PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Auralization of interactive virtual scenes containing numerous sound sources

Lukas ASPÖCK, Lucas MÖSCH, Jonas STIENEN, Michael VORLÄNDER

[1]Institute of Technical Acoustics, RWTH Aachen University, Germany, las@akustik.rwth-aachen.de

Abstract

Typical everyday situations usually contain a large number of sound sources. In virtual reality applications, where the processing demands for the acoustical rendering of a scene should be kept a low level, it is challenging to simulate and spatialize a high number of virtual sound sources. This work presents different solutions for rendering sound sources in virtual scenes with varying level of interactivity and complexity. For a binaural free-field auralization of up to hundreds of virtual sound sources, a model based on k-means clustering was recently developed, with the main objective to limit the number of required convolutions. To improve the perceptual quality of the rendering, the model was extended with an efficient correction of the interaural time difference of each virtual sound source. In addition to a brief benchmark analysis of the rendering module, this work also describes how this clustering approach was integrated in the open source auralization framework *Virtual Acoustics*.

Keywords: Auralization, Binaural simulation, Virtual acoustics

1 INTRODUCTION

Due to advances in acoustical research, signal processing and computational power, the concept and implementation of an *Auralization* [12] has been established in the acoustical research community. By simulating and auralizing a virtual sound field, controlled situations can be created for various purposes. With respect to the application of auralizations in psychoacoustical experiments, the full auralization process, involving modeling input data, simulation, signal processing and reproduction of the sound field, should ideally be indistinguishable from the corresponding real acoustic event, i.e., a virtual guitar player in a bar should be perceived acoustically in the same way as a real guitar played in the same environment.

In general, the auralization process includes the convolution of an anechoic sample with a room impulse response (RIR), describing the acoustical transfer path from the sound source to the receiver in an environment [34]. With the intention of auralization, the spatial human perception is typically modelled by using a binaural receiver, creating a binaural room impulse response (BRIR) as a simulation output. Different studies [21, 23] have shown that for simulation models based on geometrical acoustics (GA), the simulated results barely differ from measured results if the user is informed about the measurement results and the simulation results are adjusted to match the measured results using a calibration procedure. Postma showed that a methodical calibration based on parameters of RIR measurements makes it possible to *produce perceptually equivalent spatial BRIRs when compared to measured BRIRs*. The most recent round robin study on room acoustical simulation and auralization revealed [4], however, that many GA based simulations often fail to match measurements when the user is not informed about the results, especially when wave effects such as diffraction and scattering are of relevance. Informed simulations using the boundary element method [7] showed good agreement for simple scenarios, but less accurate results for a more complex room scenario, for which no reliable boundary condition data could be provided. Nevertheless, it can be stated that binaural synthesis and acoustical simulation can generate plausible, and after a thorough calibration process, also authentic acoustical signals, as demonstrated for binaural synthesis by Oberem et al. [19], for a dynamic binaural synthesis by Brinkmann et al. [5] and for BRIR simulations by Postma and Katz [22] as well as by Tommasini at al. [30]. Although these results are promising with respect to the auralization of defined scenarios, the applied methods face various challenges and limitations when *dynamic* binaural auralization of *complex* scenes should be created. The term *dynamic* corresponds to, at least, rotational movement of the receiver, which creates interactivity, while the term *complex*

indicates that three or more sound sources of the scene are simultaneously sending audible signals. Driven by the success of and tremendous progress in the gaming and Virtual Reality industry, various audio rendering tools and engines (e.g., *SteamAudio*, *Resonance Audio* or *DearVR*) have recently emerged which mostly are capable of rendering binaural feedback for dynamic complex scenes in real-time. Often even more simplified concepts than GA simulations are applied, which, in some cases, can create plausible situations (especially in combination with a photo-realistic visual feedback), but are far from being authentic, for example in case of typical everyday situations such as a noisy restaurant or a busy street in an urban area. First studies have only compared rendering engines to others [11, 2], but not to real or measured acoustic scenes. More examples for (binaural) real-time simulation tools, developed within research projects, are the open-source projects *RAZR* [38] or *3D Tune-In* [6].

When designing an experiment based on auralizations, the researcher has to define the desired complexity of the scene as well as the level of interactivity, depending on the chosen research questions and research focus before selecting the rendering technique for auralization. If no interactivity with respect to the receiver is necessary, a binaural simulation stream can fully be preprocessed completely avoiding real-time calculations (except for audio playback) during an experiment or demonstration.

When it comes to rendering scenes with a high number of sound sources, Tsingos et al. [32, 33, 31] published important pioneer work more than 15 years ago by efficiently rendering complex scenes with more than 64 sound sources. A key component of these implementations was a clustering technique, which was adapted and implemented by Hell et al. [8], who also showed that the workload of a binaural room simulation can be reduced without creating substantial perceptual differences.

In the first part of this work, a general software framework for the real-time synthesis of binaural signals is introduced, while in the second part an approach of more efficient binaural rendering for a high number of sound sources in a free-field condition is presented.

2 BINAURAL RENDERING OF VIRTUAL SOUND SOURCES

Auralizations of a virtual sound sources are usually implemented using a convolution of an anechoic sound signal $x(t)$ with a BRIR, see. Fig. 1. When GA models are considered, it is reasonable to split the BRIR into three parts: I - the direct path which accounts for the attenuation law of the point source, the air absorption and the corresponding head-related-transfer function (HRTF); II - the early reflections, consisting of a few, mostly specular reflections, typically modelled by an image source model [1]; III - a dense pattern of numerous, overlapping late reflections, the reverberation tail, which can be modelled by feedback delay networks [10] or by ray tracing models [13].

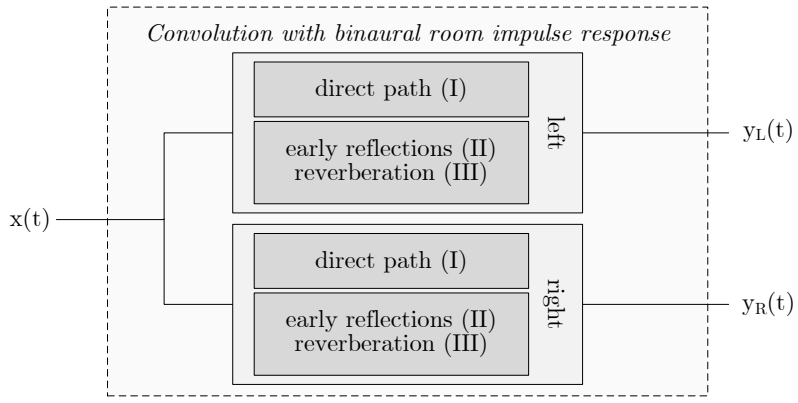


Figure 1. Concept for the binaural rendering of a virtual sound source with the anechoic source signal $x(t)$

In all three parts of the BRIR, it should be accounted for a sound source directivity as well as for a receiver directivity (HRTF). As the relevance of this data decreases for the later part of RIRs, the spatial resolution of this may be lower for the reverberation tail than for the direct sound. Extended models can additionally include further effects such as doppler shifts [37]. Depending on the virtual scene and the application, it is convenient to implement the rendering of each part separately and to select different update rates for each part of the BRIR [3, 36]. This ensures real-time updates also for more complex scenes.

2.1 Rendering modules in Virtual Acoustics

Virtual Acoustics (VA) is a real-time auralization framework [9, 29] mainly intended for scientific research. A modular concept enables the user to select from a different range of rendering and reproduction modules, e.g., a binaural renderer in combination with a crosstalk cancellation [15] reproduction module. A combination of multiple rendering modules is also possible: In addition to a real-time binaural free-field rendering of a sound source, the *GenericPath* renderer can be selected, which convolves the source signal with a filter defined by the user, e.g., a precalculated BRIR without the direct sound. Other rendering modules are capable of calculating image sources for a shoebox room or rendering a very high number of sound sources, cf. Section 3.2.

2.2 Simulation of binaural room impulse responses

RAVEN[1] is a room acoustic simulation C++ library [26, 25] based on GA models which can be applied in real-time environments [14, 3], but is mostly used for scientific purposes using a command line interface from MATLAB. In its main application, the user defines a scene to generate RIRs or BRIRs. It is also possible to define trajectories generating scenes with moving sound sources and/or the receiver. Script based processing allows the generation of sets of BRIRs, e.g., for different receiver positions or orientations. Such datasets of BRIRs can be included in a directional database in the openDAFF file format [35] and processed by the VA software, implementing a dynamic binaural synthesis.

3 EFFICIENT REAL-TIME AURALIZATION OF NUMEROUS SOUND SOURCES

Binaural rendering of many sound sources is today mostly based on a scene description in the spherical harmonics domain, either based on a virtual loudspeaker array [18, 16] or directly in spherical harmonics [24]. The renderer presented in this section uses an alternative approach, based on clustering, to spatially render more than 100 virtual sound sources simultaneously using binaural synthesis.

3.1 Concept

In contrast to a conventional binaural synthesis, which requires one HRTF convolution per sound source, this renderer applies a clustering of sound sources leading to a reduction of HRTF convolutions. For each cluster, the sound source signals are summed and then convolved with the representative HRTF of the cluster. The audibility of the angular shift caused by the clustering should be kept minimal. Another important aspect of the implementation is the efficiency, i.e., the processing time of the entire rendering based on the clustering must be substantially lower than the individual binaural rendering.

The processing of the renderer is depicted in Fig. 2. The current state of the virtual scene is analyzed and clusters based on a selected metric are created. All related individual sound sources of each cluster are summed up as two-channel source signals including the distance attenuation, the propagation delay and their corrected time delay. The corresponding cluster HRTF is then convolved with the summed cluster signal. Finally, all binaural cluster signals are summed and binaurally reproduced.

[1] RAVEN is not published open source, but it is freely available for academic purposes. Please contact the authors for further information

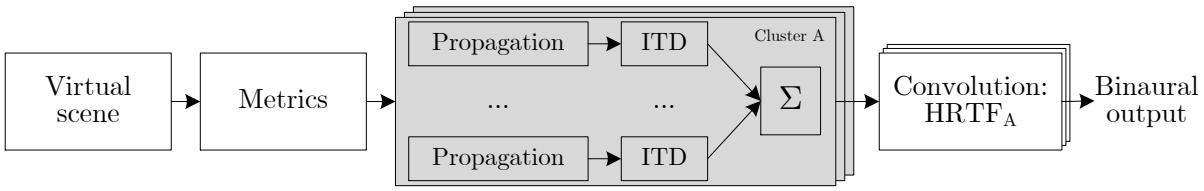


Figure 2. Concept of the clustering engine. Each sound source is individually processed based on its propagation path (attenuation and delay) and its interaural time delay (ITD). Spectral characteristics are applied by convolution with the corresponding Head-Related Transfer Function (HRTF).

3.1.1 Clustering

When sound sources are clustered based on the described concept, a spectral error is introduced. With the goal of minimizing this error, a suitable clustering approach needs to be selected. As the feature for the algorithm the spherical angle between a reference point and the sound source positions was chosen. The algorithm itself is adapted from the *k-means* clustering algorithm [20]. Instead of using a predefined number of clusters, new clusters are dynamically created when the angular distance of a sound sources exceeds a threshold distance to all existing clusters, unless the maximum number k of clusters already exists.

3.1.2 ITD correction

As a time shift of a filter is an inexpensive operation in the real-time processing chain, the interaural time difference (ITD) of all virtual sound sources can be processed individually despite the clustering. The temporal shift caused by the clustering is corrected using the difference of the ITDs of the actual and the clustered sound source position. The ITDs can either be estimated based on a rigid sphere model [40] or extracted from the HRTF data, using the maximum value in the time domain [17] or the maximum value of the correlation HRIR and its corresponding minimum-phase IR [39].

3.2 Implementation

The clustering renderer was named *VABinauralClusteringRenderer* and integrated in the open source software project *Virtual Acoustics* (VA) software [9]. VA calculates audio buffers in real time for all activated rendering modules, typically with buffer sizes of 128 or 256 samples using ASIO drivers. Propagation delays of sound sources are efficiently processed using variable delay lines (VDLs) [27].

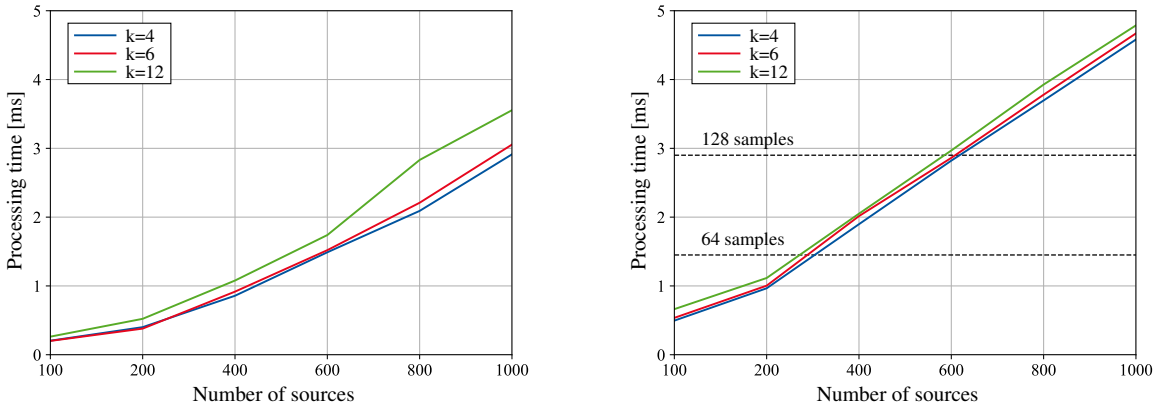
The implementation of the renderer is mainly derived from the default binaural free-field renderer of VA, but was extended by the clustering engine, which implements the concept described in Section 3.1.1. To avoid audio dropouts, the cluster engine runs as part of the scene update in a parallel thread and only triggers an output audio stream update, if the new clustering state was successfully generated for the scene. The relevant data for the ITD correction is provided by a separate estimator class or in case the ITD correction being based on the HRTF dataset, it is calculated during the initialization of the renderer.

3.3 Evaluation

As the system latency is one most critical characteristic of a real-time rendering engine, a computational benchmark (see Section 3.3.1) is an important part of its evaluation to answer questions with respect to the general real-time capability and limitations of input parameters such as the number of sound sources or the dependency from the number of clusters. The level of simplification, in case of the chosen renderer, can be described by a spectral error or, indirectly, by the angular error introduced by clustering the sound sources (see Section 3.3.2). Eventually this error has to be investigated in perception experiments, which so far, has only been done informally in listening sessions of the authors. These tests, however, have shown that very low maximum cluster sizes $k = \{4, 6, 12\}$ are a reasonable choice for a total sound source count of 100 and higher.

3.3.1 Benchmark

To benchmark the implemented renderer, the processing time of the scene- and stream update was measured for an HRTF length of 128 samples at 44100 Hz sampling rate. The virtual scene included a varying number of sound sources that were moved between every rendering step. Each sound source had a random audio signal assigned for audio playback. The benchmark was conducted on a Intel Core i7-2600K CPU @ 3.40GHz processor, 32GB Dual Channel @ 2400 MHz DDR3 memory, running Windows 10 Pro Version 10.0.17763. Benchmark results for the scene- and the stream update are shown in Figure 3a and Figure 3b, respectively.



(a) Scene update

(b) Stream update

Figure 3. Processing times for the scene- and stream update for different number of sound sources using clusters with $k = 4, 6, 12$. Dashed lines in (b) correspond to the update limit of the stream update for buffer sizes $s_b = \{64, 128\}$ @ 44100 Hz.

The processing times of both updates increase for an increasing number of sound sources. If a higher maximum number of clusters is chosen, higher processing time are observed. There is, however, no substantial increase of the processing time. For a sampling rate of 44100 Hz, the clustering rendering engine is able to process up to 240 sources for a buffer size of 64, and up to 600 sources for a buffer size of 128.

3.3.2 Angular error and perceptual analysis

To quantify the level of spatial scene simplification, the angular error introduced by the clustering can be investigated. The angular error corresponds to the spherical angle between the actual source position and its cluster representative. In this evaluation, up to 1000 sound sources were created step by step at uniform random incidence angles at 3 m distance to the receiver, repeated 100 times each for $k = \{4, 6, 12\}$. The evaluation of each step showed that the average angular errors converges for more than 200 sound sources to 36° for $k = 4$, to 29° for $k = 6$, and to 21° for $k = 12$. These values show that the average angular error decreases for higher cluster counts, but indicate only the average spatial deviation. Due to the implemented ITD correction, the angular shift only leads to a spectral deviation of the clustered sound sources and is likely to be less relevant than an actual spatial offset of a clustered sound source.

To check this hypothesis, a perceptual experiment is required. So far, only a preliminary informal listening experiment by the authors has been conducted. In this test, binaural samples including 100 common sound sources were created using different techniques: (I) full binaural synthesis, (IIa) clustering without ITD correction and (IIb) clustering with ITD correction. Conditions (IIa) and (IIb) were generated for maximum cluster counts $k = 6$ and $k = 12$. Critical listening to these samples revealed slight audible differences between (I) and

(IIa/b) but, surprisingly, no audible differences between the different number of clusters $k = 6$ and $k = 12$. Such results, however, are highly dependent on the scene design. In general, it is very challenging to create suitable and plausible virtual scenes with more than 100 mostly continuously active sound sources.

4 SUMMARY

In the first part, this work reviewed and presented the general concept and a software environment for binaural simulation and rendering of virtual sound sources, based on filter convolution with HRIRs or BRIRs. In the second part, a special rendering engine to efficiently process more than hundred sound sources in a free-field environment was described. It was shown that this efficient binaural renderer, in contrast to a normal binaural free-field rendering engine, is able to simultaneously process hundreds of sound sources for typical buffer sizes and sampling rates. The increased efficiency is achieved by reducing the number of convolutions with head-related transfer functions. The HRTF spectrum of each virtual sound source is replaced by the HRTF spectrum of a cluster representative. In how far this introduction of erroneous coloration is perceptually acceptable needs to be assessed in future studies.

Although it is uncommon that acoustical scenes applied in research contain that many sound sources all playing signals relevant and audible signals simultaneously, such a rendering concept is useful for virtual environments, which contain a high number of automatically generated individual objects, especially when these objects create sound based on a real-time synthesis, e.g., rain drops falling on the ground, or numerous vehicles in an urban environment.

The presented rendering engine is only able to generate the binaural free-field response and lacks the calculation of reflections in outdoor or indoor environments. The flexibility of the software framework in which the rendering engine was implemented, however, offers possibilities to also include binaural room impulse responses for the virtual sound sources. The concept of the clustering engine could also be adapted to room acoustics, as presented by Hell at al. [8]. Another direct application scenario of the engine is the simulation of outdoor noise [28, 27]. Here, the majority of sound sources processed by the renderer represent secondary sound sources which model the reflection or diffraction paths instead of the direct path.

REFERENCES

[1] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.*, 65(4):943–950, Apr. 1979.

[2] S. V. Amengual Garí, C. Schissler, R. Mehra, S. Featherly, and P. Robinson. Evaluation of real-time sound propagation engines in a virtual reality framework. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Mar 2019.

[3] L. Aspöck, S. Pelzer, F. Wefers, and M. Vorländer. A real-time auralization plugin for architectural design and education. In *Proc. of the EAA Joint Symposium on Auralization and Ambisonics*, pages 156–161, Berlin, Germany, Apr. 2014.

[4] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl. A round robin on room acoustical simulation and auralization. *The Journal of the Acoustical Society of America*, 145(4):2746–2760, 2019.

[5] F. Brinkmann, A. Lindau, and S. Weinzierl. On the authenticity of individual dynamic binaural synthesis. *J. Acoust. Soc. Am.*, 142(4):1784–1795, Oct. 2017.

[6] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, and A. Reyes-Lecuona. 3D Tune-In Toolkit: An open-source library for real-time binaural spatialisation. *PLOS ONE*, 14(3):1–37, 03 2019.

[7] J. A. Hargreaves, L. R. Rendell, and Y. W. Lam. A framework for auralization of boundary element method simulations including source and receiver directivity. *The Journal of the Acoustical Society of America*, 145(4):2625–2637, 2019.

[8] C. P. Hell, L. Aspöck, and M. Vorländer. Strategies for the effiecient aurlaization of complex scenes containing multiple sources. In *Fortschritte der Akustik – DAGA 2016*, pages 836–839, Aachen, Germany, Mar. 2016.

[9] Institute of Technical Acoustics, RWTH Aachen University. Virtual Acoustics (VA) - a real-time auralization framework for scientific research. Software download and source code available on http://www.virtualacoustics.org (last accessed on 2019-05-23).

[10] J.-M. Jot, L. Cerveau, and O. Warusfel. Analysis and synthesis of room reverberation based on a statistical time-frequency model. In *103rd AES Convention*, New York, USA, Sept. 1997.

[11] G. Kamaris, E. Giannatsis, K. Kaleris, and J. Mourjopoulos. Suitability of game engines for virtual acoustic experiments. In *Audio Engineering Society Convention 146*, Mar 2019.

[12] M. Kleiner, B.-I. Dalenbäck, and P. Svensson. Auralization - An overview. *J. Audio Eng. Soc.*, 41(11):861–875, Nov. 1993.

[13] A. Krokstad, S. Strom, and S. Sørsdal. Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration*, 8(1):118 – 125, 1968.

[14] T. Lentz, D. Schröder, M. Vorländer, and I. Assenmacher. Virtual reality system with integrated sound field simulation and reproduction. *EURASIP J. of Advances in Signal Processing*, pages 1–19, 2007.

[15] B. Masiero. *Individualized binaural technology. Measurement, equalization and perceptual evaluation*. Doctoral Thesis, RWTH Aachen, Aachen, Germany, Dec. 2012.

[16] D. Menzies and M. Al-Akaidi. Nearfield binaural synthesis and ambisonics. *The Journal of the Acoustical Society of America*, 121(3):1559–1563, 2007.

[17] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen. Head-related transfer functions of human subjects. *J. Audio Eng. Soc.*, 43(5):300–321, May 1995.

[18] M. Noisternig, T. Musil, A. Sontacchi, and R. Höldrich. A 3d real time rendering engine for binaural sound reproduction. In *Proceedings of the 9th International Conference on Auditory Display (ICAD2003), Boston, MA, July 7-9, 2003. Eds. Eoin Brazil and Barbara Shinn-Cunningham. International Community for Auditory Display.* Georgia Institute of Technology, 2003.

[19] J. Oberem, B. Masiero, and J. Fels. Experiments on authenticity and plausibility of binaural reproduction via headphones employing different recording methods. *Appl. Acoust.*, 114:71–78, Dec. 2016.

[20] M. G. Omran, A. P. Engelbrecht, and A. Salman. An overview of clustering methods. *Intelligent Data Analysis*, 11(6):583–605, 2007.

[21] S. Pelzer, M. Aretz, and M. Vorländer. Quality assessment of room acoustic simulation tools by comparing binaural measurements and simulations in an optimized test scenario. In *Proceedings of Forum Acusticum 2011 : 27 June - 01 July, Aalborg, Denmark*. European Acoustics Association (EAA), 2011.

[22] B. N. J. Postma and B. F. G. Katz. Correction method for averaging slowly time-variant room impulse response measurements. *The Journal of the Acoustical Society of America*, 140(1):EL38–EL43, 2016.

[23] B. N. J. Postma and B. F. G. Katz. Perceptive and objective evaluation of calibrated room acoustic simulation auralizations. *J. Acoust. Soc. Am.*, 140(6):4326–4337, Dec. 2016.

[24] J.-G. Richter, M. Pollow, F. Wefers, and J. Fels. Spherical harmonics based hrtf datasets: Implementation and evaluation for real-time auralization. *Acta Acust. united Ac.*, 100(4):667–675, July/Aug. 2014.

[25] D. Schröder. *Physically based real-time auralization of interactive virtual environments*. PhD thesis, RWTH Aachen, Aachen, Germany, 2011.

[26] D. Schröder and M. Vorländer. Raven: A real-time framework for the auralization of interactive virtual environments. In *Forum Acusticum*, pages 1541–1546. Aalborg Denmark, 2011.

[27] J. Stienen and M. Vorländer. Real-time auralization of propagation paths with reflection, diffraction and the doppler shift. In *Fortschritte der Akustik - DAGA 2018 : 44. Jahrestagung für Akustik, 19.-22. März 2018 in München*, pages 1302–1305, Berlin, Mar 2018. Deutsche Gesellsch. f. Akustik.

[28] J. P. Stienen and M. Vorländer. Geometry-based diffraction auralization for real-time applications in environmental noise. page 16 Folien. 173rd Meeting of the Acoustical Society of America, Boston, MA (USA), 25 Jun 2017 - 29 Jun 2017, Jun 2017.

[29] J. P. Stienen, F. Wefers, and M. Vorländer. The open-source Virtual Acoustics (VA) real-time auralization framework. In *23nd International Congress on Acoustics (ICA 2019) : Aachen, Germany, 9-13 September 2019*, Sep 2019.

[30] F. C. Tommasini, O. A. Ramos, M. X. Hüg, and S. P. Ferreyra. A computational model to implement binaural synthesis in a hard real-time auditory virtual environment. *Acoustics Australia*, 47(1):51–66, Apr 2019.

[31] N. Tsingos. Perceptually-based auralization. In *19th International Congress on Acoustics*, pages –, Madrid, Spain, Sept. 2007.

[32] N. Tsingos, E. Gallo, and G. Drettakis. Breaking the 64 spatialized sources barrier. *Game Audio Resource Guide 2003*, page 8, 2003.

[33] N. Tsingos, E. Gallo, and G. Drettakis. Perceptual audio rendering of complex virtual environments. *ACM Trans. Graph.*, 23(3):249–258, Aug. 2004.

[34] M. Vorländer. *Auralization. Fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer, Berlin, Heidelberg, Germany, 1st edition, 2008.

[35] F. Wefers. Opendaff - a free, open-source software package for directional audio data. In *Fortschritte der Akustik – DAGA 2010*, pages 1059–1060, Berlin, Germany, 2010.

[36] F. Wefers, J. Stienen, S. Pelzer, and M. Vorländer. Interactive acoustic virtual environments using distributed room acoustic simulations. In *Proc. of the EAA Joint Symposium on Auralization and Ambisonics*, pages 48–55, Berlin, Germany, Apr. 2014.

[37] F. Wefers and M. Vorländer. Strategies for the real-time auralization of fast moving sound sources in interactive virtual environments. In *Implementing noise control technology : 44th International Congress and Exposition on Noise Control Engineering (Internoise 2015): San Francisco, California, USA, 9 - 12 August 2015*, Red Hook, NY, Aug 2015. Curran.

[38] T. Wendt, S. van de Par, and S. D. Ewert. A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation. *J. Audio Eng. Soc*, 62(11):748 – 766, Nov. 2014.

[39] F. L. Wighman and D. J. Kistler. Measurement and validation of human hrtfs for use in hearing research. *Acta Acustica united with Acustica*, 91:429–439, 2005.

[40] H. Ziegelwanger and P. Majdak. Modeling the direction-continuous time-of-arrival in head-related transfer functions. *The Journal of the Acoustical Society of America*, 135(3):1278–1293, 2014.