Individual Three-dimensional Spatial Auditory Displays for Immersive Virtual Environments

Dr Simone Spagnol
Aalborg University, Denmark

Multimodal interfaces represent a key factor for enabling an inclusive use of new technologies by everyone. To achieve this, realistic models that describe our environment are of topical importance, particularly models that accurately describe the acoustics of the environment and communication through the auditory modality. The synthesis and manipulation of auditory spaces embodies new domains of experience that promise to change the way we think about sound, how we manipulate it and experience it, and the manner in which the role of listening is evaluated in the future of computing.

3D sound: current use and potential

3D sound technologies can provide accurate information about the relationship between a sound source and the surrounding environment, including the listener herself/himself. This information cannot be substituted by any other modality (e.g. visual or tactile). Nevertheless, today's spatial representation of audio tends to be simplistic and with poor interaction capabilities, being multimodal systems mostly focused on graphics processing and integrated with basic audio solutions.

Three important reasons why many media components lack realistic audio rendering can be identified. First, they do not exploit information about the environment in which they are working, and no adaptation onto the user is provided. Second, the ever-increasing need for bandwidth and computational resources can easily lead to system overload, especially in the case of mobile devices. Third, 3D sound technologies of the highest fidelity rely on invasive and/or costly reproduction devices (for instance, loudspeaker arrays), leading to poorly integrated experiences due to an unbridged gap between the real and virtual worlds.

Binaural technologies, i.e. headphone-based rendering of virtual sound sources as heard by our own ears in the real world, promise to bridge this gap. In order to enable authentic auditory experiences with binaural technologies, the correct sound pressure level due to one or more acoustic sources positioned in a virtual space needs to be reproduced at the eardrums of the listener.

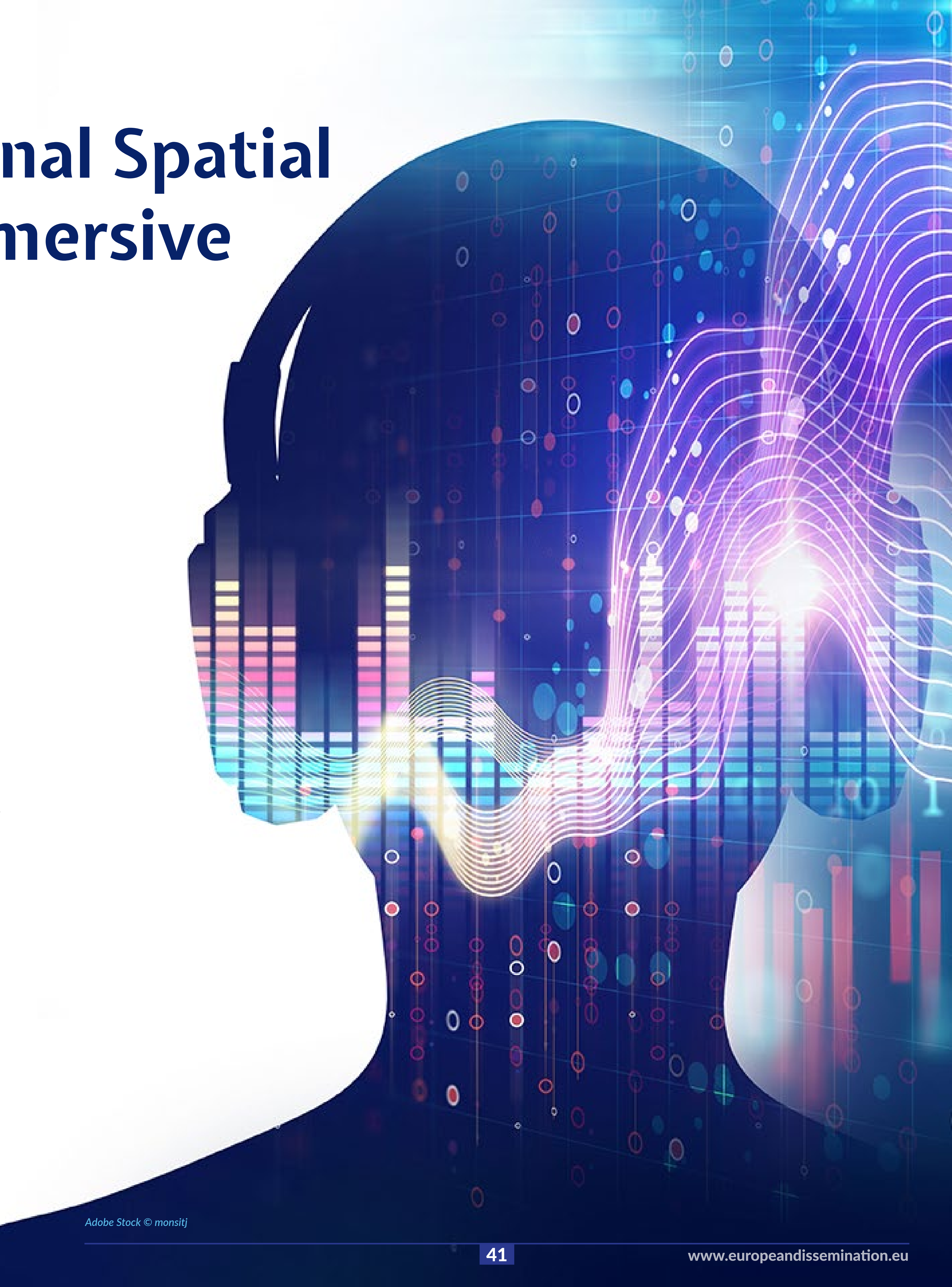
Head-related transfer functions

Typically, binaural audio technologies rely on the use of head-related transfer functions (HRTFs), specific digital filters that capture the acoustic effects of the human head. HRTFs allow faithful simulation of the sound signal arriving at the ear canal as a function of the sound source's spatial position. The classic solution to binaural audio delivery, meaning the one that best approximates real listening conditions, involves the use of individual HRTFs measured on the listener with the addition of head tracking and artificial reverberation. However, obtaining personal HRTF data for a vast number of users is only possible with expensive equipment and invasive recording procedures.

This is the reason why non-individual HRTFs acoustically measured on anthropomorphic mannequins are often preferred in practice. The drawback with non-individual HRTFs is that these transfer functions likely never match the listener's unique anthropometry—especially in the outer ear—resulting in frequent localisation errors such as front/back reversals, elevation angle misperception, and inside-the-head localisation.

IT'S A DIVE: objectives

IT'S A DIVE focuses on structural modelling of HRTFs, i.e. a family of state-of-the-art modelling techniques that has the potential to overcome most limitations of headphone-based 3D sound systems. According to this approach, the main effects involved in spatial sound perception (e.g. acoustic delays and shadowing due to head diffraction, reflections on ear edges and shoulders are isolated and modelled separately with a corresponding filtering



Adobe Stock © monsitj

element. The advantages over alternative binaural rendering techniques are twofold: adaptability to a specific subject, based on anthropometric quantities (head dimensions, ear shape, and so on); and computational efficiency, as models are structured in smaller blocks each simulating one physical effect, allowing low-cost implementation and low-latency reproduction on any device. Such an approach opens the door for a very desirable alternative to individual HRTF measurements, that is, extrapolating a personal set of HRTFs from pictures or 3D models of the user's head. Furthermore, it grants seamless fruition of realistic individual 3D audio, allowing the user to 'dive deep' into any virtual environment.

In particular, the main objective of IT'S A DIVE has been the development of a completely customisable structural HRTF model, previously not available in the literature on 3D sound, with a focus on the customisation phase. For this purpose, the following specific objectives have been pursued:

- the collection of an extensive dataset of HRTFs and anthropometric measurements focusing on the individual factor;
- the development of a technically sound methodology for HRTF analysis and synthesis with the aim of establishing a physical connection between anthropometric and acoustical data in each structural component; and
- an extensive evaluation procedure with a significant number of human participants in a virtual reality game-based setting.

Research methodology

Accordingly, the IT'S A DIVE research methodology has developed in three phases: *acquisition*, *modelling*, and *evaluation*. For what concerns the acquisition phase, a large number of public HRTF databases from worldwide research labs have been collected and fused to form a unique large set of acoustic measurements (>400 human subjects). In addition to the organisation and use of these public datasets, most resources in this phase have been allocated to the collection of a new dataset of custom acoustic measurements, named the *Viking HRTF dataset* (Spagnol *et al.*, 2019), in collaboration with the University of Iceland. This dataset includes full-sphere HRTFs measured on a dense spatial grid with a binaural mannequin with different artificial pinnae attached (Spagnol *et al.*, 2020a). Anthropometric data have been collected from pre-processing of public databases or obtained with new measurements on 2D or 3D anatomical data (e.g. ear pictures, head meshes). In particular, new features related to the shape of the ear have been automatically extracted from 3D head/ear meshes, such as depth maps (Onofrei *et al.*, 2020), *edge maps*—i.e. 2D representations of the most prominent pinna edges (Miccini and Spagnol, 2021), and *reflection maps*—i.e. selections of mesh points that theoretically produce reflections towards the ear canal entrance (Spagnol *et al.*, 2020b).

The modelling phase has focused on a blend of traditional signal processing techniques, state-of-the-art machine learning algorithms tuned to both global and local characteristics of HRTFs, and physically inspired models of sound propagation within the ear. Each structural component has been analysed through ad-hoc signal processing algorithms; this has been possible because some of the collected HRTF databases contain partial responses of head-only or earless mannequins. Then, since HRTFs are by design subject to high dimensionality issues due to the wide range of predictors, adequate dimensionality reduction and/or feature extraction techniques have been applied to partial HRTF data in order to obtain compact representations to be correlated to anthropometric data (Miccini and Spagnol, 2019; Miccini and Spagnol, 2020). Finally, the most adequate machine learning techniques, including state-of-the-art deep learning algorithms, have been applied to yield the model that better meets speed, interpretability, and accuracy requirements. This procedure has allowed the design of a complete structural HRTF model combining measured, synthesised, and selected components (Miccini and Spagnol, 2021).

In the evaluation phase, signal-related error metrics and auditory models (Spagnol, 2020a; Spagnol, 2020b) have been developed to compare the customised HRTFs obtained through the developed structural model against the original measured HRTFs of a number of database subjects. Indeed, a good objective correspondence between the two sets is the basis for performing subjective tests. The HRTF models have then been integrated into a 3D game in order to perform individual tests with the dynamic rendering of virtual sound sources (Andersen *et al.*, 2021). Collected metrics from the user tests include, among others, localisation error, degree of externalisation, and an extensive user questionnaire. These tests are being carried out at the time of writing.

Adobe Stock © aanbetta

Progress beyond the state of the art and impact

It has to be stressed that no HRTF models for customised binaural audio delivery that do not make use of pre-recorded or numerically simulated HRTFs alone have ever been successfully proposed and evaluated in previous literature. IT'S A DIVE successfully filled such a gap.

For what concerns the technological outcomes of the project, realistic 3D auditory displays represent an innovative breakthrough for a plethora of additional application areas not envisaged in IT'S A DIVE. Some examples are personal cinema, teleconferencing systems, and travel aids for the visually impaired. In particular, binaural sound technologies are expected to become more and more used in computer games. Furthermore, the techniques developed in IT'S A DIVE have minimal hardware requirements with respect to other technologies for immersive sound reproduction. This is particularly relevant for mobile applications.

References

Andersen, J.S., Miccini, R., Serafin, S. and Spagnol, S. (2021) 'Evaluation of individualized HRTFs in a 3D shooter game'. Unpublished.

Miccini, R. and Spagnol, S. (2019) 'Estimation of pinna notch frequency from anthropometry: An improved linear model based on principal component analysis and feature selection', *Proceedings of the 1st Nordic Sound & Music Computing Conference (Nordic SMC 2019)*, Stockholm, 19-20 November, pp. 5–8.

Miccini, R. and Spagnol, S. (2020) 'HRTF individualization using deep learning', *Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Workshops (VRW 2020)*, Atlanta, 22–26 March, pp. 390–395.

Miccini, R. and Spagnol, S. (2021) 'A hybrid approach to structural modeling of individualized HRTFs'. Unpublished.

Onofrei, M.G., Miccini, R., Unnthórsson, R., Serafin, S. and Spagnol, S. (2020) '3D ear shape as an estimator of HRTF notch frequency', *Proceedings of the 17th Sound & Music Computing Conference (SMC 2020)*, Torino, 24–26 June, pp. 131–137.

Spagnol, S., Purkhús, K.B., Björnsson, S.K. and Unnthórsson, R. (2019) 'The Viking HRTF dataset', *Proceedings of the 16th Sound & Music Computing Conference (SMC 2019)*, Málaga, 28–31 May, pp. 55–60.

Spagnol, S. (2020a) 'HRTF selection by anthropometric regression for improving horizontal localization accuracy', *IEEE Signal Processing Letters*, 27, pp. 590–594.

Spagnol, S. (2020b) 'Auditory model based subsetting of head-related transfer function datasets', *Proceedings of the 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2020)*, Barcelona, 4–8 May, pp. 391–395.

Spagnol, S., Miccini, R. and Unnthórsson, R. (2020a) 'The Viking HRTF dataset v2'. doi: 10.5281/zenodo.4160401 (Accessed: 6 December 2020).

Spagnol, S., Miccini, R., Onofrei, M.G., Unnthórsson, R. and Serafin, S. (2020b) 'Estimation of HRTF notches from ear meshes: Insights from a revised reflection model'. Manuscript submitted for publication.

PROJECT SUMMARY

"Individual Three-Dimensional Spatial Auditory Displays for Immersive Virtual Environments" (IT'S A DIVE) has developed innovative techniques for individualised 3D sound rendering. Based on a blend of traditional signal processing techniques, state-of-the-art deep learning algorithms, and physically inspired models of sound propagation within the ear, the proposed approach allows low-cost and real-time fruition of realistic customised 3D sound through headphones.

PROJECT LEAD PROFILE

Dr Simone Spagnol received his PhD in Information Engineering from the University of Padova in 2012. He then worked as a postdoctoral researcher in Italy and Iceland before being awarded a Marie Skłodowska-Curie fellowship at Aalborg University. He authored more than 60 peer-reviewed publications in the fields of audio and acoustics and received four best paper awards as first author.

PROJECT PARTNERS

The research has been carried out at the Department of Architecture, Design, and Media Technology (CREATE) of Aalborg University, Copenhagen. The project also saw a technical collaboration with the Faculty of Industrial Engineering, Mechanical Engineering and Computer Science of the University of Iceland.

CONTACT DETAILS

Dr Simone Spagnol

Dept. of Architecture, Design & Media
Technology, Aalborg University Copenhagen
2450 Copenhagen, Denmark

ssp@create.aau.dk

https://itsadive.create.aau.dk

FUNDING

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Individual Fellowship (MGA MSCA-IF) grant agreement No.797850.