A Robust Denoising Process for Directional Room Impulse Responses with Diffuse Reverberation Tails

Pierre Massé, Thibaut Carpentier, Olivier Warusfel, Markus Noisternig

A Robust Denoising Process for Directional Room Impulse Responses with Diffuse Reverberation Tails

Pierre Massé,[1] Thibaut Carpentier,[1] Olivier Warusfel,[1] and Markus Noisternig[1]

Acoustic and Cognitive Spaces group, Sorbonne Université, IRCAM, CNRS, STMS, 75004 Paris, France[a)]

Directional room impulse responses (DRIR) measured using spherical microphone arrays (SMA) are seeing increasingly widespread use in reproducing room reverberation effects on three-dimensional surround sound systems (e.g. Higher-Order Ambisonics) through multi-channel DRIR convolution. However, such measured impulse responses inevitably present a non-negligible noise floor, which may lead to a perceptible "infinite reverberation effect" when convolved with an input sound. Furthermore, individual sensor noise and momentary measurement artefacts may additionally corrupt the resulting impulse response. This paper presents a robust DRIR denoising procedure applicable to impulse responses with diffuse late reverberation tails, which can be modeled by a stochastic process. In such cases, the non-decaying frequency-dependent noise floor may be replaced by a synthesized diffuse tail parameterized by the DRIR's energy decay envelope. It is shown that performing such tail re-synthesis in the spherical harmonic domain (SHD), using an independent zero-mean Gaussian noise for each component, preserves not only the reverberation tail's frequency-dependent decay properties, but also its spatial incoherence. The proposed process is then evaluated through its application to DRIRs measured in real-world conditions, and finally some aspects of performance and consistency verification are discussed.

[a)] pierre.masse@ircam.fr;

I. INTRODUCTION

A. SMA analysis in the SHD

Spherical microphone arrays (SMA) enable the directional analysis of a given sound field by sampling it over the Q transducer positions on their surface. A natural choice of representation for a function defined on such a surface S^2 is the spherical harmonic domain (SHD), whose basis functions $Y_{l,m}$ are analogues of the trigonometric functions in the application of Fourier expansion theory on the sphere (Driscoll and Healy, 1994):

$$X_{l,m}(f,t) = \int_{\Omega \in S^2} x(f,\Omega,t) Y_{l,m}(\Omega) d\Omega, \tag{1}$$

where $\Omega = (\theta, \phi)$ is a point on the surface of a sphere with fixed radius $r = a$ (in conformity with ISO8000-2:2009 (E)), $x(f,\Omega,t)$ is the time-frequency domain representation of the sound field on the sphere, and $Y_{l,m}(\Omega)$ are the spherical harmonics of order $l \in \mathbb{Z}^+$ and degree $m \in [-l, l]$. This transform thus defines the SHD signal coefficients $X_{l,m}(f,t)$ for each component or mode (l,m). Using an SMA, the integral in Eq. (1) is discretized and can be approximated by a weighted sum over the microphone positions; the specific weights are chosen such that the sum approaches the ideal integral of Eq. (1), e.g. by least-squares minimization (Rafaely, 2005).

The discrete transform can be simply written in matrix form:

$$\mathbf{x}_{\mathrm{SHD}}(f,t) = \mathbf{Y}\mathbf{x}(f,t), \tag{2}$$

where $\mathbf{x}(f,t)$ is the column vector containing the time-frequency representation of the signal measured at each transducer position Ω_q, $\mathbf{Y}$ is the $(L+1)^2 \times Q$ encoding matrix of elements $y_{q,n} = \alpha_q Y_{l,m}(\Omega_q)$ (with indices $n = l^2 + l + m + 1$ up to a maximum encoding order L, and α_q the aforementioned array weights), and $\mathbf{x}_{\mathrm{SHD}}(f,t)$ is the column vector of resulting SHD coefficients. The array's sampling configuration must then lead to an encoding matrix with K non-vanishing singular values such that $K = (L+1)^2 \leq Q$ (Noisternig *et al.*, 2011), thereby effectively limiting the maximum achievable order L for a given SMA. Finally, in order to obtain an array-independent representation of the measured sound field, a subsequent correction for the so-called mode strengths (or holographic functions) of the SMA must be applied. Such is the case in the widespread Higher-Order Ambisonics (HOA) format, where the center of the sphere is used as the reference point and for which the correcting filters are determined accordingly (Daniel and Moreau, 2004).

B. Previous work

Monophonic room impulse responses (RIR) have long been modelled as an exponentially decaying stochastic process (Schroeder, 1962), which has been shown to be valid assuming sufficiently high echo density and modal overlap is achieved (Polack, 1988). These conditions lead to a lower time limit for echo density, known as the "mixing time", and a lower frequency limit for modal overlap, known as the "Schroeder frequency". Beyond these limits, the late reverberation field is considered to be fully "diffuse", i.e. it behaves as a spatially isotropic distribution of a statistically significant number of incoherent and uncorrelated plane waves. Such a field can be synthesized in the form of a zero-mean Gaussian noise filtered by an

exponentially-decaying energy envelope (Jot *et al.*, 1997). This envelope is parameterized by a frequency-dependent decay coefficient $\delta(f)$ (usually represented as the 60 dB reverberation time, $T_{60}(f) = 3\ln(10)/\delta(f)$) and an initial power spectrum $P_0(f)$; these parameters can be extracted by analysis of the energy decay relief (EDR), a time-frequency extension of the Schroeder energy decay curve (EDC) (Jot *et al.*, 1997). Non-decaying background noise present in a measured impulse response can therefore be replaced by a synthesized zero-mean Gaussian noise filtered by a prolongation of the energy decay envelope (Jot *et al.*, 1997). As a result, the final signal-to-noise ratio (SNR) is limited only by the quantization noise floor for the chosen synthesis bit depth, $P_{QN} = 20\log_{10}(2^{-d})$ dB, where d is the signal bit depth.

Guski and Vorländer (2014) have since presented a variety of other noise compensation methods, but these focus more on regularizing the broadband EDC calculation in order to improve the accuracy of extracted room acoustics parameters (e.g. T_{60}, C_{80} clarity, etc.), rather than faithfully re-synthesizing the reverberation tail for convolution applications. Some of their techniques resemble that proposed by Cabrera *et al.* (2011) for auralizing measured RIRs; all have so far only been presented in the monophonic single-slope decay case. Furthermore, eschewing tail re-synthesis for simple decay envelope adjustment places strict conditions on the content of the background noise (as noted by Cabrera *et al.* (2011)), which may easily not be verified in many "real-world" measurement conditions.

Preliminary extensions of Jot's tail re-synthesis process to the spatialized DRIR case were presented by Carpentier *et al.* (2013) (using a reference diffuse field simulated by large numbers of incident plane waves to denoise the individual SMA transducer signals) and Noisternig *et al.* (2014) (in the SHD using independent zero-mean Gaussian noise realizations

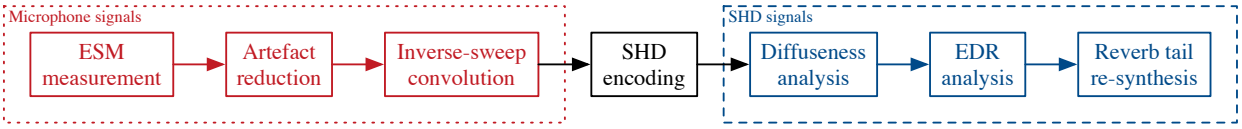


FIG. 1. Outline of the proposed DRIR denoising process, from the initial exponential sweep method (ESM) measurement through to the energy decay relief (EDR) analysis and reverberation tail re-synthesis in the spherical harmonic domain (SHD).

per component), both once again in the single-slope decay case. The current work builds upon and further details these methods, allowing for multiple-slope decays (such as those observed in certain coupled-volume configurations) and demonstrating that tail re-synthesis in the SHD guarantees preservation of the late reverberation's incoherence properties.

II. PROPOSED DENOISING PROCESS

The different parts of the proposed denoising framework are presented in this section, and the sequencing of the individual steps is outlined schematically in Fig. 1. The exponential sweep method (ESM) measurement and subsequent inverse-sweep convolution are based on Farina (2000) and performed on each SMA transducer signal independently; between these two steps we introduce an artefact reduction procedure described in section II A. The SHD encoding is based on the theory presented in the introduction, and the EDR analysis is an extension of Jot *et al.* (1997), detailed in section II B 1. Finally, the main focus of this work is on the diffuseness and mixing time analysis (section II B 2) and reverberation tail re-synthesis (section II B 3).

A. Measurement artefact reduction

Measuring impulse responses using the ESM in so-called "real-world" conditions is inevitably subject to three main risk factors: the presence of constant, stationary background noise (including transducer self-noise), any non-stationarity of the measurement conditions (temperature, humidity, etc.), and the occurence of non-stationary noise events. The first is what is assumed in previous work on the subject and what is aimed to be removed in the tail re-synthesis procedure. The second can lead to time-variance in the impulse responses which would require post-processing correction techniques using *a priori* information on the measurement conditions, and will not be considered in this study. The third is what we will refer to here as "measurement artefacts", i.e. short-term sonic events occuring during the measurement, and reducing their impact is the aim of this section.

As noted by Farina (2000), averaging a repetition of several sweeps is a simple way to increase the SNR, since the ensemble mean of any incoherent stationary noise will tend to zero as the number of repetitions increases. However, any non-stationary noise events present in the repetitions will inevitably end up in the noise floor of the average. This is especially troublesome when considering Schroeder-type reverse-integrated analysis such as the EDR, since these artefacts will not only accumulate in the reverse-integration of the noise floor, they will also deviate substantially from the theoretical profile of a reverse-integrated constant-power noise floor (see section II B 1 below).

In an attempt to minimize the influence of these non-stationary noise events, the magnitude spectrograms of the individual sweep repetitions are compared amongst each other

in order to identify artefacts, using non-negligible positive deviations from the mean magnitude spectrogram as a discriminating criterion. This maximum allowed deviation is defined as $\xi(f,t) = \mu(f,t) + \alpha\sigma(f,t)$, where $\mu(f,t)$ is the mean magnitude spectrogram, $\sigma(f,t)$ is the standard deviation over the available repetitions, and α is an empirically-set deviation factor used as a control parameter. Artefact magnitude values identified as greater than $\xi(f,t)$ in each realisation are then replaced with the corresponding mean magnitude over the remaining repetitions.

This process is applied independently to the ESM measurement signals recorded by each SMA transducer. Some example results are illustrated and discussed in section III A.

B. Reverberation tail analysis and re-synthesis

In this section, we first review the energy decay relief (EDR) analysis procedure used to extract the reverberation decay parameters, and then present a characterization of the DRIR's mixing time using a measure of the sound field's diffuseness, before showing that re-synthesizing the reverberation tail as a zero-mean Gaussian noise in the SHD preserves the late field's spatial properties.

1. EDR analysis

The EDR is a time-frequency extension of Schroeder's reverse-integrated broadband energy decay curve (EDC), from which frequency-dependent decay envelope parameters can be extracted by analyzing each frequency bin individually (Jot *et al.*, 1997). We begin our analysis by identifying the exponential decay section of the reverse-integrated curve presented

by the EDR at each frequency bin. In dB scale (such that exponential sections become linear), this curve is first segmented using an adaptive Ramer-Douglas-Peucker (RDP) algorithm (Prasad *et al.*, 2012) in order to help identify the different sections (early reflections, exponential decay, and noise floor).

The noise floor limit point $\{P_{\text{noise}}, t_{\text{lim}}\}$ can be found by fitting the theoretical dB-scale profile of a reverse-integrated constant-power noise to the curve segments (see the shaded area on Fig. 2). An additional headroom above this noise profile is then adaptively determined (see below) to ensure the limiting point $\{P_{\text{noise}}, t_{\text{lim}}\}$ belongs to the exponential decay section of the curve, thereby avoiding discontinuities when prolonging the reverberation envelope for tail re-synthesis. Finally, any non-exponentially decaying early reflection regimes are discarded by selecting an appropriate starting segmentation point (t_{start}, see Fig. 2) using a criterion on the local slopes of the curve segments up until t_{lim} (early segments to discard are assumed to be shorter and have significantly different local slopes than those belonging to exponential decays). The exponential decay section is thus delimited by t_{start} and t_{lim} and the reverberation time (T_{60}) and initial power (P_0) values can be determined by fitting an ideal decay envelope model.

In the case of a single-slope decay, the envelope parameters can be found by performing a linear regression on the identified decay section of the dB-scale curve. For multiple-slope decays, such as those observed in certain configurations of coupled volumes (Cremer *et al.*, 1982), a parameter-space search can be performed in order to fit the model to the measured

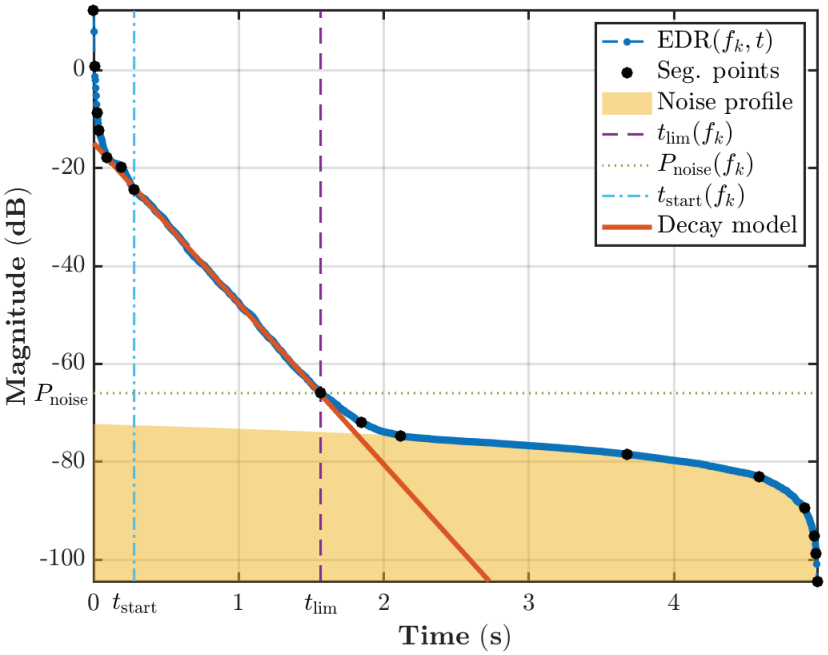


FIG. 2. (Color online.) EDR analysis schematic for a given frequency bin. The reverse-integrated decay curve is first segmented (black points). The noise floor (shaded area) is then identified, along with the noise floor limiting point $\{P_{\text{noise}}, t_{\text{lim}}\}$ (dotted and dashed lines, respectively). Early decay sections are avoided by identifying t_{start} (dash-dot line), and the exponential decay model is fitted between t_{start} and t_{lim}.

decay (Xiang *et al.*, 2011). In general, if we consider the global energy envelope of a system of C coupled volumes to be a sum of C exponential decays:

$$\text{ENV}(f, t, \mathbf{\Lambda}) = \sum_{i=1}^{C} P_{0,i}(f) e^{-2\delta_i(f)t}, \tag{3}$$

where $\delta_i(f)$ are the frequency-dependent decay coefficients, related to the $T_{60}(f)$ by $T_{60}(f) = 3\ln(10)/\delta(f)$, and $\mathbf{\Lambda}$ denotes the parameter vector containing the $P_{0,i}$ and δ_i values, then the ideal integrated decay curve is given by (see also Jot *et al.*, 1997)

$$\begin{aligned}\widehat{\mathrm{EDR}}(f,t,\mathbf{\Lambda}) &= \int_t^\infty \mathrm{ENV}(f,\tau,\mathbf{\Lambda})d\tau \\ &= \sum_{i=1}^{C} \frac{P_{0,i}(f)}{2\delta_i(f)} e^{-2\delta_i(f)t}.\end{aligned} \tag{4}$$

A *model error* ϵ_{mod} can be defined as a simple mean-squared error (MSE):

$$\epsilon_{\mathrm{mod}}(f) = \frac{1}{N_{\mathrm{fit}}} \sqrt{\sum_{n=n_s}^{n_e} \left[\mathrm{EDR}_{\mathrm{dB}}(f,t_n) - \widehat{\mathrm{EDR}}_{\mathrm{dB}}(f,t_n,\mathbf{\Lambda})\right]^2}, \tag{5}$$

where $N_{\mathrm{fit}} = n_e - n_s + 1$, with n_s the discrete time index such that $t_{n_s} = t_{\mathrm{start}}$ and similarly n_e such that $t_{n_e} = t_{\mathrm{lim}}$. This error can then be used as a loss function (or inversely as a likelihood) in order to perform the parameter search using an expectation-maximisation (EM) or maximum-likelihood (ML) algorithm. At each frequency bin, the parameter space is of dimension $2C$, since for each exponential decay both $P_{0,i}(f)$ and $\delta_i(f)$ must be estimated. To optimize the EM and avoid the detection of false local likelihood maxima, the algorithm is initialized using linear regressions performed on EDC segments defined by re-applying the adaptive RDP algorithm between t_{start} and t_{lim}.

The model error can additionally be used to adjust the headroom above the fitted ideal noise profile mentioned above. The procedure described above (segmentation, noise fitting, start point detection, and decay parameter search) is reiterated for several headroom values, and the result with the highest overall likelihood (lowest error) is chosen. The likelihood function used in this work is based on the Akaike Information Criterion (AIC) (Akaike,

172 1974) and can be written $\mathcal{L} = 2\log(1/\epsilon_{\text{mod}}) - 2C + \log(N_{\text{fit}})$, where again C is the number of
173 coupled decays, and $\log(N_{\text{fit}})$ is a regularization term used to promote fits made over longer
174 decay sections (i.e. for two fits with equal likelihood, the one made over a longer section of
175 the EDR bin will be preferred).

176 ***2. Diffuseness analysis and mixing time estimation***

177 As mentioned in section I B, replacing the non-decaying noise floor with a reverberation
178 tail synthesized as an exponentially-decaying zero-mean Gaussian noise assumes that the
179 late sound field described by the impulse response is fully diffuse. This leads to the classic
180 time-frequency limits for stochastic modelling of room reverberation, respectively the mixing
181 time and Schroeder frequency (Polack, 1988). The exploration of strategies for denoising in
182 the modal domain below the Schroeder frequency is left to future work; in this paper we will
183 apply the tail re-synthesis process across all frequencies, and note that for most reverberant
184 spaces the Schroeder frequency is low enough that the human auditory system is largely
185 insensitive to the modal reverberation below it. (This can be seen by comparing Schroeder's
186 measure $f_{\text{Sch}} \approx 2000\sqrt{\overline{T}_{60}/V}$, where $\overline{T}_{60}$ is a broadband measure of the reverberation time
187 and V is the volume of the space (Schroeder and Kuttruff, 1962), to equal-loudness contours
188 such as those given by the ISO226:2003 standard.)

189 Defining the mixing time, however, is crucial to the present work. Considering the afore-
190 mentioned requirement of a fully diffuse late sound field for synthesizing the prolongation
191 of the reverberation tail using a zero-mean Gaussian noise, we propose using a measure of
192 the sound field's diffuseness in order to estimate the moment the DRIR becomes maximally

diffuse. Furthermore, in the following section II B 3 we will show that re-synthesizing the late reverberation tail in the SHD guarantees that the resulting sound field will preserve these diffuseness properties.

Several measures of diffuseness have been proposed that directly exploit various characteristics of the SHD. The DirAc measure (Ahonen and Pulkki, 2009) uses the zeroth- and first-order components to define a sound intensity vector and analyze its temporal variation. Jarrett *et al.* (2012) use SHD inter-component coherence to define a "signal-to-diffuse ratio" (SDR) that is evaluated with respect to a directional signal with a given direction of arrival (DOA). Finally, the COMEDIE measure (Epain and Jin, 2016) exploits the eigendecomposition of the SHD signal covariance matrix, which will approach the identity matrix in the case of a fully diffuse field. The COMEDIE measure was chosen for this work due to its increase in accuracy with SHD order (whereas the DirAc measure is limited to first-order signals), its relatively lightweight implementation, and its independence from external analyses (whereas the SDR measure relies on the DOA).

In typical "large" mixing spaces, diffuseness profiles tend to quickly reach a stable maximum, as shown in Fig. 3 for the Kraftzentrale event venue in Duisburg, Germany (an industrial-era factory hall approximately 84000 m^3 in volume). Estimating the mixing time then corresponds to identifying the moment the DRIR reaches its maximum diffuseness. The idea here is to first characterize the maximum diffuseness and then find when the DRIR reaches this maximum in a definitive manner after an initial period of instability due to

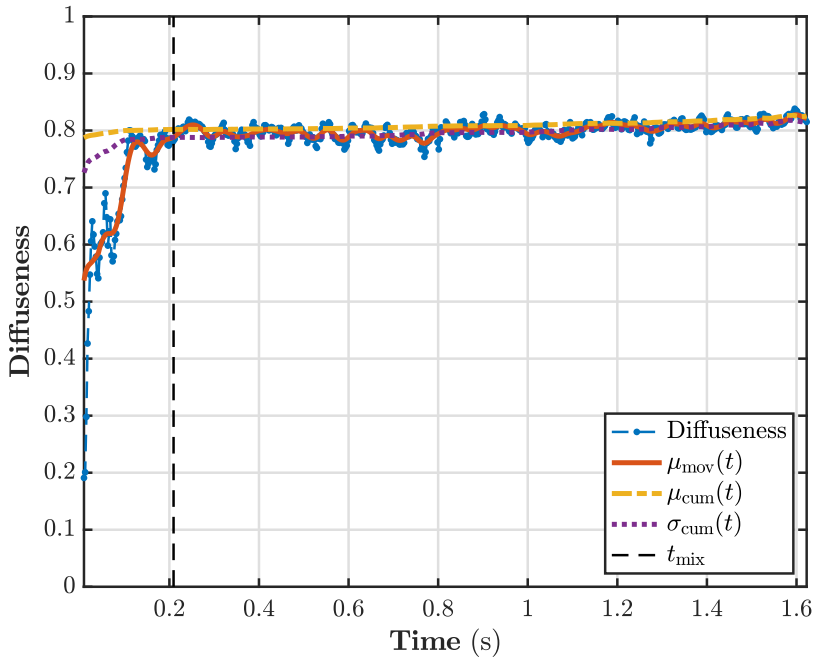


FIG. 3. (Color online.) COMEDIE diffuseness analysis (Epain and Jin, 2016) and mixing time estimation for a 4$^{\mathrm{th}}$-order SHD DRIR measured at the Kraftzentrale event venue in Duisburg, Germany, using an mh acoustics Eigenmike®. The calculated diffuseness curve is smoothed using a Gaussian kernel moving average (μ_{mov}). An inverse cumulative average (μ_{cum}) and standard deviation (σ_{cum}) are further used to identify the onset of the maximum diffuseness and thereby estimate the mixing time (t_{mix}).

coherent early reflections. This can be done by means of an appropriately-sized moving average,

$$\mu_{\mathrm{mov}}(t_i) = \sum_{n=i}^{i+N_w-1} w(t_n - t_i) d(t_n), \tag{6}$$

215 a reverse-cumulative average,

$$\mu_{\text{cum}}(t_i) = \frac{1}{N_d - i + 1} \sum_{n=i}^{N_d} d(t_n), \tag{7}$$

216 and a reverse-cumulative standard deviation,

$$\sigma_{\text{cum}}(t_i) = \sqrt{\frac{1}{N_d - i} \sum_{n=i}^{N_d} \left[d(t_n) - \mu_{\text{cum}}(t_n)\right]^2}, \tag{8}$$

217 where $i = 1, 2, ..., (N_d - N_w + 1)$ with N_d the length of the diffuseness data $d(t_i)$ and
218 w a chosen averaging kernel of length N_w. In this work (see Fig. 3), a 24-point Gaussian
219 kernel was used to calculate μ_{mov} on diffuseness data obtained using a 1024-sample, 87.5%
220 overlapping short-term Fourier transform and mathematical expectations estimated by a
221 subsequent 8-frame average (at a 48 kHz sampling rate, this corresponds to a 40.0 ms
222 average for diffuseness points and a 101 ms total average for μ_{cov}). The mixing time is then
223 determined by

$$t_{\text{mix}} = \min(t_{\text{diff}}), \tag{9}$$

224 where the time values t_{diff} satisfy

$$\sqrt{\left[\mu_{\text{mov}}(t_{\text{diff}}) - \mu_{\text{cum}}(t_{\text{diff}})\right]^2} \leq \sigma_{\text{cum}}(t_{\text{diff}}). \tag{10}$$

225 Additional checks can subsequently be performed to ensure that no μ_{mov} values are below
226 a certain threshold from μ_{cum} after this time (e.g. corresponding to late-arriving discrete
227 echoes), adjusting t_{mix} to a statisfying t_{diff} value if necessary. Further validation tests on the

value of the maximum diffuseness may also be included (e.g. a diffuseness maximum below 0.5 may not be considered "maximally diffuse").

We now need to define a condition for re-synthesizing the reverberation tail using a zero-mean Gaussian noise: if the DRIR reaches its mixing time before decaying below the noise floor, the stochastic model can be used as first proposed by Jot *et al.* (1997). However, whereas the mixing time is a broadband property, the EDR analysis described above returns a frequency-dependent noise floor limiting time $t_{\mathrm{lim}}(f)$. To get a global value for the noise floor limiting time, we use the $t_{\mathrm{lim}}(f)$ values determined for the SHD-encoded DRIR's $Y_{0,0}$ (omnidirectional) component and perform a perceptually-weighted average over the audible frequency range. This average is weighted according to the ITU-R 468 standard noise filter and then evaluated over Bark-scale frequency bands in order to avoid the over-weighting of higher-frequency bins due to the linear frequency scale of the Fourier transform.

We denote the resulting value $\bar{t}_{\mathrm{lim}}$, and the condition can then be written $t_{\mathrm{mix}} < \bar{t}_{\mathrm{lim}}$. If it is verified, tail re-synthesis may be performed using a zero-mean Gaussian noise as described below, with the perceptual considerations above ensuring that any $t_{\mathrm{lim}}(f)$ values smaller than t_{mix} should have a limited perceptual impact (future work is planned to further strengthen this aspect, e.g. by taking into account the corresponding $P_{\mathrm{noise}}(f)$ values). If the condition is not verified, however, alternative methods of noise reduction must be considered (see the conclusion in section IV below).

3. *Diffuse tail synthesis*

We now show that re-synthesizing the reverberation tail as a zero-mean Gaussian noise in the SHD preserves the spatial incoherence properties of the late reverberation field. In the SHD, the signal measured by a SMA in the presence of a perfectly diffuse field is of the form

$$X_{l,m}^{\text{diff}}(f,t) = \sqrt{P_{\text{diff}}(f,t)}b_l(f)\int_{\Omega\in S^2}\Phi(f,\Omega,t)Y_{l,m}(\Omega)d\Omega, \tag{11}$$

where $P_{\text{diff}}(f,t)$ is the diffuse field power envelope, $\Phi(f,\Omega,t) = e^{i\varphi(f,\Omega,t)}$ with $\varphi(f,\Omega,t)$ the independent and uncorrelated plane wave phase such that $|\Phi(f,\Omega,t)| = 1\ \forall\ f,\Omega,t$ and $\mathrm{E}\{\Phi(f,\Omega,t)\Phi^*(f,\Omega',t)\} = \delta_{\Omega,\Omega'}$ (with δ representing the Kronecker delta and $\mathrm{E}\{.\}$ mathematical expectation), and $b_l(f)$ are the aforementioned array mode strengths (or holographic functions). It can be shown that this leads to a spatial coherence of $\gamma_{l,m;l'm'}^{\text{diff}}(f,t) = 0\ \forall\ (l,m) \neq (l',m')$ (Jarrett *et al.*, 2012) due to the orthogonality of the spherical harmonics and the spatial independence of the plane wave phases.

On the other hand, synthesizing a zero-mean Gaussian noise of power $P_{l,m}^{\text{diff}}(f,t)$ and random phase $\Phi_{l,m}(f,t) = e^{i\varphi_{l,m}(f,t)}$ per SHD component gives a cross-power spectral density (PSD) of

$$\begin{aligned}\hat{\Psi}_{l,m;l',m'}^{\text{diff}}(f,t) &= \mathrm{E}\left\{\hat{X}_{l,m}^{\text{diff}}(f,t)\hat{X}_{l',m'}^{\text{diff}*}(f,t)\right\}\\ &= P_{l,m}^{\text{diff}}(f,t)P_{l',m'}^{\text{diff}}(f,t)\delta_{l,m;l'm'},\end{aligned} \tag{12}$$

and therefore the same diffuse field spatial coherence:

$$\begin{aligned}\hat{\gamma}^{\text{diff}}_{l,m;l'm'}(f,t) &= \frac{\hat{\Psi}^{\text{diff}}_{l,m;l',m'}(f,t)}{\sqrt{\hat{\Psi}^{\text{diff}}_{l,m;l,m}(f,t)}\sqrt{\hat{\Psi}^{\text{diff}}_{l',m';l',m'}(f,t)}} \\ &= 0 \;\forall\; (l,m) \neq (l',m').\end{aligned} \tag{13}$$

It can also be shown that synthesizing a zero-mean Gaussian noise per SHD component leads to an SHD covariance matrix that approaches the identity matrix in the same way as a diffuse field of $N \gg (L+1)^2$ independent and uncorrelated plane waves, as originally demonstrated by Epain and Jin (2016) for the COMEDIE diffuseness measure, provided a normalized covariance calculation is used (thereby imitating a coherence). Although the use of individual power envelopes per SHD component does not guarantee an ideally diffuse field, it does guarantee at least a fully incoherent field, and is furthermore necessary to account for both the order-dependent frequency response of the SHD components (Daniel and Moreau, 2004) and any deviations from perfect isotropy which could introduce continuity artefacts at $\{P_{\text{noise}}, t_{\text{lim}}\}$ when prolonging the reverberation tail.

C. Summary of denoising process

The full denoising process (outlined in Fig. 1) can thus be summarized as follows:

1. *Measurement artefact reduction.* The procedure described in section II A is applied to the raw ESM recording data of each SMA microphone channel.

2. *Inverse-sweep convolution and SHD transform.* The resulting "cleaned" ESM measurement is convolved with a time-reversed and amplitude-corrected version of the

excitation sweep signal as per Farina (2000) to obtain an IR for each microphone channel. This multi-channel IR is then transformed to the SHD according to the theory outlined in section I A.

3. *Mixing time analysis.* Diffuseness analysis is performed in the SHD, leading to an estimation of the mixing time as presented in section II B 2.

4. *EDR analysis and validation of diffuse field hypothesis.* EDR analysis is performed per SHD component in order to extract the reverberation tail decay envelope parameters ($T_{60}(f)$ and $P_0(f)$) and noise floor limit points $\{P_{\text{noise}}, t_{\text{lim}}\}(f)$. The $t_{\text{lim}}(f)$ values obtained for the omnidirectional $Y_{0,0}$ component are averaged over the audible frequency range in order to estimate the broadband noise floor limiting time and confirm (or invalidate) the diffuse field hypothesis required for tail re-synthesis using a zero-mean Gaussian noise.

5. *Tail re-synthesis.* The late reverberation tail is re-synthesized using a zero-mean Gaussian noise per SHD component, which preserves spatial incoherence as shown above. For every SHD component channel, each frequency bin of the re-synthesized tail is made to decay according to the corresponding parameters extracted from the DRIR, and is then used to replace the corresponding DRIR frequency bin starting at $t_{\text{lim}}(f)$.

III. APPLICATION TO MEASURED DRIR

In this section we show the effects of applying the denoising process described above to DRIRs measured in various locations and conditions. A qualitative overview of the results

is first presented, followed by a brief discussion of methods leading to a more quantitative assessment of the procedure's performance.

A. Measurement artefact reduction

Figs. 4 and 5 illustrate the application of the artefact reduction method described in section II to a single microphone channel of an ESM measurement performed at the Christuskirche in Karlsruhe, Germany (a late 19th-century church with a large open dome-like nave). Fig. 4 (a) shows several impulsive artefacts occuring over the course of the ESM measurement signal (averaged over four repetitions), while Fig. 5 (a) illustrates how these turn into repeated inverse-sweep artefacts when the ESM measurement signal is convolved with the time-reversed and amplitude-corrected excitation signal as per Farina (2000). Figs. 4 (b) and 5 (b) show the effect of the artefact reduction procedure on the ESM measurement signal and resulting IR, respectively. Finally, Figs. 4 (c) and 5 (c) highlight the time-frequency points identified as artefacts as well as their magnitude differences before and after reduction.

The spectrograms shown in Figs. 4 and 5 are obtained by performing a moving time average over 8 frames of short-term Fourier transform magnitudes (with 87.5% overlapping frames of 1024 samples at a 48 kHz sampling rate, this corresponds to a total averaging length of 40 ms).

The removal of the inverse-sweep-type artefacts revealed in Fig. 5 is crucial in ensuring that the reverse-integration of the IR's noise floor approaches the theoretical profile fitted to identify the noise floor limit point $\{P_{\text{noise}}, t_{\text{lim}}\}(f)$, as in Fig. 2 (see section II B 1). To

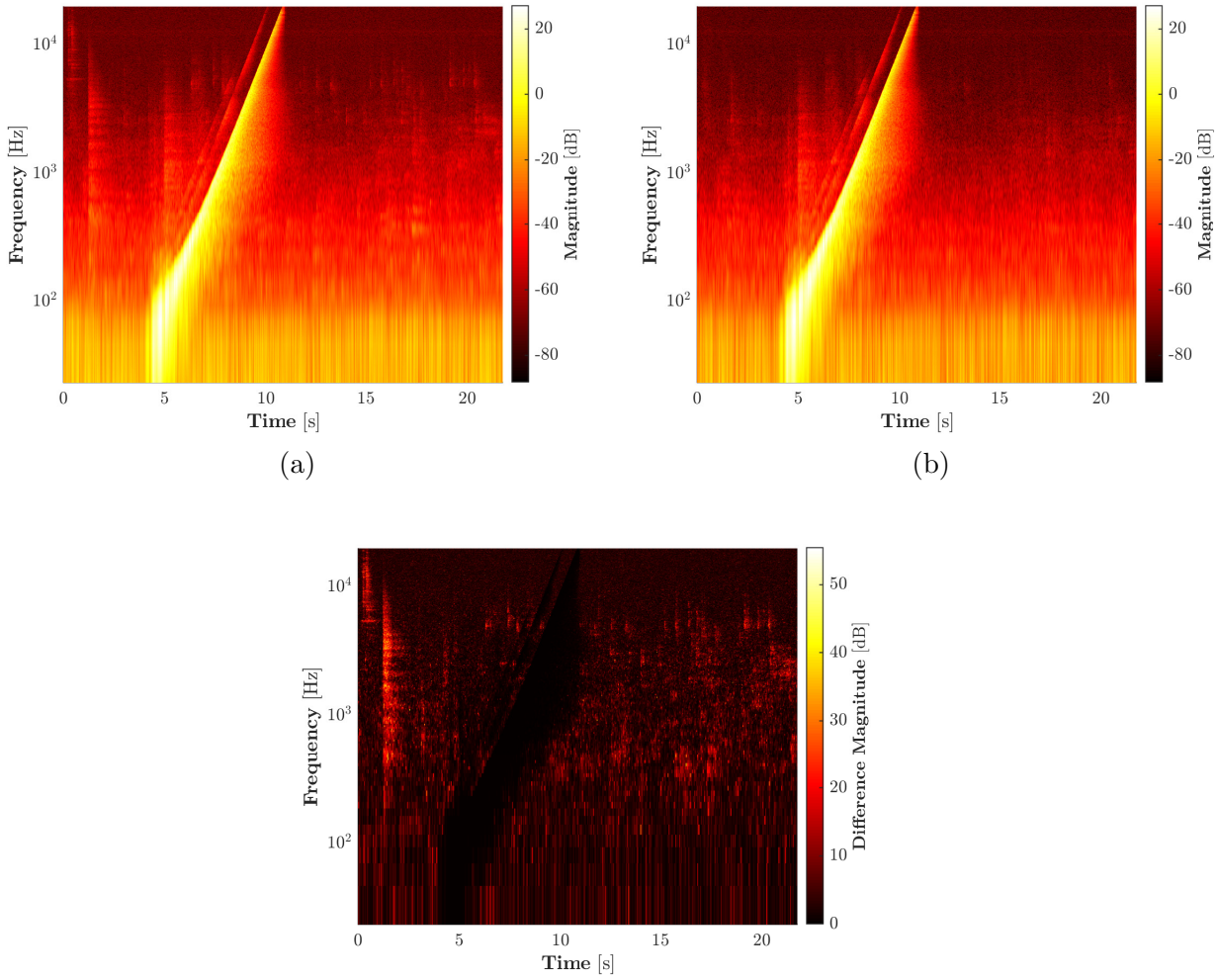


FIG. 4. (Color online.) Artefact reduction applied to a single microphone channel of an ESM measurement performed at the Christuskirche in Karlsruhe, Germany, using an mh acoustics Eigenmike®. (a) Spectrogram of the raw ESM measurement signal (averaged over four repetitions), with several impulsive sounds present. (b) Spectrogram of the ESM measurement signal after artefact reduction. (c) Spectrogram difference between (a) and (b).

further illustrate this, Fig. 6 compares the EDR profile from the Christuskirche DRIR's $Y_{0,0}$ component for one frequency (2461 Hz) before and after application of the artefact reduction process.

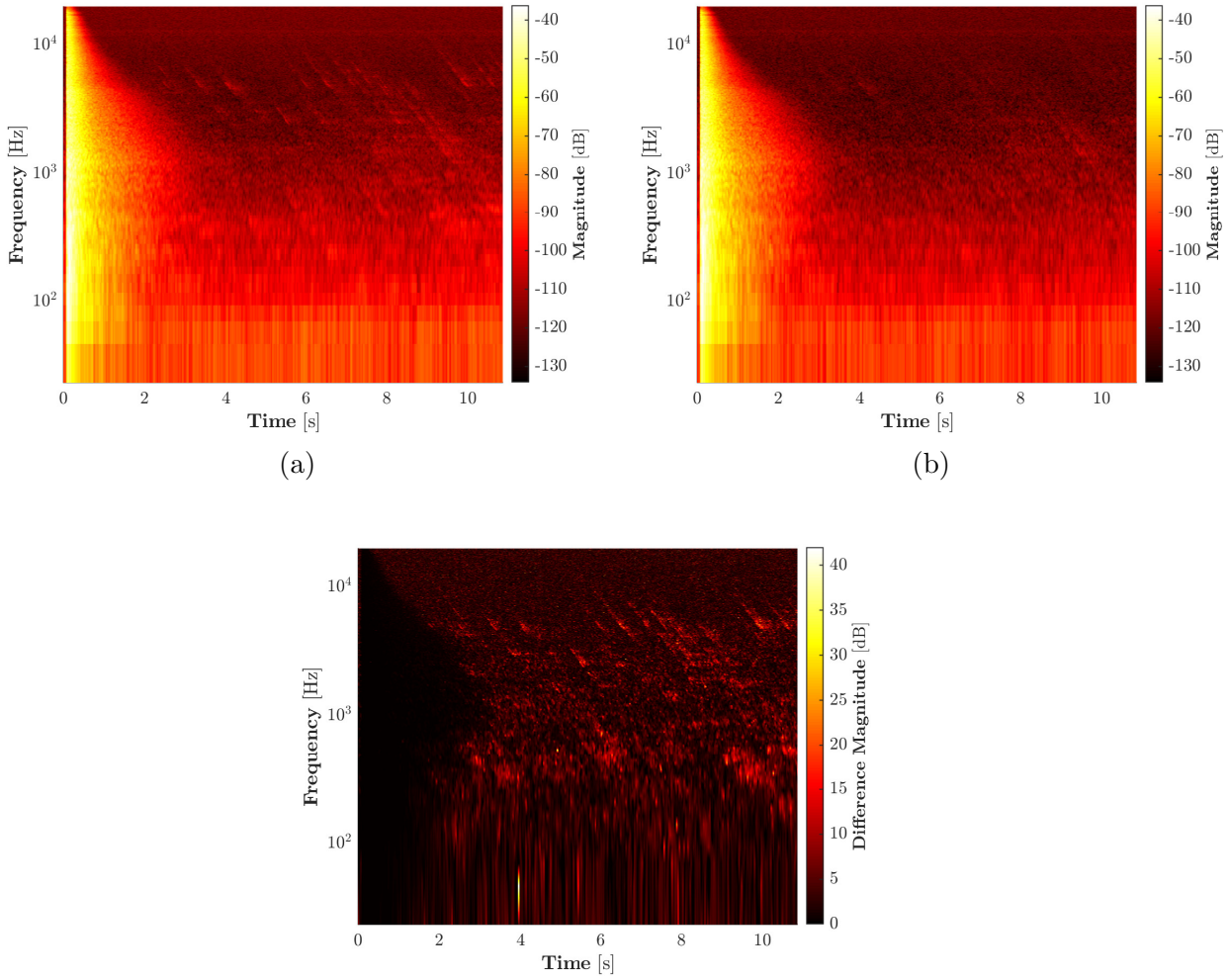


FIG. 5. (Color online.) Artefact reduction applied to a single microphone channel of an ESM measurement performed at the Christuskirche in Karlsruhe, Germany, using an mh acoustics Eigenmike®. (a) Spectrogram of the original IR, after inverse-sweep convolution with the raw ESM measurement signal (without artefact reduction). (b) Spectrogram of the IR obtained by inverse-sweep convolution with the artefact-reduced ESM measurement signal. (c) Spectrogram difference between (a) and (b).

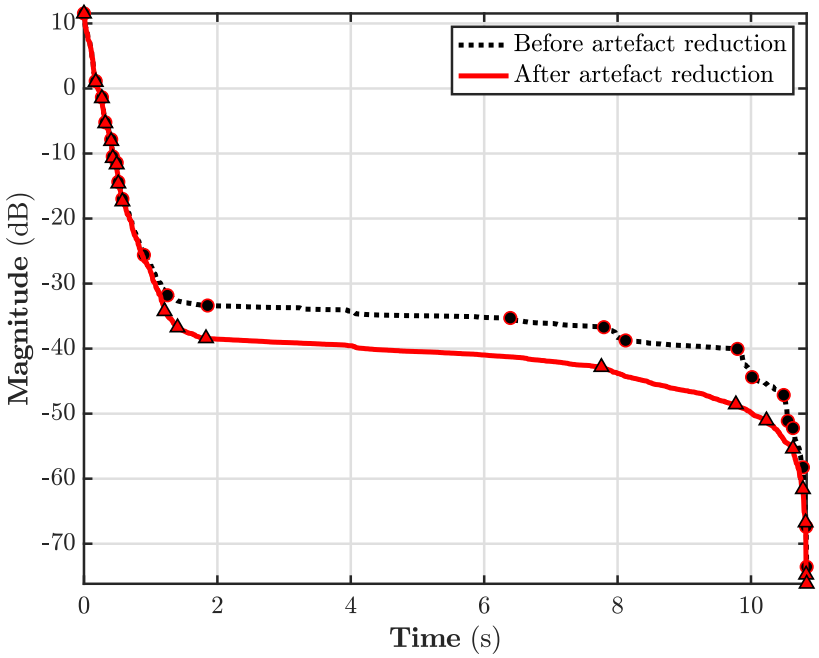
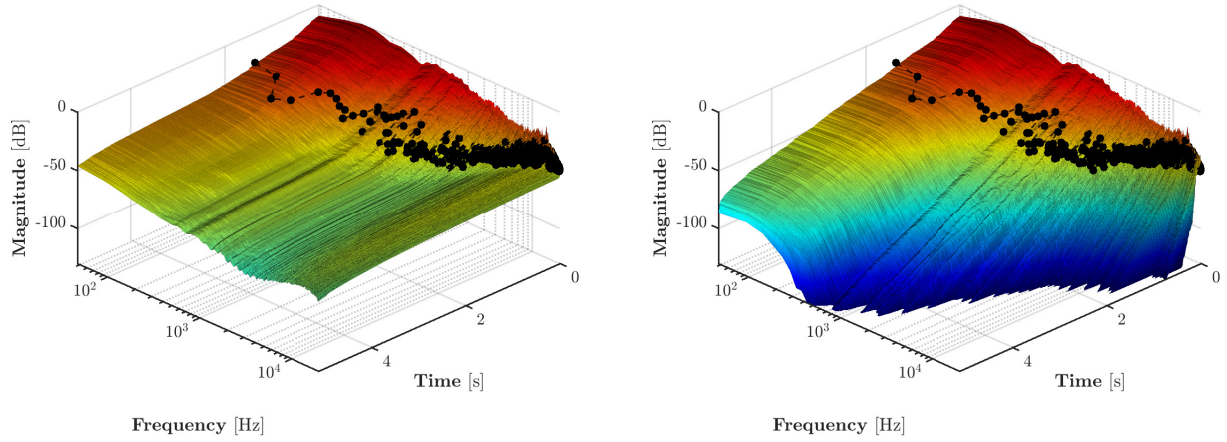


FIG. 6. (Color online.) EDR profile of the Christuskirche DRIR's omnidirectional $Y_{0,0}$ component for one frequency (2461 Hz); before (black dashed line) and after (red solid line) artefact reduction. Circle and triangle markers represent adaptive Ramer-Douglas-Peucker segmentation points (see section II B 1).

Finally, in an attempt to quantify the amount of artefact reduction, we define an *artefact-to-total-energy ratio* as the total artefact energy (i.e. the energy of spectral outliers according to the definition given in section II A) versus the total signal energy in a given frame:

$$\eta(t) = \frac{\sum_{k=0}^{K} |\tilde{X}(f_k,t)|^2}{\sum_{k=0}^{K} |X(f_k,t)|^2},$$
$$\tilde{X}(f_k,t) = \begin{cases} X(f_k,t), & |X(f_k,t)| > \xi(f,t) \\ 0 \end{cases}. \tag{14}$$



(a) (b)

FIG. 7. (Color online.) EDRs of the Kraftzentrale DRIR's omnidirectional $Y_{0,0}$ component, (a) before and (b) after reverberation tail re-synthesis. The black dotted line shows the t_{lim} value for each frequency bin.

Thus $\eta(t) = 0$ means that no outliers were found in the time frame, whereas $\eta(t) = 1$ corresponds to an entirely outlying time frame. In the current example (the Christuskirche DRIR), this measure averaged to $\bar{\eta} = 0.274$ over the four sweep repetitions.

B. Reverberation tail re-synthesis

Figure 7 illustrates the effect of the tail re-synthesis procedure on the EDR of the omnidirectional $Y_{0,0}$ component of the Kraftzentrale DRIR. The arbitrary dynamic range for synthesis is chosen to match that of the signal bit depth (193 dB at 32 bits) at the most perceptually important frequencies (again using the ITU-R 468 standard), although Fig. 7 is shown over 130 dB to match the depth of human hearing.

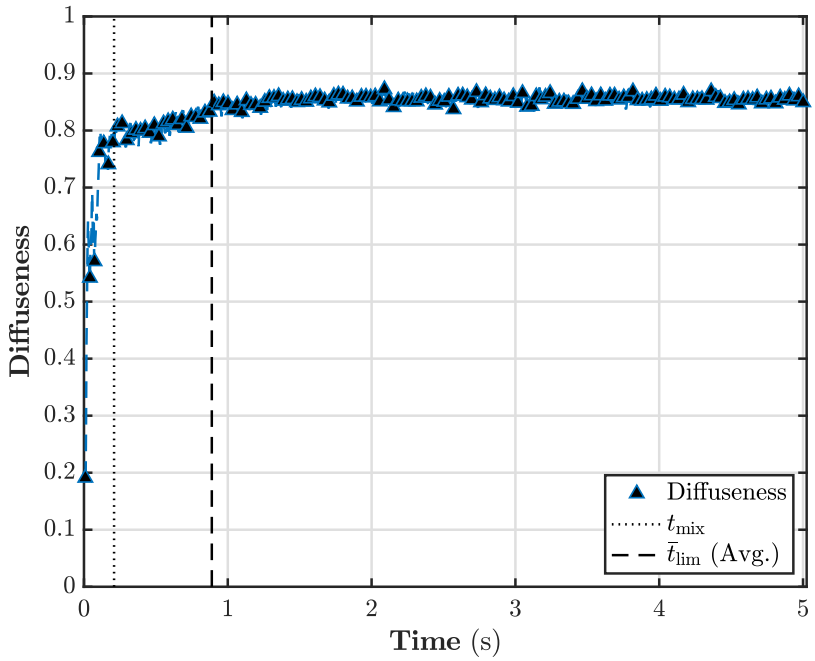


FIG. 8. (Color online.) COMEDIE diffuseness for the Kraftzentrale DRIR, after tail re-synthesis. The t_{mix} and average $\bar{t}_{\mathrm{lim}}$ values (dotted and dashed lines, respectively) are shown as temporal references.

As mentioned throughout this paper, the crucial condition for successfully denoising DRIRs by reverberation tail re-synthesis is that the late field's diffuseness properties must be preserved. To confirm that the proposed denoising procedure achieves this, Fig. 8 shows the COMEDIE diffuseness profile for the Kraftzentrale DRIR: the diffuseness maximum reached at t_{lim} (dotted line) is successfully extended and maintained beyond the average $\bar{t}_{\mathrm{lim}}$ (dashed line). Note that the COMEDIE diffuseness increases slightly from t_{mix} to $\bar{t}_{\mathrm{lim}}$, which may be due to the method's additional sensitivity to ideally diffuse signals versus large numbers of plane waves, as initially noted by Epain and Jin (2016).

IV. CONCLUSION

This paper has adressed the problem of removing the non-decaying noise floor inevitably present in DRIRs measured with SMAs and replacing it with a valid extension of the exponentially-decaying late reverberation tail. Building on previous research showing that this is possible for so-called "mixing" spaces by synthesizing the late reverberation as a zero-mean Gaussian noise and parameterizing its decay envelope by analyzing the EDR, we have demonstrated that performing this synthesis in the SHD guarantees preservation of the late field's spatial incoherence. Additionally, we have shown that including an artefact reduction step before inverse-sweep convolution of the ESM measurement signal improves identification of the noise floor during EDR analysis. As a collateral development, we have also proposed an estimate of the mixing time using measures of DRIR diffuseness in the SHD.

Further work on this topic can be organized around three main themes. First, the question of appropriately determining the number of coupled decays to consider in multi-slope cases must be adressed to avoid over-fitting and ensure that the detected model satisfies coupled-volume theory. Second, cases where the late reverberation field presents highly anisotropic energy distributions must be further investigated, as the spatial symmetry of the SHD will not enable proper re-synthesis using the method presented in this paper. Finally, techniques must be developed for cases where the reverberation tail cannot be considered diffuse before reaching the noise floor, i.e. spaces that cannot be considered traditionally "mixing" and whose late reverberation cannot be modeled as a stochastic process.

ACKNOWLEDGMENTS

This work was funded in part by a doctoral research grant from the École doctorale Informatique, Télécommunications, et Électronique (Paris) at Sorbonne Université. The authors would additionally like to thank Augustin Muller (IRCAM) and Pedro Garcia-Velazquez (Le Balcon) for their extensive DRIR measurements, as well as Franck Zagala (PhD candidate, Sorbonne Université/IRCAM) for having provided the foundations of the EDR analysis and diffuse reverberation tail re-synthesis algorithms.

Ahonen, J., and Pulkki, V. (**2009**). "Diffuseness Estimation Using Temporal Variation of Intensity Vectors," in *Proceedings of the 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, U.S.A., pp. 285–288, doi: 10.1109/ASPAA.2009.5346496.

Akaike, H. (**1974**). "A New Look at the Statistical Model Identification," IEEE Transactions on Automatic Control **AC-19**(6), 716–723, doi: 10.1109/TAC.1974.1100705.

Cabrera, D., Lee, D., Yadav, M., and Martens, W. L. (**2011**). "Decay Envelope Manipulation of Room Impulse Responses: Techniques for Auralization and Sonification," in *Proceedings of Acoustics '11*, Gold Coast, Australia, pp. 52–56.

Carpentier, T., Szpruch, T., Noisternig, M., and Warusfel, O. (**2013**). "Parametric Control of Convolution-Based Room Simulators," in *Proceedings of the 2013 International Symposium on Room Acoustics*, Toronto, Canada.

Cremer, L., Müller, H. A., and Schultz, T. J. (**1982**). *Principles and Applications of Room Acoustics, vol. 1* (Applied Science Publishers, Barking, England).

Daniel, J., and Moreau, S. (**2004**). "Further Study of Sound Field Coding with Higher Order Ambisonics," in *Proceedings of the 116th Audio Engineering Society Convention*, Berlin, Germany, http://www.aes.org/e-lib/browse.cfm?elib=12789.

Driscoll, J. R., and Healy, D. M. J. (**1994**). "Computing Fourier Transforms and Convolutions on the 2-Sphere," Advances in Applied Mathematics **15**, 202–250, doi: 10.1006/aama.1994.1008.

Epain, N., and Jin, C. T. (**2016**). "Spherical Harmonic Signal Covariance and Sound Field Diffuseness," IEEE/ACM Transactions on Audio, Speech, and Language Processing **24**(10), 1796–1807, doi: 10.1109/TASLP.2016.2585862.

Farina, A. (**2000**). "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," in *Proceedings of the 108th Audio Engineering Society Convention*, Paris, France, http://www.aes.org/e-lib/browse.cfm?elib=10211.

Guski, M., and Vorländer, M. (**2014**). "Comparison of Noise Compensation Methods for Room Acoustic Impulse Response Evaluations," Acta Acustica United with Acustica **100**(2), 320–327, doi: 10.3813/AAA.918711.

ISO226:2003 (**2003**). "Acoustics – Normal Equal-Loudness-Level Contours" (International Organization for Standardization, Geneva, Switzerland), https://www.iso.org/standard/34222.html.

ISO8000-2:2009(E) (**2009**). "Quantities and Units – Part 2: Mathematical Signs and Symbols to be Used in the Natural Sciences and Technology" (International Organization for

406 Standardization, Geneva, Switzerland), https://www.iso.org/standard/64973.html.

407 Jarrett, D. P., Thiergart, O., Habets, E. A. P., and Naylor, P. A. (**2012**). "Coherence-
408 Based Diffuseness Estimation in the Spherical Harmonic Domain," in *Proceedings of the*
409 *27th IEEE Convention of Electrical and Electronics Engineers in Israel*, Eilat, Israel, doi:
410 10.1109/EEEI.2012.6377148.

411 Jot, J.-M., Cerveau, L., and Warusfel, O. (**1997**). "Analysis and Synthesis of Room Re-
412 verberation Based on a Statistical Time-Frequency Model," in *Proceedings of the 103rd*
413 *Audio Engineering Society Convention*, New York, U.S.A., http://www.aes.org/e-lib/
414 browse.cfm?elib=7150.

415 Noisternig, M., Carpentier, T., Szpruch, T., and Warusfel, O. (**2014**). "Denoising of Direc-
416 tional Room Impulse Responses Measured with Spherical Microphone Arrays," in *Proceed-*
417 *ings of the 40th Annual German Congress on Acoustics (DAGA)*, Oldenburg, Germany, pp.
418 600–601, http://pub.dega-akustik.de/DAGA_2014/data/articles/000292.pdf.

419 Noisternig, M., Zotter, F., and Katz, B. F. G. (**2011**). "Reconstructing Sound Source Di-
420 rectivity in Virtual Acoustic Environments," in *Principles and Applications of Spatial*
421 *Hearing*, edited by Y. Suzuki, D. Brungart, Y. Iwaya, K. Iida, D. Cabrera, and H. Kato
422 (World Scientific Publishing Co. Pte. Ltd.), pp. 357–373.

423 Polack, J.-D. (**1988**). "La transmission de l'énergie sonore dans les salles," Ph.D. thesis,
424 Université du Maine.

425 Prasad, D. K., Leung, M. K., Quek, C., and Cho, S. Y. (**2012**). "A Novel Framework for
426 Making Dominant Point Detection Methods Non-Parametric," Image and Vision Comput-
427 ing **30**(12), 843–859, doi: 10.1016/j.imavis.2012.06.010.

Rafaely, B. (**2005**). "Analysis and Design of Spherical Microphone Arrays," IEEE Transactions on Speech and Audio Processing **13**(1), 135–143, doi: 10.1109/TSA.2004.839244.

Schroeder, M. R. (**1962**). "Natural-Sounding Artificial Reverberation," Journal of the Audio Engineering Society **10**(3), 219–223.

Schroeder, M. R., and Kuttruff, H. (**1962**). "On Frequency Response Curves in Rooms: Comparison of Experimental, Theoretical, and Monte Carlo Results for the Average Frequency Spacing between Maxima," The Journal of the Acoustical Society of America **34**(1), 76, doi: 10.1121/1.1909022.

Xiang, N., Goggans, P., Jasa, T., and Robinson, P. (**2011**). "Bayesian Characterization of Multiple-Slope Sound Energy Decays in Coupled-Volume Systems," The Journal of the Acoustical Society of America **129**(2), 741–752, doi: 10.1121/1.3518773.