



Zero-Shot Mono-to-Binaural Speech Synthesis

Alon Levkovitch¹, Julian Salazar², Soroosh Mariooryad², RJ Skerry-Ryan², Nadav Bar¹, Bastiaan Kleijn², Eliya Nachmani¹

¹Google Research

²Google DeepMind

alevkovitch@google.com, eliyen@google.com

Abstract

We present ZeroBAS, a neural method to synthesize binaural speech from monaural speech recordings and positional information without training on any binaural data. To our knowledge, this is the first published zero-shot neural approach to mono-to-binaural speech synthesis. Specifically, we show that a parameter-free geometric time warping and amplitude scaling based on source location suffices to get an initial binaural synthesis that can be refined by iteratively applying a pre-trained denoising vocoder. Furthermore, we find this leads to generalization across room conditions, which we measure by introducing a new dataset, TUT Mono-to-Binaural, to evaluate state-of-the-art monaural-to-binaural synthesis methods on unseen conditions. Our zero-shot method is perceptually on-par with the performance of supervised methods on previous standard mono-to-binaural dataset, and even surpasses them on our out-of-distribution TUT Mono-to-Binaural dataset.

Index Terms: mono-to-binaural, speech synthesis, zero-shot, diffusion

1. Introduction

Humans possess a remarkable ability to localize sound sources and perceive the surrounding environment through auditory cues alone. This sensory ability, known as *spatial hearing*, plays a critical role in numerous everyday tasks, including identifying speakers in crowded conversations and navigating complex environments. Hence, emulating a coherent sense of space via listening devices like headphones becomes paramount to creating truly immersive artificial experiences. Due to the lack of multi-channel and positional data for most acoustic and room conditions, the robust and low/zero-resource synthesis of binaural audio from single-source, single-channel (mono) recordings is a crucial step towards advancing augmented reality (AR) and virtual reality (VR) technologies.

Conventional mono-to-binaural synthesis techniques leverage a digital signal processing (DSP) framework. Within this framework, the head-related transfer function (HRTF), the room impulse response (RIR), and ambient noise are modeled as linear time-invariant (LTI) systems [1, 2, 3, 4]. These DSP-based approaches are prevalent in commercial applications due to their established theoretical foundation and their ability to generate perceptually realistic audio experiences. However, real acoustic propagation, unlike the one modeled by LTI systems, has nonlinear wave effects. Recent advancements in the field have witnessed a paradigm shift towards employing machine learning methods via the paradigm of supervised learning [5, 6, 7, 8, 9, 10]. The task of synthesizing binaural audio from monophonic sources presents a significant challenge for supervised learning models. This difficulty stems from two primary

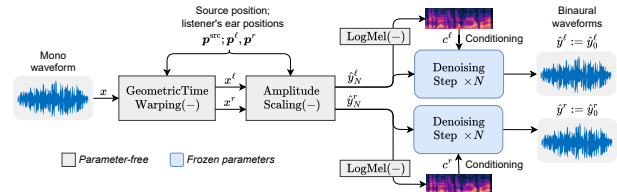


Figure 1: Our proposed ZeroBAS method. Mono waveform is binauralized with geometric time warping conditional on the speaker’s position, then the two channels’ amplitudes are scaled. Each channel is then denoised 3 times a monaural denoising vocoder.

limitations: (1) the scarcity of position-annotated binaural audio datasets, and (2) the inherent variability of real-world environments, characterized by diverse room acoustics and background noise conditions. Data collection for supervised learning necessitates specialized equipment, including tracking systems and binaural recording devices, which are both cost-prohibitive and often unavailable. Moreover, supervised models are susceptible to overfitting on the specific rooms, speaker characteristics, and languages in the training data, especially when the data is small (the standard dataset of [5] is only two hours). To address these limitations, we propose a novel zero-shot approach for monaural-to-binaural synthesis that is effective across a broader spectrum of recording scenarios by leveraging a monaural vocoder trained on tens of thousands of hours (Figure 1). Our contributions are:

- The first zero-shot method for neural mono-to-binaural speech synthesis, leveraging geometric time warping, amplitude scaling, and a (monaural) denoising vocoder [WaveFit; [11]]. Notably, we achieve natural binaural speech generation that is perceptually on par (MOS, MUSHRA) with existing supervised methods despite never seeing binaural data.
- A novel dataset-building approach and dataset, TUT Mono-to-Binaural, derived from the location-annotated ambisonic recordings of speech events in the TUT Sound Events 2018 dataset [12]. When evaluated on this out-of-distribution data, past supervised methods degrade significantly while ZeroBAS continues performing well.

2. Related Work

DSP techniques approach the mono-to-binaural problem as a stack of acoustic components, each of which is an LTI system. Accurate wave-based simulation of RIRs is computationally expensive, and thus most real-time systems rely on simplified geometrical models [13, 1]. HRTFs need to be recorded inside an anechoic chamber in about 10k locations for good results [14].

DSP approaches treat these functions as a series of convolutions that are applied to the input signal. [5] proposed one of first uses of neural networks for mono-to-binaural synthesis, composing a neural time-warping module (WarpNet) and a temporal (hyper-)convolutional neural network (CNN) to learn a direct map between mono and binaural waveforms. BinauralGrad [7] was the first to use a denoising diffusion probabilistic model (DDPM). Since then, better incorporation of the inductive biases from DSP have led to more efficient neural systems. Neural Fourier Shift [NFS; [8]] predicts delays and scaling from speaker locations and achieve close to state-of-the-art performance with a significantly smaller model. DopplerBAS [9] found that incorporating the Doppler effect into the conditioning features improved the phase loss of both the WarpNet and BinauralGrad systems. [15] used a structured state space sequence (S4) model for the mono-to-binaural task. [6] show that mono-to-binaural audio synthesis can be performed end to end with the use of audio codes. Motivated by the difficulty of collecting HRTF and RIR data, [16] showed that an implicit HRTF can be learned by a temporal CNN. [17] and [18] showed that DNNs can be used to estimate RIR filters. [19] created a model for learning an implicit representation of an acoustic field. Furthermore, a different line of work uses visual conditioning for the generation of binaural audio [20, 21, 22, 23, 24].

3. Approach

Our proposed zero-shot mono-to-binaural synthesis method utilizes a three-stage architecture. The first stage follows previous work and performs geometric time warping (GTW) to manipulate the input mono waveform into two channels based on the provided position information. Subsequently, our proposed amplitude scaling (AS) module adjusts the amplitude of the warped signal. Finally, an existing denoising vocoder iteratively refines the processed signal to generate the binaural output composed of two channels. Figure 1 provides a visual representation of this pipeline. Let x denote the mono source signal. Its position at time t is given by the 3D vector $\mathbf{p}_t^{\text{src}}$. Let ℓ and r correspond to the listener’s left and right ear. Their positions at t are given by 3D vectors $\mathbf{p}_t^\ell, \mathbf{p}_t^r$. The system first applies GTW to x conditioned on $\mathbf{p}_t^{\text{src}}, \mathbf{p}_t^\ell$ and \mathbf{p}_t^r . This warping gives left and right preprocessed channels, denoted by x^ℓ and x^r . Then, AS is employed jointly on x^ℓ and x^r , conditioning on the same data. This step aims to further enhance the spatial perception of the signal. The resulting intermediate left and right channels are denoted by \hat{x}^ℓ and \hat{x}^r , respectively. Finally, the denoising step sets its noisy inputs $\hat{y}_N^\ell, \hat{y}_N^r$ to be the outputs of the previous stage, \hat{x}^ℓ, \hat{x}^r . This replaces the typical Gaussian noise initialization used when training or sampling from denoising models. $\hat{y}_N^\ell, \hat{y}_N^r$ are fed separately into the same pretrained denoising vocoder, which treats each waveform as mono audio. The temporal sequences of conditioning vectors $\mathbf{c}^\ell, \mathbf{c}^r$ are obtained by extracting the log-mel features of \hat{x}^ℓ, \hat{x}^r . A low noise level k is also conditioned on, to reflect that we are emulating an input that is “close” to a true binaural sample. In the case of our denoising vocoder, WaveFit [11], this noise level is given by a choice of conditioning timestep; specifically, the last timestep of the WaveFit training’s denoising process. This sampling is repeated for N iterations. In the Experiments section, we show that our approach produces a binaural speech rendering whose quality approximates the ground truth binaural audio. Note that our method does not take into account room effects nor the listener’s head shape. Thus, we produce spatial audio which imputes both a generalized low RIR room (regularized by all the

Algorithm 1 ZeroBAS algorithm:

Require: Denoising vocoder \mathcal{V}_θ , iteration count N , low noise level k , mono waveform x , speaker position \mathbf{p}^{src} , listener’s ear locations $\mathbf{p}^\ell, \mathbf{p}^r$.

$$x^\ell, x^r = \text{GeometricTimeWarping}(x, \mathbf{p}^{\text{src}}, \mathbf{p}^\ell, \mathbf{p}^r)$$

$$\hat{x}^\ell, \hat{x}^r = \text{AmplitudeScaling}(x^\ell, x^r, \mathbf{p}^{\text{src}}, \mathbf{p}^\ell, \mathbf{p}^r)$$

$$\mathbf{c}^\ell, \mathbf{c}^r = \text{LogMel}(\hat{x}^\ell), \text{LogMel}(\hat{x}^r)$$

$$\hat{y}_N^\ell, \hat{y}_N^r := \hat{x}^\ell, \hat{x}^r$$
for $i \leftarrow N$ **to** 1 **do**
 $\hat{y}_{i-1}^\ell, \hat{y}_{i-1}^r = \mathcal{V}_\theta(\hat{y}_i^\ell, \mathbf{c}^\ell, k), \mathcal{V}_\theta(\hat{y}_i^r, \mathbf{c}^r, k)$
end for
return $\hat{y}^\ell, \hat{y}^r := \hat{y}_0^\ell, \hat{y}_0^r$.

data the vocoder was trained on), and an implicit HRTF.

3.1. Geometric Time Warping (GTW)

GTW aims to estimate a warfield that separates the left and right binaural signals by applying the interaural time delay (ITD) based on the relative positions of the sound source and the listener’s ears. [5] proposed GTW as a method to generate an initial estimate of the perceived signals. Let S denote the signal’s sample rate and ν_{sound} represent the speed of sound. The system employs basic GTW on the monoaural signal x . This warping is achieved by computing a warfield for both the left and right listening channels, denoted by $\rho^\ell(t), \rho^r(t)$. The values of this warfield are computed using on the source and listener ear positions $\mathbf{p}_t^{\text{src}}, \mathbf{p}_t^\ell, \mathbf{p}_t^r$:

$$\rho^{\ell/r}(t) := t - \frac{S}{\nu_{\text{sound}}} \|\mathbf{p}_t^{\text{src}} - \mathbf{p}_t^{\ell/r}\|_2 \quad (1)$$

As this function takes non-integer values, the warped left and right signals \hat{x}^ℓ, \hat{x}^r can be defined with respect to the original indexing t via linear interpolation.

3.2. Amplitude Scaling (AS)

Human spatial perception of sound relies on various factors, including the ITD, the interaural level difference (ILD), and spectral cues due to HRTFs. While prior works [25, 26] suggest that the ILD is mostly caused by scattering off of the head and is dominant in human spatial perception for sounds with high frequencies, we find that amplitude scaling based on the inverse square law has a positive effect on the perceived spatial accuracy of the processed signal. Our approach aims to leverage this amplitude manipulation to enhance the spatial realism of the generated binaural audio. Let D be the Euclidean distance from the origin of the sound waves. Then by the inverse-square law, pressure drops at a $1/D^2$ ratio [27]. In the case of microphones, pressure manifests as amplitude. Acknowledging that the left-right microphone distance of the KEMAR mannequin used in datasets like [5] is only an approximation of human heads, we define:

$$D_t^\ell = \|\mathbf{p}^{\text{src}} - \mathbf{p}_t^\ell\|_2, \quad D_t^r = \|\mathbf{p}^{\text{src}} - \mathbf{p}_t^r\|_2. \quad (2)$$

Then, at each time step we scale down the magnitude of the side furthest from the source, using the ratio of the closer side’s distance versus the further side’s distance:

$$\hat{x}_t^\ell := \min(1, (D_t^r/D_t^\ell)^2) \cdot x_t^\ell, \quad (3)$$

$$\hat{x}_t^r := \min(1, (D_t^\ell/D_t^r)^2) \cdot x_t^r. \quad (4)$$

3.3. Denoising Vocoder

GTW and AS are simple, parameter-free operations that only roughly approximate binaural audio; using the warped and scaled speech signals \hat{x}^ℓ, \hat{x}^r as-is results in acoustic artifacts and inconsistencies. Hence, there is a need for further refinement to generate natural-sounding binaural audio. To this end, we propose that a sufficiently well-trained denoising vocoder could be used on each signal *independently*. We use a WaveFit neural vocoder [11] as our denoising vocoder model. It is a fixed-point iteration vocoder combined with the discriminator of generative adversarial networks, specifically MelGAN’s [28], to learn a sampling-free iterable map that can generate natural speech from a degraded input speech signal. As a vocoder, it takes log-mel spectrogram features and noise as input and produces clean waveform output. In WaveFit’s notation, we perform the iterated application of

$$\hat{y}_{i-1} := \mathcal{V}_\theta(\hat{y}_i, \mathbf{c}, k) := \mathcal{G}(\hat{y}_i - \mathcal{F}_\theta(\hat{y}_i, \mathbf{c}, k), \mathbf{c}), \quad (5)$$

where \mathbf{c} is the spectrogram to convert and \hat{y}_{i-1} is a candidate waveform refined from \hat{y}_i . \mathcal{G} is a parameter-free gain adjustment operator and \mathcal{F}_θ is the WaveGrad architecture [29] trained for reconstruction under a discriminator. At training time, the starting noise is given by $\hat{y}_K \sim \mathcal{N}(0, \Sigma_c)$ where Σ_c is a covariance matrix initialized as in SpecGrad [30] to capture the spectral envelope of \mathbf{c} ; both k, i iterate over $K, \dots, 1$. Then, at inference time, we express our “approximation” hypothesis by iterating at the noise level of WaveFit’s final denoising step ($k = 1$). We then iteratively denoise $\hat{y}_N^\ell, \hat{y}_N^r := \hat{x}^\ell, \hat{x}^r$, conditioning on their initial log-mel spectrograms and the fixed low noise level for steps $i = N, \dots, 1$.

4. Experiments

4.1. Data and Models

For our experiments we use two datasets. The first is the Binaural Speech dataset (BSD) released by [5]. The dataset contains paired mono and binaural audio with tracking information, jointly collected in a non-anechoic room; see [5] for more details. The second dataset is an adapted version of TUT Sound Events 2018 [12] which we name “TUT Mono-to-Binaural” (TMB). It contains 1,174 recordings, each about 2 seconds long. Overall, there are 2.15 hours of recordings in the dataset. The spoken language is French, speakers are recorded in a studio, and each recording is played in a single location. Using this dataset ensures a zero-shot evaluation for all of the methods tested in this paper, as none were trained on this data. For our DSP baseline, we use the open-source Resonance Audio package. The WaveFit vocoder is described in [11]. The pretrained weights we use were trained on the 60k-hour LibriLight audio-book dataset [31] as described in [11].

4.2. TUT Mono-to-Binaural: Dataset Construction

Our purpose in creating and using the TMB dataset is threefold: (a) demonstrate a new approach for creating mono-to-binaural synthesis datasets due to their scarcity, (b) evaluate the ability of different methods to generalize to different rooms and acoustic environments, and (c) evaluate the ability of different methods to generalize to different speakers. The TUT Sound Events 2018 is build for sound event localization and is composed of ambisonic recordings from the DCASE 2016, Task 2 dataset. In TUT Sound Events 2018, mono recordings were played back using a loudspeaker at distances ranging from 1-10

Table 1: *Objective and subjective evaluations on BSD*

	Model	W $\ell_2 \downarrow$	A $\ell_2 \downarrow$	P $\ell_2 \downarrow$	$\mathcal{L}_S \downarrow$	MOS \uparrow
Zero-Shot	DSP	0.812	0.052	1.572	1.91	3.84±0.19
	ZeroBAS	0.440	0.053	1.508	1.91	4.07±0.17
Supervised	WarpNet	0.179	0.037	0.968	1.52	3.86±0.16
	BGrad	0.128	0.030	0.837	1.25	4.01±0.14
	NFS	0.172	0.035	0.999	1.29	3.99±0.15
	GT	-	-	-	-	4.30±0.12

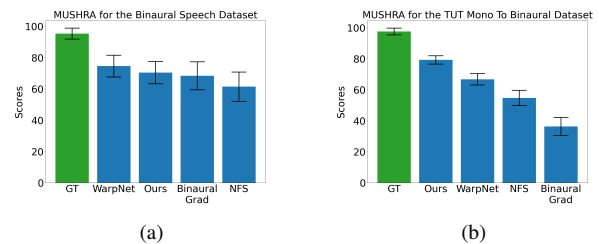


Figure 2: *MUSHRA results for (a) the BSD and (b) the TMB.*

meters and captured by an ambisonic microphone. Sound event locations are given using azimuth, elevation and distance, with each sound event having a single location. Starting from this data, we apply several processing steps: Frist, speaker location information provided in azimuth, elevation, and distance were converted into a Cartesian coordinate system (x, y, z) . Next, ground-truth metadata was leveraged to cut out speech segments from the recordings using their provided timestamps. To generate the binaural ground truth for our evaluation, ambisonic recordings are converted to binaural audio using OmniTone, a well-established DSP ambisonic decoder with a binaural renderer. Finally, the corresponding original monoaural recordings are obtained from the DCASE 2016, Task 2 dataset. Note that unlike the BSD, these mono recordings are recorded separately from their ambisonic re-recordings and binaural renderings.

4.3. Evaluations

For objective evaluations, we use metrics found in prior work. **Wave (W) ℓ_2** : mean squared error (MSE) between the ground truth and synthesized per-channel waveforms multiplied by 10^3 . **Amplitude (A) ℓ_2** : MSE between the amplitude STFTs of the ground truth and synthesized audio. **Phase (P) ℓ_2** : MSE between the left - right phase angle of the ground truth and synthesized audio. Phase is computed from the STFT. **MRSTFT \mathcal{L}_S** is the multi-resolution spectral loss. For subjective evaluations, we perform **MOS** and **MUSHRA** evaluations. For MOS, we collect mean opinion scores towards axes of naturalness. For every experiment, we use 50 random samples from each method. Every example is rated 5 times by different raters, with each experiment participated in by at least 30 raters. In the MUSHRA evaluation, we used 50 random samples from each method. Following the MUSHRA protocol, we discard raters who gave >15% of hidden references a score below 90. We used the model and code releases of WarpNet, BinauralGrad, and NFS to synthesize audio for subjective evaluations of these systems.

4.4. Binaural Speech Dataset Results

In Table 1, we observe that ZeroBAS achieves significant objective improvements over the DSP baseline, despite not mod-

Table 2: Objective and subjective evaluations on TMB.

	Model	W ℓ_2 ↓	A ℓ_2 ↓	P ℓ_2 ↓	\mathcal{L}_S ↓	MOS ↑
Zero-Shot	DSP	1.134	0.075	1.572	2.93	3.09±0.28
	ZeroBAS	0.293	0.045	1.572	2.93	3.98±0.15
Supervised	WarpNet	2.909	0.099	1.571	6.66	3.60±0.26
	BGrad	3.228	0.218	1.571	5.40	3.27±0.32
	NFS	1.574	0.085	1.571	3.06	3.79±0.23
	GT	-	-	-	-	4.08±0.11

eling additional interactions between the two generated channel streams or the RIR and HRTF. Furthermore, the performance of the ZeroBAS method approaches that of the supervised methods, even though ZeroBAS has not been trained on the BSD. Note that ZeroBAS inherently cannot model certain imperceptible environment-specific artifacts, like high-frequency recording equipment noise. Supervised methods may capture these and get superficial improvements on objectives like phase error, whereas vocoders may explicitly ignore them. In fact, subjective evaluation results in Table 1 show that ZeroBAS sounds slightly more natural to human raters than the supervised methods while being on par in comparative evaluations. MUSHRA results (Figure 2) no statistically significant preference for any of the methods WarpNet, BinauralGrad, NFS or ZeroBAS. The ZeroBAS system leverages a WaveFit model which ensures the generated audio exhibits minimal artifacts and noise compared to binaural recordings, leading to improved perceptual quality for human listeners. Despite worse objective metrics, our human evaluations suggest ZeroBAS achieves spatial fidelity and quality on par, if not better than supervised methods. Samples can be heard in our demo page: <https://alonlevko.github.io/zero-bas/>.

4.5. TUT Mono-to-Binaural Results

Although the zero-shot method underperforms supervised methods in the subjective evaluation on BSD, we argue that the supervised methods are sensitive to the room and recording conditions of BSD. To demonstrate this, we evaluated all methods on our newly constructed TMB. Table 2 demonstrates that our zero-shot method, ZeroBAS, significantly outperforms all supervised methods on TMB. Both ZeroBAS and the supervised methods struggle to capture accurate phase information, as evidenced by P ℓ_2 . The subjective evaluation results presented in Table 2 further demonstrate that ZeroBAS exhibits superior performance in terms of perceived naturalness compared to the supervised methods WarpNet, BinauralGrad, and NFS. As evidenced by the MOS, ZeroBAS surpasses these methods by notable margins. Considering the confidence intervals, these results indicate that human listeners on TMB perceive ZeroBAS as more natural-sounding than the supervised methods, with its score approaching that of the ground truth recordings. Furthermore, MUSHRA evaluations reveal a statistically significant preference for the proposed ZeroBAS method compared to supervised approaches. This suggests that human listeners perceive the spatial quality of binaural signals generated by ZeroBAS to best align with the reference. Evaluation of existing supervised learning methods on TMB revealed several limitations. BinauralGrad produced outputs with substantial Gaussian noise, hindering the diffusion process’s convergence to clean signals for out-of-distribution samples. WarpNet and NFS exhibited two key failure modes: (a) Inability to re-

Table 3: Ablation of our ZeroBAS method on BSD

Model	W ℓ_2 ↓	A ℓ_2 ↓	P ℓ_2 ↓	MOS ↑
ZeroBAS	0.440	0.053	1.508	4.07±0.17
w/o AS	0.802	0.059	1.539	2.93±0.16
w/o GTW	0.627	0.053	1.569	3.64±0.15
w/o AS, GTW	0.816	0.051	1.567	4.13±0.18
w/o WaveFit	0.539	0.044	1.572	3.52±0.16
Original Decode	0.495	0.065	1.534	2.50±0.16
Swap Order	0.474	0.072	1.277	3.85±0.19
1 iteration	0.459	0.069	1.393	3.62±0.20
2 iterations	0.450	0.061	1.492	3.83±0.24
4 iterations	0.445	0.053	1.502	3.94±0.18
5 iterations	0.449	0.053	1.494	4.05±0.15

tain speaker voice characteristics in the binaural output, leading to fidelity degradation, and (b) incorrect spatialization, manifesting as generated binaural speech with unrealistic distance cues or spatial artifacts when beyond the training range. These failures are further illustrated here: <https://alonlevko.github.io/zero-bas/>.

5. Ablation Analysis

The significance of each core component is evaluated through ablation studies (Table 3). First, AS is critical for ZeroBAS performance. Its absence leads to substantial degradation in both MOS and W ℓ_2 . AS creates a crucial perceptual difference. GTW is the second most important component. Without GTW, left-right channel time differences become misaligned, resulting in increased W ℓ_2 error and decreased MOS. The WaveFit model, when removed in isolation, has a minimal impact on objective metrics but a significant negative impact on MOS which highlights its importance. Removing both AS and GTW leads to improved MOS, albeit resulting in a monaural waveform played identically in both channels. In addition we tested the effects of modifications within WaveFit. Decoding for five iterations and initializing with Gaussian noise (Original Decode), as in the original WaveFit implementation, resulted in poor audio quality. This is because the two channels remain independent, and playing them as a binaural recording produces an unaligned and noisy output. Furthermore, applying WaveFit to the monaural input first (Swap Order), followed by AS and GTW, yielded improved performance in terms of P ℓ_2 but compromised MOS and A ℓ_2 . Finally, increasing the number of WaveFit iterations until 3 improves the objective metrics W ℓ_2 , A ℓ_2 and P ℓ_2 and improves MOS. After 3 iterations, quality is constant.

6. Conclusion

In this work, we presented a room-agnostic zero-shot method for binaural speech synthesis from monaural audio. Our results demonstrate that the method achieves perceptual performance comparable to supervised approaches on their in-distribution datasets. Furthermore, we introduce a novel dataset designed to evaluate the generalization capabilities of monaural-to-binaural synthesis methods for out-of-distribution scenarios. On this dataset, ZeroBAS exhibits superior performance compared to supervised methods, highlighting its potential for real-world applications with diverse acoustic environments.

7. References

- [1] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, “Creating interactive virtual acoustic environments,” *Journal of the Audio Engineering Society*, vol. 47, no. 9, pp. 675–705, 1999. [Online]. Available: www.aes.org/e-lib/browse.cfm?elib=12095
- [2] D. N. Zotkin, R. Duraiswami, and L. S. Davis, “Rendering localized spatial audio in a virtual auditory space,” *IEEE Trans. Multim.*, vol. 6, no. 4, pp. 553–564, 2004. [Online]. Available: doi.org/10.1109/TMM.2004.827516
- [3] K. Sunder, J. He, E. Tan, and W. Gan, “Natural sound rendering for headphones: Integration of signal processing techniques,” *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 100–113, 2015. [Online]. Available: doi.org/10.1109/MSP.2014.2372062
- [4] W. Zhang, P. N. Samarasinghe, H. Chen, and T. D. Abhayapala, “Surround by sound: A review of spatial audio recording and reproduction,” *Applied Sciences*, vol. 7, no. 5, p. 532, 2017. [Online]. Available: doi.org/10.3390/app7050532
- [5] A. Richard, D. Markovic, I. D. Gebru, S. Krenn, G. A. Butler, F. D. la Torre, and Y. Sheikh, “Neural synthesis of binaural speech from mono audio,” in *ICLR*. OpenReview.net, 2021. [Online]. Available: openreview.net/forum?id=uAX8q61EVRu
- [6] W. Huang, D. Markovic, A. Richard, I. D. Gebru, and A. Menon, “End-to-end binaural speech synthesis,” in *Interspeech*. ISCA, 2022, pp. 1218–1222. [Online]. Available: doi.org/10.21437/Interspeech.2022-10603
- [7] Y. Leng, Z. Chen, J. Guo, H. Liu, J. Chen, X. Tan, D. P. Mandic, L. He, X. Li, T. Qin, S. Zhao, and T. Liu, “BinauralGrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis,” in *NeurIPS*, 2022. [Online]. Available: papers.nips.cc/paper_files/paper/2022/hash/95f03faf3763e1b1ce2c3de62da8f090-Abstract-Conference.html
- [8] J. W. Lee and K. Lee, “Neural Fourier shift for binaural speech rendering,” in *ICASSP*. IEEE, 2023, pp. 1–5. [Online]. Available: doi.org/10.1109/ICASSP49357.2023.10095685
- [9] J. Liu, Z. Ye, Q. Chen, S. Zheng, W. Wang, Q. Zhang, and Z. Zhao, “DopplerBAS: Binaural audio synthesis addressing Doppler effect,” in *ACL*. Association for Computational Linguistics, 2023, pp. 11905–11912. [Online]. Available: doi.org/10.18653/v1/2023.findings-acl.753
- [10] C. Chen, A. Richard, R. Shapovalov, V. K. Ithapu, N. Neverova, K. Grauman, and A. Vedaldi, “Novel-view acoustic synthesis,” in *CVPR*. IEEE, 2023, pp. 6409–6419. [Online]. Available: doi.org/10.1109/CVPR52729.2023.00620
- [11] Y. Koizumi, K. Yatabe, H. Zen, and M. Bacchiani, “WaveFit: an Iterative and non-autoregressive neural vocoder based on fixed-point iteration,” in *IEEE-SLT*. IEEE, 2022, pp. 884–891. [Online]. Available: doi.org/10.1109/SLT54892.2023.10022496
- [12] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,” *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 1, pp. 34–48, 2019. [Online]. Available: doi.org/10.1109/JSTSP.2018.2885636
- [13] V. Välimäki, J. D. Parker, L. Savioja, J. O. S. III, and J. S. Abel, “Fifty years of artificial reverberation,” *IEEE Trans. Speech Audio Process.*, vol. 20, no. 5, pp. 1421–1448, 2012. [Online]. Available: doi.org/10.1109/TASL.2012.2189567
- [14] S. Li and J. Peissig, “Measurement of head-related transfer functions: A review,” *Applied Sciences*, vol. 10, no. 14, p. 5014, 2020. [Online]. Available: doi.org/10.3390/app10145014
- [15] K. Kitamura and K. Itou, “Binaural audio synthesis with the structured state space sequence model,” in *2023 9th International Conference on Computer and Communications (ICCC)*, 2023, pp. 1505–1509. [Online]. Available: doi.org/10.1109/ICCC59590.2023.10507442
- [16] I. D. Gebru, D. Markovic, A. Richard, S. Krenn, G. A. Butler, F. D. la Torre, and Y. Sheikh, “Implicit HRTF modeling using temporal convolutional networks,” in *ICASSP*. IEEE, 2021, pp. 3385–3389. [Online]. Available: doi.org/10.1109/ICASSP39728.2021.9414750
- [17] A. Richard, P. S. Dodds, and V. K. Ithapu, “Deep impulse responses: Estimating and parameterizing filters with deep networks,” in *ICASSP*. IEEE, 2022, pp. 3209–3213. [Online]. Available: doi.org/10.1109/ICASSP43922.2022.9746135
- [18] S. Lee, H. Choi, and K. Lee, “Differentiable artificial reverberation,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 2541–2556, 2022. [Online]. Available: doi.org/10.1109/TASLP.2022.3193298
- [19] A. Luo, Y. Du, M. J. Tarr, J. Tenenbaum, A. Torralba, and C. Gan, “Learning neural acoustic fields,” in *NeurIPS*, 2022. [Online]. Available: papers.nips.cc/paper_files/paper/2022/hash/151f4dfc71f025ae387e2d7a4ea1639b-Abstract-Conference.html
- [20] M. Chen, K. Su, and E. Shlizerman, “Be everywhere - hear everything (BEE): Audio scene reconstruction by sparse audio-visual samples,” in *ICCV*. IEEE, 2023, pp. 7819–7828. [Online]. Available: doi.org/10.1109/ICCV51070.2023.00722
- [21] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu, “AV-NeRF: Learning neural fields for real-world audio-visual scene synthesis,” in *NeurIPS*, 2023. [Online]. Available: papers.nips.cc/paper_files/paper/2023/hash/760dff0f9c0e9ed4d7e22918c73351d4-Abstract-Conference.html
- [22] A. Somayazulu, C. Chen, and K. Grauman, “Self-supervised visual acoustic matching,” in *NeurIPS*, 2023. [Online]. Available: papers.nips.cc/paper_files/paper/2023/hash/4cbec10b0cf25025e3f9fcfd943bb58c-Abstract-Conference.html
- [23] M. Yoshida, R. Togo, T. Ogawa, and M. Haseyama, “Binauralization robust to camera rotation using 360° videos,” in *ICASSP*. IEEE, 2023, pp. 1–5. [Online]. Available: doi.org/10.1109/ICASSP49357.2023.10096349
- [24] X. Xu, D. Markovic, J. Sandakly, T. Keebler, S. Krenn, and A. Richard, “Sounding bodies: Modeling 3D spatial sound of humans using body pose and audio,” in *NeurIPS*, 2023. [Online]. Available: papers.nips.cc/paper_files/paper/2023/hash/8c234d9c7e738a793947e0282c36eb95-Abstract-Conference.html
- [25] G. Wersényi, “Representations of HRTFs using MATLAB: 2D and 3D plots of accurate dummy-head measurements,” in *Proceedings of 20th International Congress on Acoustics, ICA 2010*, 2010, pp. 1–6. [Online]. Available: www.acoustics.asn.au/conference_proceedings/ICA2010/cdrom-ICA2010/papers/p45.pdf
- [26] F. Baumgarte and C. Faller, “Binaural cue coding - Part I: psychoacoustic fundamentals and design principles,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 509–519, 2003. [Online]. Available: doi.org/10.1109/TSA.2003.818109
- [27] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, “Auditory distance perception in humans: A summary of past and present research,” *ACTA Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, 2005.
- [28] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, “Melgan: Generative adversarial networks for conditional waveform synthesis,” in *NeurIPS*, 2019, pp. 14881–14892.
- [29] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, “WaveGrad: Estimating gradients for waveform generation,” in *ICLR*. OpenReview.net, 2021. [Online]. Available: openreview.net/forum?id=NsMLjcFaO8O
- [30] Y. Koizumi, H. Zen, K. Yatabe, N. Chen, and M. Bacchiani, “SpecGrad: Diffusion probabilistic model based neural vocoder with adaptive noise spectral shaping,” in *Interspeech*. ISCA, 2022, pp. 803–807. [Online]. Available: doi.org/10.21437/Interspeech.2022-301
- [31] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, “Libri-Light: A benchmark for ASR with limited or no supervision,” in *ICASSP*. IEEE, 2020, pp. 7669–7673. [Online]. Available: doi.org/10.1109/ICASSP40776.2020.9052942