ROOM CORRECTION FOR OBJECT-BASED AUDIO

Dylan Menzies 1 Institute of Sound and Vibration Research, University of Southampton
Filippo Maria Fazi Institute of Sound and Vibration Research, University of Southampton

Abstract

Traditional channel based room equalisation is able to improve low frequency colouration due to
interference of reflections, but is unable to address some other defects. The introduction of object-
based reproduction brings an opportunity for new correction methods. In particular the direct and
reverberant components of the production can be processed separately in order to improve the
reproduction of the direct to reverberant ratio, and late colouration. We consider this in the context
of an object-based system where the reverberant properties are parameterised as metadata, and
the correction process acts directly on this.

1 INTRODUCTION

Audio engineers have long faced the problem of compensating for the acoustic effects of the repro-
duction space, so that a reproduction over loudspeakers sounds as close as possible to the original
intended production. This is often approached by channel-based correction, in which the discrete
channels are filtered before the loudspeakers. In this way it is possible to improve the frequency and
timing response characteristics over a defined listening region to varying degrees depending on the
number of channels available. For example with a high number of channels, and precise calibration
it is possible to reliably cancel out early reflections in the reproduction room over a small region. Al-
ternatively it is possible to improve equalisation over a wide area, but without precise control over
reflections. However, even with many channels it is not practical to reduce the later diffuse part of
the room response, or equalise it independently from the earlier response. The early and late parts
may have significantly different coloration since the early part is dominated by interference effects of
reflections, whereas the later part is progressively effected by absorption from surfaces in the room.
If the original production contains reverberant material then this is modified by the reproduction room.
This is the room-in-room scenario. Even though the direct part of the reproduction may be well repro-
duced, the reverberant part may be poorly equalised. Some room acoustic may be desirable as it can
provide spatial diffuseness that cannot be generated by a loudspeaker array.

An extension to the channel approach is possible if the early and reverberant components of sounds
are separately available, which is possible within an object-based representation. Then the two com-
ponents can be filtered independently to improve both the reproduced late and early responses.
Grosse considers a related problem 1, looking at how close and far microphone signals from a source
recorded in a reverberant room can be equalised and mixed to feed a 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. The equalisation is determined by per-
ceptual properties of responses measured in the capture room, and the room-in-room response that
is the combination of both rooms. Good results are achieved in some cases, over a useful listening
region, using a small number of loudspeakers. There are obvious limitations: a dry production cannot
be reproduced in a reverberant room, because the reverberance cannot be cancelled out.

A similar path is taken here. Instead of providing a set of microphone recordings for a source in a
room, the starting point is a target production consisting of a representation of the sounds, including
the different reverberant components of each. Our aim is to find signal feeds for a loudspeaker array
in a reproduction room that reproduce the intended production as well as possible at the listener. We

1 d.menzies@soton.ac.uk
assume knowledge of the target and reproduction room responses separately, but not together as in\textsuperscript{1}. This approach lends itself to compact parametric representations, which can be processed efficiently.

\section{ROOM COMPENSATION}

A perceptual approach to reproduction aims to reproduce only perceptually relevant features of the target production. According to research in spatial audio coding it is sufficient to evaluate features for overlapping frames, each of length around 20ms, from the outputs of a perceptual filter bank applied to the binaural signals. The features include the total energy, ITD, ILD, and IACC. Transients require special representation, in order to capture the attack timing and preserve the precedence effect. Directed and diffuse panning can be used to reconstruct perceptually similar sound at the listener, within a loudspeaker array.

Perceptual room compensation is the modification of the playback signal to counteract the acoustic effects of the reproduction room in order that the reproduction is perceptually similar to the target production. To this end, the general properties of room impulse responses are now considered. A room response can be divided into the early response, containing the direct signal and early reflections, and the diffuse late response.

A late room response can be approximated as a noise signal that rises suddenly and then decays exponentially at rates that can vary across the spectrum. It is defined by the initial magnitude spectrum and the decay rates. Although the signal is not given exactly, the statistical properties are sufficient to define its perceptual properties. Consider a reverberant sound that is played into a reproduction room though a loudspeaker. This is the convolution of a dry original sound with a reverberant response, the playback response. The combined effective response at the listener is the convolution of the room response with the response contained within the played sound.

Perceptually relevant properties of the combined response are simple to estimate. The incoherence between the playback and room response leads the combined response having a well defined time profile. At each frequency there is a growth period followed by decay. The growth period is a feature of the room-in-room response that can be unnatural and may not be wanted. It can be shown that the combined growth time is approximately equal to the shorter decay time of the playback and room responses, while the combined decay time is equal to the longer decay time of the two responses: The decay time of the combined response matches the playback response provided the reproduction room decay time is less than the decay time of the original sound. As with any responses, the total magnitude spectrum for the combined response is the product of the total spectra of the component responses. If the total magnitude spectrum and decay rates of the combined response match another response, then the responses will be similar, possibly with differences in the early part.

We attempt to compensate the room perceptually by modifying the playback response so that the combined response is perceptually similar to the target response. In general there is low coherence between perceptual frames within each room response, and between each response and any sound. This implies that when the two responses have frames with the same magnitude spectra, that are then convolved with any sound, the two results will have frames with similar magnitude responses to each other. There is some additional fluctuation introduced in the magnitude spectra due to the stochastic nature of the responses, which is smoothed by the perceptual filters.

When making a perceptual comparison between the combined response with the target, we do not need to use fine frames. As discussed above, the decay rate of the combined response is set by the playback response, unless the room response decay is too slow. It is then enough to match the total spectral magnitude of the late responses to ensure the whole spectral evolution is matched. If the decays cannot be matched then it is anyway desirable to match the total spectral magnitude so that the equalisation properties of the playback response when applied to a sound are at least correct.
The early part of the response is in a separate frame to the late part. When convolved with a sound the early part produces a direct sound that stands apart from the late reverberance. This is reflected in the C50 measure of clarity, which measures the relative energy in the first 50ms. Matching the magnitude response of the early response ensures the direct colouration is correct. Ideally we should preserve the transient structure of the early response, however this is not generally possible without acoustic cancellation methods. The first transient will at least be correctly timed.

Building on the foregoing observations a compensation method is now formulated, considering at first a single reproduction channel. We denote a room impulse response by $I\Lambda = I + \Lambda$ where $f$ and $\Lambda$ are the early and late parts. Subscripts $t$, $r$ and $p$ denote the target, reproduction room, and playback responses. The physical room compensation problem is to find the playback response $I\Lambda_p$ such that the target response $I\Lambda_t$ is a discrete convolution,

$$I\Lambda_p \ast I\Lambda_r = I\Lambda_t \quad (1)$$

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

$$I\Lambda_p \ast I\Lambda_r = (I_p \ast I_t) + (A_p \ast I\Lambda_r + I_p \ast \Lambda_r) \quad (2)$$

The perceptual room compensation problem is formed by equating the energies of the perceptual filters applied to combined and target responses, separately for the early and late parts. For the early parts,

$$\epsilon_n(I_p \ast I_t) = \epsilon_n(I_t) \quad (3)$$

and for the late parts

$$\epsilon_n(A_p \ast I\Lambda_r + I_p \ast \Lambda_r) = \epsilon_n(\Lambda_r) \quad (4)$$

where $\epsilon_n$ are the energy densities in each band, estimated from the filter bank,

$$\epsilon_n(X) = |h_n \ast X|^2 / \alpha_n \quad (5)$$

for filters $h_n$, where $|x|^2 = \sum x_i^2$. $\alpha_n = 2f_n / f_s$ are the relative bandwidths of the filters, with bandwidths $f_n$. The filters are normalised so the values of the discrete time spectrum have magnitude equal to 1 in the passbands. A disadvantage of using explicit perceptual filters, such as the gammatone bank used in $^1$, is that the overlap between filters complicates equalisation calculations and resynthesis. Instead we use a bank with disjoint filters that have little overlap but with enough resolution so that the energies of the perceptual filters, such as the gammatone filters, will match when the energies of the disjoint filters match. A possible choice is the 1/3 octave filter bank.

For statistically independent signals it can be shown that the band energy densities are multiplicative under convolution, $\epsilon_n(a \ast b) = \epsilon_n(a) \epsilon_n(b)$, provided the mean spectrum level is flat across each band. This applies approximately to $I_p$ and $I_r$. Then from (3) the early playback energies are

$$\epsilon_n(I_p) = \epsilon_n(I_t) / \epsilon_n(I_t) \quad (6)$$

An early response satisfying this can be formed by equalisation using the filter bank,

$$I_p = \sum_n h_n \ast I_t / \sqrt{\epsilon_n(I_t)} \quad (7)$$

Similarly, expanding and rearranging (4), and using the incoherence between frames in $I\Lambda_p$ and $I\Lambda_r$, and between $I_p$ and $\Lambda_p$,

$$\epsilon_n(A_p) \epsilon_n(\Lambda_r) + \epsilon_n(I_p) \epsilon_n(\Lambda_r) = \epsilon_n(\Lambda_r) \quad (8)$$

Substituting from 6,

$$\epsilon_n(A_p) \epsilon_n(\Lambda_r) + \epsilon_n(I_t) \epsilon_n(\Lambda_r) / \epsilon_n(I_t) = \epsilon_n(\Lambda_r) \quad (9)$$

Solution for $\epsilon_n(A_p)$ is possible when $\epsilon_n(I_t) \epsilon_n(\Lambda_r) \geq \epsilon_n(I_t) \epsilon_n(\Lambda_r)$, since the energy terms cannot be negative. The target cannot be achieved if the room reverberant energy is already larger than the
target. Otherwise to minimise energy error the best that can be done to remove the late playback response. The energy of the least error response $\tilde{\Lambda}_p$ is then,

$$E_n(\tilde{\Lambda}_p) = \max\left(0, \frac{E_n(\Lambda_t) - E_n(\Lambda_r)}{E_n(I_r)} \frac{E_n(I_t)}{E_n(\Lambda_r)}\right)$$

(10)

When $E_n(I_r)E_n(\Lambda_r) = E_n(I_t)E_n(\Lambda_r)$ the room and target have proportional responses, and the room can then reproduce all the needed reverberance in this band without additional playback reverberance. A late playback response with these energies can be formed by equalising the target response,

$$\tilde{\Lambda}_p = \sum_n h_n \ast \Lambda_t \sqrt{\frac{E_n(\Lambda_p)}{E_n(\Lambda_t)}}$$

(11)

$\tilde{\Lambda}_p$ has the same decay times as the target $\Lambda_t$. So by the reverberation convolution properties, the late decay times of the reproduction $I \Lambda_p \ast I \Lambda_r$ will be correct provided the decay times of the room are shorter than the target. In the decay time is further controlled using the early response frame size, which works by modifying the early part of the decay. However the changes are rather small, and this involves an iterative calculation that we wish to avoid.

Rather than process the target response, it is also possible to directly process the early and late parts of the target signal $S \ast I_t$ and $S \ast \Lambda_t$, where $S$ is the signal, to give the playback signals $S \ast I_p$ and $S \ast \tilde{\Lambda}_p$, by replacing the corresponding response components $I_t$, $\Lambda_t$, $I_p$, and $\tilde{\Lambda}_p$.

For multichannel production we should take into account the early response of the loudspeakers where the target response is located and distribute the late response over the whole array. A combination of loudspeakers directed towards and away from the listener can be used, for early and late response reproduction, in order to improve control. In the Interaural Cross Correlation Coefficient (IACC) for the response is controlled by cross mixing the left and right reverberant feeds, in order to improve Apparent Source Width (ASW). This is not explored within the framework here, but should be possible.

3 COMPENSATION OF A PARAMETRIC REVERBERATION OBJECT

If the target and room reverberance are described using explicit impulse responses, then the playback response can be found from the reverberation parameters using the energy and filter bank calculation just described. If the responses are described parametrically, then the overall calculation can be simplified and made more suitable for real-time implementation. An example of such a parametric response is the Reverberation Spatial Audio Object (RSAO). The early response is encoded as a train of discrete impulses each with direction and equalisation, and the late part is encoded with magnitudes and decay times across frequency bands, representing diffuse sound coming from all directions.

If $a_{i,n}$ is the $n$th band level for the $i$th early impulse in the target $I_t$, then from (7) the playback levels are

$$a'_{i,n} = \frac{a_{i,n}}{\sqrt{E_n(I_r)}}$$

(12)

$E_n(I_r)$ can be pre-calculated directly,

$$E_n(I_r) = \sum_i a_{i,n}^2$$

(13)

since $a_{i,n}^2 = E_n(\delta_i)$ is the energy density of $i$th impulse in the $n$th band. If $b_n$ and $\beta_n$ are the amplitudes and decay rates corresponding to the late target response $\Lambda_r$, such that the magnitude spectrum
evolves as $b_n e^{-\beta_n t}$, then the amplitudes $b_n'$ for the late playback response $\Lambda_p$ are given by applying the equalisation from (11),

$$b_n' = b_n \sqrt{\frac{\varepsilon_n(\tilde{\Lambda}_p)}{\varepsilon_n(\Lambda_t)}}$$

(14)

$\varepsilon_n(\tilde{\Lambda}_p)$ can be evaluated using (10) from $\varepsilon_n(\Lambda_t)$ and $\varepsilon_n(\Lambda_r)$. It can be shown that the target energies $\varepsilon_n(\Lambda_t)$ can be evaluated from the reverberation parameters using the following expression,

$$\varepsilon_n(\Lambda_t) = b_n^2 f_s \frac{4}{\beta_n}$$

(15)

where $f_s$ is the sample rate. The room energies $\varepsilon_n(\Lambda_r)$ should be pre-calculated, directly from the definition of $\varepsilon_n()$, or possibly (15) could be used if only the parametric form is available. $\varepsilon_n(\Lambda_t)$ could also be pre-calculated and transmitted as additional metadata alongside $b_n$ and $\beta_n$.

4 EXAMPLE

The RSAO rendering framework was extended to implement a perceptual room compensation processor using the methods described here. The Python language was used because it is suited to handling the JSON-based metadata format. Each sound object transmitted to the renderer consists of a single channel of audio and metadata stream describing the object direction and target reverberant response, including the direct level. The processor modifies only the metadata to produce the appropriate playback response for each sound, for a given reproduction room. The band energies for the room were pre-calculated.

A preliminary test was made. A RSAO target response was constructed, consisting of a single direct signal and a late response with spectral variation and overall RT60 of 0.6s. A reproduction room was chosen and the response measured from each loudspeaker from a 5.0 array to the listener. The room RT60 was 0.4s. The playback response was also captured as an impulse response using a delta input to the system. The results are shown in Figs. 1 and 2. The impulses shown are the binaural responses in the left ear from a 5.0 system simulated using binaural room impulse responses (BRIRs) taken from the reproduction room. The RSAO system distributes the late energy uniformly. The figures show that corrected response $P \ast R$ resembles the target much more closely than the uncorrected response $T \ast R$, in terms of early / late balance and late colouration, particularly in the lower frequency region where the uncorrected response is over emphasized. The amplitude in $P \ast R$ appears greater than in $T$ because the convolved noise has greater standard deviation relative to the mean level, compared with the target.

5 SUMMARY

A perceptual approach to room correction has been presented, with a uniform interface operating on target room responses to produce playback responses, given the reproduction room response. Initial tests show that this is effective in balancing the early and late coloration, and late decay timing can be controlled provided the room timing does not exceed this. A more in-depth study will follow. The approach could be combined with physical inversion techniques for the early response to create a hybrid correction system that has the benefits of both.
6 ACKNOWLEDGEMENTS

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) “S3A” Programme Grant EP/L000539/1, and the BBC Audio Research Partnership. Thanks to Phil
References

