M(\theta_i)}]^T. \quad (5)$$

The noise vector, $\mathbf{n}(\omega_j)$, consists of the DFT of the ambient noise at microphones. The spatial correlation matrix, $\mathbf{R}(\omega_j) = E[\mathbf{x}(\omega_j)\mathbf{x}^H(\omega_j)]$, is then written as

$$\mathbf{R}(\omega_j) = \mathbf{A}(\omega_j)\mathbf{C}(\omega_j)\mathbf{A}(\omega_j)^H + \sigma(\omega_j)^2\mathbf{K}(\omega_j), \quad (6)$$

where $\sigma(\omega_j)^2$ and $\mathbf{K}(\omega_j)$ denote the power and the covariance of the noise $\mathbf{n}(\omega_j)$, respectively. The symbol $\mathbf{C}(\omega_j)$ denotes the cross-spectrum matrix of source defined by $\mathbf{C}(\omega_j) = E[\mathbf{s}(\omega_j)\mathbf{s}^H(\omega_j)]$. Hereafter, the matrices at the discrete frequency of ω_j, $\mathbf{A}(\omega_j)$, $\mathbf{R}(\omega_j)$, $\mathbf{C}(\omega_j)$ and $\mathbf{K}(\omega_j)$ are denoted as $\mathbf{A}_j$, $\mathbf{R}_j$, $\mathbf{C}_j$ and $\mathbf{K}_j$, respectively, for the sake of simplicity.

3. SUBSPACE METHOD

3.1. Subspace

Consider the following generalized eigenvalue problem:

$$\mathbf{R}_j\mathbf{E}_j = \mathbf{K}_j\mathbf{E}_j\mathbf{\Lambda}_j \quad (7)$$

The symbols $\mathbf{E}_j$ and $\mathbf{\Lambda}_j$ are the eigenvector matrix and eigenvalue matrix defined as

$$\mathbf{E}_j = [\mathbf{e}_1(\omega_j) \cdots \mathbf{e}_M(\omega_j)], \quad (8)$$
$$\mathbf{\Lambda}_j = \mathrm{diag}[\lambda_1(\omega_j), \cdots, \lambda_M(\omega_j)], \quad (9)$$

where $\mathbf{e}_m(\omega_j)$ and $\lambda_m(\omega_j)$ denote the eigenvector and eigenvalue of $\mathbf{R}_j$, respectively. When the number of directional sources is smaller than the number of microphones, i.e., $D < M$, the eigenvectors can be divided as follows:

$$\mathbf{E}_j^s = [\mathbf{e}_1(\omega_j) \cdots \mathbf{e}_D(\omega_j)] \quad (10)$$
$$\mathbf{E}_j^n = [\mathbf{e}_{D+1}(\omega_j) \cdots \mathbf{e}_M(\omega_j)]. \quad (11)$$

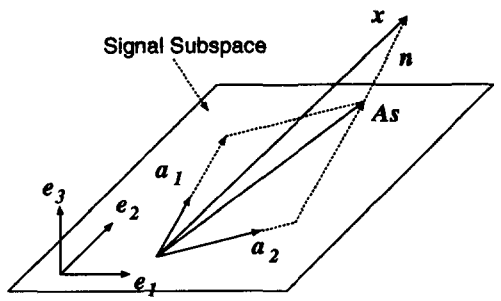

Figure 1. Relation of vectors and signal subspace.

The vector $\mathbf{e}_m \in \mathbf{E}_j^s$ becomes the basis of the signal subspace spanned by $\{\mathbf{a}_1(\omega_j), \ldots, \mathbf{a}_D(\omega_j)\}$. Also, the subspace spanned by $\mathbf{e}_m(\omega_j) \in \mathbf{E}_j^n$ is termed noise subspace and

$$\Re\{\mathbf{A}_j\} = \Re\{\mathbf{E}_j^s\} = \Re\{\mathbf{E}_j^n\}^{\perp} \tag{12}$$

where $\Re\{\cdot\}$ denotes the row space (e.g., [3]). The relation of vectors and subspace when $D = 2$ is depicted in Fig. 1.

3.2. Maximum Likelihood Estimator

To estimate the source cross-spectrum matrix, $\mathbf{C}_j$, the following maximum likelihood estimator is proposed [3].

$$\mathbf{C}_j = \mathbf{W}_j^H \mathbf{R}_j \mathbf{W}_j \tag{13}$$

where

$$\mathbf{W}_j = \mathbf{K}_j^{-1}\mathbf{A}_j[\mathbf{A}_j^H\mathbf{K}_j^{-1}\mathbf{A}_j]^{-1} \tag{14}$$

The speech spectrum from the dth source can be recovered from the dth diagonal element of $\mathbf{C}_j$, $[\mathbf{C}_j]_{dd}$.

The mechanism of noise reduction is explained as follows: As described in the previous section, any vectors in noise subspace are orthogonal to $\Re\{\mathbf{A}_j\}$. Due to this orthogonality, a component of noise that lies in the noise subspace vanishes when Eq.(13) is applied. When $\bar{\mathbf{K}}_j = \mathbf{I}$ in Eq.(14) where $\mathbf{I}$ is an identity matrix, the maximum likelihood method becomes a simple delay-and-sum beamformer.

4. ESTIMATION OF LPC COEFFICIENTS USING CSS

For the stable estimation of $\mathbf{R}_j$ for each individual frequency ω_j, averaging over a large number of frames is required [2]. This characteristic is not acceptable for frame based processing such as speech recognition. To overcome this problem, in the CSS method, averaging in the frequency domain is substituted for averaging in the time domain as

$$\bar{\mathbf{R}}_{j_0} = \sum_{j=j_0-L/2}^{j_0+L/2} \mathbf{T}_j \mathbf{R}_j \mathbf{T}_j^H, \tag{15}$$

where $\mathbf{T}_j$ is termed *focusing matrix* [2]. In Eq.(15), the correlation matrix $\mathbf{R}_j$ is averaged over the frequency index range of $[j_0 - L/2, \cdots, j_0 + L/2]$ with a center frequency of ω_{j_0}.

By using the band-averaged correlation matrix $\bar{\mathbf{R}}_{j_0}$ obtained from the CSS, the band-averaged cross-spectrum matrix can be calculated in the same manner as Eq.(13) as

$$\bar{\mathbf{C}}_{j_0} = \bar{\mathbf{W}}_{j_0}^H \bar{\mathbf{R}}_{j_0} \bar{\mathbf{W}}_{j_0} \tag{16}$$

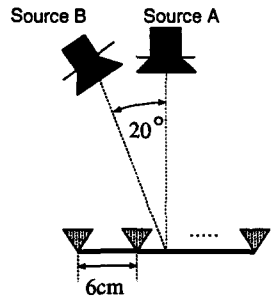

Figure 2. Configuration of the two directional sources and the microphone array.

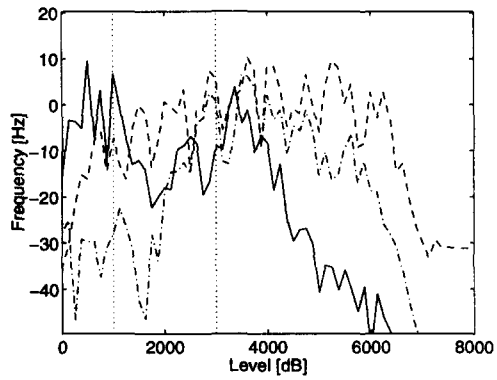

Figure 3. Input spectrum. Solid line: speech A; Dash-dotted line: speech B; Dashed line: ambient noise.

for $\omega_{j_0} = 2\pi j_0/N_a$, $j_0 = 0, \cdots, N_a - 1$. The symbol, N_a denotes the total number of the discrete center frequency. In Eq.(16),

$$\bar{\mathbf{W}}_{j_0} = \bar{\mathbf{K}}_{j_0}^{-1}\mathbf{A}_{j_0}\left[\mathbf{A}_{j_0}^H\bar{\mathbf{K}}_{j_0}^{-1}\mathbf{A}_{j_0}\right]^{-1} \tag{17}$$

$$\bar{\mathbf{K}}_{j_0} = \sum_{j=j_0-L/2}^{j_0+L/2} \mathbf{T}_j\mathbf{K}_j\mathbf{T}_j^H \tag{18}$$

The band-averaged spectrum of the dth source can be extracted from $\bar{\mathbf{C}}_{j_0}$ as

$$\bar{P}_d(\omega_{j_0}) = [\bar{\mathbf{C}}_{j_0}]_{dd} \tag{19}$$

where $[\bar{\mathbf{C}}_{j_0}]_{dd}$ denotes the (d,d) diagonal element of $\bar{\mathbf{C}}_{j_0}$. From this, the autocorrelation function can be obtained as

$$r(i) = F_{N_a}^{-1}[P_d(\omega_{j_0})] \tag{20}$$

where $F_{N_a}^{-1}[\cdot]$ shows the N_a-point inverse DFT. The LPC coefficients are obtained from the normal equation consisting of $r(i)$.

5. NOISE SUBSPACE REDUCTION USING PROJECTION

When SNR is low, however, even the classification of noise and signal subspaces is difficult, resulting in poor performance. When SNR is low, the following projection technique is useful. Assuming that the directional vector for the

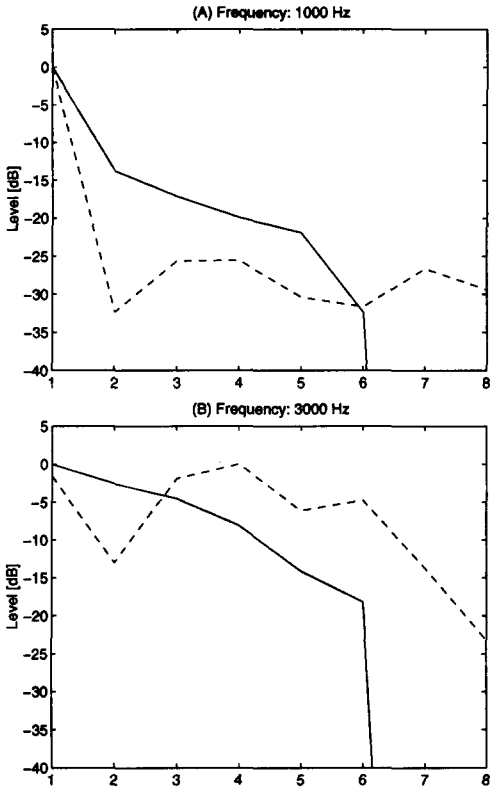

Figure 4. Eigenvalues (solid line) and the contribution c_i (dashed line).

target speech, $\hat{\mathbf{a}}_d(\omega_{j_0})$, has been obtained through the DOA estimation process, the projection of $\hat{\mathbf{a}}_d$ onto the eigenvector $\bar{\mathbf{e}}_i$ of $\bar{\mathbf{R}}_{j_0}$ is

$$\mathbf{p}_i = [\bar{\mathbf{e}}_i^H \hat{\mathbf{a}}_d]\bar{\mathbf{e}}_i = C_i \bar{\mathbf{e}}_i. \tag{21}$$

The frequency index (ω_{j_0}) is omitted for the sake of simplicity in this section. In Eq.(21), the coefficient C_i is the contribution of $\hat{\mathbf{a}}_d$ in the direction of $\bar{\mathbf{e}}_i$. When $c_i = C_i/\max\{C_i\}$ is much smaller than unity, the subspace spanned by the corresponding eigenvectors contains mostly the energy of noise or other directional signals. Therefore, the subspace where c_i is small is eliminated by weighting the eigenvalue $\bar{\lambda}_i$ of $\bar{\mathbf{R}}_{j_0}$ as

$$\bar{\mathbf{\Lambda}}^+_{j_0} = \mathrm{diag}[w_1\bar{\lambda}_1, \cdots, w_M\bar{\lambda}_M]. \tag{22}$$

The weight is given by

$$w_i = \begin{cases} c_i & \text{if } c_i \geq c_{thr} \\ 0 & \text{otherwise} \end{cases}, \tag{23}$$

where c_{thr} is an arbitrary threshold. Using this, the correlation matrix is reconstructed as $\bar{\mathbf{R}}^+_{j_0} = \bar{\mathbf{E}}_{j_0}\bar{\mathbf{\Lambda}}^+_{j_0}\bar{\mathbf{E}}^H_{j_0}$. The cross-correlation matrix $\bar{\mathbf{C}}_{j_0}$ is estimated by Eq.(16) with $\bar{\mathbf{R}}^+_{j_0}$ instead of $\bar{\mathbf{R}}_{j_0}$.

6. SIMULATION

6.1. Condition

The microphone array simulated was linearly configured with 8 microphones 6 cm apart. As depicted in Fig. 2, two

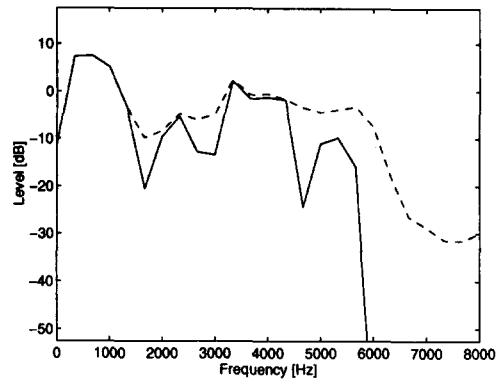

Figure 5. Band-averaged spectrum $\bar{P}_d(\omega_{j_0})$. Solid (/dashed) line: with(/without) subspace reduction using projection.

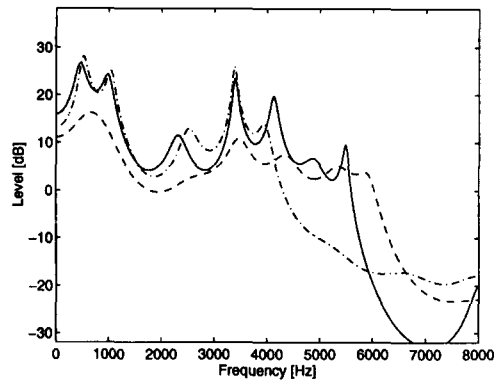

Figure 6. LPC spectrum. Dash-dotted line: original speech; Solid line: CSS-Proj; dashed line: DS.

directional sources, A (0°) and B (–20°), emitting speech were used. To simulate ambient noise, a mixture of noises coming from the directions of –90° to 90° at every 5° was used. The SNR of each directional signal was 0 dB. The spectrum of speech A, speech B, and ambient noise is depicted in Fig. 3. In CSS, the RSS focusing matrix was used [4]. The frame length was 128. N_a = 48 and N_{lpc} = 16. The sampling frequency was 16 kHz.

6.2. Case 1: one directional source

First, the case when only source A was used was investigated. Figure 4 shows the eigenvalue distribution and the contribution, c_i, at (a) 1000 Hz (high SNR) and (b) 3000 Hz (low SNR), frequencies indicated by the dotted lines in Fig. 3. In (a), the energy of the speech is concentrated in the largest eigenvalue, and the signal and noise subspace are clearly separated. On the other hand, in (b), contribution of the speech signal spreads over several subspaces.

Figure 5 shows the band-averaged spectrum estimated by Eq.(16). The solid line corresponds to the case when noise subspace reduction using projection was used, while the dotted line corresponds to the case without it. In this simulation, the threshold for the subspace reduction was set at $c_{thr} = -1$ dB. It can be seen that the power was reduced at the frequencies of low SNR by the subspace reduction.

Figure 6 shows the LPC spectrum. The solid line corresponds to the proposed method (CSS plus subspace re-

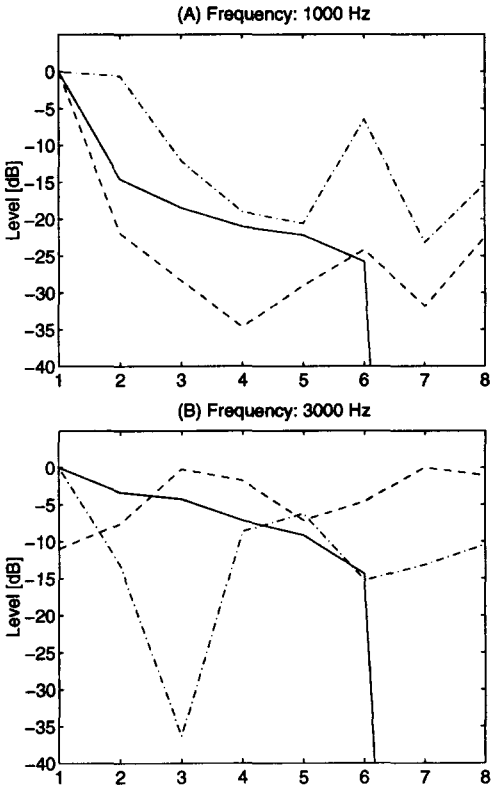

Figure 7. Eigenvalues (solid line) and the contribution c_i (Dashed line: speech A; Dash-dotted line: speech B). (A) 1000Hz and (B) 3000Hz.

duction using projection, denoted as CSS-Proj hereafter), while the dashed line corresponds to the conventional delay-and-sum beamformer (denoted as DS). In DS, Eq.(14) and Eq.(13) were calculated with $\mathbf{K}_j = \mathbf{I}$, and then the band averaged spectrum, $\bar{P}_d(\omega_{j_0})$, was calculated. The procedure for calculating the LPC coefficients was the same as the proposed CSS-Proj. From Fig. 6, it can be seen that some formants that could not be estimated by DS were clearly estimated by the proposed CSS-Proj.

6.3. Case 2: two directional sources

Next, the case of two directional sources A and B was examined. Figure 7 shows the eigenvalue spread and the contribution, c_i. At 1000 Hz where speech A is dominant, it can be seen that the energy of speech A is concentrated at the largest eigenvalue. On the other hand, at 3000 Hz, where speech B is slightly larger, the contribution of speech B to the largest eigenvalue is large, though the eigenvalue spread is rather flat.

Figure 8 shows the LPC spectrum. For speech A, the curve estimated by CSS-Proj again shows relatively good consistency with the original spectrum. On the other hand, for speech B, a large peak at the low frequency appears for both the CSS-Proj and DS methods, while formants in the middle frequency are well estimated by both methods. At lower frequencies, the difference in directional vector is small, resulting in low spatial resolution. In this simulation, the power of speech A, which has large formants at low

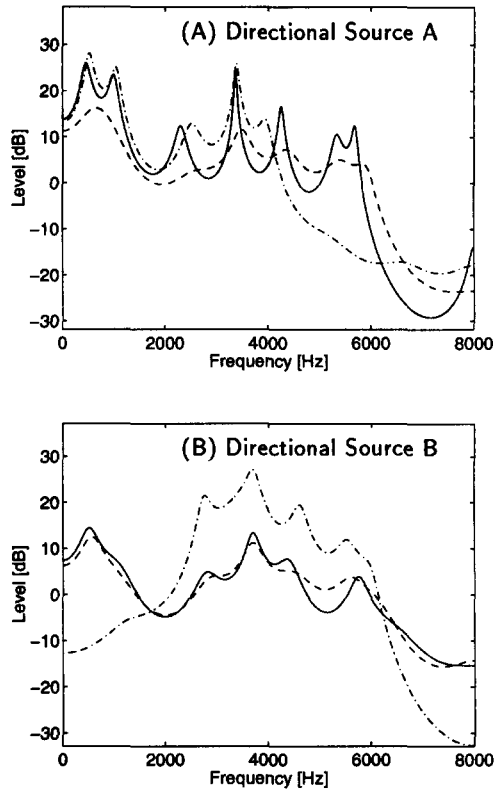

Figure 8. LPC spectrum. Dash-dotted line: original speech; Solid line: CSS-Proj; Dashed line: DS.

frequency, leaked and affected the estimation of speech B.

7. CONCLUSION

A method of estimating the LPC spectrum of speech from microphone array input corrupted by ambient noise, based on the coherent subspace method, was proposed.

REFERENCES

[1] L. J. Griffiths and K. M. Buckley, "Quiescent pattern control in linearly constrained adaptive arrays," *IEEE Trans. ASSP*, vol. ASSP-35, pp. 917–926, 1987.

[2] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. ASSP*, vol. 33(4), pp. 823–831, 1985.

[3] R. Roy and T. Kailath, "ESPRIT - Estimation of signal parameters via rotational invariance techniques," *IEEE Trans. ASSP*, vol. 37(7), pp. 984–995, 1989.

[4] H. Hung and M. Kaveh, "Focussing matrix for coherent signal-subspace processing," *IEEE Trans. ASSP*, vol. 36(8), pp. 1272–1281, 1988.