Evaluation of Auralization and Visualization Systems for Railway Noise Scenes

Pieren, Reto; Heutschi, Kurt; Aalmoes, R; Simons, Dick

Publication date
2017

Document Version
Accepted author manuscript

Published in
Proceedings of the 46th International Congress and Exposition on Noise Control Engineering Taming Noise and Moving Quiet

Citation (APA)
Pieren, R., Heutschi, K., Aalmoes, R., & Simons, D. (2017). Evaluation of Auralization and Visualization Systems for Railway Noise Scenes. In *Proceedings of the 46th International Congress and Exposition on Noise Control Engineering Taming Noise and Moving Quiet: Taming Noise and Moving Quiet, 27-30 Aug 2017 Hong Kong, China* (pp. 6555-6566)

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Postprint
Inter-Noise 2017: 46th International Congress and Exposition on Noise Control Engineering Taming Noise and Moving Quiet
27-30 August 2017
Hong Kong, China

Evaluation of Auralization and Visualization Systems for Railway Noise Scenes

Reto PIEREN[1]; Kurt HEUTSCHI[2]; Roalt AALMOES[3]; Dick G. SIMONS[4]

[1,2] Empa, Swiss Federal Laboratories for Materials Science and Technology, Switzerland

[3] NLR, The Netherlands

[1,4] Delft University of Technology, Faculty of Aerospace Engineering, The Netherlands

ABSTRACT

By directly addressing the hearing sensation, auralization is an intuitive means for the assessment and communication of noise scenarios. Authenticity can be further improved by offering visual information. Currently, various auralization and visualization systems exist, that differ with respect to their sound and image signal generation as well as their reproduction strategy. Within the past few years, the entertainment industry has launched several virtual reality (VR) products such as head-mounted displays or game engines that appear attractive for applications in environmental sound auralization. This contribution gives an overview of current VR systems and introduces evaluation criteria for comparison and assessment. Focus here is on the application to different railway noise scenes. This involves the generation of stimuli for experimental studies and the use as a demonstrator. To achieve maximal flexibility with respect to scenarios and the reproduction systems, the synthesis of sound and images on the basis of on an object-based approach is suggested.

Keywords: Auralization, Visualization, Railway Noise
I-INCE Classification of Subjects Numbers: 61, 76, 79

1. INTRODUCTION

By directly addressing the hearing sensation, auralization is an intuitive means for the assessment and communication of noise scenarios. Authenticity can be further improved by also offering visual information. Thus, through auralization and visualization, different design alternatives or noise mitigation measures can be demonstrated to the public or decision-makers. Furthermore, by using them as stimuli in experimental tests, indicators can be developed to predict noise annoyance, sound quality or acoustic comfort. Also, the use of auralization and visualization help to prepare the local community for changes in railway noise in combination with visual changes of the environment.

Currently, various auralization and visualization systems exist, that differ with respect to their sound and image signal generation, as well as their reproduction strategy. Within the past few years, the entertainment industry has launched several virtual reality (VR) products such as head-mounted displays or game engines that appear attractive for applications in environmental sound auralization and visualization.

This paper gives an overview of current visualization and auralization systems and introduces evaluation criteria for comparison and assessment with respect to railway noise scenes. On that basis, different system variants are proposed and discussed.

[1] reto.pieren@empa.ch
[2] kurt.heutschi@empa.ch
[3] roalt.aalmoes@nlr.nl
[4] d.g.simons@tudelft.nl

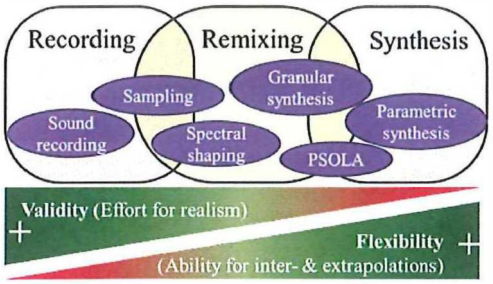

Figure 1 – Three sound generation approaches and examples of established methods with an assessment of their validity and flexibility

2. SOUND GENERATION APPROACHES

The core of an auralization system is the ability to generate and process audio data. In this section, different approaches for sound signal generation are presented with application in railway noise scenes.

Audio files may be either based on sound recordings or sound synthesis, as illustrated Figure 1. These two approaches are elucidated in the following sections. However, also hybrid methods exist, here denoted as remixing, where recorded sounds are modified in various ways. In sampling, portions of recordings (samples) are shifted in time, scaled in amplitude and summed up. This method was recently used in laboratory studies for the auralization of road traffic scenarios (1,2). With spectral shaping, the spectral content of a recorded signal is altered by filtering. This approach has been used in the project "Infopunkt Lärmschutz" of Deutsche Bahn and the Fraunhofer HHI where railway noise mitigation measures have been auralized (3). Methods which are commonly denoted as synthesis but also rely on recordings are granular synthesis and PSOLA, where recorded signals are dissected into short grains which are then manipulated and mixed together. In contrast to the above mentioned methods, parametric synthesis generates audio signals purely artificially.

As illustrated in Figure 1, relying on recordings has the advantage of an inherent high degree of realism. They however provide a low flexibility as only existing cases can be adequately represented. In contrast, parametric audio synthesis features a very high degree of flexibility and is very versatile. It allows for interpolations between known states but also for extrapolations to new, non-existing cases. It is however very challenging to synthesize realistically sounding signals.

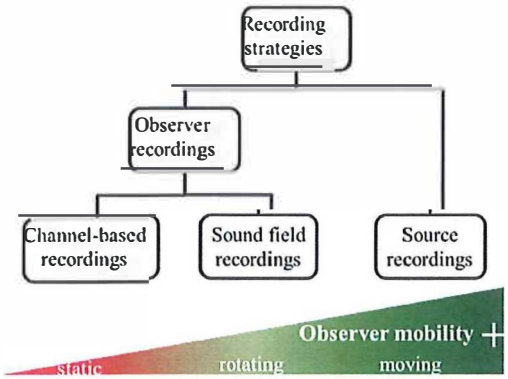

Figure 2 – Hierarchy of sound recording strategies with three types of virtual observer mobility

2.1 Sound recording

For audio recordings, a large variety of different microphone characteristics, arrangements and placement strategies exist. The preferred choice depends on the application, e.g. the virtual observer type and the reproduction system. Figure 2 shows that recording strategies can be split into two main classes. When used in an auralization system, these three recording strategies result in different degrees of virtual observer mobility. They feature either the perspective of the source or the observer. In the first class the emitted sound from a source is recorded whereas in the second class the sound at

an observer location is captured. The main difference is that in the latter case sound propagation effects are included. Further, in source recordings, emission angles are used as a descriptor, whereas in observer recordings, immission angles are relevant.

Source recordings aim at capturing the sound pressure in close proximity as radiated by a source, at a certain emission angle. To do so, a microphone is typically placed very close to an individual source, such as implemented in (4) for railway interior noise. This approach is often challenging due to practical reasons such as source motion or multiple interfering sources.

Figure 2 shows that the observer recordings can be further split into channel-based and sound field recordings. **Sound field recordings** aim at capturing the effect of the sound field at a specific location independently of the reproduction system. For that, an abstract representation is used which allows to approximatively reproduce sound pressure and particle velocity of that location. In **channel-based recordings**, the microphones are also placed at the observer location but consistent with a specific sound reproduction system. Channel-based recordings require little processing, at the cost of being inflexible with respect to the reproduction system and observer mobility.

2.2 Parametric sound synthesis

Parametric synthesis is very versatile as it gives complete control over the signals. A widely used parametric synthesis technique is referred to as Spectral Modelling Synthesis (SMS). It is a combination of additive and subtractive synthesis and is successfully used in the context of sounds from aircraft (5,6), wind turbines (7) and road traffic (8). In additive synthesis, a signal is constructed by the sum of discrete sinusoids, each having a time-varying amplitude and phase. Subtractive synthesis uses time-varying filters to shape a broadband waveform, typically white noise. However, SMS is not well suited to produce impulsive sounds.

Likewise, physical modelling can be seen as a parametric synthesis technique. It separately models the excitation and vibration of a system and successfully generates impulsive sounds from percussion, bells or footsteps (9,10). One subtype of physical modelling is modal synthesis, which uses a modal description of vibration. With this approach the complex dynamic behavior of a vibrating object is modelled by the superposition of contributions from a set of modes (11). In the Swiss research project TAURA, a synthesizer for railway rolling and impact sound was developed (12,13) that incorporates modal synthesis. Figure 3 shows simulation data that was generated using this model.

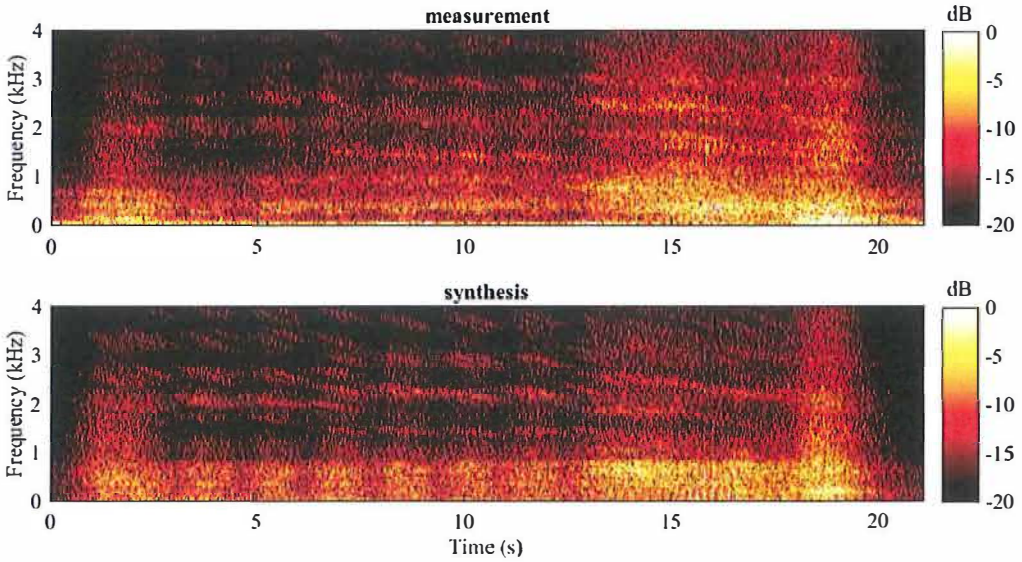

Figure 3 – Spectrogram of the sound pressure at the reference position (7.5 m/1.2 m) during a train passage with speed 60 km/h. The upper panel shows data from a measurement and the lower panel the corresponding synthesis that was produced with the physical modelling approach from (12,13).

2.3 Propagation filtering

If source signals are being synthesized or recorded, sound propagation effects yet have to be added to these signals. The propagation phenomena that have to be considered depend on the specific scene

and its application. In any case, geometrical divergence must be applied, which for the far-field sound pressure of a point source is a $1/r$ distance dependency. For quickly varying source-receiver distances the Doppler effect may become relevant and should therefore be simulated (8). For large propagation distances air absorption reduces the high frequency content (14) and atmospheric turbulences may introduce level fluctuations (15). Reflections at boundaries lead to interferences and may create diffuse sound fields, especially in room acoustical situations. Also shielding of sources may become important e.g. in the case of a noise barrier. Modern auralization models for environmental sounds simulate sound propagation with a physically-based model and process the source signals with time-varying digital filters (14,16,5,8).

3. IMAGE GENERATION APPROACHES

For highly immersive visualization purposes, a first-person view is appropriate. In this situation, the test subject will experience both the visual and audio cues as if he or she is the observer.

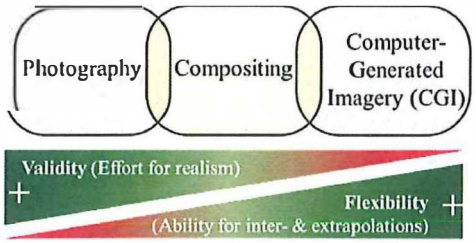

Figure 4 – Different approaches of image generation with an assessment of their validity and flexibility

To create a visualization, similar to auralization, two different generation approaches can be taken with both benefits and drawbacks. The photography approach is to capture the environment by motion picture acquisition. By using two horizontally separated lenses, stereoscopic recordings can be made. Omnidirectional (or panoramic, or 360°) capturing systems feature a 360° field of view in the horizontal plane. The photography approach limits the maneuverability within the scene, but it is a quick way to capture the environment. However, modifications of objects are difficult. Another approach is the use of Computer-Generated Imagery (CGI) to generate images by reproducing each object in a virtual 3D environment. This requires sophisticated mapping of objects in the scene. To make the scene more (photo) realistic, surfaces can be supplied by textures that match existing surfaces. Although computer graphics is improving year by year, a photorealistic CGI scene requires significant effort. A combination of both approaches is denoted as compositing.

4. SOUND REPRODUCTION SYSTEMS

The sound reproduction system transforms the generated audio data, as described in section 2, into an audible sound field. It should produce appropriate and specific sound pressures inside the left and right ear canals of the listener and thus create a credible audible impression of the virtual environment.

An important requirement is the ability to localize sound sources as this increases the credibility of an auralized scene. This task, often referred to as spatial audio or 3D audio, requires multiple acoustic transducers. The generation of the individual transducer channel signals is denoted as spatialization. Two transducer types are available: multiple loudspeakers and headphones. Both solutions have their advantages and disadvantages, which are detailed in the following sections.

4.1 Sound scene description

The sound generation module provides appropriate input for the sound reproduction module. The two modules differ in the following aspects: The sound generation module creates audible signals at the observer location which therefore include sound propagation effects. Thus, this module uses data about the virtual environment including information about the angles under which a sound wave reaches the observer point. On the other hand, the reproduction system is blind with respect to the virtual environment.

As an interface between these two modules, different sound scene description are being used. In an **object-based description**, each virtual source is represented separately by a sound pressure signal at the observer with corresponding time-variant polar angles. The concept of **Ambisonics** is to mathematically approximate the sound pressure field at the observer point using spherical harmonics.

From an object-based description, the collection of spherical harmonics (*B-format)* can be created by an ambisonic encoder. First-Order Ambisonics contains the sound pressure plus one component per spatial dimension. Ambisonics of higher than first order is denoted as Higher-Order Ambisonics.

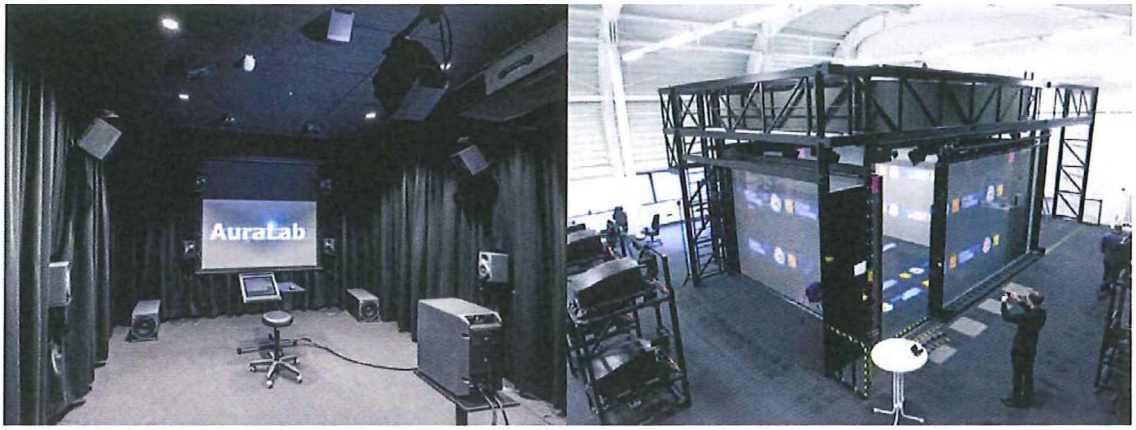

Figure 5 – Photographs of the upper hemisphere loudspeaker layout in Empa's AuraLab (left) and of the aixCAVE at RWTH Aachen University (right)

4.2 Loudspeaker reproduction

For multichannel loudspeaker reproduction, three aspects have to be considered which are

- the loudspeaker layouts, i.e. the number and placement of the loudspeakers in the room,
- the reproduction rendering, i.e. the technique to calculate individual loudspeaker feeds and
- room acoustics.

They are separately described in the following sections.

4.2.1 Loudspeaker layouts

Today, many different speaker layouts are in use. They differ with respect to the number of loudspeakers and where they are placed in the room. Typically multiple satellite speakers are arranged with differing layouts. However, for low frequencies, where sound sources cannot be localized, one or multiple subwoofers are used. Most reproduction systems are optimized for one specific listening point and require that the distances to the satellites are equal. Otherwise, loudspeaker-specific time-delays and gains may be used for compensation. A minimal listening distance of 2 meters is recommended for high-quality listening rooms. The zone around the listening point providing an adequate sound experience is denoted as sweet spot or optimum listening area.

Loudspeaker layouts may be divided into three classes depending on their used spatial dimensions:

- frontal setups
- 2D surround setups
- 3D audio

Frontal setups are 1D layouts where the loudspeakers are positioned on a line in front of the listener. In 2D surround layouts the loudspeakers are located on a horizontal plane. The height of their acoustic centre is at listener's ears, which in most cases is 1.2 m above floor for a seated person. For ambisonic reproduction, regular layouts are highly preferred. Some rendering strategies and applications also require very dense speaker arrangements. One such example is the TiME Lab at the Fraunhofer HHI in Berlin which uses a horizontal layout of 120 loudspeaker channels for a Wave Field Synthesis reproduction (3). Generally, for a given reproduction rendering strategy, the localization capability and stability grows with the number of loudspeakers.

For 3D audio even more variants exist. Full-sphere arrangements contain loudspeakers all around the listener, whereas upper hemisphere arrangements only cover the space on and above the listening plane. Regular layouts are achieved by placing loudspeakers at the corners of a platonic solid or a Lebedev grid. Irregular layouts usually consist of several horizontal loudspeaker layers with typically one being in the listening plane (see e.g. Empa's AuraLab in Figure 5). In the past few years, several 3D audio formats for cinema, such as Auro-3D, Dolby Atmos or DTS:X, have been launched. They describe irregular upper hemisphere arrangements with a total of two or three layers.

Very specific irregular layouts are used in cases where discrete loudspeakers are used to directly

simulate non-moving acoustical sources *in-situ*. With this approach the interior sound within a train vehicle mock-up was simulated (4).

4.2.2 Reproduction rendering

By reproduction rendering an input signal for each loudspeaker is derived. Various techniques to calculate signals for multiple loudspeakers, denoted as speaker feeds, exist. These techniques strongly depend on the given input type, which is a certain sound scene description, and their desired output format, i.e. a specific loudspeaker layout.

One rendering strategy may be called **virtual microphones**, where the responses of microphones within an appropriate arrangement are simulated at the observer location. A possible microphone spacing is simulated by varying time delays between the received signals and directivities by amplitude modulation. Both processes are steered by source-specific immission angles, implying that this strategy requires an object-based sound scene description. This strategy is described for a 2-channel stereo arrangement using virtual ORTF in (8).

Pair-wise amplitude panning is a strategy to obtain speaker feeds based on the creation of phantom sources. For that, different panning laws and normalizations exist to determine the speaker gains. The normalization type expresses the assumption about signal superposition at the observer point. The most widely used pair-wise panning method for 3D audio is **Vector Base Amplitude Panning (VBAP)** (17), which is a generalization of the (stereo) tangent panning law. Several modifications of the classical VBAP do exist such as a frequency dependent gain normalization (18) or the introduction of virtual speakers (19). Pair-wise amplitude panning requires an object-based sound scene description and can handle almost any irregular loudspeaker layout. As only a minimum number of feeds are simultaneously active, it produces good source localization, average sized sweet spot and only little coloration.

Ambisonic decoders are another kind of amplitude panners. In contrast to VBAP they may provide negative speaker gains which allows for sound wave cancellation at the center. Further, instead of only a few speakers, typically all speakers are simultaneously active and used to recreate the desired sound field from the spherical harmonic components. Different decoding algorithms exist to calculate speaker feeds from an ambisonic description (*B-format*). The number of loudspeakers must exceed the number of ambisonic components. In Ambisonics, regular speaker layouts are highly preferred.

With **Crosstalk-Cancellation (CTC)** the sound pressures at the ears are controlled by applying inverse filters to reduce the effect that the signal from the right speaker is picked up by the left ear, and *vice versa*. Therewith a spatial impression can be created with only two loudspeakers. This binaural rendering strategy was extended to more than two loudspeakers and a dynamic rendering (20,21). However, in practice compensation of head movements and the estimation of the crosstalk paths are difficult.

With **Wave field synthesis (WFS)** the desired sound field within a horizontal plane is reconstructed using surrounding sound sources. This technique is based on the Kirchhoff-Helmholtz integral and uses typically more than 100 loudspeakers as the upper cut-off frequency is determined by the minimal distance between individual loudspeakers. It allows for a large sweet spot, however at the cost of a high complexity and effort.

4.2.3 Room acoustics

Listening rooms with controlled room acoustical conditions are mainly achieved by use of sound absorbers to attenuate sound reflections. These rooms typically feature a very low reverberation time, e.g. Empa's AuraLab (see Figure 5) with $T_{mid} = 0.11$ s. They thus allow for the simulation of different acoustical environments and for the creation of virtual sources. For that purpose, anechoic or semi-anechoic rooms are ideal. In non-anechoic rooms, a certain amount of spectral coloration can be compensated using loudspeaker-specific equalization, such as reported for NASA's Exterior Effects Room (EER) (19). Ultimately, each loudspeaker channel has to be calibrated using a sound level meter. The ease of level calibration can be seen as a major advantage of loudspeaker reproduction compared to headphones.

4.3 Headphone reproduction

Modern headphone reproduction of spatial audio material is based on the combination of two technologies: Binaural technology and head tracking, both of which are described in the following sections. When listening to a monophonic signal through headphones, the signals which are received at the ear drums lack the influence of the own body on the sound field. Notably the head, torso and pinna significantly affect the received signals by providing shielding, reflections and time delay, which are

all a function of the incident angles. This information is essential for the localization of sound sources.

Various headphone types exist that have to be differentiated. Ear-fitting headphones should be avoided due to difficulties regarding calibration and lack of reproducibility of their mounting. Instead circumaural headphones are preferred. Closed-back headphones feature higher attenuation of ambient noise compared to open-back headphones. For some applications and environments even noise-cancelling headphones might be needed.

4.3.1 Binaural technology (HRTF rendering)

The influences of head, pinna and torso on the sound pressure at the location of the ear canal entrance are described by Head Related Impulse Responses (HRIR). For a point source in the far-field, the two HRIRs for left and right ear are a function of the immission angles, e.g. azimuth and elevation. Their Fourier transforms are known as Head Related Transfer Functions (HRTF). By convolving a monophonic signal with the corresponding HRIRs of left and right ear, a binaural (two-channel) signal is rendered.

Three major challenges in binaural technology are HRTF interpolation, individualization and calibration. Due to practical reasons measurements are best made with a spatial resolution of 5°. This is however too rough compared to the localization capabilities of humans. Therefore and in order to realize smooth transitions between the measurement points, a HRIR interpolation strategy is needed. Numerous algorithms have been proposed in the past years.

Most binaural renderers use general HRTFs which are e.g. obtained from acoustic measurements with a head-and-torso simulator or averaged over many persons. However, HRTFs are not universal, but audibly differ from person to person. Therefore high-quality binaural reproduction requires some sort of individualization. Direct acoustic measurements of HRIR are costly and difficult. Another active field of research is the development of generic models for HRTFs that use anthropometric data as input (22). HRFTs are also individualized using ear photographs as input.

Instead of rendering monophonic signals in an object-based manner, also intermediate formats are used during binaural rendering. The use of such an intermediate format can be interpreted as a form of interpolating HRTFs. One possibility is Ambisonics, where the B-format is decoded (or transcoded) to a binaural format. Ambisonics has the advantage that virtual rotations can be easily realized with a matrix mixer. The modified ambisonic channels are then convolved with static HRTF-like filters and summed up. These filters are usually designed based on a virtual loudspeaker setup.

Another possibility is to use a virtual loudspeaker layout as intermediate format. Firstly, the signal for each virtual loudspeaker is rendered. Secondly, these non-moving virtual sources are binaurally rendered using HRTFs. Such an emulation of a surround sound system through headphones is often denoted as virtualization and typically covers 5.1 or 7.1.

4.3.2 Head tracking

When listening to standard headphones, the reproduced sound field moves with the head. This is highly unnatural and thus weakens the credibility of the scene. To overcome this, an adaptation of the sound scene for (slight) head rotations is required. Binaural rendering thus becomes a function of the current head orientation which is measured by head tracking technologies. A passive measurement can be obtained by optical ("outside-in") tracking relying on a video camera in the room and a dynamic face recognition algorithm. Active optical systems use infrared lighting, wearable motion tracker markers and cameras. For optical systems unobstructed line-of-sight is necessary for continuous tracking. Another option is electromagnetic motion tracking. Source-less tracking is achieved by wearable inertial sensors, i.e. accelerometers and gyroscopes. However, inertial tracking may suffer from drift errors due to temporal integration. The data transmission from the wearable sensor to a static receiver station follows either via cable or wireless (e.g. Bluetooth).

5. VIDEO REPRODUCTION SYSTEMS

The presentation of the visualization to the audience (or participant) can be performed with a variety of means. A drawback of using a computer monitor or TV screen is that the field of view is limited: only a small angle of the capabilities of the human eye can be utilized. High immersion can be achieved by either using a large-screen display or a head-mounted display.

5.1 Large-screen displays

Large-screen displays can be realized either by direct view displays (e.g. LCD) or by projection to a screen. The latter can be either lighted from the front (reflective screen) or the rear (transmissive

screen). A powerwall consists of multiple, synchronized assembled displays in a matrix to achieve a higher image resolution. This also applies to projection where multiple projectors are combined. Screen shapes can be flat, curved or form a closed surface such as in a Cave Automatic Virtual Environment (CAVE) (e.g. aixCAVE at RWTH Aachen depicted in Figure 5).

For high-quality stereoscopic presentation, 3D glasses have to be worn. Utilized technologies to provide separate images to left and right eye include light polarization, shutter or wavelength multiplexing. Compared to monoscopy, stereoscopy thus increases immersion but also intrusion.

5.2 Head-Mounted Display (HMD)

A head-Mounted Display (HMD) is a wearable optical display which produces images close to the eyes. HMDs comprise high-resolution stereoscopic displays, combined with lenses to increase the field of view to 110°, and head tracking sensors to create the illusion of a (maximum) 360-degrees horizontal view, including 180 degrees vertical view. This technology is used in NLR's Virtual Community Noise Simulator (VCNS) which has been applied in the aerospace sector (5), but also for the presentation of new wind farms.

The main advantage of this approach is the reduction of demonstration space that is needed, and recent developments have increased the quality and reduced the costs for this solution (23). A disadvantage of HMDs is the reduction of the possibility of interaction between participants.

6. COMBINATION OF AURALIZATION AND VISUALIZATION SYSTEMS

6.1 Requirements on combined reproduction systems

When combining a spatial audio reproduction system from section 4 with a video reproduction system from section 5, some practical conflicts may arise which should thus be considered during the system design process. Some of these possible conflicts are presented in the following.

For realistic auralizations, the background noise of the system should be well below the lowest level of the scenarios in all relevant frequency bands. For spatial hearing, the cocktail party effect should be considered, leading to more rigorous background noise level requirements compared to monaural listening.

When combining loudspeakers with large screens, two aspects must be considered. Firstly, the screens should not provide disturbing sound reflections. Secondly, the loudspeakers should not obscure the screens, i.e. not be placed in the line-of-sight.

The combination of the audio and video reproduction system determines the head tracking requirements. Table 1 presents a categorization using two questions, one regarding the screen and one regarding the audio reproduction rendering. Video reproduction systems consists of static or moving displays. The latter move with the subject's head, e.g. HMDs. To dynamically render a virtual reality, they require the current subject's head orientation using head tracking. To allow for (slight) head rotations in binaural headphone reproduction, the applied HRTFs have to be dynamically adapted which requires knowledge about the current head orientation. Besides headphone reproduction, this also applies for dynamic binaural rendering using (static) loudspeakers, denoted as dynamic crosstalk cancellation. Table 1 reveals that the only system combination without the need for head tracking consists of a fixed screen reproduction, with either a non-binaural rendering with loudspeakers or a static rendering over headphones.

Table 1 – Head tracking requirements for different reproduction system combinations

Head tracking requirements		Moving screen?	
		No	Yes
Dynamic binaural rendering?	No	none	image
	Yes	sound	image and sound

6.2 Game engines

Game engines are software frameworks for the development of video games. Compared to visualization engines or image generators, they not only include image generation functionality, but also functionality related to audio, physics (modelling), interaction, and other functionality to create an interactive (entertainment) game or serious game environment. With interactivity being of central

importance, most of the processing is done in real time. Their strength lies in the image generation based on physical models. With respect to audio, they support the use of pre-recorded sounds. The signals can either be directly sent to the audio output channels or used in an object-based fashion by attributing them to point sources in the virtual space. However, only simple, non-physical sound propagation simulation is currently implemented. In general, game engines are flexible, contain many import and export options and support different image and sound reproduction systems.

Table 2 contains an assessment of different content generation modes, including real-time rendering and (offline) pre-rendering of scenes.

Table 2 – Assessment of different image and sound generation modes when using a game engine

Image generation	Sound generation	Sound quality	Interactivity
real-time	real-time	poor	high
real-time	external real-time synthesis	medium	high
pre-rendered	pre-rendered	high	low

7. SYSTEM EVALUATION

7.1 Object-based modelling approach

Three virtual observer types with respect to their mobility are distinguished, namely a static, rotating and moving observer. A static observer has a fixed position and direction. The rotating observer has a fixed position but is allowed to rotate. The moving observer enjoys full freedom by changing its position and direction. The observer type has severe implications on the modelling approach and thus determines the model complexity and flexibility. Figure 6 illustrates the three virtual observer types and the resulting modelling approaches.

In an auralization system that only supports a static observer with respect to the sound scene, it is sufficient to exclusively model this specific immission situation. It means that the immission signals may be generated in an integral way and thus sound propagation does not necessarily need to be separately treated. For a rotating observer, the system has to allow for virtual rotations of the reproduced sound. This is typically achieved by using an intermediate sound scene description at the observer point. As a result of not using a channel-based format, sound generation and reproduction are decoupled which allows for renderings in various reproduction systems.

However, an auralization that principally supports different observer locations has to be formulated as an object-based approach. Such a system may be structured into three distinct simulation modules, namely the source signal generation, the sound propagation filtering and the reproduction rendering. This approach features the highest degree of flexibility with respect to scenarios and reproduction systems.

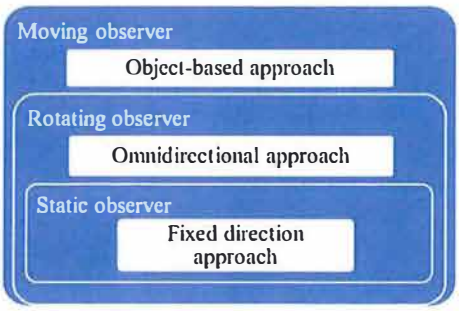

Figure 6 – Virtual observer type determines modelling approach

7.2 Evaluation criteria

To differentiate, compare and assess different auralization and visualization systems, a list of evaluation criteria was established. The majority of general criteria can be applied to auralization as well as to visualization systems and are presented in the following.

Portability: This criterion describes how easily the system can be moved from one place to another. It also describes to what extent the system depends on the specific location. For the application as a demonstrator it is e.g. important that simulations can be easily showcased at different places. On the

other hand, to conduct experimental laboratory studies, a static system has the advantage that it is well defined and controlled.

Number of simultaneous participants: The number of people who can simultaneously experience the simulation is usually small. Most presented systems have a small sweet spot and are thus restricted to one or a few persons only. However, the scalability between the systems may vary.

Interactivity and user control: The degree of interactivity and the user control by the participant define the minimal required amount of real ime signal processing. Virtual observer rotations require a real-time reproduction rendering which implies an appropriate omnidirectional format at the observer point. The option to activate and deactivate acoustical sources requires the separate rendering of their contributions and a real-time mixing. Virtual observer movements or changes of the environment (such as the introduction of a noise barrier) finally require real-time propagation simulations as well as reproductions, given that the sound sources have simple directivities. Complex sound source modifications or complex directional behavior additionally require real-time sound synthesis.

Physical correctness: Self-evidently, a simulation system has to be assessed with respect to its physical correctness. Firstly, it may be heuristically checked with respect to the physical phenomena that are accounted for and how appropriate the chosen models are. Secondly, the performance can be measured and compared, e.g. the angular coverage or the dynamic range of the sound and video reproduction system, respectively. Thirdly, deviations between ideal and reproduced values can be measured and evaluated. This calibration test involves e.g. the sound pressure level at a certain frequency and location or the brightness of an object.

Intrusion: Audio and video reproduction systems differ with respect to the degree of intrusion. If the participant has to wear some sort of equipment – e.g. glasses or headphones – this is unnatural and may have an impact on his assessment of the simulation. Besides technical aspects such as the weight of the equipment, also comfort plays a role.

Immersion: A high immersion means that the participant believes to be inside the virtual environment. The degree of immersion is influenced by the audio as well as the video reproduction system. In particular interactivity and spatial cues seem to be important for immersion, e.g. by spatial audio and stereoscopic view.

Perceived realism and plausibility: Generally, the simulations should be perceived as realistic as possible. Apart from photo-realistic images and natural sounding audio, also factors such as the degree of immersion and intrusion play a role. In general, this criterion is very difficult to quantify. For some applications, particularly if only relative differences between scenarios are in the focus, merely a plausible simulation is sufficient. Plausibility can be understood as the difference to an inner reference due to experience and expectations (24). The perceived quality of spatial audio systems may be investigated using the vocabulary developed in (25).

Appropriateness: The appropriateness of a system depends on its application. Possible applications are experimental laboratory studies on perception or the use as a demonstrator for alternative scenarios. Further, the system must be appropriate for the kind of participants, e.g. experts or laymen, and the type of relevant scenes to be simulated. A multifunctional system is appropriate for several applications, kinds of participants or scenes.

Effort: Every system requires a certain technical, financial and personnel effort. Technical efforts comprise the equipment, the room and the installation, but also the computational effort. The financial costs for equipment and implementation vary a lot across systems. The personal effort includes the system development as well as its usage and maintenance.

Complexity and flexibility: Some fundamental decisions about the modelling approach highly determine the system complexity. However, in return they often offer a greater flexibility with respect to scenarios or reproduction systems. Also modifiability, customization and expandability of the system determine its flexibility.

7.3 Loudspeaker vs. headphone reproduction

Regarding the audio reproduction system, one central question is whether to use headphones or multiple loudspeakers as transducers to create a spatial audible impression. They have their specific advantages and disadvantages which are discussed in the following.

The major advantages of headphones are their portability and lack of room influence. This permits a high flexibility regarding the choice of the location, which is especially important for a demonstrator.

However, the major disadvantages of headphones are their intrusion, the difficulty of calibration and the need for simulating the subject's influence on the sound field. The intrusion includes

discomfort due to acoustical isolation, the device's weight, contact pressure, disturbing cables or warming-up of ears. If a HMD is used together with headphones, the intrusion impact seems of lesser importance due to the (breached) intrusion by the HMD. Headphones are difficult to calibrate as their frequency response often strongly varies between listeners, but also between repeated placings on the head. The influence of head, pinna and torso on the sound field is individual and thus laborious to simulate in a proper way. Otherwise incorrect sound levels and localization confusion including in-head localization occur. And ultimately, binaural rendering requires the use of a head tracking technology.

These disadvantages are the major advantages of loudspeaker reproduction, given that a non-binaural rendering is used. Loudspeaker reproduction is non-intrusive and thus more natural. Separate channels can be calibrated easily using standard measurement equipment. Further, non-binaural loudspeaker reproduction nearly perfectly reproduces the individual's influence on the sound field. The main disadvantages however are the limited portability and the severe room influence on the reproduced sound field.

7.4 Application-specific reproduction system

The previous section already revealed that there is no single favored reproduction system. It is rather that the specific application determines the requirements and allows for an assessment of the criteria to define a system.

In the following, three different possible applications of an auralization and visualization system for railway noise scenes are discussed, namely

I. Stimuli generator for experimental laboratory study
II. Test bench for experts and engineers
III. Demonstrator for stakeholders, residents and broad public

Experimental laboratory studies are performed in highly controlled environments with well-defined, reproducible stimuli. They thus demand a maximal acoustical quality. As only a limited number of participants are involved, portability of the system is not essential. To avoid mutual influences between participants, usually there will only be one participant at the same time in the experiment and thus the system may operate with a small sweet spot. Based on that, for case I, a non-binaural loudspeaker reproduction may be recommended.

Also experts and engineers need a high-quality simulation. However, thanks to their expertise they might also deal with some sort of abstraction. A test bench is used during longer time periods and by several participants at the same time, which have to be able to interact and discuss. This sets high requirements on the intrusion of the system. Consequently, for case II, a loudspeaker reproduction and large-screen display would be recommended.

A demonstrator needs to be portable and cost-effective in order to be used at different locations and by many different persons. The main goal is to provide information in an easily understandable manner. Therefore a compromise has to be made with respect to the acoustical quality, in particular the physical correctness. Thus, for case III, headphone reproduction is recommended. A high immersion system may be built using a HMD. For large audiences, in some applications even a projector with loudspeakers may suffice, despite the limited accuracy with respect to calibration and a low immersion. A low-cost, low immersion system might also comprise of a computer screen or a smartphone display. This system would be very widely accessible by a broad public, using e.g. a web-based solution.

8. CONCLUSIONS

An overview of current auralization and visualization systems was given and evaluation criteria for comparison and assessment were introduced, with focus on the application to different railway noise scenes. To achieve maximal flexibility with respect to scenarios and reproduction systems, the synthesis of sound and images on the basis of an object-based approach is suggested. The specific application determines the requirements and allows for an assessment of evaluation criteria to finally define the reproduction system.

ACKNOWLEDGEMENTS

The DESTINATE project has received funding from the Shift2Rail Joint Undertaking under the European Union`s Horizon 2020 research and innovation programme under grant agreement No 730829.

REFERENCES

1. Schäffer B, Schlittmeier SJ, Pieren R, Heutschi K, Brink M, Graf R, et al. Short-term annoyance reactions to stationary and time-varying wind turbine and road traffic noise: A laboratory study. Journal of the Acoustical Society of America, 139 (5). 2016: p. 2949–2963.
2. Klein A, Marquis-Favre C, Champelovier P. Assessment of annoyance due to urban road traffic noise combined with tramway noise. Journal of the Acoustical Society of America, 141 (1). 2017: p. 231–242.
3. Schröter K. HHI TiME Lab, Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute. [Online].; 2017 [cited 2017 02 08. Available from: http://www.timelab-hhi.de.
4. Johansson Ö, Schönfeld S, Lindforss D. Sound sketch procedure for auralization of the interior sound of a high speed train. In Proc INTER-NOISE 2012; 2012; New York.
5. Arntzen M, Simons DG. Modeling and synthesis of aircraft flyover noise. Applied Acoustics, 84. 2014: p. 99-106.
6. Sahai A, Wefers F, Pick S, Stumpf E, Vorländer M, Kuhlen T. Interactive simulation of aircraft noise in aural and visual virtual. Applied Acoustics, 101. 2016: p. 24-38.
7. Pieren R, Heutschi K, Müller M, Manyoky M, Eggenschwiler K. Auralization of Wind Turbine Noise: Emission. Acta Acustica united with Acustica, 100. 2014: p. 25-33.
8. Pieren R, Bütler T, Heutschi K. Auralization of Accelerating Passenger Cars Using Spectral Modeling Synthesis. Applied Sciences, 6, 5. 2016: p. 1-27.
9. Cook P. Physically Informed Sonic Modeling (PhISM): Percussive Synthesis. Computer Music Journal, 21 (3). 1997: p. 38-49.
10. Turchet L. Footstep sounds synthesis: Design, implementation, and evaluation of foot-floor interactions, surface materials, shoe types, and walkers' features. Applied Acoustics, 107. 2016: p. 46-68.
11. van den Doel K, Pai DK. Synthesis of Shape Dependent Sounds with Physical Modeling. In Proc International Conference on Auditory Display (ICAD 96); 1996; Palo Alto, California.
12. Pieren R, Wunderli JM, Zemp A, Sohr S, Heutschi K. Auralisation of Railway Noise: A Concept for the Emission Synthesis of Rolling and Impact Noise. In Proc INTER-NOISE 2016; 2016; Hamburg, Germany.
13. Pieren R, Heutschi K, Wunderli JM, Snellen M, Simons DG. Auralization of Railway Noise: Emission Synthesis of Rolling and Impact Noise. Applied Acoustics (accepted). 2017.
14. Heutschi K, Pieren R, Müller M, Manyoky M, Wissen Hayek U, Eggenschwiler K. Auralization of Wind Turbine Noise: Propagation Filtering and Vegetation Noise Synthesis. Acta Acustica united with Acustica, 100. 2014: p. 13-24.
15. Rietdijk F, Heutschi K, Forssén J. Generating sequences of acoustic scintillations. Acta Acustica united with Acustica, 103. 2017: p. 331-338.
16. Rizzi SA, Aumann AR, Lopes LV, Burley CL. Auralization of Hybrid Wing-Body Aircraft Flyover Noise from System Noise Predictions. Journal of Aircraft, 51 (6). 2014: p. 1914-1926.
17. Pulkki V. Virtual sound source positioning using vector base amplitude panning. Journal of the Audio Engineering Society, 45 (6). 1997: p. 456-466.
18. Laitinen MV, J. V, Jussila K, Politis A, Pulkki. V. Gain normalization in amplitude panning as a function of frequency and room reverberance. In Proc AES 55th International Conference; 2014; Helsinki, Finland.
19. Faller KJ, Rizzi SA, Aumann AR. Acoustic Performance of a Real-Time Three-Dimensional Sound-Reproduction System. Langley Research Center, Hampton, Virginia:; 2013.
20. Lentz T, Renner C. A four-channel dynamic cross-talk cancellation system. In Proc Joint Congress CFA/DAGA '04; 2004; Strasbourg.
21. Röcher E, Kohnen M, Stienen JP, Vorländer M. Dynamic Crosstalk-Cancellation with Room Compensation for Immersive CAVE-Environments. In Proc 42nd German Annual Conference on Acoustics (DAGA 2016); 2016; Aachen, Germany.
22. Bomhardt R, Lins M, Fels J. Analytical Ellipsoidal Model of Interaural Time Differences for the Individualization of Head-Related Impulse Responses. Journal of the Audio Engineering Society, 64 (11). 2016: p. 882-893.
23. Parisi T. Learning Virtual Reality - Developing Immersive Experiences and Applications for Desktop, Web and Mobile Sebastopol, CA: O'Reilly Media; 2015.
24. Lindau A, Weinzierl S. Assessing the Plausibility of Virtual Acoustic Environments. Acta Acustica united with Acustica, 98. 2012: p. 804 – 810.
25. Lindau A, Erbes V, Lepa S, Maempel HJ, Brinkman F, Weinzierl S. A Spatial Audio Quality Inventory (SAQI). Acta Acustica united with Acustica, 100. 2014: p. 984 – 994.