SPEAR : Receiver-to-Receiver Acoustic Neural Warping Field Supplementary Material

Anonymous Author(s)
Affiliation
Address
`email`

1 Discussion on Acoustic Neural Warping Field Visualization

The acoustic neural field is represented in frequency domain, each of which in our case contains a real and imaginary one-dimensional data vector. In the main paper, we just visualize the real part due to the space limit. Here, we provide another five real/imaginary warping fields visualization in Fig. 1. From this figure, we can clearly see that our proposed framework *SPEAR* is capable of handling the warping field irregularity property.

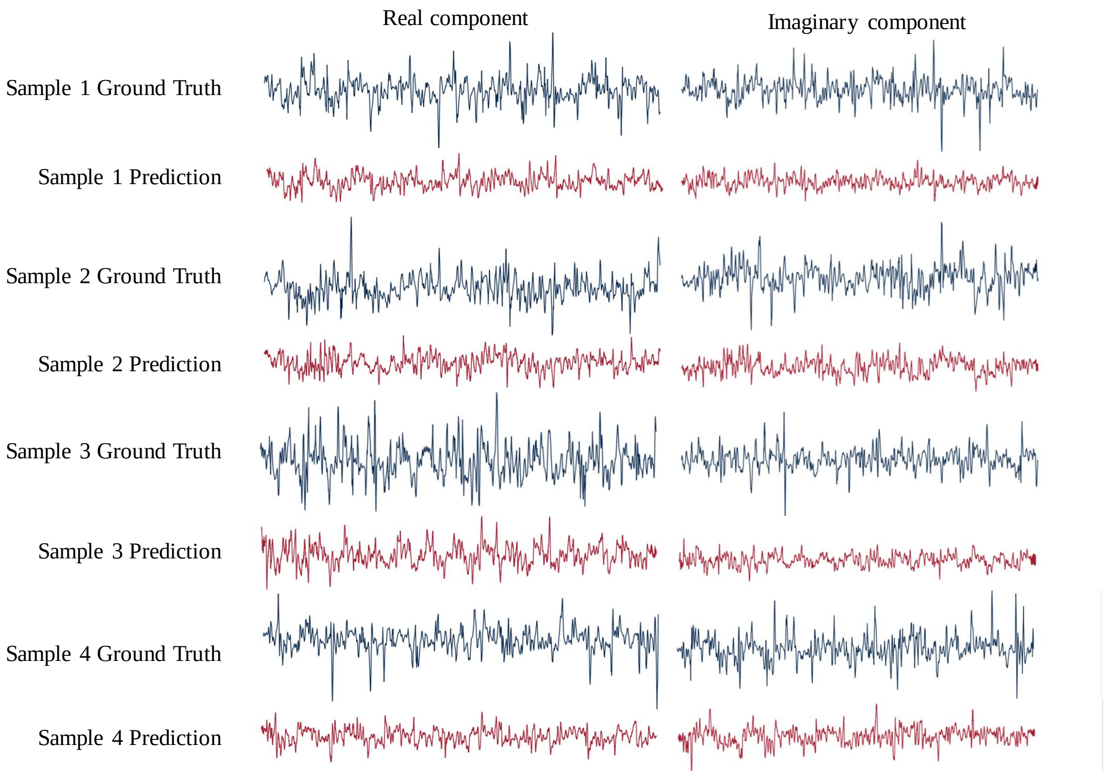

Figure 1: Visualization of real and imaginary component of the ground truth warp field in the synthetic dataset.

2 Failure Case Visualization

During the experiment process, we find all methods inevitably give failure case warping field predictions. We visualize part of some failure cases predicted by *SPEAR* on both synthetic data (Fig. 2),

Submitted to the 8th Conference on Robot Learning (CoRL 2024). Do not distribute.

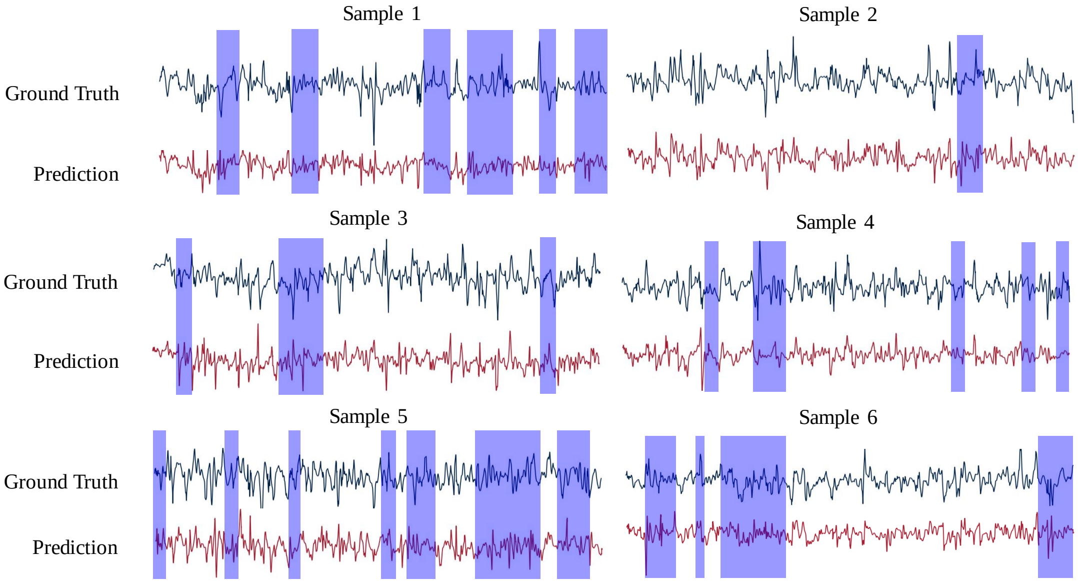

Figure 2: Examples of failure cases of SPEAR model on synthetic data. Regions with significant miss-match between the ground truth and the predicted warping field pattern are shaded in blue.

photo-realistic and real-world dataset (Fig. 3). As we discussed in the main paper, the position-sensitivity and irregularity pose challenges in the warping field prediction. We hope these failure cases will attract more investigation into this research problem.

3 Training Detail Presentation

3.1 Ground Truth Warping Field Acquirement

We obtain the ground truth by dividing the target receiver audio by the reference receiver audio in frequency domain. This division operator sometimes leads to `NaN` value or abnormally large value (> 100, when the denominator is close to zero) in the obtained warping field (see the ground truth warping field visualization in Fig. 4), resulting in the difficulty of accurately warping field learning. To handle this dilemma, we make two adjustments: first, replace `NaN` value with zero so that the whole neural network is trainable with the warping field prediction loss, which was `NaN` without the replacement. Second, clipping all warping field values to lie within $[-10, 10]$. The reason for the `clip` operation is two-fold: the abnormally large value easily allures the whole neural network to be trapped in predicting those abnormally large values, thus ignoring predicting the warping field with normal values; we further empirically verify in Fig. 4 that `clip` operation gives subtle difference in the warped target audio.

3.2 Training Configuration Presentation

We adopt AdamW optimizer [1] for training on all datasets. On the synthetic dataset, the model requires approximately 10000 epochs to converge, which takes around 11 hours on a single A10 GPU. We set the learning rate of the learnable grid feature to 1e-5, and the rest learnable parameters' learning rates to 1e-4. Using a smaller learning rate for the grid features improves the model training stability. Since the model predictions rely solely on the grid feature extracted, changes in grid feature can result in significant differences in model prediction. Therefore, setting a lower learning rate for

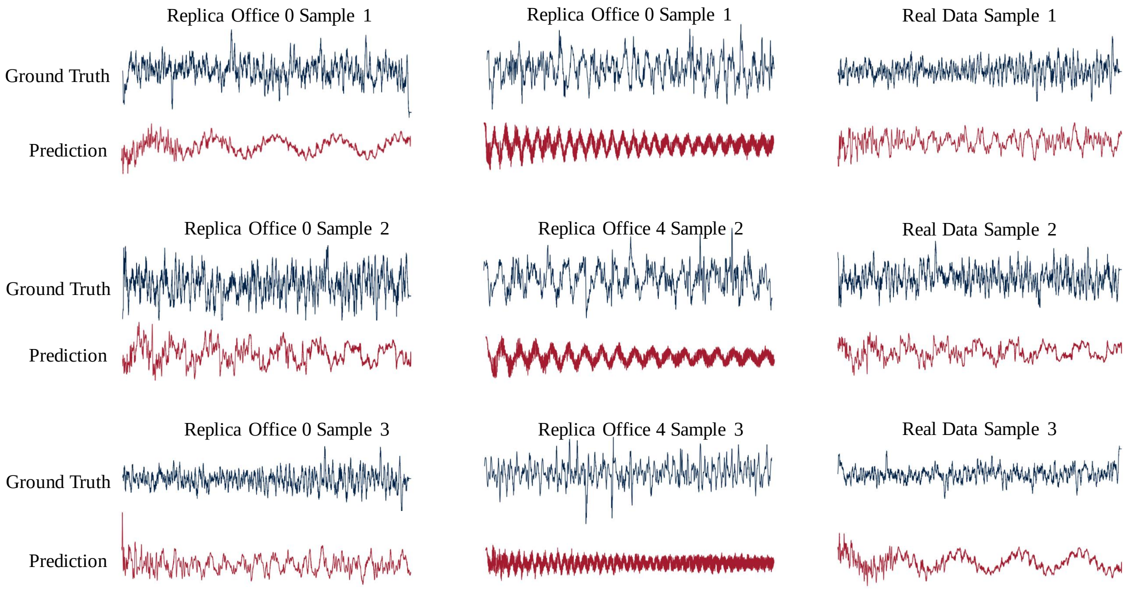

Figure 3: Failure case visualization of *SPEAR* model on both Photo-realistic and Real-world Dataset.

Method	Inf. Time	Param. Num.
NAF [2]	0.13 s	1.61 M
SPEAR	0.0182 s	27.26 M

Table 1: Model param. and Inference Time comparison. The inference time is the average of 1000 independent inferences with batch size 32 on a single A10 GPU.

the grid features prevents the model prediction from changing abruptly, and thus improves training stability.

Though our model has larger parameter size, the inference time is smaller than the NAF baseline. As shown in Tab. 1, our model has more than ten times larger parameter size than the NAF model, but inference speed is around ten times faster than the NAF model.

3.3 Data Sampling Strategy

For all three types of datasets, we construct train and test datasets by first splitting the receiver positions into two disjoint sets. The reference and target receiver positions are sampled from the same receiver position set. This means that, during both the training and testing stages, the model will not be trained/tested to predict a warping field that warps audio from a receiver position in the training set to a position in the test set, or vice versa.

In the synthetic data generation, we arrange the receivers in an 80×40 grid, with adjacent receivers spaced 0.05 meters apart. To create the test set, we select test samples so that no two test samples are adjacent in the grid. This interleaved sampling strategy ensures that each test receiver position is at least 0.05 meters away from any receiver position in the training set.

In the Photo-realistic data generation, we randomly sample 4500 and 8500 receiver positions on the two scenes' floors and select a subset of 500 receiver positions from each set as the test set. Due to the existence of furniture and other obstacles presented in the room scene, we could not employ the grid-sampling strategy used in synthetic data generation, and could only randomly sample train and test receiver positions on the scene floor.

Sample Size	SDR	MSE	PSNR	SSIM
3000	1.50	0.92	15.81	0.87
2000	1.06	0.96	15.27	0.87
1000	0.24	1.15	14.26	0.85

Table 2: Effect of different sampling density on model performance.

Layer Name	Filter Num	Output Size
Model Input: 2 position 3d coordinate: [2, 3]		
Grid Feature: concatenated 2 position feature: [1, 384]		
Transformer Encoder Input: Initial Token Representation: [43, 384]		
Transformer Layer 1 ... Transformer Layer 12	head num = 8, hidden dim = 384 ... head num = 8, hidden dim = 384	[43, 384] ... [43, 384]
Prediction Head		
Real part FC Imaginary part FC	FC, output_feat = 384 FC, output_feat = 384	[43, 384] [43, 384]
Flattern	Flattern real/imaginary token sequence. Construct complex sequence.	[16512]
Prune	Cut the sequence to 16384 length	[16384]
Mirroring	Generate the full warping field by concatenating the predicted sequence with its mirrored conjugate sequence	[32768]

Table 3: *SPEAR* Network Architecture Detail.

To train our model on the real dataset, which has a significantly smaller sampling density required by our model, we first simulate the real dataset scene using the Pyroomacoustic simulator [3] and generate 10800 samples in the scene. The receiver positions are arranged in a 90 × 120 grid, with adjacent receivers spaced 0.05 meters apart. 1000 test samples are selected from the total 10800 receiver positions. The test set sampling strategy is the same as the interleaved sampling strategy used in the synthetic data generation.

We pretrain our model on the simulated samples before fine-tuning on the real dataset. Receiver positions in the real dataset are also arranged in a grid structure, which allows us to use the interleaved sampling strategy to sample 24 test set receiver positions from the total 130 receiver positions.

3.4 Effect of Sample Size on Model Training performance

In all three datasets, reference and target audios are densely sampled from the scene. In this section, we show the necessity to sample audio at high density by training our SPEAR model on the same synthetic scene with different sampling densities. We randomly select a subset of 1000 samples and 2000 samples from the whole 3000 synthetic training data samples and show their performance metrics. Tab. 2 shows the metric of the three models. The model performance drops significantly as the sampling density decreases. In addition, we visualize the predicted warping field of the three models in Fig.5. Models trained with smaller sample sizes show significantly worse performance in higher frequency warping field prediction.

4 *SPEAR* Network Architecture

SPEAR network architecture is shown in Fig. 3, and we also provide the code in the suppementary material.

References

[1] I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[2] A. Luo, Y. Du, M. J. Tarr, J. B. Tenenbaum, A. Torralba, and C. Gan. Learning neural acoustic fields. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[3] R. Scheibler, E. Bezzam, and D. Ivan. Pyroomacoustics: A python package for audio room simulations and array processing algorithms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

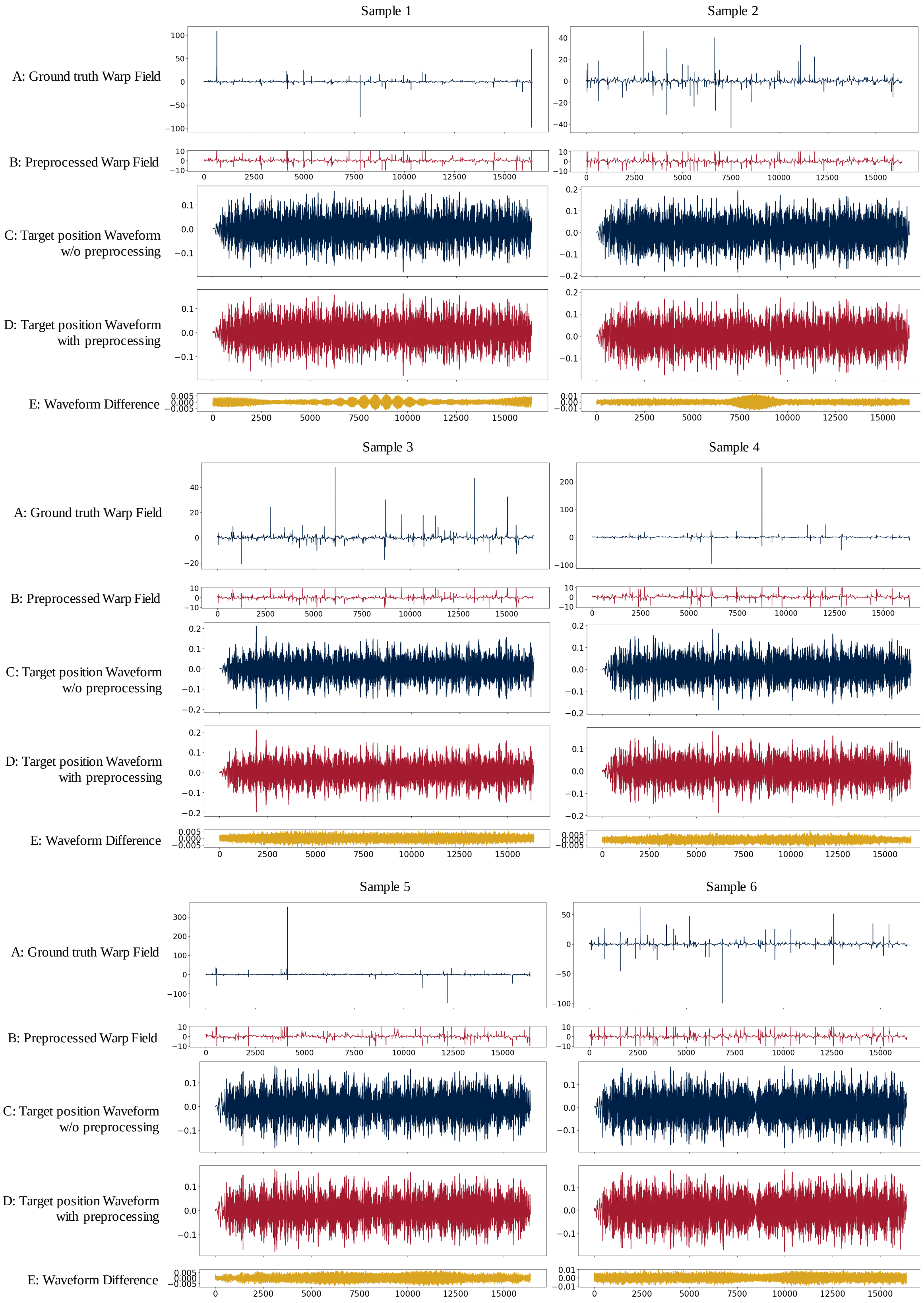

Figure 4: Visualization of the effect of preprocessing Warping field on the generated audio waveform. In plot A C, we show the warping field and waveform of the target position without warping-field preprocessing. In plot B D, we show the pre-processed warping field and the waveform at the target position after applying the preprocessed warping field. Plot E shows the difference between the two waveforms at the target position.

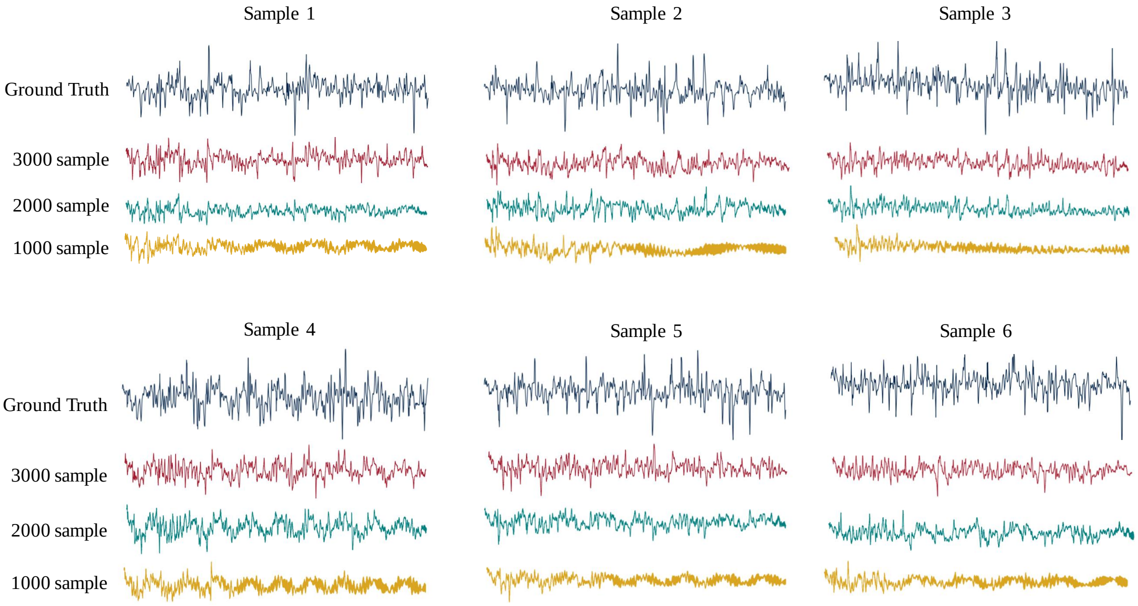

Figure 5: Predictions of models trained with different sample size.