Wed-P-OS-6-1
441

INTERSPEECH 2022

DDS: A new device-degraded speech dataset for speech enhancement

Haoyu Li[1,2], Junichi Yamagishi[1,2]
[1]National Institute of Informatics, Japan [2]SOKENDAI, Japan

Introduction

(1) Noisy speech dataset used for training neural SE model
(2) Many existing dataset generated by simulation
(3) Model trained on simulated dataset degrades on real recordings
a) Real recordings are degraded by multiple joint factors
b) RIRs cannot capture the nonlinear distortion

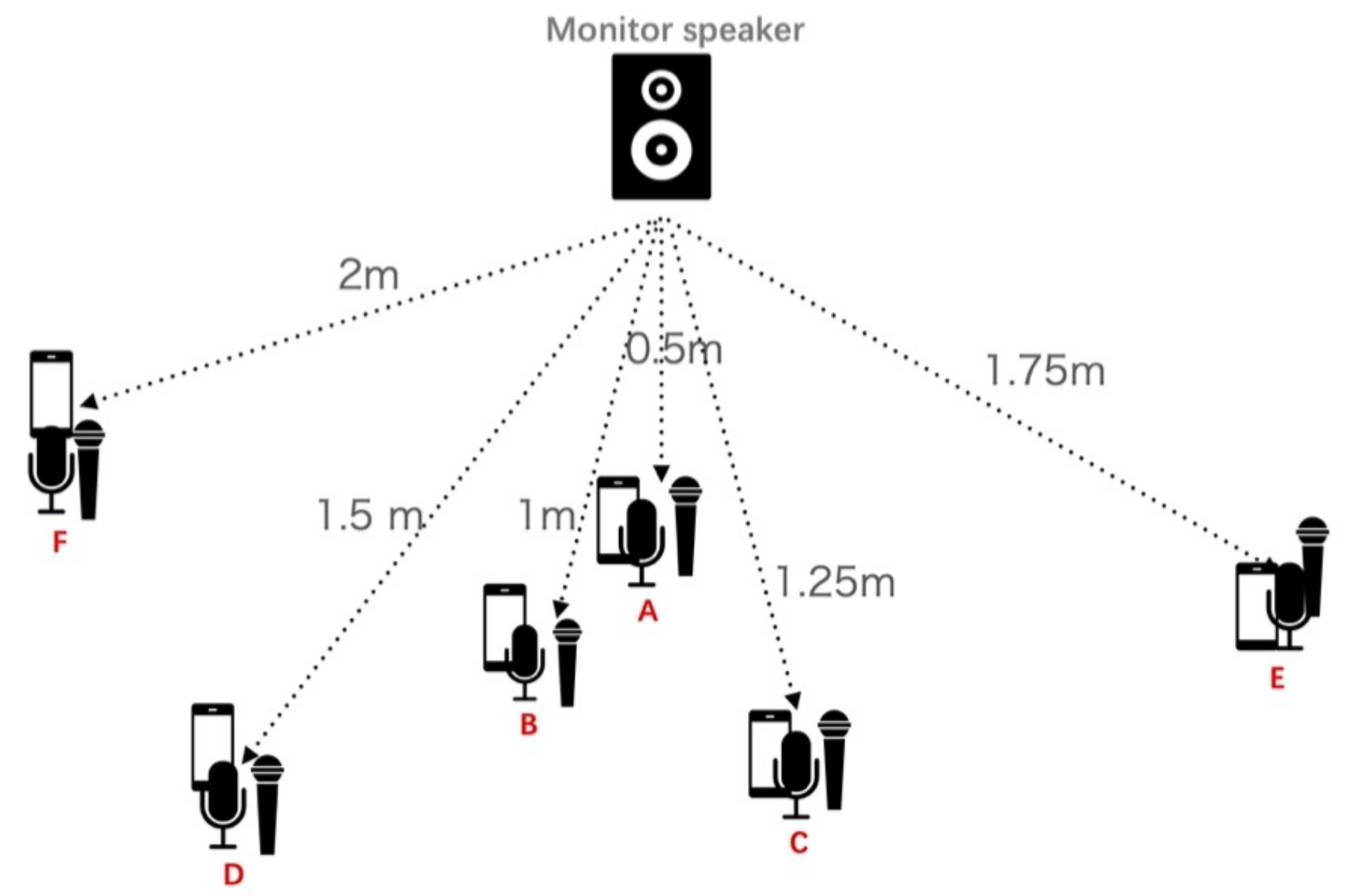
Goals

(1) Collect real noisy recordings by consumer-grade devices under uncontrolled environments (e.g., using iPhone at home)

(2) Release such a parallel noisy speech dataset to facilitate research in speech enhancement

Collection Setup

(1) Play clean speech recordings using high-quality monitor speaker and re-record waveforms on various devices and environments
(2) Perform cross-correlation to align the device-degraded speech and original clean speech



Overview of DDS

Overall settings

Nealy 2,000 hours of speech data collected in 27 conditions:

Setting	Count	Description
Speech materials	2	DAPS, VCTK clean sets
Environments	9	conference rooms (2), offices (2), studios (3), living room (1), waiting room (1)
Devices	3	iPad Air (MEMS), Uber Mic (condenser), MPM-1000 (condenser)

Device positions	6	A(50 cm, 0°), B(100 cm, 15°)
		C(125 cm, 30°), D(150 cm, 45°)
		E(175 cm, 60°), F(200 cm, 75°)

- 2 speech materials: DAPS (4 hours) and VCTK subset (8 hours)
- 9 realistic rooms
- 3 microphone devices
- 6 recording positions to collect speech at various noise and reverberation levels

Initial analysis

Average PESQ and ESTOI on different conditions

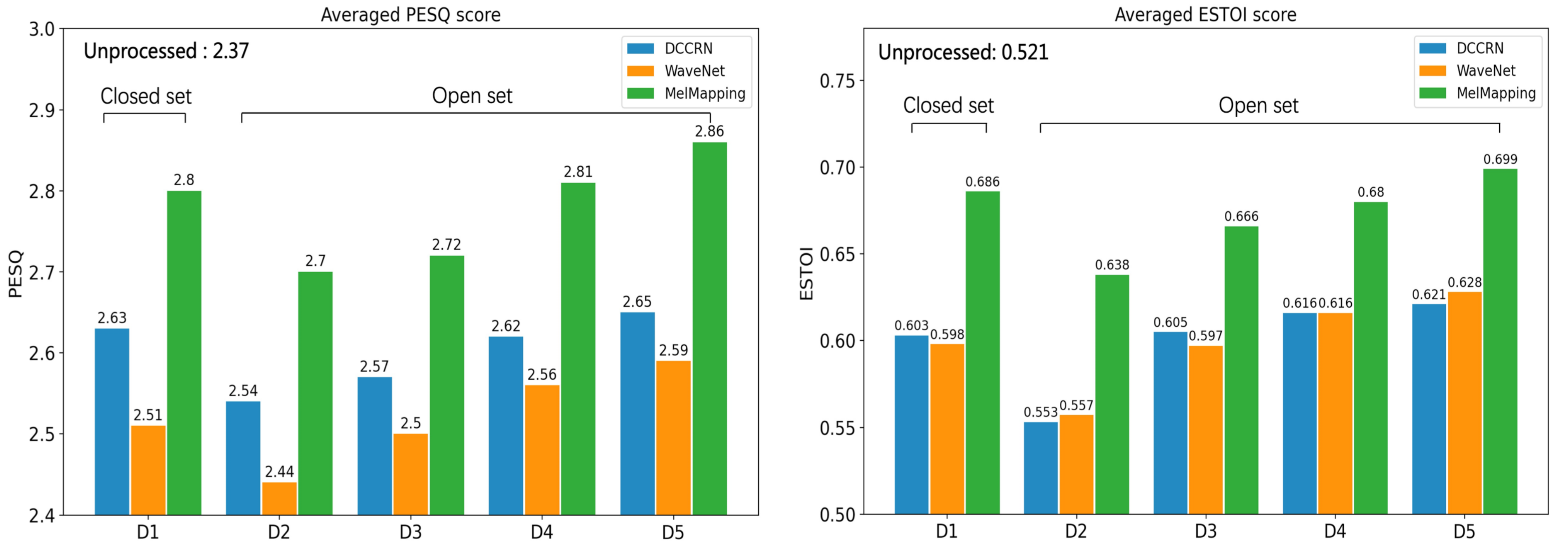
Environment	DAPS portion		VCTK portion	
	PESQ	ESTOI	PESQ	ESTOI
confroom1	2.34	0.715	2.58	0.630
confroom2	1.98	0.617	2.27	0.527
office1	2.60	0.758	2.80	0.660
office2	2.31	0.724	2.54	0.627
studio1	2.37	0.725	2.59	0.602
studio2	3.01	0.815	3.10	0.735
studio3	3.10	0.811	3.16	0.735
waitingroom1	3.02	0.796	3.13	0.722
livingroom1	2.34	0.723	2.61	0.647

Device position	DAPS portion		VCTK portion	
	PESQ	ESTOI	PESQ	ESTOI
A (50cm, 0°)	3.22	0.901	3.32	0.840
B (100cm, 15°)	2.77	0.810	2.94	0.728
C (125cm, 30°)	2.57	0.770	2.78	0.680
D (150cm, 45°)	2.44	0.720	2.65	0.624
E (175cm, 60°)	2.27	0.656	2.50	0.557
F (200cm, 75°)	2.11	0.597	2.35	0.495

Device	DAPS portion		VCTK portion	
	PESQ	ESTOI	PESQ	ESTOI
iPad	2.35	0.688	2.56	0.585
Uber Mic	2.66	0.767	2.85	0.684
MPM-1000	2.68	0.773	2.86	0.693

Baseline results on test set

- D1: 32 hours speech extracted from matched room and matched mic (closed-set)
- D2: 32 hours speech extracted from one unmatched room and one unmatched device
- D3: 32 hours speech extracted from one unmatched room and two unmatched device
- D4: 32 hours speech extracted from four unmatched room and two unmatched device
- D5: 32 hours speech extracted from eight unmatched room and two unmatched device



(a) PESQ score

(b) ESTOI score

Conclusion

➢ We release a large-scale device-degraded speech (DDS) dataset with 2,000 hours of real recordings collected under 27 conditions spanning 9 realistic rooms and 3 devices

Links

Dataset access