RAKING ECHOES IN THE TIME DOMAIN

Robin Scheibler, Ivan Dokmanić, and Martin Vetterli

School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
{robin.scheibler,ivan.dokmanic,martin.vetterli}@epfl.ch

ABSTRACT

The geometry of room acoustics is such that the reverberant signal can be seen as the same waveform emitted from multiple locations. In analogy with the rake receiver from wireless communications, we propose several beamforming strategies that exploit, rather than suppress, this additional spatio-temporal diversity. Unlike earlier work in the frequency domain, time domain designs allow to shape the impulse response of the beamformer. In particular, we can control perceptually relevant parameters, such as the amount of early echoes or the length of the beamformer response.

Relying on the knowledge of the image sources positions, we derive different optimal beamformers. Leveraging perceptual cues, we show how to improve interference and noise reduction without degrading the perceptual quality. The designs are validated through simulation. Using early echoes is shown to strictly improve the signal to interference and noise ratio. Code and speech samples are available online at `http://lcav.epfl.ch/Robin_Scheibler`.

***Index Terms*—**Beamforming, acoustic rake receiver, time domain, precedence effect, room geometry.

1. INTRODUCTION

Rake receivers for wireless communication exploit the temporal diversity of the multipath fading channel to increase the signal-to-noise ratio (SNR) [1]. The technique extends to arrays of antennas [2, 3]. One can imagine using a similar approach in acoustics, exploiting echoes in a reverberant room to improve the SNR. Indeed, such techniques have been proposed [4, 5, 6]. More recently, Dokmanić et al. developed the concept of acoustic rake receiver in more details and proposed several optimal and intuitive formulations according to the raking principle [7].

A large part of the beamforming literature tends to focus on dereverberation and room equalization and assume a detailed knowledge of the room impulse response [8, 9, 10]. This approach has two main drawbacks. First, it considers all reverberation as harmful. Second, the room impulse response is generally difficult to estimate precisely. Instead we are only interested in exploiting the early echoes to improve the desired source power versus that of an interferer or ambient noise.

Psychoacousticians demonstrated that the energy of early echoes (within 30 ms to 90 ms of the direct sound) is perceptually integrated with the direct sound [11]. Thus fully distortionless response seems not completely necessary. In fact, different works have shown that channel shortening rather than inversion leads to practical systems and better behaved filters [12, 13].

Moreover, locating just the early reflections is significantly easier than full estimation of the room impulse response (RIR). In many situations, the shape of the room can be known in advance from blueprints or measurements [14]. Then knowing the location of the real source allows to calculate the positions of the echoes. Localizing the direct sound is a well understood problem [15]. In ad-hoc deployment, recent works propose a calibration step to locate the main reflectors [14, 16, 17, 18]. Note that there is in fact no necessity to know the room geometry exactly, the positions of the image sources being sufficient. The echo sorting algorithm from [14] allows to locate the main echoes from measured RIR. Another approach is the audio camera of [4].

In [7], the beamformers are formulated in the frequency domain for narrowband sources. To extend the beamformer to wideband signals, the short time Fourier transform is applied to the signal and the optimization problem is solved for every frequency band. While the frequency domain formulation is simpler, it does not allow precise control over critical parameters of the beamforming filters. In particular, the beamforming filters might be very long and we would like to approximate them by short filters.

This paper brings together the raking principle, geometrical acoustics and perceptual criteria to optimize beamformers directly in the time-domain. We present several optimal formulations for raking beamformers. We demonstrate how the geometry of early echoes determines the minimum delay necessary for maximal raking. Conversely, we show how the delay determines the number of echoes that can be raked. Further, we show how relaxing the distortionless requirement according to psychoacoustics allows to obtain better behaved beamforming filters and higher signal to interference and noise ratio (SINR), while maintaining tight control over pre-echoes. Throughout the paper we assume the positions of the principal

Authors are with LCAV-EPFL. This work was supported by the ERC Advanced Investigators Grant: Sparse Sampling: Theory, Algorithms and Applications SPARSAM no. 247006, and a Google Doctoral Fellowship.

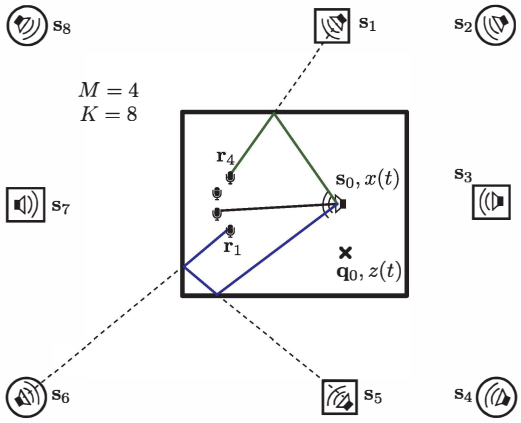

Fig. 1. Illustration of the image source model and the notation of the paper. First (□) and second (◯) order image sources of $\boldsymbol{s}_0$ are shown.

images sources to be known. In practice they can be estimated using one of the techniques mentioned earlier.

This paper is organized as follows. Section 2 introduces the notation, the signal model and basics of beamforming. Section 3 presents several time-domain formulations of rake beamformers. The beamformers are validated through numerical experiments in Section 4. Conclusions are drawn in Section 5.

2. NOTATION AND SIGNAL MODEL

We denote all matrices by bold uppercase letters, for example $\boldsymbol{A}$, and all vectors by bold lowercase letters, for example $\boldsymbol{x}$. The Euclidean norm of a vector is denoted by $\|\cdot\|$, as in $\|\boldsymbol{x}\| \stackrel{\text{def}}{=} (\boldsymbol{x}^T\boldsymbol{x})^{\frac{1}{2}}$. All vectors and matrices are real-valued.

Suppose that in a room, there is a desired source of sound located at $\boldsymbol{s}_0$. Sound from this source arrives at the microphones located at $(\boldsymbol{r}_m)_{m=1}^M$ via the direct path, but also through echoes from the walls. We model echoes, or the multipath propagation, by the image source model [19, 20]. Image sources are simply the mirror images of the real sources across the corresponding walls.

Denote the signal emitted by the source $x[n]$ (e.g. the speech signal). Then all the image sources emit $x[n]$ as well, and the signal from the image sources reaches the microphones with the appropriate delays, that correspond to delays of the echoes. In our application, the essential fact is that echoes correspond to image sources. We denote the image sources positions by $\boldsymbol{s}_k$, $1 \le k \le K$, regardless of their generation, or the sequence of walls that generates them. This is illustrated in Fig. 1. Let K denote the largest number of image sources considered.

Suppose that in addition to the desired signal, there is an interferer at the location $\boldsymbol{q}_0$. For simplicity, we consider only a single interferer, but in general there could be any number of them. The interferer emits the signal $z[n]$, and its image sources emit $z[n]$ as well. Similarly as for the desired source, $\boldsymbol{q}_k$, $1 \le k \le K'$ denote the positions of interfering image sources, where K' is the largest number of interfering image sources considered.

The signal at each microphone can thus be written

$$y_m(t) = \sum_{k=0}^{K}(a_m(\boldsymbol{s}_k,t) * x(t)) + \sum_{k=0}^{K'}(a_m(\boldsymbol{q}_k,t) * z(t)) + b_m(t) \quad (1)$$

where $a_m(\boldsymbol{s}_k,t)$ is the channel response between $\boldsymbol{s}_k$ and the $\boldsymbol{r}_m$, and $b_m(t)$ is additive white Gaussian noise (AWGN) at $\boldsymbol{r}_m$. In our simple model, we do not consider frequency selectivity of the walls and assume that

$$a_m(\boldsymbol{s}_k,t) = \frac{\alpha(\boldsymbol{s}_k)}{4\pi\|\boldsymbol{s}_k - \boldsymbol{r}_m\|}\delta\left(t - \frac{\|\boldsymbol{s}_k - \boldsymbol{r}_m\|}{c}\right)$$

where $\boldsymbol{r}_m$ is the position of the mth microphone and c is the speed of sound in air, $\alpha(\boldsymbol{s}_k)$ is an attenuation factor depending on the reflection order, and $\delta(t)$ is the Dirac delta function. We discretize the channel response into an FIR filter by convolution with an ideal low-pass filter,

$$\begin{aligned} a_m(\boldsymbol{s}_k,n) &= \int_{-\infty}^{\infty} a_m(\boldsymbol{s}_k,u)\ \mathrm{sinc}(n - F_s\,u)\,du \\ &= \frac{\alpha(\boldsymbol{s}_k)}{4\pi\|\boldsymbol{s}_k - \boldsymbol{r}_m\|}\,\mathrm{sinc}\left(n - F_s\frac{\|\boldsymbol{s}_k - \boldsymbol{r}_m\|}{c}\right). \end{aligned}$$

We assume in addition that these discrete filters can be limited to length L_h. We can now rewrite (1) in matrix form

$$\boldsymbol{y}_m = \sum_{k=0}^{K}\boldsymbol{A}_m(\boldsymbol{s}_k)\boldsymbol{x} + \sum_{k=0}^{K'}\boldsymbol{A}_m(\boldsymbol{q}_k)\boldsymbol{z} + \boldsymbol{b}_m$$

where

$$\begin{aligned} \boldsymbol{y}_m &= [y_m[n],\ldots,y_m[n-L_g+1]]^T, \\ \boldsymbol{x} &= [x[n],x[n-1],\ldots,x[n-L+1]]^T, \\ \boldsymbol{z} &= [z[n],z[n-1],\ldots,z[n-L+1]]^T, \\ \boldsymbol{b}_m &= [b_m[n],b_m[n-1],\ldots,b_m[n-L_g+1]]^T. \end{aligned}$$

and $\boldsymbol{A}_m(\boldsymbol{s}_k)$ is the $L_g \times L$ convolution matrix, with L_g the size of the beamforming filter, $L = L_h + L_g - 1$. It is a Toeplitz matrix whose first row is $a_m(\boldsymbol{s}_k,n)$, $n = 0,\ldots,L_h - 1$, padded with $L_g - 1$ zeros, and first column is $a_m(\boldsymbol{s}_k,0)$ followed by $L_g - 1$ zeros.

Stacking all the vectors and matrices, indexed by m into a single vector and matrix, and dropping the index, we obtain the following compact form

$$\boldsymbol{y} = \boldsymbol{H}_s\boldsymbol{x} + \boldsymbol{H}_q\boldsymbol{z} + \boldsymbol{b},$$

where $\boldsymbol{H}_s = \sum_{k=0}^{K}\boldsymbol{A}(\boldsymbol{s}_k)$ and $\boldsymbol{H}_q = \sum_{k=0}^{K'}\boldsymbol{A}(\boldsymbol{q}_k)$. The mth beamforming filter is $\boldsymbol{g}_m = [g_m[0],\ldots,g_m[L_g-1]]^T$ and its output at time n can be written as the inner product $\boldsymbol{g}_m^T\boldsymbol{y}_m$. Stacking all M filters in a vector, $\boldsymbol{g} = [\boldsymbol{g}_0^T\ \cdots\ \boldsymbol{g}_{M-1}^T]^T$, the sum of all filter outputs is conveniently computed as $\boldsymbol{g}^T\boldsymbol{y}$. The

responses of the beamformer towards the desired source and interferer are

$$\boldsymbol{u}_s = \boldsymbol{H}_s^T \boldsymbol{g}, \qquad \boldsymbol{u}_q = \boldsymbol{H}_q^T \boldsymbol{g},$$

respectively. Finally, the letter τ is used to denote the delay (in samples) of the beamformer.

3. TIME-DOMAIN RAKE BEAMFORMERS

3.1. Minimum Variance Distortionless Response Rake Beamformer

A time-domain flavour of the classic Capon minimum variance distortionless response (MVDR) beamformer [21] is given by[1],

$$\underset{\boldsymbol{g}}{\text{minimize}}\ \mathbb{E}|\boldsymbol{g}^T\boldsymbol{y}|^2 \ \text{ subject to } \boldsymbol{g}^T\boldsymbol{h}_\tau = 1,$$

where $\boldsymbol{h}_\tau$ is the τth column of $\boldsymbol{H}_s$. The constraint forces unit response towards the desired source. The value of τ determines the delay of the beamformer and should be larger than the latest arriving echoes that we would like to rake. The objective can be developed into $\mathbb{E}|\boldsymbol{g}^T\boldsymbol{y}|^2 = \boldsymbol{g}^T\boldsymbol{R}_{yy}\boldsymbol{g}$ where $\boldsymbol{R}_{yy}$ is the covariance matrix of $\boldsymbol{y}$,

$$\boldsymbol{R}_{yy} = \boldsymbol{H}_s\boldsymbol{R}_{xx}\boldsymbol{H}_s^T + \boldsymbol{H}_q\boldsymbol{R}_{zz}\boldsymbol{H}_q^T + \boldsymbol{R}_{bb},$$

where in turn $\boldsymbol{R}_{xx}$, $\boldsymbol{R}_{zz}$, and $\boldsymbol{R}_{bb}$ are the covariance matrices of $\boldsymbol{x}$, $\boldsymbol{z}$, and the noise. The optimization problem becomes

$$\underset{\boldsymbol{g}}{\text{minimize}}\ \boldsymbol{g}^T\boldsymbol{R}_{yy}\boldsymbol{g} \ \text{ subject to } \boldsymbol{g}^T\boldsymbol{h}_\tau = 1 \tag{2}$$

and is solved for

$$\boldsymbol{g}_{\text{R-MVDR}} = \boldsymbol{R}_{yy}^{-1}\boldsymbol{h}_\tau(\boldsymbol{h}_\tau^T\boldsymbol{R}_{yy}^{-1}\boldsymbol{h}_\tau)^{-1}.$$

Assuming samples from both sources are independent and identically normally distributed, and that the noise is AWGN, i.e. $\boldsymbol{R}_{xx} = \sigma_x^2\boldsymbol{I}$, $\boldsymbol{R}_{zz} = \sigma_z^2\boldsymbol{I}$, and $\boldsymbol{R}_{bb} = \sigma_n^2\boldsymbol{I}$, (2) can be rewritten

$$\begin{aligned}\underset{\boldsymbol{g}}{\text{minimize}} \quad & \sigma_x^2\left\|\boldsymbol{u}_s\right\|^2 + \sigma_z^2\left\|\boldsymbol{u}_q\right\|^2 + \sigma_n^2\left\|\boldsymbol{g}\right\|^2 \\ \text{subject to} \quad & u_s[\tau] = 1,\ \boldsymbol{u}_s = \boldsymbol{H}_s^T\boldsymbol{g},\ \boldsymbol{u}_q = \boldsymbol{H}_q^T\boldsymbol{g},\end{aligned}$$

where $u_s[\tau]$ is the τth element of $\boldsymbol{u}_s$. From this form, it is clear that the optimal beamformer will balance distortionless response towards desired source, interference cancellation, and noise suppression. For a fixed L_g, adding more image sources will increase L_h and consequently the number of constraints in the optimization problem. Reducing so the feasible set might decrease the noise suppression performance of the beamformer.

Finally, using our geometric interpretation it is possible to know precisely how many echoes can be exploited. Because the response is distortionless, the output of the beamformer should be the desired source with a delay τ (not considering model inaccuracies). This means that only echoes arriving within the time τ of the direct sound can be used to improve the source power. Knowing the propagation speed of sound translates into a geometrical criterion on which image sources can be included. All image sources within distance $\|\boldsymbol{s}_0 - \boldsymbol{r}_m\| + c\tau/F_s$ of the microphone array can be used, c being the speed of sound, and F_s the sampling frequency.

[1] Although the response is not truly distortionless, we follow the definition of the time-domain MVDR beamformer of Benesty et al [10].

3.2. Perceptually motivated Rake Beamformer

Psychoacoustics studies show that early echoes contribute to perceived power, and speech intelligibility. Lochner and Burger [11] describe precisely how much reverberation is perceptually beneficial. As determined for speech signals, echoes arriving within 30 ms of the direct sound are fully integrated, while those arriving within 95 ms are still partially integrated. Echoes arriving later than 35 ms are noticeable.

In regard of these results, we can partially relax the distortionless requirement. We define the perceptually motivated rake beamformer with the following four criteria.

- Minimize the interference and noise power.
- Zero response before τ (i.e. no pre-echoes).
- Unit response at τ.
- Zero response after $\tau + \kappa$, where $\kappa \sim 35$ ms.

The optimal such beamformer is found by the quadratic program,

$$\underset{\boldsymbol{g}}{\text{minimize}}\ \boldsymbol{g}^T\boldsymbol{K}_{nq}\boldsymbol{g} \ \text{ subject to } \boldsymbol{g}^T\hat{\boldsymbol{H}}_s = \boldsymbol{\delta}_\tau^T,$$

where $\boldsymbol{K}_{nq} = \boldsymbol{H}_q\boldsymbol{R}_{zz}\boldsymbol{H}_q^T + \boldsymbol{R}_{bb}$, the matrix $\hat{\boldsymbol{H}}_s$ contains the columns 1 to τ and $\kappa + 1$ to L of $\boldsymbol{H}_s$, and $\boldsymbol{\delta}_\tau$ is the vector with a one at position τ and all other entries zero. Note that an alternative formulation including all zero forcing constraints directly in the objective exists. The solution of this program is

$$\boldsymbol{g}_{\text{R-P}} = \boldsymbol{K}_{nq}^{-1}\hat{\boldsymbol{H}}_s(\hat{\boldsymbol{H}}_s^T\boldsymbol{K}_{nq}^{-1}\hat{\boldsymbol{H}}_s)^{-1}\boldsymbol{\delta}_\tau.$$

A similar criterion as for Rake MVDR beamformer applies as to which image sources can be used constructively. Thanking to the relaxation, image sources up to distance $\|\boldsymbol{s}_0 - \boldsymbol{r}_m\| + c(\tau + \kappa)/F_s$ can be included in the optimization.

3.3. Maximum SINR Rake Beamformer

The signal to interference and noise ratio (SINR) is defined as

$$\mathsf{SINR} = \frac{\mathbb{E}|\boldsymbol{g}^T\boldsymbol{H}_s\boldsymbol{x}|^2}{\mathbb{E}|\boldsymbol{g}^T(\boldsymbol{H}_q\boldsymbol{z} + \boldsymbol{b})|^2} = \frac{\boldsymbol{g}^T\boldsymbol{K}_x\boldsymbol{g}}{\boldsymbol{g}^T\boldsymbol{K}_{nq}\boldsymbol{g}}, \tag{3}$$

where $\boldsymbol{K}_x = \boldsymbol{H}_s\boldsymbol{R}_{xx}\boldsymbol{H}_s^T$. This quantity can be optimized directly by solving the generalized eigenvalue problem $\boldsymbol{K}_x\boldsymbol{g} = \lambda\boldsymbol{K}_{nq}\boldsymbol{g}$, and the maximizer is given by the generalized eigenvector corresponding to the largest generalized eigenvalue. This will however not yield a practical beamformer. Because no constraint is imposed on the response towards the desired source, its signal can be arbitrarily distorted. Nevertheless, this gives an upper bound on achievable SINR.

4. NUMERICAL EXPERIMENTS

In this section, we assess the performance of the three rake beamformers described. First, we inspect the beampatterns obtained. Then, the gain of using additional sources is evaluated in terms of output SINR. We use the same simulation setup as in [7]. For sound propagation simulation we use up to 10th order

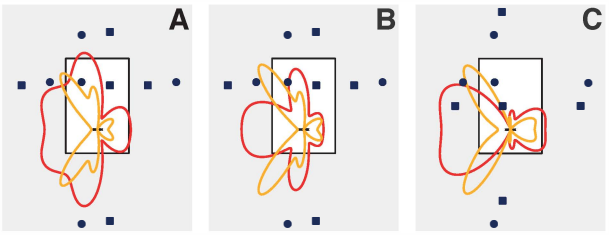

Fig. 2. Beampatterns of (A) Rake MVDR, and (B), (C) Rake Perceptual, in a 4×6 m room containing the desired source (●) and an interferer (■). In (C), the interferer is in the direct path of the desired source. First order image sources are also displayed. The darker/red and light/yellow lines are for 800 Hz and 1600 Hz, respectively.

reflections (220 image sources). The sampling frequency is 8 kHz. Samples from both sources are assumed to be zero-mean independent and identically distributed and the noise is AWGN so that

$$\boldsymbol{R}_{xx} = \sigma_x^2 \boldsymbol{I}, \quad \boldsymbol{R}_{zz} = \sigma_z^2 \boldsymbol{I}, \quad \boldsymbol{R}_{bb} = \sigma_n^2 \boldsymbol{I},$$

where $\boldsymbol{I}$ is the identity matrix and $\sigma_x^2 = \sigma_z^2 = 1$.

4.1. Beampatterns

We consider a 4 by 6 m room with a source of interest at (1, 4.5) and a linear array of eight microphones equally spaced by 8 cm, parallel to the x-axis and centered at (2,1.5), the origin being the lower left corner of the room. The beamforming filters length is 50 ms ($L_g = 400$ at 8 kHz) with a delay of 20 ms. The noise variance at the microphones is fixed at $\sigma_n^2 = 10^{-7}$. Beampatterns for both Rake MVDR and Rake Perceptual with an interferer placed at (2.8, 4.3) are shown for 800 Hz and 1600 Hz in Fig. 2. The diagram in the figure shows the beampatterns for Rake Perceptual when the interferer is placed in the direct path of the desired source at (1.5,3). We observe that in that case, the beamformer completely ignores the direct sound and focuses on the reflections. Such a scenario could not be handled by a beamformer only considering the direct sound.

4.2. SINR gain from raking

The SINR gain from raking is investigated through Monte-Carlo simulation. We consider the same room and beamforming filters length as in Section 4.1, but pick source and interferer positions uniformly at random. At each run, the SINR according to (3) is computed for Rake MVDR, Rake Perceptual, and Rake MaxSINR. Even though Rake MaxSINR is not practical, it gives an upper bound on the SINR gain that can be expected. The same number of image sources $K = K' = 0, \ldots, 9$ is used for the source and the interferer. The noise variance is fixed so that the SNR of the direct path of the desired source is 10 dB at the center of the array or $\sigma_n^2 = 10^{-1}(4\pi \|\boldsymbol{s}_0 - \bar{\boldsymbol{r}}\|)^{-2}$ where $\bar{\boldsymbol{r}} = M^{-1} \sum_{m=0}^{M-1} \boldsymbol{r}_m$ is the center of the array. The beamforming filters length is fixed to 30 ms (i.e. $L_g = 240$) and the delay is 20 ms.

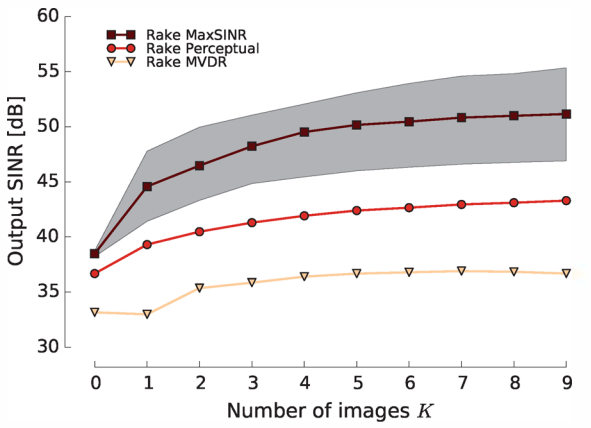

Fig. 3. Median output SINR computed according to (3) against the number of image sources K used in the optimization. The same number of image sources is used for the desired source and the interferer. The ambient noise SNR is fixed to 10 dB with respect to the direct path of the desired source and the center of the microphone array. The grey area contains 50% of the Rake MaxSINR outcomes.

The outcome of the simulation is depicted in Fig. 3. Each point is the result of 10000 outcomes. For every beamformer considered, adding more sources results in a net increase in SINR. Adding just the 1st order reflections, or 5 sources, rakes in 3.5 dB and 5.7 dB improvement in SINR for Rake MVDR and Rake Perceptual, respectively. Rake MaxSINR shows that at most 11 dB improvement can be expected. We also observe that the extra degrees of freedom of Rake Perceptual are very beneficial as it is consistently 4 to 5 dB above Rake MVDR when image sources are used.

5. CONCLUSION

Drawing inspiration from the rake receiver we developed time-domain beamforming designs exploiting temporal and spatial diversity of an acoustic signal in a reverberant environment. We proposed two beamformers, one based on the classic MVDR beamformer and another perceptually motivated with relaxed constraints on the beamformer response. We show in numerical experiments that even short filters are enough to suppress an interferer, even when it is in the direct path of the desired source. Through Monte-Carlo simulation, we show that raking signal from more sources results in a net increase of the SINR for all designs proposed, the perceptually motivated design beating the distortionless design by around 5 decibels.

Although the Rake Perceptual beamformer seems to perform well, it only minimizes the power from the interferer. In further work, we would also like to maximize the desired source power in a perceptually relevant manner. Another goal is to investigate in more details the relationship between filter length, delay, and performance. A crucial step will be to validate the designs experimentally.

6. REFERENCES

[1] R. Price and P. E. Green, "A Communication Technique for Multipath Channels," in *Proceedings of the IRE*. 1958, pp. 555–570, IEEE.

[2] B. H. Khalaj, A. Paulraj, and T. Kailath, "2D RAKE Receivers for CDMA Cellular Systems," in *Proc. IEEE GLOBECOM*. 1994, pp. 400–404, IEEE.

[3] A. F. Naguib, "Space-time receivers for CDMA multipath signals," in *Proc. IEEE ICC*, Montreal, 1997, pp. 304–308, IEEE.

[4] A. E. O'Donovan, R. Duraiswami, and D. N. Zotkin, "Automatic Matched Filter Recovery via the Audio Camera," in *Proc. IEEE ICASSP*, Dallas, 2010, pp. 2826–2829, IEEE.

[5] P. Annibale, F. Antonacci, P. Bestagini, A. Brutti, A. Canclini, L. Cristoforetti, J. Filos, E. Habets, W. Kellerman, K. Kowalczyk, A. Lombard, E. Mabande, D. Markovic, P. Naylor, and M. Omologo, "The SCENIC Project: Space-Time Audio Processing for Environment-Aware Acoustic Sensing and Rendering," in *131st Convention of the Audio Engineering Society*, New York, NY, USA, 2011, Audio Engineering Society.

[6] E.-E. Jan, P. Svaizer, and J. L. Flanagan, "Matched-Filter Processing of Microphone Array for Spatial Volume Selectivity," *Proc. IEEE ISCAS*, vol. 2, pp. 1460–1463, 1995.

[7] I. Dokmanić, R. Scheibler, and M. Vetterli, "Raking the Cocktail Party," *arXiv.org*, July 2014.

[8] M. Thomas, F. Lim, I. J. Tashev, and P. A. Naylor, "Optimal Beamforming as a Time Domain Equalization Problem with Application to Room Acoustics," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

[9] M. Miyoshi and Y. Kaneda, "Inverse Filtering of Room Acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, 1988.

[10] J. Benesty, J. Chen, Y. A. Huang, and J. Dmochowski, "On Microphone-Array Beamforming From a MIMO Acoustic Signal Processing Perspective," *IEEE Trans., Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1053–1065, Mar. 2007.

[11] J. Lochner and J. F. Burger, "The Influence of Reflections on Auditorium Acoustics," *J. Sound Vib.*, vol. 1, no. 4, pp. 426–454, 1964.

[12] W. Zhang, E. Habets, and P. A. Naylor, "On the Use of Channel Shortening in Multichannel Acoustic System Equalization," in *Proc. IWAENC*, Tel Aviv, 2010.

[13] M. R. P. Thomas, N. D. Gaubitch, and P. A. Naylor, "Application of Channel Shortening to Acoustic Channel Equalization in the Presence of Noise and Estimation Error," in *Proc. IEEE WASPAA*, New Paltz, NY, 2011, pp. 113–116, IEEE.

[14] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proc. Natl. Acad. Sci.*, vol. 110, no. 30, June 2013.

[15] D. B. Ward, E. A. Lehmann, R. C. S. Williamson, and A. P. I. T. on, "Particle Filtering Algorithms for Tracking an Acoustic Source in a Reverberant Environment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 11, no. 6, pp. 826–836, 2003.

[16] F. Ribeiro, D. A. Florencio, D. E. Ba, and C. Zhang, "Geometrically Constrained Room Modeling With Compact Microphone Arrays," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 20, no. 5, pp. 1449–1460, 2012.

[17] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of Room Geometry From Acoustic Impulse Responses," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 20, no. 10, pp. 2683–2695, 2012.

[18] I. Dokmanic, Y. M. Lu, and M. Vetterli, "Can One Hear the Shape of a Room: The 2-D Polygonal Case," in *Proc. IEEE ICASSP*, Prague, 2011, pp. 321–324.

[19] J. B. Allen and D. A. Berkley, "Image Method For Efficiently Simulating Small-room Acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[20] J. Borish, "Extension of the Image Model To Arbitrary Polyhedra," *J. Acoust. Soc. Am.*, vol. 75, no. 6, pp. 1827–1836, 1984.

[21] J. Capon, "High-Resolution Frequency-Wavenumber Spectrum Analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.