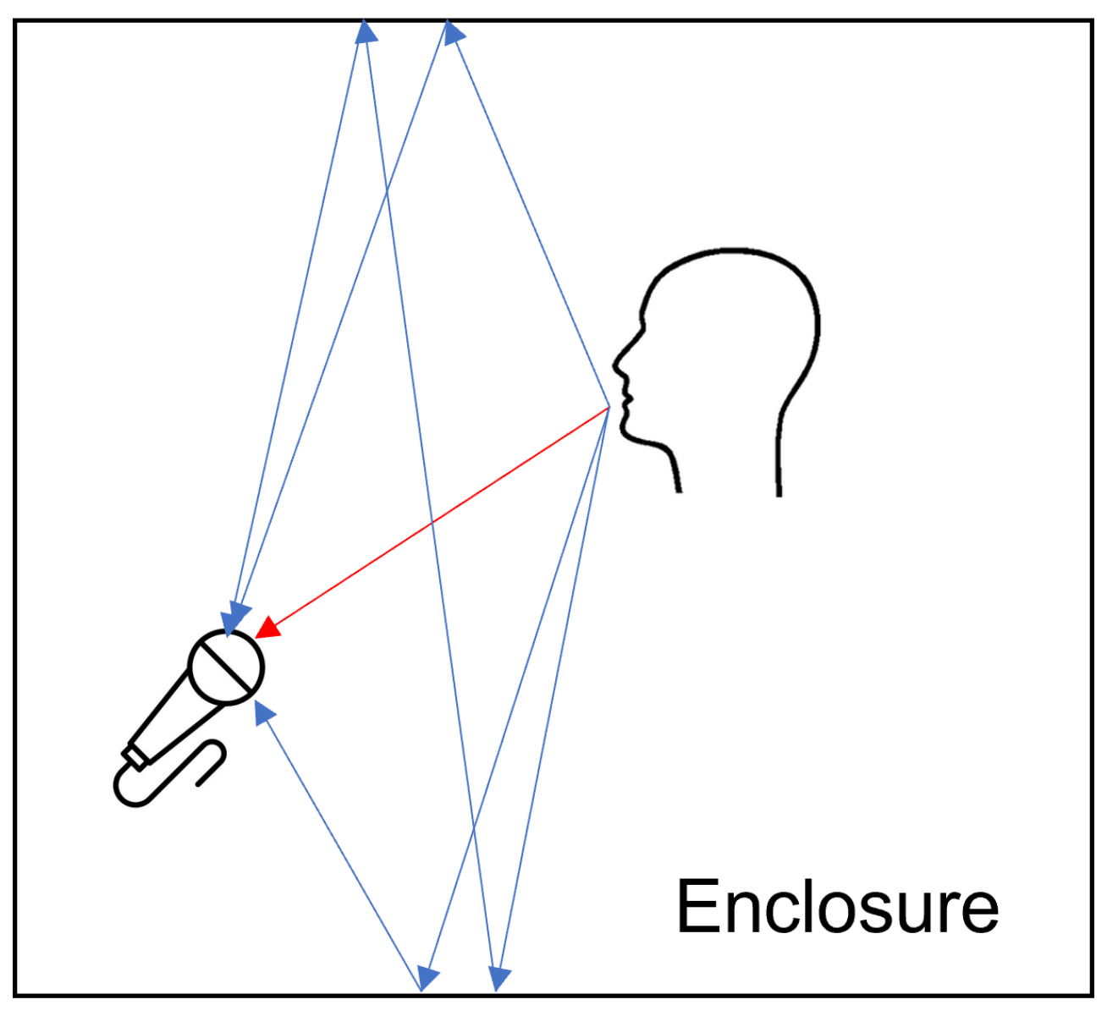
RVAE-EM: Generative speech dereverberation based on recurrent variational auto-encoder and convolutive transfer function

Pengyu Wang, Xiaofei Li
Westlake University, Hangzhou, China

Introduction

- Reverberation:

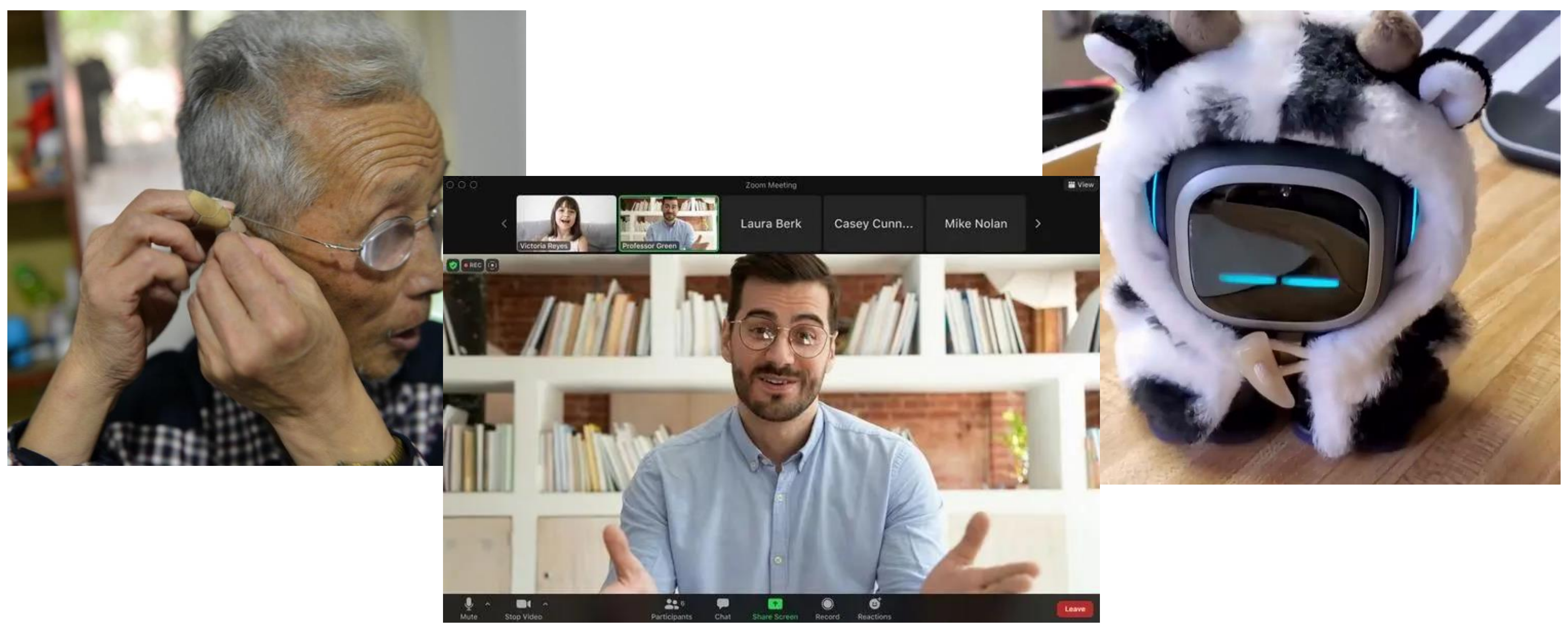
Persistence of sound in an enclosure due to multiple reflections off surfaces.



- Dereverberation:

To extract dry speech from reverberant recordings for better speech quality/intelligibility

- Applications:



Method

- Observation Model in Time-frequency Domain:

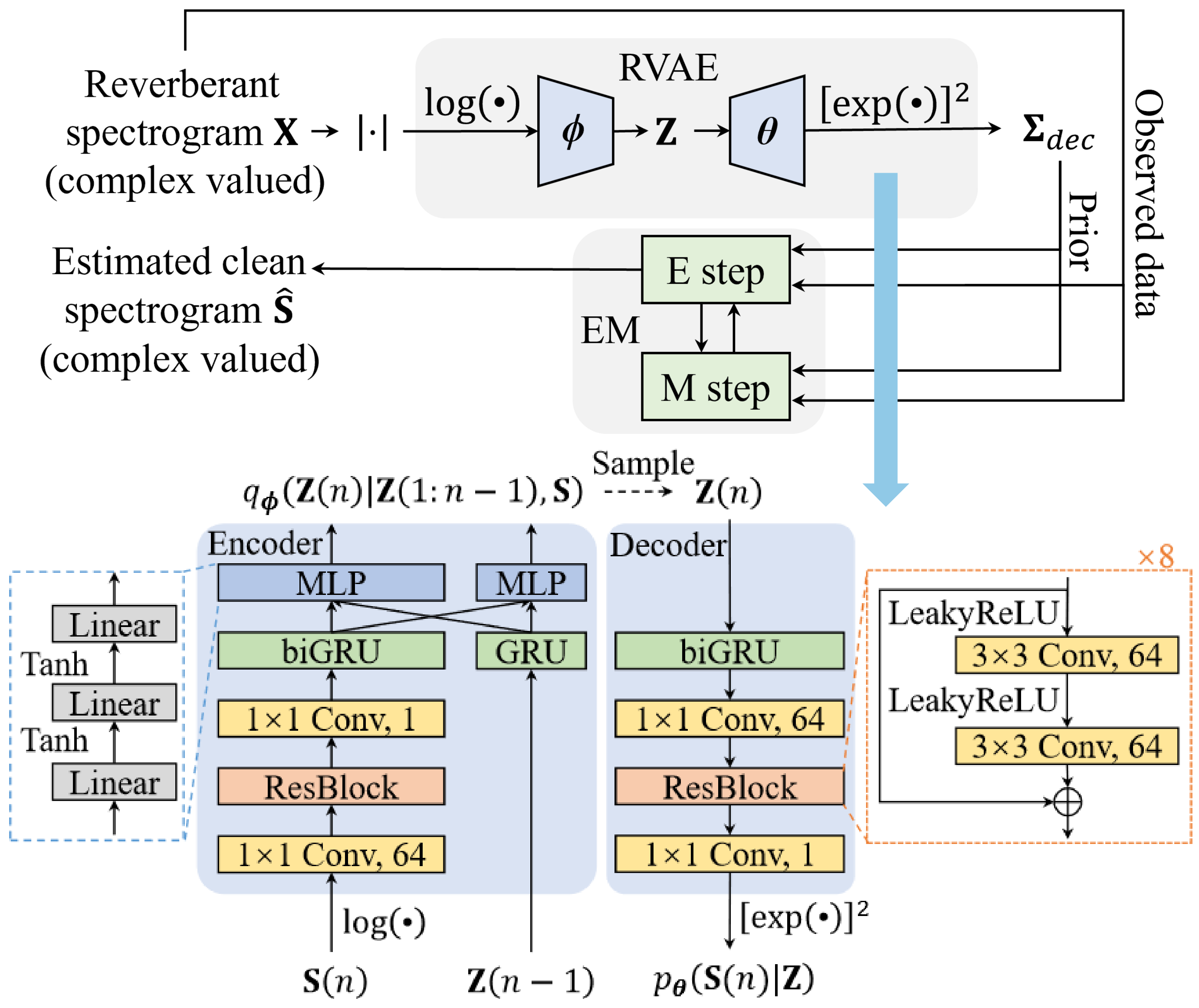
CTF (Convolutive Transfer Function) Approximation:

source speech

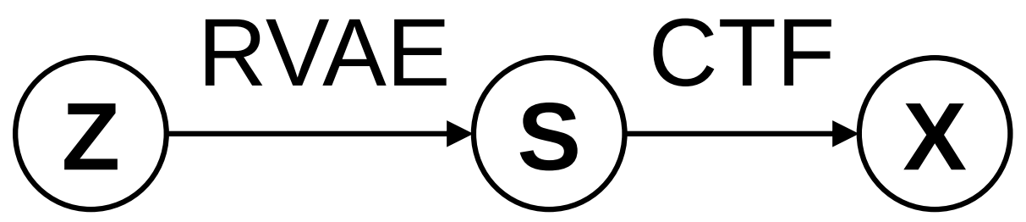
$$X_f(n) \approx \sum_{p=0}^{P} H_f(p) S_f(n-p) + W_f(n)$$

observation CTF filter noise

- RVAE-EM:



Generative-Observation Model:



Models the whole process of generating reverberant observation $\boldsymbol{X}$ from latent variables $\mathbf{Z}$.

RVAE (Recurrent Variational Auto-Encoder) Network:

Models the prior distribution of dry speech $\mathbf{S}$.

EM (Expectation Maximization) algorithm:

Solves the model parameters iteratively.

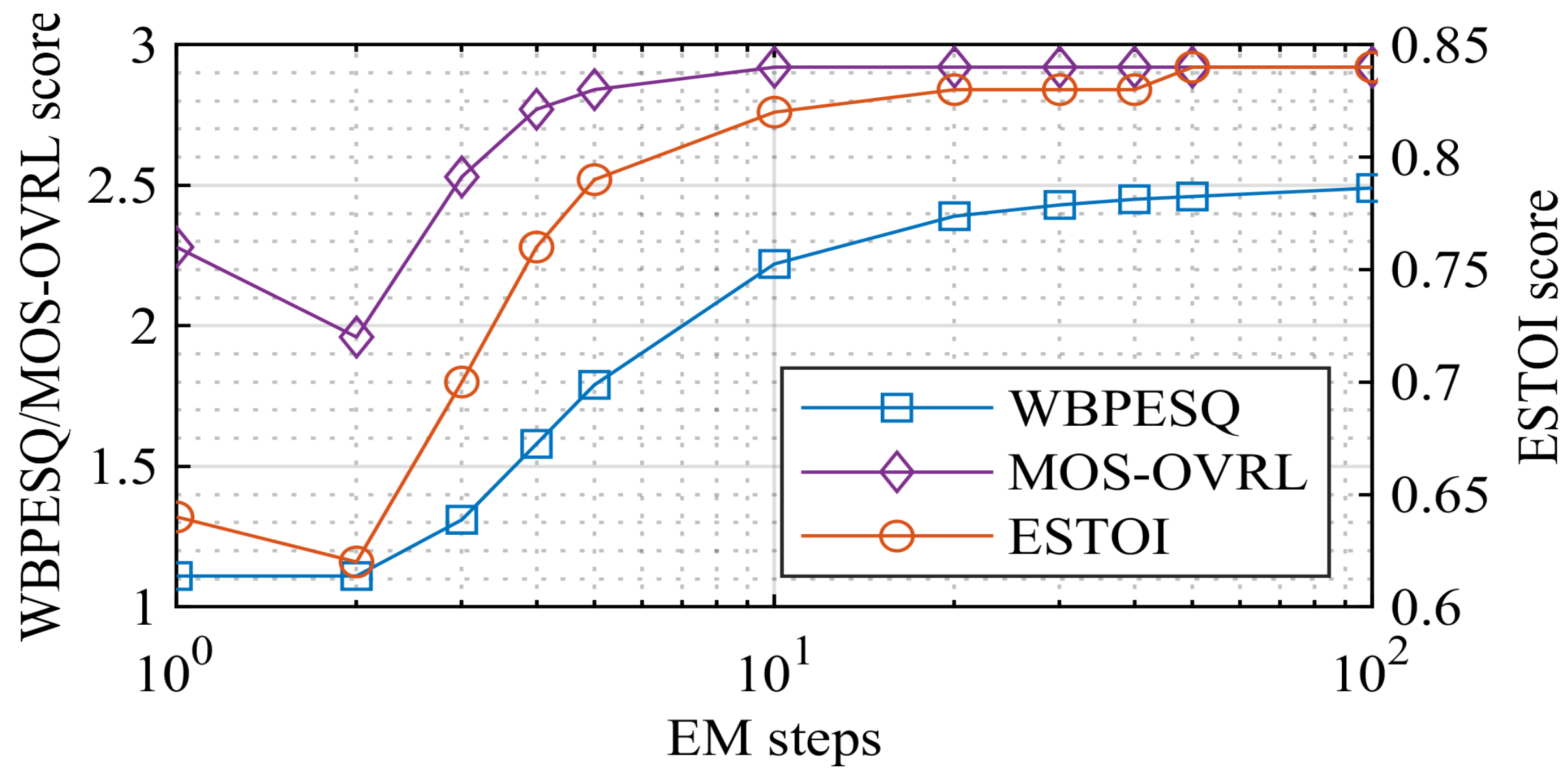
Output:

MAP (maximum a posteriori) estimation of dry speech $\mathbf{S}$.

Results

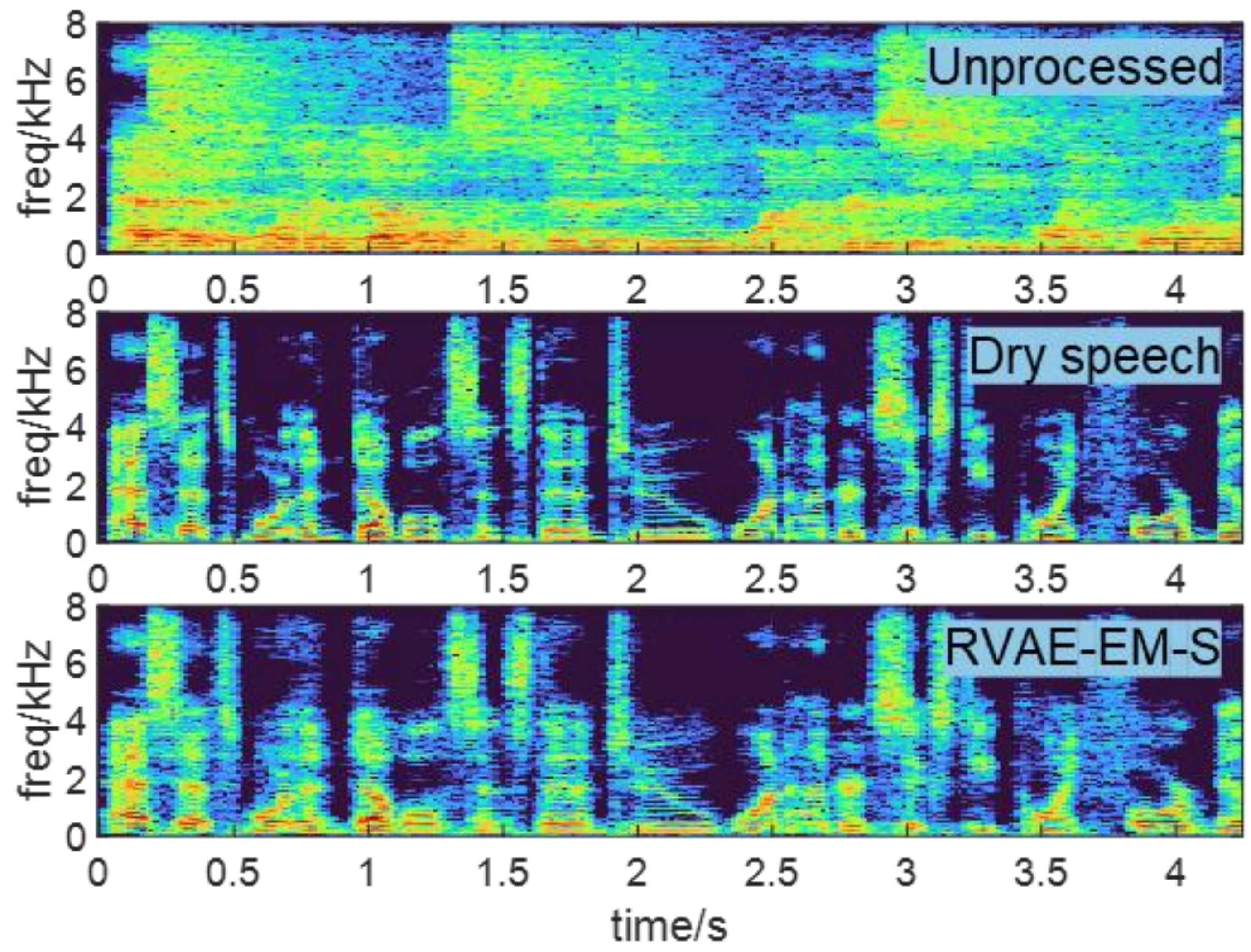
- Dataset: WSJ0 clean speech + simulated RIRs (noiseless)
- Metrics:

Method	Params	WBPESQ	ESTOI	SRMR	MOS-OVRL
Unprocessed	/	1.25	0.45	3.38	1.69
VAE-NMF	7.5M	1.36	0.52	4.34	2.05
RVAE-EM-U	7.0M	1.62	0.64	6.37	2.39
TCN-SA	4.7M	2.27	0.81	7.5	2.8
FullSubNet	14.5M	2.39	0.81	6.69	2.64
SGMSE+	65.6M	2.61	0.88	7.99	3.26
RVAE(w/o EM)	7.0M	1.97	0.75	6.43	2.64
RVAE-EM-S	7.0M	2.49	0.84	8.92	2.92



- The task assigned to DNN is simplified based on the deterministic relationship between the source speech and the observed recordings.
- EM iterations are consistently improving the estimation of clean speech and acoustic parameters.
- EM algorithm reconstructs the phase, and revises the magnitude of estimated spectrogram.
- Unsupervised or supervised trained.

- Magnitude Spectrograms in Log Scale:



- Demo: *https://audio.westlake.edu.cn/Research/RVAE.htm*
- Codes: *https://github.com/Audio-WestlakeU/RVAE-EM*
- Paper: *https://ieeexplore.ieee.org/document/10447010*

Conclusion

A speech dereverberation method is proposed.

- Advanced performance in both unsupervised and supervised categories.
- Shows the capability and potential of the proposed generative-observation model.