frontiers | Frontiers in Signal Processing

TYPE Original Research
PUBLISHED 23 August 2022
DOI 10.3389/frsip.2022.904398

OPEN ACCESS

EDITED BY
Hyunkook Lee,
University of Huddersfield,
United Kingdom

REVIEWED BY
Gavin Kearney,
University of York, United Kingdom
Kaushik Sunder,
Embody, United States

*CORRESPONDENCE
Lorenzo Picinali,
l.picinali@imperial.ac.uk

SPECIALTY SECTION
This article was submitted to Audio and Acoustic Signal Processing, a section of the journal Frontiers in Signal Processing

RECEIVED 25 March 2022
ACCEPTED 14 July 2022
PUBLISHED 23 August 2022

CITATION
Siripornpitak P, Engel I, Squires I, Cooper SJ and Picinali L (2022), Spatial up-sampling of HRTF sets using generative adversarial networks: A pilot study.
Front. Sig. Proc. 2:904398.
doi: 10.3389/frsip.2022.904398

Spatial up-sampling of HRTF sets using generative adversarial networks: A pilot study

Pongsakorn Siripornpitak[1], Isaac Engel[1], Isaac Squires[2], Samuel J. Cooper[2] and Lorenzo Picinali[1]*

[1]Audio Experience Design, Dyson School of Design Engineering, Imperial College London, London, United Kingdom, [2]Tools for Learning, Design and Research, Dyson School of Design Engineering, Imperial College London, London, United Kingdom

Headphones-based spatial audio simulations rely on Head-related Transfer Functions (HRTFs) in order to reconstruct the sound field at the entrance of the listener's ears. A HRTF is strongly dependent on the listener's specific anatomical structures, and it has been shown that virtual sounds recreated with someone else's HRTF result in worse localisation accuracy, as well as altering other subjective measures such as externalisation and realism. Acoustic measurements of the filtering effects generated by ears, head and torso has proven to be one of the most reliable ways to obtain a personalised HRTF. However this requires a dedicated and expensive setup, and is time-intensive. In order to simplify the measurement setup, thereby improving the scalability of the process, we are exploring strategies to reduce the number of acoustic measurements without degrading the spatial resolution of the HRTF. Traditionally, spatial up-sampling of HRTF sets is achieved through barycentric interpolation or by employing the spherical harmonics framework. However, such methods often perform poorly when the provided HRTF data is spatially very sparse. This work investigates the use of generative adversarial networks (GANs) to tackle the up-sampling problem, offering an initial insight about the suitability of this technique. Numerical evaluations based on spectral magnitude error and perceptual model outputs are presented on single spatial dimensions, therefore considering sources positioned only in one of the three main planes: Horizontal, median, and frontal. Results suggest that traditional HRTF interpolation methods perform better than the proposed GAN-based one when the distance between measurements is smaller than 90°, but for the sparsest conditions (i.e., one measurement every 120°–180°), the proposed approach outperforms the others.

KEYWORDS

binaural spatialisation, immersive audio, generative adversarial networks, HRTF, auditory models

1 Introduction

Spatial audio simulations allow listeners to perceive sounds as if they were coming from particular locations in the surrounding space. Applications that could benefit from this immersive 3D audio experience include extended reality (XR, comprising virtual, augmented and mixed reality), 3D video games, auditory displays and hearing assistive devices. The binaural spatialisation technique attempts to recreate the soundfield incident at the two ear canals using a simple pair of headphones, by taking into account how the sound wave is affected by the head, ears, and torso. Such information is contained in an acoustic filter known as the head-related transfer function, which includes both interaural (different between the two ears) and spectral (i.e., the same at both ears) localisation cues (Blauert, 1983). For the sake of simplicity, in this paper we will use the expression 'HRTF set' for the full transfer function and, by extension, the full set of measurements (i.e., one measurement for each possible source position). We then use the term HRTF to refer to each of the individual measurements or estimations of this function at various source locations, which together characterise the HRTF set.

A HRTF set is strongly dependent on the listener's specific anatomical structures (external ear, head, shoulders and torso) and it has been shown that virtual sounds recreated with someone else's HRTF set result in worse localisation accuracy (Wenzel et al., 1993; Møller et al., 1996) and, potentially, can alter other subjective measures such as externalisation and realism (Simon et al., 2016; Werner et al., 2016; Engel et al., 2019). Furthermore, a high spatial and frequency resolution of HRTF measurements is required in order to obtain optimal 3D audio reproduction (Wenzel et al., 1993). Therefore, the use of high-resolution individual HRTF sets is recommended when the goal is to provide the best possible experience.

Acoustically measuring spatially dense HRTFs for each individual listener is not a scalable solution, given its cost in time and resources (i.e., need for specialised equipment). Therefore, it is relevant to investigate spatial up-sampling methods, which would allow us to generate dense HRTFs from a sparse set of measurements, and therefore simplify the measurement process (Zhong and Xie, 2014). Traditionally, spatial up-sampling of HRTFs is achieved through barycentric interpolation (Cuevas-Rodríguez et al., 2019) or by employing the spherical harmonics framework (Evans et al., 1998). Such methods are very efficient from a computational point of view, but tend to perform poorly (i.e., result in inaccurate reconstructions) when the provided HRTF data is spatially very sparse. The main reason for this is due to the fact that they cannot generate missing data, but only average between existing data points. For instance, barycentric interpolation consists in calculating a weighted average of three neighbours around the target direction. The larger the distance between these neighbours, the more likely the interpolation will be inaccurate, e.g., averaging the HRTFs at azimuth angles 0°and 90°will likely not be a good approximation of the HRTF at 45°.

Generative adversarial networks (GANs) are a family of machine learning models characterised by the use of two networks competing in an adversarial game. GANs are capable of learning to generate samples from the underlying probability distribution of an input training dataset. Given the previous success of GANs when applied to the super-resolution of photographs (Dong et al., 2015), or improving astronomical images (Schawinski et al., 2017), it seemed reasonable that they could also be used to the up-sample sparse HRTFs. The potential advantage of this technique over traditional interpolation methods is that it could allow to recreate information which is missing from the sparse measurements, by using the training data from other high-resolution HRTFs. To our knowledge, this is the first time GANs, and more in general machine learning techniques, are used for up-sampling HRTF data. Other ensemble techniques (i.e., which combine several models) have been developed and employed in problems where spatial dependence in the data is very relevant; an example are the random decision forests (Hengl et al., 2018). However, in order to exploit the highly-structured nature of HRTF data, we use a super-resolution framework, which is a further element of novelty in the proposed approach.

This work presents a pilot study investigating the use of GANs to tackle the HRTF up-sampling problem, specifically looking at very sparse HRTF measurements, and offering an initial insight about the suitability of this technique and setting the path for further research in the field. It is important to underline that, due to the pilot nature of the study, several simplifications have been carried out (i.e., training and evaluating the method on 1D data only; using the same ITDs for all the HRTFs, and employing minimum-phase HRTF reconstructions) which, considering the success of this first validation, will be rectified in future research.

2 Background

2.1 Head-related transfer functions

Our auditory system is able to analyse the sound-pressure signals as they reach the two eardrums, interpret the information embedded within the signals and perceive the sound as coming from a particular position in the three-dimensional space (Blauert, 1997; Cuevas-Rodríguez et al., 2019).

When both sound source and listener are fixed, the acoustical transmission from a point source to the two ears can be regarded as a linear-time-invariant (LTI) process. HRTFs are defined as the acoustical transfer function of the following LTI system. In spherical coordinates (r, θ, ϕ) where r denotes the source distance from head centre, θ represents the azimuth angle, which varies

from 0° to 360°, and ϕ represents the elevation, varying from −90° to 90°, denoting below and above.

$$H_L(r,\theta,\phi,f,a) = \frac{P_L(r,\theta,\phi,f,a)}{P_0(r,f)}$$
$$H_R(r,\theta,\phi,f,a) = \frac{P_R(r,\theta,\phi,f,a)}{P_0(r,f)} \quad (1)$$

where P_L and P_R represent sound pressures at the left and right ears respectively, and P_0 represents the free-field sound pressure at head centre with the head absent (Zhong and Xie, 2014).

Sound signals from the source undergo scattering, diffraction, and reflection off the listener's shoulders, torso, head and ears. The disturbance to the sound signal varies depending both on the anatomy of the listener and the position of the sound source. HRTFs contain both spectral and temporal cues, and are usually measured in several different positions around the head, allowing them to be used for creating virtual free-field stimuli (Wightman and Kistler, 1989). Convolving an anechoic audio signal with a HRTF provides a listener with the impression of a free-field sound source originating from the position where that HRTF was originally measured (Wightman and Kistler, 1989). In order to accurately reproduce as many positions as possible around the listener and to enable smooth head-tracked spatial audio simulations, HRTFs should be measured with a sufficiently high spatial density.

The main directional localisation cue at low frequencies is the interaural time difference (ITD), which is the time difference between sound waves at the left and right ears. Whereas the main cue for frequencies above 1.5 kHz is the interaural level difference (ILD), the pressure level difference between sound waves at the two ears. Above 5–6 kHz, spectral cues become important for front-back discrimination, as well as vertical localisation–these are sometimes referred to as monaural cues (Blauert, 1983; Zhong and Xie, 2014). Unfortunately, spectral cues are highly dependent on the listener's anatomy, particularly their pinnae shape (Kahana and Nelson, 2006). Therefore, accurate source localisation requires for the HRTF to be not only spatially dense, but also individualised to each listener (Møller et al., 1996; Stitt et al., 2019).

Typically, a HRTF of a particular listener is obtained *via* acoustic measurements (Carpentier et al., 2014), but this process is unfortunately expensive and impractical as it requires specialised equipment and a significant time commitment by the listener. Faster measurement methods exist (Zotkin et al., 2006; Richter et al., 2016), but the equipment specifications and cost are very high. Furthermore, denser spatial resolutions require longer measurement times, which can then result in errors due to listener movements. Hence, being able to up-sample sparse measurements in order to obtain spatially dense HRTFs could alleviate those constraints and facilitate access to personalised spatial audio for more listeners.

It is important to underline that other methods and approaches exist, beyond acoustical measurement, in order to synthesize/select individual/individualised HRTFs - an overview of such methods can be found in (Picinali and Katz, 2022).

2.2 HRTF spatial up-sampling

Spatial up-sampling of HRTF sets has previously been attempted using various interpolation methods. One of them is barycentric interpolation, which calculates the HRTF for a given direction as a weighted average of the nearest three or four neighbours (Hartung et al., 1999; Cuevas-Rodríguez et al., 2019). Barycentric interpolation can be seen as a 2D extension of linear interpolation, where the HRTF is estimated as the weighted average of the closest neighbours if the direction is provided along a single axis. This method has been shown to produce a sufficiently good agreement between measured and interpolated HRTFs when a relatively large number of measurements are still present (Gamper, 2013), becoming though much less reliable with very sparse measurement grids (i.e., more than 90% of points removed).

Alternatively, HRTFs may be interpolated using the spherical harmonics (SH) framework (Evans et al., 1998). This is done by representing the HRTF set as a weighted sum of surface SH, and then sampling the resulting spherical function at the desired directions. This interpolation method has been shown to be most effective if the HRTF set is initially preprocessed to reduce its direction-dependency, such as removing the ITDs prior to calculating the SH representation (Arend et al., 2021; Engel et al., 2022). This method works well when the HRTF set is sampled evenly around the sphere.

Both of the methods above tend to perform poorly (i.e., the reconstructed data is very dissimilar from the measured one) when the provided HRTF data is spatially very sparse: in the case of barycentric, because the nearest neighbours are too far from the interpolated direction; in the case of SH, because there are large unsampled gaps on the sphere. In this paper, a method for performing the spatial up-sampling of HRTF sets *via* GANs is introduced and evaluated through a pilot study, meaning that the implementation and assessment has been carried out on single dimension cases (i.e., only horizontal, median, or frontal planes). Considering that barycentric interpolation is the most common implementation in available binaural spatialisation tools (Poirier-Quinot and Katz, 2018; Cuevas-Rodríguez et al., 2019), and that in single-dimension problems it is equivalent to linear interpolation, this has been chosen as benchmark method for comparison with the proposed GAN-based approach.

2.3 Generative adversarial networks

A GAN consists of a classifier, called a discriminator (D), and a generator (G), which learns to create fake data by incorporating

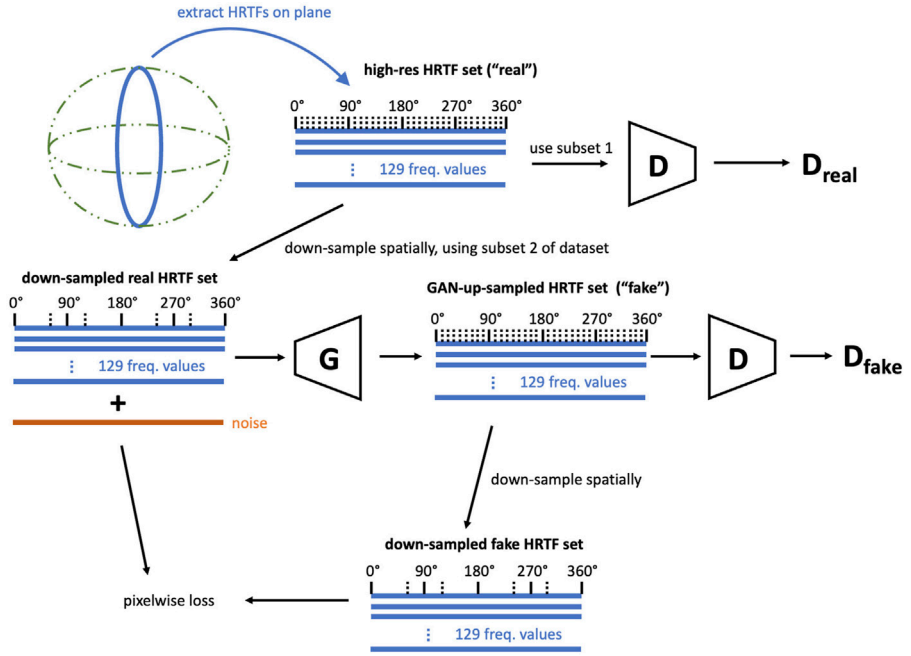

FIGURE 1
Up-sampling GAN architecture for HRTFs.

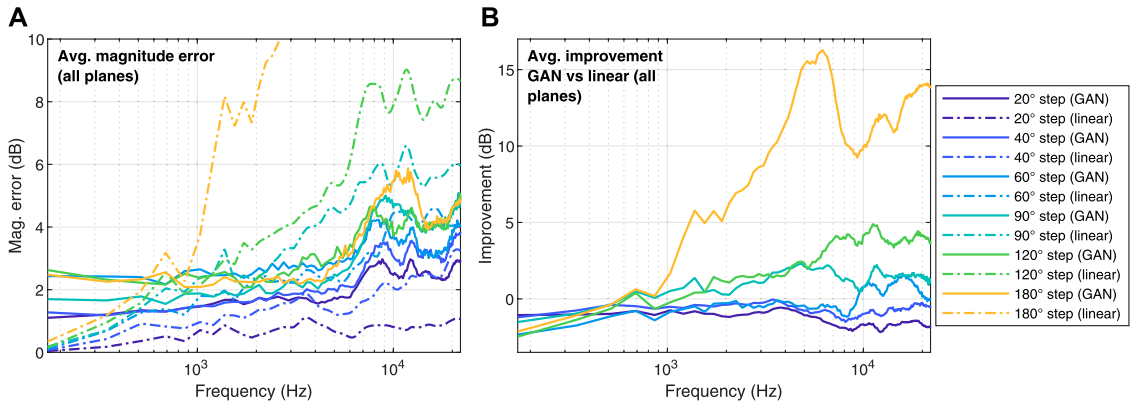

FIGURE 2
(A): magnitude error, averaged over both ears and over all available directions. **(B)**: improvement of the GAN compared with linear interpolation (values greater than zero indicate that the GAN obtained a lower magnitude error than linear interpolation). Data of all three planes were pooled together.

feedback from the discriminator. In a conventional GAN, the input to G (Figure 1) is a vector of random samples from the standard normal distribution, z, referred to as the input noise. When performing up-sampling (or super-resolution) tasks, G is also given the low-resolution (i.e., down-sampled) input. The loss function of G is adapted to include a term that punishes G for diverging from the low-resolution input.

D and G are updated iteratively and alternately in training. During training of D, this is passed a batch of real (i.e., measured) HRTFs, and then a batch of "fake" HRTFs generated by G. Each fake sample is generated by down-sampling a real HRTF, which is passed to D alongside a noise vector to G. The loss function is then calculated and the weights of D are updated using stochastic gradient descent. During training of G, a batch of measured

HRTFs is down-sampled, has noise concatenated, passed through G, and then passed through D.

3 Methods

The HUTUBS HRTF dataset was chosen for the experiment, thanks to the relatively large number of measured HRTFs, and its high spatial resolution (Brinkmann et al., 2019). Of the 94 subjects whose HRTFs sets were measured for the HUTUBS database, HRTFs sets from 89 subjects are used for training the GANs. For each subject's HRTF set, the horizontal, frontal, and median planes for each ear are separately used to train multiple independent 1D GANs. A 2D GAN would require methods to handle spatial curvature inherent in the dataset, and is not included in this pilot study. HRTF data on each plane are also spatially down-sampled and used to train each GAN. The whole process is then repeated for a range of down-sampling factors.

3.1 Preprocessing

The original HUTUBS dataset is provided in the SOFA file format, as standardised by the Audio Engineering Society (AES), and consists of head-related impulse responses (HRIRs) from 440 directions for both ears. Each HRIR is a time-domain signal consisting of 256 samples at 44,100 Hz sample rate. The ITD was removed for each HRIR using an onset threshold detection method, as described by Andreopoulou and Katz, (2017). A Hanning window was applied prior to performing the Fourier transform, ensuring that most of the HRIR energy was preserved after the windowing. Then, the HRIRs were transformed to the frequency domain using the discrete Fourier transform (DFT) to obtain the HRTFs. Then the log-magnitude of the HRTFs was calculated, which was used as input for the GANs. The HRTF phase was disregarded, assuming that the up-sampled HRTFs would be reconstructed using a minimum-phase approximation and a simple ITD model. It is known that such simplifications could have an impact on certain perceptual features of the HRTFs Andreopoulou and Katz, (2022), therefore further research beyond this pilot will probably need to also consider phase information.

The GAN is implemented in Python and takes sets of HRTF magnitudes for either training or up-sampling, once the training is complete. The output up-sampled set of HRTFs is over 36 directions for each plane.

From each of the HRTF sets we extracted data points on each of the horizontal plane, median plane, and frontal plane; for the HUTUBS dataset, there are 36 such data points on each plane, separated in 10° intervals. Data for each plane were used to train separate and independent GANs. Left and right ear HRIRs at each spatial location were also processed separately, and used to train independent GANs.

Each layer of G consists of 129 channels (linearly-spaced frequency bins). Information from any frequency can be combined in the following layer, irrespective of their frequency separation. An alternative approach would have been to also use sliding convolutional kernels in the frequency dimension (Thickstun et al., 2016).

3.2 Model architecture

Each layer in D consists of a 1D convolution layer followed by a leaky ReLU activation (Xu et al., 2015). In G, each layer consists of a 1D transpose convolution layer followed by a batch normalisation and then a ReLU activation, with the output layer having a final hyperbolic tangent activation. The ReLU and leaky ReLU were found to give better performance over the sigmoid activation function - the latter caused vanishing gradients and stopped useful training. The full structures of D and G are provided in the Supplementary Material.

The recommended Adam optimiser hyperparameters by Radford et al. (2015) were used (α = 0.0002, β_1 = 0.5 and β_2 = 0.999). This implementation of the GAN uses the canonical minimax formulation, as introduced by Goodfellow et al. (2016).

The loss functions for D and G are defined as follows:

$$D_{loss} = 0.5 \cdot \left[BCE, (\sigma(D_{real}), 1), +, B, C, E, \left(\sigma\left(D_{fake}\right), 0\right)\right] \quad (2)$$

$$G_{loss} = BCE\left(\sigma\left(D_{fake}\right), 1\right) + \alpha_{pixelwise} \cdot PWL\left(D_{real,ds}, D_{fake,ds}\right) \quad (3)$$

The loss function for G is the sum of the binary cross-entropy criterion loss, *BCE*, and a pixelwise loss term, *PWL*, which is used to anchor the output HRTFs to the down-sampled data points that were provided. Analogous to 2D images, the value at each pixel is considered to be the HRTF magnitude corresponding to one spatial location and one frequency value. The pixelwise loss is defined as the mean squared error between the down-sampled real HRTFs and the down-sampled GAN-up-sampled HRTFs, averaged across all values at each frequency value and spatial location. $\alpha_{\text{pixelwise}}$ is the pixelwise loss coefficient set at 50. In a hyperparameter sweep, using different values of $\alpha_{\text{pixelwise}}$ ranging from 0.1 to 100 did not have a large impact on the training. The losses defined above drives D to try to output 1 for real and 0 for fake, while driving G to output samples that D will classify as real (i.e., 1).

As mentioned above, three subsets of directions were extracted from each HRTF set: horizontal plane, median plane, and frontal plane. Each plane contains 36 directions, separated in 10°intervals. For each plane, five different sparse (down-sampled) subsets of directions with different angular separations were extracted: 20°, 40°, 60°, 90°, and 180°.

The dataset consisting of 94 subjects was divided into three parts of 45, 44, and 5 subjects. These were used while training the discriminator, the generator, and as a hold-out test set

respectively. The motive for splitting the data in this way was to avoid a particular kind of overfitting. The GAN is performing an up-sampling operation in which it is being asked to learn to generate full angular resolution outputs from low-resolution inputs. This is being done stochastically, as the low-resolution inputs do not contain sufficient information to perform this reconstruction exactly, which is why we concatenate noise to the input. However, if the training set was not divided, then D would have previously seen the exact full resolution output corresponding to each input and could then punish G for using the noise to add variation.

3.3 Evaluation of HRTFs

HRTFs can be evaluated experimentally by supplying real subjects with virtual sound sources and assessing their ability to accurately localise the sound (Kim et al., 2020). However, subjects should be extensively trained in order to provide a reliable level of localisation ability (Andreopoulou and Katz, 2016), and as such this validation approach was beyond the scope of this study. Other perceptual attributes, such as timbral quality or externalisation, may also be assessed (Lindau et al., 2014).

In this work, HRTFs are evaluated by predicting localisation performance using the perceptual model described in Barumerli et al. (2020), as implemented in the Auditory Modelling Toolbox (Majdak et al., 2022). This model simulates how a human listener would perform in a sound localisation task when provided with a HRTF different than their own. Therefore, the quality of both the GAN-reconstructed and linearly interpolated HRTFs can be assessed by looking at the predicted localisation error. This error is reported as the average great-circle distance between the actual and the predicted directions.

For the model-based evaluation, the HRTFs were reconstructed as minimum-phase filters from the interpolated log-magnitudes, according to Oppenheim et al. (2001). Then, ITDs were reinserted using the classic Woodworth formula (Woodworth et al., 1954). Note that this procedure was also applied to the linearly interpolated HRTFs as well as the reference full-resolution HRTFs, in order to ensure a fair comparison. The implications of this choice are discussed in Section 5.

4 Results

4.1 GAN training and output

Each GAN is trained with an early stopping criterion defined where pixelwise loss falls below 0.002. The pixelwise loss is used as a metric of similarity and indicates that the GAN is able to reproduce data at the spatial locations where HRTF magnitudes were provided to it. A batch size of 5 subjects are used in each iteration. Losses were monitored for both D and G to detect issues such as vanishing gradients.

The noise had little to no effect on the output of G suggesting mode collapse has occurred. This lack of stochasticity and reduction in the diversity of outputs could potentially impact the quality of training, as mode collapse results in D receiving the same signal from G. The suspected cause of the mode collapse is insufficient training data. Some potential methods to mitigate this include using a Wasserstein GAN with gradient penalty or training a smaller, less complex D although these were not in the scope of this project. Usually, mode collapse is to be avoided if the GAN produces a large variety of outputs. However, in this case we only require a single useful high-resolution output, therefore the model was still able to generate useful results.

4.2 Interpolation error and perceptual model output

The HRTF interpolation error was calculated as the absolute difference between the log-magnitude of the original HRTFs and the interpolated/up-sampled ones. The results, averaged across directions, across the five test subjects (i.e., the untrained HRTFs) and across the left and right channels, are shown in Figure 2 (horizontal, median and frontal planes pooled together) and Figure 3 (separated per plane). The plots show (Figure 2A and top in Figure 3) the error for both the GAN-reconstructed HRTFs and the linearly interpolated ones, as well as the difference between the two (Figure 2B and bottom in Figure 3).

According to these visualisations, the HRTFs reconstructed by the GAN displayed significantly smaller magnitude errors than the linearly interpolated ones for the sparsest conditions, such as 120°and 180°angle steps. On the other hand, this did not happen for the denser conditions, such as 20°and 40°angle steps. Furthermore, these plots suggest that the magnitude error improvement of the GAN with respect to linear interpolation was greater for the median and frontal planes than for the horizontal plane.

The results of the localisation model evaluation are displayed in Figure 4. For each test subject and each interpolated/up-sampled HRTF direction, the model simulated 100 repetitions of a perceptual experiment where a listener is asked to localise a sound source spatialised with said HRTF. The reason for the multiple repetitions is to account for the stochasticity of the model, which mimics the behaviour of listeners in real experiments. The plots show the average great-circle distance between the direction of the actual HRTF the one predicted by the model.

According to these plots, localisation errors increase monotonically with sparsity for linearly interpolated HRTFs, providing the best performance for the smallest angle step (20°) and very large errors for the largest step (180°). In comparison, GAN-up-sampled HRTFs are less affected by the sparsity, displaying comparatively small errors for the sparsest conditions. This trend seems to hold true for all three planes, although the errors are overall larger in the median plane compared to the other two. It is worth noting that consistent magnitude errors of at least 1 dB are observed

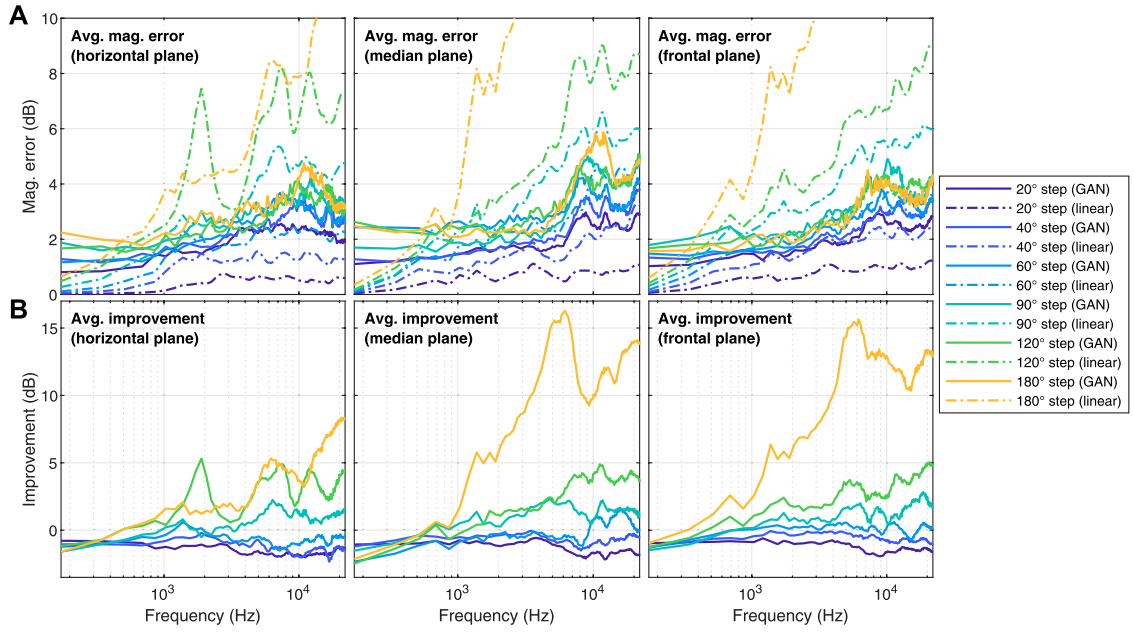

FIGURE 3
(A): magnitude error, averaged over both ears and over all available directions. **(B)**: improvement of the GAN compared with linear interpolation (values greater than zero indicate that the GAN obtained a lower magnitude error than linear interpolation). Data were split in horizontal, median and frontal planes.

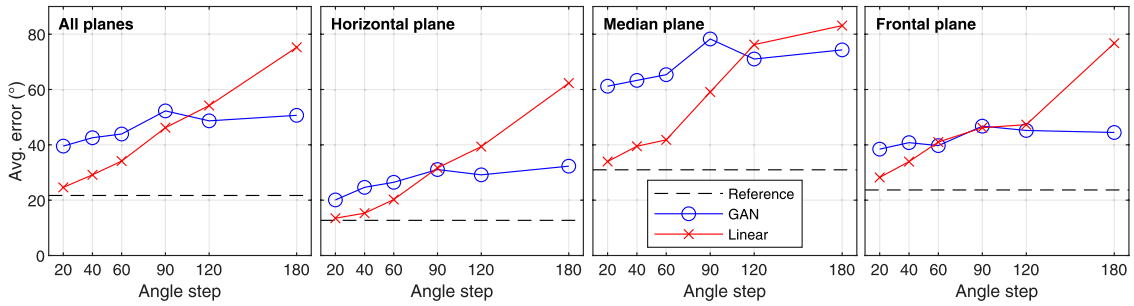

FIGURE 4
Average localisation error, as predicted by the model by Barumerli et al. (2020) (lower is better). The dashed line represents the ideal case of a real listener's performance when localising a sound source with their own HRTF.

in the sparser conditions when using the GAN method, increasing to at least 3 dB for frequencies above 6 kHz (larger errors appear when using linear interpolation). This is likely to lead to perceivable tonal differences in the binaural signals rendered with said HRTFs, and will need to be addressed in future research.

5 Discussion

Before discussing each specific output of the numerical evaluations, it is important to note that the results of the magnitude error analysis generally agree with the localisation model predictions; as expected, larger magnitude errors seem to lead to larger localisation errors.

Looking at the linear interpolation (used as benchmark method), errors increase with sparsity, as expected. For the sparsest conditions (120°and 180°steps), errors are very large, which indicates that the interpolated HRTFs will very probably provide poor spatial audio quality, likely resulting in low localisation performances and reduced realism of the rendering. Also for the GAN up-sampling errors increased with sparsity, but seem to plateau after 90°separation steps, outperforming linear interpolation for the sparsest conditions. This suggests that, for situations where only a few directions of the HRTF set are available,

the GAN is superior to linear interpolation and will likely provide higher quality spatial audio renderings in terms of localisation accuracy. In contrast, for the conditions with steps smaller than 90°, traditional interpolation methods would be a better option.

Considering specifically the perceptual model output, and the average error for the three separate planes, it is important to underline that horizontal errors are attributed to interaural cues. Since both the GAN-up-sampled and linearly interpolated HRTFs were generated as minimum-phase filters, we can safely assume that ITDs are identical in all cases, so ILDs are the only relevant interaural cue influencing horizontal localisation errors. Looking at the second subplot in Figure 4, we can infer that the GAN-up-sampled HRTFs obtained smaller ILD errors, and therefore smaller localisation errors than the linear counterpart for angle steps greater than 90°. In the case of median plane localisation, monaural cues are dominant, and we observe that GAN-up-sampled HRTFs outperformed the linear version for angle steps of 120°and 180°, but with the differences smaller in this case. Frontal plane localisation involves both interaural and monaural cues, and as expected it displays similar trends to the other two planes.

Since all HRTFs have been transformed to minimum phase, it is safe to assume that all the localisation predictions are largely based on spectral features. This is not only true for the median plane but also for the horizontal and frontal planes, where up/down and front/back reversals are still possible, and ILDs become a relevant factor. The results suggest that, when compared to linear interpolation, the GAN performs better at reconstructing interaural spectral cues than monaural ones. This would explain why the predicted localisation improvements by the GAN are larger for the horizontal and frontal planes than for the median plane. This could be due to the GAN needing a larger training dataset (e.g., more HRTFs) in order to predict monaural cues more accurately. Furthermore, we speculate that once the the GAN is trained on full-sphere rather than individual plane data, these results will significantly improve.

Considering the initial aim of this pilot study, which was to investigate the use of GANs to tackle the HRTF spatial up-sampling problem, comparing its performances with a traditional linear interpolation method, and offering an initial insight about the suitability of this technique, the results of the numerical evaluations seem to validate the proposed GAN-based method. This performs notably better than the interpolation benchmark for steps larger than 90°. This was validated both by spectral magnitude error analyses, and by computing the output of a perceptual model predicting sound sources localisation accuracy.

6 Conclusion and future work

Acoustic measurement of HRTFs is currently time consuming and expensive to obtain due to the personalised nature of HRTF sets and the need for them to be spatially dense. This study demonstrates that up-sampling a HRTF through a data-driven approach using GANs has the potential to achieve better localisation performance than when using linear interpolation. The GANs provides similar localisation results to linear interpolation when many data points are provided, but outperforms it when the spatial resolution is dramatically reduced (e.g., sampling every 120°or 180°). This opens up the potential impact of the proposed work to being able to predict high resolution individual HRTFs using only a very low number of source positions (e.g., in the order of four to six measurements), which could be measured also in uncontrolled environments and without the need for expensive hardware setups.

It has to be underlined though that this study presents only a pilot experiment; the GAN architecture is relatively simple and future research is likely to produce improvements. However, transferring this approach from a great circle to the entire surface of a sphere is non-trivial. One option for achieving this is graph neural networks, which have been used to build GANs for arbitrary particle detector topologies (Kansal et al., 2020).

The current GAN-based approach cannot yet be directly compared to other state-of-the-art HRTF interpolation methods, such as the ones described by Arend et al. (2021), since it up-samples over 1 dimension only. Introducing up-sampling over an additional dimension (e.g., azimuth, in addition to elevation) would allow a direct comparison, but would require a compression or mapping of the 3D surface to 2D, possibly using graph networks and 2D convolutional layers for both the discriminator and the generator. Furthermore, the localisation model could be added into the loss function so that the GAN can optimise for its performance directly. Additionally, future work should look into processing the HRTF phase as well as its magnitude, and training the model using both left and right ear signals on the same input data.

In the future, other perceptually relevant metrics could be evaluated in addition to localisation. For instance, a relevant one is *externalisation*, defined as the degree to which a virtual sound is perceived to come from outside of the head, which could be evaluated with appropriate auditory models (Baumgartner and Majdak, 2021), or through carrying out behavioural experiments. The latter could also be used to validate the results of the localisation model, and allow comparisons with results from other studies, aiming to ultimately validate the proposed GAN-based approach.

Data availability statement

The publicly-available HOTUBS dataset was used in this study, and can be found here: https://depositonce.tu-berlin.de/handle/11303/9429.

Author contributions

PS, SC, and LP designed the study. PS, IS, and SC developed the GAN approach. PS implemented the test and created the data. IE and LP carried out the data analysis. All authors contributed to the writing of the manuscript.

Funding

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.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frsip.2022.904398/full#supplementary-material

References

Andreopoulou, A., and Katz, B. F. G. (2016). Investigation on subjective HRTF rating repeatability. *Audio Eng. Soc. Conv.* 140, 95971–96010.

Andreopoulou, A., and Katz, B. F. G. (2017). Identification of perceptually relevant methods of inter-aural time difference estimation. *J. Acoust. Soc. Am.* 142, 588–598. doi:10.1121/1.4996457

Andreopoulou, A., and Katz, B. F. (2022). Perceptual impact on localization quality evaluations of common pre-processing for non-individual head-related transfer functions. *J. Audio Eng. Soc.* 70, 340–354. doi:10.17743/jaes.2022.0008

Arend, J. M., Brinkmann, F., and Pörschmann, C. (2021). Assessing spherical harmonics interpolation of time-aligned head-related transfer functions. *J. Audio Eng. Soc.* 69, 104–117. doi:10.17743/jaes.2020.0070

Barumerli, R., Majdak, P., Reijniers, J., Baumgartner, R., Geronazzo, M., and Avanzini, F. (2020). Predicting directional sound-localization of human listeners in both horizontal and vertical dimensions. *Audio Eng. Soc. Conv.* 148.

Baumgartner, R., and Majdak, P. (2021). Decision making in auditory externalization perception: Model predictions for static conditions. *Acta Acust. (2020).* 5, 59. doi:10.1051/aacus/2021053

Blauert, J. (1997). *An introduction to binaural technology. Binaural and spatial hearing in real and auditory environments*. Mahwah NJ: Lawrence Erlbaum.

Blauert, J. (1983). *Spatial hearing : The psychophysics of human sound localization*. Cambridge, Mass: MIT Press.

Brinkmann, F., Dinakaran, M., Pelzer, R., Grosche, P., Voss, D., Weinzierl, S., et al. (2019). A cross-evaluated database of measured and simulated hrtfs including 3d head meshes, anthropometric features, and headphone impulse responses. *J. Audio Eng. Soc.* 67, 705–718. doi:10.17743/jaes.2019.0024

Carpentier, T., Bahu, H., Noisternig, M., and Warusfel, O. (2014). "Measurement of a head-related transfer function database with high spatial resolution," in *7th forum acusticum (EAA)*.

Cuevas-Rodríguez, M., Picinali, L., González-Toledo, D., Garre, C., de la Rubia-Cuestas, E., Molina-Tanco, L., et al. (2019). 3d tune-in toolkit: An open-source library for real-time binaural spatialisation. *Plos One* 14, e0211899. doi:10.1371/journal.pone.0211899

Dong, C., Loy, C. C., He, K., and Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 295–307. doi:10.1109/tpami.2015.2439281

Engel, I., Alon, D. L., Robinson, P. W., and Mehra, R. (2019). "The effect of generic headphone compensation on binaural renderings," in Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio (Audio Engineering Society).

Engel, I., Goodman, D. F. M., and Picinali, L. (2022). Assessing HRTF preprocessing methods for Ambisonics rendering through perceptual models. *Acta Acust. (2020).* 6, 4. doi:10.1051/aacus/2021055

Evans, M. J., Angus, J. A. S., and Tew, A. I. (1998). Analyzing head-related transfer function measurements using surface spherical harmonics. *J. Acoust. Soc. Am.* 104, 2400–2411. doi:10.1121/1.423749

Gamper, H. (2013). Head-related transfer function interpolation in azimuth, elevation, and distance. *J. Acoust. Soc. Am.* 134, EL547–EL553. doi:10.1121/1.4828983

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. Cambridge, MA: The MIT Press.

Hartung, K., Braasch, J., and Sterbing, S. J. (1999). "Comparison of different methods for the interpolation of head-related transfer functions," in Audio Engineering Society 16th International Conference: Spatial Sound Reproduction (Audio Engineering Society).

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., and Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ* 6, e5518. doi:10.7717/peerj.5518

Kahana, Y., and Nelson, P. A. (2006). Numerical modelling of the spatial acoustic response of the human pinna. *J. Sound Vib.* 292, 148–178. doi:10.1016/j.jsv.2005.07.048

Kansal, R., Duarte, J., Orzari, B., Tomei, T., Pierini, M., Touranakou, M., et al. (2020). Graph generative adversarial networks for sparse data generation in high energy physics. *arXiv* [Preprint]. Available at: https://arxiv.org/abs/2012.00173.

Kim, C., Lim, V., and Picinali, L. (2020). Investigation into consistency of subjective and objective perceptual selection of non-individual head-related transfer functions. *J. Audio Eng. Soc.* 68, 819–831. doi:10.17743/jaes.2020.0053

Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F., Weinzierl, S., et al. (2014). A spatial audio quality inventory (saqi). *Acta Acustica united Acustica* 100, 984–994. doi:10.3813/aaa.918778

Majdak, P., Hollomey, C., and Baumgartner, R. (2022). Amt 1.0: The toolbox for reproducible research in auditory modeling. *Acta Acustica* 6, 19.

Møller, H., Sørensen, M. F., Jensen, C. B., and Hammershøi, D. (1996). Binaural technique: Do we need individual recordings? *J. Audio Eng. Soc.* 44, 451–469.

Oppenheim, A. V., Buck, J. R., and Schafer, R. W. (2001). *Discrete-time signal processing*, Vol. 2. Upper Saddle River: Prentice-Hall.

Picinali, L., and Katz, B. F. (2022). "System-to-user and user-to-system adaptations in binaural audio," in *Sonic interactions in virtual environments*. Editors M. Geronazzo and S. Serafin (Springer), 121–144.

Poirier-Quinot, D., and Katz, B. F. (2018). "The anaglyph binaural audio engine," in *Audio engineering society convention* (Audio Engineering Society), 144.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*

Richter, J.-G., Behler, G., and Fels, J. (2016). "Evaluation of a fast hrtf measurement system," in *Audio engineering society convention* (Audio Engineering Society), 140.

Sato, H., Morimoto, M., and Sato, H. (2020). Perception of azimuth angle of sound source located at high elevation angle: Effective distance of auditory guide signal. *Appl. Acoust.* 159, 107084. doi:10.1016/j.apacoust.2019.107084

Schawinski, K., Zhang, C., Zhang, H., Fowler, L., and Santhanam, G. K. (2017). Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit. *Mon. Not. R. Astron. Soc. Lett.* 467, slx008. doi:10.1093/mnrasl/slx008

Simon, L. S. R., Zacharov, N., and Katz, B. F. G. (2016). Perceptual attributes for the comparison of head-related transfer functions. *J. Acoust. Soc. Am.* 140, 3623–3632. doi:10.1121/1.4966115

Stitt, P., Picinali, L., and Katz, B. F. G. (2019). Auditory accommodation to poorly matched non-individual spectral localization cues through active learning. *Sci. Rep.* 9, 1063. doi:10.1038/s41598-018-37873-0

Thickstun, J., Harchaoui, Z., and Kakade, S. (2016). Learning features of music from scratch. *arXiv preprint arXiv:1611.09827*

Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am.* 94, 111–123. doi:10.1121/1.407089

Werner, S., Klein, F., Mayenfels, T., and Brandenburg, K. (2016). "A summary on acoustic room divergence and its effect on externalization of auditory events," in 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX) (IEEE), 1. doi:10.1109/QoMEX.2016.7498973

Wightman, F. L., and Kistler, D. J. (1989). Headphone simulation of free-field listening. I: Stimulus synthesis. *J. Acoust. Soc. Am.* 85, 858–867. doi:10.1121/1.397557

Woodworth, R. S., Barber, B., and Schlosberg, H. (1954). *Experimental psychology*. Holt, NY: Oxford and IBH Publishing.

Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*

Zhong, X.-L., and Xie, B.-S. (2014). "Head-related transfer functions and virtual auditory display," in *Soundscape semiotics - localisation and categorisation* (InTech). doi:10.5772/56907

Zotkin, D. N., Duraiswami, R., Grassi, E., and Gumerov, N. A. (2006). Fast head-related transfer function measurement via reciprocity. *J. Acoust. Soc. Am.* 120, 2202–2215. doi:10.1121/1.2207578