FLOW-HOA: GENERATIVE JOINT OPTIMIZATION FOR AMBISONICS ENCODING VIA FLOW MATCHING

Yuhuan You[1], *Yufan Qian*[1], *Tianshu Qu*[1*], *Bin Wang*[2], *Xueyang Lv*[3] †

[1]State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University, Beijing, China
[2]Beijing Xiaomi Mobile Software Co., Ltd, Beijing, China
[3]Xiaomi Communications Co., Ltd, Beijing, China
qutianshu@pku.edu.cn

ABSTRACT

Higher-Order Ambisonics (HOA) encoding from sparse, irregular microphone arrays remains a critical challenge for consumer spatial audio applications. We propose Flow-HOA, a generative framework that jointly optimizes for a multi-dimensional perceptual objective. By employing conditional flow matching, our method learns to map a prior distribution to a target distribution of efficient and deployable Finite Impulse Response (FIR) filters. This process is guided by a composite loss function targeting time-domain waveform error, multi-resolution spectral consistency, sub-band energy preservation, and spatial directivity. Objective evaluations demonstrate superior performance in both signal fidelity and spatial accuracy metrics. Subjective listening tests further confirm that Flow-HOA yields higher overall sound quality with reduced artifacts.

***Index Terms*—** Higher-Order Ambisonics, Flow Matching, Joint Optimization, Microphone Array Signal Processing, Spatial Audio

1. INTRODUCTION

The proliferation of virtual reality (VR) and augmented reality (AR) devices has established spatial audio capture and rendering as a key technology for enhancing user immersion. Higher-Order Ambisonics (HOA), a powerful and loudspeaker-independent format, represents the 3D sound field using a basis of spherical harmonic functions and has become a cornerstone technology in this domain [1]. Its core advantage lies in the ability to smoothly rotate the entire sound field, which is crucial for head-tracked VR/AR applications.

In theory, specially designed spherical microphone arrays allow for the direct computation of HOA signals via a plane-wave decomposition [2]. However, this approach is often impractical for consumer-grade devices like smartphones, which are typically equipped with only sparse and irregularly distributed microphones. To accommodate such arbitrary geometries, researchers have developed various methods[3, 4, 5, 6, 7, 8], broadly categorized as signal-independent and signal-dependent approaches [9].

Signal-independent methods design a universal, fixed filter bank based on the array geometry and target HOA order. Prominent baselines include signal-matching techniques like Ambisonics Signal Matching (ASM) [3], its variants [4, 5], and our prior work SHB-AE [6]. Despite their prevalence, these model-based methods are known to exhibit artifacts such as temporal smearing, spectral distortion, and spatial localization blur in practice [10].

In contrast, signal-dependent methods adapt their processing to the captured sound field's content. One branch, parametric synthesis, estimates source parameters like direction-of-arrival (DOA) to synthesize HOA signals; however, its accuracy is fundamentally bottlenecked by the challenge of robust multi-source DOA estimation with sparse arrays [7]. Another branch, end-to-end neural mapping [8], learns a direct transformation from microphone signals to HOA coefficients. While promising, these deep learning methods face significant challenges in generalization, deployment latency, and computational overhead.

A deeper analysis reveals that the limitations of prior methods stem from their reliance on a simplified, analytical optimization objective. Whether matching ideal plane wave coefficients or minimizing a simple mean squared error, these objectives fail to fully capture perceptual quality in complex acoustic scenes. To overcome this bottleneck, we argue for a shift towards a data-driven optimization paradigm based on a composite perceptual metric. Such an objective is highly non-convex and cannot be solved analytically, thus necessitating a powerful optimizer to explore its complex solution space. To this end, we propose Flow-HOA. We formulate the HOA filter design problem as a generative joint optimization task, leveraging the conditional flow matching technique. Flow Matching offers an efficient and stable method for training Continuous Normalizing Flows (CNFs) [11] and has recently achieved great success in various generative audio tasks [12, 13]. Our core contributions are as follows:

1. **A generative joint optimization framework.** We depart from traditional analytical design, instead formulating HOA filter design as a generative task of learning a probability distribution.
2. **A composite perceptual optimization objective.** This objective function jointly considers several key perceptual metrics across the time, frequency, and spatial domains, guiding the model to generate filters that excel on multiple dimensions.
3. **Efficient generation of FIR filters.** Unlike computationally expensive end-to-end models, the final output of our framework is a set of fixed FIR filter efficients, which are easy to deploy on resource-constrained devices.

The rest of this paper is organized as follows. Section 2 details the proposed Flow-HOA method. Section 3 presents the experimental setup, evaluation, and results. Finally, Section 4 provides concluding remarks.

*Corresponding author

†Thanks to the National Key Research and Development Program of China (No.2024YFB2808902), and the High-performance Computing Platform of Peking University for funding.

2. PROPOSED METHODS

This section details the Flow-HOA framework, designed to bridge the gap between simplified models and complex physical reality. We introduce a physics-informed prior, define a joint perceptual objective, and use conditional flow matching to synthesize filters. Notably, the entire framework is order-independent, ensuring its direct scalability to higher-order systems.

2.1. Physics-Informed Prior Filter Design

Ambisonics provides a complete representation of a three-dimensional sound field using a basis of Spherical Harmonics (SH) [1]. For a plane wave signal $s(t)$ arriving from a direction $\Omega_k = (\theta_k, \phi_k)$, its ideal HOA signal $b_{nm}^{\sigma}(t)$ is defined as:

$$b_{nm}^{\sigma}(t) = s(t) \cdot Y_{nm}^{\sigma}(\theta_k, \phi_k) \tag{1}$$

where Y_{nm}^{σ} is the real-valued SH of order (n, m) and type σ. In this work, we define the task of HOA encoding as designing a Finite Impulse Response (FIR) filter matrix $\mathbf{H} \in \mathbb{R}^{C \times Q \times L}$ that accurately estimates these $C = (N+1)^2$ ideal HOA signals from the signals $\mathbf{x}(t)$ captured by Q microphones.

The first step in our framework is to construct a physically plausible physics-informed prior filter, $\mathbf{h}_{\text{prior}}$, by solving a time-domain least-squares problem. This approach is considered "physics-informed" because the optimization is directly constrained by two physical principles: the measured impulse responses ($\mathbf{d}_{k,q}$) that capture the real-world acoustics of the array, and the target response ($\mathbf{y}_k$) derived from the physical theory of Spherical Harmonics. This approach, grounded in the principles of robust broadband time-domain array processing [10], aims to ensure that for a unit impulse input from any direction Ω_k, the system's filtered output approximates an ideal, delayed delta function weighted by the corresponding SH value, $\mathbf{y}_k(n) = Y_{nm}^{\sigma}(\Omega_k)\delta[n - n_0]$. This is equivalent to minimizing the following cost function:

$$J(\mathbf{H}) = \sum_{k=1}^{K} \left\| \sum_{q=1}^{Q} (\mathbf{h}_q * \mathbf{d}_{k,q}) - \mathbf{y}_k \right\|_2^2 \tag{2}$$

where $\mathbf{h}_q \in \mathbb{R}^L$ is the filter to be solved for, and $\mathbf{d}_{k,q} \in \mathbb{R}^{L_{IR}}$ is the measured impulse response. The solution, found via the Moore-Penrose pseudoinverse [14], yields excellent waveform reproduction accuracy. This filter serves as the starting point for our neural generative method, not as the final solution.

2.2. Perceptual Joint Optimization Objective

The objective of the time-domain optimization described above (i.e., matching an ideal impulse) deviates from the complexities of human auditory perception. To generate perceptually superior filters, we move beyond impulse-based optimization to a data-driven joint objective, $\mathcal{L}_{\text{joint}}$, evaluated on continuous audio signals. The optimal filter $\mathbf{h}^*$ should minimize the expected loss of this objective function over a distribution of signals $\mathcal{S}$ encompassing a wide variety of acoustic scenes:

$$\mathbf{h}^* = \arg\min_{\mathbf{h}} \mathbb{E}_{s \sim \mathcal{S}}[\mathcal{L}_{\text{joint}}(\mathbf{h}; s)] \tag{3}$$

This joint loss function is a weighted sum of four key components, designed to comprehensively characterize the perceptual attributes of a high-fidelity HOA signal:

$$\mathcal{L}_{\text{joint}} = \lambda_{\text{mse}}\mathcal{L}_{\text{mse}} + \lambda_{\text{stft}}\mathcal{L}_{\text{stft}} + \lambda_{\text{energy}}\mathcal{L}_{\text{energy}} + \lambda_{\text{spatial}}\mathcal{L}_{\text{spatial}} \tag{4}$$

To ensure fundamental waveform fidelity, the objective includes a time-domain Mean Squared Error (MSE) loss term, $\mathcal{L}_{\text{mse}}$. It suppresses temporal smearing by directly penalizing the time-domain discrepancy between the predicted signal $\mathbf{z}_{\text{est}} \in \mathbb{R}^{B \times N'}$ and the ideal signal $\mathbf{y}_{\text{ideal}} \in \mathbb{R}^{B \times N'}$, where B is the batch size and N' is the number of samples in the training audio segment.

$$\mathcal{L}_{\text{mse}} = \frac{1}{BN'} \sum_{b=1}^{B} \sum_{n=1}^{N'} (\mathbf{z}_{\text{est},c}^{(b)}[n] - \mathbf{y}_{\text{ideal},c}^{(b)}[n])^2 \tag{5}$$

To ensure the accuracy of timbre and spectral details, we introduce a multi-resolution Short-Time Fourier Transform (STFT) loss, $\mathcal{L}_{\text{stft}}$. This loss, inspired by recent advances in audio synthesis, effectively evaluates spectral similarity across different time-frequency scales [15]. It imposes fine-grained constraints on the STFT spectrogram over M different resolutions, measuring both the overall spectral magnitude and log-magnitude differences, thereby more closely aligning with the human auditory system's perception of sound spectra.

$$\begin{aligned} \mathcal{L}_{\text{stft}} = \sum_{i=1}^{M} \Bigg(& \frac{1}{B} \sum_{b=1}^{B} \frac{\| |Y_c^{(b,i)}| - |Z_c^{(b,i)}| \|_F}{\| |Y_c^{(b,i)}| \|_F + \epsilon} \\ & + \frac{1}{BT_iF_i} \sum_{b=1}^{B} \sum_{t,f} \left| \log(|Y_c^{(b,i)}|) - \log(|Z_c^{(b,i)}|) \right| \Bigg) \end{aligned} \tag{6}$$

where $|Y_c^{(b,i)}|$ and $|Z_c^{(b,i)}|$ represent the STFT magnitude spectrograms of the ideal and estimated signals for the c-th HOA channel in the b-th batch item at the i-th resolution, respectively. T_i and F_i are the number of time frames and frequency bins for that resolution.

Furthermore, to prevent the optimizer from sacrificing critical frequency bands to reduce the overall error, we have designed an energy preservation loss, $\mathcal{L}_{\text{energy}}$. This term ensures that the signal's energy distribution across P different frequency bands is consistent with the ideal signal, effectively preventing spectral holes or energy imbalances.

$$\mathcal{L}_{\text{energy}} = \frac{1}{BP} \sum_{b=1}^{B} \sum_{j=1}^{P} (\log E_{\text{est}}^{(b)}(j) - \log E_{\text{ideal}}^{(b)}(j))^2 \tag{7}$$

Finally, to guarantee the directional accuracy of the reconstructed sound field, we introduce a spatial fidelity loss, $\mathcal{L}_{\text{spatial}}$. It achieves precise matching of the target spatial directivity patterns by performing virtual beamforming in multiple directions and minimizing the log-difference between the energy of the estimated and ideal signals in each direction.

$$\mathcal{L}_{\text{spatial}} = \frac{1}{BK'} \sum_{b=1}^{B} \sum_{k=1}^{K'} (\log E_{\text{est}}^{(b)}(\Omega_k) - \log E_{\text{ideal}}^{(b)}(\Omega_k))^2 \tag{8}$$

2.3. Generative Synthesis via Flow Matching

Directly optimizing the high-dimensional, non-convex joint loss function $\mathcal{L}_{\text{joint}}$ is extremely difficult and unstable. We therefore re-frame the problem from a conventional optimization task to one of generative modeling, aiming to learn a model that can directly generate optimal filters from a simple prior distribution. To this end, we adopt the conditional flow matching paradigm. The core of this method is to train a neural network G_{θ_c}, with a U-Net backbone [16], to accurately approximate the gradient vector field of the joint

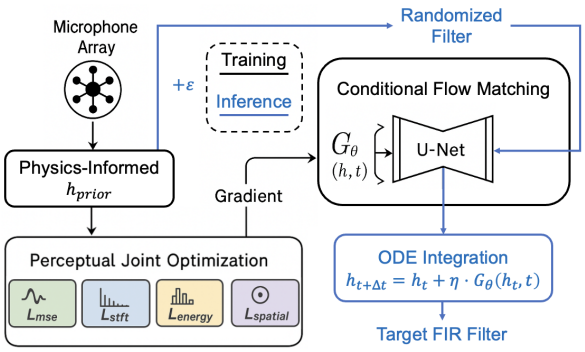

Fig. 1: Overview of Flow-HOA: training learns the perceptual-loss gradient; inference refines a physics-informed prior into FIR filters.

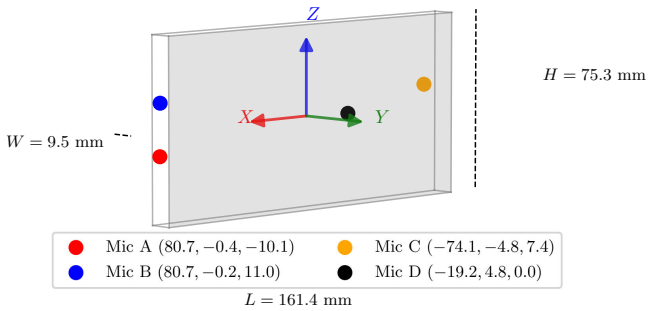

Fig. 2: Microphone array configuration of the SPMA.

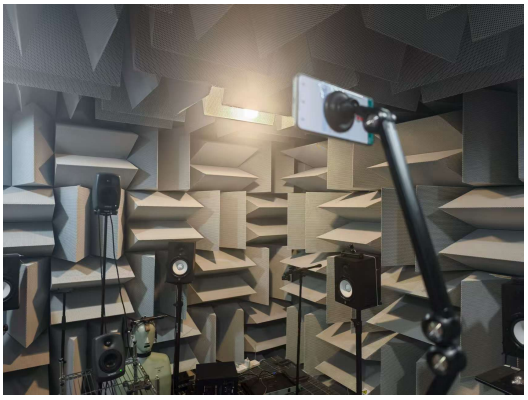

Fig. 3: Anechoic-chamber measurement setup.

loss function $\mathcal{L}_{\text{joint}}$. That is, it predicts the optimal descent direction $\mathbf{g}_c(\mathbf{h}) = -\nabla_{\mathbf{h}}\mathcal{L}_{\text{joint}}$ for any given filter $\mathbf{h}$.

Each training iteration involves two key computations. First, the target gradient is calculated: we sample a filter $\mathbf{h}_{\text{rand}}$ from a distribution centered around the physics-informed prior $\mathbf{h}_{\text{prior}}$ and compute its true task loss $\mathcal{L}_{\text{joint}}$ through a full forward pass of the physical simulation. Subsequently, using automatic differentiation, we obtain the gradient of the loss function with respect to the filter parameters. The negative of this gradient, after being smoothed via an Exponential Moving Average (EMA), constitutes the target vector $\bar{\mathbf{g}}_c$ for the training step. Concurrently, the generator network G_{θ_c} takes the filter $\mathbf{h}_{\text{rand}}$ and a random time step t as input, outputting a predicted gradient vector $\hat{\mathbf{g}}$. We update the network parameters θ_c by minimizing the mean squared error between the predicted and target vectors. This loss is known as the flow matching loss, $\mathcal{L}_{\text{FM}}$ [11]:

$$\mathcal{L}_{\text{FM}}(\theta_c) = \mathbb{E}_{\mathbf{h},t,s}\left[\|G_{\theta_c}(\mathbf{h},t) - \bar{\mathbf{g}}_c(\mathbf{h})\|_2^2\right] \tag{9}$$

This training procedure is repeated to train a separate expert network G_{θ_c} for each of the C HOA channels.

Once trained, the network G_{θ_c} defines the vector field of an Ordinary Differential Equation (ODE) that maps from a prior distribution to the distribution of optimal filters, a generation process that is theoretically linked to score-based generative models. During inference, we start from the physics-informed prior filter $\mathbf{h}_{\text{prior}}$, slightly perturbed by random noise, and iteratively solve this ODE using a numerical integrator such as the Euler method to refine it. The update rule for each step is as follows:

$$\mathbf{h}_{t+\Delta t} = \mathbf{h}_t + \eta \cdot G_{\theta_c}(\mathbf{h}_t, t) \tag{10}$$

After N steps of iteration, the final result $\mathbf{h}_N$ is the HOA filter that satisfies our complex perceptual objective.

As illustrated in Fig.1, the process starts with a prior filter computed from the array's transfer function. This prior is used to generate a perceptual loss gradient for training the network G_θ, which in turn acts as an ODE vector field during inference to refine the prior into the final FIR filter.

3. EXPERIMENTS

3.1. Experimental Setup

The target device in this study is a smartphone microphone array (SPMA) featuring four asymmetrically distributed microphones, with precise coordinates illustrated in Figure 2. The impulse responses (IRs) for filter design and evaluation were physically measured in an anechoic chamber from $K = 180$ spatial directions. These directions were formed by sampling 36 azimuth angles at 10° intervals across 5 elevation angles (0°, ±30°, ±60°). Figure 3 illustrates the measurement setup within the anechoic chamber, where a loudspeaker was excited by an exponential sine sweep (ESS) signal [17] to capture the IRs. The captured microphone signals for all experiments were then synthesized by convolving these measured IRs with anechoic source signals from the FSD50K dataset [18]. We evaluate our proposed method, Flow-HOA, against the conventional ASM baseline, which is designed analytically in the frequency domain. The training of the Flow-HOA model was conducted using the FSD50K development set. For each of the $C = 25$ HOA channels, an independent U-Net based generator was trained for 50 epochs using the AdamW optimizer with a learning rate of 10^{-5} and a cosine annealing scheduler. The training was performed with a batch size of 256 on 1-second audio segments with loss weights of $\lambda_{\text{mse}} = 50$ and $\lambda_{\text{stft}} = \lambda_{\text{energy}} = \lambda_{\text{spatial}} = 0.1$.

3.2. Evaluation Metrics

To comprehensively evaluate performance, we designed a suite of multi-dimensional, perceptually relevant objective metrics. All evaluations are conducted on a dedicated test set of anechoic audio signals not seen during training. For any given source signal $s(t)$ from a direction Ω_k, we synthesize the estimated HOA signal $\mathbf{b}_{\text{est},k}(t)$ by processing the microphone input signal with the filter bank under evaluation. This is then compared against the ground truth ideal HOA signal $\mathbf{b}_{\text{ideal},k}(t)$, which is obtained by multiplying the source

Table 1: Objective Evaluation Results.

Method	SI-SDR(dB)↑	LSD↓	SPM-KL↓	DGC(dB)↓
ASM	-13.72	11.12	1.44	2.17
Flow-HOA	**-7.31**	**5.07**	**1.14**	**0.84**

signal with the ideal spherical harmonic values. The following metrics are based on this comparison.

The signal fidelity is assessed in both the time and frequency domains. We measure the time-domain waveform fidelity using the well-established Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [19], where a higher value indicates less waveform distortion. To evaluate fidelity in the frequency domain, which is highly correlated with perceived timbre, we use the Log-Spectral Distance (LSD) [20], where a lower value signifies higher timbral fidelity.

To specifically address the crucial aspect of spatial accuracy, we define two metrics to evaluate the reconstructed sound field's structural integrity and consistency. The Kullback-Leibler divergence of the Spatial Power Map (SPM-KL) quantifies the difference between the spatial energy distribution of the estimated sound field (P_{est}) and the ideal distribution (P_{ideal}). These distributions are computed by beamforming the HOA signals onto a spherical grid of N_g points.

$$D_{KL}(P_{\text{ideal}} \| P_{\text{est}}) = \sum_{i=1}^{N_g} P_{\text{ideal}}(i) \log \left(\frac{P_{\text{ideal}}(i)}{P_{\text{est}}(i) + \epsilon} \right) \tag{11}$$

Directional Gain Consistency (DGC), defined as the standard deviation of the energy gain (in dB) across K test directions, was used to evaluate response uniformity. The gain for each direction, G_k, is the ratio of the output HOA signal energy to the input microphone signal energy, where $E(\cdot)$ denotes signal energy.

$$G_k = 10 \log_{10} \left(\frac{E(\mathbf{b}_{\text{est},k})}{E(\mathbf{x}_k)} \right), \text{DGC} = \text{std}\left(G_1, \ldots, G_K\right) \tag{12}$$

3.3. Objective Evaluation

The objective evaluation was performed on the FSD50K evaluation set. For each of the 100 randomly selected audio clips, performance was assessed at 50 randomly sampled spatial directions, resulting in 5000 unique evaluation instances. The quantitative results are summarized in Table 1.

As shown, the proposed Flow-HOA method significantly outperforms the conventional ASM baseline across all metrics. In terms of signal fidelity, Flow-HOA achieves a substantial 6.41 dB improvement in SI-SDR and more than halves the LSD score, indicating vast improvements in both time-domain waveform reconstruction and spectral fidelity. Regarding spatial accuracy, the DGC value is drastically reduced from 2.17 dB to 0.84 dB, indicating a much more uniform gain response across directions. The improvement in SPM-KL further confirms that the spatial energy pattern reconstructed by Flow-HOA is more accurate. These comprehensive improvements validate the effectiveness of our generative joint optimization framework, which achieves a synergistic enhancement of time-domain, spectral, and spatial fidelity.

3.4. Subjective Evaluation

To assess perceptual performance, we conducted a formal MUSHRA-based listening test with 16 participants (11 male, 5 female, aged 20-25), all with no self-reported hearing loss. Using the web-MUSHRA framework [21], stimuli were presented over Sennheiser HD 600 headphones in a controlled environment. The stimuli were generated from anechoic source signals physically recorded in an anechoic chamber, which were unseen by the model during training. These signals, captured at 45-degree intervals, were processed by the filters under evaluation and rendered to binaural audio using the HRTF database from [22]. Participants were asked to evaluate overall sound quality and spatial localization separately. Low-passed stimuli and opposite-angle stimuli were used for the sound quality and spatial localization tests, respectively. Post-hoc screening was applied based on each listener's correct-identification rate for the reference and anchor trials.

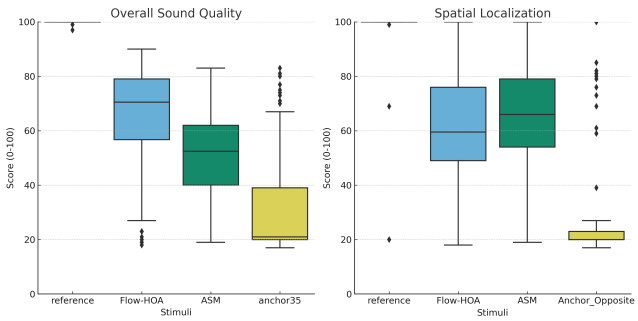

Fig. 4: Boxplots of the subjective listening test results for overall sound quality (left) and spatial localization (right).

Results (Fig. 4) were analyzed using paired-sample t-tests with Holm correction for multiple comparisons. For overall sound quality, Flow-HOA (Mean = 64.4) was rated significantly higher than the ASM baseline (Mean = 50.9), with a mean difference of 13.6 points (Holm-corrected $p < .001$), indicating a substantial improvement in perceptual audio fidelity consistent with the objective metrics. For spatial localization, no significant difference was found between Flow-HOA (Mean = 60.4) and ASM (Mean = 62.6; Holm-corrected $p = .40$). Interestingly, some participants reported a stronger in-head localization (IHL) effect with Flow-HOA. We hypothesize this is because our method's high-fidelity reproduction of the anechoic source signals made the inherent lack of externalization cues (e.g., reverberation) more apparent, which may have confounded spatial judgments [23, 24].

4. CONCLUSION

We introduced Flow-HOA, a framework that formulates the design of FIR filters for HOA encoding as a generative joint optimization task. Unlike conventional approaches based on simplified analytical objectives, our method employs conditional flow matching guided by a composite perceptual loss to learn the mapping from a prior distribution to optimal filters. Objective experiments showed that Flow-HOA outperformed the ASM baseline in signal fidelity and spatial accuracy. Subjective listening tests confirmed higher overall sound quality. The model's strong performance on real-world recorded stimuli unseen during training further underscores its generalization capabilities. They also revealed, however, that improved fidelity does not necessarily enhance spatial externalization, as in-head localization was still observed. This points to future work on integrating perceptually motivated externalization metrics into the optimization process to achieve a truly immersive experience.

5. REFERENCES

[1] Franz Zotter and Matthias Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, 01 2019.

[2] Jens Meyer and Gary Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," 06 2002, vol. 2.

[3] Yhonatan Gayer, Vladimir Tourbabin, Zamir Ben-Hur, David Alon, and Boaz Rafaely, "Ambisonics encoder for wearable array with improved binaural reproduction," 2025.

[4] Markus Zaunschirm, Christian Schörkhuber, and Robert Höldrich, "Binaural rendering of ambisonic signals via magnitude least squares," in *DAGA - Fortschritte der Akustik*, 2018, pp. 333–336.

[5] Or Berebi, Fabian Brinkmann, Stefan Weinzierl, and Boaz Rafaely, "Ambisonics binaural rendering via masked magnitude least squares," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Apr. 2025, p. 1–5, IEEE.

[6] Yuhuan You, Yufan Qian, Tianshu Qu, Bin Wang, and Xueyang Lv, "Spherical harmonic beamforming based ambisonics encoding and upscaling method for smartphone microphone array," in *Audio Engineering Society Convention 158*. Audio Engineering Society, 2025.

[7] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone Array Signal Processing*, vol. 1 of *Springer Topics in Signal Processing*, Springer, 2008.

[8] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2011–2023, 2019.

[9] Boaz Rafaely, *Fundamentals of Spherical Array Processing*, vol. 8 of *Springer Topics in Signal Processing*, Springer, 2015.

[10] Mark Poletti, "Unified theory of horizontal holographic sound systems," *J. Audio Eng. Soc.*, vol. 48, pp. 1155–1182, 12 2000.

[11] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le, "Flow matching for generative modeling," 2023.

[12] Alexander H. Liu, Matt Le, Apoorv Vyas, Bowen Shi, Andros Tjandra, and Wei-Ning Hsu, "Generative pre-training for speech with flow matching," 2024.

[13] Ziqian Wang, Zikai Liu, Xinfa Zhu, Yike Zhu, Mingshuai Liu, Jun Chen, Longshuai Xiao, Chao Weng, and Lei Xie, "Flowse: Efficient and high-quality speech enhancement via flow matching," 2025.

[14] Gene H. Golub and Charles F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, 4th edition, 2013.

[15] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, vol. 33, pp. 17022–17033.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 234–241.

[17] Angelo Farina et al., "Advancements in impulse response measurements by sine sweeps," in *Audio engineering society convention*, 2007, vol. 122.

[18] Eduardo Fonseca, Xavier Embar, Xavier Favory, Frederic Font, Alastair Porter, Annamaria Mesaros, and Xavier Serra, "FSD50K: an open dataset of human-labeled sound events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 561–565.

[19] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "Sdr–half-baked or well done?," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[20] A. H. Gray and J. D. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

[21] Michael Schoeffler, Sascha Bartoschek, Fabian-Robert Stöter, Moritz Roess, Sönke Westphal, Bernd Edler, and Jürgen Herre, "webMUSHRA—A comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, pp. 8, 2018.

[22] Benjamin Bernschütz, "A spherical far field hrir/hrtf compilation of the neumann ku 100," in *Proceedings of the 40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics*. German Acoustical Society (DEGA) Berlin, 2013, vol. 29.

[23] Jens Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, 1997.

[24] William M. Hartmann and Anthony Wittenberg, "On the externalization of sound images," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3678–3688, 1996.