Personalized Auditory Reality

Karlheinz Brandenburg[1,2], Estefanía Cano Cerón[2], Florian Klein[1],
Thomas Köllmer[2], Hanna Lukashevich[2], Annika Neidhardt[1],
Johannes Nowak[1,2], Ulrike Sloma[1], Stephan Werner[1]

[1] *Technische Universität Ilmenau, 98693 Ilmenau, Deutschland, Email: karlheinz.brandenburg@tu-ilmenau.de*

[2] *Fraunhofer Institute for Digital Media Technology IDMT, 98693 Ilmenau, Deutschland*

Introduction

In this work, we introduce Personalized Auditory Realities, a new research field that investigates methods for the manipulation of acoustic surroundings. Within such an auditory reality, users would be able to freely modify their acoustic scene by enhancing relevant sounds, suppressing irrelevant ones, or adding new ones. The perceived acoustic environment will follow the paradigm of augmented realities where real sounds are combined with added sound sources.

Figure 1 shows an exemplary scene similar to a cocktail party situation: There are different speakers (foreground objects), bubble noise and music in the background. A Personalized Auditory Reality allows you to turn down the bubble noise, but still listen to the music. You could reduce the volume of one speaker to follow the spoken word of another speaker. Another possibility would be to place an artificial audio object to your acoustic scene, for example another conversation partner who is connected via phone.

Such functionality at hand, it can be applied to a wide range of scenarios in different problem domains: improved intelligibility of talks and seamless integration of translations, "speaking" animals in the zoo, enhanced in-car audio or aids for training hearing-impaired or blind people.

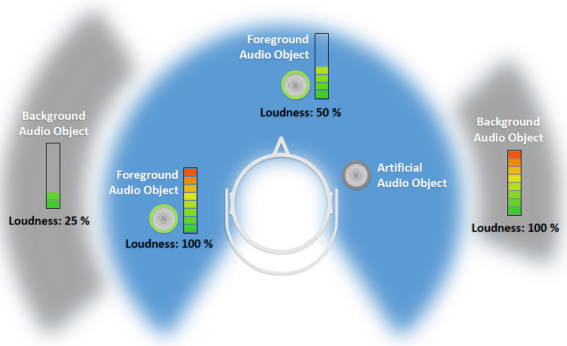

Figure 1: Personalized Auditory Reality: The user is able to dynamically mix real and virtual sound sources to adapt his current auditory situation to the ideal one.

Around the term *Virtual Reality* several other *-realities have been defined, such as augmented, diminished and mixed realities. Even if the current research is heavily focused on video, there is early research on *audio augmented realities* using binaural rendering by Cohan et. al. [5], though they are focusing on *augmenting* a perceived scene with virtual audio objects and do not try to *diminish* existing real audio sources in a controlled way. As the original authors try to "[...]overlay computer-generated imagery on top of real scenes." we are overlaying an altered representation of real sound sources on top of a virtual scene that can be generated easily nowadays.

We refer to a "Personalized Auditory Reality" as a *mixed* reality, where the user has real time control on the impact of all available virtual and real sources on his overall perceived audio experience.

Requirements and State of the Art

While previous research in the fields of video and audio analysis has addressed similar topics, no method or system exists today that allows the realization of perceptually convincing Personalized Auditory Realities. Our research tackles both headphone-based and loudspeaker-based reproduction of sound. We have to state that currently we do not see technical means to implement loudspeaker-based Personalized Auditory Realities. While part of the tasks could be realized via wave field synthesis or similar techniques, the overall problem is probably not solvable. Therefore in this work, we describe the state-of-the-art, system requirements, and first results of a system for headphone auralization. Even if there are some road blocks where it is not yet clear how to solve these issues.

To fulfill the ambitious goals described above, we combine and extend interdisciplinary research involving acoustics, digital signal processing, and data sciences (machine learning), all in close relation with auditory perception and quality. To achieve a high-quality integration of these technologies, the ideal system requires methods to: 1) decompose real-world acoustic scenes, 2) represent audio scenes as audio objects that can be manipulated, and 3) recompose scenes with added audio objects.

Scene Decomposition

The sound field has to be analysed and decomposed into distinct audio objects. The object separation should be without artefacts in order to enable a perceptually convincing scene. A high separation quality is mandatory for manipulating the objects' properties and to allow the recomposition of a scene.

One of the challenges is a high-resolution 3D sound field decomposition which usually requires microphone

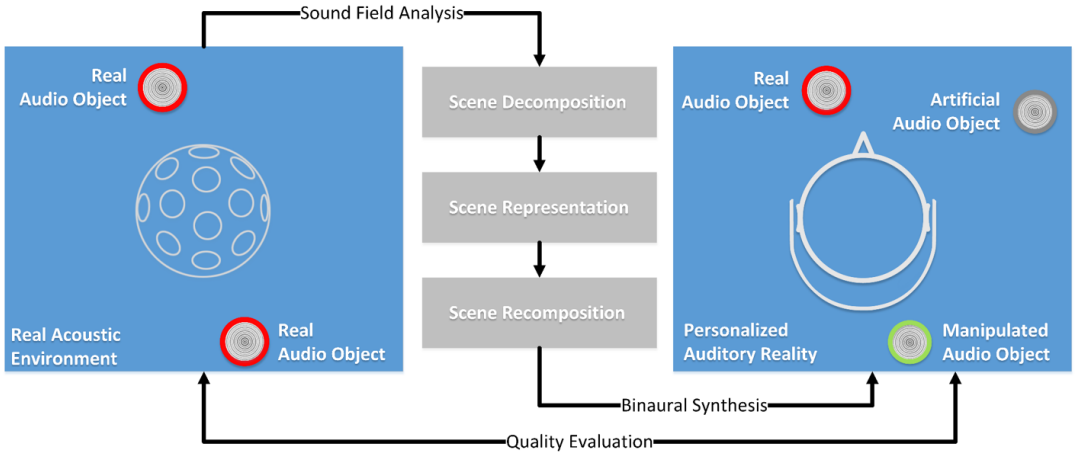

Figure 2: The framework of a processing chain and testing environment.

arrays [3]. Auralizing the recorded data with loudspeakers or headphones allows the implementation of virtual acoustic realities [9]. Currently, practical application require a high number of microphones to achieve an adequate reproduction quality [1]. After recognizing the position of audio objects, the separation of the objects has to be increased by methods known from the Music Information Retrieval research. For example, such methods are successfully used for upmixing mono recordings [6]. Throughout the years, many methods have been proposed to solve separation problems, e.g., filtering, factorization, sinusoidal modelling and sparse representation techniques. With deep neural networks [7] further advances in sound separation are expected. However, the quality evaluation of such methods is challenging, because current quantitative metrics do not correlate well with human perception [4].

In order to meet the requirements of Personalized Auditory Realities, methods for high-resolution 3D sound field analysis have to be combined with advanced sound source separation methods. Another necessity for the analysis of the acoustic environment is an internal model of the actual room. This enables to fit additional virtual sound sources into the real room later. Again, machine learning technologies will be used to get a good enough internal model of the acoustic surroundings.

Scene Representation

In the second step, all objects have to be represented in a reliable way, which includes the collection of meta data about the objects. Therefore, the decomposed audio has to be analysed and recognized in terms of "what they are". Additionally to the objects' meta data, acoustic properties of the environment, necessary for object manipulation and scene recomposition, need to be collected. One of the key challenges is the assignment and combination of sound sources to an actual object. For example, a person can speak and make walking noises at the same time, but both sound sources have to be assigned to this person.

MPEG started activities towards a new set of standards, MPEG-I, addressing coding of immersive spatial audio and video applications. The standard aims at the representation of object based audio, separating important foreground objects and not so important background objects for high coding efficiency. It is expected that Personalized Auditory Realities make use of advances made there and add problem specific tools and encodings.

Scene Recomposition

The last step reproduces the desired Personalized Auditory Reality by merging the decomposed and manipulated sound objects with the real acoustic environment and additional artificial sound objects. Two key challenges have to be addressed: First, real-time synthesis and on-the-fly manipulation of the audio objects have to be guaranteed. Second, the real acoustic environment has to be merged with manipulated and artificial sound objects while maintaining a convincing auditory illusion.

Binaural Rendering

Current binaural rendering methods lack the ability to deliver a convincing auditory illusion, especially when real and virtual sound sources are mixed [2, 10]. Bringing together sources from different origins has to account for the different auditory background they convey: most likely these sources are recordings from different rooms with different characteristics, taken with different microphones and so on. Bluntly mixing those sources without accounting for those details will most likely result in an implausible illusion. To mitigate these effects and allow the mixing of arbitrary sources, room parameters must be estimated and adapted.

Even if the sources are from a controlled set with known room parameters an adaptation is needed: In the end, the resulting signal must match the acoustic surrounding of the user of the system.

Perception and Quality

Figure 2 summarizes the building blocks of Personalized Auditory Realities presented so far. Paramount to all those steps is a quality assessment – both individually for each component and towards the overall quality of experience.

Based on listening experiments assessing the perceived quality of certain components, a suitable vocabulary for assessing the overall quality of Personalized Auditory Realities must be found, based on established vocabularies for virtual environments, such as SAQI (Spatial Audio Quality Inventory, [8]). Building on that, perceptual models describing the interlinking of the building blocks within the proposed system must be found, as there are currently no models describing the mixture of real and virtual acoustic environments.

Applications

Starting from the application depicted in figure 1 where a user is able to arbitrarily control audio objects in his surroundings, there are many application areas that benefit from techniques brought by Personalized Auditory Realities and thereby making viable business cases:

Communications Conferencing systems will profit tremendously by placing "virtual" remote participants in the same room as the real participants.

Entertainment, Arts, Education Personalized Auditory Realities allow the creation of immersive mixed reality applications, with virtual sound sources attached to real objects.

Traffic and Mobility The removal of distracting sound sources and adding desirable sounds might add comfort and increase security for passengers and pedestrians.

Medical Applications Personalized auditory environments can be used as a training help for people relying on hearing aids but also as an additional acoustical guideline for blind people, helping them navigate in their surroundings.

This wide range of applications suggests that besides the focus on signal processing and psychoacoustics, there is a need for novel input devices allowing an intuitive control over auditory realities that poses new challenges on interface design and ergonomics.

The biggest obstacles towards the realization of personalized acoustic realities are the requirements for a very low system delay to allow the seemingly simultaneuos overlay of real and virtual acoustic scenes. The requirement of real time processing calls for efficient DSP implementations, especially in the scene decomposition and scene adjustment parts. Depending on the use case there will be constraints in the size and energy usage of the hardware.

Summary & Conclusion

We presented a new research field, bringing mixed realities to the audio world with a strong focus on interactiveness: Personalized Auditory Realities. To make this possible, advances in three fields must be brought together:

1. The analysis of a real recorded scene and the assignment of parts of the signal to a trackable source.
2. The representation of the results of this scene analysis that allows for an easy distribution and further processing and mixing with other (virtual) sources.
3. Bringing the results of this process back to the users' ears, merging the real world and the enhancements made in the Personalized Auditory Reality.

We believe that the interlinking of the above techniques results in systems of high social impact and relevancy, acting as an enabling technology for novel ways of interactive education, helping impaired people, making people more productive in their daily life or simply providing a new tool to the entertainment sector. Personalized Auditory Realities will be a defining part of the now video-focused VR ecosystem.

References

[1] A. Avni et al. "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution". In: *Journal of the Acoustical Society of America* 133.5 (2013), pp. 2711–2721.

[2] K. Brandenburg et al. "Auditory illusion through headphones: History, challenges and new solutions". In: *22nd International Congress on Acoustics (ICA)* (2016).

[3] M. Brandstein and D. Ward. *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin: Springer, 2001. ISBN: 978-3-642-07547-6.

[4] E. Cano, D. FitzGerald, and K. Brandenburg. "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics". In: *24th European Signal Processing Conference (EUSIPCO)* (2016), pp. 1758–1762.

[5] M. Cohen, S. Aoki, and N. Koizumi. "Augmented audio reality: telepresence/VR hybrid acoustic environments". In: *Proceedings of the 2nd IEEE International Workshop on Robot and Human Communication* (1993), pp. 361–364.

[6] D. FitzGerald. "Upmixing from mono - A source separation approach". In: *17th International Conference on Digital Signal Processing (DSP)* (2011).

[7] E. M. Grais, M. U. Sen, and H. Erdogan. "Deep neural networks for single channel source separation". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2014), pp. 3734–3738.

[8] A. Lindau et al. "A Spatial Audio Quality Inventory (SAQI)". In: *Acta Acoustica united with Acoustica* 100 (2014), pp. 984–994.

[9] J. Nowak and S. Klockgether. "Perception and prediction of apparent source width and listener envelopment in binaural spherical microphone array auralizations". In: *The Journal of the Acoustical Society of America* 142.3 (2017), pp. 1634–1645.

[10] S. Werner and F. Klein. "Influence of Context Dependent Quality Parameters on the Perception of Externalization and Direction of an Auditory Event". In: *Proceedings of the 55th AES Conference* (2014), paper no. 6–4.