Meyer-Kahlen, Nils; Schlecht, Sebastian J.

Blind Directional Room Impulse Response Parameterization from Relative Transfer Functions

Published in:
2022 International Workshop on Acoustic Signal Enhancement (IWAENC)

DOI:
10.1109/IWAENC53105.2022.9914706

Published: 01/01/2022

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Blind Directional Room Impulse Response Parameterization from Relative Transfer Functions

Nils Meyer-Kahlen¹, Sebastian J. Schlecht¹,² *

¹Acoustics Lab, Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
²Media Lab, Department of Art and Media, Aalto University, Espoo, Finland

December 24, 2022

Abstract

Acquiring information about an acoustic environment without conducting dedicated measurements is an important problem of forthcoming augmented reality applications, in which real and virtual sound sources are combined. We propose a straightforward method for estimating directional room impulse responses from running signals. We adaptively identify relative transfer functions between the output of a beamformer pointing into the direction of a single active sound source and the complete set of spherical harmonics domain signals, representing all directions. To this end, estimation is performed with a frequency domain recursive least squares algorithm. Then, parameters such as the directions of arrival of early reflections and the reverberation time are extracted. Estimation of the direct-to-reverberant ratio requires dedicated processing. We show examples of successful estimation from speech signals, based on a simulated and a measured response. Directional Room Impulse Responses, Spatial Audio, Augmented Reality, System Identification

1 Introduction

Estimation of directional room impulse responses (DRIRs) is an important problem for acoustical rendering in augmented reality (AR). The response of a room needs to be known well enough such that all important room acoustical features are replicated when rendering virtual sound sources over headphones. If virtual source seamlessly blend in with real sources, we call the rendering transfer-plausible [1].

To solve the estimation problem of transfer-plausible rendering, we propose a method for blind DRIR estimation based on beamforming and deconvolution, or rather adaptive system identification. This straightforward, interpretable approach is an alternative to algorithms for transferring acoustics from one signal to another that are entirely based on deep neural networks [2, 3].

Figure 1 shows an overview of the complete system. The input is an Ambisonics recording, which was captured using a microphone array, and transformed into the spherical harmonics (SH) domain. First, the system detects the presence of a single active sound source like a human speaker, for example using the direct path dominance test [4]. Then, the direct signal is extracted by means of a beamformer pointing to that source. The beamformer output is subsequently used as a reference signal for deconvolving all available SH signals. From this, we obtain a set of directional relative transfer-functions (RTFs) [5], which are then regarded as an estimate of the DRIR. In practice, deconvolution should be time-adaptive. For this, we use the frequency domain least-squares algorithm (FD-RLS), which can be seen as an extension of regularized deconvolution. Once

*This research has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 812719.
estimates are obtained, parameters like the directions of arrival (DoAs) of early reflections and the spatio-temporal envelope of the late reverberation are determined from it, all with the intention of using them to render virtual sound sources. Virtual acoustic rendering using a limited set of parameters [6] is likely to yield sufficient quality for transfer-plausibility. In contrast to applications such as echo cancellation or dereverberation, where estimation methods have to be accurate enough to successfully invert the system, the goal of merely extracting sound field parameters is likely to relax the requirements on the system identification.

In the following, Section 2 summarizes related approaches. Then, Section 3 describes the central processing steps, which are beamforming, adaptive system identification and parameterization. Section 4 shows examples of estimated DRIRs.

2 Background

In contrast to the present application of reverberating a dry signal, blind identification of room impulse responses has mainly been treated in the field of echo cancellation and dereverberation. Estimation of multichannel responses has for example been attempted with adaptive methods that minimize the cross-relation between microphone pairs [7], like the family of normalized multichannel frequency-domain least mean-square (NMCFLMS) algorithms. The reason for using cross-relation is that blind system identification can not rely on a reference signal. Cross-relation based methods have been shown to perform well in small examples, and several variants incorporating different constraints were developed, e.g., [8]. These methods make use of multiple receivers, which ideally need to be far apart from each other to avoid the problem of common transfer-function zeros.

By contrast, applications that demand for transfer-plausible rendering will most likely comprise of compact arrays, for example worn on the head of a user [9]. Well-designed compact arrays allow for applying a beamformer directed to a source in the room to obtain a reference signal. So far, beamforming as a means of obtaining a reference signal for system identification seems to have been explored only rarely. In [10] a related approach is developed for denoising Ambisonics signals. Therein, a similar algorithm is used with the difference that the estimated DRIR is immediately convolved with the beamformer output signal to obtain a denoised version thereof, and evaluation focused on the quality of this signal. Here, the goal is to apply the estimated response to create a different, virtual sound source. Therefore, we focus on the estimated response itself and show parameters obtained from it. Also, while [10] uses the standard least mean squares (LMS) algorithm, we use FD-RLS and show how it is a natural extension of regularized deconvolution.

3 Method

In principle, the proposed method can be applied to any compact microphone array that is capable of beamforming. For the example below, we use a spherical microphone array (SMA). Transforming the recording to the spherical harmonics (SH) domain [11] makes single source detection, DoA estimation and broadband beamforming most convenient. Assuming that the direction of a source $\theta_s$ is known, one can use the modal beamforming weights $d$ to extract the target signal from an SH domain microphone sig-
nal $\tilde{x}$ simply by

$$x_s(f) = d^T \tilde{x}(f),$$

where $f$ is the frequency index and $[\cdot]^T$ denotes transposition. In case of the simple max-DI beamformer, which is also referred to as a plane-wave decomposition or a hyper-cardioid pattern, the vector $d_{\text{max-DI}} = y_N(\theta)$, simply contains the $Q = (N + 1)^2$ spherical harmonics of maximal order $N$, evaluated at the direction $\theta$, i.e., $y_N(\theta) = [Y_{0,0}(\theta) \ Y_{1,-1}(\theta) \ Y_{1,0}(\theta) \ldots \ Y_{N,N}(\theta)]^T$, where $Y_{n,m}$ are the real spherical harmonics of order $n$ and degree $m$, arranged according to the Ambisonics channel numbering (ACN) convention $q = n^2 + n + m$.

Other axis symmetric patterns are obtained by selecting one weight $w_n$ for all components belonging to one SH order $n$, and applying

$$d = \text{diag}_N(w_n)y_N(\theta_s),$$

where $\text{diag}_N$ creates a diagonal matrix, which has one unique weight $w_n$ for all $2n + 1$ components belonging to each SH order. Max-rF weights offer a good compromise between main lobe width and side lobe strength [12],

$$w_n \approx P_n\left(\frac{\cos(137.9^\circ)}{N + 1.51}\right).$$

Now, deconvolution can easily be computed by frequency domain division to retrieve the set of RTFs

$$\hat{h}(f) = \frac{\tilde{x}(f)}{x_s(f)},$$

whose inverse Fourier transform with respect to time is what we regard as an estimate of the desired SH domain DRIR. The consequences of this assumption will be discussed in Section 3.2. Note that henceforth, the frequency index $f$ will be omitted for the sake of readability.

It is likely that the recorded signal $\tilde{x}$ does not have energy at all frequencies, for example in the common application case in which in the input is speech. For this reason, regularization should be applied. One of the simplest choices is Tikhonov regularization, which in the present single-input multiple-output (SIMO) case is achieved by

$$\hat{h} = \frac{x_s^*\tilde{x}}{x_s^*x_s + \lambda^*},$$

where $[\cdot]^*$ denotes complex conjugation, and $\lambda$ is a regularization constant.

### 3.1 Extension to Frequency Domain RLS

While the above already captures the main idea of the beamforming and deconvolution algorithm, in practice, estimation should be adaptive, for example because the position of the source and the microphone array may change. There is a very close relationship between Tikhonov regularized deconvolution in eq.
(a) Weighted angular error  (b) Relative error in $\hat{T}_{20}$

Figure 3: Error in parameters obtained on 100 realizations of simulated rooms for different input orders. The white dot indicates the median.

(4) and the FD-RLS algorithm, given by [13]

\[
\tilde{h}(k) = \tilde{h}(k-1) + \frac{1}{\phi(k)} x_s^*(k) \varepsilon(k) 
\]

\[
\phi(k) = \lambda \phi(k-1) + x_s^*(k)x_s(k) 
\]

\[
\varepsilon(k) = x(k) - \hat{h}(k-1)x_s(k), 
\]

where the estimation is performed in blocks of $K$ samples, and $k$ denotes the block index. Note again, the iterations are performed for each frequency bin $f$. As the output of the beamformer is not a completely anechoic reference signal, but contains some reflections itself, the estimated response has an acausal component. Therefore, only the $K/2$ samples belonging to the causal part of the response are extracted.

It is easy to see that when the frequency domain signal power is initialized with $\phi(0) = 1$, and the impulse response estimate with $\hat{h}(0) = 0$, the first estimate $(k = 1)$ of the FD-RLS is exactly the same as eq. (5). The used FD-RLS is equivalent to the more common time domain RLS [14]. In general, RLS has advantages over LMS in terms of convergence speed and final error, at a higher computational cost [15]. In fact, experiments with applying standard LMS to the responses as in [10] did not lead to convergence in our test cases.

### 3.2 Parametrization

Now that an estimate of the DRIR is available, RIR parameters can be extracted. For this, a wide range of parametric DRIR methods is available, such as spatial impulse response rendering (SIRR) [16], higher-order SIRR [17], or the spatial decomposition method (SDM) [18]. All these methods comprise different sound field models using a different set of parameters extracted from the DRIR.

SDM uses a simple sound field model, and the parameters are easy to interpret. It assumes that the sound field can be described by the sound pressure $p(t)$ and one DoA $\hat{\theta}(t)$ per time instance $t$. The pressure is directly available from the omnidirectional channel $p(t) = h_0(t)$ and the DoAs can be estimated using the pseudo intensity vector (PIV)

\[
\hat{\theta}(t) = \hat{h}_0(t) [\hat{h}_4(t) \quad \hat{h}_2(t) \quad \hat{h}_3(t)]^T, 
\]

where $\hat{h}_{(4,2,3)}$ are the responses belonging to the first-order SH components, pointing along the x, y and z axis. The PIV yields good estimates of early reflection DoAs, but the assumption of one direction per time instance is inaccurate in the late tail of the response, where the PIV can at best approximate the energy distribution [19].

To parameterize the decay, we also extract the reverberation time $\hat{T}_{20}$ of the omnidirectional RIR, by using the common backwards integration method.

When obtaining a DRIR estimate via directional RTFs, several deviations from the true response are expected, because the output of the direct sound beamformer is not a perfect, anechoic reference signal. These errors depend on the spatial resolution of the beamformer. As a first consequence, the estimated response becomes acausal, but the non-zero before the largest peak is easily discarded. Another consequence is that it is impossible to estimate reflections arriving from the same direction as the direct sound, and that DoA estimates of reflections close to the direct sound will be directionally biased, as it will be seen in the examples. A further consequence of the non-ideal reference is that the direct-to-reverberant ratio will not be correct, which therefor needs to be estimated in a separate step. Here, we employed a
Figure 4: Comparison between DoAs in the early part of the response. The source was placed on the left. A first order floor reflection and backwall reflection are marked in (a), which are clearly identified with the proposed algorithm.

(a) DoAs measurement with exponential sine sweep

(b) DoAs using proposed method with speech input signal

method similar to [20], which is based on the energy of the omnidirectional channel and a pattern with a null steered towards the source.

4 Examples

4.1 Simulation

As a first step, the algorithm was applied to simulated responses of shoebox shaped rooms with random size $(8 \pm 4) \text{m} \times (5 \pm 4) \text{m} \times (3 \pm 1) \text{m}$ using up to 5th order image sources. The rooms had homogeneous, frequency-independent absorption and a randomized reverberation time between 0.1 s and 0.5 s. The late part of the response was simulated using isotropic, directionally uncorrelated noise. An ideal omnidirectional source and an ideal SH receiver of maximal SH order $N = \{1, 3, 5\}$ were randomly placed within half the maximal dimensions of each room. Estimation was done using 20s of male speech from the EBU SQUAM CD\(^1\), convolved with the response and the described frequency-domain RLS algorithm. For estimation a block size of 0.5s and an FFT size of $2^{15}$, and $\lambda = 0.99$ were used. Even though the response was static in time, the FD-RLS was used as a proof of concept, and since it provided better results than performing direct deconvolution with the complete signal.

Figure 2 shows DoA estimates from one of the simulated rooms in black and the true DRIR in red. The accuracy increases with the order. To quantify this observation, the weighted angular error is computed as

$$
\epsilon_\theta = \frac{\sum_t p^2(t) \arccos \{\theta^\top(t) \hat{\theta}(t)\}}{\sum_t p^2(t)},
$$

where $\theta(t)$ are the DoA estimates obtained from the simulated response directly. The results are shown in Figure 3a. Errors in the range of $10^\circ$ and more are not uncommon, even in non-blind simulations, for example when directional estimation is performed using spaced microphone arrays [21].

Furthermore, Fig.3b shows the deviation between the estimated reverberation time and the reverberation time obtained directly from the simulation. For third order signals, the median error is about 5.8% of the true $T_{20}$.

4.2 Measurement

For a second, more realistic example, the same speech sequence was convolved with a DRIR measured using an Eigenmike em32 and a Genelec 8331AP loudspeaker in the variable acoustics room “Arni”. The loudspeaker was on the left of the receiver, see loudspeaker 1 in [22], where also the room is shown in more detail. The result of the estimation with the

---

\(^1\)Sound Quality Assessment Material recordings for subjective tests - ©EBU SQAM CD
RIR length set to 0.4 s is shown in Figure 4. The floor and backwall reflection seen in the DoA estimates from the response can also be found in the estimated response with similar salience. The floor reflection direction is pushed away from the direction of the direct sound.

5 Conclusion and Future Work

We have presented a method for estimating DRIRs from running signals. The method directs a beamformer to the source and then identifies and estimates the relative transfer functions between this direction and all other directions. We have shown that by treating the inverse Fourier transform of the directional RTF as estimates of the DRIR, parameters such as the DoA of early reflections or \( T_{20} \) can be extracted. The performance depends on the spatial resolution of the beamformer; higher-order arrays are required for reaching reasonable errors. Future work includes using the estimated responses for rendering in order to assess if transfer-plausibility is achieved. Also, methods for single source detection and DRR estimation shall be evaluated in this context.

References


