Lorenzo Picinali, Brian FG Katz, Michele Geronazzo, Piotr Majdak, Arcadio Reyes-Lecuona, and Alessandro Vinciarelli

The SONICOM Project: Artificial Intelligence-Driven Immersive Audio, From Personalization to Modeling

Every individual perceives spatial audio differently, due in large part to the unique and complex shape of ears and head. Therefore, high-quality, headphone-based spatial audio should be uniquely tailored to each listener in an effective and efficient manner. Artificial intelligence (AI) is a powerful tool that can be used to drive forward research in spatial audio personalization. The SONICOM project aims to employ a data-driven approach that links physiological characteristics of the ear to the individual acoustic filters, which allows us to localize sound sources and perceive them as being located around us. A small amount of data acquired from users could allow personalized audio experiences, and AI could facilitate this by offering a new perspective on the matter. A Bayesian approach to computational neuroscience and binaural sound reproduction will be linked to create a metric for AI-based algorithms that will predict realistic spatial audio quality. Being able to consistently and repeatedly evaluate and quantify the improvements brought by technological advancements, as well as the impact these have on complex interactions in virtual environments, will be key for the development of new techniques and for unlocking new approaches to understanding the mechanisms of human spatial hearing and communication.

Digital Object Identifier 10.1109/MSP.2022.3182929
Date of current version: 27 October 2022

Introduction

Immersive audio is what we experience in our everyday life, when we can hear and interact with sounds coming from different positions around us. We can simulate this interactive auditory experience within virtual reality (VR) and augmented reality (AR) using off-the-shelf components such as headphones, digital signal processors, inertial sensors, and handheld controllers. Immersive audio technologies have the potential to revolutionize the way we interact socially within AR/VR environments and applications. But several major challenges still need to be tackled before we can achieve sufficiently high-quality simulations and control. This will involve not only significant technological advancements but also measuring, understanding, and modeling low-level psychophysical (sensory) as well as high-level psychological (social interaction) perception.

Funded by the Horizon 2020 FET-Proact scheme, the SONICOM project (www.sonicom.eu) started in January 2021 and, over the course of the next five years, will aim to transform auditory social interaction and communication in AR/VR by achieving the following objectives:

- It will design a new generation of immersive audio technologies and techniques, specifically looking at customization and personalization of the audio rendering.
- It will explore, map, and model how the physical characteristics of spatialized auditory stimuli can influence observable behavioral, physiological, kinematic, and psychophysical reactions of listeners within social interaction scenarios.
- It will evaluate the techniques developed and data-driven outputs in an ecologically valid manner, exploiting AR/VR simulations as well as real-life scenarios.
- It will create an ecosystem for auditory data closely linked with model implementations and immersive audio-rendering components, reinforcing the idea of reproducible research and promoting future development and innovation in the area of auditory-based social interaction.

Overview

SONICOM involves an international team of 10 research institutions and creative tech companies from six European countries, all active in areas such as immersive acoustics, AI, spatial hearing, auditory modeling, computational social intelligence, and interactive computing. The workplan is centered around three pivotal research work packages titled "Immersion," "Interaction," and "Beyond," each introduced in one of the following sections. The first looks at

immersive audio challenges dealing with the technical and sensory perspectives. On the other hand, the second focuses on the interaction between these and higher-level sociopsychological implications. Finally, the integration of core research, proof-of-concepts evaluations, and creation of the auditory data ecosystem ensures that various outputs of the project will have an impact beyond the end of SONICOM (see the "Beyond" section).

Immersion

Before reaching the listener's eardrums, the acoustic field is filtered due to shadowing and diffraction effects by the listener's body, in particular, the head, torso, and outer ears. This natural filtering depends on the spatial relationship between the source and the listener and can be described by head-related transfer functions (HRTFs), which can be acoustically measured (e.g., see Figure 1) or numerically modeled (e.g., in [1]). Everyone perceives sound differently due to the particular shape of their ears and head. For this reason, high-quality simulations should be uniquely tailored to each individual, effectively and efficiently. Within SONICOM we propose a data-driven approach linking the physiological characteristics of the ear to the individual acoustic filters employed for perceiving sound sources in space.

Our HRTF modeling research considers a variety of approaches. On the one hand, we focus on the creation of parametric pinna models (PPMs) [2], [3] and their application to create an AI-based framework for the numerical calculation of HRTFs. On the other hand, we focus on HRTF database matching, an approach based on the hypothesis that individuals can be paired with existing high-quality HRTF data sets (measured or modeled) as long as they share some relevant, predefined characteristics in the perceptual features space. To this end, we will expand the procedures based on objective similarity measures [4] and subjective listener input [5], [6]. The said measures concern geometrical variations for PPMs, perceptual deviations of the computed HRTFs, and signal-domain similarities for HRTF matching, all referenced to a project database comprising geometrical scans and associated HRTF measurements from a set of individuals.

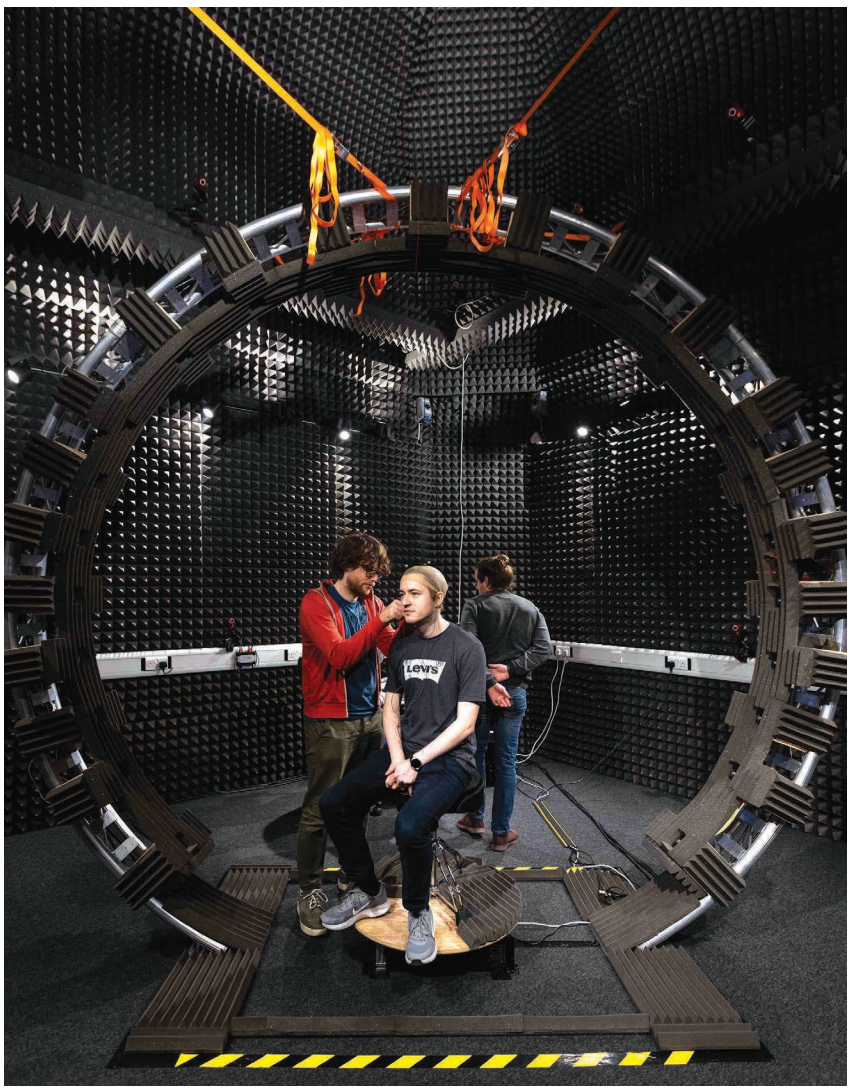

FIGURE 1. The HRTF measurement setup at Imperial College London, U.K. [24].

Being able to consistently and repeatably evaluate and quantify the improvements brought by these technological advancements will be absolutely key, not only for the development of new techniques but also for unlocking new approaches to understanding the mechanisms of human spatial hearing. Our approach to personalization presents a new perspective on this problem, linking Bayesian theories of active inference [7], [8] and binaural (i.e., related to both ears) sound reproduction to create data sets of human behavior and perceptually valid metrics modeling such behavior. By having acoustic simulations validated against acoustic measurements, and human auditory models validated against actual behavior, we will provide important tools for the development of AI-based predictors of realistic spatial audio quality.

We will also concentrate on the issue of blending virtual objects in real scenes, which is one of the cornerstones of AR. To blend the real with the virtual worlds in an AR scenario, it is essential to develop techniques for automatic estimation of the reverberant characteristics of the real environment. This will be achieved by characterizing the acoustical environment surrounding the AR user. The extracted data can then be employed to generate realistic virtual reverberation matching the real world. After a set of pilot studies looking at perceptual needs in terms

of reverberation processing (e.g., in [9]), we will employ geometrical acoustics and simplified computational models, such as scattering delay networks [10], [11], to generate real-time simulations of the real-world environment where the listener is located.

Finally, to provide ecologically valid evaluations, studied settings will not be limited to oversimplified, highly controlled, traditional "laboratory" conditions, but will seek to extend the set of evaluation scenarios to better represent a variety of real-world use cases for AR/VR technology. Combining the desire for robust evaluation techniques with realistic use cases requires a delicate balance of experimental design to present a real-world-like context while still maintaining the required laboratory controllability to obtain meaningful, exploitable data (e.g., in [12]).

Interaction

AR and VR technologies work by stimulating the senses of their users and, as a result, most of the previous research has focused on reproducing the sensory experience of the physical world. However, this is not sufficient when a virtual or augmented environment involves interaction with other agents, whether human or artificial.

For example, literature shows that a behavioral cue (smile, gesture, sentence, and so on) stimulates the same range of unconscious reactions whether it is displayed by an artificial agent or by a person [13]. Such a phenomenon, known as a *media equation* [14], can be observed in AR/VR where, in the particular case of speech, it is possible to simulate different distances between artificial speakers and listeners. This is important because social and physical distances are deeply intertwined from a cognitive and psychological point of view [15]. Therefore, it is possible to expect that VR users will tend to attribute different social characteristics (intentions, personality traits, attitudes, and so forth) to speakers rendered at different distances in the physical space. Besides being interesting from a scientific point of view, such a phenomenon can contribute to the design of personalized interaction experiences. For example, it could enable a virtual pet to develop deeper intimacy with children by sounding physically closer or help a virtual character to appear less friendly by sounding more distant. More generally, it will be possible to modulate the perceived distance between VR users and virtual agents according to roles the latter play within immersive and interactive environments. Although having attracted significant attention in recent years, the use of perceived physical distance to interface VR technology with psychology and cognition of its users is still at a pioneering stage (see, e.g., [16] and [17]). Its investigation is a part of what is planned within the SONICOM project and promises to be fruitful from both scientific and technological points of view.

Beyond

Ensuring that the project's accomplishments (algorithms, AI-based models, evaluation data, and so on) remain available to various stakeholders, including the wider research communities beyond SONICOM, is of primary importance. To facilitate this and consolidate all the developed tools within a common structure, the SONICOM Ecosystem will be created, which will include open source software modules implementing the various tools, algorithms, and models developed within the project.

The main part of the SONICOM Ecosystem will be the SONICOM Framework, consisting of the Binaural Rendering Toolbox (BRT), auditory data and models, and dedicated hardware. The rendering core of the BRT is inspired by the work that has already been done on the development of the 3D Tune-In Toolkit [18]. It will be implemented as an interchangeable module, allowing the use of numerous rendering engines for various software platforms, connected with the rendering core modules using interfaces and communication protocols. Once benchmarked and evaluated, the SONICOM Framework will become a part of the SONICOM Ecosystem, which will further include the Auditory Model Toolbox [19], [20] and toolboxes dealing with HRTFs stored in the spatially oriented format for acoustics (SOFA) [21]. SOFA, a standard of the Audio Engineering Society (AES69-2015, [22]), has received a recent upgrade and will be further extended toward the needs of SONICOM and its Ecosystem. A further component of the Ecosystem will be Mesh2HRTF, an application to numerically calculate HRTFs. Originally developed in 2015 [1], it will receive a major upgrade when integrated into the Ecosystem.

Considering the timeline of the project, the core research activities of the "Immersion" and "Interaction" work packages will progress until the first half of 2024, when the work on the SONICOM Framework and Ecosystem will commence. These efforts will be preparatory to the launch of the listener acoustic personalization (LAP) challenge, opening up to researchers across the world to contribute and compete, within various scenarios and tasks, with their state-of-the-art HRTF personalization algorithms. The recently introduced paradigm of the egocentric audio perspective [23] will guide the definition of effective evaluation measures considering the first-person point of view for embodied, environmentally situated perceivers with sensorimotor processes tightly connecting sensorimotor processes with exploratory actions. A publicly released corpus will be an integral part of the Ecosystem and will include AI-driven data, behavioral and HRTF data, and human body scans for a set of listeners. Moreover, a range of real-life scenarios of increasing complexity will be captured by microphone arrays and multimodal sensors to form the ground truth of objects and actions in the scenes. All of this is meant to simulate a digital replica of the complex listener reality system, which will allow us to create virtual/augmented listening experience. Such an integration of SONICOM's outputs aims to promote reproducible research, creating a sustainable basis for further research beyond SONICOM.

Conclusions

Although it is true that a large amount of research has been carried out in recent years looking at solutions to challenges

that are similar to what we are tackling in SONICOM, there are transformative elements within the research we are planning, which could be the key to creating a new generation of immersive audio technologies and techniques. One such element is the use of a data-driven and AI-based approach to HRTF personalization, looking not only at the physical nature of the problem (e.g., ear morphology) but also at the perceptual side of things (e.g., listener preferences and performances). The extensive use of perceptual models also presents a strong element of novelty, using existing ones as a guide during the prototyping and design stages as well as for helping to better understand the experimental research outputs. This will contribute to create new and more accurate models to be shared with the wider research communities. Within this context, the attempt to make use of collected data to model social-level processing within existing sensory models is a novel element and will enable better prediction of responses to complex tasks such as speech-in-noise understanding and, more generally, sonic interactions within AR/VR scenarios.

Finally, it seems clear that to ensure an adequate level of standardization and consistently advance the achievements of research in this area, a concerted and coordinated effort across disciplines, research institutions and industry players is absolutely essential, and this is precisely what we are trying to do within SONICOM.

Acknowledgment

The SONICOM project has received funding from the European Union's Horizon 2020 Research and Innovation Program under grant agreement number 101017743.

Authors

Lorenzo Picinali (l.picinali@imperial.ac.uk) is with Audio Experience Design, Imperial College London, London, SW7 2AZ, U.K. His research interests include spatial acoustics and immersive audio, perceptual hearing training, and ecoacoustic monitoring.

Brian FG Katz (brian.katz@sorbonne-universite.fr) is with the Institute d'Alembert, Sorbonne Université, Paris, 75252, France. His research interests include acoustics, HCI, virtual reality, spatial perception, and room acoustics.

Michele Geronazzo (michele.geronazzo@unipd.it), Imperial College London London, SW7 2AZ, and University of Padova, Padova, 35122, Italy. His research interests include binaural spatial audio modeling and synthesis, and sound in multimodal virtual/augmented reality.

Piotr Majdak (piotr.majdak@oeaw.ac.at) is with Acoustics Research Institute, Austrian Academy of Sciences, Wien, 31390, Austria. His research interests include perceptual effect of the HRTFs, their acoustic measurement, and numeric calculation.

Arcadio Reyes-Lecuona (areyes@uma.es) is with the University of Malaga, Malaga, 29071, Spain. His research interests include 3D audio and HCI in VR, including sonic interaction.

Alessandro Vinciarelli (alessandro.vinciarelli@glasgow.ac.uk) is with the School of Computing Science, University of Glasgow, Glasgow, G128QQ, U.K. His research interests include social signal processing.

References

[1] H. Ziegelwanger, W. Kreuzer, and P. Majdak, "Mesh2HRTF: An open-source software package for the numerical calculation of head-related transfer functions," in *Proc. 22nd Int. Congr. Sound Vib.*, 2015, pp. 1–8, doi: 10.13140/RG.2.1.1707.1128.

[2] K. Pollack, P. Majdak, and H. Furtado, "Evaluation of pinna point cloud alignment by means of non-rigid registration algorithms," Audio Engineering Society, New York, NY, USA, May 2021. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=21068

[3] P. Stitt and B. F. G. Katz, "Sensitivity analysis of pinna morphology on head-related transfer functions simulated via a parametric pinna model," *J. Acoust. Soc. Amer.*, vol. 149, no. 4, pp. 2559–2572, 2021, doi: 10.1121/10.0004128.

[4] A. Andreopoulou and A. Roginska, "Database matching of sparsely measured head-related transfer functions," *J. Audio Eng. Soc.*, vol. 65, no. 7/8, pp. 552–561, Jul. 2017, doi: 10.17743/jaes.2017.0021.

[5] B. F. G. Katz and G. Parseihian, "Perceptually based head-related transfer function database optimization," *J. Acoust. Soc. Amer.*, vol. 131, no. 2, pp. EL99–EL105, 2012, doi: 10.1121/1.3672641.

[6] C. Kim, V. Lim, and L. Picinali, "Investigation into consistency of subjective and objective perceptual selection of non-individual head-related transfer functions," *J. Audio Eng. Soc.*, vol. 68, no. 11, pp. 819–831, 2020, doi: 10.17743/jaes.2020.0053.

[7] K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, and G. Pezzulo, "Active inference: A process theory," *Neural Comput.*, vol. 29, no. 1, pp. 1–49, Jan. 2017, doi: 10.1162/NECO_a_00912.

[8] G. McLachlan, P. Majdak, J. Reijniers, and H. Peremans, "Towards modelling active sound localisation based on Bayesian inference in a static environment," *Acta Acust.*, vol. 5, p. 45, Oct. 2021, doi: 10.1051/aacus/2021039.

[9] I. Engel, C. Henry, S. V. Amengual Garí, P. W. Robinson, and L. Picinali, "Perceptual implications of different Ambisonics-based methods for binaural reverberation," *J. Acoust. Soc. Amer.*, vol. 149, no. 2, pp. 895–910, 2021, doi: 10.1121/10.0003437.

[10] E. De Sena, H. Haciihabiboğlu, Z. Cvetković, and J. O. Smith, "Efficient synthesis of room acoustics via scattering delay networks," *IEEE Trans. Audio, Speech, Language Process.* (2006–2013)*, vol. 23, no. 9, pp. 1478–1492, 2015, doi: 10.1109/TASLP.2015.2438547.

[11] M. Geronazzo, J. Y. Tissieres, and S. Serafin, "A minimal personalization of dynamic binaural synthesis with mixed structural modeling and scattering delay networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 411–415.

[12] D. Poirier-Quinot and B. F. G. Katz, "Assessing the impact of head-related transfer function individualization on task performance: Case of a virtual reality shooter game," *J. Audio Eng. Soc.*, vol. 68, no. 4, pp. 248–260, 2020, doi: 10.17743/jaes.2020.0004.

[13] A. Vinciarelli *et al.*, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 69–87, 2011, doi: 10.1109/T-AFFC.2011.27.

[14] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People*. Cambridge, MA, USA: Cambridge Univ. Press, 1996.

[15] E. Hall, *The Silent Language*. New York, NY, USA: Anchor Books, 1959.

[16] I. Kastanis and M. Slater, "Reinforcement learning utilizes proxemics: An avatar learns to manipulate the position of people in immersive virtual reality," *ACM Trans. Appl. Perception*, vol. 9, no. 1, pp. 1–15, 2012, doi: 10.1145/2134203.2134206.

[17] J. Williamson, J. Li, V. Vinayagamoorthy, D. Shamma, and P. Cesar, "Proxemics and social interactions in an instrumented virtual reality workshop," in *Proc. CHI Conf. Hum. Factor Comput. Syst.*, 2021, pp. 1–13, doi: 10.1145/3411764.3445729.

[18] M. Cuevas-Rodríguez *et al.*, "3D Tune-In Toolkit: An open-source library for real-time binaural spatialisation," *PLoS One*, vol. 14, no. 3, p. e0211899, Mar. 2019, doi: 10.1371/journal.pone.0211899.

[19] P. Majdak, C. Hollomey, and R. Baumgartner, "AMT 1.x: A toolbox for reproducible research in auditory modeling," *Acta Acust.*, vol. 6, p. 19, May 2022, doi: 10.1051/aacus/2022011.

[20] P. Søndergaard and P. Majdak, "The auditory modeling toolbox," in *The Technology of Binaural Listening*, J. Blauert, Ed. Berlin, Germany: Springer-Verlag, 2013, pp. 33–56.

[21] P. Majdak, F. Zotter, F. Brinkmann, J. De Muynke, M. Mihocic, and M. Noisternig, "Spatially oriented format for acoustics 2.1: Introduction and recent advances," *J. Audio Eng. Soc.*, to be published.

[22] T. Ammermann *et al.*, *AES Standard for File Exchange - Spatial Acoustic Data File Format*, Standard AES69-2015, Audio Engineering Society, New York City, NY, USA, Mar. 2015.

[23] M. Geronazzo and S. Serafin, Eds. *Sonic Interactions in Virtual Environments*, Human–Computer Interaction Series, 1st ed. Cham: Springer International Publishing, 2022.

[24] Audio Experience Design. Accessed: Dec. 20, 2021. [Online]. Available: axdesign.co.uk

SP