Wavoice: An mmWave-Assisted Noise-Resistant Speech Recognition System

TIANTIAN LIU and CHAO WANG, Zhejiang University, China
ZHENGXIONG LI, University of Colorado Denver, United States
MING-CHUN HUANG, Duke Kunshan University, China
WENYAO XU, University at Buffalo, the State University of New York, NY
FENG LIN, Zhejiang University, China

As automatic speech recognition evolves, deployment of the voice user interface (VUI) has boomingly expanded. Especially since the COVID-19 pandemic, the VUI has gained more attention in online communication owing to its non-contact property. However, the VUI struggles to be applied in public scenes due to the degradation of received audio signals caused by various ambient noises. In this article, we propose *Wavoice*, the first noise-resistant multi-modal speech recognition system that fuses two distinct voices sensing modalities (i.e., millimeter-wave signals and audio signals from a microphone) together. One key contribution is to model the inherent correlation between millimeter-wave and audio signals. Based on it, *Wavoice* facilitates the real-time noise-resistant voice activity detection and user targeting from multiple speakers. Additionally, we elaborate on two novel modules for multi-modal fusion embedded into the neural network, leading to accurate speech recognition. Extensive experiments prove the effectiveness of *Wavoice* under adverse conditions—that is, the character recognition error rate below 1% in a range of 7 m. In terms of robustness and accuracy, *Wavoice* considerably outperforms existing audio-only speech recognition methods with lower character error and word error rates.

CCS Concepts: • **Human-centered computing** → *HCI design and evaluation methods;*

Additional Key Words and Phrases: Multi-modal systems, mmWave sensing, speech recognition, biometrics

ACM Reference format:
Tiantian Liu, Chao Wang, Zhengxiong Li, Ming-Chun Huang, Wenyao Xu, and Feng Lin. 2024. Wavoice: An mmWave-Assisted Noise-Resistant Speech Recognition System. *ACM Trans. Sensor Netw.* 20, 4, Article 86 (May 2024), 29 pages.
https://doi.org/10.1145/3597457

Authors' address: T. Liu, C. Wang, and F. Lin (corresponding author), ZJU-Hangzhou Global Scientific and Technological Innovation Center, the School of Cyber Science and Technology, Zhejiang University, China, 310027; emails: tiantian@zju.edu.cn, wangchao5001@zju.edu.cn, flin@zju.edu.cn; Z. Li, Department of Computer Science and Engineering, University of Colorado Denver, Denver, CO; email: zhengxiong.li@ucdenver.edu; M.-C. Huang, Department of Data and Computational Science, Duke Kunshan University, Jiangsu, 215316, China; email: mh596@duke.edu; W. Xu, Department of Computer Science and Engineering, University at Buffalo, the State University of New York, Buffalo, NY, 14261; email: wenyaoxu@buffalo.edu.

1550-4859/2024/05-ART86 $15.00
https://doi.org/10.1145/3597457

1 INTRODUCTION

The **Voice User Interface (VUI)** plays a vital part in modern intelligent applications like smart homes [52]. A VUI serves as a hands-free and eyes-free human-machine interaction between humans and Internet of Things devices. With the aid of deep learning and natural language processes, **Automatic Speech Recognition (ASR)** allows the VUI to comprehend users' intentions accurately [93]. Thanks to such a convenient and flexible voice interaction, users can interact with various Internet of Things devices as they like. Commercial VUI products have earned popularity in recent years, such as smart speakers (e.g., Amazon Echo [4] and Google Home [22]), voice assistants in smartphones (e.g., Siri [32]), and in-vehicle voice control interactions (e.g., VUIs in Tesla Model S/X/3/Y [73]). According to a report by analysts, it is forecasted that the number of VUI-based smart speakers will reach 640 million globally [87] by 2024.

The representative of non-contact interaction (i.e., VUI) has been widely deployed in public scenes [83]. Today, VUIs tend to branch out into the smart city business [23], which gradually substitutes traditional contact interactions such as button or touch interactions [58]. Especially since the COVID-19 pandemic [35], people avoid physical contact with public facilities out of safety concerns. For instance, VUIs have been used for voice-controlled elevators [70] and ATMs [88]. Unlike relatively quiet home scenarios, VUIs ought to handle multifarious ambient noise (e.g., traffic noise, commercial noise, and nearby voices) in public places (e.g., streets, stations, or parties). However, audio-based ASR techniques using microphone arrays, including traditional statistics based [25, 89] and advanced learning based [60, 92], require clear audio signals with high **Signal-to-Noise ratios (SNRs)**. Thus, audio signals in public scenes, drowned in the unpredictable noise, become challenging to recognize. Moreover, people prefer to wear respiratory protective face masks [54] to protect themselves from the coronavirus, which degrades speech quality and further hampers speech recognition accuracy [54]. Those audio-only methods are incapable of supporting VUIs in these cases.

To address the preceding difficulties, researchers exploit multi-sensor information fusion for speech enhancement and recognition. Audio-visual methods [1, 57] integrate lip motion captured by cameras with noisy voices but are limited by lighting conditions, line-of-sight requirement, or face masks. Ultrasound-assisted speech enhancement techniques [39, 72] are merely applied into conditional scenes on account of the extremely short working distance (within 20 cm) and specific postural requirements.

We turn attention to a **Millimeter-Wave (mmWave)** radar and leverage it as a supplementary for speech recognition. Prior research demonstrates that mmWave signals enable voice information recovery with incredible ability on resistance to ambient noise and penetration [43–45]. The mmWave signal is able to capture the vocal vibration by analyzing reflective signals from target users remotely, even wearing face masks in a noisy environment. Nevertheless, the mmWave radar is not satisfactory in all respects. The mmWave signal is susceptible to both vocal vibration and user motion, due to its tiny wavelength (about 4 mm). Its vocal vibration sensing ability would be worsened by users' body movement in practice. Motion interference, ignored by prior work [91], would distort reflected signals that contain vocal information of users. Worse still, mmWave radars are possible to shake in specific scenarios (e.g., in-vehicle applications). The mmWave-based application always suffers from such motion interference from users, radars, or both. Fortunately, voice information recorded from microphones can compensate for the information loss of radars to some extent. Herein, we consider a complementary collaboration between mmWave radars and microphones. These two signals from different modalities are employed together for the sake of one common goal—accurate speech recognition.

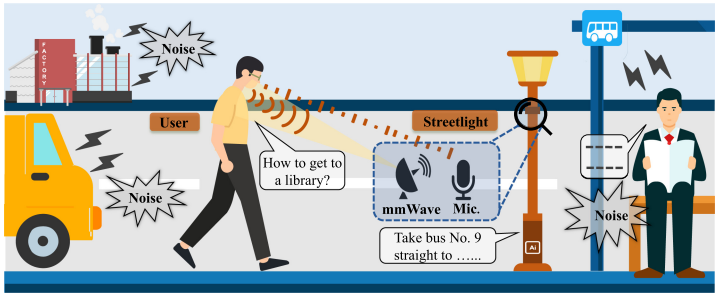


Fig. 1. An application scenario for *Wavoice* in the case of a smart city. Users can interact with a *Wavoice*-powered smart streetlight that provides services including location, navigation, emergency calling, and voice-controlled traffic lights.

To realize a multi-modal system that combines mmWave and audio signals for speech recognition in complex scenes, multiple practical challenges need to be addressed: (1) how to fuse different modality signals to support long-distance VUI applications while mmWave and audio signals may suffer from interference; (2) how to detect voice activity in an effective and real-time manner, when the user's voice is likely to be overlapped by multiple noises; and (3) how to apply this ASR system in a multi-person scene, where irrelevant conversations may intrude into users' voice commands.

We propose *Wavoice*, a multi-modal speech recognition system for public VUI applications, as illustrated in Figure 1. This system exploits mmWave radar to detect the vocal vibration of users in a noisy environment, and a microphone in case of motion interference. Furthermore, it can penetrate through face masks for semantic information extraction with the advantage of mmWave radars. We thoroughly explore the inherent correlation between mmWave and audio signals to combine their advantages. For practical applications, we conceive real-time and anti-interference voice activity detection and user targeting methods based on the frequency-dependent property between multi-modal signals. We introduce two novel modules into the neural attention mechanism for the ASR-oriented multi-modal fusion. One module is designed to exchange valid characteristics for mutual recalibration and feature enhancement, whereas the other projects separate information into a joint feature space and adjusts weight coefficients dynamically. Therefore, we integrate multi-model signals for the semantic feature enhancement to predict the utterance information accurately. Compared with audio-only or mmWave-only ASR, *Wavoice* affords long-distance, noise-resistant, and motion-robust speech recognition in public applications. We demonstrate its effectiveness in various adverse scenarios with high recognition accuracy. In particular, it can be adopted into in-vehicle applications against interferences of different practical motions.

In conclusion, our contributions are as follows:

- We design a multi-modal ASR system named *Wavoice* for a VUI's public application. It fuses mmWave and audio signals to facilitate accurate speech recognition in case of noise and motion interference under complex conditions.
- We investigate the inherent correlation between mmWave and audio signals with a mathematical model. Accordingly, we propose real-time and anti-interference methods for voice activity detection and user targeting, respectively.
- We refine the attention-based multi-modal fusion network with cross-modal recalibration. It supports the robustness of *Wavoice* and improves its sensing distance. Results show a character recognition error rate below 1% in a range of 7 m even under unfavorable conditions.

2 BACKGROUND

In this section, we briefly introduce the mechanism of mmWave sensing, especially in the field of vocal vibration sensing, and the attention mechanism for information fusion.

2.1 The mmWave Sensing Mechanism

In this study, we choose a COTS **Frequency-Modulated Continuous Wave (FMCW)** radar as a vocal vibration sensor in the proposed multi-modal system. A FMCW radar includes transmit (TX) and receive (RX) radio-frequency antennas that transmit radio-frequency signals with a wavelength of millimeter range. The chosen FMCW radar operates at 76- to 81-Ghz bandwidth, which has the ability to sense the physical movement in tiny displacement [11, 44].

Distance Estimation. The TX antenna in the FMCW radar transmits a frequency-modulated signal continuously, also called *chirp signals*. After signals are reflected by targeted objects, the RX antenna will receive the chirp signals, which are a delayed version of the original signal. Immediately, the mixer will multiply the received chirp and transmitted chirp to obtain the mixed signals. The mixed signals still include replica time-delayed versions of the transmitted signals. Herein, a low-pass filter processes on the mixed products to acquire the **Intermediate Frequency (IF)** signal. The spectrum of IF signals is capable of revealing multiple tone frequency that is proportional to the range of each object from the radar. Thus, we can estimate the distance D between the radar and the detected object with the following calculation:

$$D = \frac{c f_{IF} T_c}{2B}, \tag{1}$$

where the c denotes the speed of light, f_{IF} is the frequency of IF signals, T_c is the duration of a chirp, and B is the bandwidth of a chirp.

Angle Estimation. However, FMCW radar can efficiently measure the angle of the object with a horizontal plane. The small distance caused by the **Angle of Arrival (AoA)** from the object to each antenna results in a phase difference in the peak in spectrum. To determine the AoA, the multiple received signals from multiply RX antennas is processed by **Fast Fourier Transform (FFT)** on the spatial domain to calculate the phase difference ω. Note that the resolution of estimated angle depends on the number of antennas on the radar. Thus, we can calculate the AoA by the following formula:

$$AoA = \arcsin\left(\frac{\lambda\omega}{2\pi l}\right), \tag{2}$$

where λ denotes the wavelength of chirp signals and l is the distance between the receiving antennas of the radar.

Speech Sensing. Much research has recently emerged to exploit mmWave radar for speech sensing [10, 26, 40–42, 91], since mmWave signals own significant sensitivity to displacements. Prior studies can only detect the token speech against noise interference by using mmWave radars [77–79]. Further research recovers genuine speech signals originating from modulated vocal vibration in the reflective mmWave signal [10, 26]. Moreover, mmWave radar can be utilized to seize the vibration feature unique to acoustic organs and pronunciation habits, which can be applied into non-contact authentication [46]. However, the preceding mmWave-based systems [10, 26, 91] just have a narrow sensing distance, not more than 2 m. Additionally, all of these systems are vulnerable to motion interference. The limited sensing range and vulnerability to motion confines mmWave-assisted applications in the real world, especially public speech recognition.

2.2 The Attention Mechanism for Fusion

The insight of a multi-modal system is to maximum strengths of each modality to maintain more significant performance than uni-modal systems, even under adverse conditions such as noise, motion interference, and long-distance sensing. When users remotely call on ASR-based devices in a noisy environment, the long-distance propagation would induce the attenuation of audio SNRs, and significant multipath noise brings in additional noise mask on mmWave signals. The traditional fusion mechanism like the voting mechanism [62] cannot support such a practical speech recognition application. The newly risen fusion mechanism called *attention* may provide a possible solution dealing with tough scenarios. It has been widely developed in **Deep Neural Networks (DNNs)** to improve the learning and representation capacity of networks. Many researchers have invented various attention modules such as self-attention [76], channel attention [97], and cross-attention [27], all of which have shown significant success across multiple fields, like natural language processing [76] and computer vision tasks [27]. Inspired by research on the attention mechanism, we integrate it into classical model architectures. Furthermore, we design two attention-based modules to comprehensively fuse multiple modalities that are detailed in Section 4.3.

3 CORRELATION MODEL

In this section, we first exploit the correlation between voice signals and reflected mmWave signals through theoretical analysis. Based on their models, the voice signals coincide with the phase change of mmWave signals, which motivates us to leverage the components and property for redevelopments of voice applications.

Human voice basically depends on the vocal fold vibration. The vocal vibration process can be regarded as a one-degree-of-freedom damping system [13]. We have

$$m\ddot{x}(t) + r\dot{x}(t) + kx(t) = e^{j(2\pi f_F t + \phi_F)}, \tag{3}$$

where m, r, and k are parameters decided by the vocal fold, and $e^{j(2\pi f_F t+\phi_F)}$ is the negative Coulomb force with the frequency f_F and the initial phase ϕ_F. As a result, we obtain the vocal fold vibration velocity $x(t)$ as follows:

$$\begin{aligned} x(t) &= ke^{j(2\pi f_F t+\phi_F+\phi_k)}, \\ \dot{x}(t) &= j2\pi f_F k e^{j(2\pi f_F t+\phi_F+\phi_k)} = j2\pi f_F x(t), \end{aligned} \tag{4}$$

where k is the amplitude gain and ϕ_k is the phase lag.

Audio signals record the human voice through microphones. Typically, they are considered as a compound of series of single-frequency tones [25, 89] looking like

$$v(t) = \sum_i A_i sin(2\pi f_i t + \theta_i), \tag{5}$$

where $v(t)$ is the human voice, and A_i, f_i, and θ_i are respectively amplitude, frequency, and phase of the i-th harmonic. Its base-band frequencies are equivalent or close to the speed of vocal fold vibration [91]. The relationship can be simplified as

$$v(t) = H(\dot{x}(t)) = H(j2\pi f_F x(t)), \tag{6}$$

where $H(\cdot)$ is the transfer function from the vocal fold vibration velocity $\dot{x}(t)$ to human voice $v(t)$. The transformation function $H(\cdot)$ represents the process of generating sound by forces derived from vocal fold vibration. This transformation is not lossless, but the corresponding frequency of sound is equal to the force. Considering the impact of propagation delay on human voice $v(t)$,

Equation (6) can be formulated as follows:

$$v(t) = H\left(j2\pi f_F x\left(t + \frac{D}{s_v}\right)\right), \tag{7}$$

where D is the distance between the microphone and the target user, and s_v is the sound velocity. Since $D \ll c$, the item $\frac{D}{s_v}$ can be ignored.

mmWave-based vocal vibration sensing compares the phase difference of reflected signals for vibration measures. The reflected mmWave signals $r(t)$ from the vocal folds is represented as follows:

$$r(t) = e^{j(2\pi f_{IF} t + \phi(t))}, \tag{8}$$

where f_{IF} is the IF signal and $\phi(t)$ is the phase of the reflected signal. The displacement of vocal folds is contained in $\phi(t)$ as follows:

$$\phi(t) = \frac{4\pi f_m(t)(D + x(t))}{c}, \tag{9}$$

where $f_m(t)$ is the time-variant frequency of the mmWave signal, D is the distance between the mmWave radar and the target user, and c is the mmWave's speed. Concerning the time delay between transmission and reception, Equation (9) can also be rewritten as follows:

$$\phi(t) = \frac{4\pi f_m(t+\tau)(D + x(t+\tau))}{c}, \tag{10}$$

where $\tau = \frac{2D}{c}$ is the time delay. Whereby τ is close to zero, the τ can be ignored. Since the motion of target objects or radars, if any, is usually lower than sampling, D can be deemed a constant in a tiny time interval dt. By differentiating $\phi(t)$, we have

$$\begin{aligned} \Delta\phi(t) &= \phi(t + dt) - \phi(t) \\ &= \frac{4\pi}{c}\left(x(t)df_m(t) + f_m(t)dx(t)\right) + \frac{4\pi df_m(t)D}{c}, \end{aligned} \tag{11}$$

where $df_m(t)$ is the frequency shift of mmWave signals and $dx(t)$ is the displacement change in vocal fold. Since $4\pi df_m(t)D \ll c$, the item $\frac{4\pi df_m(t)D}{c}$ can be ignored. Here, dt and $df_m(t)$ are constant, determined by the mmWave radar's sampling rate and frequency variation rate. Therefore, $\Delta\phi(t)$ depends exclusively on $x(t)$, and we have

$$\Delta\phi(t) = \frac{4\pi}{c}(df_m(t) + j2\pi f_F f_m(t)dt)x(t). \tag{12}$$

This indicates that the phase difference of reflected mmWave signals shares the identical frequency with the vocal fold displacement. In the real measurement, all complex items are performed on their real parts, and in Equation (12), the item $x(t)$ is replaced by $Re\{x(t)\} = cos(2\pi f_F t + \phi_F + \phi_k)$. Note that the phase change only depends on the vocal vibration, since the static information like face is canceled after differentiating.

The *coherence between frequencies of different modal signals* reveals the feasibility of their fusion. Specifically, both $v(t)$ and $\Delta\phi(t)$ originate from the vocal fold displacement. According to Equations (6) and (12), $v(t)$ owns components whose frequency overlaps or approaches the frequency of $\Delta\phi(t)$. Considering the high sampling rate of both mmWave radar and microphones, the corresponding time difference between mmWave $\Delta\phi(t)$ and speech signals $v(t)$ tends to be zero. The impact of time delay can be ignored when the mmWave signal and microphone sense voice activity. In this article we entitle *Wavoice* noise-resistant voice activity detection on the basis of this frequency-dependent property and train a DNN to fusion multi-modal signals for long-distance speech recognition.

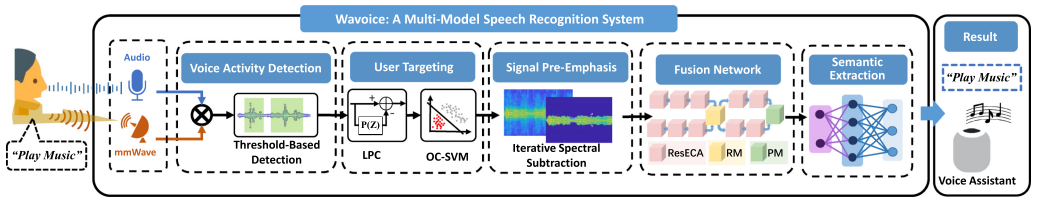


Fig. 2. *Wavoice*, a multi-modal speech recognition system that leverages an mmWave radar and a low-cost microphone to improve the resistance against noise and motion interference in a complex environment.

4 SYSTEM DESIGN

Wavoice leverages mmWave and audio signals to recognize the speech under complex conditions. It consists of five modules, as presented in Figure 2.

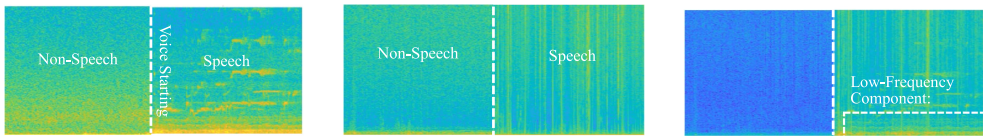
4.1 Voice Activity Detection

On the basis of the preceding frequency-dependent property, *Wavoice* employs the coherent demodulation composed of a multiplier and a filter. It has been proven to provide a noise-resistant method to detect voice activities.

Motivation. Real-time voice activity detection is a fundamental step for ASR. Without a proper detection mechanism, substantial resources would be wasted on dealing with meaningless noise. However, ambient noise tends to cover human voices with an extremely low SNR in public places. Face masks in users further blur vocal features. In such scenarios, relying solely on audio-based voice activity detection can lead to incorrect judgments, rendering the system unresponsive to user commands [37]. Users have to raise their speaking volume or take off their face masks, but this is inconvenient. Fortunately, voice activities are recorded by mmWave and audio signals simultaneously. We can exploit their coherence to intensify the distinction between noise and voice activities.

Solution. *Wavoice* draws the collective characteristic between mmWave and audio signals for accurate judgment in real time through coherent demodulation. *Wavoice* simultaneously receives signals of two modalities. These signals are segmented into 3-second frames with a 50% overlap between successive frames. We perform min-max scaling on the mmWave and audio signal, respectively. For collecting the mmWave signal, we perform range FFT on the received chirp signal to obtain the range information of objects. We leverage the classic detection method named *OS-CFAR* [66] to detect the objects (i.e., the FFT bin of the reflective object). The number of detected objects is decided by the number of people and other objects such as furniture, since the objects cannot stack together due to the radar's 4-cm range resolution. Note that the radar receives the genuine signal corresponding to voice activity and other irrelevant signals. Therefore, we design the voice activity detection to distinguish the genuine signal. Audio signals are down-sampled to 16 kHz to save computational resources, and the down-sampled voice signal $v(n)$ still retains complete human speech information. We obtain the sampling data from the object's FFT bin per chirp signal. Thus, the sampling duration of the preprocessed mmWave signal is chirp duration. Then we up-sample the preprocessed mmWave signal to 16 kHz by using linear interpolation.

We obtain the phase $\phi(n)$ by conducting FFT on the sampled mmWave signal. Then the phase difference is $\Delta\phi(n) = \phi(n) - \phi(n-1)$ $(n \in \mathbb{N}^+)$. Inspired by the frequency-dependent property between $\Delta\phi(n)$ and $v(n)$, we multiply them, followed a low-pass frequency filter for voice activity detection. If $\Delta\phi(n)$ and $v(n)$ share components of the same or similar frequency, we will obtain an energy peak at a low-frequency band after coherent demodulation [17]. We assume $H(\cdot) = 1$ here



(a) Audio signals in a noisy environment (b) mmWave signals sensing vocal vibration (c) Multiplication of mmWave and audio signals

Fig. 3. Although audio signals are noisy, the multiplication introduces an additional low-frequency component that results in a sharp distinction between noise and noisy speech.

to illustrate this method's effectiveness as follows:

$$\begin{aligned}\mathbf{F}(n) &= \mathrm{LPF}(v(n) * \Delta\phi(n)) \\ &= \mathrm{LPF}\left(\frac{4\pi}{c}(df_m(n) + j2\pi f_F f_m(n)dt)Re\{x(n)\}^2\right) \\ &= \frac{4\pi}{c}\left(df_m(n) + j2\pi\frac{f_F f_m(n)}{F_s}\right),\end{aligned} \tag{13}$$

where $\mathbf{F}$ is the residual low-frequency component, LPF$(\cdot)$ is a low-pass frequency filter, F_s is the mmWave radarâĂŹs sampling rate, and the item $\frac{4\pi}{c}(df_m(n) + j2\pi\frac{f_F f_m(n)}{F_s})$ is a known low-frequency value. When the spectral entropy of $\mathbf{F}$ is larger than a given threshold, vocal vibration is recorded simultaneously by $\Delta\phi(n)$ and $v(n)$, and this indicates that voice activities occur. Even if noise ruins audio, mmWave signals, or even worse both, the coherent demodulation still works due to the difference between noises and voice signals in the frequency domain. In a noisy environment, Equation (14) is rewritten as follows:

$$\begin{aligned}\mathbf{F}(n) &= \mathrm{LPF}((v(n) + n_v(n)) * (\Delta\phi(n) + n_\phi(n))) \\ &= \frac{4\pi}{c}\left(df_m(n) + j2\pi\frac{f_F f_m(n)}{F_s}\right),\end{aligned} \tag{14}$$

where $n_v(n)$ and $n_\phi(n)$ are the noise on mmWave and audio signals, respectively. High-frequency items $n_v(n)\Delta\phi(n)$ and $n_v(n)n_\phi(n)$ are introduced by noise but removed by the filter with little influence left. Since the duration of chirp signals is quite short (i.e., 260 μs in the experimental setting), the phase offset in the mmWave chirp duration can be considered constant. The phase offset can be counteracted when differencing the phase. Therefore, the phase offset has little effect on the multiplication results.

Detection Assessment. To investigate the effectiveness of the proposed detection module, we collect corresponding mmWave and audio signals from five subjects. During the collection, we ask each subject in four kinds of noisy environments (detailed setup in Section 5.1) to remain quiet after continuously speaking utterances. After extracting the phase difference of mmWave signals, we generate the low-frequency component $\mathbf{F}$ by multiplying the phase difference with the audio signal. As illustrated in Figure 3, $\mathbf{F}$ ranges in the low-frequency band typically within 200 Hz, whereas the multiplication corresponding to the non-speech segment cannot be seen as anything useful. Vividly, the non-speech and speech segment is explicitly divided after the coherent demodulation. In addition, the varying spectrogram of mmWave signals in Figure 3 supports mmWave signals' ability of the vocal vibration seizing. Empirically, the cut-off frequency of a low-pass filter is set to 300 Hz and the threshold of spectral entropy is set to 0.835. By comparing the spectral entropy of $\mathbf{F}$ with the given experiential threshold, we can detect voice activity with an accuracy of 97.12%. On the contrary, the voice activity detection based on individual audio or mmWave signals only has

56.48% and 88.92%, respectively. Additionally, the whole process is finished within 50 ms. *Wavoice* manages in real-time voice activity detection against various noise interference.

4.2 User Targeting

Speeches from surrounding non-target individuals may overlap users' commands. *Wavoice* proposes a targeting mechanism to derive vocal commands of target objectives against such interference.

Motivation. In a multi-person scenario, surrounding speeches would color the recognition results of ASR. These voice noises are mingled with valid vocal commands, or even cover them up in audio signals recorded by microphones due to the mask effect [17]. The audio-only ASR hardly distinguishes the target user who speaks the wake-up word for the voice interaction from others.

Solution. In *Wavoice*, we propose a user targeting mechanism. It detects the predetermined wake-up word by successively comparing each low-frequency component by multiplying mmWave signals with audio signals after voice activity detection. Notwithstanding mmWave signals sensing wake-up words, it is susceptible to motion interference and other multipath noise. In contrast, *Wavoice* can precisely target the user's command based on the correlation between mmWave and speech signals. Once finding the wake-up word, *Wavoice* separates its reflected mmWave signals and ignores other multipath signals from ambient people. It targets this objective and waits for subsequent commands.

The radar receives multiple reflected signals from nearby people, whereas the microphone records the speech mixed with other persons' voices. Multiple reflected mmWave signals can be formulated as $r_1(n), r_2(n), r_i(n), ..., r_u(n), r_m(n)$, where the subscript m is the number of received mmWave signals decided by the number of person in the sensing ranges after voice activity detection, $r_i(n)$ is the mmWave signal of the i-th person, and $r_u(n)$ is the mmWave signal caused by the wake-up word from a user. We extract the corresponding difference of phase $\Delta\phi_1(n), \Delta\phi_2(n), \Delta\phi_i(n), ..., \Delta\phi_u(n), \Delta\phi_m(n)$ from all reflected signals. We repeat the preceding coherent demodulation between each mmWave signal and audio signals. Non-vocal items are ignored.

Afterward, we leverage a **One-Class Support Vector Machine (OC-SVM)** to distinguish wake-up words from residual voice-related items. However, throwing the unprocessed multiplication production into the OC-SVM is easy to increase the risk of model overfitting substantially. Instead, we extract the **Linear Predictive Coding (LPC)** as input to the OC-SVM as follows:

$$\mathbf{F}_i(n) = -\sum_{k=1}^{p} a_i^k \mathbf{F}_i(n-k) + \varepsilon_v(n), \tag{15}$$

where p is the order of the linear prediction filter, $\varepsilon_v(n)$ is residual prediction error, and the set of a_i^k is the LPC. Benefiting from this property, we train the OC-SVM with LPC features to identify wake-up words and target users. Similar to the preceding analysis on noise cancellation, the motion influence on mmWave signals is suppressed. The LPC feature maintains rich acoustic presentation with low computation cost.

4.3 Signal Pre-Emphasis

After undergoing the preceding two modules, the received mmWave and speech signal are prone to artificial distortion caused by nonlinear signal processing.

Motivation. In the signal flow of *Wavoice*, the input signal is inevitably overlaid by artificial distortion when it is executed by nonlinear signal processing [19, 56]. These common artificial

ALGORITHM 1: Signal Pre-Emphasis

Require:
The sample of speech signals or mmWave signals x.
Ensure:
The output signals Y.
Stage 1: Signal Transformation
Compute spectrum frames from x by running discrete Fourier transform;
Compute the speech-related $s(f,\tau)$ and non-speech segments $n(f,\tau)$ in spectrum frames by running a voice activity detector, where f and τ represent the frequency subband and frame index, respectively;
Stage 2: Iterative Spectral Subtraction
Calculate the energy of non-speech segments σ;
for each f,τ in speech segment **do**
  $E(f,\tau) = \frac{1}{2K}\sum_{i=f-K}^{f+K}|n(i,\tau)|^2, K = 3$;
  if $|s(f,\tau)|^2 > \beta E(f,\tau), \beta = 0.7$ **then**
    $y(f,\tau) = \sqrt{|s(f,\tau)|^2 + \beta E(f,\tau)}e^{jarg(s(f,\tau))}$;
  else
    $y(f,\tau) = \eta s(f,\tau), \eta = 0.8$;
  end if
  if $\sigma < 0.04 * \sum_f \sum_\tau |s(f,\tau)|^2$ **then**
    break;
  else
    recalculate the energy of non-speech segments σ;
  end if
end for
$Y = \mathrm{IDFT}(y)$
return Y;

disturbances degrade the quality of signals, especially for more sophisticated signal processing systems. To address the artificial distortion problem, infinitely many studies on the analysis of artificial distortion generation are proposed to mitigate it. Unfortunately, such conventional methods ignore controlling and reducing artificial noise generation during the ongoing mitigation.

Solution. To reduce the artificial noise and simultaneously minimize new noise generation, *Wavoice* introduces the signal pre-emphasis algorithm to deal with mmWave and speech signals after being processed by two successive modules. The signal pre-emphasis algorithm utilizes spectral subtraction to efficiently enhance the quality of signals and reduce artificial noise with low computation complexity. The proposed algorithm estimates the noise energy and then iteratively performs spectral subtraction [53], leading to minimum noise introduced into denoised signals in each iteration. Concretely, the noise-to-signal ratio will be calculated as the threshold for spectral subtraction in each iteration. If the energy of signal frames is larger than the threshold, it will be suppressed so that the noise-to-signal ratio will become smaller and smaller with iterative calculations. The overall signal pre-emphasis algorithm is shown in Algorithm 1. First, the algorithm transforms the input signal (i.e., mmWave signals or speech signals) into the spectrum frames by using discrete Fourier transform. Afterward, we can separate the speech-related and non-speech segments in spectrum frames through a voice activity detector [69]. The speech-related segment is adaptively subtracted by neighboring noise energy during iterative spectrum subtraction. This

is because iterative spectral subtraction can restrain the noise generation and gradually eliminate artificial noise.

4.4 Fusion Network

The fusion network consists of **Residual Blocks with ECA (ResECAs)**, the **Recalibration Module (RM)**, and the **Projection Module (PM)** for multi-modal signals fusion, as shown in Figure 2. The fusion network refines characteristics and fuses features from different modalities to learn a joint representation from multiple domains. The extracted log-mel filter bank coefficients as network inputs follow into three successive stacked ResECAs. After that, the RM exchanges valid inherent features for mutual recalibration and characteristic enhancement, with recalibrated features flowing into two successive stacked ResECAs. Last, the PM projects respective information into a joint feature space and adjusts weight coefficients dynamically.

4.4.1 Log-Mel Filter Bank Coefficients. We extract log-mel filter bank coefficients as network inputs from audio signals and residual voice-related mmWave signals, respectively. In detail, we first apply a high-pass filter to the preprocessed audio signal. After filtering, we perform the short-time Fourier transform to measure the time and frequency domain information. Short-time Fourier transform segments the audio signal into frames of 25 ms, with an overlap of 10 ms between successive frames. During segmentation, we apply a Hamming window function to frames to reduce spectral leakage. Then, the Fourier-transformed audio signal passes through a set of band-pass triangular filters known as mel filter banks. Consequently, we calculate the logarithmically compressed filter-output energy as the log-mel filter bank coefficient. The number of coefficients is equivalent to the number of filters. In this work, the filter bank comprises 40 filters covering the frequency band within 8 kHz.

4.4.2 ResECA. We construct two branches of ResECAs [65] to integrate the features of two modalities. An ECA block is an attention-based block that is made up of convolution layers, aiming to model interdependencies among channels of convolutional features. The ECA applies **Global Average Pooling (GAP)** [28] to learn contextual information in all receptive fields of networks instead of the limited local field like traditional convolutional layers. Based on information in all channels, the ECA generates the channel attention to enable the network to focus on the more important region. Suppose the output of one convolution layer is $X = [x_1, x_2, \cdots, x_c]$, $X \in R^{H \times W \times C}$, where H, W, and C are width, height, and channel dimensions, and x_c refers to the produced channel feature of the c-th filter in the convolution layer. Then, GAP is applied to model channel-wise features $Z = [z_1, z_2, \cdots, z_c]$, $Z \in R^{1 \times 1 \times C}$, where the c-th element of Z is obtained as follows:

$$z_c = \mathrm{GAP}(x_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j). \tag{16}$$

The channel-wise feature Z contains statistical information of all channels. Then we calculate the attention feature:

$$A = \sigma(\mathrm{C1D}_k(Z)), \tag{17}$$

where $A = [a_1, a_2, \ldots, a_c]$, $A \in R^{1 \times 1 \times C}$, σ is a sigmoid activation function, and $\mathrm{C1D}_k$ represents 1D convolution with kernel size k. The final output of the ECA block $\widetilde{X}$ is obtained by channel-wise multiplication between X and A:

$$\widetilde{X} = A \odot X, \tag{18}$$

where $\odot$ indicates scalar multiplication. The attention feature A contains dynamic channel information that is continually optimized in the iteration. We concatenate a typical residual block and

an ECA block to construct a ResECA as a basic module in the network. It can be formulated as follows:

$$Y = C\left(\mathrm{ECA}(C(X, W_C)), W_C\right) + X, \tag{19}$$

where the function $C(*, W_C)$ represents multiple convolution layers to capture features, Y denotes the output of the ResECA, and $\mathrm{ECA}(\cdot)$ represents the ECA block. The operation $C + X$ represents a shortcut connection. The output from multiple successive convolution layers flows into the ECA block. After computing results through the attention procedure in ECA, a shortcut connection adds the residual block's input and the result of the ECA block to attain the final output of the ResECA.

4.4.3 Recalibration Module. A devised **Recalibration Module (RM)** is embedded into the fusion network to integrate multi-modal features from different subnetworks for multi-modal recalibration. In the following, we first describe the aim of the RM and then introduce the mechanism of the RM.

Motivation. Multi-modal recalibration is the process of combining and complementing relevant information among different modalities, leading to the performance of multi-modal fusion over using only one modality. In traditional networks, features of different modalities are processed in a separate branch composed of several ResECAs. However, stacked ResECAs only provide uni-modal features rather than multi-modal features. Yet, such a parallel-branch structure ignores the inherent correlation between mmWave and audio signals. We need to establish the interaction and collaboration of features of two modalities: mmWave and speech. More specifically, if the speech feature suffers interference and attenuation, the mmWave feature is required to guide the network framework to capture underlying representation and supply the knowledge of vocal vibration to the speech. Considering the impact of multipath noise and body motion on mmWave signals, the speech feature is obliged to recalibrate mmWave features.

Solution. We design a novel attention-based module, the RM, as an intermediate module to integrate features of two modalities. Its structure is illustrated in Figure 4. It is inserted behind the third ResECA so that features of two modalities from each branch flow into the RM for mutual recalibration. We assume that $X_W \in R^{H\times W\times C}$ and X_S are two intermediate feature maps from their own stream. The subscripts W and S individually represent the mmWave and speech features. The channel attention maps Y_W and Y_S are

$$Y_W = \sigma(W_W \mathrm{ReLU}(\mathrm{GAP}(X_W))), \quad Y_W \in R^{1\times 1\times C}, \tag{20}$$

$$Y_S = \sigma(W_S \mathrm{ReLU}(\mathrm{GAP}(X_S))), \quad Y_S \in R^{1\times 1\times C}, \tag{21}$$

where ReLU is a **Rectified Linear Unit (ReLU)** function and W indicates a learnable parameter matrix. Each stream of the channel feature maps is considered as a feature detector and filter. We implement mutual feature recalibration as follows:

$$\widetilde{X}_W = Y_S \odot X_W + X_W, \quad \widetilde{X}_W \in R^{H\times W\times C}, \tag{22}$$

$$\widetilde{X}_S = Y_W \odot X_S + X_S, \quad \widetilde{X}_S \in R^{H\times W\times C}, \tag{23}$$

where $\widetilde{X}_W$ and $\widetilde{X}_S$ are final outputs of the RM. Therefore, we obtain the multi-modal features. Aggregating the original feature map guarantees that the final output stores enough identical knowledge. The produced multi-modal features embedded in original uni-modal features will supply meaningful contexts and suppress useless ones to achieve recalibration. The RM can be flexibly placed at different levels in networks to integrate hierarchical features with different spatial

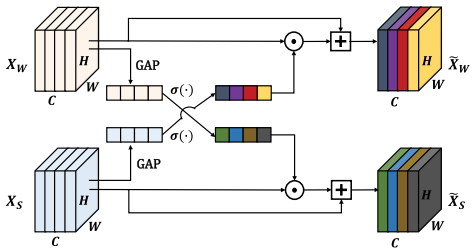


Fig. 4. Architecture of the RM. The RM recalibrates features by combining original features with those from the other modality.

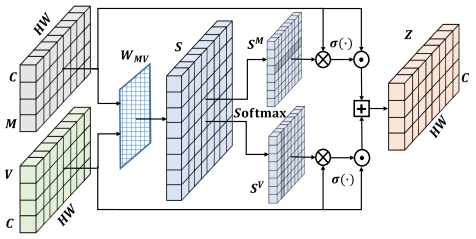


Fig. 5. Architecture of the PM. The PM constructs the similarity matrix based on features of two flattened modalities.

dimensions. Here, we place one RM in the middle to fuse mid-level features. It empirically produces comprehensive high-level features through joint recalibration [65].

4.4.4 Projection Module. The PM maps features of two modalities into a joint feature space. It adaptively selects and strengthens useful information from two isolated feature spaces and weakens irrelevant interference simultaneously.

Motivation. Due to the difference of multi-modal signals, the DNN cannot fuse these signals and transform them into semantic information directly. Traditional methods [60, 92] concatenate multiple modalities from different streams directly. They ignore the dynamic distribution of the weight across multi-modal features. Instead, the joint feature [51] using typical methods focuses on all multi-modal features equally, which requires a large amount of training data to enable the network to fully leverage the benefits of multi-modal features.

Solution. Inspired by co-attention [51], we create another novel attention-based module to project multi-modal features into a joint feature space. This module, the PM, aims to adaptively emphasize more important features and suppress less important ones in all elements of multi-modal features. Its structure is illustrated in Figure 5. The PM constructs the similarity matrix of features of two modalities to measure the correlation between each element of speech and each element of the mmWave. With the similarity matrix, we can respectively map each modality into another modality space. It induces high attention weights for the more distinct element in both modal spaces.

Given two feature maps $M \in R^{H\times W\times C}$ and $V \in R^{H\times W\times C}$ from their own stream, let M denote the mmWave feature map from the corresponding branch, and let V denote the speech feature map. We first have to flatten M and V into 2D tensors with height C and width $W \times H$. We estimate the correlations between $M \in R^{C\times HW}$ and $V \in R^{C\times HW}$ by calculating the similarity matrix S. The similarity matrix between M and V is defined as follows:

$$S = M^T W_{\text{mv}} V, \quad S \in R^{HW\times HW}, \tag{24}$$

where W_{mv} is a learnable weight matrix. Each column m^i in the flattened matrix M represents a feature vector of the C dimension at position $i \in [1, 2, \ldots, HW]$. Each entry of S reveals the correlations between the corresponding column of M and V. We perform a row-wise normalization to produce S^V with a softmax function and a column-wise normalization to produce S^M with a softmax function:

$$S^M = \text{softmax}(S), \quad S^M \in R^{HW\times HW}, \tag{25}$$

$$S^V = \text{softmax}(S^{\text{T}}), \quad S^V \in R^{HW\times HW}. \tag{26}$$

The similarity matrix S^M transfers mmWave feature space into speech feature space (vice versa for S^V). And we have

$$C^M = V \otimes S^M, \quad C^M \in R^{C \times HW}, \tag{27}$$

where $\otimes$ denotes matrix multiplication. Similarly, for the input V, we compute attention contexts of the speech feature based on every element of the mmWave, which is $C^V = M \otimes S^V$. To alleviate the underlying irrelevant interferences, we had better restrict and weigh the knowledge from features of two modalities rather than cope with all knowledge equally. Therefore, the final fusion result Z is formulated as follows:

$$Z = W_Z\{\sigma(C^M) \cdot M + \sigma(C^V) \cdot V\}, \quad Z \in R^{C \times HW}, \tag{28}$$

where $\cdot$ denotes the Hadamard product and W_Z is a learnable parameter matrix. The Z that represents features of two modalities selectively integrates informative information. The fine-grained element in Z associated with vocal vibration and acoustic characteristics occupies a dominant position. Eventually, the fusion result is fed into the *semantic extraction* to identify the speech contents.

4.5 Semantic Extraction

We utilize the typical speech-to-text translation system [80, 93] to construct the semantic extraction architecture. We choose **Listen, Attend, and Spell (LAS)** [8], a widely used end-to-end deep learning approach because of its excellent performance on small-scale training data. It does not rely on any assumptions about the probability distribution of character sequences [63].

LAS is composed of two components: an encoder called *listener* and a decoder called *speller* [8]. The listener is used to map the acoustic feature into the hidden feature through the **Pyramidal Bidirectional Long Short-Term Memory (pBLSTM)**. Each successive pBLSTM layer reduces the feature in half before feeding it to the next layer. The speller, a stacked recurrent neural network, computes the probability of output character sequences. It applies a multi-head attention mechanism to generate the context vector. Context vectors, distribution of characters, and decode states are all fed into the recurrent neural networks for the decoder state. The posterior distribution is computed based on the decoder state and context vector via a softmax function [63]. LAS is trained to maximize the logarithmic posterior probability of the correct character sequence.

Here, we stack two pBLSTM layers as the listener, whereas the speller contains two LSTM layers and an output softmax layer. With the aid of LAS, *Wavoice* extracts the semantic information from the joint features.

5 EVALUATION

We implement the prototype of *Wavoice* using off-the-shelf devices. We conduct a comprehensive evaluation on the recognition accuracy and robustness of our system.

5.1 Setup

Hardware. The proposed system is implemented on a low-cost microphone [18], a COTS IWR1642BOOST radar [33] equipped with a data collection board DCA1000EVM [74], and a laptop, as shown in Figure 6. The IWR1642BOOST equipped with DCA1000EVM is a 77-GHz mmWave radar that transmits FMCW continuously to measure range as well as angle. The mmWave radar has two transmit antennas and four receive antennas. Our commercial radar has a wide enough sensing range: it has an azimuth field of view of 120 degrees, an azimuth resolution of 15 degrees, and a high-resolution elevation view of 30 degrees. The radar transmits a 4-GHz-wide chirp signal starting from 77 GHz to 81 GHz, which yields high-ranging resolution. We configure the radar

Fig. 6. Experimental setup. An mmWave radar and a microphone receive signals from subjects sitting 7 m away.

Table 1. Configuration of the mmWave Radar

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 our experiments to transmit a chirp with 260 μs cycle time. The received channel has a 5 Msps ADC sampling rate, and each received chirp contains 1024 sample data. The detailed configuration of our FMCW radar is shown in Table 1. The configuration enables our radar to have the range resolution of 3.75 cm and displacement resolution around 300 μm.

Software. We connect and control the radar with mmWaveStudio GUI [34] running in the laptop. The mmWaveStudio GUI configures the radar parameters as described previously. We write an APP in MATLAB to control the microphone and mmWaveStudio GUI to capture the mmWave and audio signal simultaneously. The source codes are released at https://github.com/TitaniumLiu/Wavoice.

Dataset. In our experiments, we choose 40 voice commands from ok-google.io [20] and Google speech commands [85] that involve common voice commands words in all aspects. All 20 participants, including 10 males and 10 females, whose ages range from 16 to 47 years, speak all commands in their normal speech speed and volume, typically 65 dB-SPL (sound pressure level) [67]. We place the mmWave radar and microphone at a distance of 7 m from the subject. We align the mmWave radar to the subject and guarantee that the mouth and neck of subjects are within the sensing range of the mmWave radar since our commercial radar has a wide enough sensing range. The participants are asked to say all voice commands 40 times in a controlled laboratory environment. In all, we collect 32,000 pairs of samples (i.e., the mmWave and audio signal) for each situation. We randomly choose the sample from 2 males and 2 females as the test dataset. We thereby have 25,600 training data and 6,400 testing data. During the experiment, participants are required to wear various masks, undergo diverse noise, sit at different angles and distances from the mmWave radar, and perform several body motions. The experimental environments include office room, roadside, cafe, and in-vehicle scenes. Note that we explicitly tell the participants about the purpose of our experiments. Our research is approved by an institutional review board (ZJU2021-6).

5.2 Metrics and Baseline

We measure *Wavoice*'s speech recognition accuracy from the perspectives of both character and word with the two following metrics. We select **DeepSpeech2 (DS2)** [5] as our baseline system for performance comparison.

Character Error Rate. The ASR system outputs a word sequence made of characters, similar but not equal to reference transcriptions. Several characters need to be substituted, deleted, and inserted. CER is computed with the minimum number of operations [94] as follows:

$$\text{CER} = \frac{I_c + S_c + D_c}{N_c}, \tag{29}$$

where N_c represents the total number of characters and the minimum number of character insertions I_c, substitutions S_c, and deletions D_c required to transform the output into the reference transcription. Lower CER indicates better speech performance of the ASR system.

Word Error Rate. **Word Error Rate (WER)** is the standard metric to evaluate the performance of ASR systems. It computes the errors from the word level by comparing output word sequences with reference transcriptions as follows:

$$\text{WER} = \frac{I_w + S_w + D_w}{N_w}, \tag{30}$$

where N_w is the number of total words, and I_w, S_w, and D_w represent the number of insertions, substitutions, and deletions. The number of errors is the sum of substitutions, deletions, and insertions. Lower WER certainly indicates that the ASR of the system is more accurate in recognizing speech.

Baseline. We select DS2 [5], a state-of-the-art ASR for deployment into the production setting, as the baseline system to confirm *Wavoice*'s effectiveness. DS2, initially based on Baidu AI research labs, is one of the mainstreams that has changed the structure of traditional ASR. The network configuration and training parameter of DS2 are consistent with the official article [5]. We implement DS2 under three different trial conditions: (1) we directly test the well pre-trained DS2 model on our collected speech datasets, (2) we continually train the pre-trained model on our datasets and then test it, and (3) we train and test a DS2 model totally on our datasets. We observe DS2's CERs respectively are 90.60%, 71.22%, and 34.46%. Therefore, we construct the baseline results by implementing DS2 under the third condition.

5.3 Overall Performance

We evaluate the overall performance of *Wavoice* when users are in different states. Two in-lab experiments are conducted to assess whether our multi-modal system can show excellent speech recognition capacity over the standard ASR system. The factors of ambient noise and multi-person are respectively considered in the two experiments.

5.3.1 Ambient Noise. Environmental noise reduces the quality of users' voice commands when they are in interaction with voice-controlled devices. We conduct evaluations under four types of noise conditions—chatting, traffic, music, and waterflow—which widely happen in real-world situations. We request participants to speak voice commands piece by piece. Meanwhile, four loudspeakers simultaneously play noise recording with 60 dB-SPL at a distance of 40 cm from the microphone in *Wavoice*. The four loudspeakers are placed evenly around the microphone, spaced approximately 90 degrees apart. The speech recognition results of *Wavoice* and DS2 under different noise interference are shown in Figure 7. The speech recognition inaccuracy of DS2 is the average WER above 40% and the average CER above 20%. Furthermore, DS2 performs worse in noisy environments with a human voice than in white noise environments. DS2 presents poor speech recognition capability in noisy scenes, since the semantic information of received speech signals is explicitly degraded by background noise. This is because audio-only systems like DS2 are vulnerable to unstable and random noise interference. On the contrary, *Wavoice* is proved to

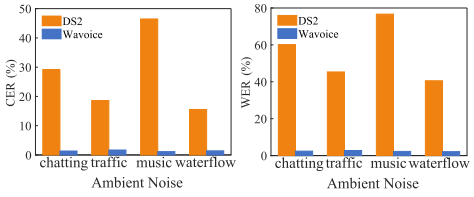


Fig. 7. Performance of *Wavoice* and DS2 under various ambient noises.

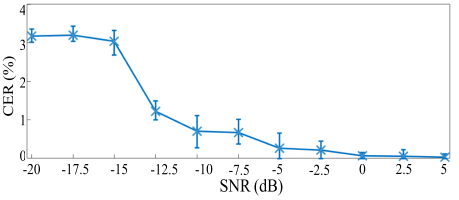


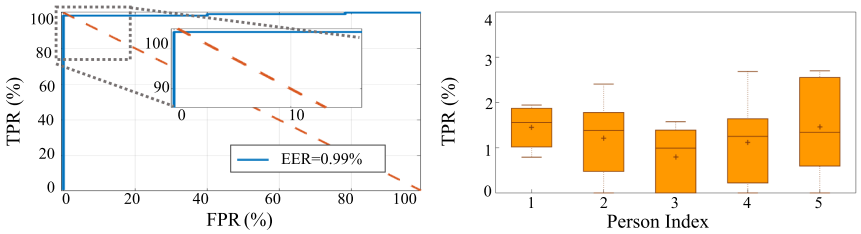
Fig. 8. Performance of *Wavoice* under various SNRs of audio signals.

own superior speech recognition ability with the average CER within 1% and the average WER about 2.5%. Additionally, faced with arbitrarily complex noise conditions, *Wavoice* maintains stable performance with high accuracy. Even in the worst case (i.e., under traffic noise), *Wavoice* still keeps CER below 1.5% and WER at about 3%. We speculate that supplementary knowledge from the mmWave modality makes up for the information loss of the voice modality. Based on the preceding comparisons and observations, it is confirmed that *Wavoice* is a noise-resistant multi-modal system integrating mmWave and speech signals.

To comprehensively investigate the ability of *Wavoice* as received SNRs decrease, we further carry on a new experiment. We modify the volume of four loudspeakers to gradually adjust the SNR of noise sources from −20 dB to 5 dB. Figure 8 presents the variation tendency of *Wavoice*'s performance. With the SNRs decreasing, the CER of *Wavoice* is rising slowly and finally becomes stable. Concretely, the CER increases a little but still maintains at lower than 1% when SNRs are higher than −10 dB. When the SNRs are above 0 dB, it appears that there is a small fluctuation slightly around 0%. It is proven that *Wavoice* can refine and fuse semantic knowledge from mmWave and speech signals to realize noise-resistant multi-modal speech recognition. It is observed that the CER of *Wavoice* tends to become steady as SNRs are under −15 dB. It is reasonable to infer that acoustic information in speech modality disappears when SNRs are extremely low, resulting in the comprehensive convergence of system performance. Under such situations, the speech recognition ability of the multi-modal system completely depends on the unaffected modality (i.e., the mmWave modality). In brief, the designed *Wavoice* owns a stable and noise-resistant speech recognition capability by sufficiently leveraging mmWave and speech modalities.

5.3.2 Multi-Person Scene. It is commonly seen that a voice-controlled device is surrounded by several people when users speak to it. The radar in the system inescapably receives multiple signals from nearby people, even worse when ambient people also make utterances. To prove the effectiveness of the proposed user targeting module, we further carry on the experiment where users are in multi-person scenes. Each of five participants (three female and two male) is asked to take turns as the target, whereas the other four subjects walk around the user who speaks commands and speak freely at a volume of 40 to 65 dB. The SNR of the signal received by the system ranges from −13 dB to −10 dB. Almost no one stands between the user and the system.

Other than uttering voice commands, each user is required to speak the wake-up word 30 times for pre-training the classifier in the user targeting module and 10 times for trial. We set the wake-up word to "Wavoice" in this experiment. We finally collect a total of 200 positive samples (i.e., mmWave and speech signals related to the wake-up word) and 8,000 negative samples related to other utterances. After data collection, we extract LPC features from the collected samples and then leverage the sample to pre-train and examine the user targeting module. According to experimental testing results, we calculate the ROC (receiver operating characteristic curve as presented in Figure 9(a). It is observed that the user targeting module can verify the genuine samples from



(a) ROC curve in identifying users (b) CER of five subjects

Fig. 9. Performance in a multi-person scene.

users with a 98.8% true positive rate and less than a 1.1% false positive rate. The equal error rate of the module is merely 0.99%, which further verifies its usefulness in multi-person scenes. Additionally, we investigate the effectiveness of speech recognition when users are encircled by other people. Figure 9(b) shows the CER of recognized utterances among five subjects. By averaging the CER of five subjects, we can obtain an overall speech recognition error rate of 1.2%. There is no significant difference between the five individuals, which further demonstrates that *Wavoice* is highly robust against interference from ambient people.

5.4 Performance Comparison

In this section, we carry out the ablation study to quantify the fusion of two modalities signals and our proposed fusion methods. In comparison, we comprehensively validate our approach by ablating specific components:

- *Speech-only*, where no mmWave is fused in our proposed network. We clip off the subnetwork of speech in our fusion network.
- *mmWave-only*, where no speech is fused in our proposed network. We clip off the subnetwork of mmWave in our fusion network.
- *Voting*, where the result is generated by voting [62] between two outputs from the preceding two modified networks: Speech-only and mmWave-only. The weight coefficient of recognized texts from the two networks will be updated during the training iteration of the majority voting. The final result is decided by the text that has higher confidence.
- *W/O Fusion*, where no proposed fusion module is performed. The two subnetworks of our fusion network still receive mmWave and audio signals separately. Then, features of two modalities are concatenated and fed into the *semantic extraction*.
- *W/O ResECA*, where no ResECA is performed. We replace ResECAs with classic residual blocks.
- *W/O RM*, where no RM is performed. The two subnetworks receive mmWave and audio signals separately. At last, the PM receives the two individual features.
- *W/O PM*, where no PM is performed. The RM still recalibrates the two features.

Moreover, except for DS2, we compare our model with another state-of-the-art speech recognition network: Wav2Letter [64]. Notably, Wav2Letter, a structured-output learning approach based on a variant of CTC, has an outstanding performance on noisy speech [64]. All of the models are fairly and fully pre-trained on our collected datasets and then validated on the same testing setup. The results of comparison are shown in Table 2.

As can be seen from Table 2, audio-only methods (i.e., Speech-only, DS2, and Wav2Letter) have poor speech recognition with high CERs and WERs, especially in dealing with noisy speech. Thus,

Table 2. Performance Comparison Among Speech Recognition Methods Under Different Conditions

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 [62]	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 [64]	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

we speculate that unpredictable ambient noise impedes the performance of audio-only methods. The mmWave-only method struggles in providing reliable results in motion cases, attributed to its sensibility to varying multipath noise and relatively coarse-grained perception. As for the information fusion method, Voting and W/O Fusion yield slightly better results over the baseline with merely a 10.78% CER and a 12.71% CER in noise, respectively. This comparison verifies that ignoring the correlation and collaboration between mmWave and audio signals cannot comprehensively exploit different modalities to achieve utmost performance in speech recognition. Conversely, the proposed fusion modules significantly improve W/O Fusion by over 12% and 5% in the CER under different cases. Our system with fusion modules is superior to Voting by 10% and 4% in terms of the CER in two different environments, respectively. Additionally, *Wavoice* outperforms WaveEar [91], whose WER is mostly more than 4%, especially under motion interference. Furthermore, we carry on an ablation study to investigate the impact of different proposed modules in the fusion network. It is observed from comparison results that each designed module plays a vital role in speech recognition abilities of the system. To sum up, our system equipped with fusion modules is superior to the methods mentioned earlier. These comparison experiments indicate that our proposed fusion modules adequately exploit the correlation between two modality signals.

5.5 Robustness Analysis

We further analyze the robustness of *Wavoice* under the influence of different distance and orientation, and body motion. Note that the sensing distance of the radar and microphone is still 7 m in the body motion circumstances.

5.5.1 Impact of Distance and Orientation. We investigate the usability of WavoID when users are at different distances and orientations from it. We thereby carry on these experiments where the mmWave radar and microphone are placed at 1- to 10-m distances and −60-degree to 60-degree orientations to users. Figure 10 displays the distribution of speech recognition ability over users' location. It can be seen that *Wavoice* can maintain extremely low CER within an 8-m sensing range from any orientations. When the sensing range is larger than 8 mm, the CER of *Wavoice* slightly increases with the range increasing. This is because the energy of speech signals gradually decays as the distance increases, leading to the low quality of recorded speech by the microphone. Whereby the supplementary information from the mmWave modality makes up for the information loss caused by distance attenuation, the mmWave-voice system still maintains significant speech recognition performance. Additionally, the efficiency of *Wavoice* shows relatively stable

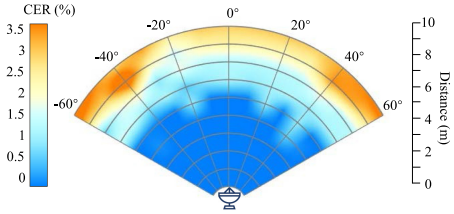


Fig. 10. Performance centered by *Wavoice*.

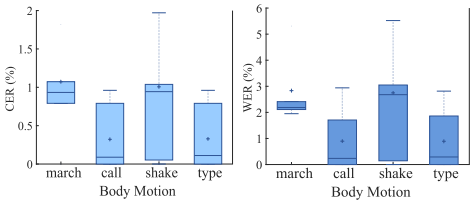


Fig. 11. Performance under body motion influence.

Table 3. Models of Involved Masks

No.	Type	No.	Type
1	Disposable medical mask	4	Scarf + N95 mask
2	Scarf	5	Gas mask
3	N95 respirator mask	6	Anti-dust mask

performance around all orientations. We speculate that the omnidirectional speech signal from the microphone is fused into the system for recalibrating and enhancing mmWave features, alleviating the limited orientation of mmWave sensing. To sum up, *Wavoice* can provide flexible speech recognition applications even if users are in remote locations.

5.5.2 Impact of Body Motion. The user is unlikely to keep still in front of smart speakers. Hence, we conduct this experiment to show how well *Wavoice* performs when the user is in body motion. Five subjects are requested to speak commands and perform body motions simultaneously. The body motions include making telephone calls, typing on phones, shaking arms, and marching on the spot. The corresponding results of five subjects across different body motions are shown in Figure 11. The average CER of calling and typing on a smartphone is 0.33% and 0.37%, respectively. The CERs of our system show a slight increase compared to those in still conditions, but they remain below 1%. When users are in motion, such as walking, the directly extracted phase difference from mmWave radars tend to be polluted with motion interference. Whereas acoustic information containing semantic contents is fused into the system, the vocal vibration masked by interference can be mined and enhanced. Therefore, the motion interference has a limited influence on the system.

5.6 Case Study

In realistic scenes, interactions between users and *Wavoice* are likely to be impeded by obstacles. We conduct three case studies where face masks, wearing accessories, or even solid obstacles cover users' acoustic organs and block the signal propagation. The experiment scene is set in an office with a noise of 40 dB. Data in case studies are collected from 20 participants (10 male and 10 female, mean 28.4, and standard deviation 7.3) who sit 7 m away from sensors.

5.6.1 Performance Under Masks. Currently, wearing masks is a growing consensus among people. Masks on users may influence the quality of spoken commands. Facial masks are frequently used for medical self-protection in public scenes. To study the impact of user-wearing masks to the system, we consider four types of commonly used masks, disposable medical masks, N95 respirator masks, gas masks, and anti-dust masks. We conduct a series of experiments where the participants wear a given mask and speak utterances. Except for face masks, the subjects are required to wear a scarf for further measuring the penetration of *Wavoice*. Table 3 lists all selected masks and their corresponding indexes. The speech recognition results in Figure 12 present that the mask worn by

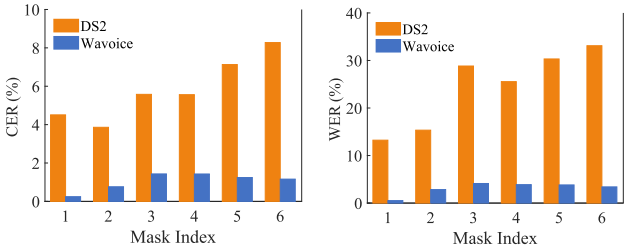


Fig. 12. Performance of *Wavoice* and DS2 influenced by masks without noise.

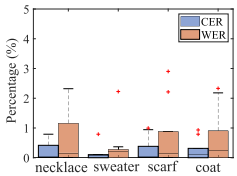


Fig. 13. Impact of clothing.

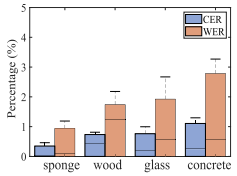


Fig. 14. Impact of obstruction.

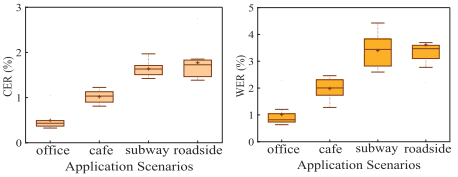


Fig. 15. Performance under environmental disturbance.

users definitely deteriorates the property and quality of speech signals, resulting in a sharp decline in system performance of a traditional ASR system like DS2. On the contrary, the proposed system significantly outperforms the baseline since its CERs are all nearly 1% under any type of mask. Our system is at least five times better than traditional systems, according to this comparison result. This comparison experiment confirms the usability of the mmWave-voice system against acoustic degradation caused by wearing masks. Thus, we validate that the multi-modal system fusing mmWave and speech signals can boost the ability of speech recognition despite users wearing masks.

5.6.2 Performance Under Clothing. It is common for users to wear clothing around the throat, which seems to affect the performance of *Wavoice*. To validate *Wavoice*'s usability considering the wearing of different accessories, we choose four types of clothing: a necklace, sweater, scarf, and coat. Figure 13 displays all of the recognition results when users wear different clothes. It can be seen that the speech recognition accuracy of the system is high under all types of clothing. The proposed system achieves a CER below 1%. Moreover, *Wavoice* performs slightly better when the throat is not covered by clothing. We envision that the wearing of accessories like a necklace and coat could block reflective mmWave signals and partly weaken the vocal vibration from the human throat.

5.6.3 Performance Under Obstructions. One challenging factor to wireless sensing is occlusion between the radar and targeting users, since both transmitted and reflected signals are impaired when penetrating an occlusion. Herein, we respectively put four universal obstacles (i.e., a sponge, wood, glass, and a concrete wall) between the radar and the user. The thickness of the sponge, wood, glass, and concrete wall are 1, 0.7, 2, and 10 cm, respectively. The recognition results for the preceding are shown in Figure 14. It is observed that the performance of *Wavoice* under dense and thick shields, especially concrete walls, is slightly lower than those under no obstruction. However, *Wavoice* still maintains accurate speech recognition with a CER below 1.3%. This is because the supplement information from the speech modality makes up for the loss of mmWave information when propagating through obstructions.

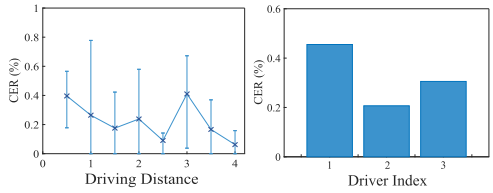


Fig. 16. Performance of *Wavoice*'s in-vehicle application.

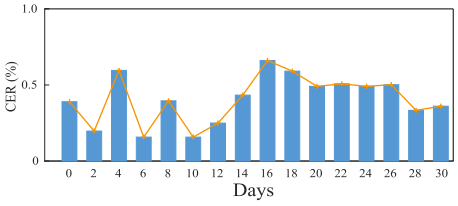


Fig. 17. A permanence study lasting 1 month.

5.7 Environmental Disturbance

Considering previous experiments set in controlled laboratory environments, we conduct comprehensive experiments in realistic scenes. Four common and noisy experiments are chosen: a filled office, a noisy cafe, a busy roadside, and a subway. Five subjects (three female and two male) are requested to speak voice commands naturally in each specific scene. The corresponding data is collected to examine the pre-trained system and verify its universality. Figure 15 displays the overall speech recognition results. The averaging CERs of different scenes are 0.49%, 1.02%, 1.64%, and 1.77%. Although speech recognition accuracy is slightly degraded, the proposed system still guarantees its superiority in arbitrary realistic scenes.

To further demonstrate the universality of *Wavoice*, we set it up in a vehicle where the mmWave radar often wobbles during driving. We ask three participants as drivers to speak voice commands as they drive the vehicle. The mmWave radar and microphone in the system are appropriately placed on the automotive center stack, which will not influence driving. Specifically, each driver drives 20 minutes at a normal speed on an urban route. While driving, we play music inside the vehicle when drivers speak commands. The collected data is fed into the system for examining the speech recognition ability of *Wavoice*.

Based on Figure 16, the averaging CER remains lower than 0.5% with a driving distance from 0 to 4 km. The CER of three drivers is 45%, 0.20%, and 0.30%, respectively. This indicates that *Wavoice* is competent for speech recognition in a vehicle despite music interference and sensor wobbling. This is logical because the system can receive enough useful mmWave and speech signals in a narrow space to produce fusion results for accurate speech recognition.

Potential Application. One potential application includes an in-vehicle voice control system. The results in Section 5.7 indicate that *Wavoice* promises good performance under a noisy and wobble in-car environment, whereas the traditional audio-only method has limited performance. Thus, *Wavoice* can act as a novel in-vehicle voice control system to facilitate the robustness of speech recognition when the user is driving. Another potential application is VUIs in public facilities. Public facilities, such as ATMs and vending machines at the subway, are often flooded by noise, which limits the application of an audio-only VUI on these machines. The facilities can benefit from *Wavoice*'s noise-resilience advantage for non-contact user interaction, which is urgent due to the growing demand for public hygiene.

5.8 Permanence Analysis

For a biometric system, it is necessary to examine the permanence of *Wavoice*. We collect 40 pieces of mmWave and speech data from 20 subjects (10 males and 10 females) every 2 days. In every data collection period, each subject is required to speak commands randomly from ok-google.io. The data collection lasts for a month. The pre-trained system takes in the collected data and outputs the recognized result. Figure 17 shows the recognition performance of the system during this month. During the 1-month experiment, the average CER is between 0.15% and 0.66%. The maximum

fluctuation in CER results is no more than 0.5%, which proves that multi-modal features fused by *Wavoice* can efficiently maintain reliable performance over a long period. Moreover, there is no notable decreasing and ascending tendency on speech recognition accuracy of *Wavoice*. This indicates that *Wavoice* is robust to the time dynamic.

6 DISCUSSION

Hardware Support. Compared with traditional ASR systems [4, 22], *Wavoice* needs an extra mmWave radar but only one low-cost microphone. Moreover, the deployment of mmWave radars has rapidly increased owing to the advanced mmWave sensing and communication technologies [68] in wireless sensing and 5G communication [82]. For instance, Google Pixel 4 [21] has been installed with a miniature mmWave radar for convenient human-machine interaction. The manufactural microphone array arranged in a special pattern aims at performing speech recognition function under ambient noises. However, these designs of microphone array demand high volume but obtain small listening coverage. In this case, it is promising that mmWave-microphone ASR systems will be migrated to voice-enabled devices in various realistic scenarios. The current prototype of *Wavoice* simply employs a microphone and a radar to support speech recognition. With the upgrade of sensor hardware, it is feasible to apply *Wavoice* on the microphone array and antenna array, improving not only perception accuracy but also signal quality. We leave the hardware upgrade and relevant performance test to future work.

Sensing Range. Extensive experimental results have demonstrated that *Wavoice* owns a sensing coverage of 7 m with a 120-degree field of view. It is feasible to transfer *Wavoice* into most public human-machine applications where users face devices within a finite field, such as voice-enabled ATMs and elevators. In terms of fully open working areas, like smart streetlights, it is suggested to equip three radars with 360-degree coverage, which is expensive. A low-cost method is to swing the mmWave radar with the help of user targeting, which has been employed into Mi Air Charge [90]. However, *Wavoice* can adopt microphone arrays instead of a single microphone for enlarging working areas.

Cost and Power Consumption. *Wavoice* needs a low-cost microphone and a COTS mmWave radar. In the aspect of hardware cost, an mmWave radar chip costs about 40 dollars [33], whereas a microphone only costs 10 cents. Given that long-distance speech recognition under a relatively low SNR, it is more worthwhile to employ *Wavoice* than microphone arrays with average costs of around 50 dollars. Additionally, there is a visible trend in which the cost of mmWave radars is on the decrease as electronic industry manufacturing grows. In terms of power consumption, the power sum of the mmWave radar and microphone stays below 20 mW. Its power consumption is so low that *Wavoice* will be acceptable for most voice-controlled devices.

Easy Interaction Without a Predetermined Wake-Up Word. The wake-up word introduced in our system is to activate the speech recognition function. Our system could accurately receive and recognize voice commands from a user by leveraging wake-up words, even in a multi-person scene. However, users are inclined to own private voice-controlled devices to protect personal information and property against audio injection attacks [71, 95]. Recent research reveals the person difference in vocal vibration independent of speech context [91] whereby we can exploit mmWave sensing vocal vibration to target the user. Therefore, our system can execute a well-trained classifier based on the uniqueness of vocal vibration for verifying and targeting the user. Then, only registered users could say voice commands without wake-up words to activate our system.

Speech Separation. Speech separation aims at separating and restoring users' speech among mixed noise, more often from a cocktail party effect. Considering the benefit of speech separation

to voice interaction, it is valuable for *Wavoice* to extend speech separation from heterogeneous speech signals. Motivated by the advantage of deep complex networks [12, 75], the proposed system has the potential for realizing speech separation. Owing to the flexible framework of *Wavoice*, we can easily replace the semantic extraction network in the system with deep complex networks. After the complex network predicting the magnitude and phase spectrogram of users' speeches, the original speech can be subsequently obtained by implementing inverse Fourier transform.

Multi-Modal Fusion. *Wavoice* has significant potential to facilitate multi-modal fusion in a large range of applications, such as mmWave-WiFi human activity recognition [36] and audio-image speech enhancement [55]. This work concretely designs multi-modal methods from voice activity detection to feature fusion, which are applicable to other applications assisted with different modalities. Beyond the specific method of *Wavoice*, this article convincingly presents that the utilization of correlation among different modalities is a critical key to multi-modal systems. To further prove it, future work will be done in replicating our techniques in other systems.

7 RELATED WORK

mmWave-based sensing promotes high-accuracy detection and perception, improving an extensive range of sensing applications from tracking and localization [84], human activity recognition [15], and vital sign monitoring [98, 99] to acoustic sensing [7, 29, 30]. Owing to optical sensors' sensitivity to environmental conditions, Chang et al. [9] fused the mmWave radar with optical sensors to overcome environmental challenges and enhance the desired results. These mmWave-vision systems can largely boost sensing resolution and ranges [14, 48, 96]. The cooperation of mmWave radar and IMU [31] can support robust ego-motion estimation [3, 50]. Furthermore, Lu et al. [49] leveraged an mmWave radar and lidar to reconstruct an indoor grid map. Similar works exhibited the collaboration between mmWave and other sensors, facilitating system stability and efficiency.

Speech enhancement aims at the advanced quality and intelligibility of desired speech under arbitrary noise and reverberation, often with the help of microphone arrays [6, 24, 38, 47, 59, 86]. Classic statistics-based approaches [25, 89] has prior knowledge of environmental noise. Learning-based speech enhancement has won popularity that leverages DNNs [60, 92] or generative adversarial networks [16, 61] but fails in long-distance speech recognition with sensing ranges. These techniques require excessive microphones (more than the number of noise sources) and special physical layouts.

Cross-modal speech recognition gives new methods of efficiently improving recognition accuracy against noise interference. The audio-visual systems integrating face landmarks [57] or lip motion [1] estimate the vibration caused by pronunciation to refine expected speeches. Meanwhile, ultrasound [39, 72], WiFi [81], and inertial signals [2] can extract semantic information or enhance speech within a limited range. Unlike other existing work, *Wavoice* fuses mmWave and speech by using a delicate network with SENet-based inter-attention, which provides long-distance speech recognition in public scenes filled with noise and motion disturbances.

8 CONCLUSION

In this article, we presented a novel system, *Wavoice*, for long-range, noise-resilient, and motion-robust speech recognition by fusing mmWave and acoustic signals. We first formulated the correlation between mmWave and acoustic signals, based on which we developed voice activity detection to combat against noise interference and a target-localizing method to separate the user from backgrounds. To achieve the noise-resistant speech recognition, we designed an attention-based network with two specialized modules leveraging the inter-attention between the multi-modal

signals to enhance recognition performance. We performed extensive experiments to evaluate the proposed system, which shows high resilience to ambient noise and face masks. The results indicated that *Wavoice* can achieve an error rate as low as 1% even in a long-range condition.

REFERENCES

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121* (2018).

[2] Omer Saad Alkhafaf, Mousa K. Wali, and Ali H. Al-Timemy. 2020. Improved prosthetic hand control with synchronous use of voice recognition and inertial measurements. *IOP Conference Series: Materials Science and Engineering* 745, 1 (2020), 012088.

[3] Yasin Almalioglu, Mehmet Turan, Chris Xiaoxuan Lu, Niki Trigoni, and Andrew Markham. 2020. Milli-RIO: Ego-motion estimation with low-cost millimetre-wave radar. *IEEE Sensors Journal* 21, 3 (2020), 3314–3323.

[4] Amazon. 2022. Amazon Echo. Retrieved May 27, 2023 from https://www.amazon.com/echo/.

[5] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, et al. 2016. Deep Speech 2: End-to-end speech recognition in English and Mandarin. In *Proceedings of the International Conference on Machine Learning*. 173–182.

[6] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. 2018. The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. *arXiv preprint arXiv:1803.10609* (2018).

[7] Chao Cai, Rong Zheng, and Jun Luo. 2022. Ubiquitous acoustic sensing on commodity IoT devices: A survey. *IEEE Communications Surveys & Tutorials* 24, 1 (2022), 432–454.

[8] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. 4960–4964.

[9] Shuo Chang, Yifan Zhang, Fan Zhang, Xiaotong Zhao, Sai Huang, Zhiyong Feng, and Zhiqing Wei. 2020. Spatial attention fusion for obstacle detection using mmWave radar and vision sensor. *Sensors* 20, 4 (2020), 956.

[10] Fuming Chen, Sheng Li, Chuantao Li, Miao Liu, Zhao Li, Huijun Xue, Xijing Jing, and Jianqi Wang. 2016. A novel method for speech acquisition and enhancement by 94 GHz millimeter-wave sensor. *Sensors* 16, 1 (2016), 50.

[11] Xingyu Chen, Zhengxiong Li, Baicheng Chen, Yi Zhu, Chris Xiaoxuan Lu, Zhengyu Peng, Feng Lin, Wenyao Xu, Kui Ren, and Chunming Qiao. 2022. MetaWave: Attacking mmWave sensing with meta-material-enhanced tags. In *Proceedings of the 30th Network and Distributed System Security Symposium (NDSS'22)*.

[12] Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee. 2019. Phase-aware speech enhancement with deep complex U-net. In *Proceedings of the International Conference on Learning Representations*.

[13] Livija Cveticanin. 2012. Review on mathematical and mechanical models of the vocal cord. *Journal of Applied Mathematics* 2012 (2012), Article 928591, 18 pages.

[14] Joseph St. Cyr, Joshua Vanderpool, Yu Chen, and Xiaohua Li. 2020. HODET: Hybrid object detection and tracking using mmWave radar and visual sensors. In *Sensors and Systems for Space Applications XIII*. Proceedings Volume 11422. SPIE, 114220I.

[15] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. 2020. RF-Net: A unified meta-learning framework for RF-enabled one-shot human activity recognition. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 517–530.

[16] Chris Donahue, Bo Li, and Rohit Prabhavalkar. 2018. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. 5024–5028.

[17] Ming Gao, Feng Lin, Weiye Xu, Muertikepu Nuermaimaiti, Jinsong Han, Wenyao Xu, and Kui Ren. 2020. Deaf-Aid: Mobile IoT communication exploiting stealthy speaker-to-gyroscope channel. In *Proceedings of the Annual International Conference on Mobile Computing and Networking*. 1–13.

[18] Gmtd. 2022. GM-A906 Microphone. Retrieved May 27, 2023 from https://item.jd.com/69500632000.html.

[19] Zenton Goh, Kah-Chye Tan, and T. G. Tan. 1998. Postprocessing method for suppressing musical noise generated by spectral subtraction. *IEEE Transactions on Speech and Audio Processing* 6, 3 (1998), 287–292.

[20] Google. 2017. Ok Google. Retrieved May 27, 2023 from https://ok-google.io/.

[21] Android Central. 2019. Here's How the Pixel 4's Soli Radar Works and Why Motion Sense Has So Much Potential. Retrieved May 27, 2023 from https://www.androidcentral.com/how-does-googles-soli-chip-work.

[22] Google. 2022. Google Home. Retrieved May 27, 2023 from https://store.google.com/product/google_home/.

[23] Surjeet and Nishu Gupta. 2021. A novel voice controlled robotic vehicle for smart city applications. *Journal of Physics: Conference Series* 1817, 1 (2021), 012016.

[24] Mary Harper. 2015. The automatic speech recognition in reverberant environments (ASpIRE) challenge. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*. 547–554.
[25] H. G. Hirsch and C. Ehrlicher. 1995. Noise estimation techniques for robust speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. 153–156.
[26] Hong Hong, Heng Zhao, Zhengyu Peng, Hui Li, Chen Gu, Changzhi Li, and Xiaohua Zhu. 2016. Time-varying vocal folds vibration detection using a 24 GHz portable auditory radar. *Sensors* 16, 8 (2016), 1181.
[27] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. 2019. Cross attention network for few-shot classification. In *Advances in Neural Information Processing Systems*. 4005–4016.
[28] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. 7132–7141.
[29] Pengfei Hu, Wenhao Li, Riccardo Spolaor, and Xiuzhen Cheng. 2023. mmEcho: A mmWave-based acoustic eavesdropping method vocabulary. In *Proceedings of the 2023 IEEE Symposium on Security and Privacy (SP'23)*.
[30] Pengfei Hu, Yifan Ma, Panneer Selvam Santhalingam, Parth H. Pathak, and Xiuzhen Cheng. 2022. MILLIEAR: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary. In *Proceedings of the 2022 IEEE Conference on Computer Communications (IEEE INFOCOM'22)*. 11–20.
[31] Pengfei Hu, Hui Zhuang, Panneer Selvam Santhalingam, Riccardo Spolaor, Parth Pathak, Guoming Zhang, and Xiuzhen Cheng. 2022. AccEar: Accelerometer acoustic eavesdropping with unconstrained vocabulary. In *Proceedings of the 2022 IEEE Symposium on Security and Privacy (SP'22)*. 1530–1530.
[32] Apple Inc. 2022. Siri. Retrieved May 27, 2023 from https://www.apple.com/au/siri/.
[33] Texas Instruments. 2022. IWR1642BOOST. Retrieved May 27, 2023 from https://www.ti.com/tool/IWR1642BOOST.
[34] Texas Instruments. 2022. MMWAVE-STUDIO. Retrieved May 27, 2023 from https://www.ti.com/tool/MMWAVE-STUDIO.
[35] Christopher I. Jarvis, Kevin Van Zandvoort, Amy Gimma, Kiesha Prem, Petra Klepac, G. James Rubin, and W. John Edmunds. 2020. Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK. *BMC Medicine* 18 (2020), 1–10.
[36] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, et al. 2018. Towards environment independent device free human activity recognition. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 289–304.
[37] Kaustubh Kalgaonkar, Rongquiang Hu, and Bhiksha Raj. 2007. Ultrasonic Doppler sensor for voice activity detection. *IEEE Signal Processing Letters* 14, 10 (2007), 754–757.
[38] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël A. P. Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, et al. 2016. A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing* 2016, 1 (2016), 1–19.
[39] Ki-Seung Lee. 2019. Speech enhancement using ultrasonic Doppler sonar. *Speech Communication* 110 (2019), 21–32.
[40] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, et al. 2020. VocalPrint: Exploring a resilient and secure voice authentication via mmWave biometric interrogation. In *Proceedings of the Conference on Embedded Networked Sensor Systems*. 312–325.
[41] Sheng Li, Ying Tian, Guohua Lu, Yang Zhang, Hui Jun Xue, Jian-Qi Wang, and Xi-Jing Jing. 2012. A new kind of non-acoustic speech acquisition method based on millimeter waveradar. *Progress in Electromagnetics Research* 130 (2012), 17–40.
[42] Sheng Li, Jian-Qi Wang, Ming Niu, Tian Liu, and Xi-Jing Jing. 2008. The enhancement of millimeter wave conduct speech based on perceptual weighting. *Progress in Electromagnetics Research* 9 (2008), 199–214.
[43] Zhengxiong Li, Baicheng Chen, Xingyu Chen, Chenhan Xu, Yuyang Chen, Feng Lin, Changzhi Li, Karthik Dantu, Kui Ren, and Wenyao Xu. 2022. Reliable digital forensics in the air: Exploring an RF-based drone identification system. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–25.
[44] Zhengxiong Li, Fenglong Ma, Aditya Singh Rathore, Zhuolin Yang, Baicheng Chen, Lu Su, and Wenyao Xu. 2020. WaveSpy: Remote and through-wall screen attack via mmWave sensing. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'20)*. 217–232.
[45] Feng Lin, Chen Song, Yan Zhuang, Wenyao Xu, Changzhi Li, and Kui Ren. 2017. Cardiac Scan: A non-contact and continuous heart-based user authentication system. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. 315–328.
[46] Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, and Kui Ren. 2021. Wavoice: A noise-resistant multi-modal speech recognition system fusing mmWave and audio signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 97–110.
[47] Philipos C. Loizou. 2013. *Speech Enhancement: Theory and Practice*. CRC Press, Boca Raton, FL.

[48] Ningbo Long, Kaiwei Wang, Ruiqi Cheng, Kailun Yang, and Jian Bai. 2018. Fusion of millimeter wave radar and RGB-depth sensors for assisted navigation of the visually impaired. In *Millimetre Wave and Terahertz Sensors and Technology XI*. Proceedings Volume 10800. SPIE, 1080006.

[49] Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A. Stankovic, Niki Trigoni, and Andrew Markham. 2020. See through smoke: Robust indoor mapping with low-cost mmWave radar. In *Proceedings of the International Conference on Mobile Systems, Applications, and Services*.

[50] Chris Xiaoxuan Lu, Muhamad Risqi U. Saputra, Peijun Zhao, Yasin Almalioglu, Pedro P. B. de Gusmao, Changhao Chen, Ke Sun, Niki Trigoni, and Andrew Markham. 2020. milliEgo: Single-chip mmWave radar aided egomotion estimation via deep sensor fusion. In *Proceedings of the Conference on Embedded Networked Sensor Systems*. 109–122.

[51] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems*. 1–9.

[52] Michal Luria, Guy Hoffman, and Oren Zuckerman. 2017. Comparing social robot, screen and voice interfaces for smart-home control. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 580–628.

[53] Rainer Martin. 1994. Spectral subtraction based on minimum statistics. In *Proceedings of the European Signal Processing Conference (EUSIPCO'94)*., 1182–1185.

[54] Youri Maryn, Floris L. Wuyts, and Andrzej Zarowski. 2023. Are acoustic markers of voice and speech signals affected by nose-and-mouth-covering respiratory protective masks? *Journal of Voice* 37, 3 (2023), 468.e1–468.e12.

[55] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. 2021. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1368–1396.

[56] Ryoichi Miyazaki, Hiroshi Saruwatari, Takayuki Inoue, Yu Takahashi, Kiyohiro Shikano, and Kazunobu Kondo. 2012. Musical-noise-free speech enhancement based on optimized iterative spectral subtraction. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 7 (2012), 2080–2094.

[57] Giovanni Morrone, Sonia Bergamaschi, Luca Pasa, Luciano Fadiga, Vadim Tikhanoff, and Leonardo Badino. 2019. Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. 6900–6904.

[58] Tomer Moscovich. 2009. Contact area interaction with sliding widgets. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*. 13–22.

[59] Mahesh Kumar Nandwana, Julien Van Hout, Mitchell McLaren, Colleen Richey, Aaron Lawson, and Maria Alejandra Barrios. 2019. The VoiCES from a Distance Challenge 2019 evaluation plan. *arXiv preprint arXiv:1902.10828* (2019).

[60] Ashutosh Pandey and DeLiang Wang. 2019. A new framework for CNN-based speech enhancement in the time domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 7 (2019), 1179–1188.

[61] Santiago Pascual, Antonio Bonafonte, and Joan Serra. 2017. SEGAN: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452* (2017).

[62] Lionel Sharples Penrose. 1946. The elementary statistics of majority voting. *Journal of the Royal Statistical Society* 109, 1 (1946), 53–57.

[63] Rohit Prabhavalkar, Kanishka Rao, Tara N. Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly. 2017. A comparison of sequence-to-sequence models for speech recognition. In *Proceedings of the Conference of the International Speech Communication Association*. 939–943.

[64] Vineel Pratap, Awni Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky, and Ronan Collobert. 2019. Wav2Letter++: A fast open-source speech recognition system. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. 6460–6464.

[65] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangment Zuo, and Qinghua Hu. 2020. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11531–11539.

[66] Hermann Rohling and Ralph Mende. 1996. OS CFAR performance in a 77 GHz radar sensor for car application. In *Proceedings of the International Radar Conference*. 109–114.

[67] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. 2018. BackDoor: Sounds that a microphone can record, but that humans can't hear. *GetMobile: Mobile Computing and Communications* 21, 4 (2018), 25–29.

[68] Kei Sakaguchi, Thomas Haustein, Sergio Barbarossa, Emilio Calvanese Strinati, Antonio Clemente, Giuseppe Destino, Aarno Pärssinen, et al. 2017. Where, when, and how mmWave is used in 5G and beyond. *IEICE Transactions on Electronics* 100, 10 (2017), 790–808.

[69] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. 1999. A statistical model-based voice activity detection. *IEEE Signal Processing Letters* 6, 1 (1999), 1–3.

[70] SoundAI. 2020. Voice-Controlled Elevator System Put into Use in Beijing. Retrieved May 27, 2023 from https://www.chinadaily.com.cn/a/202003/13/WS5e6b3fcca31012821727ef88.html.

[71] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu. 2020. Light commands: Laser-based audio injection attacks on voice-controllable systems. In *Proceedings of the 29th USENIX Security Symposium (USENIX Security'20)*. 2631–2648.
[72] Ke Sun and Xinyu Zhang. 2021. UltraSE: Single-channel speech enhancement using ultrasound. In *Proceedings of the Annual International Conference on Mobile Computing and Networking*. 160–173.
[73] Telsa. 2022. Model S/3/X/Y. Retrieved May 27, 2023 from https://www.tesla.com/.
[74] Texas Instruments. 2022. DCA1000EVM. Retrieved May 27, 2023 from https://www.ti.com/tool/DCA1000EVM.
[75] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J. Pal. 2018. Deep complex networks. In *Proceedings of the International Conference on Learning Representations*. 1–19.
[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 6000–6010.
[77] Chao Wang, Feng Lin, Zhongjie Ba, Fan Zhang, Wenyao Xu, and Kui Ren. 2022. Wavesdropper: Through-wall word detection of human speech via commercial mmWave devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–26.
[78] Chao Wang, Feng Lin, Tiantian Liu, Ziwei Liu, Yijie Shen, Zhongjie Ba, Li Lu, Wenyao Xu, and Kui Ren. 2022. mmPhone: Acoustic eavesdropping on loudspeakers via mmWave-characterized piezoelectric effect. In *Proceedings of the 2022 IEEE Conference on Computer Communications (IEEE INFOCOM'22)*. IEEE, Los Alamitos, CA, 820–829.
[79] Chao Wang, Feng Lin, Tiantian Liu, Kaidi Zheng, Zhibo Wang, Zhengxiong Li, Ming-Chun Huang, Wenyao Xu, and Kui Ren. 2022. mmEve: Eavesdropping on smartphone's earpiece via COTS mmWave device. In *Proceedings of the 28th Annual International Conference on Mobile Computing and Networking*. 338–351.
[80] Dong Wang, Xiaodong Wang, and Shaohe Lv. 2019. An overview of end-to-end automatic speech recognition. *Symmetry* 11, 8 (2019), 1018.
[81] Guanhua Wang, Yongpan Zou, Zimu Zhou, Kaishun Wu, and Lionel M. Ni. 2016. We can hear you with Wi-Fi! *IEEE Transactions on Mobile Computing* 15, 11 (2016), 2907–2920.
[82] Jie Wang, Qinhua Gao, Xiaorui Ma, Yunong Zhao, and Yuguang Fang. 2020. Learning to sense: Deep learning for wireless sensing with less training efforts. *IEEE Wireless Communications* 27, 3 (2020), 156–162.
[83] Jiwu Wang, Xuewei Hu, and Chengyu Tong. 2021. Urban community sustainable development patterns under the influence of COVID-19: A case study based on the non-contact interaction perspective of Hangzhou City. *Sustainability* 13, 6 (2021), 3575.
[84] Xuyu Wang, Mohini Patil, Chao Yang, Shiwen Mao, and Palak Anilkumar Patel. 2021. Deep convolutional Gaussian processes for mmWave outdoor localization. In *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'21)*. 8323–8327.
[85] Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209* (2018).
[86] Shinji Watanabe, Michael Mandel, Jon Barker, and Emmanuel Vincent. 2020. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. *arXiv preprint arXiv:2004.09249* (2020).
[87] Canalys. 2020. Global Smart Speaker Market 2021 Forecast. Retrieved May 27, 2023 from https://www.canalys.com/newsroom/canalys-global-smart-speaker-market-2021-forecast.
[88] Chris Wiltz. 2020. COVID-19 Giving Touchless Interfaces a Chance to Make an Impression. Retrieved May 27, 2023 from https://www.designnews.com/design-hardware-software/covid-19-giving-touchless-interfaces-chance-make-impression-0.
[89] Mingyang Wu and DeLiang Wang. 2006. A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 3 (2006), 774–784.
[90] Xiaomi. 2021. 'Not science fiction': XiaomiŠs revolutionary new wireless charging tech can charge your devices remotely. *FE TECH BYTES*. Retrieved May 27, 2023 from https://www.financialexpress.com/industry/technology/.
[91] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. 2019. WaveEar: Exploring a mmWave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of the Annual International Conference on Mobile Systems, Applications, and Services*. 14–26.
[92] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. 2020. PHASEN: A phase-and-harmonics-aware speech enhancement network. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 9458–9465.
[93] Dong Yu and Li Deng. 2016. *Automatic Speech Recognition*. Vol. 1. Springer.
[94] Li Yujian and Liu Bo. 2007. A normalized Levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 6 (2007), 1091–1095.
[95] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. DolphinAttack: Inaudible voice commands. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 103–117.

[96] Renyuan Zhang and Siyang Cao. 2019. Extending reliability of mmWave radar tracking and detection via fusion with camera. *IEEE Access* 7 (2019), 137065–137079.
[97] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision*. 286–301.
[98] Tianyue Zheng, Zhe Chen, Chao Cai, Jun Luo, and Xu Zhang. 2020. V2iFi: In-vehicle vital sign monitoring via compact RF sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–27.
[99] Tianyue Zheng, Zhe Chen, Shujie Zhang, and Jun Luo. 2022. Catch your breath: Simultaneous RF tracking and respiration monitoring with radar pairs. *IEEE Transactions on Mobile Computing*. Early access, August 9, 2022.

Received 23 December 2022; revised 17 March 2023; accepted 8 May 2023