The SONICOM project: AI-driven immersive audio, from personalisation to modelling

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

Abstract

Every individual perceives spatial audio differently due in large part to the unique and complex shape of their 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 personalisation. The SONICOM project aims to employ a data-driven approach to link the physiological characteristics of the ear to the individual acoustic filters which allow us to localise sound sources and to perceive them as being located around us. A small amount of data acquired from users could allow personalised 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 unlocking new approaches to understanding the mechanisms of human spatial hearing and communication.

Index Terms

Immersive audio, binaural spatialisation, artificial intelligence, perceptual modelling, virtual and augmented reality.

I. 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 (DSP), inertial sensors and hand-held controllers. Immersive audio technologies have the potential to revolutionise the way we interact socially within AR/VR environments and applications. But several major challenges still need to be tackled before we can achieve this level of simulation and control. This will involve not only significant technological advancements, but also employing artificial intelligence (AI) to measure, model and understand low-level psychophysical (sensory) as well as high-level psychological (social interaction) perception.

Funded by the Horizon 2020 FET-Proact scheme, the SONICOM project [1] 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:

- Design a new generation of immersive audio technologies and techniques to transform social interactions, specifically looking at customisation and personalisation of the audio rendering.
- Explore, map, and model how the physical characteristics of spatialised auditory stimuli can influence observable behavioural, physiological, kinematic, and psychophysical reactions of listeners within social interaction scenarios.
- Evaluate the techniques developed and data-driven outputs in an ecologically valid manner, exploiting AR/VR simulations as well as real-life scenarios.
- 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.

II. 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, artificial intelligence, spatial hearing, auditory modelling, computational social intelligence and interactive computing. The workplan is centred around three pivotal research work packages titled *Immersion*, *Interaction*, and *Beyond*, respectively. Sec. III summarises the aims and objectives of the first, identifying relevant technical developments and sensory aspects. On the other hand, the second focuses on the interaction between these and higher-level socio-psychological implications as briefly outlined in Sec. IV. The integration of the core research, the proof of concepts evaluations, and the creation of the auditory data ecosystem ensures that the various outputs of the project will have an impact beyond the end of SONICOM (see Sec. V).

[1] www.sonicom.eu

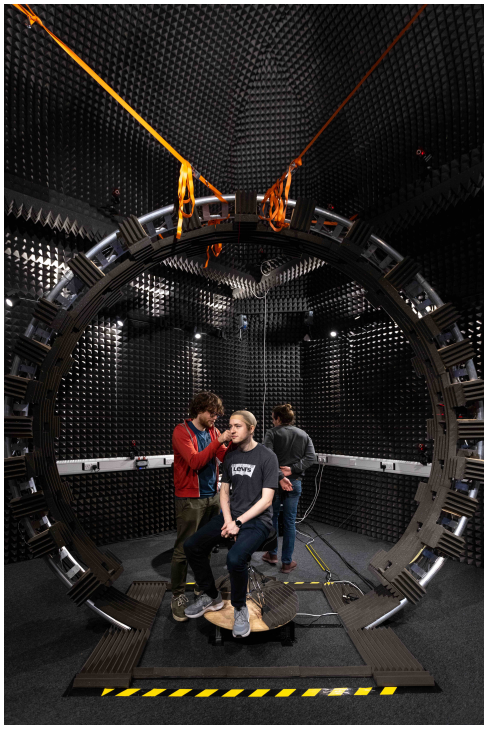

Fig. 1. HRTF measurement setup at Imperial College London (axdesign.co.uk)

III. 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 source and listener and can be described by so-called head-related transfer functions (HRTFs). These can be acoustically measured (e.g. see figure 1) or numerically modelled (e.g. [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 modelling research considers a variety of approaches. On the one hand, we focus on the creation of parametric pinna models (PPM) [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 datasets (measured or modelled), as long as they share some relevant pre-defined 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]. Said measures concern geometrical variations for parametric pinna models, perceptually signal deviations of the resulting 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.

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 personalisation 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 datasets of human behaviour and perceptually-valid metrics modelling such behaviour. By having the acoustic simulations validated against acoustic measurements, and human auditory models validated against actual behaviour, we will provide important tools for the development of AI-based algorithms predicting realistic spatial audio quality. These evaluation metrics are therefore derived through complimenting the above mentioned project database with perceptual results of the same set of individuals to a variety of perceptual tests.

We will concentrate on the issue of blending virtual objects in real scenes, which is one of the cornerstones of AR. In order to blend the real with the virtual worlds in an AR scenario, it is essential to develop techniques for the automatic estimation of the reverberant characteristics of the real environment. This will be achieved by characterising the acoustical environment surrounding the AR user, using this in-situ data to synthesize virtual sounds with matching acoustic properties. The data extracted can then be employed to generate realistic virtual reverberation matching the real world. After a set of pilot studies looking at the perceptual needs in terms of reverberation processing (e.g. [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 over-simplified highly-controlled traditional "laboratory" conditions, but will aim 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 in order to present a real-world like context while still maintaining the required laboratory controllability in order to obtain meaningful exploitable data (e.g. [12]).

IV. 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 the interaction with other agents, whether human or artificial.

For example literature shows that a behavioural cue (a smile, a gesture, a sentence, etc.) 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 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, etc.) 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 personalised 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 these latter play within the immersive and interactive environments. While 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], [17]). Its investigation is part of what is planned within the SONICOM project, and promises to be fruitful from both scientific and technological points of view.

V. Beyond

Ensuring that the project's accomplishments (algorithms, AI-based models, evaluation data, etc.) remain available to the various stakeholders including the wider research communities beyond SONICOM is of primary importance in scientific projects. In order 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 already done on the development of the 3D Tune-In Toolkit [18]. It will be implemented as an interchangeable module, allowing the use of various 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 part of the SONICOM Ecosystem. The Ecosystem will further include the Auditory Model Toolbox (AMT) [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 further extended towards the needs of SONICOM and its Ecosystem. 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 in 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 Personalisation (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 personalisation 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 connected with their exploratory actions. A publicly released corpus will be an integral part of the Ecosystem, and will include AI-driven data, behavioural 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-(virtual)reality system able to model a virtual/augmented listening experience. Such integration of SONICOM's outputs aims to promote reproducible research, creating a sustainable basis for further research beyond SONICOM.

VI. Conclusions

While 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 personalisation, 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, and at the same contributing to the creation of new and more accurate models to be shared with the wider research community. Within this context, the attempt to make use of the collected data to model social-level processing within existing sensory models is a novel element, and will allow to better predict responses to complex tasks such as speech-in-noise understanding and, more in general, sonic interactions within AR/VR scenarios.

Finally, it seems clear that to ensure an adequate level of standardisation 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 (www.sonicom.eu) has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017743.

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 *22nd International Congress on Sound and Vibration*, 2015.

[2] K. Pollack, P. Majdak, and H. Furtado, "Evaluation of pinna point cloud alignment by means of non-rigid registration algorithms," *J Aud Eng Soc*, may 2021.

[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 Am*, vol. 149, no. 4, pp. 2559–2572, 2021.

[4] A. Andreopoulou and A. Roginska, "database matching of sparsely measured head-related transfer functions," *journal of the audio engineering society*, vol. 65, pp. 552–561, july 2017.

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

[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.

[7] K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, and G. Pezzulo, "Active Inference: A Process Theory," *Neural Computation*, vol. 29, pp. 1–49, Jan. 2017.

[8] McLachlan, Glen, Majdak, Piotr, Reijniers, Jonas, and Peremans, Herbert, "Towards modelling active sound localisation based on Bayesian inference in a static environment," *Acta Acust.*, vol. 5, p. 45, 2021.

[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," *The Journal of the Acoustical Society of America*, vol. 149, no. 2, pp. 895–910, 2021.

[10] E. De Sena, H. Haciihabiboğlu, Z. Cvetković, and J. O. Smith, "Efficient Synthesis of Room Acoustics via Scattering Delay Networks," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 9, pp. 1478–1492, 2015.

[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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 411–415, 2020.

[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.

[13] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, 2011.

[14] B. Reeves and C. Nass, *The media equation: How people treat computers, television, and new media like real people*. Cambridge University Press, 1996.

[15] E. Hall, *The silent language*. 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 Transactions on Applied Perception*, vol. 9, no. 1, pp. 1–15, 2012.

[17] J. Williamson, J. Li, V. Vinayagamoorthy, D. Shamma, and P. Cesar, "Proxemics and social interactions in an instrumented virtual reality workshop," in *Proceedings of CHI*, pp. 1–13, 2021.

[18] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. d. l. Rubia-Cuestas, L. Molina-Tanco, and A. Reyes-Lecuona, "3D Tune-In Toolkit: An open-source library for real-time binaural spatialisation," *PLOS ONE*, vol. 14, p. e0211899, Mar. 2019.

[19] P. Majdak, C. Hollomey, and R. Baumgartner, "AMT 1.x: A toolbox for reproducible research in auditory modeling," *Acta Acustica*, vol. 6, p. 19, 2022. Publisher: EDP Sciences.

[20] P. Søndergaard and P. Majdak, "The auditory modeling toolbox," in *The Technology of Binaural Listening* (J. Blauert, ed.), pp. 33–56, Berlin, Heidelberg: Springer, 2013.

[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*, vol. in proof, 2022.

[22] T. Ammermann, S. Bharitkar, F. Camerer, W. D. Bruijn, T. Carpentier, M. Emerit, F. Fleischmann, K. Hamasaki, A. Harma, J.-M. Jot, M. Kelly, R. Kessler, T. Knowles, P. Majdak, F. Melchior, M. Noisternig, B. Olson, G. Pallone, C. Par, M. Parmentier, A. Pereira, C. Pike, J. Plogsties, E. Ronciere, T. Sporer, B. van Daele, W. Woszczyk, and H. Ziegelwanger, "AES standard for file exchange - Spatial acoustic data file format," Standard AES69-2015, Audio Engineering Society, New York City, United States, Mar. 2015.

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