Adaptive and Hybrid Kronecker Product Beamforming for Far-Field Speech Signals

Rajib Sharma[a,*], Israel Cohen[a,1], Jacob Benesty[b]

[a] *Andrew and Erna Viterby Faculty of Electrical Engineering, Technion - Israel Institute of Technology.*
[b] *INRS-EMT, University of Quebec, 800 de la Gauchetiere Ouest, Montreal, QC H5A 1K6, Canada.*

Abstract

This work presents a Kronecker product based methodology of frequency-domain beamforming of large sensor arrays for far-field broadband speech signals. The principal idea involves splitting up a given uniform linear array (ULA) into two smaller virtual ULAs (VULAs), using the Kronecker product. The linear system of the original ULA is bifurcated into two smaller linear systems of the VULAs. Henceforth, traditional adaptive beamformers such as the minimum-variance-distortionless-response (MVDR) beamformer may be obtained for each of the VULAs, using lesser data to estimate the statistics. The short-length beamformers, obtained from the VULAs, are finally combined by the Kronecker product to derive the full-length Kronecker product beamformer. Additionally, the VULAs allow fixed and adaptive beamforming to be implemented separately on each of them. As fixed beamformers do not employ statistical information, the Kronecker product hybrid beamformers reduce the original linear system to just a small linear system involving one VULA. Accordingly, hybrid beamformers may be implemented using traditional fixed beamformers, such as the delay-and-sum (DS) beamformer, on one VULA, and traditional adaptive beamformers, such as the MVDR, on the other. The proposed Kronecker product beamformers are observed to provide faster convergence and superior robustness with respect to the traditional beamformers.

Keywords: Adaptive beamformer, Hybrid beamformer, Kronecker product, Robust beamforming.

1. Introduction

Beamforming is the task of conserving a signal received by an array of sensors from a particular direction and source while trying to attenuate the interferences and noise-signals impinging on it from other directions and sources [1–3]. It involves applying a filter to the data received by the sensor array, resulting in a signal which is an accurate estimate of the signal-of-interest (SOI) impinging from the particular direction [1–3]. One way of doing so is to design a filter based solely on the knowledge of the direction-of-arrival (DOA) of the SOI, and sometimes also the DOAs of the interferences - this is called fixed beamforming [3]. A more robust way, additionally, involves utilizing the knowledge of the statistics of the data. Such a method is called adaptive beamforming [2–5]. When the DOA of the SOI is known, and there is a limited effect of interference, a fixed beamformer is a very useful and efficient solution. As, in reality, such situations seldom exist, adaptive beamformers are a more sensible option. Over the years, a plethora of fixed and adaptive beamformers have been developed, out of which the delay-and-sum (DS) and minimum-variance-distortionless-response (MVDR) beamformers are well-appreciated, and are utilized in this work [3].

As is apparent, the performance of both fixed and adaptive beamformers depend on the accuracy of the DOA estimate. An adaptive beamformer also depends on the accuracy of the estimated data-statistics. Therefore, there has always been a lot of focus on improving the performances of beamformers based on the better estimation of the steering vector and (or) quicker and more accurate tracking of the second-order statistics of the data [6–21]. Concurrently, recent technological advancements are driving the ever-evolving design of high-density sensor arrays, consisting of a large number of sensors, to obtain better performances [22]. Such developments have brought the challenge of accurately estimating the second-order statistics from limited data, and processing such information efficiently. These challenges have led to innovative refinements of conventional adaptive algorithms used for efficient implementation of beamforming filters, the popular ones being the multi-stage wiener (MSW), reduced-rank linearly constrained minimum variance (RRLCMV), and their widely-linear variants [2, 23–31].

It is worthwhile noting that this work is not another refinement of adaptive filtering algorithms. Rather, this work provides a theoretical framework to tackle the abovementioned challenges at a higher level of abstraction, i.e.,

*Corresponding author
Email addresses: rajibd2k@yahoo.com (Rajib Sharma), icohen@ee.technion.ac.il (Israel Cohen), benesty@emt.inrs.ca (Jacob Benesty)

[1] This research was supported by the Israel Science Foundation (grant no. 576/16) and the ISF-NSFC joint research program (grant No. 2514/17).

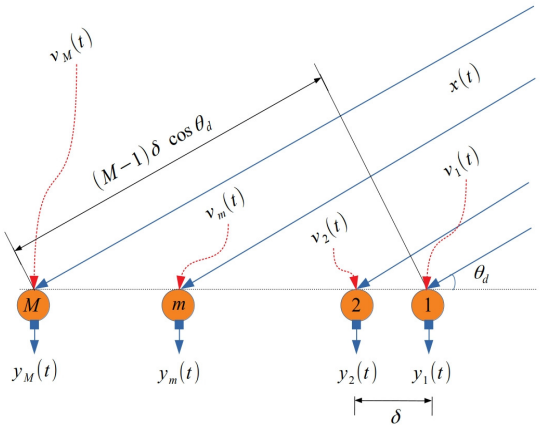

Figure 1: **An ULA of M sensors, on which a discrete-time SOI, $x(t)$, impinges upon at DOA, $\theta_{\rm d}$. The inter-sensor distance is denoted by δ. The combined effect of interferences and sensor-noise at the $m^{\rm th}$ sensor is denoted by $v_m(t)$, and $y_m(t)$ denotes the final signal detected at the sensor.**

in the beamformer/filter design level. In this work, we propose a methodology to mathematically separate a large array of sensors into two smaller arrays using the Kronecker product [32, 33]. Henceforth, we propose the methodology to obtain new beamformers for the original array by combining the traditional beamformers of the smaller sub-arrays. The practical implementation of the proposed beamformers may be carried out using suitable adaptive algorithms, or new algorithms (based on RLS, LMS, etc.) may be developed that is more suitable for our proposed framework of beamforming. That is not the scope of this work, and may be dealt with separately. As such, this work must not be confused or compared with adaptive filtering algorithms such as the MSW, RRLCMV, etc.

Using the traditional MVDR and DS beamformers as examples, we show how to utilize the proposed theoretical framework to make beamforming for large arrays more efficient and effective. As will be illustrated in this work, the proposed variants of traditional frequency domain beamformers are more robust to unstable/erroneous estimates of the data-statistics and various levels of interferences. It is assumed that the DOA of the SOI is known. The interferences are broadband noise signals taken from the NOISEX-92 database [34], and the noise in the sensors are considered as Gaussian white noise signals. Inspired by our recent works [35, 36] using the Kronecker product in differential (fixed) beamforming, this work experiments the utility of the same in adaptive beamforming. The framework also allows the combination of fixed and adaptive beamforming to generate hybrid beamformers. Uniform linear arrays (ULAs) [2–5] are used in our study.

At this juncture, we must also note that adaptive beamforming has generally been implemented on narrowband modulated signals used in communication technologies [2, 23–31]. Such signals are near-sinusoids and quasi-stationary in nature, and hence time-domain beamforming is suitable. However, in the case of non-stationary broadband signals like speech, frequency domain beamforming may be more relevant [37–39]. Moreover, while traditionally microphone array beamforming has been confined to near-field applications [40–42], the developments in sensor technology now enable us to employ large microphone arrays in far-field applications, such as surveillance in crowded environments, trading in large open halls, and other distant speech and speaker recognition tasks [43–45]. This work is devoted to beamforming for such evolving and futuristic applications.

The rest of this work is organized as follows: Section 2 formulates the beamforming problem, the various statistical and non-statistical metrics, and presents the traditional DS and MVDR beamformers. Section 3 describes the Kronecker product based beamforming methodology. Section 4 presents the experimental results, and compares the proposed beamformers with their traditional counterparts. Section 5 summarizes and concludes this work.

2. Signal Model and Conventional Filters

We consider an arbitrary ULA of M sensors, as shown in Figure 1, with the sensors located at arbitrary positions denoted by $\{\delta_m = (m-1)\delta : \ m = 1, 2, ..., M\}$, where δ is the inter-sensor distance. A discrete-time SOI, $x(t)$, where t denotes the discrete-time index, impinges on the ULA as a plane-wave in the far-field, traveling at the velocity of sound, c, through the medium. Similarly, K independent interferences, $\{u_k(t) : k = 1, 2, ..., K\}$, impinge on the ULA. We consider the DOA of the SOI as $\theta_{\rm d}$, and the DOAs of the interferences as $\{\theta_k : k = 1, 2, ..., K\}$. Each of the M sensors are afflicted with their own thermal noise, denoted by $\{w_m(t) : \ m = 1, 2, ..., M\}$. In this work, all signals are assumed to have zero statistical means, and $x(t)$, $u_k(t)$, $w_m(t)\ \forall\ k,\ m$ are uncorrelated. The interferences are considered as IID broadband signals, and so are the sensor-noise signals. The signals are sampled (sensed/detected) at the sensors at a sampling-frequency of $F_s\ (= 1/T_s)$. In the case of frequency domain beamforming, the data is processed in small blocks of samples, called frames or snapshots. Thus, the data sensed at the $m^{\rm th}$ sensor, corresponding to the $r^{\rm th}$ snapshot, may be represented as

$$\begin{aligned} y_m(t,r) &= x_m(t,r) + v_m(t,r)\ ,\ m = 1, 2, ..., M, \\ v_m(t,r) &= \sum_{k=1}^{K} u_{k,m}(t,r) + w_m(t,r), \\ x_m(t,r) &= x\left(t - \frac{\delta_m \cos\theta_d}{cT_s}, r\right), \\ u_{k,m}(t,r) &= u_k\left(t - \frac{\delta_m \cos\theta_k}{cT_s}, r\right). \end{aligned} \tag{1}$$

One must note in the last two equations of (1) that the time-delays may not be integers. As such, the discrete-time signals at the sensors are not merely sample-shifted versions of one another. The discrete-time signals are sampled versions of the continuous-time signals impinging on

the ULA. As such, if the SOI and/or interferences are non-stationary[2]signals, then, for any given snapshot, the data received at two distant sensors may have very different characteristics. Henceforth, the array size, $(M-1)\delta$, must be limited to maintain significant coherence among the data received (for any given snapshot) by the sensors. Under these conditions, the data pertaining to the r^{th} snapshot may be represented in the time-frequency domain by short-time Fourier Transform (STFT) as

$$\begin{aligned}
Y_m(f,r) &= X_m(f,r) + V_m(f,r),\\
V_m(f,r) &= \sum_{k=1}^{K} U_{k,m}(f,r) + W_m(f,r),\\
X_m(f,r) &= \exp\left(-j2\pi f \frac{\delta_m \cos\theta_d}{cT_s}\right) X(f,r)\\
&= d_{\theta_{\text{d}},m}(f) X(f,r),\\
U_{k,m}(f,r) &= \exp\left(-j2\pi f \frac{\delta_m \cos\theta_k}{cT_s}\right) U_k(f,r)\\
&= d_{\theta_k,m}(f) U_k(f,r).
\end{aligned} \tag{2}$$

Now, the data received across all the M sensors, at any frequency, f, and snapshot, r, may be represented as

$$\begin{aligned}
\mathbf{y}(f,r) &= [Y_1(f,r), ..., Y_m(f,r), ..., Y_M(f,r)]^T\\
&= \mathbf{x}(f,r) + \mathbf{v}(f,r),\\
\mathbf{x}(f,r) &= \mathbf{d}_{\theta_{\text{d}}}(f) X(f,r),\\
\mathbf{v}(f,r) &= \sum_{k=1}^{K} \mathbf{u}_k(f,r) + \mathbf{w}(f,r)\\
&= \sum_{k=1}^{K} \mathbf{d}_{\theta_k}(f) U_k(f,r) + \mathbf{w}(f,r),\\
\mathbf{d}_{\theta_{\text{d}}}(f) &= [1 \ ... \ d_{\theta_{\text{d}},m}(f) \ ... \ d_{\theta_{\text{d}},M}(f)]^T,\\
\mathbf{d}_{\theta_k}(f) &= [1 \ ... \ d_{\theta_k,m}(f) \ ... \ d_{\theta_k,M}(f)]^T.
\end{aligned} \tag{3}$$

In (3), $\mathbf{y}(f,r)$ represents the M-dimensional data sensed by the M-component sensor array. Similarly, $\mathbf{x}(f,r)$, $\mathbf{v}(f,r)$, $\mathbf{u}(f,r)$, $\mathbf{u}_k(f,r)$, and $\mathbf{w}(f,r)$ are M-dimensional vectors representing the respective components of the data across the sensor array. Again, $\mathbf{d}_{\theta_{\text{d}}}(f)$ and $\mathbf{d}_{\theta_k}(f)$ represent the M-dimensional steering vectors of the SOI and the k^{th} interference, respectively.

The objective of beamforming is to apply an M-dimensional filter, $\mathbf{h}(f,r)$, on the data-vector, $\mathbf{y}(f,r)$, so as to obtain a signal, $Z(f,r) \approx X(f,r)$.

$$\begin{aligned}
Z(f,r) &= \mathbf{h}^H(f,r)\mathbf{y}(f,r) = X_{\text{fd}}(f,r) + V_{\text{rn}}(f,r),\\
X_{\text{fd}}(f,r) &= \mathbf{h}^H(f,r)\mathbf{d}_{\theta_{\text{d}}}(f) X(f,r),\\
V_{\text{rn}}(f,r) &= \mathbf{h}^H(f,r)\mathbf{v}(f,r).
\end{aligned} \tag{4}$$

In (4), $X_{\text{fd}}(f,r)$ and $V_{\text{rn}}(f,r)$ represent the filtered-desired and the residual-noise signal, respectively. Ideally, one would want $V_{\text{rn}}(f,r) = 0$, while obtaining a distortionless estimate, i.e., $\mathbf{h}^H(f,r)\mathbf{d}_{\theta_{\text{d}}}(f) = 1$. The time-domain output, after beamforming, is obtained by applying Inverse DFT on the frequency domain output, $Z(f,r)$, followed by Overlap and Add (OLA) method [3, 46] to stitch together the samples of consecutive snapshots:

$$\begin{aligned}
&z(t,r) \longleftrightarrow Z(f,r)\\
&z(t) = \sum_{r} z(t,r) \ , \text{ using OLA.}
\end{aligned} \tag{5}$$

The variance of the output, $Z(f,r)$, is given by

$$\begin{aligned}
\phi_Z(f,r) &= E\{|Z(f,r)|^2\} = \mathbf{h}^H(f,r)\mathbf{\Phi}_{\mathbf{y}}(f,r)\mathbf{h}(f,r)\\
&= \phi_{X_{\text{fd}}}(f,r) + \phi_{V_{\text{rn}}}(f,r),\\
\phi_{X_{\text{fd}}}(f,r) &= E\{|X_{\text{fd}}(f,r)|^2\}\\
&= \mathbf{h}^H(f,r)\mathbf{d}_{\theta_{\text{d}}}(f)\phi_X(f,r)\mathbf{d}^H_{\theta_{\text{d}}}(f)\mathbf{h}(f,r),\\
\phi_X(f,r) &= E\big\{|X(f,r)|^2\big\},\\
\phi_{V_{\text{rn}}}(f,r) &= E\{|V_{\text{rn}}(f,r)|^2\}\\
&= \mathbf{h}^H(f,r)\mathbf{\Phi}_{\mathbf{v}}(f,r)\mathbf{h}(f,r).
\end{aligned} \tag{6}$$

In (6), $\phi_{X_{\text{fd}}}(f,r)$, $\phi_{V_{\text{rn}}}(f,r)$, and $\phi_X(f,r)$ represent the variances of the filtered-desired signal, the residual-noise signal, and the SOI, respectively. Similarly, $\mathbf{\Phi}_{\mathbf{v}}(f,r)$ and $\mathbf{\Phi}_{\mathbf{y}}(f,r)$ represent the $M \times M$ covariance-matrices of the disturbances (interferences plus sensor-noises), and the data, respectively. They are obtained as

$$\begin{aligned}
\mathbf{\Phi}_{\mathbf{v}}(f,r) &= E\big\{\mathbf{v}(f,r)\mathbf{v}^H(f,r)\big\}\\
&= \sum_{k=1}^{K} \mathbf{\Phi}_{\mathbf{u}_k}(f,r) + \mathbf{\Phi}_{\mathbf{w}}(f,r),\\
\mathbf{\Phi}_{\mathbf{u}_k}(f,r) &= E\big[\mathbf{u}_k(f,r)\mathbf{u}_k^H(f,r)\big]\\
&= \mathbf{d}_{\theta_k}(f)\phi_{U_k}(f,r)\mathbf{d}^H_{\theta_k}(f),\\
\phi_{U_k}(f,r) &= E\{|U_k(f,r)|^2\},\\
\mathbf{\Phi}_{\mathbf{w}}(f,r) &= E\big\{\mathbf{w}(f,r)\mathbf{w}^H(f,r)\big\},\\
\mathbf{\Phi}_{\mathbf{y}}(f,r) &= E\big\{\mathbf{y}(f,r)\mathbf{y}^H(f,r)\big\}\\
&= \mathbf{d}_{\theta_{\text{d}}}(f)\phi_X(f,r)\mathbf{d}^H_{\theta_{\text{d}}}(f) + \mathbf{\Phi}_{\mathbf{v}}(f,r).
\end{aligned} \tag{7}$$

If the signals are stationary[2], their characteristics do not vary with time, and the variances and the covariances could be averaged over the snapshots to obtain reliable estimates of their true values. The variable, r, then, may be dropped to represent the estimates of the true statistics.

[2] In signal processing, a stationary signal is one whose frequency (Fourier) spectrum does not vary with time. In practice, if the frequency spectrum of a segment/frame/snapshot of a signal is found to differ significantly from that of its another segment, the signal is deemed non-stationary. Consider that all the signals sensed/detected by the array are perfectly stationary. Then, the statistics of the data, in the frequency domain, such as its covariance and correlation matrices, will be the same for any data segment. In practice, this means that we can average the statistics of the data segments to obtain the true statistics. If only certain signals of the overall data are stationary, then, under certain circumstances, we may be able to obtain the true statistics of only those signals by averaging.

2.1. **Performance Measures**

If reliable estimates of the true statistics are available, then, various statistical measures could be evaluated for the performance evaluation of the beamformers. One such essential statistical parameter is the mean-squared-error (MSE), derived as,

$$\begin{aligned}\mathcal{E}(f,r) &= Z(f,r) - X(f,r) = \mathcal{E}_{\rm d}(f,r) + \mathcal{E}_n(f,r),\\ \mathcal{E}_{\rm d}(f,r) &= [\mathbf{h}^H(f,r)\mathbf{d}_{\theta_{\rm d}}(f) - 1]X(f,r),\\ \mathcal{E}_n(f,r) &= V_{\rm rn}(f,r) = \mathbf{h}^H(f,r)\mathbf{v}(f,r),\end{aligned} \tag{8}$$

and

$$\begin{aligned}J[\mathbf{h}(f,r)] &= E\big\{|\mathcal{E}(f,r)|^2\big\}\\ &= \phi_X(f) + \mathbf{h}^H(f,r)\mathbf{\Phi}_{\mathbf{y}}(f)\mathbf{h}(f,r)\\ &\quad - \phi_X(f)\mathbf{h}^H(f,r)\mathbf{d}_{\theta_{\rm d}}(f)\\ &\quad - \phi_X(f)\mathbf{d}^H_{\theta_{\rm d}}(f)\mathbf{h}(f,r),\\ J[\mathbf{h}(r)] &= \sum_f J[\mathbf{h}(f,r)]\ \Delta f.\end{aligned} \tag{9}$$

In (8) and (9), $\mathcal{E}_{\rm d}(f,r)$ and $\mathcal{E}_n(f,r) = V_{\rm rn}(f,r)$ represent the desired-signal-distortion and the residual-noise, respectively, after filtering. $J[\mathbf{h}(f,r)]$ and $J[\mathbf{h}(r)]$ represent the narrowband and broadband MSE, respectively. Δf represents the frequency resolution of the STFT. The narrowband MSE may also be represented as

$$\begin{aligned}J[\mathbf{h}(f,r)] &= J_d[\mathbf{h}(f,r)] + J_n[\mathbf{h}(f,r)],\\ J_d[\mathbf{h}(f,r)] &= E\big\{|\mathcal{E}_{\rm d}(f,r)|^2\big\}\\ &= \phi_X(f)|\mathbf{h}^H(f,r)\mathbf{d}_{\theta_{\rm d}}(f) - 1|^2,\\ J_n[\mathbf{h}(f,r)] &= E\big\{|\mathcal{E}_n(f,r)|^2\big\}\\ &= \mathbf{h}^H(f,r)\mathbf{\Phi}_{\mathbf{v}}(f)\mathbf{h}(f,r).\end{aligned} \tag{10}$$

The narrowband and broadband input signal-to-interference-ratios (iSIRs), input signal-to-noise-ratios (iSNRs), and input signal-to-interference-plus-noise-ratios (iSINRs) are computed at the first sensor, and given by

$$\begin{aligned}\text{iSIR}(f) &= \frac{\phi_X(f)}{\phi_U(f)},\\ \phi_U(f) &= E\big\{|U(f,r)|^2\big\} = \sum_{k=1}^{K} E\big\{|U_k(f,r)|^2\big\},\\ \text{iSIR} &= \frac{\sum_f \phi_X(f)\ \Delta f}{\sum_f \phi_U(f)\ \Delta f} = \frac{E\big\{x^2(t)\big\}}{\sum_{k=1}^{K} E\big\{u_k^2(t)\big\}},\end{aligned} \tag{11}$$

and

$$\begin{aligned}\text{iSNR}(f) &= \frac{\phi_X(f)}{\phi_{W_1}(f)} = \frac{\phi_X(f)}{E\big\{|W_1(f,r)|^2\big\}},\\ \text{iSNR} &= \frac{\sum_f \phi_X(f)\ \Delta f}{\sum_f \phi_{W_1}(f)\ \Delta f} = \frac{E\big\{x^2(t)\big\}}{E\big\{w_1^2(t)\big\}},\end{aligned} \tag{12}$$

and

$$\begin{aligned}\text{iSINR}(f) &= \frac{\phi_X(f)}{\phi_{V_1}(f)} = \frac{\phi_X(f)}{E\big\{|V_1(f,r)|^2\big\}},\\ \text{iSINR} &= \frac{\sum_f \phi_X(f)\ \Delta f}{\sum_f \phi_{V_1}(f)\ \Delta f} = \frac{E\big\{x^2(t)\big\}}{E\big\{v_1^2(t)\big\}}\\ &= \frac{E\big\{x^2(t)\big\}}{E\big\{w_1^2(t)\big\} + \sum_{k=1}^{K} E\big\{u_k^2(t)\big\}}.\end{aligned} \tag{13}$$

The narrowband and broadband output signal-to-interference-plus-noise-ratios (oSINRs) are given by

$$\begin{aligned}\text{oSINR}[\mathbf{h}(f,r)] &= \frac{E\big\{|X_{\rm fd}(f,r)|^2\big\}}{E\big\{|V_{\rm rn}(f,r)|^2\big\}}\\ &= \frac{\phi_X(f)|\mathbf{h}^H(f,r)\mathbf{d}_{\theta_{\rm d}}(f)|^2}{\mathbf{h}^H(f,r)\mathbf{\Phi}_{\mathbf{v}}(f)\mathbf{h}(f,r)},\\ \text{oSINR}[\mathbf{h}(r)] &= \frac{\sum_f \phi_X(f)|\mathbf{h}^H(f,r)\mathbf{d}_{\theta_{\rm d}}(f)|^2\ \Delta f}{\sum_f \mathbf{h}^H(f,r)\mathbf{\Phi}_{\mathbf{v}}(f)\mathbf{h}(f,r)\ \Delta f}.\end{aligned} \tag{14}$$

The narrowband and broadband gains are obtained as

$$G[\mathbf{h}(f,r)] = \frac{\text{oSINR}[\mathbf{h}(f,r)]}{\text{iSINR}(f)}\ ,\ G[\mathbf{h}(r)] = \frac{\text{oSINR}[\mathbf{h}(r)]}{\text{iSINR}}. \tag{15}$$

Apart from the aforementioned statistical measures, the beampattern will be used in this work. The beampattern of a beamformer represents its sensitivity to a plane wave impinging on the array at any arbitrary direction, θ. It is defined as

$$\begin{aligned}\beta_\theta[\mathbf{h}(f,r)] &= \frac{|\mathbf{d}_\theta^H(f)\mathbf{h}(f,r)|}{\max_\theta |\mathbf{d}_\theta^H(f)\mathbf{h}(f,r)|},\\ \beta_\theta[\mathbf{h}(f)] &= \frac{\sum_r \beta_\theta[\mathbf{h}(f,r)]}{\max_\theta \sum_r \beta_\theta[\mathbf{h}(f,r)]}.\end{aligned} \tag{16}$$

2.2. **DS Beamformer**

Irrespective of whether the true statistics is available, the narrowband gain may be expressed (not evaluated) as

$$\begin{aligned}G[\mathbf{h}(f,r)] &= \frac{|\mathbf{h}^H(f,r)\mathbf{d}_{\theta_{\rm d}}(f)|^2}{\mathbf{h}^H(f,r)\mathbf{\Gamma}_{\mathbf{v}}(f,r)\mathbf{h}(f,r)},\\ \mathbf{\Phi}_{\mathbf{v}}(f,r) &= \phi_{V_1}(f,r)\mathbf{\Gamma}_{\mathbf{v}}(f,r).\end{aligned} \tag{17}$$

Let us consider, $\mathbf{\Gamma}_{\mathbf{v}}(f,r) = \mathbf{I}_M$, where $\mathbf{I}_M$ is an $M \times M$ identity matrix. Then, we obtain the narrowband white-noise-gain (WNG) as

$$\text{WNG}[\mathbf{h}(f,r)] = \frac{|\mathbf{h}^H(f,r)\mathbf{d}_{\theta_{\rm d}}(f)|^2}{\mathbf{h}^H(f,r)\mathbf{h}(f,r)}. \tag{18}$$

One may note that the narrowband WNG is devoid of any statistical parameters. Maximizing the narrowband WNG subject to the distortionless criteria provides us with the DS beamformer, given by

$$\begin{aligned}&\min_{\mathbf{h}(f,r)} \mathbf{h}^H(f,r)\mathbf{h}(f,r)\ \text{s.t.}\ \mathbf{h}^H(f,r)\mathbf{d}_{\theta_{\rm d}}(f) = 1,\\ &\mathbf{h}(f,r) = \frac{\mathbf{d}_{\theta_{\rm d}}(f)}{M} = \mathbf{h}(f).\end{aligned} \tag{19}$$

2.3. MVDR Beamformer

The MVDR beamformer is an adaptive beamformer which attempts to minimize the variance of the residual-noise, $\phi_{V_{\rm rn}}(f,r)$, while attempting to recover the SOI without distortion. As (6) shows, this is equivalent to minimizing the variance of the final output, $\phi_Z(f,r)$, subject to the distortionless criteria:

$$\begin{aligned} &\min_{\mathbf{h}(f,r)} \phi_{V_{\rm rn}}(f,r) \text{ s.t. } \mathbf{h}^H(f,r)\mathbf{d}_{\theta_{\rm d}}(f) = 1, \\ &\mathbf{h}(f,r) = \frac{{\mathbf{\Phi_v}}^{-1}(f,r)\mathbf{d}_{\theta_{\rm d}}(f)}{\mathbf{d}^H_{\theta_{\rm d}}(f){\mathbf{\Phi_v}}^{-1}(f,r)\mathbf{d}_{\theta_{\rm d}}(f)}, \end{aligned} \tag{20}$$

and

$$\begin{aligned} &\min_{\mathbf{h}(f,r)} \phi_{Z}(f,r) \text{ s.t. } \mathbf{h}^H(f,r)\mathbf{d}_{\theta_{\rm d}}(f) = 1, \\ &\mathbf{h}(f,r) = \frac{{\mathbf{\Phi_y}}^{-1}(f,r)\mathbf{d}_{\theta_{\rm d}}(f)}{\mathbf{d}^H_{\theta_{\rm d}}(f){\mathbf{\Phi_y}}^{-1}(f,r)\mathbf{d}_{\theta_{\rm d}}(f)}. \end{aligned} \tag{21}$$

If the SOI, interferences and sensor-noises are all stationary, i.e., the data is stationary, then, reliable estimates of $\mathbf{\Phi_y}(f,r)$ may be obtained by averaging over preceding snapshots. Then, (21) is the prudent choice. If, however, the SOI is non-stationary, while the interferences and sensor-noises are stationary, then, reliable estimates of $\mathbf{\Phi_v}(f,r)$ have to be derived, and (20) has to be implemented.

3. Kronecker Product Beamforming [3]

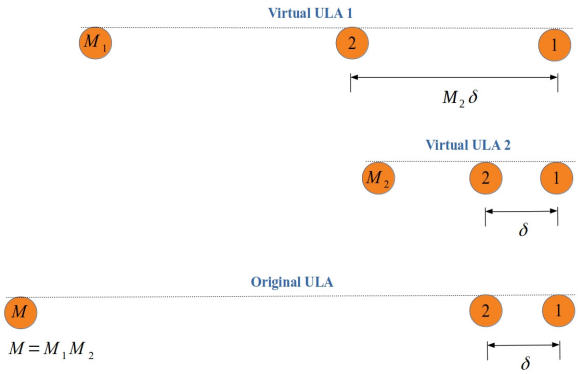

Figure 2: **An ULA of $M = M_1M_2$ sensors, and its two constituent VULAs.**

It is obvious from (20) and (21) that the performance of the MVDR beamformer depends significantly on the accuracy of the the M-dimensional square covariance matrix, $\mathbf{\Phi_v}(f,r)$ or $\mathbf{\Phi_y}(f,r)$. As the number of sensors, M, increases, more data is required to reliably estimate such matrices. If the M-dimensional linear system could be efficiently represented by one or more lower-dimensional systems, there could be improvements in the convergence and robustness of the MVDR beamformer. In this context, we visualize the original ULA as being constituted of two virtual ULAs (VULAs), as depicted in Figure 2. One of the VULAs is comprized of M_1 sensors, and the other of M_2 sensors, such that $M = M_1M_2$. The second VULA has the same inter-sensor distance, $\delta_2 = \delta$, as the original ULA, whereas the first VULA has an inter-sensor distance of $\delta_1 = M_2\delta$. Thus, the original ULA may be visualized as the periodic replication of the second VULA at positions defined by the first VULA. The sensor positions may be represented as

$$\begin{aligned} \delta_m &= (m-1)\delta,\ m = 1,2,...,M = M_1M_2, \\ \delta_{1,p} &= (p-1)\delta_1 = (p-1)M_2\delta,\ p = 1,2,...,M_1, \\ \delta_{2,q} &= (q-1)\delta_2 = (q-1)\delta,\ q = 1,2,...,M_2, \\ \delta_m &= \delta_{(p-1)M_2+q} = \delta_{1,p} + \delta_{2,q}. \end{aligned} \tag{22}$$

The steering vector (for the SOI) of the original ULA may be now represented in terms of the steering vectors of the VULAs:

$$\begin{aligned} \mathbf{d}_{\theta_{\rm d}} &= \mathbf{d}_{1,\theta_{\rm d}} \otimes \mathbf{d}_{2,\theta_{\rm d}}\ , \\ d_{1,\theta_{\rm d},p} &= \exp\left(-j2\pi f \frac{\delta_{1,p}\cos\theta_d}{cT_s}\right),\ 1 \le p \le M_1, \\ d_{2,\theta_{\rm d},q} &= \exp\left(-j2\pi f \frac{\delta_{2,q}\cos\theta_d}{cT_s}\right),\ 1 \le q \le M_2. \end{aligned} \tag{23}$$

In (23), the symbol $\otimes$ represents the Kronecker product. $\mathbf{d}_{1,\theta_{\rm d}}$ and $\mathbf{d}_{2,\theta_{\rm d}}$ are steering vectors (of lengths M_1 and M_2 respectively) of the two VULAs, and $\mathbf{d}_{\theta_{\rm d}}$ is the steering vector (of length $M = M_1M_2$) of the ULA. Henceforth, $d_{1,\theta_{\rm d},p}$ and $d_{2,\theta_{\rm d},q}$ represent the elements of the virtual steering vectors.

Our objective, now, is to derive two separate filters (of lengths M_1 and M_2 respectively) from the two VULAs, which can be used to derive the final filter (of length $M = M_1M_2$). Assuming that two such virtual filters, $\mathbf{h}_1$ and $\mathbf{h}_2$, exist, we have

$$\begin{aligned} \mathbf{h} &= \mathbf{h}_1 \otimes \mathbf{h}_2 \\ &= (\mathbf{h}_1 \otimes \mathbf{I}_{M_2})\mathbf{h}_2 = (\mathbf{I}_{M_1} \otimes \mathbf{h}_2)\mathbf{h}_1. \end{aligned} \tag{24}$$

In (24), $\mathbf{I}_{M_1}$ and $\mathbf{I}_{M_2}$ are square identity matrices of size M_1 and M_2, respectively. The distortionless constraint and narrowband WNG of the original ULA may, now, be

[3] In this section, for notational simplicity, we drop the variables within brackets, f and r, in representing the various parameters. Again, the variables with subscript $_1$ and $_2$ would be utilized in the context of the two VULAs.

factorized as:

$$\begin{aligned}\mathbf{h}^H\mathbf{d}_{\theta_\mathrm{d}} &= (\mathbf{h}_1\otimes\mathbf{h}_2)^H(\mathbf{d}_{1,\theta_\mathrm{d}}\otimes\mathbf{d}_{2,\theta_\mathrm{d}})\\ &= (\mathbf{h}_1^H\mathbf{d}_{1,\theta_\mathrm{d}})(\mathbf{h}_2^H\mathbf{d}_{2,\theta_\mathrm{d}}) = 1,\end{aligned} \tag{25}$$

$$\begin{aligned}\mathrm{WNG} &= \frac{|\mathbf{h}^H\mathbf{d}_{\theta_\mathrm{d}}|^2}{\mathbf{h}^H\mathbf{h}}\\ &= \frac{|\mathbf{h}_1^H\mathbf{d}_{1,\theta_\mathrm{d}}|^2}{\mathbf{h}_1^H\mathbf{h}_1}\times\frac{|\mathbf{h}_2^H\mathbf{d}_{2,\theta_\mathrm{d}}|^2}{\mathbf{h}_2^H\mathbf{h}_2}\\ &= \mathrm{WNG}_1\times\mathrm{WNG}_2.\end{aligned} \tag{26}$$

Using (23) and (24), the variances of the residual-noise and the final output may be represented as

$$\begin{aligned}\phi_{V_\mathrm{rn}} &= \mathbf{h}_1^H\mathbf{\Phi}_{\mathbf{v},2}\mathbf{h}_1 = \mathbf{h}_2^H\mathbf{\Phi}_{\mathbf{v},1}\mathbf{h}_2,\\ \mathbf{\Phi}_{\mathbf{v},2} &= (\mathbf{I}_{M_1}\otimes\mathbf{h}_2)^H\mathbf{\Phi}_{\mathbf{v}}(\mathbf{I}_{M_1}\otimes\mathbf{h}_2),\\ \mathbf{\Phi}_{\mathbf{v},1} &= (\mathbf{h}_1\otimes\mathbf{I}_{M_2})^H\mathbf{\Phi}_{\mathbf{v}}(\mathbf{h}_1\otimes\mathbf{I}_{M_2}).\end{aligned} \tag{27}$$

$$\begin{aligned}\phi_Z &= \mathbf{h}_1^H\mathbf{\Phi}_{\mathbf{y},2}\mathbf{h}_1 = \mathbf{h}_2^H\mathbf{\Phi}_{\mathbf{y},1}\mathbf{h}_2,\\ \mathbf{\Phi}_{\mathbf{y},2} &= (\mathbf{I}_{M_1}\otimes\mathbf{h}_2)^H\mathbf{\Phi}_{\mathbf{y}}(\mathbf{I}_{M_1}\otimes\mathbf{h}_2),\\ \mathbf{\Phi}_{\mathbf{y},1} &= (\mathbf{h}_1\otimes\mathbf{I}_{M_2})^H\mathbf{\Phi}_{\mathbf{y}}(\mathbf{h}_1\otimes\mathbf{I}_{M_2}).\end{aligned} \tag{28}$$

3.1. Kronecker Product Adaptive Beamformer

Consider $\mathbf{h}_2$ exists, and $\mathbf{h}_2^H\mathbf{d}_{2,\theta_\mathrm{d}} = 1$. Then, as is evident from (25) and (28), the variance of the final output, and the distortionless constraint may be expressed as

$$\phi_Z[\mathbf{h}_1|\mathbf{h}_2] = \mathbf{h}_1^H\mathbf{\Phi}_{\mathbf{y},2}\mathbf{h}_1,\ \mathbf{h}^H\mathbf{d}_{\theta_\mathrm{d}} = \mathbf{h}_1^H\mathbf{d}_{1,\theta_\mathrm{d}} = 1. \tag{29}$$

Minimizing $\phi_Z[\mathbf{h}_1|\mathbf{h}_2]$ with respect to $\mathbf{h}_1$, subject to the distortonless criteria, results in the MVDR beamformer for the first VULA.

Similarly, if $\mathbf{h}_1$ exists, and $\mathbf{h}_1^H\mathbf{d}_{1,\theta_\mathrm{d}} = 1$, we have,

$$\phi_Z[\mathbf{h}_2|\mathbf{h}_1] = \mathbf{h}_2^H\mathbf{\Phi}_{\mathbf{y},1}\mathbf{h}_2,\ \mathbf{h}^H\mathbf{d}_{\theta_\mathrm{d}} = \mathbf{h}_2^H\mathbf{d}_{2,\theta_\mathrm{d}} = 1. \tag{30}$$

Minimizing $\phi_Z[\mathbf{h}_2|\mathbf{h}_1]$ with respect to $\mathbf{h}_2$, subject to the distortonless criteria, results in the MVDR beamformer for the second VULA. Thus, we may formulate an iterative procedure to derive the Kronecker-product-minimum-variance-distortionless-response (KP-MVDR) beamformer.

It is evident that the second VULA is a sub-part of the ULA and consists of its first M_2 sensors. Hence, we may, initially (and independently), obtain the MVDR beamformer from it, as

$$\begin{aligned}&\mathbf{\Phi}_{\mathbf{y}_2} = E\{\mathbf{y}_2\mathbf{y}_2^H\},\ \mathbf{y}_2 = [Y_1, Y_2, ..., Y_{M_2}]^T,\\ &\mathbf{h}_2^{(0)} = \frac{\mathbf{\Phi}_{\mathbf{y}_2}^{-1}\mathbf{d}_{2,\theta_\mathrm{d}}}{\mathbf{d}_{2,\theta_\mathrm{d}}^H\mathbf{\Phi}_{\mathbf{y}_2}^{-1}\mathbf{d}_{2,\theta_\mathrm{d}}},\ [\mathbf{h}_2^{(0)}]^H\mathbf{d}_{2,\theta_\mathrm{d}} = 1.\end{aligned} \tag{31}$$

In the preceding equation-set, $\mathbf{\Phi}_{\mathbf{y}_2}$ is simply a square sub-matrix of $\mathbf{\Phi}_{\mathbf{y}}$, consisting of its first M_2 rows and columns.

Now, at any given iteration, n, $n \geq 1$, $n \in \mathbb{Z}$, and starting with $n = 1$, we have

$$\begin{aligned}&\min_{\mathbf{h_1}^{(n)}}\phi_Z[\mathbf{h}_1^{(n)}|\mathbf{h}_2^{(n-1)}]\ \text{s.t.}\ [\mathbf{h}_1^{(n)}]^H\mathbf{d}_{1,\theta_\mathrm{d}} = 1,\\ &\mathbf{h}_1^{(n)} = \frac{[\mathbf{\Phi}_{\mathbf{y},2}^{(n-1)}]^{-1}\mathbf{d}_{1,\theta_\mathrm{d}}}{\mathbf{d}_{1,\theta_\mathrm{d}}^H[\mathbf{\Phi}_{\mathbf{y},2}^{(n-1)}]^{-1}\mathbf{d}_{1,\theta_\mathrm{d}}},\end{aligned} \tag{32}$$

$$\begin{aligned}&\min_{\mathbf{h}_2^{(n)}}\phi_Z[\mathbf{h}_2^{(n)}|\mathbf{h}_1^{(n)}]\ \text{s.t.}\ [\mathbf{h}_2^{(n)}]^H\mathbf{d}_{2,\theta_\mathrm{d}} = 1,\\ &\mathbf{h}_2^{(n)} = \frac{[\mathbf{\Phi}_{\mathbf{y},1}^{(n)}]^{-1}\mathbf{d}_{2,\theta_\mathrm{d}}}{\mathbf{d}_{2,\theta_\mathrm{d}}^H[\mathbf{\Phi}_{\mathbf{y},1}^{(n)}]^{-1}\mathbf{d}_{2,\theta_\mathrm{d}}}.\end{aligned} \tag{33}$$

The full-length KP-MVDR beamformer, after $n = N$ iterations, is given by,

$$\mathbf{h} = \mathbf{h}_1^{(N)}\otimes\mathbf{h}_2^{(N)}. \tag{34}$$

We must note that, in this subsection, we have only discussed the KP-MVDR beamformer based on minimizing ϕ_Z. The subject matter is equally applicable for deriving the KP-MVDR beamformer based on minimizing ϕ_{V_rn}, and the various mathematical expressions would simply require replacing the covariance matrices corresponding to the data ($\mathbf{y}$) to those corresponding to the disturbances ($\mathbf{v}$).

3.2. Hybrid Beamformers

The KP-MVDR beamformer, discussed in the preceding sub-section, only optimizes a single parameter, ϕ_Z, and thereby implements the same filter on both the VULAs. However, it is possible to optimize two different parameters, simultaneously, by implementing a different filter on each of the VULAs. Again, in the case of the KP-MVDR beamformer, the M-dimensional linear system was reduced to two subsystems (of sizes M_1 and M_2, respectively). By considering one of the filters as a fixed beamformer, the adaptive beamforming problem is reduced to a single lower dimensional (M_1 or M_2) linear system. With this viewpoint, we combine the DS beamformer with the MVDR beamformer, to obtain two different hybrid beamformers.

As is evident from (26), the narrowband WNG of the ULA can be factorized into the individual narrowband WNGs of the VULAs. Maximizing WNG_2, similar to (19), we obtain

$$\mathbf{h}_2 = \frac{\mathbf{d}_{2,\theta_\mathrm{d}}}{M_2}. \tag{35}$$

We may, now, minimize the variance of the final output under the distortionless criteria, using the first VULA:

$$\mathbf{h}_1 = \frac{\mathbf{\Phi}_{\mathbf{y},2}^{-1}\mathbf{d}_{1,\theta_\mathrm{d}}}{\mathbf{d}_{1,\theta_\mathrm{d}}^H\mathbf{\Phi}_{\mathbf{y},2}^{-1}\mathbf{d}_{1,\theta_\mathrm{d}}}. \tag{36}$$

The resulting filter, $\mathbf{h} = \mathbf{h}_1 \otimes \mathbf{h}_2$, may be termed as the Kronecker-product-minimum-variance-distortionless-response-delay-and-sum (KP-MVDR-DS) beamformer.

Contrary to the KP-MVDR-DS beamformer, we could implement the MVDR beamformer on the second VULA, after implementing the DS beamformer on the first VULA. This results in the Kronecker-product-delay-and-sum-minimum-variance-distortionless-response (KP-DS-MVDR) beamformer:

$$\mathbf{h}_1 = \frac{\mathbf{d}_{1,\theta_\mathrm{d}}}{M_1},\ \mathbf{h}_2 = \frac{\mathbf{\Phi}_{\mathbf{y},1}^{-1}\mathbf{d}_{2,\theta_\mathrm{d}}}{\mathbf{d}_{2,\theta_\mathrm{d}}^H\mathbf{\Phi}_{\mathbf{y},1}^{-1}\mathbf{d}_{2,\theta_\mathrm{d}}},\ \mathbf{h} = \mathbf{h}_1 \otimes \mathbf{h}_2. \quad (37)$$

Finally, one must note that only the MVDR beamformer based on minimizing ϕ_Z has been implemented above. However, the KP-MVDR-DS and KP-DS-MVDR beamformers could also be obtained by utilizing the MVDR beamformer based on minimizing ϕ_{V_rn}. The expressions in (36) and (37) would simply require replacing the covariance matrices, $\mathbf{\Phi}_{\mathbf{y},1}$ and $\mathbf{\Phi}_{\mathbf{y},2}$, to $\mathbf{\Phi}_{\mathbf{v},1}$ and $\mathbf{\Phi}_{\mathbf{v},2}$, respectively.

4. Experimental Performance

In this section, we evaluate the performances of the proposed beamformers in comparison with the conventional MVDR and DS beamformers. It has been observed that the KP-MVDR beamformer saturates rapidly with respect to the number of iterations, N. Hence, $N = 5$ will be used in this work for the KP-MVDR beamformer. For our experiments, we utilize a speech signal as the SOI, and four noise signals, taken from the NOISEX-92 database [34], as interferences (of different variances corresponding to different iSIRs). Three types of noise - white (flat-band), babble (low-frequency dominant), and hfchannel (high-frequency dominant) - are considered as interferences. The DOAs of the signals are presented in Table 1. Further, all the sensors are corrupted with zero mean IID Gaussian white noise signals, with the iSNR fixed at 10 dB. The experiments are first performed for stationary synthetic speech, and then extended to natural speech.

In this work, we consider $M = 2^6 = 64$ sensors in the ULA. The two VULAs are composed of M_1 and M_2 sensors, respectively, such that:

$$M_1 = 2^l,\ M_2 = 2^{\log_2(M)-l},\ \forall\ l = 1, 2, ..., \{\log_2(M) - 1\}. \quad (38)$$

The ULA is constructed by considering the inter-sensor distance as $\delta = 1$ cm. The data is sampled at a sampling-rate of $F_s = 1/T_s = 8$ kHz, and the speed of sound is considered to be $c = 340$ ms^{-1}. The data is processed in snapshots or frames of 10 ms (80 samples), with a frameshift of 5 ms, i.e., there is 50 % overlap between any two consecutive snapshots.

Table 1: **DOAs of the SOI and Interferences.**

Signal	$x(t)$	$u_1(t)$	$u_2(t)$	$u_3(t)$	$u_4(t)$
DOA	70^o	20^o	160^o	30^o	140^o

Table 2: **Parameters of the synthetic speech SOI.**

F_0 (Hz)	F_1/B_1 (Hz)	F_2/B_2 (Hz)	F_3/B_3 (Hz)	F_4/B_4 (Hz)
150	500 / 100	1500 / 200	2500 / 300	3500 / 400

4.1. Synthetic Speech as the SOI

We consider the SOI as a speech signal synthetically generated using the source-filter theory [46–49]. Voiced speech (vowel-like sounds) is produced by the quasi-periodic vibration of the vocal cords, which excites a cascade of resonators representing the cavities of the vocal tract. The pitch frequency of adult males is around 100 Hz, whereas that of females is around 200 Hz. Again, in general, for every 1000 Hz of the frequency spectrum of any natural speech signal, a resonance peak is observed. Henceforth, in our study, a neutral gender voiced speech signal of 5 s duration is synthesized, at a sampling frequency of $F_s = 8$ kHz. The pitch frequency (F_0), resonant frequencies, $\{F_k\ :\ k = 1, 2, 3, 4\}$, and resonant bandwidths, $\{B_k : k = 1, 2, 3, 4\}$, are listed in Table 2.

In this controlled experiment, since all the signals are stationary, we assume that (3) holds perfectly true. Thus, the SOI and interferences are required to be generated only for the first sensor. The sensor-noises, of course, need to be generated for all the sensors. The current statistics of the data, at the r^{th} frame or snapshot, is given by

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(f, r) &= \frac{1}{r}\sum_{k=1}^{r}\mathbf{y}(f, k)\mathbf{y}^H(f, k)\\ &= (1 - \frac{1}{r})\mathbf{\Phi}_{\mathbf{y}}(f, r-1) + \frac{1}{r}\mathbf{y}(f, r)\mathbf{y}^H(f, r).\end{aligned} \quad (39)$$

The true statistics of the SOI and disturbances are estimated apriori, using all the snapshots:

$$\begin{aligned}\phi_X(f) &= \frac{1}{S}\sum_{k=1}^{S}|X(f, k)|^2,\\ \mathbf{\Phi}_{\mathbf{v}}(f) &= \frac{1}{S}\sum_{k=1}^{S}\mathbf{v}(f, k)\mathbf{v}^H(f, k).\end{aligned} \quad (40)$$

In (40), S denotes the total number of snapshots. The beamformers are obtained using the current statistics. The statistical performance metrics are then evaluated for the beamformers using the true statistics.

4.1.1. Optimal sizes of the VULAs

Figure 3 plots the performances of the Kronecker product beamformers as the size of the second VULA is varied.

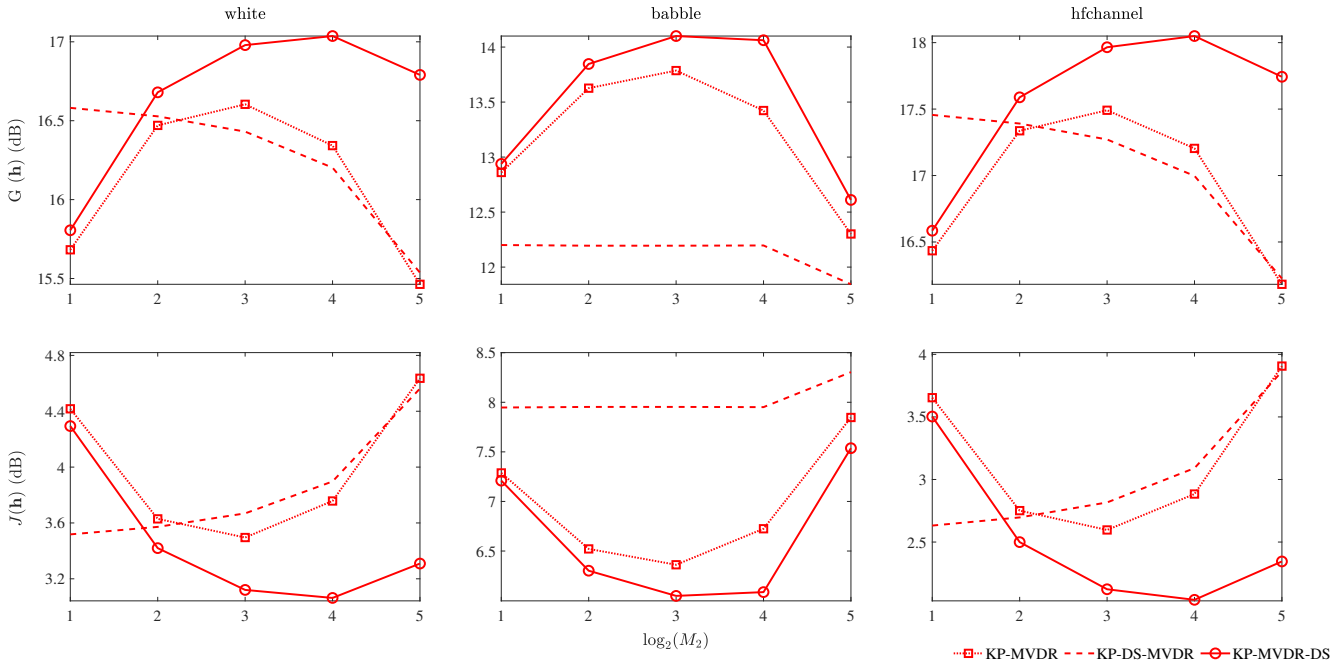

Figure 3: **Each column plots the broadband Gain (top), and broadband MSE (bottom), for varying M_2, at $r = 100$ snapshots - for a particular type (white/babble/hfchannel) of interference. Each plot has three curves, corresponding to the beamformers : KP-MVDR, KP-DS-MVDR, and KP-MVDR-DS. iSNR = 10 dB, and iSIR = 0 dB.**

As is evident, irrespective of the type of interference, the KP-MVDR and KP-MVDR-DS beamformers provide their best performances for $M_2 \approx 2^3 = 8$. On the contrary, the KP-DS-MVDR beamformer requires $M_2 = 2^1 = 2$ to obtain its best performances, which are inferior to the performances of the other two Kronecker product beamformers. However, its performances for $M_2 \lesssim \sqrt{M}$ are quite similar. Hence, from here on, we will use $M_1 = M_2 = \sqrt{M} = 8$ for all the three proposed beamformers. Under this condition, both the VULAs are provided with the same number of sensors, and thus given equal representation and importance in the final beamformer. Also, under this condition, the total number of coefficients ($= 2\sqrt{M}$) required by the Kronecker product beamformers is the least, i.e., the original M-dimensional linear system is represented by the smallest possible subsystems.

4.1.2. Robustness at varying iSIRs

Figure 4 plots the performances of the conventional and the Kronecker product beamformers as the strength of the interferences is varied, for each of the three types of interference. As is evident, the Kronecker product beamformers outperform the MVDR beamformer, under both high and low levels of interference. This demonstrates the robustness of the Kronecker product beamformers. One may also notice that the performance-gap between the DS beamformer and the other beamformers is much higher at low iSIRs, and diminishes as the iSIR increases. In fact, the DS beamformer performs better than all the other beamformers at low levels of interference (high iSIR). This demonstrates the limitations of fixed beamforming under high interference and/or noise conditions. The KP-DS-MVDR beamformer seems to closely match the performances of the DS beamformer, which indicates that the DS beamformer in the first VULA is dominant. Again, we note that the performances in the figure correspond to the 100^{th} snapshot of data. If we consider a higher snapshot (r), the performances of the MVDR beamformer will be similar to that of the KP-MVDR and KP-MVDR-DS beamformers, and superior to that of the DS beamformer, particularly at high levels of interference. This indicates that for the MVDR beamformer, the current statistical estimates must be very reliable, which is further discussed in the following sub-subsection.

4.1.3. Convergence and data requirements

As is evident from (39), as the number of snapshots increases (more data) the estimated current statistics becomes more refined. Figure 5 plots the performances of the MVDR and the Kronecker product beamformers as the number of snapshots is increased. As the DS beamformer does not rely on the statistics of the data, it is not included in the figure. The Kronecker product beamformers clearly outperform the MVDR beamformer in achieveing their optimum performances. Irrespective of the type of interference, the Kronecker product beamformers require around 50 snapshots to achieve saturation, whereas the MVDR requires over 100 snapshots. This is an indication that reducing the dimension of the linear system reduces the requirement of data. The fact that the Kronecker product beamformers are able to achieve similar steady-state performances as that of the MVDR is an additional benefit - it

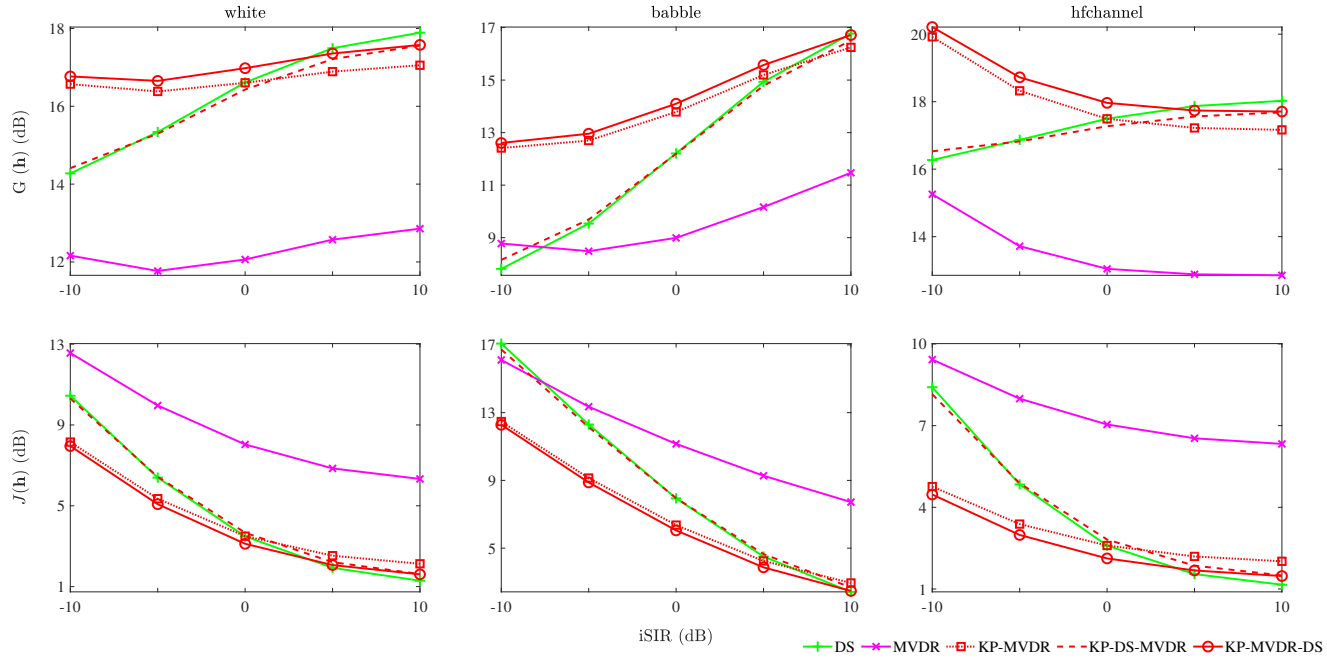

Figure 4: **Each column plots the broadband Gain (top), and broadband MSE (bottom), for varying iSIR (dB), at $r = 100$ snapshots - for a particular type (white/babble/hfchannel) of interference. Each plot has five curves, corresponding to the beamformers : DS, MVDR, KP-MVDR, KP-DS-MVDR, and KP-MVDR-DS. iSNR = 10 dB.**

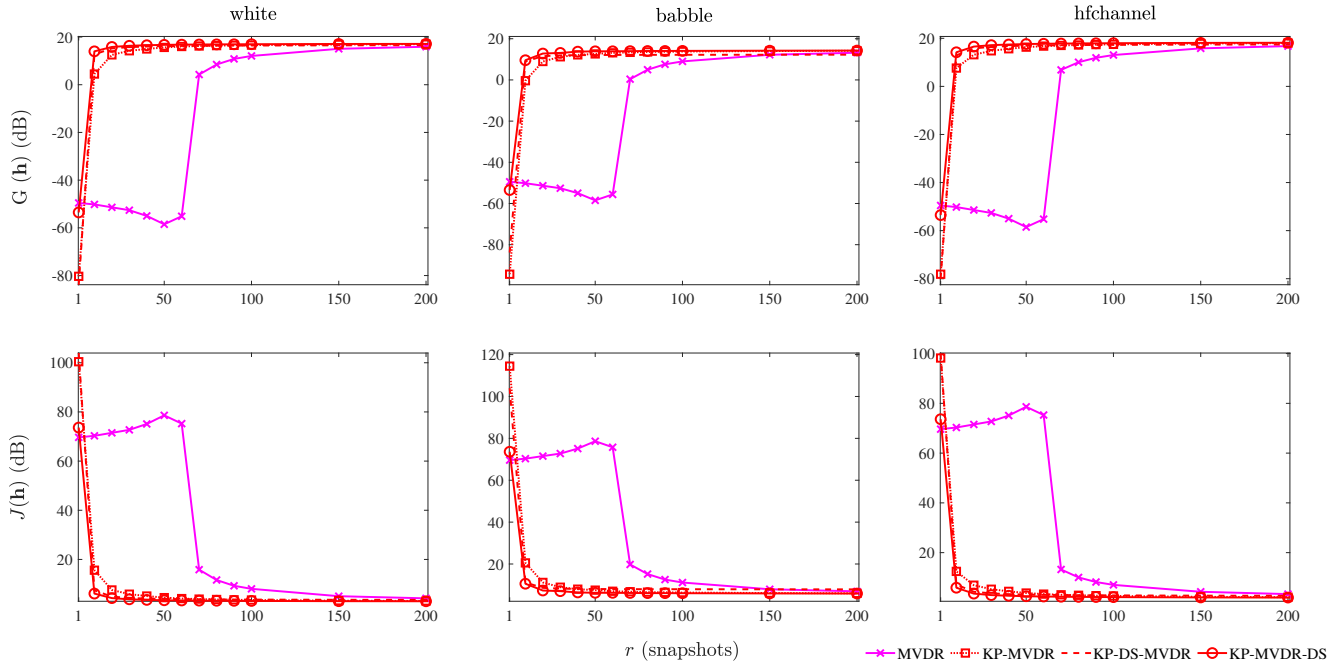

Figure 5: **Each column plots the broadband Gain (top), and broadband MSE (bottom), for varying number of snapshots (r) - for a particular type (white/babble/hfchannel) of interference. Each plot has four curves, corresponding to the beamformers : MVDR, KP-MVDR, KP-DS-MVDR, and KP-MVDR-DS. iSNR = 10 dB. iSNR = 10 dB, and iSIR = 0 dB.**

implies that segregating the original ULA into two smaller VULAs is useful ! Please note that, as mentioned in the introduction, we have not utilized any adaptive filtering algorithms to implement the beamformers, and hence must not compare these plots with the convergence of adaptive filtering algorithms [6].

4.2. Natural Speech as the SOI

Having analyzed the performances of the beamformers under controlled stationary conditions, we now apply the beamformers on non-stationary conditions, using a natural speech signal as the SOI. For this purpose, a speech signal of $\sim$ 3 s duration is arbitrarily chosen from the TIMIT [50] corpus. Using appropriate upsampling and

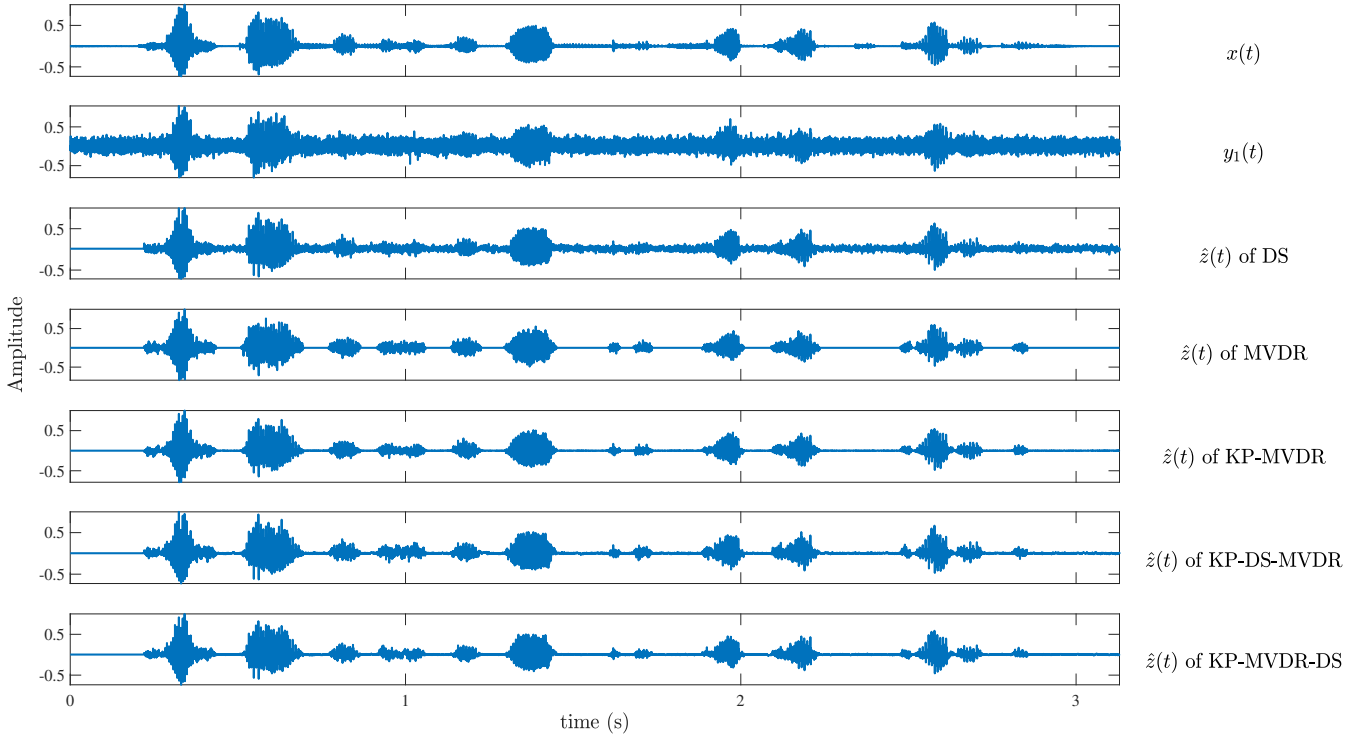

Figure 6: **Plots of (from top to bottom) the SOI, the corrupted signal (white noise interferences) at the first sensor ; the normalized outputs of the beamformers : DS, MVDR, KP-MVDR, KP-DS-MVDR, and KP-MVDR-DS. iSNR = 10 dB, and iSIR = 0 dB.**

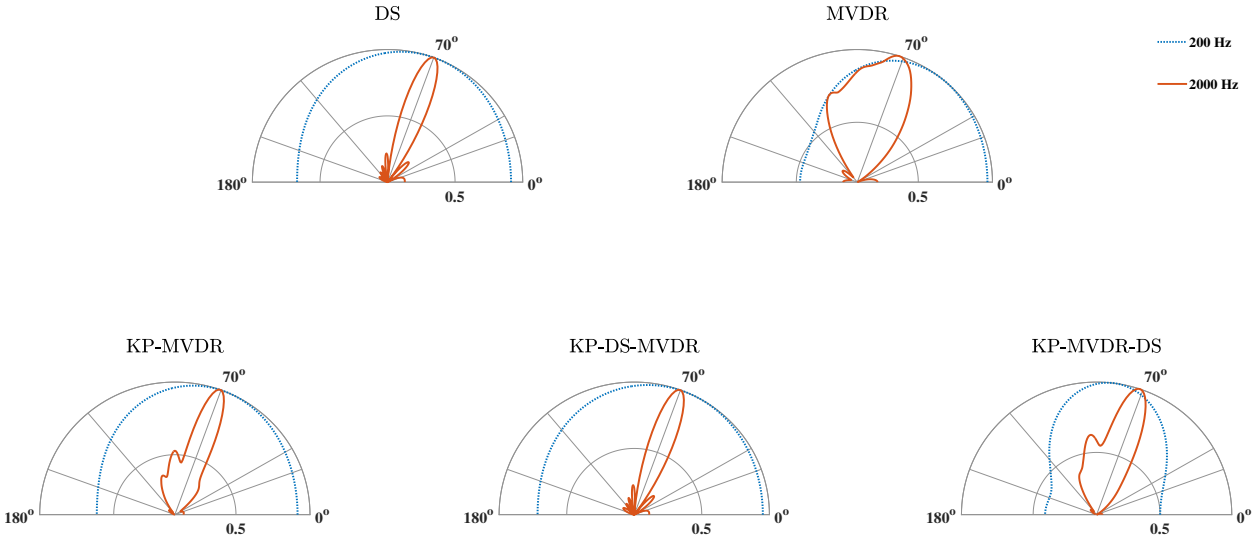

Figure 7: **Plots of $\beta_\theta[\mathbf{h}(f)]$, for $f/T_s = [200, 2000]$ Hz. iSNR = 10 dB, and iSIR = 0 dB (white noise).**

time-delays (based on the DOAs), the SOI and the interferences are generated for all the M sensors. Sensor-noises, as usual, are generated for all the sensors. Obviously, in this scenario, (3) will be only approximately true, particularly for the SOI which is non-stationary.

Under these conditions, there is no way to obtain the true statistics of the SOI, as the characteristics of the speech signal is changing dynamically from one snapshot to another. As the SOI is non-stationary, the overall data is also non-stationary, and reliable statistics cannot be estimated. However, as the disturbances are stationary, it is possible to estimate their statistics. One simple method is to identify the snapshots of the data where there is no speech, and utilize those snapshots for estimating the interference-noise statistics [3]. Discussion of sophisticated methods of real-time voice-activity-detection (VAD) from corrupted data is beyond the scope of this work. In this work, we will simply utilize energy based VAD [46–48] to apriori determine the non-speech snapshots. The

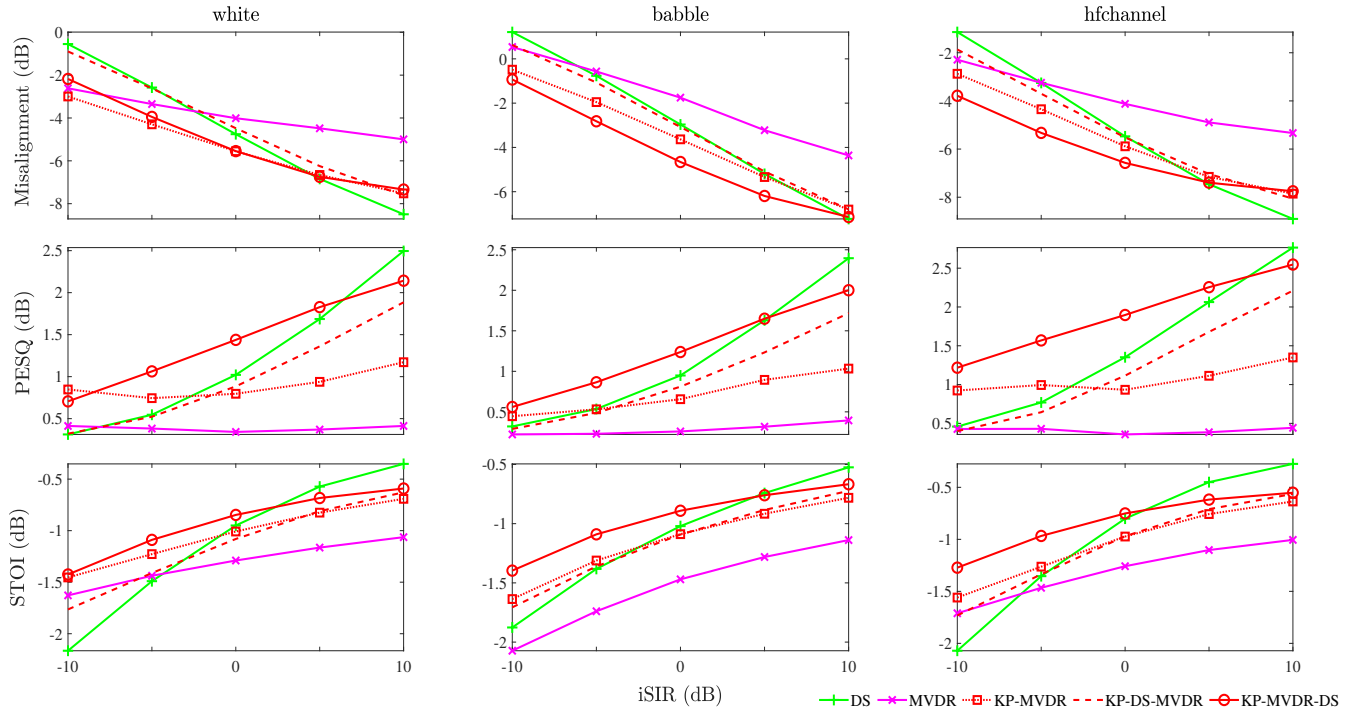

Figure 8: **Each column plots the performance metrics - Misalignment, PESQ, and STOI - for a particular type (white/babble/hfchannel) of interference. The metrics are plotted with respect to varying iSIR. Each plot has five curves, corresponding to the beamformers : DS, MVDR, KP-MVDR, KP-DS-MVDR, and KP-MVDR-DS. iSNR = 10 dB.**

interference-noise statistics is evaluated as

$$\mathbf{\Phi_v}(f,r) = \begin{cases} \mathbf{y}(f,r)[\mathbf{y}(f,r)]^H, \ r=1 \\ 0.8\mathbf{\Phi_v}(f,r-1) + 0.2\mathbf{y}(f,r)[\mathbf{y}(f,r)]^H, \ \text{no speech} \\ \mathbf{\Phi_v}(f,r-1), \ \text{contains speech} \end{cases} \tag{41}$$

The MVDR and the Kronecker product beamformers now minimize $\phi_{V_{\text{rn}}}$, instead of ϕ_Z, unlike in the case of synthetic speech. As the true statistics of the SOI is not available, the statistical performance metrics cannot be computed. Instead, we observe the final output, $z(t)$, of each of the beamformers, and their corresponding beampatterns. Figure 6 plots the outputs for the five beamformers, where the interferences are considered as white noise with iSIR = 0 dB. The beginning samples of the outputs are marked by extreme fluctuations. Hence, for better visualization, the samples till the first speech sample, r_s, have been assigned to the median value of the signal ($\mu = \text{median}[z(t)]$), and thereafter the signal is renormalized:

$$\hat{z}(t) = \begin{cases} \mu, \ t < r_s \\ z(t), \ t \geq r_s, \end{cases}, \ \hat{z}(t) \in [-1,1]. \tag{42}$$

As is evident from the figure, the DS beamformer is not very effective in mitigating disturbances. The MVDR and the Kronecker product beamformers are able to eradicate the interferences and sensor noises far more effectively. Figure 7 plots the beampatterns (averaged over all the snapshots) of the five beamformers at two different frequencies - 200 Hz and 2000 Hz. As is evident, all the beamformers have very good directivity at the higher frequency - the beam focus is narrow in the direction of the SOI. However, at the lower frequency, the beampatterns are not as good. Among the five beamformers, the KP-MVDR-DS seem to exhibit the best beampatterns, considering both the frequencies.

While the mitigation of interferences and noise is an important characteristic of a beamformer, it may also lead to loss in the waveform shape of the speech SOI, and hence in its intelligibility. Henceforth, we evaluate the Misalignment, Perceptual Evaluation of Speech Quality (PESQ) [51], and Short Time Objective Intelligibility (STOI) [52] metrics, for the output signal, in comparison with the SOI. The Misalignment is calculated as

$$\text{Misalignment} = \frac{\text{var}[e(t)]}{\text{var}[x(t)]}, \ e(t) = x(t) - z(t), \tag{43}$$

where var[.] implies the variance operator. At this juncture, we must note that STOI and PESQ are metrics used in speech enhancement. The domain of beamforming is closely related to multi-channel speech enhancement, but not the same [3]. Henceforth, we employ the PESQ and STOI metrics in our work, but for the performance evaluation of beamformers only.

Figure 8 plots the three metrics, averaged for 20 arbitrary speech SOIs taken from the TIMIT corpus. The dataset for each of the SOIs are created by corrupting

them with interferences (white/babble/hfchannel) at varying iSIRs, as we have discussed earlier. As can be observed, the KP-MVDR and KP-MVDR-DS beamformers have lower Misalignment compared to the other three beamformers, particularly at high levels of interference (low iSIRs). Similarly, in terms of PESQ, at high levels of interference, the KP-MVDR and KP-MVDR-DS beamformers provide the best performances. However, as the iSIR level increases, the DS beamformer provides the lowest Misalignment and the highest PESQ. This shows that while the DS beamformer is ineffective in mitigating disturbances, it does not negativily effect the waveform shape of the SOI. The same observations are supported by the STOI plots in the figure. At this point, one must note that the MVDR performs the worst among all the beamformers. This is because of its strong dependence on the accuracy of the interference-noise statistics. As such, if sophisticated techniques of estimating the statistics are employed, better performances may be expected from the MVDR and the Kronecker product beamformers.

5. Conclusions

We have introduced a new approach to frequency domain adaptive beamforming for large sensor arrays, with the purpose of achieving enhanced robustness to interference and statistical instability. Firstly, the original ULA is represented by two smaller VULAs, which are connected by the Kronecker product. As the VULAs are smaller than the original ULA, adaptive beamformers can be derived from them using lesser data for statistical computations. Their smaller size also makes them robust to errors in the estimated statistics associated with the much larger original ULA. Furthermore, the partitioning of the original ULA into VULAs allows the implementation of fixed and adaptive beamforming, simultaneously, incorporating the benefits of both. This leads to hybrid beamformers.

In this work, we have illustrated the utility of our proposed framework using the MVDR and DS beamformers as examples. Needless to say, the proposed methodology could be utilized for any adaptive and fixed beamformers. The choice of beamformers depends on the application at hand, and the characteristics of the signal and the interferences. Moving forward, we may investigate how to utilize the proposed methodology if the number of sensors is a prime number. Also, the interferences considered in this work are stationary, and hence experimenting with non-stationary interferences, such as background speech, may be considered. Lastly, one may also note that the proposed filters, like the conventional filters, are still quite sensitive to the DOA of the SOI. A methodology to diminish this dependency could be also explored in the future.

Acknowledgment

The authors thank Dr. Gongping Huang for his valuable inputs and suggestions.

References

[1] B. D. Van Veen, K. M. Buckley, Beamforming: A versatile approach to spatial filtering, IEEE assp magazine 5 (2) (1988) 4–24.

[2] L. Wang, Array signal processing algorithms for beamforming and direction finding, Ph.D. thesis, University of York (2009).

[3] J. Benesty, I. Cohen, J. Chen, Fundamentals of Signal Enhancement and Array Signal Processing, John Wiley & Sons, 2017.

[4] J. Benesty, Y. Huang, Adaptive signal processing: applications to real-world problems, Springer Science & Business Media, 2013.

[5] S. Chandran, Adaptive antenna arrays: trends and applications, Springer Science & Business Media, 2013.

[6] I. S. Reed, J. D. Mallett, L. E. Brennan, Rapid convergence rate in adaptive arrays, IEEE Transactions on Aerospace and Electronic Systems (6) (1974) 853–863.

[7] B. M. Asl, A. Mahloojifar, A low-complexity adaptive beamformer for ultrasound imaging using structured covariance matrix, IEEE transactions on ultrasonics, ferroelectrics, and frequency control 59 (4).

[8] Z. D. Zaharis, T. V. Yioultsis, A novel adaptive beamforming technique applied on linear antenna arrays using adaptive mutated boolean pso, Progress In Electromagnetics Research 117 (2011) 165–179.

[9] B. M. Asl, A. Mahloojifar, Eigenspace-based minimum variance beamforming applied to medical ultrasound imaging, IEEE transactions on ultrasonics, ferroelectrics, and frequency control 57 (11).

[10] Z. Zhang, W. Liu, W. Leng, A. Wang, H. Shi, Interference-plus-noise covariance matrix reconstruction via spatial power spectrum sampling for robust adaptive beamforming, IEEE Signal Processing Letters 23 (1) (2016) 121–125.

[11] Y. Feng, G. Liao, J. Xu, S. Zhu, C. Zeng, Robust adaptive beamforming against large steering vector mismatch using multiple uncertainty sets, Signal Processing.

[12] L. Landau, R. C. de Lamare, M. Haardt, Robust adaptive beamforming algorithms using the constrained constant modulus criterion, IET Signal Processing 8 (5) (2014) 447–457.

[13] W. Jia, W. Jin, S. Zhou, M. Yao, Robust adaptive beamforming based on a new steering vector estimation algorithm, Signal Processing 93 (9) (2013) 2539–2542.

[14] Y. Gu, A. Leshem, Robust adaptive beamforming based on interference covariance matrix reconstruction and steering vector estimation, IEEE Transactions on Signal Processing 60 (7) (2012) 3881–3885.

[15] Y. Gu, N. A. Goodman, S. Hong, Y. Li, Robust adaptive beamforming based on interference covariance matrix sparse reconstruction, Signal Processing 96 (2014) 375–381.

[16] A. Khabbazibasmenj, S. A. Vorobyov, A. Hassanien, Robust adaptive beamforming based on steering vector estimation with as little as possible prior information, Trans. Sig. Proc. 60 (6) (2012) 2974–2987. doi:10.1109/TSP.2012.2189389.
URL https://doi.org/10.1109/TSP.2012.2189389

[17] J. Yang, G. Liao, J. Li, Y. Lei, X. Wang, Robust beamforming with imprecise array geometry using steering vector estimation and interference covariance matrix reconstruction, Multidimensional Syst. Signal Process. 28 (2) (2017) 451–469. doi:10.1007/s11045-015-0350-7.
URL https://doi.org/10.1007/s11045-015-0350-7

[18] Y. Ke, C. Zheng, R. Peng, X. Li, Robust adaptive beamforming using noise reduction preprocessing-based fully automatic diagonal loading and steering vector estimation 5 12974–12987. doi:10.1109/ACCESS.2017.2725450.

[19] X. Yuan, L. Gan, Robust adaptive beamforming via a novel subspace method for interference covariance matrix reconstruction, Signal Processing 130 (2017) 233–242.

[20] L. Huang, J. Zhang, X. Xu, Z. Ye, Robust adaptive beamforming with a novel interference-plus-noise covariance matrix reconstruction method., IEEE Trans. Signal Processing 63 (7) (2015) 1643–1650.

[21] B. Liao, C. Guo, L. Huang, Q. Li, H. C. So, Robust adaptive beamforming with precise main beam control, IEEE Transactions on Aerospace and Electronic Systems 53 (1) (2017) 345–356.
[22] E. Weinstein, K. Steele, A. Agarwal, J. Glass, Loud: A 1020 node microphone array and acoustic beamformer, Tech. rep., Courant Institute of Mathematical Sciences New York United States (2007).
[23] M. L. Honig, J. S. Goldstein, Adaptive reduced-rank interference suppression based on the multistage wiener filter, IEEE Transactions on Communications 50 (6) (2002) 986–994.
[24] S. Burykh, K. Abed-Meraim, Reduced-rank adaptive filtering using krylov subspace, EURASIP Journal on Applied Signal Processing 2002 (1) (2002) 1387–1400.
[25] E. L. Santos, M. D. Zoltowski, On low rank mvdr beamforming using the conjugate gradient algorithm, in: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, 2004, pp. ii–173.
[26] R. C. De Lamare, R. Sampaio-Neto, Reduced-rank adaptive filtering based on joint iterative optimization of adaptive filters, IEEE Signal Processing Letters 14 (12) (2007) 980–983.
[27] R. C. de Lamare, L. Wang, R. Fa, Adaptive reduced-rank lcmv beamforming algorithms based on joint iterative optimization of filters: Design and analysis, Signal Processing 90 (2) (2010) 640–652.
[28] P. Chevalier, A. Blin, Widely linear mvdr beamformers for the reception of an unknown signal corrupted by noncircular interferences, IEEE Transactions on Signal Processing 55 (11) (2007) 5323–5336.
[29] P. Chevalier, J.-P. Delmas, A. Oukaci, Optimal widely linear mvdr beamforming for noncircular signals, in: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, IEEE, 2009, pp. 3573–3576.
[30] N. Song, W. U. Alokozai, R. C. de Lamare, M. Haardt, Adaptive widely linear reduced-rank beamforming based on joint iterative optimization, IEEE Signal Processing Letters 21 (3) (2014) 265–269.
[31] C. Zheng, A. Deleforge, X. Li, W. Kellermann, Statistical analysis of the multichannel wiener filter using a bivariate normal distribution for sample covariance matrices, IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (5) (2018) 951–966.
[32] C. F. Van Loan, The ubiquitous kronecker product, Journal of computational and applied mathematics 123 (1-2) (2000) 85–100.
[33] K. Schäcke, On the kronecker product.
[34] A. Varga, H. J. Steeneken, Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems, Speech communication 12 (3) (1993) 247–251.
[35] I. Cohen, J. Benesty, J. Chen, Differential kronecker product beamforming, IEEE/ACM Transactions on Audio, Speech, and Language Processing.
[36] W. Yang, G. Huang, J. Benesty, I. Cohen, J. Chen, On the design of flexible kronecker product beamformers withlinear microphone arrays, in: IEEE ICASSP, 2019.
[37] J. Benesty, I. Cohen, J. Chen, Fundamentals of Signal Enhancement and Array Signal Processing, Wiley Online Library, 2018.
[38] J. Benesty, J. Chen, Y. Huang, Microphone array signal processing, Vol. 1, Springer Science & Business Media, 2008.
[39] Ming Zhang, M. H. Er, Adaptive beamforming by microphone arrays, in: Proceedings of GLOBECOM '95, Vol. 1, 1995, pp. 163–167 vol.1. doi:10.1109/GLOCOM.1995.500344.
[40] M. Brandstein, D. Ward, Microphone arrays: signal processing techniques and applications, Springer Science & Business Media, 2013.
[41] J. G. Ryan, R. A. Goubran, Near-field beamforming for microphone arrays, in: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, IEEE, 1997, pp. 363–366.
[42] P. Thomas, R. Verburgh, M. Catrysse, D. Botteldooren, Design of a microphone array for near-field conferencing applications, in: Proceedings of Meetings on Acoustics 173EAA, Vol. 30, ASA, 2017, p. 055001.
[43] K. Kumatani, J. McDonough, B. Raj, Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors, IEEE Signal Processing Magazine 29 (6) (2012) 127–140.
[44] M. J. Taghizadeh, P. N. Garner, H. Bourlard, Microphone array beampattern characterization for hands-free speech applications, in: 2012 IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM), IEEE, 2012, pp. 465–468.
[45] A. AlShehhi, M. L. Hammadih, M. S. Zitouni, S. AlKindi, N. Ali, L. Weruaga, Linear and circular microphone array for remote surveillance: Simulated performance analysis, arXiv preprint arXiv:1703.02318.
[46] J. Benesty, M. M. Sondhi, Y. Huang, Springer handbook of speech processing, Springer Science & Business Media, 2008.
[47] L. R. Rabiner, R. W. Schafer, Digital processing of speech signals, Vol. 100, Prentice-hall Englewood Cliffs, 1978.
[48] L. R. Rabiner, R. W. Schafer, Introduction to digital speech processing, Foundations and trends in signal processing 1 (1) (2007) 1–194.
[49] R. Sharma, L. Vignolo, G. Schlotthauer, M. Colominas, H. L. Rufiner, S. Prasanna, Empirical mode decomposition for adaptive am-fm analysis of speech: A review, Speech Communication 88 (2017) 39 – 64. doi:http://dx.doi.org/10.1016/j.specom.2016.12.004.
URL //www.sciencedirect.com/science/article/pii/S0167639316302370
[50] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, DARPA TIMIT acoustic phonetic continuous speech corpus CDROM (1993).
URL http://www.ldc.upenn.edu/Catalog/LDC93S1.html
[51] A. W. Rix, J. G. Beerends, M. P. Hollier, A. P. Hekstra, Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs, in: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), Vol. 2, IEEE, 2001, pp. 749–752.
[52] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, A short-time objective intelligibility measure for time-frequency weighted noisy speech, in: 2010 IEEE international conference on acoustics, speech and signal processing, IEEE, 2010, pp. 4214–4217.