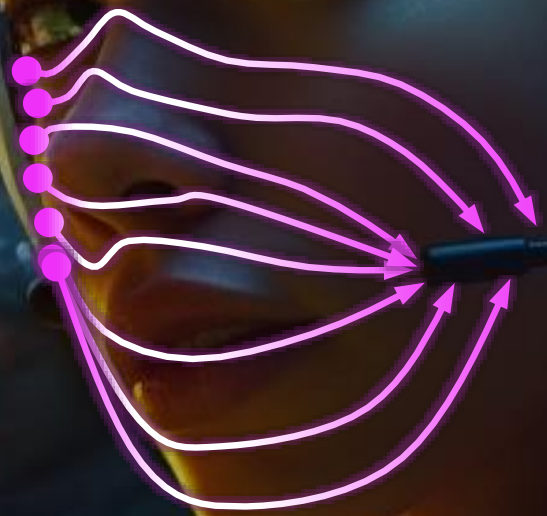
EarSE: Bringing Robust Speech Enhancement to COTS Headphones

Di Duan, Yongliang Chen, Weitao Xu, Tianxing Li

In Proceedings of UbiComp 2024

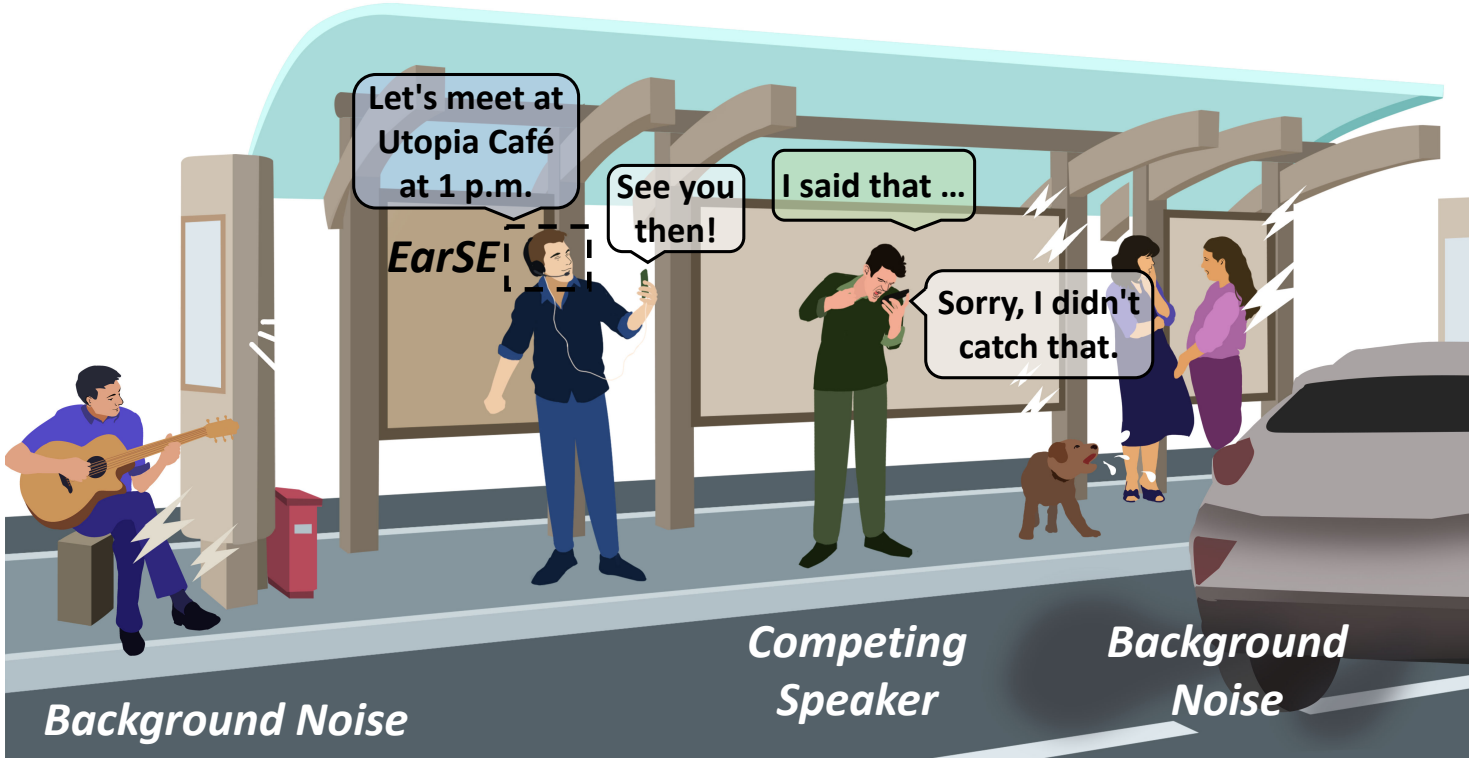


Presenter: Di Duan
Email: duandiacademic@gmail.com

香港城市大學
City University of Hong Kong

Motivation

Ambient Noise



Outline

- Existing Solutions
- Key Idea
- System Design
- Evaluation

Existing Solutions

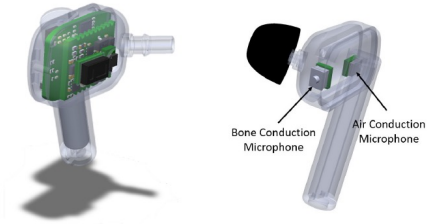
Solutions Supported by Hardware

Smartphone[1-2]

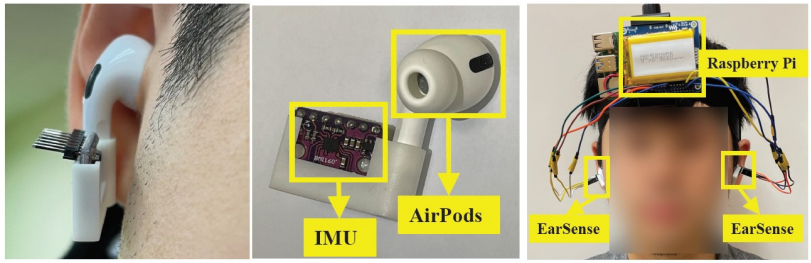
Airpods

Galaxy Buds

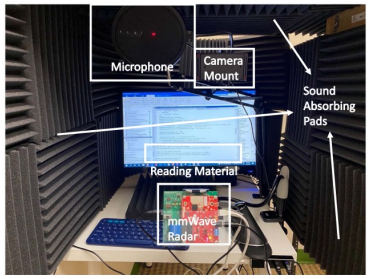
ClearBuds[3]



In-Ear-Voice[4]



EarSense[5]



RadioSES[6]

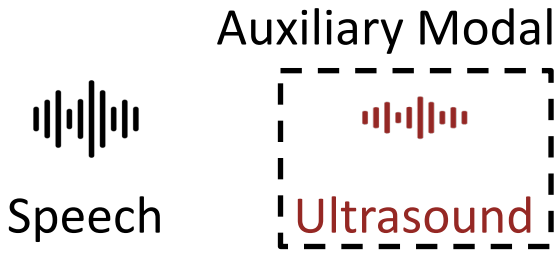
[1] Sun et al. UltraSE: single-channel speech enhancement using ultrasound.
[2] Ding et al. Ultraspeech: Speech enhancement by interaction between ultrasound and speech.
[3] Chatterjee et al. ClearBuds: wireless binaural earbuds for learning-based speech enhancement.
[4] Schilk et al. In-ear-voice: Towards milli-watt audio enhancement with bone-conduction microphones for in-ear sensing platforms.
[5] He et al. Towards Bone-Conducted Vibration Speech Enhancement on Head-Mounted Wearables.
[6] Ozturk et al. Radio SES: mmWave-Based Audioradio Speech Enhancement and Separation System.

Outline

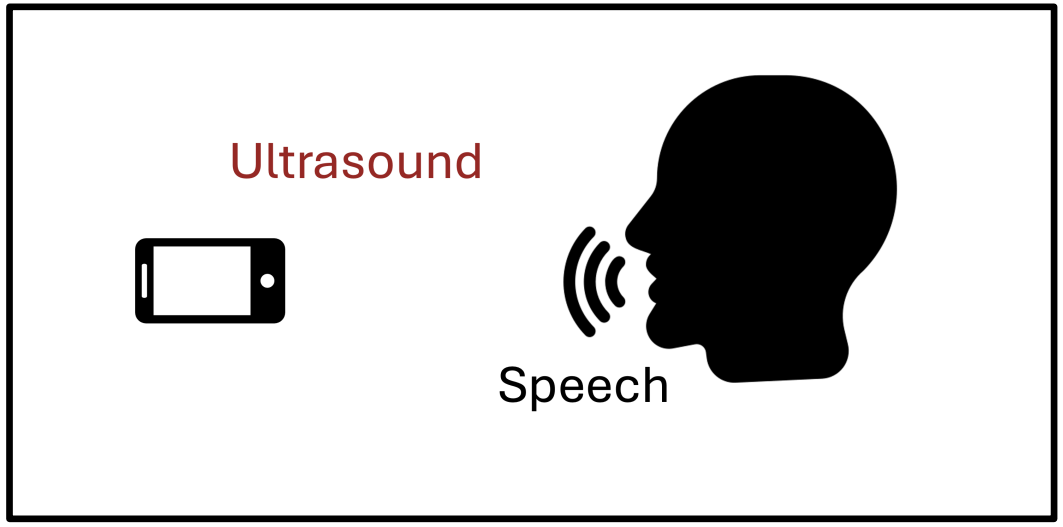
- Existing Solutions
- Key Idea
- System Design
- Evaluation

Key Idea

Embedding ultrasound into user speech using a COTS headphone

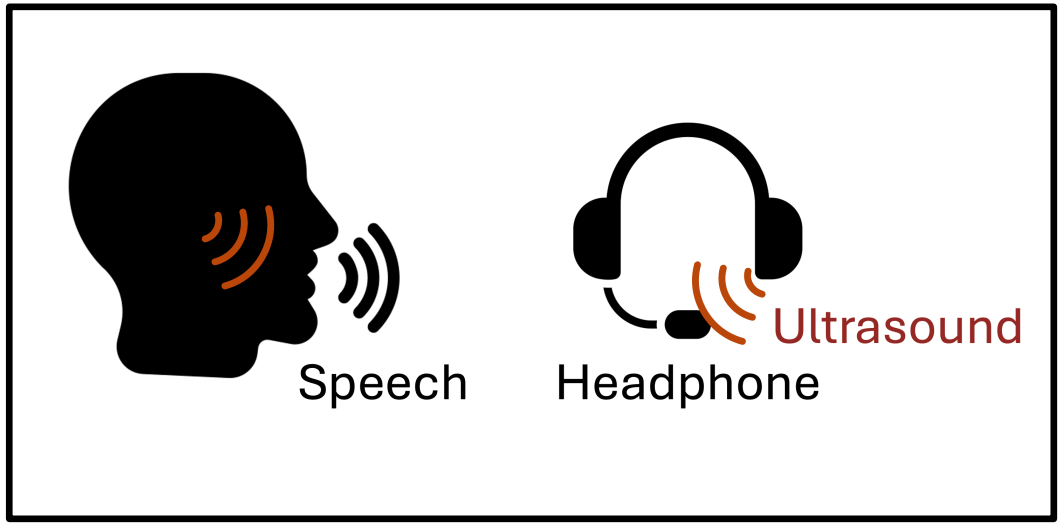


Previous Work



- Vulnerable to hand tremors
- Rely on the line of sight (LOS)

Our Idea



- Head-mounted manner provides robustness
- Leveraging the non-line-of-sight (NLOS)

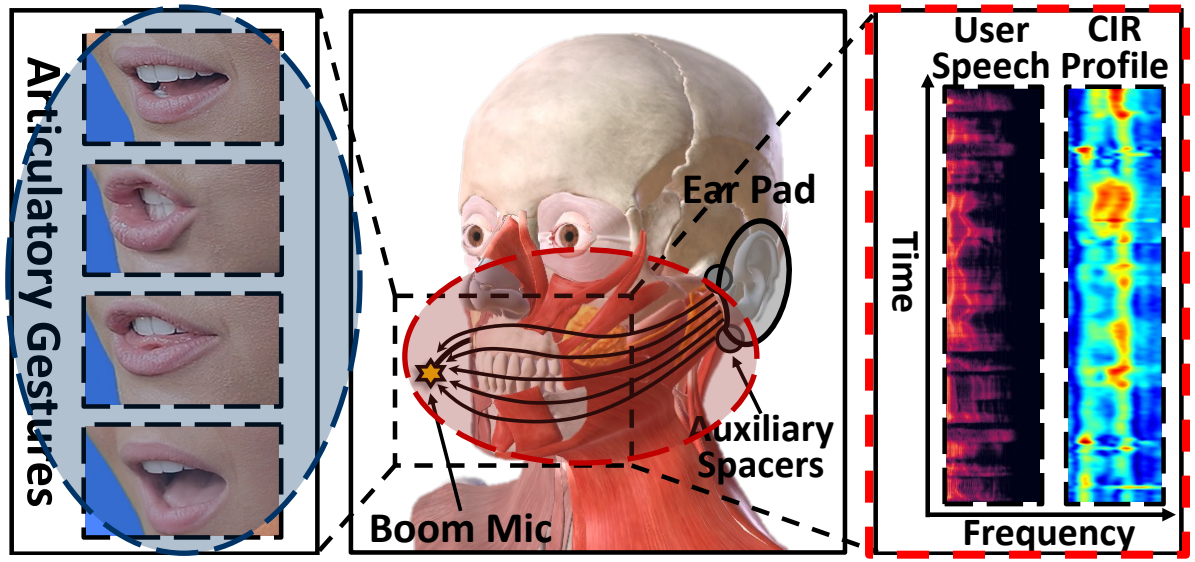
Outline

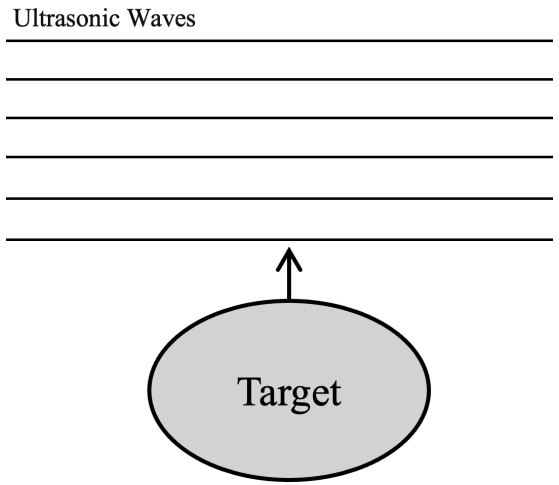
- Existing Solutions
- Key Idea
- System Design
- Evaluation

Sensing Rationale

The perturbation of an ultrasound sensing field caused by articulatory gestures

The user speech and ultrasound are aligned perfectly in a single channel manner.

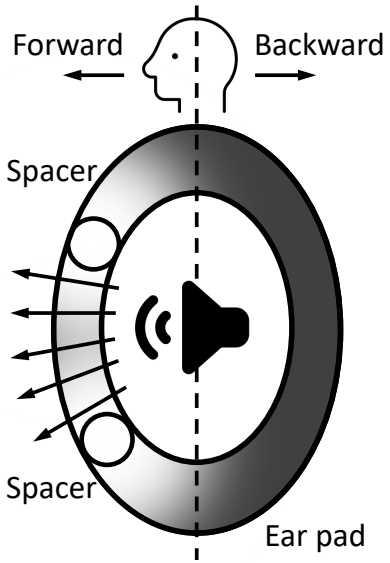




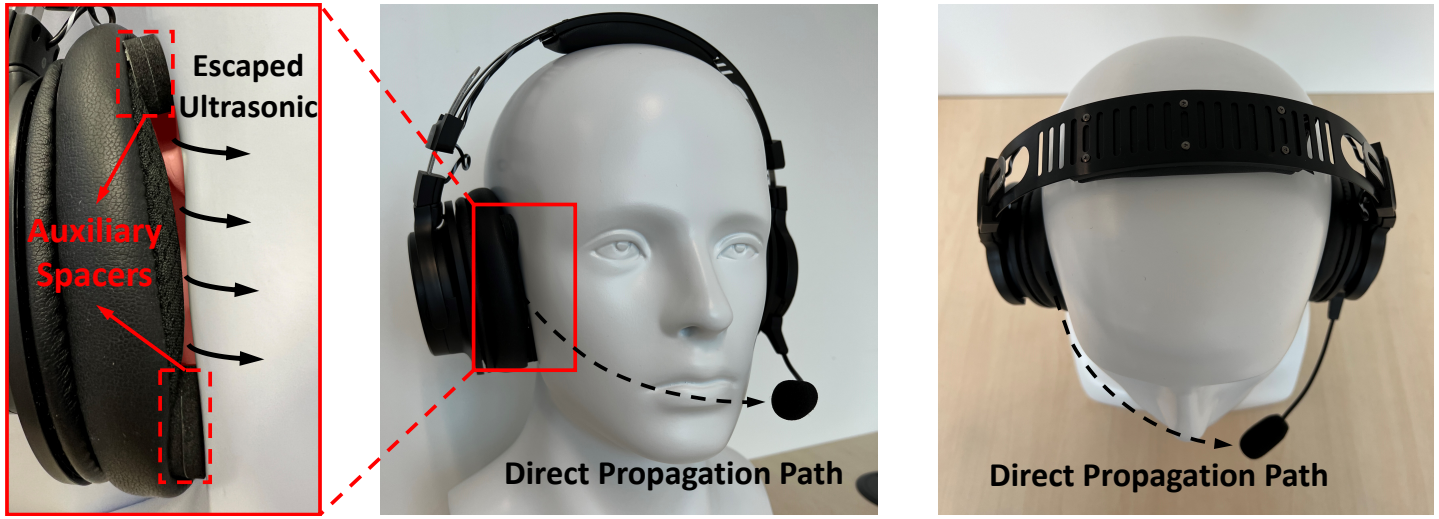
Articulatory Gestures → Perturb → Acoustic sensing field

Prototype

Increasing the SNR of escaped ultrasound

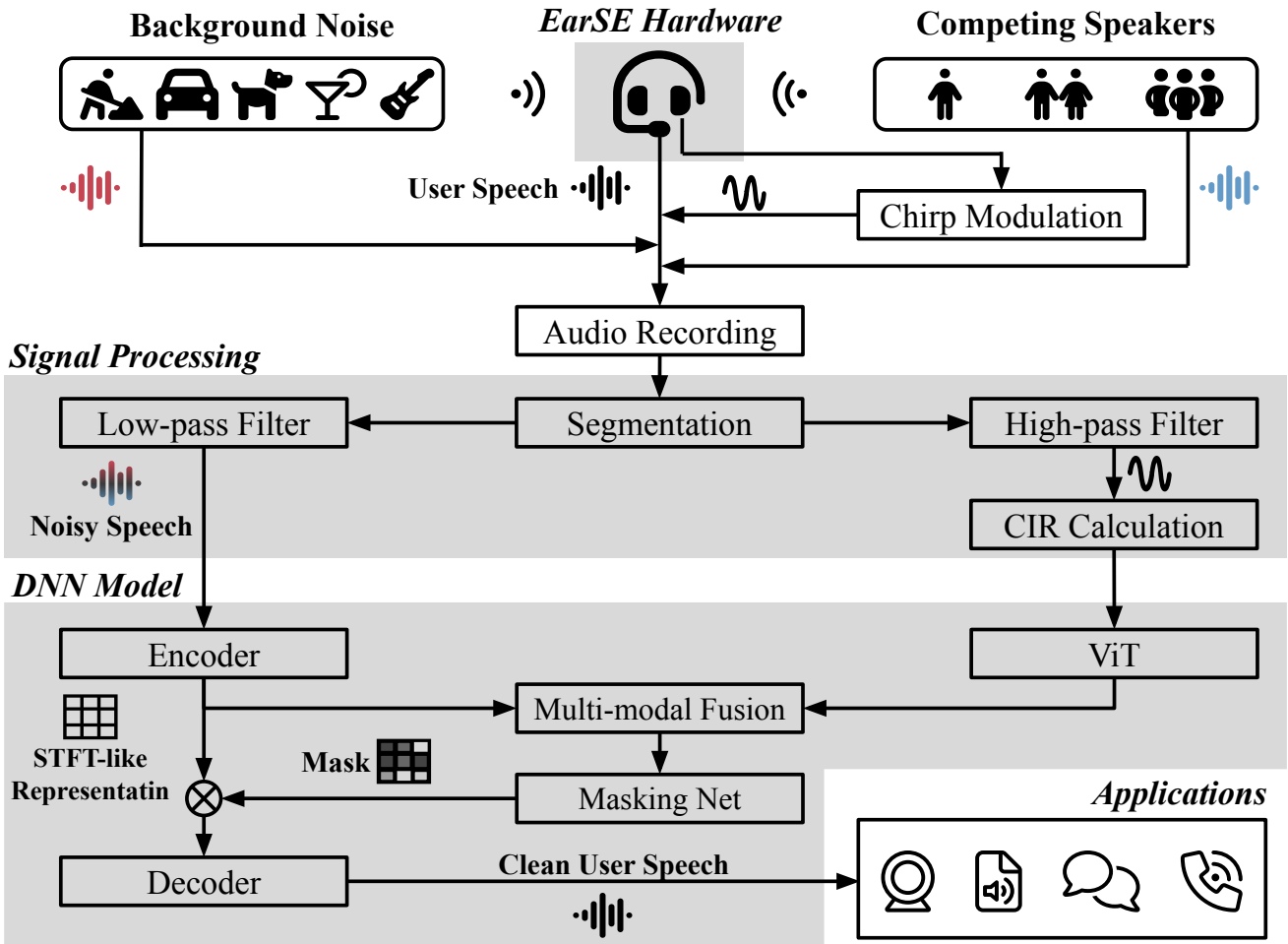


Auxiliary spacers



EarSE hardware implementation

System Overview



Outline

- Existing Solutions
- Key Idea
- System Design
- Evaluation

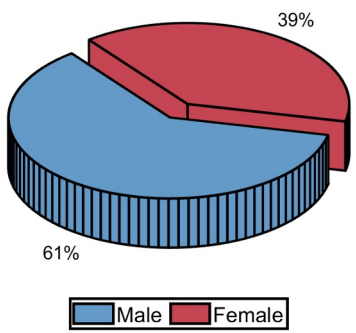
Experiment Settings

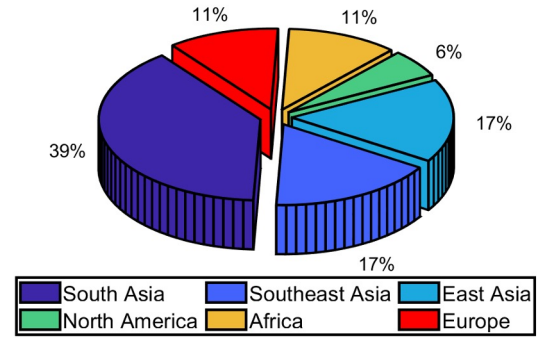
- Four common evaluation metrics: SiSNR (Clarity), SiSDR (Distortion), STOI (Intelligibility), and PESQ (Quality).
- Three devices: Logitech G733, ATH-G1WL, Sony XM4+Antlion.
- 18 native English speakers from 11 countries; the training dataset includes 273k five-second segments of noisy speech.
- 7 strong baselines, including SOTA hardware-based multi-modality solutions.
- Trained on a desktop equipped with a single NVIDIA RTX 3090 GPU.

Logitech G733

ATH-G1WL

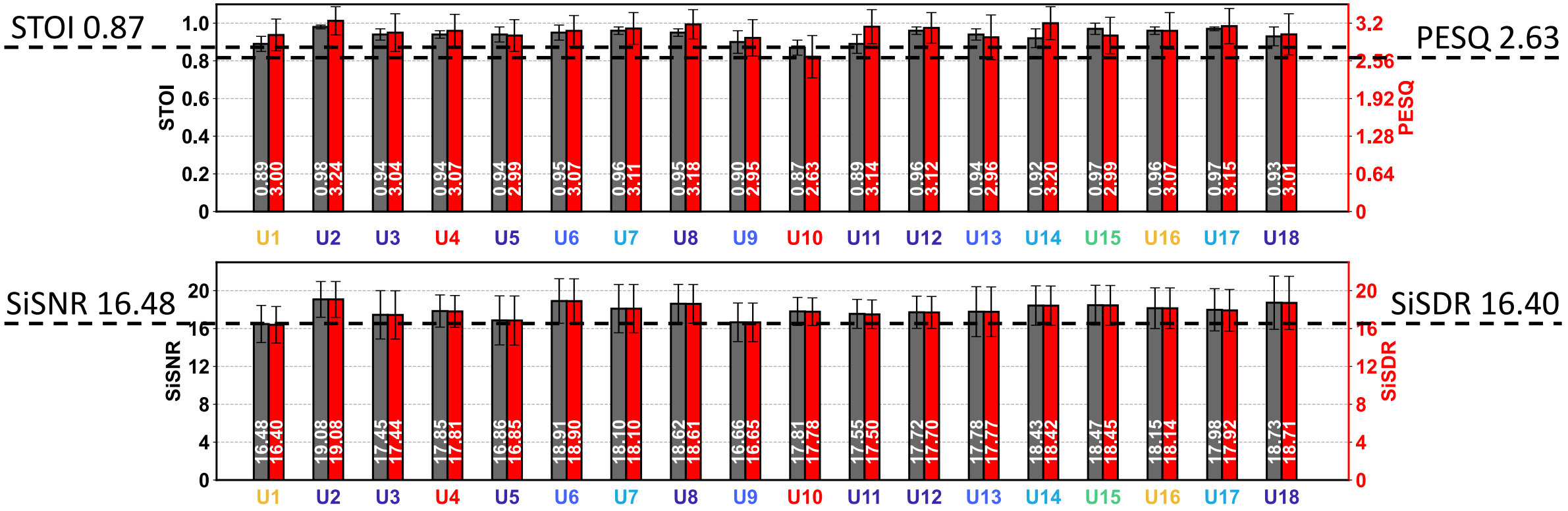
Sony XM4+Antlion





Demography

Overall Performance



Baselines

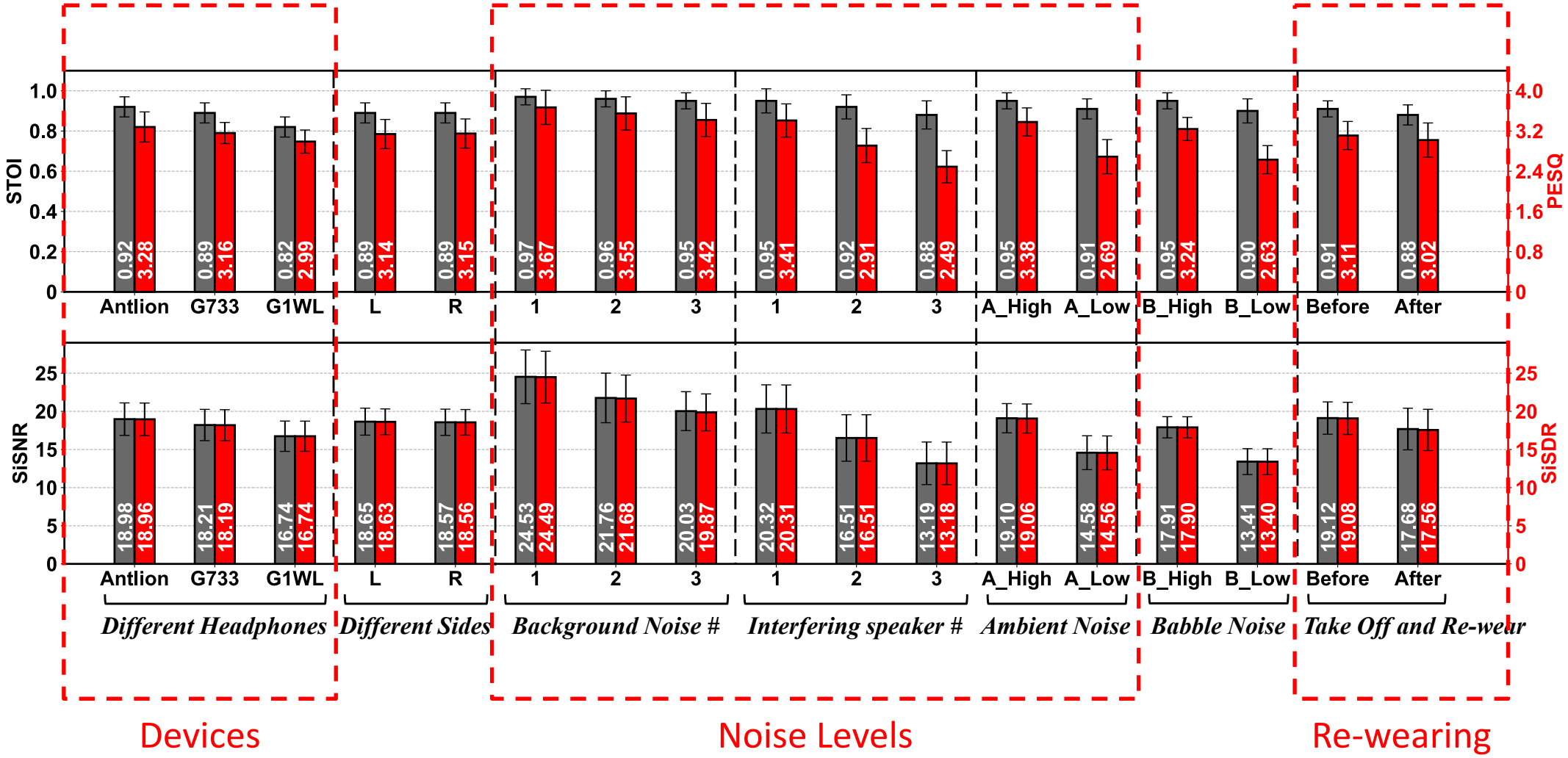
Multi-Modal Solution

Deep Learning Solution

Table 3. Performance comparison with seven baselines.

Methods	SiSNR	SiSDR	STOI	PESQ	MOS
EarSE	**19.48**	**19.47**	**0.95**	**3.32**	**4.43**
UltraSE	15.42	15.74	0.85	3.01	4.03
UltraSpeech	13.52	13.52	0.81	2.87	3.47
SepFormer	17.14	17.13	0.89	3.02	4.10
AvaTr V2	17.00	17.00	0.94	2.83	3.87
PHASEN	13.57	13.63	0.82	2.94	3.67
Conv-TasNet	12.46	12.45	0.77	2.58	2.87
VoiceFilter	11.65	11.65	0.86	2.23	2.97
Noisy speech	6.07	4.86	0.73	1.87	1.00

Micro-benchmarks



Thanks for your attention!

Email: duandiacademic@gmail.com

Personal Homepage