Audio Engineering Society

Conference Paper 37

Presented at the International Conference on Spatial and Immersive Audio
2023 August 23–25, Huddersfield, UK

Projecting source directivity variations onto an existing binaural room impulse response

Pablo Abehsera Morell, David Poirier-Quinot, and Brian F.G. Katz

Sorbonne Université, CNRS, Institut Jean Le Rond ∂'Alembert, UMR 7190, Paris, France

Correspondence should be addressed to Pablo Abehsera Morell (pablo.abehsera@dalembert.upmc.fr)

ABSTRACT

We present an approach to project source directivity variations onto a binaural room impulse response measured with an omnidirectional loudspeaker. The approach consists in applying frequency-dependent gain weightings to different time windows of the impulse response, issued from the desired source directivity pattern and orientation. The end goal of the research is to achieve plausible directivity pattern perception in auralisations, thus allowing one to mimic the rotation of a sound source with a given directivity pattern in mixed reality environments. We present the first step toward this goal, examining the perceptual threshold of analysis window size on recreating authentic room rendering of a rotating directional source. Perceptual listening tests were conducted to assess the impact of window size on the perceived authenticity of auralisations made with the proposed method. The results of this preliminary study intend to inform the ongoing development of the approach, which will next be extended to allow any arbitrarily imposed directivity pattern using perceptually motivated principles. The generalisability of this approach across different source-receiver configurations in different rooms is also discussed.

1 Introduction

In mixed reality (MR) applications that include virtual sound sources, auralisation is used to endow the virtual sound sources with desired acoustical properties. To achieve plausible auralisation, the acoustic properties of virtual sound sources must be coherent with those of the acoustic space they inhabit. In MR applications that employ binaural audio where the source and the receiver may move or rotate (i.e., allowing the user 6-degrees-of-freedom (6DOF)), the acoustic properties of specific source-receiver configurations can be captured by measuring a set of binaural room impulse responses (BRIRs), which can be convolved with an audio signal to auralise a virtual source.

To faithfully capture a real acoustic space, measurements must be carried out to accommodate for all potentially desired source-receiver configurations. With currently available technology, such a process can be excessively tedious and time-consuming. As such, a recent line of research has explored the extrapolation of non-measured configurations from already existing measurements. Pörschmann et al. [1] presented

a method to synthesise BRIRs from an existing monaural RIR, allowing to arbitrarily set both the listener head orientation and position in the room. Arend et al. [2] extended the method, allowing further control over the processing steps and including source directivity filtering for the direct sound part of the impulse response. Mittag et al. [3] presented an interpolation method using one or three existing measurements to synthesise BRIRs at new receiver positions within a grid, later improved by Sloma et al. [4] to incorporate the directivity of the sound source.

Directivity refers to how a sound source radiates sound in all directions as a function of frequency. In a reverberant space, the directivity of a sound source will interact with the acoustics of that space, resulting in audible spectral and spatial cues for the receiver. How strong these cues are will depend on the acoustic properties of the space [5]. As directivity can significantly impact spatial and spectral cues, it is important that it is realistically implemented in auralisation [6]. Some studies, such as Blau et al. [7], have already begun to explore the effect of doing so.

In this paper, we lay the foundations of an approach to project source directivity variations onto an existing BRIR measured with an omnidirectional source. The end goal is to be able to project any given directivity pattern onto a measured BRIR in a perceptually realistic manner and at a low computational cost.

Controlling the directivity of a virtual sound source arbitrarily could allow to realistically mimic source orientation shifts by rotating the pattern. This could be useful in mixed reality applications where a sound source may rotate. In an augmented reality teleconferencing system, an impulse response could be measured at the position where a virtual speaker may be located, using a mobile phone as the source and the extended reality headset as a receiver. Then, using our approach and the known directivity of the human voice, the impulse response could be dynamically modified to reflect the orientation of the speaker perceptually.

In this preliminary study, we present the first step towards this goal by evaluating the authenticity of the approach in terms of spatial and colouration perceptual attributes, examining how the temporal resolution of the directivity projection impacts both.

2 Directivity projection approach

The approach consists of identifying three time regions of the BRIR - corresponding to the direct sound, early reflections and late reverberation - and then applying individual frequency-dependent gain adjustments to each region to modify perceived directivity. The adjustments are made by octave band, with centre frequencies ranging from 62.5 Hz to 16 kHz.

The approach is conceptually similar to that of Pörschmann et al. [1] and Arend et al. [2], but we introduce a significant change. In the interest of **(1)** keeping computational expense as low as possible, and **(2)** not requiring additional information about the room (such as its dimensions), reflection detection and image-source modelling (ISM) are avoided. While the method described in [1] and [2] uses reflection detection algorithms combined with ISM to inform the adjustment of specific salient reflections, our approach treats the early reflections as a block which can be processed to an arbitrary degree of temporal resolution.

Avoiding ISM modelling and reflection detection may incur a risk of losing precision and therefore, the ability to project directivity realistically. The current study first investigates the proposed method's perceptual impact in an ideal case using measured reference data. We currently limit the study to measured BRIRs, while future work could examine extensions to Ambisonic datasets, both measured and simulated.

2.1 Processing steps

In the test case presented, the values for the adjustments are based on an analysis of the input BRIR (representing the test room measured with an omnidirectional source) and a reference BRIR (representing the same room measured with a directional source at different orientations).

2.1.1 Direct sound

The direct sound region is defined as follows. The first onset is determined using a relative peak threshold of -20 dB. The direct sound region spans from 0.25 ms before the first onset until 1 ms after it.

This process is carried out on the input and reference BRIRs. The RMS level of this region is then calculated across octave bands. To **project directivity** (*i.e.* applying the reference directivity to the input BRIR), a

per-band gain adjustment is applied to its direct sound region so that the RMS level of each band matches that of the reference BRIR.

2.1.2 Early reflections

The early reflections region spans from the end of the direct sound region until the measured mixing time of the input BRIR, estimated using the echo-density approach [8, 9] as implemented in the AKtools toolbox [10]. The mixing time is calculated individually for each channel, and the longest of the two is selected. Note that the mixing time of the input BRIR is used to define this region for both input and reference BRIRs.

The defined early reflections region is then divided into an arbitrary number of rectangular time windows. The per-band RMS analysis and adjustment method previously described for the direct sound region is then applied for each time window in the early reflections region.

Dividing the early reflections region into time windows allows one to adjust the temporal resolution of the processing: using a small window size should produce a more precise projection than applying one generalised adjustment to the whole region (*i.e.* applying a single window size spanning the entire early reflections region). However, using a larger window (or lower temporal resolution) should reduce the computational cost of the processing. The lowest resolution available is one window covering the whole early reflections region, while the highest tested resolution divides it into 1.25 ms windows.

Figure 1 illustrates the impact of using a high resolution setting over a low resolution setting, for one channel. In the reference impulse response, the reflection arriving at 20 ms is the point of highest energy. The high resolution projection is able to recreate this because it processes the early reflections region in 1.25 ms windows. On the other hand, the low resolution projection keeps a time-energy pattern similar to that of the input impulse response, where the reflection arriving at 10 ms contains more energy than the one arriving at 20 ms.

2.1.3 Late reverberation

The late reverberation region spans from the end of the early reflections region to the end of the BRIR. As with the direct sound region, the RMS level per octave band is calculated for the entire region and a per-band gain adjustment is then applied to the input BRIR region to match the per-band RMS levels of the reference BRIR region. This adjustment compensates for the late energy difference between the reference and the input BRIR that is a result of their different directivity patterns and how they excite the room's acoustics.

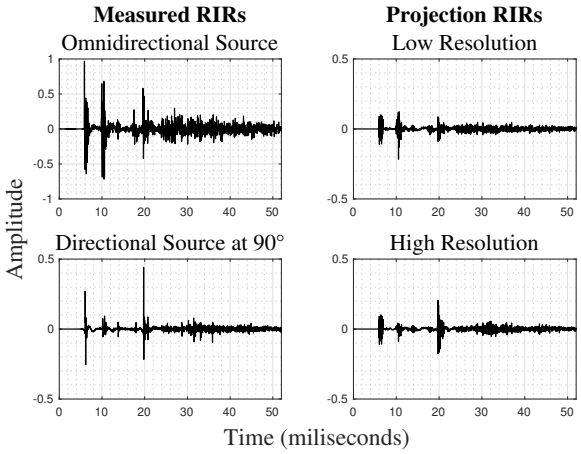

Fig. 1: Example of directivity projection at different resolution settings. Input peak normalised, reference peak normalised across all orientations.

3 Perceptual evaluation

A listening test was conducted to evaluate the perceptual validity of the proposed method across different rooms, investigating the optimal window length for achieving perceptually realistic results in the ideal case presented.

3.1 Stimuli

Input BRIRs were measured with a Neumann KU-80 binaural head fitted with DPA 4060 microphones and an omnidirectional Look Line loudspeaker in three different rooms: a small office, a meeting room, and a lecture theatre. Reference BRIRs were measured for the same source-receiver configurations with a Genelec 8331A loudspeaker. They were measured for source orientations spanning 0° to 350° in 10° steps, where 0° corresponds to the source facing the receiver. The source and receiver were always placed 2 m apart at 1.5 m height. The measurement setups are illustrated

Fig. 2: RIR measurement setups in the three rooms: (top) medium meeting room with omnidirectional source setup, (bottom-left) small office room with directional source setup, (bottom-right): large lecture theatre.

in Figure 2. In the small office, the midpoint between the source and receiver was approximately the centre of the room, skewed towards the wall behind the source by 0.5 m. In the meeting room, that midpoint was the exact centre of the room. In the lecture theatre, the source and receiver were placed towards the end of the room.

BRIR datasets were generated for each room, projecting the directivity of the directional source onto the omnidirectional source for all the measured directional source orientations. As each dataset contained source orientation values from 0° to 350° in 10° steps, each one consisted of 36 BRIRs. Datasets were generated for various directivity projection temporal resolutions, with window sizes in the early reflections region of 1.25 ms, 2.5 ms, 5 ms, 10 ms, and a single window extending to the room's mixing time. As an lower anchor candidate, a dataset was generated where the projection was applied only to the direct sound region, leaving the early and late regions unaltered.

Auralisations were made using the measured directional source BRIRs and the generated BRIRs. Two stimuli were used individually: an excerpt of anechoic female speech taken from the OpenAIR database [11] and a series of recorded anechoic hand claps. Both stimuli had an approximate length of 10 second. Each auralisation comprised a full anti-clockwise rotation

Table 1: Room information. $T30_{mid}$ is the mid-frequency reverberation time (mean over 500 Hz & 1 kHz octave bands. Mixing time (mt) is the max|mt| of the 2 channels.

ID	Description	Volume (m^3)	$T30_{mid}$ (s)	mt (s)
Small	Small Office	48	0.39	0.05
Medium	Meeting Room	324	0.47	0.05
Large	Lecture Theatre	980	1.06	0.09

by cycling through the 36 BRIRs from the perspective of the source where the initial orientation was facing straight at the receiver. Interpolation was used during transitions to avoid artefacts.

3.2 Listening test

The auralisations were presented in a MUSHRA-style test designed using webMUSHRA [12]. On each test page, participants were presented with 1 explicit reference auralisation and seven conditions to rate. The reference auralisation was made with the measured BRIRs; the seven conditions consisted of the five auralisations made with the generated BRIRs with various projection resolutions, the one anchor auralisation (direct sound modification only), and one hidden reference auralisation. At the top of each test page, participants were asked to rate one of two qualities of the auralisations in terms of similarity to the reference: **sound colour** and **spatial variations**. An information sheet containing explanations for both terms was provided.

The two rating sequences were presented twice for each combination of room and stimulus, totalling 24 test pages. To minimise participant fatigue, the test was broken down into two 12-page blocks, where each block contained all test pages for one stimulus. Participants were invited to take a break in between blocks. Half of the participants started with the speech stimulus block, the other half with the clap stimulus block. Within a block, all test pages were randomised.

For each set of auralisations corresponding to a combination of room and stimulus, the reference was normalised to −30 LUFS. Then, the adjustment needed to achieve this normalisation was applied to the other auralisations. This ensured equal playback level across test pages containing auralisations of different rooms

while maintaining the level relationships between conditions.

The auralisations were presented through Sennheiser HD650 headphones, with the playback level being calibrated to ≈65 dBA when playing the reference auralisation of the speech stimulus in the medium room. The listening experiment was conducted in a semi-isolated, acoustically treated studio room.

Once a participant was finished with the experiment, they were asked to share any thoughts in an informal interview.

3.3 Participants

20 participants, 4 female and 16 male, were recruited for the study. The mean age was 32 ± 9.7 years. 6 participants stated spending more than 3 h per week listening critically to stereo or binaural audio, 6 spent from 0 to 2 h, and 8 stated no time spent on critical listening. The average time spent on each of the two blocks across participants was 40 min. Participants were compensated 15€.

4 Results

4.1 Analysis method

Analyses of variances (ANOVAs) of participants' ratings were conducted to assess the effect of the factors: **(1)** projection method resolution, **(2)** room, and **(3)** stimulus, and the first-order interaction terms between them. Statistical significance was determined for p-values below a 0.05 threshold. The notation $p < \varepsilon$ is adopted to indicate p-values below 10^{-3}. Post-hoc pairwise comparisons for significant factors were made with Tukey-Kramer adjusted p-values, or with Wilcoxon rank-sum p-values for unbalanced comparisons.

4.2 Preamble: participants and metrics

Presumably due to the difficulty of the task, many participants failed to comply with the suggested post-screening method suggested for MUSHRA tests [13], where rating the reference condition under 90 points over 15% of the trials results in exclusion for analysis. However, in most cases, the reference received the highest rating, so it was decided not to exclude any participants in the post-screening.

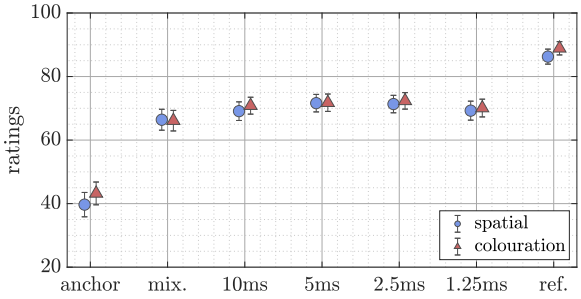

Fig. 3: Spatial rating mean and 95 % Confidence Interval (CI) across directivity projection conditions, aggregated over rooms and stimuli.

As seen in Figure 3, results for spatial and colour variations were correlated (at $r = 0.54$ and up to $r = 0.6$ for expert listeners), meaning results were similar for both metrics. In the following discussion, we refer to the results for spatial variations.

4.3 Impact of the directivity projection resolution

Window size had no overall significant impact on participants' ratings, as illustrated in Figure 3. However, there was a significant impact of the directivity projection condition on participants' ratings ($F = 82.7$, $p < \varepsilon$). The anchor condition was rated significantly lower than all the directivity projection conditions (39.7 vs. 69.5, $p < \varepsilon$), themselves rated significantly lower than the reference (69.5 vs. 86.3, $p < \varepsilon$).

4.4 Rating evolution across rooms and stimuli

In Figure 4, it can be seen that there was no significant difference observed between the various directivity projection resolutions. There was a significant impact of the stimulus condition on participants' ratings ($F = 17.4$, $p < \varepsilon$). However, there was no significant difference observed between how participants rated the anchor and the reference with each stimulus. Participants rated the other conditions significantly lower with the clap than with the speech (aggregated over projection resolutions: 65.3 vs. 73.8, $p < \varepsilon$). This result is in line with our expectation of the clap stimulus being better at revealing differences between the generated and reference impulse responses.

There was a significant impact of the room condition on ratings ($F = 23.8$, $p < \varepsilon$). As illustrated in Figure 5,

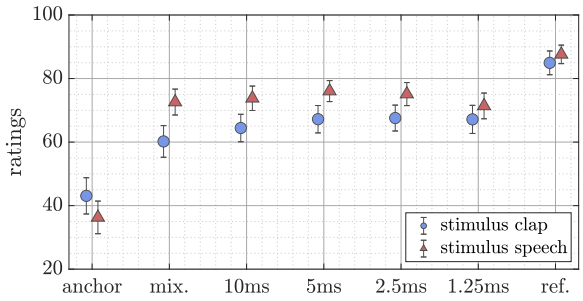

Fig. 4: Spatial rating mean and 95 % CI across directivity projection conditions and stimuli, aggregated over rooms.

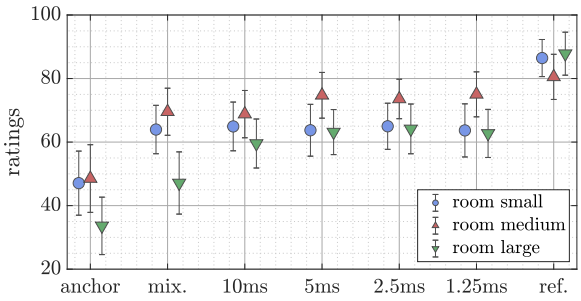

Fig. 5: Spatial rating mean and 95 % CI across directivity projection conditions and rooms, aggregated over stimuli.

renderings in the large room were rated significantly lower than those in the other two rooms (61.5 in the large room vs. 70.8 in aggregated small and medium, $p < \varepsilon$). This is mainly due to the low rating on the anchor and the **mix** conditions. The latter was rated significantly below the other projection conditions in the large room (54.4 for mix. vs. 64.7 for aggregated other projection conditions, $p = 0.006$).

Figure 6 illustrates how the interaction between the room and the stimulus led to extreme results. While the large room tended to highlight the differences between different directivity projection resolutions, the medium room tended, by contrast, to blur them. As mentioned above, projection resolution had a similar impact with both stimuli; however, the clap highlighted the directivity projection artefacts while the speech hid them. On one hand, no significant difference between the directivity projection methods and the reference can be perceived when rendering the speech in the medium room. On the other hand, there is a significant difference between the various resolutions of the projection and the reference when rendering the clap in the large room ($F = 17.7$, $p < \varepsilon$): the anchor is rated below the **mix** condition (33.6 vs. 47.1, $p = 0.032$), the **mix** below the 10ms (47.1 vs. 59.5, $p = 0.014$), and the 1.25 ms below the reference (62.7 vs. 87.8, $p < \varepsilon$). However, no significant differences are observed between the 10 ms, 5 ms, 2.5 ms and 1.25 ms projection conditions.

4.5 Interviews

A number of points were commonly raised by participants in the interviews. The listening test proved to be

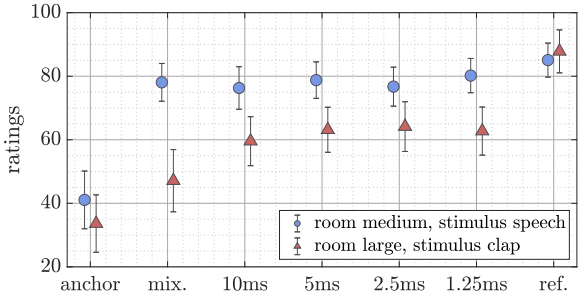

Fig. 6: Spatial rating mean and 95 % CI across directivity projection conditions for "extreme" scenarios, where the room and the stimulus both highlight (large room, clap) or hide (medium room, speech) the projection artefacts.

hard and exhausting to many. Testing only two projection conditions (such as the maximum and minimum resolution) may have prevented this; however, it was in our interest to test a wide range of resolutions.

It was often commented that differences between conditions were harder to perceive with the speech stimulus than with the clap stimulus. This is in line with our initial expectations and the test results.

Lastly, a reduced number of participants claimed to struggle to perceive the rotation of the sound source in most conditions. This could suggest that the absence of visual stimuli may have lowered the ecological validity of the task.

5 Discussion and Future Work

The results present some interesting findings. First, the proposed method was comparable to the reference in the medium room, especially in a speech auralisation context, where the lowest resolution condition seemed to perform as well as the highest resolution one. This suggests that for this kind of room and source-receiver configuration, source directivity was perceptually dominated by the overall levels of the direct, early, and late regions without the need for further time subdivisions. The performance of the various early reflection regions to the low anchor direct sound-only condition highlights the need for some adjustment of the early and late energy regions.

In contrast, in the large room, projection resolution had a perceptually significant impact. The low score of the **mix** condition suggests that, for such a room and source-receiver configuration, a resolution higher than the mixing time was required to adjust specific salient reflections individually properly. Moreover, the overall lack of success in this room suggests that the approach may not be suited to this configuration and room. Figure 7 illustrates the impact of source directivity on the resulting impulse response for the medium and large room. In the medium room, both impulse responses show a salient reflection at 10 ms. On the other hand, the impulse responses for the large room showcase vast differences in the number of salient reflections and their time of arrival (TOAs), requiring a more complex processing approach.

More surprising was the lack of success in the small room, where the projection was expected to work as well as in the medium room. In Figure 7, it can be observed that the high-resolution projection resembles the reference; however, this condition was rated significantly worse than the reference in the perceptual evaluation. This suggests that, for this room, the approach may have produced noticeable artefacts.

The planned future development of the method will seek to make it work without needing a set of reference measurements, but instead using off-the-shelf directivity data or numerical simulation. With this goal in mind, we are conducting analyses of the current measurement sets. The results of this preliminary study will help inform these analyses. In our following research, we intend to identify a model-processing concept for directivity projection, which could be used with an existing BRIR and a given directivity pattern to project directivity variations onto the BRIR. Such a concept will involve a deeper understanding of the interaction between source-receiver configuration, room acoustic parameters, and source directivity-related spatial and spectral cues. Additional investigation will examine how to improve the quality of results in the large room condition.

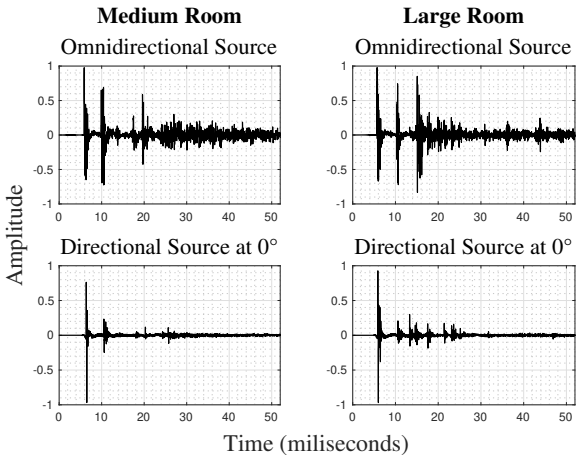

Fig. 7: Omnidirectional source and directional source impulse responses for medium and large rooms, peak level normalised.

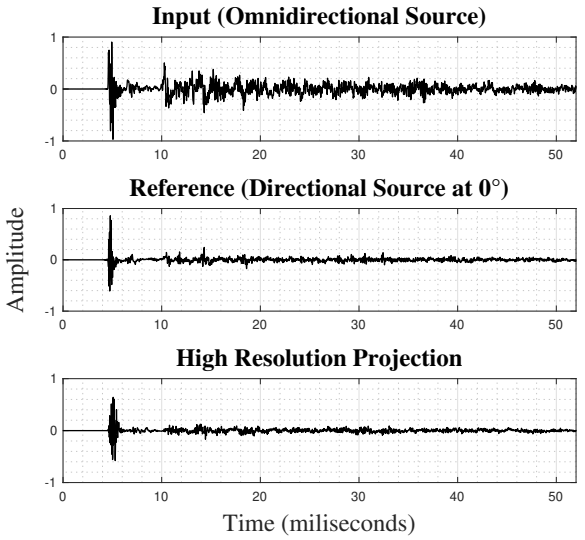

Fig. 8: Comparison of input, reference and high-resolution projection for small room.

6 Summary

We have proposed a method to project sound source directivity variations onto an existing binaural room impulse response. The method projects directivity by breaking down the BRIR into a series of time regions and applying frequency-dependent gain adjustments to each one. In the current initial development phase, reference directional BRIRs were employed.

An assessment of the relevance of temporal resolution in the processing was carried out through means of a listening test in which participants rated auralisations made with BRIRs generated using the approach. The method produced realistic projections in one of the tested rooms.

The results suggest that the method might be better suited for certain types of rooms and configurations and that temporal resolution may not be as salient as expected. We plan to develop the approach further so that it may work without needing reference measurements.

7 Acknowledgements

We would like to thank the participants who took part in the listening experiment. This work was carried out in the context of the SONICOM project (`www.sonicom.eu`) that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017743.

References

[1] Pörschmann, C., Stade, P., and Arend, J., "Binauralization of Omnidirectional Room Impulse Responses - Algorithm and Technical Evaluation," in **Proc. of the 20th Int. Conf. on Digit. Audio Effects (DAFx-17)**, pp. 345–352, 2017.

[2] Arend, J. M., Garí, S. V. A., Schissler, C., Klein, F., and Robinson, P. W., "Six-Degrees-of-Freedom Parametric Spatial Audio Based On One Monaural Room Impulse Response," **J. Aud. Eng. Soc.**, 69(7/8), pp. 557–575, 2021, doi:10.17743/jaes.2021.0009.

[3] Mittag, C., Werner, S., and Klein, F., "Development and Evaluation of Methods for the Synthesis of Binaural Room Impulse Responses based on Spatially Sparse Measurements in Real Rooms," in **43. Jahrestagung für Akustik, DAGA**, 2017.

[4] Sloma, U., Werner, S., Klein, F., and Pappachan, T., "Synthesis of binaural room impulse responses for different listening positions considering the source directivity," in **147th Aud. Eng. Soc. Conv.**, 10237, 2019.

[5] Bradley, J., Sato, H., and Picard, M., "On the importance of early reflections for speech in rooms," **J. of the Acoust. Soc. of America**, 113, pp. 3233–44, 2003, doi:10.1121/1.1570439.

[6] Steffens, H., Par, S. v. d., and Ewert, S. D., "Perceptual relevance of speaker directivity modelling in virtual rooms," in **Proc. of the 23rd Int. Congr. on Acoust.**, Aachen, 2019.

[7] Blau, M., Budnik, A., Fallahi, M., Steffens, H., Ewert, S., and Par, S., "Toward realistic binaural auralizations – perceptual comparison between measurement and simulation-based auralizations and the real room for a classroom scenario," **Acta Acustica**, 5, p. 8, 2021, doi:10.1051/aacus/2020034.

[8] Abel, J. S. and Huang, P., "A Simple, Robust Measure of Reverberation Echo Density," in **121st Aud. Eng. Soc. Conv.**, 6985, 2006.

[9] Lindau, A., Kosanke, L., and Weinzierl, S., "Perceptual Evaluation of Model- and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses," **J. Aud. Eng. Soc.**, 60(11), pp. 887–898, 2012, doi:10.14279/depositonce-15236.

[10] Brinkmann, F. and Weinzierl, S., "AKtools—An Open Software Toolbox for Signal Acquisition, Processing, and Inspection in Acoustics," in **142nd Aud. Eng. Soc. Conv.**, 309, 2017.

[11] Shelley, S. and Murphy, D., "OpenAIR: An Interactive Auralization Web Resource and Database," in **129th Aud. Eng. Soc. Conv.**, 8226, 2010.

[12] Schoeffler, M., Bartoschek, S., Stöter, F.-R., Roess, M., Westphal, S., Edler, B., and Herre, J., "webMUSHRA — A Comprehensive Framework for Web-based Listening Tests," **J. of Open Res. Softw.**, 2018, doi:10.5334/jors.187.

[13] ITU-R, "BS.1116-3 : Methods for the subjective assessment of small impairments in audio systems," 2015.