BLIND SPEECH ENHANCEMENT USING GENERALIZED EIGENVALUE DECOMPOSITION

Futoshi Asano and Hideki Asoh
AIST, Japan
f.asano@aist.go.jp

ABSTRACT

A method of speech enhancement using a microphone array and the generalized eigenvalue decomposition is proposed in this paper. This method is a blind approach which does not require array calibration. The advantage of this method over the conventional blind separation based on the ICA is its fast adaptation. This is verified by a speech enhancement experiment.

1 Introduction

For applications such as automatic speech recognition (ASR) in a real environment, separation of the target speech from environmental noise is indispensable. Array signal processing has been applied to this problem and shows a high noise reduction performance (e.g., [1]). However, the problem with array processing is the necessity of array calibration, which is often troublesome, especially for irregular arrays. To overcome this problem, the blind signal separation (BSS) approach based on independent component analysis (ICA) has been intensively studied in recent years. However, the convergence of BSS based on ICA is generally slow and is difficult to apply to dynamic environments.

In this paper, a method of blind signal enhancement using a microphone array and generalized eigenvalue decomposition (GEVD) is proposed. This method is a blind approach since it does not require any array calibration. Unlike BSS, in which all the independent components are separated, this method only separates the target signal from a mixture of the environmental noise and jammer signals. An advantage of this method is its fast adaptation characteristics. The reason of this is mainly that the proposed method utilizes only second order statistics instead of the higher order statistics used in most of the conventional BSS.

Signal enhancement using standard eigenvalue decomposition (SEVD) or, equivalently, principal component analysis (PCA) has been proposed by the authors [2, 3]. The performance of the SEVD approach is equivalent to that of the delay-and-sum beamformer and is effective for less-directional noise. However, its performance for directional or spatially-colored noise is not satisfactory. The GEVD approach makes this method adaptive to spatially colored noise and enhances its performance.

2 Method

2.1 Model of Signal

The input vector is defined as $\mathbf{x}(\omega,t) = [X_1(\omega,t), \cdots, X_M(\omega,t)]^T$, where $X_m(\omega,t)$ is a short-term Fourier transform (STFT) of the input signal in the tth time frame at the mth microphone. The input vector is modeled as a mixture of the target signal and the environmental noise as

$$\mathbf{x}(\omega,t) = \mathbf{a}(\omega)s(\omega,t) + \mathbf{n}(\omega,t). \tag{1}$$

The symbol $\mathbf{a}(\omega)$ is the transfer function vector, the m element of which is the transfer function of the direct path from the target source to the m microphone. The symbol $s(\omega,t)$ is the spectrum of the target source. The vector $\mathbf{n}(\omega,t)$ consists of the spectra of noise observed at the microphones.

The spatial correlation of the input vector is defined as

$$\mathbf{R} = E[\mathbf{x}(\omega,t)\mathbf{x}^H(\omega,t)]. \tag{2}$$

Assuming that the target signal and the environmental noise are uncorrelated, the spatial correlation matrix is written as

$$\mathbf{R} = \mathbf{a}P\mathbf{a}^H + \mathbf{K}. \tag{3}$$

The matrix $\mathbf{K}$ denotes the spatial correlation of the noise defined as $\mathbf{K} = E[\mathbf{n}(\omega,t)\mathbf{n}^H(\omega,t)]$. The symbol P is the average power of the signal, $P = E[|s(\omega,t)|^2]$.

2.2 GEVD

The first stage of GEVD is the spatial whitening of $\mathbf{R}$ as

$$\bar{\mathbf{R}} = \mathbf{W}^H\mathbf{R}\mathbf{W}. \tag{4}$$

The matrix $\mathbf{W}$ is the whitening matrix and is defined as

$$\mathbf{W} = \Phi^{-1} \quad \text{where} \quad \mathbf{K} = \Phi^H\Phi. \tag{5}$$

From (5), $\mathbf{\Phi}$ is calculated as

$$\Phi = \mathbf{\Sigma}^{1/2}\mathbf{G}^H, \tag{6}$$

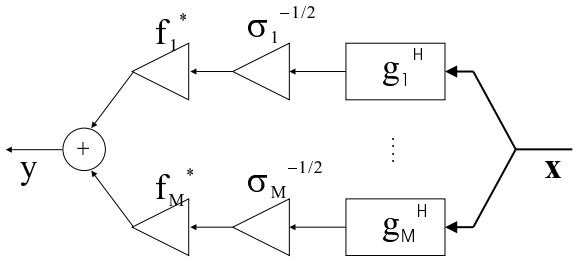


Figure 1: Block diagram.

where $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \cdots, \sigma_M)$ and $\mathbf{G} = [\mathbf{g}_1, \cdots, \mathbf{g}_M]$ are the eigenvalue matrix and the eigenvector matrix of $\mathbf{K}$, respectively. The whitening process of the input vector is then written as

$$\bar{\mathbf{x}}(\omega, t) = \mathbf{W}^H \mathbf{x}(\omega, t) = \mathbf{\Sigma}^{-1/2} \mathbf{G}^H \mathbf{x}(\omega, t). \quad (7)$$

By applying the whitening matrix, the spatial correlation of the noise, $\mathbf{K}$, becomes the identity matrix $\mathbf{I}$ as $\bar{\mathbf{R}} = \mathbf{W}^H \mathbf{a} P \mathbf{a}^H \mathbf{W} + \mathbf{I}$.

Next, SEVD of the whitened spatial correlation matrix $\bar{\mathbf{R}}$ is taken as $\bar{\mathbf{R}} = \mathbf{F} \mathbf{\Lambda} \mathbf{F}^{-1}$, where $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_M]$ and $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_M)$ are the eigenvector matrix and the eigenvalue matrix, respectively. The eigenvalues and eigenvectors are assumed to be sorted in descending order of the eigenvalues, i.e., $\lambda_1 > \cdots, > \lambda_M$. The eigenvector $\mathbf{f}_m$ is then transformed by the whitening matrix $\mathbf{W}$ as

$$\mathbf{e}_m = \mathbf{W} \mathbf{f}_m \quad (8)$$

The obtained eigenvalues $\{\lambda_1, \cdots, \lambda_M\}$ and the transformed eigenvectors $\{\mathbf{e}_1, \cdots, \mathbf{e}_M\}$ become the eigenvalues and the eigenvectors of the generalized eigenvalue problem: $\mathbf{Re} = \lambda \mathbf{Ke}$. The transformed eigenvector matrix $\mathbf{E} = [\mathbf{e}_1, \cdots, \mathbf{e}_M]$ yields the following simultaneous diagonalization properties [4]:

$$\mathbf{E}^H \mathbf{R} \mathbf{E} = \Lambda \quad (9)$$

$$\mathbf{E}^H \mathbf{K} \mathbf{E} = \mathbf{I}. \quad (10)$$

The equation (10) is related to the target enhancement while (10) is related to the noise whitening property. They are examined in detail in the next section.

The transformed eigenvector, $\mathbf{e}_1$, which corresponds to the largest eigenvalue, is then used as the spatial filter for the target signal enhancement as

$$y(\omega, t) = \mathbf{e}_1^H \mathbf{x}(\omega, t). \quad (11)$$

Using (7), the filtering process (11) can be written as

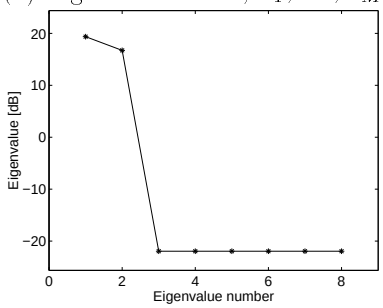
$$y(\omega, t) = \mathbf{f}_1^H \mathbf{\Sigma}^{-1/2} \mathbf{G}^H \mathbf{x}(\omega, t) \quad (12)$$

Based on (12), the block diagram of the system can be written as Fig. 1.

3 Properties of GEVD

In this section, the properties of GEVD are analyzed by using a simple example. Let us consider the case of a simple linear array with 8 microphones. In this example, the noise is a mixture of directional interference coming from 60° and spatially-white background noise. The target signal is a directional signal coming from 0° . The power ratio of the target signal, the directional noise and the spatially-white background noise are 0, 0 and -40 dB, respectively.

(a) Eigenvalues of $\mathbf{R}$, $\lambda_1, \cdots, \lambda_M$



(b) Eigenvalues of $\mathbf{K}$, $\sigma_1, \cdots, \sigma_M$

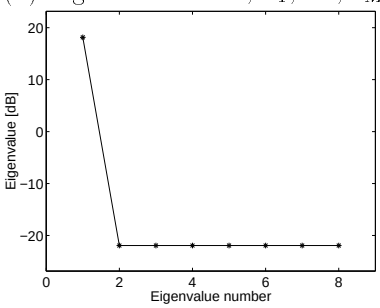


Figure 2: Eigenvalue distribution of $\mathbf{R}$ and $\mathbf{K}$.

The eigenvalue distributions of $\mathbf{R}$ and $\mathbf{K}$ are shown in Fig. 2. Two dominant eigenvalues corresponding to the directional target and the directional noise can be seen in the eigenvalue distribution of $\mathbf{R}$, while only a single dominant eigenvalue corresponding to the directional noise appears in that of $\mathbf{K}$.

Next, the spatial characteristics of the whitening filter $\mathbf{W}$ is analyzed. As depicted in Fig. 1, the column vector of $\mathbf{W}$,

$$\mathbf{w}_m = \sigma^{-1/2} \mathbf{g}_m, \quad (13)$$

functions as a filter at the mth channel. The characteristics of the filter $\mathbf{w}_m$ are depicted in Fig. 3. From this figure, it can be seen that a spatial beam focusing on the directional noise is formed in the characteristic of $\mathbf{w}_1$ (solid line), which corresponds to the dominant eigenvalue of $\mathbf{K}$. On the other hand, in the characteristics of $\mathbf{w}_m$ for $m = 2, \cdots, M$ (dashed line), a spatial notch is formed in the direction of the directional noise. Based on this, with regard to the whitening process, it can roughly be understood that the noise $\mathbf{n}(\omega, t)$ is decomposed into directional noise and spatially-white background noise, and that their gains are then normalized. The gain normalization is performed by $\sigma_m^{-1/2}$ in (13) and plays an essential role in spatial whitening. During this process, the power of the directional noise is reduced to that of the background noise. In this sense,

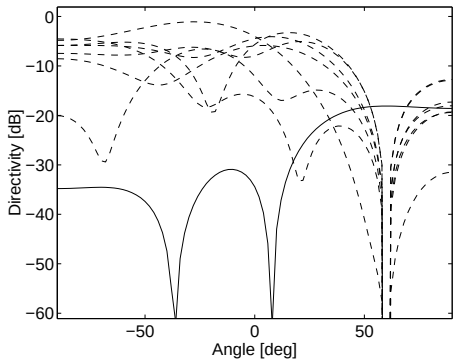


Figure 3: Spatial characteristics of the whitening filter $\mathbf{w}_m$. The solid line shows the characteristic of $\mathbf{w}_1$ corresponding to the dominant eigenvalue σ_1. The dotted line shows those of $\mathbf{w}_m$,for $m = 2, \cdots, M$.

the whitening filter $\mathbf{W}$ functions as an adaptive beamformer which reduces the directional component.

Next, the characteristics of the final GEVD filter $\mathbf{e}_1$ is analyzed. Figure 4 shows the spatial characteristics of the GEVD filter $\mathbf{e}_1$. As can be seen from this figure, a spatial notch is formed in the direction of the directional noise (60°). This is the adaptive aspect of the GEVD filter due to noise whitening.

The GEVD filter has another noise reduction aspect. For the sake of simplicity, the case of SEVD [2, 3] is considered first. The dependency of the eigenvectors can be calculated using the following projection:

$$p(m,n) = (\mathbf{e}_m^H \mathbf{e}_n)/(\mathbf{e}_n^H \mathbf{e}_n) \tag{14}$$

In the case of SEVD, since the eigenvectors for a hermitian matrix are orthogonal, this dependencies becomes

$$p(m,n) = \begin{cases} 1, & \text{for } m = n \\ 0, & \text{for } m \neq n \end{cases} . \tag{15}$$

From this orthogonality, when one of the eigenvectors of SEVD is used as a filter, this filter passes the component of the noise lying in the subspace spanned by this eigenvector and cancels out the noise component in the other subspaces. When the noise is spatially white, this filter reduces the power of the noise by $1/M$, equivalent to the performance of the delay-and-sum beamformer [2]. When GEVD is employed, the eigenvectors of GEVD are not necessarily orthogonal. Figure 5 shows the dependency of the eigenvectors. For example, the solid line in Fig. 5 shows the projection of $\mathbf{e}_1$ onto all eigenvectors, i.e., $\mathbf{e}_n$, for $n = 1, \cdots, M$. From this figure, it can be seen that the eigenvector $\mathbf{e}_1$ makes only a small contribution to the the subspace spanned by the other eigenvectors, $\mathbf{e}_n$ for $n = 2, \cdots, M$. On the other hand, the other eigenvectors makes a small contribution to the subspace spanned by $\mathbf{e}_1$. This means that the energy of the noise contained in the subspace spanned by $\mathbf{e}_n$ for $n = 2, \cdots, M$ is reduced by $\mathbf{e}_1$.

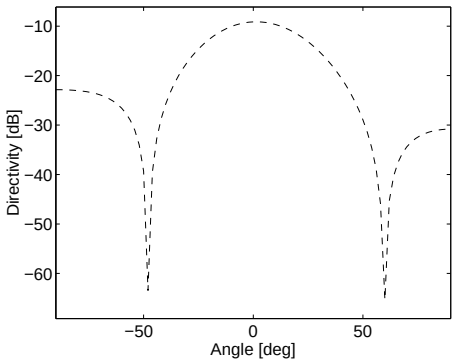


Figure 4: Spatial characteristics of GEVD filter $\mathbf{e}_1$.

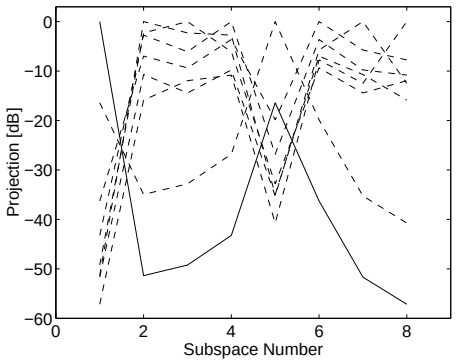


Figure 5: Dependency of the eigenvectors, $p(m,n)$

In summary, the directional noise is first reduced by the whitening process. Then the whitened noise is further reduced by the subspace selection aspect of the eigenvector $\mathbf{e}_1$.

4 Experiment

In this section, the GEVD approach is applied to the speech enhancement problem in a real environment and is evaluated via ASR.

4.1 Condition

The microphone input was generated by convolving source signals with the impulse response measured in a medium-sized meeting room. The reverberation time of this room was 0.5 s. As source signals, Japanese words (S1) and a music signal (S2) were employed. Additive background noise was not employed in this experiment. However, the reverberation of S1 and S2 behaved as omni-directional background noise. The microphone array used in this experiment was circular in shape with 8 microphones and was mounted on a mobile robot.

4.2 Results

Figure 6 shows the directivity pattern of the GEVD filter obtained in the experiment. The target source S1 is located at (0° ,1.0m) while the noise source S2 is located at (60° ,1.0m). As can be seen in the figure, A notch appears in the direction of S2. Table 1 shows the power of each component in the input/output, normalized by

Table 1: Power of components in the input/output signal (in dB). "D" and "R" denote direct sound and reflection, respectively.

Components	Input	Output	Difference
S1D	0.0	0.0	–
S2D	8.5	-14.0	-22.5
S1R	-10.3	-16.3	-6.0
S2R	-2.3	-10.2	-7.9

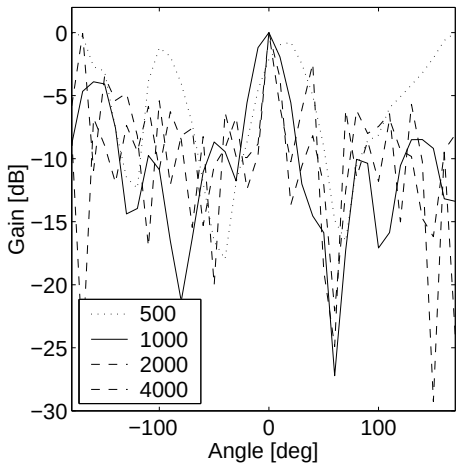


Figure 6: Directivity of GEVD filter at 0.5, 1, 2 and 4kHz.

the power of the direct sound of S1 (S1D). As can be seen in this table, the direct sound of S2 is reduced by more than 20 dB. This is the effect of the noise whitening in GEVD. The reverberation of S1 and S2 is also reduced by 6-8 dB. This is the effect of the subspace selectivity.

In the ASR experiment, four isolated words were emitted from S1. The duration of each word was approximately 1-2 s. During the emission of the four words, the direction of S1 was fixed. After these four words were processed, the direction of S1 was selected from $\{0, 120, 180, 240, 300\}$ in turn, and then another set of four words was emitted. This emission of four words was repeated 492 times. During these 492 sessions, the GEVD filter was being updated. Thus, environmental change with every four words was simulated. Figure 7 shows the ASR score for each of the four words. For the sake of comparison, the ASR score for BSS [3] is also shown. As can be seen from this figure, the score for GEVD reached a steady state at the second word while that of BSS reached its maximum at the third word. From this, the convergence of GEVD can be seen to be faster than that of BSS. Also the performance is higher for GEVD.

5 Conclusion and Discussion

In this paper, a method of speech enhancement using a microphone array and the generalized eigenvalue decomposition was proposed. The results of the speech enhancement experiment confirmed that this method has a fast adaptation property. The GEVD approach includes two-step noise reduction. The first step is a spatial whitening process, in which directional noise is reduced. The second stage is the subspace selection, in which spatially-whitened noise is reduced.

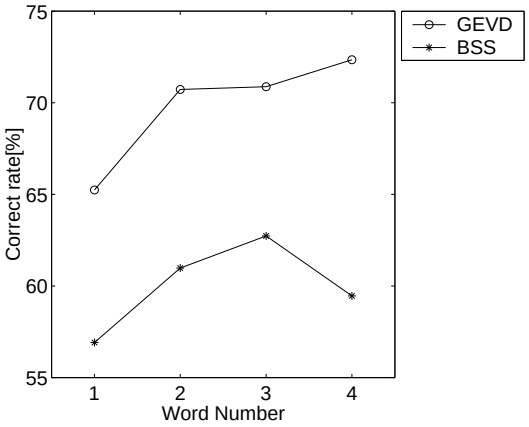


Figure 7: ASR Score.

A disadvantage of this method compared with a conventional blind separation is that the proposed method requires the spatial correlation of noise, $\mathbf{K}$. However, in the speech enhancement application, the absence of the target speech can be detected by the voice activity detector (VAD) on the assumption that the noise is not human voice. This enables us to observe $\mathbf{K}$ in the period of the absence of the target speech.

References

[1] F. Asano, *et al.*, "Real-time sound source localization and separation system and its application to automatic speech recognition," in *Proc. Eurospeech*, Sep. 2001, pp. 1013–1016.

[2] F. Asano, *et al.*, "Speech enhancement based on the subspace method," *IEEE Trans. Speech, Audio Processing*, vol. 8, no. 5, pp. 497–507, 2000.

[3] F. Asano, *et al.*, "A combined approach of array processing and independent component analysis for blind separation of acoustic signals," in *Proc. ICASSP2001*, 2001, vol. V, pp. MULT–P2.

[4] G. Strang, *Linear Algebra and Its Application*, Harcourt Brace Jovanovich Inc., 1988.