BinauralGrad: A Two-Stage Conditional Diffusion Probabilistic Model for Binaural Audio Synthesis

Yichong Leng[1*†], **Zehua Chen**[3*], **Junliang Guo**[2], **Haohe Liu**[5], **Jiawei Chen**[6], **Xu Tan**[2]
Danilo Mandic[3], **Lei He**[4], **Xiang-Yang Li**[1], **Tao Qin**[2], **Sheng Zhao**[4], **Tie-Yan Liu**[2]
[1]University of Science and Technology of China, [2]Microsoft Research Asia
[3]Imperial College London,[4]Microsoft Azure Speech,[5]University of Surrey
[6]South China University of Technology
[1]lyc123go@mail.ustc.edu.cn,xiangyangli@ustc.edu.cn
[2]{junliangguo,xuta,taoqin,tyliu}@microsoft.com
[3]{zehua.chen18,d.mandic}@imperial.ac.uk
[4]{helei,szhao}@microsoft.com
[5]hl01486@surrey.ac.uk,[6]csjiaweichen@mail.scut.edu.cn

A Appendix

A.1 Details of MOS Test

We conducted Mean Opinion Score (MOS) test to evaluate the audio quality of our model, which is a measurement of speech quality judged by human beings and usually calculated based on a human rating service similar to Amazon Mechanical Turk. Each generated sample is rated by 15 raters on a scale from 1 (bad) to 5 (excellent) with 0.5 point increments, as shown in Table 1. In our test, each rater is required to wear headphone and be native English speaker, and then 50 samples were selected for blind evaluation and scoring. The cost for a rater rating a speech is 0.07 USD. After collecting all the evaluations, the MOS score μ is estimated by averaging the scores m_k from different testers k. In addition, we also calculated the 95% confidence intervals (CIs) for the score.

$$\mu = \frac{1}{N}\sum_{k=1}^{N} m_k$$

$$CIs = [\mu - 1.96\frac{\sigma}{N}, \mu + 1.96\frac{\sigma}{N}]$$

where σ is the standard deviation of the scores collected.

Table 1: MOS criteria.

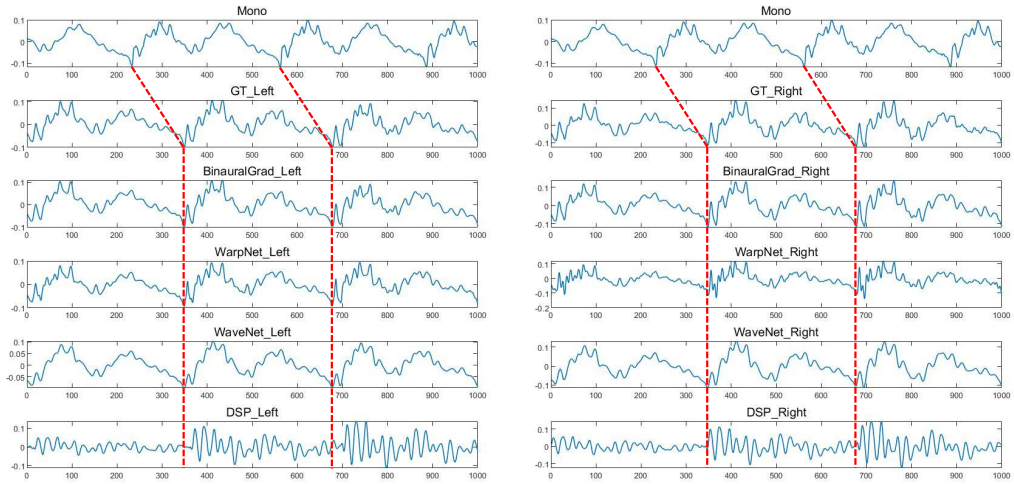
Voice Quality	Excellent	Good	Fair	Poor	Bad
Rating	5	4	3	2	1

A.2 Case Study

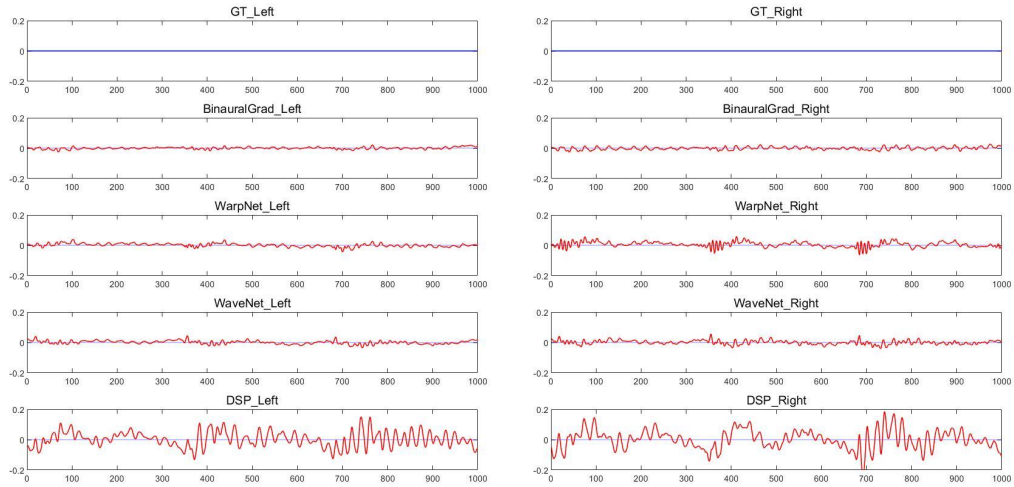
Firstly, we sincerely recommend readers to view the demo video and audio samples in the website[3].

[*]Equal contribution.
[†]This work was conducted at Microsoft. Corresponding author: Xu Tan, xuta@microsoft.com
[3]https://speechresearch.github.io/binauralgrad/

36th Conference on Neural Information Processing Systems (NeurIPS 2022).



(a) Waveform.



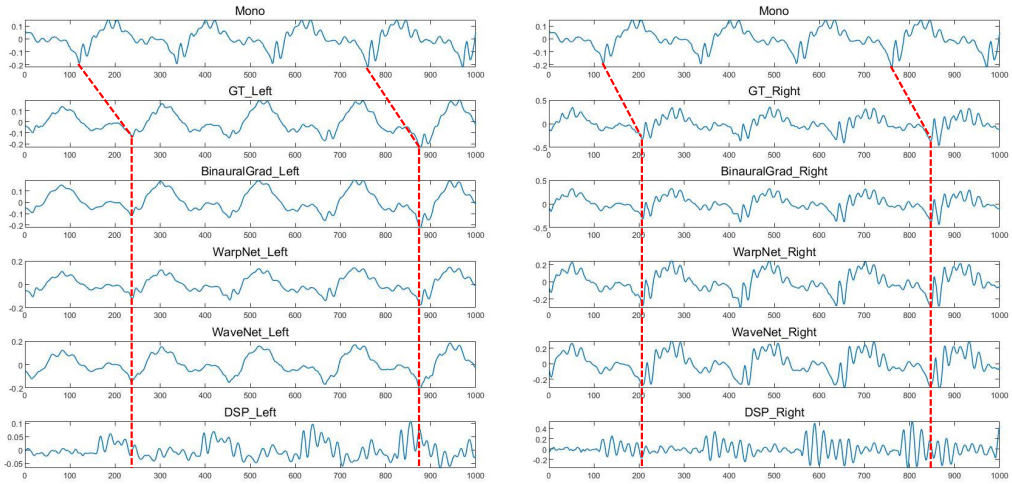
(b) Prediction Error.

Figure 1: Case 1. Sub-figure (a) shows the waveform of mono audio, ground-truth (GT) binaural audio, synthesized binaural audio from BinauralGrad and other baseline systems. Sub-figure (b) shows the prediction error between synthesized audio and GT audio.

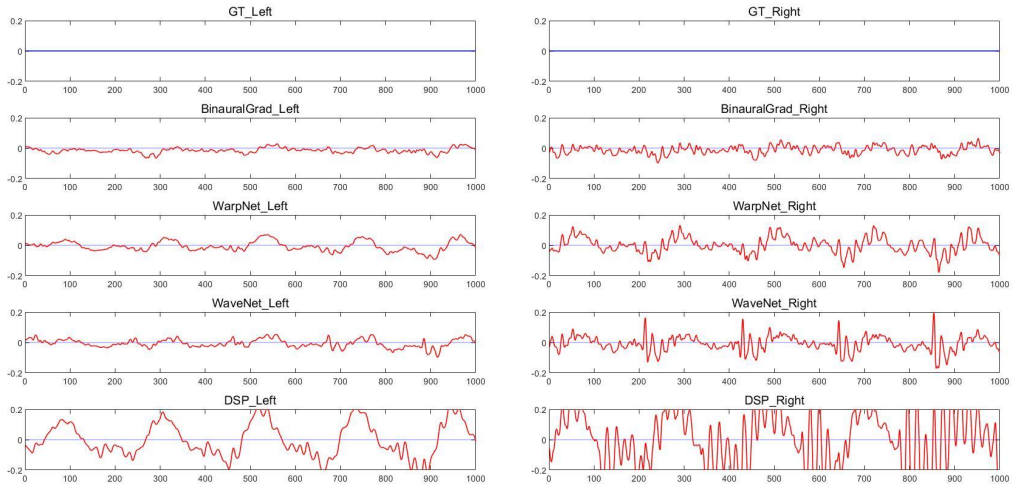
We randomly select three cases from the test set to intuitively compare the proposed BinauralGrad with other baselines. For each sample we plot 2 sub-figures, each with $N = 1000$ sampling points in a sampling rate of 48kHz, and the time length of each sample is around $0.021s$. One sub-figure (e.g., Figure 1(a),2(a),3(a)) shows the waveform of the mono audio, the ground-truth (GT) binaural audio and the generated binaural audio with amplitude information, in which we can visually check the time delay with the red dashed lines, and we can also compare the difference between different models.

Another sub-figure (e.g., Figure 1(b),2(b),3(b)) shows the prediction error between generated audio and GT audio. The blue lines show the GT results, while the red lines represent the prediction error at each sampling point. The prediction error of each model is shown in the same range $[-0.2, 0.2]$.

For these three cases, as shown in Figure 1, Figure 2 and Figure 3, we can find that BinauralGrad precisely predicts the GT waveform in both left ear audio and right ear audio. In most situations, the



(a) Waveform.

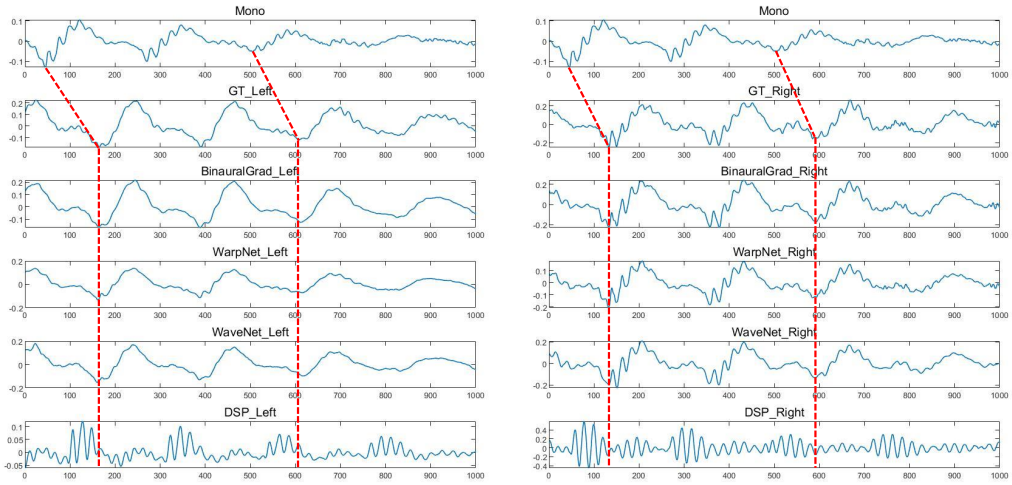


(b) Prediction Error.

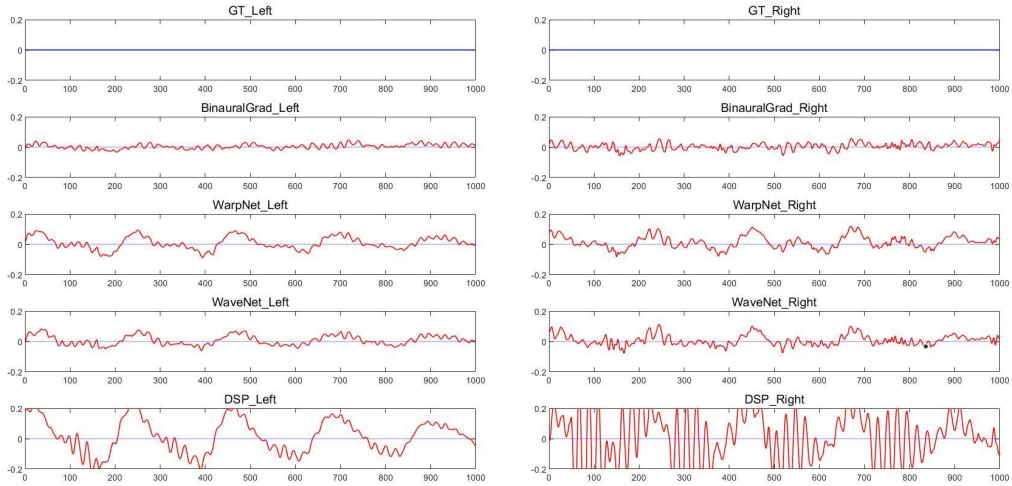
Figure 2: Case 2. Sub-figure (a) shows the waveform of mono audio, ground-truth (GT) binaural audio, synthesized binaural audio from BinauralGrad and other baseline systems. Sub-figure (b) shows the prediction error between synthesized audio and GT audio.

prediction error is close to Gaussian noise. As a comparison, WarpNet sometimes adds small artifacts on the waveform. And the prediction is not stable for binaural audios. As it can be seen in Figure 1(b), the prediction error of left ear audio is small, but the extra artifacts are especially apparent in the right ear audio when $N \in [0, 100], [350, 450], [670, 770]$. Moreover, in both Figure 2 and Figure 3, we can see obvious prediction error which even shows similarity with GT waveform in both left and right ear. For WaveNet, it may fail to accurately predict the GT waveform, which can be seen from Figure 1 and Figure 3. For DSP, the estimation results are not as good as other models either in the time delay or the amplitude prediction.

To sum up, in our test experiments, we find that our proposed BinauralGrad is advantageous in binaural audio waveform reconstruction. Compared to other models, the prediction error of our model is smaller. Especially, our results are more stable. Sometimes, other models can achieve



(a) Waveform.



(b) Prediction Error.

Figure 3: Case 3. Sub-figure (a) shows the waveform of mono audio, ground-truth (GT) binaural audio, synthesized binaural audio from BinauralGrad and other baseline systems. Sub-figure (b) shows the prediction error between synthesized audio and GT audio.

small prediction error in one ear, but they fail to accurately model the waveform of the other ear. BinauralGrad can simultaneously achieve accurate predictions in both left ear and right ear.

For other models, WarpNet sometimes adds extra artifacts on waveforms, which may be caused by their multiple methods to strengthen the phase estimation. And its prediction error is large in some time periods. WaveNet may fail to predict fine-grained details of GT waveforms because it only uses position as (a weak) condition information but not uses warping (proposed in WarpNet). DSP estimation results, where a generic (not-personalized) HRTF (head-related transfer functions) and RIR (room impulse response) is used since the dataset does not contains HRTF and RIR, are not as good as other models and its prediction error is usually much larger than other results.