

# Reverberation Modeling for Source-Filter-based Neural Vocoder

Yang Ai<sup>1</sup>, Xin Wang<sup>2</sup>, Junichi Yamagishi<sup>2,3</sup>, Zhen-Hua Ling<sup>1</sup>

<sup>1</sup>NELSLIP, University of Science and Technology of China, Hefei,  
P.R.China

<sup>2</sup>National Institute of Informatics, Japan

<sup>3</sup>CSTR, University of Edinburgh, UK

**Paper ID: 1613, INTERSPEECH 2020**



# Contents

- Background
- Review of previous work
- Theory
- Experiments
- Demos



# Background

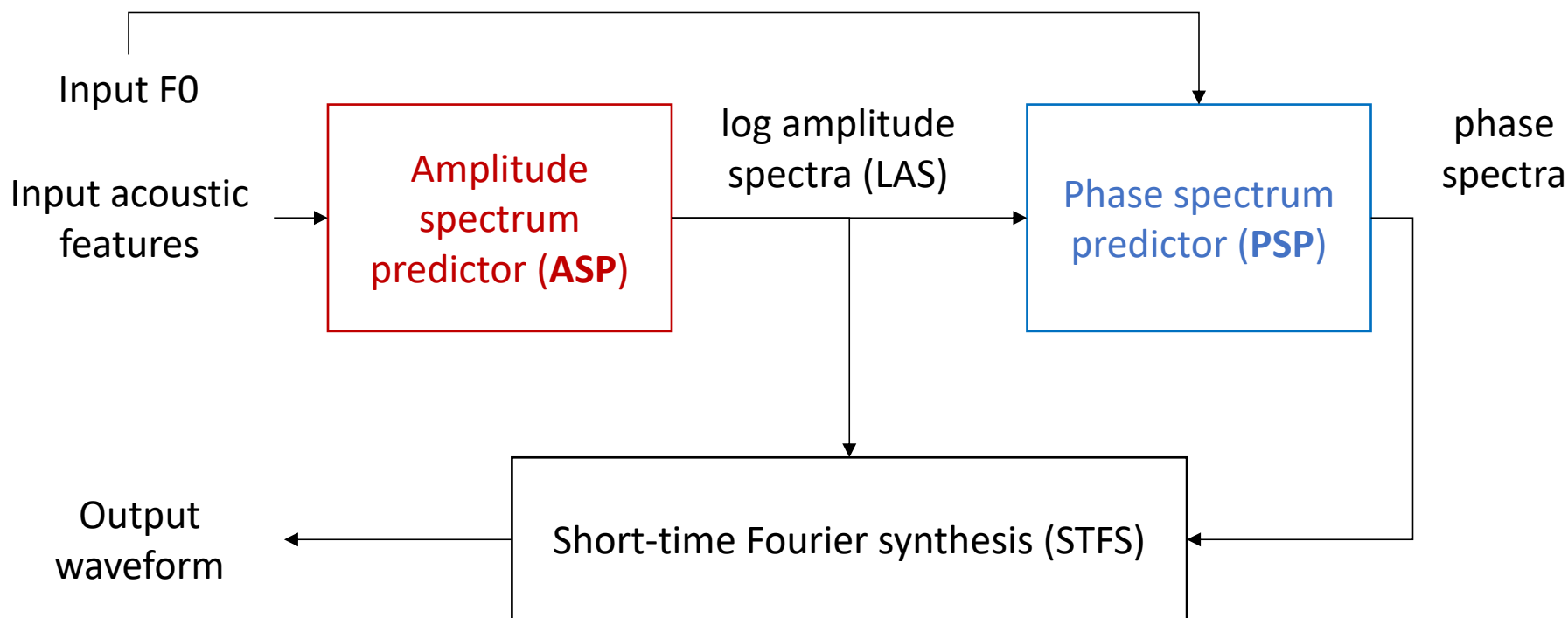
- Audio signals captured for real-life applications typically contain room reverberation;
- The reverberation poses a challenge to non-autoregressive neural vocoders, and the quality of synthesized speech usually degrades because there is no special module for reverberation modeling;
- Towards robust reverberation modeling for speech data, we proposed a trainable reverberation module for neural vocoders.



# Review of previous work

- **HiNet:**

$$x(t) \longleftrightarrow X(j\omega) = |\textcolor{red}{X}(j\omega)|e^{j\angle\textcolor{blue}{X}(j\omega)}$$



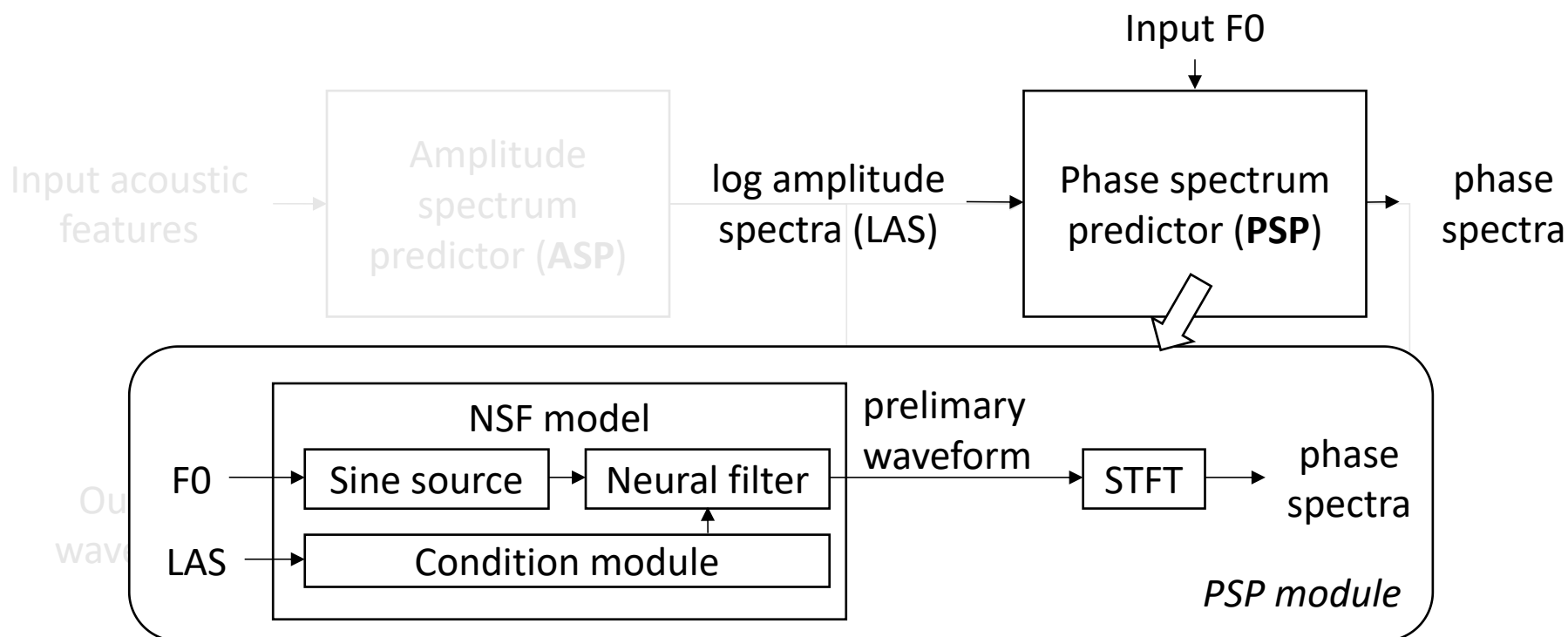
[1] Y. Ai and Z.-H. Ling, "A neural vocoder with hierarchical generation of amplitude and phase spectra for statistical parametric speech synthesis," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 839–851, 2020.



# Review of previous work

- **HiNet:**

- PSP is based on neural source-filter (NSF) model [2]



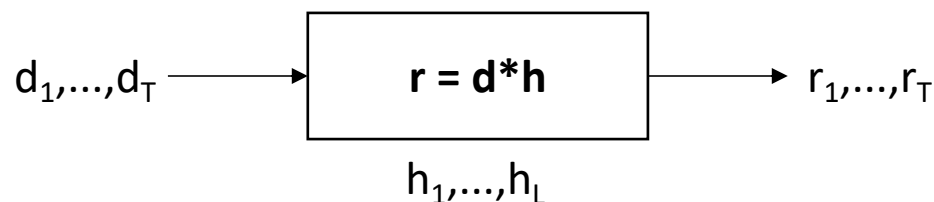
[2] X.Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in Proc. ICASSP, 2019, pp. 5916–5920.



# Proposed methods

- Reverberation theory

- A reverberant signal  $\mathbf{r}=[r_1, \dots, r_T]^T$  can be calculated by convolving clean signal  $\mathbf{d}=[d_1, \dots, d_T]^T$  with room impulse response (RIR)  $\mathbf{h}=[h_1, \dots, h_L]^T$ :



- Time domain convolution:

$$\mathbf{r} = \mathbf{d} * \mathbf{h} \quad (1)$$

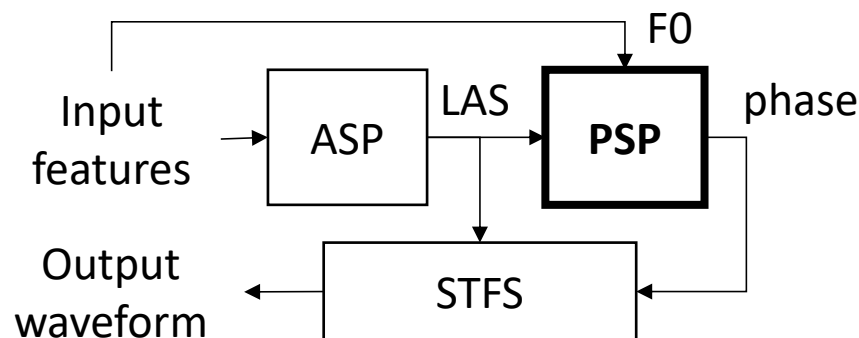
- Equivalently, frequency domain multiplication

$$\mathbf{r} = \mathcal{F}^{-1}[\mathcal{F}(\mathbf{d}) \cdot \mathcal{F}(\mathbf{h})] \quad (2)$$

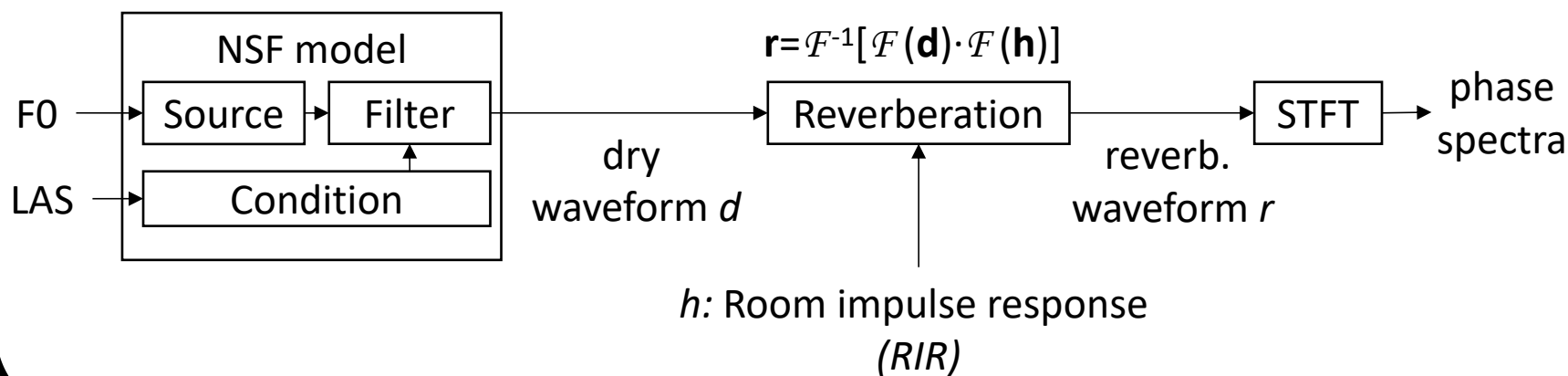


# Proposed methods

- HiNet with Reverberation
  - Only change PSP

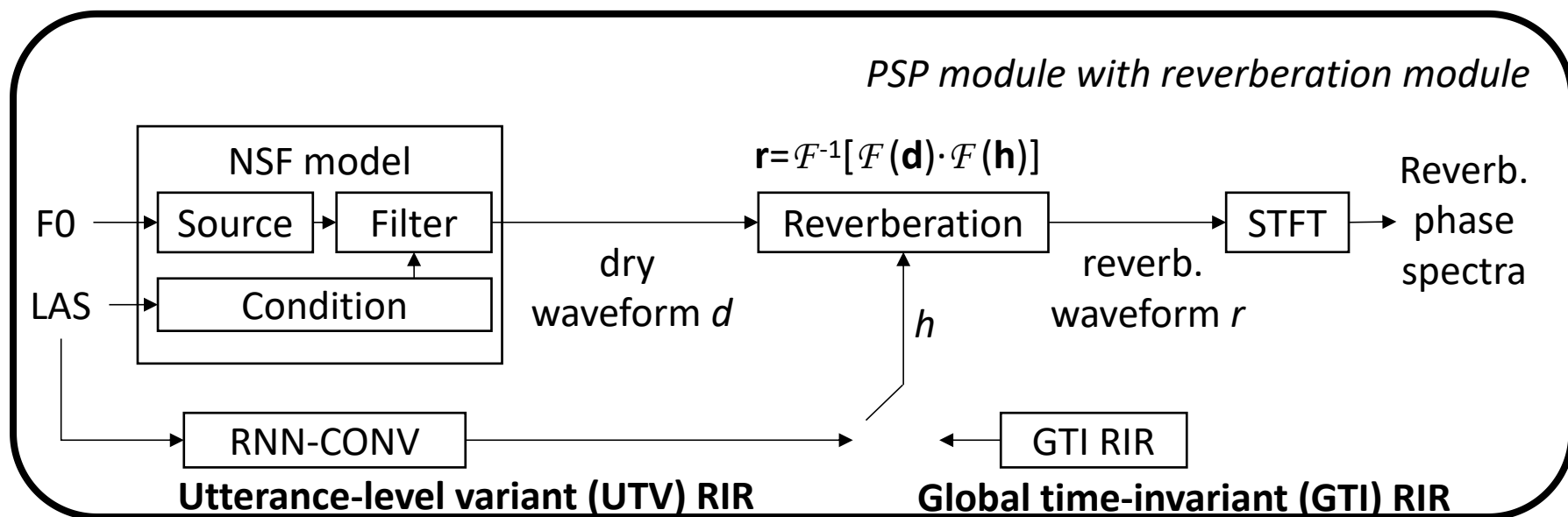
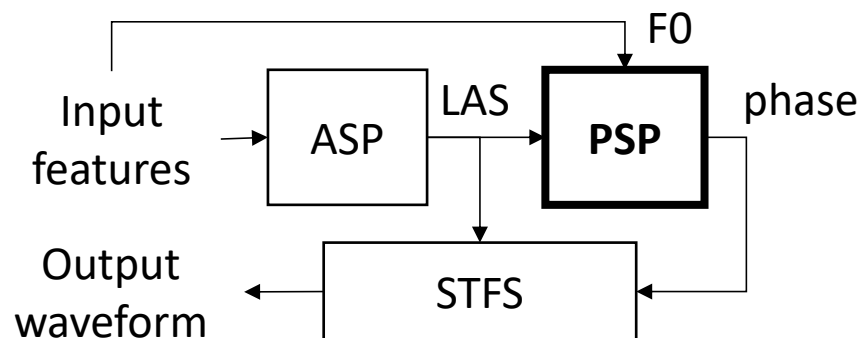


*PSP module with reverberation module*



# Proposed methods

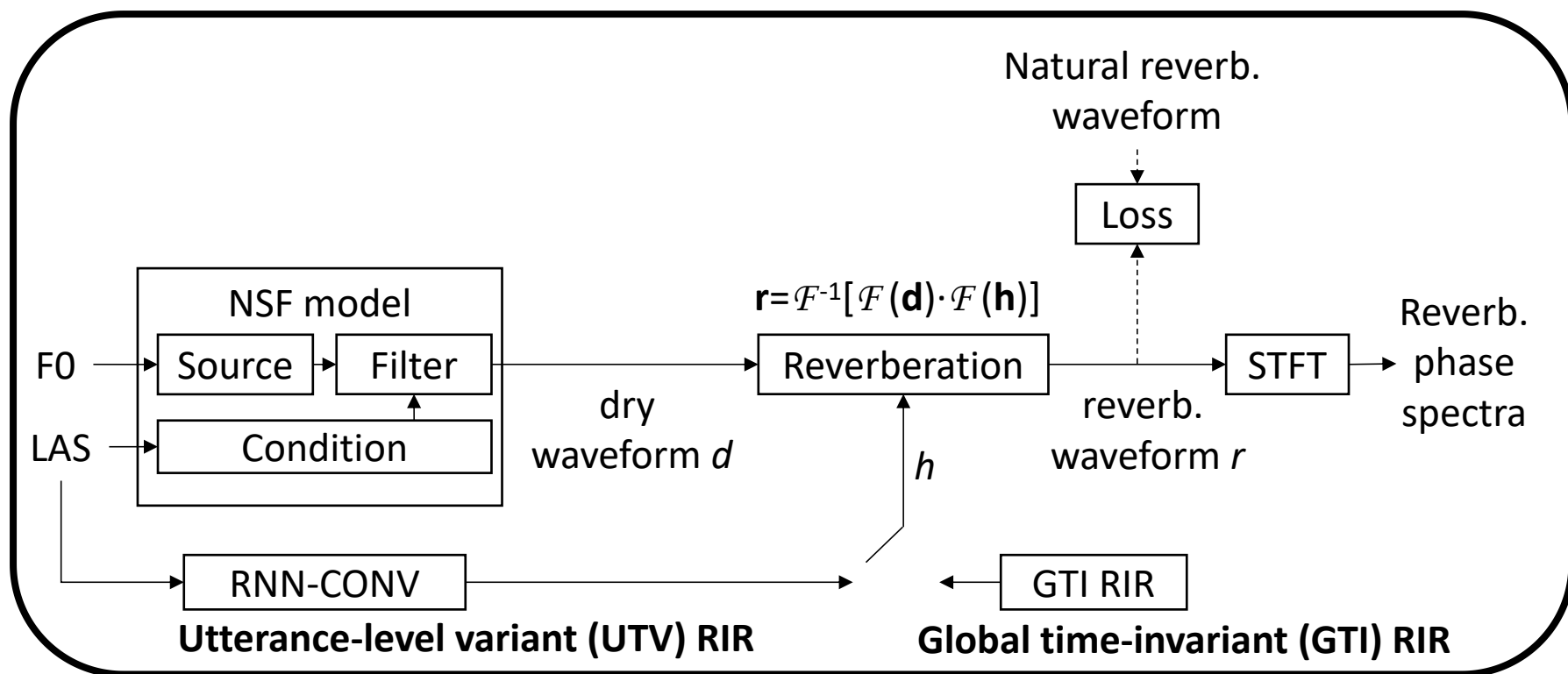
- HiNet with Reverberation
  - Only change PSP
  - Two ways to predict  $h$





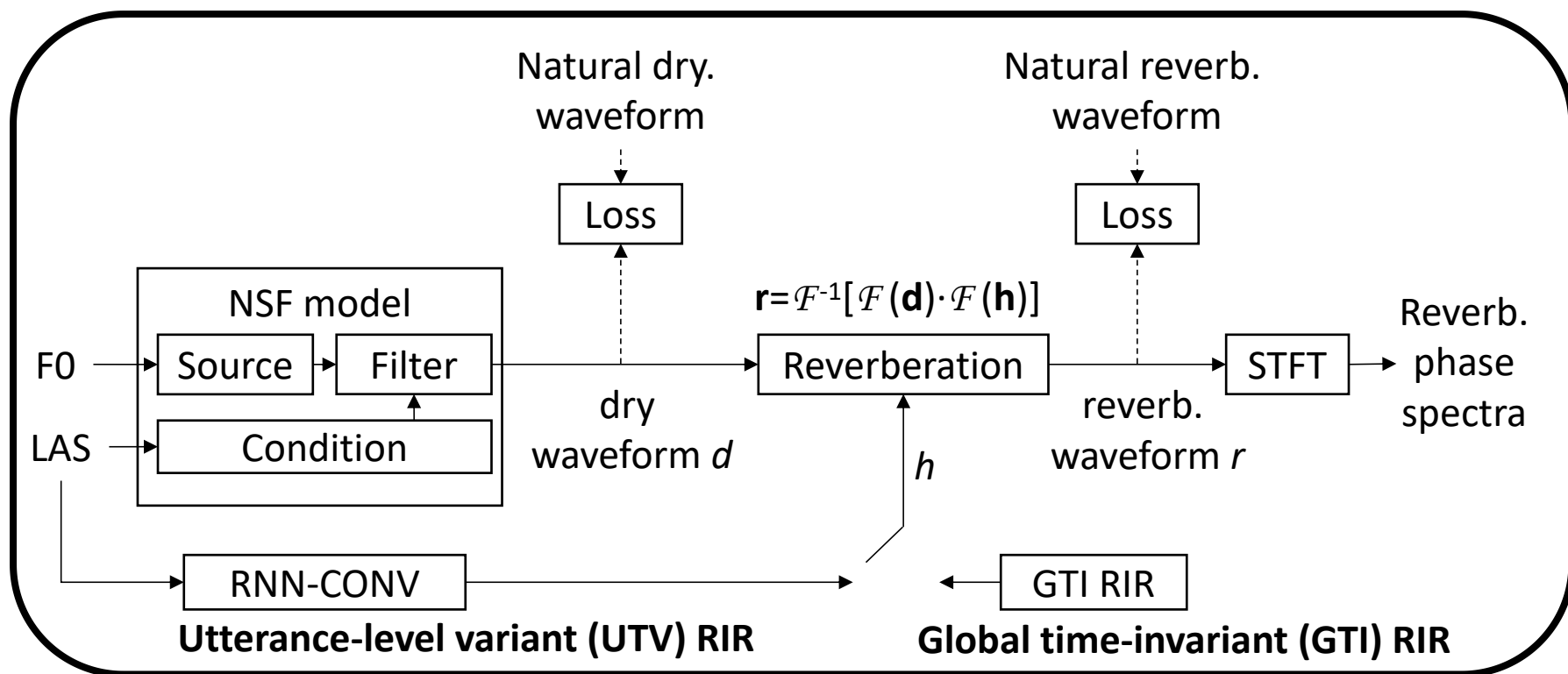
# Proposed methods

- HiNet with Reverberation
  - Normal way of training:



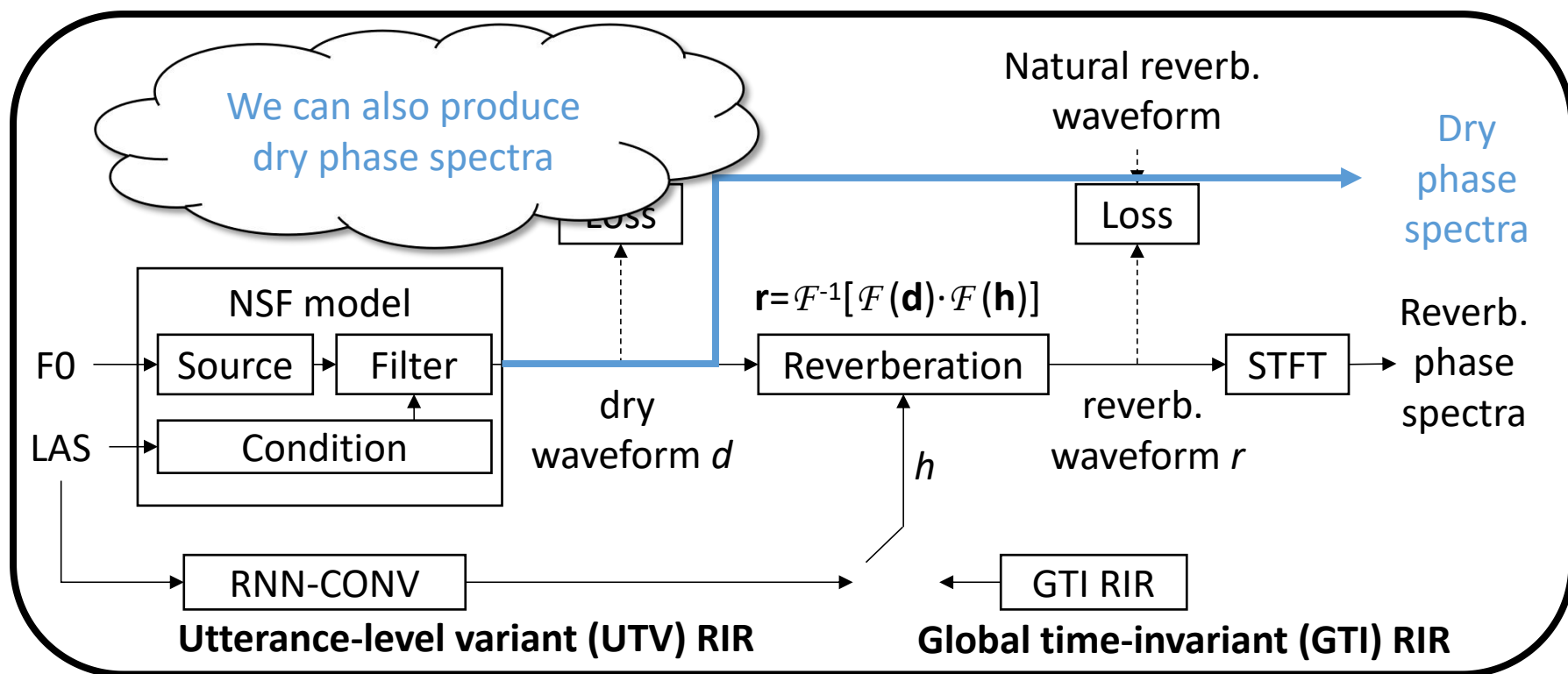
# Proposed methods

- HiNet with Reverberation
  - Multi-task training: if natural dry waveform exists



# Proposed methods

- HiNet with Reverberation
  - Multi-task training: if natural dry waveform exists



# Experiments

- Data and feature configuration
  - Datasets: A multi-speaker reverberant speech database
    - Train set (11012 utterances) + validation set (560 utterances):
      - Including 28 speakers and 18 reverberation types
    - Test set (3 test scenarios):
      - T1: Two unseen speakers' reverberant data with 6 unseen reverberation types (824 utterances);
      - T2: Two unseen speakers' reverberant data with the same 18 reverberation types as in the training set (832 utterances);
      - T3: Dry speech version of T1.
  - Acoustic features
    - Including: 80-dimensional mel-spectrogram, an F0, and a voiced/unvoiced flag.



# Experiments

- Experimental models

- **N-BL**: The harmonic-plus-noise NSF model<sup>[3]</sup> without reverberation module.
- **H-BL**: Baseline HiNet vocoder without reverberation module.
- **H-GTI**: HiNet with the GTI-RIR-based reverberation module integrated into PSP.
- **H-UTV**: HiNet with the UTV-RIR-based reverberation module integrated into the PSP.
- **H-UTV-MT**: same as **H-UTV** but with the secondary task using dry waveforms during training.
  - Note: We use **P-\*** and **P-\*(dry)** to represent the reverberant waveform and dry waveform predicted by PSP in **H-\***, respectively.

[3] X. Wang and J. Yamagishi, “Using cyclic noise as the source signal for neural source-filter-based speech waveform model,” arXiv preprint arXiv:2004.02191, 2020.



# Experiments

- Objective experiments
  - T60 estimation errors for utterances with  $T60n = 0.362s$  under test scenario **T1**
    - T60 is used to measure the reverberation effect and we used an open source toolkit<sup>[4]</sup> to blindly estimate T60 from the reverberant speech
    - T60 estimation error = estimated T60 - ground-truth T60 ( $T60n$ )

[4] M. Jeub, "Blind reverberation time estimation," 2015. [Online]. Available: <https://www.mathworks.com/matlabcentral/fileexchange/35740-blind-reverberation-time-estimation>.



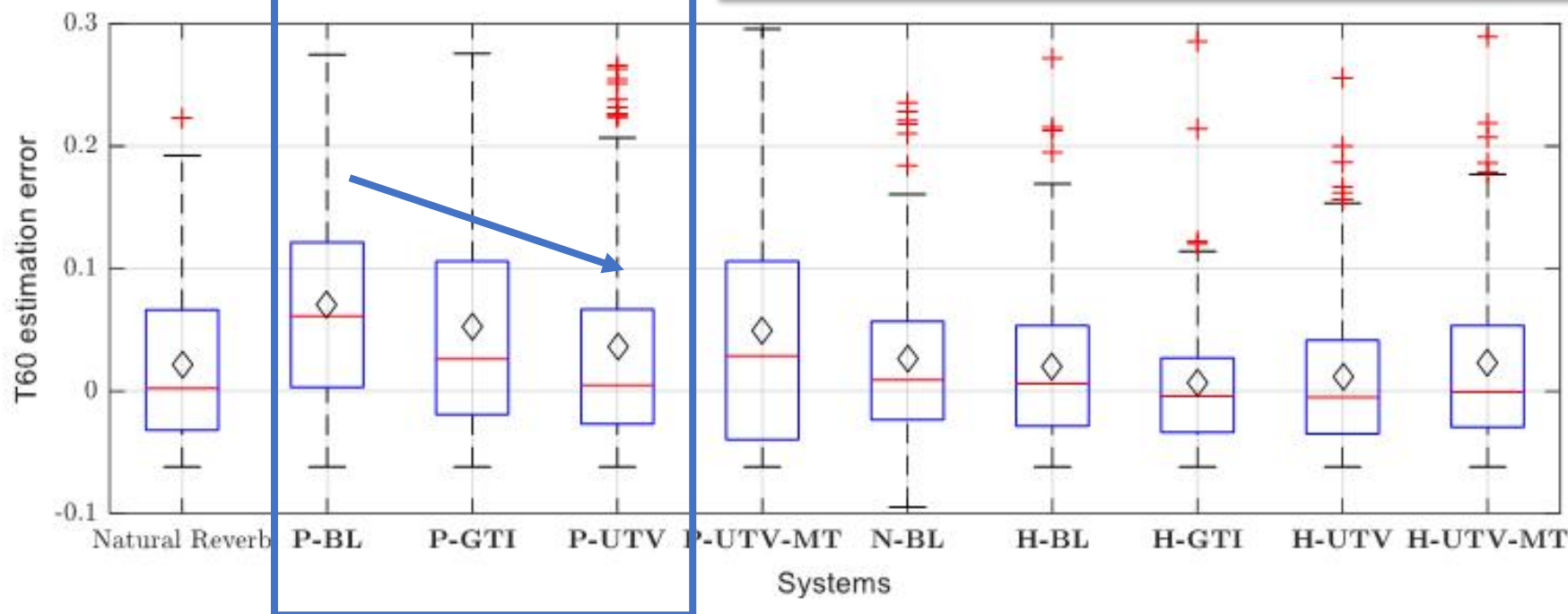
# Experiments

**P-BL:** output waveform of PSP **without** reverberation

**P-GTI:** output waveform of PSP with **global constant** reverberation

**P-UTV:** output waveform of PSP with **utterance variant** reverberation

1. Reverberation module is effective
2. Utterance variant reverberation is effective

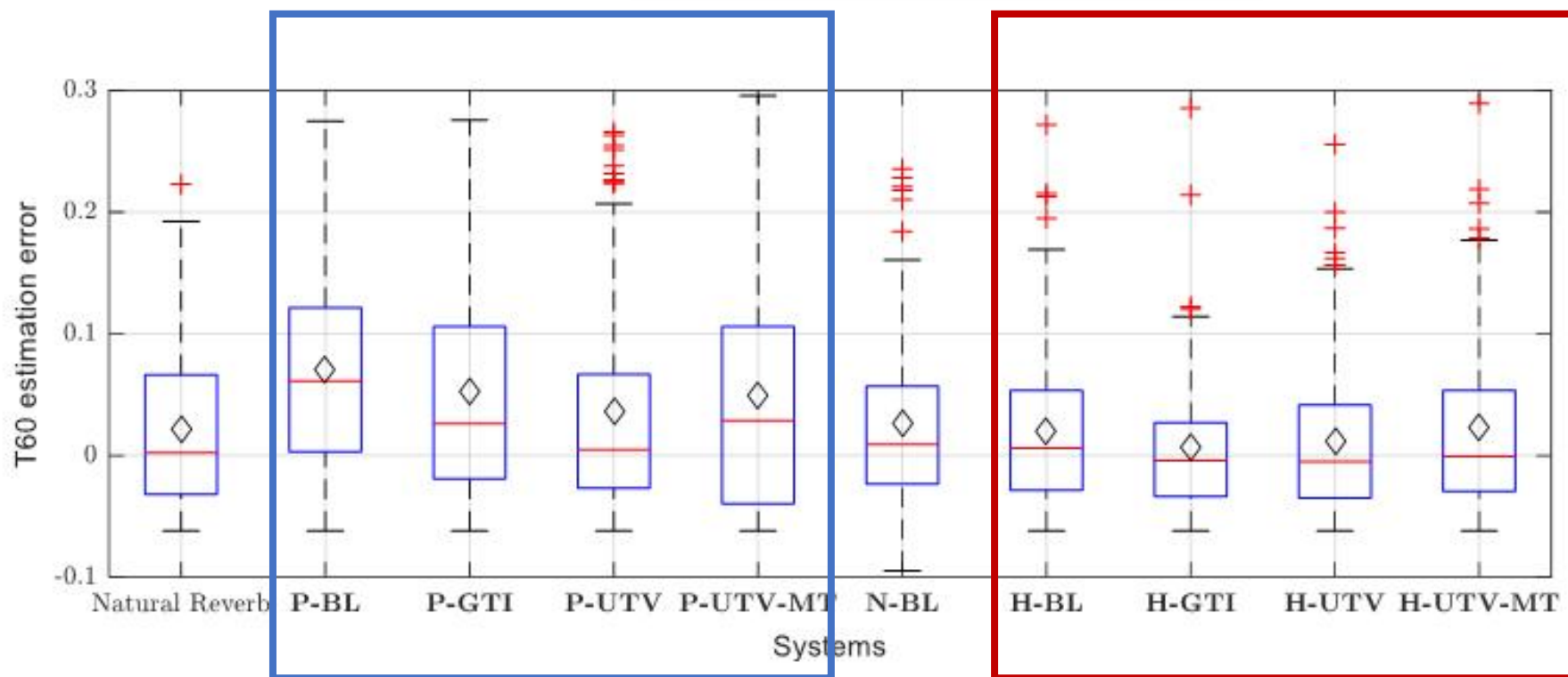


# Experiments

**P\***: output waveforms from **PSP**

**H\***: output waveforms from PSP + ASP (full HiNet models)

the ASP can also model the  
reverberation effect



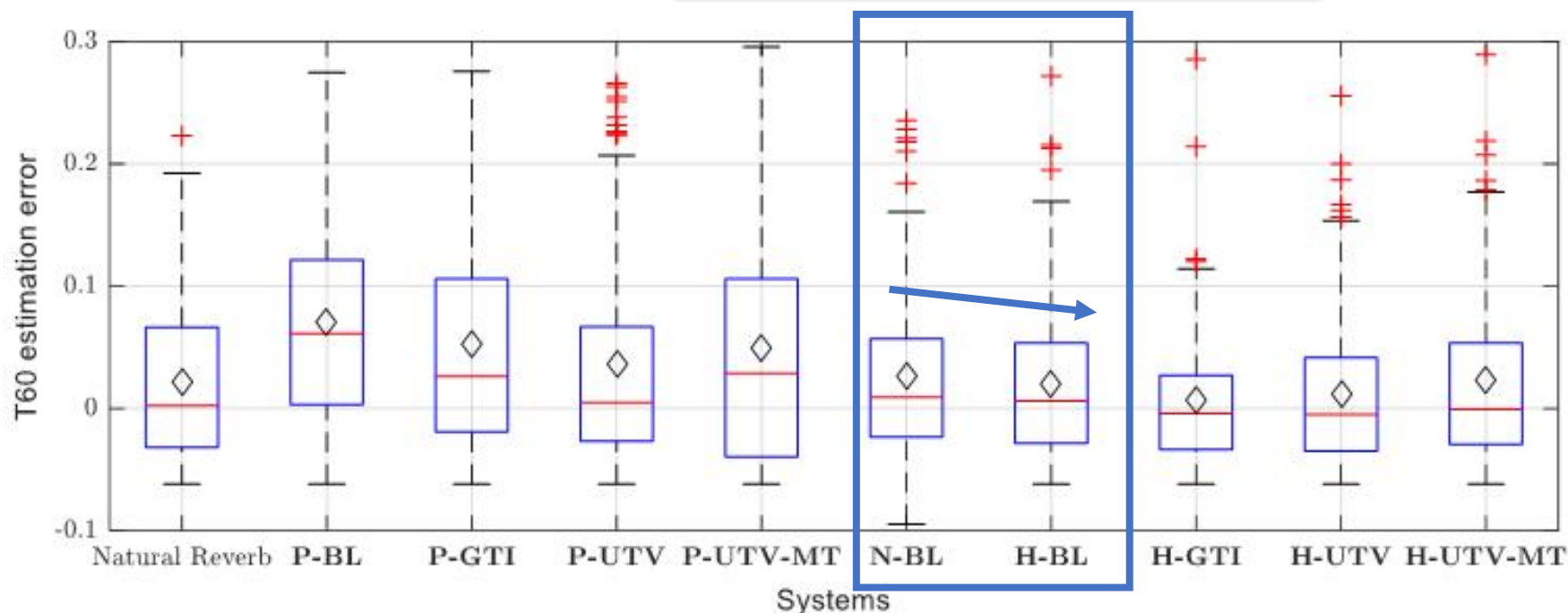


# Experiments

**N-BL:** output waveform of NSF **without** reverberation

**H-BL:** output waveform of HiNet **without** reverberation in PSP

HiNet performs better than NSF  
on this task



# Experiments

- Subjective experiments
  - Similarity test on the reverberation effect
    - Score is from 1 to 9
    - A higher score denoted a reverberation effect more similar to that in the natural reverberant audio tracks.
  - MUSHRA test on speech quality
    - Score is from 0 to 100
    - A higher score denoted higher speech quality



# Experiments-Similarity test results

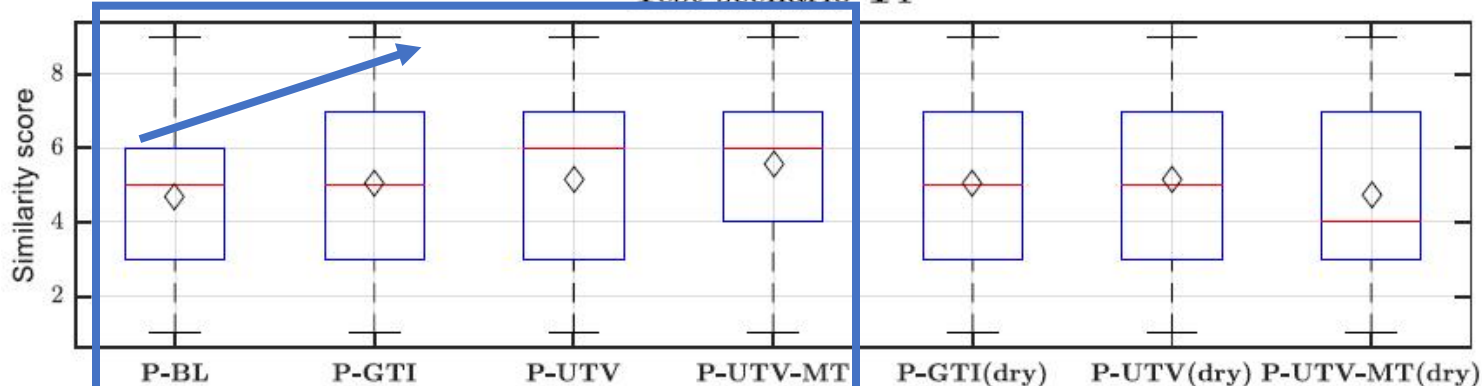
**P-BL:** output waveform of PSP **without** reverberation

**P-GTI:** output waveform of PSP with **global constant reverberation**

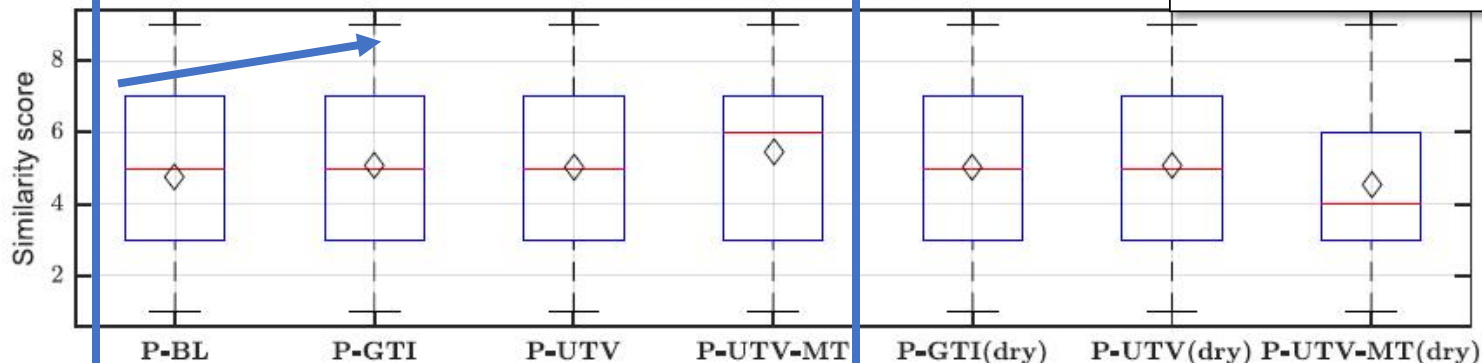
**P-UTV:** output waveform of PSP with **utterance variant reverberation**

**P-UTV-MT:** output waveform of PSP with **utterance variant reverberation** using **multi-target training**

Test scenario T1



Test scenario T2



Reverberation module is effective



# Experiments-Similarity test results

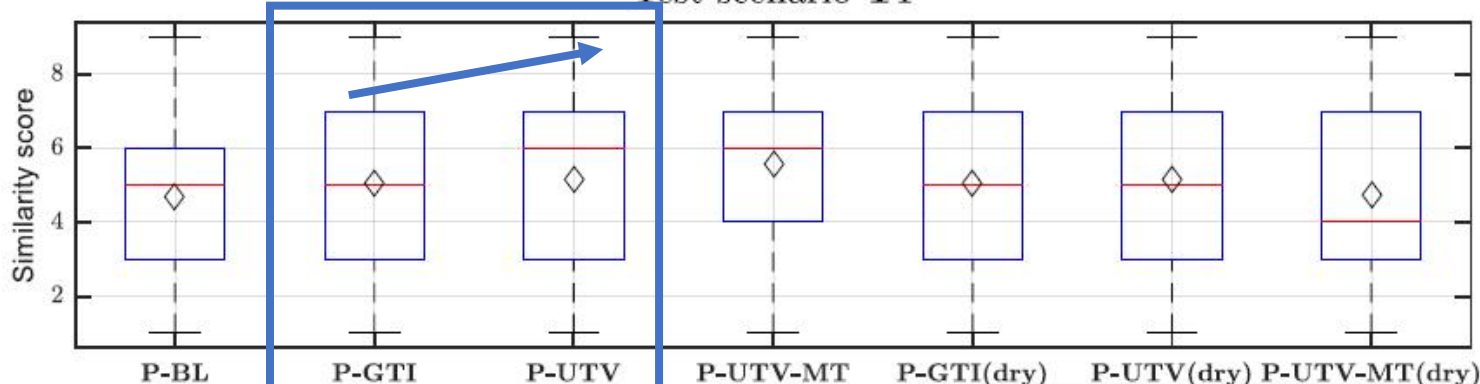
**P-BL:** output waveform of PSP **without** reverberation

**P-GTI:** output waveform of PSP with **global constant reverberation**

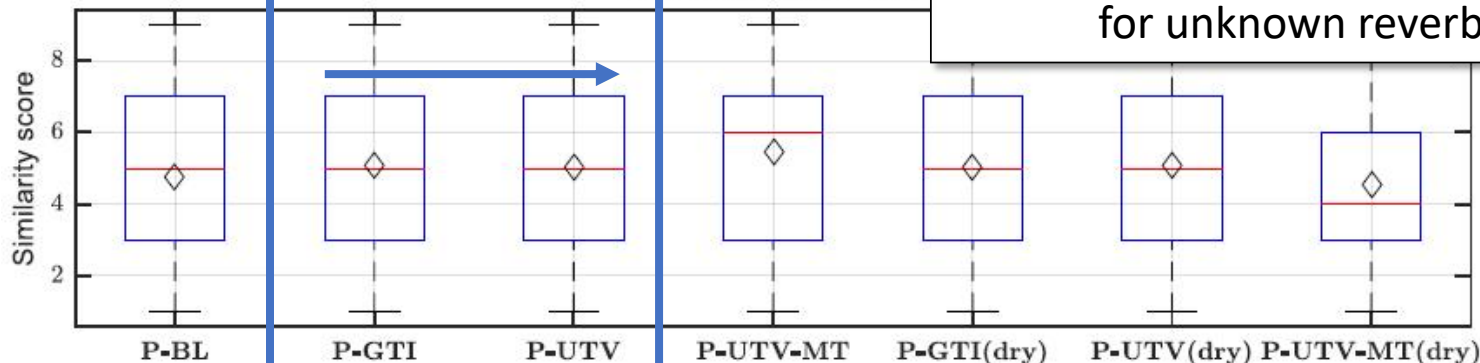
**P-UTV:** output waveform of PSP with **utterance variant reverberation**

**P-UTV-MT:** output waveform of PSP with **utterance variant reverberation** using **multi-target training**

Test scenario T1



Test scenario T2



Utterance variant reverberation is effective for unknown reverberation type



# Experiments-Similarity test results

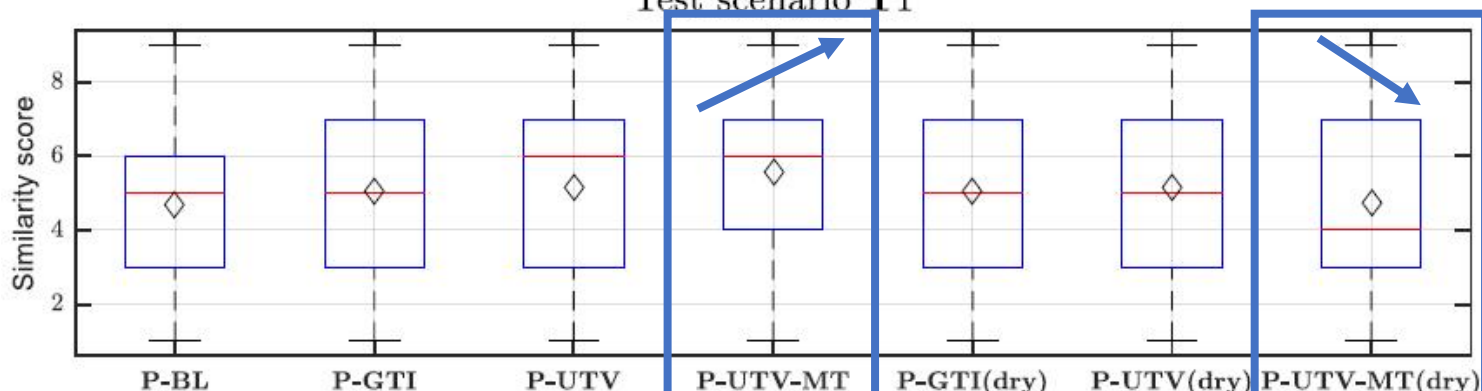
**P-BL:** output waveform of PSP **without** reverberation

**P-GTI:** output waveform of PSP with **global constant reverberation**

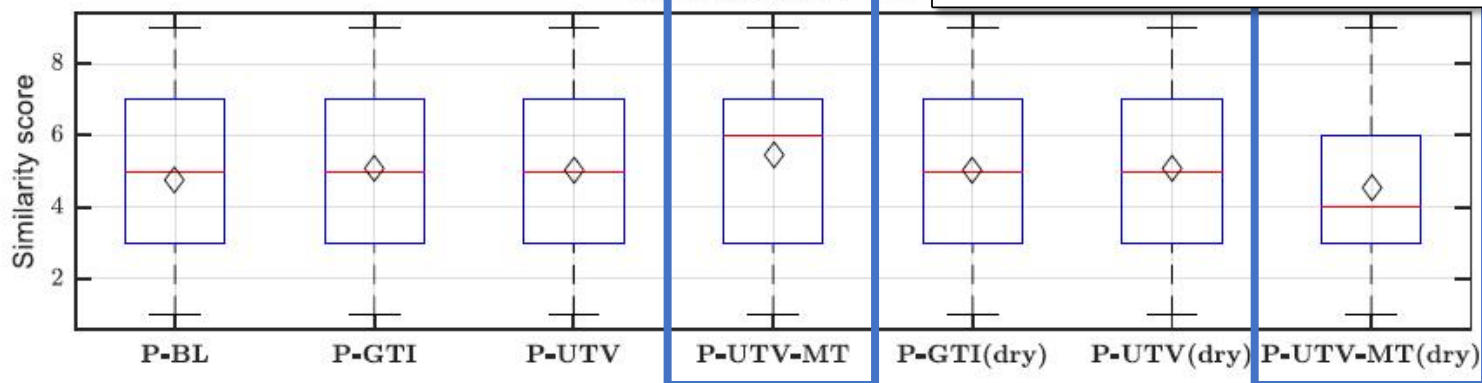
**P-UTV:** output waveform of PSP with **utterance variant reverberation**

**P-UTV-MT:** output waveform of PSP with **utterance variant reverberation** using **multi-target training**

Test scenario T1



Test scenario T2



Multi-target training is effective



# Experiments-Similarity test results

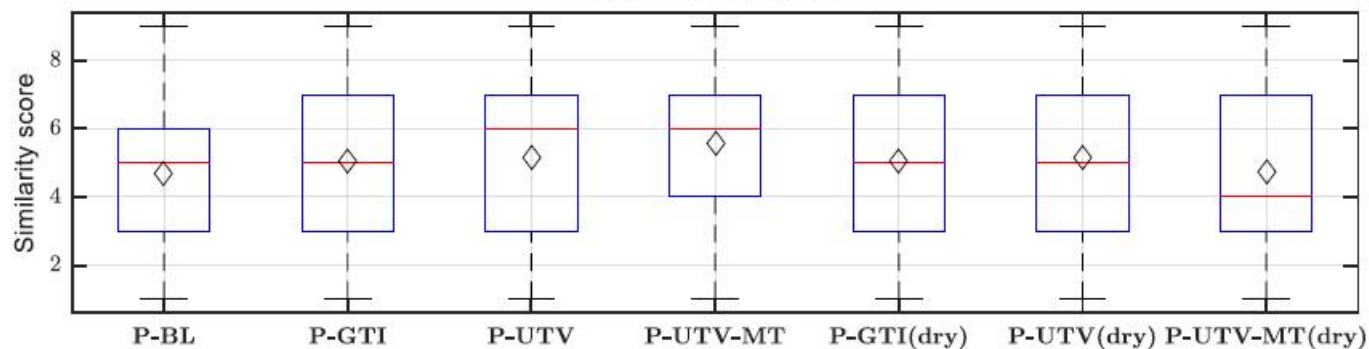
**P-BL:** output waveform of PSP **without** reverberation

**P-GTI:** output waveform of PSP with **global constant reverberation**

**P-UTV:** output waveform of PSP with **utterance variant reverberation**

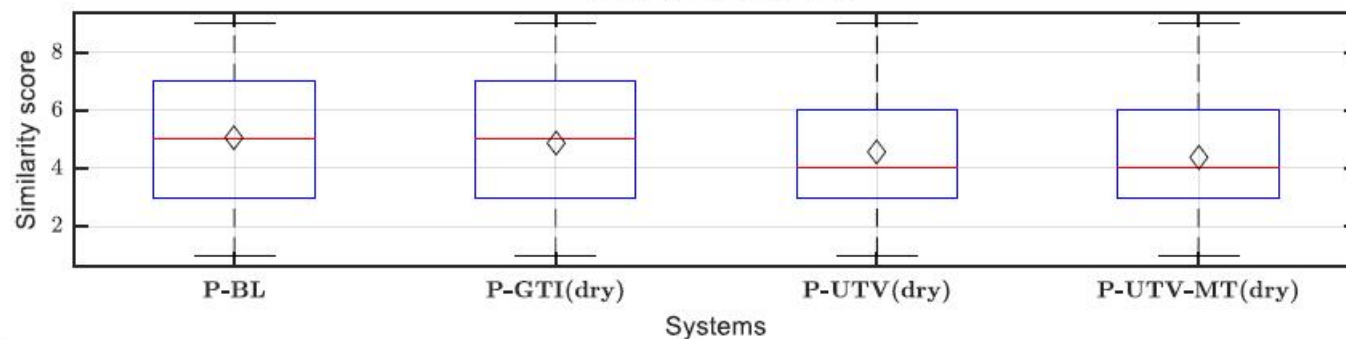
**P-UTV-MT:** output waveform of PSP with **utterance variant reverberation** using **multi-target training**

Test scenario T1



The conclusion on T3 is the same as that on T1

Test scenario T3





# Experiments-MUSHRA test results

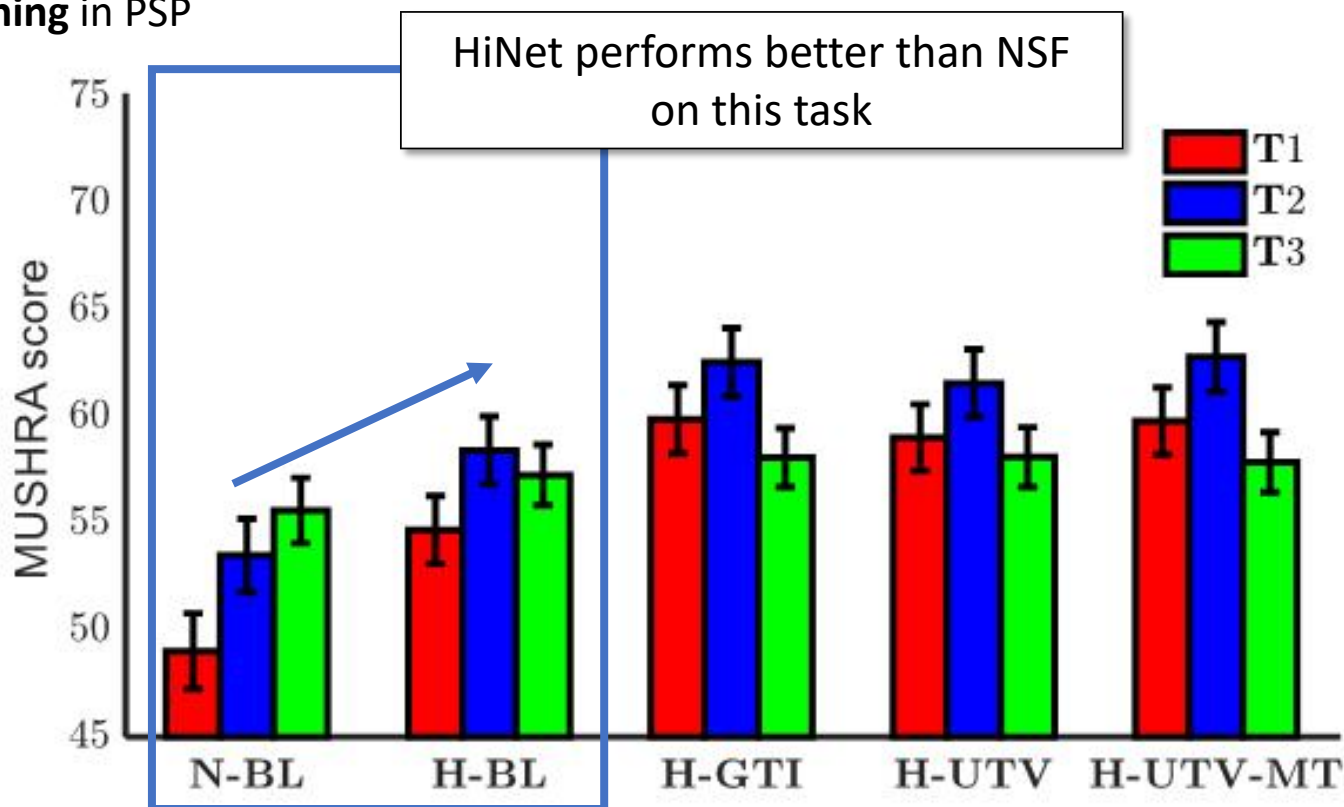
**N-BL:** output waveform of NSF **without** reverberation

**H-BL:** output waveform of HiNet **without** reverberation in PSP

**H-GTI:** output waveform of HiNet with **global constant reverberation** in PSP

**H-UTV:** output waveform of HiNet with **utterance variant reverberation** in PSP

**H-UTV-MT:** output waveform of HiNet with **utterance variant reverberation** using **multi-target training** in PSP



# Experiments-MUSHRA test results

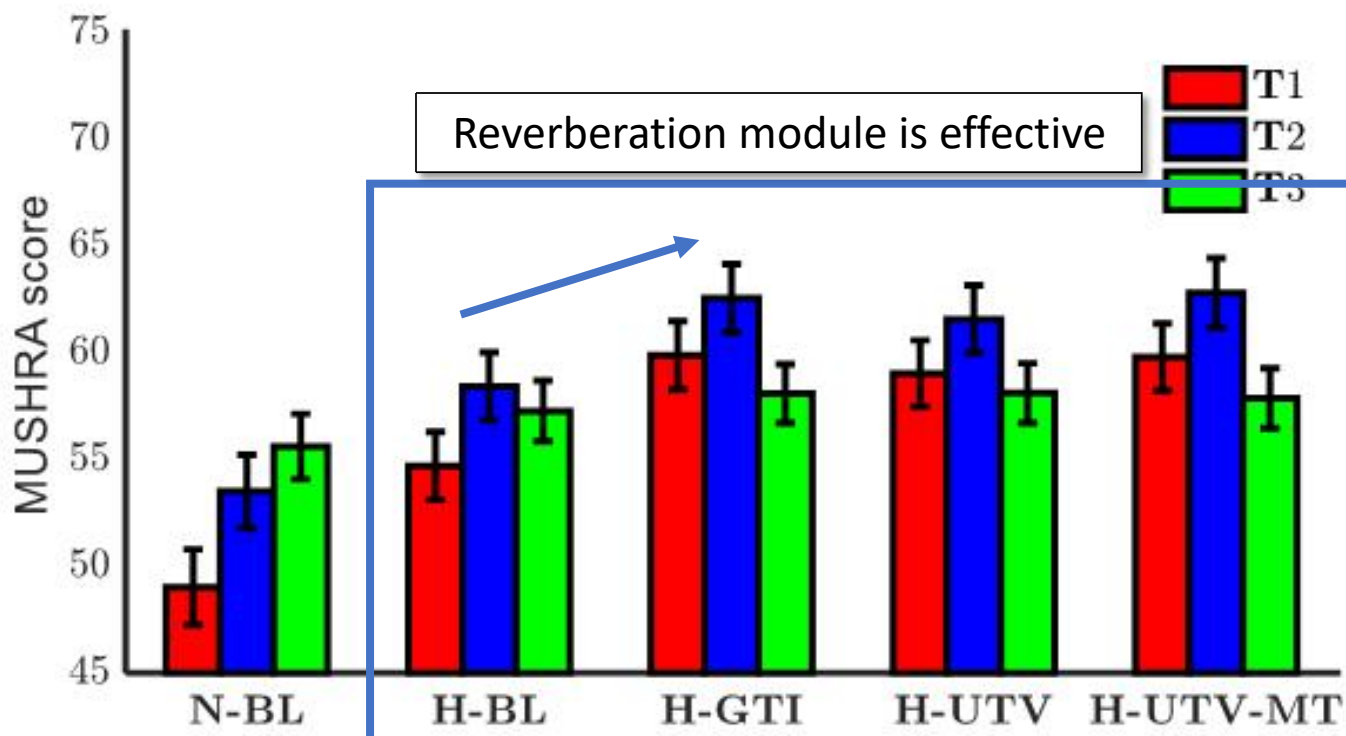
**N-BL:** output waveform of NSF **without** reverberation

**H-BL:** output waveform of HiNet **without** reverberation in PSP

**H-GTI:** output waveform of HiNet with **global constant reverberation** in PSP

**H-UTV:** output waveform of HiNet with **utterance variant reverberation** in PSP

**H-UTV-MT:** output waveform of HiNet with **utterance variant reverberation** using **multi-target training** in PSP





# Experiments-MUSHRA test results

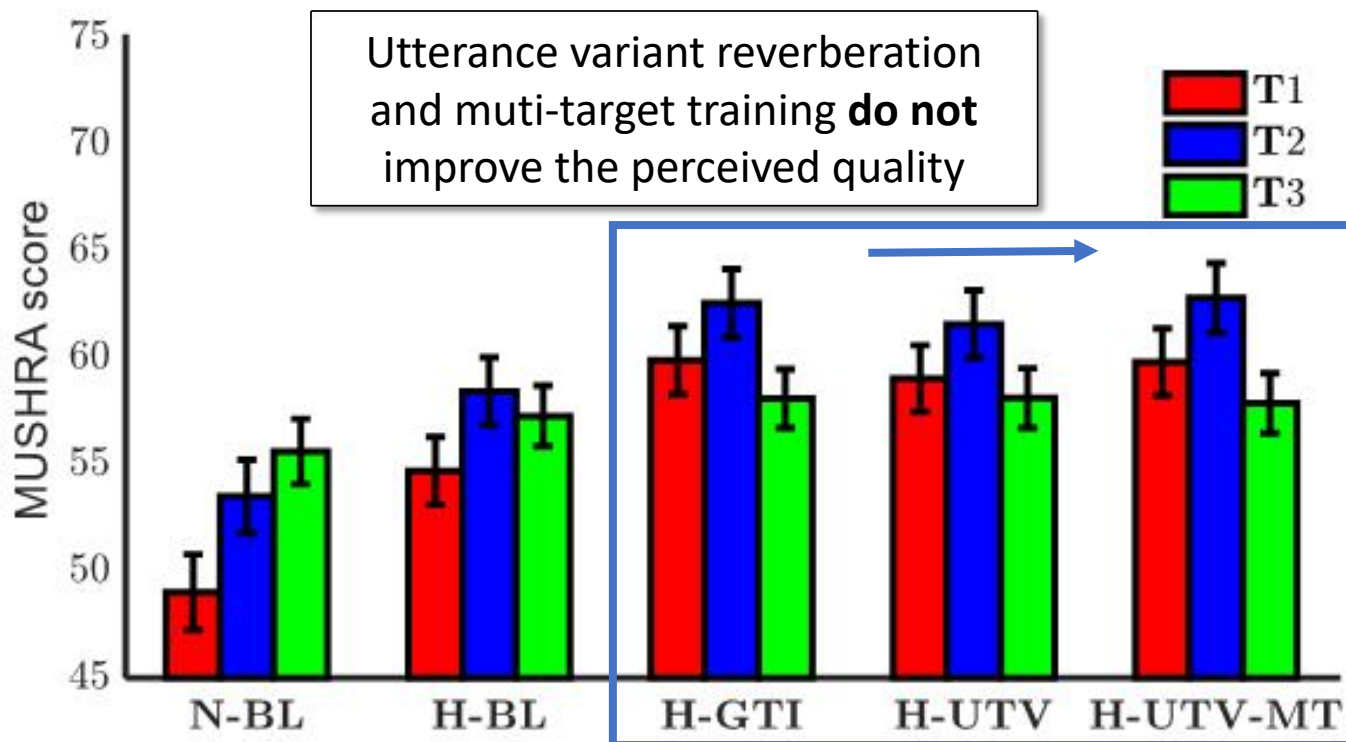
**N-BL:** output waveform of NSF **without** reverberation

**H-BL:** output waveform of HiNet **without** reverberation in PSP

**H-GTI:** output waveform of HiNet with **global constant reverberation** in PSP



















**H-UTV:** output waveform of HiNet with **utterance variant reverberation** in PSP

**H-UTV-MT:** output waveform of HiNet with **utterance variant reverberation** using **multi-target training** in PSP



# Demos

## • The output waveform of PSPs

	Test scenario T1	
Natural Reverb		
P-UTV-MT		
P-UTV		
P-GTI		
P-BL		
P-GTI(dry)		
P-UTV(dry)		
P-UTV-MT(dry)		
Natural Clean		

Note that:

1. the quality of the waveforms is not good because it is just used to extract the phase spectra (not the final waveform).
2. Please only focus on the reverberation effects.

**P-BL:** output waveform of PSP **without** reverberation

**P-GTI:** output waveform of PSP with **global constant reverberation**

























**P-UTV:** output waveform of PSP with **utterance variant reverberation**

**P-UTV-MT:** output waveform of PSP with **utterance variant reverberation** using **multi-target training**



# Demos

- The output waveform of NSF and HiNets

	Test scenario T1		Test scenario T3	
Natural				
N-BL				
H-BL				
H-GTI				
H-UTV				
H-UTV-MT				

**N-BL:** output waveform of NSF **without** reverberation

**H-BL:** output waveform of HiNet **without** reverberation in PSP

**H-GTI:** output waveform of HiNet with **global constant reverberation** in PSP

**H-UTV:** output waveform of HiNet with **utterance variant reverberation** in PSP

**H-UTV-MT:** output waveform of HiNet with **utterance variant reverberation** using **multi-target training** in PSP

More demos: <http://home.ustc.edu.cn/~ay8067/reverb/demo.html>

