Wavoice: A Noise-resistant Multi-modal Speech Recognition System Fusing mmWave and Audio Signals

Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, Kui Ren

SenSys '21

Chorom Hamm

MMLAB-SNU

crhamm@mmlab.snu.ac.kr

Dec 16, 2021

Outline

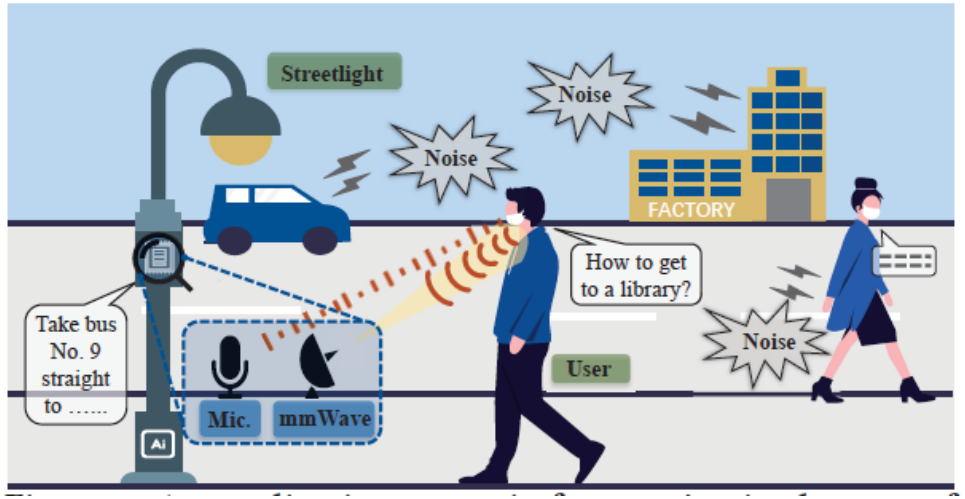
- Introduction
- Background
- System Design and details
- Evaluation
- Conclusion

Introduction

- Voice User Interface (VUI) provides a hands-free and eyes-free human-machine interaction between humans and Internet of Things devices
- It plays an essential role in smart speakers, voice assistants of smartphones, and in-vehicle voice control interactions
- Replace traditional contact interaction such as button or touch pad to VUI due to COVID-19 e.g., voice-controlled elevator and ATMs

Wavoice

- To obtain speech recognition accuracy, it requires clear audio signals with a high signal-to-noise ratio (SNR)
  - Unpredictable noise in public places
  - Degrade acoustic quality due to face masks
- Wavoice is a multi-modal speech recognition system for public VUI apps
  - Design real-time and anti-interference voice activity detection and user targeting methods
  - Exploit a mmWave radar for noisy environments and a microphone in case of motion interference



Background - mmWave

- Distance estimation
  - Frequency modulated continuous wave (FMCW) radar transmits chirp signals periodically in a specific range
  - Intermediate frequency (IF) signal can be estimated by mixing and filtering signals
- Angle estimation
  - Angle of arrival (AoA) can be estimated by employing multiple antennas which are located at differential distance
- Speech sensing
  - Vocal vibration can be captured and used for distinguishing subtle differences of users' vocal vibration

Background – Multi-modal Fusion

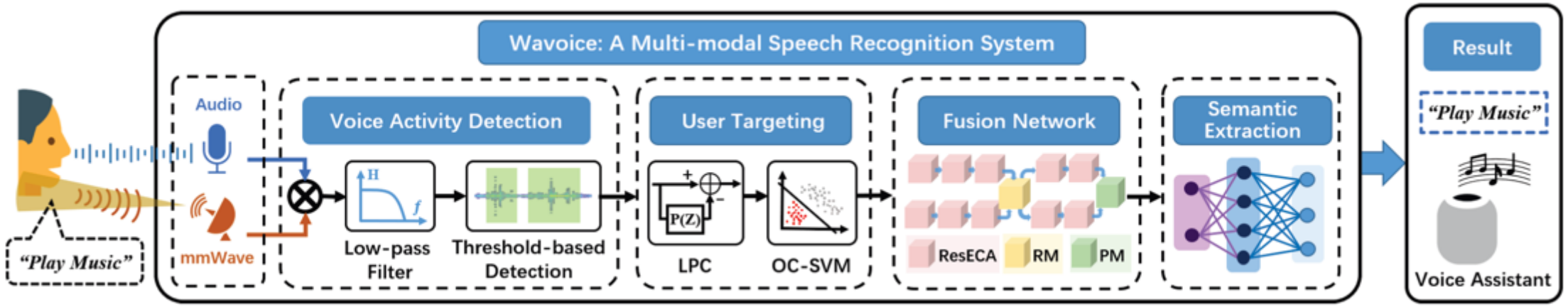
- The key issue of multi-modal fusion is how to maximize advantages of both signals to deal with complex situations
- Voting mechanism is to select the better results from simultaneous signals of different modalities
- Attention Mechanism is incorporating attentions modules into deep neural networks (DNNs) for natural language processing and computer vision tasks e.g., efficient channel attention (ECA)
- The system applies ECA blocks into classical DNN with two additional modules for the fusion of mmWave and audio signals

Correlation Model

- Human voice by microphone $v(t) = H(\dot{x}(t)) = H(j\phi_F x(t))$.
  - Its baseband frequencies are equivalent or close to the speed of vocal fold vibration
- mmWave-based vocal vibration sensing
  - The phase difference of reflected mmWave signals share the identical frequency with the vocal fold displacement
- The coherence between frequencies of different modal signals
  - Human voice and the phase difference originate from the vocal fold displacement
  - Human voice owns components whose frequency overlaps or approaches the frequency of phase difference

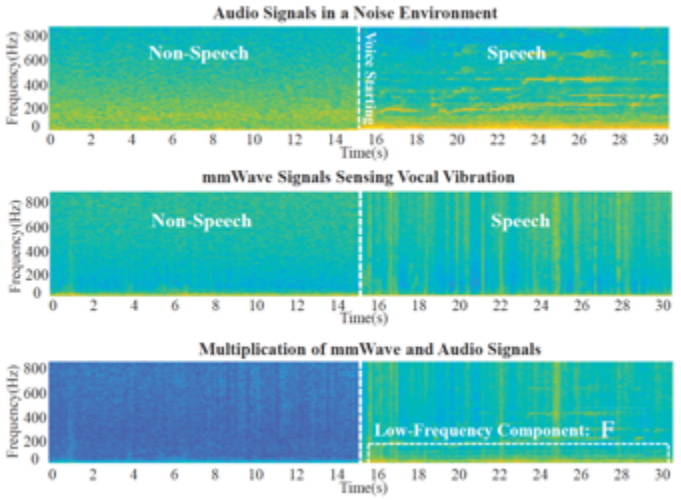
System Design

- Wavoice is the speech recognition system under complex conditions using mmWave and audio signals
- It consists of four modules including Voice Activity Detection, User Targeting, Fusion Network, and Semantic Extraction



Voice Activity Detection

- *What?* The task is to detect voice activities through the detection assessment by the coherent demodulation composed of a multiplier and a filter
- *Why?* Significant resources might be wasted to deal with intense noise which cover human voices with a low SNR in public place
- *How?*
  - Both signals are segmented with a 50% overlap
  - Signals are up or down sampled to 16kHz for preprocessing
  - An energy peak can be acquired at low-frequency band after coherent demodulation
  - The system can detect voice activity by comparing the residual low-frequency component and a given threshold



User Targeting

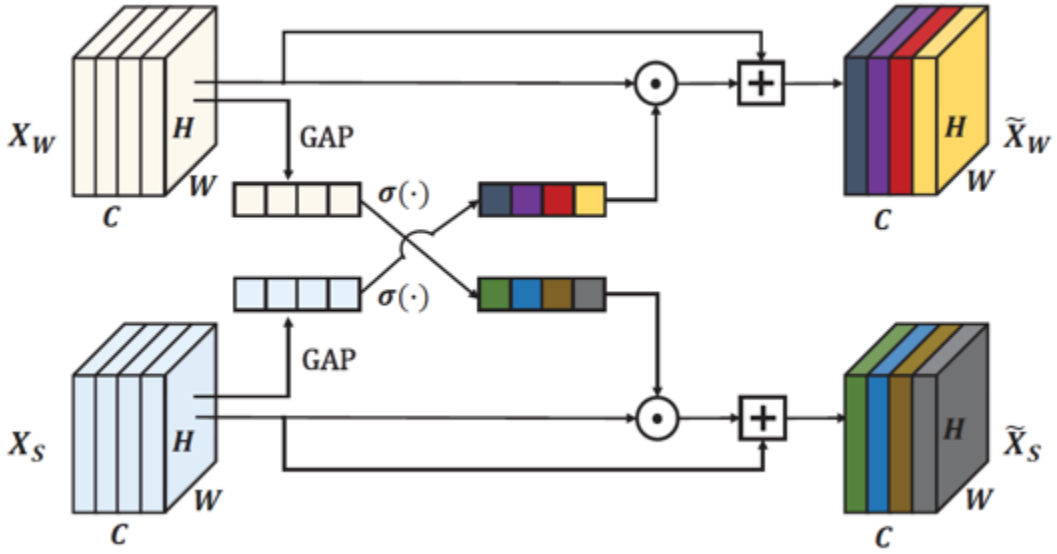
- *What?* The task is to distinguish the target user who speaks the wake-up word for the voice interaction from others
- *Why?* In a multi-person scenario, surrounding speeches should be removed for better accuracy of ASR
- *How?*
  - The radar and microphones collects multiple reflected signals and speech from people around
  - It extracts the difference of phase and repeats the coherent demodulation to ignore non-vocal items
  - It leverages a one-class support vector machine (OC-SVM) with the linear predictive coding (LPC) as input to distinguish wake-up words

Fusion Network

- *What?* The task is to refine characteristics and fuse features from different modalities
- *Why?* It needs to learn a joint representation from multiple domains
- *How?*
  - It extracts coefficients with 40 filters covering the frequency band within 8 kHz
  - The extracted log-mel fileterbank coefficients work as network inputs followed by three successive stacked residual blocks with ECA (ResECAs)
  - The Recalibration Module (RM) exchanges valid features for mutual recalibration and characteristic enhancement
  - Projection Module (PM) projects respective information into a joint feature space and adjusts weight coefficients dynamically

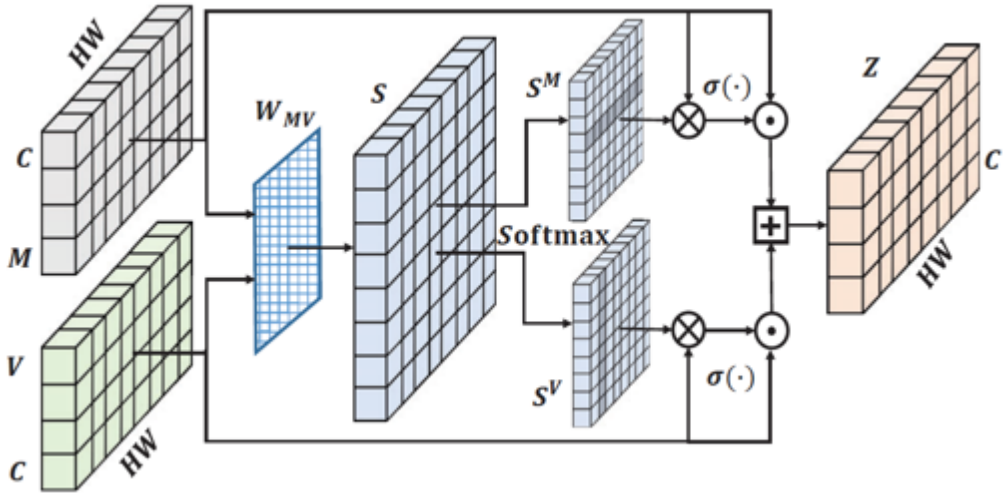
Recalibration Module

- *What?* RM is embedded into the fusion network to integrate multi-modal features from different subnetworks for multi-modal recalibration
- *Why?* It needs to establish the interaction and collaboration of features of mmWave and audio signals
- *How?*
  - It is inserted behind the third ResECA
  - RM can be flexibly placed at different levels in networks



Projection Module

- *What?* PM maps features of two modalities into a joint feature space, so it finally fuses multi-modal signals for speech recognition
- *Why?* DNN cannot fuse the multi-modal signals and transform them into semantic information directly
- *How?*
  - It constructs the similarity matrix of features of two modalities
  - It maps each modality into another modality space
  - It induces high attention weights



Semantic Extraction

- *What?* The task is to translate the meaning of human voice based on the analysis of mmWave and audio signals
- *Why?* It needs to extract semantic information from the joint features using speech-to-text translation for ASR
- *How?*
  - It applies Listen, Attend, and Spell (LAS) which consists of an encoder called listener and a decoder called speller
  - The listener maps the acoustic feature in the hidden feature through the pyramidal bidirectional long short term memory (pBLSTM)
  - The speller computes the probability of output character sequences with a multi-head attention mechanism to generate context vectors

Evaluation

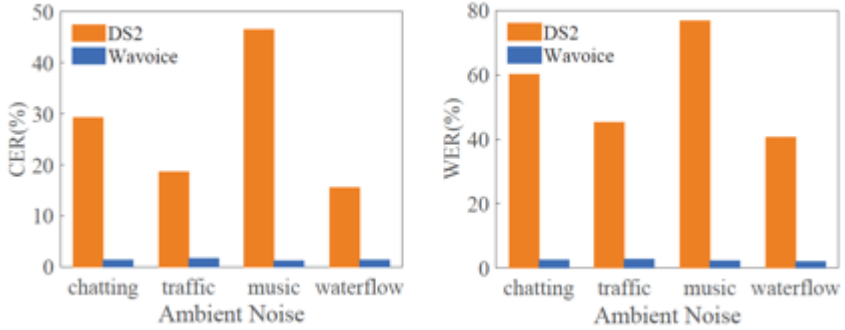
- mmWave radar is configured based on these parameter setups.

Parameter	Value	Parameter	Value
No. of frames	320	Frame periodicity	50 ms
No. of chirp	190	Frequency slope	15 MHz/ μs
Idle time	10 μs	Ramp end time	250 μs

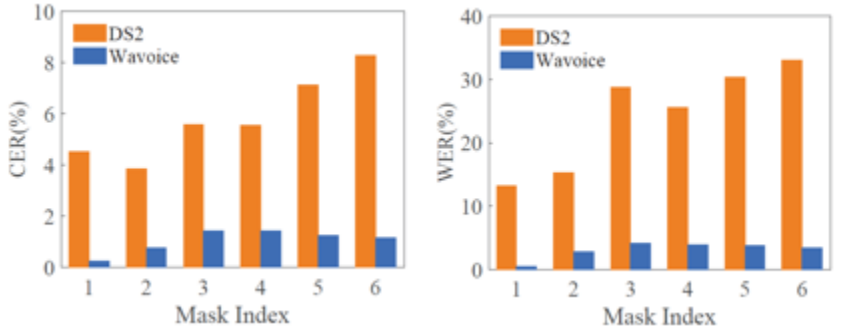
- In the experiment, 25600 training data and 6400 testing data is collected with 40 voice commands and 20 participants
- Character Error Rate (CER) and Word Error Rate (WER) are used as metrics and DeepSpeech2 (DS2) is selected as a baseline

Performance

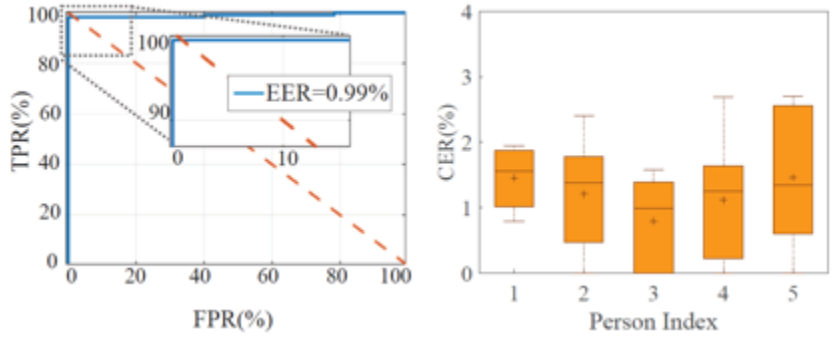
- Ambient noise



- Mask



- Multi-person scenario



- Performance Comparison

Method	Noise		Mask	
	CER(%)	WER(%)	CER(%)	WER(%)
Speech-only	45.18	73.24	8.12	29.66
mmWave-only	10.25	40.76	9.46	33.40
Voting [48]	10.78	48.20	5.37	20.21
W/O Fusion	12.71	35.38	6.43	29.20
DS2 [5]	41.12	72.70	7.13	30.32
Wav2Letter [50]	22.17	46.28	4.72	12.23
W/O ResECA	2.43	4.41	1.78	3.35
W/O RM	4.53	8.82	4.21	9.24
W/O PM	4.08	7.65	3.16	5.882
Wavoice	0.69	1.72	0.76	1.65

Robustness Analysis

- The robustness of system shows under the different distance and orientation, body motion, and environmental disturbance

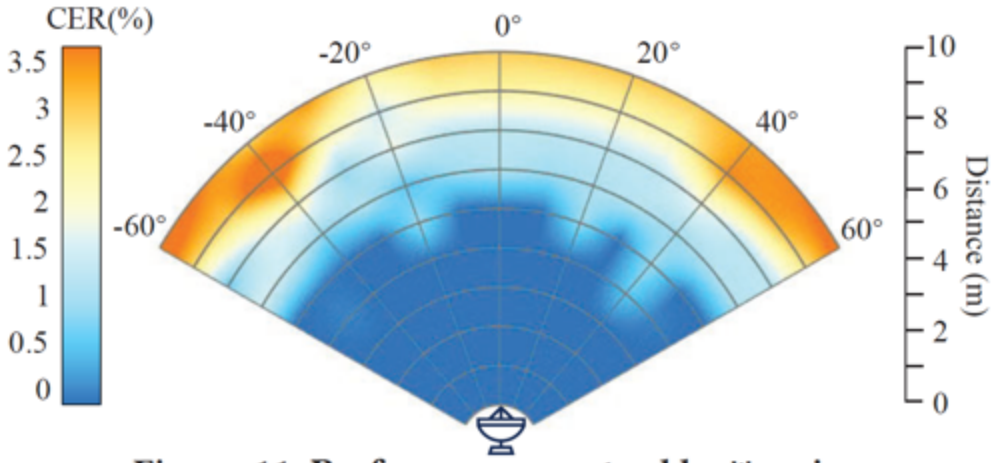
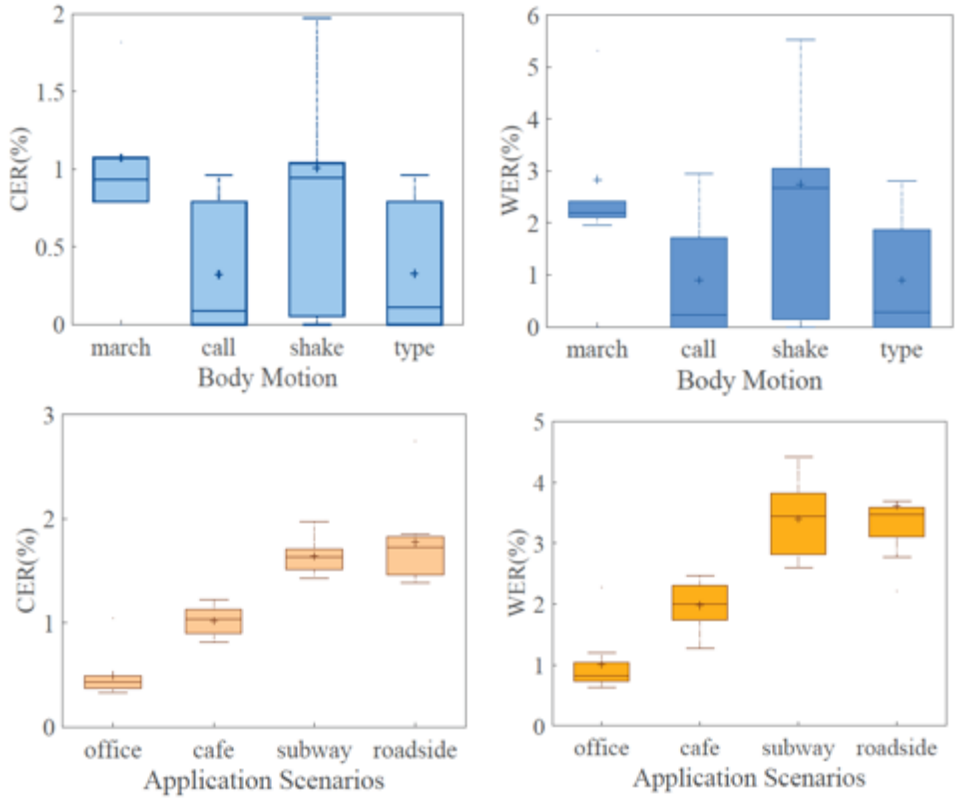
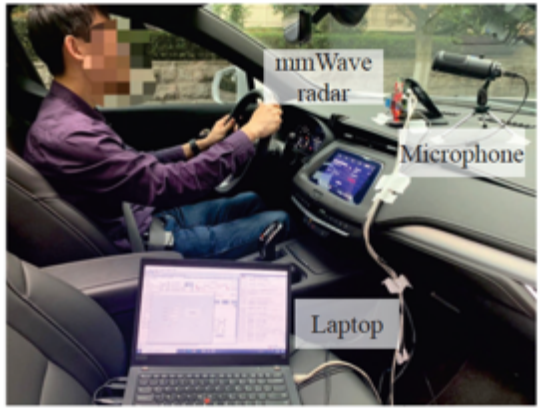


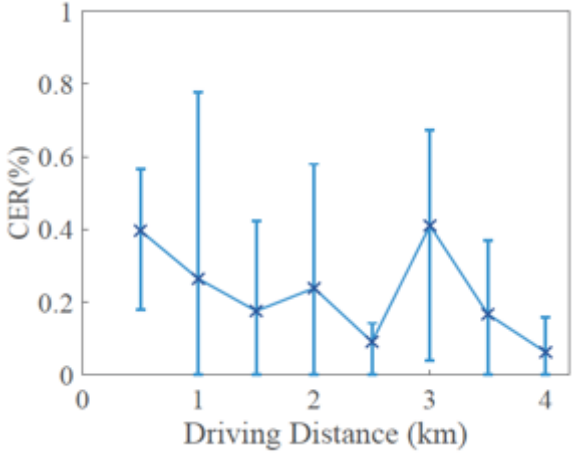
Figure 11: Performance centred by Wavoice.

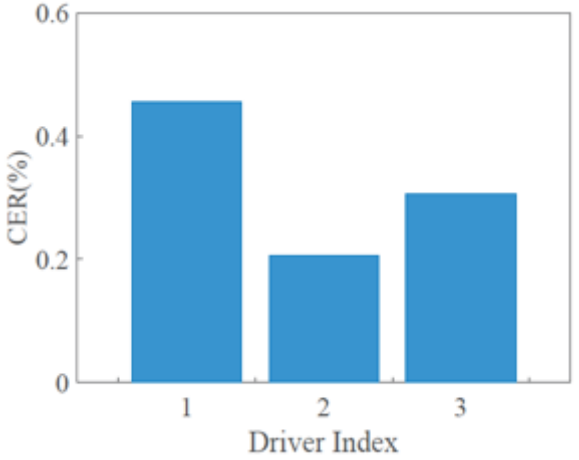


Evaluation in a Vehicle

- User drives 20 minutes at the normal speed in the urban area
- User speaks commands with playing music in the vehicle
- The average CER is 0.32% in one hour driving
- The system overcomes weakness of mmWave such as motion interferences







Conclusion

- Employ a mmWave radar and a microphone for long-distance, noise-resistant, and motion-robust speech recognition
- Formulate the correlation between mmWave and speech signals
- Propose a voice activity detection method against noise interference and a user targeting module to avoid overlaps with non-target users
- Introduce two modules based on the inter-attention between multi-modal signals
- Maintain a low error rate within 1% and its range reaches up to 7 meters

Thank you