A Room Compensation Method by Modification of Reverberant Audio Objects

Dylan Menzies, Philip Coleman and Filippo Maria Fazi

***Abstract*—Conventional channel-based room equalisation can reduce overall colouration caused by the room response, however it cannot separately correct the colouration caused by the late and early parts of the response, or consider the reverberance in the source signal. A room compensation method is developed here for a source signal in which the dry source sound and the associated target reverberant response are encoded separately, which is possible in an object-based audio framework. The target response is modified using the reproduction room response. Subject to some conditions the combined response approximates the target, with accurate early and late equalisations, reverberant balance, and decay timing. Stochastic assumptions are used to simplify the processing, enabling efficient real-time processing of the encoded audio.**

***Index Terms*—Room correction, Room compensation, Room equalisation, Object-based audio, Parametric reverberation**

I. Introduction

Audio engineers have long faced the problem of compensating for the acoustic effects of the reproduction space, so that a reproduction over loudspeakers sounds as close as possible to the original intended production. This is often addressed by filtering the loudspeaker channel feeds, improving the frequency and timing response characteristics over a defined listening region. This is referred to here as *channel-based equalisation*.

Early channel equalisers consisted of multi-band analog filters, and had limited ability to correct temporal distortion. They focused on the large equalisation errors due to low frequency room modes. Averaging equalisation over multiple points provides a way to cover a wider listening region. Digital technology has lead to attempts to invert impulse responses accurately, and reduce time distortion, for example [1; 2; 3]. Sound field synthesis methods, using many loudspeakers, have been used to cancel reflections within a region [4] and even exploit reflections to directly provide additional source freedom for target reproduction [6]. For a recent extensive recent review of room compensation methods see [5].

To achieve equalisation over a larger region, or with greater temporal accuracy generally requires more loudspeakers and more precise calibration. However, even then it is impractical to invert late diffuse high frequency energy in the room response, because it varies rapidly over space. The early and late parts of the response often have significantly different colouration, since the early part is dominated by interference and resonance effects, whereas the later part may be significantly affected by absorption from the air and surfaces in the room. The effects of the two parts on the reproduction are perceived separately because the early part is transient and the late part is smeared in time [7]. The relative magnitude of the parts, or direct to reverberant ratio, affects perception of source distance. Late reverberation compensation has not been addressed effectively by conventional methods.

Audio productions often contain reverberation that is part of natural recordings, or added using reverberation processing. When a signal with reverberation is played into a room with its own reverberation, the two reverberations combine to produce a single quasi reverberant response, called the *room-in-room response (RIRR* [8]. Typically the late part of the RIRR has increased total energy compared to the early part, is biased towards low frequency, and has greater spectral variance which contributes to colouration [9]. These negative effects are apparent for typical domestic listening conditions, and increase for larger listening spaces such as auditoriums.

If the reverberant response and the source sound are separately available at the point of reproduction then a new strategy for room compensation is possible, in which the reverberant component of the source is modified at the reproduction point. This gives more freedom than channel-based compensation, because different parts of the reverberant component can be modified independently before playback. It may then be possible to make the early and late parts of the RIRR match the target. An *object-based* audio representation is one that contains constituent parts, or objects, of an audio production, that are then combined at the point of reproduction, [10]. This provides a way to send reverberant responses that can be modified during reproduction. The *Reverberant Spatial Audio Object (RSAO)* is an example of an object encoding of a reverberant response [11], designed as part of an object-based workflow [12].

Grosse has considered a related problem [7] looking at how close and far microphone signals from a source recorded in a reverberant room can be processed to feed a special loudspeaker configuration in another room. The aim is to match the perceptual characteristics of the binaural signals of the listener in the reproduction room with those in the production room. For this system the room-in-room response cannot be made less reverberant than the room response, meaning the total early energy cannot be increased relative to the total late energy. This restriction is not severe, as reverberation is common in productions, and listening rooms have relatively little reverberance.

The scenario considered here is simpler, consisting of re-

D. Menzies and F.M. Fazi are with the University of Southampton

P. Coleman is with the Institute of Sound Recording, University of Surrey, GU2 7XH, UK

production over a single loudspeaker. This allows direct application to practical situations, although the reduced freedom poses challenges. The aim is to calculate playback responses from the target and room responses that make the reproduction at the listener perceptually similar to the original production. Following the previous observations about RIRR perception, the main criteria for comparison will be the early and late energies, and the late decay rate.

A random statistical model for reverberation is used, which greatly simplifies calculation, while giving acceptable accuracy. This approach lends itself to compact parametric representations of room reverberation, such as RSAO encoding, that can be processed efficiently, and incorporated within an object-based framework. Efficiency is an important factor for a real-time reproduction system. Such compact parametric representations have a long history of use in computer games and music production [13]. In addition, a more precise, but less convenient, correction method is presented, that takes into account covariance between responses. An early version of the object-based compensation method given here was outlined previously [14].

Faller has recently described a room compensation method [15], that works by deverberating the source production to compensate for the room reverberation that is added in playback, without need for separate reverberation objects. The method presented here has a similar effect. Because the target responses are available, there is more freedom in creating the playback signal, including separately processing multiple target responses within the target production.

In summary the contributions in this work are

1) Efficient object-based room compensation where the playback response is produced by modifying the early and late parts of each target reverberant response based on the listening room response, so that the early and late parts of the reproduced response are equalised.
2) A more precise, but less efficient, compensation method in which densities are calculated depending on both target and room responses.
3) Application of the method to parametrically encoded reverberation.
4) Improved understanding about interaction of stochastic signals

The article proceeds as follows. In Section II perceptual features of room response reproduction are reviewed. Then the physical interaction of reproduced reverberant sound with the natural room acoustic of the reproduction room is analysed using a statistical model. Based on these findings, a perceptual room compensation method is presented, in which the the playback response is produced by separately filtering the early and late parts of the target response. In Section III-A it is shown how to apply the compensation method to an object-based encoding system that includes reverberant information in parametric form. The object-based compensation method is tested objectively in V using some examples. The perceptual parameters that are used to design the method are calculated, and improvements are demonstrated in comparison with channel based equalisation. In Section VI these results are supported by subjective tests using the same examples.

Fig. 1: Schematic for the room compensation problem. Each response has an early and late part.

II. Room compensation

When a reverberant sound is played into a reproduction room through a loudspeaker, the resulting sound at the listener is the convolution of the source sound, the associated playback reverberant response, and the room response. The room compensation problem is to derive the playback response from the reproduction room response and the target reverberant response, set by the producer. This relationship is shown schematically in Fig. 1, which shows the convolution of the playback response, to be determined, with the room response, to give the target response set by the producer. Each response has an early and late part.

As discussed previously, signal based broadband compensation is not feasible for a typical room response, so instead we aim, from the outset, for perceptual equality rather than signal equality. Criteria for perceptual equality are identified in Sec. II-A. In Sec. II-B the perceptual equality condition is then used to derive a playback response by modifying the target response, based on the room response. The problem is complicated by the cross mixing of the early and late parts by the convolution shown in Fig. 1. This paper does not consider spatial reproduction using multichannel systems.

A. Room response perception

A reverberant response typically consists of an early strong transient signal followed by a late noise-like decay. When convolved with a source signal a clear direct signal is perceived from the early part, mixed with a diffuse signal from the late part. These two signals are perceived separately, so to match the whole response, the perceptual qualities of each part should be matched. This was part of the motivation for separation given by Grosse [7]. The transient response of contributing to direct perception is usually contained in 10-20 ms, so this range is choice for the early part of the response. This is at the lower end of the range used by Grosse. He varies the early duration to tune the energy decay curve, however this does not have the same effect in the method developed here. Increasing the early duration mixes in the perceived reverberance, and separation is reduced. Our priority is to choose the duration to maximise perceived separation.

This early/late division is closely related to existing acoustic measures, the direct to reverberant ratio (DRR), associated with perceived image distance, and clarity C50, [16], associated with speech intelligibility. Each is based on the ratio of the energy in the early part of the response to the remaining part. For DRR for early duration is a few ms, and for C50 it is 50 ms. The early duration used here is between these, and provides a compromise that leads to the best compensation results.

Studies of audio perception [17], and the subsequent development of perceptual audio coding methods [18], have shown that an audio signal can be approximated using sub-band signals with *critical bandwidths*, sampled every 20ms. The perceived loudness of each sub-band depends only on the total energy in the band, and not the distribution of energy within the band. Sub-bands will be used to represent perceptual content of the early and late parts of reverberant responses.

B. Room-in-room interaction

Although reverberation is a complex deterministic process, simple random statistics can be successfully used to model important features [19]. Statistical models will be used to predict features of the general room-in-room response (RIRR), and build a room compensation method.

We start by calculating the RIRR explicitly for a simple model with uniform absorption across frequency, in order to show some general time domain features of RIRRs. The responses $R_1(t)$ and $R_2(t)$ consist of exponentially decaying noise, gated zero for $t < 0$. The pressure at each time is modelled with an identically distributed random variable [19]. Any two pressures from different responses are independent, but this is not assumed within each response.

$$R_i(t) = \begin{cases} N_i(t)e^{-\beta_i t}, & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{1}$$

where $N_i(t)$ are identically distributed random variables with zero mean $E(N(t)) = 0$ and variance $V(N(t)) = \sigma^2$. Continuous time, rather than discrete time, is used here for calculation convenience. In Appendix A the convolution $R_1 \star R_2$ is shown to be normally distributed at each time, with standard deviation

$$\sigma_c(t) = \begin{cases} \sigma^2 \left[\frac{e^{-2\beta_2 t} - e^{-2\beta_1 t}}{2(\beta_1 - \beta_2)}\right]^{\frac{1}{2}}, & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{2}$$

$\sigma_c(t)$ defines the envelope of the RIRR. Without loss of generality assume $\beta_1 > \beta_2$, i.e. that the response with rate β_2 decays more slowly than that with rate β_1. Then in the late time limit the decay of $\sigma_c(t)$ becomes proportional to $e^{-\beta_2 t}$, the same as the slower decaying initial response. This means that the target decay can be reproduced by using a playback response with the same decay, provided the room decay time is shorter: The shorter decay has a smearing effect on the longer decay, stretching the sharp attack into a smooth attack over a period $\approx 1/\beta_1$ s.

The general case of non-uniform absorption can be approximated by applying the above model in sub-bands, with separate decay rates. The comments above apply for each band.

Natural room decay profiles can deviate significantly from a simple exponential. For example coupled rooms can lead to double exponential decay. However the same general result will be produced. The RIRR will resemble the longer response, whatever the profile, but smeared on the timescale given by the shorter decay.

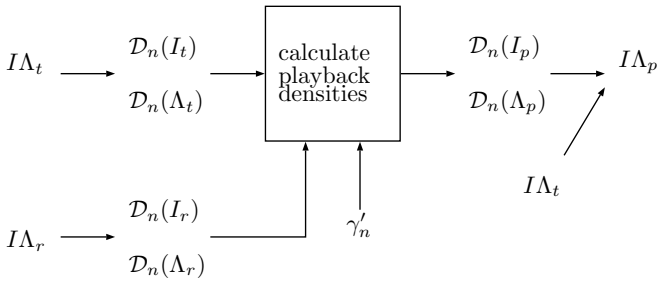


Fig. 2: Overview of the object-based room compensation method. $I\Lambda_t$, $I\Lambda_r$, $I\Lambda_p$ are the target, room and playback impulse responses. $\mathcal{D}$ indicates the densities of the early and late parts, and γ'_n are tradeoff parameters used when full compensation is not possible. The playback densities are determined by (13), (18) and (21)

C. Object-based Compensation

Based on the observations in Sections II-A and II-B, a room compensation method is now presented that simultaneously equalises both the early and late parts of the reproduced response. An overview diagram is shown in Fig. 2. A discrete-sampled time-domain room impulse response is denoted by $I\Lambda$, read as a single signal. Sample indices are omitted . The early and late parts of this response are separately written I and Λ, so that $I\Lambda = I + \Lambda$. The early duration is 20 ms, according to Section II-A. Λ is zero valued over this period. Subscripts t, r and p denote the target, reproduction room, and playback responses. The early and late parts of a convolution product are written using a double subscript, for example the late part of $I\Lambda_p \star I\Lambda_r$ is written Λ_{pr}.

The physical room compensation problem is to find the playback response $I\Lambda_p$ such that convolution with the room response $I\Lambda_r$ produces the target $I\Lambda_t$ response,

$$I\Lambda_p \star I\Lambda_r = I\Lambda_t \tag{3}$$

Expanding the left hand side, early and late parts can be identified,

$$(I_p \star I_r) + (\Lambda_p \star I\Lambda_r + I_p \star \Lambda_r) = I\Lambda_t \tag{4}$$

The aim of perceptual room compensation is to find $I\Lambda_p$ for which $I\Lambda_p \star I\Lambda_r$ is perceptually similar to $I\Lambda_t$. From Section II-A, for two reverberant signals to sound similar, at least the early and late parts should each sound similar. We use a weaker condition, that the overall band energies of the early and late parts match. Grosse takes a similar approach to measure responses in [7]. This condition does not specify anything about the decay profiles, however as shown later it is enough to reproduce approximate target decay profiles for bands where the room profile decays faster than the target. Equating separately the energies of the early and late parts in (4) gives the energy balance equations,

$$\mathcal{E}_n(I_p \star I_r) = \mathcal{E}_n(I_t) \tag{5}$$

and for the late parts

$$\mathcal{E}_n(\Lambda_p \star I\Lambda_r + I_p \star \Lambda_r) = \mathcal{E}_n(\Lambda_t) \tag{6}$$

where $\mathcal{E}_n$ is the energy in the nth band of the discrete signal x defined by,

$$\mathcal{E}_n(x) = |h_n \star x|^2 \tag{7}$$

for bandpass filter h_n responses. The norm on the signal is defined $|x|^2 = \sum_i x_i^2$. This definition of energy is dependent on the sampling rate. This won't be converted to a rate independent quantity, because it simplifies the expressions that will be produced for equalisation gains. The filters are normalised so the values of the frequency response has magnitude equal to 1 in the passbands. Gammatone filters have been used to model auditory perceptual response [7], however the overlap between such filters greatly complicates equalisation and resynthesis. Instead a filter bank is used that has little overlap between bands but with enough resolution so that when the bank energies are matched then the energies obtained from gammatone filters would match well. A 10th order recursive 1/3 octave filter bank is used in the Section V.

For convenience we define the *spectral densities*, $\mathcal{D}_n(x)$, of the band energies as,

$$\mathcal{D}_n(x) = \mathcal{E}_n(x)/\nu_n \tag{8}$$

where $\nu_n = 2f_n/f_s$ are fractional bandwidths of the filters, with bandwidths f_n and sampling frequency f_s. Correction factors, close to 1, are found and applied to the band filters, to remove excess density caused by band overlap.

To proceed in calculations using (5) and (6) it is necessary to calculate the density of signal convolutions and additions. These can be greatly simplified if the signals can be modelled with random statistics, as in Section II-B. Two results are derived in Appendix B, for addition and convolution: For any statistically independent random signals a and b

$$E(\mathcal{D}_n(a+b)) = E(\mathcal{D}_n(a)) + E(\mathcal{D}_n(b)) \tag{9}$$

where E is the expectation operator. No assumption is made about the independence of samples within either a or b. In practice we deal with definite samples of the random signals, also referred to as a and b. Then for definite densities,

$$\mathcal{D}_n(a+b) \approx \mathcal{D}_n(a) + \mathcal{D}_n(b) \tag{10}$$

The relation is now approximate because the densities are subject to statistical fluctuation that is analysed in Appendix B.

If in addition to independence, the expectation of the squared magnitude is nearly flat across each band, then the spectral densities are multiplicative under convolution,

$$E(\mathcal{D}_n(a \star b)) = E(\mathcal{D}_n(a))E(\mathcal{D}_n(b)) \tag{11}$$

And for definite densities,

$$\mathcal{D}_n(a \star b) \approx \mathcal{D}_n(a)\mathcal{D}_n(b) \tag{12}$$

The result is similar to that for the Discrete Fourier Transform of a convolution, with bands instead of frequency bins. 1/3 octave bands determined by critical band structure, also have sufficient resolution to ensure room responses are reasonably flat in the bands. For sufficiently narrow bands the statistical variance will be large enough to override the benefit of the in band flatness.

The identity (12) can be applied to the first perceptual condition (5) to find the best estimate for the early playback density,

$$\boxed{\mathcal{D}_n(I_p) = \mathcal{D}_n(I_t)/\mathcal{D}_n(I_r)} \tag{13}$$

$\mathcal{D}_n(I_p)$ may be modified later because of other requirements (Boxes are used to highlight the calculations that are needed to find the playback response). An equation for the best estimate of the late playback density $\mathcal{D}_n(\Lambda_p)$ can be found by applying (13) and (12) to (6), and using the incoherence between frames in $I\Lambda_p$ and $I\Lambda_r$, and between I_p and Λ_p,

$$\mathcal{D}_n(\Lambda_p)\mathcal{D}_n(I\Lambda_r) + \mathcal{D}_n(I_p)\mathcal{D}_n(\Lambda_r) = \mathcal{D}_n(\Lambda_t) \tag{14}$$

Substituting from (13) for $\mathcal{D}_n(I_p)$,

$$\mathcal{D}_n(\Lambda_p)\mathcal{D}_n(I\Lambda_r) + \mathcal{D}_n(I_t)\mathcal{D}_n(\Lambda_r)/\mathcal{D}_n(I_r) = \mathcal{D}_n(\Lambda_t) \tag{15}$$

The energy density terms cannot be negative, so from (15) a solution for $\mathcal{D}_n(\Lambda_p)$ is only possible when

$$\mathcal{D}_n(I_t)\mathcal{D}_n(\Lambda_r)/\mathcal{D}_n(I_r) \leq \mathcal{D}_n(\Lambda_t) \tag{16}$$

or

$$\mathcal{D}_n(I_t)/\mathcal{D}_n(\Lambda_t) \leq \mathcal{D}_n(I_r)/\mathcal{D}_n(\Lambda_r) \tag{17}$$

which is the condition that the early to late energy ratio of the room is greater than that for the target, in other words the room is *drier*. Assuming (17) holds then from (15) the solution for late playback density is

$$\boxed{\mathcal{D}_n(\Lambda_p) = \frac{\mathcal{D}_n(\Lambda_t) - \mathcal{D}_n(I_p)\mathcal{D}_n(\Lambda_r)}{\mathcal{D}_n(I\Lambda_r)}} \tag{18}$$

If (17) is an equality then the room can produce all the target reverberance, without additional playback reverberance, $\mathcal{D}_n(\Lambda_p) = 0$. However, the room-in-room decay rate is then equal to the room decay rate, which may be significantly different to the target decay rate.

If (17) does not hold then the $I_p \star \Lambda_r$ contribution, $\mathcal{D}_n(I_p)\mathcal{D}_n(\Lambda_r)$ in (14), will already exceed the late target energy $\mathcal{D}_n(\Lambda_t)$, and so additional contribution from the late playback should be suppressed, by choosing $\boxed{\mathcal{D}_n(\Lambda_p) = 0}$.

Excess late energy in a band may have a negative effect on perception of the overall room-in-room response. So it can be useful to reduce I_p so that the $I_p \star \Lambda_r$ contribution does not exceed the late target energy. This will improve the late energy and the overall early to late energy ratio that is related to transient definition. The early equalisation in this band will be worsened by reducing I_p , but this is usually a reasonable tradeoff.

It can be useful to further limit the $I_p \star \Lambda_r$ contribution to a value less than the late target energy: If the room decay rate is significantly greater than the target decay rate, then a late room-in-room response dominated by the $I_p \star \Lambda_r$ contribution would have significantly higher decay rate than the target. Limiting the $I_p \star \Lambda_r$ contribution to a value less than the late target energy ensures that some of the late energy will come from the $\Lambda_p \star I\Lambda_r$ contribution that decays like the target. However this comes at the expense of further decreasing I_p energy causing reduced transient response in the band. A subjective tradeoff is

required to decide on the appropriate limits, which will depend on several factors including the target and room responses, the source material, and the listening context.

In order to help control the I_p energy in the above cases we define a *late contribution fraction* parameter, γ_n, which is the fractional contribution of $\mathcal{D}_n(I_p)\mathcal{D}_n(\Lambda_r)$ to $\mathcal{D}_n(\Lambda_t)$ in (14),

$$\gamma_n = \mathcal{D}_n(I_p)D_n(\Lambda_r)/\mathcal{D}_n(\Lambda_t) \tag{19}$$

Using the value of $\mathcal{D}_n(I_p)$ given by (13) the condition (17) is equivalent to $\gamma_n < 1$, where

$$\gamma_n = \frac{\mathcal{D}_n(\Lambda_r)D_n(I_t)}{D_n(\Lambda_t)\mathcal{D}_n(I_r)} \tag{20}$$

Given the above considerations, our strategy will be to enforce an upper limit, γ'_n, for γ_n. If initially $\gamma_n > \gamma'_n$, then γ_n can be limited by redefining the early playback density based on (19),

$$\boxed{\mathcal{D}_n(I_p) = \gamma'_n \ \mathcal{D}_n(\Lambda_t)/\mathcal{D}_n(\Lambda_r)} \tag{21}$$

$\mathcal{D}_n(\Lambda_p)$ is then given by (18), as before.

When (21) is applied, then the value of $\mathcal{D}_n(I_p)$ will be lower than that given by (18), causing a reduction in the reproduced early energy. While the reproduced earl / late balance cannot be changed, the reproduced total energy can be restored to the target value by amplifying the playback densities, based on the predicted reproduced energy.

$$\alpha_n = \frac{\mathcal{D}_n(I\Lambda_t)}{(\mathcal{D}_n(I_p) + \mathcal{D}_n(\Lambda_p)) * \mathcal{D}_n(I\Lambda_r)} \tag{22}$$

$$\mathcal{D}'_n(I_p) = \mathcal{D}_n(I_p) * \alpha_n \tag{23}$$

$$\mathcal{D}'_n(\Lambda_p) = \mathcal{D}_n(\Lambda_p) * \alpha_n \tag{24}$$

Sometimes a band in the room response can decay slowly, or ring, relative to the target, and this stands out in an obvious way. It may then be appropriate to silence playback on this band to prevent the ringing, even though energy is lost.

The playback densities have been found, but not the profile of the playback response. The target response will be used as the basis for playback, since from Section II-B the target profile will be imprinted in the reproduced response, provided it decays more slowly. The target is equalised by removing the target gains and applying the playback gains, found in (13), (18), and (21),

$$\boxed{I_p = \sum_n h_n \star I_t \sqrt{\frac{\mathcal{D}_n(I_p)}{\mathcal{D}_n(I_t)}}} \tag{25}$$

and

$$\boxed{\Lambda_p = \sum_n h_n \star \Lambda_t \sqrt{\frac{\mathcal{D}_n(\Lambda_p)}{\mathcal{D}_n(\Lambda_t)}}} \tag{26}$$

This object-based method will be compared with conventional channel-based room compensation. Using the same framework the channel-based playback response is given by equalising the whole target response IV_t.

$$IV_p = \sum_n \frac{h_n \star IV_t}{\sqrt{\mathcal{D}_n(IV_r)}} \tag{27}$$

ensuring the overall target densities are reproduced, $\mathcal{D}_n(IV_p \star IV_r) = \mathcal{D}_n(IV_t)$.

Band overlap causes excess level, which is removed with a near-unity correction factor calibrated for a flat response. This method of parallel equalisation is not ideal, because the sub-band filters always introduce phase and transient distortion, which is particularly noticeable in the early response. A cascade type equaliser, or another which has no distortion for a flat response would be better.

Early equalisation could be treated using a more detailed inverse filtering approach that aims to solve $I_p \star I_r = I_t$ for I_p preserving temporal as well as frequency structure. Examples of this were cited in the Introduction. Band equalisation can be used if the initial value for I_p needs to be revised because of late energy over production.

The statistical assumptions imply random variation in the calculated densities. This can be measured by the standard deviation relative to the expected value, $\sqrt{V(\mathcal{E}_n(x))}/E(\mathcal{E}_n(x))$. From the analysis in Appendix B, this relative error is approximately $1/\sqrt{2f_{bw}T}$, for bandwidth f_{bw}, and sample duration T. This approximation carries through to the densities of the compensated reproduced response. For a long late response, $T = 0.5s$, and 1/3 octave band at 5000Hz, $f_{bw} = 2300$, and the relative error ≈ 0.03, which is small relative to the equalisation error being corrected. In the worst case, for the early part of a reverberant response, with $T = 0.020s$, and $f_{bw} = 50$ for a 1/3 octave band at 200Hz, the relative error ≈ 0.70, which is significant: Although a band width should be narrow enough for the local spectrum to be flat, this advantage can be lost due to statistical variation.

D. Object-based equalisation vs channel equalisation

In this section the equalisation of object-based and channel reproductions are compared theoretically. To simplify only the early/late energy ratio is compared, since the early and late common level is easily adjusted by providing gain across the whole response. To abbreviate, this ratio is written as

$$\mathcal{C}_{n,x} = \frac{\mathcal{D}_n(I_x)}{\mathcal{D}_n(\Lambda_x)} \tag{28}$$

where x stands for any response r, t, p, etc. The previous section shows that if $\mathcal{C}_{n,r} \geq \mathcal{C}_{n,t}$ then the object-based reproduction IV_{pr} can match the target early and late energies, and so also $\mathcal{C}_{n,pr} = \mathcal{C}_{n,t}$. The channel-based playback is produced using a single equalisation across the whole response, and has the same early / late ratio as the target played into the room $\mathcal{C}_{n,tr}$. Generally $\mathcal{C}_{n,tr} < \mathcal{C}_{n,t}$, which has been the motivation for equalising the early and late parts separately. If the object-based reproduction matches the target, then the performance of the channel-based reproduction can be measured relative to this using the ratio, $\mathcal{C}_{n,tr}/\mathcal{C}_{n,t}$, which can be simplified,

$$\mathcal{C}_{n,tr}/\mathcal{C}_{n,t} = \left[1 + \frac{1 + \mathcal{C}_{n,t}}{\mathcal{C}_{n,r}}\right]^{-1} \tag{29}$$

This measure falls as $\mathcal{C}_{n,r}$ falls, the room becoming less dry. When $\mathcal{C}_{n,r} < \mathcal{C}_{n,t}$ object-based reproduction is no longer exact, and $\mathcal{C}_{n,tr}/\mathcal{C}_{n,pr}$ eventually rises to 1. In Section V

practical examples are given showing significant improvement in $\mathcal{C}$ using the object-based method.

III. Compensation of parametric reverberation

Section II describes a general method for finding playback responses from impulse responses. The cost of the compensation calculation can be streamlined if the response densities are pre-calculated and transmitted in the object stream. Another possibility is to encode responses with a parametric representation. Using such a representation, it is shown how the playback response can be calculated quickly from the target and room responses.

A. Reverberant Spatial Audio Object

The Reverberation Spatial Audio Object (RSAO) [20; 11] encodes reverberation parametrically. An object-based metadata processor for room compensation using RSAO was outlined previously [14]. The early response is encoded in RSAO as a train of discrete reflection impulses each with direction and equalisation, and the late part is encoded with levels, attack and decay times across frequency bands, representing diffuse sound coming from all directions. RSAO also allows for an initial fast rise envelope in the late decay, which can blend with the early response, however this is not used here. For RSAO the early equalisation is implemented with a biquad filter cascade. Here we assume the band levels can be controlled directly.

In the following, the nth band level of the i'th early reflection is given by a parameter $a_{i,n}$. Each band of the late response is modelled as a random step exponential signal in discrete time,

$$N_n(t_m) b_n e^{-\beta_n t_m} \tag{30}$$

where $t_m = mT$ are the discrete times with sampling period T, b_n are the band amplitudes, and β_n are the decay rates. The random variables $N_n(t_m)$ are defined by bandpass filtering random white noise with standard deviation σ_N. The RSAO implementation uses 2nd order, octave bandpass filters. For accurate room compensation higher frequency resolution and less band overlap are needed. The examples use 10th order, 1/3 octave bandpass filters. The late impulse responses are formed by taking a sample of the random signal and applying the decay envelope defined in (30). This approach has the disadvantage that a sample signal can actually be anything, and so the band densities $\mathcal{D}_n(\Lambda_t)$ can vary significantly from sample to sample, with greater variance at low frequency. Any corrective equalisation would involve computationally expensive direct evaluation of densities, which defeats one advantage of using parameters. However variation can be reduced by pre-equalising noise signal samples so that they have the expected densities across the bands. Applying the envelope then leads to some variation in late density due to the non-uniform distribution of energy in the pre-equalised noise in time, for each band. The distribution could be made more uniform by decomposing the noise signal into windowed sections and equalising these separately.

B. Object-based room compensation

From (35) the playback band levels are

$$a'_{i,n} = a_{i,n} \sqrt{\frac{\mathcal{D}_n(I_p)}{\mathcal{D}_n(I_t)}} \tag{31}$$

$\mathcal{D}_n(I_t)$ can be calculated directly from RSAO parameters,

$$\mathcal{D}_n(I_t) = \sum_i a_{i,n}^2 \tag{32}$$

since $a_{i,n}^2$ is the energy density of ith impulse in the nth band of the target. $\mathcal{D}_n(I_r)$, which is needed to calculate $\mathcal{D}_n(I_p)$, can be found in a similar way.

The expected values of the randomly generated late densities, which are estimates of these densities, can be found from the late band parameters, as shown in Appendix C,

$$\mathcal{D}_n(\Lambda) = \frac{\sigma_N^2 b_n^2 f_s}{2\beta_n} \tag{33}$$

where f_s is the sampling rate. $\mathcal{D}_n(\Lambda_p)$ are evaluated from $\mathcal{D}_n(\Lambda_t)$ and $\mathcal{D}_n(\Lambda_r)$, using (18). The reverberation parameters should be chosen to that (33) predicts the original response densities accurately.

The modified amplitudes b'_n for the late playback response Λ_p are given by applying the equalisation from (26),

$$b'_n = b_n \sqrt{\frac{\mathcal{D}_n(\Lambda_p)}{\mathcal{D}_n(\Lambda_t)}} \tag{34}$$

The listening room response is information held locally at the point of reproduction, so the densities can be calculated directly from this. If density parameters are provided then other reverberation parameters are not required. In recent work the room geometry has been captured using audio visual sensors [21]. This can be used to estimate response densities.

IV. Precise compensation

Equation (12) shows that the expected value of the energy density of the convolution of random signals can be found by multiplication. However this is subject to statistical variation, as described, with the error being greater for lower bands that have smaller bandwidth. This uncertainty can be removed and the compensation made more precise, at the cost of additional calculation. As before the target response is equalised to generate the playback response, here using gains δ_n and η_n.

$$I_p = \sum_n \delta_n \; h_n \star I_t \tag{35}$$

$$\Lambda_p = \sum_n \eta_n \; h_n \star \Lambda_t \tag{36}$$

and the densities are then related as follows,

$$\mathcal{D}_n(I_p) = \delta_n^2 \mathcal{D}_n(I_t) \tag{37}$$

$$\mathcal{D}_n(\Lambda_p) = \eta_n^2 \mathcal{D}(\Lambda_t) \tag{38}$$

From (5),

$$\mathcal{D}_n(I_p \star I_r) = \mathcal{D}_n(I_t) \tag{39}$$

Substituting from (35) leads to

$$\delta_n^2 = \frac{\mathcal{D}_n(I_t)}{\mathcal{D}_n(I_t \star I_r)} \tag{40}$$

Similarly writing (6) in terms of densities,

$$\mathcal{D}_n(\Lambda_p \star I\Lambda_r + I_p \star \Lambda_r) = \mathcal{D}_n(\Lambda_t) \tag{41}$$

Splitting into parts, and using the gains (37), (38),

$$\begin{aligned}\eta_n^2 \mathcal{D}_n(\Lambda_t \star I\Lambda_r) + 2\eta_n\delta_n\mathcal{D}_n(\Lambda_t \star I\Lambda_r, I_t \star \Lambda_r) \\ +(\delta_n^2\mathcal{D}_n(I_t \star \Lambda_r) - \mathcal{D}_n(\Lambda_t)) = 0\end{aligned} \tag{42}$$

where the cross density $\mathcal{D}_n(x, y)$ is defined by

$$\mathcal{D}_n(x, y) = (h_n \star x) \cdot (h_n \star y) \ /\nu_n \tag{43}$$

where $a \cdot b$ is the vector dot product acting on the signals a, b considered as vectors. Equation (42) is a quadratic for η_n, since δ_n is already known. If a positive real solution exists the compensation is solved exactly for the band. Otherwise choosing the real value reduces the early playback and prevents over production of late energy. Although the method avoids statistical variance, it has the disadvantage that convolutions and densities have to be calculated for each new (target, room) response pair presented, whereas before only the separate room target and room densities have to be calculated. The advantages include those of the statistically based method: The late and early parts of the response are equalised independently and robustly, compared with inversion methods.

V. Objective Tests

The object-based compensation method described in the previous sections is tested here using synthetic and recorded responses, and compared with channel-based equalisation. using the perceptually important energy density ratio and decay rate measures. The synthetic examples are based on synthetic noise and provide an initial test of the method. They also provide a clear presentation of how the compensation is operating. The recorded examples are more complex, and test the statistical assumptions about real responses. The sample rate is 44100 Hz. There are 23 bands in 1/3 octave intervals with centre frequencies from 100 Hz to 15849 Hz. The early response duration is 20 ms.

The performance of the object-based method and the conventional channel-based method, are compared for 5 sets of responses, labelled *Synth1*, *Synth2*, *Synth3*, *Rec1*, *Rec2*, and *Rec3*. The results are shown in Fig. 3 and Fig. 4.*target* refers to the target response. *room* refers to the response of the listening room *object_playback* refers the object-based playback response, given by (35) and (26). *channel_playback* refers to the channel-based playback response given by (27). *playback-room* refers to the simulated RIRR made by convolving a playback response with the room. The densities and delay times are all calculated by analysing the synthesised or recorded responses: room, target, playback, and playback-room. Each response has been applied to a short sample of a woman speaking.

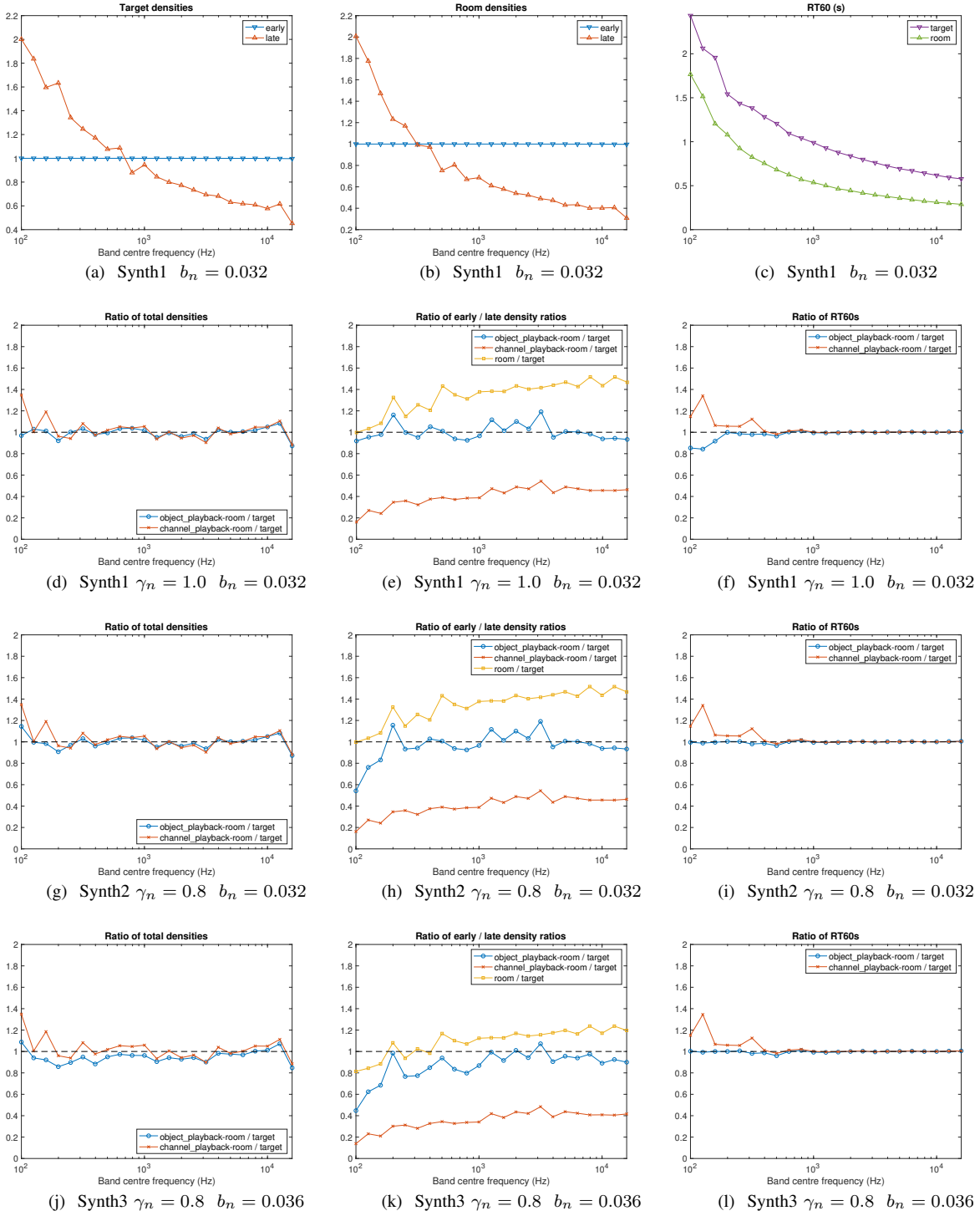
A. Synthesised responses

The synthetic responses are constructed using the RSAO model described in Section III-A. Each response has an early response consisting of a single direct impulse with amplitude 1, so that the early densities are normalised to 1. The late response begins at 20ms, with a fast attack. For Synth1 the late levels for the target response are $b_n = 0.028$. The decay rates β_n are evenly spaced across the bands from 3 to 12. The late levels and decay rates of the room response are $b_n = 0.032$ and β_n is spread evenly from 4 to 24. The late response contribution parameter is set to unity, $\gamma'_n = 1$. Fig. 3a, Fig. 3b, and Fig. 3c show the parameters of the target and room responses. The compensation results are unaffected by overall scaling of the target or room response. Fig. 3d compares the combined early and late densities of the RIRRs to the target response, and show a good match, subject to statistical fluctutation that rises at low frequency, as expected from the analysis in Section II-C. Fig. 3e shows how well the early/late ratio of the RIRRs match those of the target, $\mathcal{C}_{n,pr}/\mathcal{C}_{n,t}$. A match indicates that early and late equalisations match the target, provided the total densities match. The room is also compared to the target, $\mathcal{C}_{n,r}/\mathcal{C}_{n,t}$. Where this ratio is above 1 then object-based compensation should be possible, and this is the case, subject to statistical fluctuations. The object-based performance degrades at the lowest frequencies where $\mathcal{C}_{n,r}/\mathcal{C}_{n,t}$ is near 1. $\mathcal{C}_{n,pr}/\mathcal{C}_{n,t}$ is close to 1, within 1 dB or a factor 1.26, for most of the range, and the improvement over channel-based playback is over 3dB or a factor of 2. The improvement is significant even where the object-based reproduction does not match the target. Fig. 3f compares the absolute decay times of the RIRRs and target response, calculated using the Schroeder method [22]. Fig. 3c shows that the room decay times are smaller than the target across the frequency range, so based on Section II-B object-based and channel-based RIRR are expected to be near to the target. In Fig. 3f the object-based RIRR performs better, however there is still a dip at the lowest frequencies, consistent with $\mathcal{C}_{n,r}/\mathcal{C}_{n,t}$ being near 1 in this range. From the analysis in Section II-C, the main contribution to late energy is then from late room, not the late target, causing a reduction in decay time.

In the next example, Synth2, the late control contribution parameter has been reduced $\gamma'_n = 0.8$, forcing contribution from the late target response, to improve the RIRR decay time, at the expense of some reproduced early energy. Fig. 3i shows this is effective in this case. However this has also resulted in a reduction of $\mathcal{C}_{n,pr}/\mathcal{C}_{n,t}$ for object-based reproduction in the low frequency range, the tradeoff discussed in II-C. In example Synth3 the late room energy has been increased, causing a wider region where $\mathcal{C}_{n,r}/\mathcal{C}_{n,t} < 1$. This causes a wider suppression of the object-based ratio $\mathcal{C}_{n,pr}/\mathcal{C}_{n,t}$, and reduced clarity. Again, $\gamma'_n = 0.8$ supports decay time reproduction.

B. Measured responses

Examples Rec1 and Rec2 use recorded room and target responses, measured in two real rooms. The early responses are complex, in contrast to the previous examples. In Rec1 $\gamma_n = 1.0$. Fig. 4d shows that the overall equalisation is



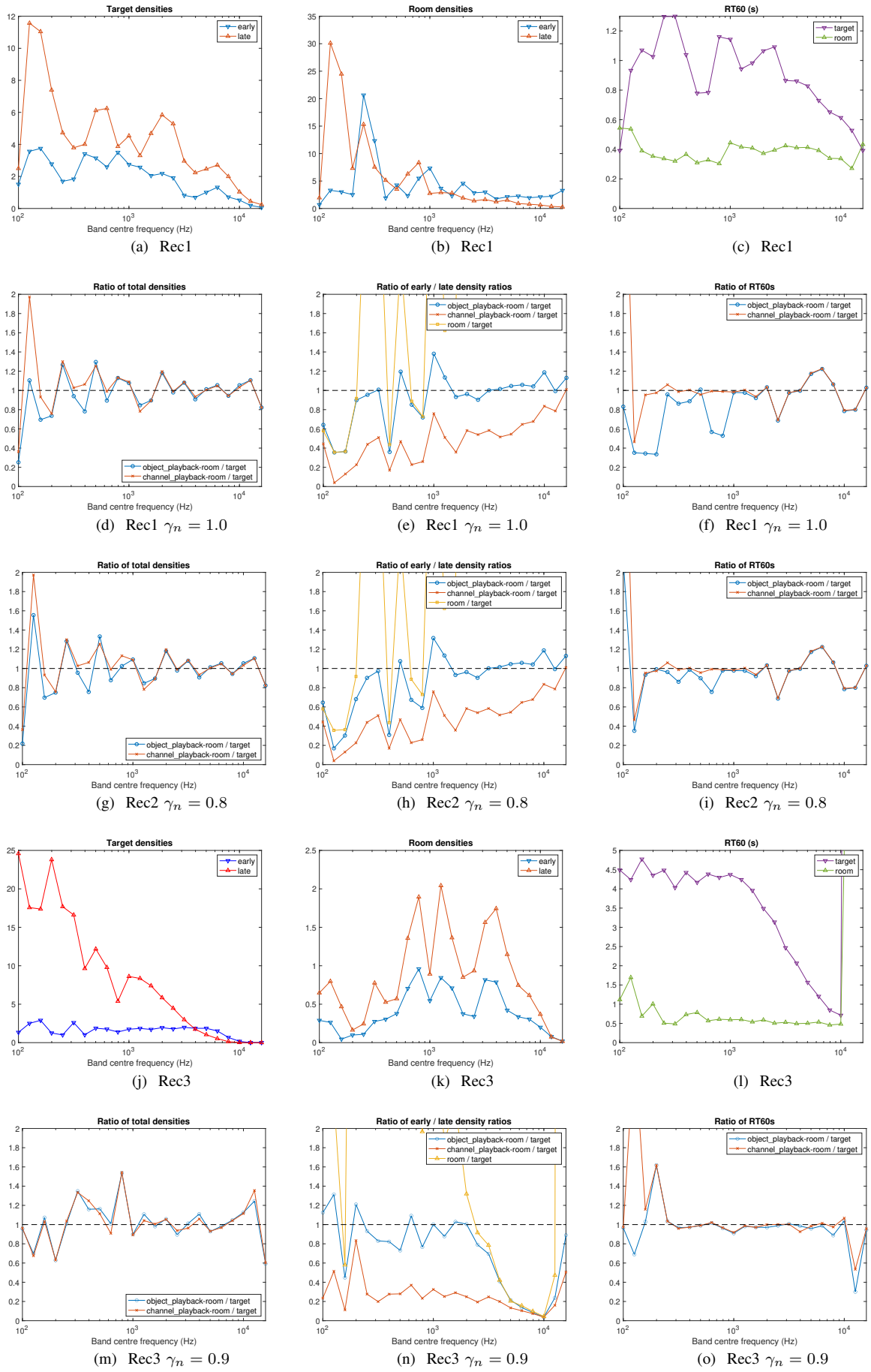
(a) Synth1 $b_n = 0.032$ (b) Synth1 $b_n = 0.032$ (c) Synth1 $b_n = 0.032$

(d) Synth1 $\gamma_n = 1.0$ $b_n = 0.032$ (e) Synth1 $\gamma_n = 1.0$ $b_n = 0.032$ (f) Synth1 $\gamma_n = 1.0$ $b_n = 0.032$

(g) Synth2 $\gamma_n = 0.8$ $b_n = 0.032$ (h) Synth2 $\gamma_n = 0.8$ $b_n = 0.032$ (i) Synth2 $\gamma_n = 0.8$ $b_n = 0.032$

(j) Synth3 $\gamma_n = 0.8$ $b_n = 0.036$ (k) Synth3 $\gamma_n = 0.8$ $b_n = 0.036$ (l) Synth3 $\gamma_n = 0.8$ $b_n = 0.036$

Fig. 3: Examples Synth1, Synth2 and Synth3, using synthetic target and room responses. The plots show the response characteristics and compare the performance of object-based and channel-based room reverberation correction methods. The plots titled *Ratio of early/late density ratios* show how well the proposed reproduction method matches the early/late ratio of the target (*playback-room / target*) vs the channel based reproduction (*channel_playbacIk-room / target*). The plots titled *Ratio of RT60s* show how well the object-based reproduction matches the delays times of the target (*playback-room / target*) vs the channel based reproduction (*channel_playback-room / target*).



(a) Rec1 (b) Rec1 (c) Rec1

(d) Rec1 $\gamma_n = 1.0$ (e) Rec1 $\gamma_n = 1.0$ (f) Rec1 $\gamma_n = 1.0$

(g) Rec2 $\gamma_n = 0.8$ (h) Rec2 $\gamma_n = 0.8$ (i) Rec2 $\gamma_n = 0.8$

(j) Rec3 (k) Rec3 (l) Rec3

(m) Rec3 $\gamma_n = 0.9$ (n) Rec3 $\gamma_n = 0.9$ (o) Rec3 $\gamma_n = 0.9$

Fig. 4: Examples Rec1, Rec2 and Rec3, using recorded target and room responses. See the Fig. 3 caption for a description.

centred around the target, but there is more fluctuation than in the synthesised examples, possible because the statistical independence assumptions are not satisfied as well. The early late density, indicated by *object_playback-room / target* in Fig. 4e, is close to the target over much of the range, and improves on channel-based compensation by over a factor of 2 for most of this. As expected, where $\mathcal{C}_{n,r}/\mathcal{C}_{n,t} < 1$ indicated by *room / target*, the early/late density drops sharply. Also the late decay is uncorrected in these regions, shown in Fig. 4f. As before, setting $\gamma'_n = 0.8$, improves the late decay, shown in Fig. 4i, but also slightly worsens the early/late density ratio shown in Fig. 4h. Similar results were found with a variety of other responses. Example Rec3 contains another two recorded responses. The target has a relatively long decay, and low early densities compared to late densities. There is a high frequency region, and a small low frequency region, where $\mathcal{C}_{n,r}/\mathcal{C}_{n,t} < 1$. With $\gamma'_n = 0.9$ the reproduced RT60 matches the target through nearly the whole frequency range. The early / late density ratio is corrected over most of the frequency range, as shown in Fig. 4n, with very significant improvements over channel-based reproduction.

C. Error from measurement and listener placement

Error can enter the compensation process when the listener position does not coincide with the position the room response was measured at. Preferably the listener has some freedom of movement without affecting compensation too much. At high frequency the responses at neighbouring locations are uncorrelated, but with similar envelopes, like samples from a random signal. The appropriate density variance is calculated in Appendix A, and applied in Section II-C to find errors expected for different bands on response lengths. At low frequencies the room modes become sparse compared with the noise model, with amplitude and phase varying between locations, further increasing variance. If the measurement points are close enough then the responses are correlated at low frequency, which reduces variance. To illustrate this practically, the early and late densities were calculated for 4 nearby measurement locations in the centre of an empty classroom 9 (wide) x 7.5 x 3.5 m [23]. The locations were arranged in a diamond shape with 1 m separating opposite corners. The source was placed at the mid front of the classroom. Fig. 5 shows the variance of the late response is roughly as expected from Section II-C. The early response, which is produced by the direct signal and nearby reflections, shows similar variance at high frequency. At low frequency the interference pattern is more structured due to the restricted range of paths from the source. A significant portion of the variation is due to the varying distance to the source. The variance increases gradually as the microphone region is increased in size. For a small enough region the error is reduced at low frequency, due to correlation, but not at high frequency. The variance overall is acceptable, although not ideal at low frequency, for producing useful object-based compensation within the , based on a single central measurement.

The same response pairs were processed using the precise compensation method from Section IV. The results generally

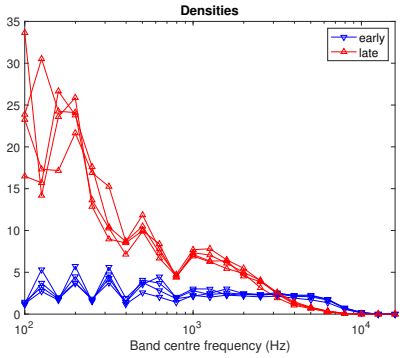


Fig. 5: Early and late response densities for 4 nearby measurement locations taken in a classroom.

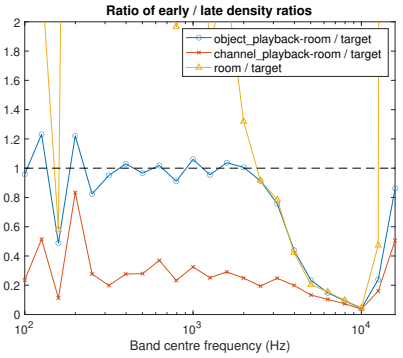


Fig. 6: Results of processing example Rec3 using the precise compensation method.

show much less fluctuation, as expected. This is illustrated by the object-based reproduction shown in Fig. 6, compared with Fig. 4n. The remaining fluctuation was found to be caused by imperfect band decomposition and reconstruction. This could be improved using more sophisticated equalisation.

VI. Listening Test

A listening test was carried out using the audio samples described in Section V, to compare the object-based compensation method to the channel-based method, using the examples from the objective tests. There were 14 subjects, all experienced with audio and testing, all reporting normal hearing, and in the age range 20-45. The last test was performed separately and had 6 subjects drawn from the first group. For each of the 5 example cases, each subject was asked to rate the difference of the object-based reproduction *playback-room.wav* to the target reference *target.wav*, relative to the channel-based reference *channel_playback-room.wav*. The difference measure is an overall impression, and reflects the choice that would be made in choosing a compensation for a real system. Although subjects may give different weight to aspects of the sound, all reported direct and reverberant tone and clarity as important factors, giving confidence that the subjects were aligned with the assumptions in the compensation method. Each sound file was produced by convolving the associated response with a short clip of female speech, and the total energy normalised to that of the original clip to minimise

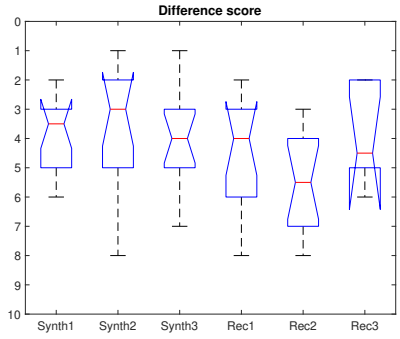


Fig. 7: Boxplots for 6 tests. In each the subjects rate the overall difference between an object-based reproduction and the target reference, relative to a channel-based anchor fixed at 10.

level selection bias. By design, the reproduced response is already normalised close to the target response. The test sounds were presented monaurally over Beyerdynamic HD650 open back headphones, in a quiet listening room. The subjects were able to listen to the samples in each test as many times as they wished, before responding using a graphic slider on a laptop computer. For each subject typically 5 minutes was taken to explain the tests and another 10 minutes to complete all the tests. The target reference was not hidden because it was needed for comparison, unlike audio quality experiments that contain a hidden high quality reference. The response scale was from 0 to 20: The test would score 0 if the test response was indistinguishable from the target reference, 10 if the test response and the channel-based reference were both judged equally different to the target reference, and 20 if the test-response was twice as different. The channel-based reference was fixed at 10, because only relative judgement is required, avoiding unnecessary absolute judgement. Boxplots for the 6 tests are shown in Fig. 7. No difference ratings above 10 were reported.

Table I contains the statistical results. The performance of object-based compensation is shown by the relative reduction of the sample mean compared to 10, the fixed channel-based difference score. The statistical significance is indicated by p_{10}, the p-value for the null hypothesis that object-based reproduction is no closer to the reference than the channel-based reproduction. To put this in more perspective the table also contains $CL99$, the upper 99% confidence limit for the population mean ($p = 0.01$), and the corresponding relative reduction R_{CL99}: With high confidence the object-based reproduction achieves $30 - 50\%$ relative reductions in difference. For the reduction of γ_n' in Example 2 the results suggest a slight decrease in difference compared with Example 1, however the same γ_n' in Example 5 causes an increase in difference compared with Example 4. This suggests the optimum choice of γ_n' depends on the particular responses considered. These comparisons have only marginal statistical significance however, and more subjects would be needed to get a clearer picture. Informal testing with a variety of responses further supports these results.

TABLE I
STATISTICAL RESULTS FROM THE LISTENING TESTS.
$\bar{x}$ - SAMPLE MEAN
$R_{\bar{x}}$ - RELATIVE REDUCTION IN DIFFERENCE
p_{10} - P VALUE FOR $\bar{x} < 10$
CL99 - 99% UPPER CONFIDENCE LIMIT FOR POPULATION MEAN
R_{CL99} - RELATIVE REDUCTION IN DIFFERENCE

	Synth1	**Synth2**	**Synth3**	**Rec1**	**Rec2**	**Rec3**
$\bar{x}$	3.8	3.6	3.9	4.6	5.6	4.0
$R_{\bar{x}}$	62%	64%	61%	54%	44%	60%
p_{10}	10^{-10}	10^{-9}	10^{-9}	10^{-8}	10^{-8}	10^{-3}
$CL99$	4.7	4.8	5.1	5.8	6.7	6.3
R_{CL99}	53%	52%	49%	42%	43%	37%

VII. CONCLUSION

A room compensation process has been presented, in which reverberant components of the target signal are available, and advantages over simple channel equalisation have been demonstrated. This can be embedded in an object-based framework. A physical-stochastic approach is taken, based on simple perceptual measures. The separate equalisations of the early and late parts of the target response can be achieved more accurately. This usually leads to reduced bass emphasis in the late response, and clearer transients due to a relative increase in early energy. The target decay rate can be reproduced providing the room decay rate is not slower than the target. The compensation process is designed to optimise perceptually relevant measures. Controls are available to limit the reproduced late energy, to limit late room influence, and to remove the influence of slow room decay. The best way to use these controls depends on the overall structure of the responses and the particular application. The objective and subjective tests show that the compensation process is a significant improvement over channel based equalisation for the cases given, in terms of separate reproduction of early and late energy. This is further supported by informal tests with a wide variety of responses.

There are a variety of possibilities for further developing the method. Extension to multichannel arrays can be made by separately compensating each loudspeaker source. In view of the statistical errors of energy density at low frequency an improved hybrid compensation could instead be formed by inverting the low frequency part of the response, which is practical. More sophisticated strategies for controlling γ_n' can be considered, including alternating emphasis between late decay time and early energy in consecutive bands where there is over contribution from late room energy. Further shaping of the late playback response may be useful to improve spectral distortion in the reproduction.

The Matlab code implementation of the compensation process, and the samples used in the listening test are available for download[1]

VIII. ACKNOWLEDGEMENTS

Thanks to James Bradley-Webster for assisting with the listening test. This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive

[1]https://doi.org/10.5258/SOTON/D1534

Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership.

APPENDIX A

In this section the convolution is calculated for two random step exponential signals that are described by,

$$R_\beta(t) = \begin{cases} N(t)e^{-\beta t}, & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{44}$$

for decay rates β = β_1 and β_2. For each time t, $N(t)$ is a random variable with zero mean $E(N(t)) = 0$ and variance $V(N(t)) = \sigma^2$. Any two variables from different responses are naturally independent, but independence is not required within each response.

First consider the simpler problem of convolving two step exponential functions,

$$\int_0^t e^{-\beta_2 \tau} e^{-\beta_1 (t-\tau)} d\tau \tag{45}$$

$$= e^{-\beta_1 t} \left[\frac{1}{\beta_1 - \beta_2} e^{(\beta_1 - \beta_2)t} \right]_0^t \tag{46}$$

$$= \frac{1}{\beta_1 - \beta_2} \left(e^{-\beta_2 t} - e^{-\beta_1 t} \right) \tag{47}$$

The convolution of the random exponential signals is

$$\int_0^t N_1(t) e^{-\beta_2 \tau} N_2(t-\tau) e^{-\beta_1 (t-\tau)} d\tau \tag{48}$$

The products $N_1(t)N_2(t-\tau)$ are identically distributed and independent random variables, so by the Central Limit Theorem, the convolution has a gaussian distribution at each time t. The variance is

$$V\left(\int_0^t N_1(t) e^{-\beta_2 \tau} N_2(t-\tau) e^{-\beta_1 (t-\tau)} d\tau \right) \tag{49}$$

$$= V(N)^2 \int_0^t e^{-2\beta_2 \tau} e^{-2\beta_1 (t-\tau)} d\tau \tag{50}$$

$$\tag{51}$$

by applying the identity $V(N_1 + N_2) = V(N_1) + V(N_2)$ to the integral, then $V(aN_1N_2) = a^2V(N)^2$ for N_1, N_2 independent, and then reforming the integral, where N_1 and N_2 are independent. The integral has the form of (45), and so the standard deviation of the convolution is

$$\sigma_c(t) = \sigma^2 \left(\frac{e^{-2\beta_2 t} - e^{-2\beta_1 t}}{2(\beta_1 - \beta_2)} \right)^{\frac{1}{2}} \tag{52}$$

APPENDIX B

The first result states that if a and b are independent random signals, and the expected value of the power spectrum of each signal is constant over each frequency band n, then the spectral energy densities are multiplicative under discrete time convolution,

$$E(\mathcal{D}_n(a \star b)) = E(\mathcal{D}_n(a)) E(\mathcal{D}_n(b)) \tag{53}$$

Independent means every sample of a is independent from every sample of b, although samples within either a or b can be dependent. In the main text the expectation function $E(\ )$ is omitted to improve readability. The variance of the densities falls with increasing frequency, and increasing signal length. To clarify the derivation, the band energies $\mathcal{E}_n$ are used rather than the densities, defined by

$$\mathcal{E}_n(x) = |h_n \star x|^2 \tag{54}$$

which are related to the band spectral density by

$$\mathcal{D}_n(x) = \mathcal{E}_n(x)/\nu_n \tag{55}$$

from (8). The energy of a signal can be expressed in terms of the components of its discrete Fourier transform (DFT), using the discrete version of Parseval's theorem. If x is a finite signal with DFT X_i, then

$$\mathcal{E}(x) = \sum_{i=1}^{I} |x_i|^2 = \frac{1}{I} \sum_{i=1}^{I} |X_i|^2 \tag{56}$$

where I is the number of samples in the sample, and bins in the DFT (Not to be confused with the use of I as the early part of the reverberant response in the main part of the text). Note that the frequency components X_i will be aliased compared with the overall sample stream. They are purely a tool for calculating the densities.

The discrete Circular Convolution Theorem states that the DFT of the circular convolution of two vectors of equal length is equal to the product of the DFT of each signal separately. The linear convolution $a \star b$ in (53) can be expressed as a circular convolution by zero padding each signal, a and b, up to a total length equal to 1 less than the sum of the signal lengths. This padding is assumed in the following. The DFT of the convolution $a \star b$ is then A_iB_i, where A_i and B_i are the DFTs of a and b, so the energy of the convolution is

$$\mathcal{E}(a \star b) = \frac{1}{I} \sum_{i=1}^{I} |A_i B_i|^2 \tag{57}$$

The energy in the nth band of a signal x is

$$\mathcal{E}_n(x) = \frac{1}{I} \sum_{i=1}^{I} |H_{n,i} X_i|^2 \tag{58}$$

$$= \frac{1}{I} \sum_{i:H_{n,i}=1} |X_i|^2 \tag{59}$$

where the response of the band filter, $H_{n,i}$, is 1 in the frequency range of band n and 0 outside. Combining these the expected energy in the nth band of the convolution of a and b is

$$E(\mathcal{E}_n(a \star b)) = E(\frac{1}{I} \sum_{i:H_{n,i}=1} |A_i B_i|^2) \tag{60}$$

$$= \frac{1}{I} \sum_{i:H_{n,i}=1} E(|A_i|^2 |B_i|^2) \tag{61}$$

Ideal band filters are chosen here, which approximate the actual filters. For independent random variables $Cov(X, Y) = E(XY) - E(X)E(Y) = 0$. If A and B are independent then the energy variables $|A_i|^2$ and $|B_i|^2$ are too, and so

$$E(\mathcal{E}_n(a \star b)) = \frac{1}{I} \sum_{i:H_{n,i}=1} E(|A_i|^2) E(|B_i|^2) \tag{62}$$

If, in addition, the expected power spectra of a and b are constant across band, then there are constants $\mathcal{A}_n$ and $\mathcal{B}_n$ so that for band n $E(|A_i|^2) = \mathcal{A}_n^2$, and $E(|B_i|^2) = \mathcal{B}_n^2$. Then

$$E(\mathcal{E}_n(a \star b)) \approx \frac{I_n}{I}\mathcal{A}_n^2\mathcal{B}_n^2 = \nu_n\mathcal{A}_n^2\mathcal{B}_n^2 \tag{63}$$

where I_n is the number of non-zero values $H_{n,i}$ in the nth band.

Working for the right side of (53),

$$E(\mathcal{E}_n(a)\mathcal{E}_n(b)) = E(\frac{1}{I}\sum_{i:H_{n,i}=1}|A_i|^2 \;\; \frac{1}{I}\sum_{i:H_{n,i}=1}|B_i|^2) \tag{64}$$

Again using the independence between a and b, and flat response in-band,

$$E(\mathcal{E}_n(a))E(\mathcal{E}_n(b)) = \frac{1}{I^2}\sum_{i:H_{n,i}=1}E(|A_i|^2)\sum_{i:H_{n,i}=1}E(|B_i|^2) \tag{65}$$

$$= \nu_n^2 \;\mathcal{A}_n^2\mathcal{B}_n^2 \tag{66}$$

Comparing with (63) ,

$$E(\mathcal{E}_n(a))(\mathcal{E}_n(b)) = \nu_n \; E(\mathcal{E}_n(a \star b)) \tag{67}$$

Then from (55),

$$E(\mathcal{D}_n(a \star b)) = E(\mathcal{D}_n(a))E(\mathcal{D}_n(b)) \tag{68}$$

which is the required result. Note that if the signals are not independent from one another generally $Cov(A, B) \neq 0$ so (62) is not true, and the required identity (53) generally does not hold, for example when the flat in-band condition is also true. Also (62) can fail when the flat in-band condition does not hold. For example if the power spectra of the two signals are non-zero only in disjoint regions of a band, then the density of the convolution is zero in this band, even though densities of the signals are non-zero.

The second result states that if random signals a and b are independent from one another, then their spectral energy densities are additive under signal addition. Applying DFT as above,

$$E(\mathcal{E}_n(a+b)) = E(\sum_{i:H_{n,i}=1}|A_i + B_i|^2) \tag{69}$$

$$= \frac{1}{I}\sum_{i:H_{n,i}=1}E(|A_i|^2) + E(|B_i|^2) \tag{70}$$

$$= E(\mathcal{E}_n(a)) + E(\mathcal{E}_n(b)) \tag{71}$$

and so,

$$E(\mathcal{D}_n(a+b)) = E(\mathcal{D}_n(a)) + E(\mathcal{D}_n(b)) \tag{72}$$

The relative error follows the same derivation as before (75), but substituting $X_i = |A_i + B_i|^2$.

The uncertainty in a signal density is as important as the expected value, and is measured by the variance $V(\mathcal{D}_n(x))$. More usefully, the error in the density relative to the expected value is given by

$$\frac{\sqrt{V(\mathcal{D}_n(x))}}{E(\mathcal{D}_n(x))} = \frac{\sqrt{V(\mathcal{E}_n(x))}}{E(\mathcal{E}_n(x))} \tag{73}$$

$$= \frac{\sqrt{\frac{I_n}{I^2}V(|X_i|^2)}}{\frac{I_n}{I}E(|X_i|^2)} \tag{74}$$

$$= \frac{1}{\sqrt{I_n}}\frac{\sqrt{V(|X_i|^2)}}{E(|X_i|^2)} \tag{75}$$

The right hand term depends on the statistics of the signal, but will be of order 1 for typical random signals encountered in this article. The left hand term $1/\sqrt{I_n}$ causes the error to increase for bands with fewer frequency samples. I_n can be found from the frequency bandwidth $f_{bw,n}$ and the duration, T, of the signal x.

$$I_n = \nu_n I = \frac{2f_{bw,n}}{f_s}I = 2f_{bw,n}T \tag{76}$$

since I is the number of time samples, equal to the number of frequency samples. In the main text this is applied to model the variation introduced in calculation, for example when $\mathcal{D}_n(a+b)$ is replaced by $\mathcal{D}_n(a)+\mathcal{D}_n(b)$ (without expectation), and also to model variation in uncorrelated room response measurements.

APPENDIX C

The energy densities are to be evaluated for the late band signals given by,

$$\Lambda_n(t_m) = N_n(t_m)b_n e^{-\beta_n t_m} \tag{77}$$

where the random variables $N_n(t_m)$ are obtained for each band by bandpass filtering a random white noise signal N with standard deviation σ_N. So $E(N^2) = V(N) = \sigma_N^2$. First consider a late signal with a single band across the full range up to Nyquist. By definition the density is

$$\mathcal{D}_n(\Lambda) = E\left(\sum_m (N(t_m)be^{-\beta t_m})^2\right) \tag{78}$$

$$= E(N^2)b^2\sum_m (e^{-2\beta T})^m \tag{79}$$

$$\approx \frac{\sigma_N^2 b^2}{1 - e^{-2\beta T}} \tag{80}$$

$$\approx \frac{\sigma_N^2 b^2 f_s}{2\beta} \tag{81}$$

The first approximation is because the sum has to be finite, and the second because βT is small for a typical sampling rate $f_s = 1/T$, and late response times. To find the density of a general sub-band, observe that the spectrum of the broadband late signal has uniform statistics across the whole frequency range. Using Parseval's theorem the total sub-band energy is then in proportion to the sub-band frequency width, which means the energy spectral density is the same,

$$\mathcal{D}_n(\Lambda) \approx \frac{\sigma_N^2 b_n^2 f_s}{2\beta_n} \tag{82}$$

REFERENCES

[1] P. G. Craven and M. A. Gerzon, "Practical adaptive room and loudspeaker equaliser for hi-fi use," in *Proc. Audio Engineering Society Conference: UK 7th Conference: Digital Signal Processing (DSP)*. Audio Engineering Society, 1992.

[2] J. N. Mourjopoulos, "Digital equalization of room acoustics," *Journal of the Audio Engineering Society*, vol. 42, no. 11, pp. 884–900, 1994.

[3] A. Mäkivirta, P. Antsalo, M. Karjalainen, and V. Välimäki, "Modal equalization of loudspeaker-room responses at low frequencies," *Journal of the Audio Engineering Society*, vol. 51, no. 5, pp. 324–343, 2003.

[4] P. Lecomte, P.-A. Gauthier, C. Langrenne, A. Berry, and A. Garcia, "Cancellation of room reflections over an extended area using ambisonics," *The Journal of the Acoustical Society of America*, vol. 143, no. 2, pp. 811–828, 2018.

[5] S. Cecchi, A. Carini, and S. Spors, "Room response equalization—a review," *Applied Sciences*, vol. 8, no. 1, p. 16, 2018.

[6] P. N. Samarasinghe, T. D. Abhayapala, and M. A. Poletti, "Room reflections assisted spatial sound field reproduction," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*. IEEE, 2014, pp. 1352–1356.

[7] J. Grosse and S. van de Par, "Perceptually accurate reproduction of recorded sound fields in a reverberant room using spatially distributed loudspeakers," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 867–880, 2015.

[8] C. Hak and R. Wenmaekers, "The impact of sound control room acoustics on the perceived acoustics of a diffuse field recording," *WSEAS Transactions on Signal Processing*, vol. 6, no. 4, pp. 175–185, 2010.

[9] A. Haeussler and S. van de Par, "Theoretischer und subjektiver einfluss des aufnahmeraumes auf den wiedergaberaum," in *DAGA '14, 40. Jahrestagung fuer Akustik*, 2014.

[10] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial sound with loudspeakers and its perception: A review of the current state," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1920–1938, 2013.

[11] P. Coleman, A. Franck, P. J. Jackson, R. J. Hughes, L. Remaggi, and F. Melchior, "Object-based reverberation for spatial audio," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 66–77, 2017.

[12] P. Coleman, A. Franck, J. Francombe, Q. Liu, T. de Campos, R. Hughes, D. Menzies, S. Galvez, Y. Tang, J. Woodcock *et al.*, "An audio-visual system for object-based audio: From recording to listening," *IEEE Transactions on Multimedia*, vol. 20, no. 8, aug 2018.

[13] J.-M. Jot, "Proc. efficient models for reverberation and distance rendering in computer music and virtual audio reality." in *ICMC*, 1997.

[14] D. Menzies and F. M. Fazi, "A perceptual approach to object-based room correction," in *Proc. Audio Engineering Society Convention 141, Los Angeles*, sep 2016.

[15] C. Faller, "Modifying audio signals for reproduction with reduced room effect," in *Audio Engineering Society Convention 147*. Audio Engineering Society, 2019.

[16] "Acoustics – Measurement of room acoustic parameters. Part 1: Performance spaces." International Organization for Standardization, Geneva, CH, Standard, 2009.

[17] E. Zwicker, G. Flottorp, and S. S. Stevens, "Critical band width in loudness summation," *The Journal of the Acoustical Society of America*, vol. 29, no. 5, pp. 548–557, 1957.

[18] J. Breebaart, S. Disch, C. Faller, J. Herre, G. Hotho, K. Kjörling, F. Myburg, M. Neusinger, W. Oomen, H. Purnhagen *et al.*, "Mpeg spatial audio coding/mpeg surround: overview and current status," in *Proc. Audio Engineering Society Convention 119*. Audio Engineering Society, 2005.

[19] M. R. Schroeder, "Statistical parameters of the frequency response curves of large rooms," *Journal of the Audio Engineering Society*, vol. 35, no. 5, pp. 299–306, 1987.

[20] L. Remaggi, P. Jackson, and P. Coleman, "Estimation of room reflection parameters for a reverberant spatial audio object," in *Proc. Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[21] H. Kim, L. Remaggi, P. Jackson, F. Fazi, and A. Hilton, "3d room geometry reconstruction using audio-visual sensors," *3DV 2017 Proceedings*, p. 4321, 2017.

[22] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965. [Online]. Available: https://doi.org/10.1121/1.1909343

[23] R. Stewart and M. Sandler, "Database of omnidirectional and b-format room impulse responses," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 165–168.

Dylan Menzies Dr Dylan Menzies is a Visiting Fellow in the Institute of Sound and Vibration, at the University of Southampton. Areas of interest include spatial audio synthesis and reproduction, sound synthesis for virtual environments, and musical synthesis and interfaces. He holds a PhD in Electronics from the University of York, an BA in Mathematics from Cambridge University, and has worked as a research engineer for several companies including Sony Professional Audio.

Philip Coleman Philip is currently a Lecturer in Audio at the Institute of Sound Recording, University of Surrey, UK. Previously, he worked in the Centre for Vision, Speech and Signal Processing (University of Surrey) as a Research Fellow on the project S3A: Future spatial audio for an immersive listening experience at home. His research interests are broadly in the domain of engineering and perception of 3D spatial audio, including object-based audio, immersive reverberation, sound field control, loudspeaker and microphone array processing, and enabling new user experiences in spatial audio. He received a Ph.D. degree in 2014 on the topic of loudspeaker array processing for personal audio (University of Surrey).

Filippo Maria Fazi Filippo Maria Fazi graduated in Mechanical Engineering from the University of Brescia (Italy) in 2005. He obtained his PhD in acoustics from the Institute of Sound and Vibration Research (ISVR) of the University of Southampton, UK, in 2010, with a thesis on sound field reproduction. In the same year, he was awarded a research fellowship by the Royal Academy of Engineering and by the Engineering and Physical Sciences Research Council. He is currently an Associate Professor at the University of Southampton. Dr Fazi' s research interests include Audio technologies, Electroacoustics and Digital Signal Processing, with special focus on acoustical inverse problems, multi-channel systems, virtual acoustics, microphone and loudspeaker arrays. He is a member of the Audio Engineering Society and of the Institute of Acoustics.