LEAST-SQUARES DISTORTIONLESS RESPONSE BEAMFORMER IN FAR-FIELD ENVIRONMENTS WITH SPATIAL CUES PRESERVATION

F. Mustière

Integrated Device Technology Canada
603 March Road, Kanata,
Ontario, Canada, K2K 2M5

M. Bouchard

School of Electrical Engineering
and Computer Science,
University of Ottawa, Canada, K1N 6N5

ABSTRACT

In this letter, some results involving a form of least-squares beamformer are derived based only on directional criteria. In far-field assumptions (i.e., plane wave propagation model for the signals captured), we show that certain factors appearing in the beamformer coefficients calculation, which are usually formulated using complicated integrals, can be computed using closed-form expressions involving familiar, physically meaningful quantities. Next, by resorting to a limiting case, we demonstrate a clear theoretical link between the resulting solution and MVDR beamforming in cylindrically isotropic noise fields. We then discuss a solution to preserve spatial cues if desired which also allows to easily control and modulate the enhancement strength of the beamformer. Some experimental results are finally given in a challenging real-world environment, showing the merits of the approach.

Index Terms— Multichannel speech enhancement, beamforming, least-squares, MVDR

1. INTRODUCTION

The celebrated Minimum-Variance-Distortionless-Response Beamformer (MVDR) is perhaps the most popular beamforming solution in speech enhancement applications. Very good results can be achieved in a variety of environments; however, one of the important issues that can affect the robustness of MVDR beamformers is the fact that they require an estimate for the noise correlation matrix (i.e., the noise power spectral density at each microphone, and also the noise cross power spectral density between microphones). The estimation of this quantity can be difficult, especially in non-stationary environments, and errors can result in poorly suppressed interfering sources [1].

For applications where the noise is unpredictable and where robustness is important, other techniques based on purely directional criteria may be preferred, for they do not require any assumption about the noises statistics. This is the case with the Least-Squares (LS) beamformer used in this paper, which can provide good performance while remaining flexible. However, the LS beamformer design involves more complex equations in general, for which numerical methods must be employed. In this paper, we first rederive the LS beamformer with and without distortionless constraints, and then show that in the far-field case the equations can be highly simplified, resulting in familiar components in the beamformer design. To deal with the remaining factors that do not admit closed-form expressions, instead of resorting to numerical methods we propose a simple approach that in turn provides a clear theoretical link between the resulting LS formulation and a type of MVDR beamforming. Next, based on the obtained LS beamformer we also derive an expression for a real-valued frequency-dependent gain that can be applied as a multiple output beamformer for scenarios where preservation of spatial impressions is important. We also explain how enhancement strength modulation can be performed. Finally, some experimental results in a real-world complex environments are reported to confirm the practical validity of the presented ideas.

2. LEAST-SQUARES DESIGN FROM AN IDEAL BEAM PATTERN

The design of the LS-beamformer at the core of this paper is done as follows (the reader can refer to [2, 3] for different derivations and slightly different final expressions). Denote by $\mathbf{w}(k)$ the length-M vector of coefficients that we seek, where M is the number of microphones in the system and k is a particular frequency bin. Next, denote by $\mathbf{h}(\theta, k)$ the length-M vector representing the frequency response from a source located at an azimuth θ to each of the microphones. Assume now that the desired directional response of the beamformer is given by $D(\theta, k)$ (a quantity that is not necessarily real). Dropping the index k to focus in a particular frequency bin,

We wish to thank NSERC for their support.

the goal is to minimize the following cost function:

$$J(\mathbf{w}) \quad = \quad \int_0^\pi \left| D(\theta) - \mathbf{w}^H \mathbf{h}(\theta) \right|^2 d\theta \qquad (1)$$

Using Wirtinger's calculus, we can differentiate the above with respect to $\mathbf{w}^H$ and equate the resulting expression to 0:

$$\left[\int_0^\pi \mathbf{h}(\theta)\mathbf{h}^H(\theta) d\theta \right] \mathbf{w} - \int_0^\pi \mathbf{h}(\theta)\overline{D(\theta)} d\theta = 0 \quad (2)$$

We may denote by $\mathbf{Q}$ the matrix $\int_0^\pi \mathbf{h}(\theta)\mathbf{h}^H(\theta)d\theta$ and by $\mathbf{p}$ the vector $\int_0^\pi \mathbf{h}(\theta)\overline{D(\theta)}d\theta$, so as to obtain an expression for the optimal value of $\mathbf{w}$:

$$\mathbf{w} \quad = \quad \mathbf{Q}^{-1}\mathbf{p} \qquad (3)$$

The inclusion of a distortionless constraint in a specified target direction θ_{target} can be done by introducing a complex Lagrange multiplier λ and solving for $\mathbf{w}$ in the augmented gradient equation:

$$\mathbf{Q}\mathbf{w} - \mathbf{p} + \lambda \mathbf{h}(\theta_{\text{target}}) \quad = \quad 0 \qquad (4)$$

which yields the solution:

$$\begin{aligned} \mathbf{w} \quad &= \quad \mathbf{Q}^{-1} \left[\mathbf{p} - \lambda \mathbf{h}(\theta_{\text{target}}) \right] \qquad (5) \\ \text{with } \lambda \quad &= \quad \frac{\mathbf{h}(\theta_{\text{target}})^H \mathbf{Q}^{-1}\mathbf{p} - 1}{\mathbf{h}(\theta_{\text{target}})^H \mathbf{Q}^{-1} \mathbf{h}(\theta_{\text{target}})} \end{aligned}$$

Note that an arbitrary angle-dependent weighting function can be immediately added to the cost function in Eqn. (1). An example of beampattern obtained with this type of beamformer is shown in Fig. 1. The main drawback that may discourage designers of real-world speech enhancement systems with limited computational resources is the required calculation of every element of $\mathbf{Q}$ and especially $\mathbf{p}$ since it is dependent on the beampattern that is currently sought (e.g., the target azimuth in a speech enhancement application). Both $\mathbf{Q}$ and $\mathbf{p}$ involve integrals that very likely do not have closed form expressions (depending on the actual form of $\mathbf{h}$). In the next Section, we show that with far-field assumptions, interesting results regarding both $\mathbf{Q}$ and $\mathbf{p}$ can be reached, both of practical and theoretical consequences.

3. FAR-FIELD FORMULATION AND RELATIONSHIP WITH MVDR BEAMFORMING

3.1. Closed-form expression for Q

Assuming a far-field approximation and a linear array, the i^{th} element of the transfer function vector $\mathbf{h}(\theta)$ can be assumed to follow:

$$\mathbf{h}(\theta)[i] \quad = \quad \exp\left(-j2\pi f \frac{d_{i-1}}{c} \cos(\theta) \right) \qquad (6)$$

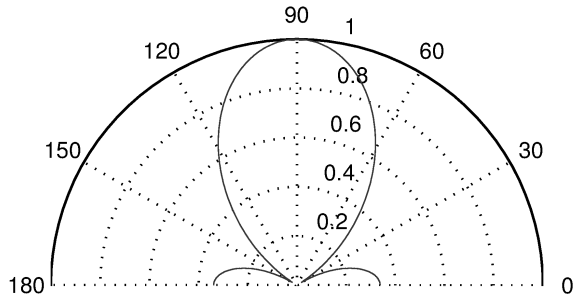

Fig. 1. Beampattern obtained at 1.5 kHz with a LS-beamformer using 4 microphones placed in a linear array, all equidistant by 3 cm, and a frontal target. The ideal beampattern was set to 1 between 70 and 110 degrees.

where by convention d_i is the distance in meters between the i^{th} microphone and the first microphone (and therefore $d_0 = 0$). Moreover, c is the speed of sound in meters per second, f corresponds to the current frequency in Hertz (f depends of course on the frequency bin index k, which was dropped earlier for clarity). The $[i,j]^{\text{th}}$ entry of the matrix $\mathbf{Q}$ can thus be written as:

$$\mathbf{Q}[i,j] \quad = \quad \int_0^\pi \exp\left(-j2\pi f \frac{d_{i-1} - d_{j-1}}{c} \cos(\theta) \right) d\theta \qquad (7)$$

The above integral can be rewritten as follows (for this, tables of integrals such as [4] can be used). Let $\mathbf{d}$ represent the vector of coordinates for the microphones, i.e., $\mathbf{d} = [d_0; d_1; d_2; \ldots; d_M]$. We can then write:

$$\mathbf{Q} = \pi J_0 \left(\frac{2\pi f}{c} \text{Toeplitz}\left(\mathbf{d}\right) \right) \qquad (8)$$

where J_0 is the Bessel function of the first kind. The above expression, which is significantly less intimidating than those appearing in previous publications (e.g. [2]), corresponds to a well-known result: it is in fact the coherence function for cylindrically isotropic noise fields [5].

3.2. Practical determination of p and relationship with MVDR beamforming

Next, the vector $\mathbf{p}$ is often more problematic in practice as it may need to be regularly updated. For a real-valued $D(\theta)$ such that:

$$D(\theta) \quad = \quad \begin{cases} 1, & \theta_1 \leq \theta \leq \theta_2 \\ 0, & \text{otherwise} \end{cases} \qquad (9)$$

Then the i^{th} element of $\mathbf{p}$ is:

$$\mathbf{p}[i] \quad = \quad \int_{\theta_1}^{\theta_2} \exp\left(-j2\pi f \frac{d_{i-1}}{c} \cos(\theta) \right) d\theta \qquad (10)$$

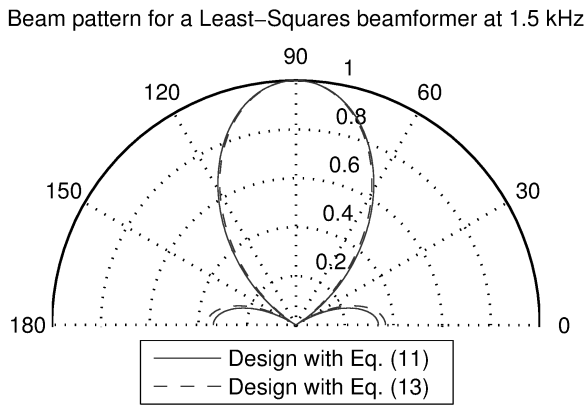

Fig. 2. Solid line was obtained from a design with Eqn. (9) and $\theta_1 = 70$, $\theta_2 = 110$, and dashed line was obtained with a design based on Eqn. (11).

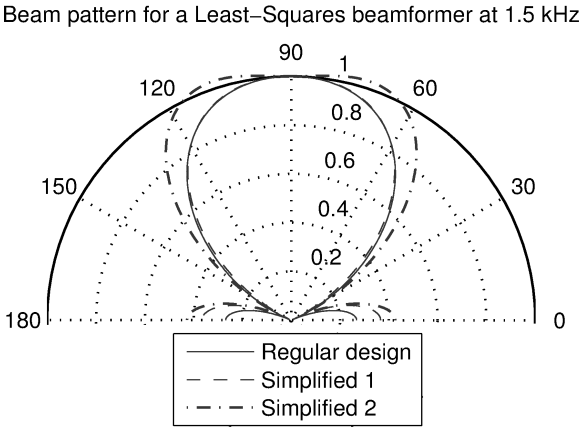

Fig. 3. Solid line obtained from a design with Eqn. (9) and $\theta_1 = 50$, $\theta_2 = 130$ (purposely wider than in the previous cases). Dashed line obtained with a design based on Eqn. (14), with Δ chosen by trial-and-error to match the pattern of the solid line. The dash-dot line is obtained with a larger value of Δ.

Unfortunately, for such a desired directional response $D(\theta)$, the values of $\mathbf{p}$ must here be computed numerically. One simple approach to obtain a closed form solution is the following. Since the goal is in many cases to isolate a certain target direction in a noisy environment, we must devise a certain desired directional response $D(\theta)$ such that not only is the target azimuth θ_{target} preserved and the energy from the other angles minimized, but also the calculation of $\mathbf{p}$ is simplified. In this context, we may propose the following function:

$$D(\theta) \quad = \quad \delta\left(\theta - \theta_{\text{target}}\right) \tag{11}$$

With the above, we immediately have:

$$\mathbf{p} \quad = \quad \mathbf{h}\left(\theta_{\text{target}}\right) \tag{12}$$

and no additional computations must be performed. An example of beampattern obtained with the above choice for $D(\theta)$ is shown in Fig. 2, along with one obtained with a choice corresponding to Eqn. (9). It appears that in this particular 4-microphones array example, the beampatterns are practically identical. This special case provides a direct link between a distortionless Least-Squares formulation and MVDR beamforming. Indeed, plugging the above equation to Eqn. (5) yields the following expression for $\mathbf{w}$:

$$\mathbf{w} \quad = \quad \frac{\mathbf{Q}^{-1}\mathbf{h}(\theta_{\text{target}})}{\mathbf{h}(\theta_{\text{target}})^H\mathbf{Q}^{-1}\mathbf{h}(\theta_{\text{target}})} \tag{13}$$

In other words, under far-field assumptions, the distortionless LS beamformer with desired directional response $D(\theta) = \delta\left(\theta - \theta_{\text{target}}\right)$ derived is identical to the MVDR beamformer derived for cylindrically isotropic noise fields. To obtain a beamformer with a wider main lobe (i.e., to increase $\theta_2 - \theta_1$ in the design based on Eqn. (9)), a similar practical solution can be used. Let $2\Delta = \theta_2 - \theta_1$, then following the same reasoning as before, a possible choice for $D(\theta)$ is:

$$\begin{aligned} D(\theta) \quad &= \quad \delta\left(\theta - \theta_{\text{target}}\right) + \delta\left(\theta - \theta_{\text{target}} - \Delta\right) \\ &+ \quad \delta\left(\theta - \theta_{\text{target}} + \Delta\right) \end{aligned} \tag{14}$$

The above corresponds to the following $\mathbf{p}$ vector:

$$\begin{aligned} \mathbf{p} \quad &= \quad \mathbf{h}\left(\theta_{\text{target}}\right) + \mathbf{h}\left(\theta_{\text{target}} - \Delta\right) \\ &+ \quad \mathbf{h}\left(\theta_{\text{target}} + \Delta\right) \end{aligned} \tag{15}$$

The resulting beamformer is not quite equivalent to an MVDR anymore. Some example designs are shown in Fig. 3. Of course Δ could be arbitrarily tuned to the desired effect – this is in fact what was done to produce the dashed line in Fig. 3 – and other additional components in $\mathbf{p}$ may be added as well. Note also that it would be straightforward to add constraints so that no distortion is present at the three directions understated by the above expression.

4. PRESERVATION OF SPATIAL CUES AND ENHANCEMENT STRENGH CONTROL

For some applications the preservation of spatial impressions is critical; for example, in audio systems designed to provide immersive environments or in hearing aids system. Preserving the spatial impressions from the M microphones can be obtained by converting the beamformer output into a single, real-valued frequency-dependent spectral gain G to be applied to all the input measurements [6]. It should be noted that

with this processing the resulting system becomes a multiple-input multiple-output (MIMO) system, as opposed to traditional beamforming systems which are multiple-input single-output systems. The gain G should be unitless and proportional to the LS-beamformer single output. With $\mathbf{z}$ representing the noisy input vector, one choice is the following [6]:

$$G = \frac{\left|\mathbf{w}^H \mathbf{z}\right|}{\left\|\mathbf{z}\right\|_1} \tag{16}$$

In this case however, the beamformer should be specifically designed with a constraint set to $\mathbf{w}^H \mathbf{h}(\theta_{\text{target}}) = \left\|\mathbf{h}(\theta_{\text{target}})\right\|_1$ (this is readily seen by assuming that $\mathbf{z}$ is only composed of the target signal and setting G to 1). Going through the beamformer derivation with the above constraint yields the same equation as Eq. (5) but with the following multiplier:

$$\lambda \quad = \quad \frac{\mathbf{c}^H \mathbf{Q}^{-1} \mathbf{p} - \left\|\mathbf{c}\right\|_1}{\mathbf{c}^H \mathbf{Q}^{-1} \mathbf{c}} \tag{17}$$

and where $\mathbf{c} = \mathbf{h}(\theta_{\text{target}})$. Another approach would be to derive a gain G expressible in terms of optimal multichannel spectral amplitude estimators (such as the multichannel Minimum Mean Squared Error Spectral Amplitude Estimator, among others [7]). In order to conveniently modulate the enhancement strength (e.g., to also control the level of output distortion), one can apply a modified gain G^α where $\alpha > 0$. Clearly, larger α values will remove more noise.

5. EXPERIMENTS

This section will present experimental results for the multiple-output beamformer described in Section 4, preserving the spatial impression. The multiple-output beamformer is only useful if no other known solution can naturally do better. For example, in an overdetermined and background-noise-free situation with fixed point sources, a well-tuned ICA-based solution could perform well and could also reconstruct the spatial acoustic images of the sources at the different microphones. The two main advantages of the multiple-output beamformer from Section 4 are that (1) the design is done the same way regardless of the noise complexity (although this is also true for ICA methods) and (2) it is fast and efficient (especially when compared to iterative ICA-based solutions). Below, we choose a challenging case to apply the multiple-output beamformer algorithm, where typical frequency domain ICA methods would struggle because the environment is not limited to point sources (i.e., presence of significant background or diffuse noise) [8].

The noisy file consists of real 4-channel recordings of 4 different speakers accompanied with background subway noise. An initial "3-speakers + subway noise" recording was picked among the test data from the SISEC 2010 evaluation campaign [9], and to increase the difficulty an extra speaker was added from the development data, coming from another direction. The respective directions-of-arrival (DOAs) of each speaker are 20, 85, 115, and 140 degrees, and are assumed to be fixed throughout the 10-seconds long recording. We use the multiple-output beamformer with spatial cues preservation of Eq. (16) via the simplified design resorting to Eq. (12). The DOAs are initially estimated using [10], and we choose to extract two of the targets, from the two first determined DOAs of approximately 19.2 and 88.3 degrees. The results as well as the noisy files are available at [11], where we have also additionally applied a frequency-domain multichannel Minimum-Mean-Squared-Error (MMSE) postfilter also preserving spatial cues [7].

Since we have no access to the clean files at DOAs 20 and 85, no objective results are reported in this paper; nevertheless the reader is invited to listen to the results at [11]. The two extracted speakers are significantly more intelligible than in the crowded, noisy mixture, confirming the usefulness of the above approach in complex situations. We also find that increasing the exponent of G does remove more noise, although as expected more artefacts appear as well. For the beamforming stage (i.e., excluding the MMSE postfilter), the 10 seconds of signal are processed with non-optimized MATLAB code in about 7 seconds on an entry-level, several-years-old laptop. This figure is to be compared with the minutes-long required processing using high-end desktop machines reported by SISEC 2010 competitors (in the simpler case of 3 sources + noise only). With the MMSE postfilter, the overall processing takes about 30 seconds on the same older laptop. Finally, while the experiments are conducted with non-moving speakers, it is interesting to note that the speakers that are considered part of the unwanted signal (i.e., the noise) could very well be in movement without affecting the algorithm's performance (unless they move too close to the extracted sources of course). This can be viewed as another advantage over some ICA-based algorithms or some other methods that must estimate the noise statistics.

6. CONCLUSION

A beamformer designed using a purely directional least-squares criterion is recapitulated in the first part of the paper. In far-field assumptions, it is shown that the beamformer design involves familiar coherence functions, as obtained in isotropic noise fields. Moreover, a direct theoretical relationship is established between a special-case of the LS beamformer and the usual MVDR beamformer. Next, a multiple-output beamformer solution is presented to address situations where it is desired to maintain spatial cues while controlling the enhancement strength. Finally, the relevance of the presented method is illustated with an experiment in a challenging real-world environment.

7. REFERENCES

[1] R.C. Hendriks and T. Gerkmann, "Noise Correlation Matrix Estimation for Multi-Microphone Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, 2012.

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

[3] S. Doclo, M. Moonen, "Design of Far-Field and Near-Field Broadband Beamformers using Eigenfilters," *Elsevier Signal Processing*, vol. 83, 2003.

[4] I. S. Gradshteyn, I. W. Ryshik, *Table of Integrals, Series, and Products*, 5th Ed., New York: Academic Press, 1994.

[5] B. Cron, C. Sherman, "Spatial Correlation Functions for Various Noise Models," *Journal of Acoustical Society of America*, vol. 34, issue 11, pp. 1732-1736, 1962.

[6] T. Lotter and H. Vary, "Dual Channel Speech Enhancement by Superdirective Beamforming", *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–14, 2006.

[7] F. Mustiere, M. Bouchard, H. Najaf-Zadeh, R. Pichevar, L. Thibault, and H. Saruwatari, "Design of multichannel frequency domain statistical-based enhancement systems preserving spatial cues via spectral distances minimization", *Signal Processing*, vol. 93, no. 1, pp. 321-325, Jan. 2013.

[8] Y. Takahashi, T.Takatani, K.Osako, H.Saruwatari, and K.Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, n. 4, pp. 650-664, May 2009.

[9] S. Araki, A. Ozerov, B.V. Gowreesunker, H. Sawada, F.J. Theis, G. Nolte, D. Lutter and N.Q.K. Duong, "The 2010 Signal Separation Evaluation Campaign (SiSEC2010): - Audio source separation -," in *Latent Variable Analysis and Signal Separation*, pp. 114-122, 2010.

[10] C. Blandin, A. Ozerov, E. Vincent, "Multi-source TDOA estimation using SNR-based angular spectra," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 2616 – 2619, 2011.

[11] Accompanying audio demonstration files: www.eecs.uottawa.ca/%7Ebouchard/papers/Icassp13%5FLSB%5Fresults.zip