# mSilent: Towards General Corpus Silent Speech Recognition Using COTS mmWave Radar

SHANG ZENG, Nanjing University, China
HAORAN WAN, Nanjing University, China
SHUYU SHI, Nanjing University, China
WEI WANG, Nanjing University, China

Silent speech recognition (SSR) allows users to speak to the device without making a sound, avoiding being overheard or disturbing others. Compared to the video-based approach, wireless signal-based SSR can work when the user is wearing a mask and has fewer privacy concerns. However, previous wireless-based systems are still far from well-studied, *e.g.* they are only evaluated in corpus with highly limited size, making them only feasible for interaction with dozens of deterministic commands. In this paper, we present mSilent, a millimeter-wave (mmWave) based SSR system that can work in the general corpus containing thousands of daily conversation sentences. With the strong recognition capability, mSilent not only supports the more complex interaction with assistants, but also enables more general applications in daily life such as communication and input. To extract fine-grained articulatory features, we build a signal processing pipeline that uses a clustering-selection algorithm to separate articulatory gestures and generates a multi-scale detrended spectrogram (MSDS). To handle the complexity of the general corpus, we design an end-to-end deep neural network that consists of a multi-branch convolutional front-end and a Transformer-based sequence-to-sequence back-end. We collect a general corpus dataset of 1,000 daily conversation sentences that contains 21K samples of bi-modality data (mmWave and video). Our evaluation shows that mSilent achieves a 9.5% average word error rate (WER) at a distance of 1.5m, which is comparable to the performance of the state-of-the-art video-based approach. We also explore deploying mSilent in two typical scenarios of text entry and in-car assistant, and the less than 6% average WER demonstrates the potential of mSilent in general daily applications.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**;

Additional Key Words and Phrases: silent speech recognition, millimeter-wave, wireless sensing

**ACM Reference Format:**
Shang Zeng, Haoran Wan, Shuyu Shi, and Wei Wang. 2023. mSilent: Towards General Corpus Silent Speech Recognition Using COTS mmWave Radar. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 39 (March 2023), 28 pages. https://doi.org/10.1145/3580838

## 1 INTRODUCTION

Automatic speech recognition (ASR) enables users to use natural language, which has strong expressive power and low learning cost, to interact with electronic devices. Moreover, ASR provides a contactless alternative to keypads to protect users from touching public devices during the pandemic of COVID-19. Also, in-car ASR can help drivers perform interaction tasks, *e.g.* set the navigation destination, without detracting their attention. Therefore, ASR has been widely used in daily tasks, such as communication, input, and voice assistants. However,

Authors' addresses: Shang Zeng, shangzeng@smail.nju.edu.cn, Nanjing University, Nanjing, Jiangsu, China, 210023; Haoran Wan, wanhr@smail.nju.edu.cn, Nanjing University, Nanjing, Jiangsu, China, 210023; Shuyu Shi, ssy@nju.edu.cn, Nanjing University, Nanjing, Jiangsu, China, 210023; Wei Wang, ww@nju.edu.cn, Nanjing University, Nanjing, Jiangsu, China, 210023.


2474-9567/2023/3-ART39 $15.00
https://doi.org/10.1145/3580838


Fig. 1. Example Application Scenario of mSilent. The user wearing a mask can contactless input complex sensitive information to a public device.

voice-based ASR requires the user to speak loudly during the interaction. In scenarios such as in a conference room or museum, voice-based ASR may disturb nearby people and the interaction can be easily interfered with by noises. Voice-based ASR also raises privacy concerns as the interaction could be overheard by others, especially when inputting sensitive information in public.

Silent Speech Recognition (SSR) recognizes speech based on articulatory gestures such as the movement of the tongue and lips, allowing the user to speak to devices without making a sound. Recent work [34] shows that people perceive silent speech as more socially acceptable than voiced speech, and are willing to tolerate more errors in exchange for privacy. Computer vision is a promising solution for contactless SSR. With the recent advancements in deep learning, the performance of video-based lip reading has been boosted [1, 31, 36], and can recognize speech in a large corpus like BBC speech. However, the fine-grained acquisition of faces for video-based sensing introduces new privacy concerns [5]. In addition, it cannot work in dimly lit scenarios or when the user is wearing a mask.

In recent years, some pioneering work explores SSR using wireless signals, such as Wi-Fi [52] and ultrasound [59]. Wireless signals can propagate through obstructions, such as the mask, and the coarse-grained sensing features reduce privacy concerns [5]. Thus, the wireless-based solution becomes a promising and favorable approach for recognizing silent speech. However, due to the limited resolution of wireless signals and the insufficient capability of recognition models, previous wireless-based systems only work in a very limited corpus. For example, the state-of-the-art ultrasound SSR [59] can only recognize 70 manually designed command sentences. Such a limited capability restricts its application to a similar scope as hand gesture recognition, which only supports simple deterministic commands. In addition, ultrasound-based SSR can only work within a short range ($< 20$ cm), which requires the user to hold the device by hand, further limiting the scope of its application. As a promising sensing technology, mmWave has a shorter wavelength and longer sensing range than ultrasound, so the mmWave-based systems can sense small human movements such as heartbeats [14] and vocal vibrations [51]. Furthermore, with the development of millimeter-wave (mmWave) communication and mmWave radars, the cost of mmWave transceivers is reducing and they are deployed in an increasing number of commercial devices [19, 29], including mobile phones[12], cars, and laptops. These features of mmWave lead to new opportunities for contactless SSR.

In this paper, we propose mSilent, a system that uses commercial off-the-shelf (COTS) mmWave radar with a novel fine-grained model combining signal processing with deep learning to recognize silent speech in the general corpus. With the fine-grained sensing capability, mSilent recognizes not only dozens of predefined commands but also thousands of general conversation sentences, such as "How long does it take you to drive home?" or "Please give me your name and address so I can send an ambulance." In addition, mSilent is hands-free and has a recognition range of more than one meter.

mSilent unleashes the representation power of silent speech recognition and enables new application scenarios, such as in-car assistants, input, and communication:

- Talking with voice assistants would be a convenient interaction method in driving or domestic scenarios, and SSR could improve the user experience. For example, when driving at night with passengers talking loudly, the driver can still set the navigation destination with a noise-proof SSR and maintain the operation of the steering wheel at the same time. Users can control the smart devices or order goods on Amazon Echo without disturbing sleeping children in the bedroom. However, the small corpus used in previous wireless works will make this scenario into a one-way order-giving experience with limited and fixed types of commands. Our case study in the in-car scenario with 500 real interaction sentences exhibits that mSilent can improve this to a real interaction experience between users and their smart voice assistants.
- On top of that, general communication with SSR can be more desirable and challenging. During the pandemic, people tend to wear masks and avoid direct contact with devices in the public. With mSilent, the user wearing a mask can enter sensitive information or personal messages into a public device, *e.g.* gifts delivery service machine, without contacting the surface of the device, as illustrated in Fig. 1. mSilent can also help users, who are unable to type or lose their speaking abilities to communicate with others. Furthermore, users can still use silent speech to send messages comfortably and with fewer privacy concerns [5] in noisy conference rooms or museums. These scenarios require the recognition capability to be even higher to support the general corpus with no domain restrictions. Our evaluations in the 1,000-sentence general conversation corpus and the case study of text entry show that mSilent is not only suitable for complex interaction with voice assistants, but also well capable of general communication and input scenarios.

With all these desirable applications being said, mSilent still faces three core technical challenges to enable SSR in the general corpus.

**(1) How to separate articulatory gestures from multi-path reflections?** The mmWave radar has a long sensing range, which eliminates the necessity for the user to touch the device. However, the long sensing range also introduces severe multi-path effects that overwhelm articulatory gestures with irrelevant movements and noise. To address this challenge, we design a clustering-selection algorithm based on our observation that we can first use coarse localization to detect the human body and then use fine search for the articulatory gestures based on the heuristic that articulatory gestures mostly appear at the top of the body.

**(2) How to extract fine-grained articulatory features from mmWave signals with noise?** Recognizing silent speech in the general corpus requires a large amount of information to eliminate uncertainty. However, extracting fine-grained articulatory features from mmWave signals is difficult, because the duration of articulatory gesture is highly variable [58] and unconscious head movements [15] introduce low-frequency noise. To address this challenge, we design a multi-scale detrended spectrogram (MSDS), which removes low-frequency noise by segmented linear detrending. We perform short-time Fourier transforms (STFT) at multiple scales to extract articulatory gestures with different movement patterns.

**(3) How to recognize silent speech in the general corpus by articulatory features?** The general corpus contains a large number of sentences with highly diversified lengths, domains, and syntax, making the task challenging. Unlike video-based approaches that can benefit from computer vision tasks, there is no general

Table 1. Comparing mSilent with existing SSR systems.

| System | HaMa et al. | Silent Speller | Ear Command | Wi Hear | Echo Whisper | Sound Lip | mSilent (Ours) |
|---|---|---|---|---|---|---|---|
| **Technology** | Video | In-mouth | Earphone | Wi-Fi | Ultrasound | Ultrasound | mmWave |
| **Corpus Size** | >100k | 1164 | 25 | 25 | 45 | 70 | 1000 |
| **Sentence-level** | Yes | No | Yes | No | No | Yes | Yes |
| **For HCI** | No | Yes | Yes | Yes | Yes | Yes | Yes |
| **Contactless** | Yes | No | No | Yes | Yes | Yes | Yes |
| **Wearing Mask** | No | Yes | Yes | Yes | Yes | Yes | Yes |
| **Hand-free** | Yes | Yes | Yes | Yes | No | No | Yes |
| **Error Rate** | 39.1% | 3% | 12.33% | 9% | 8.33% | 7.1% | 9.5% |

model that can be applied to mmWave features. To address this challenge, we design a novel end-to-end deep neural network (DNN) with a convolutional front-end and a sequential back-end. We further design a two-stage user discriminator to enable user-adaptive learning on sequential tasks.

To evaluate the sensing capability of mSilent, we design a general corpus that consists of 1,000 daily conversation sentences. We collect the dataset with two sensing modalities of mmWave and video at the same time. We collect a total of 21,404 samples from 10 users. To our best knowledge, it is the first general corpus silent speech dataset with mmWave and video bi-modality. The dataset allows us to validate the performance of mSilent in the general conversation scenarios and offers the possibility of exploring multi-modal interaction in the future. Our experimental results show that the average word error rate (WER) of mSilent is only 9.5% at a distance of 1.5 meters, which is comparable to the WER of the state-of-the-art video-based system at 7.7%. To further evaluate the performance of mSilent in the real life, we explore deploying mSilent in two typical scenarios: in-car assistant and text entry. The results show that mSilent achieves only 4.3% and 5.1% average WER in the real scenarios with task-specific corpus.

To the best of our knowledge, mSilent is the first general corpus SSR system based on wireless signals that achieve comparable performance to state-of-the-art video-based work. We make the following contributions:

- We present the first signal processing pipeline to extract articulatory features from mmWave signals. Specifically, we design a clustering-selection algorithm to separate articulatory gestures and a multi-scale detrended spectrogram to extract fine-grained features.
- We design a novel end-to-end deep neural network to recognize silent speech in the general corpus, which consists of a multi-branch convolutional front-end and a Transformer-based sequence-to-sequence back-end. We also explore user-adaptive learning on sequential tasks.
- We perform extensive evaluations in a general conversation corpus of 1,000 sentences, verifying the strong sensing capability of mSilent. We further explore deploying mSilent in two typical scenarios(in-car assistant and text entry), showing the promising applications of mSilent in daily life.

## 2 RELATED WORK

In this section, we present the existing SSR systems and mmWave-based sensing. Table 1 shows the overall comparison between mSilent and other SSR systems.

### 2.1 Wearable Device-based SSR

Specially designed wearable devices, such as EEG [11], EMG [50], and non-audible murmur microphone [40], can be used to recognize silent speech. SilentSpeller [25] uses in-mouth electropalatography (EPG) to recognize

1,164 words (sliced from 500 sentences) in MacKenzie-Soukoref corpus [44]. These sensors are highly capable of sensing articulatory gestures, making it possible to recognize them in a large corpus using a lightweight model. However, these specially designed devices are only available in clinics or laboratories and require special wear, thus not suitable for daily-life applications. There are systems that use pervasive sensors, such as magnet [38] and RFID [53], to perform SSR. SpeeChin [60] uses an infrared necklace to recognize 54 commands. EarCommand [20] uses earphones to recognize 32 word-level commands and 25 sentence-level commands. These device-based sensors still need to be actively worn by users and cannot be used for contact-less recognition.

## 2.2 Video-based Lip Reading and SSR

With the advance in computer vision, video-based lip reading has become an important research area recently. Video-based lip reading systems usually are trained on voiced speech videos, but can easily be deployed as powerful device-free SSR. Early work is commonly evaluated on a controlled corpus dataset called GRID [7]. The sentences in GRID have fixed lengths and are generated from 51 words with a fixed pattern. LipNet [2] achieves sentence-level recognition on GRID using end-to-end deep learning. WAS [43] first explores the direction of general corpus recognition, collecting a new dataset called LRS from BBC videos, where sentences do not have any topic, word, or length constraints. WAS uses the Seq2Seq structure to achieve a 3.0% WER on GRID, but the WER in LRS is 50.2%. This huge performance gap shows the difference between the difficulty of SSR in the general and controlled corpus. With the fast advance of deep learning, the performance of video-based speech recognition has been boosted by emerging neural models. Ma *et al.* [31] use Conformer [13] to achieve the state-of-the-art performance in LRS2 [43] dataset without extra training data. AV-Hubert [41] proposes a self-supervised representation learning framework, which can take advantage of the massive unlabeled video to boost performance. However these works are not designed for human-computer interaction (HCI), and the corpus from BBC Video is far from daily life. Unlike these works, Lip-Interact [45] is specially designed for silent speech interaction in smartphones and is evaluated on 20 commands. The video-based SSR can only work in good lighting conditions and without wearing masks. Furthermore, the requirements of capturing fine-grained face images also pose the risk of privacy leakage, making it unsuitable in scenarios such as bedrooms [5].

## 2.3 Wireless Signal-based SSR

Wireless signal-based solutions have become a promising direction for SSR, since they can work while the user is wearing a mask and reduce the risk of privacy leakage [5]. WiHear [52] uses Wi-Fi channel state information (CSI) to recognize 33 different words, furthermore, the low resolution of the Wi-Fi signal prevents more fine-grained recognition. To achieve better resolution, recent works use ultrasound to capture articulatory gestures. Endophasia [61] recognizes 20 word-level commands and uses transfer learning strategies to adapt the model for new users. EchoWhisper [10] uses dual microphones to recognize 45 words selected from the conversation section of the English listening test. To our best knowledge, SoundLip [59] is the only sentence-level wireless SSR, whose corpus contains only 70 manually designed commands from 4 domains. Compared to existing works, mSilent is the only wireless system that can perform SSR in the general corpus with thousands of complex sentences, enabling more general applications in daily life. In addition, the ultrasound-based work can only work at a distance of less than 20 cm, while mSilent can work at a distance of more than 1 m.

## 2.4 mmWave-based Sensing

mmWave has higher frequencies and larger bandwidths than other wireless signals, so it becomes an important method for sensing human activities in recent years. Cao *et al.* [5] use the privacy preservation feature of mmWave to do re-identification in the camera-restricted regions. mmASL [39] uses a custom antenna array to recognize 50 American Sign Language (ASL) signs. RF-SCG [14] achieves contactless seismocardiogram (SCG) recording

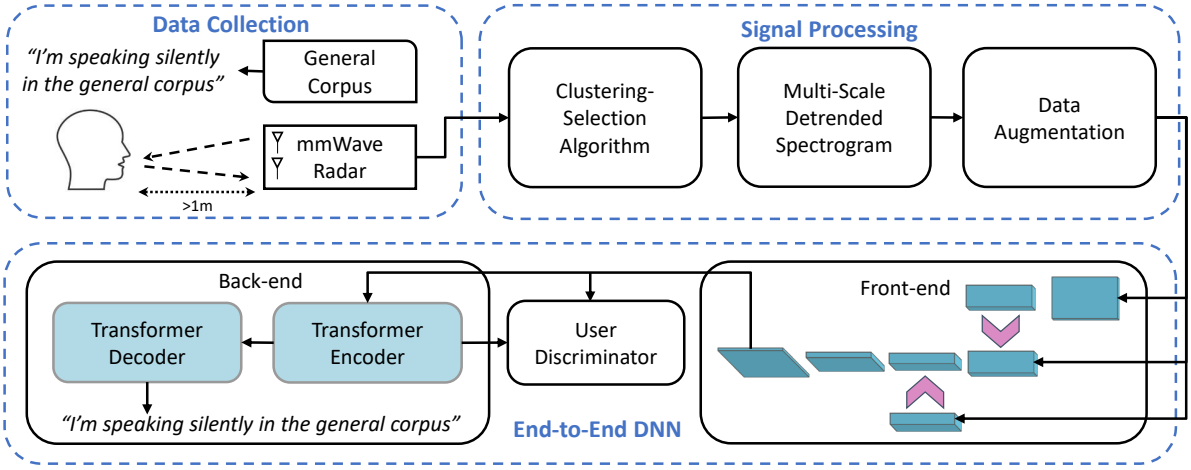



Fig. 2. System overview of mSilent.

by COTS radar. There are some systems that sense the vibrations of the vocal fold. WaveEar [57] first exploits 24GHz radar to restore the voice in noisy scenarios, and the recent work Wavesdropper [51] can recognize 57 words through the wall. mmSpy [4] uses both 77GHz and 60GHz mmWave radar to spy the phone call by sensing the vibrations of an earpiece. For small non-radial movements, such as two-finger gestures, recent works [54] can only recognize 30 gestures at a distance of 3 cm. Li *et al.* [55] shows that articulatory gestures cause signal change at a distance of 5 cm from the mouth, which is the only short-range feasibility experiment of sensing silent speech. Compared to these works, mSilent aims at recognizing silent speech in the general corpus through articulatory gestures at a distance of more than 1 m. Therefore, mSilent has to address the challenges of separating the subtle articulatory gestures from surrounding noises and extracting fine-grained articulatory features from the mmWave signal.

## 3 SYSTEM OVERVIEW

mSilent uses a hybrid architecture that combines signal processing with deep learning as shown in Fig. 2. First, mSilent uses COTS mmWave radar to collect general corpus silent speech at a distance of more than one meter, and the raw signals are converted to range-angle maps. To separate the subtle articulatory gestures from multipath reflections, we use the clustering-selection algorithm at the beginning of the pipeline to separate the articulatory gestures at a specific angle/distance. The separated signal is then transformed into the multi-scale detrended spectrogram (MSDS), which eliminates unconscious head movements and extracts the fine-grained articulatory features along the time dimension. To increase diversity, we design multiple data augmentation methods for the training process of the deep neural network (DNN). After the signal processing pipeline, we use a specifically designed end-to-end DNN to recognize silent speech in the general corpus. The front-end of the DNN is a multi-branch convolutional network, extracting high-level features from the MSDS and embedding it into a gesture sequence. The back-end is a Transformer-based sequence-to-sequence network, extracting contextual information from the gesture sequence and translating it into text outputs. In addition, we use a two-branch user discriminator to enable user-adaptive learning on sequential tasks. The overall system performance is evaluated in a general corpus dataset of 1,000 different sentences through the widely used metric of word error rate.

# 4 SIGNAL PROCESSING DESIGN

In this Section, we present the signal processing pipeline of mSilent, including the clustering-selection algorithm to separate articulatory gestures from multi-path reflections, the multi-scale detrended spectrogram to extract fine-grained features, and multiple data augmentation methods to increase data diversity.

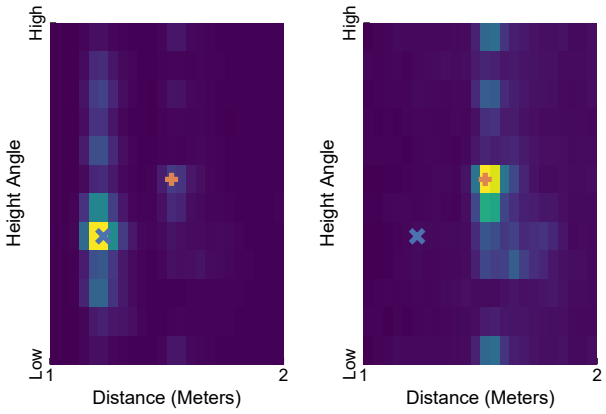
## 4.1 mmWave Signal Capture

Most COTS mmWave radars transmit frequency-modulated continuous-wave (FMCW) signal with the transmitting antenna array and receive the reflected signals with the receiving antenna array [18]. By applying Fast Fourier Transform (FFT) on the FMCW chirps, the signals reflected from different distances are separated into different range bins. In this work, we use the TI IWR1843 77 GHz radar as the sensor, which provides a bandwidth of 4 GHz and a range resolution of 4 cm. For angle resolution, the mmWave radar has three transmit antennas and four receiving antennas so that it can form an 8-antennas virtual line array in one direction, while the number of antennas in the other direction of the virtual array is only 2. Therefore, we can only achieve a suitable angle resolution in one direction. To capture the articulatory gestures, we choose to perform 1D beamforming along the vertical direction, which separates the signals reflected from different height angles, such as the head and legs. After standard mmWave signal processing, we get a 3D complex matrix $W$ of shape $T_w \times R_w \times A_w$, where $T_w$ is the number of FMCW chirps (the time dimension), $R_w$ is the number of points in chirp (the range dimension), $A_w$ is the number of beamforming bins (the angle dimension). Along the time dimension, we repeat the chirp 240 times per second so that we can capture 240 frames of range-angle maps each second. Along the range dimension, we set the number of ranges bins, $R_w$, to 256, so that we can resolve objects within 10 meters with a granularity of 4cm. Along the angle dimension, we set the number of angle bins, $A_w$, to 12 so that we have an angular resolution of 10 degrees along the height direction.

## 4.2 Clustering-Selection Algorithm

The long sensing range of mmWave radar eliminates the need for the user to hold the device, so the user can operate the steering wheel or tap the keyboard while speaking. However, long-range sensing also introduces severe multi-path effects as reflections from a wide range are mixed together. Unlike large-scale gestures [39] or periodic heartbeats [14], the movements of articulatory gestures are both weak and complex, *i.e.* without specific periodical patterns, which makes multi-path separation difficult. The subtle articulatory gestures can be easily overwhelmed by irrelevant movements or noise from nearby people. In addition, existing deep learning-based human posture tracking systems can only work in environments without obstructions such as tables, chairs, and laptops [27]. To address this problem, we design a clustering-selection algorithm by intuition that the whole body movement can be easily detected and articulatory gestures mostly appear at the top of the body. We choose to use a lightweight design so that the algorithm has few parameters and does not need training. The algorithm first clusters the peak reflection points in the dynamic spatial profile using non-maximal suppression and selects the zone of articulatory gestures by the location prior. The details of the clustering-selection algorithm are shown in Algorithm 1 and elaborated in the following.

*4.2.1 Dynamic Spatial Profile.* We first produce a dynamical spatial profile that highlights movements at different ranges and angles. The maximum range and angle resolution of 4 cm and 10 degrees of our mmWave radar are not sufficient to separate articulatory gestures very close to each other. Therefore, the signal related to the articulatory gestures may reside in a small zone consisting of several channels and we need to select a such zone to separate articulatory gestures from inferences. Since the articulatory gestures are complex and the signal is non-stationary without periodicity, it is difficult to select the articulatory zone using time-domain algorithms such as circular pattern matching [14]. Also, the small articulatory gestures cannot cause a big change in the range map, so cannot be detected by the differentially removing clutter method. To select the articulatory zone, we first




(a) Tapping the keyboard. (b) Hands Stable.

Fig. 3. Zoomed dynamic spatial profile of silent speech. Points "×" and "+" highlight the positions of hands and head, respectively.

remove the reflections from static objects by applying a bandpass filter, since the Doppler frequency of the signal represents the speed of reflector movements. We then perform amplitude accumulation along time-dimension to obtain the dynamic spatial profile $D$:

$$D(r,a) = \sum_{t} |W_f(t,r,a)|, \tag{1}$$

where $W_f$ is the filtered signal and $D$ is a real matrix of shape $R_w \times A_w$. Each peak in the profile represents a possible movement, so a trivial approach is to fine-tune the frequency band of the filter to maximize the energy of articulatory gestures in the profile. However, the speed of articulatory gestures varies greatly for different syllables and the speed range overlaps with that of other human movements [48], which makes filter design difficult. Furthermore, even with the filter gain, the weak signal of articulatory gesture can still be easily overwhelmed by other large-scale movements. Therefore, we choose to use a broadband filter to retain all possible movements and leave the zone selection problem to the clustering-selection algorithm.

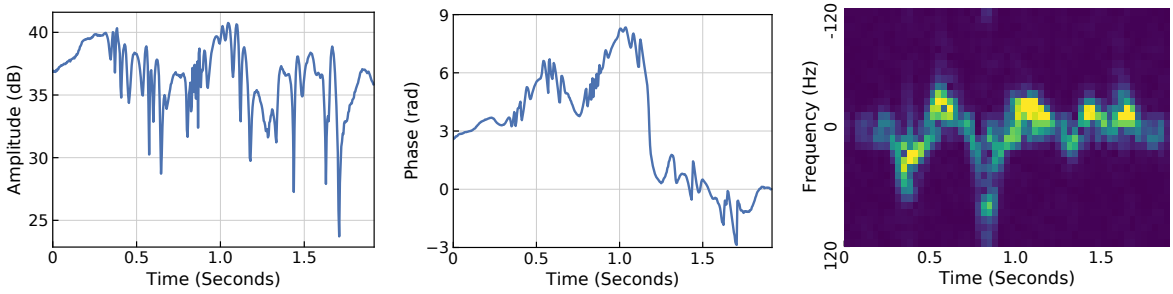
*4.2.2 Clustering Movements by Non-Maximal Suppression.* We use the non-maximal suppression (NMS) [33] algorithm, which is widely used in target detection, to cluster the movements. As shown in Fig. 3, if there are irrelevant gestures in the sensing range, such as the user typing, the subtle articulatory gestures are easily affected by environmental noise. We solve the problem with the intuition that the whole human body can be easily detected with clustering and articulatory gestures mostly appearing at the top of the body. Specifically, in each step of the clustering, we select the point with maximum energy among unclustered points as the new cluster center and gather all points in its neighborhood into this cluster. The size of the neighborhood $R_n, A_n$ can be calculated based on the spatial resolution of the radar and the average size of the human body.

After clustering, each cluster represents the body movement of one human or the environmental noise, so we calculate the sum of intra-cluster energy and remove clusters with significantly lower energy. The subtle articulatory gestures are mixed and clustered with the large-scale irrelevant gestures in the same cluster, which enhances the robustness to environmental noise. If there are still multiple clusters in the sensing range, we select the cluster at the minimum distance as the target cluster, because the user is closer to radar than irrelevant people

**Algorithm 1:** Clustering-Selection Algorithm

**Input:** The dynamic spatial profile $D$, the neighborhood size $R_n, A_n$, the threshold $\alpha$, the zone size $R_z, A_z$.
**Output:** The selected articulatory zone $Z$ of shape $R_z \times A_z$

Find all peaks $P = [(r_1, a_1), (r_2, a_2), \cdots]$ in $D$
Sort the peak list $P = [(r_i, a_i), \cdots]$ in descending order of energy $D(r_i, a_i)$
Initialize the cluster list $C$, the clustered peak list $PP$, and the energy table $E$
**for** $(r, a)$ **in** $P$ **do**
    **if** $(r, a) \in PP$ **then**
        **continue**
    **end**
    Initialize the new cluster $c$
    Put $(r, a)$ to $PP$ and $c$
    **for** $(rr, aa)$ **in** $P$ **do**
        **if** $(rr, aa) \in PP$ **then**
            **continue**
        **end**
        **if** $|r - rr| \leq \frac{R_n}{2}$ **and** $|a - aa| \leq \frac{A_n}{2}$ **then**
            Put $(rr, aa)$ to $PP$ and $c$
        **end**
    **end**
    Set $E(c)$ to the sum of energy $\sum_{(r_i,a_i)\in c} D(r_i, a_i)$.
    Put $c$ to $C$
**end**
Find the maximum energy $e_{max}$ in $E$
Remove all clusters $c_i$ from $C$ where $E(c_i) < \alpha \cdot e_{max}$
Find the cluster $c$ with the minimum average range $(\frac{1}{|c|} \cdot \sum_{(r_i,a_i)\in c} r_i)$ in $C$
Sort the peak list $c = [(r_i, a_i), \cdots]$ in descending order of energy $D(r_i, a_i)$
Initialize the peak list $cc$.
**for** $(r, a)$ **in** $c$ **do**
    Initialize the artifact flag $b = 0$
    **for** $(rr, aa)$ **in** $c$ **do**
        **if** $rr == r$ **then**
            $b = 1$
            **break**
        **end**
    **end**
    **if** $b == 0$ **then**
        Put $(r, a)$ to $cc$
    **end**
**end**
Find the zone center $(r_z, a_z)$ with the maximum height angle $a_z$ in $cc$
**return** the articulatory zone $Z$ of shape $R_z \times A_z$ centered in $(r_z, a_z)$




(a) Amplitude. (b) Phase. (c) Spectrogram.

Fig. 4. Articulatory representations of silent speech "In fact."

in typical scenarios. In this way, we can perform a coarse-grained detection of the user body and then select the articulatory zone using heuristics. In complex scenarios which have large-energy environment noise, the clusters can be further selected by well-established human localization or tracking algorithms [56], but we leave the study of these algorithms as future works.

*4.2.3 Selecting Articulatory Zone.* After obtaining the cluster of the user body, the next step is to distinguish the articulatory gestures from irrelevant gestures. We use the location prior that the articulators are at the top of the body in normal body postures. Theoretically, the angle resolution of the line array is not sufficient to distinguish between the jaw and head at a distance of more than 1 m, so the point of articulatory gestures has the maximum height angle in the cluster. However, due to the limited number of antennas available on the COTS radar, the dynamic spatial profile has severe artifacts. As shown in Fig. 3, there are artifact peaks above the point with the highest energy, which may interfere with the selection process. Therefore, we add more restrictions on the points in the cluster, and only the points with the same distance are kept with the largest amplitude. After filtering, we select the point with the maximal height in the cluster as the center of articulatory gestures. Finally, we select the $R_z \times A_z$ zone in $W$ as the multi-channel signal representing articulatory gestures, where $R_z$ and $A_z$ are the numbers of spatial channels. The number of spatial channels is a trade-off between maximizing the articulatory information and minimizing the environmental information, so the more noise sources and the closer distance, the smaller zone should be. We set $R_z = A_z = 3$ based on experiments.

## 4.3 Multi-Scale Detrended Spectrogram

Recognizing silent speech in the general corpus requires fine-grained articulatory gesture information to eliminate uncertainty. However, the duration of a single articulatory gesture is highly variable [58], while unconscious head movements [15] introduce low-frequency noises. To address this problem, we design a multi-scale detrended spectrogram (MSDS) to remove low-frequency noise by segmented linear detrending, and perform short-time Fourier transforms (STFT) at multi-scale to extract fine-grained features.

*4.3.1 Articulatory Representation.* The spatial resolution of mmWave radar is not sufficient to separate the articulatory gestures that are very close to each other, *e.g.*, the movements of the mouth and throat. Therefore, the key articulatory features can only be extracted from the time-dimensional variations in the mmWave signal. Recent ultrasound-based work, such as UltraSE [46], believes the articulatory features can be mainly represented by Doppler shifts and use STFT spectrograms as features. However, other systems, such as SoundLip [59], choose

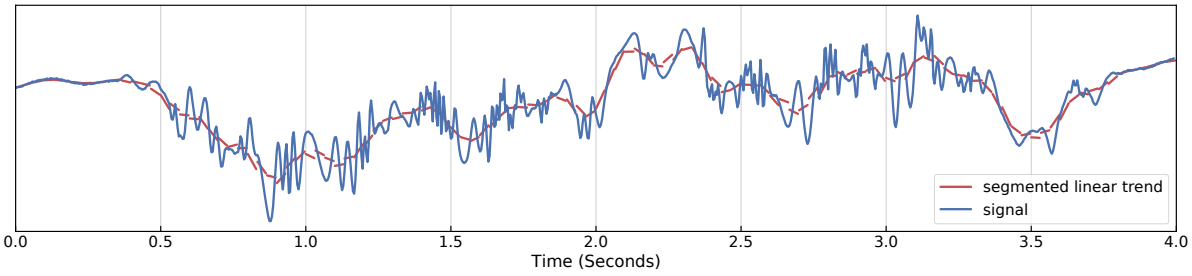



Fig. 5. Signal and segmented linear trend of silent speech "You say you like watching and playing baseball." (Non-overlapping imaginary part only.)

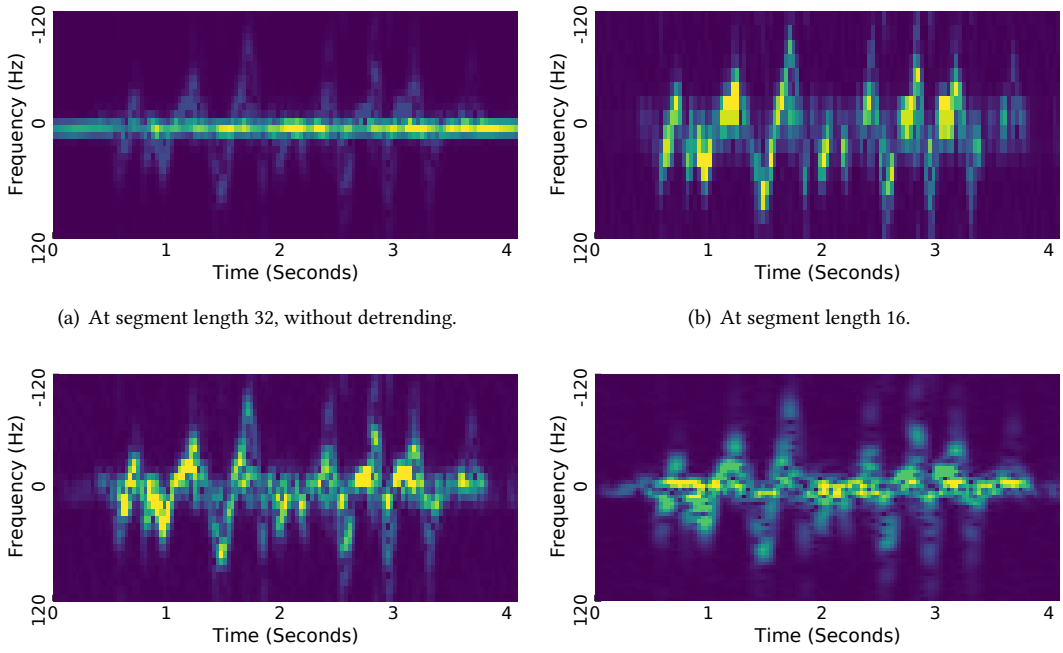
the amplitude-phase as the articulatory representation. After carefully studying the articulatory waveforms of mmWave radar, we believe that the STFT spectrogram is the better representation of SSR. The reasons are as follows:

First, the resolution of the spectrogram is sufficient for the silent speech recognition task. In mmWave sensing, the amplitude-phase features are good at recovering fine radial movements such as heartbeats [14], because the phase retains the fine-grained time and movement distance information. However, recovering the moving time and distance of each gesture is not vital for the task of recognizing silent speech texts. As we can observe from Fig. 4, the millimeter to centimeter scale movements of articulatory gestures [48] introduce large Doppler shifts in mmWave, which provide detailed information of the gesture while removing the irrelevant jitters in phase and amplitude.

Second, the spectrogram are easier to be interpreted than amplitude-phase features, therefore reducing the learning difficulty of DNN. As we can observe in Fig. 4, it is hard to distinguish the syllables of the simple sentence "In fact." from the amplitude and phase waveform, while it is easy to discern the articulatory gestures from the spectrogram. The articulatory gestures are complex and non-radial, making it difficult to extract features from obscure amplitude-phase waveforms without a massive dataset. In contrast, the STFT spectrogram is a well-designed feature extractor, which explicitly characterizes the speed distribution of the movements over time and is more easily understood by the model. It is worth mentioning that the model can acquire more information than the human eye in the spectrogram because the color mapping conceals part of the fine-grained fluctuations.

*4.3.2 Segmented Linear Detrending.* The head of the user moves unconsciously while speaking [15], and this movement is superimposed on the articulatory gestures sensed by radar. Fig. 6(a) shows that the unconscious head movements introduce high-intensity noises to the low-frequency band of the spectrogram. A common practice is to directly cut out the low-frequency band to avoid the failure of normalization [46]. However, the speeds of articulatory gestures are distributed over a wide range and also contain important low-frequency components [48]. Recognizing silent speech in the general corpus is a challenging task that requires a large number of fine-grained features. Therefore, we need to design suitable denoising methods to remove the noise while preserving the low-frequency components of articulatory gestures. Since the speed and intensity of irrelevant movements are constantly changing, calculating the difference in the spectrogram [30] cannot remove the noise. Furthermore, performing differential operations on the original signal will destroy the Doppler characteristics. In addition, there is no fixed frequency distinction between articulatory gestures and unconscious head movement to guide the design of a band-pass filter.

In mSilent, we choose to use segmented linear detrending to remove the noise. This is because we observe that the unconscious head movement is slowly varying, so the trend of the introduced signal is smooth. In contrast,




(a) At segment length 32, without detrending.

(b) At segment length 16.

(c) At segment length 32.

(d) At segment length 64.

Fig. 6. Spectrograms of silent speech "You say you like watching and playing baseball." The multi-scale detrended spectrogram consists of (b)(c)(d).

the duration of articulatory gesture is only 100 ∼ 700 ms [58], so the introduced signal trend is fast-changing and non-smooth. Therefore, we can use segmented linear detrending to remove the noise. Specifically, for each STFT segment $s$ with a length of $l$, we fit the linear trend of it using the least square method and remove the trend before performing FFT:

$$k^*, b^* = \arg\min_{k,b} |s - [k, b]\, P|, \tag{2}$$

$$s' = s - [k^*, b^*]\, P, \tag{3}$$

where $P_{2\times l} = [p_1, p_2, \cdots, p_l]$, $p_i = [i, 1]^T$, $1 \le i \le l$, $[k, b]$ is the complex coefficients, and $|\cdot|$ is the L2-norm. As shown in Fig. 5, the trend related to the unconscious movements is approximately linear in a short segment, so the least squares method can fit with it while ignoring the curved articulatory gestures. Fig. 6(c) shows the spectrogram after segmented linear detrending, and the low-frequency components are clearly visible.

*4.3.3 Multi-Scale STFT.* There is a trade-off between time and frequency resolution in the spectrogram by changing the segment length of the STFT. Empirically, the segment length should be similar to the duration of articulatory gestures to capture their variation. However, the duration of articulatory gestures varies widely, from 100 ms to 700 ms [58]. When a smaller segment length is used to recognize a short-duration gesture, it will lose information on long-duration gestures due to the low frequency resolution. Similarly, long segment length will lose information about fast changes in short-duration gestures. Therefore, we perform STFT at multiple segment lengths simultaneously to obtain the multi-scale spectrograms. To reduce the complexity of DNN models, we set

the scale of segment lengths as powers of 2. As shown in Fig. 6, we first determine the center segment length as 32 (133 ms) based on experiments. We then add two additional segment lengths of 16 and 64 to obtain a spectrogram biased toward time and frequency resolution, respectively.

We set the remaining parameters of STFT as follows: The average speed of the articulatory gestures is $40 \sim 60$ mm/s [48, 58], which introduces a Doppler shift of around 50 Hz in the 77 GHz mmWave. The maximal Doppler frequency observed in our experiments is 100 Hz, so the sample rate (FMCW chirp rate) should be more than 200 Hz. We set the sample rate to 240 Hz and the interval of STFT to 8 to obtain a frame rate of 30 Hz, which retains the fine-grained features and the compatibility with video-based solutions so that we may fuse the mmWave with videos in the future. After the signal processing pipeline, we finally obtained the MSDS of $[S_{16}, S_{32}, S_{64}]$ as the feature of articulatory gestures, and the shape of each detrended spectrogram $S_F$ is $F \times T \times 3 \times 3$, where $F \in \{16, 32, 64\}$ is the segment length, $T = \frac{T_w}{8}$ is the number of segments.

### 4.4 Data Augmentation

Complex deep neural networks usually require a huge amount of samples for training. Unlike video-based systems that can collect large-scale video samples available on the Internet, mmWave-based systems suffer from high data acquisition costs. The number of sentences in the general corpus is significantly greater than in previous works, so the available samples for each sentence spoken by the users are reduced, which exacerbates the data hunger. Furthermore, the features of multi-dimensional MSDS have different physical meanings than videos so the widely used image augmentation methods, such as crop and flip, cannot be directly applied to MSDS. To address this problem, we design multiple data augmentation methods that consider the physical meanings of MSDS.

(1) **Time warping:** We randomly warp the signal in the time-domain by a factor of $\alpha \in [0.9, 1.1]$, which introduces variation in both time and frequency similar to SoundLip [59].
(2) **Time masking:** Inspired by the idea of time masking in speech recognition [35], we randomly mask 10% of the MSDS in the time dimension, allowing the network to work properly while gesture missing.
(3) **Scale jitter:** Inspired by the idea of scale jitter [42] in compute vision, we multiply the MSDS by a random factor $a \in [0.5, 1.5]$, then add a random bias $b \in [-0.5, 0.5]$ to it.
(4) **Gaussian noise:** We introduce Gaussian noise with a standard deviation of 0.05 to enhance the robustness.
(5) **Frequency independent masking:** As the spectrograms are complementary between different scales, we design a more complex masking strategy in the frequency dimension. The frequency positions masked in different spectrograms are independent, which guides the DNN to learn the complementary relationship more efficiently. The different spectrograms characterize similar features, so the masking ratio should be higher to avoid overfitting. Specifically, for each spectrogram, we independently mask 30% of the data in the frequency dimension with a probability of 50%.

## 5 DNN DESIGN

The general corpus contains a large number of sentences with highly diversified lengths, domains, and syntax, making the task challenging. Unlike video-based approaches that can benefit from computer vision tasks, there is no existing wireless signal-based gesture recognition model that can handle such a complex task. To address this problem, we design a novel deep neural network (DNN) by the intuition that we can first recognize what articulatory gestures the MSDS represents, and then translate the gesture sequence to the final output texts. Since translating spectrograms into a sequence of gestures/texts is a common task in sensing systems, we believe that our DNN design can also benefit other similar applications.

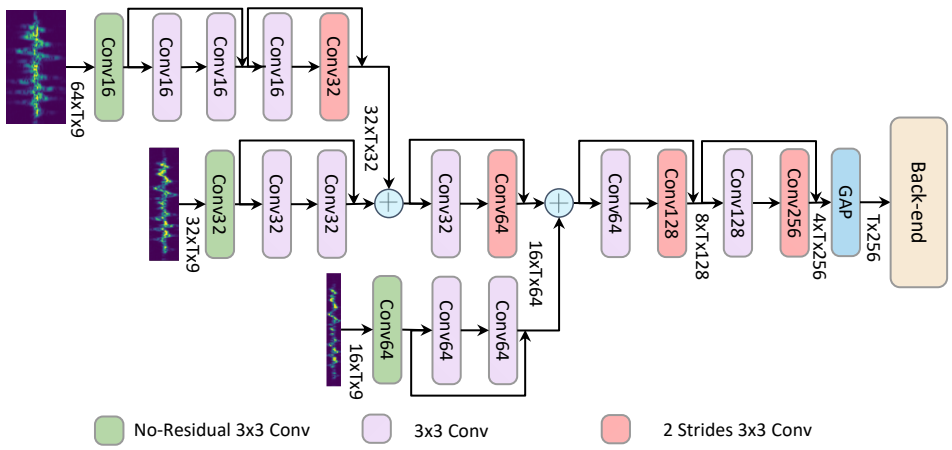



Fig. 7. Architecture of the multi-branch embedding front-end. Data flow notation: $Frequencies \times Times \times Channels$.

## 5.1 Multi-Branch Convolutional Front-end

After obtaining the MSDS, we first design a network to extract high-level short-time features, *i.e.*, what articulatory gestures the MSDS represents. The front-end design is inspired by the property of MSDS that the spectrograms characterize the speed distribution of articulatory gestures over time, and the features in the different spectrograms are complementary. The architecture of the front-end is shown in Fig. 7.

*5.1.1 Residual Time-Frequency Convolution.* We first consider a single network branch with a multi-channel spectrogram of shape $F \times T \times 3 \times 3$, where $F \in \{16, 32, 64\}$ is the number of frequency bins and $T$ is the number of STFT segments. Due to the limitation of radar spatial resolution, the input spectrogram has only $3 \times 3$ spatial channels, so we directly flatten it to $F \times T \times 9$. In the time-frequency plane, the spectrogram characterizes the speed distribution of articulatory gestures over time. Therefore, we use time-frequency 2D convolution (Conv) as the base component of the front-end, which is good at extracting high-level local features. Because the size of dimension $F$ is at most 64 and the long-term dependence is handled by the back-end, we can use a small $3 \times 3$ kernel to achieve a sufficient receptive field. We further introduce the residual connection [16] between every two Conv layers, which allows the gradient to flow directly through the network, making it easier for the network to converge.

*5.1.2 Frequency-Dimension Downsampling.* The speed information of articulatory gestures is distributed in the frequency-dimension, so we downsample in the frequency-dimension level by level to get a full view of each articulatory gesture. In contrast, the front-end should only focus on short-term features, *i.e.*, what articulatory gesture the segment of MSDS represents, instead of long-term dependencies. Therefore, unlike the usual 2D convolutional networks, we do not downsample in the time-dimension. Outputting a long sequence preserves more fine-grained features along time-dimension, which is helpful for the complex task by leaving this to the sequential back-end. The perception field is expanded after each downsampling, allowing for the extraction of larger-scale features, so the number of channels is doubled to extract more features. Through this network, the spectrogram of shape $F \times T \times 9$ is gradually transformed to a feature map of shape $4 \times T \times C$, where $C = 256$ is the number of output channels. Finally, instead of using a fully connected (FC) layer, we use the global average pooling (GAP) to aggregate the output to a gesture sequence $X$ of shape $T \times C$, which reduces the number of parameters in the network and prevent overfitting [28].

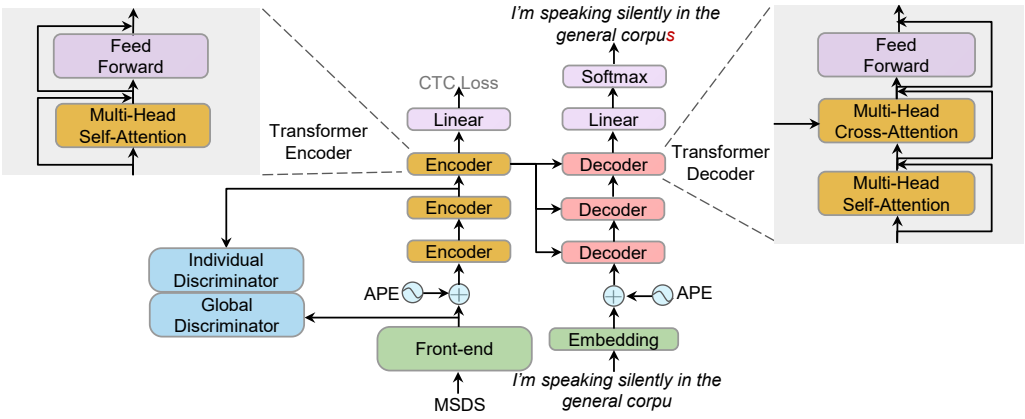



Fig. 8. Architecture of the Transformer-based Seq2Seq back-end.

*5.1.3 Fusing MSDS by Element-wise Addition.* We design a multi-branch structure to handle multi-scale inputs, where each branch processes the spectrograms of each scale independently in the early stage. Fusing features of different modalities, such as video and audio, is often done by concatenating them in the channel dimension [1, 31], so that the complex cross-modality interactions can be learned by subsequent networks. However, because the MSDS is not multi-modal data and does not have complex interactions between scales, the concatenation scheme leads to severe overfitting problems. Fortunately, the MSDS is highly interpretable, which allows us to manually design the fusion scheme. The spectrograms of different scales represent the same speed distribution, but are complementary in terms of resolution. Therefore, we fuse the multi-scale features by element-wise addition, which is a concise and natural way to aggregate complementary data. The element-wise addition forces the same channel of different branches to extract the same features, leading to the same feature space in multiple branches, which is suitable for MSDS. In addition, with the element-wise addition and the residual connection, the gradient can straightforwardly propagate back to the front-end, making the network easier to train. The spectrograms at different scales are complementary in the local details and extracting detailed local information is done by the early Conv layers. So, the fusion stage should be inserted as early as possible to preserve local features.

## 5.2 Transformer-based Seq2Seq Back-end

*5.2.1 Transformer for Long Sequences.* To retain more fine-grained features in time-dimension, the gesture sequence $X$ has a 30 Hz element rate and may have hundreds of elements in length, so the back-end should be able to handle very long-term dependencies. Traditional Recurrent Neural Networks (RNN), such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), require $O(l)$ steps to process the entire sequence, where $l$ is the length of the sequence. The long gradient propagation paths make it difficult to learn the long-term dependencies [17]. Transformer [49] breaks the temporal serial structure of traditional RNN and uses the multi-head attention (MHA) mechanism to extract contextual features. Each attention is calculated by scaled dot production:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{4}$$

where $Q$, $K$, and $V$ are projected from input sequences by Linear layers, and $d_k$ is the dimension of $K$. Attention handles the whole sequence with $O(1)$ complexity and has no problem learning long-term dependencies. After MHA, Transformer uses a feed-forward network to map the sequences to a high-dimensional space to further

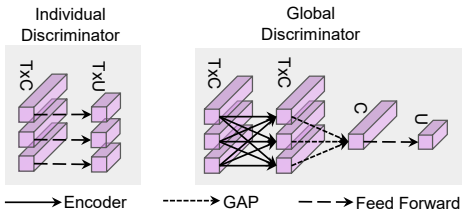



Fig. 9. Data flow of the two-stage sequential discriminator.

enhance its expressive power. Therefore, we use the powerful Transformer to build the back-end, which is expertise in extracting long-term dependencies.

*5.2.2 Seq2Seq Structure.* The sequence-to-sequence (Seq2Seq) structure contains two parts: encoder and decoder. The encoder models the contextual information of gesture sequence $X$ by the self-attention, and transforms it into a hidden sequence as one of the inputs of the decoder. The decoder is auto-correlated, *i.e.*, accepts its last output text $Y_{l-1} = [y_1, \cdots, y_{l-1}]$ as the other input at step $l$. The decoder calculates the masked self-attention of $Y$ to learn an internal language model in the corpus, which makes the Seq2Seq structure perform better [43]. The decoder then models the correlations between the two inputs by cross-attention, and outputs the $Y_l$ at step $l$, until $y_l$ is the end-of-sentence (EOS) symbol.

Due to the large vocabulary size in the general corpus, we use the character as the basic symbol unit of the output. We also introduce left-to-right alignment via the auxiliary CTC loss [23] to avoid the excessive flexibility of the Transformer [31, 59]. The best size of back-end is correlated with the size of the corpus and dataset [21], so we fine-tune it in our dataset and set it to 256-channels, 3-layers encoder, and 3-layers decoder. It is worth mentioning that the structure of back-end is scalable, and can work in a larger corpus by simply increasing the size.

## 5.3 Exploring User Adaptive Learning on Sequential Task

*5.3.1 User Adaptive Learning.* In daily conversation, users may speak more casually than giving commands to the voice assistants. This leads to a high diversity across users since different speakers have different speech speeds, different pause times, and different accents. Even if the pronunciations are the same, different users may have different mouth shapes and tongue movements. Therefore, recognizing the silent speech of new users makes the SSR task even more difficult. User-adaptive learning by gradient reversal [9] is a widely used method for the classification task, guiding the DNN to learn user-independent features through a discriminator. Through user-adaptive learning, the model can be transferred to new users using unsupervised calibration with unlabeled data, or semi-supervised calibration with a small number of labeled data. However, there is no general approach to applying user-adaptive learning to sequential tasks. We explore the extension of this approach to our sequential tasks by designing a two-stage sequential discriminator.

*5.3.2 Two-Stage Sequential Discriminator.* In classification tasks like word-level SSR [37, 61], the DNN has a heavyweight front-end for feature learning and a lightweight back-end (only one or two FC layers) for classification, so the discriminator is usually inserted after the front-end to capture high-level user information in features. However, in our sequential task, the powerful back-end is both for contextual feature learning and sequential translating. If we insert the discriminator earlier, it cannot capture the high-level user information in the context. Otherwise, when the discriminator is inserted later, the powerful back-end can translate the gestures into text before it, just as the auxiliary CTC loss [23], even though this text may be wrong due to user diversity.

Our solution is to divide the user diversity into two categories and design a two-stage sequential discriminator to handle them separately, including an individual discriminator and a global discriminator. Fig. 8 shows the position of the two-stage sequential discriminator in the DNN, and Fig. 9 shows its data flow. Each articulatory gesture contains user diversity information, such as the different shapes of mouths and moving habits. The individual discriminator processes the output from the front-end, a gesture sequence $X$ of shape $T \times C$. Each gesture vector $x$ in the sequence has a shape of $1 \times C$ and is independently passed through the feed-forward network to ensure that no vector contains user diversity information. The output of the individual discriminator has a shape of $T \times U$, where $U$ is the number of users (including an unknown user). The global discriminator is inserted after the penultimate layer of the encoder and handles user diversity in the context, such as differences in speech rate and accent. The first part of the global discriminator is a Transformer Encoder, which can extract the contextual diversity in sequence. Then, a GAP layer is used to pool the hidden sequence of shape $T \times C$ to a vector of shape $1 \times C$, and the pooled vector is fed into a feed-forward network for classification. The loss of the adaptive learning component is as follows:

$$L_{adaptive} = \frac{1}{T}\sum_{i=1}^{T} L_i + L_{global}, \tag{5}$$

where $T$ is the length of sequence, $L_i$ is the $i_{th}$ loss of individual discriminator, $1 \leq i \leq T$, and $L_{global}$ is the loss of global discriminator.

In summary, the DNN is trained in an end-to-end manner and the loss is as follows:

$$L = L_{seq2seq} + \lambda_1 L_{ctc} + \lambda_2 L_{adaptive}, \tag{6}$$

where $L_{seq2seq}$ is the Seq2Seq loss, $L_{ctc}$ is the auxiliary CTC loss, $L_{adaptive}$ is the user adaptive loss, $\lambda_1$ and $\lambda_2$ are the weights of auxiliary loss. Since the auxiliary loss should be smaller than the main loss, the factors are less than 0.5 in practice.

## 6 IMPLEMENTATION

### 6.1 Bimodal General Corpus Dataset

*6.1.1 Corpus Design.* To evaluate the performance of mSilent in the general conversation scenarios, we use the conversation section of the English listening test as the corpus source, which is general, representative, and widely used for evaluating communication skills. The pilot ultrasound SSR, EchoWhisper [10], also uses the same corpus source, but only selects dozens of words from one conversation due to the insufficient recognition capability. Instead of manually selecting, we automatically collect over 100k transcripts from multiple websites containing the corpus source, slice them into sentences by punctuation, and randomly select 1,000 sentences to form the corpus. The large corpus size with random sampling method ensures that the corpus does not contain subjective bias and is sufficiently general for the daily conversations [3]. Tab. 2 shows ten example sentences in the corpus, which are all common sentences of daily life, and vary in domains. As shown in Fig. 10, the most common words in our corpus are general words in daily conversation. Fig. 11 shows the long-tail distribution of sentence lengths in the corpus, which is as same as the natural language in real life.

*6.1.2 Bimodal Data Collection.* We develop a collecting tool, which can simultaneously collect the sensing data from a mmWave radar and a web camera with synchronized timestamps. We collect the dataset in four different scenarios, including one conference room, two pantry rooms, and one restaurant. There is at least one other person and multiple static objects such as tables, chairs, and computers in these scenarios. We invite 10 volunteers to participate in the collection, including 7 males and 3 females. The volunteers sit opposite to radar at a distance of about 1.5m. Volunteers control the recording and sentence segmentation by pressing the start and stop buttons, which introduces irrelevant hand movements from the user. Volunteers speak casually as they would in daily

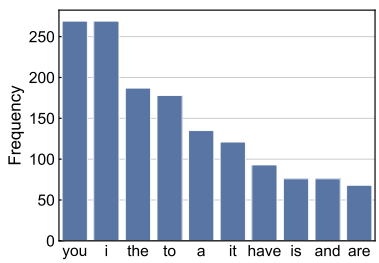
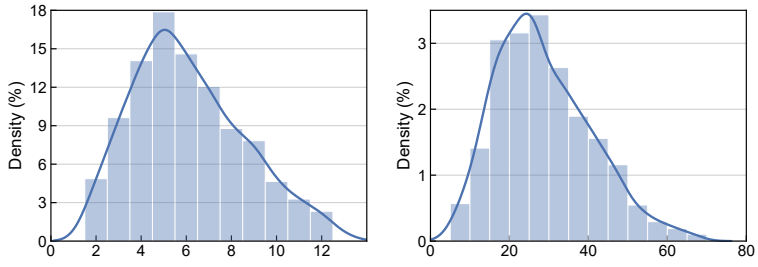



Fig. 10. Top 10 of the most common words.




(a) Words. (b) Characters.

Fig. 11. Distribution of sentence lengths.

Table 2. Examples of sentences.

| |
|---|
| We practice only one night a week. |
| Let me look at the map. |
| But i just can't spare the time. |
| What is she wearing today? |
| We have got all kinds of mobile phone here. |
| I decided to make new legs for myself |
| Do any other people in your family use the bus service? |
| Anything else? |
| You say you like watching and playing baseball. |
| No smoking is allowed in the lift. |

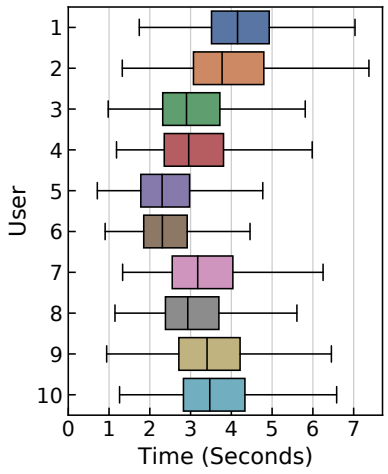



Fig. 12. Sample duration of different users.

conversation, which may increase the diversity between users. As shown in Fig. 12, the sample duration of different users varies greatly. For each volunteer, we only collect a random set of sentences that is half of the corpus, *i.e.*, 500 different sentences, to ensure the diversity of the dataset. Each volunteer speaks each sentence four times in total, including 1 normal speech sample and 3 silent speech samples. In each session, we invite a volunteer and randomly select a scenario to collect 500 $\sim$ 1000 samples. This means that the samples from each volunteer are collected in 2 $\sim$ 4 sessions with random time and scenarios. We drop out the last user as the unseen user, and randomly split the dataset above into the training set, validation set, and test set in the radio of 8:1:1. The random splitting is done by the built-in function of the Sklearn library and does not have any manual restrictions.

*6.1.3 Challenging Scenario Datasets.* To evaluate the generalizability and capability of our system, we collect five datasets in challenging scenarios, which are all not included in the training set. The five challenging scenarios include: 1) a new room, which has a completely different layout from the previous rooms; 2) a noisy meeting room, in which dozens of people are having a workshop, and the distance between the volunteer and the nearest

person is about one meter; 3) in a car, where the volunteer sits in the driver's seat and turns the steering wheel while speaking; 4) the volunteer wears a mask; 5) in a large room, collected at different distances, angles, and orientations.

We collect 21,404 samples in total. Compared to previous wireless signal-based work, our dataset has a significant improvement in corpus complexity. We reduce the repetitions of each sentence to keep the similar data size, which makes the SSR task more challenging. To the best of our knowledge, our dataset is the first bi-modality and general corpus silent speech dataset, providing the possibility to fairly compare the performance of mSilent with state-of-the-art video solutions and enabling future exploration for multi-modal systems.

### 6.2 Implementation Details

We use TI IWR1843 77GHz mmWave radar to transmit mmWave signals and the DCA1000EVM data acquisition board to collect the raw signal data. The signal processing pipeline is implemented using Pytorch and Scipy, and the DNN model is implemented by Pytorch Lightning. For the DNN model details, we choose to use ReLU as the activation function, and batch-normalization and layer-normalization for the normalization layer of front-end and back-end, respectively. The sizes of parameters for the DNN front-end and back-end are 1.3 M and 5.6 M, respectively. For the training process, we use a batch size of 64 on each GPU and use Adam [26] with a 1e-4 learning rate as the optimizer. The remaining hyperparameters used in our experiments are as follows: $R_n = 25$, $A_n = 6$, $\alpha = 0.05$, $\lambda_1 = 0.1$; $\lambda_2 = 0.1$ in user adaptive learning, otherwise $\lambda_2 = 0$.

## 7 EVALUATION

### 7.1 Evaluation Setup

*7.1.1 Baselines.* We use the state-of-the-art SSR systems based on wireless signal and video as the baselines. To our best knowledge, SoundLip [59] is the only work that explores wireless signal-based SSR at the sentence-level. To compare with ultrasound-based SoundLip, we keep our cluster-selection algorithm since the selected multi-channel signal has similar characteristics to the multi-frequency continuous wave signal in SoundLip. We further use up-sampling schemes to increase the sampling rate to 480Hz to match the ultrasound signal of SoundLip. To ensure model convergence for SoundLip, we use the same data augmentation strategies as mSilent, except for frequency masking, as there is no frequency dimension in the amplitude-phase map for SoundLip. We also reproduce state-of-the-art video-based SSR [31] systems in LRS2 dataset [43]. To improve the performance of the video-based systems in small datasets, we additionally add color jitter as the data augmentation strategy. Because the best size of the back-end is correlated with the size of corpus and dataset [21], we change the back-end size to be the same (256-channels, 3-layers encoder, and 3-layers decoder) as mSilent for both baselines. We fine-tune the remaining hyperparameters of both baselines to achieve the best performance. The training of mSilent and baselines are run on a GPU server with four Nvidia RTX3070 GPUs, and totally takes over 10 days to raise the best performance.

*7.1.2 Metrics.* We use the word error rate (WER) as the evaluation metric, which is widely used in continuous speech recognition. WER measures the minimum number of operations to transform a predicted sentence into the ground truth by inserting, deleting, and substituting words:

$$WER = \frac{I + D + S}{N}, \tag{7}$$

where $N$ is the number of words in the ground truth and $I$, $D$, and $S$ are the number of insertions, deletions, and substitutions, respectively.

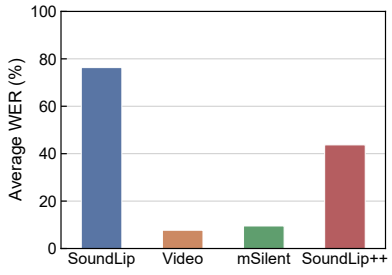



Fig. 13. Overall performance of mSilent and baselines

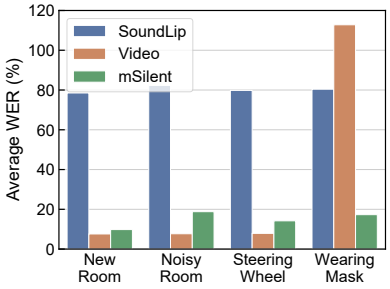



Fig. 14. Performance in new environments

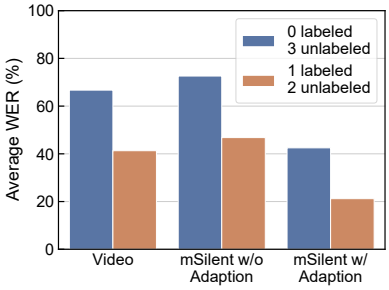



Fig. 15. Performance of user adaptive learning

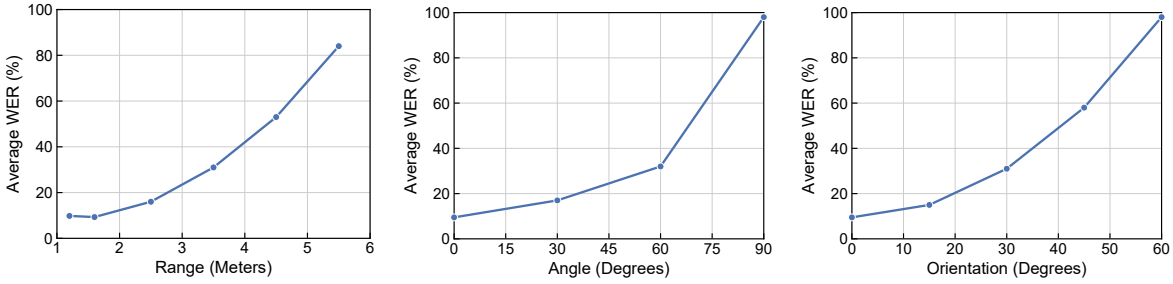
## 7.2 Performance Comparison with Baselines

The performances of mSilent and baselines in the test set are shown in Fig. 13. The average WER of mSilent is only 9.5%, which indicates that mSilent is capable to achieve accurate silent speech recognition in the general scenarios of daily conversation. For voiced and silent speech, the WER of mSilent is 9.9% and 9.4%, respectively. As a comparison, the WER of the video-based baseline is 7.7%, which is only marginally better than mSilent. Compared to the baselines, mSilent achieves wireless-based general corpus silent speech recognition and offers comparable performance with state-of-the-art video-based systems.

To our surprise, the best average WER achieved by SoundLip is 76.3%, which indicates the huge difficulty gap between general daily conversation and its limited 70 commands. As presented in 4.3.1, we believe the success of mSilent is due to the Doppler shift being the key feature in mmWave signals for SSR, which is the inter-modality feature in SoundLip's amplitude-phase map, and cannot be extracted with a fine granularity by SoundLip's heavyweight intra-modality CNN with only lightweight inter-modality CNN. To verify this, we change the structure of SoundLip to lightweight intra-modality CNN with heavyweight inter-modality CNN, and called it SoundLip++. The result shows that the SoundLip++ has 43.1% average WER, providing a side note on our mechanism, and still has a big gap between our MSDS feature and front-end. These results also confirm that, the strong capability of mSilent is not only because of the new sensing technology, but also because of our carefully designed system.

## 7.3 Performance in New Environments

Fig. 14 shows the WER of mSilent in the challenging environments that are all not included in the training set. The WER in the new room is only 9.8%, which is better than expected. This shows that mSilent is robust against variations across environments and sessions. The WER in a noisy conference room is 18.8%, which shows that the noise from irrelevant people has little effect on mSilent. We manually check the failure samples and confirm that all the articulatory zones selected by the clustering-selection algorithm are in the right place. Therefore, the small performance loss may be due to other people's articulatory gestures leaking through the FFT. The WER when the user turns the steering wheel is only 14.2%, which shows mSilent can work well with hand movements. This also illustrates the potential for mSilent to be deployed in vehicle scenarios, especially when driving at night. The WER when the user wears a mask is only 17.3%, showing that mSilent can work with masks, while the video-based SSR stops working at all. The small performance loss of mSilent may be due to the noise from the mask movements, which is unseen in the training set. Since these challenge scenarios are not included in the training set, we believe that mSilent generalizes well, and can perform better after further training in such scenarios.




(a) Ranges from radar. (b) Horizontal angles from radar. (c) Horizontal orientations of user.

Fig. 16. Performance in different positions and orientations.

## 7.4 Performance of User Adaptive Learning

Fig. 15 shows the WER for the user who is not included in the training set. Without user-adaptive learning, both mSilent and video-based SSR perform poorly for the new user with a WER higher than 60%. This shows that there is a strong user diversity in daily communication, which is prevalent in both the video and wireless-based silent speech datasets [37, 61]. With unsupervised adaptive learning, *i.e.*, training with three unlabeled samples of each sentence, the WER of mSilent reduces to 42.5%. The performance is further improved by semi-supervised adaptive learning, *i.e.*, training with one labeled sample and two unlabeled samples of each sentence, where mSilent achieves a WER of 21.2%. Without adaptive learning components, the WER of mSilent trained by the same amount of data is only 46.7%. The results show that the two-stage user discriminator successfully extended user-adaptive learning to our sequential task. In the normal dataset, each user reads each sentence four times, so the user adaptive learning reduces the amount of supervised data to 1/4. Because the transfer strategy in practical scenarios is carefully studied in word-level SSR Endophasia [61], which can be applied directly to mSilent, we do not present it.

## 7.5 Impact of Position and Orientation

There are three independent variables (range, angle, and orientation) in the spatial relationship between the user and the radar. In every experiment, we change one variable and set others to the default value, which is the same as in the training set: the user sits directly opposite to the radar (0° angle and 0° orientation) at a distance of 1.5 meters (1.5 m range). Fig. 16(a) shows the WER with different ranges from radar, and mSilent can achieve $\leq$ 20% WER within a distance of 2.5m. The performance loss may due to the insufficient number of antennas for COTS radar, which cannot separate silent speech at longer distances with a high signal-to-noise ratio. Fig. 16(b) shows the WER with different horizontal angles from radar, and mSilent can achieve $\leq$ 20% WER in $-30 \sim 30$ degrees. In addition, mSilent can also work in up to ±60 degrees, which is beyond the ±30 degrees viewing angle of our camera. Fig. 16(c) shows the WER of different orientations of the user, and mSilent can achieve $\leq$ 20% WER in $-15 \sim 15$ degrees. It is worth noting that the training set only contains samples at a distance of 1.5 meters and directly opposite the radar, so mSilent may perform better in the future with a large-scale dataset.

## 7.6 Micro Benchmark

Since the only text is available as the ground truth, we use overall WER as the metric of all micro benchmarks. As shown in Fig. 17, when replacing the element-wise addition to the widely used channel concatenation or removing the data augmentation component, the system stops working at all and has a WER of more than 100%.

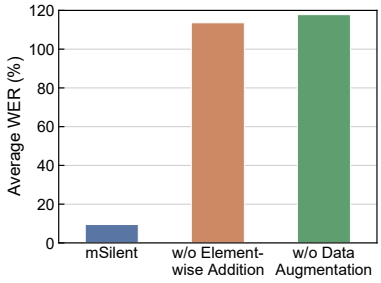



Fig. 17. Failure settings. mSilent stops work at all.

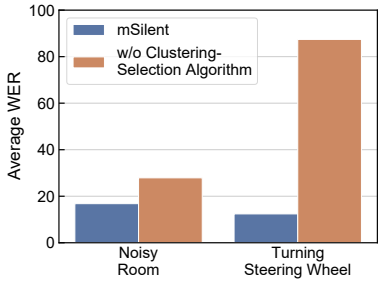



Fig. 18. Performance in complex scenarios.

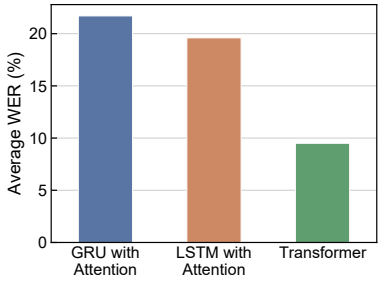



Fig. 19. Comparison with other back-ends.

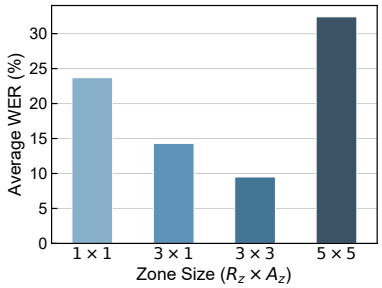



Fig. 20. Comparison with other zone sizes.

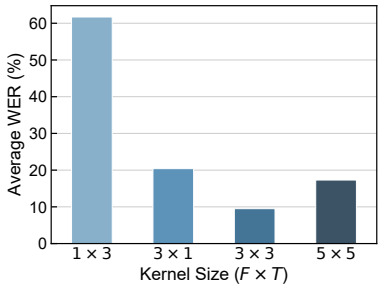



Fig. 21. Comparison with other kernel sizes.

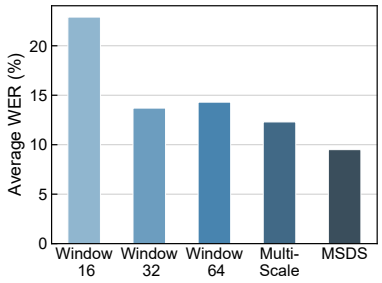



Fig. 22. Comparison with other spectrograms.

This result indicates that element-wise addition is the right way to fuse complementary features, and our multiple data augmentation strategies effectively increase the diversity of data. Fig. 18 shows the performance of mSilent in complex scenarios after removing the clustering-selection algorithm. The zone selection scheme without the clustering-selection algorithm uses a simple maximum amplitude strategy on the dynamic spatial profile. The system performs poorly without the clustering-selection algorithm, especially for the noise room and in-car scenarios. This is because the subtle articulatory gestures are easily overwhelmed by the movements of others in the conference room or by hand movements when turning the steering wheel. Fig. 20 shows the impact of different zone sizes used in articulatory zone selection. Since the range resolution is much higher than the height resolution, we also include the $3 \times 1$ zone size as a valid zone size. The result shows that the $3 \times 3$ zone size is the best trade-off between maximizing the articulatory information and minimizing the environmental information. For the MSDS feature extraction, Fig. 22 shows that the multi-scale approach is better than the single-scale with

Table 3. System Latency (ms)

| | Preprocessing | Processing | DNN | Total |
|---|---|---|---|---|
| w/o GPU | 38.9 | 12.6 | 292.7 | 344.2 |
| w/ GPU | 1.3 | 12.6 | 161.5 | 175.4 |

(a) In-car assistant. (b) Text Entry.

Fig. 23. Experiment scenarios of the case study. The radar is in the red box.

different STFT segment lengths, and the segmented linear detrending further improves the performance. Fig. 21 shows the performance of different Conv kernel sizes in front-end. The frequency-dimension 1D Conv is much better than the time-dimension 1D Conv, which shows the necessity of extracting and aggregating features in the frequency dimension. We can also observe that our time-frequency 2D Conv performs better than both 1D Conv, and a small kernel size is sufficient. As shown in Fig. 19, after replacing the Transformer with SoundLip's GRU with the attention scheme, the WER grows to 21%. Since this WER is still much lower than SoundLip++, the result also confirms that the SoundLip's performance gap is mainly due to the feature mechanism in the signal, not the back-end. The WER of LSTM is slightly lower than GRU, which is consistent with the fact that LSTM usually outperforms GRU on complex tasks. Transformer works better than these RNN structures because both the input and output sequences are long sequences in our task.

### 7.7 System Latency

To evaluate the performance and usability of mSilent in real life, we measure the system latency on a laptop with AMD Ryzen 7 5800H CPU and Nvidia RTX3060 GPU. The inference of the DNN model is based on TorchScript. We divide mSilent into three stages: Preprocessing (Section 4.1), Processing (Section 4.2 and Section 4.3), and DNN (Section 5). The duration of input data is 5 seconds ($1200 \times 256 \times 8$), and the length of output text is 100 characters. As shown in Table 3, the system latency of mSilent is only 344 ms on CPU, and 175 ms on GPU. The preprocessing stage consists of multiple rounds of FFT and beamforming operations, thus can be accelerated by GPU. In the processing stage, both our Cluster-Selection Algorithm and Multi-Scale Detrended Spectrogram are lightweight and fast. The latency of mSilent is mainly introduced by DNN. Because of the auto-correlated decoder, the inference cannot be fully parallelized, and the GPU version is only slightly faster than the CPU version.

## 8 CASE STUDY

To further evaluate the performance of mSilent in the real life, we explore deploying mSilent in two typical use cases. Different from the experiments in 7.3, the system is trained independently by the dataset collected in the

real environment with task-specific corpus. Same as the general corpus dataset, every sentence is read only four times and the ratio of the training set, validate set and test set is 8:1:1.

### 8.1 In-car Assistant

As shown in Fig. 23(a), we deploy mSilent in the vehicle as a noisy-proof and non-disturbance in-car assistant. The driver can talk with mSilent silently to control the in-car devices, such as setting the navigation destination, while the passengers are sleeping/talking. Similar to the common dashcam, the radar is placed at the left of the rearview mirror. We use the dataset collected from the real dialogue of drivers and assistants as the corpus source [8], and randomly choose 500 sentences spoken by the driver to form the corpus. The corpus includes calendar scheduling (Set an optometrist appointment for 3 pm on the 15th of this month with my aunt), weather information retrieval (Could you tell me the weather forecast in Fresno on Saturday), and point-of-interest navigation (Find a gas station that is not farther than 3 miles away), etc. Since the corpus is collected from the real dialogue, the command sentences are much more complex and variable than the previous works like SoundLip [59]. However, the domain of this corpus is focused on the in-car assistant, which is less challenging than our general corpus with no domain restrictions. The experiment result shows that mSilent achieves 4.1% average WER in the real in-car scenario. Compare to other SSR systems, the video-based SSR cannot work in the night driving and introduce more privacy concerns. The previous wireless SSR systems suffer from low recognition capability and cannot support complex interaction in the real large corpus.

### 8.2 Text Entry

As shown in Fig. 23(b), we deploy mSilent as a contact-less and hand-free text entry interface. In scenarios where the voice-based ASR is improper, the user can still use mSilent to input general sentences to the device, such as sending a message. We follow the in-mouth SSR SillentSpeller [25] to use the MacKenzie-Soukoref corpus [44], which contains 500 common phases and is widely used for evaluating text entry technology. Compare to our general corpus, the domain of this corpus is also general, but it only contains short phases with no more than 43 characters. mSilent achieves 5.3% average WER in the MacKenzie-Soukoref corpus at the distance of 1.5 m. For the speed metric of text entry, the average Word per Minute (WPM) of mSilent is 129.6, which is similar to the 115 average WPM of SillentSpeller. The results show that mSilent is not only suitable for complex interaction with voice assistants, but also well capable of general communication and input scenarios.

The results of the case study confirm the performance of mSilent in typical use cases. Since these task-specific corpora contain domain or length restrictions, mSilent can achieve better performance in them. Therefore, with the strong capability proven by the massive experiments in our challenging general corpus, we believe that mSilent would perform well in more daily scenarios.

## 9 DISCUSSION

**Limitation of Corpus:** Video-based datasets can be generated from huge amounts of videos on the Internet, such as BBC speech, so the corpus size of which can reach 100k [43]. This allows the video-based back-end to learn a generalized language model that enables end-to-end recognition for unseen sentences [43]. However, due to the limited number of 1,000 sentences in our corpus and the widely distributed language domain of the general corpus[22], mSilent currently cannot precisely estimate the probabilities of the language model. Therefore, the back-end will not consider an unseen sentence as a legitimate sentence in our corpus. However, our evaluations show that mmWave exhibits a similar articulatory gesture sensing capability as videos. We believe that with the progress in mmWave-based sensing, a larger mmWave-based dataset will emerge in the future. In that case, mSilent could recognize unseen sentences in the same way as video-based methods with a large number of training samples.

**Failure Cases:** There are some scenarios mSilent cannot cover now. When the user is lying down, mSilent on the roof may still work, but mSilent cannot work when the user raises hands above the head, because the head is not at the top of the body. We can avoid this by guiding the posture of the user. mSilent also cannot work when the head is moving fast, *e.g.*, running, or head shaking, which is a common failure case for mmWave sensing. Separating movements by contrastive learning [6] may be a solution, and we will exploit it in our future work.

**Wake-up-word Detection and Multiple Users:** Unlike audio-based systems, mmWave-based systems need to first detect and localize the articulatory zone before detecting the wake-up word. The pre-processing steps may incur a high computational cost and consume a considerable amount of energy. However, our lightweight clustering-selection algorithm can pick out the possible channels of articulatory gestures. We can design specific lightweight models to detect wake-up-word from a small number of channels, and this can draw on well-established wake-up word detection models in speech recognition [47]. mSilent can also be extended to multiple users with moderate modifications because our clustering-selection algorithm treats the human body as a whole and we can detect multiple users given that they are separable by range/angles. We leave the multiple users scenario as our future work.

**77GHz mmWave Signal:** Although mSilent works well when the user wears a mask, there may be a large performance decay when the Line of Sight (LOS) path of the signal is blocked by an interfering user or wall. The mmWave signal can propagate through the wall and may maintain the sensing capability [51], we will explore the non-LOS scenarios in our future work. If there are multiple radars within a scene, they will interfere with each other. Because a single mSilent only uses less than 10% of the time slot, we can deploy multiple radars using time-division multiplexing with a synchronization mechanism. The 77GHz band is restricted to automotive usage and cannot be used indoors in some countries. The 60GHz band is an alternate because their sensing characteristics are similar [4], and our system can be migrated to COTS 60GHz radar, *e.g.*, TI IWR6843, with small changes.

**Multimodal Interaction of mmWave and Video:** In this work, we collect a dataset with both the mmWave and the video modality. We design a mmWave-based SSR that achieves performance comparable with the video-based SSR. Our dataset also offers the possibility of cross-modal interaction in the future, where the fusion of mmWave and video may allow for accurate silent speech recognition in more complex scenarios and corpus. Another possible way is using knowledge distillation [24] to transfer information from the video-based SSR trained in a large-scale dataset, which may reduce the data hunger in mmWave. Bi-modal contrastive learning [32] is also a possible self-supervised approach to extracting articulatory features.

## 10 CONCLUSION

We present mSilent, a COTS mmWave radar-based system that enables wireless signal-based silent speech recognition in the general corpus. We propose a clustering-selection algorithm to separate articulatory gestures and a multi-scale detrended spectrogram to extract fine-grained features. We design an end-to-end DNN to handle the complexity of the general corpus, which includes a multi-branch convolutional front-end and a Transformer-based Seq2Seq back-end. We also design a two-stage discriminator that enables user-adaptive learning on sequential tasks. We build a general corpus of 1,000 different sentences in daily conversation and collect over 21K samples with mmWave and video bi-modality. The evaluation results show that mSilent achieves a comparable performance (9.5% average WER at 1.5m) with the state-of-art video-based approach. We explore deploying mSilent into two typical scenarios of in-car assistant and text entry, and mSilent achieves only 4.1% and 5.3% WER, respectively. We believe that this work demonstrates the potential of mSilent in general daily scenarios and provides the possibility for future multi-modal studies.

## ACKNOWLEDGMENTS

We would like to thank our anonymous reviewers for their valuable comments. This work was supported in part by the National Natural Science Foundation of China under Grant 62272213 and 61872173, and in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University.

## REFERENCES

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep Audio-Visual Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[2] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. 2016. Lipnet: End-to-end Sentence-level Lipreading. *arXiv preprint arXiv:1611.01599* (2016).

[3] Sue Atkins, Jeremy Clear, and Nicholas Ostler. 1992. Corpus Design Criteria. *Literary and Linguistic Computing* 7, 1 (1992), 1–16.

[4] Suryoday Basak and Mahanth Gowda. 2022. mmspy: Spying Phone Calls using mmWave Radars. In *Proceedings of 2022 IEEE Symposium on Security and Privacy (S&P '22)*. 1211–1228.

[5] Dongjiang Cao, Ruofeng Liu, Hao Li, Shuai Wang, Wenchao Jiang, and Chris Xiaoxuan Lu. 2022. Cross Vision-RF Gait Re-Identification with Low-Cost RGB-D Cameras and MmWave Radars. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3 (2022), 1–25.

[6] Zhe Chen, Tianyue Zheng, Chao Cai, and Jun Luo. 2021. MoVi-Fi: Motion-robust Vital Signs Waveform Recovery via Deep Interpreted RF Sensing. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom '21)*. 392–405.

[7] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition. *The Journal of the Acoustical Society of America* 120, 5 (2006), 2421–2424.

[8] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-Value Retrieval Networks for Task-Oriented Dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL '17)*. 37–49.

[9] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *International Conference on Machine Learning (ICML '15)*. PMLR, 1180–1189.

[10] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-based Silent Speech Interface for Smartphone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3 (2020), 1–27.

[11] P Ghane, G Hossain, and A Tovar. 2015. Robust Understanding of EEG Patterns in Silent Speech. In *Proceedings of 2015 National Aerospace and Electronics Conference (NAECON '15)*. IEEE, 282–289.

[12] Google. 2022. Google Soli Products. https://www.atap.google.com/soli/products/

[13] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH '20)*.

[14] Unsoo Ha, Salah Assana, and Fadel Adib. 2020. Contactless Seismocardiography via Deep Learning Radars. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking (MobiCom '20)*. 1–14.

[15] Uri Hadar, Timothy J Steiner, EC Grant, and F Clifford Rose. 1983. Kinematics of Head Movements Accompanying Speech during Conversation. *Human Movement Science* 2, 1-2 (1983), 35–46.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16)*. 770–778.

[17] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. 2001. Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-term Dependencies.

[18] Cesar Iovescu and Sandeep Rao. 2017. The Fundamentals of Millimeter Wave Sensors. *Texas Instruments* (2017), 1–8.

[19] Shekh MM Islam, Naoyuki Motoyama, Sergio Pacheco, and Victor M Lubecke. 2020. Non-Contact Vital Signs Monitoring for Multiple Subjects using a Millimeter Wave FMCW Automotive Radar. In *Proceedings of 2020 IEEE/MTT-S International Microwave Symposium (IMS '20)*. 783–786.

[20] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: "Hearing" Your Silent Speech Commands In Ear. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2 (2022), 1–28.

[21] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2022. Scaling Laws for Neural Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR '22)*.

[22] Adam Kilgarriff and Tony Rose. 1998. Measures for Corpus Similarity and Homogeneity. In *Proceedings of the Third Conference on Empirical Methods for Natural Language Processing (EMNLP '98)*. 46–52.

[23] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint CTC-attention based End-to-end Speech Recognition using Multi-task Learning. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '17)*. 4835–4839.

[24] Yoon Kim and Alexander M Rush. 2016. Sequence-level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP '16)*. The Association for Computational Linguistics, 1317–1327.

[25] Naoki Kimura, Tan Gemicioglu, Jonathan Womack, Richard Li, Yuhui Zhao, Abdelkareem Bedri, Zixiong Su, Alex Olwal, Jun Rekimoto, and Thad Starner. 2022. SilentSpeller: Towards Mobile, Hands-free, Silent Speech Text Entry using Electropalatography. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. 1–19.

[26] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR '15)*.

[27] Hao Kong, Xiangyu Xu, Jiadi Yu, Qilin Chen, Chenguang Ma, Yingying Chen, Yi-Chao Chen, and Linghe Kong. 2022. M3Track: mmWave-based Multi-User 3D Posture Tracking. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '22)*. 491–503.

[28] Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network in Network. *arXiv preprint arXiv:1312.4400* (2013).

[29] Fan Liu, Yuanhao Cui, Christos Masouros, Jie Xu, Tony Xiao Han, Yonina C Eldar, and Stefano Buzzi. 2022. Integrated Sensing and Communications: Towards dual-Functional Wireless Networks for 6G and beyond. *IEEE Journal on Selected Areas in Communications* (2022).

[30] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Linghe Kong, and Minglu Li. 2019. Lip Reading-based User Authentication through Acoustic Sensing on Smartphones. *IEEE/ACM Transactions on Networking* 27, 1 (2019), 447–460.

[31] Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2021. End-to-end Audio-Visual Speech Recognition with Conformers. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '21)*. 7613–7617.

[32] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. 2021. Active Contrastive Learning of Audio-Visual Video Representations. In *Proceedings of the International Conference on Learning Representations (ICLR '21)*.

[33] Alexander Neubeck and Luc Van Gool. 2006. Efficient Non-maximum Suppression. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*. IEEE, 850–855.

[34] Laxmi Pandey, Khalad Hasan, and Ahmed Sabbir Arif. 2021. Acceptability of Speech and Silent Speech Input Methods in Private and Public. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. 1–13.

[35] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proceedings of the 20st Annual Conference of the International Speech Communication Association (INTERSPEECH '19)*.

[36] KR Prajwal, Triantafyllos Afouras, and Andrew Zisserman. 2022. Sub-word Level Lip Reading with Visual Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '22)*. 5162–5172.

[37] Ahmed Rekik, Achraf Ben-Hamadou, and Walid Mahdi. 2016. An Adaptive Approach for Lip-reading using Image and Depth data. *Multimedia Tools and Applications* 75, 14 (2016), 8609–8636.

[38] Himanshu Sahni, Abdelkareem Bedri, Gabriel Reyes, Pavleen Thukral, Zehua Guo, Thad Starner, and Maysam Ghovanloo. 2014. The Tongue and Ear Interface: a Wearable System for Silent Speech Recognition. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers (ISWC '14)*. 47–54.

[39] Panneer Selvam Santhalingam, Al Amin Hosain, Ding Zhang, Parth Pathak, Huzefa Rangwala, and Raja Kushalnagar. 2020. mmASL: Environment-Independent ASL Gesture Recognition using 60 GHz Millimeter-wave Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1 (2020), 1–30.

[40] Neil Shah, Nirmesh J Shah, and Hemant A Patil. 2018. Effectiveness of Generative Adversarial Network for Non-Audible Murmur-to-Whisper Speech Conversion. In *Proceedings of the 19st Annual Conference of the International Speech Communication Association (INTERSPEECH '18)*.

[41] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. In *International Conference on Learning Representations (ICLR '22)*.

[42] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-scale Image Recognition. In *Proceedings of the International Conference on Learning Representations (ICLR '15)*.

[43] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip Reading Sentences in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '17)*. 6447–6456.

[44] R. William Soukoreff and I. Scott MacKenzie. 2003. Metrics for Text Entry Research: An Evaluation of MSD and KSPC, and a New Unified Error Metric. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. 113–120.

[45] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. 2018. Lip-interact: Improving Mobile Device Interaction with Silent Speech Commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. 581–593.

[46] Ke Sun and Xinyu Zhang. 2021. UltraSE: Single-channel Speech Enhancement using Ultrasound. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom '21)*. 160–173.

[47] Raphael Tang, Jaejun Lee, Afsaneh Razi, Julia Cambre, Ian Bicking, Jofish Kaye, and Jimmy Lin. 2020. Howl: A Deployed, Open-Source Wake Word Detection System. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS '20)*. Association for Computational Linguistics, 61–65.

[48] Kristin J Teplansky, Brian Y Tsang, and Jun Wang. 2019. Tongue and Lip Motion Patterns in Voiced, Whispered, and Silent Vowel Production. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS '19)*. 1–5.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems (NeruIPS '17)*.

[50] Michael Wand, Christopher Schulte, Matthias Janke, and Tanja Schultz. 2013. Array-based Electromyographic Silent Speech Interface. In *Biosignals*. 89–96.

[51] Chao Wang, Feng Lin, Zhongjie Ba, Fan Zhang, Wenyao Xu, and Kui Ren. 2022. Wavesdropper: Through-wall Word Detection of Human Speech via Commercial mmWave Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2 (2022), 1–26.

[52] Guanhua Wang, Yongpan Zou, Zimu Zhou, Kaishun Wu, and Lionel M. Ni. 2014. We Can Hear You with Wi-Fi!. In *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking (MobiCom '14)*. 593–604.

[53] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I Hong, Carmel Majidi, and Swarun Kumar. 2019. Rfid Tattoo: A Wireless Platform for Speech Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4 (2019), 1–24.

[54] Haowen Wei, Ziheng Li, Alexander D Galvan, Zhuoran Su, Xiao Zhang, Kaveh Pahlavan, and Erin T Solovey. 2022. IndexPen: Two-Finger Text Input with Millimeter-Wave Radar. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2 (2022), 1–39.

[55] Li Wen, Changzhan Gu, and Jun-Fa Mao. 2020. Silent Speech Recognition based on Short-range Millimeter-wave Sensing. In *Proceedings of 2020 IEEE/MTT-S International Microwave Symposium (IMS '20)*. 779–782.

[56] Chenshu Wu, Feng Zhang, Beibei Wang, and KJ Ray Liu. 2020. mmTrack: Passive Multi-Person Localization using Commodity Millimeter wave Radio. In *Proceedings of 2020 IEEE Conference on Computer Communications (INFOCOM '20)*. 2400–2409.

[57] 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 17th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '19)*. 14–26.

[58] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing Your Voice is not Enough: An Articulatory Gesture based Liveness Detection for Voice Authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)*. 57–71.

[59] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2021. SoundLip: Enabling Word and Sentence-level Lip Interaction for Smart Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1 (2021), 1–28.

[60] Ruidong Zhang, Mingyang Chen, Benjamin Steeper, Yaxuan Li, Zihan Yan, Yizhuo Chen, Songyun Tao, Tuochao Chen, Hyunchul Lim, and Cheng Zhang. 2021. SpeeChin: A Smart Necklace for Silent Speech Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4 (2021), 1–23.

[61] Yongzhao Zhang, Wei-Hsiang Huang, Chih-Yun Yang, Wen-Ping Wang, Yi-Chao Chen, Chuang-Wen You, Da-Yuan Huang, Guangtao Xue, and Jiadi Yu. 2020. Endophasia: Utilizing Acoustic-based Imaging for Issuing Contact-free Silent Speech Commands. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1 (2020), 1–26.