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

Rajib Sharmaa,∗, Israel Cohena,1, Jacob Benestyb

aAndrew and Erna Viterbi Faculty of Electrical Engineering, Technion - Israel Institute of Technology.
bINRS-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 [4]. 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 above-mentioned challenges at a higher level of abstraction, i.e.,
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. 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 interfering signals are broadband noise signals, on which 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 $\theta_d$, 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 $m$th sensor, corresponding to the $r$th snapshot, may be represented as

$$y_m(t, r) = x_m(t, r) + v_m(t, r), \quad m = 1, 2, ..., M,$$

$$v_m(t, r) = \sum_{k=1}^{K} u_{k,m}(t, r) + w_m(t, r),$$

$$x_m(t, r) = \sum_{k=1}^{K} \left( t - \frac{\delta_{m,k} \cos \theta_d}{cT_s}, r \right),$$

$$u_{k,m}(t, r) = \frac{\delta_{m,k} \cos \theta_k}{cT_s}.$$

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

Figure 1: An ULA of $M$ sensors, on which a discrete-time SOI, $x(t)$, impinges upon at DOA, $\theta_d$. The inter-sensor distance is denoted by $\delta$. The combined effect of interferences and sensor-noise at the $m$th sensor is denoted by $v_m(t)$, and $y_m(t)$ denotes the final signal detected at the sensor.
the ULA. As such, if the SOI and/or interferences are non-stationary\footnote{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 reliable estimates of their true values. The variable, $r$, then, may be dropped to represent the estimates of the true statistics.} 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

$$Y_m(f, r) = X_m(f, r) + V_m(f, r),$$

$$V_m(f, r) = \sum_{k=1}^{K} U_{k,m}(f, r) + W_m(f, r),$$

$$X_m(f, r) = \exp\left(-j2\pi f \frac{\delta_m \cos \theta_d}{c T_s}\right) X(f, r)$$

$$= d_{\theta_m}(f) X(f, r),$$

$$U_{k,m}(f, r) = \exp\left(-j2\pi f \frac{\delta_m \cos \theta_k}{c T_s}\right) U_k(f, r)$$

$$= d_{\theta_k}(m) U_k(f, r).$$

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

$$y(f, r) = [Y_1(f, r), ..., Y_M(f, r)]^T$$

$$= x(f, r) + w(f, r),$$

$$x(f, r) = d_{\theta_k}(f) X(f, r),$$

$$w(f, r) = \sum_{k=1}^{K} u_k(f, r) + w(f, r)$$

$$= \sum_{k=1}^{K} d_{\theta_k}(f) U_k(f, r) + w(f, r),$$

$$d_{\theta_k}(f) = [1 ... d_{\theta_k,m}(f) ... d_{\theta_k,M}(f)]^T,$$

$$d_{\theta_k}(f) = [1 ... d_{\theta_k,m}(f) ... d_{\theta_k,M}(f)]^T.$$

In \cite{[6]}, $y(f, r)$ represents the $M$-dimensional data sensed by the $M$-component sensor array. Similarly, $x(f, r)$, $w(f, r)$, $u_k(f, r)$, and $w(f, r)$ are $M$-dimensional vectors representing the respective components of the data across the sensor array. Again, $d_{\theta_k}(f)$ and $d_{\theta_k}(f)$ 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, $h(f, r)$, on the data-vector, $y(f, r)$, so as to obtain a signal, $Z(f, r) \approx X(f, r)$.

$$Z(f, r) = h^H(f, r)y(f, r) = X_{id}(f, r) + V_{rn}(f, r),$$

$$X_{id}(f, r) = h^H(f, r)d_{\theta_d}(f) X(f, r),$$

$$V_{rn}(f, r) = h^H(f, r)v(f, r).$$

In \cite{[1]}, $X_{id}(f, r)$ and $V_{rn}(f, r)$ represent the filtered-desired and the residual-noise signal, respectively. Ideally, one would want $V_{rn}(f, r) = 0$, while obtaining a distortion-less estimate, i.e., $h^H(f, r)d_{\theta_d}(f) = 1$. 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 \cite{[3][4]} to stitch together the samples of consecutive snapshots:

$$z(t, r) \leftrightarrow Z(f, r)$$

$$z(t) = \sum_r z(t, r) , \text{ using OLA. (5)}$$

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

$$\phi_Z(f, r) = E\{|Z(f, r)|^2\} = h^H(f, r)\Phi_y(f, r)h(f, r)$$

$$= \phi_{X_{id}(f, r)} + \phi_{V_{rn}(f, r)},$$

$$\phi_{X_{id}(f, r)} = E\{|X_{id}(f, r)|^2\}$$

$$= h^H(f, r)d_{\theta_d}(f)\phi_X(f, r)d_{\theta_d}(f)^H h(f, r),$$

$$\phi_{V_{rn}(f, r)} = E\{|V_{rn}(f, r)|^2\},$$

$$\phi_{X_{id}(f, r)} = E\{|V_{rn}(f, r)|^2\}$$

$$= h^H(f, r)\Phi_y(f, r)h(f, r).$$

In \cite{[6]}, $\phi_{X_{id}(f, r)}$, $\phi_{V_{rn}(f, r)}$, and $\phi_{X_{id}(f, r)}$ represent the variances of the filtered-desired signal, the residual-noise signal, and the SOI, respectively. Similarly, $\Phi_y(f, r)$ and $\Phi_y(f, r)$ represent the $M \times M$ covariance-matrices of the disturbances (interferences plus sensor-noises), and the data, respectively. They are obtained as

$$\Phi_{y}(f, r) = E\{|v(f, r)v^H(f, r)\}$$

$$= \sum_{k=1}^{K} \Phi_{u_k}(f, r) + \Phi_{w}(f, r),$$

$$\Phi_{u_k}(f, r) = E\{u_k(f, r)u_k^H(f, r)\}$$

$$= d_{\theta_k}(f)\Phi_{u}(f, r)d_{\theta_k}(f)^H,$$

$$\Phi_{w}(f, r) = E\{|w(f, r)w^H(f, r)\}$$

$$= d_{\theta_k}(f)\Phi_{w}(f, r)d_{\theta_k}(f)^H + \Phi_{y}(f, r).$$

If the signals are stationary\footnote{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 reliable estimates of their true values. The variable, $r$, then, may be dropped to represent the estimates of the true statistics.} 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.
2.1. 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,
\[
\mathcal{E}(f,r) = Z(f,r) - X(f,r) = \mathcal{E}_d(f,r) + \mathcal{E}_n(f,r),
\]
\[
\mathcal{E}_d(f,r) = |h^H(f,r)\mathbf{d}_{\Phi}(f) - 1|^2 X(f,r),
\]
\[
\mathcal{E}_n(f,r) = V_n(f,r) = h^H(f,r)\mathbf{v}(f,r),
\]
and
\[
J[h(f,r)] = E[|\mathcal{E}(f,r)|^2] = \phi_X(f) + h^H(f,r)\Phi_X(f)h(f,r) - \phi_X(f)h^H(f,r)\mathbf{d}_{\Phi}(f) - \phi_X(f)\mathbf{d}_{\mathbf{v}}(f)\mathbf{v}(f,r),
\]
\[
J[h(r)] = \sum_f J[h(f,r)] \Delta f.
\]

In [8] and [9], \(\mathcal{E}_d(f,r)\) and \(\mathcal{E}_n(f,r) = V_n(f,r)\) represent the desired-signal-distortion and the residual-noise, respectively, after filtering. \(J[h(f,r)]\) and \(J[h(r)]\) represent the narrowband and broadband MSE, respectively. \(\Delta f\) represents the frequency resolution of the STFT. The narrowband MSE may also be represented as
\[
J[h(f,r)] = J_d[h(f,r)] + J_n[h(f,r)],
\]
\[
J_d[h(f,r)] = E[|\mathcal{E}_d(f,r)|^2] = \phi_X(f)h^H(f,r)\mathbf{d}_{\Phi}(f) - |h^H(f,r)\mathbf{d}_{\Phi}(f) - 1|^2,
\]
\[
J_n[h(f,r)] = E[|\mathcal{E}_n(f,r)|^2] = h^H(f,r)\Phi_{\mathbf{v}}(f)\mathbf{v}(f,r).
\]

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
\[
\text{iSIR}(f) = \frac{\phi_X(f)}{\phi_{\mathbf{v}}(f)},
\]
\[
\text{iSNR}(f) = \frac{\phi_X(f)}{\phi_{W_1}(f)} = \frac{\phi_X(f)}{E[|W_1(f,r)|^2]},
\]
\[
\text{iSIR}(f) = \frac{\phi_X(f)}{\phi_{W_1}(f)} = \frac{\phi_X(f)}{E[|W_1(f,r)|^2]},
\]
\[
\text{iSNR}(f) = \frac{\phi_X(f)}{\phi_{W_1}(f)} = \frac{\phi_X(f)}{E[|W_1(f,r)|^2]}.
\]

The narrowband and broadband output signal-to-interference-plus-noise-ratios (oSINR) are given by
\[
o\text{SINR}[h(f,r)] = \frac{E[|X_{id}(f,r)|^2]}{E[|V_n(f,r)|^2]} = \frac{\phi_X(f)h^H(f,r)\mathbf{d}_{\Phi}(f)^2}{\mathbf{h}^H(f,r)\Phi_{\mathbf{v}}(f)\mathbf{h}(f,r)}.
\]
\[
o\text{SINR}[h(r)] = \frac{\sum_f \phi_X(f)h^H(f,r)\mathbf{d}_{\Phi}(f)^2 \Delta f}{\sum_f \mathbf{h}^H(f,r)\Phi_{\mathbf{v}}(f)\mathbf{h}(f,r) \Delta f}.
\]

The narrowband and broadband gains are obtained as
\[
G[h(f,r)] = \frac{o\text{SINR}[h(f,r)]}{i\text{SIR}(f)}, \quad G[h(r)] = \frac{o\text{SINR}[h(r)]}{i\text{SIR}(r)}.
\]

Apart from the aforementioned statistical measures, the beam pattern 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, \(\theta\). It is defined as
\[
\beta_{\theta}[h(f,r)] = \frac{|d_{\theta}^H(f,h(f,r))|}{\max_{\theta} |d_{\theta}^H(f,h(f,r))|},
\]
\[
\beta_{\theta}[h(r)] = \frac{\sum_{\theta} \beta_{\theta}[h(f,r)]}{\max_{\theta} \sum_{\theta} \beta_{\theta}[h(f,r)]}.
\]

2.2. DS Beamformer

Irrespective of whether the true statistics is available, the narrowband gain may be expressed (not evaluated) as
\[
G[h(f,r)] = \frac{|h^H(f,r)\mathbf{d}_{\Phi}(f)|^2}{\mathbf{h}^H(f,r)\Phi_{\mathbf{v}}(f)\mathbf{h}(f,r)},
\]
\[
\text{WNG}[h(f,r)] = \frac{|h^H(f,r)\mathbf{d}_{\Phi}(f)|^2}{\mathbf{h}^H(f,r)\mathbf{h}(f,r)}.
\]

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
\[
\min_{h(f,r)} \mathbf{h}^H(f,r)\mathbf{h}(f,r) \text{ s.t. } \mathbf{h}^H(f,r)\mathbf{d}_{\Phi}(f) = 1,
\]
\[
\mathbf{h}(f,r) = \frac{\mathbf{d}_{\Phi}(f)}{M} = \mathbf{h}(f).
\]
2.3. MVDR Beamformer

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

\[
\begin{align*}
\min_{\mathbf{h}(f, r)} & \; \phi_{\text{VUL}}(f, r) \text{ s.t. } \mathbf{h}^H(f, r) \mathbf{d}_{\delta_1}(f) = 1, \\
\mathbf{h}(f, r) & = \frac{\Phi_{\nu}^{-1}(f, r) \mathbf{d}_{\delta_1}(f)}{\mathbf{d}_{\delta_1}(f) \Phi_{\nu}^{-1}(f, r) \mathbf{d}_{\delta_1}(f)},
\end{align*}
\]

and

\[
\begin{align*}
\min_{\mathbf{h}(f, r)} & \; \phi_Z(f, r) \text{ s.t. } \mathbf{h}^H(f, r) \mathbf{d}_{\delta_2}(f) = 1, \\
\mathbf{h}(f, r) & = \frac{\Phi_{\nu}^{-1}(f, r) \mathbf{d}_{\delta_2}(f)}{\mathbf{d}_{\delta_2}(f) \Phi_{\nu}^{-1}(f, r) \mathbf{d}_{\delta_2}(f)}.
\end{align*}
\]

If the SOI, interferences and sensor-noises are all stationary, the variables with subscript \( \nu \) 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 \( \Phi_{\nu}(f, r) \) have to be derived, and (20) has to be implemented.

3. Kronecker Product Beamforming

If 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 comprised of \( M_1 \) sensors, and the other of \( M_2 \) sensors, such that \( M = M_1 M_2 \). The second VULA has the same inter-sensor distance, \( \delta_2 = \delta_1 \), 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

\[
\begin{align*}
\delta_m & = (m - 1)\delta, \; m = 1, 2, ..., M = M_1 M_2, \\
\delta_{1,p} & = (p - 1)\delta_1 = (p - 1)M_2\delta, \; p = 1, 2, ..., M_1, \\
\delta_{2,q} & = (q - 1)\delta_2 = (q - 1)\delta, \; q = 1, 2, ..., M_2, \\
\delta_m & = \delta_{(p-1)M_2+q} = \delta_{1,p} + \delta_{2,q}.
\end{align*}
\]

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

\[
\begin{align*}
\mathbf{d}_{\theta_1} & = \mathbf{d}_{1,\theta_1} \otimes \mathbf{d}_{2,\theta_1}, \\
d_{1,\theta_1,p} & = \exp \left( -j2\pi f \frac{\delta_{1,p} \cos \theta_1}{cT_s} \right), \; 1 \leq p \leq M_1, \\
d_{2,\theta_1,q} & = \exp \left( -j2\pi f \frac{\delta_{2,q} \cos \theta_1}{cT_s} \right), \; 1 \leq q \leq M_2.
\end{align*}
\]

In (23), the symbol \( \otimes \) represents the Kronecker product, \( \mathbf{d}_{1,\theta_1} \) and \( \mathbf{d}_{2,\theta_1} \) are steering vectors (of lengths \( M_1 \) and \( M_2 \) respectively) of the two VULAs, and \( \mathbf{d}_{\theta_1} \) is the steering vector (of length \( M = M_1 M_2 \)) of the ULA. Henceforth, \( d_{1,\theta_1,p} \) and \( d_{2,\theta_1,q} \) 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 \( M = M_1 M_2 \)). Assuming that two such virtual filters, \( \mathbf{h}_1 \) and \( \mathbf{h}_2 \), exist, we have

\[
\mathbf{h} = \mathbf{h}_1 \otimes \mathbf{h}_2 = (\mathbf{h}_1 \otimes \mathbf{I}_{M_2}) \mathbf{h}_2 = (\mathbf{I}_{M_1} \otimes \mathbf{h}_2) \mathbf{h}_1.
\]

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

![Figure 2: An ULA of \( M = M_1 M_2 \) sensors, and its two constituent VULAs.](image)
factorized as:

\[
\begin{align*}
    h^H d_{\phi} &= (h_1 \otimes h_2)^H (d_{1,\theta_i} \otimes d_{2,\theta_i}) \\
    &= (h_1^H d_{1,\theta_i})(h_2^H d_{2,\theta_i}) = 1,
\end{align*}
\]

WNG = \[\frac{|h_1^H d_{1,\theta_i}|^2}{h_1^H h_1} \times \frac{|h_2^H d_{2,\theta_i}|^2}{h_2^H h_2} = \text{WNG}_1 \times \text{WNG}_2.\]

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

\[
\begin{align*}
    \phi_{V,n} &= h_1^H \Phi_{\nu,2} h_1 = h_2^H \Phi_{\nu,1} h_2, \\
    \Phi_{\nu,2} &= (I_{M_1} \otimes h_2)^H \Phi_{\nu}(I_{M_1} \otimes h_2), \\
    \Phi_{\nu,1} &= (h_1 \otimes I_{M_2})^H \Phi_{\nu}(h_1 \otimes I_{M_2}). \\
    \phi_Z &= h_1^H \Phi_{y,2} h_1 = h_2^H \Phi_{y,1} h_2, \\
    \Phi_{y,2} &= (I_{M_1} \otimes h_2)^H \Phi_{y}(I_{M_1} \otimes h_2), \\
    \Phi_{y,1} &= (h_1 \otimes I_{M_2})^H \Phi_{y}(h_1 \otimes I_{M_2}).
\end{align*}
\]

### 3.1. Kronecker Product Adaptive Beamformer

Consider \( h_2 \) exists, and \( h_1^H d_{1,\theta_i} = 1 \). Then, as is evident from (25) and (28), the variance of the final output, and the distortionless constraint may be expressed as

\[
\phi_Z[h_1, h_2] = h_1^H \Phi_{y,2} h_1, \quad h_1^H d_{\phi} = h_1^H d_{1,\theta_i} = 1. \quad (29)
\]

Minimizing \( \phi_Z[h_1, h_2] \) with respect to \( h_1 \), subject to the distortionless criteria, results in the MVDR beamformer for the first VULA.

Similarly, if \( h_1 \) exists, and \( h_1^H d_{1,\theta_i} = 1 \), we have

\[
\phi_Z[h_2, h_1] = h_2^H \Phi_{y,1} h_2, \quad h_2^H d_{\phi} = h_2^H d_{2,\theta_i} = 1. \quad (30)
\]

Minimizing \( \phi_Z[h_2, h_1] \) with respect to \( h_2 \), subject to the distortionless 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

\[
\begin{align*}
    \Phi_{y,2} &= E[y_2 y_2^H], \quad y_2 = [Y_1, Y_2, ..., Y_{M_2}]^T, \\
    h_2^{(0)} &= \Phi_{y,2}^{-1} d_{2,\theta_i}, \quad [h_2^{(0)}]^H d_{2,\theta_i} = 1.
\end{align*}
\]

In the preceding equation-set, \( \Phi_{y,2} \) is simply a square sub-matrix of \( \Phi_{y} \), 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

\[
\begin{align*}
    \min_{h_1^{(n)}} \phi_Z[h_1^{(n)}, h_2^{(n-1)}] \text{ s.t. } [h_1^{(n)}]^H d_{1,\theta_i} &= 1, \\
    h_1^{(n)} &= \left[\Phi_{y,2}^{(n-1)}\right]^{-1} d_{1,\theta_i}, \\
    \min_{h_2^{(n)}} \phi_Z[h_2^{(n)}, h_1^{(n)}] \text{ s.t. } [h_2^{(n)}]^H d_{2,\theta_i} &= 1, \\
    h_2^{(n)} &= \left[\Phi_{y,1}^{(n)}\right]^{-1} d_{2,\theta_i}.
\end{align*}
\]

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

\[
\begin{align*}
    h &= h_1^{(N)} \otimes h_2^{(N)},
\end{align*}
\]

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

### 3.2. Hybrid Beamformers

The KP-MVDR beamformer, discussed in the preceding sub-section, only optimizes a single parameter, \( \phi_Z \), 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 \( (M_1 + M_2) \) linear system. With this viewpoint, we combine the DS beamformer with the MVDR beamformer, to obtain two different hybrid beamformers.

As is evident from (20), the narrowband WNG of the ULA can be factorized into the individual narrowband WNGs of the VULAs. Maximizing WNG2, similar to (19), we obtain

\[
\begin{align*}
    h_2 &= \frac{d_2,\theta_i}{M_2},
\end{align*}
\]

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

\[
\begin{align*}
    h_1 &= \frac{\Phi_{y,1}^{-1} d_{1,\theta_i}}{d_{1,\theta_i}^H \Phi_{y,1}^{-1} d_{1,\theta_i}}.
\end{align*}
\]
The resulting filter, \( h = h_1 \otimes h_2 \), 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.

\[
h_1 = \frac{d_{1,\theta_1}}{M_1}, \quad h_2 = \frac{d_{2,\theta_2}}{d_{2,\theta_1}} \Phi_{y,1}^H d_{2,\theta_1}, \quad h = h_1 \otimes h_2. \tag{37}
\]

Finally, one must note that only the MVDR beamformer based on minimizing \( \phi_2 \) 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 \( \phi_{v,a} \). The expressions in (36) and (37) would simply require replacing the covariance matrices, \( \Phi_{y,1} \) and \( \Phi_{y,2} \), to \( \Phi_{v,1} \) and \( \Phi_{v,2} \), 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 iSNRs). 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:

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

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>
<thead>
<tr>
<th>Signal</th>
<th>( x(t) )</th>
<th>( u_1(t) )</th>
<th>( u_2(t) )</th>
<th>( u_3(t) )</th>
<th>( u_4(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOA</td>
<td>70°</td>
<td>20°</td>
<td>160°</td>
<td>30°</td>
<td>140°</td>
</tr>
</tbody>
</table>

### 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 \( t^{th} \) frame or snapshot, is given by

\[
\Phi_y(f, r) = \frac{1}{r} \sum_{k=1}^{r} y(f, k)y^H(f, k) = (1 - \frac{1}{r})\Phi_y(f, r - 1) + \frac{1}{r}y(f, r)y^H(f, r). \tag{39}
\]

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

\[
\Phi_X(f) = \frac{1}{S} \sum_{k=1}^{S} |X(f, k)|^2, \tag{40}
\]

\[
\Phi_V(f) = \frac{1}{S} \sum_{k=1}^{S} v(f, k)v^H(f, k).
\]

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 100th 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 demonstrates 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 Figure 5, 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. The number of snapshots is increased. The Kronecker product beamformers clearly outperform the MVDR beamformer in achieving 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...
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 ~ 3 s duration is arbitrarily chosen from the TIMIT [50] corpus. Using appropriate upsampling and
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
interference-noise statistics is evaluated as

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

(41)

The MVDR and the Kronecker product beamformers now minimize \(\phi_{V_{rn}}\), instead of \(\phi_Z\), 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 \(\text{iSIR} = 0 \text{ dB}\). 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:

\[
\hat{z}(t) = \begin{cases} 
\mu, & t < r_s \\
z(t), & t \geq r_s, \quad \hat{z}(t) \in [-1, 1]
\end{cases}
\]

(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

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

(43)

where \(\text{var}[.]\) 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 negatively affect 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
