Audio Engineering Society

Conference Paper

Presented at the Conference on
Immersive and Interactive Audio
2019 March 27 – 29, York, UK

Aesthetic modification of room impulse responses for interactive auralization

Keith W. Godin[1], Hannes Gamper[2], and Nikunj Raghuvanshi[2]

[1]*Microsoft Corporation*
[2]*Microsoft Research Redmond, United States*

Correspondence should be addressed to Keith W. Godin (kegodin@microsoft.com)

ABSTRACT

Interactive auralization workflows in games and virtual reality today employ manual markup coupled to designer-specified acoustic effects that lack spatial detail. Acoustic simulation can model such detail, yet is uncommon because realism often does not perfectly align with aesthetic goals. We show how to integrate realistic acoustic simulation while retaining designer control over aesthetics. Our method eliminates manual zone placement, provides spatially smooth transitions, and automates re-design for scene changes. It proceeds by computing perceptual parameters from simulated impulse responses, then applying transformations based on novel modification controls presented to the user. The result is an end-to-end physics-based auralization system with designer control. We present case studies that show the viability of such an approach.

1 Introduction

A practical interactive auralization system for games and virtual reality must enable designers to realize their aesthetic goals. At the same time, it must render physically plausible acoustics and dynamic changes as sources and listeners move through complex 3D environments, such as occlusion caused by walls and doorways. The key difficulty in interactive audio is that sound events and motion are not predetermined, requiring tools and techniques fundamentally different from traditional linear media like film.

Existing tools such as AudioKinetic's *Wwise* [1] have evolved to expose powerful design controls that influence real-time signal processing based on game logic, such as changing the room response when the player enters manually drawn reverberation volumes. But this control requires extensive manual markup, placing the burden of specifying detailed acoustic behavior entirely on the designer. The process is tough to scale with the increasing complexity of virtual worlds. Markup also suffers an essential difficulty: it must remain 3D to be intuitive yet the acoustic response is a 6D function varying with both 3D source and 3D listener location.

Automatic physics-based auralization has been studied extensively [2, 3]. While architectural walk-through applications afford consuming all system resources (CPU and/or RAM) [4], audio-visual applications like games and VR require orders of magnitude less usage, prompting recent systems geared for these use cases [5, 6] that trade off accuracy. User control is not a central consideration in current physics-based

auralization systems.

We show that the two approaches above may be combined so that physics-based acoustics provides automation and realism, which is then modified with novel aesthetic design controls to express designer intent maintaining perceptual plausibility without tedious markup. We observe that perceptual parameters derived from impulse responses form an ideal space for such controls. We extend the work by Raghuvanshi and Snyder [7] as a case study since it already employs a parametric analysis-synthesis approach, although originally motivated by efficiency concerns. We present example transformations and the perceptual principles behind their design. We identify which aspects of the simulation should be retained and how to factor out the parts that should be under designer control. Finally, we discuss the future research problems inherent in selecting and transforming a parametrization.

2 Related Work

The vast majority of auralization systems are based on the high-frequency geometric-acoustic approximation assuming ray propagation of sound. A detailed survey of recent techniques is presented by Savioja and Svensson [8]. Savioja et al. [9] proposed DIVA, one of the earliest end-to-end auralization systems. For each source, an impulse response is computed and then parameterized physically via delays and amplitudes for discrete reflections, each filtered and spatialized individually along with a single statistical late reverberation for a room assumed to contain the source and listener. A more recent system is RAVEN [4] which is designed for a more general range of spaces and employs computation and convolution of each source's impulse response at the listener. Processing may consume the entire compute power of one or several workstations. This fits the target application of interactive walk-through in computer-aided-design applications.

Games and VR on the other hand impose far tighter CPU budgets of $\sim 0.1\%$ of a CPU core for acoustic calculation per moving source. Rather than exact acoustical prediction, the goal is to produce perceptually convincing cues that vary smoothly on motion and have the expected correspondence to visual scene geometry. Many geometric acoustic systems such as Steam Audio [5] and Google Resonance [6] have been recently proposed with these applications in mind, accelerating computation by sacrificing accuracy, such as by ignoring diffraction modeling, but CPU usage still remains a concern for wide practical adoption. Our results would apply to such geometric acoustic systems as well, by first encoding the generated per-source impulse responses to perceptual parameters which are modified with our techniques and then using an impulse response synthesis procedure similar to [7].

Systems based on wave acoustics avoid high-frequency approximation [7, 10] but the high cost of numerical wave simulation requires precomputation on static scenes. The system proposed by Raghuvanshi and Snyder [7] meets the performance and robustness goals for interactive applications, enabling recent adoption in games [11] and a VR operating system shell [12]. To limit memory usage, lossy compression is performed by transforming simulated spatial fields of impulse responses to perceptual parameter fields. Runtime computation is reduced to interpolated lookup in the parametric field data for any source/listener location. The set of parameters is then synthesized into an impulse response for efficient application on the source sound.

Coleman et al. [13] describes a method for incorporating reverberation as a parameterized data stream into object-based spatial audio formats. The focus is on encoding for distribution and playback on traditional linear media like movies rather than design controls for interactive audio, like our work.

Parametric artificial reverberators have been studied extensively [2], and Feedback Delay Networks [14] are particularly commonly used. Although physically-inspired, these signal processing techniques are designed primarily for efficient natural-sounding reverberation with perceptual control. We draw inspiration from their idea of "perceptual orthogonalization," namely the controls should ideally affect independent dimensions of the auditory experience. This ensures that as the number of controls increases, the design process does not suffer from combinatoric explosion. Our work differs in that our proposed controls are meant to *modify* a dynamic perceptual parametrization of the impulse response derived from simulation, rather than constituting direct specification of the acoustics, for instance, by attaching particular hand-tuned parameter settings to a reverberation volume.

3 Perceptual parametrization

The acoustic impulse response (AIR) describes the acoustic path between a source and a receiver. Here we

look at the more specific case of the IR between source and a receiver inside an acoustic enclosure, referred to as the room impulse response (RIR). In room acoustics, the source is typically a loudspeaker, a musical instrument, or human voice, and the receiver consists of a microphone or human listener. The analysis of RIRs often focuses on perceptual aspects, i.e., models or parameters describing a human's auditory experience at the receiver position in a particular room. These perceptual characteristics are derived from the properties of the human auditory system, including the ability to determine the location and spectro-temporal fine structure of sounds. As an example, the ISO3382-1 standard defines acoustic parameters describing the perceptual properties of performance spaces [15], including perceived reverberance, clarity, and source width. Lokki et al. [16, 17] extended these standardized objective parameters with subjective attributes obtained directly from expert listeners using individual vocabulary profiling, to explain listener preference in concert halls.

We use three perceptually-motivated parameters estimated directly from a RIR $h[n]$ and the Euclidean distance d from the source to the listener. While the parameters can be calculated as a function of frequency, here we assume frequency independence.

- **Direct-path gain** G: The RMS amplitude of initial arriving wavefronts at the listener is extracted as:

$$G = \sqrt{\sum_{n=n_1}^{n_2} h^2[n]}, \tag{1}$$

  where n_1 and n_2 are the time samples at the beginning and end of the first-arriving wavefront, and $(n_2 - n_1)$ corresponds to a time difference of about 2–10 ms. We then define an obstruction-based gain as,

$$G_{obs} \equiv dG. \tag{2}$$

  Multiplication with d compensates for distance-dependent attenuation. Assuming a monopole source and that the RIR is normalized so that $\sum_n h^2[n] = 1$ at a distance of 1 m from the source in free field, $G_{obs} = 1$ throughout space in the absence of geometry. It thus isolates the effect of geometry from distance attenuation on the amplitude of initial arriving propagation paths, such as due to diffracted losses from propagation around obstructions or amplification from ground reflection.

- **Reverberation gain** F: Energy of RIR after the first-arriving wavefront:

$$F = \sqrt{\sum_{n=n_2}^{\infty} h^2[n]}. \tag{3}$$

- **Reverberation time** R_{T60}: The time it takes for the reverberation to decay to -60 dB. It can be calculated from an RIR using the method by Karjalainen et al. [18].

4 Aesthetic transformations

For each sound source, transformations are applied to simulation-derived parameters. These transformations have their own parameters that form the control surface used by the designer; to distinguish them from the impulse response parameters we refer to them as hyperparameters. The hyperparameters are a distance-based attenuation function $\alpha(d)$, occlusion factor γ, decay-time multiplier λ, and DRR warp factor p. They support the following aesthetic transforms that we discuss in detail shortly:

- adjust **direct-path gain**:

$$\tilde{G} = \alpha(d) G_{obs}^{\gamma} \tag{4}$$

- adjust **reverberation gain**:

$$\tilde{F} = \alpha(d) d^p G_{obs}^{\gamma-1} F \tag{5}$$

- adjust **reverberation time**:

$$\tilde{R}_{T60} = \lambda R_{T60}. \tag{6}$$

4.1 Distance-based attenuation function $\alpha(d)$

Distance-based attenuation is one of the most important acoustics design tools provided in interactive media toolkits. In typical implementations, a graphical user interface is used to draw piece-wise curves using elements such as linear, logarithmic, and other curve primitives, where the x-axis is Euclidean distance from source to listener and the y-axis is source gain. An example from AudioKinetic's *Wwise* is shown in Fig. 2. This allows the designer some control over the spatial aspects of source audibility; goals in this regard can

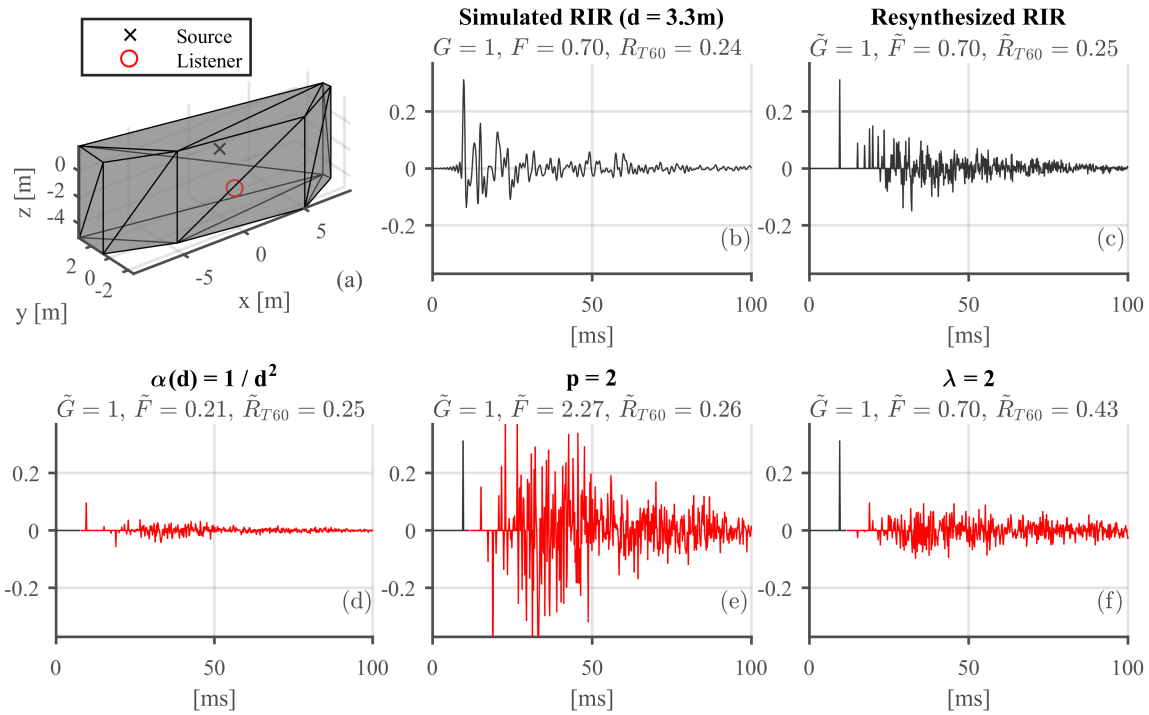

Fig. 1: (a) Example scene and (b) simulated impulse response [7]; (c) resynthesized impulse response according to the synthesis method described in Section 5.2; (d–f) example hyperparameter variations of $\alpha(d)$, p, and λ. The resynthesis stages generalize to any impulse response for which G, F, and RT_{60} are a useful parametrization.

arise from a range of gameplay, rendering cost, dynamic range, and aesthetic considerations. We leverage this existing hyperparameter, referring to it as $\alpha(d)$.

Adjustment to the hyperparameter $\alpha(d)$ does not affect the direct-to-reverberant ratio (DRR) $\frac{\tilde{G}}{\tilde{F}}$, which is always equal to the simulation-derived DRR $\frac{G}{F}$ unless the DRR warping parameter $p \neq 1$.

4.2 Occlusion factor γ

Occlusion factor adjusts the dynamic range of the attenuation effect caused by objects occluding the sound source. The default value of $\gamma = 1$ means the attenuation level is derived from simulation. When $0 \leq \gamma < 1$, moving behind a wall causes a less dramatic attenuation effect than derived from simulation, but still retains the smooth progression between acoustically shadowed and unoccluded regions. When $\gamma > 1$, the occlusion effect becomes more dramatic than reality.

4.3 DRR warping p

The direct-to-reverberant ratio (DRR) plays an important role in distance perception [19]. The hyperparameter p adjusts the DRR derived from simulation to create the perception of a sound being farther or closer, independently of loudness variation due to distance attenuation via $\alpha(d)$ or adjusted obstruction via γ. Using Eqs. 1 through 5 it can be shown that:

$$\frac{\tilde{G}}{\tilde{F}} = d^{1-p}\frac{G}{F}. \tag{7}$$

Thus a value of $p = 1$ results in simulation-derived DRR regardless of transformations via $\alpha(d)$ and γ. Decreasing p towards zero increases the DRR throughout the simulated space. This can help the designer achieve a variety of aesthetic goals; for example, this can increase intelligibility while retaining the immersive effects of spatially-dynamic reverberation levels, or to increase acoustic intimacy in a virtual reality chat room. Adjusting p upwards to reduce the DRR helps

to make an event sound distant and reverberant, which can add to the drama of a sound source.

4.4 R_{T60} multiplier λ

The multiplicative factor λ is applied to the R_{T60}, because differences in reverberation time are perceived in ratios. Adjustment of the R_{T60} through adjustments to λ can make spaces sound larger or smaller, while the simulation still drives the spatial R_{T60} dynamics so that listeners perceive smooth changes throughout the simulated space.

5 Rendering and resynthesis

After the RIR parameters are adjusted according to the designer's specified transforms, they are applied to each source's emitted audio signal either by driving the parameters of a real-time audio engine which implicitly applies an RIR, or through explicit re-synthesis and convolution with the impulse response.

5.1 Implicit resynthesis

In practical systems, the proposed framework acts to automate some of the parameters of the real-time audio engine used in the designer's pre-existing workflow. This precludes the need for explicit RIR re-synthesis, instead it is applied implicitly by the audio engine's signal processing graph. The designer also retains their control over other acoustic parameters that are not computed from simulation.

The audio signal flow graph in an interactive experience will typically comprise filter effects such as gains, mixers, and reverberation filters. The application of the above RIR parameters to this graph will depend on the specific set of effects and software, but a common setup is described here. A typical graph comprises a 'dry path' and a 'wet path', where the dry path is the input audio source with some gain applied, and the wet path is a group of reverberation filters. 'Send levels' from each wet path determine the input gain to each reverberation filter. Referring to Section 4, the dry-path gain is $\tilde{G}$, the wet send level is $\tilde{F}$, sending to a reverberation filter whose decay time is set to $\tilde{R}_{T60}$.

Some reverberation effect implementations support dynamic R_{T60} rendering, but most do not. A method to approximate per-source dynamic R_{T60} using blends of fixed R_{T60} ("canonical") filters is described in [7]. The cited method describes how to compute send levels for each source into each reverberation filter by splitting up $\tilde{F}$. This offers significant efficiency gains by first mixing the scaled signals at the input to each fixed filter, requiring a small, fixed number of convolutions regardless of source count.

5.2 Explicit resynthesis

An RIR suitable for convolution with the source signal can be synthesized from the RIR parameters by directly employing any of a variety of artificial reverberation techniques such as those described in [2]. To synthesize the RIR without applying modifications, we set: $\alpha(d) = d^{-1}$, $\gamma = 1$, $\lambda = 1$, and $p = 1$.

This synthesis method is used to generate Figure 1. The accuracy of the result will depend on the reverberation technique chosen. For example, in panel (f), the RT_{60} is scaled by $\lambda = 2$, resulting in an RT_{60} input to the resynthesis stage of 0.48. The measured RT_{60} from the reverberator is 0.43. The advantage of the parametric design approach is that such rendering efficiency versus accuracy tradeoffs can be made independently of the acoustic modeling technique employed.

6 Case studies

By attaching simulation output to the parameters of an audio engine, the proposed framework inserts a layer of abstraction between the designer and those parameters. This facilitates automation of certain aspects of the design process with the intention that the hyperparameters are more powerful than the parameters they replace. To explore whether and how the proposed hyperparameters facilitate design, we compare the traditional and proposed regimes in their process to achieve common design goals. As part of this exploration, these controls were implemented in a software package available for download [20].

6.1 Distance-based audibility cutoff

A particular 3D map of an interactive title may comprise hundreds of total sound sources that play and stop based on the systems and behaviors that make up the experience. Many of the sound sources may have these behaviors expressed terms of the Euclidean distance from source to player.

In both the traditional and proposed regimes, the designer uses a graphical user interface (an example is

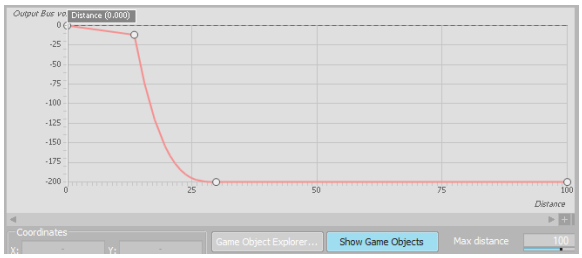

Fig. 2: Controlling distance-based source audibility with the Wwise audio design tools

shown in Figure 2) provided by their choice of audio engine to construct a gain function over Euclidean source distance, d, using curve pieces, and chooses a distance beyond which the source is inaudible. In the proposed regime, this gain function is denoted $\alpha(d)$, and Eq. 4 and 5 show that as with the traditional regime, when $\alpha(d) = 0$, the source is inaudible.

In the traditional regime adjustments to distance-based attenuation also affect the DRR. The proposed regime separates DRR adjustments into an orthogonal hyperparameter, p. This separation aids designer productivity because source audibility and DRR are often adjusted according to different classes of design goals. Source audibility is often used to create distinct ambiences in different rooms of a scene, while DRR may be adjusted to achieve a desired ambience or atmosphere in a given scene or for a particular source type.

6.2 Smooth transitions at doorways

In the everyday world, acoustical properties including reverberation level and decay time are spatially smooth. In the traditional design regime the designer specifies zones in which these are piece-wise constant. If the designer desires a spatially smooth transition, several small zones must be drawn to form a transition region. In the proposed design regime, all acoustical properties are spatially smooth to the extent supported by the chosen simulation method.

For high accuracy wave-based simulations that reproduce natural diffraction and occlusion effects (such as [7], which shows plots of spatially-smooth parameter fields) the acoustical properties will exhibit smooth transitions throughout the space. Changing the hyperparameters only modifies the acoustical contrast between various spatial positions. For instance, consider a sound source inside a room. The listener walks from near the source, through a door to outside. Reducing the occlusion factor, γ, would cause a smaller reduction in loudness as the listener walks out. But the smooth loudness transition near the doorway is preserved.

6.3 Partially-diegetic dialogue

In many interactive experiences, scripted dialogue (such as instructions from non-player characters) convey crucial information necessary to proceed through the experience. Applying realistic acoustics to these recordings could result in a low DRR, low overall level, or long R_{T60}, threatening intelligibility. One possible solution is to disable acoustics for speech, or, in the traditional regime, to specify fixed source-specific values for reverberation gain and R_{T60}. If all speech recordings share the same DRR, level, and R_{T60}, they would not vary spatially and thus would no longer sound as if they're part of the same virtual environment as other sounds. Applying separate settings to each recording to make them similar to nearby sounds but adjusted to ensure intelligibility could represent a significant task.

With our technique, all speech recordings that are meant to convey critical information could share the same hyperparameters designed to increase intelligibility, such as an $\alpha(d)$ curve with only slight attenuation, a low γ to reduce occlusion, and a low p to increase DRR. Despite sharing these parameters, the acoustical processing applied to these sources would still vary spatially, consistent with the visual geometry, such that the DRR, although scaled upwards, would still decrease when walking from close to a speaker to further away.

7 Summary

We have described a framework to integrate acoustics simulation into audio design for interactive auralization. In contrast to directly rendering impulse responses, the proposed framework retains the designer's control over the final aesthetic outcome using spatially-invariant hyperparameters. Our approach uses simulation to augment, rather than replace, the tools used in existing sound design workflows.

The proposed set of aesthetic transformations modify perceptual acoustic parameters commonly employed in room acoustics. We have shown these cover a range of important design scenarios. Additional parameters and transformations could expand the scope of application. For example, frequency-dependent reverberation time

is a straightforward extension. Initial explorations suggest that employing shortest propagation path length in place of the Euclidean distance might result in more intuitive design of distance attenuation.

References

[1] AudioKinetic Inc., "Wwise," https://www.audiokinetic.com/products/wwise/, 2018, accessed Nov 2018.

[2] Välimäki, V., Parker, J. D., Savioja, L., Smith, J. O., and Abel, J. S., "Fifty Years of Artificial Reverberation," *IEEE Trans. on Audio, Speech, and Lang. Process.*, 20(5), pp. 1421–1448, 2012.

[3] Vorländer, M., *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality (RWTHedition)*, Springer, 1 edition, 2007, ISBN 3540488294.

[4] Schröder, D. and Vorländer, M., "RAVEN: A Real-Time Framework for the Auralization of Interactive Virtual Environments," in *Forum Acusticum*, 2011.

[5] Valve, "Steam Audio," https://valvesoftware.github.io/steam-audio/, 2018, accessed Nov 2018.

[6] Google Inc., "Resonance Audio," https://developers.google.com/resonance-audio/, 2018, accessed Nov 2018.

[7] Raghuvanshi, N. and Snyder, J., "Parametric Wave Field Coding for Precomputed Sound Propagation," *ACM Trans. on Graphics*, 2014.

[8] Savioja, L. and Svensson, U. P., "Overview of geometrical room acoustic modeling techniques," *J. of the Acoustical Soc. of Am.*, 138(2), pp. 708–730, 2015.

[9] Savioja, L., Lokki, T., and Huopaniemi, J., "Auralization applying the parametric room acoustic modeling technique - The DIVA auralization system," in H. Ryohei Nakatsu, editor, *The 8th Intl. Conf. on Auditory Display, Kyoto, Japan, 2. - 5.7.2002*, pp. 219–224, Advanced Telecommunications Research Institute, 2002.

[10] Mehra, R., Rungta, A., Golas, A., Lin, M., and Manocha, D., "WAVE: Interactive Wave-based Sound Propagation for Virtual Environments." *IEEE Trans. on Visn. and Comp. Graph.*, 21(4), pp. 434–442, 2015.

[11] Raghuvanshi, N., Tennant, J., and Snyder, J., "Triton: Practical pre-computed sound propagation for games and virtual reality," *J. Acoustical Soc. of Am.*, 141(5), pp. 3455–3455, 2017.

[12] Godin, K. W., Rohrer, R., Snyder, J., and Raghuvanshi, N., "Wave Acoustics in a Mixed Reality Shell," in *2018 AES Intl. Conf. on Audio for Virt. and Augmented Reality*, 2018.

[13] Coleman, P., Franck, A., Jackson, P., Hughes, R., Remaggi, L., and Melchior, F., "On object-based audio with reverberation," in *AES 60th Intl. Conf.*, 2016.

[14] Jot, J.-M. and Chaigne, A., "Digital Delay Networks for Designing Artificial Reverberators," in *Audio Eng. Soc. Conv. 90*, 1991.

[15] Gade, A., "Acoustics in Halls for Speech and Music," in T. Rossing, editor, *Springer Handbook of Acoustics*, chapter 9, Springer, 2007 edition, 2007, ISBN 0387304460.

[16] Lokki, T., Pätynen, J., Kuusinen, A., Vertanen, H., and Tervo, S., "Concert hall acoustics assessment with individually elicited attributes," *J. Acoustical Soc. of Am.*, 130(2), pp. 835–849, 2011.

[17] Lokki, T., Pätynen, J., Kuusinen, A., and Tervo, S., "Disentangling preference ratings of concert hall acoustics using subjective sensory profiles," *J. Acoustical Soc. of Am.*, 132(5), pp. 3148–3161, 2012.

[18] Karjalainen, M., Antsalo, P., Mäkivirta, A., Peltonen, T., and Välimäki, V., "Estimation of Modal Decay Parameters from Noisy Response Measurements," *J. Audio Eng. Soc*, 50(11), p. 867, 2002.

[19] Zahorik, P., Brungart, D. S., and Bronkhorst, A. W., "Auditory Distance Perception in Humans: A Summary of Past and Present Research," *Acta Acustica united with Acustica*, 2005.

[20] Microsoft Corp., "Project Acoustics," https://aka.ms/acoustics, 2018, accessed Jan 2019.