Simplified Maximum SNR Beamformers with Spatial Coherence Matrix Modeling

Fan Zhang[1], Chao Pan[1], Jacob Benesty[2], and Jingdong Chen[1]
[1]CIAIC, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[2]INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada

***Abstract*—The maximum signal-to-noise ratio (SNR) beamformer is useful in a wide range of applications to enhance speech signals of interest and attenuate as much as possible the noise. But robust implementation of this beamformer is challenging in practical applications as it requires to know the signal and noise covariance matrices. This paper investigates how to simplify the beamformer for use in small-spacing microphone arrays. Indeed, with small-spacing arrays, a practical parametric model can be used to model the covariance matrix of the observations, which is closely related to the front-to-back ratio (FBR) in differential beamforming. With this parametric model, we derive two simplified maximum SNR beamformers, which depend on the signal power spectral density (PSD) only. We then propose an estimator based on Frobenius-norm minimization to estimate the PSD. Since PSDs are usually easier to estimate than covariance matrices, the developed beamformers have great advantage over its traditional counterparts in terms of implementation in practical systems. The performance of the developed beamformers are validated in a simulated classroom environment.**

***Index Terms*—Microphone arrays, adaptive beamforming, maximum SNR beamformer, parametric covariance matrix modeling.**

I. Introduction

Adaptive beamforming with microphone arrays has been widely used in a large number of speech applications for desired signal acquisition and extraction over the last few decades [1]–[3]. The most representative algorithms on this topic include the minimum variance distortionless response (MVDR) [4]–[8], maximum signal-to-noise (SNR) [9]–[11], and Wiener [12]–[17] beamformers.

Generally, the implementation of adaptive beamformers requires to estimate the signal and noise statistics and/or direction of arrival information of the sources of interest. For example, the MVDR beamformer depends on the covariance matrix of the array observations or noise and steering vector of the desired source. The maximum SNR beamformer depends on the covariance matrices of the desired signal and the noise. The Wiener beamformer is a function of the covariance matrices of the observations and desired source. Since it is difficult to reliably estimate the signal and noise statistics in practical acoustic environments due to the nonstationary nature of the speech as well as the noise signals, adaptive beamformers often suffer from severe performance degradation, i.e., signal cancellation and distortion in practice [18]–[21].

In some applications, the spatial distributions of the desired and/or noise sources are approximately known. For example, in classroom public address (PA) systems, if the sound acquisition microphone array is placed on the top of the teaching desk, the desired source (e.g., a teacher) and the interference sources (e.g., students) are from different sides of the array. With such *a priori* information and small-spacing microphone arrays, a parametric model can used to model the covariance matrix of the observations as in differential microphone arrays [22], [23]. With this model, we derive two simplified maximum SNR beamformers, which depend on the signal power spectral density (PSD) only. We then present a PSD estimator based on the Frobenius-norm minimization. Since it is easier and more reliable to estimate PSD than covariance matrices, the developed beamformers are preferred in practical systems. Simulations in a classroom environment also demonstrate the advantages of the developed beamformers in terms of performance as compared to two widely used beamformers.

II. Signal Model, Problem Formulation, and Covariance Matrix Modelling

We consider to use a small-spacing uniform linear array (ULA) consisting of M omnidirectional microphones to pick up a speech signal of interest in some noise field. The interelement spacing of the ULA is denoted as δ. In the rest of this paper, we use $(\cdot)^T$ and $(\cdot)^H$ to denote, respectively, the transpose and conjugate-transpose of a vector/matrix, $E[\cdot]$ to denote mathematical expectation, and $\mathrm{tr}\{\cdot\}$ to denote the trace of a matrix. Using the short-time Fourier transform (STFT), we can express the mth channel signal in the STFT domain as [24]

$$Y_m(k,n) = X_m(k,n) + V_m(k,n), \;\; m = 1, 2, \ldots, M, \quad (1)$$

where k is the frequency-bin index, n is the frame index, and $X_m(k,n)$ and $V_m(k,n)$ are the STFTs of the desired speech and noise signals, respectively. The signals in (1) can be rearranged into the following vector form:

$$\begin{aligned}\mathbf{y}(k,n) &\triangleq \left[\begin{array}{cccc} Y_1(k,n) & Y_2(k,n) & \cdots & Y_M(k,n) \end{array} \right]^T \quad (2)\\ &= \mathbf{x}(k,n) + \mathbf{v}(k,n),\end{aligned}$$

where the speech signal vector $\mathbf{x}(k,n)$ and the noise vector $\mathbf{v}(k,n)$ are defined in a similar way as $\mathbf{y}(k,n)$. We assume that $\mathbf{x}(k,n)$ and $\mathbf{v}(k,n)$ are of zero mean and they are uncorrelated.

This work was supported in part by the National Key Research and development Program of China under Grant 2021ZD0201502, in part by the NSFCs under Grants 61901318, 62171373, and 61831019, and in part by the Fundamental Research Funds for the Central Universities under Grant 3102021HHZY030001.

Consequently, the covariance matrix of $\mathbf{y}(k, n)$ can be written as

$$\begin{aligned}\mathbf{C}_{\mathbf{y}}(k, n) &= E\left[\mathbf{y}(k, n)\mathbf{y}^H(k, n)\right] \\ &= \mathbf{C}_{\mathbf{x}}(k, n) + \mathbf{C}_{\mathbf{v}}(k, n),\end{aligned} \tag{3}$$

where $\mathbf{C}_{\mathbf{x}}(k, n)$ and $\mathbf{C}_{\mathbf{v}}(k, n)$ are the covariance matrices of $\mathbf{x}(k, n)$ and $\mathbf{v}(k, n)$, respectively.

The objective of this work is to recover the noise-free speech signal at any one of the microphones in the ULA. Without loss of generality, we consider to formulate the problem as one of recovering the noise-free speech signal at the first microphone, i.e., $X_1(k, n)$, which will be called the desired signal. The estimation of $X_1(k, n)$ is achieved by applying a beamforming filter, $\mathbf{h}(k, n)$, to the observation vector, $\mathbf{y}(k, n)$, i.e.,

$$Z(k, n) = \mathbf{h}^H(k, n)\mathbf{y}(k, n), \tag{4}$$

where $Z(k, n)$ is the beamformer's output. Then, the problem becomes one of identifying an optimal filter so that $Z(k, n)$ is a good or an optimal estimate of $X_1(k, n)$. A number of algorithms have been developed over the past few decades and the typical ones include the MVDR, maximum SNR, and Wiener beamformers. Generally, implementation of those beamformers relies on the estimation of the signal and noise covariance matrices and/or the steering vector for the desired source, which is difficult to achieve in practical acoustic conditions. Inaccurate estimation of those parameters often leads to severe performance degradation [18]–[21]. Fortunately, with proper array configuration, some of the parameters can be modeled with some *a priori* information, which would significantly simplify adaptive beamformers as will be demonstrated in this work.

To start with, we consider the farfield case and neglect the gain difference among channels. In this case, the steering vector for the ULA in the STFT domain can be written as [24]

$$\mathbf{d}_\theta(k) = \left[\begin{array}{cccc} 1 & e^{-\jmath 2\pi f_k \frac{\delta}{c}\cos\theta} & \cdots & e^{-\jmath(M-1)2\pi f_k \frac{\delta}{c}\cos\theta}\end{array}\right]^T, \tag{5}$$

where $\theta \in [0, \pi]$ is the source incident angle, $\jmath$ is the imaginary unit, f_k denotes the frequency corresponding to the kth STFT bin, and c is the sound speed in air. Furthermore, we assume that the noise source is located in the back-half plane of the ULA while the desired source is located in the front-half plane of the ULA. In this case, the covariance matrix of $\mathbf{y}(k, n)$ in (3) can be modelled as

$$\mathbf{C}_{\mathbf{y}}(k, n) = \lambda_{X_1}(k, n)\mathbf{\Gamma}_{\mathrm{F}}(k) + \lambda_{V_1}(k, n)\mathbf{\Gamma}_{\mathrm{B}}(k), \tag{6}$$

where $\lambda_{X_1}(k, n)$ and $\lambda_{V_1}(k, n)$ are the PSDs of the desired signal and the noise at the first microphone, respectively, and

$$\mathbf{\Gamma}_{\mathrm{F}}(k) = \int_0^{\pi/2} \mathbf{d}_\theta(k)\mathbf{d}_\theta^H(k)\sin\theta \mathrm{d}\theta, \tag{7}$$

$$\mathbf{\Gamma}_{\mathrm{B}}(k) = \int_{\pi/2}^{\pi} \mathbf{d}_\theta(k)\mathbf{d}_\theta^H(k)\sin\theta \mathrm{d}\theta. \tag{8}$$

The elements of the two matrices $\mathbf{\Gamma}_{\mathrm{F}}(k)$ and $\mathbf{\Gamma}_{\mathrm{B}}(k)$ can be expressed in analytic forms [22], [23]. The model in (6) is very useful in many application scenarios. For example, in a typical classroom, the teacher may stand on one side of the desk while the students sit on the other side. A uniform linear microphone array is placed on the desk to pick up the sound from the teacher and meanwhile attenuating the sounds and noise from the student side. In such scenarios, (6) is a reasonable model for modeling the covariance matrix of the array observation vector.

III. Simplified Maximum SNR Beamformers with Known Coherence Matrices

In this section, we derive two simplified maximum SNR beamformers corresponding to the model described previously.

A. Derivations of Simplified Maximum SNR Beamformers

The subband input SNR at the nth frame is defined as

$$\xi_{\mathrm{in}}(k, n) \triangleq \frac{\lambda_{X_1}(k, n)}{\lambda_{V_1}(k, n)}. \tag{9}$$

With the beamforming output given in (4) and the parametric model given in (6), the subband output SNR of the beamformer at the nth frame can be written as

$$\begin{aligned}\xi_{\mathrm{out}}\left[\mathbf{h}(k, n)\right] &= \frac{\mathbf{h}^H(k, n)\mathbf{C}_{\mathbf{x}}(k, n)\mathbf{h}(k, n)}{\mathbf{h}^H(k, n)\mathbf{C}_{\mathbf{v}}(k, n)\mathbf{h}(k, n)} \\ &= \xi_{\mathrm{in}}(k, n)\frac{\mathbf{h}^H(k, n)\mathbf{\Gamma}_{\mathrm{F}}(k)\mathbf{h}(k, n)}{\mathbf{h}^H(k, n)\mathbf{\Gamma}_{\mathrm{B}}(k)\mathbf{h}(k, n)}.\end{aligned} \tag{10}$$

Using (9) and (10), we deduce that the subband SNR gain of the beamformer is

$$\mathcal{G}\left[\mathbf{h}(k, n)\right] = \frac{\xi_{\mathrm{out}}\left[\mathbf{h}(k, n)\right]}{\xi_{\mathrm{in}}(k, n)} = \frac{\mathbf{h}^H(k, n)\mathbf{\Gamma}_{\mathrm{F}}(k)\mathbf{h}(k, n)}{\mathbf{h}^H(k, n)\mathbf{\Gamma}_{\mathrm{B}}(k)\mathbf{h}(k, n)}, \tag{11}$$

which corresponds to the well-known front-to-back ratio (FBR) [25], [26]. Therefore, the simplified maximum SNR beamformer with the assumed model in (6) is the maximum FBR beamformer, which is of the following form:

$$\mathbf{h}_{\mathrm{FBR}}(k, n) = \alpha(k, n)\mathbf{t}_1(k), \tag{12}$$

where $\alpha(k, n)$ is any non-zero complex number and $\mathbf{t}_1(k)$ is the eigenvector corresponding to the maximum eigenvalue of $\mathbf{\Gamma}_{\mathrm{B}}^{-1}(k)\mathbf{\Gamma}_{\mathrm{F}}(k)$. Note that $\alpha(k, n)$ has no impact on the subband SNR gain of the maximum FBR beamformer, but it may have significant impact on speech distortion and the fullband SNR gain because the observation signals at different frames and frequency bins are processed independently, leading to inconsistent values of $\alpha(k, n)$ at different frames and frequency bins. Consequently, it is necessary to determine the proper value for $\alpha(k, n)$. In what follows, we will present two methods to determine this value. To simplify the notation, we will drop the index k in the rest of this section.

The first method to determine the value of $\alpha(n)$ is through minimizing the mean-squared error (MSE) between the desired signal, $X_1(n)$, and its estimate, $Z(n)$, i.e.,

$$\begin{aligned}J\left[\mathbf{h}_{\mathrm{FBR}}(n)\right] &\triangleq E\left[\left|\mathbf{h}_{\mathrm{FBR}}^H(n)\mathbf{y}(n) - X_1(n)\right|^2\right] \\ &= J_{\mathrm{d}}\left[\mathbf{h}_{\mathrm{FBR}}(n)\right] + J_{\mathrm{n}}\left[\mathbf{h}_{\mathrm{FBR}}(n)\right]\end{aligned} \tag{13}$$

where $J_{\mathrm{d}}\left[\mathbf{h}_{\mathrm{FBR}}(n)\right] \triangleq E\left[\left|\mathbf{h}_{\mathrm{FBR}}^H(n)\mathbf{x}(n) - X_1(n)\right|^2\right]$ is the term associated with speech distortion and $J_{\mathrm{n}}\left[\mathbf{h}_{\mathrm{FBR}}(n)\right] \triangleq E\left[\left|\mathbf{h}_{\mathrm{FBR}}^H(n)\mathbf{v}(n)\right|^2\right]$ is the variance of the residual noise. Taking the derivative of $J\left[\mathbf{h}_{\mathrm{FBR}}(n)\right]$ with respect to $\alpha^*(n)$ and forcing the result to zero, one can obtain

$$\alpha_{\mathrm{MMSE}}(n) = \frac{\lambda_{X_1}(n)\mathbf{t}_1^H\mathbf{\Gamma}_{\mathrm{F}}\mathbf{u}_1}{\lambda_{X_1}(n)\zeta_0 + \lambda_{V_1}(n)\zeta_1}, \tag{14}$$

where $\mathbf{u}_1$ is the first column of the $M \times M$ identity matrix $\mathbf{I}_M$, $\zeta_0 \triangleq \mathbf{t}_1^H\mathbf{\Gamma}_{\mathrm{F}}\mathbf{t}_1$, and $\zeta_1 \triangleq \mathbf{t}_1^H\mathbf{\Gamma}_{\mathrm{B}}\mathbf{t}_1$. Substituting the result in (14) into (12) gives the first maximum FBR beamformer:

$$\mathbf{h}_{\mathrm{FBR,MSE}}(n) = \frac{\lambda_{X_1}(n)\mathbf{t}_1\mathbf{t}_1^H\mathbf{\Gamma}_{\mathrm{F}}\mathbf{u}_1}{\lambda_{X_1}(n)\zeta_0 + \lambda_{V_1}(n)\zeta_1}, \tag{15}$$

which depends on the PSDs of speech and noise. Another convenient way to write (15) is

$$\mathbf{h}_{\mathrm{FBR,MSE}}(n) = \frac{\xi_{\mathrm{in}}(n)\mathbf{t}_1\mathbf{t}_1^H\mathbf{\Gamma}_{\mathrm{F}}\mathbf{u}_1}{\xi_{\mathrm{in}}(n)\zeta_0 + \zeta_1}, \tag{16}$$

which is a function of the subband input SNR only.

In some situations, we want to make a tradeoff between the degree of speech distortion and the amount of noise attenuation. Following the idea from the so-called speech distortion weighted multichannel Wiener filter (SDW-MWF) [12], [13], [27], we propose to determine the value of $\alpha(n)$ by minimizing the following cost function:

$$J_{\mu}\left[\mathbf{h}_{\mathrm{FBR}}(n)\right] = J_{\mathrm{d}}\left[\mathbf{h}_{\mathrm{FBR}}(n)\right] + \mu J_{\mathrm{n}}\left[\mathbf{h}_{\mathrm{FBR}}(n)\right], \tag{17}$$

where $\mu \geq 0$ is a weighting coefficient. We deduce that the optimal value of $\alpha(n)$ is

$$\alpha_{\mu}(n) = \frac{\lambda_{X_1}(n)\mathbf{t}_1^H\mathbf{\Gamma}_{\mathrm{F}}\mathbf{u}_1}{\lambda_{X_1}(n)\zeta_0 + \mu\lambda_{V_1}(n)\zeta_1}. \tag{18}$$

Substituting (18) into (12), we obtain the second maximum FBR beamformer:

$$\begin{aligned}\mathbf{h}_{\mathrm{FBR},\mu}(n) &= \frac{\lambda_{X_1}(n)\mathbf{t}_1\mathbf{t}_1^H\mathbf{\Gamma}_{\mathrm{F}}\mathbf{u}_1}{\lambda_{X_1}(n)\zeta_0 + \mu\lambda_{V_1}(n)\zeta_1} & (19)\\ &= \frac{\xi_{\mathrm{in}}(n)\mathbf{t}_1\mathbf{t}_1^H\mathbf{\Gamma}_{\mathrm{F}}\mathbf{u}_1}{\xi_{\mathrm{in}}(n)\zeta_0 + \mu\zeta_1}. & (20)\end{aligned}$$

If the value of μ is equal to one, we have $\mathbf{h}_{\mathrm{FBR},\mu=1}(n) = \mathbf{h}_{\mathrm{FBR,MSE}}(n)$. Consequently, the first maximum FBR beamformer can be viewed as a particular case of $\mathbf{h}_{\mathrm{FBR},\mu}(n)$.

B. Power Spectral Density Estimation Based on Frobenius-Norm Minimization

Implementation of the derived simplified beamformers requires the power spectral densities (PSDs) of the speech and noise signals. There are numerous PSD estimation methods in the literature and they can be broadly categorized into single-channel [28]–[32] and multichannel methods [33]–[39]. The single-channel ones are generally based on the sparsity property of the speech signal in the time-frequency domain and stationarity assumption of the noise, while the multichannel methods utilize the spatial coherence of the desired source and the noise. In this work, we will adopt the principle in the multichannel approach and propose the following estimation method.

The unknown parameter vector is defined as $\boldsymbol{\lambda}(n) \triangleq \left[\begin{array}{cc}\lambda_{X_1}(n) & \lambda_{V_1}(n)\end{array}\right]^T$. We propose to estimate the signal's PSD by minimizing the Frobenius-norm of the error matrix between the sample covariance matrix of the observations $\mathbf{S}_{\mathbf{y}}(n)$ and the modelled covariance matrix $\mathbf{C}_{\mathbf{y}}(n)$ in (6), i.e.,

$$g\left[\boldsymbol{\lambda}(n)\right] = \left\|\mathbf{S}_{\mathbf{y}}(n) - \lambda_{X_1}(n)\mathbf{\Gamma}_{\mathrm{F}} - \lambda_{V_1}(n)\mathbf{\Gamma}_{\mathrm{B}}\right\|_{\mathcal{F}}^2. \tag{21}$$

Taking the derivative of $g\left[\boldsymbol{\lambda}(n)\right]$ with respect to $\lambda_{X_1}(n)$ and $\lambda_{V_1}(n)$, respectively, and setting the results to zeros, we obtain a system of linear equations:

$$\begin{aligned}&\underbrace{\left[\begin{array}{cc}\mathrm{tr}\left\{\mathbf{\Gamma}_{\mathrm{F}}^2\right\} & \mathrm{tr}\left\{\mathbf{\Gamma}_{\mathrm{F}}\mathbf{\Gamma}_{\mathrm{B}}\right\}\\ \mathrm{tr}\left\{\mathbf{\Gamma}_{\mathrm{F}}\mathbf{\Gamma}_{\mathrm{B}}\right\} & \mathrm{tr}\left\{\mathbf{\Gamma}_{\mathrm{B}}^2\right\}\end{array}\right]}_{\mathbf{A}}\underbrace{\left[\begin{array}{c}\widehat{\lambda}_{X_1}(n)\\ \widehat{\lambda}_{V_1}(n)\end{array}\right]}_{\widehat{\boldsymbol{\lambda}}(n)}\\ &= \underbrace{\left[\begin{array}{c}\mathrm{tr}\left\{\mathbf{\Gamma}_{\mathrm{F}}\mathbf{S}_{\mathbf{y}}(n)\right\}\\ \mathrm{tr}\left\{\mathbf{\Gamma}_{\mathrm{B}}\mathbf{S}_{\mathbf{y}}(n)\right\}\end{array}\right]}_{\mathbf{b}(n)}, \qquad (22)\end{aligned}$$

where $\widehat{\lambda}_{X_1}(n)$ and $\widehat{\lambda}_{V_1}(n)$ are the estimates of $\lambda_{X_1}(n)$ and $\lambda_{V_1}(n)$, respectively. Using the fact that $\left|\mathrm{tr}\left\{\mathbf{B}_1\mathbf{B}_2^H\right\}\right|^2 \leq \mathrm{tr}\left\{\mathbf{B}_1^H\mathbf{B}_1\right\}\mathrm{tr}\left\{\mathbf{B}_2^H\mathbf{B}_2\right\}$ where the equality holds if and only if $\mathbf{B}_1 = \mathbf{B}_2$, one can prove that the determinant of $\mathbf{A}$ in (22) is greater than zero. So, $\mathbf{A}$ is invertible and the estimate of $\boldsymbol{\lambda}(n)$ is obtained as $\widehat{\boldsymbol{\lambda}}(n) = \mathbf{A}^{-1}\mathbf{b}(n)$. Since $\lambda_{X_1}(n)$ and $\lambda_{V_1}(n)$ represent PSDs, their estimates should be nonnegative. So, in real implementation, if the estimate $\widehat{\lambda}_{X_1}(n)$ or $\widehat{\lambda}_{V_1}(n)$ is negative, one can apply the half-wave rectifying idea to force the negative estimate to be zero.

IV. Simulations

In this section, we will evaluate the performance of the two developed beamformers in a simulated classroom environment.

A. Setup

The length, width and height of the simulated classroom are, respectively, 6.00 m, 4.00 m, and 3.00 m. The walls, ceiling and floor of the room are assumed to have the same reflection coefficient, 0.8. With this setup, the reverberation time of the room is approximately 300 ms. The desired source (simulate a teacher) is located at (5.30,2.10,1.60) and four noise sources (simulate four students) are located at (1.80,0.50:1.00:3.50,1.60), respectively. The desired source is a loudspeaker playing back a female speech signal recorded in an anechoic chamber. The signal length is approximately 30 s and the sampling rate is 16 kHz. The noise source signals are computer generated white Gaussian sequences. We consider to use a small ULA with $M = 3$ omnidirectional microphones whose locations are (4.50:$-$0.02:4.46,2.10,1.60), respectively. The array is placed in such a way that the desired source is located in the endfire direction of the ULA. We consider the microphone close to the desired source as the first microphone and also the reference. The acoustic impulse responses from the sources to the microphones are simulated with the well-known image-model method [40].

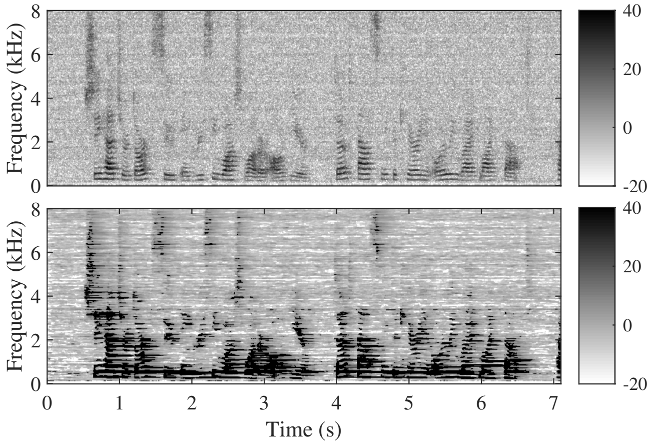

Fig. 1. The top panel plots the spectrogram (in dB) of the observation signal at the first microphone when the input SNR is 10 dB and the bottom panel plots the subband input SNR (in dB) calculated with the proposed PSD estimator.

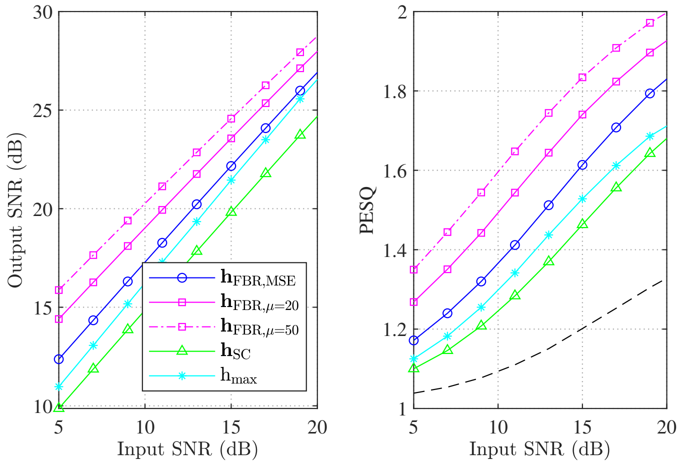

Fig. 2. The output SNR and PESQ of all considered beamformers as a function of the input SNR. The dashed line in the right panel represents the PESQ value of the first microphone signal.

B. Implementation

The signals are partitioned into overlapping short frames with a frame length of 512 and an overlapping ratio of 75%. A Hamming window of length 512 is applied to every frame and the windowed frame is then transformed to the STFT domain using the fast Fourier transform (FFT) of size 512. The sample covariance matrix of the observations, $\mathbf{S}_{\mathbf{y}}(k,n)$, is obtained with the following recursive averaging method:

$$\mathbf{S}_{\mathbf{y}}(k,n) = \eta \mathbf{S}_{\mathbf{y}}(k,n-1) + (1-\eta)\mathbf{y}(k,n)\mathbf{y}^H(k,n), \quad (23)$$

where $0 \leq \eta < 1$ is a forgetting factor. In our simulations, we set the forgetting factor as $\eta = 0.85$. With the estimated PSDs, we can compute $\xi_{\mathrm{in}}(k,n)$ according to (9). In order to control the level of speech distortion in the beamformer's output, the value of $\xi_{\mathrm{in}}(k,n)$ is restricted to the range between 10^{-2} and 10^4.

C. Results

Figure 1 plots the spectrogram of the first microphone signal at a 10-dB input SNR and the subband input SNR calculated with the estimated PSDs. As seen, the estimated subband input SNR has very small values in the noise-only periods while larger values in the periods where speech is present. This visualization shows that the proposed estimation method works pretty well on estimation of the PSDs of the speech and noise signals.

Now, we consider to compare the developed beamformers with the supercardioid (SC) [24], [26] and conventional maximum SNR [11] beamformers. SC is a fixed beamformer that maximizes the FBR and has a unit response at the direction $\theta = 0$. The maximum SNR beamformer relies on the covariance matrices of speech and noise. Since the noise in our simulations is stationary, we estimate the noise covariance matrix with the observations at the first 40 noise-only frames and keep it constant during the entire processing. The speech covariance matrix is then obtained by subtracting the estimated noise covariance matrix from the sample covariance matrix of the observations. Figure 2 plots the fullband output SNR (see [24] for its definition) and the Perceptual Evaluation of Speech Quality (PESQ) [41], [42] of the studied beamformers as a function of the input SNR. To calculate the PESQ, the clean speech source signal is assumed to be known and used as the reference signal. From Fig. 2, one can see that the SC beamformer yields the lowest output SNR and the lowest PESQ under all the studied input SNR conditions. Both the two developed beamformers outperform the two baseline beamformers in terms of the output SNR and PESQ in this simulated classroom environment. This validates the effectiveness of the covariance matrix model in (6) as well as the developed beamformers. We also observe that between the two developed beamformers given in, respectively, (16) and (20), the second one with $\mu = 50$ leads to the best results in both evaluation measures. It indicates that setting a proper value of μ can help control the amount of noise attenuation and the overall speech quality.

V. Conclusions

Implementation of the conventional maximum SNR beamformer requires reliable estimates of the signal and noise covariance matrices, which are challenging to obtain in practical acoustic environments. In this paper, we first presented a parametric model with small-spacing microphone arrays, which exploits the *a priori* information of the spatial distributions of the desired and the noise sources to model the covariance matrices. Two simplified maximum SNR beamformers were then deduced, which require only the signal and noise PSDs. Since PSDs are easier to estimate than the covariance matrices, the deduced maximum SNR beamformers are more convenient to use than their traditional counterparts in practical applications. Moreover, simulations demonstrated that the developed beamformers are able to produce better performance in terms of SNR and PESQ improvement in comparison with the conventional maximum SNR and supercardioid beamformers.

References

[1] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Berlin, Germany: Springer-Verlag, 2001.

[2] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[3] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, Inc., 2018.

[4] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, pp. 1408–1418, Aug. 1969.

[5] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 158–170, Jan. 2010.

[6] C. Pan, J. Chen, and J. Benesty, "Performance study of the MVDR beamformer as a function of the source incidence angle," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, pp. 67–79, Jan. 2014.

[7] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. IEEE ICASSP*, 2016, pp. 5210–5214.

[8] Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, and D. Yu, "ADL-MVDR: All deep learning MVDR beamformer for target speech separation," in *Proc. IEEE ICASSP*, 2021, pp. 6089–6093.

[9] H. L. Van Trees, *Optimum Array Processing*. John Wiley & Sons, Inc., 2002.

[10] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1529–1539, July 2007.

[11] G. Huang, J. Benesty, T. Long, and J. Chen, "A family of maximum SNR filters for noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2034–2047, Dec. 2014.

[12] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, pp. 2230–2244, Sep. 2002.

[13] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260–276, Feb. 2010.

[14] R. Serizel, M. Moonen, B. Van Dijk, and J. Wouters, "Low-rank approximation based multichannel Wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 785–799, Apr. 2014.

[15] J. R. Jensen, J. Benesty, and M. G. Christensen, "Noise reduction with optimal variable span linear filters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 631–644, Apr. 2016.

[16] N. Ito, R. Ikeshita, H. Sawada, and T. Nakatani, "A joint diagonalization based efficient approach to underdetermined blind audio source separation using the multichannel Wiener filter," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1950–1965, 2021.

[17] F. Zhang, C. Pan, J. Benesty, and J. Chen, "A simplified Wiener beamformer based on covariance matrix modelling," in *Proc. IEEE ICASSP*, 2021, pp. 796–800.

[18] A. Spriet, M. Moonen, and J. Wouters, "Robustness analysis of multichannel Wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications," *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 4, pp. 487–503, July 2005.

[19] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel Wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1368–1381, Jul. 2011.

[20] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *J. Acoust. Soc. Am.*, vol. 54, no. 3, pp. 771–785, Jan. 1973.

[21] L. Ehrenberg, S. Gannot, A. Leshem, and E. Zehavi, "Sensitivity analysis of MVDR and MPDR beamformers," in *IEEE 26th Convention of Electr., and Electron., Eng., in Israel*, 2010, pp. 416–420.

[22] J. Benesty and J. Chen, *Study and Design of Differential Microphone Arrays*. Berlin, Germany: Springer-Verlag, 2012.

[23] J. Benesty, J. Chen, and C. Pan, *Fundamentals of Differential Beamforming*. Berlin, Germany: Springer-Verlag, 2016.

[24] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of Signal Enhancement and Array Signal Processing*. John Wiley & Sons, Inc., 2018.

[25] R. N. Marshall and W. R. Harry, "A new microphone providing uniform directivity over an extended frequency range," *J. Acoust. Soc. Am.*, vol. 12, no. 4, pp. 481–498, 1941.

[26] X. Wang, J. Benesty, I. Cohen, and J. Chen, "Microphone array beamforming based on maximization of the front-to-back ratio," *J. Acoust. Soc. Am.*, vol. 144, no. 6, pp. 3450–3464, Dec. 2018.

[27] D. A. Florencio and M. S. Malvar, "Multichannel filtering for optimum noise reduction in microphone arrays," in *Proc. IEEE ICASSP*, 2001.

[28] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

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

[30] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[31] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[32] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1404–1415, 2020.

[33] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE ICASSP*, 1988, pp. 2578–2581.

[34] I. A. Mccowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[35] H. Q. Dam, S. Nordholm, H. H. Dam, and S. Y. Low, "Maximum likelihood estimation and Cramer-Rao lower bounds for the multichannel spectral evaluation in hands-free communication," in *2005 Asia-Pacific Conference on Communications*, 2005, pp. 961–964.

[36] J. Jensen and M. S. Pedersen, "Analysis of beamformer directed single-channel noise reduction system for hearing aid applications," in *Proc. IEEE ICASSP*, 2015, pp. 5728–5732.

[37] Y. A. Huang, A. Luebs, J. Skoglund, and W. B. Kleijn, "Globally optimized least-squares post-filtering for microphone array speech enhancement," in *Proc. IEEE ICASSP*, 2016, pp. 380–384.

[38] A. Zahedi, M. S. Pedersen, J. Østergaard, L. Bramsløw, T. U. Christiansen, and J. Jensen, "A constrained maximum likelihood estimator of speech and noise spectra with application to multi-microphone noise reduction," in *Proc. IEEE ICASSP*, 2020, pp. 6944–6948.

[39] C. Pan, J. Chen, and G. Shi, "On estimation of time-varying variances of source and noise for sensor array processing," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2865–2879, 2020.

[40] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[41] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[42] P. C. Loizou, *Speech Enhancement: Theory and Practice (2nd ed.)*. CRC Press, 2013.