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

Pierre Massé, Thibaut Carpentier, Olivier Warusfel, Markus Noisternig

HAL Id: hal-02443679
https://hal.science/hal-02443679v2

Submitted on 20 Apr 2020

A robust denoising process for spatial room impulse responses with diffuse reverberation tails

Pierre Massé, Thibaut Carpentier, Olivier Warusfel, and Markus Noisternig

A robust denoising process for spatial room impulse responses with diffuse reverberation tails

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

Sciences et Technologies de la Musique et du Son (STMS) – Sorbonne Université, IRCAM, CNRS, 75004 Paris, France

ABSTRACT:

Spatial room impulse responses (SRIRs) measured using spherical microphone arrays are seeing increasingly widespread use in reproducing room reverberation effects on three-dimensional surround sound systems (e.g., higher-order ambisonics) through multi-channel SRIR 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 SRIR 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 incoherent tail parameterized by the SRIR's energy decay envelope. It is shown that performing such tail re-synthesis in the spherical harmonic domain, using an independent zero-mean Gaussian noise for each component, preserves both the reverberation tail's frequency-dependent decay as well as its spatial coherence properties. The proposed process is then evaluated through its application to SRIRs measured in real-world conditions, and finally some aspects of performance and consistency verification are discussed. © 2020 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1121/10.0001070

I. INTRODUCTION

A. SMA analysis in the spherical harmonic domain (SHD)

Spherical microphone arrays (SMAs) 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 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),

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

where $\Omega = (\theta, \phi)$ is a point on the surface of a sphere with fixed radius $r = a$ [in conformity with ISO 8000-2:2009(E) (2009)], $x(f,\Omega,t)$ 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 a 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

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

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 $\mathbf{x}_{\text{SHD}}(f,t)$ 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 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 (RIRs) have long been modelled as an exponentially decaying stochastic

[a] Electronic mail: pierre.masse@ircam.fr

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 stochastically independent (Kuttruff, 2000) and therefore incoherent (Cremer et al., 1982) 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 case of monophonic RIRs with single-slope decays. Furthermore, such decay envelope adjustment methods require strict conditions on the content of the background noise (as noted by Cabrera et al., 2011): since the signal is re-used as is, it must approach a constant-power stationary white noise. This is typically not the case in many "real-world" measurement conditions.

RIRs measured with a SMA are commonly known as spatial room impulse response (SRIR) or directional room impulse response (DRIR), although the latter usually refers to RIRs representing a particular direction (e.g., through beamforming). Preliminary extensions of Jot's tail re-synthesis process to the SRIR 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 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 spatial coherence 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 Sec. II A. The SHD encoding is based on the theory presented in Sec. I, and the EDR analysis is an extension of Jot et al. (1997), detailed in Sec. II B 1. Finally, the main focus of this work is on the spatial coherence and mixing time analysis (Sec. II B 2) and reverberation tail re-synthesis (Sec. 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 occurrence of non-stationary noise events. The first is what is assumed in previous work on noise reduction 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 occurring 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 short-term non-stationary noise events present
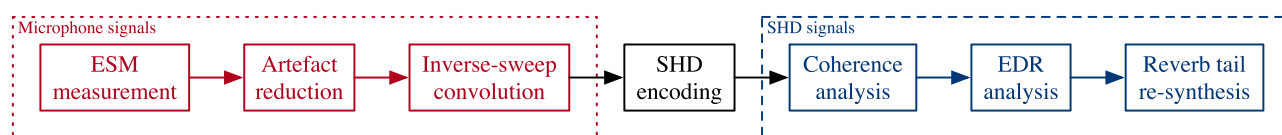

FIG. 1. (Color online) Outline of the proposed SRIR denoising process, from the initial ESM measurement through to the EDR analysis and reverberation tail re-synthesis in the SHD.

J. Acoust. Soc. Am. **147** (4), April 2020

Massé et al.   2251

in the repetitions will inevitably end up in the noise floor of the average, and therefore also in the noise floor of the RIR obtained by convolution with the time-reversed and amplitude-corrected excitation signal. 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 Sec. II B 1 below).

Attempting to reduce the relative amplitude of the artefacts by greatly increasing the number of repetitions is not only impractical but also increases the risk of breaking the long-term stationarity condition on the measurement environment. Therefore, we seek to minimize the influence of these short-term non-stationary noise events by comparing the magnitude spectrograms of the individual sweep repetitions among each other. Non-negligible positive deviations from the mean magnitude spectrogram are thus used as a discriminating criterion in order to identify artefacts. 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 realization 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 Sec. III A.

B. Reverberation tail analysis and re-synthesis

In this section, we first review the EDR analysis procedure used to extract the reverberation decay parameters, and then present a characterization of the SRIR's mixing time using a measure of the sound field's coherence, 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 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). Developed for dominant point detection in digital image processing, the adaptive RDP algorithm determines the point with the maximum deviation from a linear regression over the curve, and calculates a related tolerance factor.

If the maximum deviation is greater than the tolerance factor, the algorithm is recursively applied to either side of the maximally deviating point. Such adaptive segmentation is crucial to the robustness of the sectioning and fitting procedures described below.

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 In Fig. 2). 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 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 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

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} (dashed-dotted line), and the exponential decay model is fitted between t_{start} and t_{lim}.

2252   J. Acoust. Soc. Am. **147** (4), April 2020

Massé *et al.*

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

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 Λ 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)

$$\widehat{\text{EDR}}(f, t, \Lambda) = \int_t^{\infty} \text{ENV}(f, \tau, \Lambda) d\tau$$
$$= \sum_{i=1}^{C} \frac{P_{0,i}(f)}{2\delta_i(f)} e^{-2\delta_i(f)t}. \qquad (4)$$

A *model error* ϵ_{mod} can be defined as a simple mean-squared error,

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

where $N_{\text{fit}} = n_e - n_s + 1$, with n_s is the discrete time index such that $t_{n_s} = t_{\text{start}}$ and similarly n_e such that $t_{n_e} = t_{\text{lim}}$. 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 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 (Akaike, 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 coupled decays, and $\log(N_{\text{fit}})$ is a regularization term used to promote fits made over longer decay sections (i.e., for two fits with equal likelihood, the one made over a longer section of the EDR bin will be preferred).

2. Coherence analysis and mixing time estimation

As mentioned in Sec. I B, replacing the non-decaying noise floor with a reverberation tail synthesized as an exponentially-decaying zero-mean Gaussian noise assumes that the late sound field described by the impulse response is fully diffuse. This leads to the classic time-frequency limits for stochastic modeling of room reverberation, respectively,

the mixing time and Schroeder frequency (Polack, 1988). The exploration of strategies for denoising in the modal domain below the Schroeder frequency is left to future work; in this paper we will apply the tail re-synthesis process across all frequencies, and note that for most reverberant spaces the Schroeder frequency is low enough that the human auditory system is largely insensitive to the modal reverberation below it. [This can be seen by comparing Schroeder's measure $f_{\text{Sch}} \approx 2000\sqrt{\bar{T}_{60}/V}$, where $\bar{T}_{60}$ is a broadband measure of the reverberation time and V is the volume of the space (Schroeder and Kuttruff, 1962), to equal-loudness contours such as those given by the ISO 226:2003 (2003) standard.]

Defining the mixing time, however, is crucial to the present work. Since a diffuse field must be spatially incoherent (Cremer *et al.*, 1982), we propose using a measure of the sound field's spatial coherence, or rather its "level of incoherence," in order to estimate the moment the SRIR becomes maximally incoherent. Furthermore, in Sec. 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 coherence 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 eigen-decomposition 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 since it can be adapted to analyze spatial coherence without regard to the underlying spatial power distribution (which the DirAc measure, for example, assumes to be perfectly isotropic), by using a normalized covariance calculation analogous to a coherence function. This ensures that analysis accuracy is maintained even when the late reverberation field deviates from ideal isotropic diffuseness conditions, and as such we will refer to the result as a measure of "incoherence" rather than diffuseness. Finally, the COMEDIE measure also increases in accuracy with SHD order (whereas the DirAc measure is limited to first-order signals), has a relatively lightweight implementation, and is independent from additional analyses (whereas the SDR measure requires an estimation of the DOA).

In typical "large" mixing spaces, the COMEDIE incoherence 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 $84\,000\,\text{m}^3$ in volume). Estimating the mixing time then corresponds to identifying the moment the SRIR reaches its maximum incoherence. The idea here is to first characterize the maximum incoherence and then find when
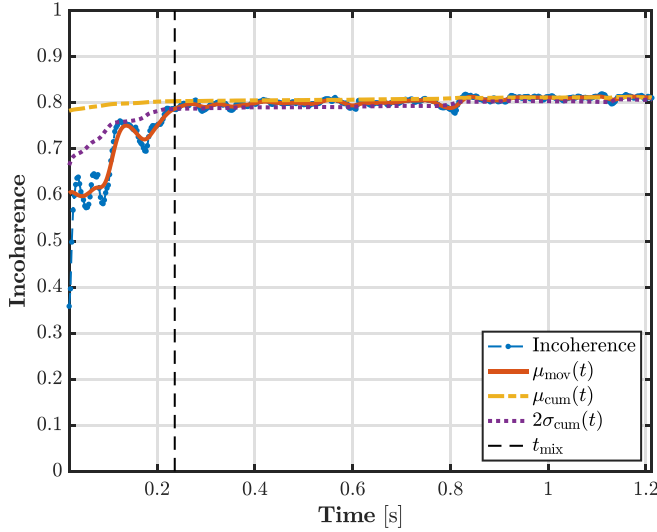
FIG. 3. (Color online) COMEDIE incoherence analysis (Epain and Jin, 2016) and mixing time estimation for a 4th-order SHD SRIR measured at the Kraftzentrale event venue in Duisburg, Germany, using mh Acoustics Eigenmike®. The calculated incoherence curve is smoothed using a Gaussian kernel moving average (μ_{mov}). An inverse cumulative average (μ_{cum}) and standard deviation (σ_{cum}) is further used to identify the onset of the maximum incoherence and thereby estimate the mixing time (t_{mix}).

the SRIR reaches this maximum in a definitive manner after an initial period of instability due to coherent early reflections. This can be done by means of an appropriately-sized moving average,

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

a reverse-cumulative average,

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

and a reverse-cumulative standard deviation,

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

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

$$t_{\text{mix}} = \min(t_{\text{inc}}), \qquad (9)$$

where the time values t_{inc} satisfy

$$\sqrt{[\mu_{\text{mov}}(t_{\text{inc}}) - \mu_{\text{cum}}(t_{\text{inc}})]^2} \leq 2\sigma_{\text{cum}}(t_{\text{inc}}). \qquad (10)$$

Additional checks can subsequently be performed to ensure that no μ_{mov} values are below a certain threshold from μ_{cum} after this time (e.g., corresponding to late-arriving discrete echoes), adjusting t_{mix} to a satisfying t_{inc} value if necessary. Further validation tests on the value of the maximum incoherence may also be included (e.g., an incoherence maximum below 0.5 may not be considered "maximally incoherent").

We now need to define a condition for re-synthesizing the reverberation tail using a zero-mean Gaussian noise: if the SRIR 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 $t_{\text{lim}}(f)$. To get a global value for the noise floor limiting time, we use the $t_{\text{lim}}(f)$ values determined for the SHD-encoded SRIR'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 $\bar{t}_{\text{lim}}$, and the condition can then be written $t_{\text{mix}} < \bar{t}_{\text{lim}}$. 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 $t_{\text{lim}}(f)$ 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 $P_{\text{noise}}(f)$ values]. If the condition is not verified, however, alternative methods of noise reduction must be considered (see the conclusion in Sec. IV below).

3. Incoherent tail synthesis

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

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

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 $\mathrm{E}\{\cdot\}$ 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 of

$$\widehat{\Psi}_{l,m;l',m'}^{\text{diff}}(f,t) = \mathrm{E}\left\{\widehat{X}_{l,m}^{\text{diff}}(f,t)\widehat{X}_{l',m'}^{\text{diff}*}(f,t)\right\}$$
$$= P_{l,m}^{\text{diff}}(f,t)P_{l',m'}^{\text{diff}}(f,t)\delta_{l,m;l'm'}, \qquad (12)$$

and therefore the same field spatial coherence as a diffuse field,

$$\widehat{\gamma}_{l,m;l'm'}^{\text{diff}}(f,t) = \frac{\widehat{\Psi}_{l,m;l',m'}^{\text{diff}}(f,t)}{\sqrt{\widehat{\Psi}_{l,m;l,m}^{\text{diff}}(f,t)}\sqrt{\widehat{\Psi}_{l',m';l',m'}^{\text{diff}}(f,t)}}$$
$$= 0 \,\forall\, (l,m) \neq (l',m'). \qquad (13)$$

In other words, a diffuse field is fully incoherent in the SHD, and subsequently synthesizing a zero-mean Gaussian noise per SHD component produces an identically incoherent field.

It can also be shown (provided again that a normalized covariance is used) that synthesizing a zero-mean Gaussian noise per SHD component leads to a SHD covariance matrix that approaches the identity matrix in the same way as an incoherent field of $N \gg (L+1)^2$ independent plane waves, as originally demonstrated by Epain and Jin (2016) for the COMEDIE diffuseness measure using an isotropic field. As such, the use of individual power envelopes per SHD component does not guarantee an ideally isotropic diffuse field in and of itself, but it does guarantee at least a fully incoherent field. Individual power envelopes are furthermore necessary to account for both the order-dependent frequency response of the SHD components (Daniel and Moreau, 2004) as well as any deviations from perfect isotropy, in which case imposing fully diffuse power envelopes could introduce discontinuity artefacts at the noise floor limit points when prolonging the reverberation tail (see Sec. III B below).

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 Sec. II A is applied to the raw ESM recording signal 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 a RIR for each microphone channel. This multichannel RIR is then transformed to the SHD according to the theory outlined in Sec. I A.

(3) *Mixing time analysis.* Coherence analysis is performed in the SHD, leading to an estimation of the mixing time as presented in Sec. 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 resynthesis using a zero-mean Gaussian noise. Note that since the coherence analysis does not account for the isotropy condition of diffuse fields, a verification of P_0 values per order may be necessary to ensure that their variance is not "too large."
(5) *Tail re-synthesis.* The late reverberation tail is resynthesized 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 SRIR, and is then used to replace the corresponding SRIR frequency bin starting at $t_{\text{lim}}(f)$.

III. APPLICATION RESULTS

In this section we show the effects of applying the denoising process described above to SRIRs 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

Figure 4 illustrates the application of the artefact reduction method described in Sec. 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). Figure 4(a) shows several impulsive artefacts occurring over the course of the ESM measurement signal (averaged over four repetitions), while Fig. 4(b) 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). Figures 4(c) and 4(d) show the effect of the artefact reduction procedure on the ESM measurement signal and resulting RIR, respectively. Finally, Figs. 4(e) and 4(f) highlight the time-frequency points identified as artefacts as well as their magnitude differences before and after reduction.

The spectrograms shown in Fig. 4 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).

J. Acoust. Soc. Am. **147** (4), April 2020
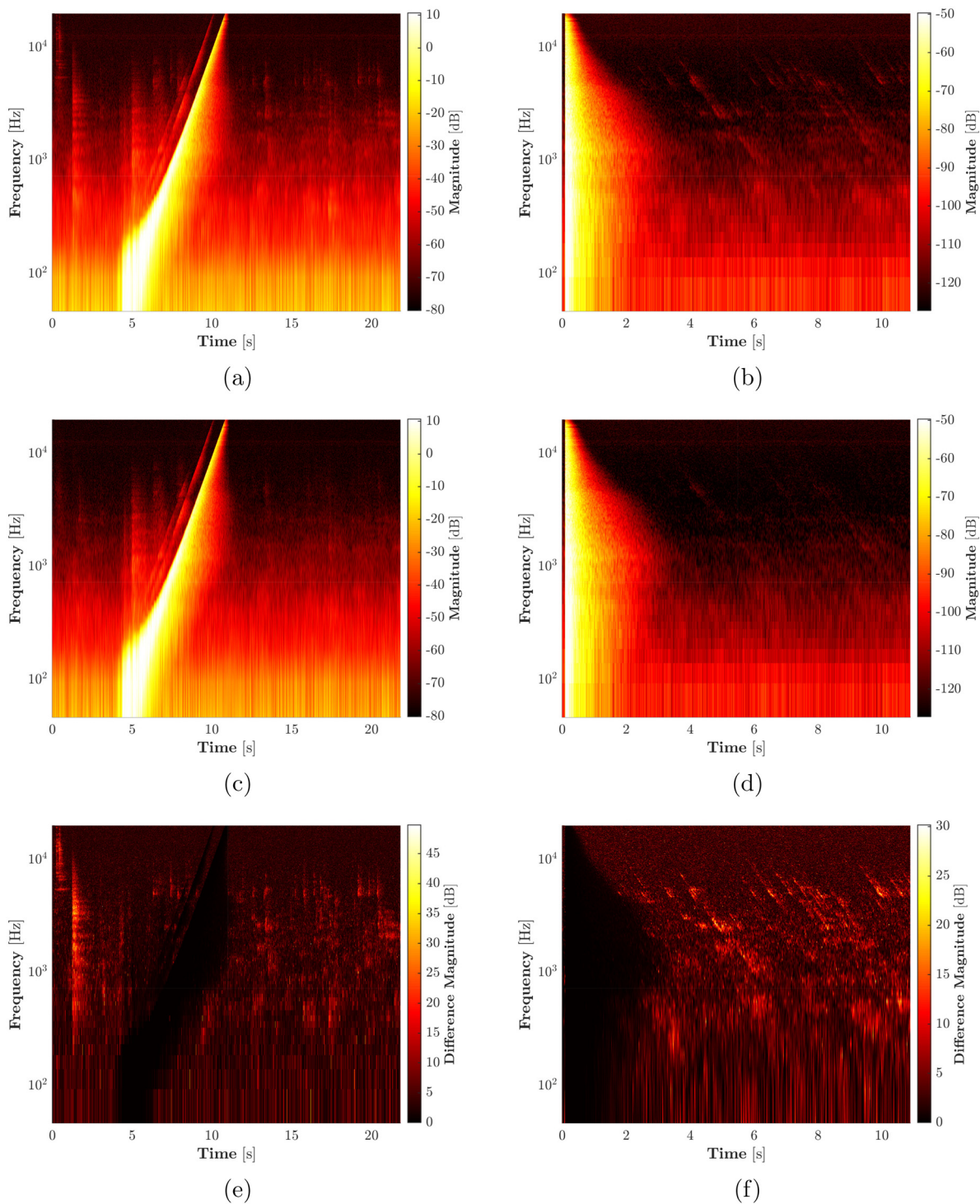
Massé *et al.* 2255

FIG. 4. (Color online) Artefact reduction applied to a single microphone channel of an ESM measurement performed at the Christuskirche in Karlsruhe, Germany, using mh Acoustics Eigenmike®. (a) Spectrogram of the raw ESM measurement signal (averaged over four repetitions), with several impulsive sounds present. (b) Spectrogram of the original RIR, after inverse-sweep convolution with the raw ESM measurement signal (without artefact reduction). (c) Spectrogram of the ESM measurement signal after artefact reduction. (d) Spectrogram of the RIR obtained by inverse-sweep convolution with the artefact-reduced ESM measurement signal. (e) Spectrogram difference between (a) and (c). (f) Spectrogram difference between (b) and (d).

The removal of impulsive measurement noises and the subsequent reduction in inverse-sweep-type artefacts revealed in Fig. 4 is crucial in ensuring that the reverse-integration of the RIR's noise floor approaches the theoretical profile fitted to identify the noise floor limit point $\{P_{\mathrm{noise}}, t_{\mathrm{lim}}\}(f)$, as in Fig. 2 (see Sec. II B 1). To further illustrate this, Fig. 5 compares the EDR profile from the Christuskirche SRIR's $Y_{0,0}$ component for one frequency (2461 Hz) before and after application of the artefact reduction process: not only is the SNR increased but the number
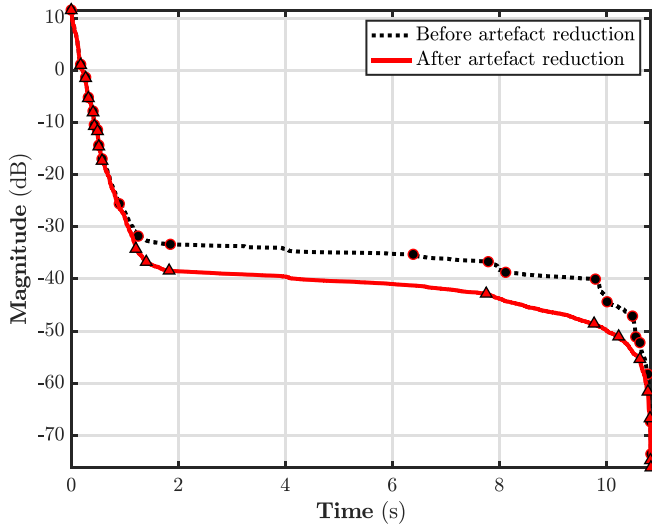
FIG. 5. (Color online) EDR profile of the Christuskirche SRIR's omnidirectional $Y_{0,0}$ component for one frequency (2461 Hz); before (black dashed line) and after (red solid line) artefact reduction. Circle and triangle markers represent adaptive RDP segmentation points (see Sec. II B 1).

of "accidents" in the curve (or deviations from the theoretical reverse-integration of a constant power) is also decreased.

Finally, in an attempt to quantify the amount of artefact reduction, we define an *artefact-to-total-energy ratio* as the total artefact energy removed by replacing detected outlying points with the mean magnitude value (according to the definition given in Sec. II A) versus the total signal energy in a given frame

$$\eta(t) = \frac{\sum_{k=0}^{K} |\tilde{X}(f_k, t)|^2}{\sum_{k=0}^{K} |X(f_k, t)|^2},$$

$$\tilde{X}(f_k, t) = \begin{cases} X(f_k, t) - \mu(f_k, t), & |X(f_k, t)| > \xi(f_k, t) \\ 0. \end{cases} \quad (14)$$

Thus $\eta(t)$ represents the relative total energy removed in each time frame during artefact reduction. In the current example (the Christuskirche SRIR), this measure averaged to $\bar{\eta} = 0.129$ or -17.8 dB over the four sweep repetitions.

B. Reverberation tail re-synthesis

The re-synthesis of the late reverberation tail is evaluated here in two steps: the consistency of the EDR analysis is first verified on deliberately noised simulations, and then results obtained on an SRIR measured in real-world conditions are subsequently presented.

1. Simulated late reverberation tails

The consistency of the EDR analysis process presented in Sec. II B 1 can be simply verified by applying it to simulated late reverberation tails to which a constant-power

diffuse noise floor has been deliberately added. To simplify the evaluation of the analysis results, the simulation and subsequent analysis are performed in a broadband manner, i.e., without introducing or taking into account any frequency dependence in either P_0 or T_{60}. In a filter-bank view, this can be seen as reducing the evaluation to that of a single frequency bin.

The reverberation tails are simulated by first synthesizing a zero-mean Gaussian noise per SHD component, to which an exponential energy decay envelope is then applied. Two types of reverberation tails are created: a perfectly diffuse late field is obtained by setting equal P_0 and T_{60} values over all SHD components, and a slightly anisotropic deviation from ideal diffuseness is simulated by introducing random fluctuations of up to 10 dB over the P_0 values and up to 10% over the T_{60} values.

Table I displays the error measurements obtained between the EDR analysis results and the original values used in the simulations ($P_0 = 0$ dB, $T_{60} = 3$ s, $\{P_{\text{noise}}, t_{\text{lim}}\} = \{-50 \text{ dB}, 2.5 \text{ s}\}$). These show that not only does the proposed process guarantee preservation of the late reverberation tail's spatial coherence properties, as demonstrated in Sec. II B 3, it will additionally tend to preserve a perfectly diffuse field (to within $3 \times 10^{-4} \pm 0.05$ dB initially, subject to an evolution of $0.3 \pm 0.7\%$ over a 60 dB decay). Furthermore, the procedure is able to reproduce deviations from ideal isotropy without substantial increases in errors; this ensures that discontinuity artefacts will be avoided when replacing the noise floor from t_{lim} onwards in real-world SRIR measurements, where these types of deviations are inevitable.

Finally, it should be noted that the large errors in the t_{lim} results are due to the fact that the EDR analysis process is deliberately conservative with respect to the detection of the noise floor (see Sec. II B 1), both in order to avoid the influence of the reverse-integration and to further avoid discontinuities when replacing the noise floor.

2. Measured SRIR

We now present results obtained with the SRIR measured at the Kraftzentrale venue presented in Sec. II B 2. Figure 6 illustrates the effect of the tail re-synthesis procedure on the EDR of the omnidirectional $Y_{0,0}$ component; 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

TABLE I. Error measurements for broadband EDR analysis performed on simulated noisy late reverberation tails. The "anisotropic" tail is obtained by introducing random fluctuations over the SHD of up to 10 dB on P_0 and 10% on T_{60}. The noise power is set at -50 dB.

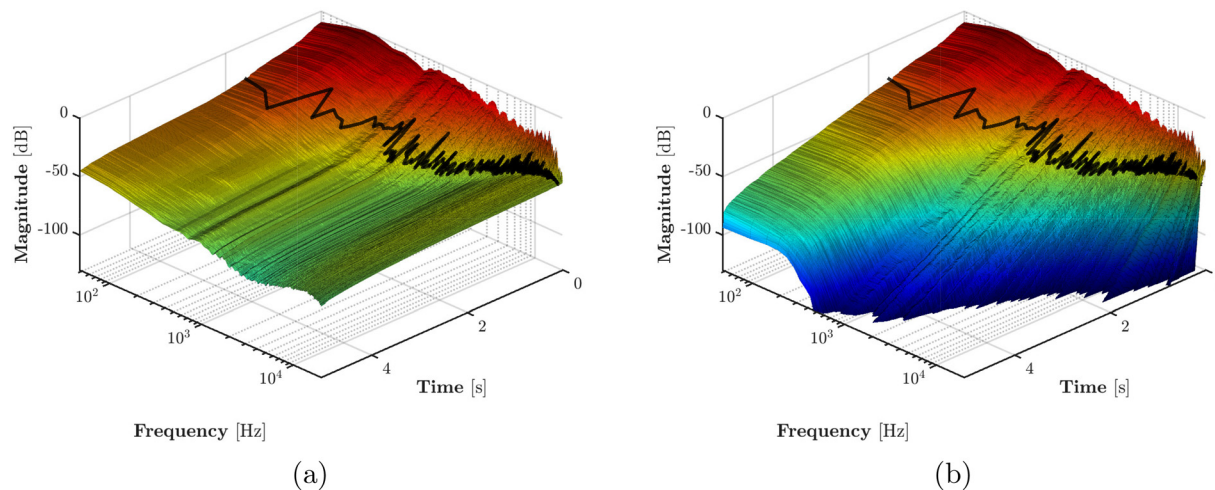|	P_0 [dB]		T_{60} [%]		t_{lim} [%]	
	Iso.	Aniso.	Iso.	Aniso.	Iso.	Aniso.
Mean Error	3.05×10^{-4}	5.94×10^{-3}	0.317	0.592	-4.5	-5.3
St. Dev.	0.0525	0.231	0.682	1.04	6.69	5.58

FIG. 6. (Color online) EDRs of the Kraftzentrale SRIR'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.

468 standard), although Fig. 6 is shown over 130 dB to match the depth of human hearing.

As mentioned throughout this paper, the crucial condition for successfully denoising SRIRs by reverberation tail re-synthesis is that the late field's coherence properties must be preserved. To confirm that the proposed denoising procedure achieves this, Fig. 7 shows the COMEDIE incoherence profile for the Kraftzentrale SRIR: the incoherence maximum reached at t_{lim} (dotted line) is successfully extended and maintained beyond the average $\bar{t}_{lim}$ (dashed line). Note that the COMEDIE incoherence increases slightly from t_{mix} to to $\bar{t}_{lim}$, 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).

The model decay error ϵ_{mod} used in the EDR analysis procedure [Eq. (5), Sec. II B 1], can be examined in order to assess the performance of the denoising process on a given

SRIR measurement, at least in terms of the quality of the fitted envelope which prolongs the reverberation tail. Indeed, if the measurement artefact reduction successfully ensures that the noise floor of the SRIR approaches an ideal constant-power background, the procedure should accurately detect the noise floor limit point $\{P_{noise}, t_{lim}\}$ per frequency bin and per SHD component, thereby allowing the decay error to be minimized. Conversely, should any of these steps perform less than ideally (indicating perhaps that the measurement does not satisfy some or all of the modeling assumptions made along the way), the decay error can be expected to increase.

Figure 8 shows the maximum and mean ϵ_{mod} over all SHD components at each frequency bin for the SRIR measured at the Kraftzentrale. Most frequency bins below ~10 kHz have a maximum error of less than 1 dB/frame, and the mean error only becomes important when
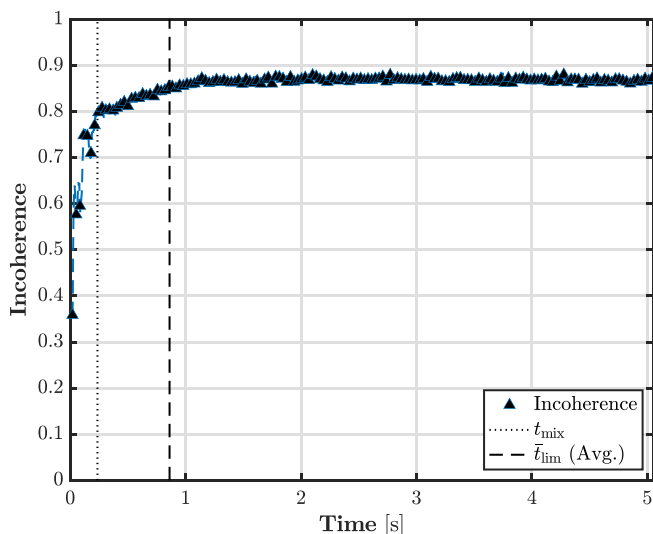

FIG. 7. (Color online) COMEDIE incoherence for the Kraftzentrale SRIR, after tail re-synthesis. The t_{mix} and average $\bar{t}_{lim}$ values (dotted and dashed lines, respectively) are shown as temporal references.
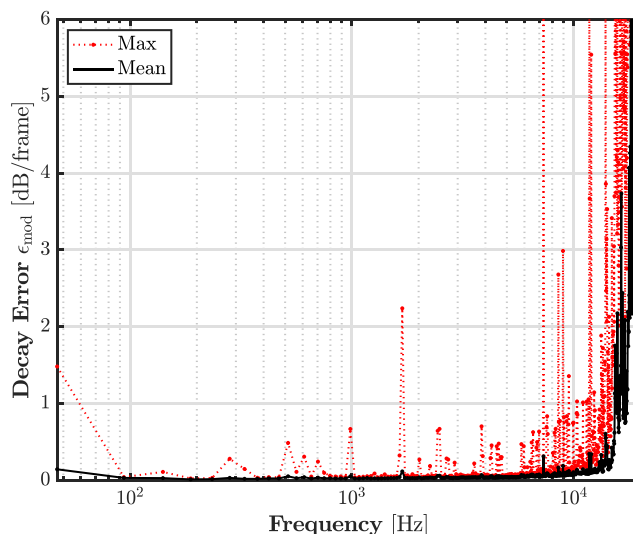
FIG. 8. (Color online) Model decay error ϵ_{mod} for the Kraftzentrale SRIR. The red dotted and black solid lines, respectively, show the maximum and mean error over all SHD components at each frequency bin.

approaching 20 kHz. As can be seen from Fig. 6(a), this corresponds to frequencies where the SRIR's SNR rapidly decreases, and consequently fitting an exponential decay becomes increasingly difficult (reasons for this include the Eigenmike transducers' inherent frequency response and SNR at high frequencies, as well as the effect of air absorption on the RIR).

The model decay error therefore provides a general overview of the proposed procedure's performance and rapidly reveals any debilitating issues in a given application. However, it does not give an accurate representation of the perceptual quality of the final denoised SRIR, as it cannot predict the effect a certain error value (e.g., the $\sim$2 dB/frame error at $\sim$1.75 kHz in Fig. 8) will have on the resulting perceived "sound" of the SRIR. Further work should attempt to combine the model decay error with the time-frequency decay envelope in order to assess the perceptual "importance" of a given ϵ_{mod} value (e.g., errors at frequencies with high initial powers and slow decays are much more perceptually important than errors at frequencies with low initial powers and fast decays).

Finally, it should also be noted that some ambiguity may arise when working with multiple-slope decays, in which case the "knee" between the exponential decay and constant-power noise floor (due to the EDR's reverse-integration formalism) may be mistaken as a separate curve, resulting in an erroneously low ϵ_{mod}. Improvements with respect to this point are currently being pursued, including verifying the results of the modeling process against coupled-volume theory (e.g., Cremer et al., 1982).

IV. CONCLUSION

This paper has addressed the problem of removing the non-decaying noise floor inevitably present in SRIRs 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 a measure of SRIR incoherence in the SHD, and briefly discussed error and performance metrics and analysis for the proposed method.

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 addressed 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 incoherent 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.

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, NY, pp. 285–288.

Akaike, H. (**1974**). "A new look at the statistical model identification," IEEE Trans. Auto. Control **AC-19**(6), 716–723.

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* (Applied Science Publishers, Barking, England), Vol. 1.

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.

Driscoll, J. R., and Healy, D. M. J. (**1994**). "Computing Fourier transforms and convolutions on the 2-sphere," Adv. Appl. Math. **15**, 202–250.

Epain, N., and Jin, C. T. (**2016**). "Spherical harmonic signal covariance and sound field diffuseness," IEEE/ACM Trans. Audio, Speech, Lang. Process. **24**(10), 1796–1807.

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.

Guski, M., and Vorländer, M. (**2014**). "Comparison of noise compensation methods for room acoustic impulse response evaluations," Acta Acust. Acust. **100**(2), 320–327.

ISO 226:2003 (**2003**). "Acoustics—Normal equal-loudness-level contours" (International Organization for Standardization, Geneva, Switzerland).

ISO 8000-2:2009(E) (**2009**). "Quantities and units—Part 2: Mathematical signs and symbols to be used in the natural sciences and technology" (International Organization for Standardization, Geneva, Switzerland).

Jarrett, D. P., Thiergart, O., Habets, E. A. P., and Naylor, P. A. (**2012**). "Coherence-based diffuseness estimation in the spherical harmonic domain," in *Proceedings of the 27th IEEE Convention of Electrical and Electronics Engineers in Israel*, Eilat, Israel.

Jot, J.-M., Cerveau, L., and Warusfel, O. (**1997**). "Analysis and synthesis of room reverberation based on a statistical time-frequency model," in *Proceedings of the 103rd Audio Engineering Society Convention*, New York, NY.

J. Acoust. Soc. Am. **147** (4), April 2020

Massé *et al.*   2259

Kuttruff, H. (**2000**). *Room Acoustics*, 4th ed. (Spon Press, London, UK).

Noisternig, M., Carpentier, T., Szpruch, T., and Warusfel, O. (**2014**). "Denoising of directional room impulse responses measured with spherical microphone arrays," in *Proceedings of the 40th Annual German Congress on Acoustics (DAGA)*, Oldenburg, Germany, pp. 600–601.

Noisternig, M., Zotter, F., and Katz, B. F. G. (**2011**). "Reconstructing sound source directivity in virtual acoustic environments," in *Principles and Applications of Spatial Hearing*, edited by Y. Suzuki, D. Brungart, Y. Iwaya, K. Iida, D. Cabrera, and H. Kato (World Scientific Publishing Co. Pte. Ltd., Singapore), pp. 357–373.

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

Prasad, D. K., Leung, M. K., Quek, C., and Cho, S. Y. (**2012**). "A novel framework for making dominant point detection methods non-parametric," Image Vision Comput. **30**(12), 843–859.

Rafaely, B. (**2005**). "Analysis and design of spherical microphone arrays," IEEE Trans. Speech Audio Process. **13**(1), 135–143.

Schroeder, M. R. (**1962**). "Natural-sounding artificial reverberation," J. Audio Eng. Soc. **10**(3), 219–223, see http://www.aes.org/e-lib/browse.cfm?elib=343.

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," J. Acoust. Soc. Am. **34**(1), 76–80.

Xiang, N., Goggans, P., Jasa, T., and Robinson, P. (**2011**). "Bayesian characterization of multiple-slope sound energy decays in coupled-volume systems," J. Acoust. Soc. Am. **129**(2), 741–752.