The SEUEE System for the CHiME-8 MMCSG Challenge – Neural Directional Speech Extraction for ASR on Smart Glasses

Cong Pang[1,2], Feifei Xiong[2], Ye Ni[1], Lin Zhou[1], Jinwei Feng[2]

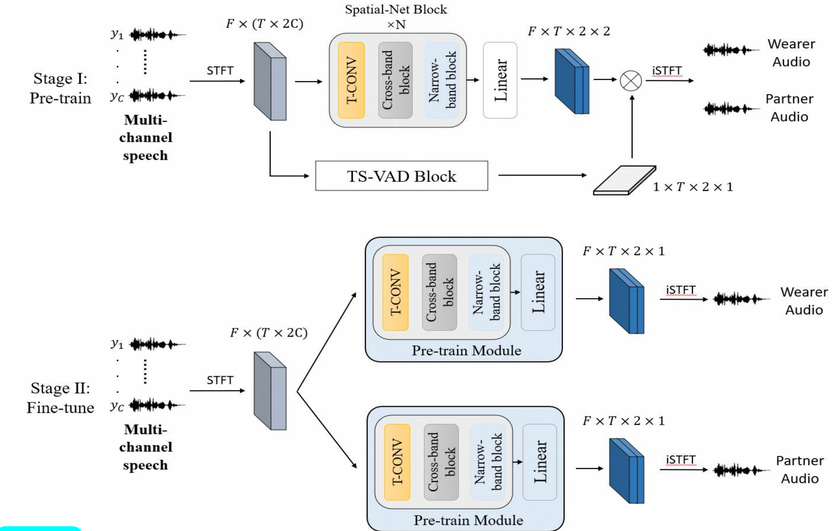
[1]*Southeast University, Nanjing, China*

[2]*Hummingbird Audio Lab, Alibaba Group, Hangzhou, China*

Introduction

- We introduce our directional speech extraction (DSE) system for MMCSG task to extract the wearer and the partner audio
- *Submitted system*: based on SpatialNet and target-speaker voice activity detection (TS-VAD), we introduce a two-stage training strategy to stabilize the individual DSE models
- *Extension work*: we introduce direction features (DFs) and ASR-inspired loss function to constrain the DSE model

Submitted System for MMCSG

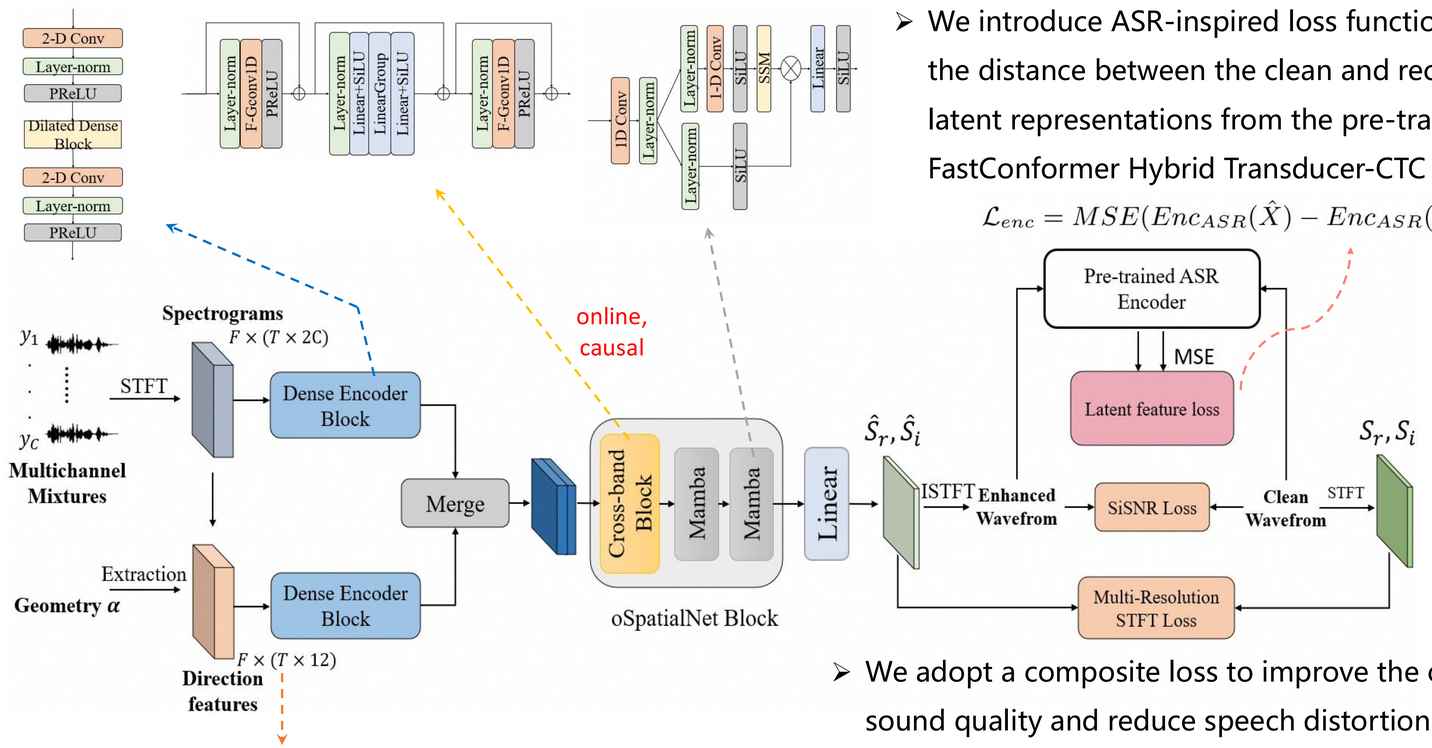


dev set

Method	Latency [s]	SELF					OTHER				
		WER	INS	DEL	SUB	ATTR	WER	INS	DEL	SUB	ATTR
Baseline	0.15	17.9	1.7	4.2	10.5	1.6	24.4	2.6	7.3	12.3	2.2
	0.34	15.0	1.4	3.9	8.4	1.4	21.4	2.2	7.2	10.1	1.8
	0.62	14.3	1.3	3.8	7.9	1.3	20.3	2.1	7.1	9.6	1.6
SEUEE	>1.0	**12.0**	1.4	3.9	6.3	0.4	**20.2**	3.0	6.5	10.2	0.5

- TS-VAD is helpful to obtain a robust pre-trained model
- SEUEE achieved a relative WER improvement of 16.43% (Self) and 0.49% (Other) over the baseline on development set

Extension Work



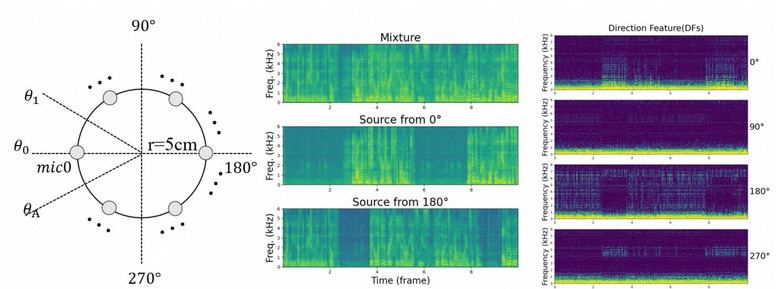
- We introduce ASR-inspired loss function to reduce the distance between the clean and reconstructed latent representations from the pre-trained FastConformer Hybrid Transducer-CTC model

$$\mathcal{L}_{enc} = MSE(Enc_{ASR}(\hat{X}) - Enc_{ASR}(X))$$

- We adopt a composite loss to improve the output sound quality and reduce speech distortion

- We introduce direction features (DFs) and visualize their values w.r.t. input from different directions

$$DF(\theta_i, t, f) = \sum_p \left\langle \mathrm{K}^{IPD^{(p)}(t,f)}, \mathrm{K}^{TPD^{(p)}(\theta,f)} \right\rangle, i = 1, 2, ..., M,$$



- DFs trained to learn the directional knowledge representations to further assist the spectrograms to extract the speech from a specific direction (a-prior)

eval set

Method	latency [s]	Overall WER	SELF WER	SELF ATTR	OTHER WER	OTHER ATTR
Baseline	0.14	22.1	17.8	2.5	26.3	2.5
	0.33	18.9	15.0	2.4	22.9	2.2
	0.62	17.9	14.1	2.3	21.7	2.1
SEUEE	>1.0	16.3	11.1	1.1	21.5	**0.8**
Extended	0.34	**15.8**	**10.7**	**1.0**	**20.9**	0.9

- Extended system: real-time, causal system, latency 0.34s
- Compared to our previously submitted system, a relative WER improvement of 3.6% (Self) and 2.8% (Other) is achieved on evaluation test set
- More analysis/results will be presented in our final paper