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

Yichong Leng

University of Science and Technology of China

Microsoft Research Asia

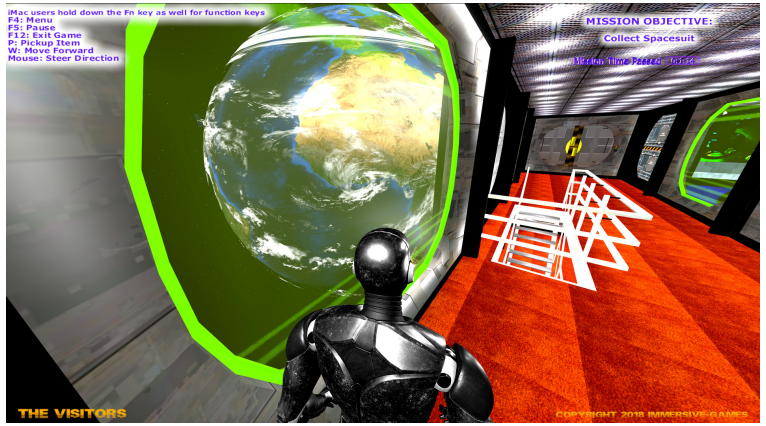
lyc123go@mail.ustc.edu.cn

Outlines

- Background
- BinauralGrad
- Future Directions

Spatial Audio

- Spatial audio can give listener a sense of direction of sound source.
- Spatial audio plays an important role in
  - virtual realities (VR);
  - immersive games;
  - metaverse.



Spatial Audio

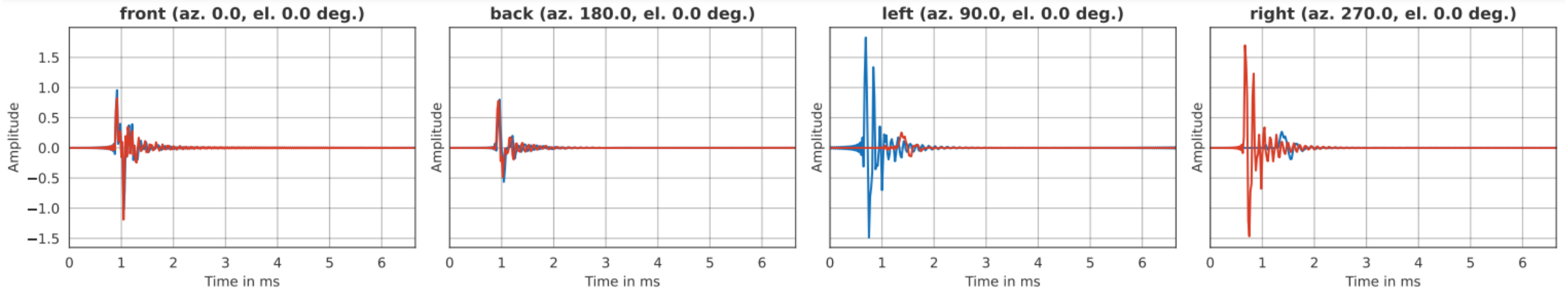
- Spatial audio can give listener a sense of direction of sound source.
- Spatial audio plays an important role in
  - virtual realities (VR);
  - immersive games;
  - metaverse.
- Spatial audio can be perceived with
  - headphones (*Binaural audio*);
  - home theater.

Binaural Audio

- Binaural audio contains
  - two channels: left ear and right ear
- Binaural audio synthesis
  - with a mono audio;
  - with a relative position of source and listener.

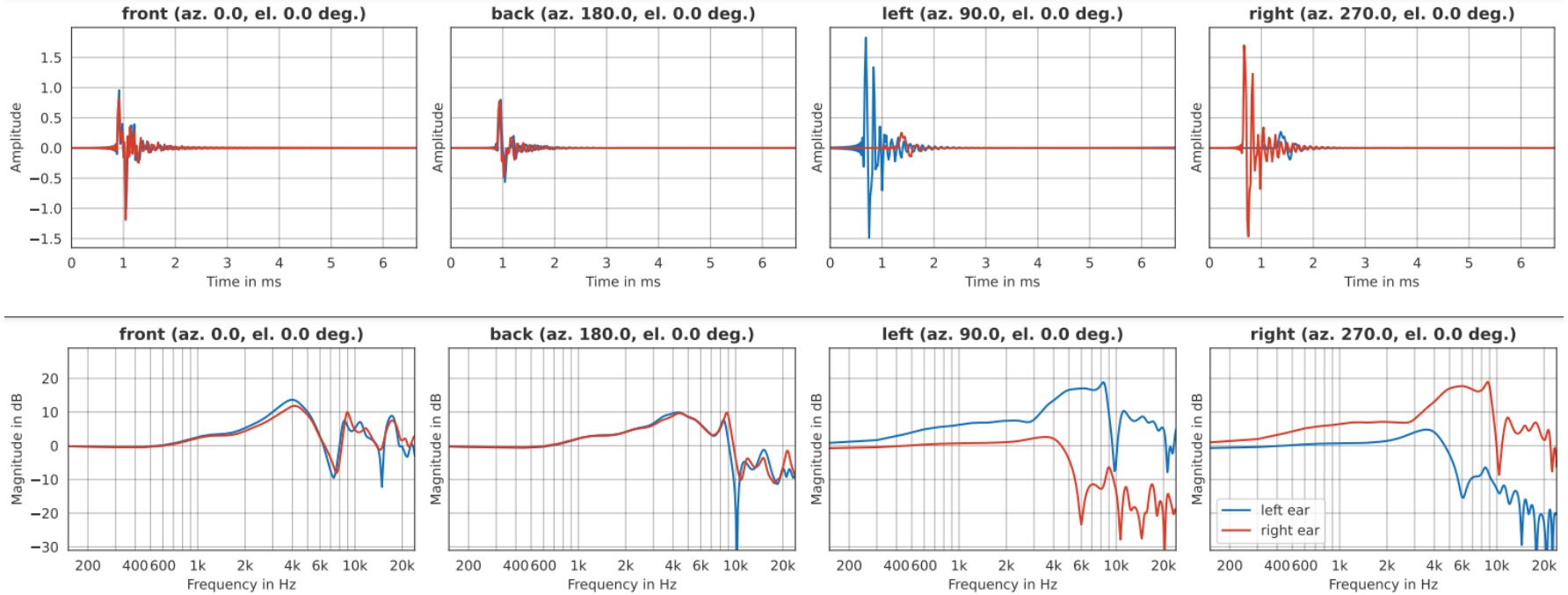
Spatial perception

- Listener (human) can discriminate the direction of sound by:
  - Interaural time difference (ITD)
    - Sound from the right side reaches the right ear earlier than the left ear.
  - Interaural level difference (ILD)
    - Sound from the right side has a higher level at the right ear than at the left ear.



Spatial perception

- Listener (human) can discriminate the direction of sound by:
  - Interaural time difference (ITD)
  - Interaural level difference (ILD)
  - Spectral information
    - Spectral difference of sound hearing by two ears.

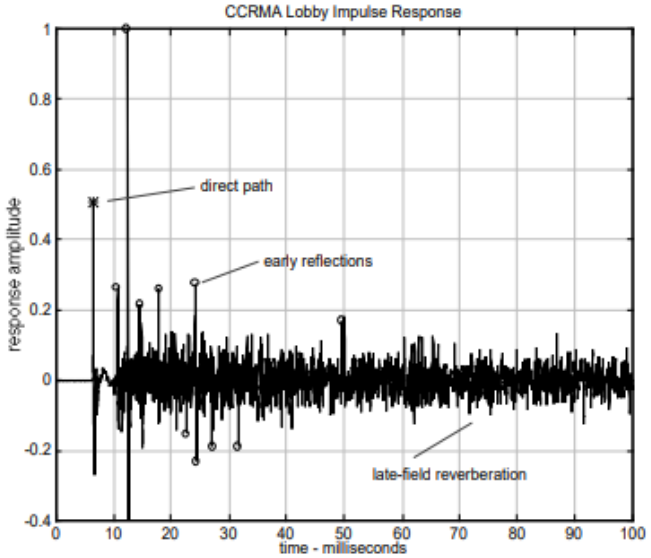


Generation of Spatial Audio

- The generation of spatial audio consists of two process.
  - Model environment:
    - Room Impulse Response (RIR)
  - Model human:
    - Head-Related Impulse Response (HRIR)
  - Based on a linear time-invariant system (LTI) assumption.

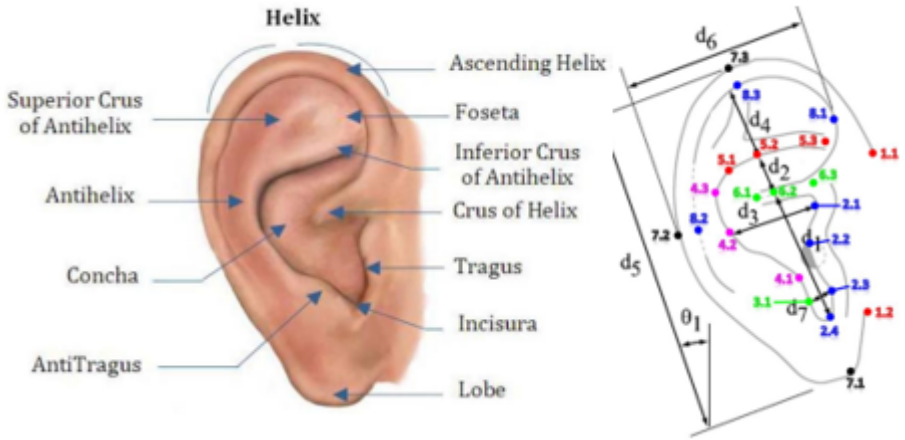
Generation of Spatial Audio

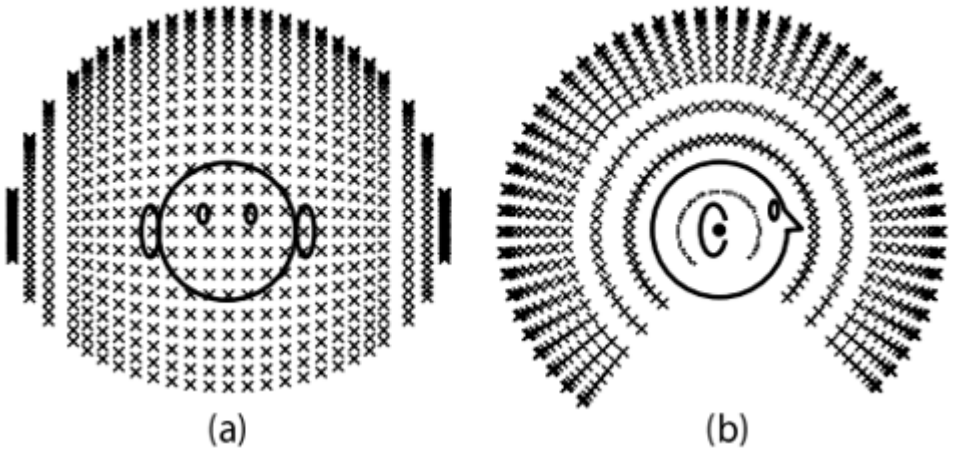
- Model environment: Room Impulse Response (RIR)
  - Interaction of sound with environment.
  - Given the location of source and listener, the sound arrives to the listener
    - directly (direct path),
    - after reflection (early reflections),
    - after many reflection (late-field reverberation)



Generation of Spatial Audio

- Model human: Head-Related Impulse Response (HRIR)
  - how an ear receives a sound from a point in space.
  - determined by morphology of listener (torso, head, and pinna).
  - measured on many direction (e.g., 720) surrounding the listener.





Generation of Spatial Audio

- Digital Signal Processing (DSP) method
  - With the mentioned notation, the spatial audio can by synthesized with

$$y^{(r|l)}(t) = h_{hrir}^{(r|l)}\left(t, \theta_t^{(src,lstn)}\right) * h_{rir}\left(t, \theta_t^{(src)}, p_t^{(lstn)}\right) * x(t) + w^{(r/l)}(t)$$

Outlines

- Background
- BinauralGrad
- Future Directions

Previous works

- Previous works on binaural audio synthesis:
  - Traditional digital signal processing (DSP) method.
  - End-to-end neural network method.

Previous works

- Previous works on binaural audio synthesis:
  - Traditional digital signal processing (DSP) method.
    - Linear time-invariant system with
      - Room impulse response (RIR) and head-related impulse response (HRIR).
  - End-to-end neural network method.

Previous works

- Previous works on binaural audio synthesis:
  - Traditional digital signal processing (DSP) method.
    - Linear time-invariant system with
      - Room impulse response (RIR) and head-related impulse response (HRIR).
    - Advantage:
      - Consistency with physical process of audio generation;
      - Good sense of direction.
  - End-to-end neural network method.

Previous works

- Previous works on binaural audio synthesis:
  - Traditional digital signal processing (DSP) method.
    - Linear time-invariant system with
      - Room impulse response (RIR) and head-related impulse response (HRIR).
    - Advantage:
      - Consistency with physical process of audio generation;
      - Good sense of direction.
    - Disadvantage:
      - Moderate audio quality;
      - Cannot model non-linear effect.
  - End-to-end neural network method.

Previous works

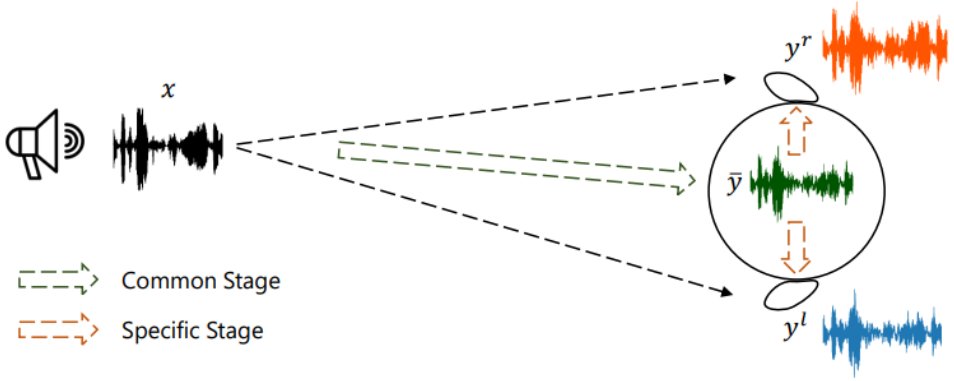
- Previous works on binaural audio synthesis:
  - Traditional digital signal processing (DSP) method.
  - End-to-end neural network method.
    - Learning to synthesize binaural audio with
      - mono audio and position of source and listener.

Previous works

- Previous works on binaural audio synthesis:
  - Traditional digital signal processing (DSP) method.
  - End-to-end neural network method.
    - Learning to synthesize binaural audio with
      - mono audio and position of source and listener.
    - Advantage:
      - Non-linear effect modelling.
      - Better audio quality.
    - Disadvantage:
      - Lack of inductive bias on physical process of audio generation.

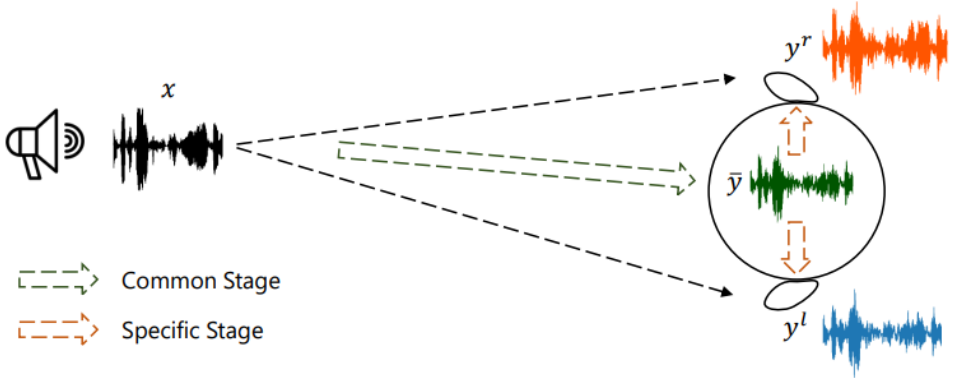
BinauralGrad

- Decomposing the synthesis of binaural audio into two stages:
  - Common stage
  - Specific stage



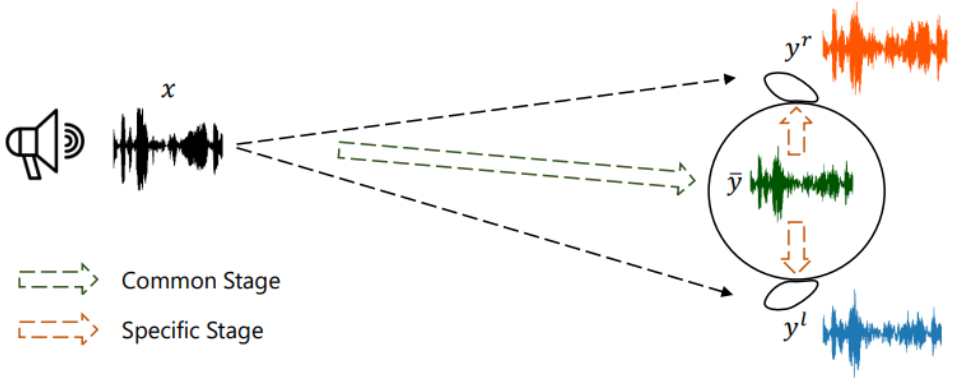
BinauralGrad

- Decomposing the synthesis of binaural audio into two stages:
  - Common stage
    - models the factor affecting both the left and right channel.
      - Distance of source and listener;
      - Interaction of audio with environment.
  - Specific stage



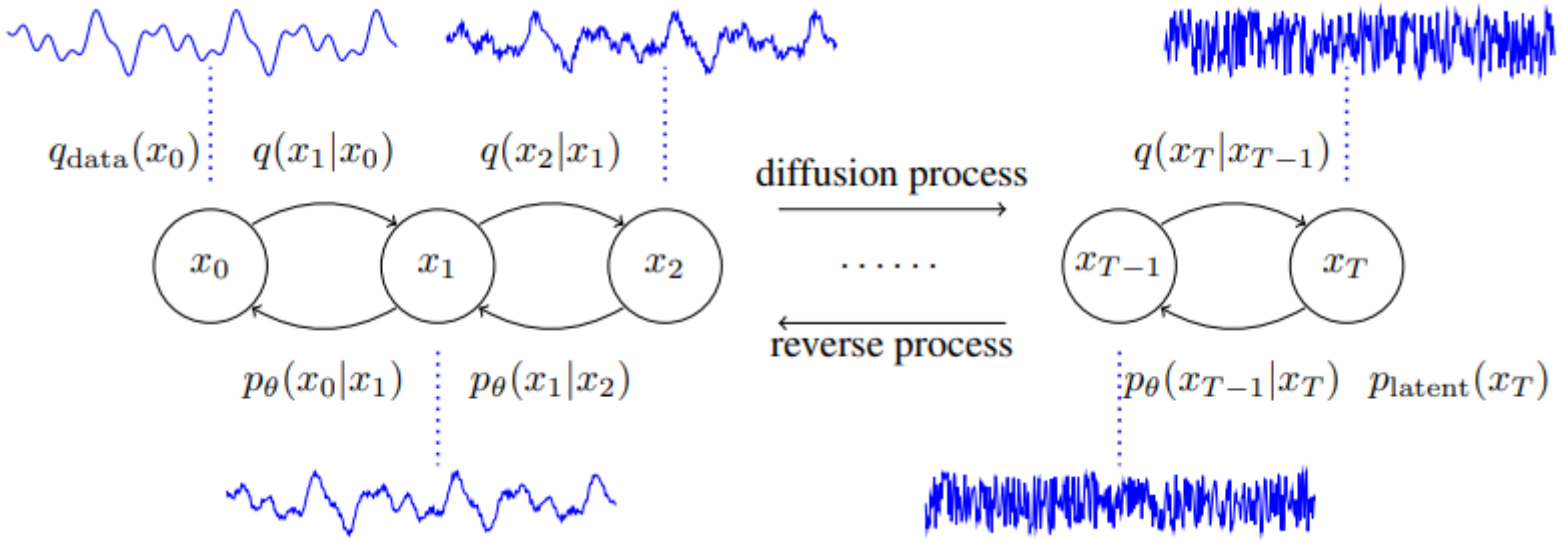
BinauralGrad

- Decomposing the synthesis of binaural audio into two stages:
  - Common stage
    - models the factor affecting both the left and right channel.
      - Distance of source and listener;
      - Interaction of audio with environment.
  - Specific stage
    - models the difference of the left and right channel.
      - Marginal distance difference of two ears;
      - Difference of impulse response of two ears.



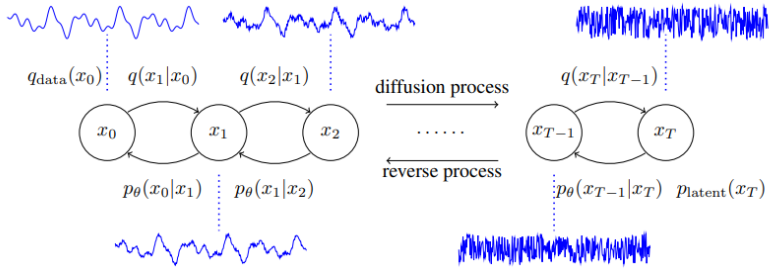
BinauralGrad

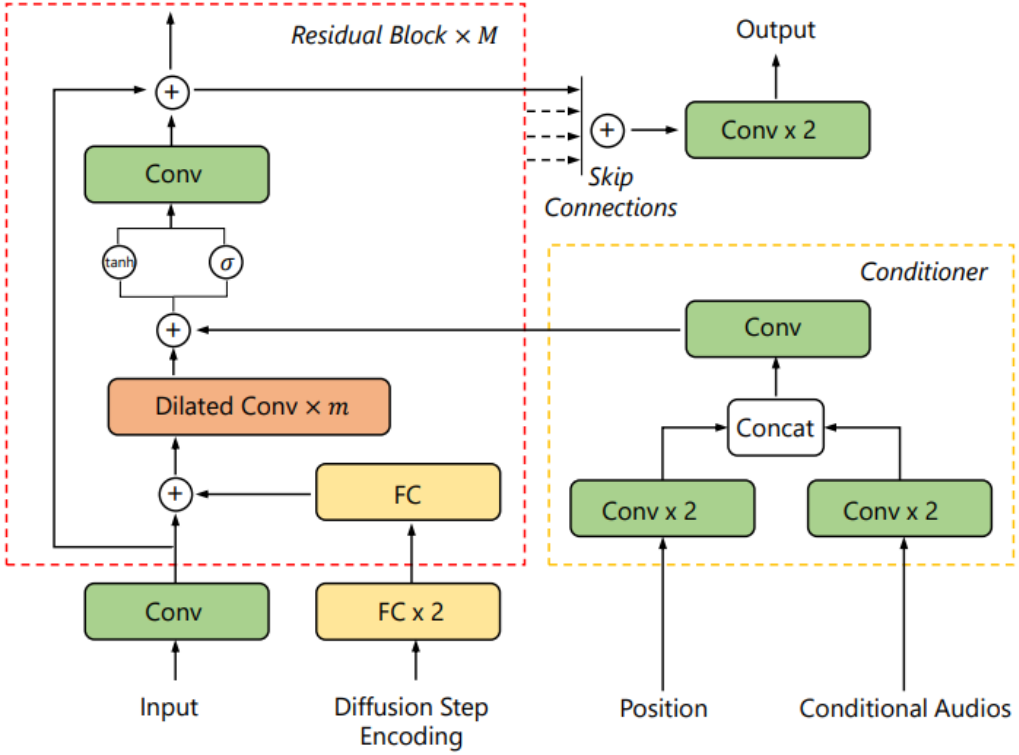
- We utilize diffusion model to model the two stages
  - to learn the high-dimensional data space of waveform.



BinauralGrad

- We utilize diffusion model to model the two stages
  - to learn the high-dimensional data space of waveform.





BinauralGrad

- Physical warping the mono audio x to binaural audio ${x_{warp}}^{l}, {x_{warp}}^{r}$
  - based on the position of source and ears.
  - Calculate the delay:

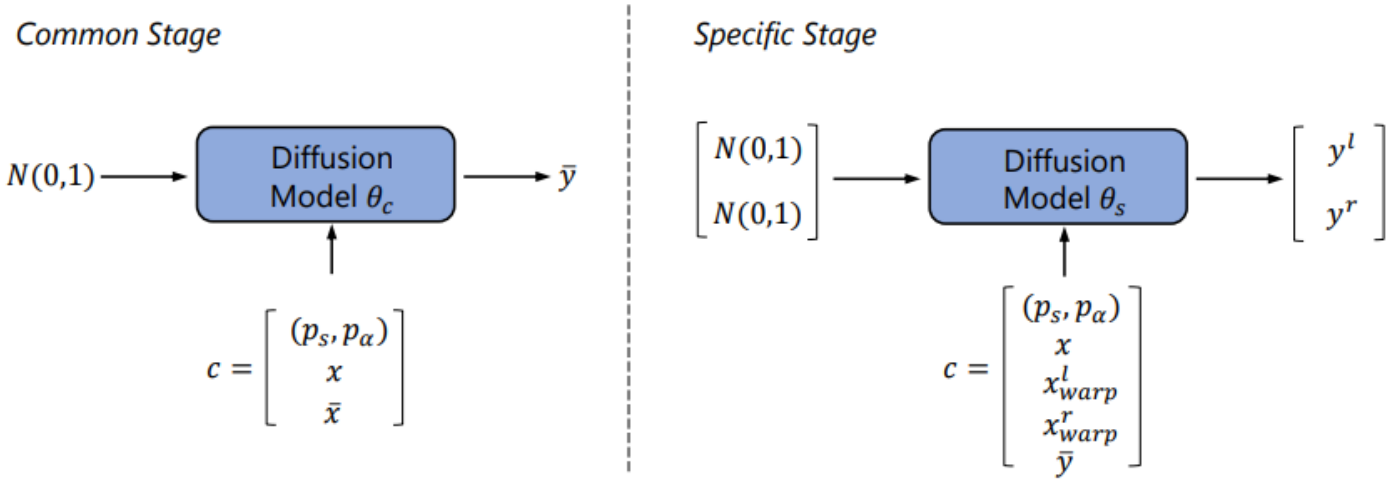
$$\rho(n) = n - C \cdot \|p_{src}(n) - p_{lstn}(n)\|,$$

  - Warp the audio:

$$x_{warp}(n) = (\lceil \rho(n) \rceil - \rho(n)) \cdot x_{\lfloor \rho(n) \rfloor} + (\rho(n) - \lfloor \rho(n) \rfloor) \cdot x_{\lceil \rho(n) \rceil},$$

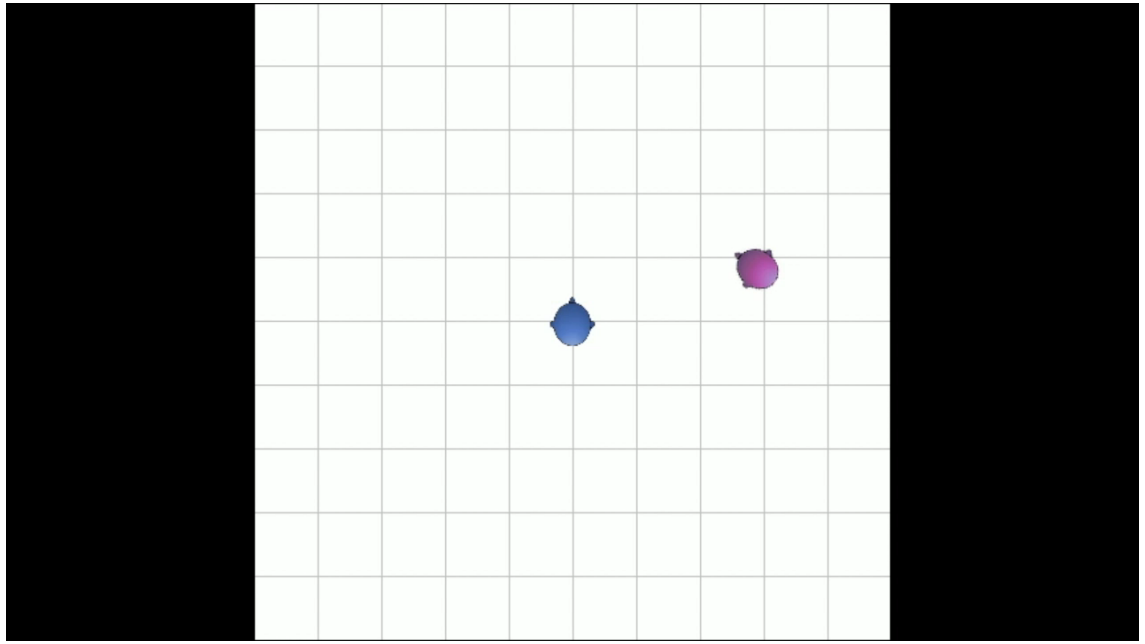
BinauralGrad

- Warping the mono audio x to binaural audio ${x_{warp}}^l, {x_{warp}}^r$
- Common information is the average of two channels.
  - Warping: $\bar{x} = ({x_{warp}}^l + {x_{warp}}^r)/2$.
  - Target: $\bar{y} = (y^l + y^r)/2$.



Dataset

- Open-source dataset from Meta.
- Eight different subjects.
  - four male and four female.
- Talking to a mannequin while walking around.
  - 15 minutes each.
  - Mono audio
  - Binaural audio
  - Relative position of source and listener.



Baselines

- Digital Signal Processing (DSP) method
- WaveNet
  - the positions of source and listener are used as condition signals.
  - generate binaural audio from the mono one.
- WarpNet
  - stacking a neural time warping module with a temporal ConvNet to generate binaural audio

Experiments

Table 1: The comparison regarding binaural audio synthesis quality. For all baselines, we report the results based on our re-implementation.

Model	Wave L2 ($\times 10^{-3}$) ↓	Amplitude L2 ↓	Phase L2 ↓	PESQ ↑	MRSTFT ↓
DSP	1.543	0.097	1.596	1.610	2.750
WaveNet [24]	0.179	0.037	0.968	2.305	1.915
WarpNet [29]	0.157	0.038	0.838	2.360	1.774
BinauralGrad	**0.128**	**0.030**	**0.837**	**2.759**	**1.278**

- BinauralGrad achieves
  - better waveform quality;
  - better perception score;
  - better spectrum quality.

Ablation Study

Table 2: The comparison with the single stage diffusion model.

Model	Wave L2 ($\times 10^{-3}$) ↓	Amplitude L2 ↓	Phase L2 ↓	PESQ ↑	MRSTFT ↓
Single Stage	0.144	0.031	0.877	2.544	**1.269**
BinauralGrad	**0.128**	**0.030**	**0.837**	**2.759**	1.278

- Verifying the effectiveness of two stage.

Subjective Test

Table 3: The MOS test results on the synthesized binaural audio. Best scores over baselines are marked bold.

Model	MOS	Similarity MOS	Spatial MOS
Recording	3.69 ± 0.24	4.52 ± 0.23	4.01 ± 0.13
DSP	3.41 ± 0.23	3.97 ± 0.37	3.75 ± 0.15
WaveNet [24]	3.37 ± 0.20	4.14 ± 0.29	3.41 ± 0.15
WarpNet [29]	3.61 ± 0.24	4.25 ± 0.25	3.77 ± 0.13
BinauralGrad	$\mathbf{3.80} \pm 0.23$	$\mathbf{4.43} \pm 0.20$	$\mathbf{3.79} \pm 0.16$

- BinauralGrad achieves
  - better quality;
  - better similarity with corresponding binaural recording;
  - Good spatial quality.

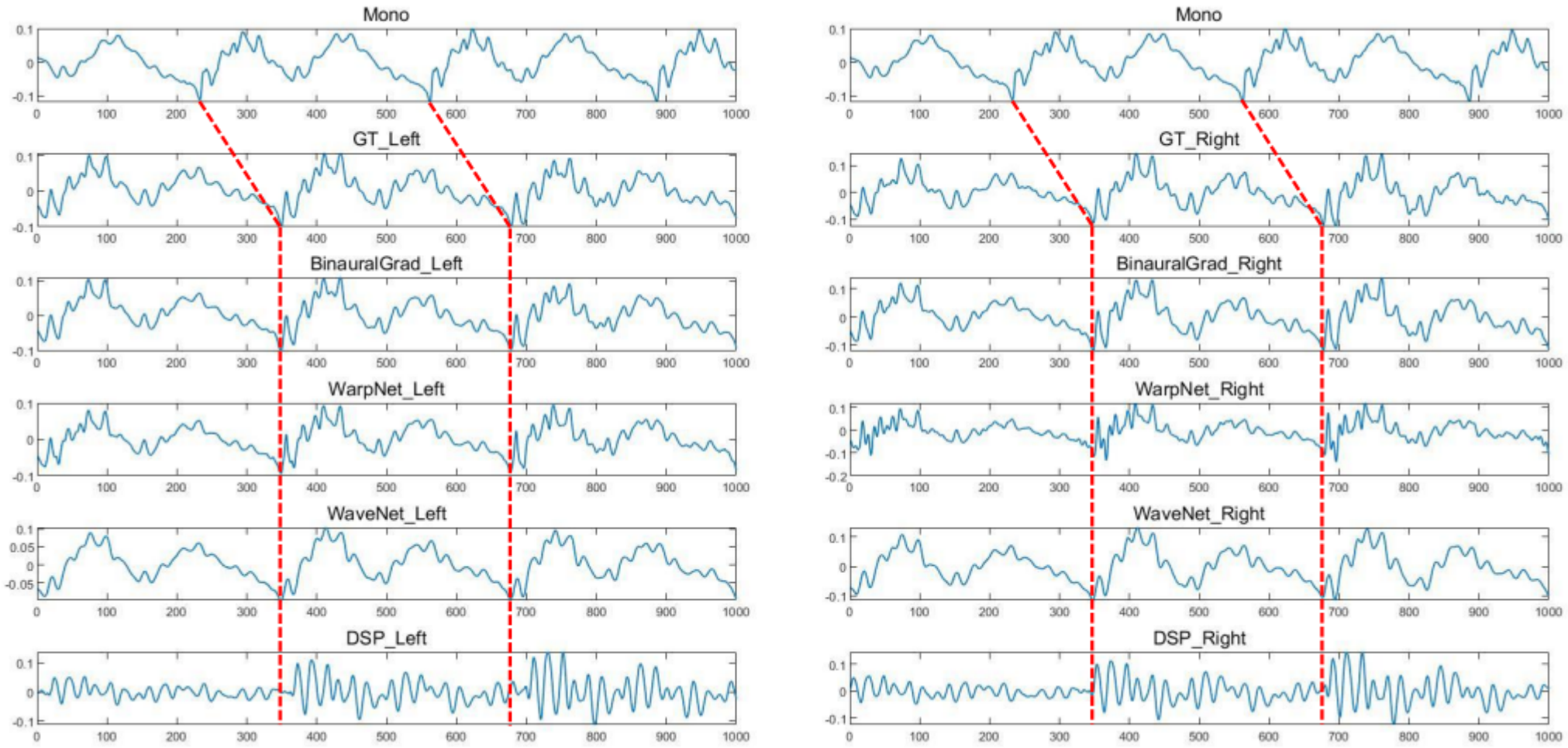
Analysis

Table 4: Analysis of BinauralGrad. Stage 1 stands for the first/common stage and stage 2 stands for the second/specific stage in BinauralGrad.

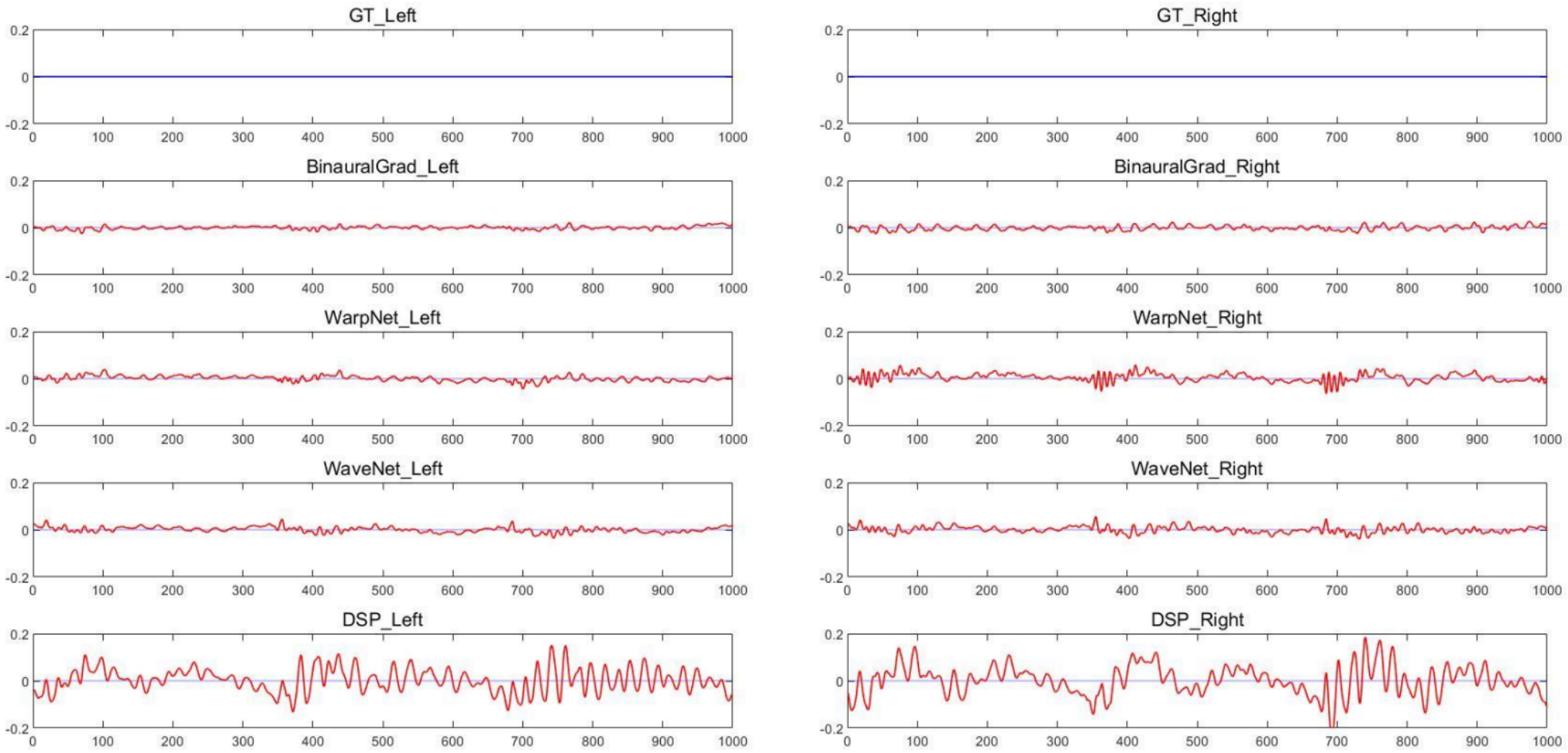
ID	Setting	Wave L2 ($\times 10^{-3}$) ↓	Amplitude L2 ↓	Phase L2 ↓	PESQ ↑	MRSTFT ↓
1	Mono	1.340	0.063	1.564	1.410	2.159
2	Physical Warp	0.725	0.060	1.584	1.411	2.140
3	Stage 1	0.300	0.043	1.070	2.000	1.682
4	$(\bar{y}, \bar{y})$	0.200	0.031	0.598	3.031	1.020
5	Stage 2	0.128	0.030	0.837	2.759	1.278
6	Stage 2 on ID 4	0.013	0.013	0.372	4.212	0.697

- 1-2: Physical warping is not enough.
- 3-4: First stage is not satisfying.
- 5-6: Second stage is well-trained.

Case Study


Mono
GT_Left
BinauralGrad_Left
WarpNet_Left
WaveNet_Left
DSP_Left
Mono
GT_Right
BinauralGrad_Right
WarpNet_Right
WaveNet_Right
DSP_Right

Case Study


GT_Left
GT_Right
BinauralGrad_Left
BinauralGrad_Right
WarpNet_Left
WarpNet_Right
WaveNet_Left
WaveNet_Right
DSP_Left
DSP_Right
0.2
0
-0.2
0
100
200
300
400
500
600
700
800
900
1000

Conclusion

- BinauralGrad: a two-stage diffusion model for binaural audio synthesis.
  - Common stage to model the factor affecting both ears.
  - Specific stage to model the difference of left ear and right ear.
- Demo Page
  - https://speechresearch.github.io/binauralgrad
- Code
  - https://github.com/microsoft/NeuralSpeech/tree/master/BinauralGrad

Outlines

- Background
- BinauralGrad
- Future Directions

Future Directions

- Better corporation of two-stage models.
  - The first stage model is not satisfying.

Future Directions

- Better corporation of two-stage models.
  - The first stage model is not satisfying.
- Faster sampling for diffusion models.
  - With the strong condition.

Future Directions

- Better corporation of two-stage models.
  - The first stage model is not satisfying.
- Faster sampling for diffusion models.
  - With the strong condition.
- Using knowledge of room or head.
  - With the help of mesh.

Future Directions

- Application on real scenario
  - Such as Hololens, Teams, Virtual Cube

Future Directions

- Application on real scenario
  - Such as Hololens, Teams, Virtual Cube

- Development of data
  - Larger recording dataset.
  - Utilization of simulated dataset (SoundSpace 2.0, GWA)

BinauralGrad

- Demo Page
  - https://speechresearch.github.io/binauralgrad
- Code
  - https://github.com/microsoft/NeuralSpeech/tree/master/BinauralGrad

Thank you