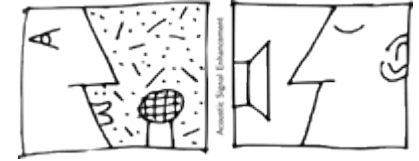


L-02



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



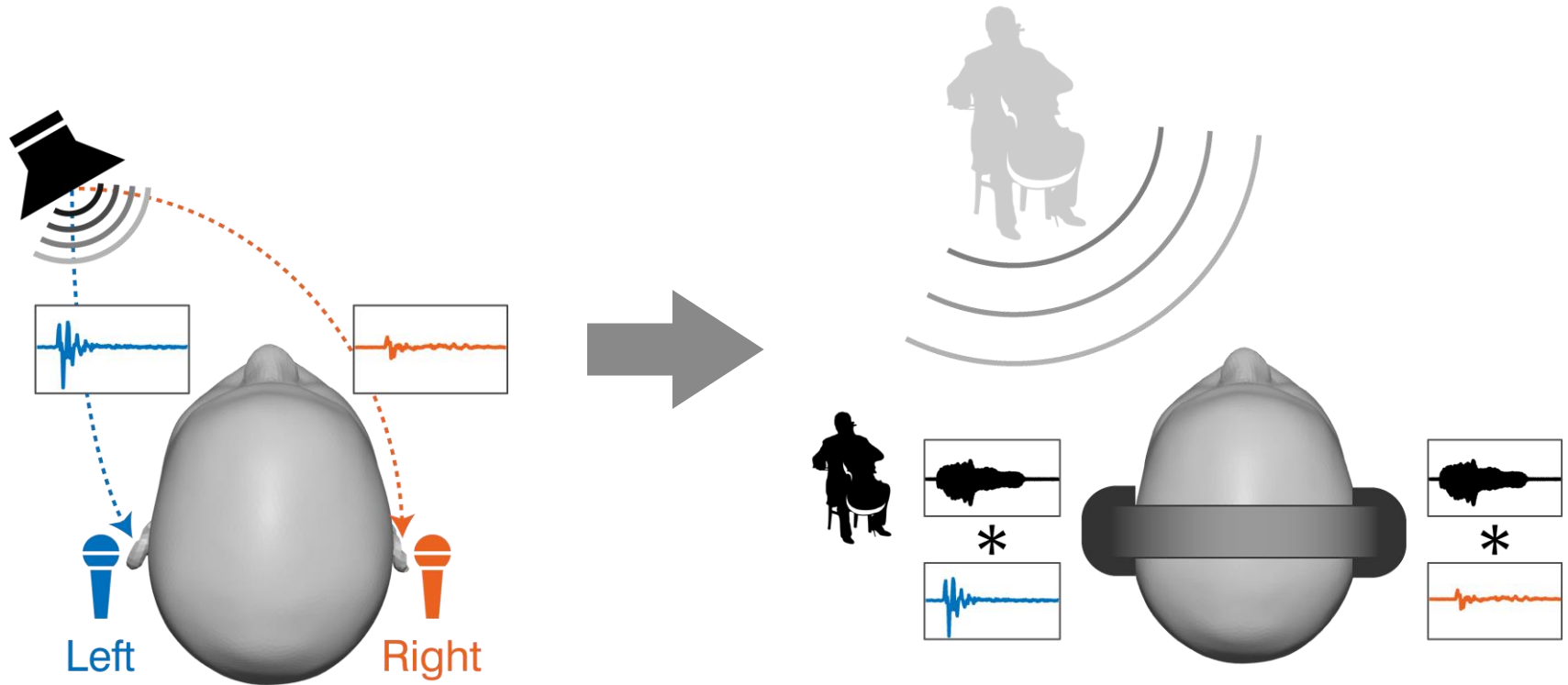
Demo



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

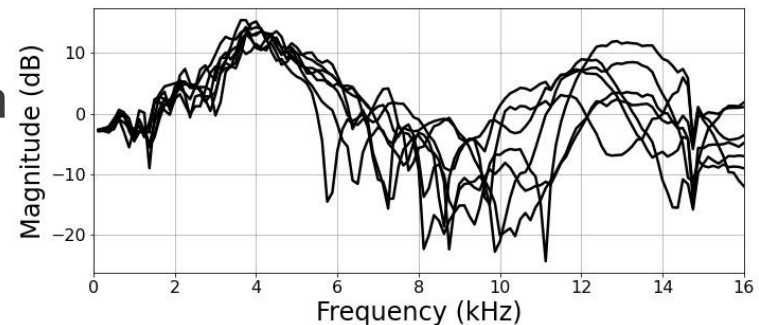
■ HRTF measurement is costly

- **Necessity of person-by-person measurement**

Because of sensitivity to individual differences for sound image localization

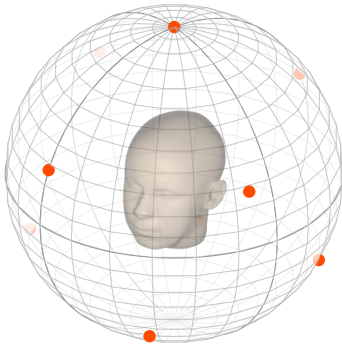
- **Long measurement time**
(60-90 min/pers [Watanabe+14])

Hard to measure HRTF “casually”

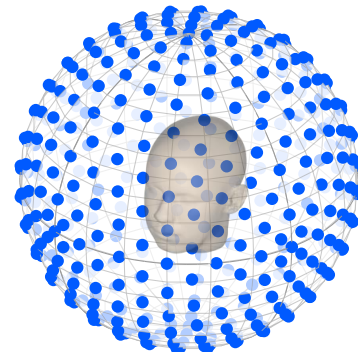


HRTF interpolation can reduce measurement costs!

Sparse
HRTFs

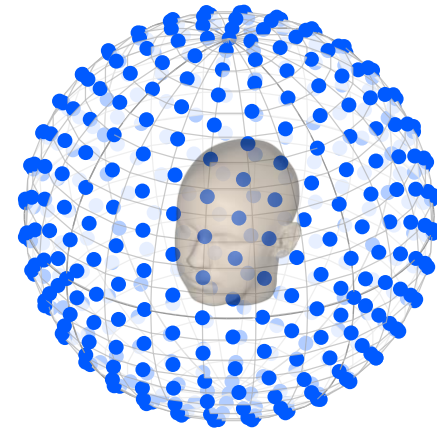
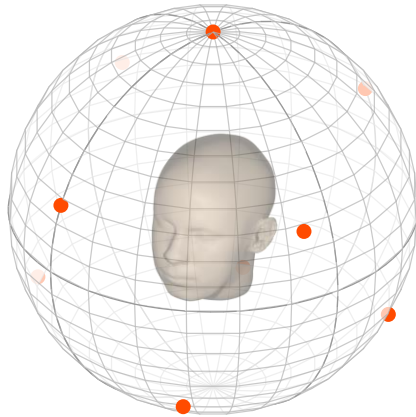


Spatial upsampling



Dense
HRTFs

HRTF Interpolation Problem



Sparse HRTFs at
 “**measurement positions**”

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

($\mathcal{B}' \subset \mathcal{B}$, $|\mathcal{B}'| =: B'$)

Dense HRTFs at
 “**target positions**”

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

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

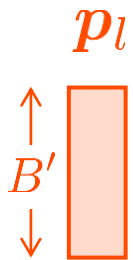
- └ Frequency bin $l \in \{1, \dots, L\}$
- └ Source position $b \in \{1, \dots, B\} =: \mathcal{B}$

Regularized-Linear-Regression(RLR)-based Method

[Duraiswami+04]

Resynthesizes **HRTFs at target positions**
from **expansion coefficients**

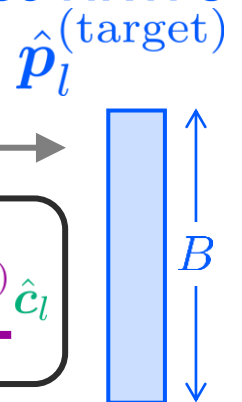
Sparse HRTFs



Expansion coefficients

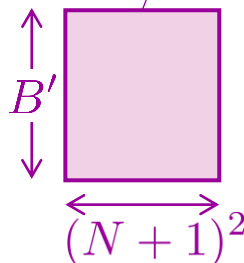


Dense HRTFs



[Step 1] solve RLR problem

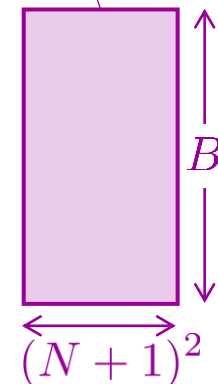
$$\hat{c}_l = \underset{c_l \in \mathbb{C}^{(N+1)^2}}{\operatorname{argmin}} \left\| p_l - \Phi_l c_l \right\|_2^2 + \lambda \underbrace{\left\| D^{1/2} c_l \right\|_2^2}_{\text{Regularizer}}$$
$$= (\Phi_l^H \Phi_l + \lambda D)^{-1} \Phi_l^H p_l$$



Matrix consisting of spherical wavefunctions
 N : truncation order

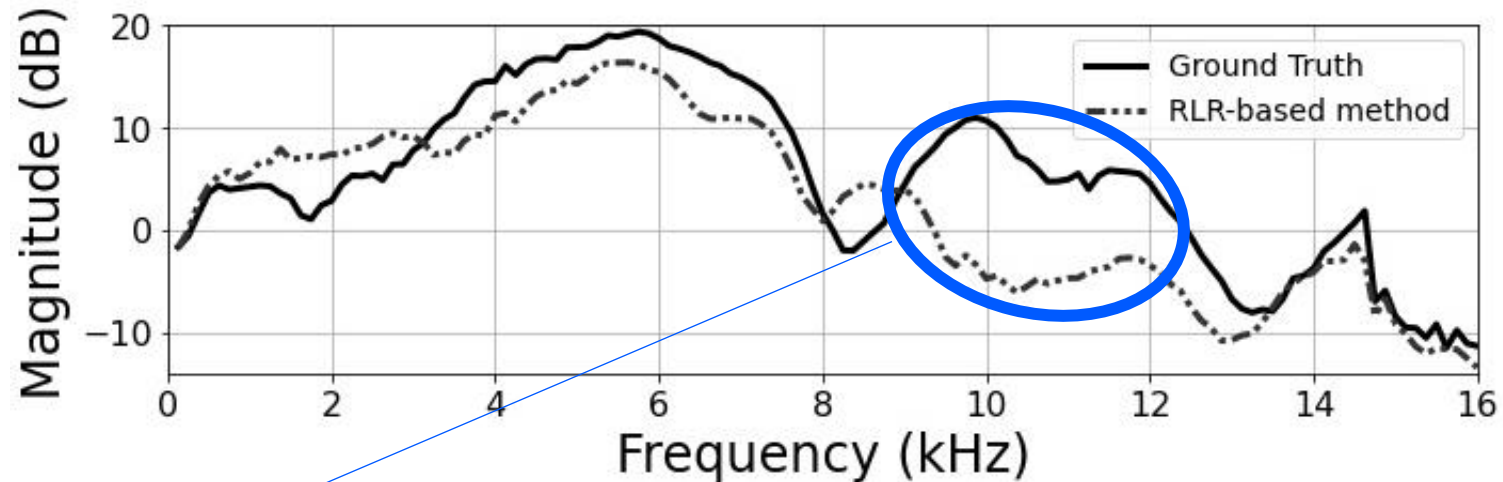
[Step 2]

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



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!

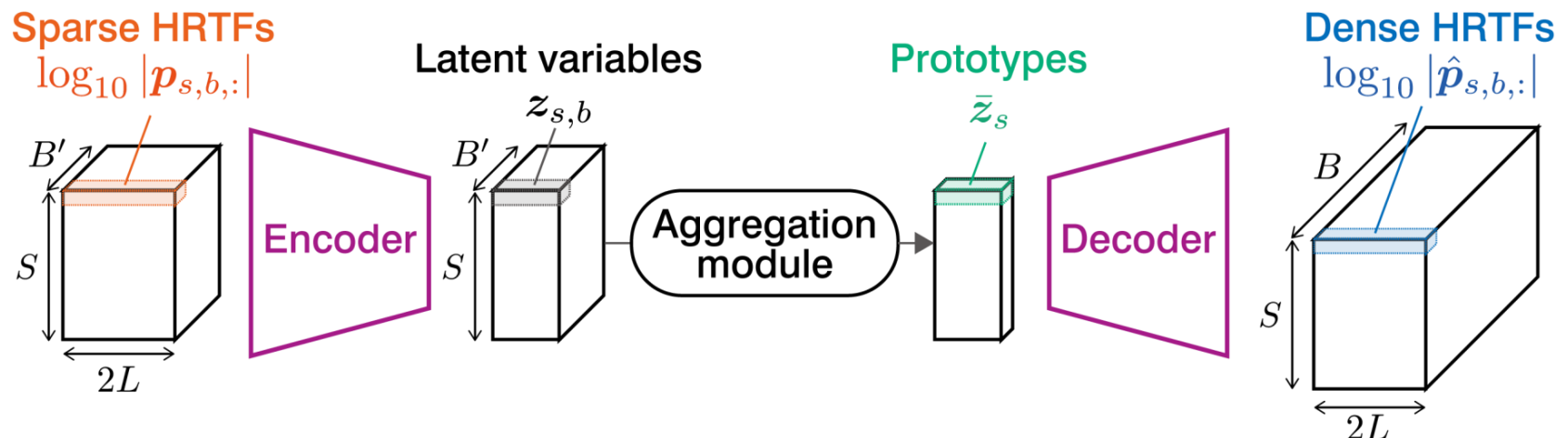
When measurement positions are quite sparse, i.e. B' is small, RLR-based method can perform badly...

Our challenge:
HRTF interpolation from highly sparse measurements

Proposed Method

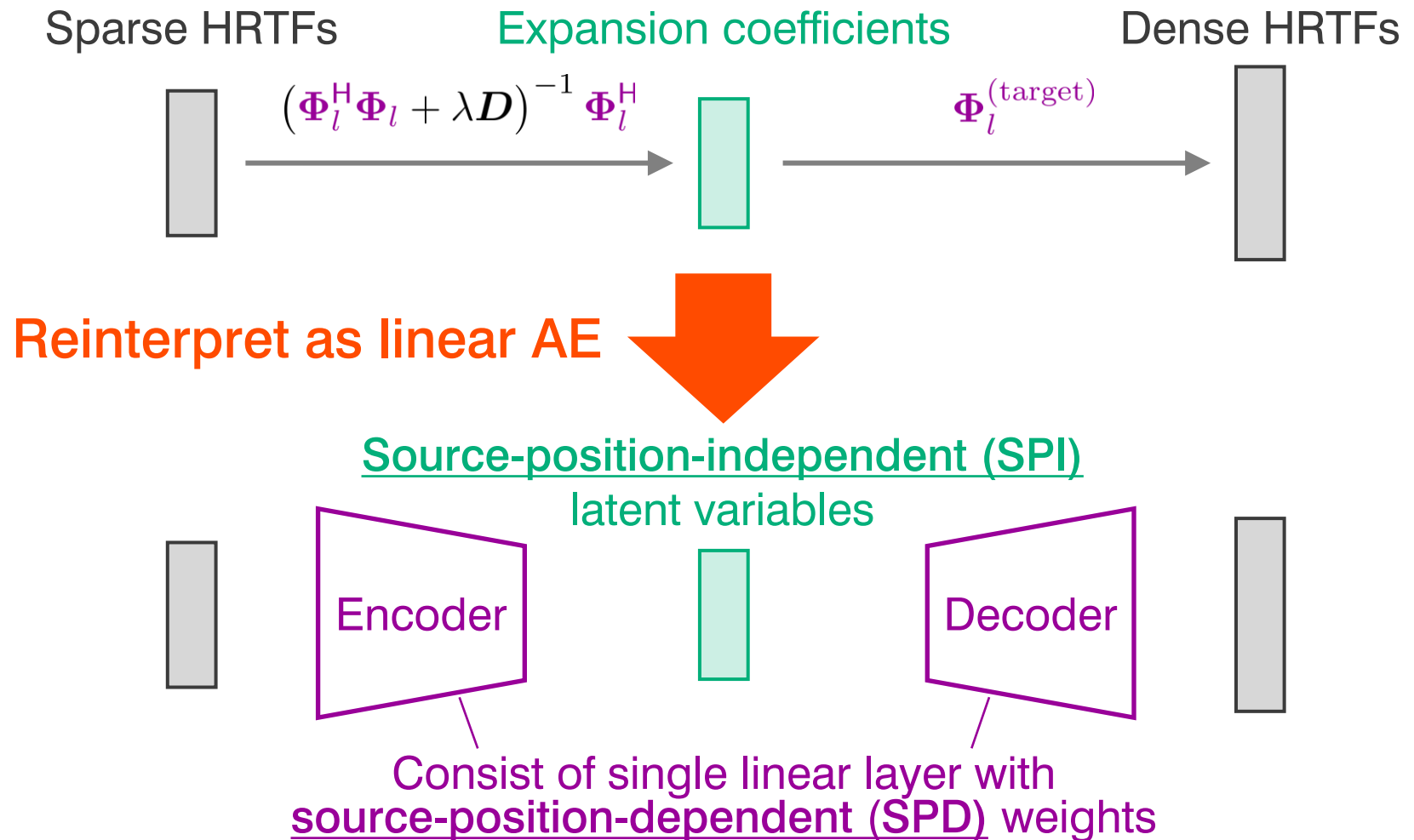
■ Our strategy: training-based method using **deep neural network (DNN)**

- Promising results for other HRTF-related tasks
- 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

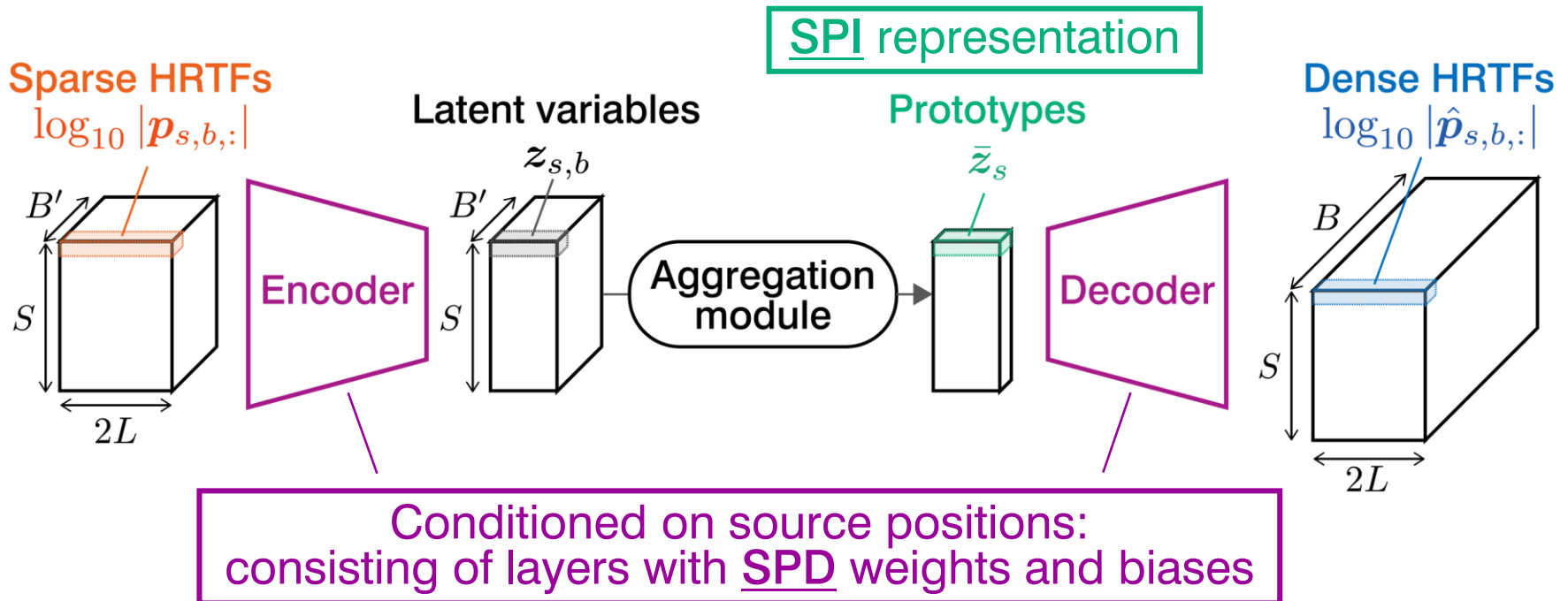
■ RLR-based method as linear autoencoder (AE)



Proposed Model Architecture

■ Overview

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

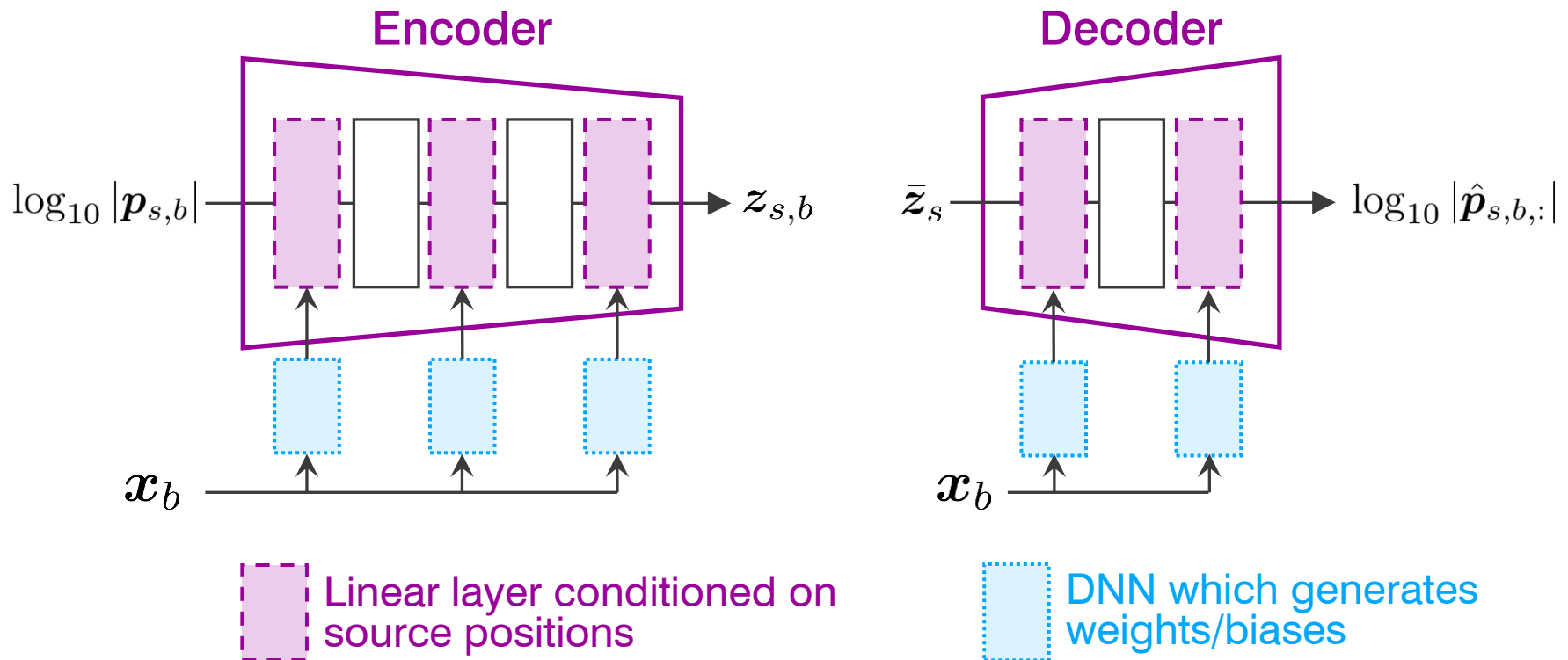


* $s \in \{1, \dots, S\}$: index of subjects

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 \mathbf{x}_b as auxiliary information.



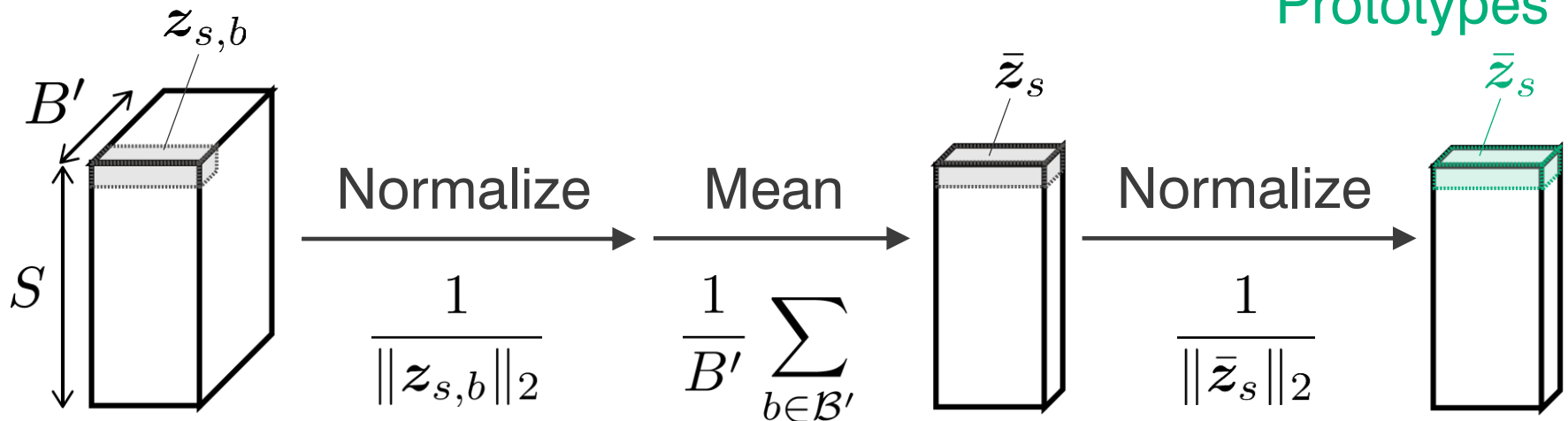
Aggregation Module

■ Create SPI “prototype”

- **Prototype** can be used as representation of subject
- Inspired by prototypical networks for few-shot learning

[Snell+17]

Latent variables



Proposed Loss Function

■ Loss function

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

Hyperparameter (≥ 0)

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

$$\text{LSD} := \frac{1}{SB} \sum_{s,b} \sqrt{\frac{1}{L} \sum_l \left(20 \log_{10} \frac{\overbrace{|\hat{p}_{s,b,l}|}^{\text{Estimated}}}{\underbrace{|p_{s,b,l}|}_{\text{Ground Truth}}} \right)^2}$$

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

$$\text{CosDist} := \sqrt{\frac{1}{SB'} \sum_{s,b} \left(1 - \frac{z_{s,b}^\top \bar{z}_s}{\|z_{s,b}\|_2 \|\bar{z}_s\|_2} \right)^2}$$

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

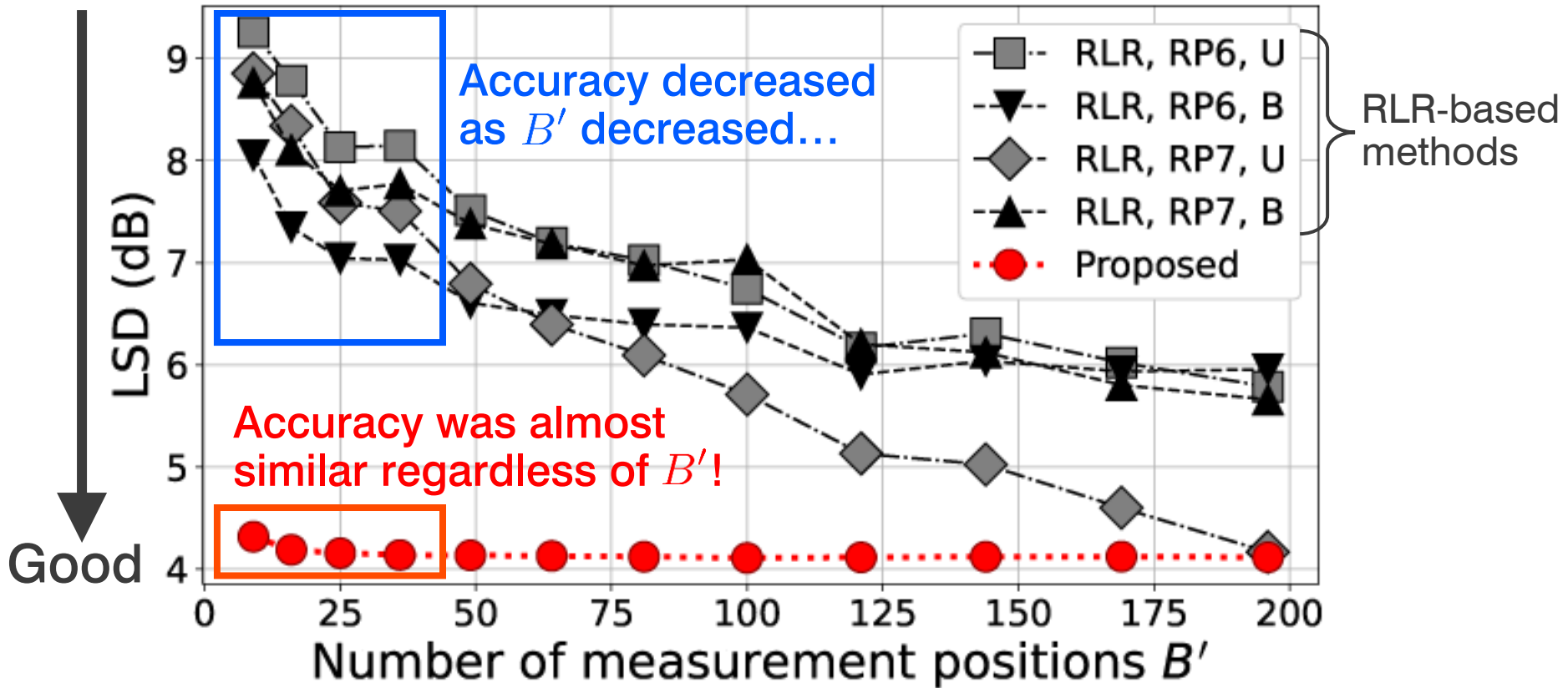
■ Tasks

- Estimation of HRTFs at 440 target positions from HRTFs at $B' = 9, 16, \dots, 196$ measurement positions for test data
- Evaluation metric: LSD  nearly-uniformly sampled

■ Compared methods

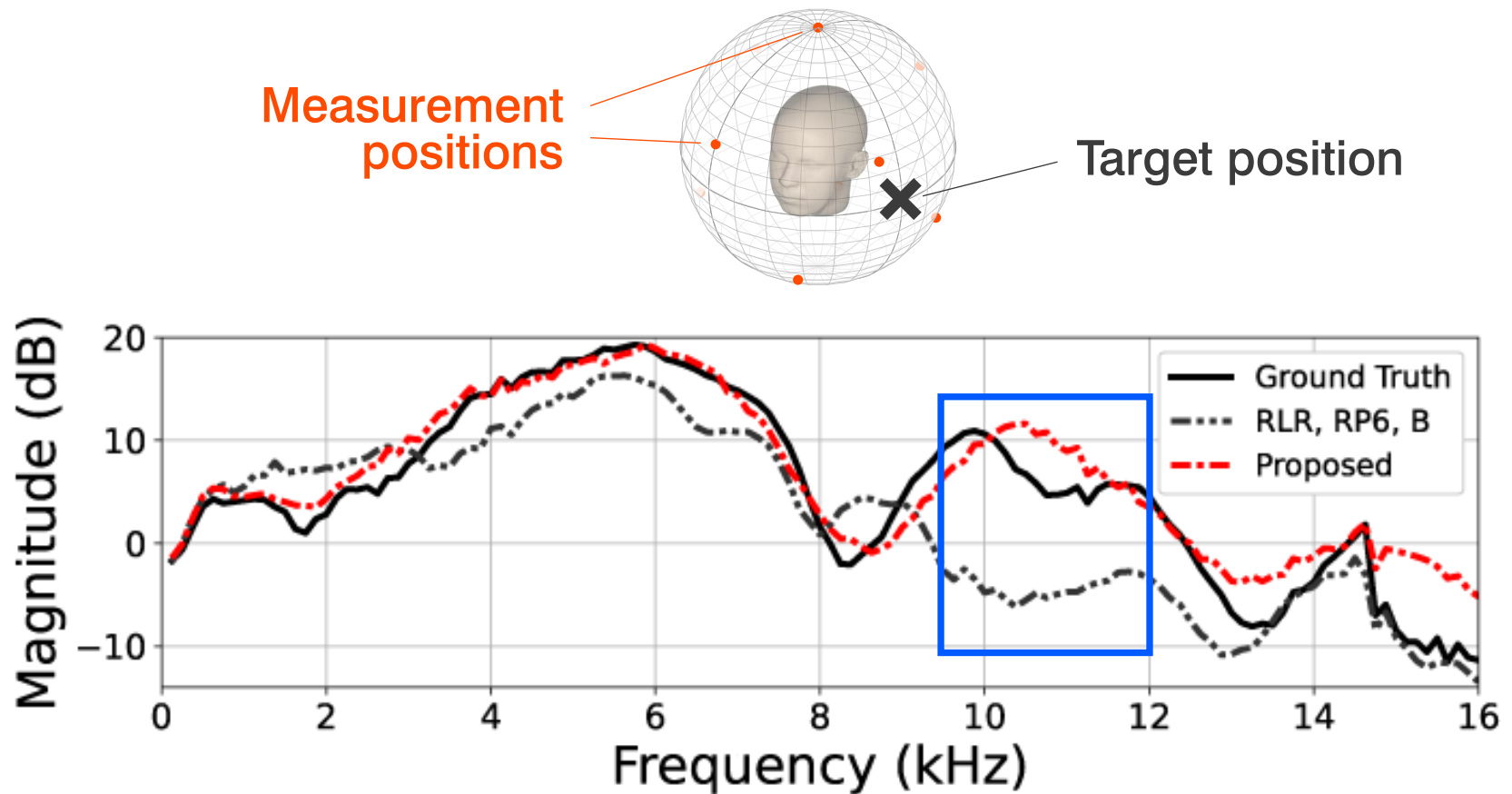
- RLR-based method: 4 configurations
- Proposed method
 - Trained with Adam for 1000 epochs, learning rate: 1e-3, early stopping
 - $B' = 440$ during training and validation, hyperparameter $\alpha = 1$

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