Audio Engineering Society

Convention Express Paper 117

Presented at the 155th Convention
2023 October 25–27, New York, USA

Comparison of synthesized Virtual Sound Environments with validated Hearing Aid experiments

Umut Sayin Saraç[1]

[1]*Eurecat, Centre Tecnològic de Catalunya, Multimedia Technologies Group, Barcelona, Spain*

Correspondence should be addressed to Umut Sayin Saraç (umut.sayin@eurecat.org)

ABSTRACT

Real-life situations are hard to replicate in the laboratory and often discarded during hearing aids optimisation, leading to performance inconsistencies and user dissatisfaction. As a solution, the authors propose a tool set to incorporate real-life conditions in the design, test and fitting of hearing aids. This tool set includes a spatial audio simulation framework for generating large number of realistic situations, a machine learning algorithm focused on prominent hearing aids problems trained with the newly generated data, and a low-cost spatial audio solution for audiological clinics for improved fitting of hearing aids. The current article presents the first results of the spatial audio simulation framework compared to a reference scenario and other existent solutions in literature. First findings demonstrate that synthesized impulse responses with arbitrary source directivity combined with using hearing aid head related transfer functions, with spatial upsampling and Ambisonic domain optimizations, to generate simulated binaural audio can be a powerful tool for generating several real-life situations for further hearing aids research.

1 Introduction

In recent years, Virtual Sound Environments(VSEs) became important to research groups in the field of signal processing for hearing aids (HA) [1, 2, 3] due the discrepancy between the laboratory environment and real-life. Such differences make it difficult to assess the real life performance of developed algorithms, as well as reduce the perceived benefits of such algorithms due to the lack of proper fitting sessions involving more realistic environments and elongate the adaptation period for the end user. In wake of such discrepancy, different methods are applied to alleviate the aforementioned problems. Using a combination of ray-tracing and Ambisonics [4] or image source methods [5], acoustics of different spaces are simulated and later on presented to listeners over multi-channel loudspeaker systems or headphones. The advantages and shortcomings of these efforts are well documented. One prominent problem to these solutions is the compromise between flexibility and realism. Obtaining realism through simulations sometimes requires a tedious setup that might consume a similar time frame with executing a binaural reference measurement, whilst flexibility and ease of setup might

reduce realism so that the benefits obtained using any algorithm in such studies might not apply to real-life situations. Based on the findings by [6], the authors propose a flexible system that is easy to setup up whilst adhering to realism in acoustical scenes. In addition, the authors would like to include new functionalities that can improve the following shortcomings: Ray-tracing applications require a lot attention for correct acoustics and lack the possibility of providing any movement capabilities for the sources. Solutions that depend on image source methods, such as TASCAR [5] implements a flexible solution with an excellent interface to interact with environments. However there are still certain features missing, such as arbitrary higher order Ambisonics support and arbitrary directivity patterns for both source and receiver. In addition to applications in hearing science, our application also facilitates the construction of a large number of variations of different stimuli for the purpose of artificial intelligence based algorithm training. Section 2 describes how the simulations are constructed. Measurements done to validate the simulations are presented in Section 3. The results and the implications of such simulations are discussed on Section 4 and 5. Section 6 concludes the work and reflects upon next steps.

2 Methods

The proposed method for this study requires the construction of an impulse response(IR) of a previously known environment as realistic as possible for validation. Even-though certain ray-tracing applications do an excellent job for such a task, they lack the flexibility and ease of setup. Our application requires to be able to generate many situations in the same or in several new environments. Image source method(ISM), as applied by applications such as TASCAR or others, provide an easy to setup and efficient alternative. In this study, the authors have taken the shoe-box room simulator developed by [7] and added certain modifications to successfully represent any environment with an arbitrary number of sources and receivers. The set of scripts given in [7], is capable of representing different room sizes with an arbitrary number of sources and receivers as well as setting up the RT60 values for different octave bands and even calculating the air absorption. Our modifications mostly concern source and receiver directivity patterns, with the scripts now being able to support arbitrary source directivity patterns from polar maps and simulated patterns for the receivers. The authors have also included slight modifications to deploy third octave band filters for better frequency resolution in reverberation time calculation and more accurate representations of the IRs. All of this information can be either encoded into a set of Ambisonic IRs or HRTF receivers directly. Since the scripts allow an arbitrary order of Ambisonics, it is possible to create acoustically faithful IRs for an extended range of frequencies. Since the shoe-box room simulator has been developed for several years and is free to use under the BSD-3 license, this article will focus more on the modifications done to the scripts rather than the toolbox itself. We expect to publish the code at the end of the year with all the aforementioned modifications in a separate repository.

One of the main problems of Ambisonic signals for representing different environments is that encoding a point source assumes a plane wave, impinging to the center of the sphere where the receiver is located. Neither the directivity of the source nor the orientation is regarded while encoding the signal. Since the IR of any scenario requires the existence of reflections with their corresponding amplitudes, the directivity and the orientation of the source becomes an important factor. The shoe-box room simulator successfully calculates the arrival time, direction and amplitude of each reflection in a three dimensional, limited space and is capable of converting this information into Ambisonic IRs, encoding each reflection due to ISM as a separate source with a different time of arrival and amplitude. Since each reflection has an angle of arrival depending on the source position, orientation and directivity, it is important to calculate the angle of departure from the source so that the directivity pattern and orientation factors can be deployed. This process can be quickly done by interchanging the position of the receiver and the source, and re-calculating the angle of arrival for the new receiver position which now will correspond to angle of departure of the source in the original position due to the reciprocity in acoustics of both situations. Once the angle of departures are calculated, an arbitrary directivity pattern as well as the change in amplitude for each frequency band due to the orientation of the loudspeaker can be applied. Since all this information can be represented as a spherical surface function of gain multipliers around the source, the resulting reflections in the original receiver position with new gains (the time and the directions of arrival are not affected by this process) can be calculated by an additional

multiplier factor for each band in question. Luckily, such information is also publicly available for several loudspeakers under the open Common Loudspeaker Format(CLF). In addition the CLF group recently published the Common Instrument Format(CIF), which ensures the availability of such data for instruments and voices. Thus, it is possible to construct scenarios with an arbitrary number of sources with an arbitrary number of receivers, both with their corresponding directivity pattern. The resulting echograms (the set of reflections with corresponding time and direction of arrivals, and the gain for each band resulting from different absorption and directivity factors) can be consolidated to IRs of arbitrary order in Ambisonics, which can be decoded for binaural or loudspeaker array reproduction. It is important at this stage to assess the performance of the generated IRs compared to a reference situation or previously tested methods involving Ambisonics recordings. The four situations provided as a solution to execute measurements can be seen in Figure 1. The reference situation consists of setting up the scenario that is to be recorded with loudspeakers and head and torso simulator(HATS) with or without HA in place in a controlled environment. The recording and playback method requires the same setup as reference situation but the HATS replaced with an Ambisonic microphone and the resulting audio is played back through a calibrated multichannel loudspeaker setup for spatial audio to be recorded again with the HATS. The advantages and short-comings of the recording and playback method is well documented within [4]. The synthesized and simulated setup requires for generating the IR of the reference situation in a computer program and convolving it with the desired audio, and either playing back through the multichannel loudspeaker setup (Synthesized) or convolving with the HRTFs of the HATS with or without HA (Simulated). Note that with available HRTFs, this latter method requires no laboratory setup.

3 Measurements

To test the accuracy of the generated IRs, a reference situation must be taken into account. The comparison must measure certain parameters, such as change in Signal to Noise Ratio(SNR), SNR Gain for different settings on HA devices and deviations in time of arrival to assure that the synthesized IRs can be an alternative to real-life situations for HA research. In other studies, the comparison between a reference situation recorded by a HATS and the representation of same scene by ray-tracing apps and Ambisonics recordings demonstrate the advantages and short-comings of those solutions [2]. With this information in mind, the reference situation that is represented in Figure 2 is recorded by a Neumann KU100 (Listener), while using Genelec 8020s for playback, due to their size and availability of their directivity response in a CLF file format.

The listener was either used as is or equipped with two HA devices provided by Amplifon Inc. The HA devices in question were set-up to either have omnidirectional (OMNI), cardioid (CARD) or adaptive beamforming (ADM) settings with either an obstructed ear canal (CLOSED) or a semi-open fitting (OPEN) and all other processing settings turned off. The disturbances are chosen to be located in three different positions compared to the listener (the loudspeaker pairs always facing each other) to test both HA devices in different scenarios (ipsi-lateral or contra-lateral disturbance, opposite side or similar direction disturbance) and the disturbances are aligned with loudspeaker positions of the outer multichannel loudspeaker setup to test the possibility of using the closest loudspeaker as a representation as it is done in other studies.

A series of situations are recorded for signal to noise ratio comparisons: only the target(TGT), only the each pair of disturbance(DTB1, DTB2, DTB3), all of the disturbances together(NOS) and all of the loudspeakers together(ALL). As recording material, a white noise signal is used (WBD) as well as speech signal(SPCH) proposed as ISTS[8] as a band limited excitation signal. These scenarios allow the measurements to be analyzed to calculate the SNR ratio in different scenarios, as well as SNR Gain for different settings of HA devices for different listening material. As a last measurement, the IR(SWP) from each loudspeaker is measured using the sweep-method [9] to extract the ITDs for different scenarios. Since Ambisonic recordings are presented as an alternative in [2, 3] the same setup was recorded also using an Eigenmike(SCN2_REC) to test the viability of using multi-channel recordings and playback through loudspeaker arrays as an intermediate solution. As advised by [3], a matrix of shape-matching filters are used. Adding the modifications proposed by [10], a fast frequency dependent deconvolution of the 32 microphone signals to 25 loudspeakers can be achieved. The output is recorded again(SCN2_PLA) using the Neumann KU100 dummy head microphone,

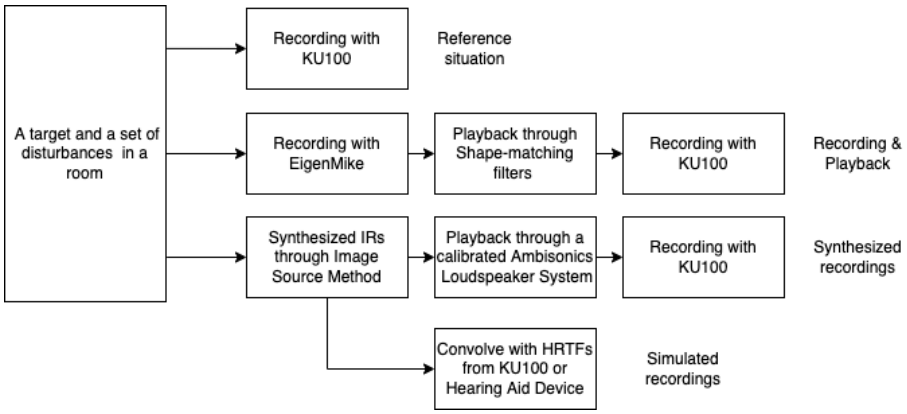


Fig. 1: Different propositions to generate VSEs. Reference situation refers to recording the auditory scenario directly with a HATS while other methods use a combination of Ambisonic microphones, HATS and premeasured HRTFs.

with and without the HA devices, applying the same settings in each case. For synthesized recordings, all the possible signal excitations(WBD_TGT, SPCH_TGT, WBD_DTB1,....,SWP) are generated in the shoe box room simulator with corresponding positions and orientations. These are then reproduced through the 25 loudspeaker setup and to be recorded with the KU100 and the HA devices again(SCN3_SYN). Our proposition, which requires the measured HRTFs of either the KU100 or a HA device, convolves the generated IRs with the HRTFs to generate a simulated version of the audio recordings(SCN3_SIM) which is tested against the other scenarios.

Eventhough it is possible to have arbitrary RT60 values for comparison purposes, a measurement was done to correctly consider the third-octave band reverberation times for the synthesized IRs. Once the room size and the RT60 values are defined, the reflections for each frequency band are calculated within the reverberation time of that band with corresponding time, and direction of arrival as well as amplitude. With the angle of departure from each source, the gains of the echograms are modified accordingly with absorption values corresponding to the RT60 and source directivity patterns. Once the echograms are converted to 10^{th} order Ambisonic IRs, the signals are then processed with a double-band ALLRAD decoder [11] to ensure amplitude preservation for low frequencies while preserving the energy for high frequencies (maxRE) for SCN3_SYM. At this order of Ambisonics and decoding scheme, correct construction of the acoustical field up-to 6 kHz can be ensured. In addition, a BiMagLS decoder [12] using Bilateral Ambisonics where a separate Ambisonic signal is generated per ear for the SCN3_SIM recordings. These signals are convolved with the spherical harmonic(SH) equivalent of the HRTFs measured on a fifty point Lebedev grid with magnitude least square error minimization, diffuse field EQ and higher order tapering. This decoder is chosen as the minimum order Ambisonics decoder to still maintain a sweet spot big enough to engulf an average human head at its center (18cm in our case). Prior to the SH optimization, the HRTFs are spatially up-sampled using SUpDEq [13] to deploy HOA successfully at such order. One should note that for the second scenario SCN2 where the reference situation is recorded with 32 capsules and played back with 25 loudspeakers this limit stands at around 3 kHz [3]. Since there is no Ambisonics involved in the shape-matching method, a psycho-acoustical extension of this scenario would require more studies and processing, which is out of the scope of this paper. Once the Ambisonics signal of SCN3_SYN is prepared, they are played back by 25 equalized and gain-delay calibrated loudspeakers.

All the measurements are done in a room with a size of 7.06m by 5.13m by 3.16m. The $RT60$ values ranged from 0.398 seconds at 125 Hz to 0.253 seconds at 8 kHz. The $RT60_{500Hz}$ and $RT60_{1000Hz}$ average was 0.293s. The KU100 is positioned at the center of the loudspeaker array at a height of 123 cm with a head diameter of approximately 18cm for recording SCN1, 2 and 3. The Eigenmike that is used for the SCN2 is

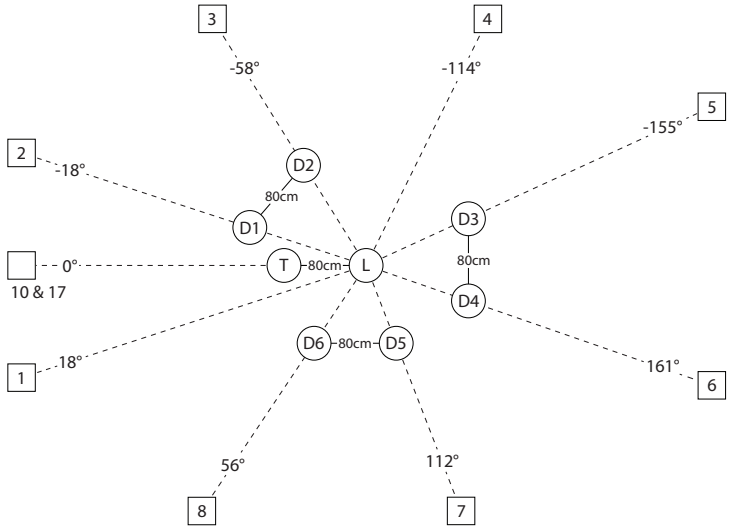


Fig. 2: Reference scenario. L represents the listener and T represents the listening target. In addition, three pairs of disturbances (D1-D6) are positioned around the supposed conversation. The positions of the disturbances are chosen deliberately to test certain HA characteristics. The numbered boxes along the perimeter represent the loudspeakers in the multiple loudspeaker setup, where certain attention is paid to align the disturbances with the loudspeakers of the calibrated multichannel spatial audio setup.

positioned at a height of 127 cm with a diameter of 8 centimeters. The target loudspeaker is positioned at 119 cm at the bottom of its tweeter to simulate a person sitting down, while all disturbance loudspeakers are positioned at 155 cm to simulate standing up people. All pairs are kept at 80 cm away from each other to represent an average conversation distance. The calibration of the microphones is done in the following fashion:

- First a measurement microphone is calibrated with an acoustic calibrator to 94 dBSPL at 1 kHz.
- Then the measurement microphone is placed next to the left ear canal and a test signal of 1kHz is emitted to measure the level at both the measurement microphone and the ear to match the levels.
- The same process is repeated for the right ear with the calibration microphone.
- Finally both ears are measured with signals of level -3 dBFS, -9 dBFS and -15 dBFS to check for levels and linearity on both ears.
- The Eigenmike is also checked against the measurement microphone for ensuring similar SPL levels for both recordings

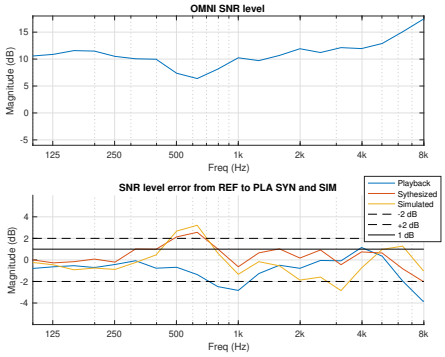
It should be noted that for the SCN2 and SCN3_PLA playback the first two loudspeakers of the loudspeaker array are used instead of the target loudspeaker due to the lack of the latter. The 25 loudspeaker array for Ambisonics playback has been calibrated with the same measurement microphone in the center to adjust the delays, gains and equalization.

The measured speech signal levels correspond to different gains adjusted to have approximately 6 dB of SNR from the NOS to ALL situation in different scenarios whilst they are kept at similar level and adhere to nominal speech levels of between 50 dBSPL to 65 dBSPL.

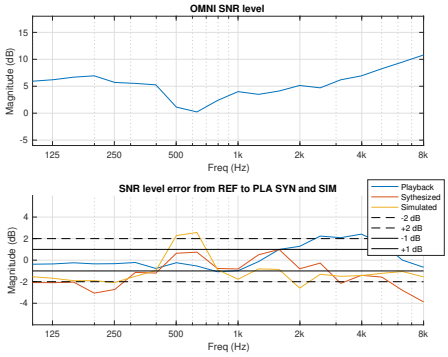
4 Results

The SNR levels in dBs and the difference in SNR from CARD to OMNI setting are presented in Table 1. While simulated IRs demonstrated the closest overall results to the reference situation, the playback and synthesized method deviated further while calculating the SNR gain. The amount of SNR Gain from the CARD to OMNI setting for third octave bands is presented in Figure 3.

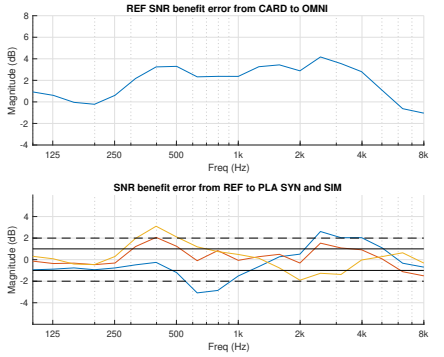
As it can be observed from the plots, both scenarios have preserved comparable SNR values to the reference situation with the error margin not superseding



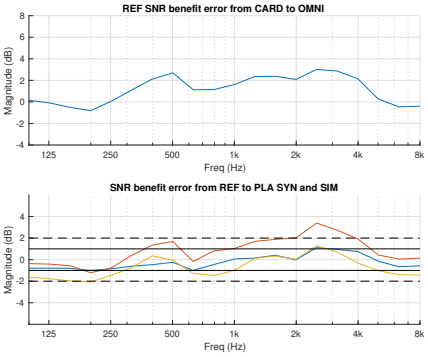
(a) SNR levels for OMNI setting in the DTB2 situation (upper) and the SNR error for PLA, SYN and SIM compared to the reference situation (lower)



(b) SNR levels for the OMNI setting in the NOS situation (upper) and the SNR error as in (a)



(c) SNR benefit for the CARD setting compared to the OMNI in DTB2 situation (upper) and the SNR benefit error as in (a)



(d) SNR benefit for the CARD setting compared to the OMNI in NOS situation (upper) and the SNR benefit error as in (a)

Fig. 3: SNR levels for OMNI setting and benefit from CARD to OMNI for NOS and DTB2 situations and the corresponding SNR errors from the Reference recordings to Playback and Synthesized and Simulated IR recordings. Dashed lines represent the ± 2 dB and the continuous lines represent the ± 1 dB error margin.

	REF	PLA	SYN	SIM
SNR_O (OMNI)	9.79	10.35	9.71	9.32
SNR_C (CARD)	14.05	13.13	12.58	13.54
ΔSNR	4.26	2.78	2.86	4.22

Table 1: SNR gain in dBs from TGT to DTB2 for OMNI and CARD options.

± 2 dB in most of the frequency range. Both the difference in-between NOS and DTB2 as well as OMNI and CARD are coherent with the measured reference situation. The increase in SNR due to the setting change is also apparent in the mid-high frequency region.

5 Discussion

For most of the measurements the SNR Gain error from the PLA, SYN and SIM recordings to REF recordings stayed in the ± 2 dB region between 125 Hz to 6.3 kHz. It should be noted that the frequency limit for the Playback scenario is around 3 kHz for 25 loudspeakers and 32 microphones for a listener area comparable to the size of the human head [3], whilst the same size of listening area can be achieved using a 10^{th} order Ambisonics signal up to 6 kHz. Beyond this limit, the accuracy of the PLA and SYN methods can not be assured. Nonetheless, it should be noted that the useful and functional region of most of the HA devices are

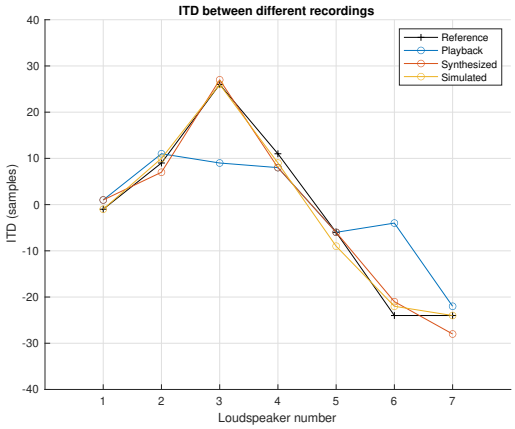


Fig. 4: Calculated ITDs for REF, PLA, SYN and SIM conditions

below 8 kHz due to the fact that the devices are mostly designated for human speech.

One of the important aspects of spatial sound for synthesized or Ambisonic recordings to replace reference situations are the perceived ITDs for different cases. The measured ITDs for different recordings (REF, PLA and SYN and SIM) can be observed in Figure 4. While the SYN and SIM IRs demonstrate a good match for the reference scenario, there is a certain error for the playback and recording setup. This could be due to the size mismatch of the size of the microphone and the area where the ITD measurements are made, especially for higher frequencies, as explained before. It should be noted that a certain sensitivity to the EQ and time alignment measurement done for the Synthesized recording scenario was detected, which requires precaution while utilizing this method.

The preliminary results demonstrate that the synthesized IRs can be a plausible alternative to recording scenes in either HATS or microphone arrays. It is a promising solution where lots of different scenarios can be generated. In fact, in a recent study, using VSEs in similar fashion, the objective performance of five high-end commercially available Hearing Aid (HA) devices were compared to DNN-based speech enhancement algorithms in complex acoustic environments [14]. These tools provide a flexible way to generate IRs with arbitrary number of sources with their own directivity and orientation, within different sized and differently treated rooms and the planned inclusion of predetermined movements to further study spatial and spectral features and performance in complex acoustic scenes.

6 Summary

Synthesized Ambisonic impulse responses are a flexible and realistic alternative to create virtual sound environments (VSEs) that can replace many other stimulus types used for hearing aid research to help close the gap between laboratory and real-life situations. The inclusion of arbitrary source directivity data helps the generated impulse responses provide a more realistic audio when compared to previous image source methods. Our experiments for validation of the proposed method involved comparing a reference situation to three different methods involving VSEs. The results demonstrated that the artificially generated scenarios combined with the measured HA HRTFs, optimized for the Ambisonic domain through addition of certain psychoacoustic optimizations, resulted in an error less than two decibels in several frequency bands in the region of interest and also matched interaural time differences with the reference stimuli for different HA directivity settings. These experiments open the way to possibly use Ambisonic HRTFs both for head and torso simulators in our facilities and several HA devices in order to create a workflow that would require no extra measurement or complicated laboratory setup besides the one-time measurement that is required for the Ambisonic HRTF decoder extraction per HA device. In our experience, the bilateral magnitude least squares method that is used to generate binaural decoders with a measurement of fifty points on a Lebedev grid that is post processed through spatial upsampling and directivity dependent equalization can generate coherent results for any device/setting combination that is suitable for simulating these situations. It should be noted that Ambisonics allows using both point sources and diffuse sounds through sampling points that are used in the decoder and can be combined within the same scenario to improve upon simpler image source method solutions. For research purposes, it is possible to generate a large number of complex scenarios with either solely the listening target and the noisy environment including the same target as problem-solution pairs as labeled data using this method. The vast amount of data generated can be then used to train artificial intelligence (AI) algorithms to help hearing loss patients and

potentially detecting other directivity related hearing loss problems.

HA users can also benefit from these VSEs during user fitting. Thanks to decreasing costs in multi channel installations, VSEs can be both used in laboratory environments and clinics. Use of these complex scenes created in Ambisonics also allows testing these situations both in multi-channel setups or over binaural recordings with few modifications for hearing loss patients to evaluate several devices in a short period of time. As a result, these scenarios could be used for detecting more complex hearing problems or help the user adjust to their device faster in clinics. It is expected that both the AI algorithms trained with spatial audio and listening tests regarding their performance in these generated situations will help study the effects of spatial cues and other directivity related elements on speech comprehension and further investigation of direction related hidden hearing loss symptoms.

Acknowledgements

The research leading to these results have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101017884 – GuestXR Project

References

[1] Lundbeck, M., Grimm, G., Hohmann, V., Bramsløw, L., and Neher, T., "Effects of directional hearing aid settings on different laboratory measures of spatial awareness perception," *Audiology Research*, 8(2), 2018.

[2] Oreinos, C. and Buchholz, J. M., "Evaluation of Loudspeaker-Based Virtual Sound Environments for Testing Directional Hearing Aids," *Journal of the American Academy of Audiology*, 27(07), pp. 541–556, 2016.

[3] Minnaar, P., "Reproducing real-life listening situation in the laboratory for testing hearing aids," *Journal of Audio Engineering Society*, p. 10, 2013.

[4] Oreinos, C. and Buchholz, J. M., "Objective analysis of ambisonics for hearing aid applications: Effect of listener's head, room reverberation, and directional microphones," *The Journal of the Acoustical Society of America*, 137(6), pp. 3447–3465, 2015.

[5] Grimm, G., Luberadzka, J., Herzke, T., and Hohmann, V., "Toolbox for acoustic scene creation and rendering (TASCAR): Render methods and research applications," in *Linux Audio Conference*, p. 8, 2015.

[6] Oreinos, C. and Buchholz, J., "Validation of realistic acoustic environments for listening tests using directional hearing aids," in *14th IWAENC*, pp. 188–192, IEEE, 2014.

[7] Politis, A., *Microphone array processing for parametric spatial audio techniques*, G5 artikkeliväitöskirja, Aalto University, 2016.

[8] Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B., "Development and analysis of an International Speech Test Signal (ISTS)," *International journal of audiology*, 49, pp. 891–903, 2010.

[9] Farina, A., "Advancements in Impulse Response Measurements by Sine Sweeps," *Journal of The Audio Engineering Society*, 2007.

[10] Kirkeby, O., Nelson, P., Hamada, H., and Orduna-Bustamante, F., "Fast deconvolution of multichannel systems using regularization," *IEEE Transactions on Speech and Audio Processing*, 6(2), pp. 189–194, 1998.

[11] Zotter, F. and Frank, M., *Ambisonics*, Springer, 2019, ISBN 7101112131415.

[12] Engel, I., Goodman, D., and Picinali, L., "Improving Binaural Rendering with Bilateral Ambisonics and MagLS," in *DAGA 2021 Proceedings*, 2021.

[13] Pörschmann, C., Arend, J. M., and Brinkmann, F., "Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(6), pp. 1060–1071, 2019.

[14] Gusó, E., Luberadzka, J., Baig, M., Sayin, U., and Serra, X., "An Objective Evaluation of Hearing AIDS and DNN-Based Binaural Speech Enhancement in Complex Acoustic Scenes," in *IEEE WASPAA*, pp. 1–5, 2023.