Supplementary Material

Rajib Sharma*, Israel Cohen, and Baruch Berdugo

I. Diagramatic Representation of a CCA and the circular co-ordiante system

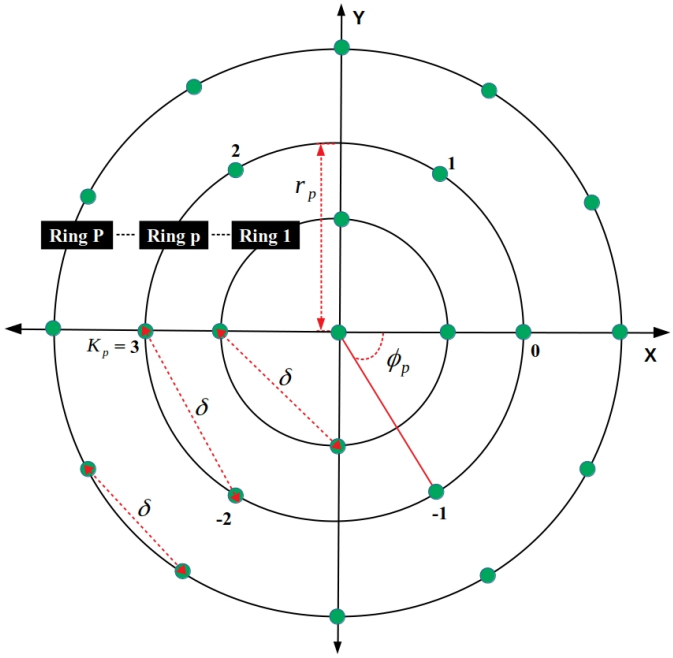

Fig. 1: **Diagrammatic representation (top-view) of a CCA. δ denotes the inter-sensor distance. r_p denotes the radius of the p^{th} ring in which the angular separation of the sensors is ϕ_p.**

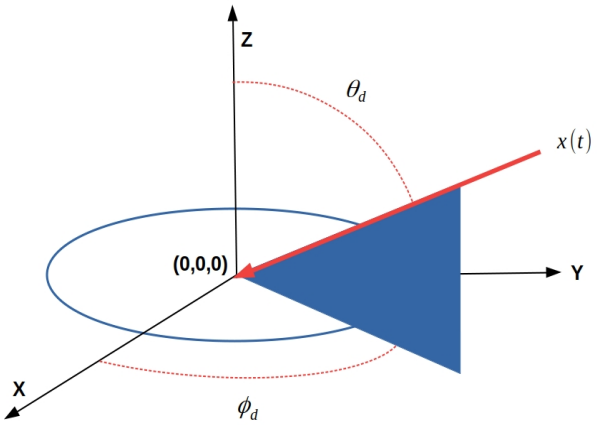

Fig. 2: **The circular co-ordinate system. A signal, $x(t)$, impinges at the center with a direction-of-arrival, $(\theta_{\text{d}}, \phi_{\text{d}})$. The center of the co-ordinate system may represent the center of a CCA.**

R. Sharma (*Corresponding author), I. Cohen, and B. Berdugo are with Andrew and Erna Viterby Faculty of Electrical Engineering, Technion – Israel Institute of Technology, Technion City, Haifa 3200003, Israel (e-mail: rajibd2k@yahoo.com, icohen@ee.technion.ac.il, bbaruch@technion.ac.il).

II. Convergence of the gradient-descent algorithm

We drop the frequency variable, f, for brevity and readability, in our proposed algorithm. Thus, $\mathcal{L}$ is the optimization function, and $\mathbf{s}^{(n)} = [s_1^{(n)}, s_2^{(n)}, ..., s_{2P+1}^{(n)}]^T$ is the parameter-vector at the n^{th} iteration. Similarly, $\Delta^{(n)}s$ and $\mu^{(n)}$ are the step-size and learning-rate at the n^{th} iteration.

Kindly note that in conventional gradient-descent where the mathematical expression for the gradients, $\{\frac{\partial \mathcal{L}}{s_l},\ l = 1, 2, ..., 2P+1\}$, are available, there is no need for step-size. In fact, the terms 'step-size' and 'learning-rate' are used interchangeably and are one and the same. However, in our case, since we do not know the mathematical expressions for the gradients, we use a brute-force gradient-descent algorithm, where the step-size is used to calculate the gradients. A gradient so calculated at a particular point is only an approximation of the true gradient and represents the slope of the line connecting two neighbouring points on either side of the concerned point. We can modify the step-size if the gradient is too-small or too-large to avoid non-change or large fluctuations of the parameters (equivalently that of the optimization function) between iterations. The adaptive learning-rate also serves the same purpose in practice.

In Fig.2 of the manuscript, we can observe the complicated non-linear relationship between the Kaiser-window parameter, $\beta_1(f)$, and the elevation-beamwidth, azimuth-beamwidth, and DF of a circular array. Notice that $\beta_1(f)$ is plotted in the $\log_{10}(.)$ scale. In the case of a CCA with multiple rings, such relationships are likely to be even more complicated. The same may be expected in the case of the ring-weights. Nevertheless, we will assume that the gradient of the optimization function is Lipschitz continuous with constant $L > 0$, i.e.,

$$\nabla_{\boldsymbol{s}}^2 \mathcal{L} \leq L\mathbf{I}_{2P+1},\ \nabla_{\boldsymbol{s}} \mathcal{L} = [\frac{\partial \mathcal{L}}{s_1}, \frac{\partial \mathcal{L}}{s_2}, ..., \frac{\partial \mathcal{L}}{s_{2P+1}}]^T. \tag{1}$$

Lipschitz continuity is not a strong assumption and is believed to hold true even for large-scale machine-learning and neural-network algorithms. In the above, $\mathbf{I}_{2P+1}$ is a square identity matrix of size $2P+1$, and $\mathbf{s}$ is an arbitrary parameter-vector which represents the point at which the gradient of the optimization function is being calculated. From hereon, we will use the notation, $\mathcal{L}(\mathbf{s})$, to indicate the value of the optimization function at the point $\mathbf{s}$. From the above equation, we have,

$$\begin{aligned} \boldsymbol{s}_2^T\ \nabla_{\boldsymbol{s}}^2 \mathcal{L}(\boldsymbol{s}_1)\ \boldsymbol{s}_2 &\leq \boldsymbol{s}_2^T\ L\mathbf{I}_{2P+1}\ \boldsymbol{s}_2 \\ &\leq L||\boldsymbol{s}_2||^2, \end{aligned} \tag{2}$$

where $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ are two arbitrary points.

Consider that the optimization function is quadratic. Then, using multivariate Taylor series expansion, we have,

$$\begin{aligned} \mathcal{L}(\mathbf{s}_2) &= \mathcal{L}(\mathbf{s_1}) + \ (\nabla_{\boldsymbol{s}} \mathcal{L}(\boldsymbol{s}_1))^T(\boldsymbol{s}_2 - \boldsymbol{s}_1)\ +\ \frac{1}{2}\ (\boldsymbol{s}_2 - \boldsymbol{s}_1)^T\ (\nabla_{\boldsymbol{s}}^2 \mathcal{L}(\boldsymbol{s}_1))\ (\boldsymbol{s}_2 - \boldsymbol{s}_1) \\ &\leq \mathcal{L}(\mathbf{s}_1) + \ (\nabla_{\boldsymbol{s}} \mathcal{L}(\boldsymbol{s}_1))^T(\boldsymbol{s}_2 - \boldsymbol{s}_1)\ +\ \frac{L}{2}||\boldsymbol{s}_2 - \boldsymbol{s}_1||^2. \end{aligned} \tag{3}$$

Now, consider that, $\mathbf{s}_2 = \mathbf{s}^{(n+1)}$, and $\mathbf{s}_1 = \mathbf{s}^{(n)}$, and the learning-rate is fixed, i.e., $\mu^{(n)} = \mu$. We have,

$$\boldsymbol{s}^{(n+1)} = \boldsymbol{s}^{(n)} - \mu \nabla_{\boldsymbol{s}} \mathcal{L}(\boldsymbol{s}^{(n)}),\ \mu > 0 \tag{4}$$

$$\boldsymbol{s}_2 = \boldsymbol{s}_1 - \mu \nabla_{\boldsymbol{s}} \mathcal{L}(\boldsymbol{s}_1) \tag{5}$$

Thus, we have,

$$\begin{aligned} \mathcal{L}(\mathbf{s}^{(n+1)}) &\leq \mathcal{L}(\mathbf{s}^{(n)}) + \ (\nabla_{\boldsymbol{s}} \mathcal{L}(\boldsymbol{s}^{(n)}))^T(\boldsymbol{s}^{(n+1)} - \boldsymbol{s}^{(n)})\ +\ \frac{L}{2}||\boldsymbol{s}^{(n+1)} - \boldsymbol{s}^{(n)}||^2 \\ &\leq \mathcal{L}(\mathbf{s}^{(n)}) + (\frac{L}{2} - \mu)||\nabla_{\boldsymbol{s}} \mathcal{L}(\boldsymbol{s}^{(n)})||^2. \end{aligned} \tag{6}$$

The above condition is guaranteed to be true for $\mu \leq \frac{L}{2}$. Thus, $0 < \mu \leq \frac{L}{2}$.

In other words, the gradient-descent algorithm is guaranteed to converge for small learning-rate values. Computing the value of L is quite expensive and difficult. Thus, in practice, the learning-rate is determined by trial-and-error. Using an adaptive learning-rate facilitates this process. Generally, as we have implemented in our work, a relatively large learning-rate is initially chosen, which is then decreased based on certain checks and balances as the algorithm iterates.

References: https://www.cs.ubc.ca/~schmidtm/Courses/540-W18/L4.pdf
https://www.stat.cmu.edu/~ryantibs/convexopt-F13/scribes/lec6.pdf.

III. Final parameters' values of the KW beamformer applied on CCA-I

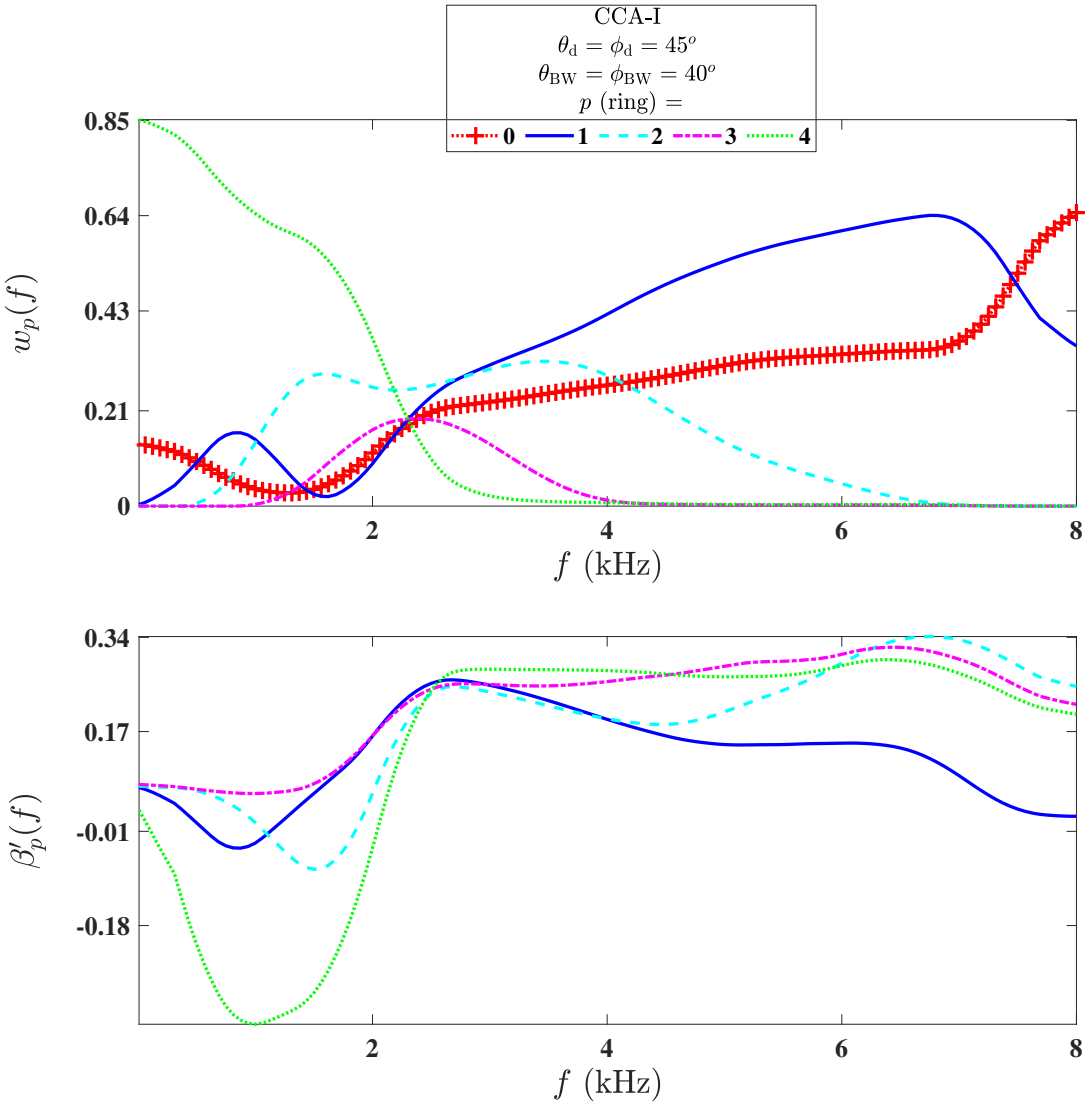

Fig. 3: **Final optimized parameters vs. frequency for the KW beamformer. A CCA-I with $P = 4$ rings and a central sensor is considered.**

The above figure shows the optimal parameters of the KW beamformer for the entire frequency spectrum. There are five ring-weight parameters - one for the central sensor ($p = 0$), and the rest for the four rings. Each of the four rings, additionally, have a Kaiser-window width parameter. It is easier to intuitively make sense of the ring-weight parameters. As we have mentioned in Section III of the manuscript, higher frequencies generally have lesser beamwidths, and larger rings further accentuate the situation. Hence, the algorithm assigns larger weights to the central sensor and the smaller rings, as the frequency increases. Contrarily, the algorithm assigns the largest weight to the outermost ring ($p = 4$) at low frequencies. In the case of the Kaiser-window width parameters, it is difficult to intuitively make sense of the optimal parameters as they have a complicated relationship with the metrics (kindly refer to Fig.2 in the manuscript and the corresponding subject matter). One might simply view them as lending more control over the characteristics of the beamformer, in conjunction with the ring-weight parameters.

IV. Performance on CCA-I (without the central sensor)

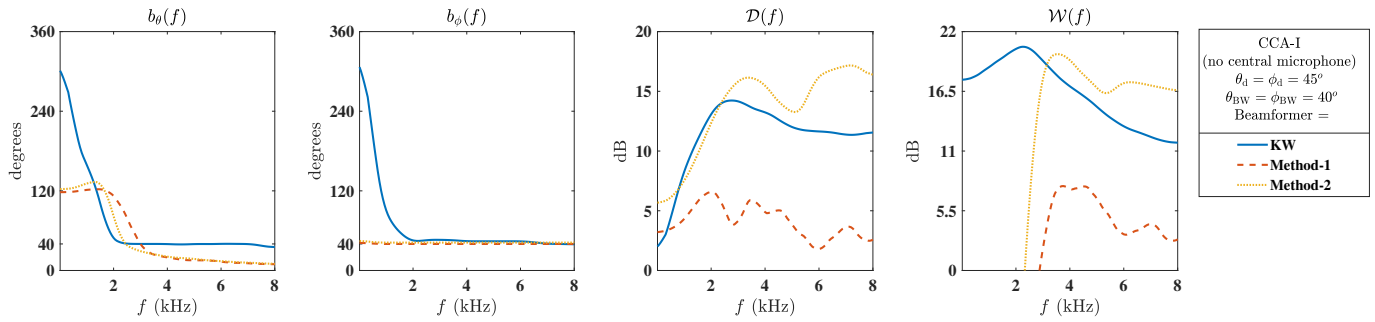

Fig. 4: **Comparison of the performance metrics of the KW, Method-1, and Method-2 beamformers applied on a CCA-I with $P = 4$ rings and without a central microphone.**

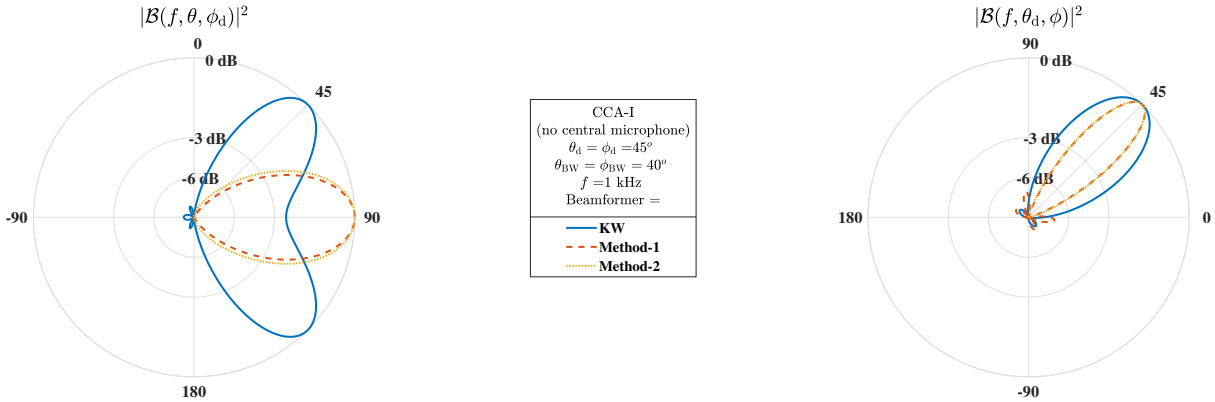

Fig. 5: **Comparison of the elevation and azimuth powerpatterns (normalized) at 1 kHz frequency for the KW, Method-1, and Method-2 beamformers. A CCA-I of $P = 4$ rings without a central microphone is considered.**

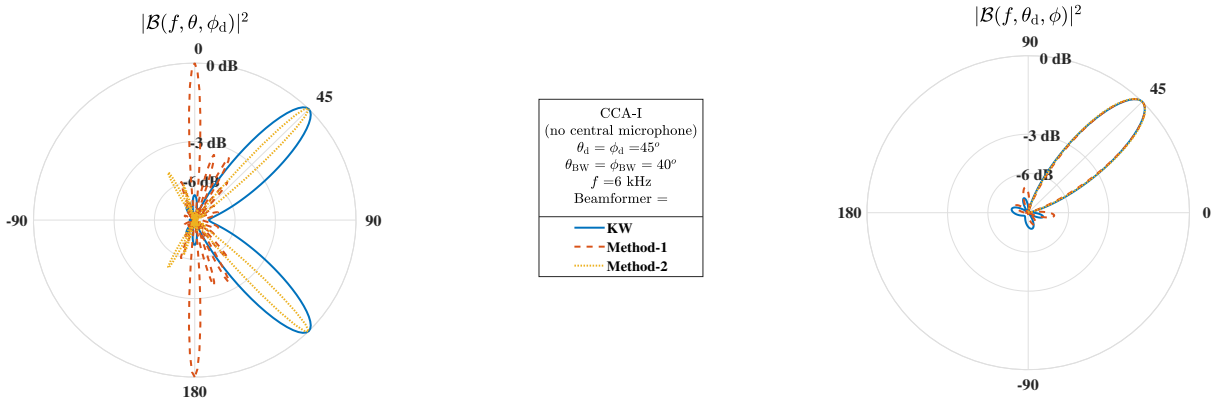

Fig. 6: **Comparison of the elevation and azimuth powerpatterns (normalized) at 6 kHz frequency for the KW, Method-1, and Method-2 beamformers. A CCA-I of $P = 4$ rings without a central microphone is considered.**

We consider the CCA-I design with $P = 4$ rings, as described in the manuscipt, but with no central microphone. The above three figures show the performance metrics and beampatterns for the KW, Method-1, and Method-2 beamformers. We may compare the figures with Figs. 4, 5, and 6 in the manuscript to investigate if the proposed KW beamformer is independent of the presence of the central microphone. The near-identical plots show that that is indeed the case. The above figures also show that the limitations of the Method-1 and Method-2 beamformers persist even if the central microphone is absent.

At this juncture, one must note that had there been only a couple of rings in the CCA-I design, then, the central sensor could have been more important, particularly at high frequencies (kindly refer to the previous section). When multiple rings of large and small sizes are present, then, the central sensor is not so influential. Nevertheless, the proposed KW beamformer, and the assoicated gradient-descent algorithm, is design-independent. It simply does its best for the available array design.