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

Tiantian Liu[1], Ming Gao[1], Feng Lin[1,2*], Chao Wang[1], Zhongjie Ba[1], Jinsong Han[1], Wenyao Xu[3], Kui Ren[1]

[1]Zhejiang University, Hangzhou, China

[2]Key Laboratory of Blockchain and Cyberspace Governance of Zhejiang Province, Hangzhou, China

[3]University at Buffalo, the State University of New York, Buffalo, New York, USA

{tiantian,gaomingppm,flin,wangchao5001,zhongjieba,hanjinsong,kuiren}@zju.edu.cn,wenyaoxu@buffalo.edu

ABSTRACT

With the advance in automatic speech recognition, voice user interface has gained popularity recently. Since the COVID-19 pandemic, VUI is increasingly preferred in online communication due to its non-contact. Additionally, various ambient noise impedes the public applications of voice user interfaces due to the requirement of audio-only speech recognition methods for a high signal-to-noise ratio. In this paper, we present Wavoice, the first noise-resistant multi-modal speech recognition system that fuses two distinct voice sensing modalities, i.e., millimeter-wave (mmWave) signals and audio signals from a microphone, together. One key contribution is that we model the inherent correlation between mmWave and audio signals. Based on it, Wavoice facilitates the real-time noise-resistant voice activity detection and user targeting from multiple speakers. Furthermore, we elaborate on two novel modules into the neural attention mechanism for multi-modal signals fusion, and result in accurate speech recognition. Extensive experiments verify Wavoice's effectiveness under various conditions with the character recognition error rate below 1% in a range of 7 meters. Wavoice outperforms existing audio-only speech recognition methods with lower character error rate and word error rate. The evaluation in complex scenes validates the robustness of Wavoice.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Speech recognition, voice user interface, mmWave sensing, multi-modal fusion

ACM Reference Format:

Tiantian Liu[1], Ming Gao[1], Feng Lin[1,2*], Chao Wang[1], Zhongjie Ba[1], Jinsong Han[1], Wenyao Xu[3], and Kui Ren[1]. 2021. Wavoice: A Noise-resistant Multi-modal Speech Recognition System Fusing mmWave and Audio Signals. In *The 19th ACM Conference on Embedded Networked Sensor Systems (SenSys'21), November 15–17, 2021, Coimbra, Portugal.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3485730.3485945

*Feng Lin is the corresponding author.

SenSys'21, November 15–17, 2021, Coimbra, Portugal

ACM ISBN 978-1-4503-9097-2/21/11…$15.00

https://doi.org/10.1145/3485730.3485945

1 INTRODUCTION

Voice user interface (VUI) plays an essential role in intelligent scenes, e.g. smart homes [41]. It provides a hands-free and eyes-free human-machine interaction between humans and Internet of Things devices. Benefiting from the development of deep learning and natural language process, the automatic speech recognition (ASR) entitles VUI to the capacity of accurate comprehension on users' intentions [77]. With such a convenient and flexible service, users can interact with various devices as they please. Commercial VUI products have gained in popularity over recent years, such as smart speakers (e.g., Amazon Echo [4] and Google Home [17]), voice assistants in smartphones (e.g. Siri [25]), and in-vehicle voice control interactions (e.g. VUIs in Tesla Model S/X/3/Y [58]). Analysts forecast that by 2024, the deployment of VUI-based smart speakers will reach 640 million globally [70].

Nowadays, VUI tends to branch out into the smart city business [19]. Non-contact interaction, represented by VUI, has been widely deployed in public places [67]. It gradually replaces traditional contact interaction such as button or touch interactions [44]. Especially due to the corona virus disease-19 (COVID-19) pandemic [26], people avoid physical contact with public facilities for safety reasons. For example, VUIs have been exploited for voice-controlled elevators [55] and ATMs [71]. Different from home scenes, VUI needs to address more multifarious ambient noise (e.g. traffic noise, commercial noise, and nearby voices) in public places (e.g. streets, stations, halls, or parties). However, audio-based ASR techniques based on microphone arrays, including traditional statistic-based [21, 72] and advanced learning-based [46, 76], require clear audio signals with a high signal-to-noise ratio (SNR). Hence in public applications, audio signals, drowned in the unpredictable noise, become difficult to identify. Additionally, to protect themselves from the corona virus, people prefer to wear respiratory protective face masks [42], which further degrades acoustic quality and encumbers speech recognition accuracy [42]. Audio-only methods are incompetent to support VUI in these cases.

To address these difficulties above, researchers exploit multi-sensor information fusion for speech enhancement and recognition.

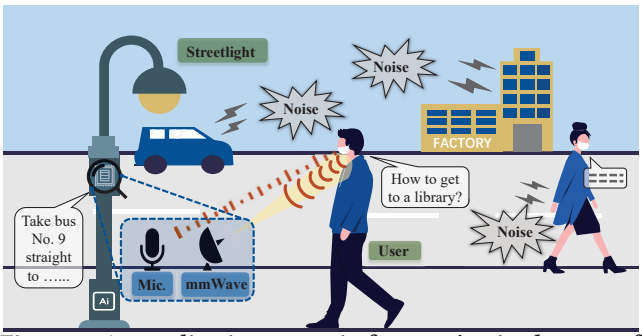


Figure 1: An application scenario for Wavoice in the case of smart city. Users can interact with a Wavoice-powered smart streetlight that provides services including location, navigation, emergency calling, and voice-controlled traffic lights.

Audio-visual methods [1, 43] 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 [30, 56] are merely applied into conditional scenes on account of the extremely short working distance (within 20cm) and specific postural requirements.

We turn attention to millimeter wave (mmWave) radars, and leverage it as a supplementary for speech recognition. It has been demonstrated that mmWave signals contribute to voice information recovery with excellent performance on resistance to ambient noise and penetration [35, 75]. Those signals can detect the vocal vibration by analyzing reflected signals of remote target users even wearing face masks in a noisy environment. Nevertheless, mmWave radars do not always perform satisfactorily. Due to the tiny wavelength (about 4mm), mmWave signals are sensitive to both vocal vibration and motion. They would be affected by users' body movement in practice. To make matters worse, mmWave radars are likely to shake in specific scenarios, e.g. vehicle applications. Motion interference, ignored by prior work [75], would distort reflected signals that contain vocal information of users. mmWave-based applications always suffer from such motion interference from users, radars, or both. Fortunately, microphone-based voice collection can compensate for the loss of information to some extent. Therefore, we consider the complementary collaboration between a mmWave radar and a microphone. These two signals of different modalities are employed together for accurate speech recognition. Here, the mmWave signal encourages noise-resistant speech sensing in spite of face masks, while the audio signal collected by a microphone serves as a guide to calibrate speech features in the mmWave signal under motion interference.

To realize the 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 signals of different modalities to support long-distance VUI applications, while mmWave and audio signals may suffer from noise. (2) How to detect voice activity in an effective and real-time manner, when user's voice is probable to be overlapped by multiple noises. (3) How to apply this ASR system in a multi-person scene, where irrelevant conversation may disturb users' voice commands.

We propose Wavoice, a multi-modal speech recognition system for public VUI applications, as illustrated in Figure 1. It exploits a mmWave radar for detecting users' vocal vibration in noisy environments, and a microphone in case of the motion interference. Moreover, it is able to penetrate through face masks for speech information extraction. To combine their advantages, we investigate the inherent correlation between mmWave and audio signals. For practical application, we design real-time and anti-interference voice activity detection and user targeting methods based on the frequency-dependent property between these multi-modal signals. Furthermore, we introduce two novel modules into the neural attention mechanism for the ASR-oriented multi-modal fusion. One module exchanges valid features for mutual recalibration and characteristic enhancement, while the other module projects respective information into a joint feature space and adjusts weight coefficients dynamically. Therefore, we integrate multi-model signals for semantic features enhancement. As a result, the utterance information is predicted. 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 scenarios with a low recognition error rate. Particularly, it can be adopted into in-vehicle applications against interference of various practical motions.

In conclusion, our contributions are as follows:

- We design a multi-modal ASR system named Wavoice for VUI's public application. It fuses mmWave and audio signals to facilitate accurate speech recognition in case of noise, 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 the character recognition error rate below 1% in a range of 7 meters 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 mmWave Sensing Mechanism

The frequency modulated continuous wave (FMCW) radar is widely used to transmit mmWave signals for the perception of the physical world to capture. It performs well in tiny displacement measurement as well as cover penetration [35, 75].

Distance Estimation. The FMCW radar transmits chirp signals, whose frequency changes linearly in a specific range periodically. The receive antenna in the radar captures the reflective chirp signal from the object. The received chirp is immediately mixed with the transmitted chirp by a mixer to obtain the mixed signal. The mixed signal, including a replica of the transmitted signal, is filtered out by a low-pass filter to obtain the intermediate frequency (IF) signal. Analysing the spectrum of IF signals, we can estimate the distance d between the object and the radar as follows:

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

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

Angle Estimation. An FMCW radar can estimate the angle of arrival (AoA), the elevation angle of reflected signals. It employs multiple antennas where the differential distance from the detected object to each antenna results in a phase change ω. We can obtain AoA as follows,

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

where λ is the wavelength and l represents the distance between the receiving antennas.

Speech Sensing. Due to the sensitivity to displacements, mmWave signals are exploited for speech sensing [9, 22, 31, 33, 34, 75]. Researchers started from the mmWave-based vocal vibration detection against noise interference [33, 34]. Further research [9, 22, 75] leveraged mmWave to capture vocal vibration for the reconstruction of genuine speech. Additionally, a mmWave radar can distinguish subtle differences of users' vocal vibration, whose uniqueness supports a mmWave-assistant non-contact voice authentication. However, above systems [9, 22, 33, 34, 75] have a limited sensing distance of at most 2 meters. Moreover, they are vulnerable to motion influence. The short sensing range and vulnerability to motion restrict mmWave-based systems' application in practice, especially the public speech recognition.

2.2 Attention Mechanism for Fusion

We aim at a multi-modal speech recognition system based on the mmWave and audio signals. The key issue is to maximum advantages of both signals to deal with complex scenes such as ambient noise and long-distance sensing. voting mechanism [48] seems to be a convenient assistant to the multi-modal fusion. It selects the better results from simultaneous signals of different modalities as final ones. It can compensate for information loss if one kind of signal is ruined. However, mmWave and audio signals are likely to be corrupted simultaneously. For example, users may fidget while calling on ASR-based devices in a noisy environment. Furthermore, in long-distance sensing tasks, the acoustic attenuation would induce a further cut in the audio SNR, while significant multipath effect of mmWave signals introduces additional noise masking valid information.In this case, a simple voting mechanism cannot afford a long-distance speech recognition.

Attention-based networks may provide a possible solution. Attention mechanism has been widely used in the information fusion [32, 40, 57]. Incorporating attention modules into deep neural networks (DNNs) has shown significant success across multiple fields, such as natural language processing [63] and computer vision tasks [23]. Various attention modules [24, 40, 51] are proposed for the better fusion. In particular, efficient channel attention (ECA) [51] performs well in guiding networks to notice important knowledge. Inspired by this, we integrate ECA blocks into classical DNN with two additional novel modules (See Section 4.3) for the fusion of mmWave and audio signals.

3 CORRELATION MODEL

In this section, we exploit the relationship between voice signals and reflected mmWave signals with a theoretical model. It is fundamental for the fusion of multi-modal signals.

Human voice basically depends on the vocal fold vibration. The vocal vibration process can be regraded as a one-degree-of-freedom damping system [11]. 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,

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

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

Audio signals record human voice without distortion through microphones. Typically, they are considered as a compound of series of single-frequency tones [21, 72] 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 baseband frequencies are equivalent or close to the speed of vocal fold vibration [75]. The relationship can be simplified as

$$v(t) = H(\dot{x}(t)) = H(j\phi_F x(t)), \tag{6}$$

where $H(\cdot)$ is the transfer function from the vocal fold vibration velocity $\dot{x}(t)$ to human voice $v(t)$.

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:

$$r(t) = e^{j(2\pi f_{IF} t+\phi(t))}, \tag{7}$$

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

$$\phi(t) = \frac{4\pi f_m(t)(d + x(t))}{c}, \tag{8}$$

where $f_m(t)$ is the time-variant frequency of mmWave signal, d is the distance between the mmWave radar and the target user, and c is mmWave's speed. 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

$$\begin{aligned} \Delta\phi(t) &= \phi(t + dt) - \phi(t) \\ &= \frac{2\pi}{c}\left(x(t)df_m(t) + f_m(t)dx(t)\right), \end{aligned} \tag{9}$$

where $df_m(t)$ is the frequency shift of FMCW mmWave signals, and $dx(t)$ is the displacement change in vocal fold. Here, dt and $df_m(t)$ are fixed, determined by the mmWave radar's sampling rate and frequency variation rate respectively. Therefore, $\Delta\phi(t)$ depends exclusively on $x(t)$, and we have

$$\Delta\phi(t) = \frac{2\pi}{c}(df_m(t) + f_m(t)j\phi_F)x(t). \tag{10}$$

It indicates that the phase difference of reflected mmWave signals share the identical frequency with the vocal fold displacement.

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 Eq.

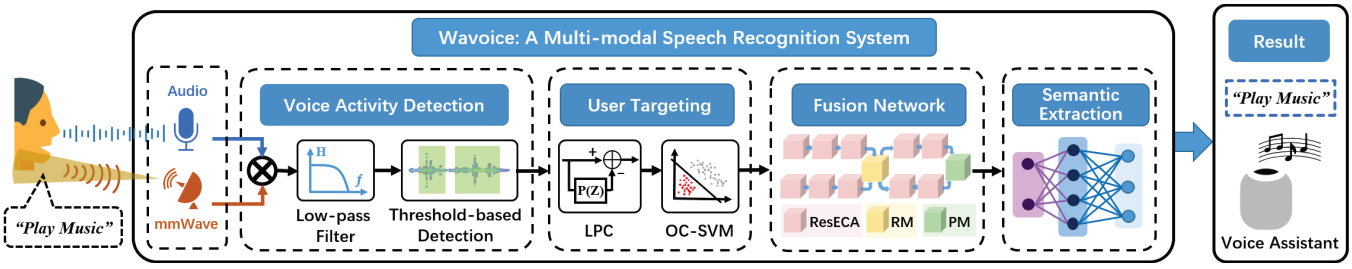


Figure 2: Wavoice, a multi-modal speech recognition system that leverages a mmWave radar and a low-cost microphone to improve the resistance against noise and motion interference in complex environment.

6 and Eq. 10, $v(t)$ owns components whose frequency overlaps or approaches the frequency of $\Delta\phi(t)$. Furthermore, once the transfer function $H(\cdot)$ is determined, we can calculate signals of one modal directly by the other one. In this paper, 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.

4 SYSTEM DESIGN

In this section, we introduce Wavoice, which leverages mmWave and audio signals to recognize the speech under complex conditions. It consists of four modules, i.e., *Voice Activity Detection*, *User Targeting*, *Fusion Network*, and *Semantic Extraction*, as presented in Figure 2.

4.1 Voice Activity Detection

On the basis of the above 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 through the detection assessment.

Motivation. The real-time voice activity detection is a fundamental step for ASR. Without a proper detection mechanism, significant resources would be wasted on dealing with meaningless noise. However, intense noise is likely to cover human voices with an extremely low SNR in public places. Face masks further blur vocal features. Under these circumstances, audio-only voice activity detection would make a wrong judgement and be not responsive to users' commands [27]. Users have to raise their voices or take off their face masks, but this is inconvenient. Fortunately, voice activities are recorded by mmWave and audio signals simultaneously. We can leverage their coherence to amplify the difference between noise and voice activities.

Solution. Wavoice draws the collective characteristic between mmWave and audio signals for the accurate judgement in real time through the coherent demodulation. Wavoice simultaneously receives signals of two modalities. These signals are segmented into 3s 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[52] 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 fast Fourier transform on the sampled mmWave signal. Then the phase difference is

$$\Delta\phi(n) = \phi(n) - \phi(n-1)\ (n \in \mathbb{N}^+). \tag{11}$$

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 low-frequency band after coherent demodulation [14]. We assume $H(\cdot) = 1$ here to illustrate this method's effectiveness ad follows

$$\begin{aligned}\mathbb{F}(n) &= \mathrm{LPF}(v(n) * \Delta\phi(n)) \\ &= \mathrm{LPF}(\frac{2\pi}{c}(df_m(n) + f_m(n)j\phi_F)x^2(n)) \\ &= \frac{2\pi}{c}(df_m(n) + f_m(n)j\phi_F),\end{aligned} \tag{12}$$

where $\mathbb{F}$ is the residual low-frequency component, $\mathrm{LPF}(\cdot)$ is a low-pass frequency filter and the item $\frac{2\pi}{c}(df_m(n) + f_m(n)j\phi_F)$ is a known low-frequency value. When the spectral entropy of $\mathbb{F}$ is larger than a given threshold, vocal vibration is recorded simultaneously by $\Delta\phi(n)$ and $v(n)$ and it 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 noisy environment, Eq. 13 is rewritten as follows,

$$\begin{aligned}\mathbb{F}(n) &= \mathrm{LPF}((v(n) + n_v(n)) * (\Delta\phi(n) + n_\phi(n))) \\ &= \frac{2\pi}{c}(df_m(n) + f_m(n)j\phi_F),\end{aligned} \tag{13}$$

where $n_v(n)$ and $n_\phi(n)$ are the noise on mmWave and audio signals respectively. High-frequency items $x(n)*((\frac{2\pi}{c}(df_m(n)+f_m(n)j\phi_F))*n_v(n)+n_\phi(n))$ 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 very short, i.e., 260s 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 4 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 $\mathbb{F}$ by multiplying the phase difference with the audio signal. As illustrated in Figure 3, $\mathbb{F}$ ranges in the low-frequency band typically within 200Hz, while 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 300Hz and the threshold of spectral entropy is set to 0.835. By comparing the spectral entropy of $\mathbb{F}$ 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 the real-time voice activity detection against various noise interference.

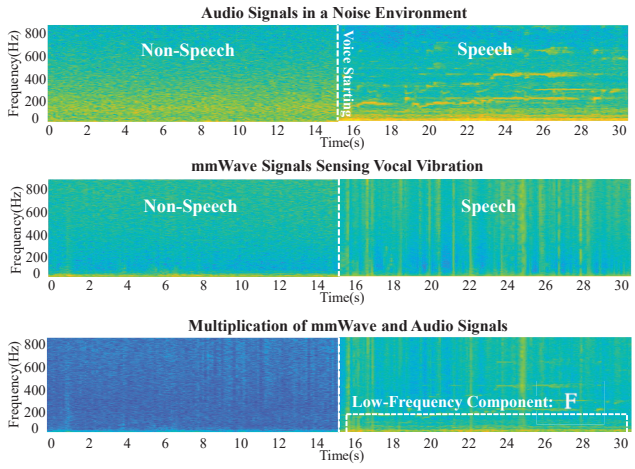


Figure 3: Though audio signals are noisy, the multiplication introduces an additional low-frequency component that results in a sharp distinction between noise and noisy speech.

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 colour the recognition results of ASR. These voice noises are mingled with valid vocal commands, or even cover up them in audio signals recorded by microphones due to the mask effect [14]. 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 people around, while 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 above coherent demodulation between each mmWave signal and audio signals. Non-vocal items are ignored. Afterwards, 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 OC-SVM as follows,

$$\mathbb{F}_i(n) = -\sum_{k=1}^{p} a_i^k \mathbb{F}_i(n-k) + \varepsilon_v(n), \tag{14}$$

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. LPC features of different words have a remarkable difference. Benefiting from this property, we train the OC-SVM with LPC features to identify wake-up words and target users. Similar to the above analysis on noise cancellation, the motion influence on mmWave signals is suppressed. LPC yields high accuracy and robustness with low computational cost.

4.3 Fusion Network

The fusion network comprises residual blocks with ECA (ResECAs), Recalibration Module (RM), and Projection Module (PM) for multimodal 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 filterbank coefficients as network inputs followed by three successive stacked ResECAs. The RM exchanges valid features for mutual recalibration and characteristic enhancement, with recalibrated features flowing into two successive stacked ResECAs. Lastly, PM projects respective information into a joint feature space and adjusts weight coefficients dynamically.

4.3.1 *Log-mel Filterbank Coefficients.*

We extract log-mel filterbank coefficients as network inputs from audio signals and residual voice-related mmWave signals respectively. In detail, we first apply a pre-emphasis filter to the preprocessed audio signal. After pre-emphasis, we perform the short-time

fourier transform (STFT) to measure the time and frequency domain information. The STFT segments the audio signal into frames of 25 ms, with an overlap of 10 ms between successive frames. During segmentation, we need to 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 log-mel filterbank coefficient. The number of coefficients is equivalent to the number of filters. In this paper, the filter bank comprises 40 filters covering the frequency band within 8 kHz.

4.3.2 *ResECA*.

We construct two branches of ResECAs [51] 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 the global average pooling (GAP) [24] 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 dimension, 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

$$z_c = \mathrm{GAP}(x_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j). \tag{15}$$

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

$$A = \sigma(\mathrm{C1D}_k(Z)), \tag{16}$$

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

$$\widetilde{X} = A \odot X, \tag{17}$$

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:

$$Y = \mathcal{C}\left(\mathrm{ECA}(\mathcal{C}(X, W_C)), W_C\right) + X, \tag{18}$$

where the function $\mathcal{C}(*, W_C)$ represents multiple convolution layers to capture features, Y denotes the output of the ResECA, and $\mathrm{ECA}(\cdot)$ represents the ECA block. The operation $\mathcal{C}+X$ 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.

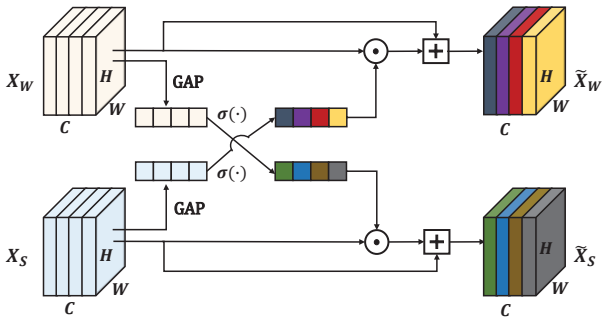


Figure 4: Architecture of Recalibration Module (RM). RM generates recalibrated features by combining original features with these from the other modality.

4.3.3 *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 will first describe the aim of RM and then introduce the mechanism of 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 by several ResECAs. However, stacked ResECAs only provide unimodal features rather than multi-modal features. However, 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. 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, 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 subscript W and S individually represent the mmWave and speech feature. The channel attention map Y_W and Y_S are

$$Y_W = \sigma(W_W \mathrm{ReLU}(\mathrm{GAP}(X_W))), \quad Y_W \in R^{1\times 1\times C}, \tag{19}$$

$$Y_S = \sigma(W_S \mathrm{ReLU}(\mathrm{GAP}(X_S))), \quad Y_S \in R^{1\times 1\times C}, \tag{20}$$

where ReLU is a rectified linear unit (ReLU) function and W indicates learnable parameter matrix. Each stream of channel feature maps is considered as a feature detector and filter. We implement mutual feature recalibration as follows,

$$\widetilde{X}_W = Y_S \odot X_W + X_W, \quad \widetilde{X}_W \in R^{H\times W\times C}, \tag{21}$$

$$\widetilde{X}_S = Y_W \odot X_S + X_S, \quad \widetilde{X}_S \in R^{H\times W\times C}, \tag{22}$$

where $\widetilde{X}_W$ and $\widetilde{X}_S$ are final outputs of RM. Therefore, we obtain the multi-modal features. Aggregating the original feature map

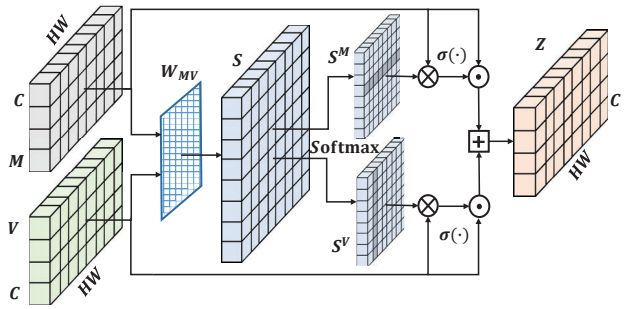


Figure 5: Architecture of Projection Module (PM). PM constructs the similarity matrix based on features of two flattened modalities to learn joint representation.

guarantees that the final output stores enough identical knowledge. The produced multi-modal features embedded in original unimodal features will supply meaningful contexts and suppress useless ones to achieve recalibration. RM can be flexibly placed at different levels in networks to integrate hierarchical features with different spatial dimensions. Here, we place one RM in the middle to fuse mid-level features. It empirically produces comprehensive high-level features through joint recalibration [51].

4.3.4 *Projection Module*.

PM maps features of two modalities into a joint feature space. It finally fuses multi-modal signals for speech recognition.

Motivation. Due to the difference of multi-modal signals, DNN cannot fuse these signals and transform them into semantic information directly. Tradition methods [46, 76] concatenate multiple modalities from different streams directly. They ignore the dynamic distribution of the weight across multi-modal features. Instead, the joint feature [40] using typical methods focuses on all multi-modal features equally, which costs large-scale training data to allow a network to take full advantage of multi-modal features.

Solution. Inspired by the co-attention [40], we create another novel attention-based module to project multi-modal features into a joint feature space. This module called projection module (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. PM constructs the similarity matrix of features of two modalities to measure the correlation between each element of speech and each element of 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 denotes the mmWave feature map from the corresponding branch, and V denotes the speech feature map. We firstly have to flatten M and V into 2D-tensors with height C and width W × 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:

$$S = M^T W_{\text{mv}} V, \quad S \in R^{HW\times HW}, \tag{23}$$

where W_{mv} is a learnable weight matrix. Each column m^i in the flattened matrix M represents a feature vector of C dimension at position $i \in [1, 2, \cdots, HW]$. 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:

$$S^M = \text{softmax}(S), \quad S^M \in R^{HW\times HW}, \tag{24}$$

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

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

$$C^M = V \otimes S^M, \quad C^M \in R^{C\times HW}, \tag{26}$$

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: $C^V = M \otimes S^V$. In order to alleviate the underlying irrelevant interferences, we had better restrict and weigh the knowledge from features of two modalities than cope with all knowledge equally. Therefore, the final fusion result Z is formulated as:

$$Z = W_Z\{\sigma(C^M) \cdot M + \sigma(C^V) \cdot V\}, \quad Z \in R^{C\times HW}, \tag{27}$$

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.

4.4 Semantic Extraction

We utilize the typical speech-to-text translation system [64, 77] to build the semantic extraction architecture. We choose *Listen, Attend, and Spell (LAS)* [7], 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 [49].

LAS is composed of two components: an encoder called listener and a decoder called speller [7]. The listener maps 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 context vectors. Context vectors, distribution of characters, and decode states are all fed into the RNNs for the decoder state. The posterior distribution is computed based on the decoder state and context vector via a softmax function [49]. LAS is trained to maximize the logarithmic posterior probability of the correct character sequence.

Here, we stack two pBLSTM layers as the listener while 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 under various condition.

5.1 Setup

Hardware. We implement our system on a low-cost microphone [15], a COTS IWR1642BOOST radar [60] equipped with a data collection board DCA1000EVM [59], and a laptop, as shown in Figure 6. The IWR1642BOOST equipped with DCA1000EVM is a 77 GHz mmWave radar that transmits FMCW continuously in order 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 degree, an azimuth resolution of 15 degree, and a high-resolution elevation view of 30 degree. 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 in our experiments to transmit a chirp with 260μs cycle time. The received channel has a 5000k 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 300μm.

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

Software. We connect and control the radar with mmWaveStudio GUI [61] running in the laptop. The mmWaveStudio GUI configures the radar parameters as described above. 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 [18] and Google speech commands [68] that involve common voice commands words in all aspects. All 20 participants, including ten females and ten females, whose ages range from 16 to 47, speak all commands in their normal speech speed and volume, typically 65 dB sound pressure level (db-SPL) [53]. We place the mmWave radar and microphone at a distance of seven meter from the subject. We align the mmWave radar to the subject and guarantee the mouth and neck of subjects 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 32000 pairs of samples (i.e., the mmWave and audio signal) for each situation. We randomly choose the sample from two males and two females as the test dataset. We thereby have 25600 training data and 6400 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 scenes include an office room, a roadside, a cafe, and an in-vehicle. Note that we explicitly tell the participants about the purpose of our experiments. Our research is approved by IRB: ZJU2021-6.

Figure 6: Experimental setup. A mmWave radar and a microphone receive signals from subjects sitting 7 meters away.

5.2 Metrics and Baseline

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

Character Error Rate (CER). 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 [78] as follows,

$$\text{CER} = \frac{I_c + S_c + D_c}{N_c}, \tag{28}$$

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 the better speech performance of the ASR system.

Word Error Rate (WER). 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,

$$\text{WER} = \frac{I_w + S_w + D_w}{N_w}, \tag{29}$$

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

Baseline. We select DeepSpeech2 (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 the 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. (3) We train and test a DS2 model totally on our datasets. We observe DS2's CERs are respectively 90.60%, 71.22% and 34.46%. Therefore, we construct the baseline results by implementing DS2 under the third condition.

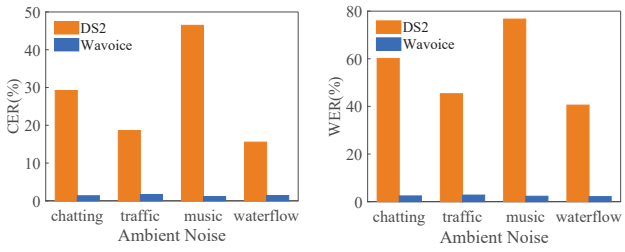


Figure 7: Performance of Wavoice and DeepSpeech2 (DS2) under various ambient noises.

5.3 Overall Performance

We evaluate the overall performance of Wavoice when users are in different states. Three in-lab experiments are conducted to assess whether our multi-modal system can show excellent speech recognition capacity over the standard ASR system. The following factors: (1) Ambient Noise, (2) Mask, and (3) Multi-Person are considered respectively in the three experiments.

5.3.1 *Ambient Noise*.

Ambient noise reduces SNRs of received voice commands and interferes with the recognition accuracy. We evaluate the speech recognition performance of Wavoice under four types of noise conditions, i.e., chatting, traffic, music, and waterflow. When subjects speak required voice commands, four loudspeakers play noises with 60 db-SPL at 40 cm from the microphone of Wavoice, with SNR of recorded audio signals within 5dB. Figure 7 presents the performance of Wavoice and DS2 under different noise interference. DS2 obtains the low recognition accuracy with the average CER above 20% and the average WER above 40%. The background noise explicitly degrades the speech recognition accuracy of traditional ASR systems.This is because audio-only methods like DS2 are sensitive to unpredictable and unknown noise. On the contrary, Wavoice yields superior performance with the average CER within 1% and the average WER about 2.5%. Even in the worst case (i.e. traffic noise interference), Wavoice maintains WER about merely 3% and CER below 1.5%. With all comparisons and observations above, we conclude that Wavoice is extremely stable and effective against ambient noise.

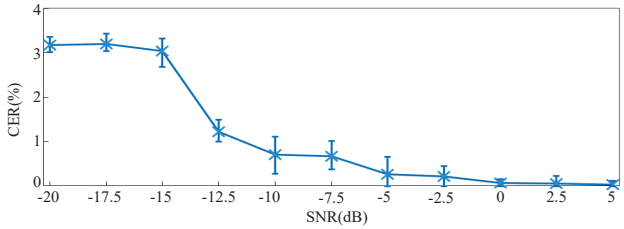


Figure 8: Performance of Wavoice under various SNRs of audio signals.

We further investigate the speech recognition capability of Wavoice under different SNR conditions. We adjust the source intensity of noise here to modify SNR from -20 dB to 5 dB, with Wavoice's CER shown in Figure 8. When the SNR is above 0 dB, Wavoice maintains a tiny error rate that is nearly constant as SNR changes. When the SNR is above -10 dB, the CRE increases a little but keeps lower than 1 %. Wavoice can extract and fuse semantic information from mmWave and audio signals to achieve noise-resistant multi-modal speech recognition. We observe that the CER of speech recognition is stable as SNRs decrease under -15 dB. This is because acoustic information in audio signals vanishes in adversely low SNRs, leading to the convergence of the multi-modal system. In this case, the performance of Wavoice depends on mmWave radar merely. In short, Wavoice obtains an accurate and noise-resistant speech recognition by fusing mmWave and audio signals.

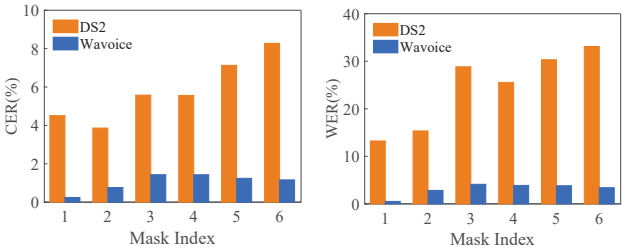


Figure 9: Performance of Wavoice and DeepSpeech2 (DS2) influenced by masks without noise.

5.3.2 *Mask*.

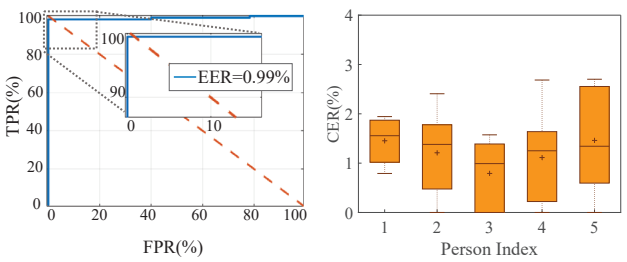
We study the speech recognition capacity of Wavoice when users wear face-masks and speak voice commands. We select some typical masks: disposable medical masks, N95 respirator masks, gas masks, and anti-dust masks. A series of experiments are conducted where participants put on a given mask and speak words. To further measure the proposed system's penetration, we additionally require the subject to wear a scarf in one experiment. All selected masks and their indexes are listed in Table 2. The speech recognition results in Figure 9 show that diverse masks worn by subjects degrade the acoustic properties and voice quality in different extent. The speech recognition capacity of DP2 is dramatically impacted by mask conditions, particularly when the air tightness of the mask is relatively high. The results in Figure 9 show that our system consistently outperforms the baseline when the subject wears different masks. We observe that the CER of Wavoice is all nearly 1% while the baseline is mostly above 5%, which confirms our system's effectiveness against acoustic degradation caused by masks. Through the comparison, we validate that the fusion of mmWave and audio signals can significantly enhance the speech recognition performance regardless of mask conditions.

Table 2: 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

5.3.3 *Multi-Person Scene*.

In a multi-person scene, the radar in the system tends to receive various reflected signals from people around. Those reflected signals contain components unrelated to the user's vocal vibration. To investigate the effectiveness of the proposed targeting module, we further conducted experiments in a multi-person scene. We asked each of five subjects to take turns as the target user and the other four subjects walked around and spoke freely in the meantime.



(a) The ROC curve in identifying user. (b) The CER of five subjects.

Figure 10: Performance in a multi-person scene.

Except for speaking voice commands, each subject acting as the user is requested to say the wake-up word 30 times for training the classifier and 10 times for testing. The wake-up word is set to "Wavoice". We thus collect the positive sample (i.e., the mmWave and audio signal related to the wake-up word) and the negative sample related to other utterances. To verify the performance of targeting the user, we preprocess the sample to produce the LPC feature and then use testing samples to examine the trained classifier. We derive the receiver operating characteristic (ROC) curve as shown in Figure 10(a). We observe that the user targeting module achieves more than 98.8% true positive rate (TPR) and less than 1.1% false positive rate (FPR) with an equal error rate (EER) of 0.99%, which confirms the effectiveness of targeting the user in a multi-person scene. Moreover, we evaluate the performance of speech recognition under multi-person conditions. We measure the CER of the system on recognizing speech as shown in Figure 10(b). By averaging the recognition result of commands from five subjects, we get an overall speech recognition accuracy of 1.2%. The results in Figure 10(b) demonstrate that the system is highly effective against interference from people around.

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 [48] between two outputs from the two modified networks above, i.e., **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 which 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 [50]. Notably, Wav2Letter, a structured-output learning approach based on a variant of CTC, has an outstanding performance on noisy speech [50]. 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 3.

Table 3: 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 [48]	10.78	48.20	5.37	20.21
W/O Fusion	12.71	35.38	6.43	29.20
DS2 [5]	41.12	72.70	7.13	30.32
Wav2Letter [50]	22.17	46.28	4.72	12.23
W/O ResECA	2.43	4.41	1.78	3.35
W/O RM	4.53	8.82	4.21	9.24
W/O PM	4.08	7.65	3.16	5.882
`Wavoice`	0.69	1.72	0.76	1.65

As shown in Table 3, audio-only methods (i.e., Speech-only, DS2, and Wav2Letter) present high CERs and WERs, especially in dealing with noisy speech. Therefore, we speculate that unpredictable ambient noise impedes the performance of audio-only methods. The mmWave-only method struggles in providing reliable results, attributed to its susceptibility to varying multipath noise and relatively coarse-grained perception. However, Voting and W/O Fusion yield slightly better results over the baseline with merely 10.78% CER and 12.71% CER in noise, which verifies that ignoring the correlation and collaboration between mmWave and audio signals is unable to exploit two modalities signals for utmost performance in speech recognition. Meanwhile, our fusion modules improve W/O Fusion by over 12% and 5% in terms of CER under different conditions. Our system with fusion modules is superior to the Voting by 10% and 4% in terms of CER in two different environments, respectively. Besides, Wavoice outperforms WaveEar[75] whose WER is mostly more than 4% , especially under motion interference. Moreover, we conduct an ablation study by considering the different proposed modules of the fusion network. The comparison result shows that every module we propose plays a vital role in speech recognition performance. In summary, our system with fusion modules outperforms the aforementioned methods. These experiments indicate that our proposed fusion modules adequately utilize the correlation between two modalities signals.

5.5 Robustness Analysis

We further analyze the robustness of `Wavoice` under the influence of different distance and orientation, body motion, and environmental disturbance. Note that the sensing distance of the radar and

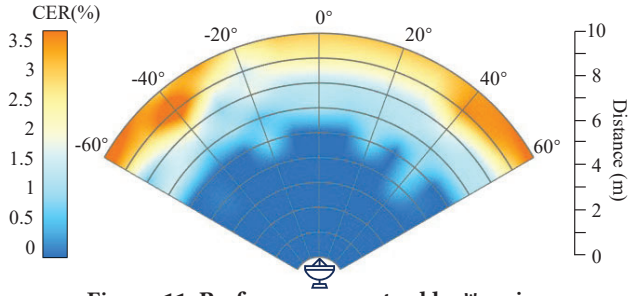


Figure 11: Performance centred by Wavoice.

microphone is still 7 meters in the body motion and environmental disturbance circumstances.

5.5.1 *Distance and Orientation.*

We compare the performance of Wavoice when the user is located at different distances and orientations of the mmWave radar. In this experiment, the sensors, including the mmWave radar and microphone, are set at different distances (from 1 m to 10 m) and different orientations (from -60° to 60°) to the subjects' mouth and throat. The overall results are shown in Figure 11. When the distance is larger than 9 m, the CER slightly increases as the distance increases. This is because when the energy of speech decays rapidly, especially at exceedingly long distances, the microphone thereby captures the raw speech from subjects. As for the orientation, speech recognition results are less than 1.5% in all orientations when the distance is less than 9 meters. Our system's speech recognition performance is relatively stable and excellent as the orientation changes. We envision that recorded omnidirectional signals by the microphone are exploited to recalibrate and enhance coarse-grained mmWave features in the proposed fusion network. Our system can support flexible and convenient speech recognition even though the user is in a remote location.

5.5.2 *Body Motion.*

We evaluate the robustness of Wavoice when users are in body motion. We request five subjects to speak commands and perform body motions, including making telephone calls, typing on phones, shaking arms, and marching on the spot. We test the CER of five subjects across different body motions and the corresponding results in Figure 12. As shown in Figure 12, the average CERs are 0.33% and 0.37% in the calling and typing smartphone, respectively, while CERs are slightly high but are mostly less than 1% in other body motions. The results further prove that Wavoice is robust to the common body motion. When the user is in motion such as march, the directly extracted phase across multiple FFT bins is mixed up with motion interference. However, acoustic signals are fused to recalibrate mmWave features and compensate for the loss of information in the proposed system. Thus, motion interference has a limited impact on the performance of the system.

5.5.3 *Environmental Disturbance.*

Since our experiments are set in our controlled laboratory environment, we also evaluate the system in the real-world environment. We conduct experiments on four types of scenes: a filled office, a noisy cafe, a busy roadside, and a subway. We request five participants to speak voice commands naturally and comfortably. The collected data is fed into the pre-trained Wavoice modal to verify

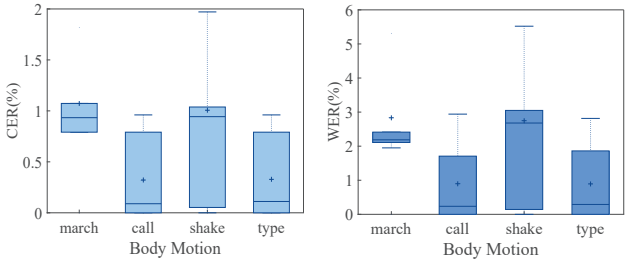


Figure 12: Performance under the body motion influence.

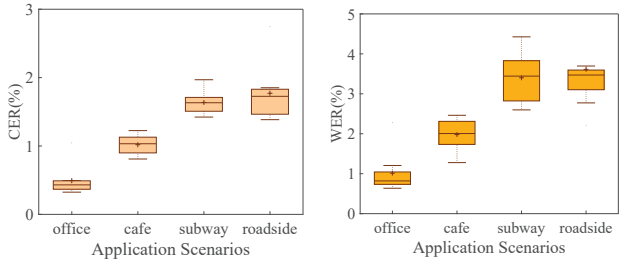
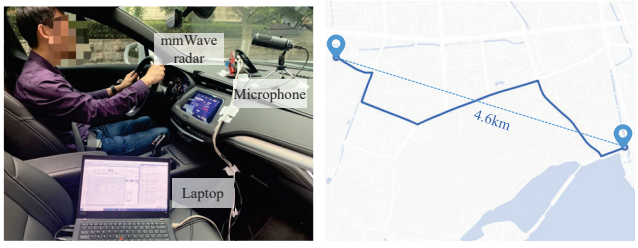


Figure 13: Performance under environmental disturbance.

the universality. The results of speech recognition are shown in Figure 13. The average CERs are 0.49%, 1.02%, 1.64%, and 1.77%, respectively. Although the accuracy of speech recognition is slightly degraded, the results in Figure 13 demonstrate the universality of Wavoice in arbitrary realistic scenes.

We also study the generalization of Wavoice in a vehicle, where the mmWave radar tends to wobble during driving. three subjects are asked to speak commands as driving a vehicle. The mmWave radar and microphone are appropriately placed on the automotive center stack, which does not affect the subject's driving. Figure 14(a) shows the experimental setup. Each subject drives 20 minutes following the route shown in Figure 14(b) at the normal speed in the urban area. To fully validate the generalization, when the subject speaks commands, we play music in the vehicle during driving. After attaining the two modalities signals during driving, we examine the speech recognition capacity of Wavoice.



(a) In-vehicle setup. (b) The driving route.

Figure 14: The setup in a vehicle for collecting data and the corresponding driving route.

As we can see in Figure 15(a), the average CER stays below 0.5% as the driving distance rangs from 0 to 4 km. Figure 15(b) shows that the CER of three subjects are 0.45%, 0.20%, and 0.30% respectively, which indicates that the average CER is 0.32% in one hour of driving time. The results demonstrate that our system is competent for speech recognition in vehicles regardless of the wobble of mmWave radar. This is reasonable because the mmWave

radar and microphone receive enough useful signals in a narrow space to generate fusion features for accurate speech recognition.

6 DISCUSSION

Hardware Support. Compared with traditional ASR systems [4, 17], Wavoice requires an additional mmWave radar but merely one microphone. Nevertheless, mmWave radars diffuse rapidly with the development of mmWave technologies in wireless sensing [54] and 5G communication [66]. For example, Pixel 4 [16] has carried the miniature mmWave radar for man-machine interaction. Furthermore, various ambient noises requires excessive microphones. Under a specific layout, microphone arrays demand a lager volume but obtain a small coverage. In this case, it is foreseeable that Wavoice will be applied on voice-controlled devices for speech recognition in various scenes.

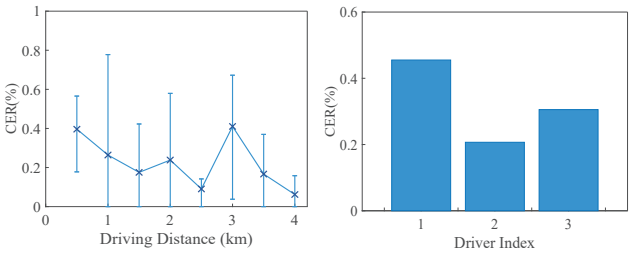
Sensing Range. It has been demonstrated that Wavoice has a range coverage of 10 meters with 120° field-of-view. It can deal with most applications where users face sensors within a certain deviation, such as voice-controlled elevators and ATMs. As for applications in the fully open space, such as smart streetlights, it requires at least three radars for a 360° coverage but increases costs. A possible way is to rotate the mmWave radar with the aid of user targeting, which has been applied in commercial wireless chargers [74]. Furthermore, we can employ a microphone array rather than a single one for the further sensing range extension.

Cost and Power Consumption. Wavoice requires a mmWave radar and a low-cost microphone. Here, a mmWave radar chip costs about 40 dollars [60] and a microphone costs about 10 cents. Considering the long sensing distance and the resistance against noise, Wavoice is more affordable than the high-cost directional microphone array, at an average price of around 50 dollars. Furthermore, the cost of mmWave radars will reduce as its popularity and mass production. As for the power Consumption, both mmWave radars and microphones perform well. Their power consumption both keep below than 20mW, which is acceptable for most VUIs.

Speech Separation. Speech Separation is the task of separating and recovering the target speech from background interference such as the cocktail party effect. Due to the benefit of speech separation to VUIs, it is worth extending Wavoice to separate speech from noisy signals. Motivated by the research on deep complex networks [10, 62], Wavoice has the potential to achieve speech separation. Due to the flexibility of Wavoice, the complex network can replace our semantic extraction network in the system to predict the magnitude and phase spectrogram of target speech. Then, the original speech can be estimated by performing the inverse Fourier transform on the estimated spectrogram.

7 RELATED WORK

mmWave-based Sensing benefits high precision sensing in complex environments [8, 38, 39]. Chang et al. [8] leveraged a spatial attention fusion method for obstacle detection by integrating data from mmWave radar and vision sensor. The mmWave-vision fusion can improve resolution and expand measuring ranges [12, 37, 79]. The mmWave radar also has comprehensive cooperation with IMU to estimate ego-motion [3, 39]. Furthermore, Lu et al. [38] reconstructed an indoor grid map with a mmWave radar and a lidar. Similar works displayed that the mmWave-lidar collaboration benefits system stability [29, 73].



(a) CERs varies in driving distance. (b) CERs of three drivers.

Figure 15: Performance of Wavoice's in-vehicle application.

Speech Enhancement aims at improving the quality and intelligibility of degraded speech in adverse listening conditions with the aid of microphone arrays [6, 20, 28, 36, 45, 69]. Classic statistic-based methods [21, 72] require prior knowledge about noise characteristics. Learning-based techniques have gained in popularity which leverage DNN [46, 76] or generative adversarial networks (GANs) [13, 47] but fail in long-distance speech recognition.These techniques demand excessive microphones (more than the number of noise sources) and a particular layout. These requirements may lead to a too large volume to be integrated into public application.

Cross-modal Speech Recognition provides a new idea against noise interference. Audio-visual means detect lip motion [1] or face landmarks [43], while ultrasound-assisted techniques [30, 56] measure vocal vibration to extract target speeches. Moreover, WiFi signals [65] and inertial signals [2] can recover semantic information. Different from existing work, we fuse mmWave and audio signals through the improved network with SENet-based inter-attention. Wavoice supports long-distance speech cognition (up to 7 meters) in public places full of noise and motion interference.

8 CONCLUSION

In this paper, we employ a mmWave radar and a microphone for long-distance, noise-resistant, and motion-robust speech recognition. We formulate the correlation between mmWave and speech signals. Benefiting from this correlation, we propose a voice activity detection method against noise interference and a user targeting mean to avoid overlaps with non-target users. Two novel modules are introduced into an attention-based network based on the inter-attention between multi-modal signals. Here, mmWave signal improves recognition accuracy despite ambient noise or face masks, while audio signals rectify errors caused by motions. Wavoice maintains a low error rate within 1% and its range reaches up to 7 meters. It provides a comprehensive solution to public applications of VUIs.

9 ACKNOWLEDGMENTS

This paper is supported by the National Key R&D Program of China (Grant No. 2020AAA0107700), National Natural Science Foundation of China (Grant No. 62032021, 61972348, 61772236, and 61872285), Zhejiang Key R&D Plan (Grant No. 2019C03133), Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (Grant No. 2018R01005), Research Institute of Cyberspace Governance in Zhejiang University, Alibaba-Zhejiang University Joint Institute of Frontier Technologies.

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 (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.com. 2021. https://www.amazon.com/echo/, title = Amazon echo.

[5] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*. PMLR, 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] 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 International Conference on Acoustics, Speech and Signal Processing*.

[8] 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.

[9] 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.

[10] 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 International Conference on Learning Representations*.

[11] Livija Cveticanin. 2012. Review on Mathematical and Mechanical Models of the Vocal Cord. *J. Appl. Math.* 2012 (2012), 928591:1–928591:18.

[12] Joseph St Cyr, Joshua Vanderpool, Yu Chen, and Xiaohua Li. 2020. HODET: Hybrid object detection and tracking using mmWave radar and visual sensors. In *Proceedings of the Sensors and Systems for Space Applications XIII*, Vol. 11422. International Society for Optics and Photonics, 114220I.

[13] 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*.

[14] 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*. ACM.

[15] Gmtd. 2021. GM-A906 microphone. [online].

[16] Google. 2019. https://www.androidcentral.com/how-does-googles-soli-chip-work, title = Here's how the Pixel 4's Soli radar works and why Motion Sense has so much potential,.

[17] Google. 2021. https://store.google.com/product/google_home/, title = Google home.

[18] Google. 2021. ok-google.io. https://ok-google.io/

[19] Nishu Gupta et al. 2021. A Novel Voice Controlled Robotic Vehicle For Smart City Applications. In *Journal of Physics: Conference Series*, Vol. 1817. IOP Publishing, 012016.

[20] Mary Harper. 2015. The automatic speech recogition in reverberant environments (ASpIRE) challenge. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*. 547–554.

[21] 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*.

[22] 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.

[23] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. 2019. Cross Attention Network for Few-shot Classification. In *Proceedings of Advances in Neural Information Processing Systems*.

[24] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. In *Proceedings of Conference on Computer Vision and Pattern Recognition*.

[25] Apple Inc. 2021. https://www.apple.com/au/siri/, title = Siri - Apple.

[26] 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.

[27] Kaustubh Kalgaonkar, Rongquiang Hu, and Bhiksha Raj. 2007. Ultrasonic doppler sensor for voice activity detection. *IEEE Signal Processing Letters* 14, 10 (2007), 754–757.

[28] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël AP Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, 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.

[29] Aldebaro Klautau, Nuria González-Prelcic, and Robert W Heath. 2019. LIDAR data for deep learning-based mmWave beam-selection. *IEEE Wireless Communications Letters* 8, 3 (2019), 909–912.

[30] Ki-Seung Lee. 2019. Speech enhancement using ultrasonic doppler sonar. *Speech Communication* 110 (2019), 21–32.

[31] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, 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*.

[32] Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[33] 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.

[34] 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.

[35] 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 IEEE Symposium on Security and Privacy*.

[36] Philipos C Loizou. 2013. *Speech enhancement: theory and practice*. CRC press.

[37] 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 *Proceedings of the Millimetre Wave and Terahertz Sensors and Technology XI*, Vol. 10800. 1080006.

[38] 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*.

[39] Chris Xiaoxuan Lu, Muhamad Risqi U Saputra, Peijun Zhao, Yasin Almalioglu, Pedro PB 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*.

[40] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *Proceedings of Advances in Neural Information Processing Systems*.

[41] 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*.

[42] Youri Maryn, Floris L Wuyts, and Andrzej Zarowski. 2021. Are Acoustic Markers of Voice and Speech Signals Affected by Nose-and-Mouth-Covering Respiratory Protective Masks? *Journal of Voice* (2021).

[43] 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*.

[44] Tomer Moscovich. 2009. Contact Area Interaction with Sliding Widgets. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*.

[45] 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).

[46] 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.

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

[48] Lionel Sharples Penrose. 1946. The Elementary Statistics of Majority Voting. *Journal of the Royal Statistical Society* 109, 1 (1946), 53–57.

[49] 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*.

[50] Vineel Pratap, Awni Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky, and Ronan Collobert. [n.d.]. Wav2Letter++: A Fast Open-source Speech Recognition System. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP, year = 2019,*.

[51] Pengfei Zhu Peihua Li Wangmeng Zuo Qilong Wang, Banggu Wu and Qinghua Hu. 2020. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural

Networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.
[52] Hermann Rohling and Ralph Mende. 1996. OS CFAR performance in a 77 GHz radar sensor for car application. In *Proceedings of International Radar Conference*. IEEE, 109–114.
[53] 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.
[54] Kei Sakaguchi, Thomas Haustein, Sergio Barbarossa, Emilio Calvanese Strinati, Antonio Clemente, Giuseppe Destino, Aarno Pärssinen, Ilgyu Kim, Heesang Chung, Junhyeong Kim, et al. 2017. Where, when, and how mmWave is used in 5G and beyond. *IEICE Transactions on Electronics* 100, 10 (2017), 790–808.
[55] SoundAI. 2020. https://www.chinadaily.com.cn/a/202003/13/WS5e6b3fcca31012821727ef88.html, urldate = March 13, 2020, title = Voice-controlled Elevator System Put into Use in Beijing.
[56] 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*.
[57] Lorenzo Tarantino, Philip N. Garner, and Alexandros Lazaridis. 2019. Self-Attention for Speech Emotion Recognition. In *Proceedings of the Conference of the International Speech Communication Association*.
[58] Telsa. 2021. https://www.tesla.com/, title = Model S/3/X/Y.
[59] TI. 2021. DCA1000EVM. https://www.ti.com/tool/DCA1000EVM.
[60] TI. 2021. IWR1642. https://www.ti.com/tool/IWR1642BOOST.
[61] TI. 2021. mmWave Studio. https://www.ti.com/tool/MMWAVE-STUDIO.
[62] 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*.
[63] 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 *Proceedings of Advances in Neural Information Processing Systems*.
[64] Dong Wang, Xiaodong Wang, and Shaohe Lv. 2019. An overview of end-to-end automatic speech recognition. *Symmetry* 11, 8 (2019), 1018.
[65] 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.
[66] 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.
[67] 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.
[68] Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209* (2018).
[69] 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).
[70] Canalys website. 2020. https://www.canalys.com/newsroom/canalys-global-smart-speaker-market-2021-forecast, urldate = October 22, 2020, title = Global Smart Speaker Market 2021 Forecast.
[71] Chris Wiltz. 2020. https://www.designnews.com/design-hardware-software/covid-19-giving-touchless-interfaces-chance-make-impression-0, urldate = June 03, 2020, title = COVID-19 Giving Touchless Interfaces a Chance to Make an Impression.
[72] 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.
[73] Henk Wymeersch, Gonzalo Seco-Granados, Giuseppe Destino, Davide Dardari, and Fredrik Tufvesson. 2017. 5G mmWave positioning for vehicular networks. *IEEE Wireless Communications* 24, 6 (2017), 80–86.
[74] Xiaomi. [n.d.]. 'Not science fiction': Xiaomi's revolutionary new wireless charging tech can charge your devices remotely. https://www.financialexpress.com/industry/technology/not-science-fiction-xiaomis-revolutionary-new-wireless-charging-tech-can-charge-your-devices-remotely/2181661/.
[75] 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.
[76] 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*.
[77] Dong Yu and Li Deng. 2016. *AUTOMATIC SPEECH RECOGNITION*. Springer.
[78] Li Yujian and Liu Bo. 2007. A normalized Levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence* 29, 6 (2007), 1091–1095.
[79] Renyuan Zhang and Siyang Cao. 2019. Extending reliability of mmwave radar tracking and detection via fusion with camera. *IEEE Access* 7 (2019), 137065–137079.