uSee: Unified Speech Enhancement And Editing with Conditional Diffusion Models

Muqiao Yang[*], Chunlei Zhang[†], Yong Xu[†], Zhongweiyang Xu[‡], Heming Wang[#], Bhiksha Raj[*], Dong Yu[†]

[*]Cargegie Mellon University, [†]Tencent AI Lab, [#]The Ohio State University, [‡]University of Illinois Urbana-Champaign

https://muqiaoy.github.io/usee

Introduction

Background

This work proposes Universal Speech Enhancement and Editing (uSee) model to perform generative speech **denoising, dereverberation and controlled speech editing**, given specific **acoustic and textual** prompts as conditions.

Problem Formulation

The backbone of our generation model is conditional diffusion model. The conditions can be given as a combination of acoustic and textual prompts:

1. Acoustic prompts as SSL embeddings from source speech (HuBERT)
2. Textual prompts as fine-grained instructions to instruct the model with task types/background sound type/SNR/RT60s.

Method

Our proposed uSee model has two objectives:

- Objective I: Speech Enhancement. We would like a universal speech enhancement (SE) model to handle various speech distortions in a generative manner.
- Objective II: Speech Editing. As a reverse process of Speech Enhancement, we perform controllable speech editing with adding background sound/reverberation/other operations to source speech.

The overall architecture of the proposed uSee model is depicted in Fig.1 which comprises two main parts:

- Target Spectrogram, used as input to the score-based conditional diffusion model (u-Net-based structure).
- Acoustic/Textual Prompts, used as conditions to steer the process to generate the audio containing specific information. Acoustic prompts are concatenated with a frame-by-frame correspondence, while textual prompts are applied to the input with cross attention.

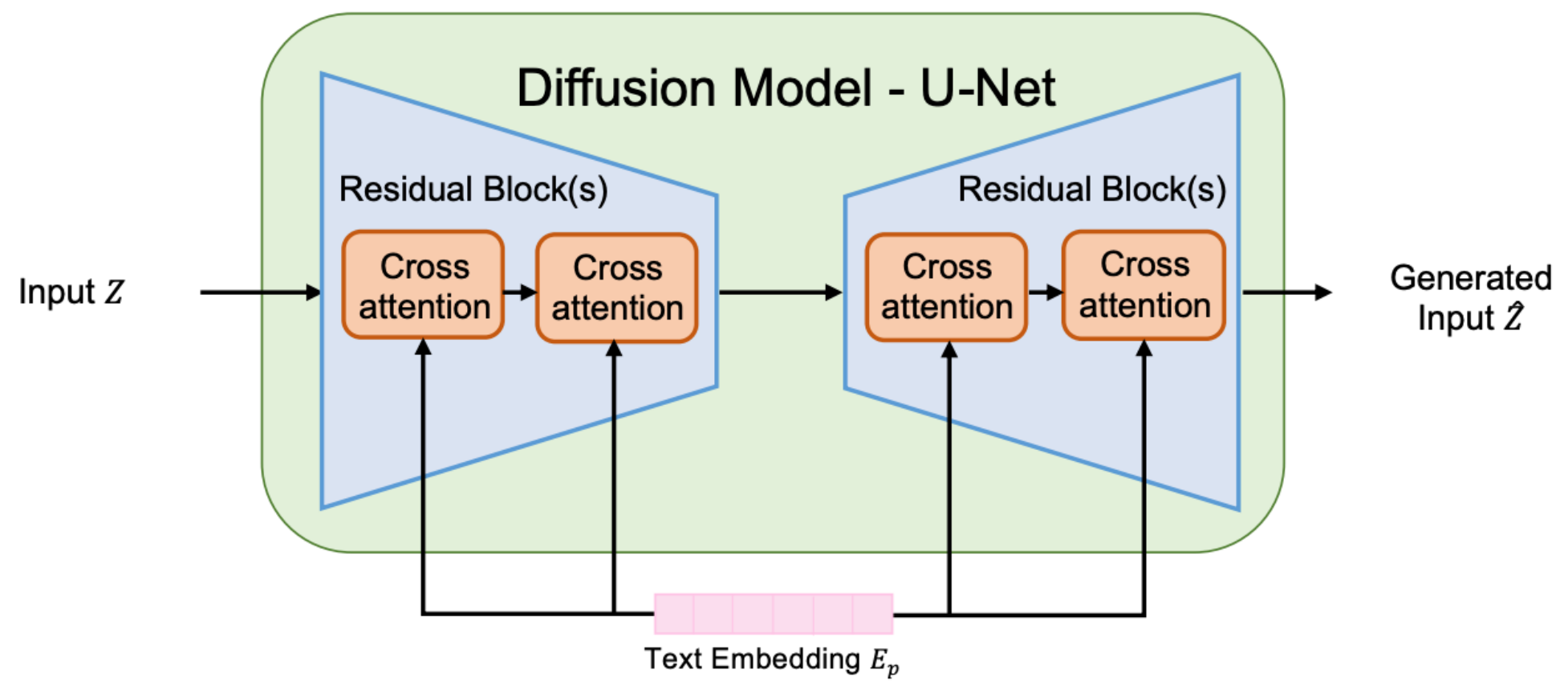

Fig. 2: Illustration of cross-attention layers for textual prompts as conditions in the diffusion model.

The format of text prompt conditions may include: task type, noise types, noise/reverberation levels.
During training, clean speech and noisy speech are reversed as source and target compared to speech enhancement in one universal model.

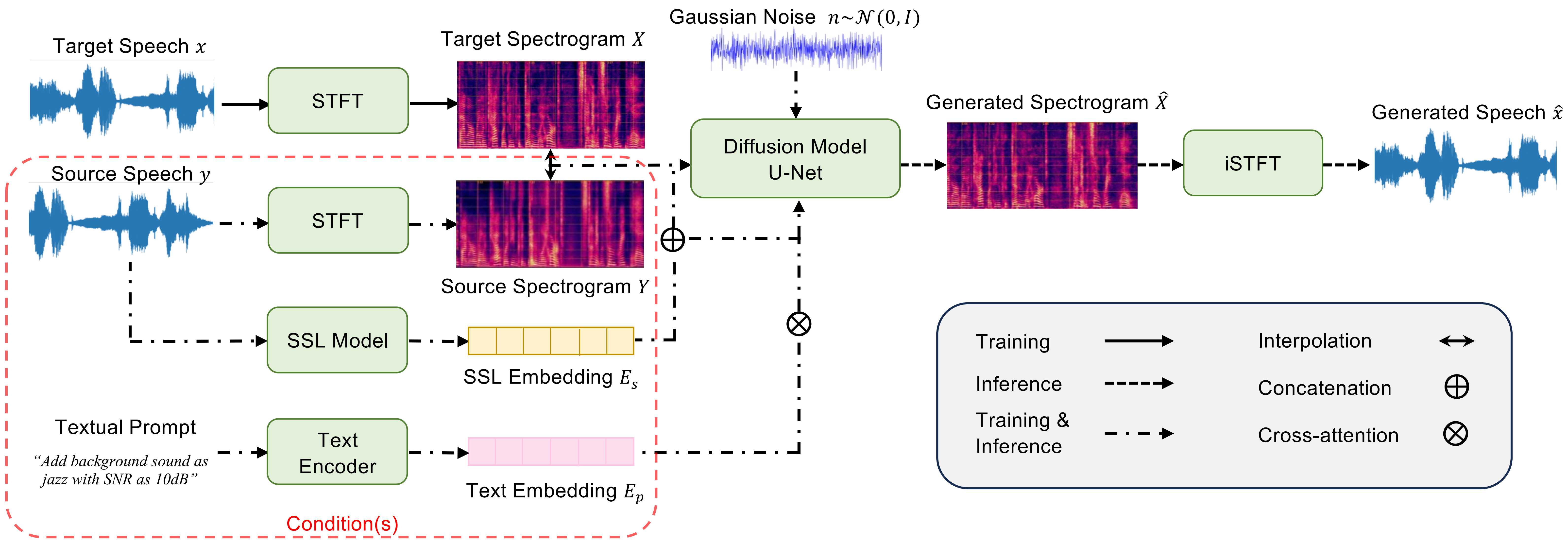

Fig. 1. Overview of the proposed unified speech enhancement and editing (uSee) model.

Experiments

Datasets

Clean speech: LibriTTS (resampled to 16kHz)
Noise set: Balanced AudioSet
RIR set: Room Impulse Response and Noise Database

Simulation

During speech simulation, we add to the clean speech with either a background sound or a room impulse response.

(1) Background sound:
Mixed with SNR ranging from [0, 15]dB;
If the clean speech is shorter than the background sound, instead of clipping the sound, we will randomly insert the clean speech into one interval of the background sound. The motivation is that we will utilize text information to generate speech with specific sound types, and in this way, the information representing the sound type will be fully preserved;
If the clean speech is longer, the background sound will be repeated until it reaches the same length as the clean speech.

(2) RIR:
Selected from rooms with different volumes including large, medium, and small sizes.

Example prompt: Add background sound as dog barking with SNR as 10dB.

SSL model: HuBERT
Text encoder: T5

Results

Speech Enhancement

Quantitative results of speech enhancement is shown in Table 1. We report evaluation metrics including both intrusive metrics such as PESQ and STOI, and non-intrusive metric DNSMOS. The results show that both acoustic (SSL embedding from HuBERT) and textual prompts (text embeddings from T5) have an effect on boosting the speech enhancement performance of tuSee model individually.

Model	WB-PESQ	NB-PESQ	STOI	DNSMOS [28]
Noisy	1.46	2.05	0.61	3.26
cDiffuSE [12]	1.75	2.47	0.69	3.48
SGMSE [22]	1.91	2.66	0.72	3.65
uSee (ours)				
linear interp.	1.76	2.49	0.67	3.50
exponential interp.	2.01	2.72	0.71	3.68
+ acoustic prompts	2.15	2.86	0.77	3.80
+ textual prompts	**2.20**	**2.88**	**0.80**	**3.86**

Table 1: Quantitative evaluation results on joint speech denoising and dereverberation of our uSee model with different conditions.

Speech Editing

Qualitative results of the generated speech with fine-grained control is shown in Fig. 2. We can observe that the lengths of reverberant tails in the spectrograms are different according to the text prompts. Examples of reverberant tail differences are highlighted in the green circles. In larger rooms, the length of the reverberant tails are increasingly longer compared to small room size.

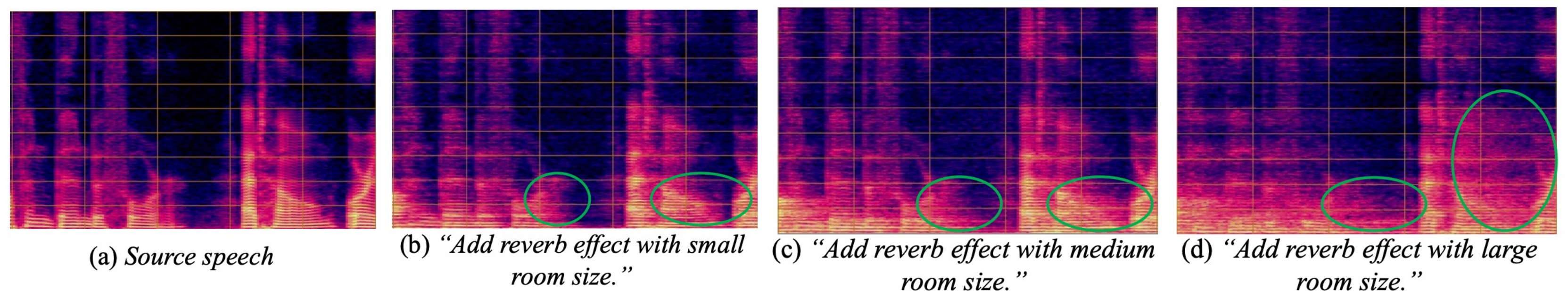
(a) *Source speech* (b) *"Add reverb effect with small room size."* (c) *"Add reverb effect with medium room size."* (d) *"Add reverb effect with large room size."*

Fig. 2. Qualitative demos of edited speech generated by uSee model. Controllable generation is enabled by the textual prompts in the captions to control the room size of RIRs.

Conclusions

This project proposes a unified Speech Enhancement and Editing framework with conditional diffusion model. By providing the uSee model with conditions including both acoustic and textual prompts, we show the capability of controllable generation for both speech enhancement and editing tasks.