Perceptually-motivated Environment-specific Speech Enhancement

Jiaqi Su[1], Adam Finkelstein[1], Zeyu Jin[2]
[1]Princeton University, [2]Adobe Research

Introduction

Many factors in a typical environment can diminish the quality of a recording, including **noise**, **reverberance**, and **undesirable equalization**.

Existing SE methods:

- Spectral methods,require target phase information to recover waveform, which introduces noticeable artifacts.
- Popular sample-based loss functions for waveform methods are not in line with human perception, and are brittle to misalignment in real recordings.

We propose:

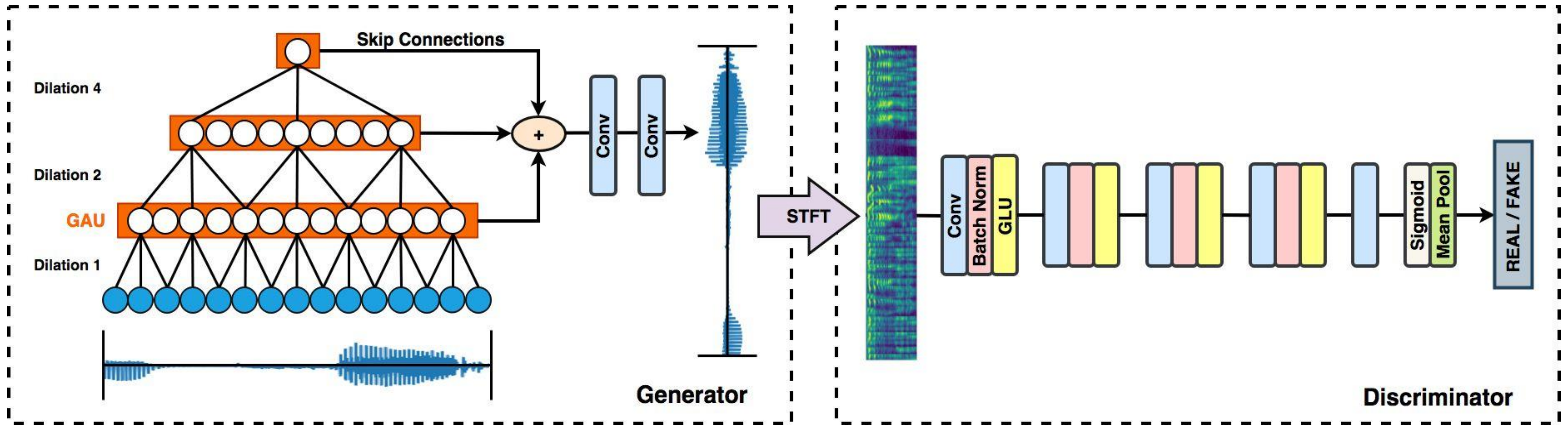
- A **perceptually-motivated loss function** that combines adversarial loss with spectrogram features.
- A **waveform SE method** that works with synthetic and real data.

Both objective and subjective evlauation results show:

- improved performance over previous methods for real and synthetic data.
- capability to ameliorate several types of recording artifacts.

Method

Feedforward WaveNet with a combination of L2 loss on log spectrogram and adversarial loss on log mel spectrogram:



Perceptually Motivated Loss

Generator:
$$L_G(x,x') = \alpha||LogSpec(G(x)) - LogSpec(x')||^2 + (1-\alpha)(1-D(LogMel(G(x)))$$

Discriminator:
$$L_D(x,x') = D(LogMel(G(x))) + 1 - D(LogMel(x'))$$

Spectrogram Loss:
(1) Allow misalignment between input and target during training.
(2) In accordance with human perception.

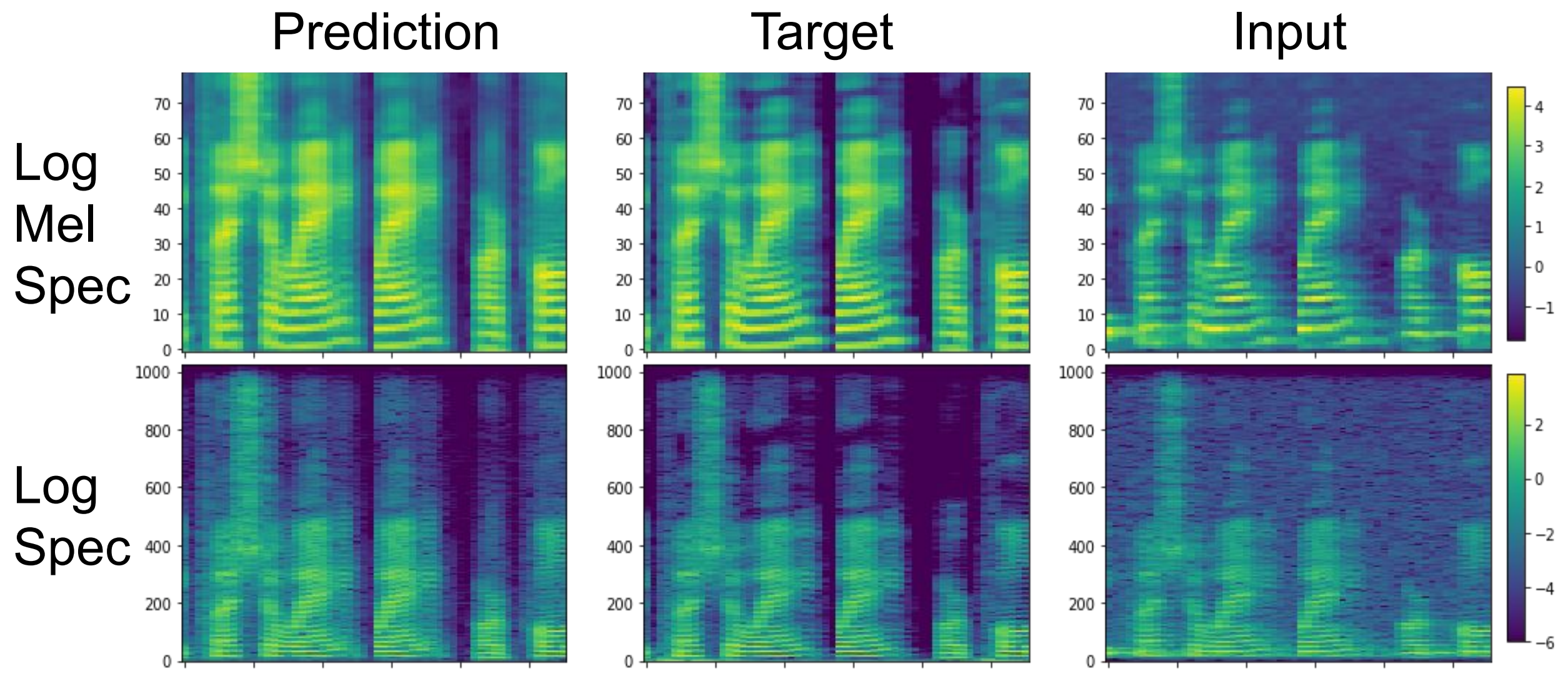
Adversarial Loss:
Offer variable discriminitive power to refine details.

Conclusions

- Designing an appropriate objective function that reflects human perception is key to data driven methods on audio.
- Acoustic feature-based losses, such as log spectrogram loss
  - ○ achieve better perceptual qualities than sample-level loss function;
  - ○ enable learning on real world sloppy recordings.
- Adversarial training mimics human perception to some extent, and reduces noise and artifacts introduced by SE process.

Experiments

Example Results

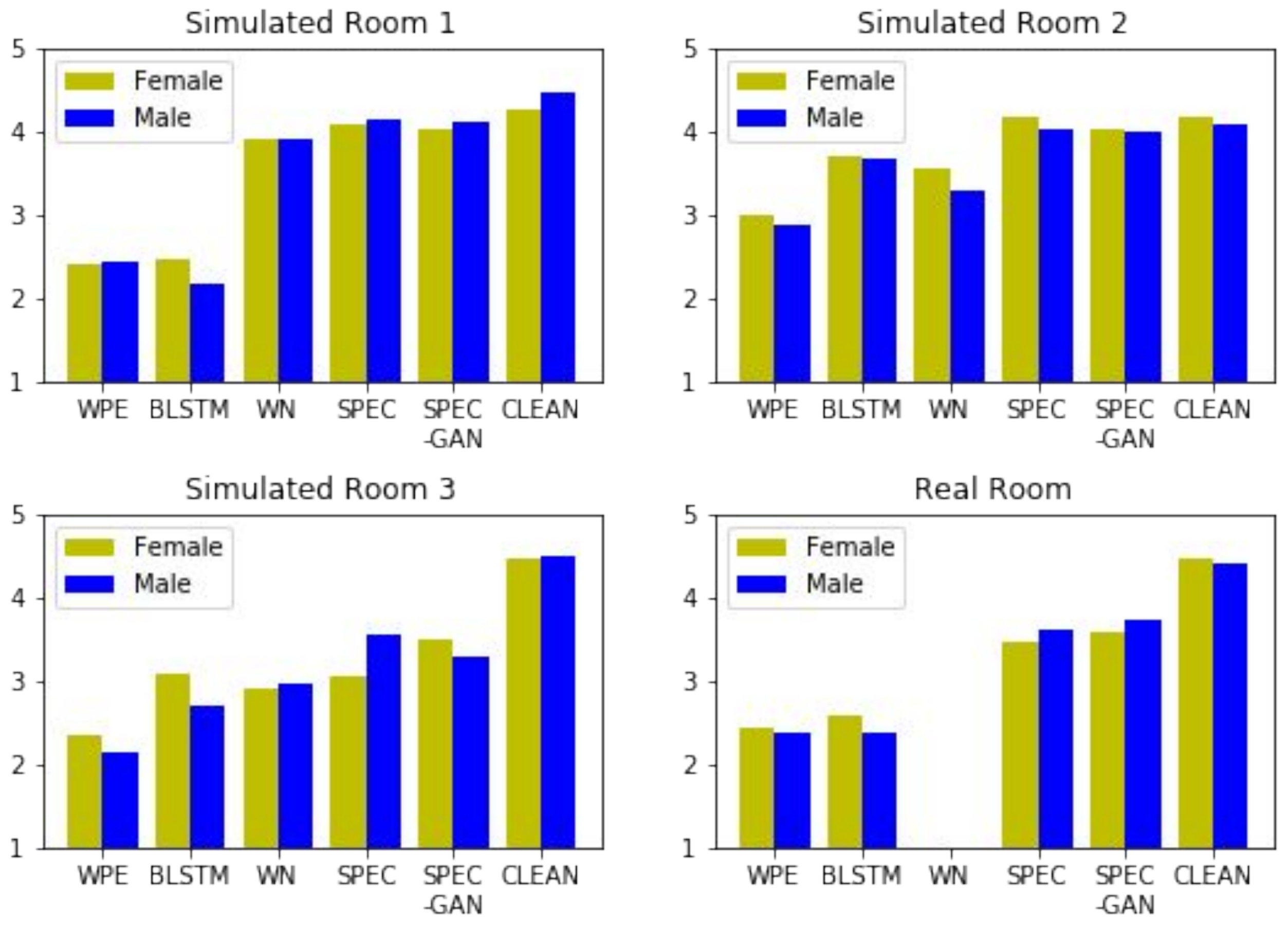


Objective Evaluation

Method	PESQ	FWSEGSNR	SRMR	CD
CLEAN	4.64	35.0	8.41	0.0
REVERB	1.24	-0.63	5.82	7.02
DN-WN	2.17	-1.55	**8.18**	6.94
BLSTM	2.10	5.87	6.90	3.87
WPE	1.39	0.01	7.03	6.98
SPEC	2.45	6.34	7.45	4.70
S-GAN	**2.61**	**12.53**	8.17	**3.12**

- **SPEC:** Ours, spectragram loss only
- **S-GAN:** SPEC + adversarial loss
- **WPE [8]:** traditional inverse filtering method
- **BLSTM [14]**: Spectral masking with Bi-LSTMs
- **DN-WN [20]:** Speech Denoising WaveNet

Subjective Evaluation



Mean Opinion Scores (1=Bad, 5=Excellent) collected from Amazon Mechanical Turk