DOI: 10.22032/dbt.39968

DOI (proceedings): 10.22032/dbt.39936

Verband Deutscher Tonmeister e.V.

Abstract Reviewed Paper at ICSA 2019

Presented * by VDT.

Real-time Estimation of Reverberation Time for Selection of suitable binaural room impulse responses

F. Klein[1], A. Neidhardt[1], M. Seipel [1]

[1] *Technische Universität Ilmenau, Germany, Email: florian.klein@tu-ilmenau.de*

Abstract

The aim of auditory augmented reality is to create a highly immersive and plausible auditory illusion combining virtual audio objects and scenarios with the real acoustic surrounding. For this use case it is necessary to estimate the acoustics of the current room. A mismatch between real and simulated acoustics will easily be detected by the listener and will probably lead to In-head localization or an unrealistic acoustic envelopment of the virtual sound sources. This publication investigates State-of-the-Art algorithms for blind reverberation time estimation which are commonly used for speech enhancement algorithms or speech dereverberation and applies them to binaural ear signals. The outcome of these algorithms can be used to select the most appropriate room out of a room database for example. A room database could include pre-measured or simulated binaural room impulse responses which could directly be used to realize a binaural reproduction. First results show promising results combined with low computational effort. Further strategies for enhancing the used method are proposed in order to create a more precise reverberation time estimation.

1. Introduction

The aim of Auditory Augmented Realities (AAR) is to enrich the real acoustic environment of the listener with additional sound objects. Thus, it is inevitable to match the acoustic of the augmented sound objects with the acoustics of the real room or acoustic surrounding. Previous research has shown that a mismatch of the virtual and real acoustics can lead to in-head localization which means, that the sound objects are not located outside the head as usual for natural sound sources [12]. This effect was named room-divergence effect [9]. The current study focuses on the reverberation time as room dependent value, but of course a room has much more characteristics and properties which contribute to the "identity of a room". The pattern of the first reflections, timbre of the reverb, room acoustical modes and others are maybe equally important. Additionally, many of these parameters change when the listener moves through the room which gives him or her an impression how the room sounds.

In order to realize a listening room dependent acoustic simulation, it is possible to create a 3D model of room and conduct a simulation [11] or to select an appropriate set of binaural room impulse responses (BRIRs) out of a database. On the one hand, approaches which generate or measure BRIRs beforehand require an extensive database but on the other hand the computational load during the rendering of the ear signals is low, because only the BRIR selection has to be done in real time. In conjunction with methods for the interpolation and extrapolation of BRIRs of a moving listener [10], the required memory for a BRIR database can be reduced. This approach is the motivation behind this publication.

2. State of the Art

Auditory Augmented Realities is a emergent research field and there are no state of the art approaches to include the real acoustics into the simulation. But of course there are several ways to do so. A database of rooms and acoustic surroundings can be used in conjunction with artificial neural networks for

	Lecture Room	Meeting Room	Office Room	H2505
Size [m]	10.8 x 10.9 x 3.15	8 x 5 x 3.1	5 x 6.4 x 2.9	9.9 x 4.7 x 3.1
Meas. Dist. [m]	4, 5.56, 7.1, 8.68, 10.2	1.45, 1.7, 1.9, 2.25, 2.8	1, 2, 3	1, 2, 3, 4, 5, 6, 7, 8

Tab. 1: Dimensions of the rooms from which the BRIRs were used in this study. Lecture, meeting and the office rooms are located at the RWTH Aachen and the H2505 room at the TU Ilmenau. The measurement positions indicate the different distances between sound source and artificial head.

acoustic scene classification [14]. Those methods may also be able to estimate room acoustic parameters or room geometries in the future. By including optical information a room size or reverberation estimation might be improved. When multiple cameras or time of flight cameras are available, a 3D model of a room can be created [15]. Based on this, an acoustic room simulation can be realized which synthesizes new binaural room impulse responses. This paper focuses on blind T60 estimation methods known from communication engineering. Blind T60 estimations are utilized for dereverberation in order to enhance speech perception, for automatic speech recognition or for acoustical scene analysis [4]. Because of these popular applications several methods for blind reverberation time estimation from speech exist [6]. In a comparative study by Eaton [5] the algorithms by Prego [7] and Löllmann [8] delivered the best results. Both algorithms use framewise subband analysis of free decaying regions. Eaton states, that algorithms which use features based on the decay rate are most accurate.

3. Reference and test data

Two different datasets of binaural room impulse responses (BRIRs) were used for this work. A freely available data set from Jeub et al. [3] includes BRIR measurements of three different rooms at the RWTH Aachen University. Another data set was recorded at the TU Ilmenau in room H2505, which is a seminar room. Thus, four different rooms with different dimensions and acoustical properties were available (see table 1 for details). In each room, measurements with different distances to the sound source were conducted. For the measurements in the lecture, meeting and office room the artificial head HMS2 from HEAD Acoustics was used. For BRIR measurements in Aachen, pseudo-random noise was used to excite the space [3]. The measurements in room H2505 were conducted with an artificial head type KEMAR 45BA of the manufacturer G.R.A.S. The room H2505 was excited with a sweep from 60Hz to 20kHz. For all positions BRIRs with head direction of 0° azimuth were used. The default orientation of the sound source is frontal towards the dummy head. For room H2505 there was an additional condition with the sound source turned to the opposite direction.

A reference T60 dataset was derived from these measurements. At first the T30 is calculated on basis of the energy decay curve of the BRIRs and then the T60 value is calculated from the T30 value. This calculation is done for left and right BRIR and the T60 valued are averaged afterwards. This is justifiable, because the head is oriented towards the source and therefore left and right ear signals should be similar. However, there were small differences because the room acoustics were not symmetric or the measurement positions were not centered in the rooms.

For the preparation of the test data, the BRIRs are convolved with a $30s$ speech signal. Three additional variants were created by adding white noise to create a signal with $30dB$, $20dB$ and $10dB$ SNR next to the basic version without artificially added noise. This was done to simulate more realistic and less sophisticated recording equipment like microphones typically used in consumer electronics.

4. T60 estimation algorithm

An algorithm proposed by Löllman et al. [2] was used to estimate the reverberation time. In contrast to the aforementioned algorithm Löllman (see section 2) this algorithms makes no use of a subband decomposition. The advantages are a low computational complexity and robustness against background noise [5]. The algorithm uses a statistical decay model of the transient signal components from a mono speech signal and their selection for a blockwise T60 calculation. The selection of interesting frames is done via a maximum value, minimum value, and energy comparison of successive subframes within every frame. The calculated values are compared with those of the previous subframe and it is determined whether the entire frame has a descending structure. If this is the case, the frame is considered in the T60 calculations. If this is not the case, the current frame is discarded and the next frame is analyzed. For each subframe a decay parameter is determined which can be used for T60 calculation. With the help of a maximum-likelihood estimation, the most probable reverberation time is determined by the probability density of the decay parameter of the selected frame, thereby calculating the T60 for each frame. The calculated values are stored in a T60 histogram. The variance of the stored T60 data is last reduced by a recursive smoothing to get a more reliable final estimation. The binsize of the histogram describes the quantization step in which the T60 values are calculated. For the calculations, the smoothing factor α and the binsize were adapted to achieve the best possible results for the T60 values in this case. The settings of the smoothing factor and binsize have an effect on the accuracy of the estimation for different reverberation times. A smaller binsize and a smaller smoothing factor is beneficial for shorter reverberation times and respectively, the algorithm archives better results for longer reverberation times with a larger binsize and a higher smoothing factor. A higher downsampling factor speeds up the process. The settings for this study were determined empirically. Table 2 shows the selected values.

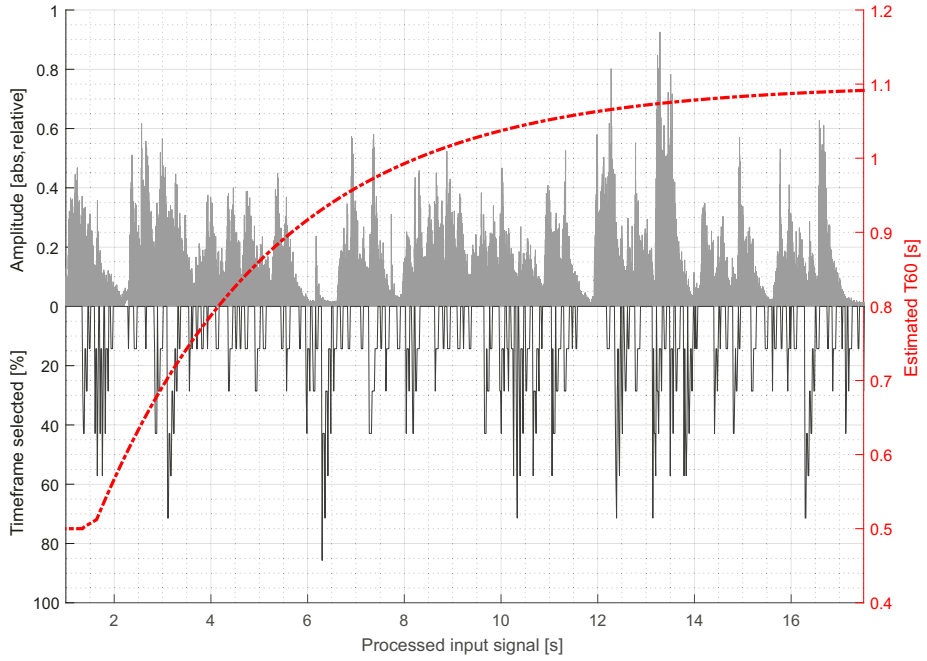

Fig. 1: Example of the frame wise processing of the input signal. Figure shows the waveform of a speech signal in room H2505, the percentage of each frame which was used for calculation as well as the progression of the T60 estimation.

frame size [n]	5740
subframe size [n]	820
binsize	0.15
smoothing factor α	0.996
downsampling factor D	2
input size [s]	30

Tab. 2: Empirically selected parameters for the T60 estimation algorithm by Löllman.

The algorithm works in an interval of the reverberation time from $0.2s$ to $1.2s$ and with a recommended minimum input signal duration of $10s$. The quality of the estimate changes with the SNR of the input signal. The decaying structure of the transients is distorted by the noise components and fewer frames are considered during the preselection of frames, which leads to an inaccurate T60 estimate. Therefore, input signals with a low SNR need to be preprocessed with an noise reduction algorithm to ensure a good estimate. Noise reduction was performed using a wavelet denoising process. The denoising was designed adaptively for various SNR. For a $SNR > 20dB$ no noise reduction is carried out, for $20dB \leq SNR > 10dB$ a moderate denoising and for $SNR \leq 10dB$ a strong denoising. Since white noise is assumed, an orthogonal wavelet (Daubechies10-Wavelet) was used.
Figure 1 shows the progression of the T60 estimation. The figure shows the input signal (speech signal in room H2505 with frontal condition) as absolute amplitude in the upper half of the plot. The lower half of the plot shows how many of the subframes of a frame were selected for the T60 estimation of the corresponding frame or if the frame was used at all (corresponds to 0%). A minimum of three out of seven subframes have to be selected in order to assume a sound decay structure in a frame - this corresponds to 43% of the length of a frame. The red line shows the estimated T60 over time. The start value for the T60 value defaults to $0.5s$. While the algorithm is real-time capable it takes some time to approach the final T60 estimation.

5. Experimental Results

Figures 2, 3 and 4 show the estimation results for the speech signals with different amounts of noise added. In each of the figures estimation results are separated by the rooms. Each marker corresponds to a estimation using one position in the respective room. The amount of available positions is noted in brackets and varies with rooms. The colored areas indicate the range of the T60 values derived from the BRIRs directly, thus indicating the reference range of the reverberation time.
Figure 2 shows the condition without additional noise. For the meeting room and both conditions of the H2505 room, all estimations are inside or close to the reference range. For the other two rooms we see several outliers. However, in a classification task those rooms would be easily distinguishable.
However with raising noise levels the estimation gets worse. Figure 3 shows the condition with $30dB$ SNR and without denoising applied. The estimations for the H2505 room are still within the reference range. For the other rooms the estimations become more scattered.
Figure 4 shows the condition with $20dB$ SNR and with denoising applied. There is a strong tendency towards longer reverberation times for the lecture, office and meeting room. Surprisingly, the results for the H2505 are still very good. The results for the condition with $10dB$ SNR and with denoising are not shown here, because they are very similar to the $20dB$ SNR condition and shows the same tendencies. The spread of the estimations is increased again, especially for the smaller

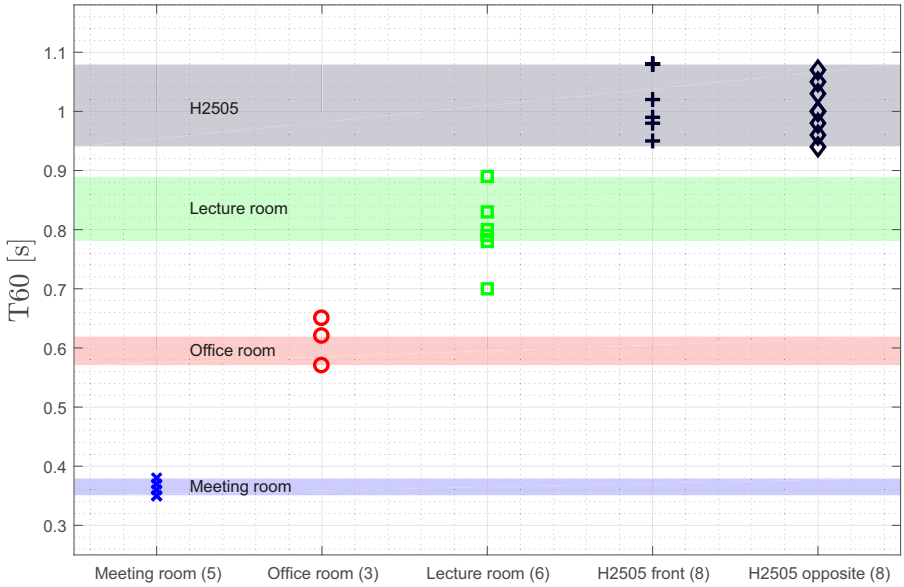

Fig. 2: T60 Estimation performance without additional noise. Markers show estimation results for each position in the corresponding rooms. The colored areas indicate maximum and minimum reverberation times in each room derived directly from the BRIRs.

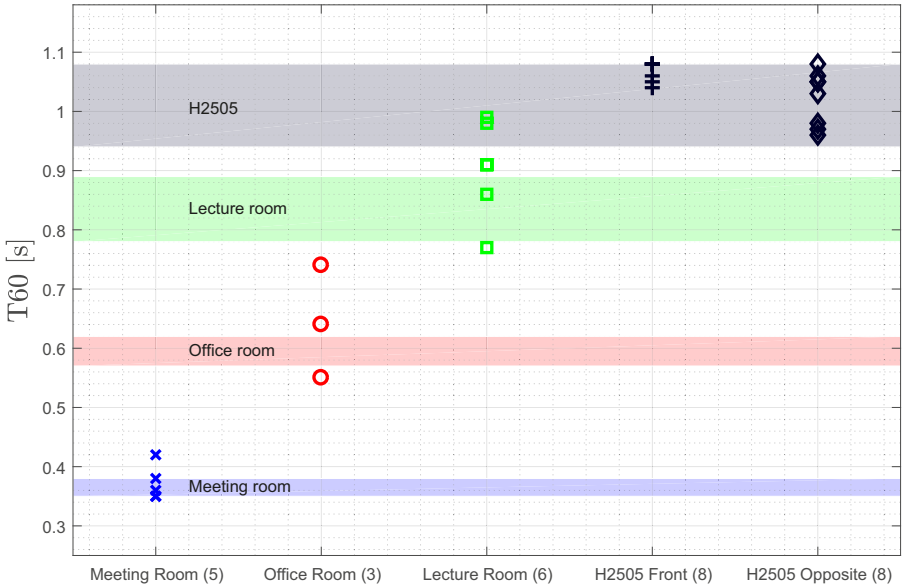

Fig. 3: T60 Estimation performance with $30dB$ SNR. Markers show estimation results for each position in the corresponding rooms. The colored areas indicate maximum and minimum reverberation times in each room derived directly from the BRIRs.

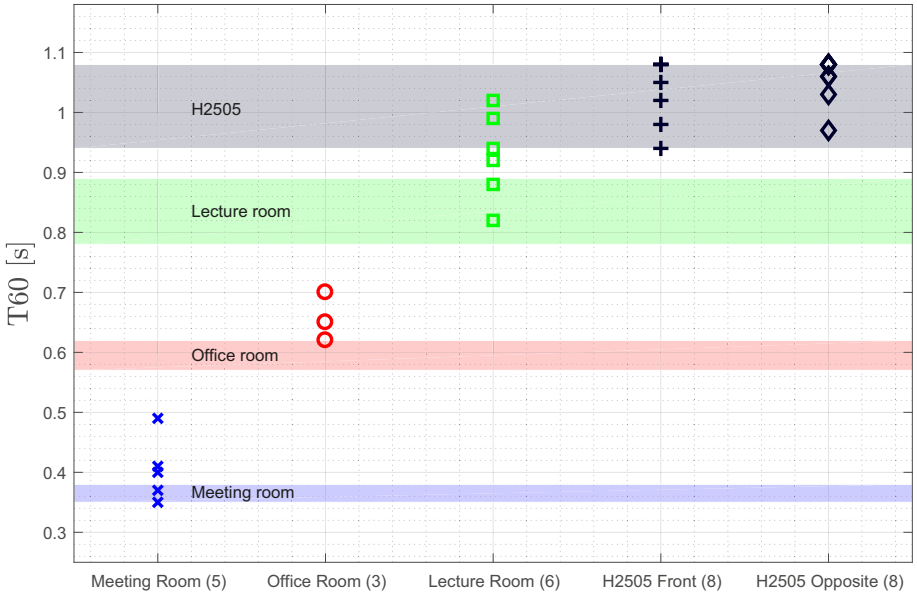

Fig. 4: T60 Estimation performance with $20dB$ SNR and denoising. Markers show estimation results for each position in the corresponding rooms. The colored areas indicate maximum and minimum reverberation times in each room derived directly from the BRIRs.

meeting and office room. Raising noise levels modify the energy decay curve and thus lead to an overestimation of the reverberation time.

5.1. Analysis of the DRR

The direct to reverberant ratio is a position and room depended value and therefore a correlation with the T60 estimation was suspected. The DRR is calculated on the basis of the BRIRs and corresponds to the average from left an right. The DRR values in the meeting room range between $8.3dB$ and $6.8dB$, which means a high proportion of direct sound at all positions. In the office room the position depended variation is significant higher. The DRR decreases rapidly with increasing distance from $8.2dB$ to $-1dB$. This is due to the many differently reflective surfaces of the room furnishing. This results in a sharp increase in diffuse sound, even with small changes in distance of one meter each. As the diffuse sound component increases, the results of the estimation improves. However, this is only the case for test sounds without additional noise and it must be considered that there were only three measurement positions in this room. In the lecture room the DRR drops sharply from $7.2dB$ to $-5.7dB$ as the distance from speaker to microphone increases. Accordingly, the direct sound component at the outermost position is considerably lower than at closest position. However, no dependency between DRR and T60 estimation performance can be observed. The T60 estimations in room H2505 were the best among the rooms in this study and it is also the room with lowest DRR values. The DRR in the H2505 room front condition changes from $6dB$ to $-8.5dB$ with increasing distance. Not surprisingly, the DRR values for the H2505 opposite direction range between $-0.5dB$ and $-5.4dB$.
The DRR analysis could not show a clear dependency between DRR and T60 estimation performance. In some rooms and noise conditions such a tendency could be observed but other measurements disprove this hypothesis. The estimation performance in room H2505 in connection with the low DRR values leaves room for speculations whether or not there may be an interconnection. A study with more rooms would have to prove this.

6. Discussion and Conclusion

In this paper a state of the art approach for blind reverberation time estimation was evaluated for binaural test signals. The estimation for the selected rooms is good to distinguish the rooms. However, this task is not particular difficult given the fairly large differences between the rooms. For test signals without artificial noise the results are close or within the T60 values obtained from the BRIRs directly. With more noise the estimation gets worse, even when denoising is applied. It remains unclear why the estimation performance in room H2505 was the best and the most robust. The DRR could be an explanation, but this could not be confirmed with certainty. The recording equipment and procedure was different between room H2505 and the others, but the authors are in doubt whether this is a decisive factor.
The tuning of the parameters of the algorithm require some a priori knowledge about the rooms in order to perform best. Here parameters were chosen which should result in better estimations for rooms with a reverberation time over $0.5s$. Another restriction of the algorithm is the delay of the estimations, because the algorithm needs several seconds to approach a stable T60 estimation. It will depend on the application if that is an issue. The algorithm is tuned to speech, because a speech based decay model is used. By combining sound recognition with this approach, different decay models might be applied based on the type of the sound source. Furthermore in future researches additional

acoustical parameters can be used for a better selection of suitable BRIRs. The current state of research can not explain which acoustical parameter is most important with respect to the room divergence effect. A study by Werner [13] found no significant effect of DRR manipulation on externalization in situations of room divergence. This finding draws attention to other features like the temporal structure of BRIRs. If temporal structures are indeed relevant, a possible solution would be to combine an BRIR selection based on T60 with an optical or acoustical geometry estimation. This way BRIRs would be selected from a room which is similar to the actual listening room in terms of T60 and geometry.
Overall, the performance in light of the low computational complexity make this, and other similar algorithms, attractive for further investigations in the scope of augmented auditory realities.

7. References

[1] H. Löllmann and P. Vary, “Estimation of the reverberation time in noisy environments”, in Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC), Seattle, Washington, USA, 2008

[2] H. Löllmann, E. Yilmaz, M. Jeub and P. Vary, “An improved algorithm for blind reverberation time estimation”, inf Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC), Tel Aviv, Israel, 2010

[3] M. Jeub, M. Schafer and P. Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms”, in Proc. of International Conference on Digital Signal Processing (DSP), Santorini, Greece, IEEE, July 2009, pp. 1-4, https://doi.org/10.1109/ICDSP.2009.5201259

[4] F. Lim, M. Thomas, and I. Tashev. “Blur kernel estimation approach to blind reverberation time estimation”, IEEE International Conference on Acoustics, Speech and Signal Processing, 2015, https://doi.org/10.1109/ICASSP.2015.7177928

[5] J. Eaton, N. Gaubitch, A. Moore, and P. Naylor. “Estimation of room acoustic parameters: The ace challenge”, IEEE Transactions on Audio, Speech, and Language Processing, 2016, https://doi.org/10.1109/TASLP.2016.2577502

[6] N. Gaubitch, H. Loellmann, and M. Jeub. “Performance comparison of algorithms for blind reverberation time estimation from speech”, IEEE International Workshop on Acoustic Signal Enhancement, 2012.

[7] M. Prego, A. Lima, and S. Netto. “Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition”, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2015, https://doi.org/10.1109/WASPAA.2015.7336954

[8] H. Löllmann and P. Vary. “Single-channel maximum - likelihood t60 estimation exploiting subband information”. ACE Challenge Workshop, satellite of IEEE-WASPAA, New Paltz, NY, USA, 2015, https://arxiv.org/abs/1511.04063

[9] S. Werner, F. Klein, T. Mayenfels und K. Brandenburg. “A Summary on Acoustic Room Divergence and its Effect on Externalization of Auditory Events”. In Proc. of 8th International Conference on Quality of Multimedia Experience (QoMEX). https://doi.org/10.1109/QoMEX.2016.7498973. 2016

[10] K. Brandenburg, E. Cano, F. Klein, T. Köllmer, H. Lukashevich, A. Neidhardt, U. Sloma, and S. Werner, “Plausible Augmentation of Auditory Scenes Using Dynamic Binaural Synthesis for Personalized Auditory Realities”, in Proc. of: Conference of the Audio Engineering Society (AES) Audio for Virtual and Augmented Reality, USA, 2018.

[11] A. Zimmermann and A. Lorenz. 2008. “LISTEN: a user-adaptive audio-augmented museum guide”. User Modeling and User-Adapted Interaction 18, 5 (2008), 389-416, https://doi.org/10.1007/ s11257-008-9049-x

[12] G. Plenge, “Über das Problem der Im-Kopf-Lokalisation [On the Problem of In Head Localization]”, In: Acustica 26.5 (1972), S. 241-252

[13] Werner, S., Klein, F., and Sporer T., “Adjustment of the Direct-to-Reverberant-Energy-Ratio to Reach Externalization within a Binaural Synthesis System”, AES Conference on Audio for Virtual and Augmented Reality, Los Angeles, CA, USA, 2016.

[14] Z. Ren, K. Qian, Z. Zhang, V. Pandit, A. Baird, and B. Schuller, “Deep Scalogram Representations for Acoustic Scene Classification,” IEEE/CAA Journal of Automatica Sinica, vol. 5, no. 3, pp. 662-669, 2018, https://doi.org/10.1109/JAS.2018.7511066

[15] M. Markovic, S. K. Olesen and D. Hammershøi, “Three-dimensional point-cloud room model for room acoustics simulations”, in Proc. of Meetings on Acoustics, 19/1, 2013, https://doi.org/10.1121/1.4800237