Audio Engineering Society

Conference Paper 4

Presented at the AES International Conference on Machine Learning and Artificial Intelligence for Audio
2025 September 8–10, London, UK

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Extraction and neural synthesis of low-order spectral components for Head-Related Transfer Functions

Mary Pilataki[1, 2], Chris G. Buchanan[1], and Cal Armstrong[1]

[1] *PlayStation, Sony Interactive Entertainment, London UK*
[2] *Queen Mary University of London, London, UK*

Correspondence should be addressed to Cal Armstrong (calum.armstrong@sony.com)

ABSTRACT

Audio spatialisation is a crucial component of immersive technologies, shaping the quality of experiences in VR, AR and gaming. Spatial audio rendering relies on Head-Related Transfer Functions (HRTFs), which encode both spatial localisation cues and listener-dependent spectral characteristics associated with perceptual attributes such as sound colouration. However, research has largely prioritised localisation accuracy, leaving the remaining subject-specific spectral features underexplored. Meanwhile, 3D audio renderers often employ HRTFs with entangled spectral colouration profiles and localisation cues, leading to cases where localisation is accurate, but the resulting sound coloration is perceived as unnatural or undesirable. In this paper, we introduce a novel signal processing method, which we term the Delta Spectral Transfer Function (DSTF) which aims to separate low-order spectral characteristics from directional cues. We also present a conditional autoencoder model trained to reconstruct and synthesise new DSTF profiles. To ensure disentanglement between localisation and DSTF characteristics, we incorporate adversarial learning. The resulting model enables (1) synthesis of novel location-agnostic spectral characteristics, (2) neural compression of databases of such features, and (3) analysis and exploration of HRTF variation beyond spatial encoding.

1 Introduction

Head-Related Transfer Functions (HRTFs) are frequency- and time-dependent filters that represent the stereo anechoic acoustic transfer function between a positional sound source and a listener's ears. In the time domain, these are referred to as Head-Related Impulse Responses (HRIRs). An individual's HRTF is commonly measured at various angles around their head, typically referenced with respect to azimuth and elevation. The left and right ear responses differ and are both encoded into the HRTF. Examples of the specialised equipment used to measure HRTFs can be found in [1, 2, 3].

HRTFs provide critical spatial cues that enable listeners to localise sounds in three-dimensional (3D) space. These cues include interaural time differences (ITD) and interaural level differences (ILD), primarily associated with lateral localisation, and the frequency-dependent peaks and notches in the ear's response, which are key for elevatory localisation [4]. By convolving an audio signal with an HRIR and presenting the result directly to a listener's ears (typically via

headphones), a sound source can be simulated as if it originates from the direction in which the HRTF was measured [1]. This binaural rendering process has become a fundamental technique for simulating immersive acoustic environments in augmented reality (AR), virtual reality (VR), gaming, and entertainment.

In addition to localisation, HRTFs are associated with other perceptual attributes, including realism, externalisation, and colouration [4, 5]. While early efforts in HRTF personalisation predominantly focused on improving localisation accuracy, subject-specific spectral characteristics associated with such perceptual attributes, particularly those related to perceived coloration, have been under-explored. When an HRTF is applied to an audio signal, distinct changes in sound colouration are perceived. This colouration is primarily shaped by low-order spectral components that vary across listeners, influenced by factors such as the shape of the head, pinnae, and torso, as well as early reflections from the shoulders and torso [4]. Clearly, these physical variations not only affect the localisation cues within the HRTF but also impact sound coloration. Furthermore, listening test results have shown that the sound colouration introduced by a different set of HRTF may be more pleasant for some people than that of their own HRTF [1]. Spatially accurate but perceptually unpleasant renderings can disrupt immersion, suggesting that non-localisation-related spectral variation should be treated as a first-class perceptual attribute in spatial audio design.

We propose a novel signal processing method to analyse HRTFs and extract listener-dependent low-order spectral features. Using these features, we train a conditional autoencoder model in an adversarial learning framework to disentangle localisation and non-spatial spectral information such that the latent space exclusively encodes direction-independent spectral coloration components. The decoder module can then be used as a synthesiser for novel spectral characteristics, providing separate control over localisation and coloration effects. This enables exploration of the non-directional spectral variation in HRTFs. By sampling latent embeddings, we can generate a set of spectral filters across a 3D sphere, corresponding to different directions. Our model can be employed as a tool for personalising HRTFs by substituting spectral colouration filters within existing HRTFs to modify their perceptual character without altering localisation cues.

Finally, the learned latent space can serve as a neural compression tool, enabling efficient storage of a pre-calculated selection of such low-order spectral filters. The paper is organised as follows. Section 2 reviews state-of-the-art research on HRTF personalisation and spectral modelling using autoencoders. In Section 3, we describe the Sonicom HRTF database used in our experiments. Section 4 details the experimental setup, including feature extraction, dataset creation, and the deep learning model architecture and training details. Section 5 presents the results, discussing spectral reconstruction, synthesis, and neural compression, followed by the conclusion in Section 6.

2 Related Work

2.1 Head-Related Transfer Functions

Although HRTF personalisation is an active area of research, the analysis and synthesis of non-localisation-related spectral characteristics as an exclusive feature has not been extensively explored in the literature. Marentakis and Saitis [6] analyze the changes in the timbre descriptors *log attack time* and *spectral centroid* between anechoic and HRTF-filtered sounds. Some comparison is also presented between an individual's left and right ear HRTFs; however, variations in HRTFs between different subjects are beyond the scope of that work. Marggraf-Turley et al. [7] propose a method for HRTF recommendation based on Predicted Binaural Colouration (PBC) [8], where the goal is to minimise the colouration perceived by a target subject when compared to their own individual HRTF. Although this method does account for sound colouration, it is only relative to an individual's HRTF, does not allow for further personalisation, and requires the estimation of an individualised HRTF based on anthropometric features.

In this work, we focus on the perceptual attribute of sound coloration by removing dominant localisation-related information and considering only the residual subject-variant elements.

2.2 Spectral modeling with autoencoders

Variational autoencoders (VAEs) have been proposed for music audio coding, synthesis of novel timbres, and timbre exploration, where a low-dimensional latent space is learned through a guided process to achieve specific goals. For example, Esling et al. [9] propose

regularisation to ensure the latent space matches timbre qualities identified in timbre research. Colonel and Keene [10] propose conditioning using harmonic features so that the latent space becomes uniform in terms of pitch and timbre. Puche and Lee [11] also use conditioning combined with adversarial learning to disentangle pitch and timbre, ensuring that the latent space exclusively encodes non-pitch-related spectral information.

Inspired by this work and motivated by the findings in [12], which highlight the effectiveness of disentangling qualities in an autoencoder's latent space, we guide the learning process of our framework so that the latent space exclusively encodes the low-order, location-invariant spectral components of HRTFs that mainly affect coloration. Following a similar approach to [11], we employ adversarial learning to separate direction-dependent localisation cues from subject-dependent spectral coloration. This framework allows us to generate novel spectral features and provides separate control over both directional and subject-specific spectral components, enabling targeted manipulation of perceptual coloration in HRTFs.

3 HRTF Database

For our experiments, we use the Sonicom database [3], which consists of 200 subjects, each with 828 HRTF measurements. Azimuths are sampled in 5° increments between -175°to 180°. Elevations are sampled every 10° between -30° to 30°, and every 15° below and above that (-45°to 90°).

We use free-field compensated HRTFs, recorded at a sampling rate of 48 kHz. Each HRTF in the database is preprocessed as follows: 1) We remove unwanted floor reflections by applying a 75-sample unity pad followed by a 50-sample half-Hann closing window, 2) we apply diffuse field equalisation using the method proposed in [1], 3) we apply loudness normalisation according to EBU R128 [13], 4) we re-interpolate the measurement grid to a new set of azimuths ranging from -172.5° to 180° and elevations between -60° to 60°. The interpolation is performed in the frequency domain using the barycentric weighted Delaunay triangle method.

4 Method

4.1 Dataset

Each ear in the database is treated as a distinct data instance, with a unique Delta Spectral Transfer Function (DSTF) feature associated with every ear. To enforce left-right symmetry prior to feature extraction (detailed in Section 4.2), all right-ear measurements are horizontally flipped to represent left-ear equivalents. This results in a total of $200 \times 2 = 400$ unique DSTFs across the dataset. As described in Section 4.2 and shown in Figures 1 and 4, each DSTF feature corresponds to a single ear at a specific azimuth and covers 17 elevation angles from -60°to 60°. With 48 azimuth angles and 400 DSTFs, the full dataset comprises $48 \times 400 = 19{,}200$ DSTF features.

The dataset is split into training and testing subsets using an approximate 85%–15% ratio. To construct the test set, we randomly select 10 azimuth angles for 146 unique left-ear instances and 146 right-ear instances (flipped to left-ear equivalents), resulting in 2,920 DSTF features for evaluation. This yields $2 \times 146 \times 10 = 2{,}920$ DSTF features for evaluation, with the remaining features used for training.

Each DSTF feature is annotated with one-hot encoded labels for both subject identity and azimuth. Subject labels are 400-dimensional binary vectors, where a single 1 indicates which subject (i.e., ear) the feature belongs to. Azimuth labels are 48-dimensional vectors representing azimuths in ascending order: index 0 corresponds to an azimuth of -172.5°, index 1 corresponds to -167.5°, and so on, up to index 47, which corresponds to 180°.

4.2 Feature extraction

In this work, we focus on capturing low-order spectral colourations in HRTFs, which we refer to as the *Delta Spectral Transfer Function* (DSTF). The goal of DSTFT feature extraction is twofold: (1) to remove the diffuse-field response, which may account for residual measurement equalisation or approximate physiological constants such as ear canal resonance and (2) to eliminate features that are known from prior literature to primarily support sound localisation. The remaining components form a filter set that represents the uncharacterised residual elements of the HRTF. We consider the inter-subject variation of this filter set and propose interpreting it as subject-specific sound coloration not directly related to localisation cues.

To extract the DSTF, we apply the following steps:

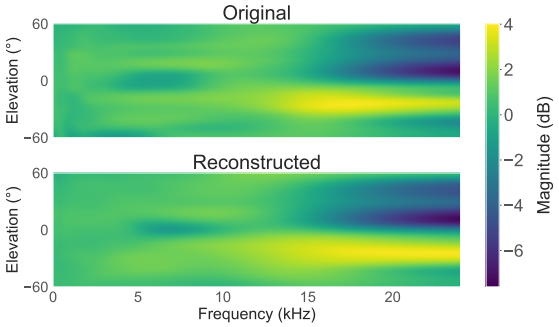


Fig. 1: Comparison between input feature (Original) and model output (Reconstructed) for subject P0001 at azimuth 0°. The x-axis represents frequency (0 Hz up to 24 kHz) and the y-axis represents elevation (-60 up to 60°). Colour represents magnitude values in dB.

1. **Notch Removal**: We first remove spectral notches from each HRTF measurement (per ear, per angle). We use monaural magnitude spectra to ensure ITDs are excluded. Approximate notch frequencies are identified using a Linear Predictive Coding (LPC)-based signal processing method, as described in [14] and [15]. These notch frequencies are determined by detecting local minima (the actual notch frequencies), inverting the spectrum and using a peak prominence detection algorithm (e.g *scipy.peak_prominences* [16]), which returns the left and right bounds of each peak. To finetune this process, we can alternatively calculate various notch width frequencies (e.g. *scipy.peak_widths* [16]), which calculates the notch width, allowing for more accurate definition of the start and end of each notch. After identifying the notch regions, we remove the corresponding frequency bins (by setting them to NaN) and reinterpolate the data over these samples using the *pchip* interpolation algorithm [17]. We refer to the output of the notch removal process as an HRTF′.

2. **ASTF calculation**: We compute the average HRTF′ at each spatial direction across all subjects, yielding the Average Spatial Transfer Function (ASTF). It is important here to distinguish ASTF from Diffuse Field Equalisation. While Diffuse Field Equalisation calculates the average filter for a single subject across all angles, the ASTF is the average filter for a single angle, across all subjects. As a result, the ASTF effectively captures ILDs and average head-related features such as head shadowing.

3. **DSTF calculation**: The Delta Spectral Transfer Function (DSTF) for each subject is defined as the difference between their HRTF′ and the ASTF, as shown in Equation 1, where ϕ denotes the azimuth angle.

$$DSTF_\phi = HRTF'_\phi - ASTF_\phi \tag{1}$$

Table 1: Glossary of key terms used in the feature extraction process.

Term	Definition
HRTF′	HRTF after notch removal and interpolation.
ASTF	Average notchless HRTF across subjects at each azimuth.
DSTF	Deviation from ASTF representing low-order spectral characteristics per subject at each azimuth.
PCHIP	Interpolation algorithm used to fill in spectral notches after removal [17].

For input into our autoencoder model, we treat the DSTF as a series of 2D matrices, where each matrix corresponds to the DSTF at a single azimuth ϕ. Each matrix represents the frequency-domain magnitude data of the DSTF across elevation and frequency. Figures 1 and 4 show example DSTFs at 0° azimuth for subjects P0001 and P0005. Each matrix has dimensions $[257 \times 17]$, corresponding to 257 frequency bins (0 to 24 kHz) and 17 elevation angles (from -60° to 60°). For training, all DSTF features are normalized in $[-1, 1]$ dB.

4.3 Model architecture and training details

The proposed model is a conditional autoencoder composed of four main components: a convolutional Encoder, a convolutional Decoder, a Fully Connected (FC) Subject Identity Classifier, and an FC Azimuth Discriminator. As shown in Figure 2, the architecture is designed to reconstruct DSTF features while disentangling subject identity (unique perceptual attributes of an HRTF, mainly sound coloration) from spatial direction (azimuth). The model is trained using a multi-task

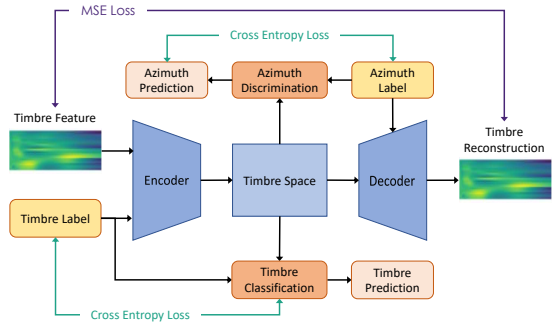


Fig. 2: Autoencoder architecture. The model consists of an Encoder, a Decoder, a Subject Identity Classifier and an Azimuth Discrimination module.

objective combining Mean Squared Error (MSE) for reconstruction and Cross-Entropy (CE) losses for classification. The overall objective is expressed as:

$$Loss = L_{AE} + L_{ID} - L_{AD} \tag{2}$$

L_{AE} denotes the MSE between the Decoder's output and the Encoder's input, both in dB-scaled spectral feature space. L_{TC} is the CE between the predicted and ground truth subject identity, and L_{AD} is the CE between the predicted and ground truth azimuth. During training, the goal is to minimize L_{AE} and L_{ID}, while maximizing L_{AD}, encouraging the Encoder to discard azimuth-specific cues and retain only low-order spectral information in the latent space, associated with perceptual attritrubes, maily sound coloration.

The Encoder is an 8-layer convolutional neural network that compresses input features of size $[257 \times 17](frequency \times elevation)$ into a 64-dimensional latent vector. Except for the first and last layers, each convolution is followed by a LeakyReLU activation with a slope of 0.2 and Batch Normalization.

The Decoder mirrors this structure using transposed convolutions to reconstruct DSTF features from the latent space, again applying LeakyReLU and Batch Normalization on all layers except the final two.

The Subject Identity Classifier is a 4-layer FC network. It takes the 64-dimensional latent vector as input and outputs a 400-dimensional one-hot subject prediction vector. The two hidden layers contain 128 neurons each. The Azimuth Discriminator has a similar structure, with the final output layer containing 48 neurons corresponding to the number of azimuth directions.

During training, input DSTF features are passed to the Encoder, which produces a latent representation. This latent embedding, along with azimuth labels, is fed to the Decoder to reconstruct the original feature. The MSE loss, L_{AE}, is calculated between the input features and the Decoder's reconstructions. The Identity Classifier attempts to predict the correct subject identity, while the Azimuth Discriminator tries to infer the azimuth. Adversarial training encourages the encoder to encode only information relevant to unique perceptual attributes such as sound coloration and suppress localisation cues.

The use of azimuth discrimination is specific to the structure of the features used, which encode data for a single azimuth across multiple elevations. Alternatively, if the dataset encoded a single elevation across multiple azimuths, an elevation discrimination module could be used analogously, by maximizing the classification error on elevation while minimizing reconstruction and subject identity classification losses.

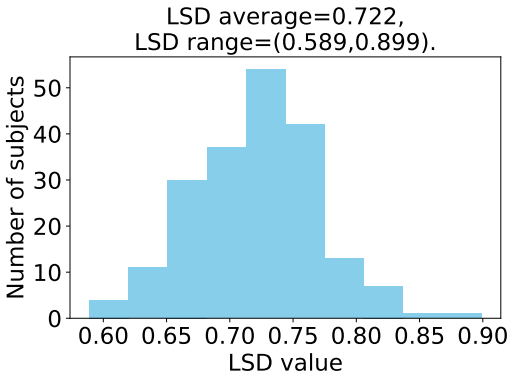
The model is trained for 2,000 epochs and evaluation is performed using the checkpoint of the last epoch. The batch size is set to 12 and the learning rate is set to 0.00005.

5 Results and Discussion

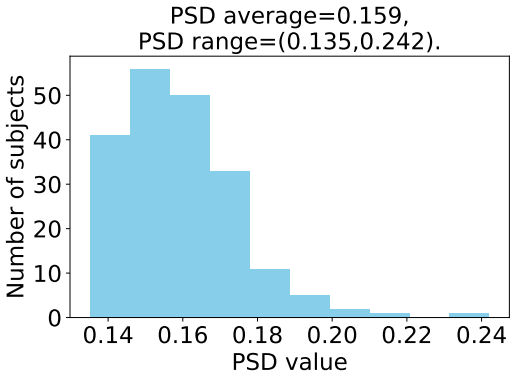
5.1 DSTF reconstruction and synthesis

To evaluate the model's performance, we first assess its reconstruction capability. Prior to evaluation, we smooth each FFT magnitude spectrum in the model's output using a moving average filter, where the window size increases linearly with respect to frequency and the window samples are weighted by a Hann function. Figure 1 shows a typical reconstruction of the left ear of subject P0001 at azimuth 0 (across all elevations). For each subject, we compute the Log Spectral Distance (LSD) [18, 19] and Perceptual Spectral Distance (PSD) [20] between the input and the reconstructed features at all spatial points.

LSD is a commonly used metric for evaluating spectral distortion. It measures the difference between two spectra on a logarithmic scale, reflecting perceptual similarity. As defined in [18], LSD is computed as the



(a) LSD histogram.



(b) PSD histogram.

Fig. 3: Evaluation of the model's reconstruction capability using the Log Spectral Distance (LSD) and Perceptual Spectral Distance (PSD) metrics.

root mean square (RMS) of the frame-wise differences between the log-magnitude spectra of a reference and a test signal, as expressed in Equation 3.

$$LSD(X,\hat{X}) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left[20\log_{10}\left(\frac{|X(k)|}{|\hat{X}(k)|}\right)\right]^2} \quad (3)$$

In the equation above, K is the total number of frequency bins, X(k) is the magnitude of the reference spectra at frequency bin k and $\hat{X}(k)$ is the magnitude of the test spectra at frequency bin k.

PSD is a metric designed to quantify spectral differences between binaural audio signals by mimicking human auditory perception. It measures the perceptual dissimilarity between the reference and test signals

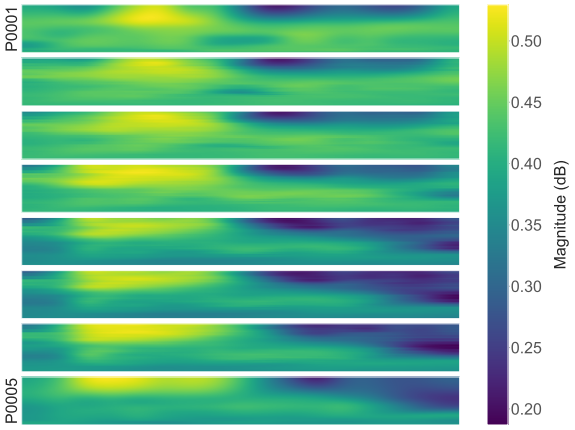


Fig. 4: DSTF feature interpolation between subjects P0001 and P0005 using the proposed model. The top image shows the DSTF feature of subject P0001, and the bottom image shows that of P0005. Intermediate features are generated by the model using linear interpolation between latent embeddings of the two subjects. The x-axis represents frequency (0 Hz to 24 kHz), and the y-axis represents elevation (-60 up to 60°).

by comparing their spectral envelopes after processing through an auditory-inspired model that accounts for frequency selectivity and binaural hearing effect [20]. This makes PSD sensitive to perceptually relevant spectral changes rather than simple numerical differences.

The LSD and PSD histograms for the entire database are shown in Figure 3. For 95.5% of subjects, the LSD values are lower than 0.775, and the PSD values are below 0.178.

In terms of synthesising novel DSTFs, Figure 4 illustrates an indicative example where such features are synthesised through linear interpolation between subjects P0001 and P0005. By leveraging the learned latent space, a range of novel DSTFs can be generated that combine DSTF features from multiple subjects within the database, showcasing the model's ability to generate new location-agnostic perceptual characteristics beyond the original data.

5.2 Latent space

The latent space is designed to capture the perceptual characteristics of each subject's HRTF while discarding azimuthal cues. To visualize the structure of this space,

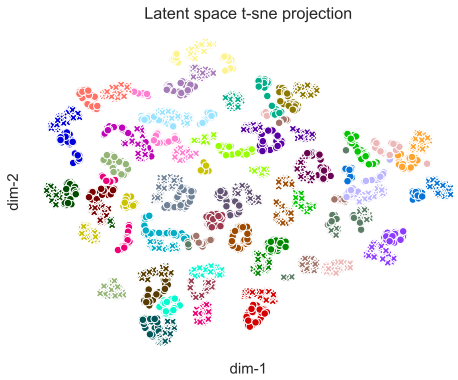


Fig. 5: Latent space visualisation using T-SNE dimensionality reduction. A subset of the subjects included in the database is depicted. Each point represents the latent vector of an DSTF feature at a specific azimuth. Different colours represent different subjects. Circles denote that a feature belongs to the left ear of a subject and crosses denote that a feature belongs to the right ear of a subject.

we apply T-SNE [21] to reduce the 64-dimensional latent vectors to two dimensions, as shown in Figure 5. Each point corresponds to the latent embedding of a DSTF feature measured at a specific azimuth. Colours indicate subject identity, while marker shape distinguishes ear: circles represent left-ear features, and crosses represent right-ear features.

The visualization reveals that features from the same subject are grouped into well-defined clusters, demonstrating that the model effectively encodes unique subject-specific perceptual characteristics while suppressing azimuth-related information. Interestingly, although left and right ears are treated as separate classes during training, the latent vectors for both ears of a subject tend to be clustered close together. This indicates that the unique perceptual characteristics of a subject's HRTF are similar for both ears, and the model is able to capture these relationships, even though it is not explicitly guided to do so. These observations support the claim that the encoder produces a latent representation that preserves subject identity while discarding spatial direction cues, a key goal of our adversarial training strategy.

5.3 Azimuth and DSTF disentanglement

Since DSTF features are calculated at specific azimuth locations, they inherently encode spatial information. To achieve independent control over localisation when synthesizing novel DSTF filters from the autoencoder's latent space, the model must disentangle azimuth qualities from the given features. The Encoder module must discard any correlation with azimuthal information, while the Decoder must rely solely on azimuth conditioning to synthesize a DSTF filter at the desired azimuth. This is accomplished by training the model in a multi-task learning framework, as described in Equation 2, and using adversarial learning for azimuth discrimination, as detailed in Section 4.3.

As outlined in Section 4.1, our dataset consists of 48 unique azimuths, and thus a complete DSTF is composed of 48 features (or DSTF filters). Given that our model can successfully disentangle DSTF and azimuth, a single vector sampled from any point in the latent space is sufficient to generate features at all 48 azimuths around the head. This can be achieved by feeding a latent vector to the Decoder along with the desired azimuth location and repeating this process for all 48 azimuths while keeping the latent vector fixed.

Although reconstructing an existing DSTF feature requires a specific latent vector, synthesising novel DSTFs does not impose this restriction. While azimuthal information is discarded from the latent space, some elevation-specific information remains, as it is encoded within the DSTF features themselves. The further disentanglement of elevation is left for future work.

5.4 Neural compression

The learned DSTF space can be utilised as a compressed representation of a predefined selection of DSTF profiles. An example database of 100 profiles, 1000 measurements per profile and 256-sample stereo filters per measurement with 32-bit floating point precision would require approximately 204.8 MB of disk space. The components required for real-time use of our model are the Decoder module (2.60 MB) and the latent space encodings (2.46 MB). In this example we offer a x40 compression ratio.

6 Conclusion

This work addresses the sensitive and often-overlooked issue of sound coloration introduced by 3D audio processing, advocating for greater emphasis on perceptual modeling in spatial audio, beyond localization accuracy alone. While the boundary between localisation and spectral colouration is inherently blurred, our findings suggest that HRTFs contain rich, underexplored features unrelated to spatial localisation that remain an open area for research.

We introduce a novel signal processing method for extracting Delta Spectral Transfer Function (DSTF) features from HRTFs, and present a conditional autoencoder that compresses and reconstructs these features as well as synthesising novel ones. Through multitask and adversarial training, our model successfully disentangles azimuth information from the perceptual characteristics encoded in DSTFs, enabling independent control of spatial localisation during synthesis. This disentanglement leads to a more structured latent space, where DSTF features from the same subject are consistently clustered, demonstrating the model's ability to preserve subject-specific perceptual characteristics. These results prove the potential of DSTFs as a perceptually meaningful representation and offer a step towards more customisable and perceptually optimised 3D audio experiences.

7 Acknowledgment

This research was conducted during an internship at PlayStation London and is supported by the EPSRC UKRI Centre for Doctoral Training in Artificial Intelligence and Music (Grant no. EP/S022694/1).

References

[1] Armstrong, C., Thresh, L., Murphy, D., and Kearney, G., "A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database," *Applied Sciences*, 8(11), 2018, ISSN 2076-3417, doi:10.3390/app8112029.

[2] Armstrong, C., Chadwick, A., Thresh, L., Murphy, D., and Kearney, G., "Simultaneous HRTF Measurement of Multiple Source Configurations Utilizing Semi-Permanent Structural Mounts," in *Audio Engineering Society Convention 143*, 2017.

[3] Engel, I., Daugintis, R., Vicente, T., Hogg, A. O. T., Pauwels, J., Tournier, A. J., and Picinali, L., "The SONICOM HRTF Dataset," *J. Audio Eng. Soc*, 71(5), pp. 241–253, 2023.

[4] Simon, L. S. R., Zacharov, N., and Katz, B. F. G., "Perceptual attributes for the comparison of head-related transfer functions," *The Journal of the Acoustical Society of America*, 140(5), pp. 3623–3632, 2016, ISSN 0001-4966, doi:10.1121/1.4966115.

[5] Lindau, A., "Spatial Audio Quality Inventory (SAQI). Test Manual. v1.2," *Acta Acustica united with Acustica*, 2014, doi:https://doi.org/10.3813/AAA.918778.

[6] Marentakis, G. and Saitis, C., "Timbre in Binaural Listening: A Comparison of Timbre Descriptors in Anechoic and HRTF Filtered Orchestral Sounds," in *Forum Acusticum*, pp. 1047–1052, Lyon, France, 2020, doi:10.48465/fa.2020.0992.

[7] Marggraf-Turley, N., Lovedee-Turner, M., and De Sena, E., "HRTF Recommendation Based on the Predicted Binaural Colouration Model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1106–1110, 2024, doi:10.1109/ICASSP48485.2024.10448092.

[8] McKenzie, T., Armstrong, C., Ward, L., Murphy, D. T., and Kearney, G., "Predicting the Colouration between Binaural Signals," *Applied Sciences*, 12(5), 2022, ISSN 2076-3417, doi:10.3390/app12052441.

[9] Esling, P., Chemla-Romeu-Santos, A., and Bitton, A., "Generative timbre spaces with variational audio synthesis," in *Proceeedings of the 21st International Conference on Digital Audio Effects (DAFx-18)*, 2018.

[10] Colonel, J. and Keene, S., "Conditioning Autoencoder Latent Spaces for Real-Time Timbre Interpolation and Synthesis," in *Proceeedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2020, doi:10.1109/IJCNN48605.2020.9207666.

[11] Puche, A. V. and Lee, S., "Caesynth: Real-Time Timbre Interpolation and Pitch Control with Conditional Autoencoders," *2021 IEEE 31st Interna-*

tional Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6, 2021.

[12] Hawley, S. H. and Steinmetz, C. J., "Leveraging Neural Representations for Audio Manipulation," in *Audio Engineering Society Europe Convention*, 2023.

[13] European Broadcasting Union, "Loudness normalisation and permitted maximum level of audio signals," EBU Recommendation R 128, 2023, version published 21 Nov 2023.

[14] Raykar, V. C., Duraiswami, R., and Yegnanarayana, B., "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses," *The Journal of the Acoustical Society of America*, 118(1), pp. 364–374, 2005, ISSN 0001-4966, doi:10.1121/1.1923368.

[15] Spagnol, S. and Avanzini, F., "Frequency Estimation Of The First Pinna Notch In Head-Related Transfer Functions With A Linear Anthropometric Model," in *Proceeedings of the 18th International Conference on Digital Audio Effects (DAFx-15)*, 2015.

[16] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, 17, pp. 261–272, 2020, doi: 10.1038/s41592-019-0686-2.

[17] Fritsch, F. N. and Butland, J., "A Method for Constructing Local Monotone Piecewise Cubic Interpolants," *SIAM J. Sci. Stat. Comput.*, 5(2), p. 300–304, 1984, ISSN 0196-5204, doi:10.1137/0905021.

[18] Gray, A. H. and Markel, J., "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24, pp. 380–391, 1976.

[19] Lee, G. W. and Kim, H. K., "Personalized HRTF Modeling Based on Deep Neural Network Using Anthropometric Measurements and Images of the Ear," *Applied Sciences*, 8(11), 2018, ISSN 2076-3417, doi:10.3390/app8112180.

[20] Armstrong, C., McKenzie, T., Murphy, D., and Kearney, G., "A Perceptual Spectral Difference Model for Binaural Signals," in *Audio Engineering Society Convention 145*, 2018.

[21] van der Maaten, L. and Hinton, G., "Visualizing High-Dimensional Data Using t-SNE," *Journal of Machine Learning Research*, 9(nov), pp. 2579–2605, 2008, ISSN 1532-4435, pagination: 27.