DOI: 10.61782/fa.2023.0266

EXPLORING THE IMPACT OF TRANSFER LEARNING ON GAN-BASED HRTF UPSAMPLING

Aidan O. T. Hogg* **He Liu** **Mads Jenkins** **Lorenzo Picinali**
Audio Experience Design Team, Dyson School of Design Engineering
Imperial College London, UK

ABSTRACT

Individualised head-related transfer functions (HRTFs) are essential for creating realistic virtual reality (VR) and augmented reality (AR) environments and interactions. Performing acoustic measurements is the most accurate way to capture these individualised HRTFs. However, one of the main challenges is acoustically capturing high-quality HRTFs without the need for expensive equipment and a lab-controlled setting. To make these measurements more feasible on a large scale, HRTF upsampling has been exploited in the past, where a high-resolution HRTF is created from a low-resolution measurement. However, as the world shifts to more data-driven methods, upsampling HRTFs using machine learning (ML) has become more prevalent. The main limitation is the lack of HRTF data available for model training. This paper explores the use of transfer learning (TL) on a new synthetic HRTF dataset generated from three-dimensional (3D) meshes using a parametric pinna model and the performance improvements that can be achieved. The performance is also compared against using a small acoustically measured dataset for TL, with the aim to start answering the question: 'Is it better to have more data, or is it better for the data to be of higher quality?'.

Keywords: *transfer learning, generative adversarial network, head-related transfer function, super-resolution, upsampling.*

1. INTRODUCTION

We live in a world where communication is vital, and this could not be more evident when it comes to online interactions. We all desire seamless remote connectivity, whether it be in an online meeting or the latest VR experience. The ability to create realistic, immersive audio scenarios is needed to achieve this. Immersive audio is what we experience in everyday life; some sounds are close, some are far away, some are moving, and all come from different directions. In most real-life situations, the lack of immersive audio leads to frustration often felt when communicating remotely [1].

One way to achieve high realism in immersive audio is to exploit our two ears and generate realistic sounds at these two sensors. The question is how sounds that mimic real-world 3D audio can accurately be created [2] and, more specifically, how this can be adapted for individual listeners. This individualisation has resulted in a large amount of research focusing on HRTFs, which capture the interaural (i.e. differences heard between the listener's two ears) and monaural localisation cues [3].

It has been shown in the past that many approaches can be deployed for this HRTF individualisation task, and an overview of some of the most common methods can be found in [4]. However, taking an acoustic measurement [5] is still considered to be the gold standard among these different approaches. The downside is the time it takes and the expensive custom setup required. To overcome these issues and to make the method scalable, spatial upsampling methods have been proposed in the past that can generate spatial high-resolution HRTFs from low-resolution ones [6]. This process is commonly referred to as HRTF upsampling and can be achieved using various approaches.

The main aim of this paper is to investigate the use

**Corresponding author: aidan@aidanhogg.uk.*

of TL on an ML approach that uses a super-resolution generative adversarial network (SRGAN) to tackle the HRTF up-sampling problem. The paper builds on a pilot study that was undertaken in [7] and the SRGAN approach of [8] and [9]. The advantage of ML approaches over traditional upsampling is that it is able to recreate the missing information in the sparse measurements using the knowledge learnt from a training set that contains many high-resolution HRTFs. The main drawback, currently, is the limited amount of acoustically measured HRTF data available for training. A common solution to this problem of limited data is TL, which has been shown to be successful in many domains, including image recognition, natural language processing (NLP) and speech recognition [10–12]. In [13], TL was used in conjunction with GANs for image reconstruction of under-sampled magnetic resonance imaging (MRI) data. [14] also showed the benefits of TL when a GAN used for image generation was pre-trained with ImageNet. There is no standard implementation of TL; [15] described over 20 different strategies. A common approach is to train the entire network with pre-training data and then retrain the entire network with task-training data, which will be exploited in this work. This paper will present the results of TL using a new synthetic HRTF dataset generated from 3D meshes using a parametric pinna model and the performance improvements that can be obtained.

2. METHOD

In the past, SRGANs have performed well on the task of up-sampling images. However, when applying this approach to HRTFs up-sampling, the main challenge is that the data occupies an extra dimension in physical space where it is not uniformly distributed. One way to solve this problem is to modify the network to handle this non-uniformly distributed HRTF data, such as graph neural networks (GNNs) [16]. However, in this work, a pre-processing step is used to convert the spherical HRTF data into a form that all of the image upsampling literature can exploit.

2.1 Pre-processing

This pre-processing consists of two main steps 1) the spherical HRTF data is projected onto a two-dimensional (2D) surface using a gnomonic equiangular projection to remove the extra dimension. 2) barycentric interpolation [17] is utilised to shift the irregularly spaced impulse responses (IRs) onto an evenly spaced Cartesian grid. A detailed description of this approach is given in [8]. One of the big advantages of this pre-processing is that it becomes possible for the SRGAN network to exploit many different HRTF datasets for training, which is essential for TL. This is true even when two datasets contain spatially very different measurements because the points will all be mapped onto the same Cartesian grid.

2.2 GAN Architecture

A GAN architecture, similar to that of [18], is exploited where the generator network G, in this case, aims to generate high-resolution HRTFs from low-resolution HRTF inputs. On the other hand, the discriminator network D aims to discriminate whether a HRTF is real or generated by the network G. The loss function used is a weighted sum of the content loss, which compares the output of G using the log-spectral distortion (LSD) to the high-resolution ground truth, with an adversarial loss, which measures how frequently G successfully fools D.

2.3 Transfer Learning

This paper aims to explore improvements that can be achieved from the use of TL and attempts to answer the question: 'When it comes to TL for HRTF upsampling, is it better to have more data, or is it better for the data to be more realistic?'. To achieve this, TL is employed by first pre-training the SRGAN before retraining on the HRTF task data. A comparison of two types of pre-training data is presented: acoustically measured data from a different dataset to the task data and a newly created synthetic dataset. Due to the synthetic dataset being much larger than the acoustically measured dataset, this should help in beginning to address the question of whether quality or quantity is more important.

2.4 Post-processing

The full HRTFs were reconstructed after upsampling using a minimum-phase approximation and a simple ITD model to perform this perceptual evaluation.

3. TRAINING

The high-resolution 1280 target is generated by processing the ARI HRTF dataset (which contains 1550 positions for each listener, see Section 4.2.1) using the pre-processing described in Section 2.1.

To obtain the low-resolution HRTFs from their high-resolution counterparts, the HRTFs are downsampled

Table 1. A comparison of the mean log-spectral distortion (LSD) and standard deviation (SD) error across all source positions in the ARI dataset for different upsampling factors where the results are given in decibels [dB]. The 'best' performance of each upsampling factor has been highlighted.

Method	Upsample Factor (No. original → upsampled) [dB]			
	320 → 1280	80 → 1280	20 → 1280	5 → 1280
SRGAN (No TL)	3.28 (0.13)	4.86 (0.24)	4.99 (0.27)	5.30 (0.35)
TL (Synthetic)	3.02 (0.15)	4.07 (0.19)	4.75 (0.28)	5.16 (0.39)
TL (SONICOM)	3.05 (0.14)	4.21 (0.17)	4.79 (0.25)	5.39 (0.33)
Baseline	2.50 (0.20)	3.71 (0.22)	5.18 (0.23)	7.30 (0.33)

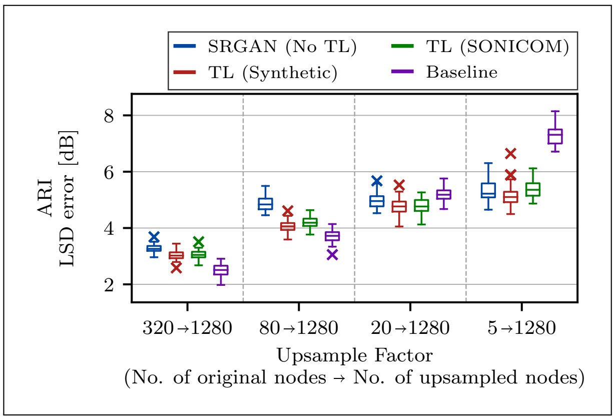

Figure 1. Log-spectral distortion (LSD) evaluation.

using the `torch.nn.functional.interpolate` function. To evaluate the performance of the GAN, results from four different downsampling rates are given, where the number of original positions kept is 320, 80, 20 and 5. The high-resolution HRTF target of 1280 positions was selected as it is comparable to the 1550 positions measured in the ARI HRTF dataset.

It should also be noted that while HRTFs possess both magnitude and phase components, only the magnitude component is used as an input to the SRGAN. The model was trained using the *Adam* optimiser [19], with hyperparameter values of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rates for the generator, G, and discriminator, D, networks were set as 2.0×10^{-4} and 1.5×10^{-6}, respectively. In training, D and G were alternately updated with different frequencies, with D being updated four times for every G update. G contains 8 residual blocks and 512 hidden features. This model was implemented using a PyTorch framework and trained on an NVIDIA Quadro RTX 6000 graphics processing unit (GPU).

4. RESULTS

The code to reproduce these results is provided in [9].

4.1 Experimental Setup

Results are given for the ARI dataset; the SONICOM dataset is used for TL and represents an acoustically measured dataset for comparison. These results are compared against using the SONICOM Synthetic dataset for TL as well as the performance without the use of TL. Barycentric interpolation, as implemented in [17], is also used as a baseline as it is one of the most common methods for HRTF upsampling.

4.2 Experimental Data

4.2.1 SONICOM and ARI

The SONICOM dataset [5, 20] is a publicly-released HRTF dataset which aims at facilitating reproducible research in the spatial acoustics and immersive audio domain by including in a single database HRTFs measured from an increasingly large number of subjects (200 subjects were used in this work). The ARI HRTF database [21], on the other hand, contains HRTF measurements on 221 subjects, making it one of the largest measured HRTF datasets available.

4.2.2 SONICOM Synthetic

This newly created dataset was generated from 3D meshes using the boundary element method (BEM) [22] and the open-source tool Mesh2HRTF [23]. The parametric pinna model (PPM) [24] was used to generate these meshes. The 9 control bone parameters and 18 shape key parameters (which customised 134 overall dimensions) used in the PPM are selected randomly from a normal distribution where the mean and standard deviation are chosen by taking the average parameters from 15 pinna models. To introduce differences between the left and right ear, the left ear is first generated, and then for each parameter, a maximum of 20% difference of the parameter range is added, with the amount chosen being randomly selected from a normal distribution. The left and right ear meshes are then stitched onto a dummy head [25]. To speed up the computation, the meshes (containing approximately 63,000 triangles) are downscaled to between 16,000

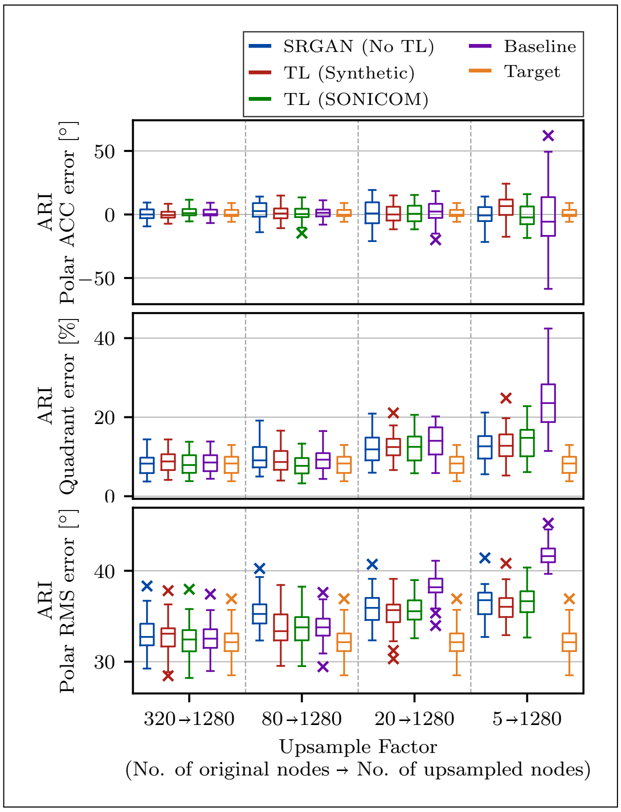

Figure 2. Perceptual evaluation.

to 20,000 triangles using the mesh grading plugin on OpenFlipper [26]. This mesh grading used the fourth-order COSalpha grading function with a minimum and maximum global target edge length of 0.0015 m and 0.01 m, respectively.

Finally, the graded meshes are passed as the input into Mesh2HRTF, where the output frequency range of the HRTF was set to 100 - 16,000 Hz (128 samples), the speed of sound was set to 346.18 (m/s), and the density of air was set to 1.1839 (kg/m^3). The magnitude of the generated synthetic HRTF data was then scaled to match the ARI and SONICOM datasets more closely.

4.3 Comparative Evaluation

4.3.1 Log-spectral distortion (LSD) evalulation

The LSD metric is defined as

$$\mathrm{LSD} = \tfrac{1}{N}\textstyle\sum_{n=1}^{N}\sqrt{\tfrac{1}{W}\textstyle\sum_{w=1}^{W}\left(20\log_{10}\dfrac{|H_{\mathrm{HR}}(f_w,x_n)|}{|H_{\mathrm{US}}(f_w,x_n)|}\right)^2}, \quad (1)$$

Table 2. The mean and standard deviation (SD) values of the perceptual evaluation across the subjects in the ARI test set for the different upsampling factors. The 'best' performance of each upsampling factor has been highlighted.

(a) Polar accuracy error comparison where the results are given in degrees [°].

Method	Upsample Factor (No. original → upsampled) [°]			
	320 → 1280	**80 → 1280**	**20 → 1280**	**5 → 1280**
SRGAN (No TL)	0.46 (4.50)	4.38 (6.63)	1.04 (8.46)	-1.95 (8.71)
TL (Synthetic)	-0.17 (4.30)	3.48 (4.55)	-0.58 (7.18)	3.05 (8.39)
TL (SONICOM)	0.60 (4.11)	0.36 (5.42)	0.50 (7.28)	3.38 (7.28)
Baseline	1.17 (3.84)	1.57 (4.32)	2.22 (8.36)	-2.54 (23.84)
Target	0.93 (3.72)			

(b) Quadrant error comparison where the results are given as a percentage [%].

Method	Upsample Factor (No. original → upsampled) [%]			
	320 → 1280	**80 → 1280**	**20 → 1280**	**5 → 1280**
SRGAN (No TL)	8.33 (2.84)	9.96 (3.34)	12.39 (3.79)	12.84 (3.71)
TL (Synthetic)	8.59 (2.72)	9.01 (2.90)	12.35 (3.19)	12.67 (4.09)
TL (SONICOM)	8.30 (2.68)	7.78 (2.42)	12.29 (3.58)	14.15 (4.44)
Baseline	8.50 (2.68)	9.15 (2.73)	13.79 (3.76)	24.65 (7.28)
Target	8.03 (2.65)			

(c) Polar root mean square (RMS) error comparison where the results are given in degrees [°].

Method	Upsample Factor (No. original → upsampled) [°]			
	320 → 1280	**80 → 1280**	**20 → 1280**	**5 → 1280**
SRGAN (No TL)	32.97 (1.83)	35.46 (1.77)	35.89 (1.70)	36.51 (1.64)
TL (Synthetic)	32.84 (1.83)	33.75 (1.87)	35.27 (1.79)	35.93 (1.65)
TL (SONICOM)	32.48 (1.89)	33.64 (1.97)	35.69 (1.58)	36.66 (1.54)
Baseline	32.61 (1.70)	33.75 (1.68)	38.24 (1.36)	41.79 (1.22)
Target	32.26 (1.73)			

where $|H_{\mathrm{HR}}(f_w, x_n)|$ and $|H_{\mathrm{US}}(f_w, x_n)|$ represent the magnitude responses of the high-resolution and upsampled HRTF sets, W is the number of frequency bins in the HRTF, N is the number of locations, f_w is the frequency, and x_n is the location. In these results, the LSD is calculated for every measurement source position and then averaged over all the source positions.

Table 1 and Fig. 1 show the average results for the LSD evaluation over the ARI test set. In Fig. 1, the TL approaches show clear benefits over standard training. The most interesting observation, however, is that using the SONICOM Synthetic dataset of 1000 subjects for TL slightly outperforms using the acoustically measured SONICOM dataset of 200 subjects. This result is potentially highly advantageous as synthetic HRTF data is much quicker and cheaper to generate than undertaking costly acoustic measurements.

4.3.2 Model-based perceptual evaluation

To compare the localisation performance, a Bayesian model [27] for predicting human localisation performance was also used. Unlike the LSD metric, this model can evaluate the different approaches based on attributes that matter to human perception. This is important as not all errors in the LSD impact human localisation performance in the same way. To perform an effective comparison, the results for the original high-resolution measured HRTFs are provided as the 'Target' results. These 'Target' results are the best performance that can be achieved as it effectively compares the localisation performance of the original high-resolution HRTF with itself. Therefore the proposed method and the baseline need to be benchmarked against the 'Target' performance.

Similarly to the results shown for the LSD evaluation, Table 2 and Fig. 2 show the positive impact of using TL for HRTF upsampling. The TL approach using the acoustically measured data (SONICOM) can even outperform the barycentric baseline at every upsampling factor. It should also be noted that TL with both synthetic and acoustically measured data leads to similar performance gains, with the acoustically measured only slightly outperforming synthetic data in a few instances. This result could be explained considering the fact that the training uses more realistic data and should therefore lead to better perceptual performance but not necessarily a better performance in terms of LSD, which is the case here.

5. CONCLUSION

In this paper, it has been shown that SRGAN HRTF upsampling can exploit the use of TL to improve performance. Furthermore, it has also been shown that synthetic data works well for TL, sometimes outperforming acoustically measured data.

6. ACKNOWLEDGMENTS

This research is supported by the SONICOM project (EU Horizon 2020 RIA grant agreement ID: 101017743).

7. REFERENCES

[1] T. Lokki, H. Nironen, S. Vesa, L. Savioja, A. Härmä, and M. Karjalainen, "Application scenarios of wearable and mobile augmented reality audio," in *Proc. Audio Eng. Soc. (AES) Conv.*, May 2004.

[2] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. I: Stimulus synthesis," *J. Acoust. Soc. Am.*, vol. 85, no. 2, Feb. 1989.

[3] J. Blauert, "An introduction to binaural technology," in *Binaural and Spatial Hearing in Real and Virtual Environments*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, 1997, pp. 593–609.

[4] L. Picinali and B. F. G. Katz, "System-to-user and user-to-system adaptations in binaural audio," in Sonic interactions in virtual environments," in *Sonic Interactions in Virtual Environments*, M. Geronazzo and S. Serafin, Eds. Springer, 2022, pp. 121–144.

[5] I. Engel, R. Daugintis, T. Vicente, A. O. T. Hogg, J. Pauwels, A. J. Tournier, and L. Picinali, "The SONICOM HRTF dataset," *J. Audio Eng. Soc. (AES)*, June 2023.

[6] X.-L. Zhong and B.-S. Xie, *Head-Related Transfer Functions and Virtual Auditory Display*. Plantation, FL, USA: InTech, Mar. 2014.

[7] P. Siripornpitak, I. Engel, I. Squires, S. J. Cooper, and L. Picinali, "Spatial up-sampling of HRTF sets using generative adversarial networks: A pilot study," *Front. in Signal Process.*, vol. 2, 2022.

[8] A. O. T. Hogg, J. Mads, H. Liu, I. Squires, S. J. Cooper, and L. Picinali, "HRTF upsampling with a generative adversarial network using a gnomonic equiangular projection," *Submitted to IEEE/ACM Trans. Audio, Speech, Language Process.*, 2023. [Online]. Available: https://arxiv.org/abs/2306.05812

[9] A. O. T. Hogg, J. Mads, and H. Liu, 2023. [Online]. Available: https://github.com/ahogg/HRTF-upsampling-with-a-generative-adversarial-network-using-a-gnomonic-equiangular-projection

[10] F. Yu, X. Xiu, and Y. Li, "A survey on deep transfer learning and beyond," *Mathematics*, vol. 10, no. 19, p. 3619, Jan. 2022.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," in *Proc. Int. Joint Conf. on Learning Representations (ICLR)*, Apr. 2023.

[12] S. Ntalampiras, "Transfer learning for generalized audio signal processing," in *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity*, E. R. Miranda, Ed. Cham: Springer, 2021, pp. 679–691.

[13] J. Lv, G. Li, X. Tong, W. Chen, J. Huang, C. Wang, and G. Yang, "Transfer learning enhanced generative adversarial networks for multi-channel MRI reconstruction," *Comput. in Biology and Medicine*, vol. 134, p. 104504, July 2021.

[14] Y. Wang, C. Wu, L. Herranz, J. van de Weijer, A. Gonzalez-Garcia, and B. Raducanu, "Transferring GANs: Generating images from limited data," in *European Conf. on Comput. Vision (ECCV)*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Springer, 2018, pp. 220–236.

[15] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. of IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[16] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, Jan. 2020.

[17] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la 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, no. 3, p. e0211899, 2019.

[18] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, July 2017.

[19] D. P. Kingma and L. J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. Int. Joint Conf. on Learning Representations (ICLR)*, May 2015.

[20] L. Picinali, B. F. Katz, M. Geronazzo, P. Majdak, A. Reyes-Lecuona, and A. Vinciarelli, "The SONICOM Project: Artificial Intelligence-Driven Immersive Audio, From Personalization to Modeling [Applications Corner]," *IEEE Signal Process. Mag.*, vol. 39, no. 6, pp. 85–88, Nov. 2022.

[21] P. Majdak, "ARI HRTF database," June 2022. [Online]. Available: http://www.kfs.oeaw.ac.at/hrtf

[22] B. F. G. Katz, "Measurement and calculation of individual head-related transfer functions using a boundary element model including the measurement and effect of skin and hair impedance," Ph.D. dissertation, The Pennsylvania State University, 1998. [Online]. Available: https://hal.sorbonne-universite.fr/tel-02641309

[23] H. Ziegelwanger, W. Kreuzer, and P. Majdak, "Mesh2HRTF: An open-source software package for the numerical calculation of head-related transfer functions," in *Proc. Int. Cong. on Sound and Vibration (ICSV)*, July 2015, pp. 1–8.

[24] K. Pollack and P. Majdak, "Evaluation of a parametric pinna model for the calculation of head-related transfer functions," in *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, Sept. 2021, pp. 1–5.

[25] H. Ziegelwanger, P. Majdak, and W. Kreuzer, "Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization," *J. Acoust. Soc. Am.*, vol. 138, no. 1, pp. 208–222, 2015.

[26] J. Möbius and L. Kobbelt, "OpenFlipper: An open source geometry processing and rendering framework," in *Proc. Int. Conf. on Curves and Surfaces*, ser. Lecture Notes in Computer Science, J.-D. Boissonnat, P. Chenin, A. Cohen, C. Gout, T. Lyche, M.-L. Mazure, and L. Schumaker, Eds. Berlin, Heidelberg: Springer, 2012, pp. 488–500.

[27] R. Barumerli, P. Majdak, M. Geronazzo, D. Meijer, F. Avanzini, and R. Baumgartner, "A Bayesian model for human directional localization of broadband static sound sources," *Acta Acust.*, vol. 7, p. 12, 2023.