HRTF upsampling with a generative adversarial network using a gnomonic equiangular projection

Aidan O. T. Hogg, Mads Jenkins, He Liu, Isaac Squires, Samuel J. Cooper, and Lorenzo Picinali

***Abstract*—An individualised head-related transfer function (HRTF) is very important for creating realistic virtual reality (VR) and augmented reality (AR) environments. However, acoustically measuring high-quality HRTFs requires expensive equipment and an acoustic lab setting. To overcome these limitations and to make this measurement more efficient HRTF upsampling has been exploited in the past where a high-resolution HRTF is created from a low-resolution one. This paper demonstrates how generative adversarial networks (GANs) can be applied to HRTF upsampling. We propose a novel approach that transforms the HRTF data for direct use with a convolutional super-resolution generative adversarial network (SRGAN). This new approach is benchmarked against three baselines: barycentric upsampling, spherical harmonic (SH) upsampling and an HRTF selection approach. Experimental results show that the proposed method outperforms all three baselines in terms of log-spectral distortion (LSD) and localisation performance using perceptual models when the input HRTF is sparse (less than 20 measured positions).**

***Index Terms*—generative adversarial network, head-related transfer function, super-resolution, upsampling, interpolation.**

I. Introduction

REMOTE interaction has grown in use in recent years, however, there are still many unsolved problems with remote connectivity. A common issue is the lack of immersive audio in these virtual interactions. Immersive audio is what people experience in their everyday lives; some sounds are close, some are far away, some are moving, some are static, and all come from different directions. The loss of the acoustic spatial dimension, as well as the physical interactions with sound, can lead to the frustration people often feel when communicating remotely (e.g. [1]). This is immediately apparent in online meetings when multiple participants try to speak simultaneously or struggle to hear what others are saying. This problem worsens when some participants are present in person and others online. However, this need not be the case, and realistic immersive audio attempts to contribute to solving this problem by making the transition from real to virtual audio seamless.

Furthermore, improvements to online communication are not the only aim. In recent years, the prevalence of virtual reality (VR) and augmented reality (AR) devices, as well as three-dimensional (3D) video games, auditory displays and hearing assistive devices [2], has led to a demand for more realistic 3D audio rendering. Therefore, the need for better immersive audio solutions is becoming increasingly relevant to the modern world.

One way to achieve high realism in immersive audio is to place the listener in the centre of a large spherical loudspeaker array and play different sounds from different directions in space. Unfortunately, although this solution works well in an acoustic lab and could be used in cinemas and other large venues, it is not practical beyond these controlled settings. It is also costly and only works well for the small number of participants at the centre of the array. A more practical solution exploits the fact that humans have two auditory sensors (i.e. two ears); in theory, we should only need two speakers (i.e. in-ear headphones) to generate the correct sound at those sensors, performing what is commonly referred to as binaural (i.e. involving the two ears) spatialisation.

A. The matter of individualisation

The main challenge with binaural spatialisation is understanding how sounds at the entrance of the two ear canals can be realistically generated to mimic real-world 3D audio accurately [3] and, more specifically, how this can be adapted for individual listeners. This individualisation has resulted in a large amount of research focusing on head-related transfer functions (HRTFs), which capture the filtering effects related to the anatomy of different listeners. This filtering is caused by the sound wave being reflected and scattered off the head, torso, and pinnae before it enters the ear canal of a given listener. HRTFs are, therefore, able to capture interaural (i.e. the difference heard between the listener's two ears) and monaural localisation cues [4]. It is customary to refer to HRTF when considering the impulse response (IR) in the frequency domain and head-related impulse response (HRIR) in the time domain. In this paper, we also use the term HRTFs to refer to the complete set of IR measurements corresponding to a full set of source positions around each listener.

This study was made possible by support from SONICOM (www.sonicom.eu), a project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017743.

A. O. T. Hogg is with both the Centre for Digital Music, Queen Mary University of London, 327 Mile End Road, London, E1 4NS, UK and the Audio Experience Design Group, Dyson Sch. of Design Engineering, Imperial College London, Exhibition Road, SW7 2AZ, UK (e-mail: aidan@aidanhogg.uk).

Lorenzo Picinali is with the Audio Experience Design Group, Dyson Sch. of Design Engineering, Imperial College London, Exhibition Road, SW7 2AZ, UK (e-mail: l.picinali@imperial.ac.uk).

Mads Jenkins and He Liu were with the Audio Experience Design Group, Dyson Sch. of Design Engineering, Imperial College London, Exhibition Road, SW7 2AZ, UK.

Isaac Squires and Samuel J. Cooper are with the Tools for Learning, Design and Research Group, Dyson Sch. of Design Engineering, Imperial College London, Exhibition Road, SW7 2AZ, UK (e-mail: i.squires20@imperial.ac.uk and samuel.cooper@imperial.ac.uk).

It has been shown in the past that using non-individualised HRTFs can significantly affect the individual's sound source localisation accuracy [5]–[7] as spectral cues are highly dependent on the listener's anatomy, particularly the shape of their pinnae [8]. This non-individualisation can also affect perceptual attributes such as externalisation, immersion, colouration, realism and relief/depth [9]–[11]. Furthermore, the choice of HRTF can significantly impact an individual's ability to understand speech in a cocktail party scenario [12].

As a result, capturing individual listeners' personalised HRTF remains an important area of active research. It has been shown in the past that many approaches can be deployed for this task of HRTF individualisation, including acoustic measurements [13], 3D scans [14], modelling the morphological geometric information of the listener's ears [15], [16] and selecting the best-fitting HRTF from a database of already-measured ones. This best-fitting HRTF selection is often made using morphology-based methods [17], [18] or perceptual-based methods (e.g. using individual preference [19] and/or localisation accuracy [20], [21]). An overview of some of the most common methods can be found in [22].

The acoustic measurement [23] is still considered the gold standard of these different approaches. The downside to performing this acoustic measurement is the expensive custom setup required and the time it takes. This is because numerous IRs need to be measured around the individual, with numbers ranging anywhere from 200 to 3000 [13]. This process can be sped up by taking advantage of interlaced sine sweeps [24], but this often only makes the elevation measurements faster. Other methods do exist [25], [26] that aim to improve the time performance of the HRTF measurement, but the equipment specifications and cost are usually very high.

B. Spatial upsampling of HRTFs

To reduce the time required and the complexity of the HRTF setup and to make the method scalable, spatial upsampling methods have been proposed in the past that can generate high-resolution HRTFs, i.e. HRTFs that contain many (normally over 300) IRs from many directions, from low-resolution HRTFs, i.e. HRTFs that include very few IRs from very few directions [27]. This process is commonly referred to as HRTF upsampling and can be achieved using various approaches.

The most common HRTF upsampling method is barycentric interpolation [28]–[30]. This method has been shown to produce good results when the HRTFs contain a relatively large number of IRs [31], for example, with an angular distance of 10-15°between measurements; however, it becomes much less reliable when interpolating sparser measurements (e.g. each 30-40°). Another common approach is spherical harmonic (SH) interpolation [32]–[36], but again results in poor reconstruction when the low-resolution HRTF input is spatially sparse. This is because these methods rely on averaging between existing data points based on prior information. For example, barycentric interpolation uses the three nearest neighbours around the point to be interpolated to calculate the weighted average. Therefore, as the distances between the neighbours grow larger, the upsampling becomes more and more inaccurate.

More recently, machine learning (ML) methods have started to become the focus of research on HRTF personalisation. In the past, ML techniques have been shown to be effective at estimating HRTFs from just the anthropometric measurements of the listener. In [37], a deep neural network (DNN)-based approach is used to synthesise a personalised HRTF using the anthropometric features of the user and was able to achieve a performance of 3.2 dB log-spectral distortion (LSD). This approach consisted of using the encoder part of an autoencoder to reduce the dimensionality of the raw HRTFs, which can be used as a set of input features for training. This aims to minimise overfitting as HRTF datasets are usually small. The decoder part of the autoencoder then estimates the HRTF magnitudes using the output of a DNN that is trained to output the latent representation given the anthropometric features and the target azimuth. This type of approach is explored further in [38], where two autoencoders are exploited: one that reduces the dimensionality of a feature vector containing the azimuth and anthropometric parameters and another that reduces the dimensionality of the magnitudes of the full HRTF measurement. These two autoencoders are combined to estimate the HRTF magnitudes from anthropometric features and achieve a performance of 4.3 dB LSD.

It has also been shown in the past that autoencoders can be used to upsample low-resolution HRTFs. In [39], a method is proposed that uses an autoencoder and is an extension of a regularised linear regression (RLR) approach that makes use of the spherical wavefunction presented in [40]. This method's key feature is that it decomposes HRTFs into source position-dependent and source position-independent factors, i.e. the spherical wavefunction expansion and expansion coefficients, respectively. The autoencoder is conditioned on source positions and obtains the source-position-independent representation by using an aggregation module between the encoder and decoder, aggregating latent variables of a given source position. This approach was able to notably achieve 4.4 dB in LSD when upsampling from 9 to 440 positions. Other ML methods also exist that are able to perform the task of HRTF upsampling, including [41], which exploits a deep belief network (DBN). This method accomplishes an LSD average of less than 3 dB; however, results are only given for upsampling from 125 source positions to 1250, which is still relatively dense. Another method in [42] uses a convolutional neural network (CNN) and has been shown to yield a good performance of 4.4 dB LSD when upsampling from 23 positions to 1250 and 3.8 dB when upsampling from 105 positions. However, in this case, the sphere is sliced into planes to create a two-dimensional (2D) representation rather than considering the sphere as a whole. ML techniques have also been used in the past as part of a postprocessing step of the spherical harmonic transform (SHT) interpolation [43].

C. HRTFs and GANs - our proposed solution

The main advantage of ML approaches over traditional upsampling is that they can extrapolate patterns from the data

rather than these patterns being hard-coded, making it possible to recreate the missing information in the sparse measurements using the knowledge learnt from a training set that contains many high-resolution HRTFs. The aim of this paper is to investigate the use of generative adversarial networks (GANs) to tackle the HRTF upsampling problem, explicitly looking at very sparse HRTF measurements and offering insight into the practicality of this approach. In the past, GANs have been successfully applied to many audio applications, including WaveGAN [44], which applies GANs to the unsupervised synthesis of raw-waveform audio. GANs were also used in [45] for speech super-resolution, which aims to upsample a given speech signal by generating the missing high-frequency content. These applications of GANs are motivated by the fact that GANs have been successfully applied to the task of upsampling photos [46], [47] and astronomical images [48] and are often referred to as super-resolution generative adversarial networks (SRGANs). SRGANs [49] are a family of ML models characterised by the use of two networks that compete in an adversarial manner. These models have been shown to work well on upsampling very low-resolution images; however, apart from a pilot study [50], they have not been exploited for the task of HRTF upsampling.

A novel approach is proposed here using the SRGAN framework, as introduced by [47], to allow the generation of accurate high-quality HRTFs from sparsely measured ones, thus making this personal acoustic data available faster and at a lower cost, albeit requiring a small number of measurements anyway. The paper builds on a pilot study that was undertaken in [50], which explored using an SRGAN for upsampling HRTFs across single planes in space, e.g. the horizontal, median and vertical planes. This limitation is overcome in the study presented here, where the full 3D HRTF is employed for the SRGAN training and prediction. The next steps to further validate this technique, extend it, and ultimately integrate it within a tool to be openly released are outlined at the end of the paper.

The first challenge that was tackled was to transform the original HRTF data into something more suitable for the SRGAN, and this was achieved through various transformations and resampling operations. The main transformation is that of a gnomonic equiangular projection [51]–[53], often referred to as a cubed sphere. The reason this type of projection was selected is that it does not produce singularities at the poles [54], and the distortion is quasi-uniform over the whole sphere [55].

The transformed HRTF was then used to train the SRGAN, for which an updated loss function was actually designed. Finally, an evaluation was carried out by spatially upsampling a certain number of low-resolution HRTFs and comparing the results with various benchmark techniques (e.g. barycentric and spherical harmonics interpolations). The comparison relied on both signal-level metrics and model-based perceptual evaluations.

This paper is structured as follows: Section II introduces the method, including the pre- and post-processing steps along with the GAN architecture. Section III explains the experimental setup, that is, the data used, how the GAN was trained, and an explanation of the baselines that were used for comparison. In Section IV, spectral and perceptual model-based results are presented. Finally, Section V provides the conclusions drawn.

II. METHOD

A. Data pre-processing

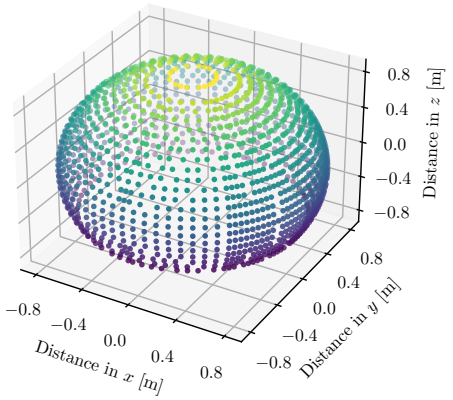
SRGANs have been shown to perform well on the task of upsampling images. The main challenge when it comes to upsampling HRIRs is that the data occupies an extra dimension in space compared to a 2D image. Another issue is that CNNs are designed for applications using uniformly spaced data (such as the pixels in conventional 2D images). In contrast, the IRs in an HRIR are spaced non-uniformly on the surface of a sphere. In particular, HRIRs often contain no measurements at lower elevations, and the number of measurements is denser around the horizontal plane.

Various approaches exist for processing non-uniformly distributed spatial data, such as graph neural networks (GNNs) [56]. However, in order to exploit the vast literature that exists for the upsampling of images, in this study, we apply a pre-processing step to convert the spherical data into a form that a standard CNN can process.

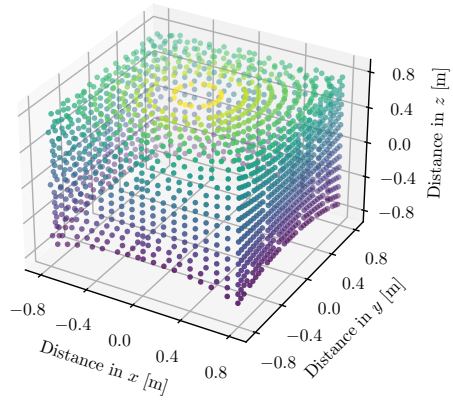
Two main steps are required to convert the HRIR data into a format that can be exploited by a CNN architecture. First, the spherical HRIR data needs to be projected onto a 2D surface to remove the extra dimension (see Section II-A1). Second, an interpolation is utilised to shift the irregularly spaced IRs onto an evenly spaced Cartesian grid (see Section II-A3). This strategy has the advantage of mapping any HRIR dataset to the same cartesian grid, and as a result, any dataset can be deployed for training and testing the SRGAN [57].

In addition to these two steps, the phase and interaural time differences (ITDs) were disregarded by taking the HRIR into the frequency domain, referred to as the HRTF, and only considering the magnitude component from each IR in the HRIR. These additional simplifying pre-processing steps can be performed because the up-sampled HRTFs can be effectively reconstructed using a minimum-phase approximation and a simple ITD model [30]. However, it should be noted that such simplifications could have an impact on certain perceptual features of the HRTFs [58]. For this reason, future advancements in this technique should aim to include phase information when performing the upsampling.

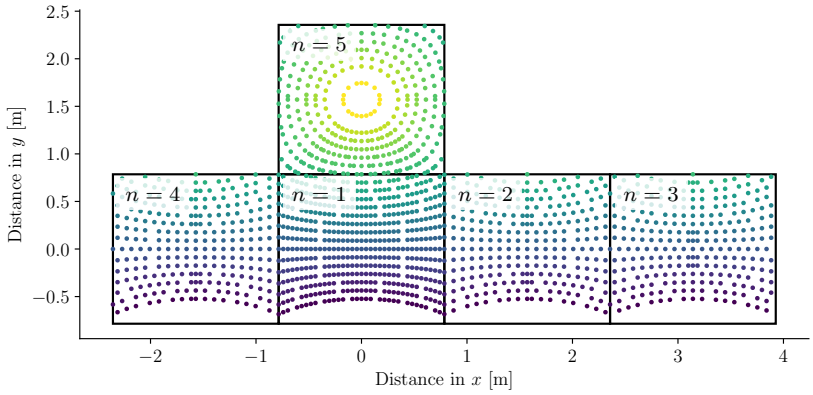
1) Gnomonic equiangular projection: A gnomonic equiangular projection [59] is used to project the locations of the IRs, seen in Fig. 1(a), in the HRIR to a cube, shown in Fig. 1(b), which can then be flattened as shown in Fig. 1(c). This process creates five panels where identical local curvilinear coordinates are constructed for each panel [51]. It should be noted that the 6^{th} (bottom) panel is removed as it contains no IR measurements as the HRIRs are usually not measured below the listener. More information about this choice can be found later on.



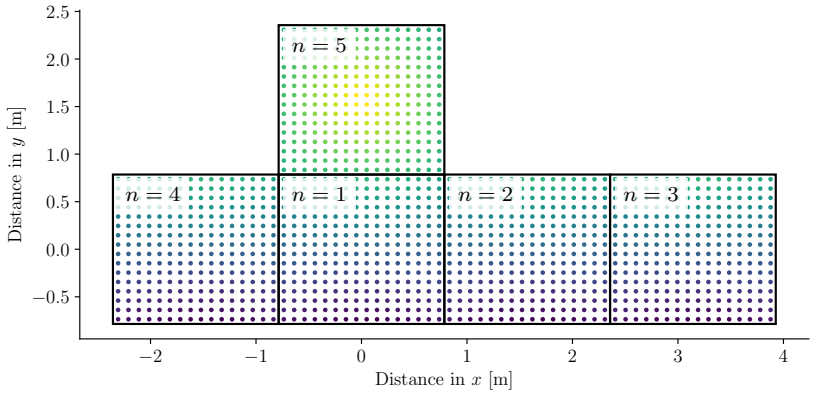
(a) Original positions of the IRs.



(b) Gnomonic equiangular projected positions of the IRs.



(c) Flattened cube of projected positions of the IRs



(d) Barycentric interpolation of flattened cube projected positions of the IRs.

Fig. 1. The four steps that project the impulse response (IR) locations for a given head-related transfer function (HRTF) in the ARI dataset onto a flattened uniform cube.

The gnomonic equiangular projection [59] transforms the location of any point on the sphere into Cartesian coordinates x and y using

$$x = r\left(\theta - \frac{(n-1)}{2}\pi\right) , \tag{1}$$

$$y = r\arctan\left[\tan(\phi)\sec\left(\theta - \frac{(n-1)}{2}\pi\right)\right] . \tag{2}$$

where n represents the equatorial panel on the horizontal plane that the point on the sphere would map onto, which is determined based on the original elevation and azimuth and can only take the values 1, 2, 3 and 4. However, the location of any point on the top of the sphere, i.e. when $n = 5$, is mapped to Cartesian coordinates x and y using

$$x = r\arctan\left(\sin(\theta)\cot(\phi)\right) , \tag{3}$$

$$y = r\arctan\left(-\cos(\theta)\cot(\phi)\right) , \tag{4}$$

where θ, ϕ and r correspond to the azimuth, elevation and radius of each point on the original sphere. More precisely, the azimuth θ is defined as the angle between the projection of the source direction in the horizontal plane and the front axis in the range of $[\pi, -\pi]$ going from left to right, and elevation ϕ is defined as the angle between the horizontal plane and the position of the source.

2) Interaural time difference removal: Due to the anatomy of the head, sounds from different directions will inevitably arrive with different time delays. These delays are referred to as the ITDs. These delays can be removed and then reconstructed after the upsampling has occurred using a simple ITD model. [58].

In this paper, a Kalman filter [60] is used to detect the onset of each IR in the HRTFs so that the ITD can be removed. This onset detection method is similar to that of [61], [62] and works on the assumption that the amplitude behaviour of the noise floor is predictable and, therefore, if the error in the prediction is large, then the amplitude could not be predicted, and the onset of IR has occurred.

The amplitude of the IR, x_n, for the time index, n, is modelled here as a random walk with zero-mean, normally distributed increments such that

$$x_n = x_{n-1} + w, \quad w \in \mathcal{N}(0, \sigma_w^2) , \tag{5}$$

where the amplitude at n deviates from the amplitude at $n-1$ with a variance of σ_w^2, and x_0 is defined as zero. Observations of the IR, z_n, are modelled conditionally on x_n as

$$z_n = x_n + v, \quad v \in \mathcal{N}(0, \sigma_v^2) \,, \tag{6}$$

where the measurement noise, v, in this case, models the uncertainty in the observations.

The Kalman filter estimates the system's state and then acquires feedback from noisy measurements using a prediction and update step. The predicted amplitude estimate, $\hat{x}_{n|n-1}$, and predicted estimate variance, $P_{n|n-1}$, are given by

$$\hat{x}_{n|n-1} = \hat{x}_{n-1|n-1} \,, \tag{7}$$

$$P_{n|n-1} = P_{n-1|n-1} + \sigma_w^2 \,. \tag{8}$$

The updated amplitude estimate, $\hat{x}_{n|n}$, and updated estimate variance, $P_{n|n}$, are given by

$$\hat{x}_{n|n} = \hat{x}_{n|n-1} + K_n(z_n - \hat{x}_{n|n-1}) \,, \tag{9}$$

$$P_{n|n} = (1 - K_n)^2 P_{n|n-1} + K_n^2 \sigma_v^2 \,. \tag{10}$$

Where the innovation variance, S_n, and optimal Kalman gain, K_n, are given by

$$S_n = P_{n|n-1} + \sigma_v^2 \,, \tag{11}$$

$$K_n = \frac{P_{n|n-1}}{S_n} \,. \tag{12}$$

The error between measurement and prediction can, therefore, be calculated as

$$\tilde{y}_{n|n} = z_n - \hat{x}_{n|n} \,. \tag{13}$$

If this error, $\tilde{y}_{n|n}$, is above a threshold, η, then that implies that the error is large and the value, x_n, could not be predicted. This is indicative of the onset of the IR, which is not predictable by the Kalman filter. Therefore, once the onset has been located, the IR is trimmed before and after the onset so that all the IRs in the HRTF possess the same delay, thus removing the ITD.

3) Barycentric interpolation: The gnomonic equiangular projection (shown in Fig. 1(c)) has transformed the 3D space into a 2D plane; however, the issue of the measurements being spaced at irregular intervals still remains as the Cartesian points lie along curves. This is a problem as the convolution kernels used by the CNN require a uniform grid to function correctly. Therefore, barycentric interpolation [30] is used to project the data onto a regular Cartesian grid. For simplicity, the barycentric interpolation is performed on the sphere of IRs before the IRs are mapped using the gnomonic projection.

In previous work using barycentric interpolation [30], the three nearest measurement points to the interpolated point are calculated. However, suppose the measurement points are not evenly spaced, which is the case here. In that case, this can lead to the issue of the three selected measurement points not forming a spherical triangle around the point to be interpolated. Therefore, to solve this problem, we take the three closest measurement points, forming a spherical triangle around the point to be interpolated. This is similar to [31], which proposes a barycentric interpolation among the HRIRs at the points of a 3D tetrahedron conformed by four measurement points which surround the point to be interpolated.

First, in order to perform barycentric interpolation, it is necessary to find the three closest points that form a spherical triangle (P_1, P_2 and P_3) around the interpolated point (P_i), where a spherical triangle can be defined as a curved surface on a sphere which is bounded by the arcs of three great circles. Second, the barycentric coordinates α, β, and γ need to be calculated. These coordinates represent the ratio of the areas of the three smaller triangles ($P_iP_2P_3$, $P_1P_iP_3$, and $P_1P_2P_i$) relative to the larger triangle ($P_1P_2P_3$), such that $\alpha+\beta+\gamma = 1$. The coefficients α, β, and γ correspond to weights applied to the IRs at points P_1, P_2, and P_3, respectively. Ultimately, these coefficients will be used to find the interpolated HRIR for point P_i.

In [30], elevation (ϕ) and azimuth (θ) are treated as Cartesian coordinates, and the following formulas are used to find the ratio of the areas

$$\alpha = \frac{(\phi^{P_2} - \phi^{P_3})(\theta^{P_i} - \theta^{P_3}) + (\theta^{P_3} - \theta^{P_2})(\phi^{P_i} - \phi^{P_3})}{(\phi^{P_2} - \phi^{P_3})(\theta^{P_1} - \theta^{P_3}) + (\theta^{P_3} - \theta^{P_2})(\phi^{P_1} - \phi^{P_3})} \,, \tag{14}$$

$$\beta = \frac{(\phi^{P_3} - \phi^{P_1})(\theta^{P_i} - \theta^{P_3}) + (\theta^{P_1} - \theta^{P_3})(\phi^{P_i} - \phi^{P_3})}{(\phi^{P_2} - \phi^{P_3})(\theta^{P_1} - \theta^{P_3}) + (\theta^{P_3} - \theta^{P_2})(\phi^{P_1} - \phi^{P_3})} \,, \tag{15}$$

$$\gamma = 1 - \alpha - \beta \,. \tag{16}$$

Then, in order to treat elevation and azimuth as spherical coordinates rather than Cartesian coordinates, L'Huilier's Theorem [63] is used. L'Huilier's Theorem states that the surface area of a spherical triangle is given by $A = r^2E$, where r is the radius of the sphere, and E is the excess angle. The excess angle, E, is defined by

$$E = 4 \times \arctan\left(\sqrt{\begin{aligned}&\tan\left(\tfrac{1}{2}s\right)\tan\left(\tfrac{1}{2}(s-a)\right)\\ &\quad\times\tan\left(\tfrac{1}{2}(s-b)\right)\tan\left(\tfrac{1}{2}(s-c)\right)\end{aligned}}\right), \tag{17}$$

where a, b, c represent the side arc lengths of the triangle calculated using the haversine distance between two points on a sphere and $s = (a+b+c)/2$. The equations (14) to (16) can then be modified using (17) to use spherical coordinates to obtain the new weights

$$\alpha = \frac{E^{P_iP_2P_3}}{E^{P_1P_2P_3}} \,, \quad \beta = \frac{E^{P_1P_iP_3}}{E^{P_1P_2P_3}} \,, \quad \gamma = 1 - \alpha - \beta \,. \tag{18}$$

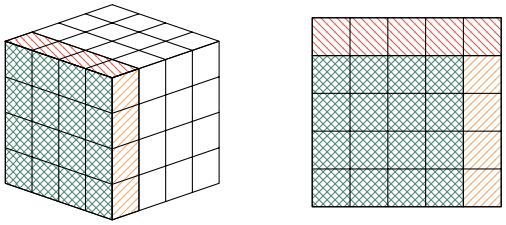
4) Magnitude spectrum extraction: After removing the ITD, the HRIRs are interpolated for each point of interest, P_i, using the barycentric coordinates α, β, and γ (see Section II-A3) along with

$$\text{HRIR}^{P_i} = \alpha\text{HRIR}^{P_1} + \beta\text{HRIR}^{P_2} + \gamma\text{HRIR}^{P_3} \,. \tag{19}$$

Following interpolation, the HRIR is transformed into the HRTF via the discrete Fourier transform (DFT). The magnitude of the HRTF is then used as an input to the GANs.

B. GAN architecture

In this work, a GAN architecture [49] is used to generate high-resolution HRTFs, H_{HR}, from their low-resolution counterparts, H_{LR}. This is achieved through a supervised learning approach where the network has access to the high-resolution H_{HR} target during training.



(a) (b)

Fig. 2. Each face of the gnomonic equiangular projection (in green) is padded with data from the adjacent faces (in red). This is displayed both for the 3D cube (a) and for the flattened 2D surface (b). In the corner, the value is ambiguous, therefore, values are taken from the top panel [64].

To generate the low-resolution H_{LR}, the high-resolution H_{HR} is downsampled by a factor of r. The dimensions of H_{HR} are $B \times X \times W \times H$, where B, X, W and H represent respectively the number of frequency bins, the number of cube sphere panels, the height of each cube sphere panel, and the width of each cube sphere panel. Therefore, after downsampling, considering that we are using $X = 5$ panels for this implementation, the dimensions of H_{LR} are $B \times 5 \times \frac{W}{r} \times \frac{H}{r}$. It is important to note that the downsampling is only spatial; therefore, only W and H are downsampled by r. In contrast, frequency resolution and the number of cube sphere panels remain the same.

The GAN architecture that is exploited in this paper is similar to that of [47] and is commonly referred to as SRGAN. The SRGAN architecture relies on two networks competing in a minimax game, with the generator consisting of residual layers followed by upsampling layers with a low-resolution image as the input and the discriminator taking a high-resolution input and then performing a series of convolutions. This network was chosen as the foundation for this work as it has proven successful at a diverse range of super-resolution tasks. The novelty in this work is that the 2D convolutional layers are adapted to be able to handle the gnomonic equiangular projection input data where one set of weights is learned for the equatorial panels ($n = 1$ to 4) and a separate set of weights is learned for the top panel ($n = 5$), i.e. the convolution takes an input of size $B \times 1 \times \frac{W}{r} \times \frac{H}{r}$ and learns two sets of weights, one for the top panel and a shared weight set for the equatorial panels. This differs from previous approaches that learn a single set of convolutional weights for all panels of the gnomonic equiangular projection, e.g. [64]. In addition, a novel gnomonic equiangular projection padding layer is added before each convolutional layer in the discriminator and generator models; this layer pads each panel in the gnomonic equiangular projection based on its adjacent panels, as shown in Fig. 2. In the corner, the value is ambiguous, therefore, values are taken from the top panel (see Fig. 2(b)). The GnomonicProjConv layer is defined as the gnomonic equiangular projection padding layer followed by the adapted convolutional layer.

In cases where there is no adjacent face, i.e. the lower edge of the equatorial panels, that edge is just padded by repeating the values that are closest to that edge. Note that because the padding is added repeatedly throughout the generator and discriminator networks, the networks can learn from points that stretch around the corners of the gnomonic equiangular projection.

GANs consist of a discriminator network D, shown in Fig. 3(a), which is optimised alongside a generator network G, shown in Fig. 3(b), in an alternating manner to find a solution to the adversarial minimax problem

$$\min_G \max_D \mathbb{E}_{H_{\text{HR}} \sim p_{\text{train}}(H_{\text{HR}})}[\log D(H_{\text{HR}})] + \mathbb{E}_{H_{\text{LR}} \sim p_{\text{G}}(H_{\text{LR}})}[\log(1 - D(G(H_{\text{LR}})))] \, . \tag{20}$$

1) Generator network: In this work, network G aims to generate high-resolution HRTFs from low-resolution HRTF inputs. The network G consists of B identical residual blocks, each containing two convolutional layers. A batch normalisation layer follows each of these convolutional layers. These batch normalisation layers are followed by a PReLU activation layer [65] after the first batch normalisation and an element-wise sum after the second batch normalisation. These element-wise sum units function as an additive residual (skip) connection.

To increase the HRTF 's resolution, R upsampling blocks are added after. Each block has an upsampling factor of 2; therefore, the number of needed blocks, R, is related to the downsampling factor, r, using $r = 2^R$. The spatial upsampling is performed via a standard pixel shuffle operation, which compresses channels and expands spatial extent by rearranging pixels. This is mathematically equivalent to, but more computationally efficient than, a transposed convolution.

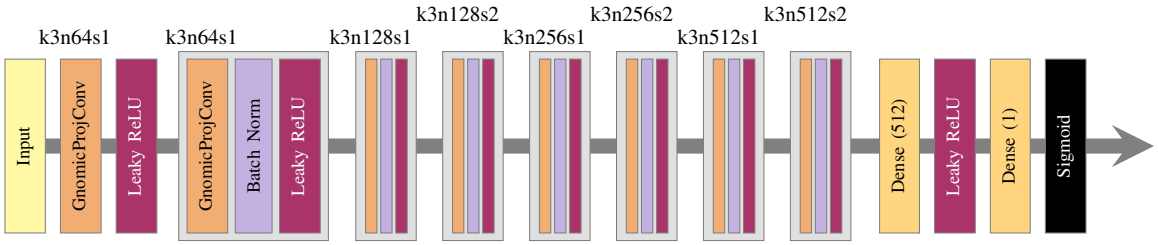
Another convolutional layer then follows these upsampling blocks before a final activation layer. The primary requirement of the activation layer is that the output is constrained to be positive, as the magnitude responses contained in an HRTF cannot be negative. There are multiple candidates for this, such as ReLU and Sigmoid. However, in this work, a softplus activation was selected as it is smoother near the origin than a ReLU and has shown better stabilisation and performance properties [66], [67].

2) Discriminator network: In this work, network D aims to discriminate whether an HRTF is real or generated by the network G. The network D consists of eight convolutional layers that are immediately followed by batch normalisation with the exception of the first layer.

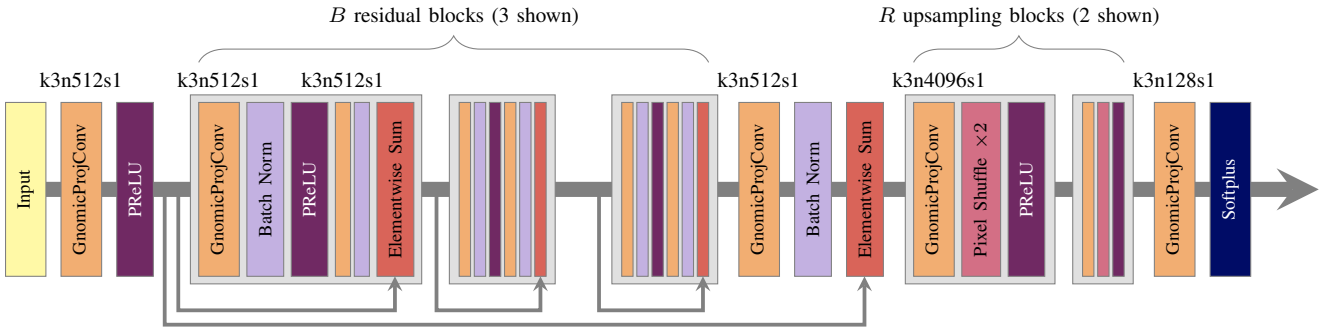
Two dense layers finally follow these convolutional layers, and then a Sigmoid activation function provides the discriminator network's output. Apart from the last layer, a leaky rectified linear unit (ReLU) is used as the activation function throughout, just as in [68]. The Leaky ReLU activation function is similar to that of the PReLU activation function in that both of them are defined as

$$f(x) = \max(ax, x) \, , \tag{21}$$

but in the case of the leaky ReLU activation function, a is a hyper-parameter that is set prior to training, while for PReLU a is a parameter that is learned during training [65].



(a) Discriminator network D.



(b) Generator network G.

Fig. 3. The architecture of the discriminator and generator networks, where each convolutional layer contains k kernels, n feature layers, and s stride. Acronyms: Leaky rectified linear unit (Leaky ReLU), parametric rectified linear unit (PReLU).

C. Cost functions

The total loss function l^{US} used in the generator network is key to its performance. This function is a weighted sum of the content loss $l_{\mathrm{C}}^{\mathrm{US}}$, which compares the upsampled generator output to the high-resolution ground truth, with an adversarial loss $l_{\mathrm{A}}^{\mathrm{US}}$, which measures how frequently the generator successfully fools the discriminator network. Therefore, l^{US} is defined as

$$l^{\mathrm{US}} = \lambda_{\mathrm{C}} \times l_{\mathrm{C}}^{\mathrm{US}} + \lambda_{\mathrm{A}} \times l_{\mathrm{A}}^{\mathrm{US}} \,, \tag{22}$$

where multipliers λ_{C} and λ_{A} represent the weight assigned to $l_{\mathrm{C}}^{\mathrm{US}}$ and $l_{\mathrm{A}}^{\mathrm{US}}$.

1) Content loss: The content loss, $l_{\mathrm{C}}^{\mathrm{US}}$, is the combination of the LSD metric and the interaural level difference (ILD) metric defined as

$$l_{\mathrm{C}}^{\mathrm{US}} = \mathrm{LSD} + \mathrm{ILD} \,. \tag{23}$$

The LSD metric [69] is used in order to score the difference between the target spectrum H_{HR} and the generated spectrum H_{US}.

$$\mathrm{LSD} = \tfrac{1}{N}\sum_{n=1}^{N} \sqrt{\tfrac{1}{B}\sum_{b=1}^{B}\left(20\log_{10}\frac{|H_{\mathrm{HR}}(f_b,x_n)|}{|H_{\mathrm{US}}(f_b,x_n)|}\right)^2} \,, \tag{24}$$

where $|H_{\mathrm{HR}}(f_b,x_n)|$ and $|H_{\mathrm{US}}(f_b,x_n)|$ represent the magnitude responses of the high-resolution and up-sampled HRTF sets, B is the number of frequency bins in the HRTF, N is the number of locations, f_b is the frequency, and x_n is the location.

The ILD metric [35], [70] is defined as

$$\mathrm{ILD} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{B}\sum_{b=1}^{B}\left|\left(20\log_{10}\frac{|H_{\mathrm{HR}}^{\mathrm{Left}}(f_b,x_n)|}{|H_{\mathrm{HR}}^{\mathrm{Right}}(f_b,x_n)|}\right) - \left(20\log_{10}\frac{|H_{\mathrm{US}}^{\mathrm{Left}}(f_b,x_n)|}{|H_{\mathrm{US}}^{\mathrm{Right}}(f_b,x_n)|}\right)\right| \,, \tag{25}$$

where $|H^{\mathrm{Left}}(f_b,x_n)|$ and $|H^{\mathrm{Right}}(f_b,x_n)|$ represent the magnitude responses of the left and right ear, respectively.

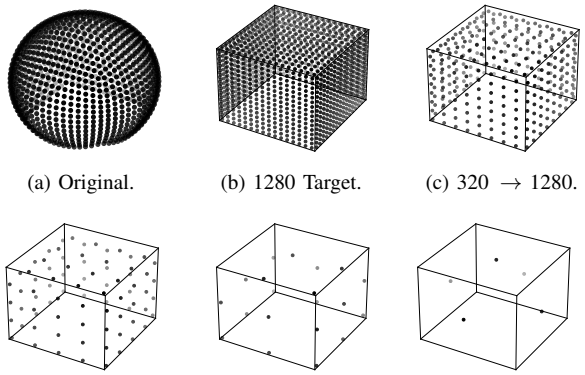
The LSD and ILD metrics then are both z-score normalised where the mean and standard deviation for both the ILD and LSD were calculated by comparing each HRTF in the training set to every other HRTF. The LSD and ILD are then summed to form the content loss function, $l_{\mathrm{C}}^{\mathrm{US}}$. This normalisation is to avoid either the LSD or ILD dominating the gradients during backpropagation if its loss is significantly greater.

2) Adversarial loss: The original GAN loss is used as the adversarial loss component of the total loss outlined in [49], which relates the generator network's training to the discriminator's output. The adversarial loss is defined over all training samples, M, as the binary cross-entropy loss

$$l_{\mathrm{A}}^{\mathrm{US}} = -\frac{1}{M}\left[\sum_{m=1}^{M}\Big(y_m \log\left(D(G(H_{\mathrm{LR}}^{m}))\right) + (1-y_m)\log\left(1-D(G(H_{\mathrm{LR}}^{m}))\right)\Big)\right] \,, \tag{26}$$

D. Data post-processing

To carry out some of the evaluations described in the following sections, the full HRTFs needed to be reconstructed, including the additional phase information that was disregarded in the pre-processing step. This phase information was removed on the assumption that



(a) Original. (b) 1280 Target. (c) 320 $\rightarrow$ 1280.

(d) 80 $\rightarrow$ 1280. (e) 20 $\rightarrow$ 1280. (f) 5 $\rightarrow$ 1280.

Fig. 4. The source positions for each downsampling factor.

the up-sampled HRTFs can be reconstructed using a minimum-phase approximation and a simple ITD model.

This minimum-phase approximation is achieved by calculating the minimum-phase function, $m(\omega)$, which can be uniquely determined by the magnitude spectrum of each transfer function, $H(\omega)$, at every source position in the HRTF through the Hilbert transform [71]

$$m(\omega) = \mathscr{H}\{-\ln(|H(\omega)|)\} \,, \tag{27}$$

where $\mathscr{H}\{\cdot\}$ denotes the Hilbert transform, and ω represents the frequency bin.

The simple model used to calculate the ITDs is based on the radius, r, of the listener's head (set to 8.75 cm, which corresponds to an average adult head size), the speed of sound, c, (approximately 343 m/s) and the interaural azimuth, θ_I, (in radians, from 0 to $\frac{\pi}{2}$ for sources on the listener's left, and from $\frac{\pi}{2}$ to π for sources on the listener's right) using

$$\begin{aligned} \text{ITD} &= \frac{r}{c}\big(\theta_I + \sin(\theta_I)\big) \,, \\ \text{where} \quad \theta_I &= \arcsin\big(\sin(\theta)\cos(\phi)\big) \,. \end{aligned} \tag{28}$$

It should be noted that a final interpolation could also be performed to map the points generated from the gnomonic equiangular projection back to an even spherical distribution. The main reason why this final interpolation is not performed is because it is unnecessary. As sofa files are all measured on different grids, most, if not all, software that utilises sofa files re-interpolates them onto a uniform grid [30] before deployment. In the proposed method, the grid of the upsampled HRTFs is very dense. Therefore, any re-interpolation that may take place in spatial acoustic software would only introduce errors that would not be perceived perceptually.

III. Experimental Setup

A. Data

The network was trained and validated on the HRTF dataset measured in Austria by the Acoustic Research Institute (ARI) and, throughout this paper, will be referred to as the ARI HRTF database [72]. The dataset contains HRTF measurements on 221 subjects for both the left and right ear (442 HRTF in total for both ears), making it one of the largest measured HRTF datasets available. Each measured HRTF contains IRs for 1550 directions around the listener; these range from 0° to 360° in terms of azimuth and from -30° to 80° in terms of elevation (a measurement for the top position was not available in this specific dataset). The number of measurements near the horizontal plane was of a higher density, which is common in HRTF measurement systems to reflect that humans can localise sounds in this space more accurately. The ARI HRTF dataset is read using the `Hartufo` toolkit [73], [74], which was developed for HRTF data management with a specific focus on deep learning.

After processing all HRTFs in the ARI HRTF dataset as described in Section II-A, the 442 HRTFs are split so that 352 HRTFs are used for training and 90 HRTFs are used for validation. This represents an 80-20 split between training and validation sets where the left and right ear for the same individual are not split between sets, i.e. a subject may not contribute one ear to training and another ear to the validation. This ensures that the generator network is tested on unseen data from a given individual.

B. Training

The high-resolution 1280 target is generated by pre-processing the ARI HRTF dataset (which contains 1550 positions for each listener) into 5 panels, which contain 16 by 16 source positions (i.e. 1280 positions $= 5 \times 16 \times 16$, shown in Fig. 1(d)) where the Kalman filter parameters where set to $\eta = 5{\times}10^{-3}$, $\sigma_w^2 = 1{\times}10^{-4}$, $\sigma_v^2 = 400$ and the IRs where trimmed to 10 samples before the onset and to a length of 128 samples.

To obtain the low-resolution HRTF inputs from the high-resolution targets, the HRTFs are downsampled by selecting one IR in every r. This means each high-resolution HRTF target, generated from the gnomonic equiangular projection, whose dimensions are $256 \times 5 \times 16 \times 16$, are downsampled to $256 \times 5 \times \frac{16}{r} \times \frac{16}{r}$ to create each input. In the case where only 5 source positions are available, the centre position of each panel is selected with coordinates (8,8). Therefore, the generator network aims to preserve the points of the low-resolution HRTF given at the input while interpolating all the other points to match the target.

It should be noted that the 256 dimension refers to the concatenation of the two left and right ear 128-point IRs. Fig. 4 shows the positions of the sources for r values 2 (320 $\rightarrow$ 1280), 4 (80 $\rightarrow$ 1280), 8 (20 $\rightarrow$ 1280), 16 (5 $\rightarrow$ 1280). These positions were selected using the `torch.nn.functional.interpolate` function, where the '`scale_factor`' was set to r. The high-resolution HRTF target of 1280 positions (i.e. $5 \times 16 \times 16$) was selected as it is comparable to the 1550 positions measured in the ARI HRTF dataset.

It should also be noted that the 128-point fast Fourier transform (FFT) magnitude inputs are also not scaled or normalised, as the LSD metric used in the content loss for the generator, G, requires non-normalised magnitudes.

The GANs hyperparameters were adjusted in order to find the best-performing model in terms of the cost function on

TABLE I
THE HYPERPARAMETER VALUES SELECTED FOR THE FOUR NETWORKS WITH DIFFERENT UPSAMPLING FACTORS.

Hyperparameter	**Upsample Factor [No. orginal → No. upsampled]**			
	320 → 1280	**80 → 1280**	**20 → 1280**	**5 → 1280**
No. Epochs	300	300	300	300
LR - Generator	2.0×10^{-4}	8.0×10^{-4}	2.0×10^{-4}	2.0×10^{-4}
LR - Discriminator	1.5×10^{-6}	1.5×10^{-6}	1.5×10^{-6}	1.5×10^{-6}
Content Weight (λ_C)	0.1	0.01	0.001	0.01
Adversarial Weight (λ_A)	0.001	0.1	0.001	0.01

the training data. To achieve this hyperparameter tuning, a grid search was deployed using [75], where the search space consisted of 'Learning rate (LR) - Generator': $\{2.0\times10^{-4}, 4.0\times10^{-4}, 6.0\times10^{-4}, 8.0\times10^{-4}\}$, 'LR - Discriminator': $\{1.5\times10^{-6}, 3.0\times10^{-6}, 4.5\times10^{-6}, 6.0\times10^{-6}\}$, 'Number of epochs': {300, 250, 200, 150}, 'Adversarial weight (λ_A)': {0.1, 0.01, 0.001}, 'Content weight (λ_C)': {0.1, 0.01, 0.001}. This resulted in a variation of 15.5% in terms of the cost function with the selected values for each of the four networks given in Table I. Various batch sizes were also explored informally; a relatively small batch size of 8 was found to give the best performance.

The model was trained using the *Adam* optimiser [76], with hyperparameter values of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. In training, D and G were alternately updated with different frequencies to improve stability, with D being updated four times for every G update. This was found by informally testing different ratios inspired by [77]. The kernel weights were initialised using Kaiming initialisation [65]. This model was implemented using a PyTorch framework with custom modules for gnomonic equiangular projection padding and convolution layers and trained on an NVIDIA Quadro RTX 6000 graphics processing unit (GPU).

C. Convergence

In Fig. 5, the convergence of both the generator and the discriminator networks is given for the 20 → 1280 network. It should be noted that owing to the fact that a GAN consists of two networks competing against each other, an improvement in the generator will lead to a higher loss in the discriminator and vice versa. Hence, both networks converge to a stable value over time (although this is not expected to be zero). It can also be seen in Fig. 5 that the generator network can converge quicker than the discriminator. Although not shown here, similar loss curves can be observed for the three other networks (5 → 1280, 80 → 1280 and 320 → 1280).

D. Baselines

Three baseline methods have been used as a benchmark for the comparison of the results from the experimental evaluation.

1) Baseline-1 - SH interpolation: An approach that has shown to yield good performance in the domain of HRTF upsampling is SH interpolation. It works by projecting the HRTF onto a set of spherical basis functions, known as spherical harmonics, which produces a continuous representation of an HRTF [78]. This work

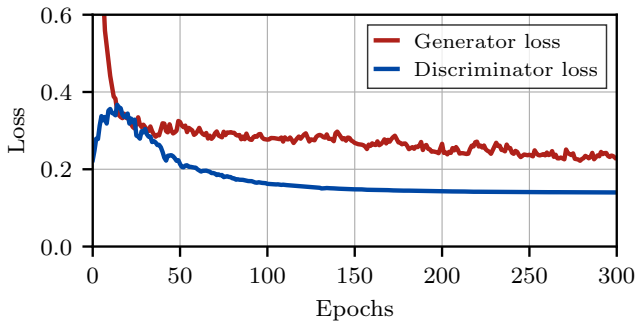


Fig. 5. Illustrative example of overall loss curves for 20 → 1280 network.

uses a magnitude-corrected and time-aligned SH interpolation presented in [36] as a baseline.

2) Baseline-2 - Barycentric interpolation: The most common method for upsampling is that of barycentric interpolation, which is used in this section as a baseline and described in Section II-A3. It should be noted, however, that the approach in Section II-A3 is modified when the number of source positions is only 5 in the low-resolution HRTF. This is because if no triangle can be formed around the point to be interpolated, which is sometimes the case at this low resolution, then the three closest points need to be used instead.

3) Baseline-3 - non-individual HRTF selection: As mentioned in Section I, instead of interpolating a low-resolution HRTF, we can just select an HRTF from a database. In this baseline, instead of randomly selecting an HRTF, two HRTFs are selected from the training dataset based on their average LSD error when comparing them against all other HRTFs in the training set. The subject whose HRTF produces the lowest average LSD error is considered the most generic (Selection-1), and the subject whose HRTF produces the largest average LSD error is considered the most unique (Selection-2). It must be highlighted, however, that this selection is only based on the LSD, and not all LSD errors have the same perceptual relevance.

IV. EXPERIMENTAL EVALUATION

This section will compare four different levels of SRGAN upsampling against the two baselines described earlier for 45 test subjects. These upsampling levels include 320, 80, 20 and 5 source positions to 1280 source positions.

The complete SRGAN implementation and pre-processing code to reproduce these results are available at [79].

A. LSD metric evaluation

The LSD metric, defined in (24), can be calculated for every measurement source position and then averaged over all the source positions. Table II and Fig. 6 show the average results for this LSD evaluation over the 45 subjects in the test set. In Fig. 6, it is clear to see the benefit of using the proposed SRGAN over the barycentric interpolation (Baseline-2) when the input HRTF is spatially sparse. The SRGAN outperforms the barycentric method when the input contains 20 or fewer

TABLE II
A COMPARISON OF THE MEAN LOG-SPECTRAL DISTORTION (LSD) AND (STANDARD DEVIATION (SD)) ERROR ACROSS ALL SOURCE POSITIONS FOR DIFFERENT UPSAMPLING FACTORS. THE 'BEST' PERFORMANCE OF EACH UPSAMPLING FACTOR HAS BEEN HIGHLIGHTED.

Method	Upsampling [No. orginal → No. upsampled]			
	320 → 1280	80 → 1280	20 → 1280	5 → 1280
SRGAN	3.28 (0.13)	4.86 (0.24)	4.99 (0.27)	5.30 (0.35)
SH	3.54 (0.15)	4.94 (0.20)	5.90 (0.25)	10.36 (0.74)
Barycentric	2.50 (0.20)	3.71 (0.22)	5.18 (0.23)	7.30 (0.33)
Selection-1	6.96 (0.47)			
Selection-2	8.20 (0.61)			

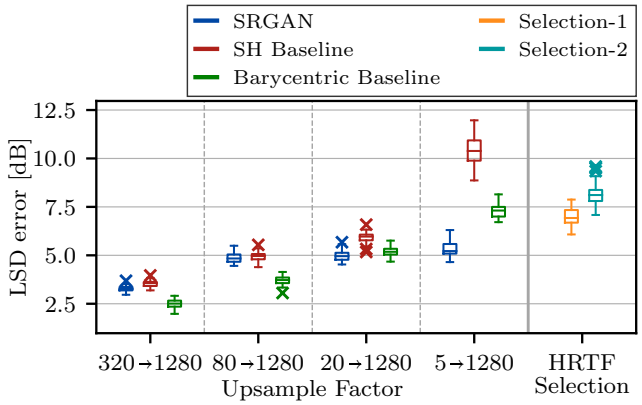


Fig. 6. Log-spectral distortion (LSD) error comparison.

different IR source positions. The most striking result is upsampling 5 positions to 1280 positions (5 → 1280 network), where the barycentric method only achieves an LSD error of 7.30. This makes sense as barycentric interpolation can only average the three closest IRs, enclosing the interpolated point. The further away these closest points are that form the enclosing triangle, the more inaccurate the barycentric approach becomes.

In terms of the SH interpolation baseline (Baseline-1), it performs slightly worse than barycentric interpolation across all upsampling factors and never outperforms the SRGAN. This result is likely due to the distribution of the source positions not being uniformly spaced around the sphere after the gnomonic equiangular projection. It is expected that the performance of SH and barycentric interpolation will be comparable as both methods effectively rely on a weighted sum of existing points to generate new ones having no prior knowledge of the HRTF data.

The results of HRTF selection (Baseline-3) are also shown to give poor performance in terms of LSD error when compared against the proposed method and barycentric interpolation (Baseline-2). What is interesting is that the HRTF Selection-1 method slightly outperforms barycentric interpolation when only 5 source positions are used as the low-resolution input (with a mean LSD error of 6.96 for Selection-1 compared with 7.30 for barycentric), however, as is expected, Selection-2 does perform worse (where the mean LSD error is 8.20 for Selection-2 compared with 7.30 for barycentric).

TABLE III
THE MEAN AND (STANDARD DEVIATION (SD)) VALUES OF THE MODEL-BASED PERCEPTUAL EVALUATION ACROSS THE SUBJECTS IN THE TEST SET FOR THE DIFFERENT UPSAMPLING FACTORS. THE 'BEST' PERFORMANCE OF EACH UPSAMPLING FACTOR HAS BEEN HIGHLIGHTED.

Method	Upsampling [No. orginal → No. upsampled]			
	320 → 1280	80 → 1280	20 → 1280	5 → 1280
SRGAN	0.46 (4.66)	2.73 (6.82)	1.06 (10.02)	-0.70 (8.33)
SH	-0.39 (4.38)	-1.92 (4.54)	-4.17 (5.02)	-28.61 (17.55)
Barycentric	1.17 (3.84)	1.57 (4.32)	2.22 (8.36)	-2.54 (23.84)
Selection-1	2.05 (17.17)			
Selection-2	22.18 (21.75)			
Target	0.86 (3.74)			

(a) Polar accuracy error comparison.

Method	Upsampling [No. orginal → No. upsampled]			
	320 → 1280	80 → 1280	20 → 1280	5 → 1280
SRGAN	8.64 (2.78)	9.83 (3.25)	16.53 (4.59)	11.28 (3.60)
SH	8.43 (2.89)	10.10 (2.99)	16.04 (3.66)	19.17 (8.25)
Barycentric	8.50 (2.68)	9.15 (2.73)	13.79 (3.76)	24.65 (7.28)
Selection-1	22.36 (7.59)			
Selection-2	22.17 (12.14)			
Target	7.99 (2.76)			

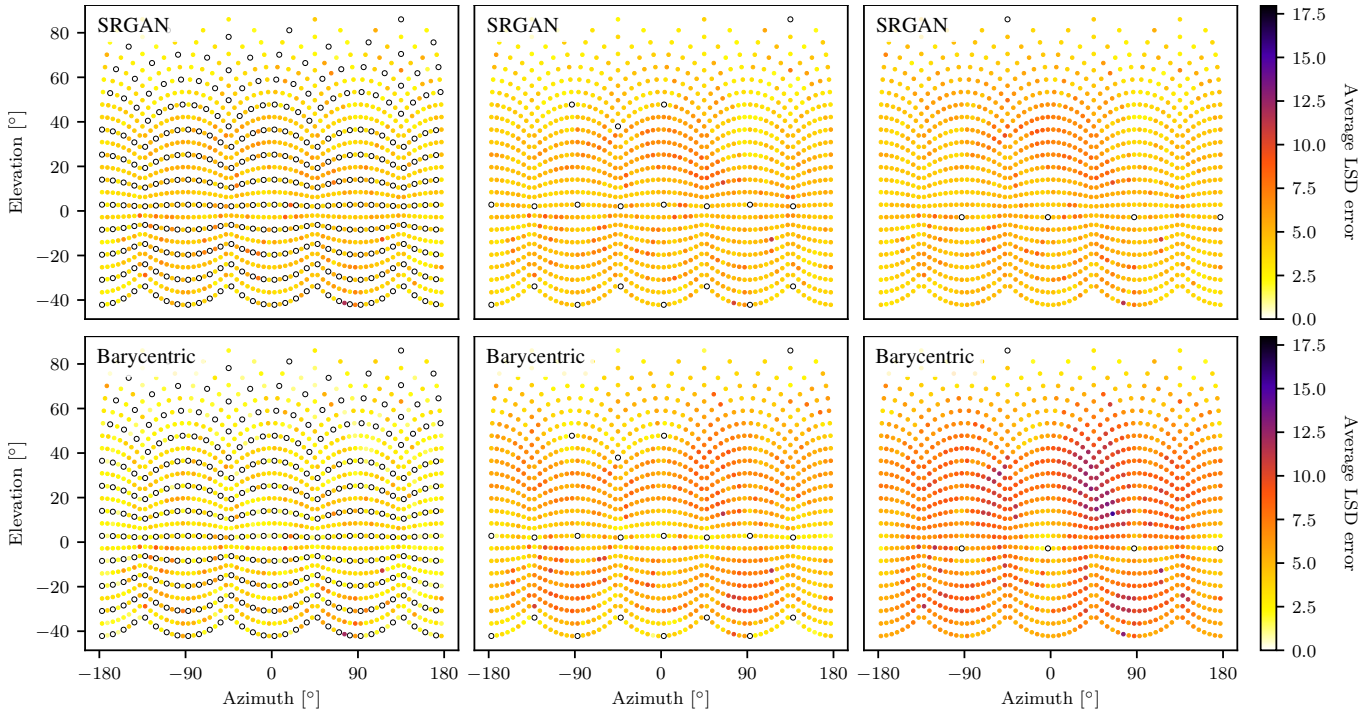
(b) Quadrant error comparison.

Method	Upsampling [No. orginal → No. upsampled]			
	320 → 1280	80 → 1280	20 → 1280	5 → 1280
SRGAN	32.96 (1.83)	35.46 (1.73)	36.50 (1.65)	36.03 (1.87)
SH	33.02 (2.04)	33.83 (2.21)	34.71 (2.24)	41.98 (2.52)
Barycentric	32.61 (1.70)	33.75 (1.68)	38.24 (1.36)	41.79 (1.22)
Selection-1	38.89 (2.43)			
Selection-2	39.90 (2.32)			
Target	32.11 (2.10)			

(c) Polar root mean square (RMS) error comparison.

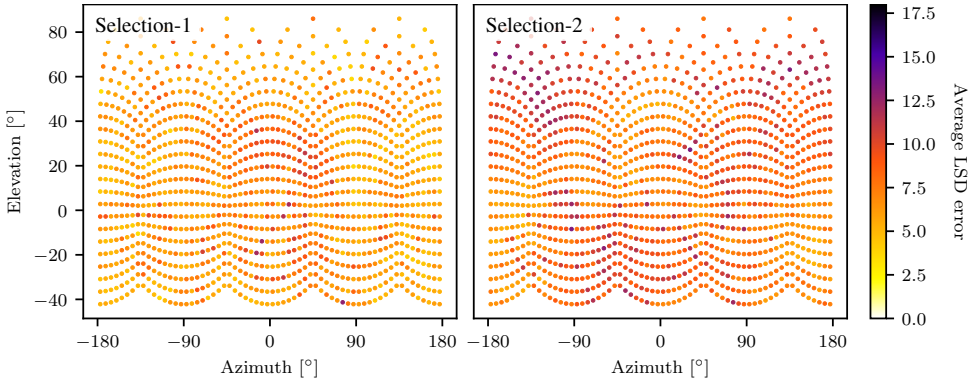
To better understand where these errors occur for the proposed SRGAN and barycentric interpolation, an illustrative example of a random individual in the test set (SubjectID 868) is given in Fig. 7, where the LSD errors for all the interpolated source positions are shown for the different levels of upsampling. The LSD errors for Baseline-3 are also given in Fig. 8 for the same subject.

It can be seen in Fig. 7(a) for upsampling factor 320 → 1280 that the source positions with larger errors over 7 dB are the same for both methods. However, as the input becomes more spatially sparse, the errors between both methods start to diverge, which can be seen in Fig. 7(b) and Fig. 7(c). This is due to the fact that barycentric interpolation is unable to interpolate source positions that are not close to measured points. On the other hand, the proposed SRGAN has learnt general patterns across the training data and the relationships between these and the low-resolution input. It is, therefore, able to perform equally across all source positions regardless of where they are located and their proximity to measurement points. This is most clearly seen when comparing



(a) 320 → 1280. (b) 20 → 1280. (c) 5 → 1280.

Fig. 7. Comparison of the proposed super-resolution generative adversarial network (SRGAN) (top) and barycentric (bottom) in terms of log-spectral distortion (LSD) errors at different levels of upsampling SubjectID 868 (all source positions). The original source positions before interpolation are outlined with a black circle.



(a) The most 'generic' HRTF. (b) The most 'unique' HRTF.

Fig. 8. The log-spectral distortion (LSD) of Basline-2 showing the comparison between Selection-1 and Selection-2 for SubjectID 868 (all source positions).

the SRGAN approach in Fig. 7(b) and Fig. 7(c). The errors are almost identical across all source positions (see also Table II), although the number of original points has been reduced by a factor of four. In contrast, when comparing the barycentric interpolation for the same upsampling levels, the errors increase substantially from Fig. 7(b) to Fig. 7(c) at the points where the input measurements have been discarded (e.g. above and below the equator).

B. Model-based perceptual evaluation

In this section, we use a Bayesian model, Barumerli2022, introduced in [80], to compare the localisation performance. Unlike Section IV-A, where the LSD metric is used to compare the performance of the different methods, the evaluation in this section is able to differentiate the different techniques based on errors that matter to human perception. This is important as some minor errors in the LSD could significantly impact human localisation performance. Likewise, some significant errors in the LSD may not affect localisation performance nearly as much. The Barumerli2022 model was chosen as it has already been successfully employed in the past for evaluating different binaural rendering methods [35].

The Barumerli2022 model was fed features that were obtained from the directional transfer functions (DTFs), which

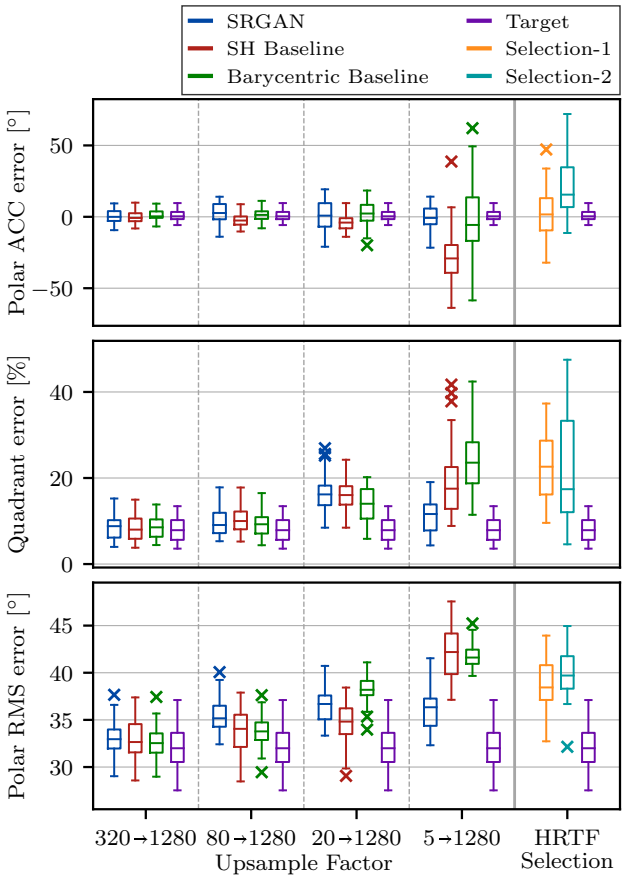


Fig. 9. Results from the model-based perceptual evaluation.

is the directional component directly extracted from the HRTF [81]. The model's predefined parameters were set to: 'estimator': maximum a posteriori (MAP), 'num_exp': 300, 'sigma_itd': 0.569, 'sigma_ild': 1, 'sigma_spectral': 4, 'sigma_prior': 11.5, 'sigma_motor': 10. The 'targ_az' and 'targ_el' are set to the source positions measured in the HRTF.

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 they effectively compare the localisation performance of the original high-resolution HRTF with itself. Therefore, the proposed method and the two other baselines must be benchmarked against the 'Target' performance.

The results obtained from the Barumerli2022 model are shown in Table III with a graphical representation given in Fig. 9. Table IVa, Table IVb and Table IVc show the mean and SD of the polar accuracy error, quadrant error and RMS error, respectively. These metrics [82], [83] can be defined mathematically for N localisation trials, where each target source direction ϕ_i has an associated response direction $\tilde{\phi}_i$, for $i = 1, 2, \ldots, N$. If a set of local responses is then defined as $\mathcal{A} = \{i : \text{wrap}|\tilde{\phi}_i - \phi_i| < 90°\}$ the three error metrics are defined as follows:

$$\text{Polar Accuracy} = \frac{\tilde{\phi}_i - \phi_i}{|\mathcal{A}|}\,, \tag{29}$$

$$\text{Polar RMS Error} = \sqrt{\frac{\sum_{i\in\mathcal{A}}(\text{wrap}(\tilde{\phi}_i - \phi_i))^2}{|\mathcal{A}|}}\,, \tag{30}$$

$$\text{Quadrant Error} = \left(1 - \frac{|\mathcal{A}|}{N}\right) \times 100\,, \tag{31}$$

where the polar accuracy shows the local response bias in the polar angles for responses within 90° of the target. The polar RMS error is the aggregated error in the polar dimension for responses within 90° of the target. The quadrant error rate corresponds to the percentage of polar errors larger than 90° and accounts for top-down and front-back confusions. To avoid highly distorted polar errors on the far left and right sides of the listener, polar and quadrant errors are only defined for targets within a lateral angle $|\theta| \leq 30°$.

It can be seen that when the input provides 320 or 80 source positions, the SRGAN performance is comparable to SH and barycentric interpretation, only performing marginally worse. However, when the low-resolution HRTF contains 20 or fewer source positions, the proposed SRGAN performs significantly better than all baselines.

It is also interesting that the SRGAN at upsample factor $5 \rightarrow 1280$ outperforms the SRGAN at $20 \rightarrow 1280$. This is because even though the LSD is worse when only 5 points are used, that does not mean that those spectral differences are perceptually relevant. This shows the need to incorporate these perceptual models into the loss function going forward by overcoming the need for them to be differentiable. It should also be noted that when only 5 source positions are available, the centre position of each panel is selected, and therefore, the source positions are not an exact subset of the 20. This highlights, understandably, the likelihood that certain positions could be more meaningful for localisation when compared with other positions.

C. Limitations and future work

While the results presented in the previous sections are relatively positive and confirm the potential suitability of the proposed approach for spatially upsampling very sparse HRTF measurements, there are some evident limitations which should be addressed in future research.

Firstly, this study didn't assess the impact of the pre-processing transformations on numerical and model-based perceptual metrics. The limitation of the current method is that to achieve optimal performance and reduce the Barycentric interpolation errors while upsampling an existing dataset; the approach currently relies on the model being retrained with the input positions that are closest to any given point on the high-resolution cartesian grid. In future work, it would be good to explore the possibility of passing the position of each point as an additional input feature to the model so that this retaining is not needed.

In order to simplify the SRGAN architecture, for this first study, we have disregarded the phase information and used only the magnitude component of each HRTF. Upsampled HRTFs were then reconstructed using minimum-phase approximation and an ITD model. Future research should explore the possibility of including phase information when

training the network, and further evaluations should outline how this may result in perceptually relevant improvements.

Another simplification was performed when looking at the number of panels on which the projection was performed during the HRTF pre-processing stages. The choice of 5 panels was mainly dictated by the dataset used for the training. Still, it is, of course, possible to extend this to 6 panels, therefore, to include all HRTFs measured below the subject both as training and output data. Future research could also look at removing this limitation, allowing an arbitrary number (and position) of HRTFs to be used for the upsampling (currently, this is limited to a minimum of 5 positions, one per panel) and exploring the impact of having non-uniform measurements as a starting point for the upsampling process.

The design of the loss function could also benefit from further investigations. For example, looking at the weight of each of the two components, ILD and LSD, increasing it for the first in lateral positions and for the second in positions on the sagittal planes. Furthermore, perceptual models could also be employed, but relevant challenges should be addressed first, especially related to the computational complexity of such models.

It is important to underline that this paper describes the second milestone (after [50]) in the process of designing, validating, improving and ultimately releasing this method as an openly available tool. Research following these future directions is already being conducted.

V. Conclusion

In this paper, it has been shown that an SRGAN can be used effectively for the task of upsampling low-resolution HRTFs. Furthermore, this work has extended the pilot study from [50], modifying the SRGAN so that it can upsample the HRTFs in 3D across the entire sphere. It has been demonstrated that when the low-resolution HRTF input is very sparse and consists of less than 20 source positions, the SRGAN outperforms both SH and barycentric interpolation in terms of LSD error. The same applies to localisation performance using perceptual models when the low-resolution HRTF contains 5 source positions. Therefore, in the case where the low-resolution HRTF contains 320 or more source positions, it is preferable to use barycentric interpolation; however, if the HRTFs are very sparse and contain less than 20 measurements, then the SRGAN approach produces significantly lower errors. SH interpolation, on the other hand, performs slightly worse than both barycentric interpolation and the SRGAN, which is likely due to the distribution of the source positions not being uniformly spaced around the sphere. Non-individual HRTF selection also never performs best; however, it does outperform SH and barycentric interpolation for very sparse HRTFs, according to the employed metrics.

In order to reinforce the idea of reproducible research and promote future development and innovation in this specific domain, the complete SRGAN architecture and pre-processing code can be found in our public repository [79].

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] D. Vickers, M. Salorio-Corbetto, S. Driver, C. Rocca, Y. Levtov, K. Sum, B. Parmar, G. Dritsakis, J. Albanell Flores, D. Jiang, *et al.*, "Involving children and teenagers with bilateral cochlear implants in the design of the BEARS (Both EARS) virtual reality training suite improves personalization," *Front. Digit. Health*, vol. 3, Nov. 2021.

[3] 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.

[4] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. Cambridge, Mass: MIT Press, 1983.

[5] P. Stitt, L. Picinali, and B. F. G. Katz, "Auditory accommodation to poorly matched non-individual spectral localization cues through active learning," *Scientific Reports*, vol. 9, no. 1, pp. 1063:1–14, Jan. 2019.

[6] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111–123, July 1993.

[7] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?" *J. Audio Eng. Soc. (AES)*, vol. 44, no. 6, pp. 451–469, June 1996.

[8] Y. Kahana and P. A. Nelson, "Numerical modelling of the spatial acoustic response of the human pinna," *J. of Sound and Vibration*, vol. 292, no. 1, pp. 148–178, Apr. 2006.

[9] L. S. R. Simon, N. Zacharov, and B. F. G. Katz, "Perceptual attributes for the comparison of head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 140, no. 5, pp. 3623–3632, Nov. 2016.

[10] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *Proc. Int. Conf. on Quality of Multimedia Experience (QoMEX)*, June 2016, pp. 1–6.

[11] I. Engel, D. L. Alon, P. W. Robinson, and R. Mehra, "The effect of generic headphone compensation on binaural renderings," in *Proc. Audio Eng. Soc. (AES) Int. Conf. on Immersive and Interactive Audio*, Mar. 2019.

[12] M. Cuevas-Rodriguez, D. Gonzalez-Toledo, A. Reyes-Lecuona, and L. Picinali, "Impact of non-individualised head related transfer functions on speech-in-noise performances within a synthesised virtual environment," *J. Acoust. Soc. Am.*, vol. 149, no. 4, pp. 2573–2586, Apr. 2021.

[13] 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.

[14] 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.

[15] B. F. G. Katz, "Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation," *J. Acoust. Soc. Am.*, vol. 110, no. 5, pp. 2440–2448, Nov. 2001.

[16] P. Stitt and B. F. 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, Apr. 2021.

[17] M. Geronazzo, E. Peruch, F. Prandoni, and F. Avanzini, "Applying a single-notch metric to image-guided head-related transfer function selection for improved vertical localization," *J. Audio Eng. Soc. (AES)*, vol. 67, no. 6, pp. 414–428, June 2019.

[18] DYN. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "HRTF personalization using anthropometric measurements," in *Proc. IEEE Workshop on Appl. of Signal Process. to Audio and Acoust. (WASPAA)*. IEEE, Oct. 2003, pp. 157–160.

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

[20] 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. (AES)*, vol. 68, no. 11, pp. 819–831, Dec. 2020.

[21] F. Zagala, M. Noisternig, and B. F. Katz, "Comparison of direct and indirect perceptual head-related transfer function selection methods," *J. Acoust. Soc. Am.*, vol. 147, no. 5, pp. 3376–3389, May 2020.

[22] 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, Oct. 2022, pp. 121–144.

[23] T. Carpentier, H. Bahu, M. Noisternig, and O. Warusfel, "Measurement of a head-related transfer function database with high spatial resolution," in *Proc. EAA Forum Acusticum, Eur. Congress on Acoust.*, Krakow, Poland, Sept. 2014.

[24] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. Audio Eng. Soc. (AES) Conv.*, Feb. 2000, pp. 1–23.

[25] D. N. Zotkin, R. Duraiswami, E. Grassi, and N. A. Gumerov, "Fast head-related transfer function measurement via reciprocity," *J. Acoust. Soc. Am.*, vol. 120, no. 4, pp. 2202–2215, Oct. 2006.

[26] J.-G. Richter, G. Behler, and J. Fels, "Evaluation of a fast HRTF measurement system," in *Proc. Audio Eng. Soc. (AES) Conv.*, vol. 140, May 2016, p. 9498.

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

[28] 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*, Mar. 1999.

[29] D. Poirier-Quinot and B. F. G. Katz, "The anaglyph binaural audio engine," in *Proc. Audio Eng. Soc. (AES) Conv.*, ser. 144, May 2018.

[30] 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, Mar. 2019.

[31] H. Gamper, "Head-related transfer function interpolation in azimuth, elevation, and distance," *J. Acoust. Soc. Am.*, vol. 134, no. 6, pp. EL547–EL553, Dec. 2013.

[32] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Analyzing head-related transfer function measurements using surface spherical harmonics," *J. Acoust. Soc. Am.*, vol. 104, no. 4, pp. 2400–2411, June 1998.

[33] C. Pörschmann, J. M. Arend, and F. Brinkmann, "Directional equalization of sparse head-related transfer function sets for spatial upsampling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 6, pp. 1060–1071, June 2019.

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

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

[36] J. M. Arend, C. Pörschmann, S. Weinzierl, and F. Brinkmann, "Magnitude-corrected and time-aligned interpolation of head-related transfer functions," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 3783–3799, Sept. 2023.

[37] 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)*, May 2019, pp. 271–275.

[38] D. Yao, J. Zhao, L. Cheng, J. Li, X. Li, X. Guo, and Y. Yan, "An individualization approach for head-related transfer function in arbitrary directions based on deep learning," *JASA Express lett. (JASA-EL)*, vol. 2, no. 6, p. 064401, June 2022.

[39] Y. Ito, T. Nakamura, S. Koyama, and H. Saruwatari, "Head-related transfer function interpolation from spatially sparse measurements using autoencoder with source position conditioning," in *Proc. Int. Workshop on Acoust. Signal Enhancement (IWAENC)*, Sept. 2022, pp. 1–5.

[40] R. Duraiswami, D. N. Zotkin, and N. A. Gumerov, "Interpolation and range extrapolation of HRTFs," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 4, May 2004, pp. iv–45.

[41] G. Kestler, S. Yadegari, and D. Nahamoo, "Head related impulse response interpolation and extrapolation using deep belief networks," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2019, pp. 266–270.

[42] Z. Jiang, J. Sang, C. Zheng, A. Li, and X. Li, "Modeling individual head-related transfer functions from sparse measurements using a convolutional neural network," *J. Acoust. Soc. Am.*, vol. 153, no. 1, pp. 248–259, Jan. 2023.

[43] B. Tsui, W. A. P. Smith, and G. Kearney, "Low-order spherical harmonic HRTF restoration using a neural network approach," *Appl. Sci.*, vol. 10, no. 17, p. 5764, Jan. 2020.

[44] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *Proc. Int. Joint Conf. on Learning Representations (ICLR)*, Feb. 2018.

[45] S. E. Eskimez, K. Koishida, and Z. Duan, "Adversarial training for speech super-resolution," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 347–358, May 2019.

[46] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[47] 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.

[48] K. Schawinski, C. Zhang, H. Zhang, L. Fowler, and G. K. Santhanam, "Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit," *Mon. Not. Royal Astron. Soc. Lett.*, vol. 467, no. 1, pp. L110–L114, May 2017.

[49] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Neural Inform. Process. Conf*, vol. 2. Curran Associates, Inc., Dec. 2014, pp. 2672–2680.

[50] 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, Aug. 2022.

[51] C. Ronchi, R. Iacono, and P. S. Paolucci, "The "cubed sphere": A new method for the solution of partial differential equations in spherical geometry," *J. Computational Physics*, vol. 124, no. 1, pp. 93–114, Mar. 1996.

[52] M. Rančić, R. J. Purser, and F. Mesinger, "A global shallow-water model using an expanded spherical cube: Gnomonic versus conformal coordinates," *J. of Royal Meteorological Soc.*, vol. 122, no. 532, pp. 959–982, Apr. 1996.

[53] R. J. Purser and M. Rančcić, "Smooth quasi-homogeneous gridding of the sphere," *J. of Royal Meteorological Soc.*, vol. 124, no. 546, pp. 637–647, Jan. 1998.

[54] R. Purser and M. Rancic, "A standardized procedure for the derivation of smooth and partially overset grids on the sphere, associated with polyhedra that admit regular griddings of their surfaces. Part I: Mathematical principles of classification and construction," National Weather Service (NWS), NOAA/NCEP Office Note 460, Dec. 2011. [Online]. Available: https://www.emc.ncep.noaa.gov/officenotes/FullTOC.html

[55] W. M. Putman and S.-J. Lin, "Finite-volume transport on various cubed-sphere grids," *J. Computational Physics*, vol. 227, no. 1, pp. 55–78, Nov. 2007.

[56] 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.

[57] A. Hogg, H. Liu, M. Jenkins, and L. Picinali, "Exploring the impact of transfer learning on GAN-based HRTF upsampling," in *Proc. EAA Forum Acusticum, Eur. Congress on Acoust.*, Sept. 2023.

[58] A. Andreopoulou and B. F. G. Katz, "Perceptual impact on localization quality evaluations of common pre-processing for non-individual head-related transfer functions," *J. Audio Eng. Soc. (AES)*, vol. 70, no. 5, pp. 340–354, May 2022.

[59] J.-H. Jung, C. S. Konor, and D. Randall, "Implementation of the vector vorticity dynamical core on cubed sphere for use in the quasi-3-D multiscale modeling framework," *J. Adv. Model. Earth Syst.*, vol. 11, no. 3, pp. 560–577, Feb. 2019.

[60] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. of the ASME J. of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, Mar. 1960.

[61] A. O. T. Hogg, P. A. Naylor, and Christine. Evers, "Speaker change detection using fundamental frequency with application to multi-talker segmentation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2019.

[62] A. O. T. Hogg, C. Evers, A. H. Moore, and P. A. Naylor, "Overlapping speaker segmentation using multiple hypothesis tracking of fundamental frequency," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1479–1490, Mar. 2021.

[63] D. Zwillinger, Ed., *CRC Standard Mathematical Tables and Formulas*, 33rd ed. Boca Raton: Chapman and Hall/CRC, Jan. 2018.

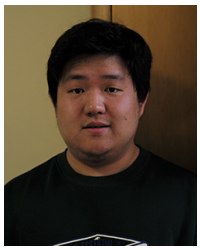
[64] J. A. Weyn, D. R. Durran, and R. Caruana, "Improving Data-Driven Global Weather Prediction Using Deep Convolutional Neural Networks on a Cubed Sphere," *J. Adv. Model. Earth Syst.*, vol. 12, no. 9, p. e02109, Aug. 2020.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. on Comput. Vision (ICCV)*, Dec. 2015, pp. 1026–1034.

[66] H. Zheng, Z. Yang, W. Liu, J. Liang, and Y. Li, "Improving deep neural networks using softplus units," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN)*, July 2015, pp. 1–4.
[67] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation functions in deep learning: A comprehensive survey and benchmark," *Neurocomputing*, vol. 503, pp. 92–108, Sept. 2022.
[68] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Joint Conf. on Learning Representations (ICLR)*, Y. Bengio and Y. LeCun, Eds., Nov. 2015.
[69] P. Gutierrez-Parera, J. J. Lopez, J. M. Mora-Merchan, and D. F. Larios, "Interaural time difference individualization in HRTF by scaling through anthropometric parameters," *EURASIP J. on Audio, Speech, and Music Process.*, vol. 2022, no. 1, p. 9, May 2022.
[70] T. McKenzie, D. T. Murphy, and G. Kearney, "Interaural level difference optimization of binaural ambisonic rendering," *Appl. Sci.*, vol. 9, no. 6, p. 1226, Jan. 2019.
[71] S. Mehrgardt and V. Mellert, "Transformation characteristics of the external human ear," *J. Acoust. Soc. Am.*, vol. 61, no. 6, pp. 1567–1576, June 1977.
[72] P. Majdak, "ARI HRTF database," June 2022. [Online]. Available: http://www.kfs.oeaw.ac.at/hrtf
[73] J. Pauwels, "The Hartufo toolkit for machine learning with HRTF data," in *Proc. Audio Eng. Soc. (AES) Conf. on Spatial and Immersive Audio*, Aug. 2023.
[74] "AES69-2015: AES standard for file exchange - Spatial acoustic data file format," Sept. 2022.
[75] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica, "Tune: A research platform for distributed model selection and training," *arXiv preprint arXiv:1807.05118*, July 2018.
[76] D. P. Kingma and L. J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. Int. Joint Conf. on Learning Representations (ICLR)*, May 2015.
[77] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein GANs," in *Proc. Neural Inform. Process. Conf*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 5769–5779.
[78] J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing spherical harmonics interpolation of time-aligned head-related transfer functions," *J. Audio Eng. Soc. (AES)*, vol. 69, no. 1/2, pp. 104–117, Feb. 2021.
[79] 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
[80] 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, May 2023.
[81] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1480–1492, Sept. 1999.
[82] ——, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *J. Acoust. Soc. Am.*, vol. 106, no. 3 Pt 1, pp. 1493–1510, Sept. 1999.
[83] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 791–802, Sept. 2015.

Aidan O. T. Hogg is a Lecturer in Computer Science at the Centre for Digital Music (C4DM) at Queen Mary University of London. He received an M.Eng. degree in electronic and information engineering and a PhD degree from Imperial College London in 2017 and 2022, respectively. He previously worked as a Research Associate in spatial audio and virtual reality with the Audio Experience Design group at Imperial College London, where he is now an Honorary Research Associate. He has also worked in various engineering roles with Broadcom, Dialog Semiconductor, and Nuance Communications. His current research focuses on using deep learning to capture head-related transfer functions and, more generally, spatial acoustics and immersive audio. Other research interests include speaker diarization and statistical signal processing for audio applications. More information about current research projects can be found here: https://aidanhogg.uk/

Mads Jenkins received BSc degrees in Computer Science and in Economics from the Massachusetts Institute of Technology in 2017, and received an MSc degree in Advanced Computing from Imperial College London in 2022. She conducted this spatial audio project with the Audio Experience Design group at Imperial College London, and now works as a machine learning engineer at Cohere, where she develops large language models for conversational applications.



He Liu received the BA (Mod) Honors degree in Computer Science from Trinity College Dublin in 2021 and MSc degree in Advanced Computing from Imperial College London in 2022. He has worked on projects related to machine learning and deep learning, including human position recognition and video/audio processing.

Isaac Squires is a Ph.D. student in the Dyson School of Design Engineering at Imperial College London. He is supervised by Dr. Samuel J. Cooper, and his research focuses on machine-learning-driven characterization and optimization of battery materials, alongside electrochemical modelling of Li-ion batteries.

Samuel J. Cooper is a Senior Lecturer in the Design of Energy Materials at Imperial College London. His team focus on the development of open-source tools for the characterisation and optimisation of energy storage devices using simulations and machine learning. https://tldr-group.github.io/

Lorenzo Picinali is a Reader in Audio Experience Design at Imperial College London. In the past years he worked in Italy, France, and the UK on projects related with 3D binaural sound rendering, interactive applications for visually and hearing impaired individuals, audiology and hearing aids technology, audio and haptic interaction and, more in general, acoustical virtual and augmented reality. More information about the projects in which Lorenzo is involved can be found here https://www.axdesign.co.uk/