applied sciences

MDPI

Article

DIFFBAS: An Advanced Binaural Audio Synthesis Model Focusing on Binaural Differences Recovery

Yusen Li, Ying Shen * and Dongqing Wang

School of Software Engineering, Tongji University, Shanghai 201804, China; 2211247@tongji.edu.cn (Y.L.); wangdongqing@tongji.edu.cn (D.W.)
* Correspondence: yingshen@tongji.edu.cn

Abstract: Binaural audio synthesis (BAS) aims to restore binaural audio from mono signals obtained from the environment to enhance users' immersive experiences. It plays an essential role in building Augmented Reality and Virtual Reality environments. Existing deep neural network (DNN)-based BAS systems synthesize binaural audio by modeling the overall sound propagation processes from the source to the left and right ears, which encompass early decay, room reverberation, and head/ear-related filtering. However, this end-to-end modeling approach brings in the overfitting problem for BAS models when they are trained using a small and homogeneous data set. Moreover, existing losses cannot well supervise the training process. As a consequence, the accuracy of synthesized binaural audio is far from satisfactory on binaural differences. In this work, we propose a novel DNN-based BAS method, namely DIFFBAS, to improve the accuracy of synthesized binaural audio from the perspective of the interaural phase difference. Specifically, DIFFBAS is trained using the average signals of the left and right channels. To make the model learn the binaural differences, we propose a new loss named Interaural Phase Difference (IPD) loss to supervise the model training. Extensive experiments have been performed and the results demonstrate the effectiveness of the DIFFBAS model and the IPD loss.

Keywords: binaural speech synthesis; binaural differences; interaural phase difference; deep neural network

Citation: Li, Y.; Shen, Y.; Wang, D. DIFFBAS: An Advanced Binaural Audio Synthesis Model Focusing on Binaural Differences Recovery. *Appl. Sci.* **2024**, *14*, 3385. https://doi.org/10.3390/app14083385

Academic Editor: Masayuki Takada

Received: 18 March 2024
Revised: 11 April 2024
Accepted: 12 April 2024
Published: 17 April 2024

1. Introduction

Humans can locate sound sources due to their ability to perceive differences in sound signals transferred to their left and right ears [1]. However, audio signals are usually recorded in mono channels, which cannot provide realistic spatial cues for the listener without any processing. Binaural Audio Synthesis (BAS) aims to synthesize binaural audio from mono audio based on given spatial information, such as a relative position of the sound source to the listener so that the synthesized binaural audio can provide immersive auditory experiences to the listener [2]. It is crucial, especially in the fields of virtual reality (VR) and augmented reality (AR) because it can restore real-world sound environments [3].

Most BAS methods rely on traditional digital signal processing (DSP) techniques [4] to model the true propagation process from the sound source to each ear, producing binaural audio with spatial information. As shown in Figure 1a, the propagation modeling can be decomposed into a series of components, including HRTF (head-related transfer function), room reverberation (e.g., RIR(room impulse response)), and environmental noise, each of which is modeled as a linear time-invariant system (LTI) [5–7]. To generate binaural audio, the DSP-based BAS methods often perform a series of convolutions with these systems. Among them, HRTF represents the transmission path of sound from a certain position in space to the listener's eardrum in the free field, which is a key component for binaural audio to have a sense of space. Due to its high computational cost for accurate numerical simulation [8–10], people tend to collect general discrete HRTF databases[11,12] in anechoic chambers.

Traditional DSP methods are easy to understand and relatively easy to model mathematically. However, the actual sound propagation process has non-linear characteristics that are not appropriately modeled by LTI systems [13]. In addition, since existing HRTF databases are discrete, inaccurate interpolation algorithms [14] will cause distortion of continuously moving sound sources. Therefore, in order to synthesize more realistic and natural binaural audio, DSP-based methods are far from satisfactory.

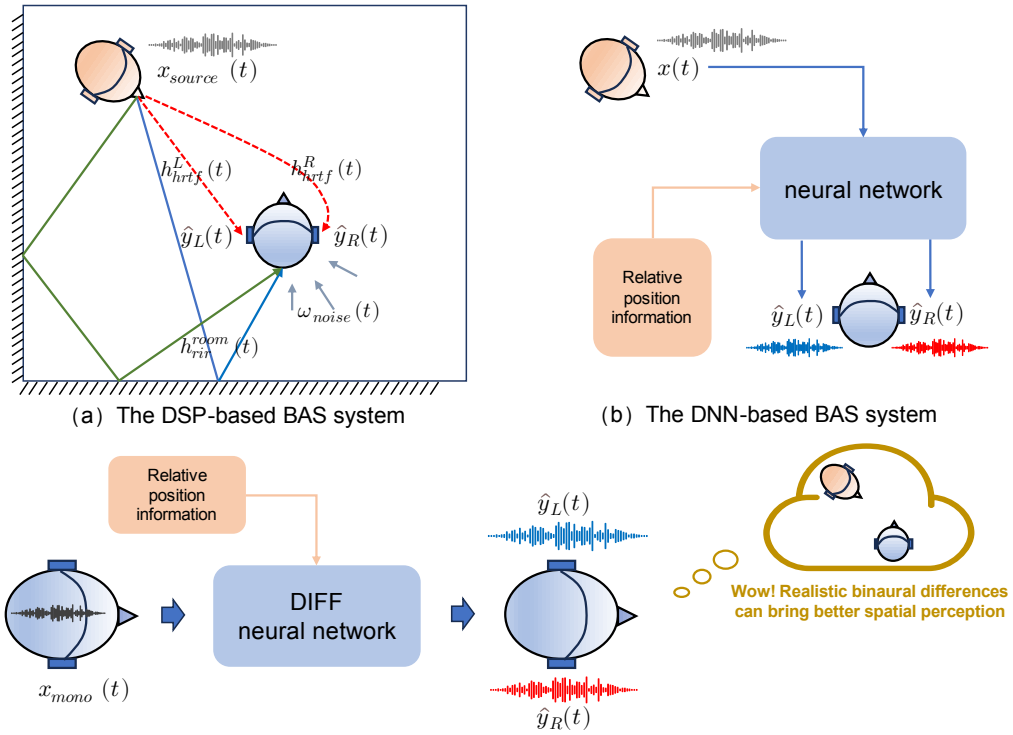


Figure 1. Binaural audio synthesis systems. (**a**) The DSP-based BAS model. Input mono signal from source $\boldsymbol{x}_{source}(t)$ is filtered by HRTF $h^{L/R}_{hrtf}(t)$ and room impulse response $h^{room}_{rir}(t)$ respectively and then superimposed with environmental noise $\omega_{noise}(t)$ to produce the binaural signals $\hat{\boldsymbol{y}}_L(t)$ and $\hat{\boldsymbol{y}}_R(t)$; (**b**) The DNN-based BAS method directly models the propagation process from the sound source to the left and right ears to produce $\hat{\boldsymbol{y}}_L(t)$ and $\hat{\boldsymbol{y}}_R(t)$; (**c**) The framework of proposed DIFFBAS model.

With the emergence of deep learning techniques, researchers set to utilize deep neural networks (DNN) to address the limitations of DSP-based methods. An end-to-end BAS model utilizing geometric and neural warping techniques was first proposed by Richard et al. [15]. It takes mono audio and spatial information (e.g., relative position information between the sound source and listener) as inputs to produce binaural audio which correlates with people's hearing experience (see Figure 1b). The model has been trained and tested on the first and only continuous binaural dataset, comprising paired monaural and binaural audio recordings captured in the real world. Based on this dataset, considering the limited frequency distribution of training samples, Lee et al. [16] rendered speech in the Fourier domain to improve the model's generalization to out-of-distribution audio. Although the introduction of DNN has brought significant advantages in addressing the nonlinear effects of sound propagation, the complexity of the sound propagation process and the supervised end-to-end training process may lead to model overfitting on small binaural datasets.

In order to simplify the sound propagation process, Leng et al. [17] suggested that the sound propagation process can be modeled in two stages which include a common path modeling (i.e., early sound propagation from the source to the listener's front head) and specific path modeling (i.e., the sound propagation from the listener's front head

to left/right ears). They also suggested using the generative model DDPM (Denoising Diffusion Probabilistic Model) [18] to enhance the naturalness of synthesized audio. However, they jointly trained these two stages, and the model may still cause overfitting in the common path modeling part, thus affecting the accuracy of the specific path modeling.

In fact, the spatial sense of human hearing mainly lies in the binaural differences perceived by the listeners [19] which is closely related to the specific path modeling instead of common path modeling. The overfitting problem in common path modeling may favor simulations of early attenuation and reverberation establishment, thus affecting the modeling of binaural differences and deviating from the original intention of the BAS task. In view of this, we propose the DIFFBAS method to decompose existing DNN-based methods (see Figure 1c). The design of DIFFBAS is intended to eliminate the modeling of common paths in the sound propagation process (such as reverberation effects), and focus on the modeling of specific paths, especially the filtering effects of the head and ears. This method not only reduces dependence on the dataset but also more accurately reproduces differences between ears. Specifically, DIFFBAS takes the mono audio at the point of the front head as input and adopts the newly proposed interaural phase difference (IPD) loss to supervise the model training. Such an operation is also reasonable in practice: the common path model can be trained separately and output a decayed mono signal which can be used as the input of DIFFBAS. Based on the decayed mono signals, DIFFBAS only has to focus on modeling the specific path to generate more accurate binaural audio which contains better spatial information. Meanwhile, the addition of IPD loss can better supervise the modeling of binaural differences based on the original loss supervision.

Our contributions are summarized as follows:

- A novel DNN-based BAS method, DIFFBAS is proposed. By focusing on modeling the specific path, DIFFBAS can avoid the overfitting problem caused by environmental and sound source diversity. The model employs new loss supervision for supplementary supervision and consequently improves the accuracy of synthesized binaural audio referring to the interaural phase difference between two ears.
- To better train the DIFFBAS model, a new loss, namely IPD loss, is proposed. This loss fits the perceptual characteristics of the human ear and ensures that the model will not lose too much spatial information during the training process. In addition, IPD is regarded as a new evaluation metric that reflects the accuracy of binaural phase difference in synthesized audio. It is the first objective metric that can quantify the spatialization degree of binaural audio.
- The performance of DIFFBAS has been validated on the only binaural dataset captured in the real world [15]. Sufficient contrast and ablation experimental results further demonstrate the effectiveness of DIFFBAS and IPD loss.

The source code and the datasets are available at https://github.com/tongjiRain/DIFFBAS.

2. Methodology

2.1. Problem Definition

Given a single-channel source audio $\boldsymbol{x}_{source} \in R^{1\times T}$ of length T and the relative position $\boldsymbol{p} = (P_s, R_\alpha)$ between the source and the listener, the objective of a BAS system is to generate binaural audio as

$$(\hat{\boldsymbol{y}}_L(t), \hat{\boldsymbol{y}}_R(t)) = f(\boldsymbol{x}_{source}(t), \boldsymbol{p}) \tag{1}$$

where $\hat{\boldsymbol{y}}_L$ and $\hat{\boldsymbol{y}}_R \in R^{1\times T}$, f denotes the transformation function parameterized by a DNN, $P_s = (x, y, z)$ is the spatial position that is indicated by the coordinates, and $R_\alpha = (q_x, q_y, q_z, q_w)$ is a quaternion that indicates the head orientation from the listener to the source. Function f models the overall sound propagation from the sound source to the left ear and right ear, which mixes modeling of the common path (e.g., early decay during audio transmission and the build of reverberation) and the specific path (e.g., the head/ear filtering). However, since the sound propagation model of the common path can be trained separately and

shared by left and right channels in the two-stage solution, we remove the common path modeling from f. Accordingly, we reformulate the problem as directly modeling the specific path of the sound propagation, i.e., the sound transfer from the front head to each ear, as follows

$$(\hat{\boldsymbol{y}}_L(t), \hat{\boldsymbol{y}}_R(t)) = \tilde{f}(\boldsymbol{x}_{mono}(t), \boldsymbol{p}) \tag{2}$$

where $\boldsymbol{x}_{mono}$ is no longer the mono audio collected from the sound source, but the mono audio collected in front of the listener's head. $\tilde{f}$ is in fact equivalent to a head/ear-related filter.

2.2. Framework of DIFFBAS

As shown in Figure 2, DIFFBAS accepts the mono audio $\tilde{\boldsymbol{x}}_{mono}$ and the relative position $\boldsymbol{p}$ as the inputs, and warps the mono audio to produce two synthesized audio for the left and right ears. The backbone network of DIFFBAS can be any DNN. To illustrate, select two networks to implement DIFFBAS, including WarpNet [15] which encodes audio in the temporal domain and NFS [16] which encodes audio in Fourier space. The networks are trained under the supervision of a loss function composed of three terms.

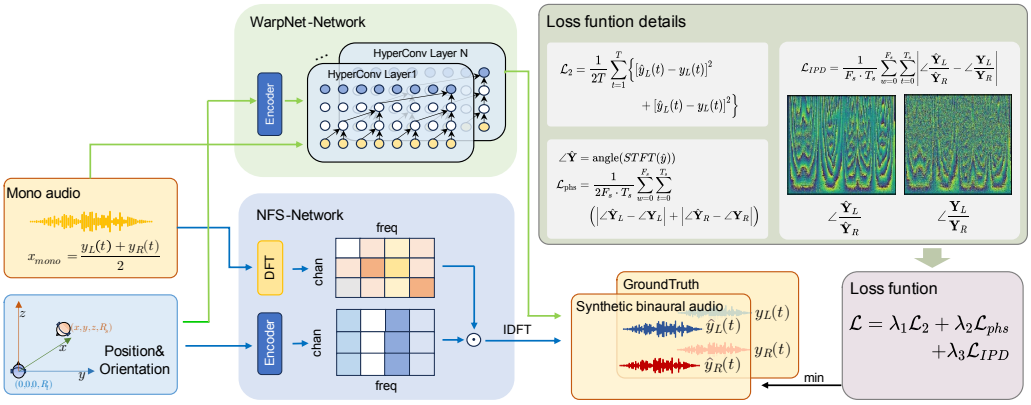


Figure 2. The overall framework of DIFFBAS. Our method takes mixed monaural audio (the mean value of binaural audio) and sound source location information as input. We adopt a joint loss for supervision, where the IPD is designed to enhance the recovery of interaural differences. We apply our method on two open-source backbone networks, including WarpNet [15], which encodes position information as time-domain convolution weights, and NFS [16], which encodes position information as frequency-domain masks.

The first term is the $\mathcal{L}_2$ which is defined as

$$\mathcal{L}_2 = \frac{1}{2T}\sum_{t=1}^{T}\left\{[\hat{\boldsymbol{y}}_L(t) - \boldsymbol{y}_L(t)]^2 + [\hat{\boldsymbol{y}}_R(t) - \boldsymbol{y}_R(t)]^2\right\} \tag{3}$$

where T is the audio length, $\boldsymbol{y}_L$ and $\boldsymbol{y}_R$ are the ground truth signals from left and right channels. $\hat{\boldsymbol{y}}_L$ and $\hat{\boldsymbol{y}}_R$ are the synthesized binaural audio. $\mathcal{L}_2$ guarantees the synthesized left and right ear audio is close to the ground truth in the overall waveform.

The second term $\mathcal{L}_{phs}$ is the phase spectrum difference between the synthesized audios and ground truth, which is defined as follows

$$\mathcal{L}_{phs} = \frac{1}{2F_s \cdot T_s}\sum_{w=0}^{F_s}\sum_{t=0}^{T_s}\left(\left|\angle\hat{\mathbf{Y}}_L - \angle\mathbf{Y}_L\right| + \left|\angle\hat{\mathbf{Y}}_R - \angle\mathbf{Y}_R\right|\right) \tag{4}$$

where $\mathbf{Y}_L$ and $\mathbf{Y}_R$ are the results of Short Term Fourier Transform (STFT) on $\boldsymbol{y}_L$ and $\boldsymbol{y}_R$, $\hat{\mathbf{Y}}_L$ and $\hat{\mathbf{Y}}_R$ are the results of STFT on $\hat{\boldsymbol{y}}_L$ and $\hat{\boldsymbol{y}}_R$, and F_s and T_s are the frequency and time resolutions of STFT. $\angle$ means performing phase calculations on the spectrogram to obtain the phase spectrum. $\mathcal{L}_{phs}$ helps to better model sound time delay.

The third term is a newly proposed loss, namely $\mathcal{L}_{IPD}$, which is used to supervise model training from the perspective of the interaural phase difference to retain better spatial information. It has been explained in more detail in Sections 2.3 and 2.4.

We want to stress that during the training process, the input of DIFFBAS $\boldsymbol{x}_{mono}$ is approximated by the average signals of the collected signals from the left and right ears, i.e., $\boldsymbol{x}_{mono} = \frac{\boldsymbol{y}_L(n)+\boldsymbol{y}_R(n)}{2}$. Such approximation is reasonable, because Leng et al. [17] pointed out that the signals collected at the front of the head can be represented by the average signals from the left and right ear channels. In addition, Gao et al. [20] also used the average signals from the left and right channels as the approximation of the mono audio collected near the head. Experimental results have demonstrated that by replacing the training input signals and incorporating the new IPD loss term, the synthesized binaural audio exhibits more accurate spatial cues to the listener.

2.3. Definition of IPD

The human's ability to localize sound sources predominantly relies on interaural time differences(ITD) between the signals received by the left and right ears. Different incident angles of sound waves will lead to different time differences, allowing humans to perceive relevant spatial information. Taking the horizontal plane as an example, as shown in Figure 3a, if the role of the head is omitted, the ears are approximated as two points $2a$ apart (a is approximately the radius of the head). For plane sound waves incident in direction θ (can be regarded as generated by a point sound source at infinity), ITD can be calculated as

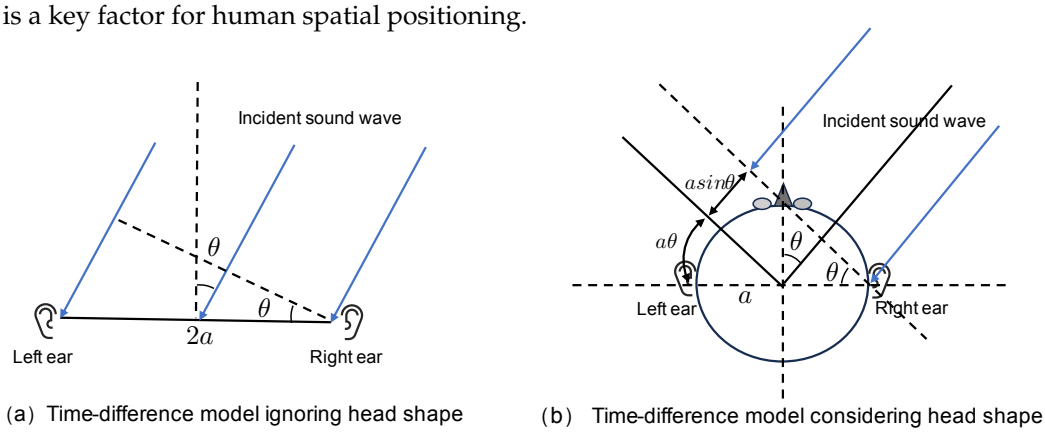
$$ITD(\theta) = \frac{2a}{c} sin\theta \tag{5}$$

where c is the speed of sound, and if ITD is greater than 0, it means it reaches the right ear first. If ITD is less than 0, it means it reaches the left ear first.

If we further consider the shape of the head, as shown in Figure 3b, the head is approximated as a sphere with a radius a, and the ears are approximated as two opposite points on the spherical surface. The formula for the binaural time difference is approximately

$$ITD(\theta) = \frac{a}{c}(sin\theta + \theta) \tag{6}$$

From Equations (5) and (6), it can be seen that ITD is related to the azimuth angle and is a key factor for human spatial positioning.



(a) Time-difference model ignoring head shape (b) Time-difference model considering head shape

Figure 3. Schematic diagram of horizontal ITD (interaural time difference) calculation.

The moment a sudden or transient sound reaches the basilar membrane of the human ear, the hair cells in the ear become excited and sensitive, so that the human ear can keenly distinguish the difference in arrival time [21]. However, due to the human ear's adaptability to sound, a continuous audio sequence may cause subsequent sounds to arrive at the same ear and mask previous sounds, thus weakening the significance of temporal differences. In this case, the human ear mainly infers the interaural time difference by analyzing the

interaural phase difference (IPD) of the sounds in the left and right ears [22]. According to Equation (6), IPD can be defined as

$$IPD(\theta, \omega) = 2\pi\omega\frac{a}{c}(sin\theta + \theta) \tag{7}$$

where ω is the frequency of the sound wave. The human auditory system realizes the perception of sound direction by comprehensively processing the information of perceived sound frequency and interaural phase difference. Therefore, in the digital signal of binaural audio, we can use STFT to analyze the phase spectrum difference of short-term audio to quantify the spatial information contained in it.

2.4. IPD Calculation of Binaural Audio

Given binaural audio $\boldsymbol{y}(t) = [\boldsymbol{y}_L(t), \boldsymbol{y}_R(t)]$, the corresponding spectrograms $[\mathbf{Y}_L, \mathbf{Y}_R] \in C^{F_s \times T_s}$ can be computed as

$$\begin{aligned} \mathbf{Y}_L(\omega, t) &= \mathbf{A}_L(\omega, t) \cdot e^{j\boldsymbol{\phi}_L(\omega,t)} \\ \mathbf{Y}_R(\omega, t) &= \mathbf{A}_R(\omega, t) \cdot e^{j\boldsymbol{\phi}_R(\omega,t)} \end{aligned} \tag{8}$$

where $\mathbf{A}_L(\omega, t)$ and $\mathbf{A}_R(\omega, t)$ denote the amplitudes of the left and right ear signals at time t and frequency ω, $\boldsymbol{\phi}_L(\omega, t)$ and $\boldsymbol{\phi}_R(\omega, t)$ represent their initial phases. The IPD $\Delta\boldsymbol{\phi}$ between two ears can be computed as

$$\Delta\boldsymbol{\phi}(\omega, t) = \boldsymbol{\phi}_L(\omega, t) - \boldsymbol{\phi}_R(\omega, t) \in [-2\pi, 2\pi] \tag{9}$$

We need to limit the phase difference within one cycle. Leveraging Euler's theorem and Equation (8), the spectral disparity between the left and right ears can be computed as

$$\frac{\mathbf{Y}_L(\omega, t)}{\mathbf{Y}_R(\omega, t)} = \frac{\mathbf{A}_L(\omega, t)}{\mathbf{A}_R(\omega, t)} \cdot (\cos\Delta\boldsymbol{\phi}(\omega, t) + j\sin\Delta\boldsymbol{\phi}(\omega, t)) \tag{10}$$

Furthermore, $\Delta\boldsymbol{\phi}$ can be obtained by calculating the phase angle of this complex matrix

$$\Delta\boldsymbol{\phi}(\omega, t) = \angle\left(\frac{\mathbf{Y}_L(\omega, t)}{\mathbf{Y}_R(\omega, t)}\right) \in [-\pi, \pi] \tag{11}$$

where $\angle$ is the depression angle for calculating the complex matrix, and $\Delta\boldsymbol{\phi}$ represents the initial phase difference between the left and right ears that we ultimately calculate.

Therefore, for an arbitrary binaural audio $\boldsymbol{y}(t) = [\boldsymbol{y}_L(t), \boldsymbol{y}_R(t)]$, the IPD can be represented by a complex matrix

$$IPD(\boldsymbol{y}(t)) = \angle\left(\frac{\mathbf{Y}_L}{\mathbf{Y}_R}\right) \in C^{F_s \times T_s} \tag{12}$$

where F_s and T_s are the frequency and time resolutions of STFT. Using this formula, the phase difference of binaural signal $\boldsymbol{y}_L(t)$ at a specific time and frequency after STFT transformation can be calculated. Combined with Equation (7), it can be shown that the phase difference contains the orientation information of the sound source at a certain moment.

2.5. Overall Loss Function

In addition to the ITD mentioned above, the human ear's ability to localize sound sources also depends on the ILD (interaural level difference) and the spectral cues generated by the pinna. Under ideal circumstances, by accurately modeling the time delays and energy attenuation during the sound source propagating to both ears, the synthesized binaural audio can have ITD, ILD and spectral cues consistent with the real audio, thereby achieving higher sound localization accuracy. Richard et al. [15] underscored the significance of accurate phase estimation for binaural audio by introducing a phase loss term $\mathcal{L}_{phs}$. They also revealed that the $\mathcal{L}_2$ loss on waveforms combined with $\mathcal{L}_{phs}$ loss

can effectively supervise the time delays and energy reductions in sound propagation. However, these two losses are only used to supervise the training of propagation models for each ear. Since these two losses cannot be completely optimized to 0, even if the loss is well optimized, it cannot ensure that the difference between the two ears is fully restored. Therefore, additional loss is needed to compensate for this deficiency to intuitively compute disparity cues between ears

According to Sections 2.3 and 2.4, we infer that the IPD of binaural audio contains spatial information and can be quantified by calculating the phase spectral difference. Ideally, the predicted IPD of binaural audio should be highly consistent with the actual IPD of binaural audio to obtain a consistent spatial experience. Therefore, we introduce an additional loss named $\mathcal{L}_{IPD}$ to better supervise the model to learn more accurate binaural differences. According to Equation (12), $\mathcal{L}_{IPD}$ can be defined as

$$\mathcal{L}_{IPD} = \frac{1}{F_s \cdot T_s} \sum_{w=0}^{F_s} \sum_{t=0}^{T_s} |IPD(\hat{\boldsymbol{y}}(t)) - IPD(\boldsymbol{y}(t))| \tag{13}$$

where $\hat{\boldsymbol{y}}(t)$ is the predicted binaural audio signal and $\boldsymbol{y}(t)$ is the actual binaural audio, and F_s and T_s are the frequency and time resolutions of STFT. According to Equation (7), we calculated that the binaural phase difference may exceed 2π at high frequencies, causing confusion in human positioning. Therefore, it is necessary to focus on the low-frequency range and filter the frequency distribution with lower energy when calculating IPD, since real sound is usually concentrated at specific frequencies with high energy.

The backbone network is trained by minimizing the following loss:

$$\mathcal{L}(\hat{\boldsymbol{y}}(t), \boldsymbol{y}(t)) = \lambda_1 \mathcal{L}_2 + \lambda_2 \mathcal{L}_{phs} + \lambda_3 \mathcal{L}_{IPD} \tag{14}$$

where λ_1, λ_2 and λ_3 are the hyperparameters during training.

3. Experiments and Results

3.1. Experiment Settings

3.1.1. Dataset

The binaural speech benchmark dataset is used in all the validation experiments. This is the only paired binaural audio dataset currently collected in dynamic scenes. It comprises approximately 2 h of monaural-binaural paired audio data with a sampling rate of 48 kHz. Data set characteristics and training, validation and testing splitting are shown in Table 1. The binaural audio was recorded using the KEMAR anthropomorphic model in a semi-anechoic room, featuring source speech from eight distinct subjects (different speakers). The 'acoustically adapted room' is an acoustically treated room that has partially anechoic walls and floors, but does not achieve the level of complete anechoics. Figure 4 shows the detailed dataset capture layout. The KEMAR is located in the center of the room and is set as the coordinate origin in OptiTrack's motion capture settings. Different speakers can walk within a 1.5-m radius circle around KEMAR and move their heads up and down to simulate the area coverage of a normal social conversation and the OptiTrack system tracked the position and orientation of the speaker at a frequency of 120 Hz, synchronized with the audio timestamps. The gender ratio of different speakers is the same, and the collection time of each of them is about 16 min. We adhere to the initial training-validation-testing splits as Richard et al. [15] and Lee et al. [16] for a fair comparison.

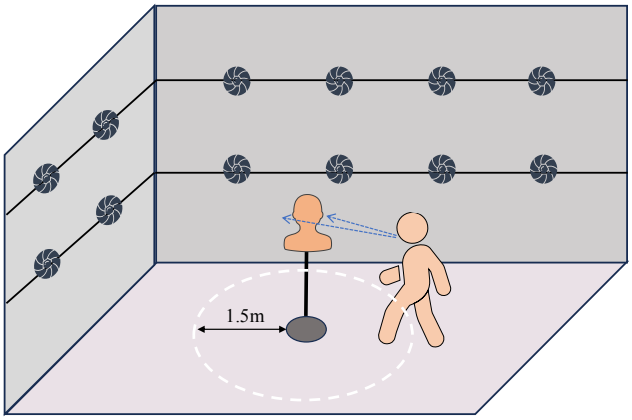


Figure 4. Side view of capture layout. The speaker speech is recorded with a head-mounted microphone and the binaural audio is captured with binaural microphones on the ears of the mannequin. Mannequin and participant positions are tracked with OptiTrack cameras mounted on the walls of the room.

Table 1. Dataset characteristics and training, validation and testing splits. The unit of time length is minutes: seconds.

Speaker	Listener	Gender	Room Characteristics	Time Length	
				Train	Test
subject1	KEMAR	male	acoustically adapted room	13:17	1:55
subject2		female		14:24	1:59
subject3		female		13:26	1:55
subject4		female		13:12	1:54
subject5		male		13:29	1:59
subject6		male		13:16	1:58
subject7		male		13:51	1:57
subject8		female		13:14	1:49
validation		male		1:24	

3.1.2. Baselines

The proposed framework is denoted as DIFFBAS. In order to prove the superiority of this method in modeling binaural differences among neural network-based methods, we select two baseline models with obvious differences in network structure. The first one is **WarpNet** [15], which uses the idea of dilated convolution in WaveNet [23] and mainly encodes position information as time-domain convolution weights. The second is **NFS** [16], which uses the Fourier shift theorem [24] to model the time delay and energy attenuation of the sound propagation process in the frequency domain space. Both networks use geometric warping modules in the early stages to roughly model the time warp from the sound source to the listener's left and right ears while respecting physical properties.

3.1.3. Evaluation Method

In our experiments, IPD is introduced as an additional evaluation metric for assessing the spatial performance of the synthesized audio in addition to the previously used metrics, i.e., $\mathcal{L}_2$, amplitude spectrum, and $\mathcal{L}_{phs}$ [15–17]. Therefore, four metrics are adopted to evaluate the models' performances:

- $\mathcal{L}_2$, which represents the mean square error between the synthesized binaural audio and the golden binaural recording.
- Amp, which measures the mean absolute error between the synthesized binaural audio and the binaural recording on the amplitude after performing STFT on the wave.
- $\mathcal{L}_{phs}$, which captures the mean absolute error between the synthesized binaural audio and the binaural recording on the phase after performing STFT on the wave.

- $\mathcal{L}_{IPD}$, which quantifies the mean absolute error on IPD between the synthesized binaural audio and the binaural recording. The smaller the metric, the closer the spatial information carried by the synthesized binaural audio is to the binaural recording.

3.1.4. Training Details

The backbone network of DIFFBAS utilizes two open-source BAS systems WarpNet and NFS separately. Their geometric warp module used for early delay modeling is removed. Therefore, two versions of DIFFBAS were implemented, i.e., DIFFBAS + WarpNet and DIFFBAS + NFS. Table 2 shows the configurations of these models used in the experiments. They all use the Adam optimizer to optimize the model and use the same learning rate, with a value of 0.001. However, the learning rate decay strategy is slightly different. DIFFBAS + WarpNet will adaptively decay at a decay rate of 0.5 according to gradient changes, while DIFFBAS + NFS will decay at a decay rate of 0.1 after each epoch training. For performance comparison, WarpNet and NFS were also trained and tested on the same dataset. The training details are as follows. WarpNet and DIFFBAS + WarpNet were trained using two NVIDIA 3090Ti GPUs with a batch size of 24 for 100 epochs, and NFS and DIFFBAS + NFS were trained using two NVIDIA 3090Ti GPUs with a batch size of 16 for 100 epochs.

Table 2. Configuration of two binaural audio synthesis models.

		DIFFBAS + WarpNet	**DIFFBAS + NFS**
loss function hyperparameters	λ_1	1	1
	λ_2	0.01	0.001
	λ_3	0.01	0.001
Optimizer	activation	Tanh	Relu
	type	Adam	Adam
	learning_rate	1.0×10^{-3}	1.0×10^{-3}
	learning_rate_decay	0.5	0.1
	weight_decay	0	0

3.2. Experiment Results

3.2.1. Loss Evaluation

In order to verify the effectiveness of our added IPD loss in modeling binaural differences, we train WarpNet with $\mathcal{L}_2 + \mathcal{L}_{phs}$ losses and with the losses proposed in Equation (14). The results are shown in Figure 5, which illustrates the development of phase error and interaural phase difference (IPD) error as L2 decreases during training. The model trained with $\mathcal{L}_2 + \mathcal{L}_{phs}$ (see Figure 5a) shows this phase is actively optimized during training, but IPD is optimized slowly at first. In Figure 5b, it can be seen that adding in $\mathcal{L}_{IPD}$ loss, the model of IPD error decreases more smoothly and reaches lower value under consistent changes of $\mathcal{L}_{phs}$ and $\mathcal{L}_2$ losses. This indicates that the supervision of $\mathcal{L}_{IPD}$ does not conflict with other losses, and can effectively supplement the supervision of the binaural phase difference. Various audio losses have been proposed over time, ranging from optimizing the power spectrum only and copying the input's phase [25,26] over a multiscale STFT loss for high frequency and high time resolution [27] to optimization of the scale-invariant signal to distortion ratio (Si-SDR) [28–30]. In order to better illustrate the superiority of IPD loss in BAS tasks, we make a full comparison with different loss functions in Table 3. As can be seen from the results, all the loss functions mentioned above fail to predict an accurate phase. Although people later use $\mathcal{L}_2 + \mathcal{L}_{phs}$ losses to greatly reduce the phase error [15–17], these two losses are only used to supervise the training of propagation models for each ear. They cannot guarantee the correct relative differences between signals in two ears. Overall, our additionally introduced IPD loss preserves accurate phase estimation and far outperforms other methods in terms of error in interaural phase difference.

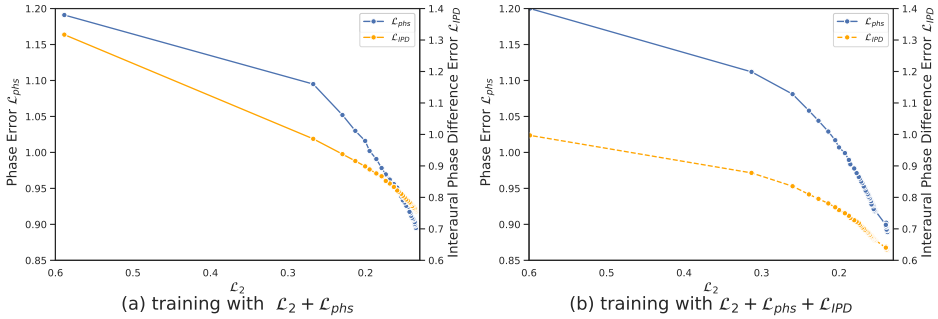


Figure 5. Development of phase error ($\mathcal{L}_{phs}$) and interaural phase difference error ($\mathcal{L}_{IPD}$) as the $\mathcal{L}_2$-loss decreases during training on WarpNet [15] using different loss functions. The blue line represents the changing trend of $\mathcal{L}_{phs}$, and the orange line represents the trend of $\mathcal{L}_{IPD}$.

Table 3. Comparison of commonly used losses for audio modeling to our proposed loss. The ↓ symbol indicates that the lower the metric, the better the effect. The optimal value of each metric is shown in bold.

	$\mathcal{L}_2 \times 10^3 \downarrow$	$Amp \downarrow$	$\mathcal{L}_{phs} \downarrow$	$\mathcal{L}_{IPD} \downarrow$
power spectrum+phase copy [25]	1.276	0.048	1.563	-
multiscale STFT [15]	2.279	0.043	1.996	-
Si-SDR [28]	0.798	0.222	1.507	-
$\mathcal{L}_2$	0.146	**0.033**	0.894	1.217
$\mathcal{L}_2 + \mathcal{L}_{phs}$	0.145	0.036	**0.815**	1.064
$\mathcal{L}_2 + \mathcal{L}_{phs} + \mathcal{L}_{IPD}$ (Equation (14))	**0.142**	0.036	0.819	**0.967**

3.2.2. Quantitative Evaluation

Comparative experiment

We compared the quality of binaural audio synthesized by the following systems:

- **Geo-Warp**, which performs time translation and amplitude alteration on the left and right ear signals based on the geometric information and attenuation formula of the distance from the sound source to both ears.
- **WarpNet**, which transforms the time-domain waveform to obtain binaural audio through geometric warp and neural warp.
- **NFS**, which generates frequency domain masks based on position information to obtain binaural audio.
- **DIFFBAS + WarpNet**, the backbone network of DIFFBAS is WarpNet.
- **DIFFBAS + NFS**, the backbone network of DIFFBAS is NFS.

To demonstrate the deficiency of previously used metrics, we add two dummy BAS systems for comparison:

- **Mono-Mono**, which produces binaural audio simply by copying the mono audio of the sound source to the left and right channels.
- **Avg-Avg**, which produces binaural audio by copying the average signals of left and right channels.

The results of the comparative experiments are summarized in Table 4.

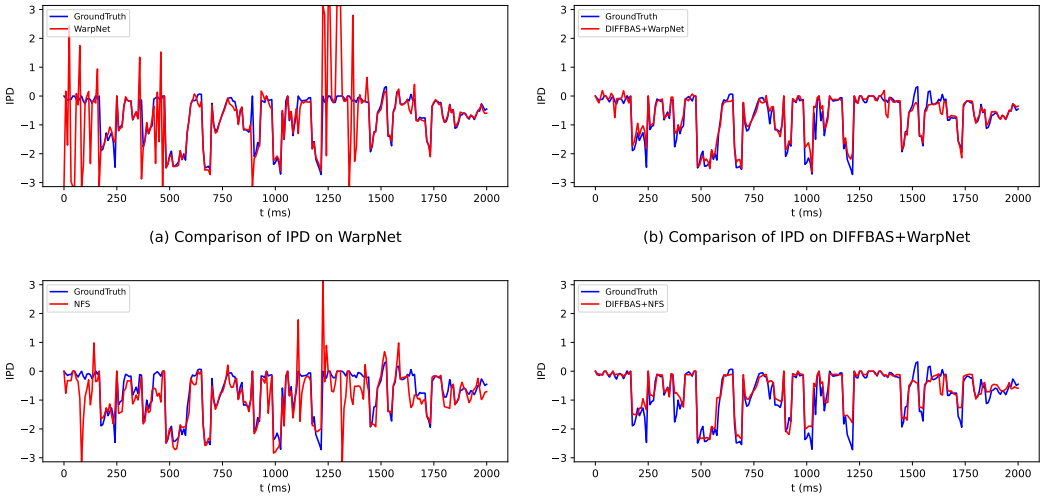
It can be seen that DNN-based methods are better than Geo-Warp, which shows that rough time translation and amplitude alteration based only on geometric information cannot model accurate and real binaural audio, and can only retain fuzzy spatial information. In terms of binaural difference recovery, compared with WarpNet, the IPD error of DIFFBAS + WarpNet is greatly reduced from 1.064 to 0.834. Similarly, the IPD error of DIFFBAS + NFS is also reduced from 1.250 to 1.083. Therefore, DIFFBAS can significantly improve the accuracy of the synthesized audio referring to its interaural phase

differences. At the same time, the losses of other metrics are also reduced, which means that the introduction of IPD loss will not impact the supervision of other losses.

Table 4. Performances of six BAS models on the four evaluation metrics. We reproduced the results using the official codes for WarpNet and NFS. The ↓ symbol indicates that the lower the metric, the better the effect. The optimal value of each metric is marked with an asterisk $(\cdot)^*$ and bolded, and the suboptimal value is marked with a $(\cdot)^\dagger$ and bolded.

Model	$\mathcal{L}_2 \times 10^3 \downarrow$	$Amp \downarrow$	$\mathcal{L}_{phs} \downarrow$	$\mathcal{L}_{IPD} \downarrow$
Mono-Mono	1.340	0.063	1.564	1.472
Avg-Avg	0.200	0.031	0.598	1.472
Geo-Warp	0.408	0.052	1.156	1.414
WarpNet [15]	0.145	0.036	0.815	**1.064** †
DIFFBAS + WarpNet	**0.033** *	**0.018** *	**0.358** *	**0.834** *
NFS [16]	0.163	0.040	0.869	1.250
DIFFBAS + NFS	**0.049** †	**0.018** *	**0.431** †	1.083

In addition, the human auditory system is most sensitive to the phase difference information of low-frequency signals, especially for the high-energy parts. Therefore, we randomly selected a 2-second audio segment and performed STFT, screened out the low-frequency parts, and calculated the frequencies with peak energy at different time points. We compute the IPD of synthesized audio at the frequency with the peak energy at each time. Then, the IPD values of synthesized audio are compared with the IPD values of the ground truth audio, the results of which have been shown in Figure 6. It can be seen that DIFFBAS can synthesize binaural audio whose IPD is almost consistent with the ground truth at the frequency with the highest energy.



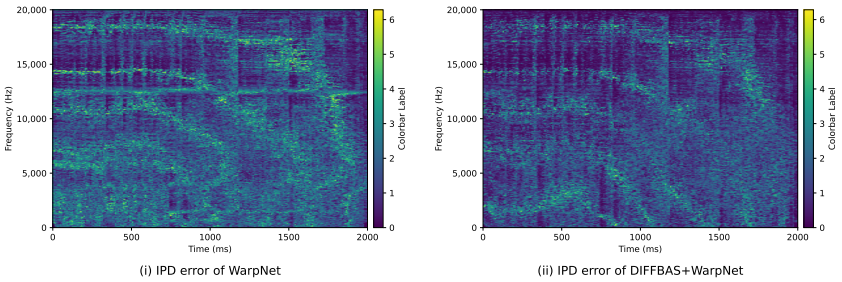
(a) Comparison of IPD on WarpNet (b) Comparison of IPD on DIFFBAS+WarpNet

(c) Comparison of IPD on NFS (d) Comparison of IPD on DIFFBAS+NFS

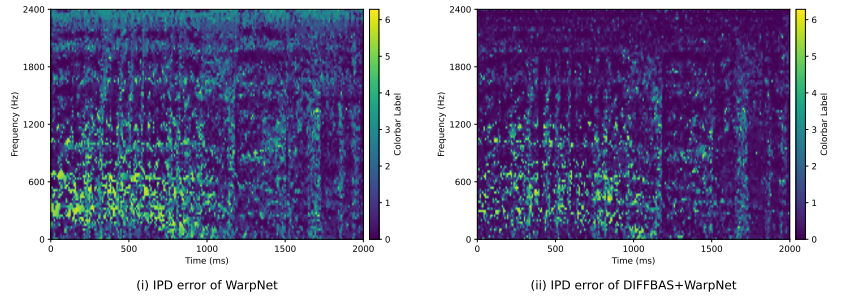
Figure 6. Comparison of IPD(interaural phase difference) at frequencies with peak energy over a two-second period compared to recording audio. The range of IPD is $[-\pi, \pi]$. The blue line represents the changing trend of recording audio, and the red line represents the changing trend of binaural audio synthesized by the model.

In addition, Figure 7 shows the IPD error heatmaps at different frequencies for the same four models as above during the same time period. The color of each position represents the model's IPD error in the corresponding frequency/time pair. A darker color indicates a smaller IPD error. It can be seen that Figure 7a–d(ii) are darker than Figure 7a–d(i), which indicates that DIFFBAS can reduce IPD errors in the entire time-frequency domain. People usually use IPD at lower frequencies for spatial localization because high frequencies

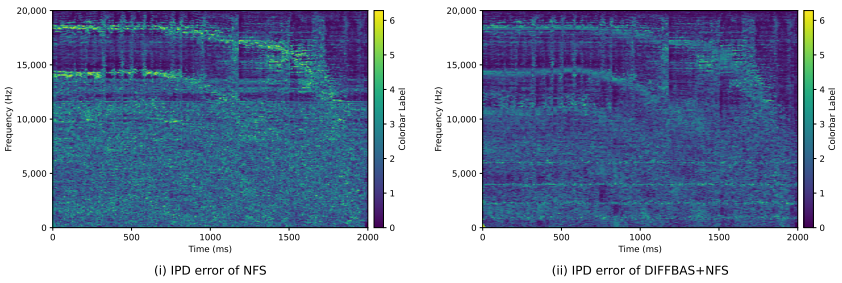
can cause perceptual confusion. Further observations of Figure 7b,d shows that in the low-frequency band of 0–2000 Hz, DIFFBAS can effectively reduce IPD errors and outliers, which is more in line with human perception characteristics.



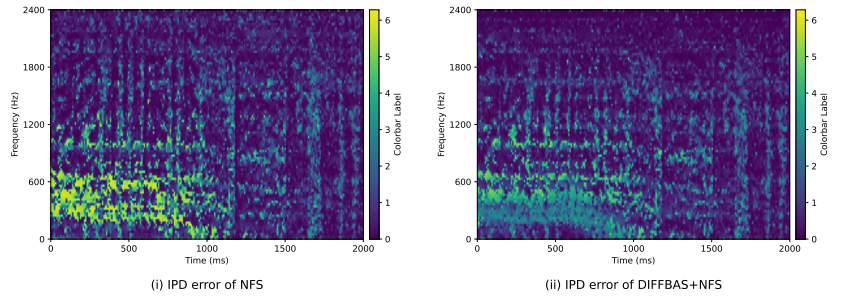
(**a**)



(**b**)



(**c**)



(**d**)

Figure 7. IPD error heatmaps of four BAS models in the time-frequency domain. The darker the point, the lower IPD error of the model. (**a**) IPD error heatmap across the entire frequency range. (i) IPD error of WarpNet. (ii) IPD error of DIFFBAS + WarpNet. (**b**) Heat map of IPD error in low-frequency range. (i) IPD error of WarpNet. (ii) IPD error of DIFFBAS + WarpNet. (**c**) IPD error heatmap across the entire frequency range. (i) IPD error of NFS. (ii) IPD error of DIFFBAS + NFS. (**d**) Heat map of IPD error in low-frequency range. (i) IPD error of NFS. (ii) IPD error of DIFFBAS + NFS.

Other evaluation metrics, such as $\mathcal{L}_2$, Amp, and $\mathcal{L}_{phs}$, cannot be used alone to evaluate the performance of BAS systems. Take the dummy Avg-Avg system as an example. Its

performance on the first three metrics has been significantly improved. In Table 4, compared with WarpNet, the values of Amp and $\mathcal{L}_{phs}$ have been reduced from 0.036 and 0.815 to 0.031 and 0.598, respectively. However, the synthesized audio has no spatial information at all because its left and right channel signals are always the same and listeners cannot use it to localize the sound source. However, the IPD metric can measure such deficiency of spatial information. The IPD error of the Avg-Avg model is 1.472, which is much larger than the IPD error of WarpNet 1.064. As a comparison, the IPD error of DIFFBAS + WarpNet is 0.834, which means the binaural audio synthesized by DIFFBAS + WarpNet provides more accurate spatial information. Similar results are also found when comparing Avg-Avg, NFS, and DIFFBAS + NFS models.

Ablation experiment

We conducted two ablation experiments to evaluate the impact of the newly added IPD loss and different mono audio inputs on the model performances separately, the results of which have been shown in Table 5.

Table 5. Comparison of binaural audio synthesis quality under different modules. The ↓ symbol indicates that the lower the metric, the better the effect. The optimal value of each metric is shown in bold.

	WarpNet [15]				**NFS [16]**			
Condition	$\mathcal{L}_2 \times 10^3 \downarrow$	$Amp \downarrow$	$\mathcal{L}_{phs} \downarrow$	$\mathcal{L}_{IPD} \downarrow$	$\mathcal{L}_2 \times 10^3 \downarrow$	$Amp \downarrow$	$\mathcal{L}_{phs} \downarrow$	$\mathcal{L}_{IPD} \downarrow$
Baseline	0.145	0.036	0.815	1.064	0.163	0.040	0.869	1.250
x_{source}+$\mathcal{L}_{IPD}$	0.142	0.036	0.819	0.967	0.177	0.047	0.824	1.194
x_{mono}-$\mathcal{L}_{IPD}$	0.033	0.018	0.358	0.865	0.049	0.019	0.434	1.087
DIFFBAS	**0.033**	**0.018**	**0.358**	**0.834**	**0.049**	**0.018**	**0.431**	**1.083**

In Table 5, x_{source}+$\mathcal{L}_{IPD}$ represents the model using mono audio from the source as the input and using the IPD loss in the training process. x_{mono}-$\mathcal{L}_{IPD}$ means using average signals from the two channels as the input and removing the IPD loss from the training process. It can be seen that, as an additional supervision loss, IPD can significantly improve the modeling ability of binaural differences with almost no reduction in other metrics. For example, using WarpNet, the IPD error of x_{source}+$\mathcal{L}_{IPD}$ is reduced from 1.064 to 0.967 compared with the baseline. For x_{mono}-$\mathcal{L}_{IPD}$, its IPD error is also reduced from 1.064 to 0.865 compared with the baseline, which indicates that using the average of the binaural audio as the training inputs, models are able to model binaural differences in the sound propagation more accurately. It demonstrates the effectiveness of specific path modeling.

4. Discussion

DNN-based modeling technology has been proven effective for binaural audio modeling and captures nonlinear effects in dynamic sound propagation that traditional DSP methods cannot handle. Currently, there are three neural network types used for BAS, including WarpNet [15], NFS [16] and BinauralGrad [17], which have shown that synthesized binaural audio is superior to traditional DSP methods in terms of clarity and realism in subjective and objective evaluations. However, the potential of neural networks in modeling binaural audio spatial information has not yet been fully explored. In order to further improve the ability of the neural network-based BAS method to synthesize binaural audio spatial information, we propose DIFFBAS. This method avoids over-fitting problems caused by the diversity of environments and sound sources by focusing on modeling sound propagation paths, thereby improving the accuracy of synthesizing binaural audio. Combined with the characteristics of human auditory perception, we introduced the IPD loss as a training loss for the first time in related tasks to effectively supervise the learning of interaural phase difference information, which is highly related to the spatial information in binaural audio.

Our method has significant improvements over existing baseline models. In terms of binaural difference recovery, compared with WarpNet, the IPD error of DIFFBAS + WarpNet

is significantly reduced from 1.064 to 0.834. Similarly, the IPD error of DIFFBAS + NFS is also reduced from 1.250 to 1.083. These results not only prove the effectiveness of our method but also show that it is suitable for different backbone networks, providing new research ideas for DNN-based binaural audio synthesis tasks.

Furthermore, our study is also limited by the scarcity of dynamic continuous binaural audio datasets, which makes it difficult for the model to account for the generalization ability of audio rendering beyond the frequency domain of the training set. Additionally, we are unable to objectively assess sounds distributed across different frequency ranges. However, we have tested different types of sounds such as music, dog barking, and car horns into the trained model. The synthesized binaural audio can obtain a sense of space consistent with the input position information in the auditory experience. This shows to a certain extent that our method also has a certain generalization ability when dealing with other types of sounds.

5. Conclusions

In this work, we propose a novel DNN-based BAS method, DIFFBAS. By focusing on modeling the specific path of the sound propagation process, DIFFBAS can avoid overfitting problems caused by the diversity of environments and sound sources, and improve the accuracy of synthesized binaural audio. A new training loss, namely IPD loss, is first proposed in our method to effectively supervise the learning of interaural phase difference information, which is highly related to the spatial information contained in binaural audio. Furthermore, IPD can be used as a new objective quality assessment metric for the level of spatial information contained in the synthesized binaural audio. This provides a better quantification standard for BAS tasks. Experimental results have demonstrated that DIFFBAS performs exceptionally well across different types of backbone networks and achieves a significant improvement in binaural difference learning. Future work will focus on two aspects: first, collecting a more general binaural audio dataset to improve network training effect and generalization ability; second, exploring personalized binaural audio synthesis, using the listener's head and pinnae photometric information as an additional conditioning parameter.

Author Contributions: Supervision, Y.S.; Writing—original draft, Y.L. Writing—review & editing, Y.S. and D.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Fundamental Research Funds for the Central Universities.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available dataset is analyzed in this study. The dataset can be found here: https://github.com/facebookresearch/BinauralSpeechSynthesis/.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hendrix, C.; Barfield, W. The sense of presence within auditory virtual environments. *Presence Teleoper. Virtual Environ.* **1996**, *5*, 290–301. [CrossRef]
2. Hammershøi, D.; Møller, H. Binaural technique—Basic methods for recording, synthesis, and reproduction. In *Communication Acoustics*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 223–254.
3. Hoeg, E.R.; Gerry, L.J.; Thomsen, L.; Nilsson, N.C.; Serafin, S. Binaural sound reduces reaction time in a virtual reality search task. In Proceedings of the 2017 IEEE 3rd VR Workshop on Sonic Interactions for Virtual Environments (SIVE), Los Angeles, CA, USA, 19 March 2017; pp. 1–4.
4. Blauert, J.; Braasch, J. Binaural signal processing. In Proceedings of the 2011 17th International Conference on Digital Signal Processing (DSP), Corfu, Greece, 6–8 July 2011; pp. 1–11.
5. He, J.; Tan, E.L.; Gan, W.S. Natural sound rendering for headphones: Integration of signal processing techniques. *IEEE Signal Process. Mag.* **2015**, *32*, 100–113.
6. Zotkin, D.N.; Duraiswami, R.; Davis, L.S. Rendering localized spatial audio in a virtual auditory space. *IEEE Trans. Multimed.* **2004**, *6*, 553–564. [CrossRef]

7. Zhang, W.; Samarasinghe, P.N.; Chen, H.; Abhayapala, T.D. Surround by Sound: A Review of Spatial Audio Recording and Reproduction. *Appl. Sci.* **2017**, *7*, 532. [CrossRef]
8. Katz, B.F.G. Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. *J. Acoust. Soc. Am.* **2001**, *110*, 2440–2448. [CrossRef] [PubMed]
9. Xiao, T.; Huo Liu, Q. Finite difference computation of head-related transfer function for human hearing. *J. Acoust. Soc. Am.* **2003**, *113*, 2434–2441. [CrossRef] [PubMed]
10. Salvador, C.D.; Sakamoto, S.; Treviño, J.; Suzuki, Y. Dataset of near-distance head-related transfer functions calculated using the boundary element method. In *Audio Engineering Society Conference: 2018 AES International Conference on Spatial Reproduction-Aesthetics and Science*; Audio Engineering Society: New York, NY, USA, 2018.
11. Algazi, V.R.; Duda, R.O.; Thompson, D.M.; Avendano, C. The cipic hrtf database. In Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575), New Platz, NY, USA, 24 October 2001; pp. 99–102.
12. Li, S.; Peissig, J. Measurement of Head-Related Transfer Functions: A Review. *Appl. Sci.* **2020**, *10*, 5014. [CrossRef]
13. Brinkmann, F.; Lindau, A.; Weinzierl, S. On the Authenticity of Individual Dynamic Binaural Synthesis. *J. Acoust. Soc. Am.* **2017**, *142*, 1784–1795. [CrossRef] [PubMed]
14. Lemaire, V.; Clerot, F.; Busson, S.; Nicol, R.; Choqueuse, V. Individualized HRTFs from few measurements: A statistical learning approach. In Proceedings of the Proceedings. 2005 IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 31 July–4 August 2005; Volume 4, pp. 2041–2046.
15. Richard, A.; Markovic, D.; Gebru, I.D.; Krenn, S.; Butler, G.A.; Torre, F.; Sheikh, Y. Neural synthesis of binaural speech from mono audio. In Proceedings of the ICLR, Virtual Event, Austria, 3–7 May 2021.
16. Lee, J.W.; Lee, K. Neural Fourier Shift for Binaural Speech Rendering. In Proceedings of the ICASSP, Rhodes, Greece, 4–10 June 2023 ; pp. 1–5.
17. Leng, Y.; Chen, Z.; Guo, J.; Liu, H.; Chen, J.; Tan, X.; Mandic, D.; He, L.; Li, X.; Qin, T.; et al. Binauralgrad: A Two-stage Conditional Diffusion Probabilistic Model for Binaural Audio Synthesis. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 23689–23700.
18. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
19. Culling, J.F.; Akeroyd, M.A. Spatial hearing. In *Oxford Handbook of Auditory Science: Hearing*; OUP Oxford: Oxford, UK, 2010; pp. 123–144.
20. Gao, R.; Grauman, K. 2.5D Visual Sound. In Proceedings of the CVPR, Long Beach, CA, USA, 16–20 June 2019; pp. 324–333.
21. Blauert, J.; Jekosch, U. Sound-quality evaluation—A multi-layered problem. *Acta Acust. United Acust.* **1997**, *83*, 747–753.
22. Culling, J.F.; Lavandier, M. Binaural Unmasking and Spatial Release from Masking. In *Binaural Hearing: With 93 Illustrations*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 209–241.
23. Rethage, D.; Pons, J.; Serra, X. A WaveNet for Speech Denoising. In Proceedings of the I2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5069–5073.
24. Tancik, M.; Srinivasan, P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J.; Ng, R. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 7537–7547.
25. Zhao, H.; Gan, C.; Rouditchenko, A.; Vondrick, C.; McDermott, J.; Torralba, A. The Sound of Pixels. In Proceedings of the ECCV, Munich, Germany, 8–14 September 2018; pp. 570–586.
26. Gao, R.; Grauman, K. Co-Separating Sounds of Visual Objects. In Proceedings of the ICCV, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3879–3888.
27. Yamamoto, R.; Song, E.; Kim, J.M. Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6199–6203.
28. Le Roux, J.; Wisdom, S.; Erdogan, H.; Hershey, J.R. SDR–Half-Baked or Well Done? In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 626–630.
29. Heitkaemper, J.; Jakobeit, D.; Boeddeker, C.; Drude, L.; Haeb-Umbach, R. Demystifying TasNet: A Dissecting Approach. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6359–6363.
30. Luo, Y.; Mesgarani, N. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Trans. Audio, Speech Lang. Process.* **2019**, *27*, 1256–1266. [CrossRef] [PubMed]