Data-based spatial audio processing

Citation for the original published paper (version of record):
Cobos, M., Ahrens, J., Kowalczyk, K. et al (2022). Data-based spatial audio processing. Eurasip Journal on Audio, Speech, and Music Processing, 2022(1).
http://dx.doi.org/10.1186/s13636-022-00248-5

Cobos *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2022) 2022:13
https://doi.org/10.1186/s13636-022-00248-5

EURASIP Journal on Audio, Speech, and Music Processing

EDITORIAL Open Access

Data-based spatial audio processing

Maximo Cobos[1*†], Jens Ahrens[2*†], Konrad Kowalczyk[3*†] and Archontis Politis[4*†]

Spatial audio today is a mature field with an active research community involving both industry and academia. Spatial audio processing is a key technology for many sound-related applications, ranging from personalized audio experiences to the creation of the sound metaverse. The present special issue concerns data-based spatial audio processing. The term "data" is defined very broadly in this case, and any processing method that involves data of any sort—be it measured data or data captured in real-time—qualifies. The topics that are covered by the articles that this special issue comprises are just as broad.

The past decade has shown a growing interest in such data-based methods. This may be attributed to the increased availability of advanced multichannel measurement and capture setups and, of course, the increased availability of processing resources. The accomplishment of tasks such as the measurement of head-related transfer functions (HRTFs) or sound source directivities has become considerably easier, and the data continues to become publicly available on a much larger scale than previously. This allows also for researchers who do not have access to according measurement resources to be active in the domain in a comprehensive manner. Along with this availability of data, the exponential growth of deep learning and its application to most research areas has also contributed to this interest, where data-based methods in the classical sense have evolved into data-driven approaches fully based on machine learning (ML) techniques. The article co-authored by the guest editors of this special issue provides an extensive overview over data-based methods in spatial audio capture, processing, and reproduction. The article's quintessence is the identification of the role that ML-based methods play in this regard, whereby the authors use the term data-based in a narrower interpretation than this Editorial. The authors assume the so-called data-based representation of the audio scene that is being processed where the spatial information such as the positions of the sound sources and the acoustic response of the virtual environment are encoded into the audio signals. This is opposed to model-based representations where the audio scenes are described by a set of individual source signals as well as metadata such as the sources' positions. Sound radiation and propagation are described by physical models in this case. The authors categorize the available methods based on the task that they implement and sort them based on their location inside the processing pipeline, which is generally composed of a capture stage, a processing stage, and a reproduction stage. The capture stage is dominated by model-based methods, whereby recent works have successfully used ML to perform tasks such as the identification of auditory localization cues in binaural signals or scene decomposition into foreground and background. The situation is similar for the reproduction stage where ML is popular primarily only in the individualization of HRTFs. A number of tasks in terms of spatial audio processing have been accomplished with ML such as source localization and the extraction of acoustic metrics. All methods that produce a high-resolution representation of the spatial audio scene are dominated by classical signal processing. The authors identify the following aspects to be hindering a breakthrough in this regard: (1) The

†Maximo Cobos, Jens Ahrens, Konrad Kowalczyk and Archontis Politis contributed equally to this work.

*Correspondence: maximo.cobos@uv.es; jens.ahrens@chalmers.se; konrad.kowalczyk@agh.edu.pl; archontis.politis@tuni.fi

[1] Computer Science Department, Universitat de València, 46100 Burjassot, Valencia, Spain
[2] Division of Applied Acoustics, Chalmers University of Technology, 412 96 Gothenburg, Sweden
[3] Institute of Electronics, AGH University of Science and Technology, 30-059 Krakow, Poland
[4] Department of Information Technology and Communication Sciences, Tampere University, FI-33720 Tampere, Finland

lack of suitable psychoacoustic models that are applicable to a broad range of audio scenes, and (2) the very limited availability of content and the associated immense data volume that requires computation resources beyond the typical. The breakthrough of ML in the creation and processing of high-fidelity spatial audio scenes is therefore yet to occur.

A good example of the growing availability of spatial audio data is the article of Di Carlo and co-authors, which presents "dEchorate," a dataset of multichannel room impulse responses (RIRs) measured in a cubic room. This is the first publicly available dataset of RIRs including annotations of the timings of early echoes as well as positions from microphones, real, and image sources for different configurations. The dataset features 1800 RIRs obtained from 6 arrays of 5 microphones each, 6 sound sources, and 11 different acoustic conditions. Such data along with its accompanying information and software utilities can be an excellent research resource for echo-aware audio processing, room-geometry estimation, learning-based RIR estimation, and algorithm validation considering a variety of realistic conditions.

Binaural technology has experienced a surge in popularity in recent years. In particular, the perception of the spatial location is an important aspect of binaural rendering in a breadth of audio and multimedia applications, including gaming, virtual, and augmented reality. For natural spatial hearing, personal HRTFs should be used as they depend on the anthropometric parameters of an individual. The most common approach to obtain personal HRTFs is to directly measure the binaural responses of a listener. Alternative approaches include synthesizing HRTFs based on an individual's morphological features and an assumed model or the adaptation of the relevant spatial hearing cues for a particular listener. The article of Gutierrez-Parera et al. treats the latter issue of adjusting generic HRTFs to those of an individual, which the authors achieve by scaling inter-aural time differences (ITDs) through anthropometric parameters. The authors perform a listening test to infer the relationship between the measured anthropometric parameters and listeners' spatial perception when using scaled versions of ITDs of generic HRTFs. Based on the results of an exploratory perceptual test, they propose a method to predict an individual's scaling factor that is applied to adapt ITDs for localization in a horizontal plane and validate their approach with another perceptual test and objective measures. An outcome of this study includes a practical method to fit ITDs of the widely used binaural dummy heads of Brüel & Kjær and Neumann.

While the localization of sources from binaural signals has been and continues to be a problem that attracts great interest, few works have focused on the localization mechanisms involved in the localization of "source ensembles" in binaural music recordings. Indeed, front-back confusions are known to be an important issue in this regard. The article by Zieliński et al. addresses the automatic disambiguation between front and back audio sources in binaural music recordings. The work is developed considering a large dataset of binaural excerpts (22,496) generated by convolving multi-track recordings with 74 sets of HRTFs. A traditional ML method and a deep learning-based method using convolutional neural networks (CNNs) are proposed and compared on the discrimination task. The article analyses the design choices of both systems and compares their performance over both HRTF-dependent and HRTF-independent scenarios, identifying also those features that provide high discrimination power within the considered framework. The result of this work may lead to a better understanding of the front-back confusion problem, paving the way towards future location-aware binaural music search and retrieval systems.

Binaural rendering of real spatial sound scenes is a core component of immersive audio and audio for virtual and augmented reality. It also enables to auralize concert recordings and sports events with a high level of perceptual immersion. Traditionally, such a sound reproduction task has been dominated by recording devices equipped with spherical microphone arrays and the associated spatial transformation of their recordings to integrate HRTF information, as done, e.g., in Ambisonics. The basic approach consists in capturing the spatial sound field, transforming the recorded multichannel signals into the spherical harmonic domain, referred to as encoding, and reproducing the encoded signals over headphones using inverse spherical harmonic transform followed by weighting with HRTFs for the corresponding directions. This processing pipeline has been studied extensively in the literature, with bounds in performance depending on the size and number of microphones, and optimization of the spatial transform stage informed by perceptual spatial cues. The article of Arend et al. presents a computationally efficient method for real-time binaural rendering in which instead of the encoding and decoding steps, the signals of a spherical microphone array are directly convolved with a set of precomputed FIR filters that model a linear time-invariant system. The real-time operation even for spherical arrays of high orders is possible with this approach; however, it comes at a cost of a lower flexibility in sound field manipulation such as rotation by an angle. As a consequence, the set of linear filters needs to be precomputed and stored for every possible head orientation of a listener. The method has been validated on two working examples.

Traditional methods for binaural reproduction of the recorded sound scenes have been developed and formulated in the spherical harmonic domain, tailored for

spherical microphone arrays. As more general array configurations, such as wearable or head-mounted ones, are becoming more important for immersive media and immersive communications, binaural methods from such array recordings will also need to be developed and studied. The article by Ifergan and Rafaely is a comprehensive step in that direction. The authors formulate binaural rendering for arbitrary arrays as a beamforming problem of a discrete set of beamformer signals rendered with corresponding HRTFs from the respective directions. Parallels with spherical microphone arrays and spherical harmonic processing for binaural rendering are drawn, and useful performance metrics are transferred from the spherical microphone array case to the general array one.

Sound source directivities are usually determined from recordings of a given source such as a musical instrument in an open spherical microphone array in a free field. The main difficulties with such measurements are the fact that the source signal is not known, and the resolution of the angular sampling of the sound field is very low. As a consequence, usually, only the frequency-dependent magnitude of the directivity is determined, and interpolation over the direction is performed. Ackermann et al. compare the accuracy of the three different methods of spherical harmonic interpolation, thin plate pseudo-spline interpolation, and piece-wise linear, spherical triangular interpolation. The test data are obtained from four different musical instruments, whereby an automatic excitation was used to be able to achieve a measurement with a high angular resolution by using only a moderate number of microphones and rotating the sound source. The test data are downsampled to 32 directions, and the accuracy of the three methods under consideration is determined by comparing the interpolation results to the ground truth measured at the given direction. The accuracy turns out to be strongly dependent on the source type and on the sampling grid. It is in the same order of magnitude for all methods under consideration, whereby the smallest average global error occurred for thin plate pseudo-spline interpolation. The authors also provide a number of guidelines for further processing of the interpolated directives.

Multichannel source separation methods have a strong potential in spatial audio applications, such as in spatial re-mixing, enhancement, or modification of spatial content. In recent years, large performance gains have been achieved using deep neural network (DNN) separation models, such as the Conv-TasNet and related multichannel extensions. Similar gains have been reported by learning-based dereverberation, training DNN models to perform a complex spectral mapping from reverberant source spectrograms to dry or anechoic ones. The article by Chen et al. combines skillfully such advances for the task of simultaneous multichannel source separation and dereverberation. A Conv-TasNet-inspired DNN is a construction operating as a trainable beamformer and giving an estimate of separated source signals. These separated signals are then post-processed by a dereverberation network, producing the final enhanced source signals. The work conducts a comprehensive comparison against traditional dereverberation and spatial filtering combinations, as well as traditional source separation approaches. Furthermore, it compares against a recent multichannel Conv-Tasnet extension, and, tested in reverberant speech mixtures, it shows solid improvements in separation performance, speech intelligibility, and enhanced speech quality. The performance is additionally analyzed with respect to the number of microphones, varying reverberation times, and spatial separation of the speakers.

The above contributions are significant examples of ongoing work in data-based spatial audio and reflect very well its current status. The availability of more datasets (RIRs, HRTFs, Ambisonics recordings, etc.) is facilitating open research that efficiently exploits the acoustic knowledge provided by such data, showing a very interesting situation where traditional signal processing and emerging ML-based methods coexist and interact in a wide range of applications. Echo-aware signal processing, binaural perception analysis, HRTF individualization, efficient and flexible binaural rendering, interpolation methods for source directivities, or multichannel source separation, which have been covered in this special issue, are only some of these applications.

Maximo Cobos
Jens Ahrens
Konrad Kowalczyk
Archontis Politis

Authors' contributions
All authors contributed equally to this work. The authors read and approved the final manuscript.

Funding
This work received funding from Grant RTI2018-097045-B-C21 funded by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe." Additionally, from the National Science Centre of Poland under grant number DEC-2017/25/B/ST7/01792 and Generalitat Valenciana under grants AICO/2020/154 and AEST/2020/012.

Declarations

Competing interests
The authors declare that they have no competing interests.

Published online: 08 June 2022

Publisher's Note