Raking the Cocktail Party

Ivan Dokmanić, Student Member, IEEE, Robin Scheibler, Student Member, IEEE, and Martin Vetterli, Fellow, IEEE

Abstract—We present the concept of an acoustic rake receiver (ARR)—a microphone beamformer that uses echoes to improve the noise and interference suppression. The rake idea is well-known in wireless communications. It involves constructively combining different multipath components that arrive at the receiver antennas. Unlike typical spread-spectrum signals used in wireless communications, speech signals are not orthogonal to their shifts, which makes acoustic raking a more challenging problem. That is why the correct way to think about it is spatial. Instead of explicitly estimating the channel, we create correspondences between early echoes in time and image sources in space. These multiple sources of the desired and interfering signals offer additional spatial diversity that we can exploit in the beamformer design.

We present several “intuitive” and optimal formulations of ARRs, and show theoretically and numerically that the rake formulation of the maximum signal-to-interference-and-noise beamformer offers significant performance boosts in terms of noise suppression and interference cancellation. We accompany the paper by the complete simulation and processing chain written in Python. The code and the sound samples are available online at http://lcav.epfl.ch/people/ivan.dokmanic.

Index Terms—Room impulse response, beamforming, echo sorting, acoustic rake receiver

I. INTRODUCTION

RAKE receivers are habitually used in wireless communications to take advantage of multipath propagation, instead of trying to mitigate it. The original scheme was designed for single-input-single-output systems [1], and it was later extended to arrays of antennas [2], [3] that exploit spatial diversity. By using multiple antennas and spatial processing, multipath components that would otherwise not be resolvable because they arrive at similar times, become resolvable because they arrive from different directions. The basic idea is to coherently add all the multipath components, and thus increase the effective signal-to-noise ratio (SNR) and the probability of correct symbol detection.

Considering the success of rake receivers in wireless communications, it is somewhat remarkable that the principle has not received more attention in room acoustics. Nevertheless, constructive use of echoes in rooms to improve beamforming has been mentioned [4]–[6]. In particular, the term acoustic rake receiver has been used in the SCENIC project proposal [4]. The list of ingredients is the same as in wireless: We have a wave (acoustic instead of electromagnetic) propagating in space. Reflections and scattering cause the wave to arrive at the receiver through multiple paths in addition to the direct path, and each of the multipath components contains essentially the same information as any other.

Things become clearer when we look at the typical parameter values in the two scenarios. In CDMA, the delay spread—time between the first and last echoes arriving at the receiver—is typically around 10 μs [7]. The signal is designed so that it is near-orthogonal to its time-shifted copies, which makes for cleaner signal processing, in particular streamlining the multipath channel estimation. On the other hand, in room acoustics, typical times between consecutive echoes are in milliseconds, but the archetypal signal of interest is speech. Speech segments are always several orders of magnitude longer than the delay spread of the channel, and speech lacks the orthogonality property, so constructing a rake receiver for speech in analogy with wireless communications is not possible. Even if the two problems are parametrized by the same set of parameters, acoustic and wireless raking behave distinctly.

It is well established that exploiting the temporal structure of single-channel speech convolved with a room impulse response (RIR) is difficult. This is epitomized in the non-invertibility of RIRs [8], [9]. On the contrary, there are no significant differences from wireless as far as the spatial structure goes. If we know where the echoes are coming from, we can try to design spatial processing algorithms, such as beamformers, that use multiple copies of the same signal arriving from different directions.

In order to do so, we need to localize the echoes. Suppose first that we know the room geometry. Then it suffices to localize the direct path signal, and we can find the image sources using straightforward geometric rules [10], [11]. Localizing the direct signal in a reverberant environment is a well-understood problem [12]. The room geometry does not have to be known in great detail for this to work. Locations of the most important reflectors (ceiling, floor, walls) are enough. In many cases this knowledge is readily available through floor plans or measurements, and we can take advantage of it when deploying the system. Sometimes, mostly in ad-hoc deployments, the room geometry may be difficult to obtain, in which case we could perform a simple calibration step in order to learn it. This step can be performed acoustically, as was demonstrated recently [13]–[16]. Finally, assume that the room geometry estimation is not possible due to various non-idealities that are challenging to model. To accommodate for this scenario, note that we are not after the room geometry itself. Rather, we only need to know where the echoes are coming from, or equivalently, where the image sources are. Image source localization by a microphone array requires the disambiguation of the echo ordering, and the removal of the falsely detected echoes. This problem can be solved by echo sorting as described in [16]. As an alternative, O’Donovan and coauthors [5] propose to use an audio camera comprising...
a large number of microphones to find the images. Once the image sources are localized (in a calibration phase or otherwise), we can predict their movement using geometrical rules, as detailed in Section V.

In this paper, we present the concept of acoustic rake receivers (ARR). We caution the reader that it is somewhat of a misnomer, as we primarily exploit the spatial structure of the multipath, unlike the original work on rake receivers (and the origin of the noun adjunct rake\(^1\)) that was exploiting the temporal structure. They achieve this suppression by listening to the echoes (image sources) in addition to, or instead of the real desired source. Listening to echoes is further useful when a directive source is not facing the microphone array. This is typical of human talkers [17], [18], as speech production is a directive phenomenon.

The motivation for ARRs borrows from perception. An important perceptual insight is that the early echoes improve speech intelligibility [19], [20]. In fact, adding energy in the form of early echoes (approximately within the first 50 ms of the RIR) is equivalent to adding the same energy in the direct sound [19]. This is akin to deciding on how we assess the performance, and what we choose as a reference signal. A detailed discussion of this topic is given by Habets and coauthors [21], notably, by examining the tradeoff between dereverberation and denoising in beamforming. There is a certain similarity between ARRs and channel shortening, where the idea is to not try to equalize the channel to a Dirac, but rather to just shorten it. Channel shortening results in solvable systems and filters that are much better behaved [22], [23] than complete inversion, e.g. by the multiple-input-output-theorem (MINT) [24], [25].

Unlike in channel shortening and approaches that assume the knowledge of the acoustic channels from the sources to the microphones [24], [26], we never attempt to estimate the corresponding impulse responses. This estimation is generally a difficult problem. Our task is simpler—we only need a couple of early echoes, lifted to 3D space as image sources.

A. Main Contributions

We define the acoustic rake receiver (ARR) as the echo-aware microphone beamformer. We present several formulations with different properties, and analyze their behavior both theoretically and numerically. The analysis shows that ARRs lead to significantly improved SNR and interference cancellation when compared with standard beamformers that only extract the direct path. ARRs can suppress interference in cases when conventional beamforming is bound to fail, for example when an interferer is occluding the desired source (for a sneak-peak fast forward to Fig. 5). In contrast with earlier works, we explicitly focus on image sources. Moreover, we present optimal formulations that significantly outperform earlier delay-and-sum (DS) approaches, especially when interferers are present. This improved ability to suppress interference by leveraging echoes of both the desired source and the interferers suggests that the proposed methods may help to address the cocktail party problem [27].

We design and apply the beamformers in the frequency domain. While certainly time-domain designs [28] are important, and even essential in some cases, conciseness of the frequency domain approach enables us to focus on the objective gains from acoustic raking. For clarity, the numerical experiments are presented in a 2D “room”, and as such are directly applicable to planar (e.g. linear, circular) arrays. Algorithmic extension to 3D is straightforward.

Let us also mention some limitations of our results. We do not treat robust formulations that deal with uncertainties in the array manifold and source locations. Microphones are assumed to be ideally omni-directional with a flat frequency response. Except for Section V, we assume that the locations of the image sources are known. We explain how to find the image sources when the room geometry is either known or unknown, but we do not provide a deep overview of geometry estimation techniques. To this end, we suggest a number of references for the interested reader. We consider the walls to be flat-fading. In reality, they are frequency selective. We do not discuss the estimation of various covariance matrices [29].

The results in this paper are reproducible. Python (NumPy) [30] code for the beamforming routines, for the STFT processing engine, and to generate the figures is available online at http://lcav.epfl.ch/people/ivan.dokmanic. The link also contains sound samples.

B. Paper Outline

In Section II we explain the notation and the signal model used in the paper. A brief overview of relevant beamforming techniques and performance analysis is given in Section III. We formulate the acoustic rake receiver in Section IV, and we present a theoretical and numerical analysis of the corresponding beamformers. Section V explains how to locate the image sources, and comments on localizing the direct source. Numerical experiments are presented in Section VI.

II. NOTATION AND SIGNAL MODEL

We denote all matrices by bold uppercase letters, for example \( \mathbf{A} \), and all vectors by bold lowercase vectors, for example \( \mathbf{x} \). The Hermitian transpose of a matrix or a vector is denoted by \((\cdot)^H\), for example \( \mathbf{A}^H \), and the Euclidean norm of a vector by \( \| \cdot \| \), as in \( \| x \| \overset{\text{def}}{=} (x^H x)^{1/2} \).

Suppose that in a room, there is a desired source of sound located at \( s_0 \). Sound from this source arrives at the microphones located at \( \{ r_m \}_{m=1}^M \) via the direct path, but also through echoes from the walls. We model echoes, or the multipath propagation, by the image source model [10], [11]. Image sources are simply the mirror images of the real sources across the corresponding walls. A more thorough discussion of the image source model is given in Section V.

Denote the signal emitted by the source \( \tilde{x}[n] \) (e.g. the speech signal). Then all the image sources emit \( \tilde{x}[n] \) 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

\(^1\)A rake is a broom for outside use; a horticultural implement consisting of a toothed bar fixed transversely to a handle, and used to collect leaves, hay, grass, etc. [Wikipedia].
to image sources. We denote the image source positions by \( s_k \), \( 1 ≤ k ≤ K \), 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 \( q_0 \). For simplicity, we consider only a single interferer, but in general there could be any number of them. The interferer emits the signal \( \hat{z}[n] \), and its image source emit \( \hat{z}[n] \) as well. Similarly as for the desired source, \( q_k \), \( 1 ≤ k ≤ K' \) denote the positions of interfering image sources, where \( K' \) is the largest number of interfering image source considered.

The signal received by the \( m \)th microphone is then a sum of convolutions

\[
y_m[n] = \sum_{k=0}^{K} (\hat{a}_m(s_k) * \hat{x})[n] + \sum_{k=0}^{K'} (\hat{a}_m(q_k) * \hat{z})[n] + \hat{n}_m[n],
\]

where \( \hat{a}_m(s_k) \) denotes the impulse response of the channel between the source located at \( s_k \) and the \( m \)th microphone—
in this case a delay and scaling.

We design and analyze all the beamformers in the frequency domain. That is, we will be working with the DTFT of the discrete-time signal \( \hat{x} \),

\[
x(e^{jω}) \overset{\text{def}}{=} \sum_{n \in \mathbb{Z}} \hat{x}[n] e^{-jωn}. \tag{2}
\]

In practical implementations, we use the discrete-time short-time Fourier transform (STFT). More implementation details are given in Section VI.

With the established notation, the signal picked up by the \( m \)th microphone is given as

\[
y_m(e^{jω}) = \sum_{k=0}^{K} a_m(s_k, \Omega)x(e^{jω}) + \sum_{k=0}^{K'} a_m(q_k, \Omega)z(e^{jω}) + n_m(e^{jω}), \tag{3}
\]

where \( n_m(e^{jω}) \) contains all unmodeled phenomena and noise, and \( a_m(s_k, \Omega) \) denotes the \( m \)th component of the steering vector corresponding to the source \( s_k \). The steering vector is the Fourier transform of the continuous version of the impulse response \( \hat{a}(s_k) \), evaluated at \( \Omega \). Continuous-time domain frequency is denoted by \( \omega \), while \( \omega = \Omega T_s \) denotes the discrete-time domain frequency. The steering vector is then simply \( a(s_k) = [a_m(s_k)]_{m=0}^{M-1} \).

We can write out the entries of the steering vectors explicitly for a point source in free field. They are given as the appropriately scaled Green’s function for a Helmholtz equation at the frequency \( \Omega \) [31],

\[
a_m(s_k, \Omega) = \frac{α_k}{4π∥r_m - s_k∥} e^{-jκ∥r_m - s_k∥}, \tag{4}
\]

where we define the wavenumber \( κ \overset{\text{def}}{=} \Omega/c \), and \( α_k \) is the flat-fading coefficient corresponding to \( s_k \). The far field steering vector does not depend on the distance of the source from the microphones. The underlying assumption is that the source is far enough so that the impinging wavefronts are planar,

\[
π_m(s_k, \Omega) = e^{-jκ∥r_m, s_k∥/∥s_k∥}. \tag{5}
\]

To compute (5), we place the origin of the coordinate system at the center of the microphone array \( \frac{1}{M} \sum_{m=1}^{M-1} r_m \).

The microphone signals can be written jointly in a vector form as

\[
y(e^{jω}) = \sum_{k=0}^{K} a(s_k, \Omega)x(e^{jω}) + \sum_{k=0}^{K'} a(q_k, \Omega)z(e^{jω}) + n(e^{jω}), \tag{6}
\]

where \( A_s(e^{jω}), A_q(e^{jω}) \) are matrices whose columns are \( a(s_k, \Omega), a(q_k, \Omega) \), respectively, and \( 1 \) is the all-one vector of length \( K \). Depending on the focus, we either make explicit, or omit the interference term, absorbing it in \( n(e^{jω}) \). From here onward, we suppress the frequency dependency of the steering

![Fig. 1. Illustration of notation and concepts.](image-url)
vectors and the beamforming weights to reduce the notational clutter. Wherever essential, we make it explicit again.

III. BEAMFORMING PRELIMINARIES

Microphone beamformers combine the outputs of multiple microphones in order to achieve spatial selectivity, or more generally, to suppress noise and interference [32]. At a given frequency, beamforming is achieved by taking a linear combination of the microphone outputs. We compute the output of a beamformer as a linear combination of microphone outputs at a given frequency,

\[ u = w^H y = w^H A_s 1 x + w^H A_q 1 z + w^H n, \]

where the vector \( w \in \mathbb{C}^M \) contains the beamforming weights.

The weights \( w \) are selected so that they optimize some design criterion. Common examples are the minimum-variance-distortionless-response (MVDR) beamformer, maximum-signal-to-interference-and-noise (Max-SINR) beamformer, and minimum-mean-squared-error (MMSE) beamformer.

A. Delay-and-Sum Beamformer

The simplest and often quite effective beamformer is the delay-and-sum beamformer [32]. It compensates the propagation delays from the source to the microphones. Assume that we want to listen to a source at \( s \). We have

\[ w_{DS} = a(s) / \| a(s) \|, \]

so that

\[ u_{DS} = w_{DS}^H y \approx \frac{1}{M} \sum_{m=0}^{M-1} y_m e^{-j\pi \| r_m - s \|} \approx \frac{x}{4\pi \| r - s \|} + \frac{1}{M} \sum_{m=0}^{M-1} n_m, \]

where we used the definition of \( y_m \) (6), the definition of the steering vector (4) with the far field approximation, i.e. \( a_m(s, \Omega) \approx (4\pi \| r - s \|)^{-1} e^{-j\pi \| r_m - s \|}, \) \( r \) being the center of the array. If \( n \sim \mathcal{N}(0, \sigma^2 I_M) \), then the output noise is distributed according to \( \mathcal{N}(0, \sigma^2 / M) \), that is, we obtain an \( M \)-fold decrease in the noise variance at the output with respect to any reference microphone.

B. Minimum Variance Distortionless Response

The MVDR beamformer [33] minimizes the total power of the output signal, with a constraint that the output is unity for the desired source. Therefore, the power is minimized by reducing the sidelobes, effectively reducing the influence of noise and interference. The weights are computed by solving the following optimization problem,

\[
\begin{align*}
\text{minimize} & & w^H K w \\
\text{subject to} & & w^H a_s = 1,
\end{align*}
\]

where \( K \) denotes the covariance matrix of the total signal (desired signal + error). The solution is readily obtained as

\[ w_{MVDR} = \frac{K^{-1} a_s}{a_s^H K^{-1} a_s}, \]

by using Lagrange multipliers.

MVDR beamformers are known to be particularly sensitive to mismatches in the knowledge of the array manifold, i.e. to calibration errors [34]. Estimating the covariance matrix is also a potential problem. Wrong estimates often lead to cancellation of the desired signal.

C. Signal-to-Interference-and-Noise Ratio

To assess the performance of ARR and compare it against standard non-raking beamformers, we use the SINR. In particular, we compare Max-SINR beamformers that are optimal in terms of this metric.

The signal-to-interference-and-noise ratio (SINR) is generally given as

\[ \text{SINR} \equiv \frac{\text{power of the desired output signal}}{\text{power of undesired output signal}}. \]

The desired output signal is the output signal due to the desired source, while the undesired signal is the output signal due to interferers and noise. For a desired source at \( s \), and an interfering source at \( q \), we can write

\[ \text{SINR} \equiv \frac{\mathbb{E} | w^H a(s) x |^2}{\mathbb{E} | w^H (a(q) z + n) |^2} = \sigma_s^2 w^H a(s) a(s)^H w / w^H K_{nq} w, \]

where \( K_{nq} \) is the covariance matrix of the noise and the interference.

D. Maximum Signal-to-Interference-and-Noise Ratio Beamformer

It is compelling to pick \( w \) that maximizes the SINR (13) [32]. The maximization can be solved by noting that rescaling the beamformer weights leaves the SINR unchanged, and setting the numerator to any constant. The latter variant leads to exactly the same type of problem as (10), but with only the noise covariance matrix \( K_{nq} \) instead of the full covariance matrix \( K \). Thus the solution is immediately seen to be

\[ w_{\text{SINR}} = \frac{K_{nq}^{-1/2} a_s}{a_s^H K_{nq}^{-1/2} a_s}. \]

Interestingly, it can be shown that \( w_{\text{MVDR}} = \gamma w_{\text{SINR}} \) for some scalar constant \( \gamma \): The two beamformers are equivalent up to a scaling of the weights [32]. Note that the scaling factor is frequency dependent and thus the equivalence only holds for in narrow-band sense.

Using the definition (13) we can derive for the Max-SINR beamformer

\[ \text{SINR} = \sigma_s^2 a_s^H K_{nq}^{-1} a_s. \]

Because \( K_{nq}^{-1} \) is a positive definite matrix, it has an eigenvalue decomposition as \( K_{nq}^{-1} = U^H \Lambda U \), where \( U \) is unitary, and \( \Lambda \) is diagonal with positive entries. We can write \( a_s^H K_{nq}^{-1} a = (U a)^H \Lambda (U a) \). Since \( \| U a \|^2 = \| a \|^2 \), and because \( \Lambda \) is positive, increasing \( \| a \|^2 \) typically leads to an increased SINR, although we can construct counterexamples.
E. Useful-to-Detrimental Sound Ratio

The next measure, useful-to-detrimental sound ratio (UDR), is perceptually motivated [19], [20]. It expresses the fact that adding early reflections (up to 50 ms in the RIR) is as good as adding the energy to the direct sound as far as the speech intelligibility goes. The useful signal is then a coherent sum of direct and early reflected speech energy, so that

$$\text{UDR} = \frac{\mathbb{E} \sum_{k=0}^{K} |w^H a(s_k, \omega) x_k|^2}{\mathbb{E} |w^H n|^2},$$

(16)

In applications $K$ is rarely large enough to cover all the reflections occurring within 50 ms, simply because it is too optimistic to assume we know all the corresponding image sources. Therefore, (16) typically underestimates the UDR. Alas, because (16) is specified in the frequency domain, it is difficult to control whether the reflections in the numerator arrive before or after the direct sound.

IV. ACOUSTIC RAKE RECEIVER

In the acoustic rake receiver (ARR), we aim to constructively use the echoes, instead of considering them to be detrimental. In this section we present four ARR designs ideas that intuitively make sense.

A. Delay-and-Sum Raking

If we had access to every individual echo separately, we could align them all to achieve very sizable performance improvements. Unfortunately this is not the case: Each microphone picks up the convolution of speech with the impulse response, which is effectively a sum of echoes. If we only wanted to extract the direct path, we would use the standard DS beamformer (8). To build a Rake DS receiver, we create a DS beamformer for every image source, and average the outputs,

$$\frac{1}{K+1} \sum_{k=0}^{K} \alpha_k' \frac{1}{M} \sum_{m=0}^{M-1} y_m e^{j k m} ||r_m - s_k||,$$

(17)

where $\alpha_k' \equiv \alpha_k/(4\pi ||r_m - s_k||)$. Note that we can rescale the weights arbitrarily, without altering the performance in terms of SINR. We read out the beamforming weights as

$$w_{R-DS} = \frac{1}{\sum_{k=0}^{K} a(s_k)} \sum_{k=0}^{K} a(s_k) = \frac{A_s 1}{\|A_s 1\|},$$

(18)

where we chose the scaling in analogy with (8). We see that this is just a sum of the steering vectors for each image source, with the appropriate scaling.

B. One-Forcing Raking

An alternative criterion is to design a beamformer that listens to all $K$ image sources with the same power, and at the same time minimizes the noise and interference energy.

$$\text{minimize} \mathbb{E} \left| \sum_{k=0}^{K'} w^H a(q_k) z + w^H n \right|^2$$

subject to $w^H a(s_k) = 1, \forall \ 0 \leq k \leq K$.

(19)

Alternatively, we may choose to null the interfering source and its image sources. This is an instance of the standard linearly-constrained-minimum-variance (LCMV) beamformer [35]. Collecting all the steering vectors in a matrix, we can write the constraint as $w^H A_s = 1^T$. The solution can be found in closed form as

$$w_{OF} = K_{nq}^{-1} A_s (A_s^H K_{nq}^{-1} A_s)^{-1} 1_M.$$  

(20)

A few remarks are in order. First, with $M$ microphones, $K$ can be at most $M$, as otherwise we end up with more constraints than degrees of freedom. Second, using this beamformer is a bad idea if there is an interferer along the ray through the microphone array and any of image sources.

As with all LCMV beamformers, adding linear constraints uses up degrees of freedom that could be used for noise and interference suppression. As we demonstrate in the following section, it is better to let the “beamformer decide” or “the beamforming procedure decide” on how to maximize a well-chosen cost function.

C. Max-SINR Raking

The main workhorse of the paper is the Rake-Max-SINR.

$$\text{maximize} \mathbb{E} \left| \sum_{k=0}^{K} w^H a(s_k) x_k \right|^2$$

$$\mathbb{E} \left| \sum_{k=0}^{K} w^H a(q_k) z + w^H n \right|^2.$$

(21)

The logic behind this expression is as follows: We present the beamforming procedure with a set of good sources, whose influence we aim to maximize at the output, and with a set of bad sources, whose power we try to minimize at the output. Interestingly, this leads to the standard Max-SINR beamformer with a structured steering vector and covariance matrix. Define $A_s \equiv [a(s_0), \ldots, a(s_K)]$, and

$$K_{nq} \equiv K_n + \sigma_z^2 \left( \sum_{k=0}^{K'} a(q_k) \right) \left( \sum_{k=0}^{K'} a(q_k) \right)^H,$$

(22)

where $K_n$ is the covariance matrix of the noise term, and $\sigma_z^2$ is the power of the interferer at a particular frequency.

Then the solution to (21) is given as

$$w_{R-SINR} = \frac{K_{nq}^{-1} A_s 1}{1^H A_s^H K_{nq}^{-1} A_s 1}.$$  

(23)

It is interesting to note that when $K_{nq} = \sigma_z^2 I_M$ (e.g. no interferers and iid noise), the Rake-Max-SINR beamformer reduces to $A_s 1/\|A_s 1\|$, which is exactly the Rake-DS beamformer.
TABLE II
SUMMARY OF BEAMFORMERS.

| Acronym    | Description                                                                 | Beamforming Weights                                                                 |
|------------|-----------------------------------------------------------------------------|------------------------------------------------------------------------------------|
| DS         | Align delayed copies of signal at the microphone                            | \( w_{DS} = a(s) / ||a(s)|| \)                                                   |
| MVDR       | \( \text{min. } w^H K w \), s.t. \( w^H a_s = 1 \)                         | \( w_{MVDR} = K^{-1} a_s / (a_s^H K^{-1} a_s) \)                                 |
| Max-SINR   | \( \text{max. } w^H A_s a_s^H w / (w^H K_{nq} w) \)                        | \( w_{SINR} = K_{nq}^{-1} a_s / (a_s^H K_{nq}^{-1} a_s) \)                      |
| Rake-DS    | Weighted average of DS beamformers over image sources                        | \( w_{R-DS} = A_s / ||A_s|| \)                                                   |
| Rake-OF    | \( \text{min. } E \sum_{k=0}^{K} w^H a(q_k) z + w^H n \)                    | \( w_{OF} = K_{nq}^{-1} A_s (A_s^H K_{nq}^{-1} A_s)^{-1} I_M \)                 |
| Rake-Max-SINR | \( \text{max. } E \sum_{k=0}^{K} |w^H a(q_k)|^2 / E \sum_{k=0}^{K} |w^H a(q_k)|^2 \) | \( w_{R-SINR} = K_{nq}^{-1} A_s / (1 + w^H A_s^H K_{nq}^{-1} A_s) \)          |
| Rake-Max-UDR | \( \text{max. } E \sum_{k=0}^{K} |w^H a(q_k)|^2 / E w^H \sum_{k=0}^{K} a(q_k) z + w^H n \) | \( w_{R-UDR} = C^{-1} \tilde{w}_{max} ((C^{-1})^H A_s A_s^H C^{-1}) \) |

D. Max-UDR Raking

Finally, it is interesting to investigate what happens if we choose the weights that optimize the UDR. Concretely, we solve the following optimization program.

\[
\text{maximize } \frac{E \sum_{k=0}^{K} |w^H a(q_k)|^2}{E \sum_{k=0}^{K} |w^H a(q_k)|^2}, \quad (24)
\]

Using the earlier results we see that this amounts to maximizing the following generalized Rayleigh quotient,

\[
\frac{w^H A_s A_s^H w}{w^H K_{nq} w}. \quad (25)
\]

Assuming that \( K_{nq} \) has a Cholesky decomposition as \( K_{nq} = C^H C \) we can write this quotient as

\[
\bar{w}^H (C^{-1})^H A_s A_s^H C^{-1} \bar{w}, \quad (26)
\]

\( \bar{w} \overset{\text{def}}{=} C w \). The maximum of this expression is

\[
\lambda_{\text{max}} ((C^{-1})^H A_s A_s^H C^{-1}), \quad (27)
\]

where \( \lambda_{\text{max}} (\cdot) \) denotes the largest eigenvalue of the argument matrix. The maximum is achieved by the corresponding eigenvector \( \bar{w}_{\text{max}} \). Then the optimal weights are given as

\[
w_{\text{R-UDR}} = C^{-1} \bar{w}_{\text{max}}. \quad (28)
\]

E. SINR Gain from Raking

Intuitively, if we have multiple sources of the desired signal scattered in space, and we account for it in the design, at least we cannot do worse than if we ignore the copies. Let us see how large the gain can be.

We have for the Rake-Max-SINR beamformer

\[
\text{SINR} = (A_s 1)^H K_{nq}^{-1} A_s 1. \quad (29)
\]

Equivalently, we ask whether \( a(s_k) \) sum coherently or they cancel out. This is answered by the following theorem. It is stated for a linear array, but the described behavior is universal.

**Theorem 1.** Assume that there are \( K + 1 \) image sources located at \( s_k = r_k [\cos \theta_k \sin \theta_k]^T \) where \( \theta_k \sim \mathcal{U}(0, 2\pi) \) and \( r_k \sim \mathcal{U}(a, b) \) are all independent, for some \( 0 < a < b \) such that the far-field assumption holds. Let \( A_s \) collect the corresponding steering vectors for a uniform linear microphone array. Then \( E ||A_s 1||^2 \geq (1 + \beta) E ||a(s_0)||^2 \), where \( \beta = \sum_{k=1}^{K} (\alpha_k / \alpha_0) \), and \( \alpha_k \) are attenuations of the steering vectors, assumed independent from the source locations. Furthermore \( E ||A_s 1||^2 = (1 + \beta) E ||a(s_0)||^2 + O(1 / \Omega^3) \).

A couple of remarks are in order:

1) This result is in expectation. In the worst case we can have cancellations. But the numerical experiments suggest that this is very rare in practice, and we can on the other hand observe large gains.

2) This means that for large enough frequencies (in fact not that large as we show below), summing the phasors behaves as a two-dimensional random walk. It is known that the RMS distance of a 2D random walk in \( n \) steps is \( \sqrt{n} [36] \).

3) Due to the far-field assumption, the attenuations \( \alpha_k \) are independent of the microphones. In reality they do depend on the source locations. However, they also depend on the number of additional factors, for example wall attenuations and radiation pattern of the sources. Therefore, for simplicity, we consider them independent from source locations. It is not difficult to verify that this assumption does not change the described trend.

**Proof:** Thanks to the far-field assumption, we can decompose the steering vector into a factor due to the array, and a phase factor due to different distances of different image sources. We have that

\[
a_m = (A_s 1)_m = \sum_{k=0}^{K} \alpha_k e^{-j \kappa m d \sin \theta_k} e^{-j \delta_k / c}, \quad (31)
\]

where \( d \) is the microphone spacing and \( \kappa \overset{\text{def}}{=} \Omega / c \). Without loss of generality we assume that \( \delta_k \sim \mathcal{U}(a, b) \). We can further...
Theorem 1 and Fig. 2 show that acoustic raking indeed improves the SNR. As we show later in the experiments section, the gains increase in the presence of interference.

V. FINDING AND TRACKING THE ECHOES

Thus far we assumed that the locations of the image sources are known. In this section we briefly describe some methods of obtaining the locations of image sources when they are a priori unknown. We assume that we can localize the direct path, or at least one image source. Coupled with the knowledge of the room geometry, this suffices to find the locations of other image sources [38].

A. Known room geometry

In many cases, for example for fixed deployments, the room geometry is known. This knowledge could be obtained at the time of the deployment, or from blueprints. In most indoor environments, we encounter a large number of planar reflectors. These reflectors correspond to image sources. With reference to Fig. 3, we can easily find the image source locations [10].

Suppose that the real source is located at \( s \). Then the image source with respect to wall \( i \) is computed as,

\[
\mathbf{i}_i(s) = s + 2 \left( \mathbf{p}_i - s, \mathbf{n}_i \right) \mathbf{n}_i,
\]

where \( i \) indexes the wall, \( \mathbf{n}_i \) is the outward normal associated with the \( i \)th wall, and \( \mathbf{p}_i \) is any point belonging to the \( i \)th wall. Analogously, we compute image sources corresponding to higher order reflections,

\[
\mathbf{i}_j(\mathbf{i}_i(s)) = \mathbf{i}_i(s) + 2 \left( \mathbf{p}_j - \mathbf{i}_i(s), \mathbf{n}_j \right) \mathbf{n}_j.
\]

The above expressions are valid regardless of the dimensionality, concretely in 2D and 3D.

B. Acoustic geometry estimation

When the room geometry is not known, it is possible to estimate it using the same array we use for beamforming. Recently a number of different methods appeared in the literature that propose to use acoustics to estimate the shape.
of a room. For example, in [13] the authors use a dictionary of wall impulse responses recorded with a particular array. In [14] the authors use tools from projective geometry together with the Hough transform to estimate the room geometry. In [16] the authors derive an echo sorting mechanism that finds the image sources, from which the room geometry is then derived. All these methods rely on a model where the room is bounded by planar reflectors.

C. Without Estimating the Room Geometry

In many scenarios the room geometry is difficult to estimate. This is where echo sorting can be particularly useful. Observe that we do not really need to know how the room looks like, at least not exactly—we only need to know where the major echoes are coming from in order to apply our ARR principle. If we locate the image sources once, and then devise a mechanism that tracks them, we have solved the problem. Again with reference to Fig. 3, we can state the following simple proposition.

Proposition 1. Suppose that the room has only right angles so that the walls are parallel with the coordinate axes. Let the source move from $s$ to $s + t$. Then any image source $s_k$, moves to a point given by

$$s_k + Tt,$$

where $T = \text{diag}(\pm 1, \mp 1)$ for odd generations, and $T = \pm I_2$ for even generations.

Proof: The proof follows directly from the figure. The displacement of the image source is the same as the displacement of the true source, passed through a series of reflections. Reflection matrices are diagonal matrices with $\pm 1$ on the diagonal, and determinant equal to $-1$, hence the result. ■

The usefulness of this proposition is that it gives us a tool to track image sources even when we do not know the room geometry (as long as it has right angles). A possible use scenario is to start with a calibration procedure with a controlled source, and perform the echo sorting to find multiple image sources. Then if possible, we assign to each image source a generation (this is in fact a by-product of echo sorting), or we try different hypotheses using Proposition 1, and choose the one that maximizes the output SINR.

VI. NUMERICAL EXPERIMENTS

In this section, we validate all the theoretical results described so far through numerical experiments. In particular, we will first analyze the beampatterns produced by the ARR, then simulate the propagation of sound from two sources and test the efficiency of the ARR in suppressing an interferer.

A. Simulation Setup

We use a simple room acoustic framework written in Python, relying on Numpy and Scipy for matrix computations [30]. We limit ourselves in this paper to 2D geometry and rectangular rooms. In all experiments the sampling frequency $F_s$ is 8 kHz. An overview of the simulation setup is shown in Fig. 4.

![Flowchart of the simulation setup used for numerical experiments.](image)

Starting from the room geometry and the positions of sources and microphones, we first compute all images sources up to ten generations. The RIR between source $s_0$ and microphone $r_m$ is convolved with an ideal low-pass filter in the continuous domain and then sampled at $F_s$.

$$\hat{a}_m(s_0)[n] = \sum_{k=0}^{K} \frac{\alpha_k}{4\pi ||r_m - s_k||} \text{sinc} \left( n - F_s \frac{||r_m - s_k||}{c} \right),$$

where $K$ is the number of image sources considered. We pick the limits of $n$ such that the sinc can decay sufficiently to avoid sharp discontinuities. Then, the signals (sampled at $F_s$) from all sound sources are convolved with their respective RIR and summed up to compose the $m$th microphone signal.

The beamforming weights are computed in the frequency domain. We use STFT processing with a frame size of $L = 4096$, 50% overlap and zero padding on both sides of the signal by $L/4$. A real fast Fourier transform of size $2L$ and a Hann window are used in the analysis. By exploiting the conjugate symmetry of the real FFT we only need to compute $L + 1$ beamforming weights, one for every positive frequency bin. The length $L$ is dictated by the length of the beamforming filters in the time-domain and was set empirically to avoid any aliasing of the filter responses. Synthesis of the output signal is performed using conventional overlap-add method [39].

B. Results

1) Beampatterns: We first inspect the beampatterns created by Rake-Max-SINR and Rake-Max-UDR in a number of scenarios. We consider a 4 by 6 meters rectangular room with a source of interest at point $(1, 4.5)$ and a linear microphone array centered at $(2, 1.5)$, parallel to the $x$-axis, and 8 cm spacing between microphones. In Fig. 5, the beampatterns in four different scenarios are displayed. We consider a scenario without interferer, one with an interferer placed relatively favorably at $(2.8, 4.3)$, and finally one where the interferer is placed half-way directly in front of the source of interest at $(1.5, 3)$. This last scenario is the least favorable and we observe that the Rake-Max-SINR beampattern completely ignores the direct path in order to concentrate on reflections of the source of interest, thus validating the intuition. Note that such a pattern cannot be achieved by a beamformer that only takes into account the direct path. We further note that while the beampatterns only show the magnitude, with multiple
Fig. 5. Beam patterns in different scenarios. The rectangular room is 4 by 6 metres and contains a source of interest (■) and an interferer (○) ((B), (C), (D) only). The first order image sources are also displayed. The weight computation of the beamformer includes the direct source and the first order image sources of both desired source and interferer (when applicable). (A) Rake-Max-SINR, no interferer, (B) Rake-Max-SINR, one interferer, (C) Rake-Max-UDR, one interferer, (D) Rake-Max-SINR, interferer is in direct path.

Fig. 6. SINR gain plotted against the number of image sources used in the design for different beamformers, at a frequency $f = 1$ kHz. The shaded area contains the Rake-Max-SINR output SINRs for 50% of the 20000 Monte Carlo runs.

Fig. 7. UDR gain plotted against the number of image sources used in the design for different beamformers, at a frequency $f = 1$ kHz. The shaded area contains the Rake-Max-UDR output UDRs for 50% of the 20000 Monte Carlo runs.

Fig. 8. SINR gain as a function of frequency for different beamformers and $K = K' = 10$. The curves show the average of 20000 runs, with averaging performed in the dB domain.

Sources present, the phase of the beamformer’s response plays an important role.

2) SINR Gains from Raking: In the experiments in this subsection, we set the power of the desired source and of the interferer to be equal, $\sigma_x^2 = \sigma_z^2 = 1$. The noise covariance matrix is set to $10^{-3} \cdot I_M$. We use a circular array of $M = 12$ microphones with a diameter of 30 cm, and randomize the position of the desired source and the interferer inside the room. The resulting curves show median performance out of 20000 runs.

Fig. 6 shows SINR gains from raking for different beamformers. The one-forcing beamformer is left out because it performs poorly in terms of SINR, as predicted earlier. Clearly, the Rake-Max-SINR beamformer outperforms all others. The output SINR for beamformers using only the direct path (Max-SINR and DS) remains approximately constant. UDR against the number of image sources for various beamformers is shown in Fig. 7. The Rake-Max-UDR beamformer performs well in terms of the two measures, but as we also mention
in the next section, its output is perceptually unpleasing due to audible pre-echoes. In informal listening tests, the Rake-Max-SINR beamformer did not produce such artifacts. It is interesting to note that the Rake-Max-SINR also performs well in terms of UDR. Similar SINR gains to those in Fig. 6 are observed in Fig. 8 over frequencies. It is therefore justified to extrapolate the results at one frequency in Fig. 6 to the wideband SINR.

3) Sound Samples: Next, we evaluate the sound output quality of Rake-Max-SINR and compare it to that of the conventional Max-SINR beamformer (using only the direct sounds of desired signal and interferer). We consider the same room, source, interferer, and microphone array as in Fig. 5B. The source signal is an extract of an opera singer (Fig. 9A). The interfering signal is a speech extract. The two signals are normalized to have unit maximum amplitude and mixed using the simulation framework described in Section VI-A (Fig. 9B). White Gaussian noise is added to the microphone signal with a power such that the SNR of the direct sound of the desired source is 20 dB at the center of the microphone array. All signals are high-pass filtered with a cut-off frequency of 300 Hz. The Rake-Max-SINR beamformer weights are computed using the direct source and three generations of image sources for both the desired sound source (singing) and the interferer (speech). The output of the conventional Max-SINR beamformer (Fig. 9C) is compared to that of Rake-Max-SINR (Fig. 9D). We can observe from the spectrogram that Rake-Max-SINR reduces very effectively the power of the interfering signal, at all frequencies, but particularly in the mid to high range, even when it overlaps significantly with the desired signal. Informal listening tests confirmed that the Rake-Max-SINR maintains high quality in the desired signal while providing significant reduction of the interferer power. The Rake-Max-UDR beamformer provides good interference suppression, but it produces audible pre-echoes that render it unsuitable for speech processing applications. The sound clips can be found online along the code.

VI. CONCLUSION

We investigated the concept of acoustic rake receivers—beamformers that use echoes. Unlike earlier related work, we presented optimal formulations that outperform the delay-and-sum style approaches by a large margin. This is especially true in the presence of interferers, hence the title “Raking the Cocktail Party”. We demonstrate theoretically that the ARR improves the SINR, and the numerical simulations agree well with these predictions.

Beyond theoretical and numerical evaluations of the performance measures, we demonstrated in informal listening tests the improved interference suppression by the ARR. A particularly illustrative example is when the interferer is in the direct path of the desired signal—the optimal ARR takes care of this simply by listening to the echoes.

Perhaps the most important aspect of ongoing work is the design of robust formulations of ARRs. It seems that this may involve some heuristics and combinatorial optimization. Furthermore, we only demonstrated the acoustic rake receiver in the static case, tracking methods need to be devised in order to tackle the dynamic case where sound sources are moving.

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] A. F. Naguib, “Space-time receivers for CDMA multipath signals,” in Proc. IEEE ICC, Montreal, 1997, pp. 304–308, IEEE.
[3] B. H. Khalaj, A. Paulraj, and T. Kailath, “2D RAKE Receivers for CDMA Cellular Systems,” in Proc. IEEE GLOBECOM, 1994, pp. 400–404, IEEE.
[4] P. Annibale, F. Antonacci, P. Bestagini, A. Bruni, A. Cancolini, L. Cristofoletti, J. Filos, E. Habets, W. Kelleman, K. Kowaleczky, 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.
[5] A. E. O’Donovan, R. Duraeswami, and D. N. Zotkin, “Automatic Matched Filter Recovery via the Audio Camera,” in Proc. IEEE ICASSP, Dallas, 2010, pp. 2826–2829, IEEE.
[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] S. Y. Seidel, T. S. Rappaport, and R. Singh, “Path Loss and Multipath Delay Statistics in Four European Cities for 900 MHZ Cellular and Microcellular Communications,” Electron. Lett., vol. 26, no. 20, pp. 1713–1715, 1990.
[8] S. T. Neely and J. B. Allen, “Invertibility of a Room Impulse Response,” J. Acoust. Soc. Am., vol. 66, no. 1, pp. 165–169, 1979.
[9] J. Mourjopoulos, “On the variation and invertibility of room impulse response functions,” J. Sound Vib., vol. 102, no. 2, pp. 217–228, 1985.
[10] 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.
[11] J. Borish, “Extension of the Image Model To Arbitrary Polyhedra,” J. Acoust. Soc. Am., vol. 75, no. 6, pp. 1827–1836, 1984.
[12] 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.
[13] 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.

[14] 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.

[15] 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.

[16] I. Dokmanic, 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.

[17] W. T. Chu and A. C. C. Warnock, “Detailed Directivity of Sound Fields Around Human Talkers,” Tech. Rep. RR-104, NRC Institute for Research in Construction; National Research Council Canada, 2002.

[18] H. K. Dunn and D. W. Farnsworth, “Exploration of Pressure Field Around the Human Head During Speech,” J. Acoust. Soc. Am., vol. 10, no. 3, pp. 184–199, Jan. 1939.

[19] J. S. Bradley, H. Sato, and M. Picard, “On the Importance of Early Reflections for Speech in Rooms,” J. Acoust. Soc. Am., vol. 113, no. 6, pp. 3233, 2003.

[20] J. Lochner and J. F. Burger, “The Influence of Reflections on Auditorium Acoustics,” J. Sound Vibration, vol. 1, no. 4, pp. 426–454, 1964.

[21] E. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, “New Insights Into the MVDR Beamformer in Room Acoustics,” IEEE Trans. Audio, Speech, Language Process., vol. 18, no. 1, pp. 158–170, Jan. 2010.

[22] 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.

[23] 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.

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

[25] K. Furuya, “Noise Reduction and Dererberation using Correlation Matrix Based on the Multiple-Input/Output Inverse-Filtering Theorem (MINT),” in Proc. Intl. Workshop on HSC, Kyoto, Japan, 2001, pp. 59–62.

[26] 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.

[27] S. Haykin and Z. Chen, “The Cocktail Party Problem,” Neural Comput., vol. 17, no. 9, pp. 1875–1902, 2005.

[28] W. Herbet, and W. Keller, “Adaptive Beamforming for Audio Signal Acquisition,” in Adaptive Signal Processing, pp. 155–194. Springer, Berlin, Heidelberg, Feb. 2003.

[29] B. D. Carlson, “Covariance Matrix Estimation Errors and Diagonal Loading in Adaptive Arrays,” IEEE Trans. Aerosp. Electron. Syst., vol. 24, no. 4, pp. 397–401, July 1988.

[30] T. E. Oliphant, “Python for Scientific Computing,” Computing in Science & Engineering, vol. 9, no. 3, pp. 10–20, 2007.

[31] D. Duffy, Green’s Functions with Applications, Chapman and Hall/CRC, 2001.

[32] B. D. Van Veen and K. M. Buckley, “Beamforming: A Versatile Approach to Spatial Filtering,” IEEE ASSP Mag., vol. 5, no. 2, pp. 4–24, 1988.

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

[34] I. J. Tashev, Sound Capture and Processing, Practical Approaches. John Wiley & Sons, Chichester, UK, July 2009.

[35] W. McCrea and F. Whipple, “Random Paths in Two and Three Dimensions,” in Proc. Roy. Soc. Edinburgh, 1940, vol. 60, pp. 281–298.

[36] M. Abramowitz and I. A. Stegun, Handbook of Mathematical Functions, Courier Dover Publications, 1972.

[37] O. O. Frost, “An Algorithm for Linearly Constrained Adaptive Array Processing,” in Proc. IEEE. 1972, pp. 926–935, IEEE.

[38] W. McCrea and F. Whipple, “Random Paths in Two and Three Dimensions,” in Proc. Roy. Soc. Edinburgh, 1940, vol. 60, pp. 281–298.

[39] M. Abramowitz and I. A. Stegun, Handbook of Mathematical Functions, Courier Dover Publications, 1972.

[40] O. Ocal, I. Dokmanic, and M. Vetterli, “Source Localization and Tracking in Non-Convex Rooms,” Proc. IEEE ICASSP, 2014.

[41] J. J. Shynk, “Frequency-Domain and Multirate Adaptive Filtering,” IEEE Signal Process. Mag., vol. 9, no. 1, pp. 14–37, 1992.