DOI: 10.61782/fa.2023.0682

# SELF-ADAPTIVE MIXING SYSTEM FOR SPATIAL AUDIO IN VIRTUAL REALITY

**Xiyao Jin**[1] **Zhaorui Liu**[2] **Xiaorong Shao**[3] **Haiyan Li**[1*] **Zijin Li**[4]

[1] Communication University Of China, Beijing, China
[2] China Conservatory of Music, Beijing, China
[3] Shanghai Conservatory of Music, Shanghai, China
[4] Central Conservatory of Music, Beijing, China

## ABSTRACT

The application of spatial audio in VR plays an important role in creating a sense of immersion. However, there are still some challenges in using spatial audio in VR: It is hard to design a purely virtual soundscape in VR. And if the virtual content is out of the ordinary, it may cause auditory disharmony and conflicts among sound sources if spatial audio design is only based on the purpose of realism. This study establishes a prototype of VR spatial audio mixing system, which sound elements can be adaptively adjusted according to the content viewed by users. In order to verify the actual effect of the system, the study further conducted corresponding user experiments, evaluating the perception of spatial audio in a VR prototype from five dimensions: "Naturalness", "Presence", "Preference", "Localization", and "Source Envelopment", and carried out structured interviews. The experiment showed that the application of the system could effectively improve the user's perception effect of VR spatial audio and bring a better sound experience. Based on this research, we hope to further develop an intelligent spatial audio system that can autonomously adjust and process sound effects based on the viewing behavior of VR users in the future.

**Keywords:** *spatial audio, sound design, virtual reality, soundscape*

## 1. INTRODUCTION

The sound experience has always been a key aspect to consider in VR content design. Currently, research on sound experience design in VR environments often focuses on aspects such as realistic acoustic environments and efficient sound rendering. A significant amount of research has focused on how to create spatial audio environments in VR [1,2].

However, there were also a series of problems of using spatial audio techniques in VR. The typical problem is that virtual sound sources may not be perceived as coming from the intended location, leading to a less immersive experience [3]. Or the accuracy of listening will decrease as the number of sound sources in the virtual environment increased [4]. It can be observed that the main problem of using spatial audio in VR is the issue of orientation. In addition to solving the problem by optimizing the algorithms, helping users locate the points of listening interest by proper cues is also a way to improve the situation [5].

In traditional film productions, directors rely on camera shots to set special cues and match sonic information to visuals. However, the way viewers watch VR undermines this typical sound-image pairing. Viewers can freely shift their viewpoints and pay attention to what interests them. Furthermore, when VR aims to display content that differs from real-life scenarios, creating a comfortable experience may be challenging if sound design just focuses solely on producing a realistic auditory sensation of space.

To address these issues, this study designs a VR adaptive spatial audio mixing system centered around the VR viewer's attention. The system takes the relative findings of "cocktail party effect"[6] and soundscape theory as the basis

[*] ***Corresponding author***: *hyli@cuc.edu.cn*

for the design framework, and proposes a corresponding adaptive mixing scheme based on user viewing behavior for different sound types. The prototype system design used Max/MSP, and we applied it and evaluated its actual utility in a specific VR experiment. We provided two VR experiences, the visual content of VR is identical, one with a real-time mixing system for VR spatial audio turned on, and another turned off. From the self-rating scale questionnaire and structured interviews, the results show that the adaptive mixing system provided an improved evaluation for the VR sound experience. This study has laid the foundation for designing more complex sound systems in the future. And we aim to design a more intelligent system that can be applied to other specific virtual reality scenarios based on this groundwork.

## 2. RELATIVE WORKS

For the design and creation of VR image content, sound design is an important factor in creating immersion and enhancing the sense of presence in the virtual environment. Specialized sound design can improve the feeling of being in a specific place [7]. Traditional film sound design emphasizes the "added-value" of sound to the image, which means that sound can enrich a given image to produce a specific impression [8]. The so-called "audio-visual language" of films involves creating the desired aesthetic effect of a director through a combination of visual and auditory content. Early sound design for VR content was mainly focused on the challenge of achieving realistic listening experiences. But current challenges in this field include finding a balance between accuracy and plausibility in sound simulations, improving the efficiency of sound rendering algorithms, and addressing the issue of listener-specific sound perception [9].

The soundscape theory that guides sound design in traditional film production has also been considered as a powerful scaffold in VR sound design. The challenges of soundscapes and virtual worlds include the difficulty in choosing how the soundscape is delivered, as there is a gap between what can be programmed or recorded and what is actually perceived by the user through arrays of speakers or headphones [10]. But virtual soundscape also provide the opportunity to make real-time adjustments to sound elements based on user needs and states, which can lead to improved experiences. Further research had been conducted on sound design for virtual environments, utilizing adaptive or generative techniques. This study showed that sound in a virtual environment was dynamic and responsive to changes in the environment, enhancing the listener's engagement and immersion in the sonic environment [11]. Furthermore, related studies adopted a method of selective masking of certain sounds in the acoustic environment to enhance the acoustic atmosphere. The effects of the “masker” in that paper were evaluated to determine the perceived improvement in pleasantness of the soundscape in the presence of pleasing sounds [12]. And a study also described the development of an interactive binaural soundscape that responds to user 3D displacement [13]. What can be observed is that we could actively adjust and manipulate the auditory experience in VR through the design of dynamic and variable soundscapes.

For immersive experience, the spatial audio system is the main technology used in virtual reality sound design. There is currently a large amount of research focusing on how to arrange spatial audio system and render spatial audio in VR experience scenarios. For example, a study provided a method for manipulating sound and music in virtual space using spatial audio through VR head-mounted display and hand-held motion controller [14]. Further research had also been conducted on a system to estimate room acoustic for plausible reproduction of spatial audio using CNN to estimate a 360° image's room geometry. The reconstructed scenes are rendered with synthesised spatial audio as VR/AR content [15]. Furthermore, related studies tested the behavioral response of children with autism spectrum disorder to spatial audio in a multimodal VR environment and created a safe platform for treating symptoms related to this condition by using spatial audio rendering technology in VR [16].

From the perspective of artistic creation, merely simulating acoustic environment for VR content may not necessarily achieve the desired experiential effect. Moreover, strictly following the realistic logic for sound simulation in the VR experience of displaying content out of the ordinary may instead lead to disharmony or "distortion" in sound experience. Based on the commonly used spatial audio technique in VR audio-visual design, we hope to integrate the experiential characteristics of spatial audio with a real-time soundscape adjustment approach, and design a VR spatial audio adaptive mixing system which can adjust the sound in real-time based on the user's viewing actions.

## 3. SELF-ADAPTIVE MIXING SYSTEM DESIGN FOR SPATIAL AUDIO IN VR

### 3.1 Psychoacoustic Fundations and Design Framework

The cocktail party effect refers to the ability of the human sense of hearing to extract a specific target sound source from a mixture of background noises in complex acoustic scenarios [17]. It has revealed the abilities that people can make a perceptual separation between a signal sound and a competing sound. Some early applications like "Visual Resonator"[18] provided an interactive realization of the cocktail party phenomenon, which allowed the user to hear a voice or auditory information only from the direction in which they are facing and send their voice only in the direction towards which they are facing. One major explaination of cocktail party effect is that the difference in sound level between the two ears allows the brain to separate speech from other sounds [19]. The finding provides a strategy for mixing sounds in VR: During the changes of user's field of view and the movement of head, artificially altering the frequency response range of various audio sources in spatial audio can enhance a user's perception of specific sound sources, thereby creating an auditory attention. In other words, we can create an active cocktail party effect in VR environment.

Auditory attention is the mechanism that allows us to focus on specific sounds while ignoring others in the environment [20]. Examples of sensitivity control include directing eye movement or changing the orientation of the head can effectively influence the auditory attention. According to the theory of soundscape, there are three types of sounds included in a soundscape: keynote sounds, signals, and soundmarks [21]. The categorization of sounds based on Soundscape provides a basis for sound content designers to organize and arrange sound sources. By analyzing the features of different sounds, we can sort out sound elements in VR environment, and the connection between sounds and virtual objects can also be established. Based on that connection, the system can adaptively adjust the sound according to the user's focus point and attention weight in the auditory channel by obtaining the virtual objects in the user's field of vision at a certain moment. This implies that we can artificially enhance the cocktail party effect in auditory experiences, thereby prompting users to develop a stronger sense of auditory attention. Starting from this concept, we can start building the basic framework of the VR spatial audio adaptive mixing system, which can dynamically adapt to the user's watching behavior, the process is shown in Fig. 1.

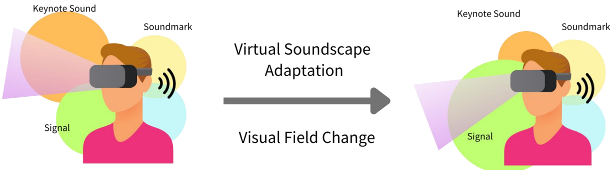

**Figure 1**. Soundscape adaption in virtual environment.

### 3.2 The Prototype of the Self-adaptive Mixing System

Based on the basic framework of the adaptive mixing system mentioned above, this study used tools such as Oculus Quest 2, SteamVR, Godot Engine 3.5, and Max/MSP to design a prototype of the system, the system structure is shown in Fig. 2. The VR content was developed in Godot Engine 3.5, with a focus on displaying spectacular and hyper-realistic scenes set in an ocean environment. The scene includes multiple fixed-position sound sources, as well as real-time moving sound sources around the center of the user: four static light balls with metal rings distributed equally around the user's perspective; purple fish moving continuously around the user, serving as attractors for the blue fish swarm's collective movement. The position of these virtual objects in the space, the distance and angle data between the objects and the user (virtual world origin), and user's head rotation angle are sent to Max/MSP through the OSC protocol for further processing, which are set as control parameters for real-time adjustment of spatial audio. In Max/MSP, the system uses externals ICST Ambisonics Tools to encode the adaptive-mixed audio sources into Ambisonics signals and decode them into binaural audio for monitoring through headphones.

The adaptive mixing system is mainly realized through the viewport object detection program in Godot Engine and the real-time audio processing program in Max/MSP. With the adaptive system turned on, Godot Engine will continuously detect whether there are light balls and attractors in the VR viewport, as well as the distance and angle parameters of each virtual object relative to the viewport's location, and then send the detection data to Max/MSP via the OSC protocol. While rendering spatial audio, the program will adaptively mix the sound according to whether objects appear in the viewport and the type of corresponding sound source. When the adaptive system is turned off, the audio objects corresponding to fish schools and light balls will only be rendered at the corresponding Ambisonics sound image positions, and fish schools and light balls in virtual space will be allocated corresponding basic sounds.

The fish and attractors within the scene serve as the core visual content, and sound is dynamically adjusted through an adaptive mixing method of the soundmark. The Max/MSP program detects the appearance of attractors every four seconds within the VR viewport, and generates a random Chinese National Pentatonic Scale melody of four seconds in length based on the quantity of attractors present within the viewport. The more attractors that appear in the

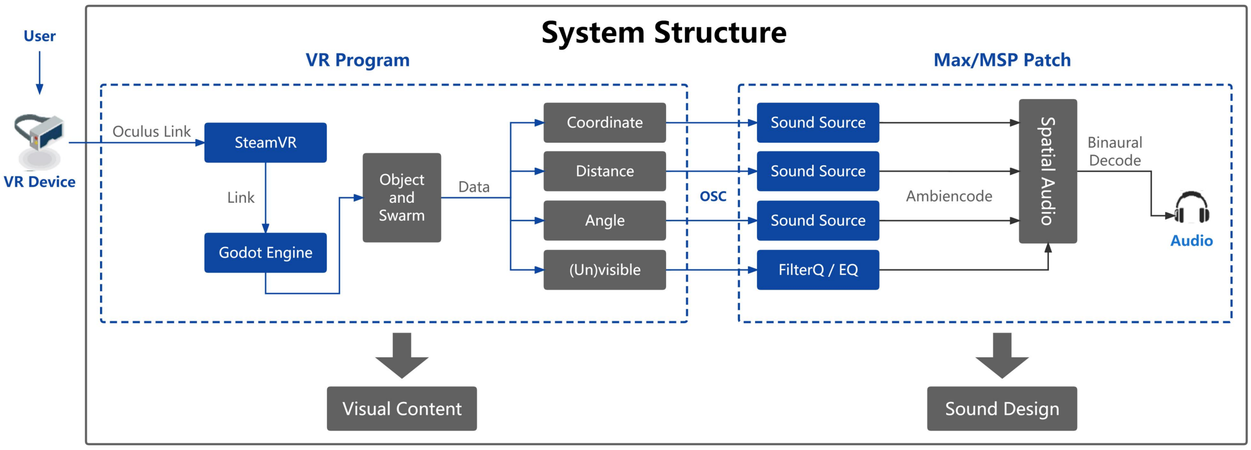

**Figure 2**. Technological realization of self-adaptive mixing system.

user's field of view, the more notes are generated within the 4-second span, resulting in a faster rhythm and drawing additional attention from the user.

The sound corresponding to the light ball in the scene is used for auditory reminders and to attract the attention of users in conjunction with the visual focus. Therefore, the light ball adjusts dynamically according to the self-adaptive mixing method of the signal sound. In Max/MSP, the program no longer limits the frequency of user's viewport detection, and the sound of the light ball will be mixed through a bandpass filter. When an object appears within the viewport, the Q value of the filter is at its minimum, and the frequency attenuation within the frequency domain range of the sound source is not obvious, making it more prominent in terms of auditory perception. When the object moves out of the viewport range, the Q value of the filter linearly increases, and the frequency outside of the cutoff frequency range of the light ball sound source is significantly attenuated, and there is no "auditory focus" effect. The screen in VR is shown in Fig. 3.

In addition to the sound sources corresponding to the core visual objects mentioned above, a keynote sound that reveals environmental features is added to the VR scene. A looped sound sample of ocean flow variations is used here, and an adaptive adjustment method is used to match the keynote sound. When the visual content in the user's viewport is limited, the system will increase the volume of the keynote sound appropriately and release some filtering frequencies to balance the overall auditory experience.

## 4. USER EXPERIENCE EVALUATION

### 4.1 Experiment design and process

To evaluate the actual effectiveness of the spatial audio processing system mentioned above, this study conducted a user experience evaluation experiment based on the design of the prototype system. The experiment used a within-subjects design method. The users in experience were 23 university students aged between 20 and 28 years old. The VR device used in this experiment was Oculus Quest2, and the earphones were Sony WH-1000XM3. The subjects observed from a first-person perspective, and the content of the VR display was the same for both groups, with the experience order of the two schemes randomly assigned. Group A1 used the previously described system for adaptive spatial audio mixing, while Group A2 used only ordinary, unprocessed spatial audio. Every subject experienced the system for 2 minutes each time, and after each experience, the subject filled out a sound experience evaluation questionnaire to self-assess their perception of the spatial audio effect in VR. After the two experiments, the experimenters conducted structured interviews with the subjects to collect qualitative data on their perceptions of the two experiences.

### 4.2 The results of the questionnaire

In terms of sound experience evaluation, we chose five core dimensions to evaluate spatial audio: Naturalness, Presence, Preference, Localization, and Source Envelopment [22].

**Figure 3**. The screen in VR.

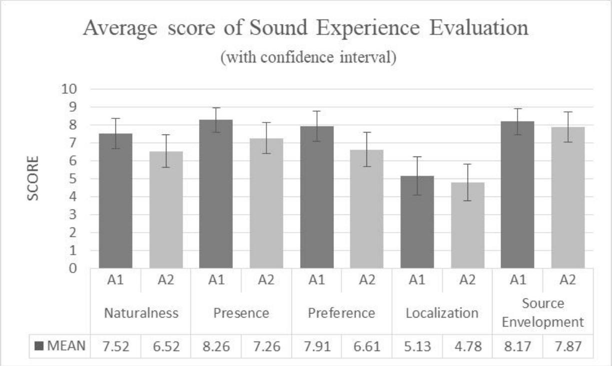

**Figure 4**. Average score of sound experience evaluation.

Each dimension was evaluated using a 10-point scale. After data analysis, the average scores and corresponding confidence intervals of the A1 and A2 groups' experiences can be obtained as shown in Fig. 4. From the data results, the experience scores of the processed spatial audio group were better than those of the unprocessed group. Therefore, it can be concluded that the design framework of adaptive processing for spatial audio mentioned earlier has a substantial role in enhancing the experience of spatial audio in VR environment.

### 4.3 The results of the Structured interviews

To further evaluate the actual perception of the subjects towards two sets of VR spatial audio, a structured interview was conducted after the experience, including three aspects: "differences in sound experience," "sound experience preferences," and "immersion." A total of 23 qualitative materials were collected from the interviews, from which we expect to obtain evidence of the improvement of spatial audio experience. The three groups of questions are shown below:

*Q1: Could you notice the difference in the sound between the two virtual environments during the two experiences? If so, can you please describe these differences in detail?*
*Q2: Which of the two virtual reality sound experiences did you find more enjoyable, and why? Or, which one did you find less enjoyable, and what were your reasons for feeling this way?*
*Q3: In your opinion, which of the two experiences provided a stronger sense of immersion? Can you please talk specifically about why?*

In response to the interview results, we conducted further coding and theme analysis. For the first question, four topics can be summarized from group A1 and group A2 (as shown in Tab. 1). From the analysis results, it can be seen that since the sound of the two groups of VR is spatial, there is a description of the perception of the position change of the sound source in both sets of experiences. However, the treated A1 group significantly had more positive depictions, and only in the A1 group did participants report hearing the ambient sound of "ocean waves", and most participants thought that the A1 group had richer sound levels and more comfortable hearing. In the A2 group, the most described word was "Boring", and although the pitch setting of the A2 sound was no different from that of the A1 group, many participants reported that the A2 group sounded higher and sharper.

For the second question, 17 participants reported better feelings towards Group A1, 4 participants thought that the A2 group was better. Furthermore, 2 participants reported no clear preference. The reason statement basically coincided with the expression of the first question. For the last question, the participants' preference for immersion was not completely consistent with the answer to the second question, with 14 participants believing that the A1 group had a stronger sense of immersion, and 9 participants thought that the A2 group had a stronger sense of immersion. Two of the participants said that although the experience in the A2 group was less comfortable, it was more emotionally impactful and therefore more immersive. Additionally, 2 participants reported that their assessment of immersion may have been influenced by the order in which they experienced the groups, but it should be noted that the orders of these participants were not the same, making it impossible to confirm if immersion preference was related to experience order. The majority of participants believed that spatial audio processed through adaptive mixing better matched the atmosphere portrayed in the VR environment.

An analysis of comprehensive qualitative materials shows that when adaptive mixing spatial audio is used in VR

**Table 1.** Themes of the interview results.

| Group | Themes | Examples |
|---|---|---|
| **A1** | Melody | "There are differences in pitch, variations in melody." |
| | Comfortable | "Pleasant, relaxed atmosphere." |
| | Environmental sound | "The sound of the sea could be heard." |
| | Panoramic sound | "Can feel the change in the position of the sound source." |
| **A2** | Monotone | "It sounds monotonous and repetitive." |
| | High tone | "The pitch is higher and sharper." |
| | Uncomfortable | "Feeling uneasy and eerie." |
| | Panoramic sound | "The sound has a three-dimensional surround feeling." |

experiences, it can significantly enhance the experience for most participants and aligns with our initial expectations for aesthetic experiences. Participants can perceive a more varied sound hierarchy and are more sensitive to changes in the spatial location of sound sources. At the same time, the spatial audio using this system can also enhance the immersion of the participants in the VR experience, making the viewing process more interesting and reduce the sense of boredom.

## 5. CONCLUSION

Aiming at the artistic expression demands of audio-visual coordination in VR, this study has designed a system that can perform real-time mixing of spatial audio according to the viewing content of VR users and audiences, using design framework based on the soundscape's classification and the cocktail party effect in psychoacoustic. The system divides the spatial audio sounds in VR scenes into corresponding adaptive mixing mechanisms based on the classification of "Keynote Sound", "Signal" and "Soundmark". And the frequency response range of various sound elements can be dynamically adjusted according to user behavior. The basic prototype of the system was built using Godot Engine 3.5 and Max/MSP. To verify the actual effectiveness of the system in VR experience, this study designed an A/B experiment within two groups, and evaluated the VR sound experience from five aspects: naturalness, presence, preference, localization, and source envelopment, also with structured interviews conducted simultaneously. The results of the experiment questionnaire showed that the adaptive mixing system provided participants with a better auditory experience. And the structured interviews indicated that most participants believed that the adaptive mixing system was more in line with the current VR visual content, thus achieving the expected aesthetic experience.

## 6. DISCUSSION

As a sound processing technology that shapes the realism of sound space, "spatial audio" can shape the sense of real space, providing better auditory content for VR experience. Through the design and user experience evaluation of the VR spatial audio adaptive mixing system described in this article, we emphasize the design idea of re-creation of sound experience based on spatial audio technology: the content creation of virtual reality originates from the simulation of real environment, but it is not a complete virtual reproduction of the real environment. Instead, it can be based on the simulation of basic sensory experience in reality and artfully processed the elements of experience. This paper points out that VR content designers can also shape the sound experience of VR content to achieve corresponding artistic expression. And due to the interactive nature of VR experience, the sound experience is also affected by the user's viewing behavior in VR, so we need to design a sound system that can adapt to the changes of user behavior.

The adaptive mixing system designed in this study still cannot truly understand users' viewing behaviors from a "semantic" perspective. The ideal system should be able to do so. Currently, some related studies have used machine learning algorithms to assist in constructing soundscapes [23]. It is believed that in the future, we can use AI-related algorithms to design adaptive mixing systems with true "context-awareness" capabilities to better match the user's viewing behavior and the designer's design intent.

## 7. ACKNOWLEDGMENTS

Funded by: the Fundamental Research Funds for the Central Universities (CUC23GZ006); Key Laboratory of Intelligent Processing Technology for Digital Music (Zhejiang Conservatory of Music), Ministry of Culture and Tourism (2022DMKLB003).

## 8. REFERENCES

[1] S. Yong, H.C. Wang, "Using spatialized audio to improve human spatial knowledge acquisition in

virtual reality," in *Proc. of the 23rd International Conference on Intelligent User Interfaces Companion*, pp. 1-2, 2018.

[2] S. Bhide, E. Goins, J. Geigel, "Experimental analysis of spatial sound for storytelling in virtual reality," in *Proc. of 12th International Conference on Interactive Digital Storytelling*, (Little Cottonwood Canyon, UT, USA), pp. 3-7, 2019.

[3] T.C. Tanner, J.P. Lester: *System and method for localization of virtual sound.* U.S. Patent No.6,307,941. 2001.

[4] D.L. Guettler, R.S. Bolia, W.T. Nelson: "Monitoring and localizing simultaneous real-world sounds: Implications for the design of spatial audio displays," *The Journal of the Acoustical Society of America*, vol. 108, no.5, pp. 2573-2573, 2000.

[5] P. Bala, R. Masu, V. Nisi, et al, "'When the Elephant Trumps' A Comparative Study on Spatial Audio for Orientation in 360º Videos," in *Proc. of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-13, 2019.

[6] I. Pollack, J.M. Pickett: "Cocktail Party Effect," *Journal of the Acoustical Society of America*, pp. 1262-1262, 1957.

[7] G. Serafin, S. Serafin, "Sound design to enhance presence in photorealistic virtual reality," *Georgia Institute of Technology*, 2004.

[8] M. Chion: L'audio-vision-5e éd: Son et image au cinéma. *Armand Colin*, 2021.

[9] S. Serafin, M. Geronazzo, C. Erkut, et al.: "Sonic interactions in virtual reality: State of the art, current challenges, and future directions," *IEEE computer graphics and applications*, vol. 38, no. 2, pp. 31-43, 2018.

[10] C. Rajguru, M. Obrist, G. Memoli: "Spatial soundscapes and virtual worlds: challenges and opportunities," *Frontiers in Psychology*, vol. 11, pp. 569056, 2020.

[11] M. Nazemi, D. Gromala, "Sound design: A procedural communication model for VE," in *Proc. of the 7th Audio Mostly Conference: A Conference on Interaction with Sound*, pp. 16-23, 2012.

[12] Z.T. Ong, J.Y. Hong, B. Lam, et al., "Effect of masker orientation on masking efficacy in soundscape applications," in *Proc. of INTER-NOISE and NOISE-CON Congress and Conference*, pp: 4916-4922, 2017.

[13] I. Batista, F. de Paula Barretto, "Developing an Synthetic Binaural Interactive Soundscape Based on User 3D Space Displacement Using OpenCV and Pure Data," in *Proc. of 20th HCI International 2018–Posters' Extended Abstracts*, (Las Vegas, NV, USA), pp: 231-236, 2018.

[14] D. Jordan, F. Müller, C. Drude, et al.: "Spatial audio engineering in a virtual reality environment," *Mensch und Computer*, 2016.

[15] H. Kim, L. Remaggi, P.J.B. Jackson, et al., "Immersive Spatial Audio Reproduction for VR/AR Using Room Acoustic Modelling from 360° Images," in *Proc. of 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp: 120-126, 2019.

[16] D. Johnston, H. Egermann, G. Kearney: "Measuring the behavioral response to spatial audio within a multi-modal virtual reality environment in children with autism spectrum disorder," *Applied Sciences*, vol. 9, no. 15, pp: 3152, 2019.

[17] T. Fischer, M. Caversaccio, W. Wimmer: "Multichannel acoustic source and image dataset for the cocktail party effect in hearing aid and implant users," *Scientific data*, vol. 7, no. 1, pp: 440, 2020.

[18] J. Watanabe, H. Nii, Y. Hashimoto, et al., "Visual resonator: Interface for interactive cocktail party phenomenon," in *Proc. of CHI'06 Extended Abstracts on Human Factors in Computing Systems*, pp: 1505-1510, 2006.

[19] A.W. Bronkhorst: "The cocktail party effect: Research and applications," *The Journal of the Acoustical Society of America*, vol. 105, no. 2, pp: 1150-1150, 1999.

[20] D. Oldoni, B. De Coensel, M. Boes, et al.: "A computational model of auditory attention for use in soundscape research," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp: 852-861, 2013.

[21] R.M. Schafer: *The soundscape: Our sonic environment and the tuning of the world*, Simon and Schuster, 1993.

[22] J. Berg, F.Rumsey, "Systematic evaluation of perceived spatial quality," in Proc. of AES International Conference: Multichannel Audio, The New Reality, pp: 184-198, 2003.

[23] M. Thorogood, J. Fan, P. Pasquier: "A framework for computer-assisted sound design systems supported by modelling affective and perceptual properties of soundscape," Journal of New Music Research, vol. 48, no. 3, pp: 264-280, 2019.