PSZ U-Net Overview

Tanay Mannikar
8/27/24

Outline

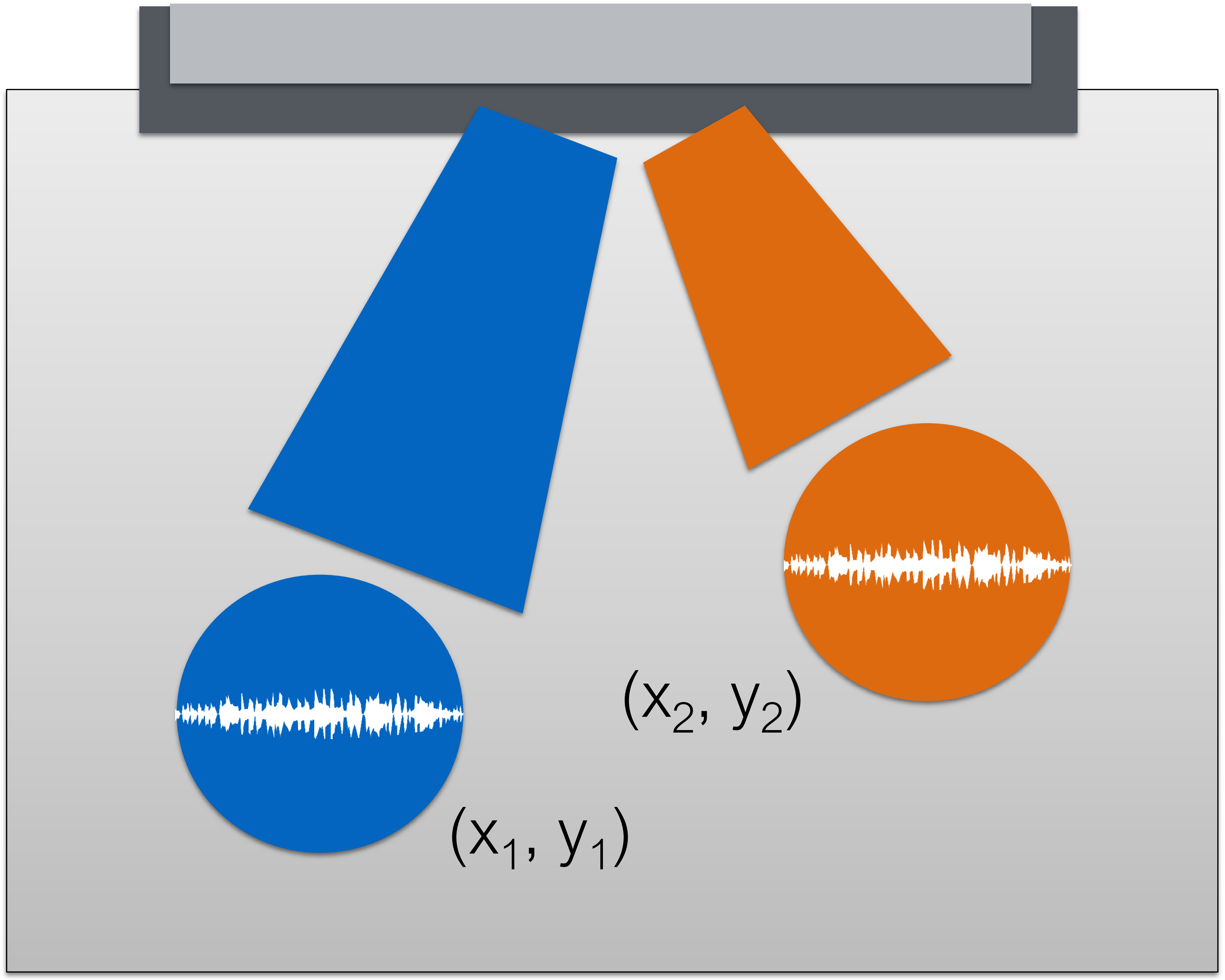
- Audio “End-To-End” Generation Model
- RTF & Audio Dataset
- Loss Optimization Scheme
- Auditory Filter Bank & Perceptual Speech Quality Loss
- Next Steps

Overview

Personal Sound Zones (PSZs)

- Use loudspeaker arrays to generate isolated audio programs in two individual zones via PSZ filtering

- Current approach: filters generated for each loudspeaker via deep neural network for each program/zone



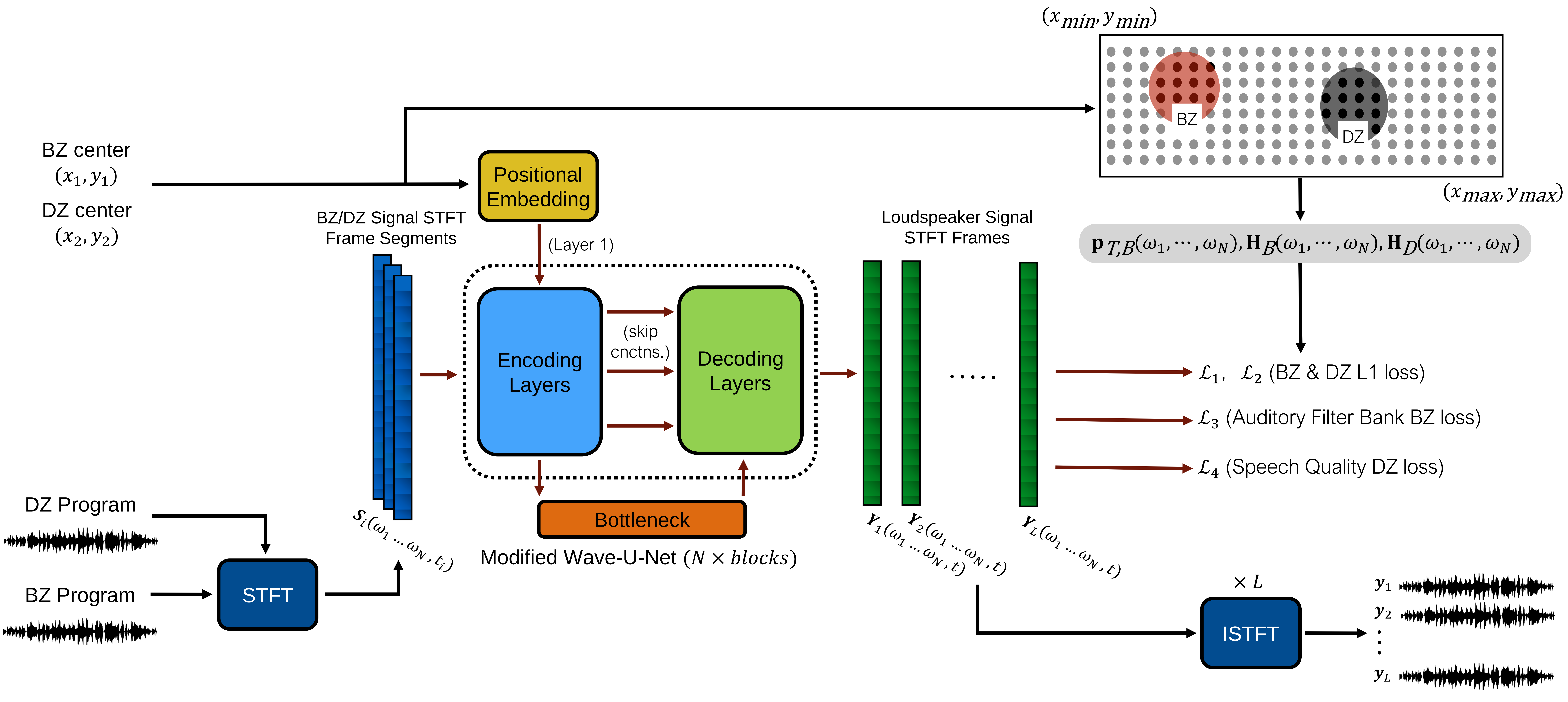
Audio “End-To-End” Generation Model

Modified *Wave-U-Net* architecture replaces previous filter generation model

- Learns a **multichannel nonlinear filter** independent of input audio
- Removes additional convolution step and can output **simultaneous zones** in a single model instance
- Allows for use of **perceptual loss metrics** during training for more effective PSZ filtering
- **Complex-valued inputs/outputs** offer simpler implementation and reduced model size

Block processing – adapts output in real-time to user programs and head centers

- Trains on multiple time-frequency frames for time-dependent features

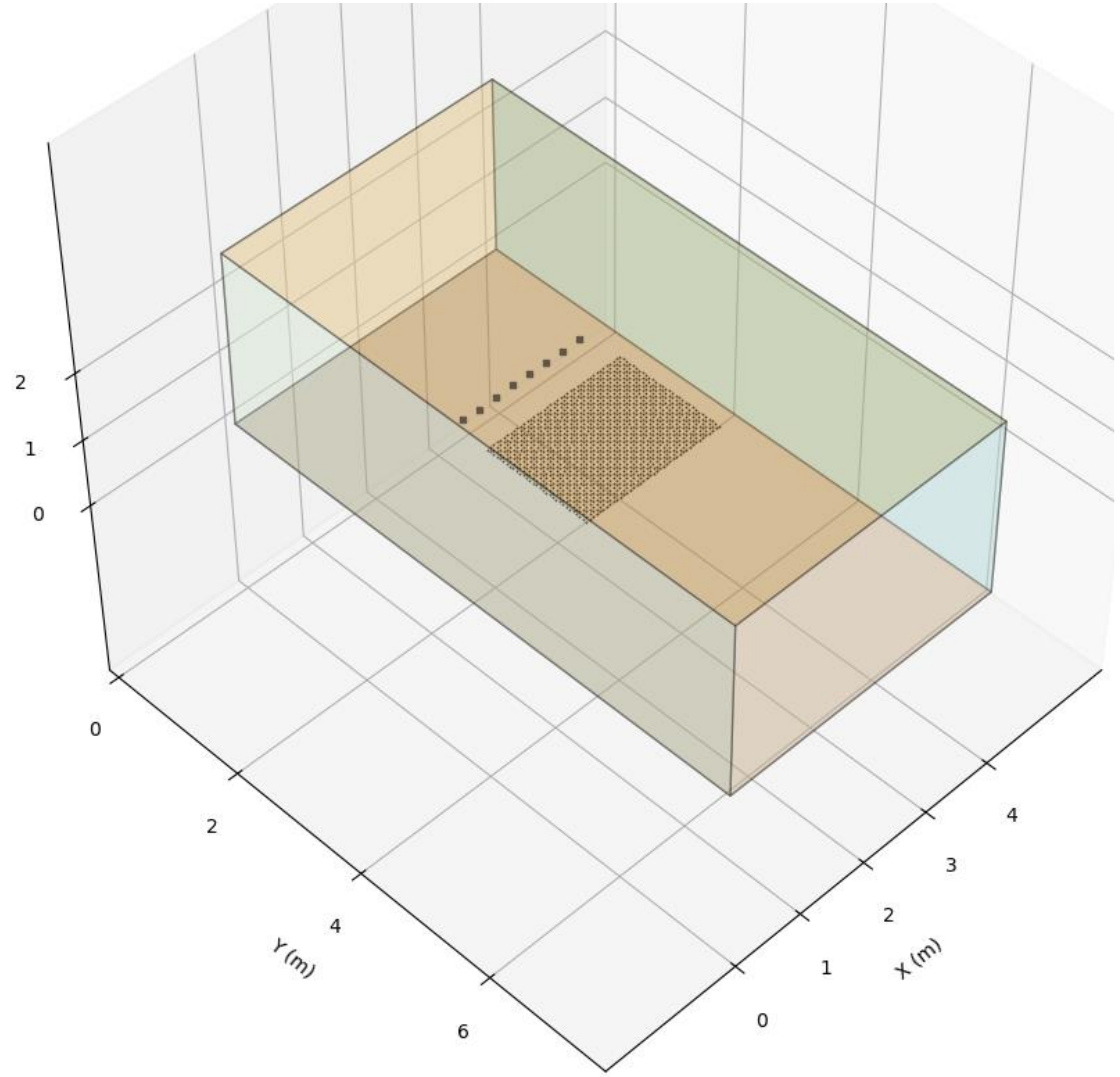


BZ center
(x_1, y_1)
DZ center
(x_2, y_2)
Positional Embedding
BZ/DZ Signal STFT Frame Segments
(Layer 1)
Encoding Layers
(skip cnctns.)
Decoding Layers
Bottleneck
Modified Wave-U-Net ($N \times blocks$)
DZ Program
BZ Program
STFT
$S_i(\omega_1 \dots \omega_N, t_i)$
Loudspeaker Signal STFT Frames
$Y_1(\omega_1 \dots \omega_N, t)$
$Y_2(\omega_1 \dots \omega_N, t)$
$Y_L(\omega_1 \dots \omega_N, t)$
(x_{min}, y_{min})
(x_{max}, y_{max})
BZ
DZ
$\mathbf{p}_{T,B}(\omega_1, \cdots, \omega_N), \mathbf{H}_B(\omega_1, \cdots, \omega_N), \mathbf{H}_D(\omega_1, \cdots, \omega_N)$
$\mathcal{L}_1$, $\mathcal{L}_2$ (BZ & DZ L1 loss)
$\mathcal{L}_3$ (Auditory Filter Bank BZ loss)
$\mathcal{L}_4$ (Speech Quality DZ loss)
$\times L$
ISTFT
y_1
y_2
y_L

RTF Dataset

gpuRIR

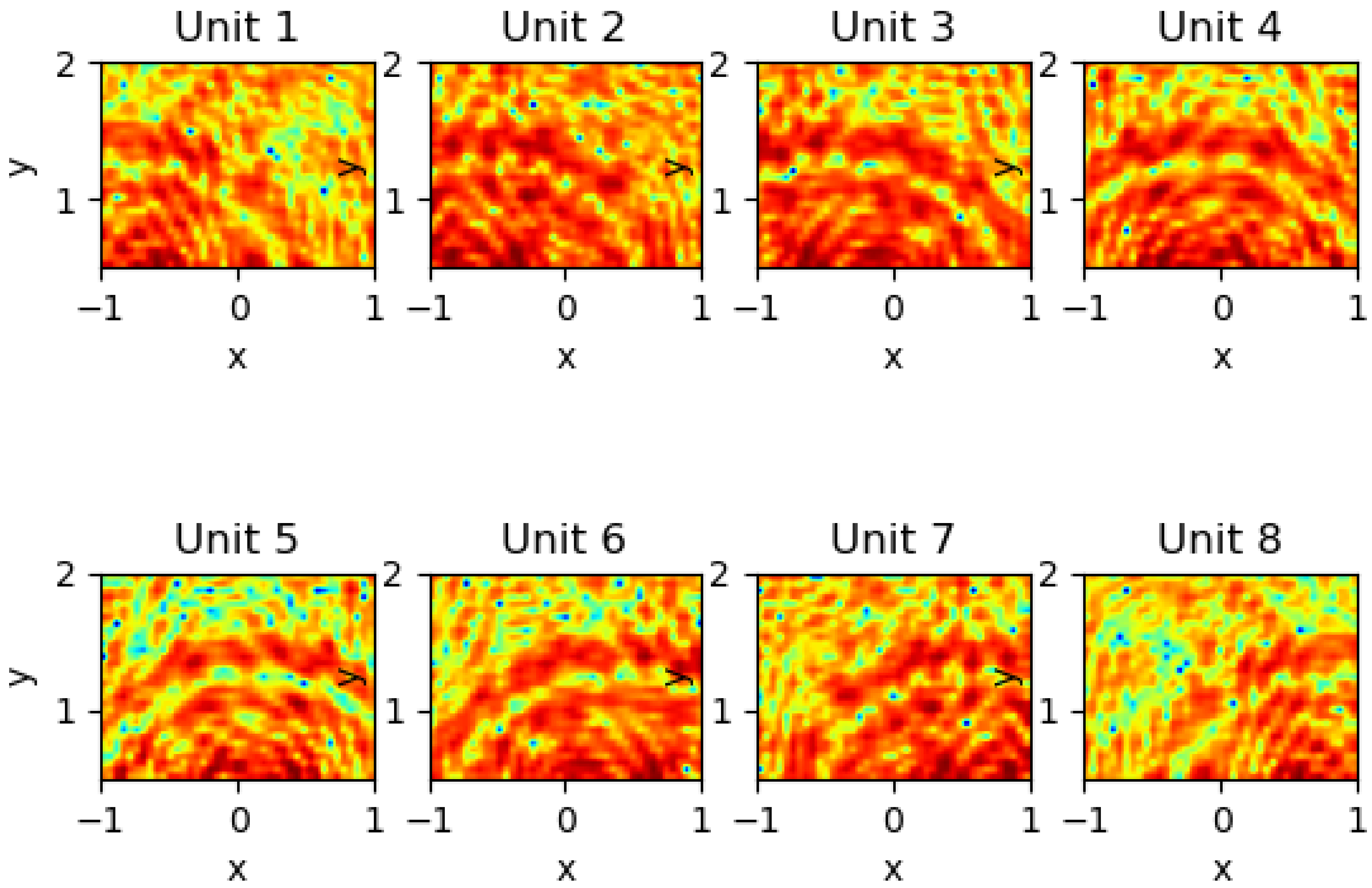
- Parallel GPU-accelerated multichannel RIR generation at desired coordinates using image source method
- Requires use of university HPCs (e.g., Adroit/Della) for data generation
- Diverse RTF dataset precomputed by Yue factoring in various room dimensions, RT60s, zone positions, etc.



RTF Dataset

Usage

- Model output multiplied with RTF data in frequency domain at loss computation stage to simulate pressure in target zones
- RTFs from center/reference speakers 3 & 4 used to promote program perception at center of array



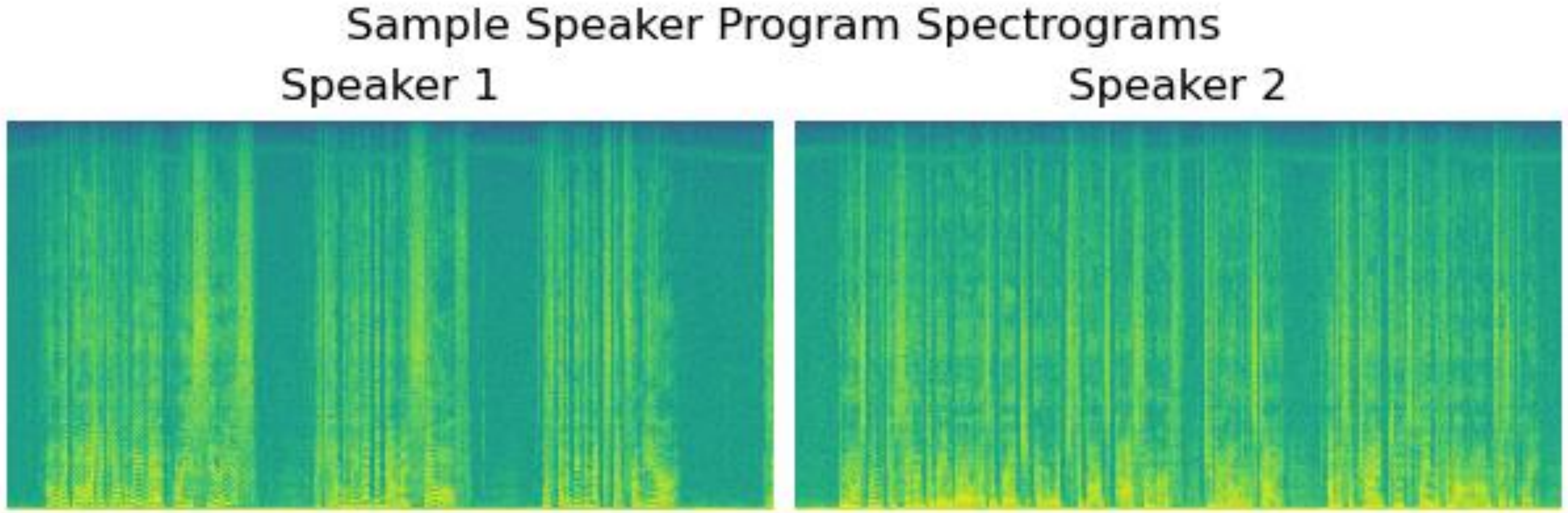
Audio Dataset

EARS speech dataset

- 48 kHz dataset consisting of 108 different speakers with varying inflections, emotions, vocal ranges, genders, etc.
- Model currently trained using single speaker data

ARCA23K audio dataset

- 44.1 kHz corpus consisting of various music, speech, and noise samples



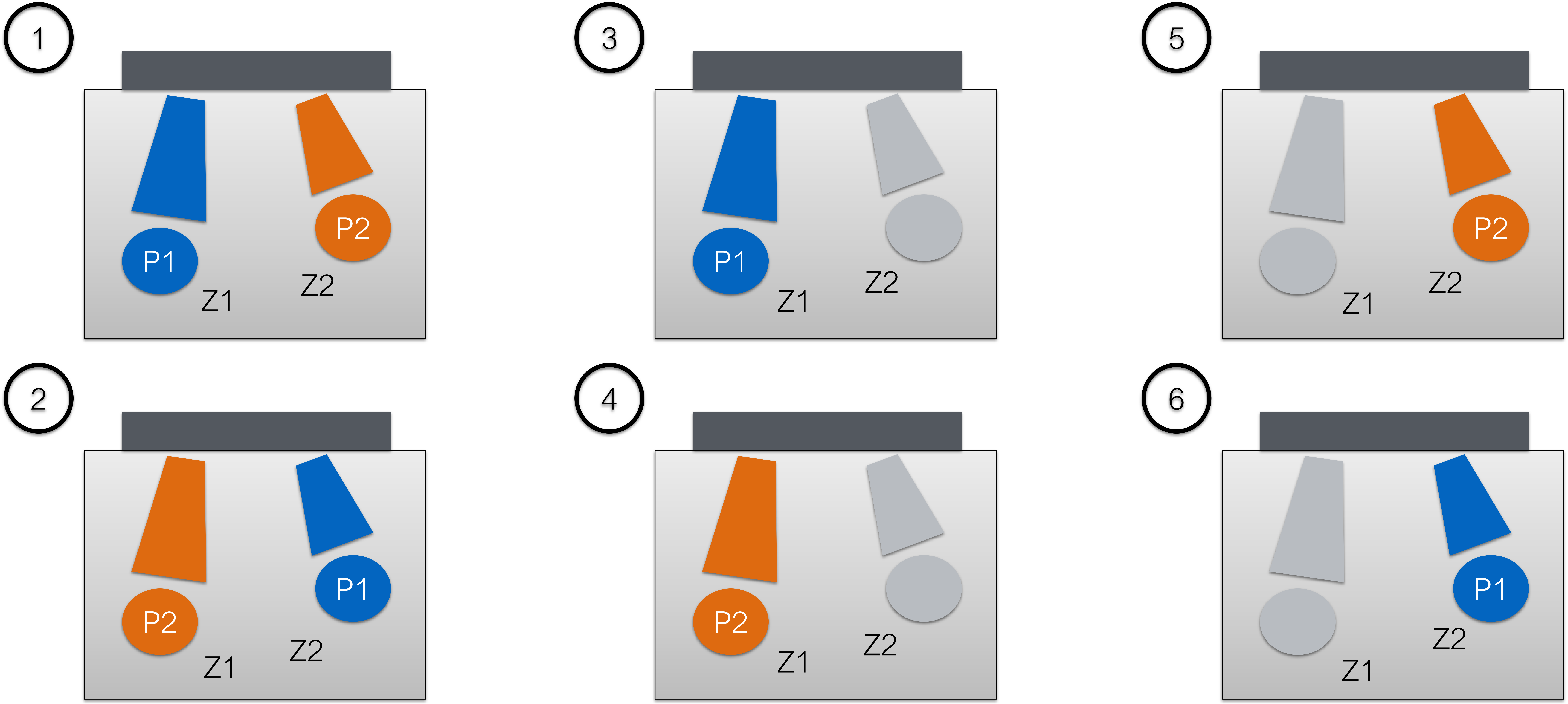
Loss Optimization Scheme

Six possible scenarios optimized when computing loss given two audio programs P1, P2 and two zone positions Z1, Z2

- Each instance of the model generates truncated frequency spectrum according to woofer/tweeter frequency response limits (i.e., 100 – 1500 Hz model and 1500 – 8000 Hz model)

- L1 (magnitude) loss computed only on truncated/band-limited STFT and RTF frequency data

- Reduces number of model parameters and total instances of model needed to generate target zones


1
P1
Z1
P2
Z2
2
P2
Z1
P1
Z2
3
P1
Z1
Z2
4
P2
Z1
Z2
5
Z1
P2
Z2
6
Z1
P1
Z2

Auditory Filter Bank Loss

Targets BZ loss by **minimizing irrelevant frequency information**

Dynamic Compressive Gammawarp filter bank $(\mathbf{F})$

- Emphasizes auditory **frequency selectivity** and cochlear nonlinearity/compressive gain structure of the inner ear (cochlea) – fine loudness perception
- Nonlinear frequency scale of *warped* filter bank improves **robustness** against reverberant audio compared to octave-resolution filter banks
- Efficient implementation and **better interpretability** than other nonlinear filter bank models due to linear FIR structure and invertibility/orthogonality

Static pre-emphasis filter $(\mathbf{W})$

- Static A-weighted FIR filtering of model output prior to loss computation accounts for outer-middle ear frequency masking – coarse loudness perception

Auditory Filter Bank Loss

$$\mathcal{L}_3 = \frac{1}{KLN} \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{l=1}^{L} \sum_{k=1}^{K} \mathbf{W}(\omega_n) \left| \left| \mathbf{F}_m(\omega_n) \mathbf{R}_{k,l_{ref}}(\omega_n) \mathbf{X}(\omega_n) \right| - \left| \mathbf{F}_m(\omega_n) \mathbf{R}_{k,l} \mathbf{Y}_l(\omega_n) \right| \right|$$

$\mathbf{F}_m(\omega_n) = m^{th}$ bandpass filter in perceptual filterbank $\mathbf{F}$
$\mathbf{W}(\omega_n) =$ outer & middle – ear weighting filter
$\mathbf{Y}_l(\omega_n) = l^{th}$ model output STFT frame
$\mathbf{X}(\omega_n) =$ anechoic target program STFT frame
K zone control points, L loudspeakers, M bandpass filters, N frequency bins

Perceptual Speech Quality Loss

Targets DZ loss by **minimizing speech intelligibility** in silent or opposing zone

PESQ (Perceptual Evaluation of Speech Quality)

- Commonly-used loss metric for speech intelligibility enhancement
- Must be combined with *scale-invariant SDR* (SI-SDR) for use in noise suppression

STOI (Short-Time Objective Intelligibility)

- Loss based on relative speech degradation compared to clean speech
- Although targets noise suppression, does not take inner-ear filtering into consideration

HASQI/HAAQI (Hearing Aid Speech/Audio Quality Index)

- Loss based on relative degradation compared to clean speech or audio
- Takes inner-ear filtering into consideration

Next Steps

- Finish IPI/IZI evaluation for non-perceptually optimized model
- Train model using larger and broader *ARCA23K* dataset
- Add filter bank and speech/audio quality metric to loss computation as opposed to solely L1 loss
- Numerical evaluation and subjective human testing using woofer/tweeter array setup

Questions?