EarSpeech:

Exploring In-Ear Occlusion Effect on Earphones for Data-efficient Airborne Speech Enhancement

Feiyu Han[1,2], Panlong Yang[1], You Zuo[2], Fei Shang[2], Fenglei Xu[3], and Xiang-Yang Li[2].

[1] Nanjing University of Information Science and Technology (NUIST), China
[2] University of Science and Technology of China (USTC), China
[3] Suzhou University of Science and Technology (SUST), China

Presenter: Feiyu Han

Motivation

- **Voice Input on Earphone Slides**

The quality of recorded speech on earphones drops extremely in **noisy** environments, negatively impacting the user experience.

Speech Enhancement Technology Toward Earphones

Audio-only

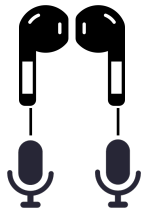
Spectral subtraction, filtering, decomposition, DL-based denoising

- Performance limit

Multi-modality

Dual Earphones

[Mobisys'21]



- Distortion in angle of users
- Unable to be applied to single-earphone scenarios

Single Earphone

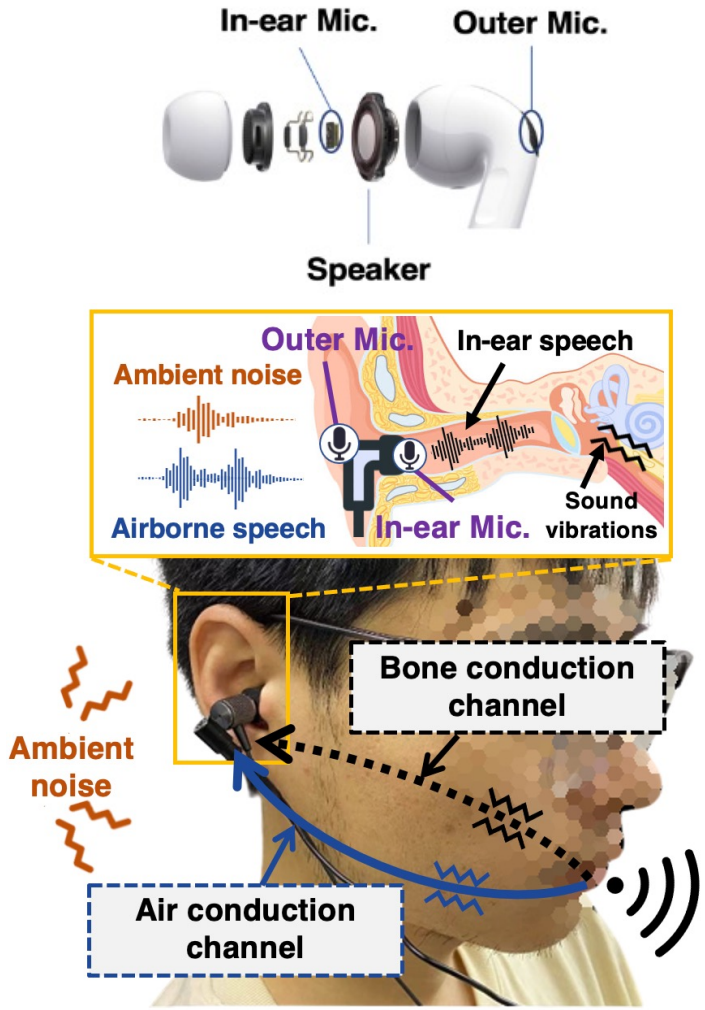
[Mobisys'22,TASLP'22]

- IMU with the high sampling rate
- Bone-conduction Mic.

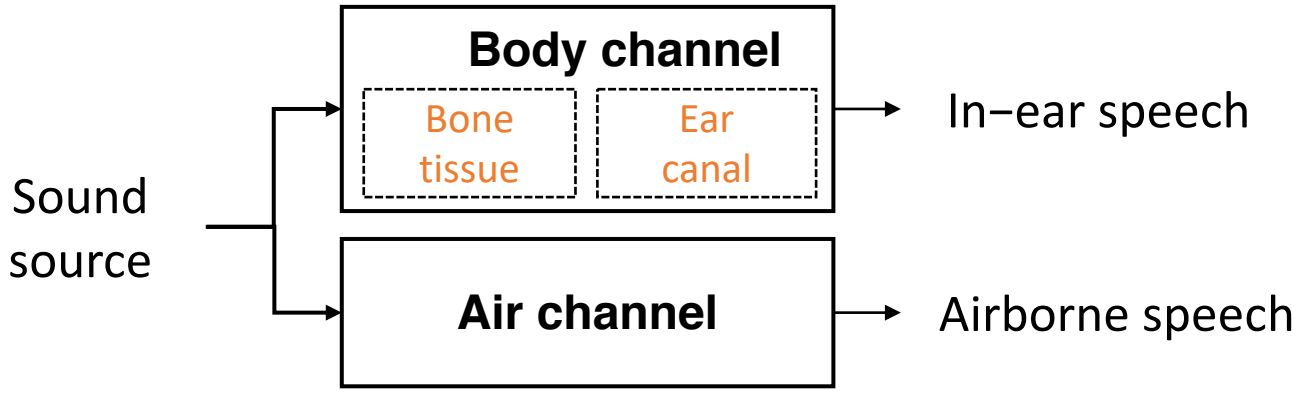
Meets the single-earphone usage scenario and has low requirements for onboard sensor performance

Opportunity

- **Dual-microphone structure**



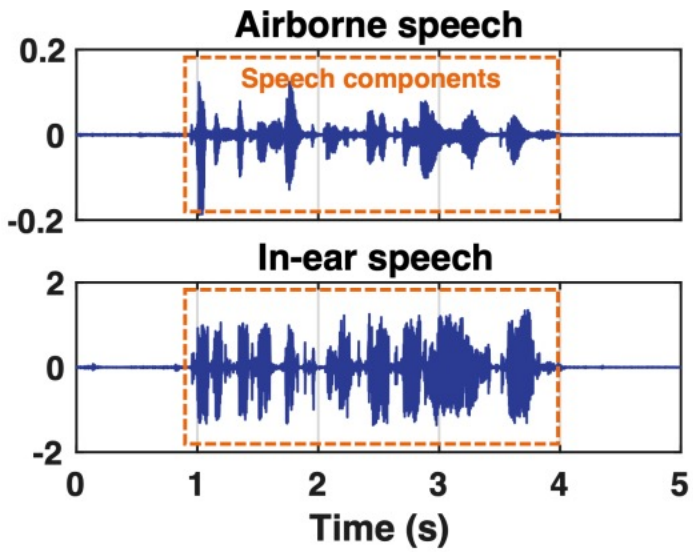
- **Two sound propagation channels**



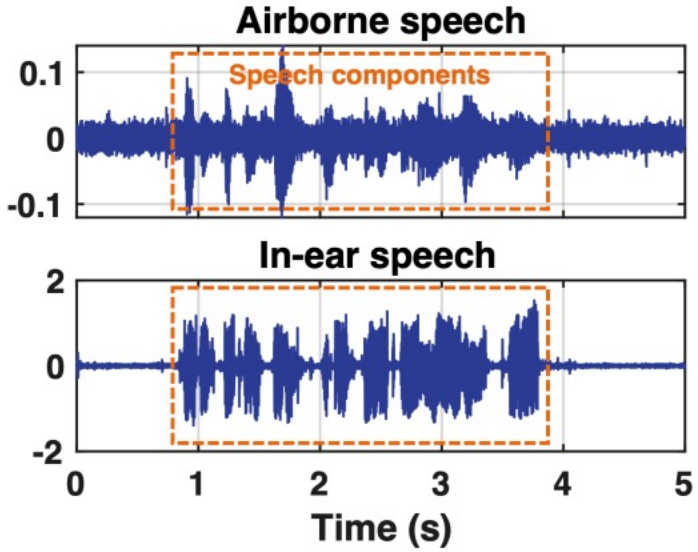
Can we take advantage of body channel to enhance the airborne speech?

Airborne Speech *vs.* In-ear Speech

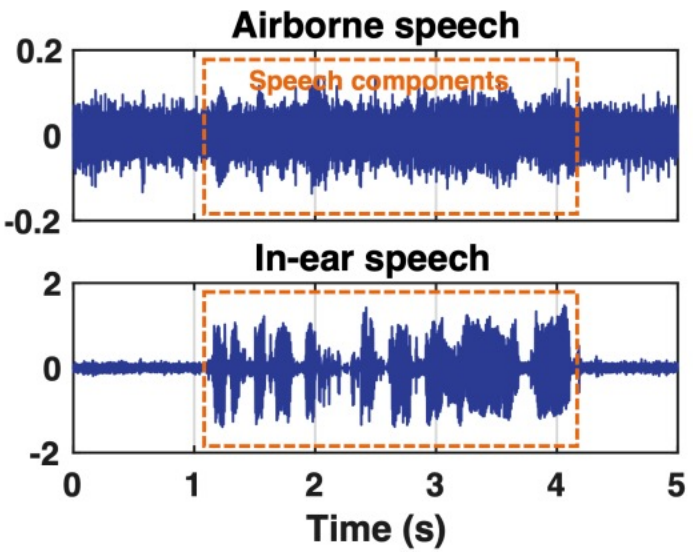
- **Noise Resistance Study**



(a) Noise SPL is 30.45 dB.



(b) Noise SPL is 52.25 dB.



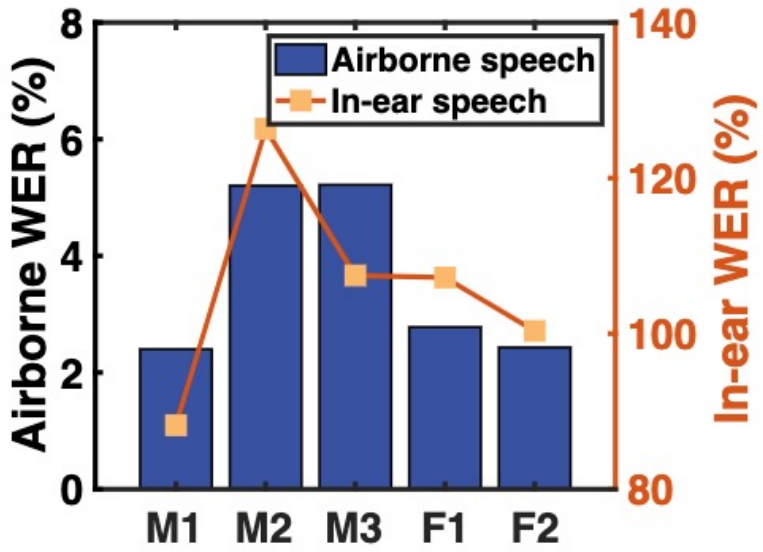
(c) Noise SPL is 61.96 dB.

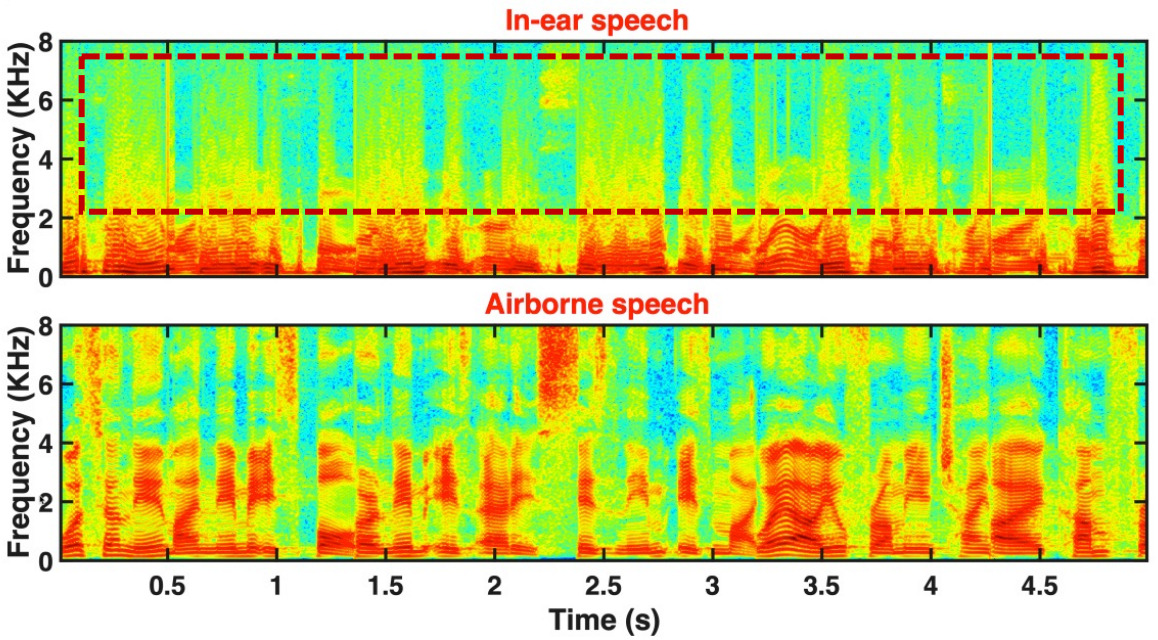
a) Air-channel ambient noise has subtle impact on in-ear speech that propagating the body channel.

b) The ear canals fit well with earplugs, blocking noise from entering the ear canal.

Airborne Speech *vs.* In-ear Speech

- **Speech Intelligibility**



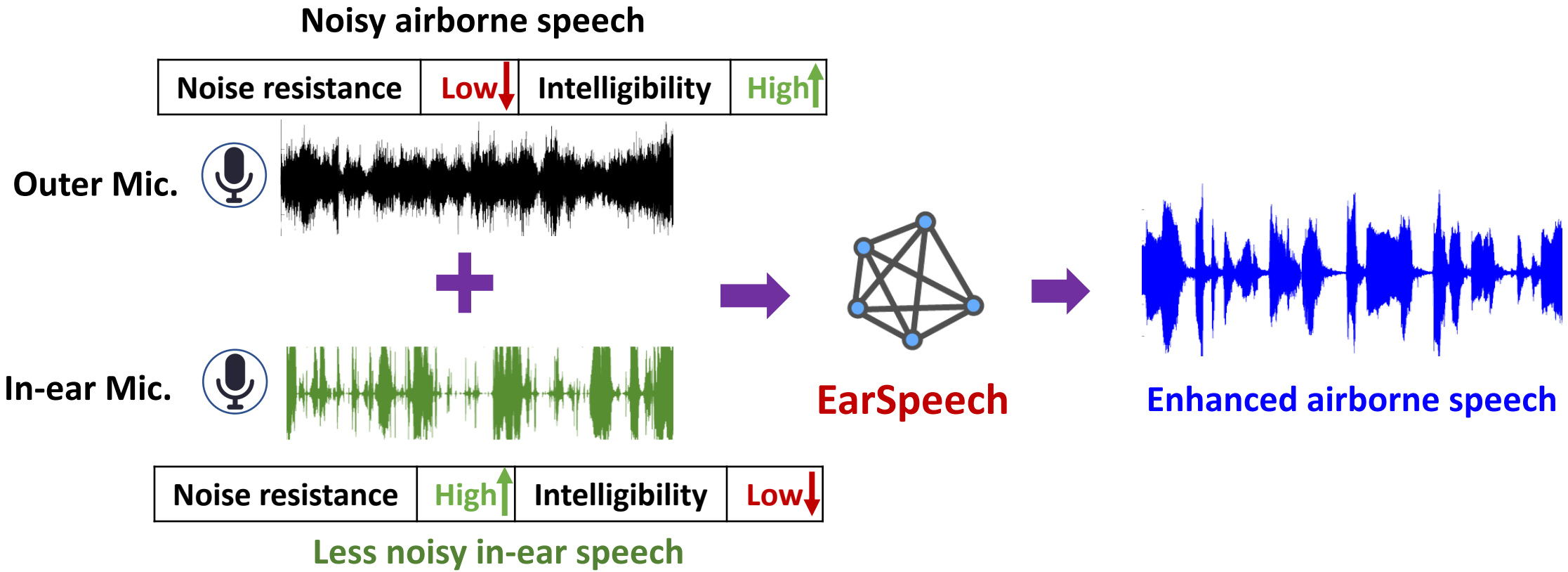


Airborne

In-ear

Bone conduction and occlusion effect make the loss of high-frequency components

Our Solution



Utilizing **in-ear speech** as the supplemental modality to enhance the quality of airborne speech in noisy environments

Technical Challenge-1

- **A sufficient dataset of paired airborne and in-ear speech with labels is still lacking.**

Airborne speech dataset **Large-scale**

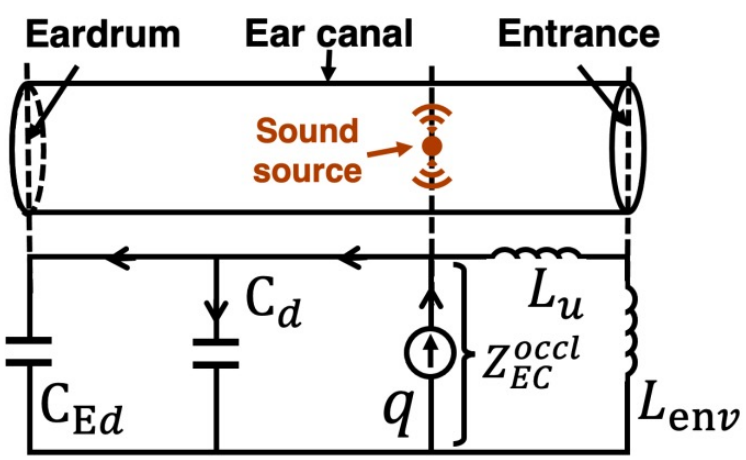
- ✓ LibriSpeech (292 000 utterances, 2000+ speakers)
- ✓ LibriVox (180 000 utterances, 9000+ speakers)
- ✓ VoxCeleb 1&2 (1 000 000+ utterances, 7000speakers)

In-ear speech dataset **Small-scale**

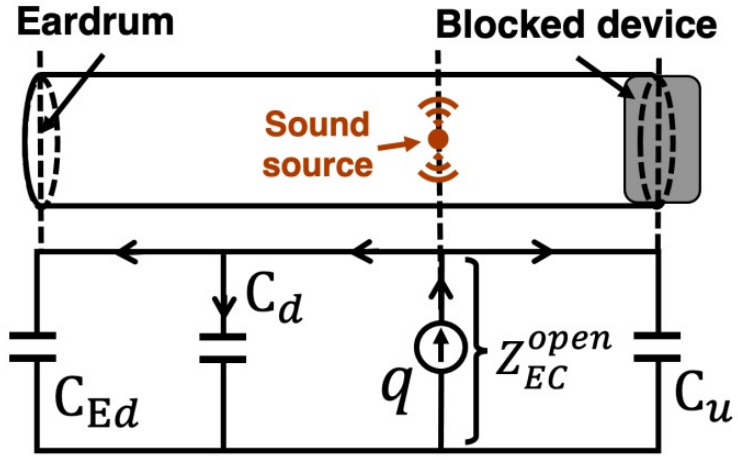
An intuitive way: manually collect airborne speech samples and corresponding in-ear speech samples in the lab environment.

EA-based Cross-channel Correlation Analysis

- Utilizing electro-acoustic (EA) model to conduct cross-channel correlation analysis



(a) Open case.



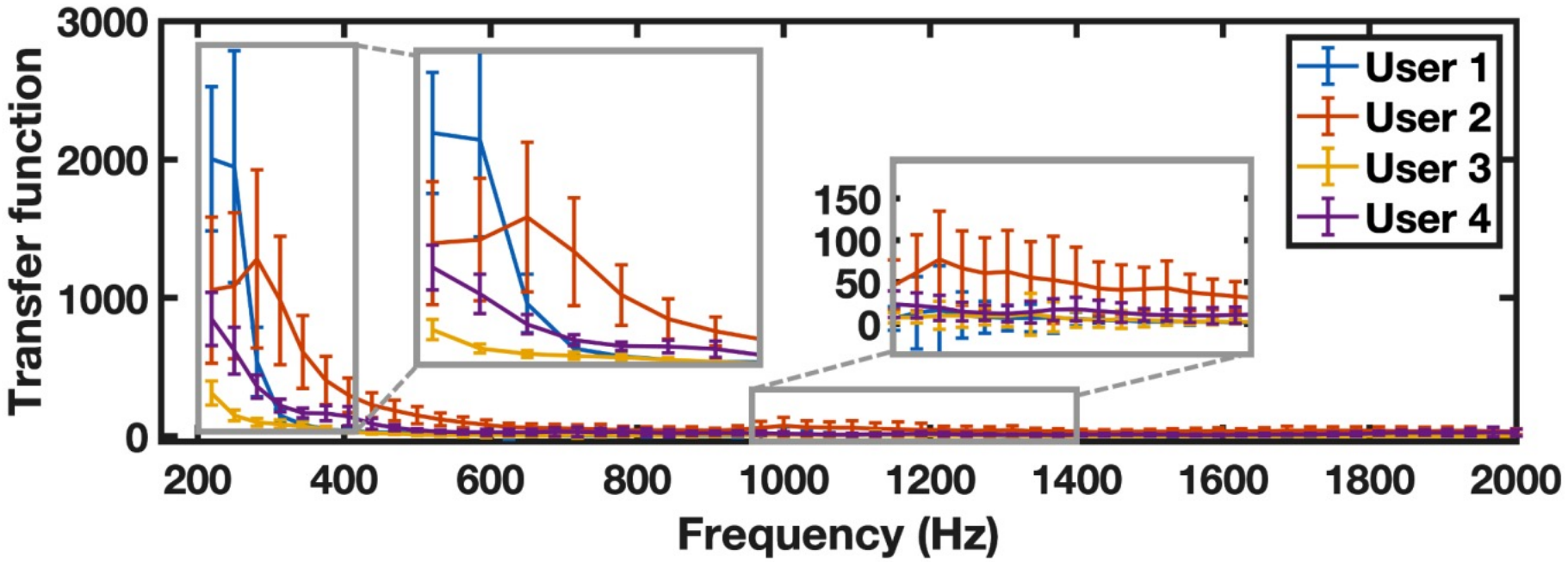
(b) Occluded case.

$$F_{OE} = P_{occl}/P_{open} = q * Z_{EC}^{occl}/q * Z_{EC}^{open} = \frac{\omega^2(C_{Ed}+C_d)(L_u+L_{env})-1}{\omega^2(C_{Ed}+|C_d+C_u)(L_u+L_{env})}$$

$$\Rightarrow \quad F_{tf} = \frac{s_{ie}(f)}{s_{air}(f)} = \frac{F_{OE}(f) * H_{bone}(f) * \cancel{s(f)}}{H_{air} * \cancel{s(f)}}$$

Transfer Function: a cross-channel correlation

Measurement of Transfer Function

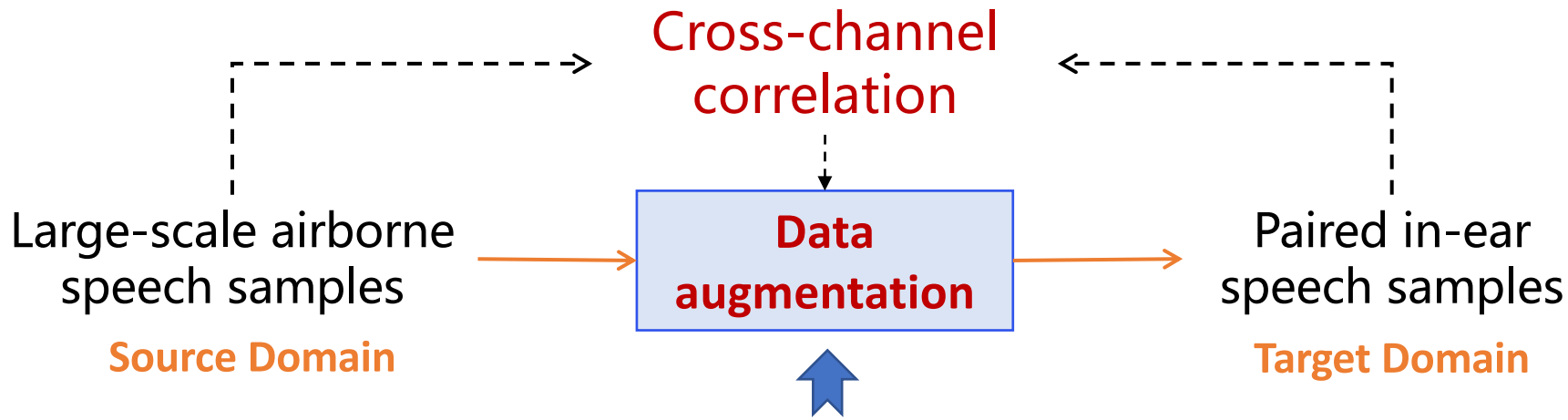


Observation: Transfer function varies person from person

$$F_{tf} = \frac{s_{ie}(f)}{s_{air}(f)} = \frac{F_{OE}(f) * H_{bone}(f) * \cancel{s(f)}}{H_{air} * \cancel{s(f)}}$$

Related to the geometry structure of the ear canal and skulls .

GMM-based Data Augmentation

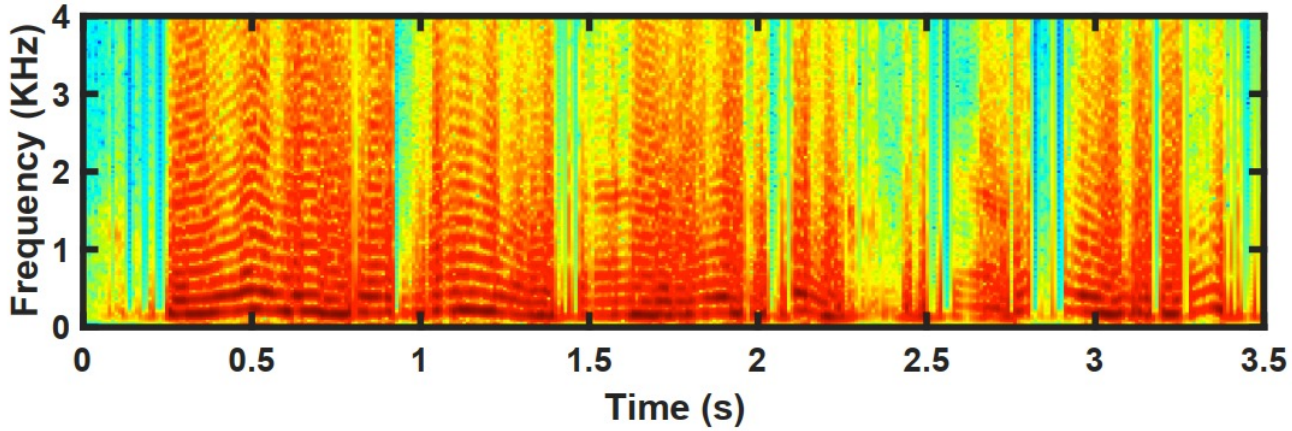


GMM-based Transfer Function Estimation

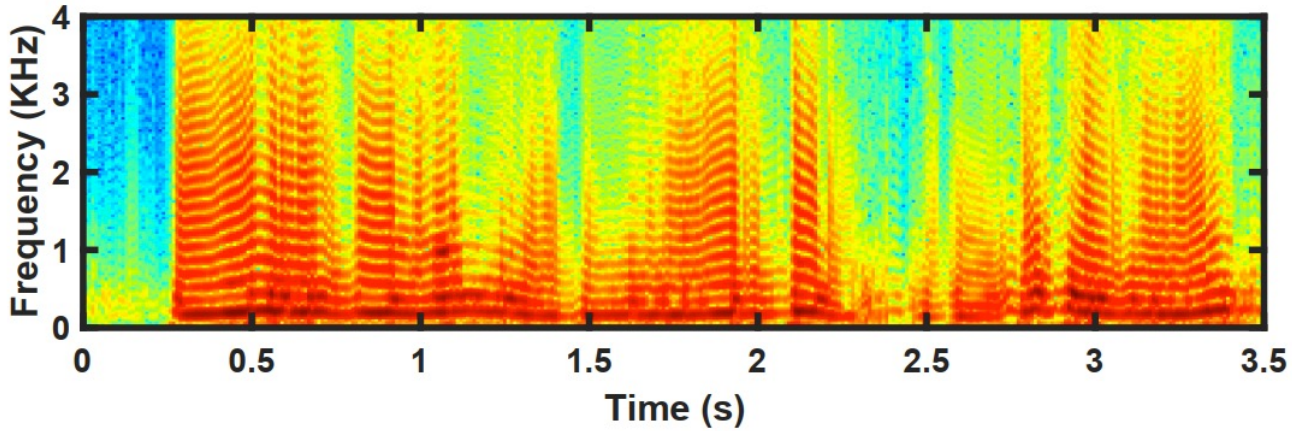
Advantages:

- Compared with hard mapping, it has higher generalization
- Compared with DL networks, less training samples are required

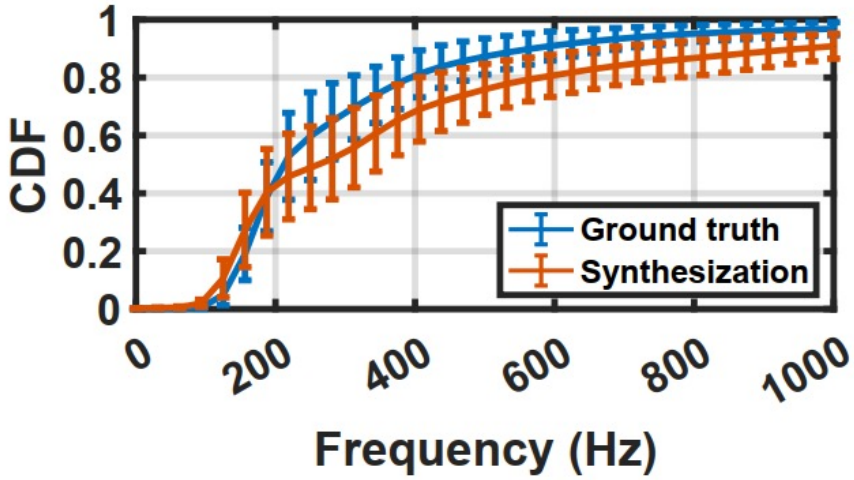
Examples of Data Augmentation



(a) Ground truth of in-ear speech.



(b) synthetic in-ear speech.



(b) Cumulative distribution.

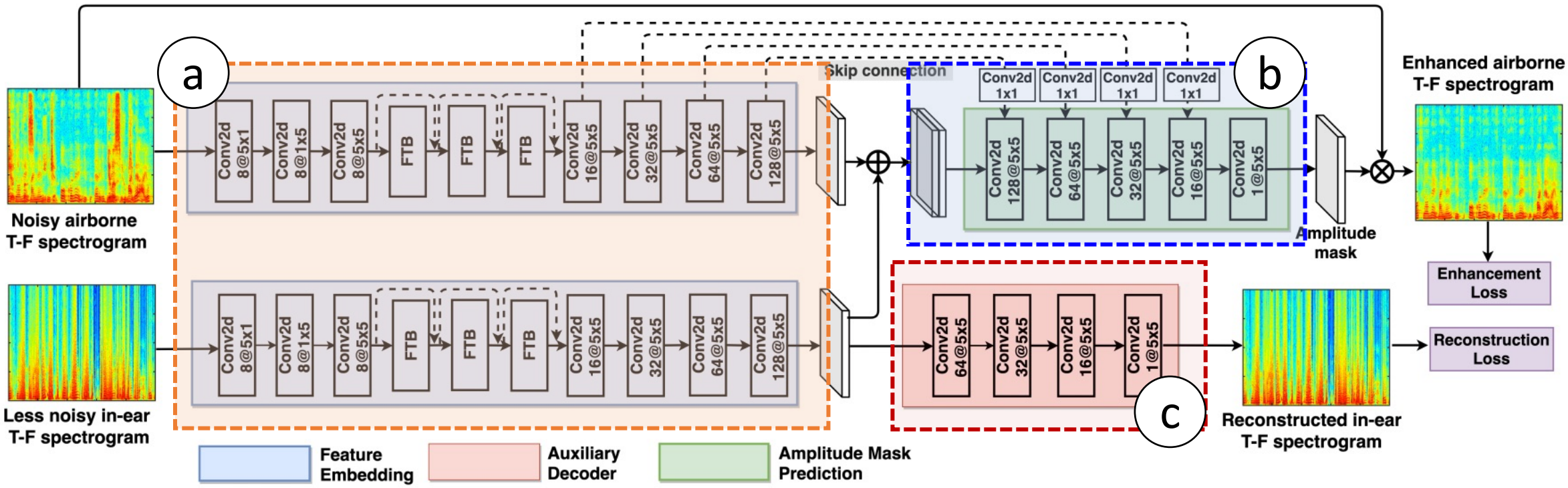
- The synthetic in-ear speech has a similar spectral structure with ground truth

Technical Challenge-2

- The heterogeneity in speech signals caused by diversities (e.g., different channels and speakers) makes it difficult to extract effective and generalized features from different speech channels for speech enhancement.

- **Diversities**
  - ✓ Channel difference
  - ✓ Body structure
  - ✓ Pronunciation habits
  - ✓ ……

Dual-channel speech enhancement network

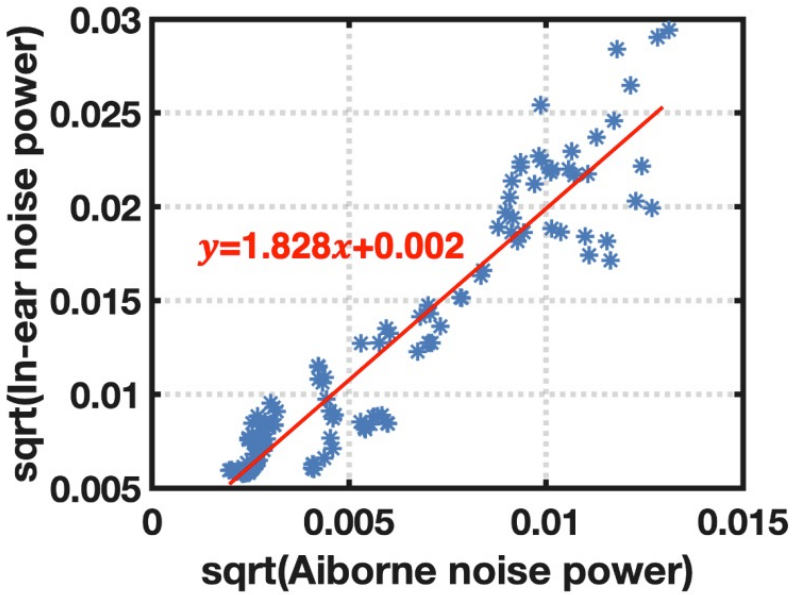


a) Extracting high-level representations in the same feature space to represent the cross-channel correlation.
b) Fusing dual-channel representations with **element-wise skip connections** to predict amplitude mask.
c) Enforcing model to learn the information of in-ear channel. Only participating in **training process**.
d) Convolutional network structure, the total parameters of the model are about **3.8 MB.**

Training Methodology

- **Dual-channel Noise mixture scheme.**

In extremely noisy environments, ambient noise indeed affects in-ear speech. Directly ignoring the impact of ambient noise on in-ear speech decreases the robustness of our system in the real world.



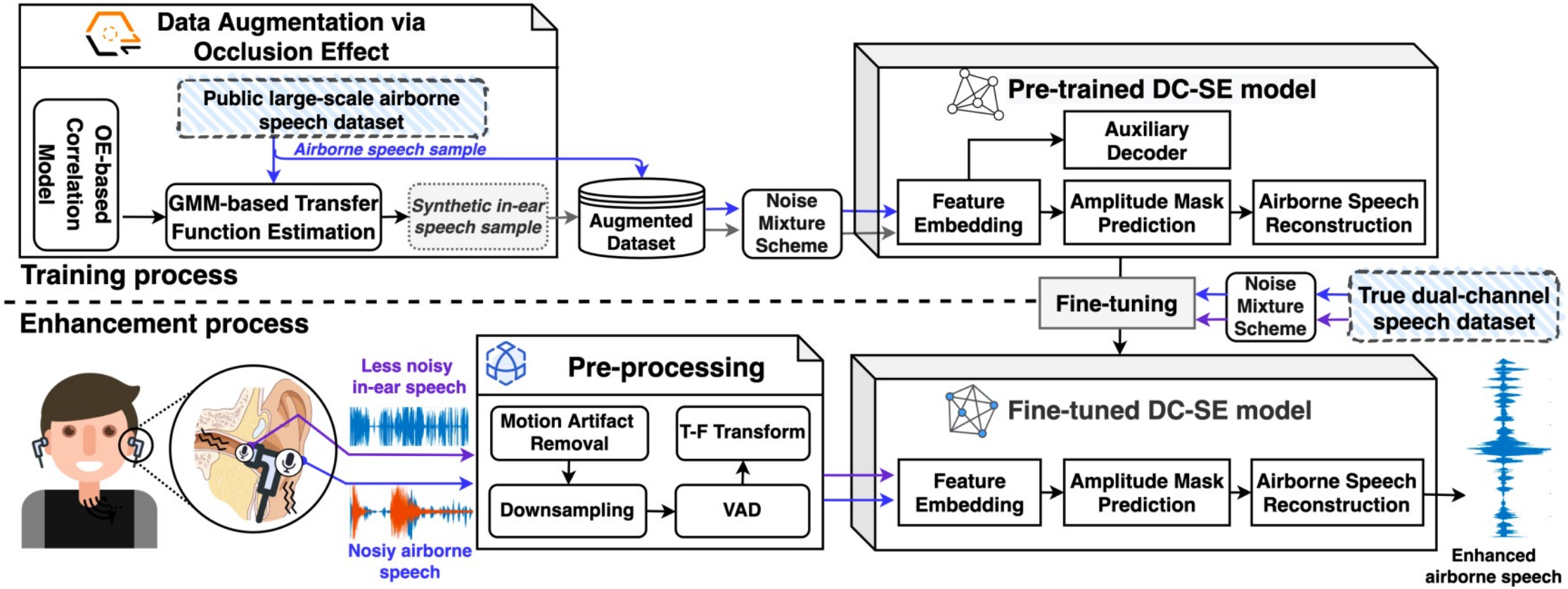
Field study: the linear relationship between in-ear noise and ambient noise

① Fitting a **linear function** between in-ear noise power and ambient noise power.

② Calculating the in-ear power noise according to ambient noise power.

③ Simultaneously adding corresponding noise to in-ear speech.

System Overview of EarSpeech



More technical details can be found in our paper.

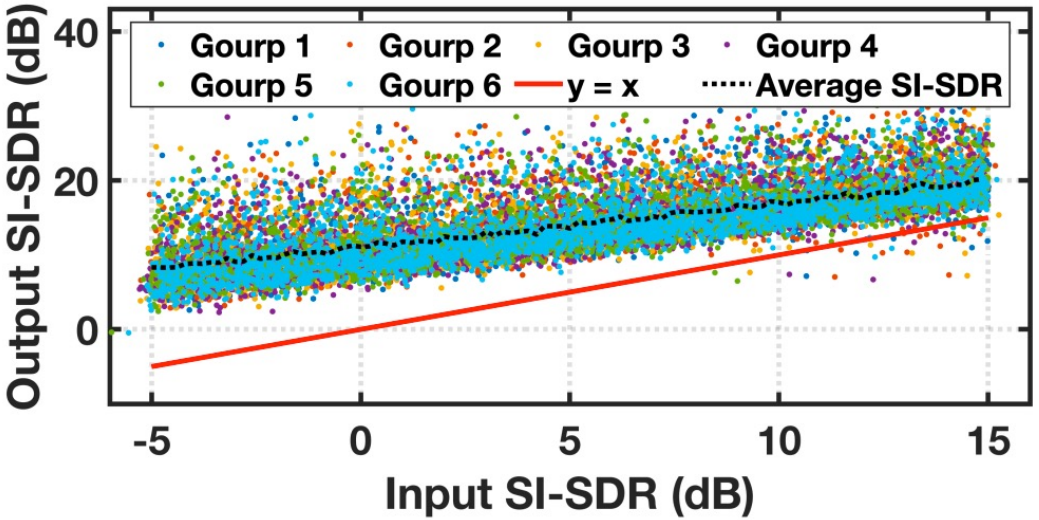
Performance Evaluation

- **Experimental Setup**

  Client-server Mode: Earphones are connected to a Laptop via a 3.5 mm jack adapter

- **Overall Performance:**

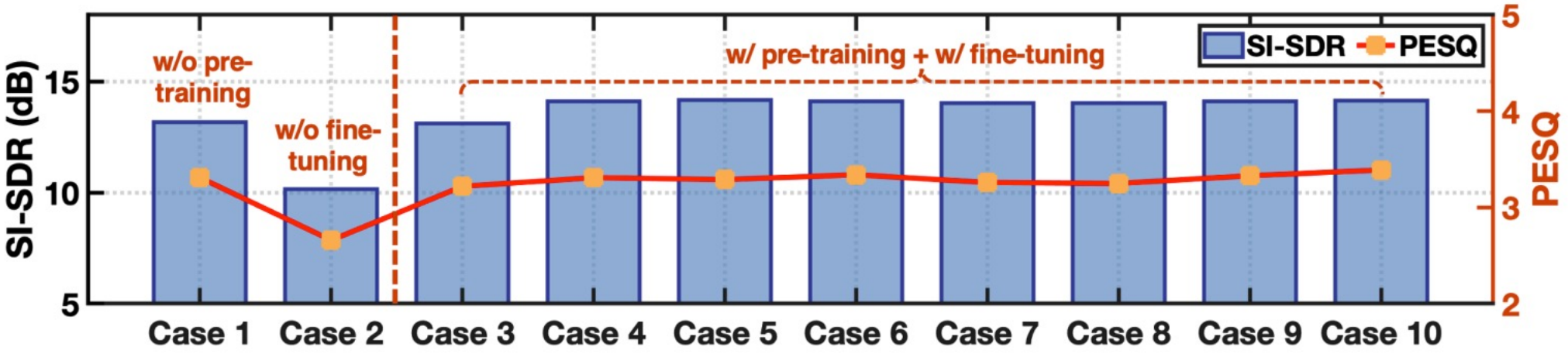
  18 participants, divided into 6 groups, leave-one-group-out cross validation.



Method	PESQ				STOI				SI-SDR (dB)			
	EN	MN	SN	Avg	EN	MN	SN	Avg	EN	MN	SN	Avg
Noisy speech	2.65	2.25	2.29	2.46	0.84	0.75	0.74	0.79	5.08	5.05	5.09	5.07
Phasen	3.24	3.00	2.91	3.05	0.86	0.85	0.80	0.84	11.05	9.93	9.36	10.11
EARSPEECH	**3.25**	**3.06**	**2.97**	**3.13**	**0.91**	**0.89**	**0.88**	**0.90**	**15.16**	**12.89**	**12.73**	**13.98**
-w/o FTB	3.08	2.78	2.70	2.91	0.88	0.85	0.84	0.87	13.80	11.66	10.96	12.55
-w/o SK	3.11	2.87	2.79	2.97	0.89	0.87	0.86	0.88	13.87	11.80	11.31	12.70
-w/o AD	3.16	2.93	2.78	3.01	0.90	0.88	0.86	0.88	14.09	12.16	11.52	12.96
-w/o IC	2.32	2.23	1.98	2.21	0.76	0.76	0.68	0.74	5.93	5.88	3.15	5.24

Performance Evaluation

- **Data Augmentation Effectiveness**



1984 pairs of real samples

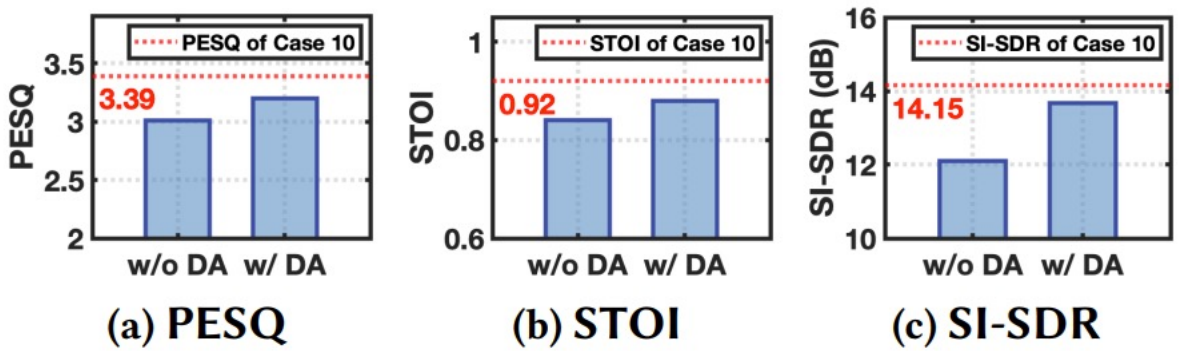
100 pairs of real samples + data augmentation

about 20x Data efficiency!

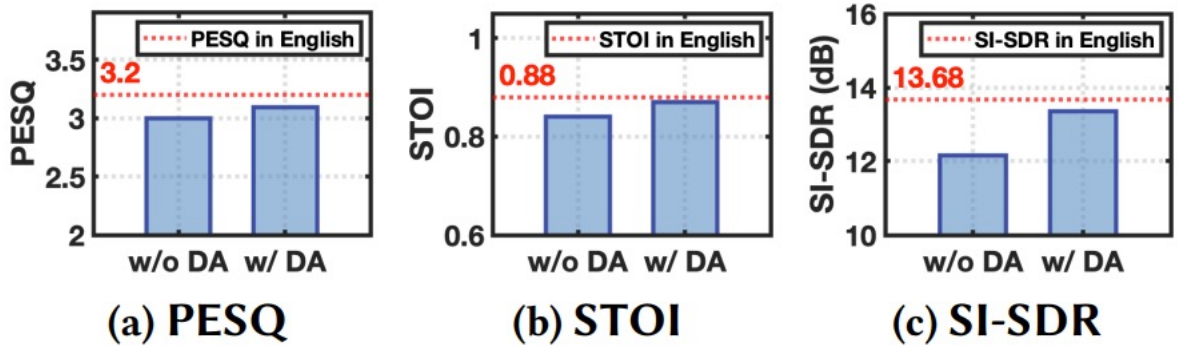
Performance Evaluation

Generalization Capability

a) New sentences.

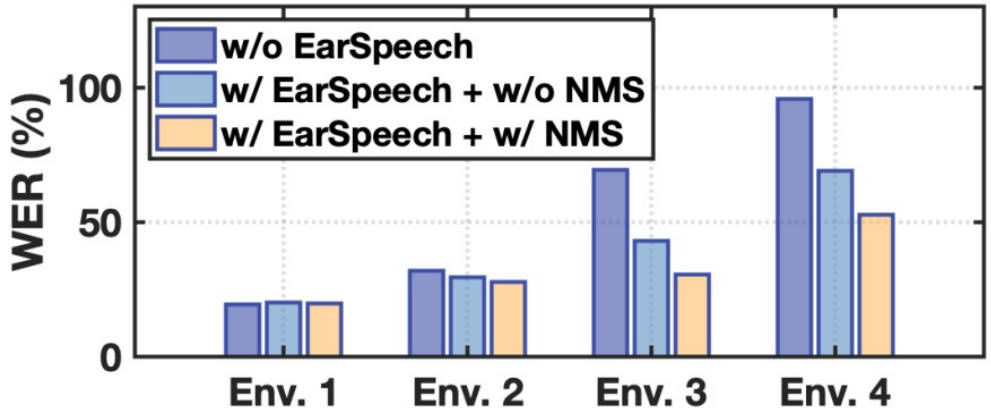


b) Mandarin language



Real-world Study

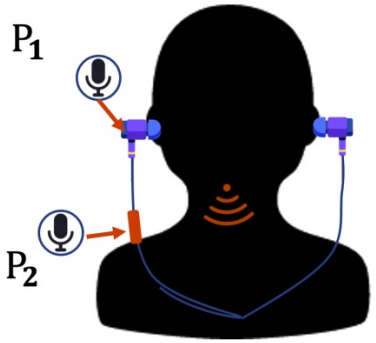
Reference	In your high school, most of the teachers there are helpful and friendly.
Noisy speech	In miao miao high school, most of the teachers they are here for are friends.
Enhanced speech	In your high school, the post of the teachers they are helpful and friendly.



Performance Evaluation

- **Robustness Study In Real World**

a) Impact of audio length.
b) Wearing position of earphones.
c) Impact of music playing.
d) Impact of earbud types.
e) ……



	(ET 1)	(ET 2)	(ET 3)	(ET 4)
Material	Memory foam	Silicone	Silicone	Silicone
Geometry	Single flange	Single flange	Double flange	Wingtips

- **Run-time Latency**

Platform	Pre-processing	Inference	Total
Laptop GPU	4.79 ms (± 0.72 ms)	38.39 ms (± 0.06 ms)	36.51 ms (± 7.35 ms)
Laptop CPU	7.80 ms (± 0.78 ms)	1.64 s (± 0.16 s)	1.71 s (± 0.13 s)

Audio Demo and Source Code are available on https://github.com/EarSpeech/earspeech.github.io

Discussion

1. Onboard Deployment

- OS Audio Interface’s Public Access.
- Lightweight Computing.

2. Music Replayed with a higher volume.

1. Music played from on-earphone speakers with the normal volume has subtle impact
2. "Speak-to-Chat” mode

……

Thanks for your listening

EarSpeech:

Exploring In-Ear Occlusion Effect on Earphones for Data-efficient Airborne Speech Enhancement

Feiyu Han[1,2], Panlong Yang[1], You Zuo[2], Fei Shang[2], Fenglei Xu[3], and Xiang-Yang Li[2].

[1] Nanjing University of Information Science and Technology (NUIST), China
[2] University of Science and Technology of China (USTC), China
[3] Suzhou University of Science and Technology (SUST), China

Email: fyhan@mail.ustc.edu.cn **Personal website:** https://fyhancs.github.io/