PERCEPTUAL OPTIMISATION OF LOUDSPEAKER-BASED REPRODUCTION

Antoine Souchaud[*], *Pedro Lladó*[*], *Rapolas Daugintis*[*], *Annika Neidhardt*[†], *Zoran Cvetković*[‡], *Enzo De Sena*[*]

[*]Institute of Sound Recording, University of Surrey, Guildford, United Kingdom
[†]Fakultät Medien, Hochschule-Mittweida, Mittweida, Germany
[‡]Department of Engineering, King's College, London, United Kingdom

ABSTRACT

This paper proposes POLAR, a framework for the optimisation of loudspeaker signals using end-to-end differentiable perceptual loss functions. The framework optimises multiple perceptual attributes across multiple listeners, offering a versatile method for a range of problems. This versatility stems from the ability to customise the number of loudspeakers, listeners, and the weighting applied to different perceptual attributes. Here, we apply the method to four problems: (a) source panning for a single listener in stereo reproduction, (b) single-listener colouration matching in stereo reproduction, (c) extended sweet spot using stereo pairs beamforming, and (d) multi-attribute perceptually driven panning in stereo reproduction. The first three problems are evaluated against solutions traditionally used for these tasks: solutions of (a) are shown to be similar to those obtained with tangent panning law and vector-base amplitude panning (VBAP), solutions of (b) are shown to be similar to those obtained for cross-talk cancellation, and solutions of (c) are shown to be similar to those obtained in earlier work on directivity pattern optimisation for sweet spot widening. Each of these solutions was previously obtained using fundamentally different methodologies, demonstrating the flexibility and broad applicability of the proposed framework.

1. INTRODUCTION

Spatial audio reproduction comprises a wide range of techniques and methodologies that aim to recreate sound fields to maintain or enhance the listener's perception of immersion. Applications span VR/AR, gaming, cinema, teleconferencing, and architectural acoustics, all demanding accurate and predictable spatial sound reproduction over loudspeakers. While some spatial audio techniques, such as higher order ambisonics (HOA) [1, 2] or wave-field synthesis (WFS) [3] can recreate a sound field with great accuracy, they are often difficult to use in practice due to practical constraints [4], including a limited number of loudspeakers, the assumption of a perfect environment, listener position, or computational cost.

Trying to circumvent the practical limitations, recent progress in spatial audio research has focused on optimising towards perceptual attributes [5, 6]. Such perceptually-motivated optimisation strategies leverage current theories of psychoacoustics to minimise the perceptual degradation caused by the inevitable physical errors [7, 8, 9, 10]. Historically, this type of optimisation was largely restricted to simple perceptual metrics. Ambisonic design often relied on criteria such as Gerzon's velocity and energy vectors, considered proxies for perceived localisation at low and high frequencies, respectively [11], and later the rE vector [12]. Although widely used, these metrics do not capture all aspects of auditory perception; for example they do not account for listener orientation, even though localisation depends on source direction relative to the listener's head [13]. More recently, perceptually motivated optimisation of loudspeaker signals has been formulated using monaural auditory models [8]. This paper extends this line of research to more complex models, including binaural ones.

We present POLAR, a differentiable[1] framework for Perceptual Optimisation of Loudspeaker-based Audio Reproduction. POLAR aims to optimise loudspeaker signals to yield a desired auditory experience for one or more listeners in different positions. This is done using perceptual loss functions derived from differentiable auditory models [14, 15], combined with loss functions that induce certain properties of the desired loudspeaker signals. We demonstrate the versatility of the framework by applying it to three use cases with established solutions in the literature: amplitude panning filters for stereo [16, 17, 18], cross-talk cancellation filters [19] and stereo loudspeaker beamforming for a wider sweet spot [20]. Finally, we show the ability of the framework-derived solutions to trade-off or balance between multiple attributes.

2. METHOD

Fig. 1 illustrates the proposed framework, which consists of three end-to-end differentiable stages. The first is the rendering stage (Section 2.1), generating the loudspeaker signals, which are then fed to the propagation stage (Section 2.2), producing the ear signals. These signals are subsequently passed to the loss stage (Section 2.3), in which a weighted loss is built by combining differentiable perceptual loss functions. The framework and the perceptual loss functions are implemented using PyTorch[2].

2.1. Rendering

We formulate the problem in the time domain and we seek to find optimal loudspeaker signals, $y_l(n)$, where n is the discrete-time variable and $l = 1, ..., N_{\mathrm{L}}$ is the loudspeaker index. For simplicity, we focus on the case of reproducing a single virtual sound source and assume that multiple virtual sources can be rendered via superposition. We consider the optimisation independent of the input signal, and thus, instead of optimising the loudspeaker signals $y_l(n)$, we aim to find optimal loudspeaker Finite Impulse Response (FIR) filters, $h_l(n)$ such that $y_l(n) = s(n) * h_l(n)$,

[1]Here, "differentiable" refers to the use of automatic differentiation for optimisation through the full processing chain, rather than to the use of neural networks.

[2]Github: https://github.com/IoSR-Surrey/POLAR.

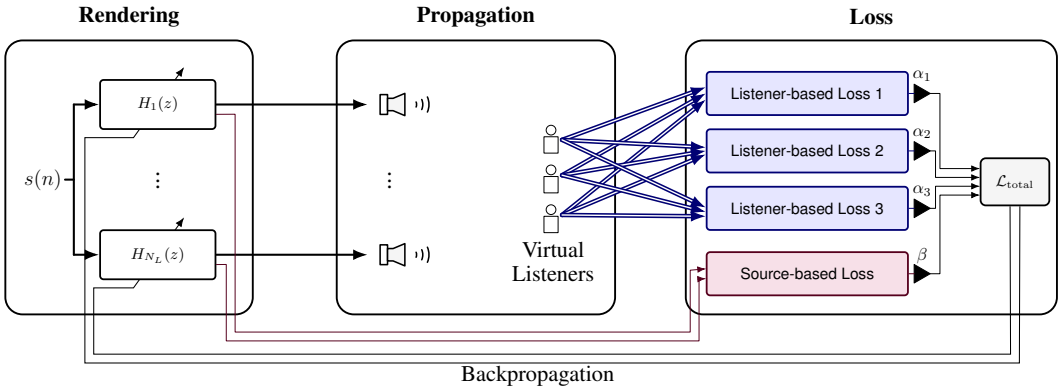


Figure 1: Illustration of the optimisation framework. All stages are end-to-end differentiable.

where $s(n)$ is the input signal and $*$ denotes convolution. The optimisation problem thus becomes:

$$\mathbf{h}_{\text{opt}} = \arg\min_{\mathbf{h}} \mathcal{L}_{\text{total}}(\mathbf{h}), \tag{1}$$

where $\mathcal{L}_{\text{total}}$ is an appropriate (differentiable) loss function defined in Sec. 2.3, $\mathbf{h} = [\mathbf{h}_1, \ldots, \mathbf{h}_{N_\text{L}}]$ is an $N_{\text{taps}} \times N_\text{L}$ matrix formed of the column vectors, $\mathbf{h}_l = [h_l(1), \ldots, h_l(N_{\text{taps}})]^T$, and N_{taps} is the filter length. Filters are optimised using gradient descent and automatic differentiation, which is efficient and scales very favourably with the number of parameters, i.e., $N_\text{L} N_{\text{taps}}$ filter coefficients. All convolutions in the pipeline are causal and use no zero-padding which preserves the input tensor shape.

2.2. Propagation

Consider N_VL Virtual Listeners (VLs), at given positions and with given orientations. The signal reaching the left (L) and the right (R) ears of the k-th listener is

$$x_k^{\text{L/R}}(n) = \sum_{l=1}^{N_\text{L}} \left(y_l * g_{l,k}^{\text{L/R}} \right)(n), \tag{2}$$

where $g_{l,k}^{\text{L/R}}(n)$ is the impulse response between the l-th loudspeaker and the ears of the k-th VL, including the loudspeaker response, the propagation path between loudspeaker and listener (e.g., delay, attenuation, and room reflections), and the listener's Head-Related Impulse Response (HRIR).

2.3. Loss

We consider two types of losses, which we refer to as "listener-based losses" and "source-based losses". Listener-based losses take the ear signals $x_k^{\text{L/R}}(n)$ as input and output a scalar value representing a deviation of specific perceptual attributes (e.g. localisation or colouration) from a specific target. Source-based losses take the loudspeaker filters, $\mathbf{h}$, as input, and output a scalar value representing a penalty associated with a desired property of the loudspeaker filters. The overall loss function is the combination of the listener-based and loudspeaker-based loss functions:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{listener}} + \mathcal{L}_{\text{lspk}}, \tag{3}$$

where $\mathcal{L}_{\text{listener}}$ and $\mathcal{L}_{\text{lspk}}$ are the listener- and loudspeaker-based loss, respectively, defined in Sections 2.3.1 and 2.3.2, and where the dependency on $\mathbf{h}$ is omitted.

2.3.1. Listener-based losses

Consider a perceptual loss function that translates the ear signals of the k-th VL, $x_k^{\text{L/R}}$, into an estimated value $\hat{\Phi}_{A,k}$ of a perceptual attribute A (e.g. estimated lateral localisation, apparent source width or colouration). Let $\Phi_{A,k}$ denote the target value associated with this attribute. We define the corresponding loss as

$$\mathcal{L}_A = \frac{1}{N_\text{VL}} \sum_{k=1}^{N_\text{VL}} f_A(\hat{\Phi}_{A,k}, \Phi_{A,k}), \tag{4}$$

where f_A is the loss function associated with the attribute A (e.g., ℓ_2 or ℓ_1 norm). Losses associated with different attributes can be combined. For controlling the weighting between the different loss functions, it is convenient to normalise the losses between 0 and 1 using what we call the Expected Maximum Loss (EML), i.e. the maximum loss that is expected from the given setup, problem and perceptual attribute. For example, in a stereo loudspeaker configuration with horizontal-plane localisation, the EML corresponds to the loss associated with an angular error equal to the span of the stereo setup. Since this is only an expected maximum, the normalised loss may exceed 1 in some cases.

The combined loss function $\mathcal{L}_{\text{listener}}$ is defined as

$$\mathcal{L}_{\text{listener}} = \sum_A \alpha_A \frac{\mathcal{L}_A}{\text{EML}_A}, \tag{5}$$

where α_A is the weighting factor for the attribute A. Specific listener-based losses used in the validation of the framework are detailed in Sec. 2.4.

2.3.2. Loudspeaker-based losses

Loudspeaker-based losses are introduced to control certain properties of the solutions, and can serve as regularisation. The regularisation loss $\mathcal{L}_\text{B}$ is computed as

$$\mathcal{L}_\text{B}(\mathbf{h}) = \frac{1}{N_\text{L}} \sum_{l=1}^{N_\text{L}} f_B(\mathbf{h}_l) \tag{6}$$

where $f_B(\cdot)$ is the loss applied to the l-th filter (e.g., ℓ_2 or ℓ_1 norm). The regularisation losses can be combined through the same scheme as in Eq. 5. We define $\mathcal{L}_{\text{lspk}}$ as

$$\mathcal{L}_{\text{lspk}} = \sum_B \beta_B \frac{\mathcal{L}_B}{\text{EML}_B}, \tag{7}$$

having once again omitted the dependency on $\mathbf{h}$. The loudspeaker-based loss used in the evaluation is detailed in Sec. 2.4.

2.4. Examples of losses

We introduce here example loss functions, all of which are end-to-end differentiable.

Lateral localisation loss

The perceptually-motivated lateral localisation loss is built upon a differentiable perceptual localisation model, introduced by Lladó et al. [21, 22], that estimates the perceived lateral angle of the incoming sound source based on binaural signals at the ears. The model computes frequency-dependent interaural time and level differences (ITDs and ILDs) as the delay that maximises the cross-correlation and the energy ratio, respectively, between the left and right ears. The localisation estimates are then determined from minimising the distance between the target binaural sound and a set of templates, which consist of the interaural cues computed for different directions in a given set of HRIRs [22]. The relative contribution of the ITD and ILD cues is controlled by frequency-dependent weights, which reflect the perceptual importance of these cues across different frequency regions. This model was selected due to the availability of a differentiable implementation [21], its interpretability, and its close agreement with perceptual data ($R^2 = 0.88$) [22], including in off-centre listening positions–a key aspect in further validation steps of this study. The loss is defined as

$$\mathcal{L}_{\mathrm{loc}} = \frac{1}{N_{\mathrm{VL}}} \sum_{k=1}^{N_{\mathrm{VL}}} (\hat{\theta}_k \, - \, \theta_k)^2, \tag{8}$$

where $\hat{\theta}_k$ and θ_k are predicted and target lateral angles within $(-90°, 90°)$ for the k-th VL, respectively. The $\mathrm{EML}_{\mathrm{loc}}$ depends on the use case, e.g. for a stereo setup with a spanning angle of 60°, this corresponds to an $\mathrm{EML}_{\mathrm{loc}} = 60^2$ (i.e., the desired localisation is towards one of the loudspeakers and the estimated localisation points at the other loudspeaker).

Localisation uncertainty loss

A localisation uncertainty loss is introduced based on the model by De Sena et al. [23]. This attribute quantifies how easily a listener can determine where a sound source is located. It is related to apparent source width and is therefore relevant in assessing the quality of loudspeaker reproduction methods [23]. The model shares most of the processing stages with the localisation model in [22]. To compute the uncertainty, it relies on the spread of the estimates for the localisation cues. The model presents the same advantages as the localisation model, i.e. an available differentiable implementation, interpretability and generalisability to off-centre listening. The associated loss is defined as

$$\mathcal{L}_{\mathrm{unc}} = \frac{1}{N_{\mathrm{VL}}} \sum_{k=1}^{N_{\mathrm{VL}}} \left(\hat{\mathrm{U}}_k \, - \, \mathrm{U}_k \right)^2, \tag{9}$$

where $\hat{\mathrm{U}}_k$ and U_k the predicted and target localisation uncertainty for the k-th VL, respectively. The estimates from the localisation uncertainty model are already within 0 and 1, thus we set $\mathrm{EML}_{\mathrm{unc}} = 1$. For all the simulations in this paper, we set $\mathrm{U}_k = 0$, i.e., optimising for minimal uncertainty.

Colouration loss

The colouration loss is constructed via a modification of the log-spectral distortion (LSD) [24]:

$$\mathrm{LSD}_{\mathrm{k}}^{\mathrm{L/R}} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} w_m \left(20\log_{10} \frac{|X_k^{\mathrm{L/R}}[m]|}{|X_{k,\mathrm{target}}^{\mathrm{L/R}}[m]|} \right)^2}, \tag{10}$$

where $X_k^{\mathrm{L/R}}[m]$ and $X_{k,\mathrm{target}}^{\mathrm{L/R}}[m]$ are the discrete Fourier Transform (FT) coefficients of the ear signals and target ear signals, respectively, for the k-th VL, $m \in 1, ..., M$ is the frequency bin and $w_m = 24.7(4.37 \cdot 10^{-3} \cdot f_m + 1)$ [25] are Equivalent Rectangular Bandwidth (ERB)-based weights that compensate for the linear spacing of Fast Fourier Transform (FFT) bins to better reflect the cochlea's logarithmic frequency resolution, with f_m the frequency of the m-th bin in Hertz. The loss is defined as

$$\mathcal{L}_{\mathrm{col}} = \frac{1}{N_{\mathrm{VL}}} \sum_{k=1}^{N_{\mathrm{VL}}} \frac{1}{2} (\mathrm{LSD}_{\mathrm{k}}^{\mathrm{L}} \, + \, \mathrm{LSD}_{\mathrm{k}}^{\mathrm{R}}), \tag{11}$$

where we assumed the target colouration to be 0. $\mathrm{EML}_{\mathrm{col}}$ is chosen as the loss between the reference signal and the same signal low-pass filtered with a cutoff at 7 kHz (adopted from the mid-quality anchor described in [26]). For white noise or impulses, this results in an $\mathrm{EML}_{\mathrm{col}} = 4.87\,\mathrm{dB}$.

Interaural phase difference (IPD) loss

The IPDs are defined as $\mathrm{IPD}_k[m] = \angle\big(X_k^{\mathrm{L}}[m]/X_k^{\mathrm{R}}[m]\big)$ and $\mathrm{IPD}_{k,\mathrm{ref}}[m] = \angle\big(X_{k,\mathrm{ref}}^{\mathrm{L}}[m]/X_{k,\mathrm{ref}}^{\mathrm{R}}[m]\big)$ respectively, where $\angle(\cdot)$ denotes the principal phase in the interval $[-\pi, \pi)$. The associated loss, a commonly used metric for spatial perception [27, 28], is defined as

$$\mathcal{L}_{\mathrm{IPD}} = \frac{1}{N_{\mathrm{VL}} M} \sum_{k=1}^{N_{\mathrm{VL}}} \sum_{m=1}^{M} \left(\mathrm{IPD}_k[m] - \mathrm{IPD}_{k,\mathrm{ref}}[m] \right)^2, \tag{12}$$

where we assumed the target IPD loss to be 0. The EML for this loss is $\mathrm{EML}_{\mathrm{IPD}} = \pi$ rad. The same ERB-weighting scheme as in Eq. 10 is applied here.

Signal spread loss

As a loudspeaker-based loss, we use a measure of the temporal spread of the energy of the filters [29]:

$$\mathcal{L}_{\mathrm{spread}} = \frac{1}{N_{\mathrm{L}}} \sum_{l=1}^{N_{\mathrm{L}}} \frac{\sum_n (n - \mu_l)^2 \, |h_l(n)|^2}{\sum_n |h_l(n)|^2}, \tag{13}$$

where $\mu_l = \frac{\sum_n n |h_l(n)|^2}{\sum_n |h_l(n)|^2}$ is the centroid of the energy of the l-th filter. The EML for this loss is assumed to be the one obtained with white noise, its value depends on the number of taps in the filters.

3. EVALUATION

The proposed method is validated against different problems with known solutions. We hypothesise that by replicating the respective setup of those problems and choosing the right set of losses, the framework would derive optimal filters that are comparable to

accepted existing solutions from the literature. Then, the flexibility of the framework is analysed by combining multiple perceptual losses, showing its ability to trade off several perceptual attributes.

The simulation setup is as follows. The sampling frequency is set to 48 kHz. Sound propagation is simulated in the free field (no room reflections) with ideal, omnidirectional, point-source loudspeakers. The propagation to the VLs is modelled as a delay introduced by the distance between the loudspeaker and the listener, r, and attenuation of the signal amplitude proportional to $1/r$. The delay is applied in the time domain, assuming the speed of sound of 343 m/s, without fractional delays (maximal quantisation error = 11 μs at a sampling rate of 48 kHz). Once delayed and attenuated, the signals are filtered using the appropriate HRIR from the KU 100 dummy head from the SADIE II database [30]. The HRIRs are made symmetric such that the left and right ear responses at an angle θ are equal to the right and left responses, respectively, at the opposite angle $-\theta$. In all simulations below, we use the Adaptive Moment Estimation (ADAM) optimiser. We initialise the loudspeaker filters with an impulse in the centre of the window. Most simulations presented here take a few seconds to complete (Macbook Pro M3 Max; CPU-based processing).

3.1. Validation against amplitude panning

Amplitude panning is the most widely used solution for sound source panning [16, 17, 18]. Different frequency-independent gains are applied to the two loudspeakers, whose energy ratio defines the Inter-Channel Level Difference (ICLD). We hypothesise that the framework would derive filters with energy ratios of the left and right loudspeakers that resemble amplitude panning when they are optimised for localisation of a single virtual listener in the centre of the sweet spot.

3.1.1. Simulation setup

We simulate a stereo setup, $N_{\mathrm{L}} = 2$, with a spanning angle of 60° at 2 m from a single VL oriented towards the direction between the two loudspeakers. The input signal, $s(n)$, is a white noise burst $N_{\mathrm{samples}} = 480$ samples long. We set the FIR filter length to $N_{\mathrm{taps}} = 21$, learning rate $l_r = 1 \cdot 10^{-2}$, and the number of gradient-descent steps to 600.

As we are aiming to control the perceived location of a sound source, we utilise the perceptual lateral localisation loss introduced in Eq.8. Furthermore, to regularise the optimisation problem, the spread of the filter (Eq. 13) is combined with the localisation loss. We set $\alpha_{\mathrm{loc}} = 1 \cdot 10^{0}$ and $\beta_{\mathrm{spread}} = 1 \cdot 10^{1}$.

3.1.2. Results and Discussion

Since the framework estimates FIR filters, in this setup, we compare the energy ratio of the filters estimated using the proposed method and the ICLD values prescribed by Vector Base Amplitude Panning (VBAP) [18]. Figure 2 shows the ICLD of the optimised filters. The results show a strong correlation ($R^2 = 0.97$) between the ICLDs obtained using the two methods. Due to the relatively high weighting of the signal spread loss and initialisation scheme, the filters are aligned in phase. These results extend the findings of [21], which presented a model for finding the optimal individual ICLD values (i.e. $N_{\mathrm{taps}} = 1$). Here, finding 21-tap FIR filters with similar energy ratios indicates method robustness beyond producing single output values and could pave the way for frequency-dependent panning curves.

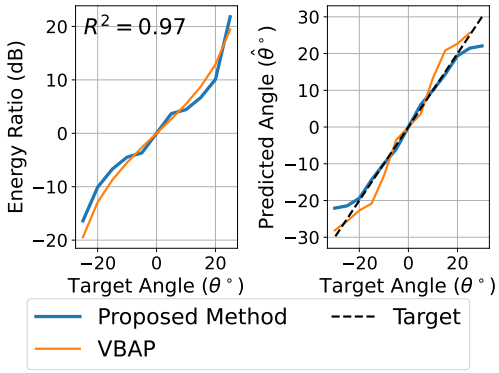


Figure 2: Energy ratios between the left and the right loudspeakers (left) and predicted angle using the localisation model [22] (right) across different panning angles for the proposed method and VBAP.

3.2. Validation against cross-talk cancellation

Cross-Talk Cancellation (CTC) is a method to deliver binaural sound via loudspeakers [31]. In its simplest form, the method involves collecting a 2×2 plant matrix with the transfer functions between left/right loudspeakers and left/right ears, which is then inverted, producing the 2×2 inverse matrix, $\mathbf{H}_{CTC}(z)$. We hypothesise here that the proposed method will estimate similar solutions to that of CTC given the proper optimisation setup. To render a single sound source in a specific direction on the horizontal plane, θ_{target}, the desired ear signal is given by

$$\begin{bmatrix} X^{\mathrm{L}}(z) \\ X^{\mathrm{R}}(z) \end{bmatrix} = \begin{bmatrix} H^{\mathrm{L}}(z;\theta_{\mathrm{target}}) \\ H^{\mathrm{R}}(z;\theta_{\mathrm{target}}) \end{bmatrix} S(z), \tag{14}$$

where $X^{\mathrm{L/R}}(z)$ are the z-transforms of the left/right ear signals, $H^{\mathrm{L/R}}(z;\theta_{\mathrm{target}})$ are the target Head-Related Transfer Functions (HRTFs) of the left/right ears for the azimuth θ_{target}, $S(z)$ is the z-transform of the input signal $s(n)$, and where we omitted the subscript k since there is a single VL. By multiplying the target HRTF vector with $H_{CTC}(z)$, we obtain the equivalent left and right loudspeaker filters:

$$\begin{aligned} \begin{bmatrix} Y^{\mathrm{L}}(z) \\ Y^{\mathrm{R}}(z) \end{bmatrix} &= \mathbf{H}_{CTC}(z) \begin{bmatrix} X^{\mathrm{L}}(z) \\ X^{\mathrm{R}}(z) \end{bmatrix} \\ &= \underbrace{\mathbf{H}_{CTC}(z) \begin{bmatrix} H^{\mathrm{L}}(z;\theta_{\mathrm{target}}) \\ H^{\mathrm{R}}(z;\theta_{\mathrm{target}}) \end{bmatrix}}_{\text{equivalent filters}} S(z). \end{aligned} \tag{15}$$

These equivalent filters derived using the CTC matrix inversion can be compared to the ones estimated using our proposed method.

3.2.1. Simulation setup

The propagation setup follows Section 3.1, with the listener positioned at the centre of the sweet spot. The input signal $s(n)$ is an impulse zero-padded to $N_{\mathrm{samples}} = 480$. The filter length is chosen to match the number of frequency bins in the equivalent CTC filters, which is determined by the HRTF length, yielding $N_{\mathrm{taps}} = 256$. We employ a learning rate of $l_r = 5 \cdot 10^{-4}$ and 5000 gradient-descent steps. We set $\theta_{\mathrm{target}} = 45°$, i.e., a target

angle outside the span of the stereo loudspeakers, and $X_{\text{ref}}^{\text{L/R}}(z) = H^{\text{L/R}}(z;\theta_{\text{target}})S(z)$, the reference binauralised signal.

To optimise the filters, we considered two cases–one with the LSD loss alone, and the second with both the LSD and the IPD losses ($\alpha_{LSD} = 1$, $\alpha_{IPD} = 5$).

3.2.2. *Results and Discussion*

Figure 3 shows the magnitude spectra obtained at the left and the right ear, indicating a near exact match for all methods against the target HRTF, $H^{\text{L/R}}(\theta_{\text{target}})$. Since the proposed method minimises the distance with the target HRTF via the colouration (Log Spectral Distortion (LSD)) loss, this result is expected. Importantly, however, it demonstrates convergence of the proposed method to desired solutions.

The figure also shows the magnitude response of the filters optimised with the proposed method using the two loss models, compared to the equivalent CTC filters in Eq. 15. Here, it can be observed that the optimisation using the colouration loss leads to a dissimilar solution to CTC for both loudspeakers. This is due to the fact that phase relations are left uncontrolled. Once the IPD loss is included, the optimised loudspeaker filters closely match the CTC filters, with an RMSE of only 0.57 dB. If one assumes invertibility of the plant matrix, this result is also expected. Imposing a good match with the magnitude spectra at the two ears and of the phase relation at the ears, leaves the only degree of freedom to be a common phase term for the loudspeakers filters, which is not shown here and is less relevant perceptually.

Nevertheless, this result shows the framework's flexibility, possibly paving the way for hybrids involving CTC-like solutions in the future. Also, the two methods arrive at closely similar solutions through fundamentally different procedures. One is frequency-domain matrix inversion followed by an inverse FFT, which may produce non-causal solutions, while the proposed framework directly optimises finite-length causal FIR filters in the time domain using gradient descent and automatic differentiation.

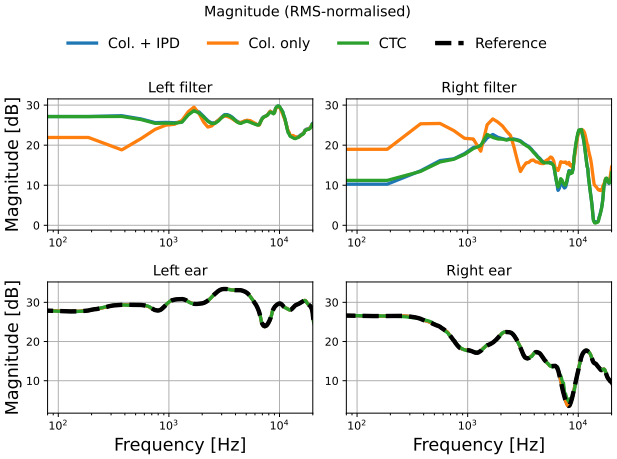


Figure 3: Comparison between the proposed methods and CTC. The first row presents the magnitude spectrum of the loudspeaker filters obtained by the CTC and the proposed method with and without IPD loss. The filters were estimated for a sound coming from $\theta_{\text{target}} = 45°$ (i.e. outside the range of the two loudspeakers at $\pm 30°$). The second row represents the corresponding ear signals (all lines, including the proposed method, are essentially overlapping). The dashed black curve represents the reference log magnitude response of $H^{\text{L/R}}(\theta_{\text{target}})$.

3.3. Validation against stereo pair beamforming

The use cases presented so far are limited to only one listener; however, the versatility of the proposed method allows optimisation of the filters for multiple listeners simultaneously, making it possible to optimise the size of the listening area.

Figure 4a depicts a classic problem in loudspeaker-based reproduction. Two loudspeakers with an omnidirectional directivity pattern are positioned at a distance of 2.5 m and with a base angle of 60° and are playing back identical signals. The arrows represent the perceived direction as predicted by the model of Lladó et al. [22] for virtual listeners oriented such that they are facing the direction of the positive x-axis. While a virtual listener in the centre of the sweet spot perceives a sound in the desired direction (directly ahead, in between the two loudspeakers), as soon as they move to the side, the perceived auditory image collapses in the direction of the closest loudspeaker. This is due to the fact that the closer loudspeaker becomes louder than the other and its signal arrives earlier, dominating spatial perception.

In [20], Rodenas et al. introduced a method to increase the size of the listening area without explicitly tracking the listener/s, through optimisation of loudspeaker directivity pattern. The idea is to design the directivity pattern in such a way that for off-centre listeners, the effect of the signal of the closer loudspeaker arriving earlier is counterbalanced by the other loudspeaker being louder.

They derive an optimal directivity pattern from existing psychoacoustics time/intensity trading curves indicating what is the ICLD necessary to counterbalance an ICTD in opposite direction. Loudspeaker beamforming is then achieved via appropriate design of the FIR filters applied to a pair of close loudspeakers, via mean squared error optimisation. Using the same loudspeaker array setup as in [20], we postulate that the proposed method can estimate similar solutions.

3.3.1. *Simulation setup*

For this simulation, we set up the same loudspeaker configuration as in [20]. Rodenas et al. mention the distance between pairs to be between 1 cm and 10 cm [20], but they provide neither specific distance values they used to obtain their results, nor the type of loudspeaker used for their real-world measurements. We set the distance to 10 cm. The filter length is set to $N_{samples} = 40$, as in [20]. To optimise the beamforming pattern over an area, we position $N_{\text{VL}} = 51$ VLs equally spaced virtual listeners along the line in the figure. The learning rate l_r is set to 0.05 and the optimisation is performed over 200 gradient-descent steps. We employ the localisation loss described in Eq. 8. The target angles θ_k are set such that the direction from the k-th listener is towards a source located at equal distance in between the loudspeaker arrays.

3.3.2. *Results and Discussion*

Fig. 4b presents the directivity pattern resulting from the FIR filter optimisation for each pair of loudspeaker and averaged over octave bands between 250 Hz and 8000 Hz. This is compared to two directivity patterns proposed by [20]: the optimal one derived

directly from the psychoacoustic panning curves, and the one that can practically be obtained through mean squared error optimisation for two loudspeakers in each pair (only the 1 kHz curve is shown here, with the others being similar [20]). The figure shows that the loudspeaker directivity pattern resulting from our proposed method is very similar to the one obtained in [20], and is, in fact, closer in certain directions to the directivity pattern that they define as optimal. Overall, the estimated directivity pattern has a mean absolute distance to the optimal and 1000 Hertz curves of 2.68 and 1.75 dB within the target region, respectively. After optimisation, the mean absolute localisation angle error is of only 6.4°.

This is a noteworthy result. It shows that the proposed method, only relying on a binaural localisation model as loss function, can generate FIR filters that together yield an appropriate loudspeaker beamformer. This is notable because beamforming is not imposed explicitly. It also implies that the localisation model of Llado et al. [22] embeds information similar to the psychoacoustic curves used in [20]. This result could pave the way to broader applications of the proposed framework, for example in-situ soundbar optimisation in reflective environments.

3.4. Analysis of trade-off between attributes

One of the main advantages of the proposed method is that it allows for the optimisation of multiple attributes simultaneously. In this section, we show the result of trading off lateral localisation, localisation uncertainty and colouration.

3.4.1. Simulation setup

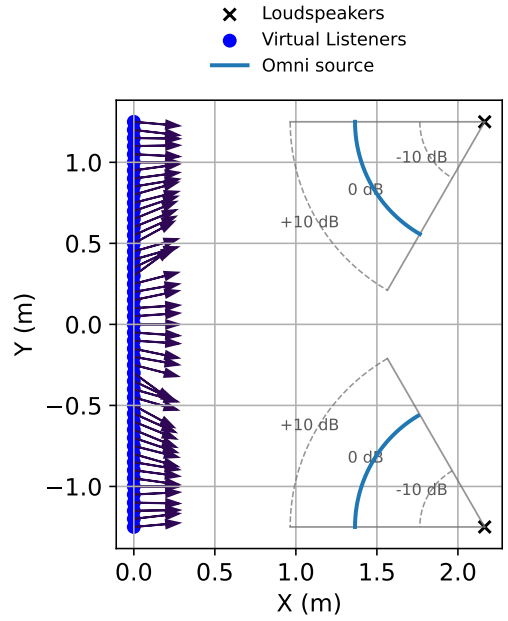
We set up the same loudspeaker configuration as in 3.1. For the VLs we define a 1 m line centred on the sweet spot and oriented parallel to the axis connecting the two loudspeakers. In order to assess how the method generalises to positions that the filters have not been explicitly optimised for, $N_{\mathrm{VL}} = 30$ VLs are distributed on the line with the exception of three 10 cm segments centred around $y = 0, \pm 25$ cm, i.e., the only points shown in Figure 5. The 10 cm gap was chosen as a trade-off between keeping the VLs within a reasonable area, and a large enough gap such that nearby optimised positions do not also effectively optimise the evaluation locations. At low frequencies, however, some overlap is still expected due to the smoother spatial variation of the sound field.

During the optimisation, the VLs are processed in batches of five listeners. The optimisation is performed over 300 gradient-descent steps with the learning rate l_r set to $1 \cdot 10^{-1}$. The target for the localisation and colouration loss is defined by placing a target source in the far-field at $\theta_{\mathrm{target}} = 10°$ to the left of each VL. The HRTF of the same angle serves as the colouration reference.

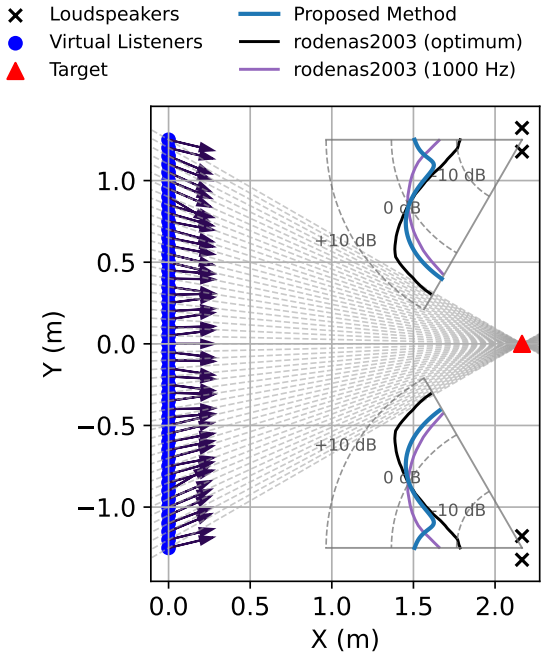
For this trade off analysis, the lateral localisation loss (Eq. 8), localisation uncertainty loss (Eq. 9) and colouration loss (Eq. 11) are combined. Table 1 presents the different α weightings applied for the different optimisations. β_{spread} is set to $1 \cdot 10^{-1}$.

3.4.2. Results and Discussion

The first three rows of Fig.,5 show results for optimising localisation, uncertainty, and colouration individually; the last row combines all three losses. Table 1 shows the corresponding weights and errors. All results here only consider the three positions that haven't been explicitly optimised for.



(a) Omni loudspeakers playing back identical signals (no optimisation).



(b) Proposed method with two pairs of loudspeakers.

Figure 4: Figure (a) shows the perceived directions as estimated by the localisation model [22] with no optimisation applied and omnidirectional stereo loudspeakers. Figure (b) shows the perceived directions obtained with the proposed method with two pairs of loudspeakers, along with the resulting directivity pattern of each loudspeaker pair, compared to results from [20]. The dashed lines indicate the target perceived direction.

Table 1: Summary of the results for the trade-off assessment introduced in Sec. 3.4. The first three columns show the loss weights. The last three columns show the model-based errors for localisation, uncertainty and colouration for the corresponding simulations in the three evaluation positions. The spread loss weight is $\beta_{\text{spread}} = 1$ for all simulations. Values in bold indicate best performing ones. The colouration loss is computed as in 11.

sim. #	α_{loc}	α_{unc}	α_{col}	$\epsilon_{\text{loc}}(^\circ)$	ϵ_{unc}	$\epsilon_{\text{col}}(\text{dB})$
Loc. only	1e3	0	0	**3.43**	0.48	9.33
Unc. only	0	1e-1	0	18.68	**0.25**	14.0
Col. only	0	0	1e-2	10.09	0.42	**3.16**
Combined	2e2	1e-2	5e-2	4.03	0.28	3.21

As shown in Fig. 5 and Table 1, optimising for a single loss results in that loss having lowest error, which is expected. When using localisation uncertainty only, the method yields a solution where essentially only one loudspeaker is active, as one would expect. The combined simulations shows results comparable to when optimising for the single attributes only. This is a surprising results as trying to localise a source in between two loudspeakers should increase localisation uncertainty. Further analysis could determine whether these attributes are complementary or reveal inconsistencies between the different models. These results are achievable through the careful selection of the weights associated to the perceptual attributes. As of now, this choice stems from trial and error; however, we are currently looking into a more systematic way to choose these parameters. Future work will also involve incorporating a more advanced colouration model like the one in [25], which has been shown to have a stronger correlation with perceptual data, and, as noted in [10], is differentiable.

As opposed to the three use cases presented above, the results here have only been evaluated with the same perceptual models used for optimisation, which makes the evaluation circular. Given the promising results in the previous use cases, one may hypothesise that the trends shown in these plots might also hold perceptually, but that remains to be tested in formal listening experiments.

4. CONCLUSIONS

The optimisation of loudspeaker reproduction on the basis of perceptual models has been an aim of many researchers for many decades. This paper demonstrated that what was previously only possible for simple models like Gerzon's velocity and energy vectors is now possible for more complex models, including full-fledged binaural ones. We owe this advance to (a) the fact that some of these models can be expressed in an end-to-end differentiable form, making gradient-based optimisation over a larger number of parameters computationally practical via automatic differentiation, and (b) the availability of parallelised and accelerated optimisation frameworks such as PyTorch.

The paper demonstrated the potential of the proposed framework, POLAR, through a selection of fundamental scenarios. Using a classic stereo setup, we showed that the framework yields filters comparable to VBAP or similar to CTC, if optimised for localisation or colouration and IPD, respectively. In a scenario optimising localisation with two pairs of closely spaced loudspeakers, the proposed method generated FIR filters that resulted in an appropriate loudspeaker beamformer [20], even though beamforming was not imposed explicitly.

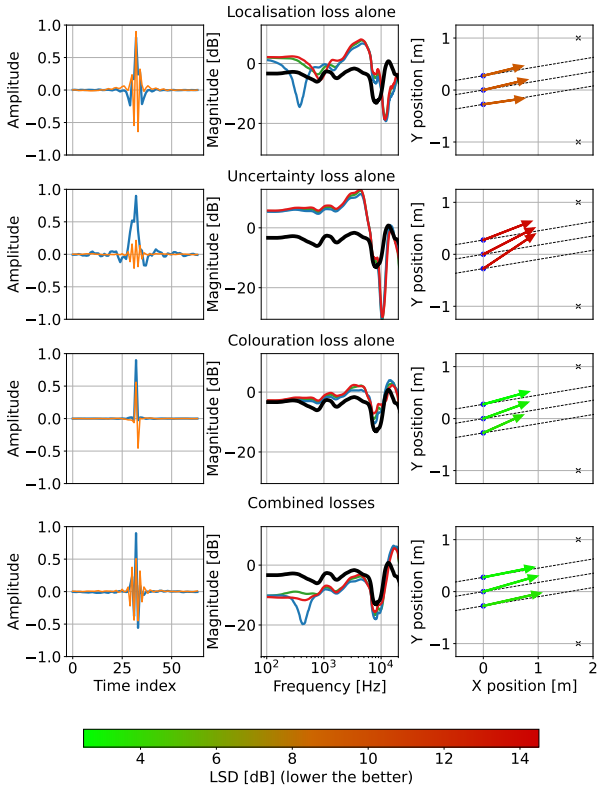


Figure 5: Effect of the optimisation using different weightings. Each row corresponds to one set of weights given in Table 1. The first column shows the IRs of the left (blue) and right (orange) loudspeaker filters. The middle column shows the magnitude spectrum at the left (blue), centre (green), and right (red) listening positions, with the black curve as the reference. The right column shows, for three positions not considered in the optimisation, arrow direction, length, and colour indicating estimated localisation, localisation certainty, and colouration respectively. Black crosses mark the loudspeakers and dashed black lines the target localisation directions.

These results pave the way for optimisation beyond what is currently possible with traditional spatial audio methods, e.g., hybrids of cross-talk cancellation and other reproduction methods, or in-situ optimisation of soundbar reproduction.

POLAR was also used for reproduction based on a combination of perceptual attributes, including localisation, localisation uncertainty and colouration. Future work will involve formal listening experiments. Furthermore, understanding the relative importance and role of each quality attribute and their interaction in overall quality perception is necessary for deriving efficient perceptual multi-attribute models.

5. ACKNOWLEDGMENT

This work was funded by the UKRI Engineering and Physical Sciences Research Council (EPSRC) under the 'Challenges in Immersive Audio Technology (CIAT)' grants EP/X032914/1 and EP/X032981/1.

6. REFERENCES

[1] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1973.

[2] F. Zotter and M. Frank, *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. Springer Nature, 2019.

[3] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2764–2778, 1993.

[4] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial sound with loudspeakers and its perception: A review of the current state," *Proc. IEEE*, vol. 101, no. 9, pp. 1920–1938, 2013.

[5] B. Rafaely, S. Weinzierl, O. Berebi, and F. Brinkmann, "Loss functions incorporating auditory spatial perception in deep learning – a review," in *2025 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, 2025.

[6] H. Hacihabiboglu, E. De Sena, Z. Cvetkovic, J. Johnston, and J. O. Smith III, "Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics," *IEEE Signal Process. Mag.*, vol. 34, no. 3, pp. 36–54, 2017.

[7] E. De Sena, H. Hacıhabiboğlu, and Z. Cvetković, "Analysis and design of multichannel systems for perceptual sound field reconstruction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 8, pp. 1653–1665, 2013.

[8] P. Izquierdo Lehmann, R. F. Cádiz, and C. A. S. Long, "Towards maximizing a perceptual sweet spot for spatial sound with loudspeakers," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 3304–3319, 2023.

[9] O. Berebi, Z. Ben-Hur, D. L. Alon, and B. Rafaely, "BSM-iMagLS: ILD informed binaural signal matching for reproduction with head-mounted microphone arrays," *IEEE Trans. Audio Speech Lang. Process.*, vol. 33, pp. 2705–2718, 2025.

[10] Y. Zhang, A. Francl, R. Gao, P. Calamia, Z. Duan, and I. Ananthabhotla, "Towards perception-informed latent hrtf representations," in *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2025.

[11] M. A. Gerzon, "General Metatheory of Auditory Localisation," in *Proceedings of the 92nd Audio Engineering Society Convention*. Vienna, Austria: Audio Engineering Society (AES), 1992, p. 2206.

[12] P. Stitt, S. Bertet, and M. Van Walstijn, "Extended Energy Vector Prediction of Ambisonically Reproduced Image Direction at Off-Center Listening Positions," *Journal of the Audio Engineering Society*, vol. 64, no. 5, pp. 299–310, May 2016.

[13] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[14] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, and S. Weinzierl, "A spatial audio quality inventory (SAQI)," *Acta Acust. United Ac.*, vol. 100, no. 5, pp. 984–994, 2014.

[15] P. Lladó, A. Neidhardt, F. Brinkmann, and E. De Sena, "Spatial audio models' inventory to cover the attributes from the spatial audio quality inventory," in *11th Convention of the EAA Forum Acusticum EuroNoise*, 2025, pp. 3457–3464.

[16] B. B. Bauer, "Phasor analysis of some stereophonic phenomena," *J. Acoust. Soc. Am.*, vol. 33, no. 11, pp. 1536–1539, 1961.

[17] J. C. Bennett, K. Barker, and F. O. Edeko, "A new approach to the assessment of stereophonic sound system performance," *J. Audio Eng. Soc.*, vol. 33, no. 5, pp. 314–321, 1985.

[18] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.

[19] M. Schroeder and B. Atal, "Computer simulation of sound transmission in rooms," *Proc. IEEE*, vol. 51, no. 3, pp. 536–537, 1963.

[20] J. A. Ródenas, R. M. Aarts, and A. Janssen, "Derivation of an optimal directivity pattern for sweet spot widening in stereo sound reproduction," *J. Acoust. Soc. Am.*, vol. 113, no. 1, pp. 267–278, 2003.

[21] A. Souchaud, P. Lladó, A. Neidhardt, Z. Cvetković, and E. De Sena, "White-box differentiable model of perceived localisation," in *2025 IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2025, pp. 264–268.

[22] P. Lladó, A. Neidhardt, A. Souchaud, Z. Cvetkovic, and E. De Sena, "Perceptually-driven panning for an extended listening area," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2025.

[23] E. De Sena, Z. Cvetković, H. Hacıhabiboğlu, M. Moonen, and T. van Waterschoot, "Localization uncertainty in time-amplitude stereophonic reproduction," *IEEE/ACM Trans. Audio Speech Lang. Proc.*, vol. 28, pp. 1000–1015, 2020.

[24] T. McKenzie and F. Brinkmann, "Toward an improved auditory model for predicting binaural coloration," *J. Audio Eng. Soc.*, vol. 73, no. 3, pp. 115–126, 2025.

[25] T. McKenzie, C. Armstrong, L. Ward, D. T. Murphy, and G. Kearney, "Predicting the Colouration between Binaural Signals," *Applied Sciences*, vol. 12, no. 5, p. 2441, Feb. 2022.

[26] ITU-R BS.1534-3, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," International Telecommunication Union, Tech. Rep., 2015.

[27] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Commun.*, vol. 53, no. 5, pp. 592–605, 2011.

[28] Y. Li, Y. Shen, and D. Wang, "Diffbas: An advanced binaural audio synthesis model focusing on binaural differences recovery," *Appl. Sci.*, vol. 14, no. 8, p. 3385, 2024.

[29] O. Selvi, "A note on digital filtering with the second moment norm," *Geophysics*, vol. 62, no. 4, pp. 1315–1320, 1997.

[30] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A perceptual evaluation of individual and non-individual hrtfs: A case study of the SADIE II database," *Appl. Sci.*, vol. 8, no. 11, 2018.

[31] B. Masiero, J. Fels, and M. Vorlander, "Review of the crosstalk cancellation filter technique," in *International Conference on Spatial Audio (ICSA)*, Detmold, Germany, 2011, pp. 112–116.