A Machine Learning Approach for Denoising and Upsampling HRTFs

Xuyi Hu*, Jian Li*, Lorenzo Picinali* and Aidan O. T. Hogg†*
*Audio Experience Design, Dyson School of Design Engineering, Imperial College London, UK
†Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

***Abstract*—The demand for realistic virtual immersive audio continues to grow, with Head-Related Transfer Functions (HRTFs) playing a key role. HRTFs capture how sound reaches our ears, reflecting unique anatomical features and enhancing spatial perception. It has been shown that personalized HRTFs improve localization accuracy, but their measurement remains time-consuming and requires a noise-free environment. Although machine learning has been shown to reduce the required measurement points and, thus, the measurement time, a controlled environment is still necessary. This paper proposes a method to address this constraint by presenting a novel technique that can upsample sparse, noisy HRTF measurements. The proposed approach combines an HRTF Denoisy U-Net for denoising and an Autoencoding Generative Adversarial Network (AE-GAN) for upsampling from three measurement points. The proposed method achieves a log-spectral distortion (LSD) error of 5.41 dB and a cosine similarity loss of 0.0070, demonstrating the method's effectiveness in HRTF upsampling.**

***Index Terms*—Head-Related Transfer Function, Generative Adversarial Network, Upsampling, Denoising.**

I. Introduction

We live in an ever more digital world where the need to create realistic, immersive audio is becoming ever more essential. The implications of being able to create convincing immersive audio virtually are broad, not only in helping enhance augmented and virtual meetings or video games, but also playing an essential role in improving assistive technologies. These include but are not limited to hearing aids [1] and advancing speech intelligibility algorithms [2].

One of the main challenges of immersive audio is adapting to individual listeners [3]. This individualization has resulted in a large amount of research focusing on Head-Related Transfer Functions (HRTFs). HRTFs describe how sound waves are modified as they propagate from a sound source to the listener's ears. This filtering is affected by a number of factors, including sound diffraction, reflection, and absorption by the listener's head and torso, as well as the resonances and shape of the listener's outer ears (pinnae). As a result, HRTFs contain both binaural cues, including interaural level differences (ILDs) and interaural time differences (ITDs), and monaural spectral cues, which are crucial for sound localization. Therefore, when the sound is appropriately filtered by the HRTF and presented at the entrance of the listener's ear canals, the listener should be able to perceive the sound originating from a specific source location [4]. This underscores the significant role HRTFs play in creating a realistic and immersive audio environment, particularly in applications such as virtual reality (VR) [5], [6] and augmented reality (AR) [7], [8].

Personalized HRTFs are closely tied to an individual's anatomy, with each person possessing a unique HRTF. Utilizing non-individualized HRTFs in virtual simulations often leads to poor sound source localization performance [9], [10]. To ensure an optimal user experience, acquiring a personalized HRTF is crucial. One common approach involves acoustic measurements [11], where sine sweeps are played from specific source locations, recorded at the listener's ears, and analyzed to extract impulse responses for generating the HRTF. However, this process requires specialized equipment and controlled environments, making it time-intensive [12].

To improve the efficiency and scalability of HRTF personalization, spatial up-sampling has been introduced to address the limitations of low-resolution HRTF data, which typically includes sparse measurements from limited directions. This technique generates high-resolution HRTFs by increasing measurement density, enhancing accuracy and coverage. Two common approaches are Barycentric interpolation [13] and spherical harmonic (SH) interpolation [14]. Barycentric interpolation uses weighted averages of known points to estimate values in unmeasured areas, while SH interpolation projects the HRTF onto spherical basis functions for smooth spatial representation. These methods have significantly advanced HRTF up-sampling, enabling more accurate and individualized sound localization.

In recent years, machine learning (ML) methods have also emerged as promising approaches for HRTF personalization [15]–[17]. Techniques such as variational autoencoders (VAEs) [18] encode HRTFs into a latent space to reduce dimensionality while preserving key features, enabling reconstruction and upsampling of low-resolution HRTFs by filling in missing details. Generative adversarial networks (GANs) [19], [20] use a generator to produce synthetic HRTFs and a discriminator to differentiate between real and synthetic data, learning data distributions to generate detailed high-resolution HRTFs. The SONICOM Listener Acoustic Personalization (LAP) Challenge 2024 demonstrated the superiority of ML techniques over traditional signal processing approaches [21].

Despite their advancements, these ML approaches still rely on clean and high-resolution data, necessitating noise-free environments for recording accurate HRTF measurements. To

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

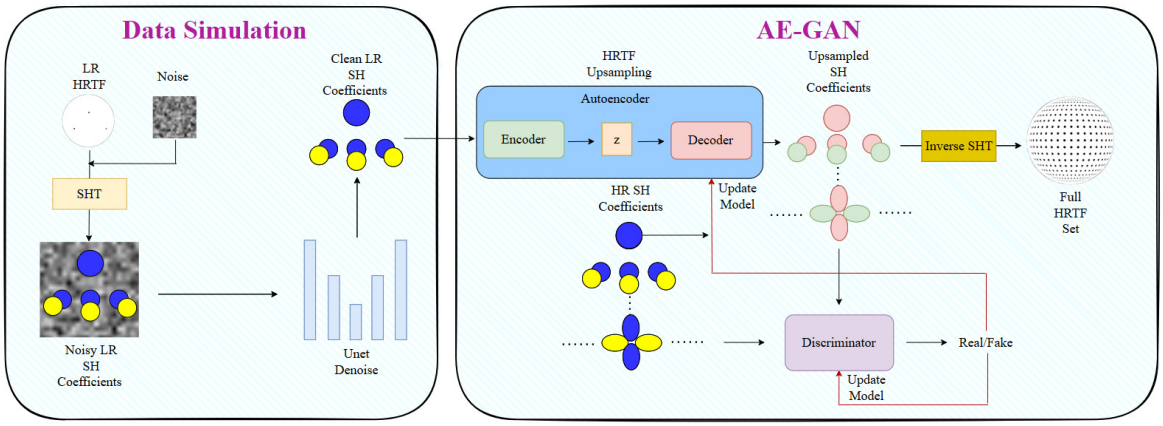

Fig. 1. HRTF-DUNet Flowchart. The left panel illustrates the data simulation pipeline, where noisy HRTF data is generated, segmented, and transformed using spherical harmonic analysis, resulting in low-resolution noisy coefficients stored in a dataset. The Denoisy U-Net then reconstructs clean SH coefficients from these inputs. The right panel presents the overall model framework. Red arrows indicate the AE-GAN training process, including the feedback loop for parameter updates, while black arrows represent the feedforward process through the model.

overcome this constraint, denoising (as well as upsampling) is needed, as real-world measurements are often degraded by background noise and room interactions, particularly in non-acoustically treated settings. Addressing these challenges is critical to improving accessibility, enabling accurate HRTF measurements, and expanding the adoption of immersive audio technologies and their applications. The contributions of this paper can, therefore, be broken down as follows:

1) We enhance the AE-GAN approach from the authors' previous work [20] on HRTF upsampling.
2) We employ an HRTF Denoisy U-Net for the task of denoising HRTFs measured in simulated noisy conditions.
3) We propose a novel end-to-end framework (HRTF-DUNet) and evaluate its performance against four baselines (AE-GAN without DUNet, Barycentric interpolation, SH interpolation, and HRTF selection) in terms of the LSD on the SONICOM HRTF dataset [11].

II. METHOD

A. HRTF Denoisy U-Net

The proposed approach improves SH interpolation by using an AE-GAN to increase the SH order. Therefore, the noisy HRTF data first needs to be transformed into the SH domain using the Spherical Harmonic Transform (SHT) as part of a pre-processing step [22]. The noisy, low-resolution SH coefficients are then passed to the HRTF Denoisy U-Net, which outputs the denoised low-resolution SH coefficients.

The Denoisy U-Net architecture is shown in Fig. 2 and consists of an initial convolutional block that processes the noisy input SH coefficients. Each convolutional block in the network applies a 1D convolutional layer followed by batch normalization and a ReLU activation function. The final layer of the network maps the refined features back to the original low-resolution SH coefficients.

B. AE-GAN

Next, the AE-GAN model from [20] is employed to upsample the denoised SH coefficients. The autoencoder consists of an encoder and a decoder network, where the encoder extracts the latent representation, z, of the low-degree SH coefficients, and the decoder reconstructs the high-resolution coefficients. As a refinement, we incorporate channel attention blocks within the residual blocks of the encoder, allowing the model to adaptively focus on important frequency components by dynamically weighting channel-wise features. Additionally, a discriminator is integrated into the model to distinguish between the SH coefficients produced by the generator and those from the real data, ensuring the authenticity of the generated outputs. To further enhance the diversity and sparsity of the generated SH coefficients and improve the realism of individualized HRTFs, we integrate minibatch discrimination [23] into the original discriminator network. This mechanism allows the model to assess not just individual samples but their variations within a batch, promoting higher sparsity of input and greater diversity in the generated outputs.

III. EXPERIMENTAL SETUP

A. Data Generation for Training

The training and testing data come from the SONICOM HRTF dataset [11], employing 793 positions per HRTF.

To simulate real-world HRTF measurement noise, we add both white and pink noise in the time domain to the clean HRTFs from the SONICOM dataset. Noise is added to the left and right sides of the HRTF independently before the impulse responses from the two ears are concatenated together. The noise component, $N(t)$, which can comprise either of white noise or pink noise, is defined as follows,

$$N_{\text{white}}(t) \sim \mathcal{N}(0, \sigma^2)\,, \tag{1}$$

where $N_{\text{white}}(t)$ represents the white noise at time t and $\mathcal{N}(0, \sigma^2)$ indicates that the white noise follows a gaussian distribution with Mean 0 and variance σ^2.

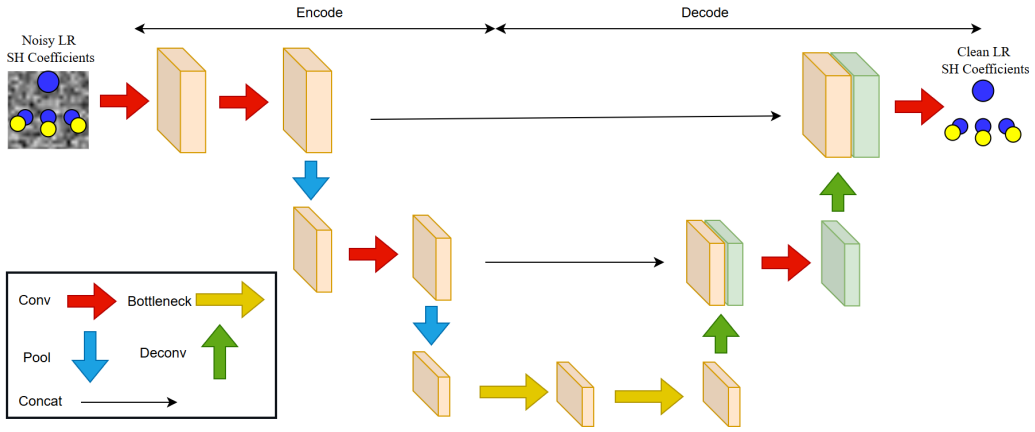

Fig. 2. Scheme of the proposed HRTF Denoisy U-Net.

$$N_{\text{pink}}(t) = \frac{1}{M} \sum_{i=1}^{M} \text{randn}_i(t) \,, \tag{2}$$

where M is the number of random sources, typically set to 16 in the Voss-McCartney algorithm. Noise is added at the desired signal-to-noise ratio (SNR) using the following approach.

$$\text{HRTF}_{\text{noisy, white}}(t) = \text{HRTF}(t) + N_{\text{white}}(t) \,, \tag{3}$$

where $\text{HRTF}_{\text{noisy, white}}(t)$ is the white noisy HRTF signal in the time domain and $\text{HRTF}(t)$ is the original HRTF signal.

$$\text{HRTF}_{\text{noisy, pink}}(t) = \text{HRTF}(t) + \frac{1}{M} \sqrt{\frac{P_{\text{signal}}}{\text{SNR}_{\text{linear}} \cdot P_{\text{noise}}}} \sum_{i=1}^{M} \text{randn}_i(t) \,. \tag{4}$$

where $\text{HRTF}_{\text{noisy, pink}}(t)$ is the pink noisy HRTF signal in the time domain, $P_{\text{signal}} = \frac{1}{T} \sum_{t=1}^{T} \text{HRTF}(t)^2$ and $P_{\text{noise}} = \frac{1}{T} \sum_{t=1}^{T} N(t)^2$ denote the average power of the original HRTF and noise, respectively. The signal-to-noise ratio in linear scale is given by $\text{SNR}_{\text{linear}}$, where $\text{randn}_i(t)$ represents the i-th random noise source and M is the total number of noise sources. The noisy HRTFs are then downsampled to generate the low-resolution noisy data for training.

B. Training

1) Denoisy U-Net Training: The model is trained using a combination of loss functions, primarily L1 loss,

$$\mathcal{L}_{L1} = \frac{1}{N} \sum_{i=1}^{N} \left| \text{SH}_{\text{denoised},i} - \text{SH}_{\text{target},i} \right| \,, \tag{5}$$

where $\text{SH}_{\text{target},i}$ is the ground truth clean SH coefficient, $\text{SH}_{\text{denoised},i}$ is the denoised SH coefficient, and N is the number of coefficients.

Additionally, a cosine similarity loss (CSL) is used to ensure that the angular similarity between the denoised SH coefficients and target coefficients is maximized. The cosine similarity loss is defined as,

$$\mathcal{L}_{\cos} = 1 - \frac{\sum_{i=1}^{N} \text{SH}_{\text{denoised},i} \cdot \text{SH}_{\text{target},i}}{\sqrt{\sum_{i=1}^{N} \text{SH}_{\text{denoised},i}^2} \cdot \sqrt{\sum_{i=1}^{N} \text{SH}_{\text{target},i}^2}} \,. \tag{6}$$

2) AE-GAN Training: The discriminator is trained via supervised learning, utilizing both generated and real HRTF data, and aims to guide the autoencoder to produce high-fidelity results. In this study, we further extend the application of AE-GAN by expanding the range of sparcity levels to include 4 points and 3 points, thus demonstrating the model's robustness and scalability across a broader spectrum of resolutions.

3) HRTF-DUNet Training: We employ cascaded backpropagation for end-to-end training, allowing smooth gradient flow between the U-Net denoiser and AE-GAN upsampler. Trained on 162 noisy subjects, this setup ensures the U-Net effectively denoises SH coefficients to support accurate upsampling by AE-GAN, particularly under extreme sparsity conditions.

C. Baselines

The performance of the proposed approach is compared against four baselines: AE-GAN without DUNet, barycentric interpolation, SH interpolation, and non-individual HRTF selection. The AE-GAN approach is presented in [20]. Barycentric interpolation estimates unknown values by computing weighted averages of known points using three barycentric coordinates. SH interpolation, widely used for HRTF upsampling [22], projects HRTF data onto SH for smooth spatial representation. An alternative to personalized HRTF modeling selects the closest match from a database. Following [19], Selection-1 represents the most 'generic' HRTF, while Selection-2 is the most 'distinct'.

D. Evaluation Metrics

Three metrics are used for evaluating the performance.

1) Interaural level difference (ILD): The ILD represents the interaural level difference, which is the difference in sound pressure level between the two ears for a given frequency f_b, number of spatial locations N, total number of frequency bins B and direction x_n, calculated by,

$$\text{ILD} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{B} \sum_{b=1}^{B} \left| \left(20 \log_{10} \frac{|H_{\text{LR}}^{\text{Left}}(f_b, x_n)|}{|H_{\text{LR}}^{\text{Right}}(f_b, x_n)|} \right) - \left(20 \log_{10} \frac{|H_{\text{DN}}^{\text{Left}}(f_b, x_n)|}{|H_{\text{DN}}^{\text{Right}}(f_b, x_n)|} \right) \right| \,, \tag{7}$$

The terms $|H^{\text{Left}}(f_b, x_n)|$ and $|H^{\text{Right}}(f_b, x_n)|$ denote the magnitude responses for the left and right ears, respectively. Similarly, $|H_{\text{LR}}(f_b, x_n)|$ and $|H_{\text{DN}}(f_b, x_n)|$ represent the magnitude responses for the low-resolution and denoised HRTF sets.

2) Interaural time difference (ITD): The ITD, quantifies the arrival time gap of a sound wave between the left and right ears for the same frequency and direction, given by,

$$\text{ITD} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{B}\sum_{b=1}^{B}\left|\left(\frac{\phi_{\text{LR}}^{\text{Left}}(f_b, x_n) - \phi_{\text{LR}}^{\text{Right}}(f_b, x_n)}{2\pi f_b}\right) - \left(\frac{\phi_{\text{DN}}^{\text{Left}}(f_b, x_n) - \phi_{\text{DN}}^{\text{Right}}(f_b, x_n)}{2\pi f_b}\right)\right|, \quad (8)$$

where $\phi_{\text{LR}}^{\text{Left}}(f_b, x_n)$ and $\phi_{\text{LR}}^{\text{Right}}(f_b, x_n)$ represent the phase responses of the low-resolution HRTF for the left and right ears. Similarly, $\phi_{\text{DN}}^{\text{Left}}(f_b, x_n)$ and $\phi_{\text{DN}}^{\text{Right}}(f_b, x_n)$ correspond to the denoised HRTF phase responses.

3) Log-spectral distortion (LSD): The LSD [24] is an evaluation metric utilized to assess the quality of a synthesized audio signal relative to a reference audio signal. In this context, LSD is employed to evaluate the denoising and upsampling performance of HRTFs using the proposed HRTF-DUNet framework. The LSD loss quantifies this comparison by evaluating the discrepancy between the target magnitude spectrum H_{HR} and the generated spectrum H_{G}. This computation can be expressed in the following way,

$$\text{LSD} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{W}\sum_{w=1}^{W}\left(20\log_{10}\frac{|H_{\text{HR}}(f_w, x_n)|}{|H_{\text{G}}(f_w, x_n)|}\right)^2}, \quad (9)$$

where N represents the overall count of positions, and x_n corresponds to a specific position.

IV. Experimental Results

Two experiments were performed to evaluate the newly proposed HRTF-DUNet model. These experiments utilised 41 test subjects (HRTFs) not seen in training from the SONICOM dataset, where white noise was added at an SNR of 5dB.

A. Denoising Evaluation

The proposed Denoisy U-Net model for HRTF denoising is evaluated against three baseline methods: Spectral Subtraction [25], Wavelet [26] and Kalman [27] Filtering. Table I presents the performance comparison across three evaluation metrics: CSL (detailed in Section III-B1), ILD (outlined in Section III-D1), and ITD (described in Section III-D2). Fig. 3 shows results of the HRTF Denoisy U-Net.

The results demonstrate that the HRTF Denoisy U-Net model outperforms the baselines across all metrics. For example, CSL, which measures the similarity between the denoised and target HRTFs, is significantly lower for the HRTF Denoisy U-Net model (0.007), indicating a higher degree of similarity and better denoising capability. Additionally, the U-Net model shows superior performance in preserving ILDs and ITDs, with the lowest deviations of 19.757 and 1.301, respectively. These results indicate that the denoising process effectively preserves critical spatial cues that are contained within the HRTFs and which are needed for realistic, immersive audio.

TABLE I
A COMPARISON OF HRTF DENOISY U-NET AND BASELINES WITH DIFFERENT EVALUATION METRICS ('BEST' RESULT HIGHLIGHTED).

Method	CSL	ILDs	ITDs
HRTF Denoisy U-Net	0.007	19.757	1.301
Wavelet Filtering with dB7	0.283	24.591	2.783
Wavelet Filtering with Gaus3	0.213	22.178	2.846
High-Pass Spectral Subtraction	0.339	23.936	2.946
Kalman Filtering	0.206	20.491	2.152

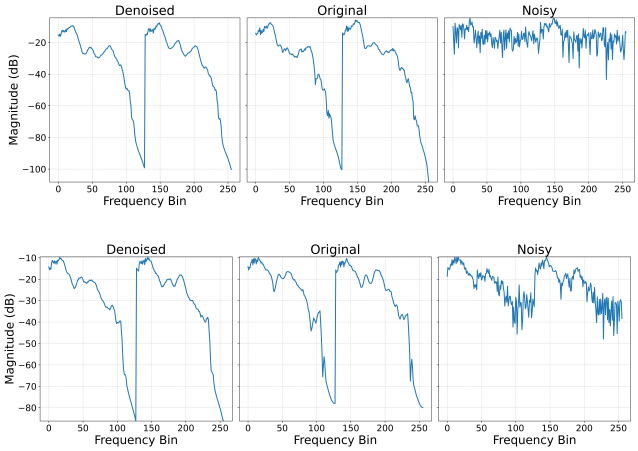

Fig. 3. Two illustrative examples (top and bottom) showcasing the HRTF Denoisy U-Net's performance on two different subjects at the same measurement location, with additive white Gaussian noise applied at an SNR of 5 dB.

B. LSD Evaluation

Second, the HRTF-DUNet model is evaluated using the LSD metric (see Section III-D3), averaging the LSD across all measurement positions. Table II presents the results for 41 noisy test subjects, with a visualization in Fig. 4.

The results show that the HRTF-DUNet model consistently achieves a lower LSD error at sparcity levels $4 \rightarrow 793$ and $3 \rightarrow 793$, where it significantly outperforms other baselines. This suggests that HRTF-DUNet effectively denoises and learns patterns in noisy HRTF features, even in extremely sparse conditions. Unlike traditional interpolation methods, which rely on a predefined spatial structure, the deep-learning-based approach can generalize from the available data and reconstruct a more accurate HRTF. Barycentric and SH interpolation yield higher LSD errors at extreme sparsity levels as their geometric assumptions break down. However, with sufficient initial points, they perform well by leveraging spatial smoothness and structured mathematical formulations for accurate interpolation. AE-GAN struggles at high sparsity levels due to insufficient spatial information and noise in the limited initial HRTF points, preventing effective feature learning. However, as sparsity decreases, AE-GAN benefits from more input data, enabling better feature extraction and improved high-resolution HRTF reconstruction. The HRTF selection approach performs poorly, with LSD errors of 6.31 and 8.33 for Selection-1 and Selection-2, respectively. This reinforces the limitation of non-individualized HRTFs, emphasizing the need for personalization to achieve realistic virtual audio.

TABLE II
A COMPARISON OF THE MEAN LSD ERROR (STANDARD DEVIATION) FOR DIFFERENT SPARSITY LEVELS ('BEST' PERFORMANCE HIGHLIGHTED).

Method	Upsampling [No. intial → No. upsampled]				
	27 → 793	**18 → 793**	**8 → 793**	**4 → 793**	**3 → 793**
HRTF-DUNet	5.23 (0.19)	5.58 (0.28)	6.06 (0.32)	5.43 (0.45)	5.41(0.41)
AE-GAN	7.74 (0.41)	8.20 (0.49)	8.76 (0.55)	9.70 (0.56)	9.89 (0.51)
SH	5.12 (0.27)	5.54 (0.31)	7.54 (0.37)	12.46 (0.39)	12.41 (0.44)
Barycentric	4.89 (0.24)	5.46 (0.27)	7.22 (0.35)	10.07 (0.43)	11.69 (0.47)
Selection-1	6.31 (0.59)				
Selection-2	8.33 (0.47)				

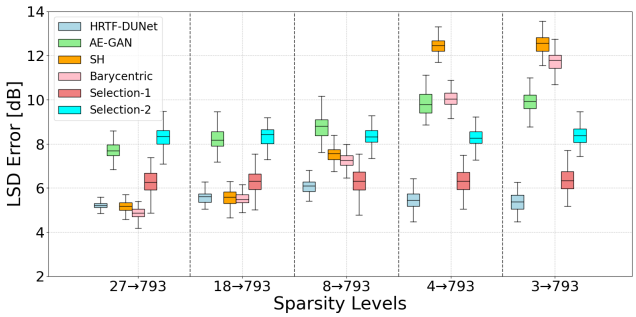

Fig. 4. Log-spectral distortion (LSD) error comparison.

V. CONCLUSION AND FUTURE WORK

This paper introduces a novel framework using the HRTF-DUNet model for simultaneous HRTF denoising and upsampling, simplifying the measurement process for personalised HRTFs. To the best of our knowledge, this is the first work to address the problem of HRTF denoising, demonstrating its feasibility and advantages in improving HRTF quality. The proposed method outperforms other approaches by effectively denoising and upsampling three measurement points with 5 dB of additive white noise into a high-resolution, clean HRTF.

This work serves as a proof of concept, showing that denoising is both feasible and helpful for HRTF upsampling. Future work will move beyond simulated white and pink noise, applying the method to more realistic synthetic and recorded noise to better reflect real-world conditions.

Additionally, challenges still remain in measuring HRTFs in uncontrolled environments, particularly problems with reverberation and frequency range limitations. These issues arise due to the reverberation present when measuring in untreated rooms and the frequency limitations imposed by the speakers used for recording the HRTFs. We aim to address these challenges through future work, including perceptual evaluations with auditory models and listening tests to validate the model's effectiveness in practical immersive audio settings.

REFERENCES

[1] J. Foerster, H. Janning, S. Nagel, *et al.*, "Reduced-complexity binaural source localization for headphones and hearing aids using low-rank DRTF approximations," in *Speech Commun.* VDE, 2023, pp. 91–95.

[2] E.-L. Tan, S. Peksi, and W.-S. Gan, "Implementing continuous HRTF measurement in near-field," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2023, pp. 1–5.

[3] L. Picinali and B. F. Katz, "System-to-user and user-to-system adaptations in binaural audio," in *Sonic interactions in virtual environments*. Springer International Publishing Cham, 2022, pp. 115–143.

[4] J. Blauert and R. A. Butler, "Spatial hearing: The psychophysics of human sound localization," *J. Acoust. Soc. Am.*, vol. 77, no. 1, pp. 334–335, Jan 1985.

[5] D. Brahimaj, E. Vezzoli, F. Giraud, *et al.*, "Enhancing object localization in VR: Tactile-based HRTF and vibration headphones for spatial haptic feedback," *IEEE Trans. on Haptics (ToH)*, 2024.

[6] M. Ramírez, J. M. Arend, P. von Gablenz, *et al.*, "Toward sound localization testing in virtual reality to aid in the screening of auditory processing disorders," *Trends in Hearing*, vol. 28, p. 23312165241235463, 2024.

[7] A. G. Privitera, F. Fontana, and M. Geronazzo, "On the effect of user tracking on perceived source positions in mobile audio augmented reality," in *ACM SIGCHI Italian Chapter Int. Conf. on Comput.-Human Interaction*, 2023, pp. 1–9.

[8] A. G. Privitera, M. Noro, M. Geronazzo, *et al.*, "Preliminary evaluation of the auralization of a real indoor environment for augmented reality research," in *Conv. Eur. Acoust. Assoc.*, 2023.

[9] P. Stitt, L. Picinali, and B. F. Katz, "Auditory accommodation to poorly matched non-individual spectral localization cues through active learning," *Scientific Reports*, vol. 9, no. 1, p. 1063, 2019.

[10] E. M. Wenzel, M. Arruda, D. J. Kistler, *et al.*, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111–123, 1993.

[11] I. Engel, R. Daugintis, T. Vicente, *et al.*, "The SONICOM HRTF dataset," *J. Audio Eng. Soc. (AES)*, vol. 71, no. 5, pp. 241–253, 2023.

[12] S. Li and J. Peissig, "Measurement of head-related transfer functions: A review," *Appl. Sci.*, vol. 10, no. 14, p. 5014, 2020.

[13] K. Hartung, J. Braasch, and S. J. Sterbing, "Comparison of different methods for the interpolation of head-related transfer functions," in *Proc. Audio Eng. Soc. (AES) Conf. on Spatial Sound Reproduction*. Audio Engineering Society, 1999.

[14] I. Engel, D. F. Goodman, and L. Picinali, "Assessing HRTF preprocessing methods for ambisonics rendering through perceptual models," *Acta Acust.*, vol. 6, p. 4, 2022.

[15] T.-Y. Chen, T.-H. Kuo, and T.-S. Chi, "Autoencoding HRTFs for dnn-based HRTF personalization using anthropometric features," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2019, pp. 271–275.

[16] P. Siripornpitak, I. Engel, I. Squires, *et al.*, "Spatial up-sampling of HRTF sets using generative adversarial networks: A pilot study," *Front. in Signal Process.*, p. 54, 2022.

[17] A. O. T. Hogg, H. Liu, M. Jenkins, *et al.*, "Exploring the impact of transfer learning on GAN-based HRTF upsampling," in *Proc. EAA Forum Acusticum, Eur. Congress on Acoust.*, 2023.

[18] R. Lopez, J. Regier, M. I. Jordan, *et al.*, "Information constraints on auto-encoding variational bayes," *Advances in Neural Inform. Process. Systems*, vol. 31, 2018.

[19] A. O. T. Hogg, M. Jenkins, H. Liu, *et al.*, "HRTF upsampling with a generative adversarial network using a gnomonic equiangular projection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, 2024.

[20] X. Hu, J. Li, L. Picinali, *et al.*, "HRTF spatial upsampling in the spherical harmonics domain employing a generative adversarial network," *Proc. Conf. on Digital Audio Effects*, 2024.

[21] A. O. T. Hogg, R. Barumerli, R. Daugintis, *et al.*, "Listener acoustic personalisation challenge - LAP24: Head-related transfer function upsampling," *Open J. Signal Process.*, vol. (submitted), 2025.

[22] J. M. Arend, F. Brinkmann, and C. Porschmann, "Assessing spherical harmonics interpolation of time-aligned head-related transfer functions," *J. Audio Eng. Soc. (AES)*, vol. 69, no. 1/2, pp. 104–117, 2021.

[23] T. Salimans, I. Goodfellow, W. Zaremba, *et al.*, "Improved techniques for training GANs," *Proc. Neural Inform. Process. Conf*, vol. 29, 2016.

[24] P. Gutierrez-Parera, J. J. Lopez, J. M. Mora-Merchan, *et al.*, "Interaural time difference individualization in HRTF by scaling through anthropometric parameters," *EURASIP J. on Audio, Speech, and Music Process.*, vol. 2022, no. 1, pp. 1–19, 2022.

[25] R. Martin, "Spectral subtraction based on minimum statistics," *power*, vol. 6, no. 8, pp. 1182–1185, 1994.

[26] J. D. Villasenor, B. Belzer, and J. Liao, "Wavelet filter evaluation for image compression," *IEEE Transactions on image processing*, vol. 4, no. 8, pp. 1053–1060, 1995.

[27] G. Welch, G. Bishop, *et al.*, "An introduction to the kalman filter," 1995.