forum acusticum 2026
8-12 September · Graz, Austria

LEARNING-BASED MULTI-LISTENER SOUND FIELD REPRODUCTION USING BOUNDARY VELOCITY TARGETS

Vlad S. Paul[1], Nara Hahn[1], Philip A. Nelson[1]

[1] *Institute of Sound and Vibration Research, University of Southampton, United Kingdom*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

This paper studies learning-based sound field control for a prototype large-audience scenario with ten loudspeakers and multiple listeners distributed over a listening region. Two identical complex-valued multilayer perceptrons are compared. One is trained on a boundary normal-velocity objective and one is trained directly on binaural ear pressures. The velocity-based network is first validated against the analytical controller based on previous work by Shin et al [1] and shown to closely reproduce its region-based behavior. The two models are then compared using common region-based and listener-based metrics, including pressure error and binaural cue errors based on Interaural Level Difference (ILD) and Interaural Phase Difference (IPD). For ten listeners, direct ear pressure training defines an overdetermined 20-ear/10-loudspeaker problem, whereas the velocity-based model provides a more robust trade-off between regional accuracy and perceptual performance. The results indicate that region-based velocity training is a promising approach for overdetermined multi-listener sound field reproduction.

Keywords: *sound field control, complex-valued neural networks, multi-listener reproduction*

1. Introduction

Sound field reproduction aims to synthesize a desired acoustic field over a listening region using an array of loudspeakers. Established approaches include Wave Field Synthesis (WFS) [2] and Higher-Order Ambisonics (HOA) [3]. In parallel, optimization-based methods such as pressure matching have become widely used because they can accommodate practical loudspeaker layouts and directly optimize the reproduced field at selected control points [4], [5].

***Corresponding author**: v.s.paul@soton.ac.uk.*

For large audience reproduction, however, matching pressure only at a discrete set of positions may not provide sufficient control of the field throughout the surrounding region, especially when the array is sparse or non-uniform [6]. Shin et al. addressed this issue by replacing boundary pressure control with control of the normal particle velocity on the boundary [1]. For sparse and irregular loudspeaker layouts, they reported that this formulation is easier to regularize, produces smoother loudspeaker weights, and better preserves the sound-intensity flow relevant to localization.

Recent work has also moved toward more region-aware and learning-based formulations. Weighted pressure and mode matching have been used to better relate discrete control-point optimization to control over a broader target region [6], while deep learning approaches have shown that neural networks can be used either to estimate loudspeaker driving signals directly [7] or to compensate sound field synthesis for irregular loudspeaker arrays [8]. These developments suggest that, for sound field control over extended listening areas, a region-based physical objective may generalize better than a purely listener-specific pointwise objective, while a compact neural model may approximate a useful controller without requiring an online inversion solution [6], [7], [8].

In this work, we study sound field control for a prototype large-audience scenario using a ten-loudspeaker circular array and ten listeners distributed over a listening region. We train a compact complex-valued multilayer perceptron (cMLP) to predict loudspeaker gains from source angle with the surface particle velocity objective used by Shin et al [1], and compare it with an architecturally identical cMLP trained directly on listener ear pressures. The two approaches are evaluated using region-based reconstruction error together with binaural cue errors based on ILD and IPD, allowing the comparison of region-based and listener-based learning objectives in

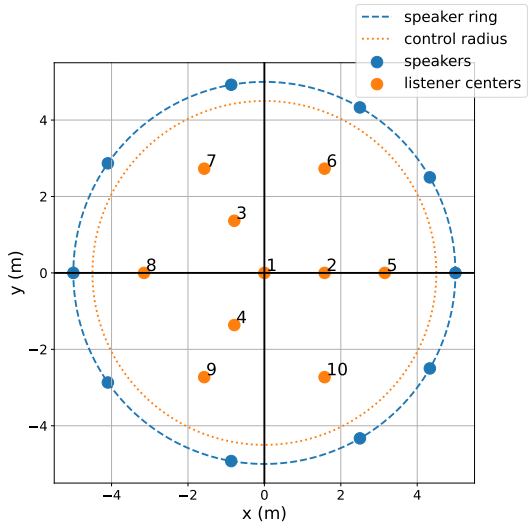

Figure 1: Geometry of the reproduction problem. Ten loudspeakers are placed on a circular array of radius 5 m around a control region of radius 4.5 m. Ten listeners are distributed inside the region, with one at the center, three on an inner ring, and six on an outer ring.

an overdetermined multi-listener setting.

2. Methods

2.1. Reproduction problem

We consider a two-dimensional multi-listener sound field reproduction scenario with a circular array of ten loudspeakers placed at a radius of 5 m. The listening region is defined by a circular control area of radius 4.5 m centered within the array. To emulate a prototype large-audience scenario, ten listeners are distributed inside this region. One listener is placed at the center, three on an inner ring, and six on an outer ring, as shown in Fig. 1. Each listener is modeled by a pair of ears separated by an interaural distance of 0.18 m.

Loudspeaker-to-ear transfer functions are constructed from measured head-related transfer functions (HRTFs) [9] in the horizontal plane. For each loudspeaker-listener pair, the nearest measured HRTF is selected according to the relative azimuth and the transfer function is adjusted according to the propagation distance from the loudspeaker to each ear by applying both amplitude scaling and a corresponding phase delay. This yields a frequency-domain plant that maps the ten loudspeakers to the 20 ears of the ten listeners.

Since the number of ear constraints exceeds the number of loudspeaker controls, the multi-listener ear pressure problem is overdetermined in this arrangement. Exact reconstruction of all target binaural signals is therefore not generally achievable, making the problem a useful test case for comparing region-based and listener-based control objectives. Virtual sources are placed on a circle of radius $d_{\mathrm{src}} = 5$ m, i.e., on the loudspeaker ring. All frequency-domain quantities are represented using a 2048-point FFT with the DC bin removed.

2.2. Velocity-based controller

Following Shin et al. [1], the reproduced sound field is described by a continuous distribution of secondary sources on the boundary of the reproduction region. Under the usual assumption of time-harmonic fields, the reproduced pressure on the boundary of the control region can be written as a single-layer potential

$$\hat{p}(\mathbf{x}) = j\omega\rho \int_{\partial\Lambda} G(\mathbf{x},\mathbf{y}) q(\mathbf{y}) dS(\mathbf{y}), \tag{1}$$

where $G(\mathbf{x},\mathbf{y})$ is the free-field Green's function and $q(\mathbf{y})$ denotes the source strength distribution on the loudspeaker boundary $\partial\Lambda$. Shin et al. then derive the reproduced particle velocity from Euler's relation and define the normal velocity as the component of the particle velocity along the inward normal to the control boundary ∂V.

Rather than controlling the acoustic pressure on the boundary, Shin et al. propose matching this normal particle velocity. In the discrete implementation, the continuous boundary integral is approximated using N loudspeakers located at positions $\mathbf{y}_n$ and M control points $\mathbf{x}_m$ on the boundary of the control region. The reproduced normal velocity at control point m is then written as

$$\hat{v}(\mathbf{x}_m) = \sum_{n=1}^{N} h_v(\mathbf{x}_m,\mathbf{y}_n)\, q_n, \tag{2}$$

where q_n is the n-th loudspeaker weight and

$$h_v(\mathbf{x}_m,\mathbf{y}_n) = jk\left(1+\frac{1}{jkr_{m,n}}\right) G(\mathbf{x}_m,\mathbf{y}_n)\left[\mathbf{a}_{m,n}\cdot\mathbf{u}_r(\mathbf{x}_m)\right] \tag{3}$$

is the transfer function from loudspeaker n to the normal particle velocity at control point m, with $\mathbf{a}_{m,n} = (\mathbf{x}_m - \mathbf{y}_n)/||\mathbf{x}_m - \mathbf{y}_n||$, $\mathbf{u}_r(\mathbf{x}_m)$ denoting the inward unit normal on the control boundary and $k = 2\pi f/c$ is the wavenumber. In matrix form, this gives $\hat{\mathbf{v}} = \mathbf{H}_v\mathbf{q}$, with $\mathbf{H}_v \in \mathbb{C}^{M\times N}$, $\mathbf{q} \in \mathbb{C}^N$, and $\hat{\mathbf{v}} \in \mathbb{C}^M$. In the present implementation, the control boundary is sampled using $M = 40$ control points uniformly distributed on the circular control region. Following the spatial bandwidth rule introduced by Shin et al., the control radius decreases in inverse proportion to frequency, $r(f) = c(N-1)/(2\omega)$, clipped to $[0, 4.5]$ m. This gives approximately 0.25 m at 1 kHz and 0.025 m at 10 kHz. [1].

At each frequency, the loudspeaker weights are obtained from a regularized least-squares problem

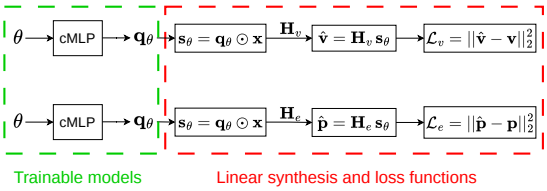

Figure 2: Overview of the two learning-based controllers. Two separate cMLPs with identical architecture map the source angle θ to complex loudspeaker gains $\mathbf{q}_\theta$. These gains are combined with the broadband probe spectrum $\mathbf{x}$ to form the loudspeaker driving spectra $\mathbf{s}_\theta$. The velocity-based model is trained through the velocity plant $\mathbf{H}_v$, while the ear pressure-based model is trained through the ear pressure plant $\mathbf{H}_e$.

using the closed-form solution

$$\mathbf{q}_{\text{opt}} = \left(\mathbf{H}_v^{\text{H}}\mathbf{H}_v + \beta\mathbf{I}\right)^{-1}\mathbf{H}_v^{\text{H}}\mathbf{v}, \tag{4}$$

where $\mathbf{v}$ is the target normal velocity vector. The regularization parameter follows $\beta = \beta_0\sigma_1^2$, with $\beta_0 = 10^{-4}$, where σ_1 is the largest singular value of $\mathbf{H}_v$ at each frequency.

In the present work, this formulation is used in two ways. First, it provides the analytical velocity-based baseline. Second, it defines the physical objective for the velocity-based complex-valued neural network, which is trained to predict loudspeaker gains from source angle without requiring a matrix inversion.

2.3. Learning-based controllers

To compare region-based and listener-based learning objectives under the same loudspeaker configuration, two compact cMLPs are considered. The velocity-based and ear pressure-based models use the same compact cMLP architecture. The input is a single complex scalar $e^{j\theta}$ representing the source angle. This is processed by two hidden complex-valued fully connected layers with 300 neurons each, followed by a complex cardioid activation [10], and a final linear complex-valued output layer. Using a 2048-point FFT with the DC bin removed, the frequency representation contains 1024 bins. For a 10-loudspeaker array, the output dimension is therefore $10 \times 1024 = 10240$ complex gains. Thus, the two models are architecturally identical. They differ only in the fixed acoustic plant and in the training objective used to optimize the predicted loudspeaker gains. Figure 2 illustrates the training pipelines of the two cMLP models.

For a given source angle θ, the network predicts a vector $\mathbf{q}_\theta$ containing the complex gains of all loudspeakers over all frequency bins. During training, these gains are multiplied elementwise by a broadband probe spectrum $\mathbf{x}_i$ to form the loudspeaker driving spectra

$$\mathbf{s}_{\theta,i} = \mathbf{q}_\theta \odot \mathbf{x}_i. \tag{5}$$

For both trained models, the same set of $N_x = 200$ independent white-noise probe spectra $\{\mathbf{x}_i\}_{i=1}^{N_x}$ is used at every source angle. Thus, each spatial target is paired with the same collection of white noise spectra, while the source angle remains the control parameter of interest.

The dataset is split by noise realization rather than by source angle. For each angle, 80% of the noise realizations are used for training and the remaining 20% for validation. As a result, the training and validation sets contain the same angular conditions but different probe spectra. The validation loss therefore reflects generalization to unseen spectral realizations rather than to unseen source angles.

The first model is a velocity-based cMLP. Its objective follows the velocity-based formulation of Shin et al., described in the previous section. Let $\mathbf{H}_v$ denote the transfer matrix from loudspeaker driving spectra to the normal particle velocity at the control points. The predicted boundary velocity is then

$$\hat{\mathbf{v}}_\theta = \mathbf{H}_v\mathbf{s}_\theta, \tag{6}$$

and the network is trained to minimize the mismatch with the target normal velocity

$$\mathcal{L}_v = \|\hat{\mathbf{v}}_\theta - \mathbf{v}_\theta\|_2^2, \tag{7}$$

where $\mathbf{v}_\theta$ is the target normal-velocity for the virtual source at angle θ. This model can therefore be interpreted as a learned surrogate of the velocity control in the manner of Shin et al, in which the mapping from source angle to loudspeaker gains is learned directly rather than obtained by solving a regularized inverse problem.

The second model is an ear pressure-based cMLP. Here, the objective is defined at the listener ears rather than on the boundary of the control region. Let $\mathbf{H}_e$ denote the transfer matrix from loudspeaker driving spectra to the binaural ear pressures of all listeners, and let $\mathbf{p}_\theta$ denote the target ear pressure vector for source angle θ. The binaural pressure vector $\mathbf{p}_\theta$ is formed by concatenating the left- and right-ear pressure spectra of all listeners for source angle θ. More precisely, $\mathbf{p}_\theta = [\mathbf{p}_{L,\theta}^{(1)}, \mathbf{p}_{R,\theta}^{(1)}, \ldots, \mathbf{p}_{L,\theta}^{(L)}, \mathbf{p}_{R,\theta}^{(L)}]^{\text{T}}$, with L denoting the number of listeners. The reproduced ear pressures are written as

$$\hat{\mathbf{p}}_\theta = \mathbf{H}_e\mathbf{s}_\theta, \tag{8}$$

and the corresponding training objective is

$$\mathcal{L}_e = \|\hat{\mathbf{p}}_\theta - \mathbf{p}_\theta\|_2^2. \tag{9}$$

In the present setup, this model attempts to match the binaural target signals of all listeners simultaneously.

3. Results

3.1. Evaluation metrics

Since the two controllers are trained with different objectives, their training losses are not directly comparable. We therefore evaluate both methods using common region-based and listener-based metrics: ear pressure error, region pressure error, region velocity error, and binaural cue errors based on ILD and IPD. All quantities are represented in the frequency domain, and frequency-bin indices are omitted for compactness unless explicitly needed.

First, the normalised ear pressure error is defined as

$$E_p = \mathbb{E}_\theta \left[\frac{\|\hat{\mathbf{p}}_\theta - \mathbf{p}_\theta\|_2^2}{\|\mathbf{p}_\theta\|_2^2} \right]. \tag{10}$$

where θ is the virtual source angle and $\mathbb{E}_\theta$ corresponds to an average over all azimuth angles. Next, the region based pressure error over a dense evaluation grid is defined as

$$E_{\Omega,p} = \mathbb{E}_\theta \left[\frac{\|\hat{\mathbf{p}}_{\Omega,\theta} - \mathbf{p}_{\Omega,\theta}\|_2^2}{\|\mathbf{p}_{\Omega,\theta}\|_2^2} \right], \tag{11}$$

where $\mathbf{p}_{\Omega,\theta}$ denotes the target pressure evaluated on the dense grid Ω.

In addition, we evaluate the reproduced full in-plane particle velocity field over the same dense grid. Let

$$\mathbf{v}_{\Omega,\theta} = \begin{bmatrix} \mathbf{v}_{x,\Omega,\theta} \\ \mathbf{v}_{y,\Omega,\theta} \end{bmatrix} \tag{12}$$

denote the target particle velocity vector field over Ω, and let $\hat{\mathbf{v}}_{\Omega,\theta}$ denote its estimate. The corresponding region based velocity error is defined as

$$E_{\Omega,v} = \mathbb{E}_\theta \left[\frac{\|\hat{\mathbf{v}}_{\Omega,\theta} - \mathbf{v}_{\Omega,\theta}\|_2^2}{\|\mathbf{v}_{\Omega,\theta}\|_2^2} \right]. \tag{13}$$

Here, the particle velocity at each grid point is obtained from the reproduced or target sound field by evaluating its horizontal components v_x and v_y, so that $E_{\Omega,v}$ measures the error in the full in-plane velocity field rather than only in the normal component used in the boundary control formulation. The dense evaluation grid Ω is a uniform mesh over the square region $x, y \in [-4.5, 4.5]$ m at 100×100 resolution, retaining only points that fall within the 4.5 m control radius (approximately 7,850 evaluation points).

The ILD is defined as

$$\mathrm{ILD}_\theta = 20 \log_{10} \frac{|\mathbf{p}_{L,\theta}|}{|\mathbf{p}_{R,\theta}|}, \tag{14}$$

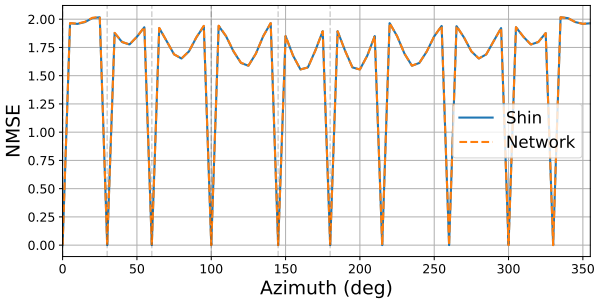

Figure 3: $E_{\Omega,p}$ as a function of source azimuth for the cMLP compared to the Shin et al baseline approach for a virtual source at 5 m radius. The reported NMSE is computed over the full evaluation frequency range from 20 Hz to 20 kHz.

and is evaluated for frequencies $f \geq 1500$ Hz. The IPD is defined as

$$\mathrm{IPD}_\theta = \angle \mathbf{p}_{L,\theta} - \angle \mathbf{p}_{R,\theta}, \tag{15}$$

and is evaluated for frequencies $f \leq 1500$ Hz, with phase wrapped to $(-\pi, \pi]$, where $\angle$ denotes the phase. Reported ILD and IPD errors are computed as absolute differences between reproduced and target values, averaged over the corresponding frequency bins and over all validation samples.

3.2. Velocity-guided learning versus analytical baseline

The velocity-based cMLP closely matches the analytical controller under the common evaluation metrics. Averaged over all source azimuths, both methods yield the same region-based pressure and particle-velocity errors, with $E_{\Omega,p} = 1.6$ and $E_{\Omega,v} = 1.7$. This indicates that the learned controller successfully reproduces the behavior of the analytical baseline while avoiding a matrix inversion.

A more detailed view is shown in Fig. 3, which reports the region error $E_{\Omega,p}$ as a function of source azimuth for both methods. The two curves follow identical trends over angle. This close agreement is expected given the conditioning of $\mathbf{H}_v$ in this geometry. Across the full evaluated band (20 Hz–20 kHz), the condition number of $\mathbf{H}_v$ remains below 4.14, well within the range where $\beta_0 = 10^{-4}$ has negligible regularizing effect. The analytical and learned controllers therefore converge to essentially the same solution. The pronounced minima occur when the virtual source position coincides with a loudspeaker position on the circular array. In the present evaluation, both the loudspeaker radius and the virtual source distance are 5 m, so source directions aligned with loudspeaker azimuths correspond to exact source-loudspeaker coincidence. In these cases, the target field can be reproduced directly by the corresponding loudspeaker, leading to near-zero NMSE for both the analytical Shin controller and the learned cMLP.

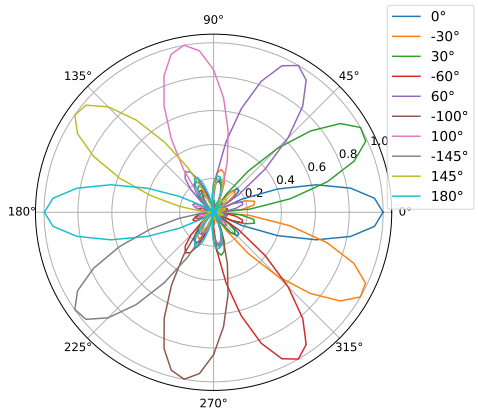

Figure 4: RMS loudspeaker gain vs. source azimuth (velocity cMLP). Each curve is one loudspeaker, labeled by its array position.

Table 1: Comparison between the velocity-based and ear pressure-based cMLPs in the ten-listener setup. Lower values are better. ILD and IPD are reported as mean absolute errors, in dB and radians, respectively.

Method	E_p	$E_{\Omega,p}$	ILD	IPD
Velocity cMLP	1.5	1.6	6.0	0.7
Ear pressure cMLP	0.4	2.4	7.5	0.8

Figure 4 shows that the velocity-based cMLP learns a smooth angle-to-gain mapping. For each source direction, the dominant gain is assigned to the nearest loudspeaker, with weaker contributions from neighboring loudspeakers. This is consistent with a panning-like control law on the circular array.

Based on these results, the velocity-based cMLP is used in the remainder of the paper as a learned surrogate of the analytical velocity controller. This allows the subsequent comparison with the ear pressure-based cMLP to focus on differences in training objective rather than on differences in network architecture.

3.3. Comparison of velocity- and ear pressure-based training objectives

We next compare the two cMLPs in the ten-loudspeaker, ten-listener arrangement. The ear pressure model solves a 20-ear/10-speaker pressure-matching problem, whereas the velocity model is trained on a region-based boundary-velocity objective. For the comparison, both models are evaluated on the same test set, built using a common loudspeaker-to-ear plant and a fixed set of $N_x = 100$ white noise signals reused across all source angles.

Table 1 summarizes the comparison using common evaluation metrics. As expected, the ear pressure model achieves the lower ear spectrum reconstruction error E_p, since it is optimized directly for binaural pressure matching. However, this advantage does not translate into lower binaural cue or region-based error in the ten listener case. The velocity-based model yields lower ILD and IPD error, and also lower region pressure error $E_{\Omega,p}$.

This shows that lower overall ear spectrum mismatch does not necessarily imply better binaural cue reconstruction. In particular, ILD and IPD depend on the relative left-right structure of the reproduced ear signals rather than on absolute reconstruction accuracy alone. In the present overdetermined setting, direct ear pressure training therefore becomes too listener specific, whereas the velocity objective provides a more robust compromise between regional reconstruction and listener side perceptual accuracy.

Representative listener side examples are shown in Fig. 5 and Fig. 6. Although the ear pressure model achieves lower overall ear spectrum error, the velocity-based model follows more closely the target binaural cues in the selected example, particularly in IPD and, to a lesser extent, ILD. The region-based comparison in Fig. 7 further shows that the velocity-based cMLP yields lower pressure error across source azimuths. As in the comparison with the work of Shin et al., both methods exhibit pronounced minima at source directions that coincide with loudspeaker positions, where the target field can be reproduced directly by a single loudspeaker. However, away from these directions, the velocity-based cMLP maintains consistently lower region error than the ear pressure model.

4. Conclusion

This paper investigated learning-based sound field control for a prototype multi-listener reproduction scenario using a ten-loudspeaker circular array. A complex-valued MLP trained on a boundary normal-velocity objective was first shown to reproduce the analytical velocity controller closely. This validated the use of the learned model as a surrogate for the original inverse formulation.

The main comparison then considered velocity-based and ear pressure-based training objectives under the same ten listener arrangement. Although the ear pressure model achieved the lower overall ear spectrum reconstruction error, the velocity-based model yielded lower binaural cue error and lower region pressure error. These results indicate that, in the overdetermined multi-listener regime considered here, direct ear pressure matching becomes too listener specific, whereas a region-based velocity objective provides a more robust compromise between spatial reconstruction and perceptual accuracy. Future work will examine this trade-off as a function of listener count and extend the evaluation to additional listener configurations and source conditions.

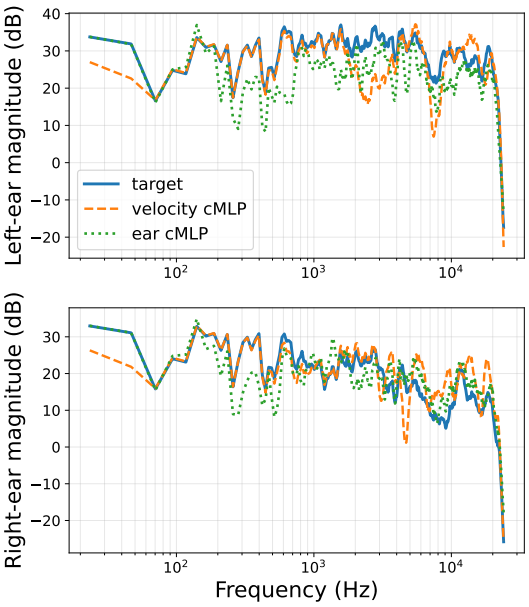

Figure 5: Representative ear pressure spectra for Listener 1 and a virtual source at $\theta = 45°$. For visual clarity, the ear pressure magnitude spectra are shown after smoothing on the log-frequency axis. All reported metrics are computed from the original spectra.

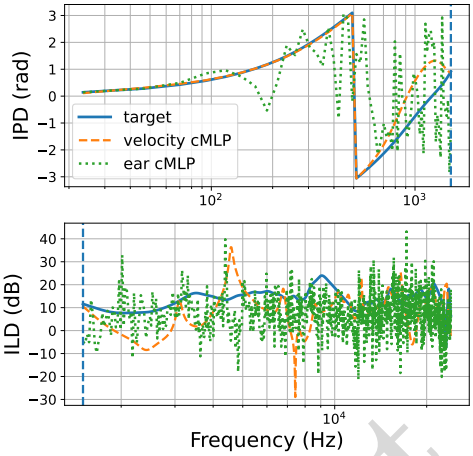

Figure 6: Representative binaural cues for Listener 1 and a virtual source at $\theta = 45°$. Performance is compared using IPD for $f \leq 1500$ Hz and ILD for $f \geq 1500$ Hz.

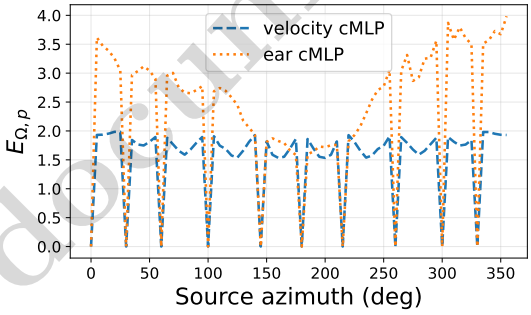

Figure 7: Region pressure error $E_{\Omega,p}$ as a function of source azimuth for the velocity-based and ear pressure-based cMLPs. The velocity-based model yields lower region error across most source directions.

References

[1] M. Shin, P. A. Nelson, F. M. Fazi, and J. Seo, "Velocity controlled sound field reproduction by non-uniformly spaced loudspeakers," *Journal of Sound and Vibration*, vol. 370, pp. 444–464, 2016.

[2] S. Spors, R. Rabenstein, and J. Ahrens, "The theory of wave field synthesis revisited," in *124th AES convention*, Audio Engineering Society (AES), 2008, pp. 17–20.

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

[4] O. Kirkeby and P. A. Nelson, "Reproduction of plane wave sound fields," *The Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 2992–3000, 1993.

[5] S. Koyama and K. Arikawa, "Weighted pressure matching based on kernel interpolation for sound field reproduction," *arXiv preprint arXiv:2210.14711*, 2022.

[6] S. Koyama, K. Kimura, and N. Ueno, "Weighted pressure and mode matching for sound field reproduction: Theoretical and experimental comparisons," *Journal of the Audio Engineering Society*, vol. 71, no. 4, pp. 173–185, 2023.

[7] X. Hong, B. Du, S. Yang, M. Lei, and X. Zeng, "End-to-end sound field reproduction based on deep learning," *The Journal of the Acoustical Society of America*, vol. 153, no. 5, pp. 3055–3055, 2023.

[8] L. Comanducci, F. Antonacci, and A. Sarti, "Synthesis of soundfields through irregular loudspeaker arrays based on convolutional neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 17, 2024.

[9] B. Bernschütz, *Spherical far-field hrir compilation of the neumann ku100*, Jul. 2020. DOI: 10.5281/zenodo.3928297. [Online]. Available: https://doi.org/10.5281/zenodo.3928297.

[10] P. Virtue, S. X. Yu, and M. Lustig, "Better than real: Complex-valued neural nets for mri fingerprinting," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3953–3957. DOI: 10.1109/ICIP.2017.8297024.