SPATIAL HIGH FREQUENCY EXTRAPOLATION METHOD FOR ROOM ACOUSTIC AURALIZATION

Alex Southern & Lauri Savioja, *

Virtual Acoustics Team, Department of Media Technology
School of Science, Aalto University, Espoo, Finland
alexander.southern@aalto.fi

ABSTRACT

Auralization of numerically modeled impulse responses can be informative when assessing the geometric characteristics of a room. Wave-based acoustic modeling methods are suitable for approximating low frequency wave propagation. Subsequent auralizations are perceived unnaturally due to the limited bandwidth involved. The paper presents a post-processing framework for extending low-mid frequency band limited spatial room impulse responses (SRIR) to include higher frequency signal components without the use of geometric modeling methods. Acoustic parameters for extrapolated RIRs are compared with reference measurement data for existing venues and a Finite Difference Time Domain modeled SRIR is extrapolated to produce a natural sounding full-band SRIR signal. The method shows promising agreement particularly for large venues as the air absorption is more dominant than the boundary absorption at high frequencies.

1. INTRODUCTION

Auralization is the audible rendering of a modeled soundfield leading a listener to a simulated but natural binaural listening experience, after [1]. Auralization of acoustic models has been achieved using geometric-based methods e.g. RAVEN and DIVA systems [2][3]. This is in addition to the auralization facilities offered in commercial geometric modeling products such as Odeon (odeon.dk). Geometric-based auralizations are inherently *full-band* meaning, in the context of this work, they are performed over the full audible frequency band. This is because reflections are represented in the time-domain as broadband pulses and subsequently full-band auralizations inevitably sound natural. Generally, the pulses are weighted according to the air absorption, boundary characteristics and associated reflection arrival time at the receiver. The reflection arrival time describes the total distance travelled by the reflection and therefore the incurred energy attenuation due to the inverse square law of acoustics. While geometric methods inherently offer full-band room impulse response auralization, their low frequencies are also contrived due to the underlying assumption that sound waves propagate in a ray-like manner.

In room acoustics it is well known that the low frequencies can be particularly problematic. Therefore geometric methods are not the optimal choice when the auralization of low-mid frequencies is of greater importance. Wave-based modeling techniques discretely implement a numerical solution to the wave equation with some appropriate boundary conditions. Generally, an aspect of the room geometry, either the air volume or the enclosing boundary, is represented by dividing it into a discrete number of sections. The chosen discrete solution of the wave equation is then implemented for each section. As a result of this physically founded approximation, the wave-based acoustic simulations offer greater accuracy in the low frequency bands and can model wave phenomena including wave interference, occlusion and diffraction inherently. Wave-based techniques include Finite Difference Time Domain (FDTD) e.g. [4] and most recently Adaptive Rectangular Decomposition (ARD) [5]. In ARD a simple method for extending the low frequency response was discussed, although this did not preserve the spatial cues for auralization. In wave-based techniques computation times increase as a function of the volume and/or desired frequency band. Subsequently, for large room acoustics in particular, the modeled frequency band is limited leading to unnatural sounding auralizations. One solution is to employ a suitable geometric model for the high frequencies and combine the resulting SRIRs as was considered previously [6, 7]. This however can be sub-optimal when it is low-mid frequencies that are of interest.

Therefore this paper presents a novel approach for producing full-band auralizations from low-mid frequency band-limited room impulse responses. In this context, a method for estimating temporally varying directional and diffuse parts of measured or synthesized SRIRs is also discussed.

The structure of this paper is as follows: Section 2 the directly relevant previous work is discussed. Section 3 outlines the proposed extrapolation method and then discusses the analysis and synthesis parts in detail. Section 4 provides some objective assessment as an indicator of the subjective reliability of the extrapolated signal followed by some conclusions in Section 5.

2. SPATIAL IMPULSE RESPONSE RENDERING (SIRR)

Related work, SIRR presented by Merimaa and Pulkki, is the foundation for the work presented here [8]. SIRR is a technique for rendering a measured soundfield with an arbitrary number and positioning of loudspeaker reproduction channels. The general approach is to measure physical properties of the soundfield in an analysis stage and then relate these to human localization cues for soundfield rendering. The time-frequency dependent nondiffuse and diffuse components of the soundfield are identified in the analysis stage. Specifically, the analysis output is a pressure signal and a meta-data stream carrying the direction of arrival and the diffuseness of each time-frequency component. In the SIRR synthesis the soundfield rendering is performed by amplitude panning nondiffuse components to the estimated direction in the meta-data, while diffuse parts are produced by all reproduction channels but are decorrelated to avoid forming directional cues within a listeners auditory system. In this work we intend for the auralization to be carried out using the previously established SIRR approach.

The instantaneous sound intensity is defined as the product of a particles' instantaneous acoustic pressure at a point and its

* Thanks to Samuel Siltanen for the image-source algorithm.

associated velocity through that point in a given direction [9]. In practice the particle velocity cannot be directly measured and so a related quantity, the pressure gradient, is measured instead. It has been shown that by measuring the sound intensity in the directions of the Cartesian coordinate system that the average direction of arrival can be estimated after [8]:

$$\theta(\omega) = \tan^{-1}\left[\frac{-I_y(\omega)}{-I_x(\omega)}\right] \quad (1)$$

$$\phi(\omega) = \tan^{-1}\left[\frac{-I_z(\omega)}{\sqrt{I_x^2(\omega)+I_y^2(\omega)}}\right] \quad (2)$$

where $\theta(\omega)$ and $\phi(\omega)$ denote the azimuth and elevation of the arrival direction as a function of radial frequency and $I_x(\omega)$, $I_y(\omega)$ and $I_z(\omega)$ are the measured sound intensity vectors in the frequency domain.

Equations (1) and (2) can produce ambiguous direction of arrival values such as in the late reverberation tail of an RIR where the soundfield is more diffuse. The diffuseness estimate $\psi(\omega)$ provides an indication of when the arrival directions in (1) and (2) can be considered as nondiffuse or diffuse energy and is defined as the ratio of the sound intensity with the energy density after [8]. When $\psi(\omega) = 0$ it is ideally nondiffuse and $\psi(\omega) = 1$ is ideally diffuse.

3. THE SPATIAL HIGH FREQUENCY EXTRAPOLATION METHOD (SHEM)

The proposed SHEM can be conceptually divided into two separate parts that in practice are carried out simultaneously, namely the SHEM analysis and SHEM synthesis. In the analysis, the general purpose is to process the existing low frequency portion of the SRIR to generate time varying meta-data describing the directionality/diffusivity of the soundfield. The meta-data consists of four time dependent parameters, *directivity*, *direction of arrival*, *average direction of arrival* and *average energy*. These meta-data parameters are obtained by using any appropriate co-incident and/or spaced array of microphones, although in this paper only one such array and processing scheme will be discussed and demonstrated.

In the synthesis portion the aim is to retain the existing low frequencies of the RIR and introduce temporally weighted high frequency energy to produce a new full-band RIR for auralization with SIRR. The early high-frequencies are also directionally rendered into the final auralization by generating a corresponding high frequency meta-data. The energy weighting is determined based on three contributing factors, those being 1) the meta-data stream from the analysis portion, 2) the known physical characteristics of wave propagation and 3) the known, or intended, acoustic properties of the room geometry. The remainder of this section will discuss one specific implementation of SHEM.

3.1. Analysis

B-Format is an appropriate microphone array input to the processing algorithm as it represents the soundfield as captured by a co-incident array of three pressure gradient microphones orientated along the *xyz* axes and a centrally located pressure microphone. This is convenient for computing the sound intensity as discussed in Section 2. The B-Format signals are processed according to the scheme outlined in Figure 1.

The input signals are processed using the short time fourier transform (STFT). The fast transient nature of the SRIR dictates that the reflection detection must be carried out using high resolution analysis frames in the time domain. Therefore short windowed and overlapping time frames are employed in the analysis, in this case 16 sample 50% overlapped hanning windows. After windowing, the frames are zero padded as required by the overlap-add method of convolution before the frequency transformation. Each frame is convolved with FIR low pass filter coefficients b_L to remove any aliasing or higher frequency artefacts in the modeled SRIR. An appropriate quantity for the average energy in each frame is taken for each input channel, in this case the sum of the mean energy over a range of frequencies from the current and past analysis frame. The sound intensity vectors are then used to calculate the direction of arrival meta-data using (1) and (2) as in SIRR. The *resultant vector* is proposed in the following for producing an average angle of arrival and directivity estimate.

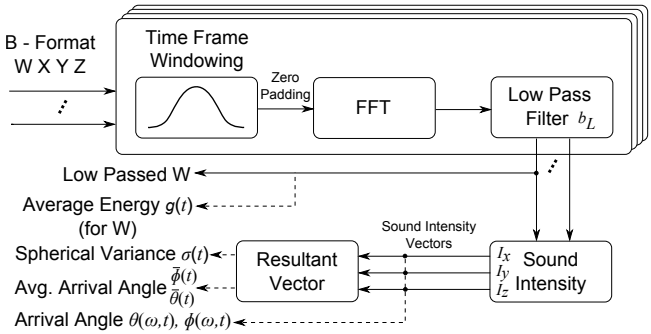

Figure 1: *The analysis portion of the SHEM implementation presented in this paper.*

3.1.1. The Resultant Vector

The sound intensity vectors across the frequency bins of a single analysis frame will point in the same direction when one or more coherent reflections arrive at the microphone array. On the other hand, as the soundfield becomes more chaotic with the arrival of high order reflections, the directions of these vectors will be distributed more randomly. This is a property that was exploited in SIRR by the diffuseness estimate discussed in Section 2. In this work a similar quantity is employed for this purpose, the *spherical variance*, and is calculated using the magnitude of the mean resultant vector.

The mean resultant vector has previously been employed in circular statistics for analyzing circular distributions [10]. For the purposes of this work, the circular formulation is extended to a spherical one as the sound intensity vectors may point in elevatory directions. Note that the spherical variance was previously employed for localizing reflections in [14]. The magnitude of each sound intensity vector in each frequency bin is regarded to have a magnitude of unity although the associated azimuth and elevation angles, calculated using (1) and (2), are maintained as in (3).

$$\mathbf{I_i} = [\cos\theta_i \cos\phi_i \quad \sin\theta_i \cos\phi_i \quad \cos\phi_i]^T \quad (3)$$

where $\mathbf{I_i}$ is the unit vector for θ_i and ϕ_i for the i^{th} frequency bin. The average of multiple sound intensity vector angles for a given analysis time frame is then calculated as in (4). Note that this could also be achieved by first normalizing the length of each sound intensity vector in each frequency bin to unity and then applying (4) on the result.

$$\bar{\mathbf{I}}(t) = \frac{1}{X}\sum_{i}^{X}\mathbf{I_i}(t) \quad (4)$$

where $\bar{\mathbf{I}}$ is defined as the mean resultant vector and X is the total number of frequency bins in the passband of the low pass filter

b_L. The mean angular direction for the time frame $[\bar{\theta}(t),\bar{\phi}(t)]$ is then computed by converting $\bar{\mathbf{I}}$ back to its polar co-ordinate representation using (1) and (2), by substituting $[I_x\ I_y\ I_z]$ with $[\bar{I}_x\ \bar{I}_y\ \bar{I}_z]$ and disregarding the frequency dependence. The spherical variance is defined using the magnitude of the resultant vector as in (5):

$$\sigma(t) = 1 - \left\|\bar{\mathbf{I}}(t)\right\| \tag{5}$$

where $\sigma(t)$ is an indicator of the reliability of the angular mean $[\bar{\theta}(t),\bar{\phi}(t)]$. The operation $\|\cdot\|$ determines the magnitude/Euclidean norm of the enclosed vector. When $\sigma(t) = 1$ the sound intensity vectors are distributed evenly implying a diffuse field, and $\sigma(t) = 0$ meaning that all vectors are pointing in the same direction and that the time frame contains a clear reflection.

3.2. Synthesis

In this implementation the synthesis can be considered generally as a cross-fading in the frequency domain between the existing low frequency W channel RIR with an extrapolated high frequency RIR. The cross fade is performed using a matching pair of low and high pass FIR filter coefficients b_L (used in the analysis) and b_H of equal tap length, i.e. MATLAB `fir1`. Each analysis frame is characterized as being either directional or diffuse by testing the spherical variance meta-data against a preset threshold σ_t. When $\sigma(t) \leq \sigma_t$ the frame is directional but must also contain a local maxima in the average energy meta-data before any high frequency energy is introduced to the frame. If this is the case then a pulse, determined by b_H, is weighted and added to the low passed input signal in the centre of the current time frame according to (6):

$$\hat{Z}(\omega,t) = Z(\omega,t) + g(t)\alpha(\omega,t)\beta(\omega)B_H(\omega)e^{-j\omega d} \tag{6}$$

where $\hat{Z}(\omega,t)$ is the processed the complex frequency domain signal in analysis frame t. $Z(\omega,t)$ is the low frequency input signal, $g(t)$ and $\alpha(\omega,t)$ are the averaged energy and normalized air absorption function respectively. Previous work by Bass *et al.* defined an analytical expression to predict air absorption [11]. It has been adapted here for application to overlap-add based convolution so that the absorption weighting changes as a function of the current time frame as this implies a distance travelled, see Figure 2(a). The term $\beta(\omega)$ is an arbitrary weighting function for representing the high frequency energy loss due to the boundaries. $B_H(\omega)$ is the frequency domain transform of FIR coefficients b_H and $d = 0.5N$ where N is the length of the time frame and shifts the pulse to the frames centre.

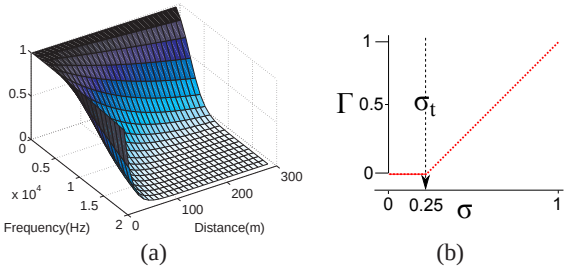

Figure 2: *(a) The normalized air absorption weighting as a function of distance travelled and frequency. (b) The diffuse weighting profile which determines the gain of the diffuse high frequency energy in a frame based on the spherical variance.*

When $\sigma(t) < \sigma_t$ the frame is potentially diffuse and the amount of high frequency diffuse energy that is added is determined by a number of contributing factors as in (7):

$$\hat{Z}(\omega,t) = Z(\omega,t) + g(t)\alpha(\omega,t)\beta(\omega)B_H(\omega)P(\omega)\Gamma(\sigma) \tag{7}$$

where $P(\omega)$ is a frequency domain generated random pulse train of the frame length whose amplitudes are uniformly distributed between -1 and 1 i.e. MATLAB `2(0.5-rand(1,length))`. $\Gamma(\sigma)$ is the diffuse weighting profile (DWP) which is defined here as in Figure 2(b). The DWP ensures that high gain diffuse energy is added gradually as $\sigma(t)$ rises above the threshold σ_t. Note that for this work, $\sigma_t = 0.25$ was chosen experimentally. In addition, directional frames $\sigma(t) \leq \sigma_t$ that are not a local maxima in $g(t)$ will not have any diffuse energy applied as $\Gamma(\sigma) = 0$ under this condition.

An overview of SHEM synthesis described in this section is given in Figure 3. In addition to estimating high frequency energy, the synthesis also requires that direction of arrival meta-data is generated. This is to attribute appropriate directional information to the new energy in each time-frequency bin for auralization with SIRR synthesis and is discussed in the following section.

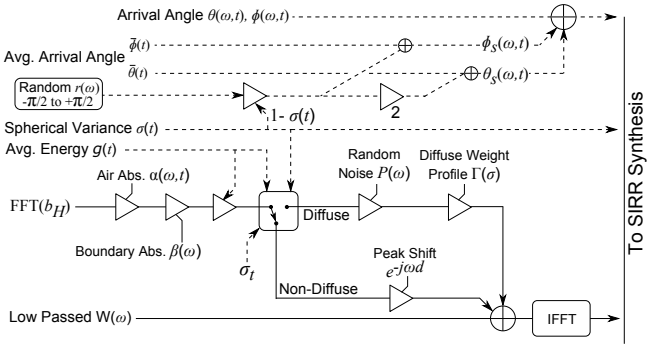

Figure 3: *The synthesis portion of the SHEM implementation. The dashed and solid lines represent meta-data and audio respectively. In the top half, high frequency direction of arrival meta-data is generated and added to the existing low frequency meta-data. The bottom portion synthesizes the high frequency energy and adds it to the existing low frequency limited signal.*

3.2.1. *Generating new directional meta-data*

The high frequency meta-data is generated based on the spherical variance $\sigma(t)$ and the average direction of arrival $(\bar{\theta}(t), \bar{\phi}(t))$:

$$\theta_s(\omega) = 2r(\omega)(1-\sigma) + \bar{\theta} \tag{8}$$

$$\phi_s(\omega) = r(\omega)(1-\sigma) + \bar{\phi} \tag{9}$$

where $\theta_s(\omega)$ and $\phi_s(\omega)$ are the randomly generated azimuth and elevation angles for frequency ω in time frame t. Note that the time dependence is omitted here for brevity and r is a uniformly distributed random number between $-\pi/2$ and $\pi/2$. The $2r$ in (8) ensures the random r values have the range $-\pi$ and π. It can be observed that for ideally directional time frames $\sigma(t) = 0$, that $\theta_s(\omega) = \bar{\theta}$ and $\phi_s(\omega) = \bar{\phi}$. When $\sigma(t) = 1$ the angles are randomly distributed in all directions.

Equations (8) and (9) can be used to generate high frequency direction of arrival meta-data. This ensures that directional components are rendered with a panning function and diffuse components are rendered in all directions in a decorrelated fashion as defined in SIRR [8].

4. VALIDATION & RESULTS

A B-Format SRIR is calculated for a 10m,6m,3m shoebox using the image-source (IS) method and FDTD method with a valid response up to around 2.5kHz. Figure 4 shows the overall response for the FDTD extrapolated case as well as the early part without the DWP in the synthesis. In (a) the overall effect of the presented SHEM approach is apparent from the spectrograms, the technique adds temporally weighted high frequency energy to the existing low frequency signal. From (b) and (c) the need for the DWP is clear - the early part shown in (b), even the direct sound, is incorrectly flooded with high frequency diffuse energy. Comparing (c) and (d) it can be observed that the early reflections have been identified by using the mean resultant vector parameters discussed in Section 3.1.1. The first reflection has greater relative amplitude in (c) compared with the direct sound. This could also be seen in the input and it occurs because multiple reflections correctly sum at the receiver location, something that does not occur in the IS model.

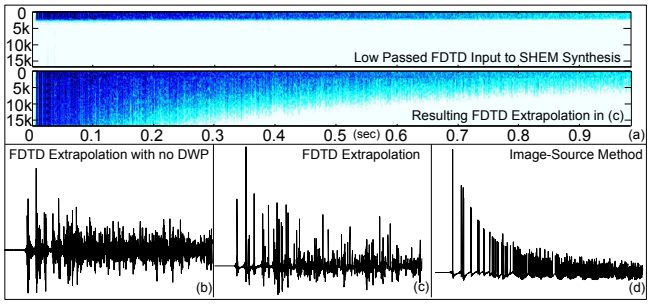

Figure 4: *(a) FDTD extrapolated W channel for the shoebox. (b) Image-source response for the shoebox (c) Result in (a) without DWP (d) Spectrogram of synthesis input and the result in (a).*

Real B-Format SRIRs have been measured previously in three Finnish concert halls [12] and York Minster [13] and are used here as they provide an ideal target extrapolation. These SRIRs are low pass filtered around 2.2kHz to represent the acoustic modeling output. The actual and extrapolated W channel signals are compared in terms of RT_{60} in 1/3 octave bands $\geq$ 1kHz, as in Figure 5, providing an objective indication of the subjective accuracy in an auralization. For clarity only the extreme real case results are shown, the best being the Minster and the worst being a concert hall. The average high frequency boundary absorption β is unknown for the real venues and was subsequently set with an arbitrary high frequency roll-off characteristic. This comparison also includes the extrapolated signal RT_{60} values with the air absorption omitted from the synthesis. It is clear that the air absorption contributes significantly in the extrapolation even for the concert hall cases. It is reasonable to conclude that the Minster extrapolation is most accurate as the air absorption is more dominant than β due to the larger volume involved.

5. CONCLUSIONS & FUTURE WORK

SHEM was proposed for extending low frequency band limited SRIRs to a natural sounding fullband auralization using the mean resultant vector for estimating the average direction of arrival, as well as diffuseness, in a given time frame. The synthesis portion of the established SIRR approach was taken as the method of auralization. The low frequency modeled portion of the SRIR is not altered and can therefore be reliably auralized by SIRR. The extrapolated high frequencies offer a more natural listening experience and a method for generating appropriate high frequency spatial meta-data for SIRR was also given. RT_{60} provided an objective measure of the subjective reliability of the method and indicated that the technique is most accurate for larger spaces where the air absorption dominates over the high frequency boundary absorption.

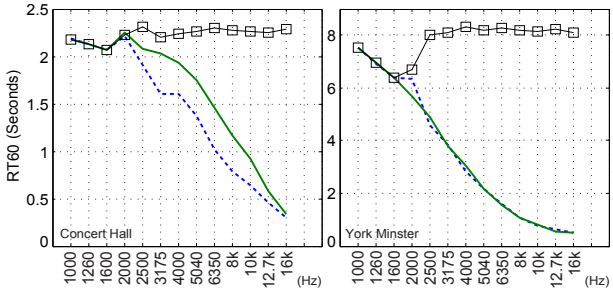

Figure 5: *The RT_{60} in seconds as a function of one-third octave bands $\geq$ 1kHz for the extrapolated and reference cases. The solid lines are the real hall reference cases and the dashed lines are the extrapolated results. The solid lines with square markers indicate no air absorption cases.*

Further work should include an analytical formulation of high frequency boundary characteristic β and a fullband FDTD reference SRIR should be computed to provide a reference/target response. The method could potentially be extended to extrapolate the XYZ channels directly for reproduction with Ambisonic decoders, though avoiding processing artefacts would require greater attention. The SHEM analysis could also be tested with any microphone arrangement where the direction of arrival and diffuseness may be estimated. The performance of the spherical variance as a diffuseness parameter should be compared with other diffuseness estimators. Finally, the objective results presented here are promising, although informal listening clearly suggests the extrapolated cases offer a more natural listening experience. Comparing the reference and extrapolated cases, critical differences are very difficult to hear for orchestral music but faster transients are easier to discriminate, e.g at http://www.tml.tkk.fi/~aps/dafx12.html. Auralization based listening tests should be carried out to further assess the performance of the proposed SHEM technique.

6. References

[1] M. Kleiner, B.I. Dalenbäck, and P. Svensson, "Auralization-an overview," *J. Audio Eng. Soc.*, vol. 41, no. 11, pp. 861–875, 1993.

[2] D. Schröder, F. Wefers, S. Pelzer, D. Rausch, M. Vorländer, and T. Kuhlen, "Virtual reality system at RWTH Aachen university," in *ISRA 2010*, 2010.

[3] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Virtual environment simulation - advances in the DIVA project," in *(ICAD'97), Palo Alto CA, USA*, 1997, pp. 43–46.

[4] K. Kowalczyk and M. van Walstijn, "Room acoustics simulation using 3-D compact explicit FDTD schemes," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 1, pp. 34–46, 2011.

[5] N. Raghuvanshi, R. Narain, and M.C. Lin, "Efficient and accurate sound propagation using adaptive rectangular decomposition," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, pp. 789–801, 2009.

[6] M. Beeson, A. Moore, D. Murphy, S. Shelley, and A. Southern, "Renderair - room acoustics simulation using a hybrid digital waveguide mesh approach," in *Audio Eng. Soc. (AES) Convention 124*, 2008.

[7] A. Southern, S. Siltanen, and L. Savioja, "Spatial room impulse responses with a hybrid modeling method," in *AES 130 Conv.*, 2011.

[8] J. Merimaa and V. Pulkki, "Spatial impulse response rendering I: Analysis and synthesis," *J. Audio Eng. Soc*, vol. 53, no. 12, pp. 1115–1127, 2005.

[9] F.J. Fahy, *Sound Intensity*, Elsevier Applied Science, London, 1989.

[10] P. Berens, "Circstat: A MATLAB toolbox for circular statistics," *J. Statistical Software*, vol. 31, no. 10, 2009.

[11] S. Tervo, T. Korhonen, and T. Lokki, "Estimation of reflections from impulse responses," in *ISRA 2010*, 2010.

[12] H. E. Bass, H.J. Bauer, and L.B. Evans, "Atmospheric absorption of sound: Analytical expressions," *J. Acoust. Soc. of Am.*, vol. 52, no. 3B, pp. 821–825, 1972.

[13] T. Lokki, J. Pätynen, A. Kuusinen, H. Vertanen, and S. Tervo, "Concert hall acoustics assessment with individually elicited attributes," *J. Acoust. Soc. of Am.*, vol. 130, no. 2, pp. 835–849, August 2011.

[14] "OpenAIR, Audiolab, University of York," Online at www.openairlib.net.