mmMIC: Multi-modal Speech Recognition based on mmWave Radar

Long Fan, Lei Xie, Xinran Lu, Yi Li, Chuyu Wang, Sanglu Lu
State Key Laboratory for Novel Software Technology, Nanjing University, China
fanl@smail.nju.edu.cn, lxie@nju.edu.cn, {luxinran,yili}@smail.nju.edu.cn, {chuyu,sanglu}@nju.edu.cn

IEEE INFOCOM 2023 - IEEE Conference on Computer Communications | 979-8-3503-3414-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/INFOCOM53939.2023.10229085

***Abstract*—With the proliferation of voice assistants, microphone-based speech recognition technology usually cannot achieve good performance in the situation of multiple sound sources and ambient noises. In this paper, we propose a novel mmWave-based solution to perform speech recognition to tackle the issues of multiple sound sources and ambient noises, by precisely extracting the multi-modal features from lip motion and vocal-cords vibration from the single channel of mmWave. We propose a difference-based method for feature extraction of lip motion to suppress the dynamic interference from body motion and head motion. We propose a speech detection method based on cross-validation of lip motion and vocal-cords vibration so as to avoid wasting computing resources on nonspeaking activities. We propose a multi-modal fusion framework for speech recognition by fusing the signal features from lip motion and vocal-cords vibration with the attention mechanism. We implemented a prototype system and evaluated the performance in real test-beds. Experiment results show that the average speech recognition accuracy is 92.8% in realistic environments.**

I. Introduction

Nowadays, with the practical use of speech recognition technology, voice assistants have been widely used in various application scenarios to bring more convenience to our lives. In particular, speech recognition in voice assistants is used to improve the efficiency of human-computer interaction in smart driving [1] [2], smart home [3], and smart medical care [4]. For example, for intelligent meeting minutes, voice assistants can be deployed in the meeting room to convert human voice to text. Moreover, for safe driving, voice assistants can also be used in intelligent driving to recognize the instructions from the drivers without manual touch. Current speech recognition technology in voice assistants is mainly based on a microphone to collect voice signals from human subjects. The microphone-based speech recognition [5] [6] works well in situations without other voice interference and environmental noises. However, in the situation of multiple sound sources and ambient noises, the speech recognition performance based on a microphone decreases dramatically. For example, when multiple passengers in the car are speaking simultaneously, multiple voices mix together in the voice assistant, which prevents the driver from effectively interacting with voice assistants. Therefore, new approaches are essentially required to collect voice-related signals from the human subject so as to ensure the performance of speech recognition.

There are two main approaches for speech recognition technology to improve speech recognition performance in the situation of multiple sound sources and ambient noises. *The first approach is based on single-sensor speech recognition*. It uses a single sensor, such as a wireless antenna or camera, to collect voice-related signals. Considering the wide coverage of wireless signals, researchers leverage the wireless signals such as WiFi [7]–[10], RFID [11]–[14], and mmWave [15]–[18] to perform speech recognition. Either the lip motion or the sound vibration is collected via a wireless channel for speech recognition. Besides, the camera-based solutions [19] are also investigated to collect the sound vibrations according to ultra-high-frame-rate video streams. However, the single-sensor-based approach usually fails to achieve good performance in speech recognition since they usually cannot recover the voice-related signals from the single channel well. *The second approach is based on multi-modal fusion from multiple sensors for speech enhancement and recognition*. This approach uses multiple sensors with different modalities, such as audio, video, and wireless signals to collect the voice-related signals simultaneously. The complementarity among different modalities is investigated and leveraged to enhance speech recognition performance. However, the existing multi-modal fusion-based approaches either have privacy issues, e.g., the audio-visual fusion [20] [21], or have limitations in sensing range, e.g., the Ultrasound [22] [23]. Moreover, the multiple sensor-based approaches incur additional hardware costs and require essential synchronization in both time and space.

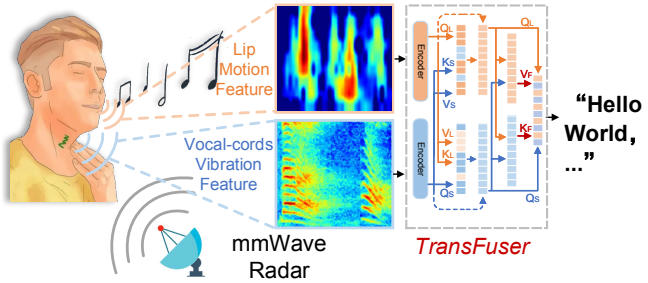


Fig. 1. Speech recognition based on single-channel multi-modal fusion

In this paper, we propose a novel speech recognition solution called *mmMIC* based on mmWave, by extracting voice-related signal features from multiple modalities, i.e., lip motion and vocal-cords vibration, as shown in Fig.1. Different from the previous approaches, *mmMIC* extracts the features of the above two modalities from a single channel, i.e., mmWave. We use mmWave radar to transmit frequency-modulated continuous waves (FMCW) to focus on the region of the lip and vocal cords. To extract the features from lip motion, we propose a feature extraction scheme for *macro-motion* based

on *Doppler velocity* to represent the lip motion. Moreover, since the reflected signals from lip motion usually include other dynamic interference, e.g., the signals from body motion and head motion, we propose a difference-based method to suppress the dynamic interference from body motion and head motion. To extract the features from vocal-cords vibration, we propose a feature extraction scheme for *micro-vibration* based on *frequency-time spectrogram* to represent the vocal-cords vibration. Moreover, to efficiently detect the speaking activity of the human subject, we propose a speech detection method based on cross-validation of lip motion and vocal-cords vibration so as to avoid wasting computing resources on nonspeaking activities. To effectively fuse the features from the two modalities for speech recognition, we propose *TransFuser*, a multi-modal fusion transformer that leverages the attention mechanism to fuse the spectrograms of lip motion and vocal-cords vibration. In this way, the mutually complementary features can be leveraged to further improve speech recognition performance.

There are three challenges to be addressed in this paper. *The first challenge is to extract the signal features from lip motion with the dynamic interference from the body motion and head motion.* Due to the dynamic interference from body and head motion, the reflected signals from lip motion are severely suppressed by the interference signals, since the amplitudes of body motion and head motion are usually much greater than the lip motion. To address this challenge, we propose a difference-based method to remove the dynamic interference. Specifically, according to the range bins of mmWave, we extract the signals from the adjacent bins next to the lip motion-related bin, and then average the signals in the adjacent bins to estimate the dynamic interference to the lip motion. After that, we obtain the lip motion-related spectrogram by canceling the averaged signals in the original spectrogram.

The second challenge is to effectively distinguish the speaking activities from nonspeaking activities so as to avoid unnecessary wastes of computing resources on the nonspeaking activities. Considering that there always exist nonspeaking activities in real scenarios, e.g., the human subject is eating or chewing with obvious lip motion and minor vocal-cords vibration. If these signals of nonspeaking activities cannot be effectively distinguished, much computing resource can be wasted in feature extraction/fusion and speech recognition. To address this challenge, we propose a cross-validation-based speech detection method. We first use lip motion-related signals for pre-detection to determine whether there exists a possibility of the speaking activity. Then, we further verify the vocal-cords vibration-related signals from the pre-detection results. Considering that when the human subject is actually speaking, the vowels usually lead to obvious vocal-cords vibration, and the vowels and consonants usually appear alternately, thus we can distinguish the speaking and nonspeaking activities by further verifying the amplitudes of vocal-cords vibration, especially from pronouncing the vowels.

The third challenge is to fuse multi-modal features to improve speech recognition performance. While we use mmWave as a single channel to perceive the lip motion and vocal-cords vibration signals simultaneously, there exist inherent correlation and complementarity in the features of the two modalities. For example, both features are temporally and spatially synchronized, and lip motion features can effectively compensate for the vocal-cords vibration features when the latter is not so obvious. To effectively fuse the corresponding features to improve speech recognition performance, it is challenging to explore the inherent correlation and complementarity between the two modalities. To address this challenge, we propose *Transfuser*, a multi-modal fusion framework that uses the attention mechanism, i.e., *cross-attention* and *merged-attention*, to fuse the Doppler-velocity spectrogram and vocal-cords vibration spectrogram. In this way, the inherent correlation and complementarity can be effectively quantified by leveraging the attention mechanism.

This paper makes the following contributions. First, we propose a novel mmWave-based solution to perform speech recognition in the situation of multiple sound sources and ambient noises, by precisely extracting the multi-modal features from lip motion and vocal-cords vibration from the single channel of mmWave. Second, we propose a multi-modal fusion framework for speech recognition by fusing the signal features from lip motion and vocal-cords vibration with the attention mechanism. Third, we implemented a prototype system and evaluated the performance in real test-beds, including multiple sound sources and ambient noises. Experiment results show that the average speech recognition accuracy is 92.8% in realistic environments.

II. Related Work

Speech recognition is widely used in real-world scenarios. Speech recognition based on one single sensor is called single-channel speech recognition, while speech recognition based on multiple sensors is called multi-model speech recognition.

Single-channel Speech Recognition. Microphone-based automatic speech recognition(ASR) system is widely used in our lives [24] [25]. The microphone is the most commonly used but not suitable for all scenarios, for example, through-wall perception. To solve the problem, some researchers propose other channels for ASR. The wireless perception technology obtains the characteristics of the signals channel by analyzing the changes of the wireless signals during the propagation process to realize the scene's perception. Wang et al. [7] propose a Wi-Fi signal-aware-based lip reading by detecting and analyzing fine-grained radio reflections from mouth motion. Li et al. [15] proposed a flexible mmWave interrogation system that directly captures and analyzes sound vibrations for user authentication. Wang et al. [11] proposed Tag-Bug, which focuses on human voices with complex frequency bands and eavesdrops on speakers through walls by capturing sub-millimeter vibrations. However, the single-sensor-based approach usually fails to achieve good performance in speech recognition since they usually cannot recover voice-related signals from the single channel well.

Multi-Channel Speech Recognition. Multi-channel wireless sensing technology is widely used because of its comple-

mentarity between channels. Therefore, vision and wireless signal perceptions are introduced to improve speech recognition performance [26] [21]. Ding et al. [26] proposed a self-supervised audio-video synchronous learning method to solve the speaker classification problem without extensive labeling work. With the widespread of deep learning in speech signals processing, Yu et al. [21] proposed an attention mechanism to realize multi-modal fusion speech recognition. However, the audio-visual method requires an additional camera in addition to the microphone. In addition, cameras are inconvenient in many typical scenarios due to privacy concerns. Therefore, UltraSE [22] used ultrasound Doppler shifts caused by facial motion to enhance noisy speech. Nevertheless, this algorithm limits the distance from the mouth to the microphone to 20cm. Therefore, speech recognition work based on radio frequency sensing, which has a broader sensing area, has also attracted much attention. Liu et al. [27]integrated the speech perception signals of mmWave and microphone modalities and proposed an anti-noise multi-modal speech recognition system. Therefore, multi-modal fusion methods can effectively improve speech recognition performance by utilizing the complementarity between modalities. Nevertheless, the multiple sensor-based approaches incur additional hardware costs and require essential synchronization in both time and space.

We are considering the significant shortcomings of single-channel speech recognition and the complementarity of multi-modal features. Therefore, we propose a single-channel multi-modal speech recognition algorithm. Specifically, we utilize a single-channel mmWave radar that can reduce costs and effectively avoid synchronization problems between multiple sensors. In addition, the complementarity of lip motion and vocal-cords vibration can effectively improve speech recognition performance.

III. Empirical Study

This section briefly introduces the feasibility of simultaneously sensing human subjects' micro-vibration and macro-motion using FMCW mmWave radar.

A. Principle of Vibration Perception

mmWave radar senses the motion of a target by transmitting the FMCW signal and receiving the reflected signal. Assuming that the transmitted and received signals are defined as:

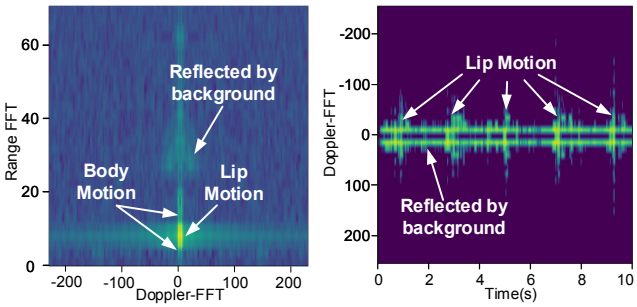
$$S_{tx}(t) = \exp(j2\pi f_c t + j\pi B t^2/T), \tag{1}$$

$$S_{rx}(t) = A\exp[j2\pi f_c(t-\tau) + j\pi B(t-\tau)^2/T]. \tag{2}$$

Here, f_c represents the start frequency, B represents the bandwidth of the chirp, T represents the period of the chirp, and τ represents the delay of the received signal. Assuming that the distance between the vibrating target and the mmWave radar is d. We use a mixer and filter to eliminate the carrier wave in the received signal $S_{rx}(t)$ and obtain the intermediate-frequency (IF) signal $S_{IF}(t)$ as:

$$S_{IF}(t) = A\exp[j2\pi(\frac{B2d}{Tc}t + \frac{2\pi}{\lambda}d)]. \tag{3}$$

Here, A is the path loss, c is the speed of light, and λ is the wavelength. Suppose the displacement Δd of the moving target is 1mm (for 60GHz radar λ=5mm, B=4GHz), and the frequency of the IF signal changes in the observation window. This corresponds to only an additional $0.026T$, i.e., $\Delta fT = \frac{B2\Delta d}{c}$. Since this change is below the frequency resolution T, it will not be captured effectively in the spectrum. Moreover, the phase of IF signal changes is given by:



(a) Range-doppler spectromgram (b) Doppler-time spectromgram

Fig. 2. Lip motion feature extraction

$$\Delta\phi = \frac{4\pi}{\lambda}\Delta d. \tag{4}$$

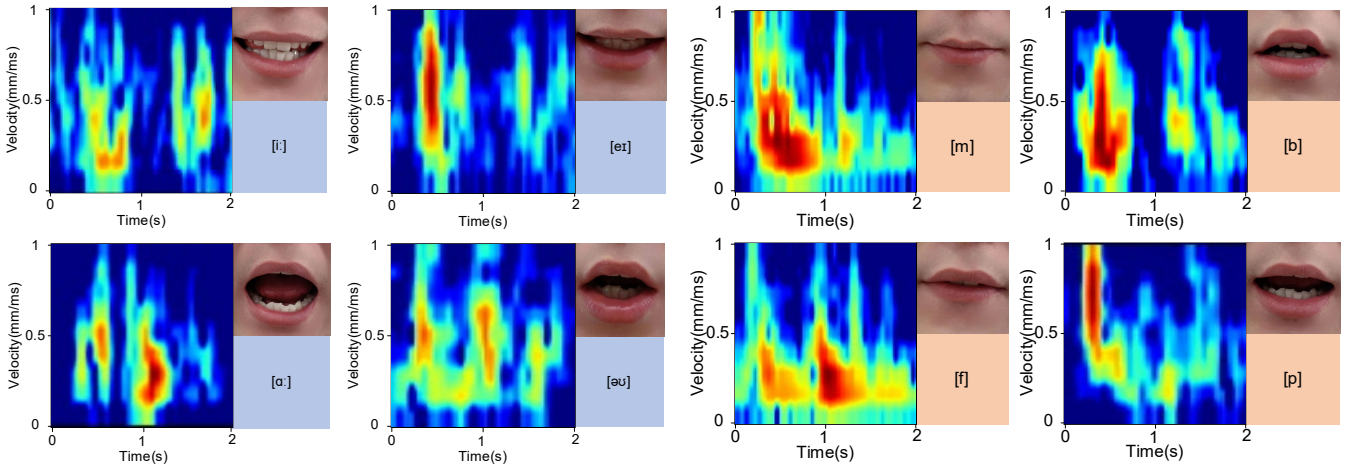
If the displacement of the moving target is 1mm (for 60GHz radar λ=5mm, B=4GHz), the phase of the IF signal changes is 0.8π, i.e., $144°$, this change is much larger than the phase angle resolution. Therefore, according to the range resolution $d_{res} = \frac{c}{2B}$ of mmWave radar, it is difficult to distinguish the position of lips and vocal-cords even with a minimum range resolution. Fortunately, the phase of the IF signal is sensitive to small changes. Therefore, we can simultaneously sense lip motion and vocal-cords vibration using the phase change within the same range bin.

B. Sensing Lip Motion

Lip language is composed of a series of lip motions. Therefore, we can obtain lip language information by sensing lip motion signals. Here, we explore whether we can use mmWave radar to perceive lip motion and obtain lip language information. Notably, we only moved the lips without real pronunciation to avoid the disturbance of the vocal-cords vibration. To confirm the location of the lip motion, we use Range-FFT and Doppler-FFT [28] on the received signal to obtain the range bin where the lip motion is located, as shown in Fig.2(a). Although the frequency of lip motion is minor, its motion amplitude is large. Moreover, the velocity changes obviously, and the lip motion information can be perceived with the phase velocity feature, as shown in Fig.2(b). According to Eq.(4), the angular frequency w of the phase-amplitude changing with time is given by:

$$w = \frac{\Delta\Phi}{\Delta t} = \frac{4\pi v_r}{\lambda} = 2\pi f_d. \tag{5}$$

Here, v_r is the target velocity, and f_d is the Doppler shift. Therefore, we can obtain the lip motion information by calculating the Doppler frequency of the reflected signals. Specifically, the frequency of lip motion is usually below 20Hz. We consider performing down-sampling (sampling rate is 76.9Hz) of the original phase signals while retaining more phase information. According to Eq.(5), the velocity information of the lip motion can be obtained. Fig.3 shows the



(a) Vowvels (b) Conconants

Fig. 3. Macro-motion spectrogram of lip motion

velocity-time spectrograms of lip motion for different phonetic symbols which are perceived by mmWave radar. Note that there are significant differences in the velocity-time spectrograms among different lip motions, no matter whether it is a vowel phoneme or consonant phoneme. Thus, mmWave radar can obtain lip language information by sensing lip motion.

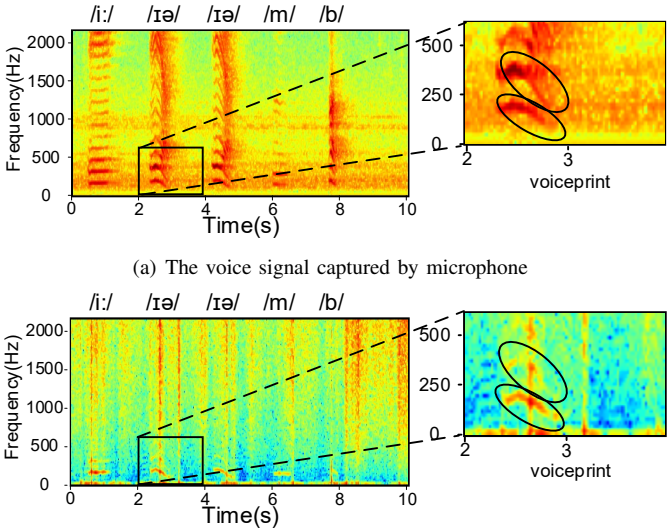
C. Sensing Vocal-cords Vibration

The human voice mainly depends on the vibration of the vocal-cords. Therefore, we designed feasibility experiments to explore the correlation between vocal-cords vibration and speech signals. Specifically, we placed a mmWave radar in front of the person to sense vocal-cords vibration signals and a microphone to collect speech signals. The mmWave radar measures distance with FMCW chirp signals.

When the human subjects speak, they are located in the radial direction of the mmWave radar. Therefore, according to Eq.(4), we can obtain the vocal-cord displacement by calculating the phase change of the mmWave radar signals. Since lip motion is low-frequency and has a large amplitude, we can easily remove the interference of lip motion by using a high-pass filter. We can observe that the phase change of mmWave radar can depict the vibration of the vocal-cords caused by a human voice.

The distance between the subject and the mmWave radar is estimated by calculating the frequency difference between the transmitted signal and the received signal. We use Range-FFT and Doppler-FFT on the reflected signals of mmWave radar to estimate the location of vocal-cords vibration. Fig.4(a) and Fig.4(b) respectively show the speech signals collected by the microphone and the vocal-cords vibration signals collected by the millimeter wave radar. Comparing Fig.4(a) with Fig.4(b), the mmWave signal perceived vocal-cords vibration signal has a significant correlation with the speech signal collected by the microphone. For example, two voiceprints can be seen at 200Hz and 300Hz in the time-frequency spectrogram.

Thus, mmWave radar can obtain speech by sensing the vocal-cords vibration and lip language by sensing the lip motion. We can draw the following conclusions: Lip motion has the characteristics of *low frequency*, *large amplitude*, and



(a) The voice signal captured by microphone

(b) The vocal-cords vibration captured by mmWave

Fig. 4. Comparison of spectrogram of vocal-cords vibration signal and microphone speech signal

high velocity. We can use the doppler velocity spectrogram to extract lip motion features. The vocal-cords vibration has the characteristics of *high frequency* and *small vibration amplitude*. We can extract the vocal-cords vibration signal frequency-time spectrogram by sensing the phase change.

IV. System Design

We design mmMIC, a mmWave-based multi-modal fusion lip-vibration speech recognition approach, which simultaneously senses the speaker's lip macro-motion signal and vocal-cords micro-vibration signal using mmWave radar. As shown in Fig.5, we propose the following three modules:

Extracting Macro-motion Features. To extract the Macro-motion spectrogram of lip motion, we utilize mmWave radar to transmit beamforming chirp signals, then extract the doppler velocity spectrogram from the reflected signal. Consider that body and head motions are inevitable when humans speak. We propose a difference-based motion disturbance removal algorithm to reduce the impact of body and head motion.

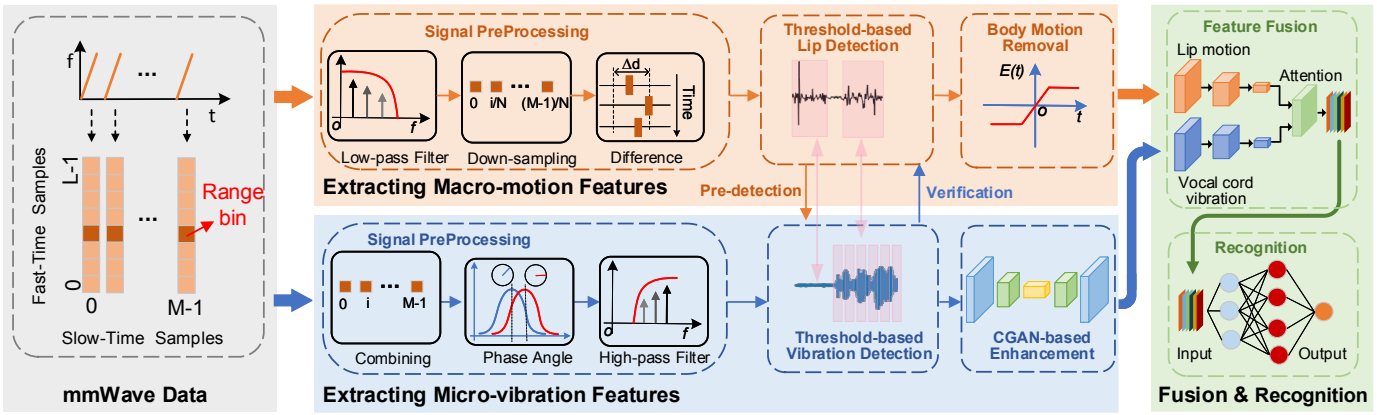


Fig. 5. System overview

Extracting Micro-vibration Features. We obtain the vocal-cords vibration signal from the phase change of the reflected signal. However, many resources will be wasted on processing meaningless noise without proper detection mechanisms. Therefore, we propose a speech detection method based on lip motion and vocal-cords vibration cross-verification to solve this problem.

Attention-based Multi-modal Fusion and Recognition. Vowels produce significant vocal-cords vibration when humans speak, and all phonetic symbols produce lip motion. We propose *TransFuser*, a novel multi-modal fusion transformer that leverages attention to integrate lip motion spectrogram and vibration spectrogram representations.To realize speech recognition and verify the effectiveness of feature fusion, we connect a recognition module to the fusion network output.

A. Extracting Macro-motion Features

Signal Preprocessing. To extract lip motion features, we first utilize mmWave radar to transmit FMCW signals and employ beamforming to focus the signal energy on the lip region, as shown in Fig.2. Then, to estimate the lip motion, we perform Range-FFT operation at the fast-time windows and Doppler-FFT operation at the slow-time windows on the received signal. For example, as shown in Fig.2(a), we can see that the Doppler velocity changes significantly at the position where the range-FFT bin is 8. Therefore, the range bin corresponding to the target is found by finding the max value within the user-specified range limit in the range profile. After that, we can obtain the phase signal containing the lip motion from the range bin. As shown in Fig.6, we use the vector of the signals to provide an intuitive representation of the signal superimposition:

$$\mathbf{S} = \mathbf{S}_v + \mathbf{S}_l + \mathbf{S}_d + \mathbf{S}_b. \tag{6}$$

Here, $\mathbf{S}$ represents all reflected signals at the receiver, $\mathbf{S}_b$ represents the background static interference signals, $\mathbf{S}_d$ represents the dynamic interference signals generated by the speaker's body motion and head motion, $\mathbf{S}_l$ represents the lip motion signals, $\mathbf{S}_v$ represents the vocal-cords vibration signals we expect to obtain. Since S_b is a static background component, we directly cancel it according to the reflected signal variability. Due to the low frequency and significant

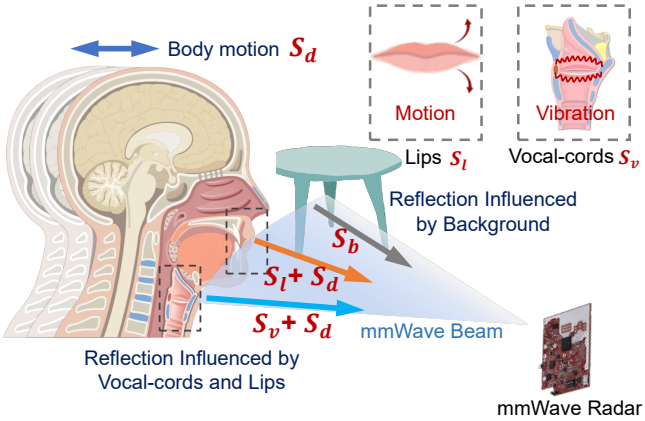


Fig. 6. A multi-model speech recognization in mmMIC

amplitude variation of lip motion, we apply a low-pass filter and resample the phase signal in the range bin to remove vocal-cord interference. So the receiver signals can be expressed as:

$$\mathbf{S}_{l+d} = \mathbf{S}_l + \mathbf{S}_d. \tag{7}$$

Here, $\mathbf{S}_{l+d}$ represents all reflected signals, $\mathbf{S}_d$ represents the dynamic interference signals generated by the speaker's body and head motion, $\mathbf{S}_l$ represents the lip motion signals.

Dynamic Interference Removal. At present, we have obtained the lip motion signal S_{l+d}, i.e., Eq(7) without vocal-cords vibration. Considering that during the process of lip motion, the human's body and head motion, which is called dynamic interference, also exists simultaneously. Therefore, we need to consider the effects of dynamic interference when extracting lip motion features. Even if we use beamforming to reduce the beam width of mmWave radar, it cannot altogether avoid dynamic interference. As shown in Fig.2(a), we find that the dynamic interference in the Range-doppler spectrogram is contained in the range bin where the lip motion is located and has apparent changes in its adjacent bins.

To solve this problem, we propose a difference-based method to remove the dynamic interference. Specifically, we can estimate dynamic interference signals $\mathbf{S}_d$ from the adjacent range bins where the lip motion is located, expressed as:

$$\hat{\mathbf{S}}_d = \frac{1}{2N} \sum_{n=-N}^{N} \alpha_n \cdot \mathbf{S}_{l+d}(R+n) \; n, N \in Z, n \neq 0. \tag{8}$$

Here, $\hat{\mathbf{S}}_d$ is the estimated signal of dynamic interference,

$\mathbf{S}_{l+d}(R)$ is the lip motion signal, including dynamic interference at range bin R, $2N$ is the number of adjacent bins. α_n is the weight of the signal in the n-th bin and $\sum_{n=-N}^{N} \alpha_n = 1$. After that, we use Short-Time Fourier Transform (STFT) [29] to calculate the Doppler spectrogram of the lip motion-related signals and the dynamic interference signals, respectively:

$$STFT(\mathbf{S}_l) = STFT(\mathbf{S}_{l+d}) - STFT(\hat{\mathbf{S}}_d). \quad (9)$$

Finally, we obtain the Doppler velocity spectrogram of lip motion-related without dynamic interference.

B. Extracting Micro-vibration Features

Signal Preprocessing. Based on the empirical study, we can obtain the phase change signal caused by the vibration of the vocal-cords. According to Eq.(4), the displacement Δd of the measured object is related to the phase $\Delta\Phi$ of the reflected signal. Since phase is directly related to vocal-cords vibration, we combine the phase variation over time within each frame into a waveform. In addition, the phase of the reflected signal also changes when the sensed human subject has other motions, i.e., body and head motions, as shown in Fig.6. Fortunately, since the lip motion component S_l, the body motion component S_d, and the background component S_b are low-frequency signals. Therefore, we can use a high-pass filter to suppress these low-frequency components effectively. According to Eq.(6), the vocal-cords vibration signal can be expressed as $S = S_v$. We have obtained the vocal-cords vibration without lip motion signal and dynamic signal.

Speech Activity Detection based on Cross-validation. Considering that there always exist nonspeaking activities in real scenarios, e.g., the human subject is eating or chewing with obvious lip motion and minor vocal-cords vibration. If these signals of nonspeaking activities cannot be effectively distinguished, much computing resource can be wasted in feature extraction/fusion and speech recognition. Fortunately, mmWave radar record both vocal-cords vibration signals and lip motion signals. During the process of speech activity, lip motion and vocal-cords vibration usually coincide. Therefore, we can exploit the multi-modal correlation to distinguish the speech and non-speech activity. To solve this problem, we propose a method based on cross-validation of lip motion and vocal-cords vibration using support vector machines (SVMs), as shown in Fig.7. This method mainly contains three steps:

Lip Motion Pre-detection. Considering that human speech activities must contain lip motion, we use a threshold-based energy detection algorithm to estimate the energy intensity of lip motion signals within the window. Specifically, we split the lip motion signals into segments of $\Delta\tau$ window length and then calculate the energy intensity $E_l(t, t+\Delta\tau)$ within the window. If $E_l(t, t+\Delta\tau)$ is greater than the threshold of lip motion, we obtain the energy $E_l(t, t+\Delta\tau)$ and further verify it with vocal-cords vibration features.

Verification of Vocal-cords Vibration. We use the sliding window dt to perform sliding detection with an overlap is $dt/2$ on the time window of $\Delta\tau$ in the signal of vocal-cords vibration. Note that not all phonetic symbols vibrate the vocal cords, for example, unvoiced consonants. Fortunately, since

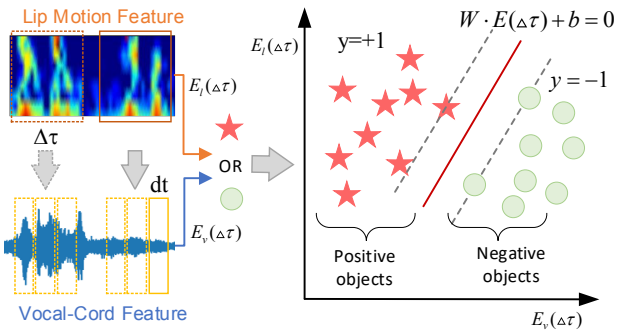


Fig. 7. Speech activity detection based on cross-validation

consonants rarely occur alone, we can perceive speech activity based on the consonant and vowel alternation principle in human speech. We compute the signal energy $E_v(t+\Delta\tau) = \sum E_v^i(t+dt)$ if the energy of each window dt is greater than the threshold in the vocal-cords vibration signal.

Decision based on SVM. We can obtain the energy $E_l(\tau)$ of lip motion spectrogram and the energy $E_v(\tau)$ of vocal-cords vibration signal in the τ window. To comprehensively consider the energy of lip motion and vocal-cords vibration, we can obtain the final energy $E(\Delta\tau)$ using **Concat** operation to concatenate the lip motion's energy $E_l(\Delta\tau)$ and the vocal-cords vibration's energy $E_v(\Delta\tau)$, express as:

$$E(\Delta\tau) = \textbf{Concat}(E_l(\Delta\tau), E_v(\Delta\tau)). \quad (10)$$

Finally, we utilize SVM to discriminatively classify the energy $E(\Delta\tau)$ of speech detection. Specifically, the positive samples are signals with a speech activity window t, and the negative samples are signals without a speech activity signal.

CGAN-based Enhancement. The vocal-cord only generates the fundamental frequency of speech signals, and the high-frequency harmonics are mainly generated by the resonance of the oral and nasal cavities [30]. Therefore, the high-frequency harmonic components are lacking in the speech signal collected by mmWave radar, as shown in Fig.4. Fig.4(a) compares the frequency band between the original sound from the microphone and the extracted sound from mmWave radar in the spectrogram. The low-frequency band is the fundamental frequency for the human voice, while the high-frequency band is the harmonic frequency. Therefore, we need to recover the harmonic band from the fundamental frequency.

To solve the problem, we use the cycle-consistent adversarial networks(Cycle-GAN) [31] to generate all frequency coefficients from the fundamental frequency signal. We use the STFT to convert the vocal-cords vibration signal and the speech signal to the spectrogram, respectively. For the generator network, we use nine blocks for 128×128 images in the generator, and each residual block consists of two convolutional layers with kernel size 3×3. The discriminator network contains two convolutional blocks, six residual blocks, and a fully connected network to predict true or false.

In order to make the generated spectrograms consistent with the natural speech spectrogram features, we utilize a similarity loss, i.e., L1 loss, to constrain the correspondence between the vibration and speech spectrum. Fig.9 shows that

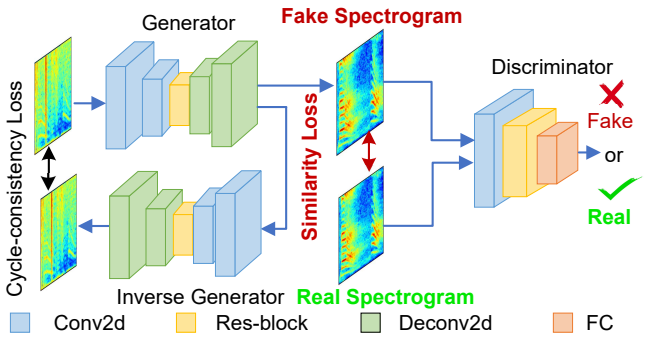


Fig. 8. CGAN framework.

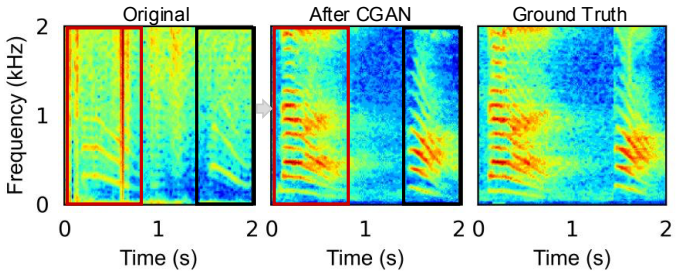


Fig. 9. Spectrogram comparison

the spectrogram of the original vocal-cord signal collected by mmWave radar only contains the fundamental frequency signal, while the vocal-cords vibration signal we generated contains rich high-frequency harmonics. Note that they are consistent by comparing the spectrogram after CGAN and the ground truth. Finally, we use inverse STFT [32] to transform the spectrogram to achieve a better human voice.

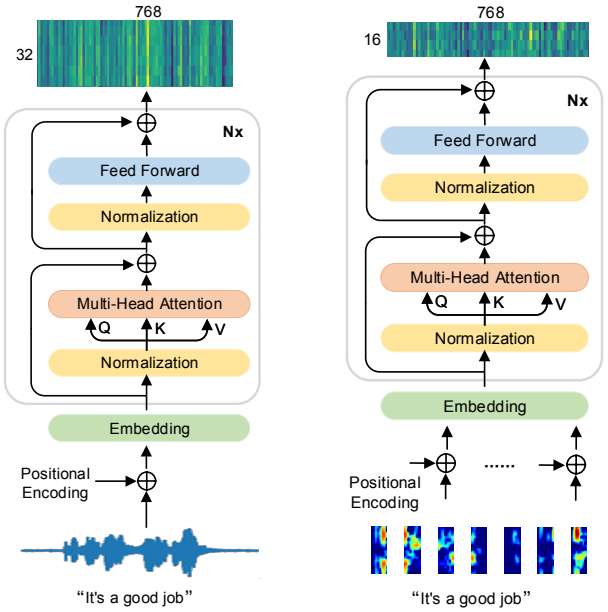
C. Attention-based Multi-modal Fusion and Recognition

Lip motion and vocal-cords vibration represent human speech signals from macro and micro perspectives, respectively. Since mmWave radar can sense lip motion and vocal-cords vibration simultaneously, we fuse these two modalities to improve speech recognition accuracy. We propose *TransFuser*, which improves the complementarity of multi-modal features by fusing vocal-cord micro-vibration signals and lip macro-motion features. As shown in Fig.11, to fuse the vocal-cords vibration feature and lip motion feature, our key idea is to leverage the self-attention mechanism [33] of transformers.

Vocal-cords Vibration Encoder. The feature extraction of the speech signal is the first step of automatic speech recognition(ASR) [34], so the speech encoder based on the attention mechanism is widely used. Fig.10(a) is the feature encoding module of vocal-cords vibration. We use learned embeddings to transform input segments $x = (x_1, ..., x_N)$ into continuous representation output segments $y = (y_1, ..., y_N)$, with $x_i, y_i \in \mathbb{R}^d$. As shown in Fig.10(a), for the embedded segmentation vector, we compute the attention function

$$\textbf{Attention}(Q, K, V) = \textbf{softmax}(\frac{QK^T}{\sqrt{d_k}})V. \quad (11)$$

Here, Q represents a set of query matrices, the keys and values are packed together into matrices K and matrices V, $\frac{1}{\sqrt{d_k}}$ represents the scaling factor. Specifically, all attention outputs



(a) Extracting vocal-cords vibration feature based on encoder

(b) Extracting lip motion feature based on encoder

Fig. 10. Feature encoding module

will be concatenated into a vector and passed through a feed-forward network to a predefined output dimension.

Lip Motion Encoder. Unlike the speech encoder, the lip encoder inputs a lip motion spectrogram. Therefore, we focus on patch features and study the use of vision transformers (ViTs) [35] for vision encoder. In addition, Fig.10(b) is the feature encoding module of lip motion. Considering that the lip motion spectrograms that we collected contain important temporal information. As a result, we segment the spectrograms $L \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened 2D patches $L_p \in \mathbb{R}^{M \times (H \times P \times C)}$ according to the time dimension, where (H, W) is the resolution of an original spectrogram, (H, P) is the resolution of each patch, C is the number of channels, and $M = W/P$ is the resulting number of patches. Finally, to preserve the temporal logic in the lip motion spectrogram, we add standard learnable 1D position embeddings.

Multi-modal Fusion. Although lip motion and vocal-cords vibration represent speech signals from different perspectives, their speech information has a significant correlation and complementarity. For example, vocal-cords vibration is weak for consonants, lip motion is significant, while vocal-cords vibration and lip movement are significant for vowels. Since the human voice is completed by vocal-cords vibration and lip motion, they describe speech information from two aspects and have a significant correlation. Therefore, we can exploit the correlation and complementarity between the two modalities to improve speech recognition performance. To reasonably and effectively fuse lip motion and vocal-cords vibration features, we propose an attention-based multi-modal fusion approach, *TransFuser*, to improve the recognition performance. As shown in Fig.11, *TransFuser* is mainly connected by *cross-attention* [36] and *merged-attention* [37] alternately. Specifically, the essence of *cross-attention*, i.e.,$\mathbf{A}_S =$

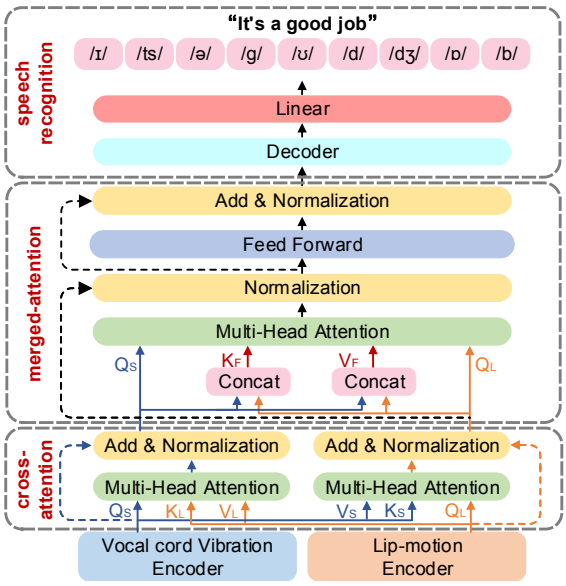


Fig. 11. Attention-based multi-modal fusion and recognition.

$\mathbf{att}((\mathbf{K}_L, \mathbf{V}_L), \mathbf{Q}_S)$ and $\mathbf{A}_L = \mathbf{att}((\mathbf{K}_S, \mathbf{V}_S), \mathbf{Q}_L)$, is to achieve features correlation between lip motion and vocal-cords vibration by exchanging the attention information. The essence of *merged-attention*, i.e., $\mathbf{A}_F = \mathbf{att}((\mathbf{K}_F, \mathbf{V}_F), \mathbf{Q}_S) + \mathbf{att}((\mathbf{K}_F, \mathbf{V}_F), \mathbf{Q}_L)$, is to achieve features complementarity between lip motion and vocal-cords vibration by merging the attention information.

Speech Recognition. To realize speech recognition and verify the effectiveness of feature fusion, we connect a recognition module to the fusion network output. Inspired by the existing ASR methods [5], we design a decoder to recognize the fused features. Specifically, similar to the encoder structure in Fig.9, the decoder is mainly connected by a multi-head attention sublayer and a feed-forward sublayer alternately. In addition, a normalization layer is added before each block, and a residual connection is applied after each block.

V. Performance Evaluation

A. Experiment Setup

Hardware. The platform to implement our algorithm is based on TI's IWR6843BOOST radar, high-speed data collection board DCA1000EVM, and laptop, as shown in Fig.12. The IWR6843BOOST is a mmWave radar that continuously transmits FMCW with a 60GHz carrier to measure distance and angle. Specifically, the chirp transmits period of the radar configuration is 50μs. The received channel has a 4000k ADC sampling rate, and each received chirp contains 64 samples.

Software. We use a laptop to run a python script to connect and control the radar. We write python scripts to control the microphone and mmWave radar to capture both mmWave radar signals and audio signals. The vocal-cords vibration encoder, the lip encoder, and the decoder for recognition consist of N = 6 identical layers, the hidden layer size is 384, and the number of attention heads is 6. We apply dropout to the output of each sub-layer with a dropout rate of 0.2. We use the Adam optimizer with $\beta = 0.9$, $l_{rate} = 1e-4$. Weight decay is 100, and batch size is set to 32.

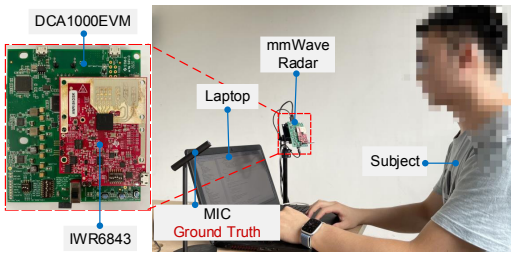


Fig. 12. The evaluation setup.

Dataset. We choose 48 phonetic symbols as recognition objects. We invited eight male and seven female volunteers to speak all the phonetic symbols. We collected 21,600 sets of data, including speech signals, mmWave vocal-cords vibration signals, and lip motion signals. We choose 1/5 of them as the test, which covers all the phonetic symbols of each volunteer.

Metrics. To comprehensively evaluate the overall performance of our algorithm, we use two metrics, *accuracy* and *recall*. The accuracy rate represents the proportion of correctly identified samples in all samples. We use the recall rate to evaluate the speech recognition performance to reduce the missed recognition rate. The recall rate is the proportion of correctly identified samples to the total samples for each class.

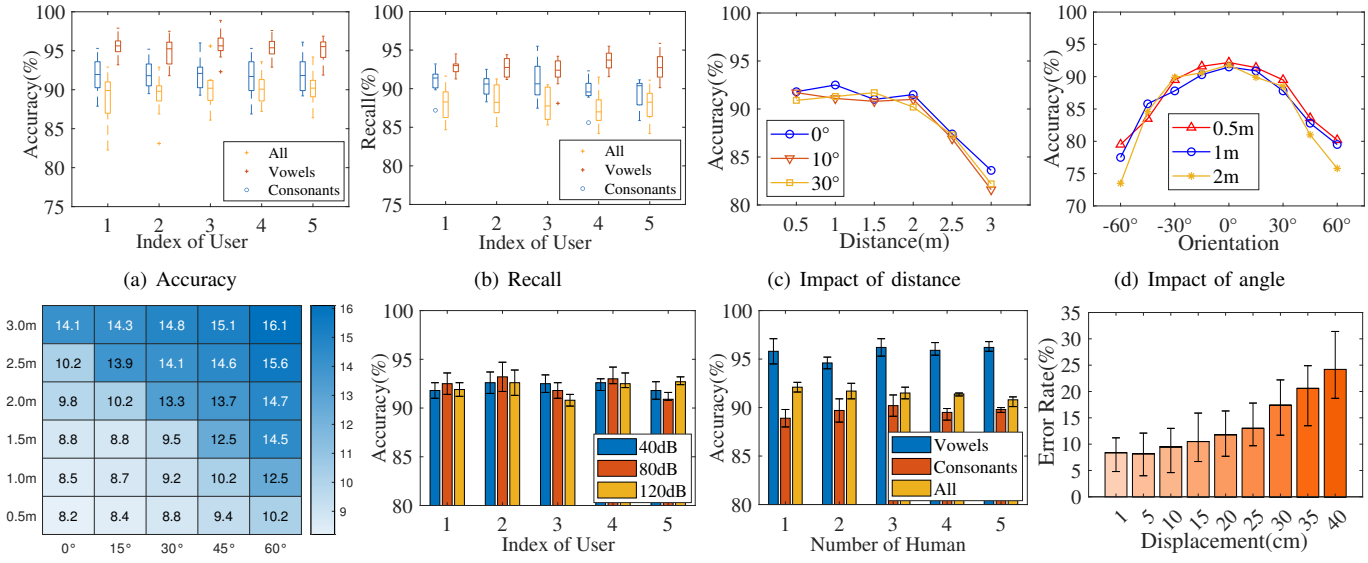
B. Performance

Overall Performance. Our solution achieves the best performance in speech recognition in the presence of speech interference. Fig.13(a) illustrates the accurate measurement of different phonetic symbols for five volunteers. Our algorithm achieves over 90% accuracy for all phonetic symbols. Especially for vowels, the accuracy rate reaches 95%. Even for consonants, we can exploit the complementarity of multi-modality to improve accuracy in speech recognition. In addition, as shown in Fig.13(b), our recall rate is also higher than 85% in speech recognition.

Impact of Distance and Orientation. Our system can still perform speech recognition efficiently when the distance is within 2m, and the orientation is less than 30 degrees. In our experiments, we set the angle to change from 0 degrees to 60 degrees and the distance to change from 0.5m to 3m. Fig.13(c), Fig.13(d) and Fig.13(e) indicate that when the direction angle is 30 degrees and the distance is within 2m, and the speech recognition accuracy is more than 90%.

Impact of Ambient Noises. Our system can perform speech recognition efficiently in music noise. We evaluated the speech recognition performance of five volunteers in different decibel music noise scenarios. Fig.13(f) shows that speech recognition accuracy under different decibels remains above 90% without apparent attenuation. Therefore, our system is hardly disturbed by ambient noise.

Impact of Multi-human. Our system can perform speech recognition efficiently in multi-human speech noise. We separately evaluated the speech recognition performance of volunteers in the presence of different amounts of human speech noise. Fig.13(g) shows that as the number of humans producing speech noise increases, speech recognition performance does not significantly impact.



(a) Accuracy (b) Recall (c) Impact of distance (d) Impact of angle

(e) Impact of distance and angle (f) Impact of ambient noises (g) Impact of multi-human (h) Impact of body motion

Fig. 13. Impact of Environment.

Impact of Body Motion. Our system can effectively perform speech recognition when the human subject's speaking body motion displacement is less than 10cm. We individually assessed five volunteers with a range of body motion within 40 cm. Fig.13(h) shows that the error rate increases as the vibration amplitude of the body motion increases.

C. Ablation Study

In this section, we conduct an ablation study to quantify the fusion of the two modal signals and our proposed fusion method. In contrast, we comprehensively validate our method by ablating specific components:

Lip Motion, where no vocal-cord micro-vibration signal is fused in our fusion recognition network, we replace the *TransFuser* sub-network with a transformer of the same depth to extract features. As shown in Table I, we find that whether it is vowels or consonants, the recognition accuracy of lip movements is gathering more than 73.35%.

Vocal-cords Vibration, where no lip macro-motion signal is fused in our fusion recognition network, we replace the *TransFuser* sub-network with a transformer of the same depth to extract features. As shown in Table I, we find that the recognition accuracy of vowels with noticeable vocal cord vibration was 90.8%, but the recognition rate of consonants with slight vocal-cord vibration was 75.3%.

Fusion (Lip Motion and Vocal-cords Vibration), which uses *TransFuser* sub-network to fuse lip macro-motion and vocal-cord micro-vibration to perform speech recognition on the fused features. As shown in table I, we find that after fuse lip motion and vocal-cords vibration features, our recognition accuracy for both vowels and consonants is higher than 90%.

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT FEATURES.

Dataset	Methods		
	Lip	Vibration	Fusion
Monophthongs	73.2	91.3	95.89
Diphthongs	80.64	93.2	97.61
All Vowels	75.21	90.8	**95.5**
Unvoiced Consonants	75.51	73.1	90.56
Voiced Consonants	77.1	76.2	93.31
All Consonants	73.35	75.3	**91.3**
All Phonetics	71.83	88.9	**92.8**

In Table I, we find that even though lip motion and vocal-cords vibration have lower recognition performance in vowels and consonants, after *TransFuser* sub-network fusion, its recognition performance is 20.97% higher than lip motion and 3.9% higher than vocal-cords vibration.

VI. Conclusion

In this paper, we propose a novel mmWave-based solution to perform speech recognition in the situation of multiple sound sources and ambient noises by precisely extracting the multi-modal features from lip motion and vocal-cords vibration from the single channel of mmWave. To extract the signal features from lip motion with the dynamic interference from the body and head motions, we propose a difference-based method to remove the dynamic interference. We propose a cross-validation-based speech detection method to distinguish speaking activities from nonspeaking activities so as to avoid wasting computing resources on nonspeaking activities. We implemented a prototype system and evaluated the performance in real test-beds. Experiment results show that the average speech recognition accuracy is 92.8% in realistic environments.

Acknowledgments

This work is supported in part by National Key Research and Development Program of China under Grant No.2022YFB3303900; National Natural Science Foundation of China under Grant Nos. 62272216, 61832008, 61872174, 61902175; Collaborative Innovation Center of Novel Software Technology and Industrialization. This work is supported in part by the program A for Outstanding Ph.D. candidate of Nanjing University No.202201A015 and Postgraduate Research & Practice Innovation Program of Jiangsu Province No.KYCX22_0151. Lei Xie is the corresponding author.

REFERENCES

[1] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7073–7083.

[2] S. Wang, R. Chen, L. Zhao, and C. Liu, "Millimeter wave integrated sensing and communication with hybrid architecture in vehicle to vehicle network," in *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, 2021, pp. 01–06.

[3] M. Luria, G. Hoffman, and O. Zuckerman, "Comparing social robot, screen and voice interfaces for smart-home control," in *Proceedings of the 2017 CHI conference on human factors in computing systems*, 2017, pp. 580–628.

[4] S. Tian, W. Yang, J. M. Le Grange, P. Wang, W. Huang, and Z. Ye, "Smart healthcare: making medical care more intelligent," *Global Health Journal*, vol. 3, no. 3, pp. 62–65, 2019.

[5] Y. Xie, C. Shi, Z. Li, J. Liu, Y. Chen, and B. Yuan, "Real-time, universal, and robust adversarial attacks against speaker recognition systems," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1738–1742.

[6] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6134–6138.

[7] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M.-S. Ni, "We can hear you with wi-fi!" in *Proceedings of the Annual International Conference on Mobile Computing and Networking, MOBICOM*, 2014, p. 593.

[8] D. Xia, X. Zheng, F. Yu, L. Liu, and H. Ma, "Wira: Enabling cross-technology communication from wifi to lora with ieee 802.11ax," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, 2022, pp. 430–439.

[9] Y. Xie, Z. Li, and M. Li, "Precise power delay profiling with commodity wi-fi," *IEEE Transactions on Mobile Computing*, vol. 18, no. 6, pp. 1342–1355, 2019.

[10] Z. Chen, G. Zhu, S. Wang, Y. Xu, J. Xiong, J. Zhao, J. Luo, and X. Wang, "m^3m3: Multipath assisted wi-fi localization with a single access point," *IEEE Transactions on Mobile Computing*, vol. 20, no. 2, pp. 588–602, 2021.

[11] C. Wang, L. Xie, Y. Lin, W. Wang, Y. Chen, Y. Bu, K. Zhang, and S. Lu, "Thru-the-wall eavesdropping on loudspeakers via rfid by capturing sub-mm level vibration," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–25, 2021.

[12] Y. Zheng, Y. He, M. Jin, X. Zheng, and Y. Liu, "Red: Rfid-based eccentricity detection for high-speed rotating machinery," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018, pp. 1565–1573.

[13] C. Zhao, Z. Li, H. Ding, G. Wang, W. Xi, and J. Zhao, "Rf-wise: Pushing the limit of rfid-based sensing," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, 2022, pp. 1779–1788.

[14] C. Zhao, Z. Li, T. Liu, H. Ding, J. Han, W. Xi, and R. Gui, "Rf-mehndi: A fingertip profiled rf identifier," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 2019, pp. 1513–1521.

[15] H. Li, C. Xu, A. S. Rathore, Z. Li, H. Zhang, C. Song, K. Wang, L. Su, F. Lin, K. Ren *et al.*, "Vocalprint: exploring a resilient and secure voice authentication via mmwave biometric interrogation," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 312–325.

[16] H. Li, C. Xu, A. S. Rathore, Z. Li, H. Zhang, C. Song, K. Wang, L. Su, F. Lin, K. Ren, and W. Xu, "Vocalprint: A mmwave-based unmediated vocal sensing system for secure authentication," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2021.

[17] C. Wang, F. Lin, T. Liu, Z. Liu, Y. Shen, Z. Ba, L. Lu, W. Xu, and K. Ren, "mmphone: Acoustic eavesdropping on loudspeakers via mmwave-characterized piezoelectric effect," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, 2022, pp. 820–829.

[18] Z. Li, B. Chen, X. Chen, H. Li, C. Xu, F. Lin, C. X. Lu, K. Ren, and W. Xu, "Spiralspy: Exploring a stealthy and practical covert channel to attack air-gapped computing devices via mmwave sensing," in *The 29th Network and Distributed System Security (NDSS) Symposium 2022*. The Internet Society, 2022.

[19] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: Passive recovery of sound from video," *ACM Transactions on Graphics*, vol. 33, no. 4, July 2014. [Online]. Available: https://doi.org/10.1145/2601097.2601119

[20] J. Yu, S.-X. Zhang, J. Wu, S. Ghorbani, B. Wu, S. Kang, S. Liu, X. Liu, H. Meng, and D. Yu, "Audio-visual recognition of overlapped speech for the lrs2 dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6984–6988.

[21] W. Yu, S. Zeiler, and D. Kolossa, "Fusing information streams in end-to-end audio-visual speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3430–3434.

[22] K. Sun and X. Zhang, "Ultrase: single-channel speech enhancement using ultrasound," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 2021, pp. 160–173.

[23] Y. Chen, M. Gao, Y. Li, L. Zhang, L. Lu, F. Lin, J. Han, and K. Ren, "Big brother is listening: An evaluation framework on ultrasonic microphone jammers," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, 2022, pp. 1119–1128.

[24] Y. Wu, J. Liu, Y. Chen, and J. Cheng, "Semi-black-box attacks against speech recognition systems using adversarial samples," in *2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, 2019, pp. 1–5.

[25] Z. Li, C. Shi, T. Zhang, Y. Xie, J. Liu, B. Yuan, and Y. Chen, "Robust detection of machine-induced audio attacks in intelligent audio systems with microphone array," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 1884–1899. [Online]. Available: https://doi.org/10.1145/3460120.3484755

[26] Y. Ding, Y. Xu, S.-X. Zhang, Y. Cong, and L. Wang, "Self-supervised learning for audio-visual speaker diarization," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4367–4371.

[27] T. Liu, M. Gao, F. Lin, C. Wang, Z. Ba, J. Han, W. Xu, and K. Ren, "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*, 2021, pp. 97–110.

[28] C. Jiang, J. Guo, Y. He, M. Jin, S. Li, and Y. Liu, "mmvib: micrometer-level vibration measurement with mmwave radar," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–13.

[29] N. Kehtarnavaz, "Frequency domain processing," in *Digital Signal Processing System Design (Second Edition)*, second edition ed. Burlington: Academic Press, 2008, pp. 175–196.

[30] R. T. Sataloff, "The human voice," *Scientific American*, vol. 267, no. 6, pp. 108–115, 1992.

[31] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[32] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, p. 6000–6010.

[34] Y. zhao, J. Li, X. Wang, and Y. Li, "The speechtransformer for large-scale mandarin chinese speech recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7095–7099.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[36] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng *et al.*, "An empirical study of training end-to-end vision-and-language transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 166–18 176.

[37] L. A. Hendricks, J. Mellor, R. Schneider, J.-B. Alayrac, and A. Nematzadeh, "Decoupling the role of data, attention, and losses in multimodal transformers," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 570–585, 2021.