L-02

Head-Related Transfer Function Interpolation from Spatially Sparse Measurements Using Autoencoder with Source Position Conditioning

Yuki Ito, Tomohiko Nakamura, Shoichi Koyama, and Hiroshi Saruwatari

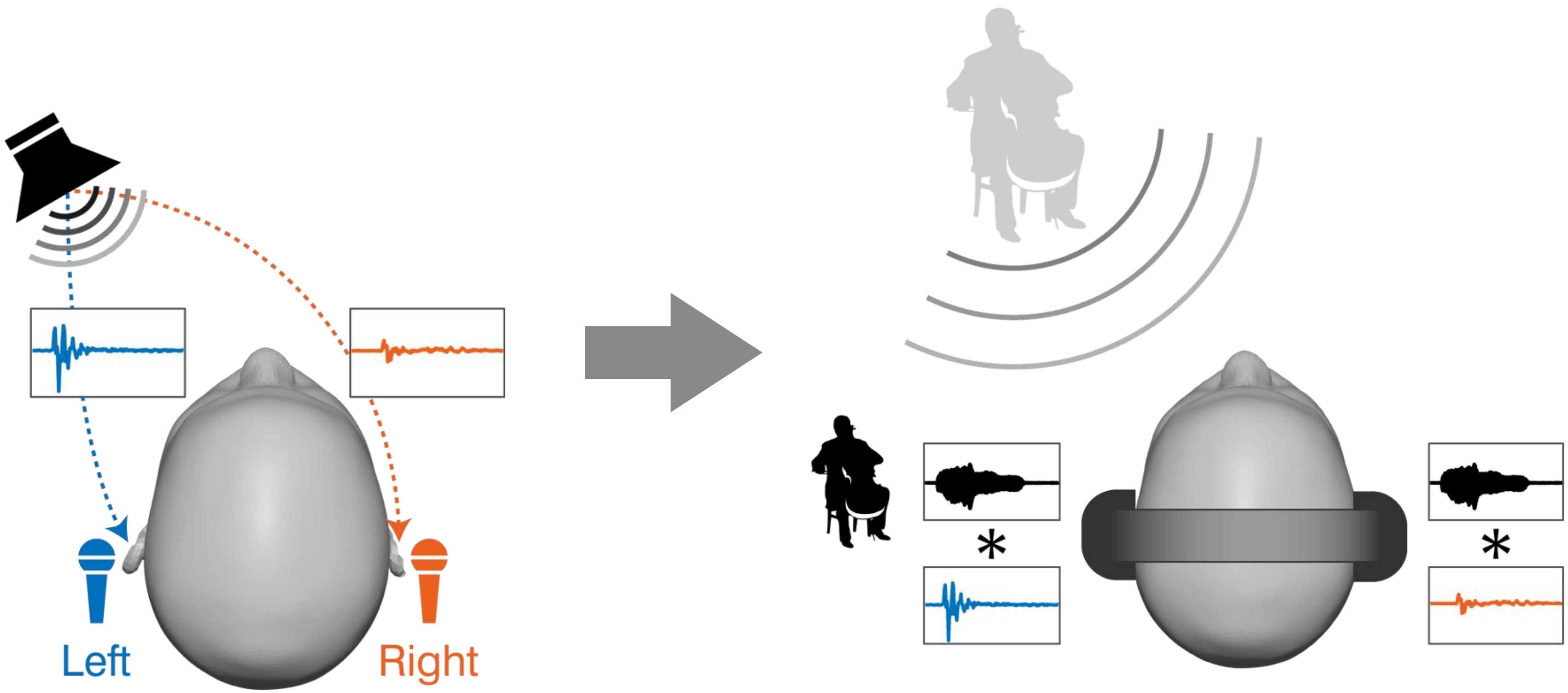
The University of Tokyo

Demo

Slides

Head-Related Transfer Function (HRTF)

- Transfer characteristics from sound source to both ears
- Contains auditory cues for sound image localization
- Applicable to synthesize binaural signals for VR/AR audio

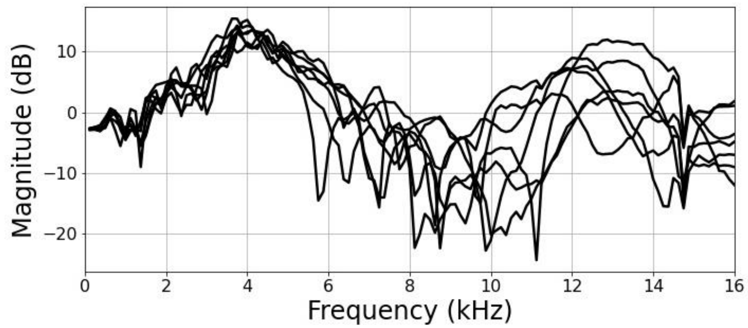


Motivation: Reduction of HRTF Measurement Cost

■ HRTF measurement is costly

– Necessity of person-by-person measurement

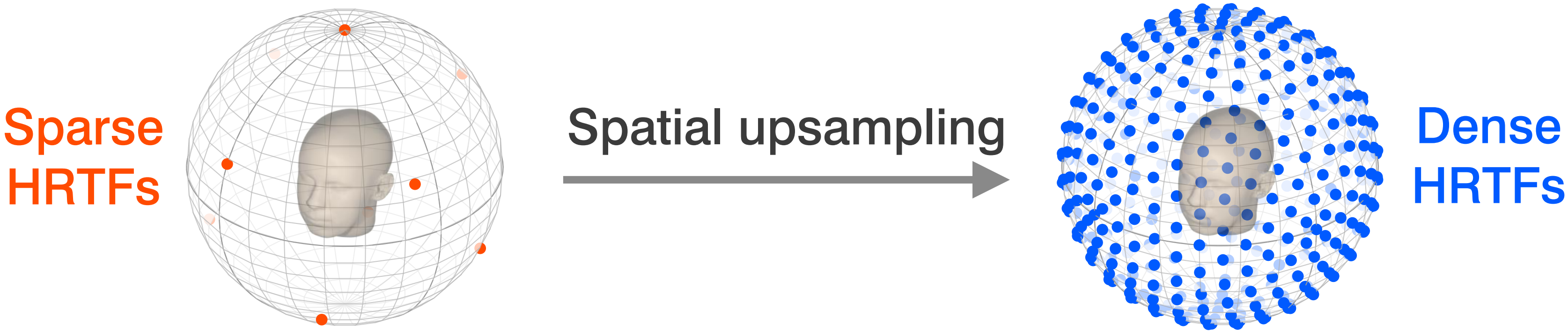
Because of sensitivity to individual differences for sound image localization



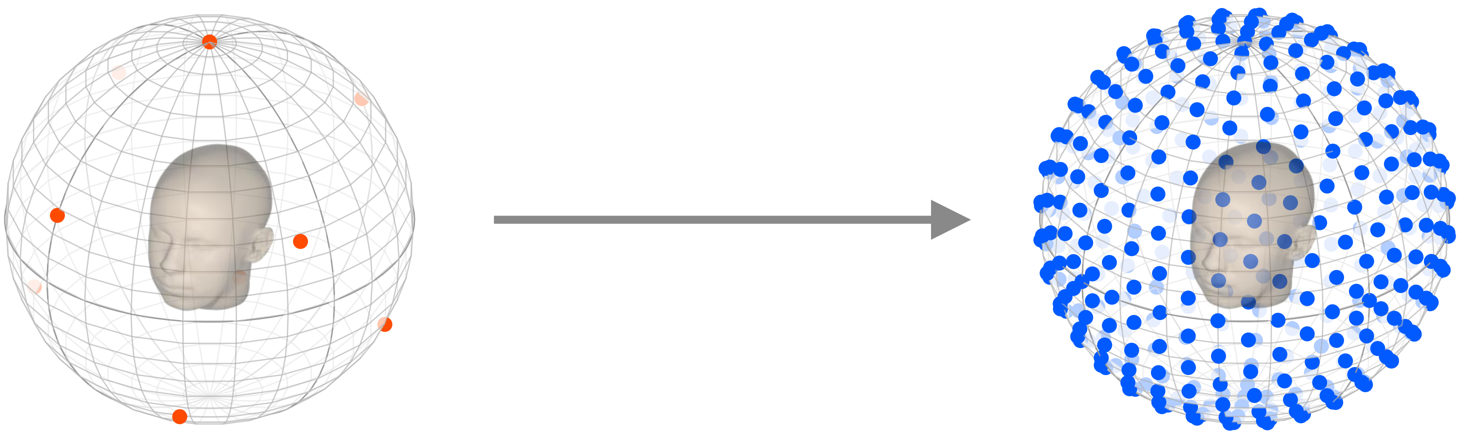
– Long measurement time (60-90 min/pers [Watanabe+14])

Hard to measure HRTF “casually”

HRTF interpolation can reduce measurement costs!



HRTF Interpolation Problem



Sparse HRTFs at
"measurement positions"

$$\{p_{b,l}\}_{\boxed{b\in\mathcal{B}'}} \quad (\mathcal{B}' \subset \mathcal{B},\ |\mathcal{B}'| =: B')$$

Dense HRTFs at
"target positions"

$$\{p_{b,l}\}_{\boxed{b\in\mathcal{B}}}$$

$p_{b,l} \in \mathbb{C}$: HRTFs; acoustic transfer functions from source to ears

Frequency bin $l \in \{1, \ldots, L\}$

Source position $b \in \{1, \ldots, B\} =: \mathcal{B}$

Regularized-Linear-Regression(RLR)-based Method

[Duraiswami+04]

Resynthesizes HRTFs at target positions from expansion coefficients

Sparse HRTFs $\boldsymbol{p}_l$ (B')

Expansion coefficients $\hat{\boldsymbol{c}}_l$

Dense HRTFs $\hat{\boldsymbol{p}}_l^{(\text{target})}$ (B)

[Step 1] solve RLR problem

$$\hat{\boldsymbol{c}}_l = \underset{\boldsymbol{c}_l \in \mathbb{C}^{(N+1)^2}}{\operatorname{argmin}} \|\boldsymbol{p}_l - \boldsymbol{\Phi}_l \boldsymbol{c}_l\|_2^2 + \underbrace{\lambda \|\boldsymbol{D}^{1/2} \boldsymbol{c}_l\|_2^2}_{\text{Regularizer}}$$

$$= \left(\boldsymbol{\Phi}_l^{\mathsf{H}} \boldsymbol{\Phi}_l + \lambda \boldsymbol{D}\right)^{-1} \boldsymbol{\Phi}_l^{\mathsf{H}} \boldsymbol{p}_l$$

[Step 2]

$$\hat{\boldsymbol{p}}_l^{(\text{target})} = \boldsymbol{\Phi}_l^{(\text{target})} \hat{\boldsymbol{c}}_l$$

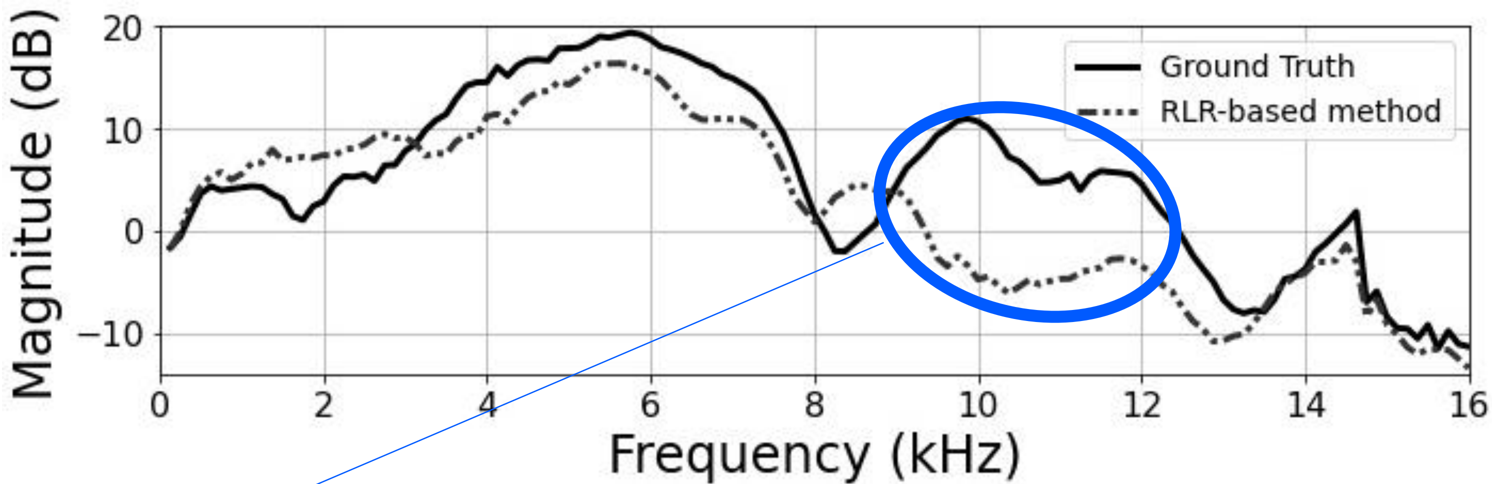
$\boldsymbol{\Phi}_l$: $B' \times (N+1)^2$; $\boldsymbol{\Phi}_l^{(\text{target})}$: $B \times (N+1)^2$

Matrix consisting of spherical wavefunctions

N : truncation order

Limitation of RLR-based Method

HRTF estimated from $B' = 9$ measurement positions by RLR-based method



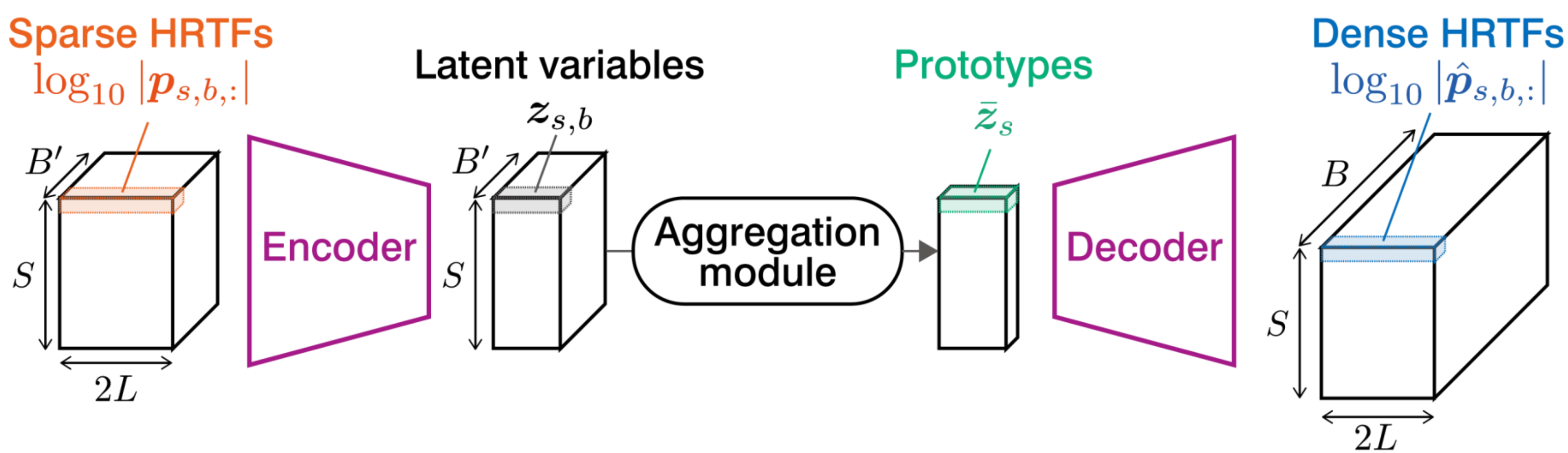
Loss of peak may lead to failure of sound image localization!

When measurement positions are quite sparse, i.e. B' is small, RLR-based method can perform badly…

Our challenge:
HRTF interpolation from highly sparse measurements

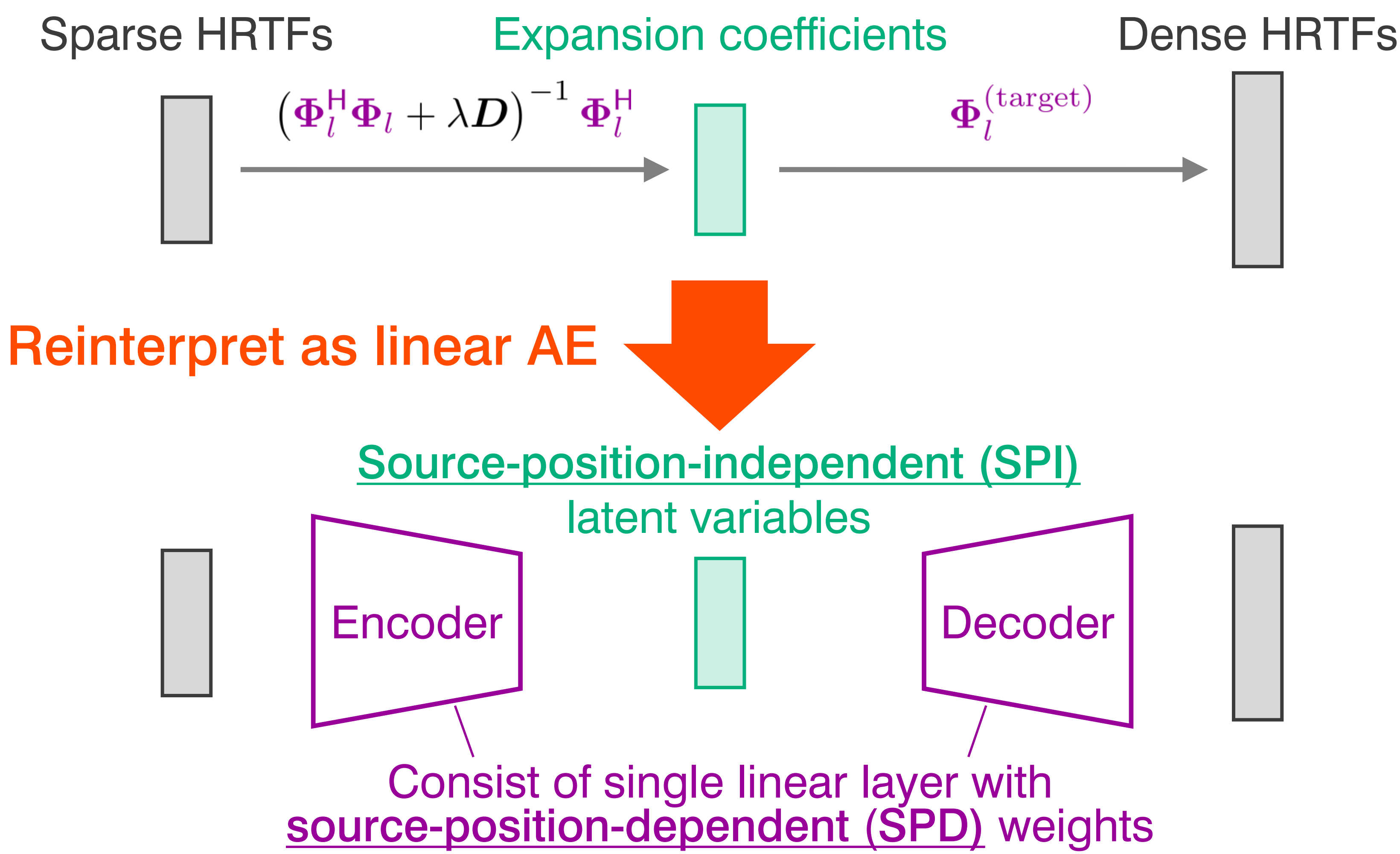
Proposed Method

- Our strategy: training-based method using **deep neural network (DNN)**
  - Promising results for other HRTF-related tasks [Hu+08, Chun+17, Chen+19, Zhang+20, Xi+21]
  - Consists of 2 steps:
    1. Train DNN using HRTFs of subjects in training data
    2. Interpolate HRTFs of subjects of interest, unseen for model
  - Question: How should we design network architecture?
    → **Our focus: analogy with RLR-based method**



Our Focus: Reinterpretation of RLR-based method

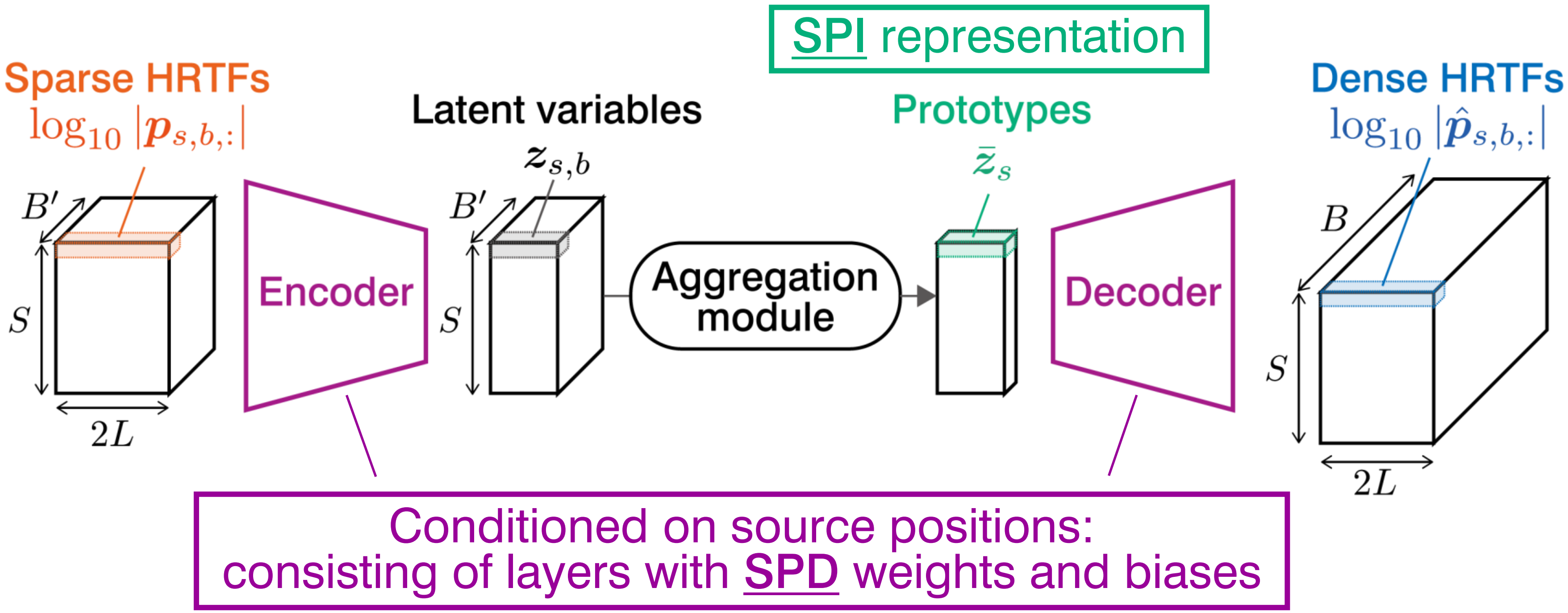
■ RLR-based method as linear autoencoder (AE)



Proposed Model Architecture

■ Overview

– Operates in magnitude domain, similarly to DNN-based methods for HRTF-related tasks [Chen+19, Xi+21]

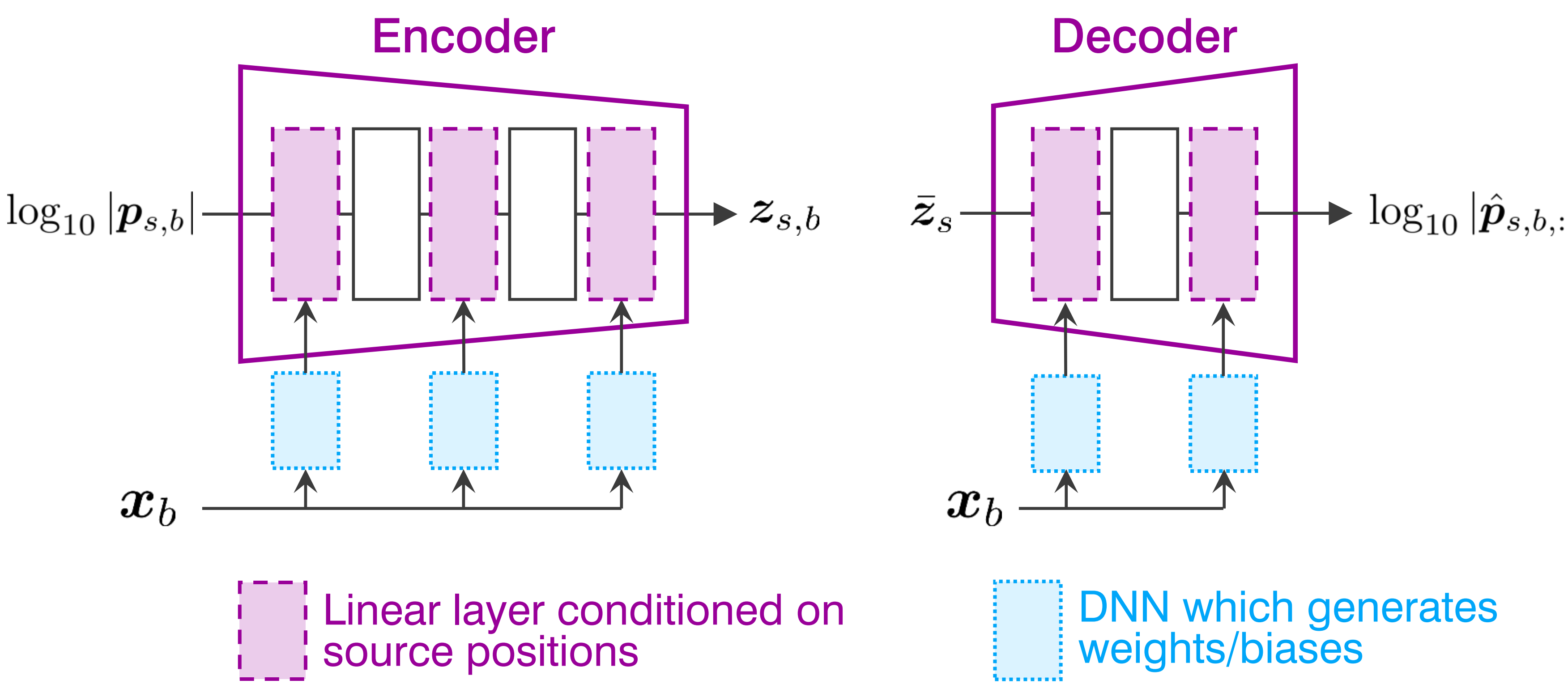


* $s \in \{1, \ldots, S\}$: index of subjects

Encoder & Decoder

■ Both conditioned on source positions

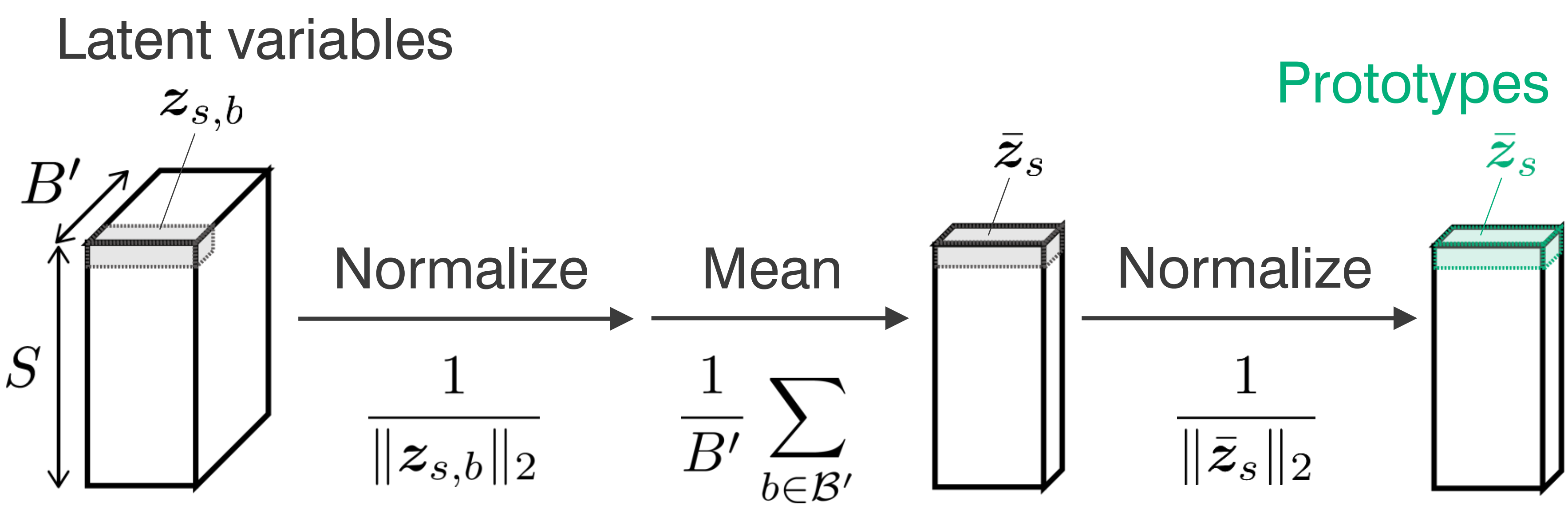
- Hypernetworks [Ha+17]: generate weights/biases of layers from auxiliary information aside associated with input
- We can use measurement positions in 3D Cartesian coordinates $\boldsymbol{x}_b$ as auxiliary information.



Aggregation Module

- Create SPI "prototype"
  - Prototype can be used as representation of subject
  - Inspired by prototypical networks for few-shot learning

[Snell+17]



Proposed Loss Function

Loss function

Hyperparameter (≥ 0)

$$\mathcal{L} = \mathrm{LSD} + \alpha\ \mathrm{CosDist}$$

- Log-spectral distortion: term to make estimated HRTFs closer to ground truth

Estimated

$$\mathrm{LSD} := \frac{1}{SB} \sum_{s,b} \sqrt{\frac{1}{L} \sum_{l} \left(20 \log_{10} \frac{|\hat{p}_{s,b,l}|}{|p_{s,b,l}|} \right)^2}$$

Ground Truth

- Term to promote latent variables $\boldsymbol{z}_{s,b}$ to be distributed near prototype $\bar{\boldsymbol{z}}_s$

$$\mathrm{CosDist} := \sqrt{\frac{1}{SB'} \sum_{s,b} \left(1 - \frac{\boldsymbol{z}_{s,b}^{\mathsf{T}} \bar{\boldsymbol{z}}_s}{\|\boldsymbol{z}_{s,b}\|_2 \|\bar{\boldsymbol{z}}_s\|_2} \right)^2}$$

Experimental Setting

Objective
- To evaluate effectiveness of proposed method

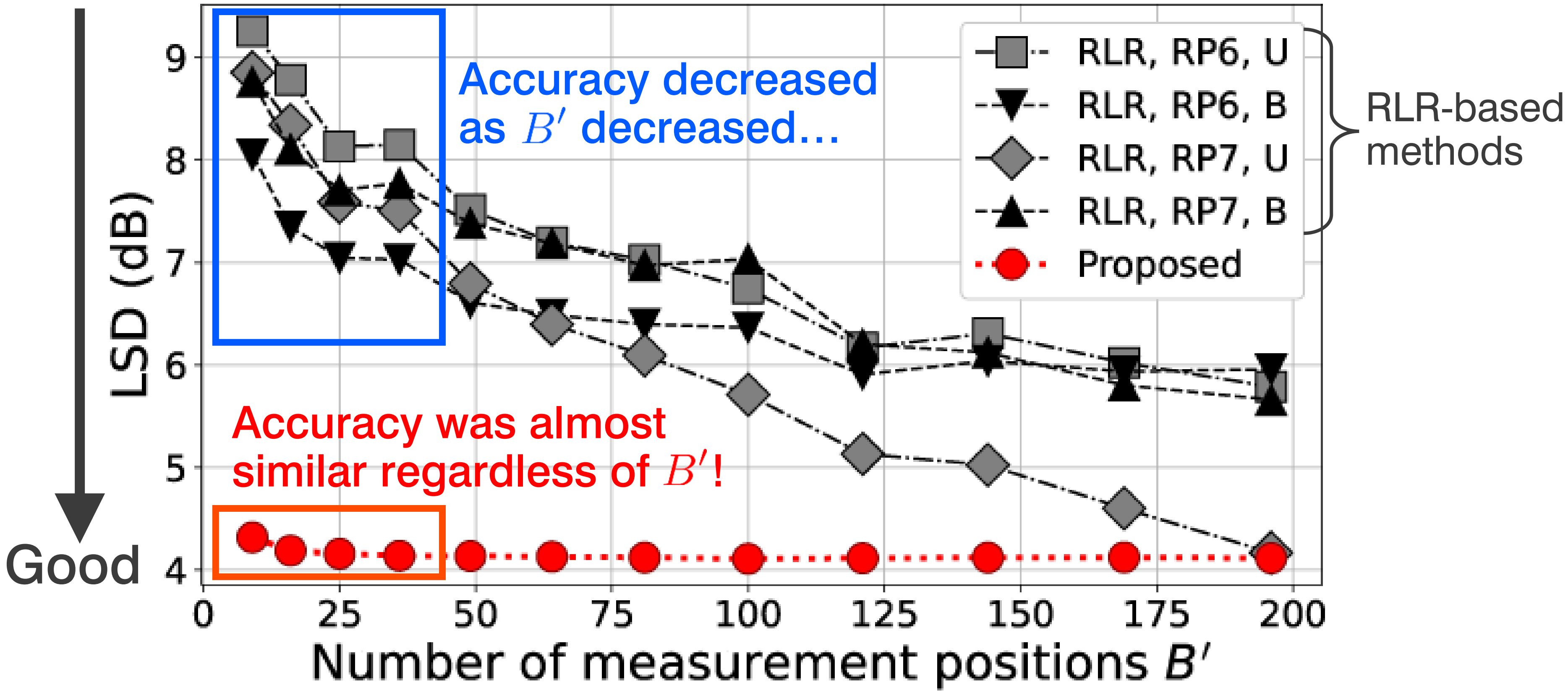
Data
- Dataset: HUTUBS [Brinkmann+19]
- Head-related impulse responses (HRIRs) at 440 points on sphere (radius: 1.47 m)
- Converted to HRTFs at 128 frequency bins 125 Hz, …, 16 kHz
- Used 77, 10, and 7 subjects for training, validation, and test

Tasks
- Estimation of HRTFs at 440 target positions from HRTFs at $B' =$ 9, 16,…, 196 measurement positions for test data

  nearly-uniformly sampled
- Evaluation metric: LSD

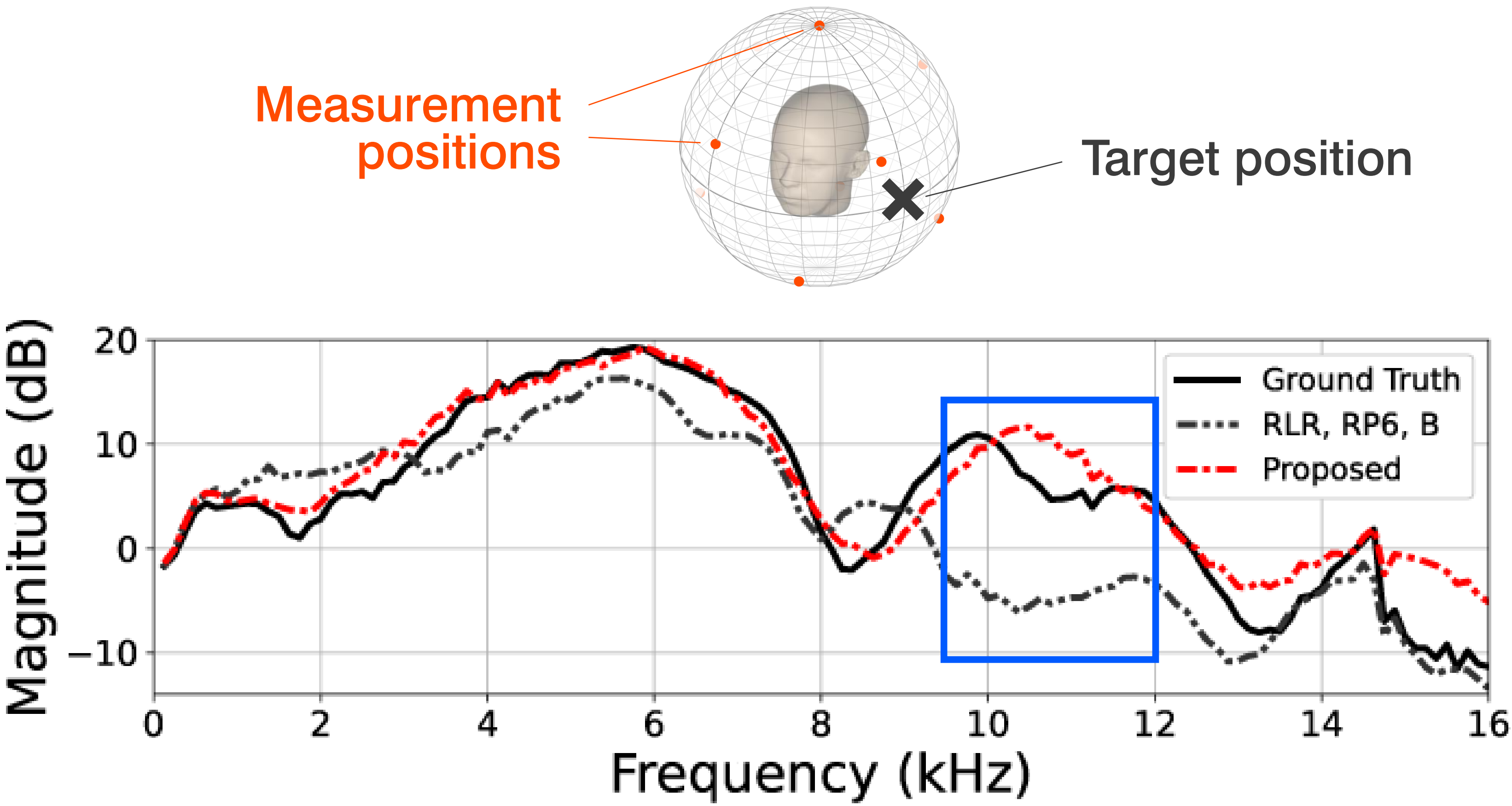
Compared methods
- RLR-based method: 4 configurations
- Proposed method
  - Trained with Adam for 1000 epochs, learning rate: 1e-3, early stopping
  - $B' = 440$ during training and validation, hyperparameter $\alpha = 1$

Results: LSDs



Proposed method can interpolate HRTFs at only $B' = 25$ as accurately as RLR-based method at $B' = 196$

Example of Obtained HRTFs with $B'=9$



– Proposed: appropriately captured peaks and notches
– RLR-based: failed to capture peak around 10∼12 kHz

Conclusion

- **Objective: to interpolate HRTFs accurately from spatially sparse measurements**
  - If achieved, it will reduce measurement time greatly
  - RLR-based method can perform badly in such situations
- **Proposed DNN-based HRTF interpolation method**
  - Architecture designed based on our finding: analogy with RLR-based method
  - Autoencoder with source position conditioning
    - Encoder / decoder: source-position-dependent
    - Latent variables: source-position-independent
- **Numerical Experiments**
  - Proposed method can work well for unseen subjects
  - Proposed method outperforms RLR-based methods, especially when measurement positions are quite sparse

Demo

Slides