Strategies for the efficient auralization of complex scenes containing multiple sound sources

Christian Philip Hell, Lukas Aspöck, Michael Vorländer
Institute of Technical Acoustics, RWTH Aachen University, Germany, Email: christian.hell@rwth-aachen.de

Introduction

Simulation models for geometrical acoustics are often based on a image source method [1], a ray tracing algorithm or a hybrid model combining these methods in an efficient way [2]. Modern computer architectures made it possible to apply these models in real-time auralization software. The typical auralization process consists of the simulation of the propagation paths, a binaural filter synthesis and a convolution with an anechoic sound source file. While these steps can be processed for one sound source in real-time applications on a standard computer [3], update rates have to be reduced as soon as the scene contains multiple sound sources. For each source, the computational workload increases proportionally. To simplify such scenes by reducing the computational complexity while maintaining the same auditory impression for the listener, a clustering method similar to the concept proposed by Tsingos [4][5] was developed. This method replaces groups of closely spaced sources by a single representative source by taking advantage of spectral masking. Another approach was discussed by Herder [6], who also analyzed the effects of clustering and removing less salient sources (*Culling*). Both of these methods aim at reducing the number of propagation paths which have to be calculated during the update process of the virtual scene. The presented algorithm creates sound source clusters and culls inaudible sources based on geometrical and psychoacoustical properties. To adjust the source count of the virtual scene, the clustering intensity can be varied. This paper first presents the concept of the algorithm and then evaluates the concept based on a performance analysis and a listening test which was conducted for two example scenes.

Concept

In the context of this research, a virtual scene is considered to be a complex scene if it has a sound source count of more than five sound sources. The process of reducing the number of sound source positions or sound propagation paths which have to be calculated by the simulation module is called *scene simplification*. The scene might still contain a higher number of sound sources, but the count has been reduced in comparison to the original scene. The software module for the scene simplification is structured into five parts (Fig. 1. In the first step (*scene analysis*), the current scene is analyzed and relevant data for the following steps is processed. This includes the positions and orientations of sources and receivers as well as the type of scene (outdoor or indoor) and the types of the signals assigned to the sound sources. As additional information for the prioritization, the signals used in each scene were roughly categorized into main (e.g., speech) and background signals (e.g., steady wind noise) beforehand. In the second step of the process (*prioritization of sound sources*), a priority level is calculated for each sound source. This level is mainly based on the position of the sound source relative to the receiver, which is used to calculate the expected localization accuracy for the relative positions. Using these values as well as the sound pressure level (based on distance, sound power level and source directivity), a priority is determined which describes the importance of a sound source regarding the auditory perception of the scene. As soon as all relevant data about the current scene is available, the processing steps three and four, *sound source culling* and *sound source clustering*, reduce the number of sound propagation paths of the virtual scene. Based on the result of the clustering, a new set of source positions is generated, composed of the cluster representatives (step 5) and the remaining sources which were not clustered or culled. The whole process of step 1 to 5 is meant to be executed each time there is a significant scene change, e.g., in case of a source movement of more than 20 cm. As the scene simplification is supposed to be applied in a real-time auralization system, the amount of computational operations should be kept at a minimum level and the software implementation should be realized in an efficient way. Details about the processing steps three, four and five are described in the following sections.

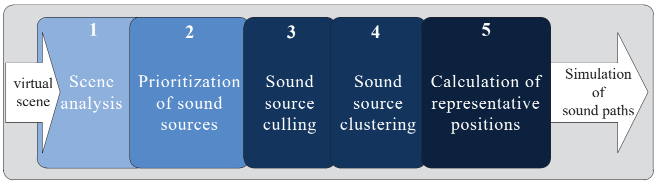
virtual scene
1
Scene analysis
2
Prioritization of sound sources
3
Sound source culling
4
Sound source clustering
5
Calculation of representative positions
Simulation of sound paths

Figure 1: Concept of the scene simplification algorithm showing five processing steps.

Source culling

Prior to the clustering of the sound sources, a simple sound source culling is done by calculating the sound pressure levels at the receiver position based on their distance, the source directivity and their sound power. For indoor scenarios, the user has the possibility to set, if known, an average reverberation time of the room. In this case a sound pressure level of the diffuse sound field can be added to the previously calculated free field level of each sound source. Thresholds for the culling can either be absolute, e.g., the hearing threshold, or relative, e.g., a 50 dB threshold as a maximum level difference between current and loudest sound source of the scene).

Source clustering

The clustering algorithm is based on a method developed by Tsingos called *adaptive positional clustering* [5]. The processing steps one and two provide processed data about the scene, geometrical information and relative positions and orientations as well as a prioritized list of sound sources for the clustering process. Although the implemented concept can account for the sound power level and allows the integration of a signal analysis of the sound source, the presented clustering algorithm in the scope of this work mainly focuses on the geometrical properties of the scene. For this, the key aspects in the clustering process are the human localization accuracy [7] and the positions of the sources relative to the receiver. For both criteria, the angle criteria and the distance criteria, a set of parameters can be varied to control the intensity of the clustering.

The human localization accuracy differs depending on the direction of the incoming sound. For this algorithm, sectors in the azimuth- and elevation angle are defined. The specific localization accuracy for any angle is interpolated between the basic experimental values determined by Blauert (see Fig. 2). To be grouped in a cluster, the difference in the azimuth- and elevation angle of two or more sources must not exceed the respective localization accuracy values (Fig. 3). Furthermore the difference in the distance of these sources to the receiver (*maxDistanceDifference*) has to be below a defined threshold. This value depends a lot on the boundaries of the scene: As long as in a free-field situation the effect of the air attenuation can be neglected (e.g., for distances below 15 m), even sources with a high distance to each other (e.g., 3 m and 10 m) can be assigned to a cluster if the angle criteria is fulfilled. In this case, the consequence of the clustering is a shared directional rendering (application of HRTF filters), the levels of each sources are still treated individually by a simple multiplication according to the distance law. As soon as the simulated sound field is influenced by strong early reflections or room - the parameter *maxDistanceDifference* has to be set to a lower value as the change of the reflections might lead to an audible difference. For outdoor scenes, the distance criterion was treated as less relevant due to the fact that the missing reverberation makes a shift in distance more difficult to perceive if the source level is adjusted accordingly. This however is meant as an experimental assumption and would require a listening experiment for validation. To create a cluster, at least two sound sources have to fulfill both criteria, the angle criterion as well as the distance criterion. By widening or narrowing the angles of the localization accuracies, different sizes of clusters can be created with a variation of the intensity of the source reduction.

Another factor which is relevant for the clustering approach is the calculated priority which was calculated in step two. Sources which are close to the receiver or in a sector with a high localization accuracy get assigned a higher priority than sound sources, which are distant, quiet or which are more difficult to localize. This is done to assure that the least significant sound sources are clustered first. Currently the model does not consider the orientation of the sources. If the sound sources of the scene do not have an omnidirectional character and distribute the energy in a concentrated direction, the orientation has to be included as another criteria for the clustering process. For the first analysis of the clustering concept, all sound sources of the scene were treated as omnidirectional sources.

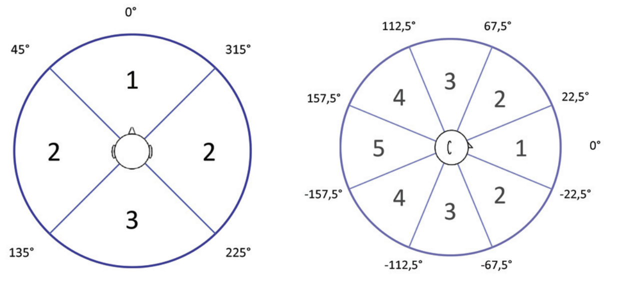

Figure 2: Sectors defined for the azimuth (left) and elevation angle (right) based on the experimental values determined by Blauert.

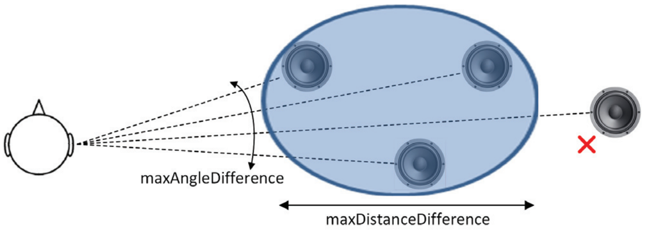

Figure 3: Depiction of the distance and angle criteria. Three sources with similar angles and distances relative to the receiver are within the respective thresholds and will be clustered. The fourth source fails the distance criterion and thus will not be added to the cluster.

Cluster Representative

The representative position of a cluster is determined based on the geometrical positions and priorities of the sources. An average of all source positions is calculated in spherical coordinates, the positions are however also weighted according to the priority of each source. Due to this method, the change in position of the sources with more salient signals will be comparably small. To compensate for the change in the signal levels at the receiver position caused by the relocation of the sources, the level of each source signal is adjusted according to the change in distance to the receiver. The software realization of the scene simplification also contains an option to render the direct sound of all the original sources and only calculate the reflections for the representative positions, which leads to an auralization closer to the original scene. The orientation of the cluster representative is calculated in the same way as the average position. The source directivity of the most salient source of each cluster is also assigned to the representative source. In the scope of this work, source directivities did not impact the results of the simplification model as only omnidirectional sources were used in the example scenarios.

Implementation

The software module was implemented in C++ and used in combination with the room acoustics simulation software RAVEN [7], which applies a hybrid geometrical acoustics simulation model to generate BRIRs. Interfaces for the integration into a real-time auralization system were prepared. Example scenes for the listening test, an indoor restaurant scene and a outdoor scenario containing a bus stop, were created with the 3D modeling software *SketchUp*.

Listening Test

A 3AFC-listening test with 30 participants was conducted to investigate if the impact of the scene manipulation can be identified by the test persons. In this initial experiment, a comparison of three different selected configurations for the clustering approach should indicate suitable parameters for the simplification of the investigated scenes. *Cluster intensity 1* represents a careful clustering (average source count reduction: 31.2 %) while *Cluster intensity 3* uses parameters leading to a strong sound source clustering (average source count reduction: 72.5 %). *Cluster intensity 2* used the same parameter configuration as *Cluster intensity 3*, the cluster representatives were only used for the calculation of the reflections while the direct sound of each sources was renderer according to the original position. Detection rates of *Cluster intensity 3* were hypothesized as being the highest, while *Cluster intensity 1* and *Cluster intensity 2* were expected to result in lower detection rates. The participants had to listen to the original and the three simplified auralizations in a pairwise comparison. The sound samples were reproduced by equalized headphones (*Sennheiser HD-600*) in a hearing booth. For each comparison, test persons were allowed to replay the samples unlimited times. The task was to identify a different sample of two scenes each with two different receiver positions: An indoor restaurant scene (Fig. 5) and outdoor scene (see Fig. 6). Two sequences with a duration of 10 s of both scenes, containing around 10-12 active sound sources playing typical signals according to the situation, were played back to the test persons. No visual feedback was given.

The results show that even in the laboratory situation without visual feedback the differences are difficult to notice: Mean values for correctly identified samples are 51.7% 56.3% and 62.5% for the cluster intensities 1, 2 and 3. An ANOVA ($p < 0.05$) showed no significant differences for the comparison of the cluster intensities (p=0.124, F=2.14). However the boxplot (see Fig. 4) shows a tendency towards higher detection rates for an increased cluster intensity. Test persons reported that the auralizations could be distinguished especially when a scene contained a very salient source (e.g., a barking dog in the bus stop scene). The comparison of the two scenes shows that the audible effect of the clustering in case of the restaurant was harder to detect than in the outdoor scene. The variance analysis showed a significant difference (p=0.007, F=7.82).

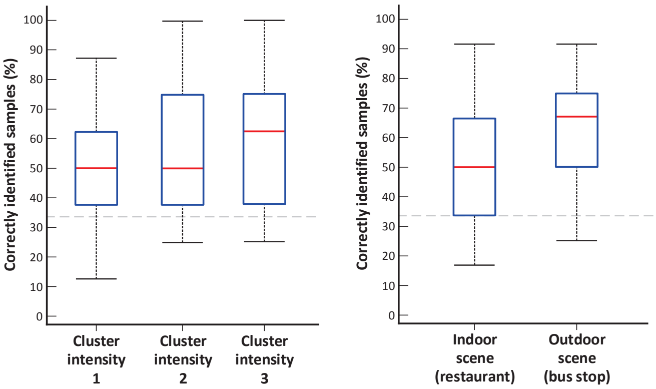

Figure 4: Box plots of the listening test results. Comparison of correctly identified different samples for three clustering intensities (left) and the two example scenes (right). Red line shows the median values, the box ranges from the upper to the lower quartiles, whiskers (in black) indicate the minimum and maximum values. The dashed grey line shows the probability to guess the correct sample (33%).

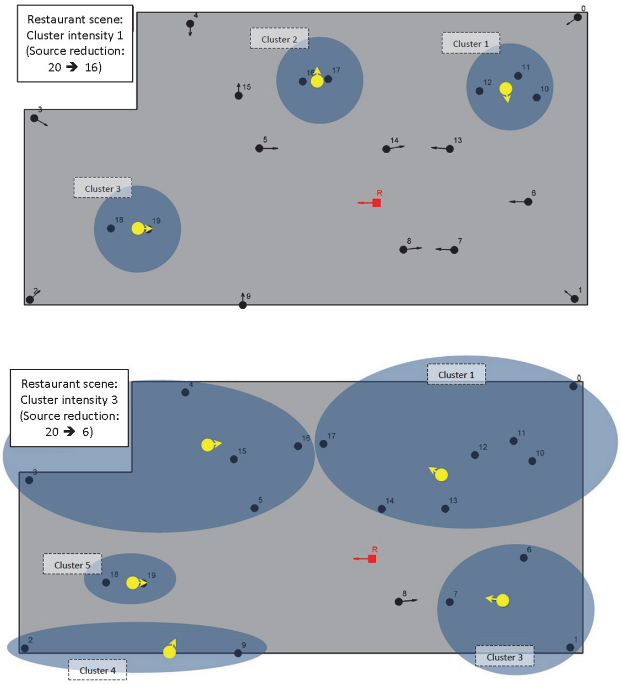

Figure 5: Top view of clusters created for the restaurant scene. Sources (black dots) are grouped in clusters (blue) and replaced by a representative source (yellow dots). The lower picture shows a higher clustering intensity with greater source reduction.

Performance

The algorithm was applied for the bus stop scene with different numbers of sources (20, 50 and 100) and for

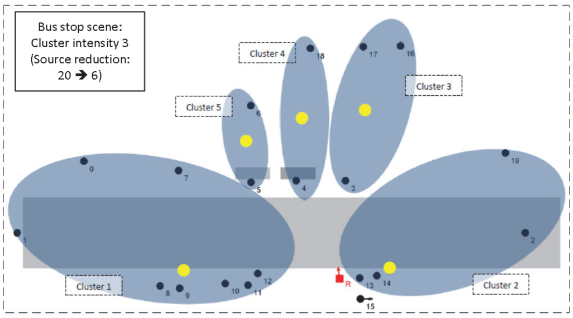

Figure 6: Top view of clusters created for the bus stop scene with a high clustering intensity.

three clustering intensities. The resulting clusters were visualized and the calculation times were averaged over 10 runs on a notebook pc (*Core i5*, 2.5 GHz) and a desktop pc (*Core i7*, 3.4 GHz).

Overall the calculation times increase with the number of sources, because for each source combination the angle and distance criteria are examined. For 20 sources, the calculation times (see Tab. 1) are always below 1 ms and do not exceed 4 ms for 100 sources. In this case, the number of sources was lowered by 20% - 86% depending on the clustering intensity and in this way drastically reducing the computational workload of the simulation engine. The analysis (caluclation times around 1 ms) indicates that in a real-time simulation environment, the scene simplification process could run at an update rate of up to 100 Hz, leaving sufficient computing time for the acoustic rendering of the scene.

Sources	Clustering intensity	Remaining sources	Reduction [%]	Calc. time notebook [ms]	Calc. time desktop PC [ms]
20	1	12	40	0.271	0.157
	2	8	60	0.264	0.132
	3	6	70	0.253	0.118
50	1	19	62	1.205	0.653
	2	11	78	1.101	0.488
	3	3	82	1.060	0.486
100	1	33	67	3.879	1.639
	2	18	82	3.053	1.234
	3	14	86	2.668	1.159

Table 1: Calculation times for the simplification algorithm for the outdoor scene (bus stop) with up to 100 sources.

Conclusion

For large scenes containing numerous virtual sound sources an automated identification of irrelevant sound propagation paths is an efficient method to significantly reduce the number of simulation tasks in a real-time auralization environment. By considering psychoacoustical and geometrical aspects, it is possible to render simplified acoustic scenes without an audible effect for the listener. The concept was tested and investigated by conducting a listening tests for two example situations. Results showed that listeners could not reliably detect the simplified situations and that the detection rates strongly depend on the simulated scene environment and/or the type of signals used in the scenario. A general solution for an adequate scene simplication is however hard to find. Nevertheless, for interactive multimodal simulations without the possibility to repeat the samples it is likely that the detection rates are lower than in the unimodal laboratory environment and the auralized scenes appear convincing to the users. The next step of this research will include the consideration of source directivities and further validation of the implemented concept. This will be followed by the integration of the scene simplication module into a real-time auralization framework, in which, as shown by the performance analysis, the simplfication algorithm can be run at high update rates and lead to an overall reduction of the computational workload.

References

[1] Allen, J.; Berkley, D.: Image method for efficiently computing small-room acoustics. In: Journal of the Acoustic Society of America 1979 (65), Nr. 4, S. 934–950

[2] Vorländer, M.: Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm. In: Journal of the Acoustic Society of America 86 (1989), Nr. 1, S. 172–178

[3] Aspöck, L.; Pelzer, S.; Wefers, F.;Vorländer, M.: A Real-Time Auralization Plugin for Architectural Design and Education Proceedings of the EAA Joint Symposium on Auralization and Ambisonics, 3-5 April 2014 in Berlin, Germany.

[4] Tsingos, N.; Gallo, E.; Drettakis, G.: Breaking the 64 spatialized sources barrier. In: Gamasutra (2003)

[5] Tsingos, N.; Gallo, E.; Drettakis, G.: Perceptual audio rendering of complex virtual environments. In: Marks, J. (Hrsg.): ACM SIGGRAPH 2004 Papers, p. 249

[6] Herder, J.: Optimization of Sound Spatialization Management through Clustering. In: Journal of the 3D-Forum Society (Sep. 1999), No 13 (3)

[7] Blauert, J.: Spatial Hearing: The Psychophysics of Human Sound Localization: 2nd Edition. Cambridge, Massachusetts : The MIT Press, 1997

[8] Schröder, D.: Physically based real-time auralization of interactive virtual environments. PhD thesis, RWTH Aachen University. Bd. 11. Logos Verlag Berlin GmbH, 2011