EarHear: Enabling the Deaf to Hear the World via Smartphone Speakers and Microphones

Zhanjun Hao, *Member, IEEE*, Yuejiao Wang, Zhenyi Zhang, and Xiaochao Dang, *Member, IEEE*

***Abstract*—Sign language plays a vital role in communication and learning for individuals with hearing and speech disabilities, serving as a common language for the deaf. Current state-of-the-art sign language recognition methods primarily rely on computer vision techniques, but they have certain limitations, including susceptibility to light interference and privacy concerns. Ubiquitous acoustic sensing provides new possibilities for sign language recognition, leveraging its high resistance to interference and cost effectiveness. However, existing methods face challenges in achieving satisfactory results due to environmental interference and the complexity of sign language recognition contexts. In this work, we propose EarHear, a robust contactless Chinese Sign Language Recognition and translation system. EarHear adopts a differential-Doppler data preprocessing method to cleverly mitigate the interference caused by the environment. To further identify differences in the morphology, speed, and direction of sign language actions and distinguish similar gestures, we propose the vision transformer for sign language recognition, which is able to model the context dependence of long-range features and output indeterminate long sign language sequences using an attention mechanism. As a result, computational speed and recognition accuracy are improved. Moreover, we explore a large-scale language-model-based sign language translation, which enables sign language recognition results to follow natural language standards, thus realizing a true sense of sign language recognition. The evaluation results based on 15 Chinese sentences show that our system achieves an average recognition rate of 93.38% and a BLEU-1 score of 80.73% for sign language translation, reaching the most advanced level in terms of accuracy and robustness.**

***Index Terms*—Acoustic sensing, Chinese Sign Language Recognition (CSLR), Chinese Sign Language Translation (CSLT), human–computer interaction.**

I. Introduction

IN CHINA, about 30 million hearing-impaired individuals use Chinese sign language to communicate, accounting for 2.14% of the country's total population. Sign language is essential for facilitating communication and education for individuals with hearing and speech difficulties, serving as a common language for the deaf. However, there is limited knowledge of sign language among the able-bodied, which makes it challenging for deaf people to face social isolation and communication barriers in their daily lives. To bridge the huge communication gap, sign language recognition is attracting more and more attention from researchers. In this article, we concentrate on Chinese Sign Language Recognition (CSLR) and Chinese Sign Language Translation (CSLT), which aims to automatically generate natural language sequences that correspond to verbal expressions, as shown in Fig. 1.

With the rapid development of IoT sensing technology, sign language recognition systems have emerged. Common sensing techniques include camera based [1], [2], wearable sensor based [3], [4], and RF based [5], [6], [7]. However, these techniques have certain limitations. Camera-based sensing technology is susceptible to light interference and also invades the user's privacy. Wearable sensor-based sensing technology requires users to wear external devices for extended periods, incurring both cost and inconvenience. RF-based sensing technology is susceptible to electromagnetic interference, which reduces the accuracy of behavior recognition.

Compared to the above sensing techniques, acoustic-based sensing has the following three advantages: 1) *contactless:* acoustic sensors do not require skin contact, ensuring a user-friendly experience; 2) *affordable:* data acquisition requires only speakers and microphones, without other sensor devices or additional sensing modules; and 3) *resistant to interference:* compared to camera-based sensing, acoustic-based sensing is less susceptible to light environments. Recent works have shown acoustic signals can be effectively utilized for accurate hand motion tracking and gesture recognition. For example, SonicASL [8] uses the Doppler effect to extract spectral information from reflected echoes, processes it with a CNN+LSTM+CTC deep learning technique to convert it into text, and generates speech for the earphone audio feed. UltrasonicGS [9] employs generative adversarial neural networks as a data augmentation (DA) technique. And by incorporating residual neural networks and Bi-LSTM, UltrasonicGS effectively considers information in both the feature and temporal dimensions, resulting in high-precision gesture recognition. However, these methods ignore the signal fading caused by static environment and other moving object interference. They also fail to design a specialized method

Manuscript received 31 May 2023; revised 9 August 2023 and 9 September 2023; accepted 7 October 2023. Date of publication 31 October 2023; date of current version 7 March 2024. The work of Zhanjun Hao was supported in part by the National Natural Science Foundation of China under Grant 62262061 and Grant 62162056; in part by the Gansu Provincial Department of Education: Industry Support Program Project under Grant 2022CYZC-12; and in part by the Lanzhou City Technology Innovation and Entrepreneurship Talent Project under Grant 2021-RC-81 and Grant 2020-RC-116. *(Zhanjun Hao and Yuejiao Wang contributed equally to this work.) (Corresponding author: Zhanjun Hao.)*

This work involved human subjects or animals in its research. Approval for all ethical and experimental procedures and protocols was granted by the Institutional Review Board (IRB) at Northwest Normal University.

The authors are with the School of Computer Science and Engineering, Northwest Normal University, Lanzhou 730000, China (e-mail: haozhj@nwnu.edu.cn; yuejiaowangiot@126.com; Z467718583@126.com; dangxc@nwnu.edu.cn).

Digital Object Identifier 10.1109/JIOT.2023.3323631

for context dependence and sequencing of sign language recognition. As a result, directly using existing methods cannot obtain satisfactory performance for CSLR and CSLT.

Implementing a robust acoustic CSLR and CSLT system is a nontrivial task due to complicated movements of fingers and interference in signal acquisition. There are three main challenges in designing the system. The first challenge is that the system is vulnerable to noise and other factors, how to effectively reduce the impact of environmental factors to improve system performance. The second challenge is how to build recognition models with long-range context modeling capabilities for accurate gesture behaviors recognition, considering the differences in morphology, speed, and direction of different CSL gestures, and the presence of interference from similar gestures. The third challenge is the inconsistency problem between sign language gestures and natural language sequences, and how to translate the sign language recognition results into natural language sequences corresponding to the verbal descriptions.

In this article, we propose EarHear, a robust CSLR and CSLT system based on acoustic signals transmitted by the smartphone. EarHear can achieve high recognition accuracy under various practical factors and accurately recognize gestures at long distances of up to 120 cm from the smartphone.

In our solution, to eliminate the effect caused by the static environment and other moving object interference, we propose the differential-Doppler (D-Doppler) strategy. Specifically, by analyzing the variance of each frame in the original spectrogram, we find that the smaller the variance, the smaller the energy fluctuation of the frame. Therefore, we select the frame with the smallest variance as the static frame, after that, the difference between each frame in the spectrum and the static frame is processed. In this way, the Doppler image only contains the sign language component for further recognition. Second, we propose vision transformer (ViT) for sign language recognition (ViT4SLR), a sign language recognition model based on ViT, to take into account the long-range context dependencies. We also use the attention mechanism instead of the commonly used CTC loss function to tackle the prediction of variable-length sequences, further optimizing the computational speed and recognition accuracy. Finally, to overcome the challenge of the inconsistency of order between sign language gestures and natural language, we add a CSLT model after ViT4SLR, which aims to translate the gloss sequences obtained from ViT4SLR into natural language sequences that correspond to the verbal descriptions. In addition, we also exploit the pretraining-fine-tuning paradigm not only to alleviate the insufficient acoustic sign language translation data set but also to substantially improve the performance of the CSLT model.

In a nutshell, our main contributions are summarized as follows.

1) We design a novel acoustic-based CSLR and CSLT system for smartphones. To the best of our knowledge, EarHear is the first smartphone-based CSLR and CSLT system using acoustic sensing.
2) We propose the D-Doppler strategy to eliminate the interference of static environment and other moving objects, as well as to enhance gesture-related information.
3) We apply ViT to the field of acoustic sign language recognition for the first time, achieving high-performance results. To enhance speed and accuracy, we utilize an attention mechanism instead of CTC loss for variable length sequence prediction. By fine-tuning the Chinese Grammatical Error Correction (MuCGEC) pretraining model on the acoustic sign language data set, we address sequence inconsistency between sign language actions and natural languages, demonstrating the benefits of the two-stage model.
4) We implement EarHear and conduct extensive evaluations. The experimental results show that our method outperforms state-of-the-art work in terms of accuracy and robustness under various practical impact factors.

II. Related Work

In this section, we review the existing research works relevant to this article. Since our system is an acoustic-based sign language recognition system, the works we review focus on the following two topics: 1) acoustic sensing systems and 2) sign language recognition systems.

A. Acoustic Sensing Systems

In recent years, there has been a large number of works on acoustic sensing. We organize state-of-the-art systems in terms of their applications.

Motion Tracking: LLAP [10] enables trajectory tracking of a finger by extracting the phase changes from a single-frequency continuous-wave (CW) signal. FingerIO [11] achieves accurate tracking of moving objects by transmitting acoustical signals modulated with orthogonal frequency-division multiplexing (OFDM). Covertband [12] is a full-body tracking system that works similarly to FingerIO. The difference is that FingerIO uses the built-in speaker and microphone of a mobile to track fingers while Covertband leverages the speaker and microphone in a home audio system to track human motions. Strata [13] estimates the channel impulse response (CIR) and measures phase changes in each channel tap to track finger motions. CTrack [14] measures the chirp's time of flight to accurately determine the distance from the hand to a built-in speaker array. DMT [15] proposes a hand motion tracking system that uses a Fourier fitting-based method to accurately detect the Doppler shift.

Detection: BreathListener [16] is a fine-grained breathing monitoring system for drivers, capturing breathing procedures using the acoustic signal's energy spectrum density (ESD). OmniResMonitor [17] fully leverages abundant acoustic multipath reflection to monitor a single target's respiration focusing on the challenge scenarios where there are no signals directly reflected from target's chest. MultiResp [18] leverages abundant acoustic signals indirectly reflected from subjects' chests, considering subject difference in terms of respiration rate and respiration phase. UFA [19] is an acoustic-based upper facial action recognition system. It applies time–frequency analysis to derive the time–frequency-domain signal from

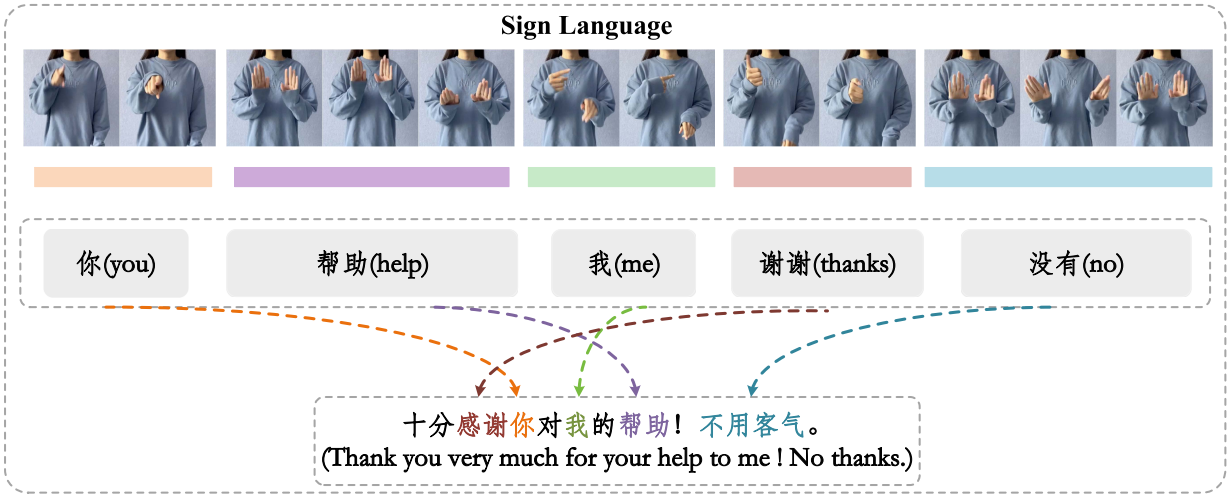


Fig. 1. Overview of CSLR and CSLT.

the channel state information (CSI) for phase change measurement. HearFire [20] consists of a collocated commodity speaker and microphone pair, which remotely senses fire by emitting inaudible sound waves. HearLiquid [21] is a low-cost and nonintrusive liquid fraud detection system that extracts the liquid's acoustic absorption and transmission curve (AATC) over multiple frequencies of the acoustic signal.

Gesture Recognition: RobuCIR [22] solves the frequency selective fading problem caused by multipath effects by periodically transmitting acoustic signals of different frequencies. UltrasonicG [23] utilizes a pair of built-in speakers and microphones to detect the Doppler shift caused by hand gestures. They employ data extensions to overcome the challenge of insufficient training data. EchoWrite 2.0 [24] proposes a lightweight and zero-shot text-entry system for unseen users based on acoustic sensing. However, these systems focus on dynamic activity recognition of single gestures and cannot be extended to complex sign language gesture recognition or applied to special user groups.

B. Sign Language Recognition Systems

Existing sign language recognition applications can be generally divided into three major categories: 1) vision based; 2) wearable sensor based; and 3) wireless based.

Vision-based approaches usually rely on videos captured by cameras to analyze sign language gestures. Papastratis et al. [25] employed a generative adversarial network to enhance recognition performance in sign language conversations by evaluating video encoder predictions and incorporating contextual information. Luqman and El-Alfy [26] evaluated the fusion of spatial and temporal features of different modalities of sign language gestures and facial expressions for sign language recognition using state-of-the-art deep learning techniques. Kraljević et al. [27] proposed a real-time sign language dynamic gesture recognition system for smart home environments, enabling custom sign language gesture commands to interact with the environment. Zhou et al. [28] introduced a sign back-translation approach to address the scarcity of parallel data in Sign Language Translation, presenting a large-scale Sign Language Translation benchmark.

Wearable sensor-based sensing techniques provide less intrusive and privacy-preserving approaches for sign language recognition tasks than vision-based solutions. Lee and Bae [29] employed an artificial intelligence-enabled sign language recognition and communication system comprising sensing gloves, a deep learning block, and a virtual reality interface. Qaroush et al. [30] fused information from 3-axis accelerometer and gyroscope inertial measurement units (IMUs) sensors to automatically detect significant gesture segments for recognition. Lee and Bae [29] developed a data glove-based real-time sign language recognition system that covers both repetitive and nonrepetitive gestures. Glove-based approaches utilize IMU sensors to capture hand sign motions. However, the shape of gloves will bring additional inconvenience to users, which limits the availability of glove shape gadgets in many application scenarios.

Wireless-based method (such as Wi-Fi, millimeter wave, and acoustic wave) has attracted the attention of many researchers in recent years. In particular, the acoustic wave has emerged as a promising wireless technology that requires no additional hardware or complex background environment, while ensuring user privacy. SonicASL [8] utilizes the deep learning technique to convert the spectrums into text and also generates speech as the feed source to the earphone. UltrasonicGS [9] employs a DA method with GAN to feed the multiscale semantic features extracted by residual neural networks into the Bi-LSTM. HearASL [31] uses CIR to represent each sign language gesture and pay attention to conversion movements between two words. However, to the best of our knowledge, the above works utilize acoustic sensing technology for sign language recognition, limited to simple Chinese isolated words or English sign language sentences. EarHear can not only achieve high-performance CSLR with ViT4SLR but also generate natural language sequences matching spoken language descriptions, which realizes acoustic CSLR in a real sense.

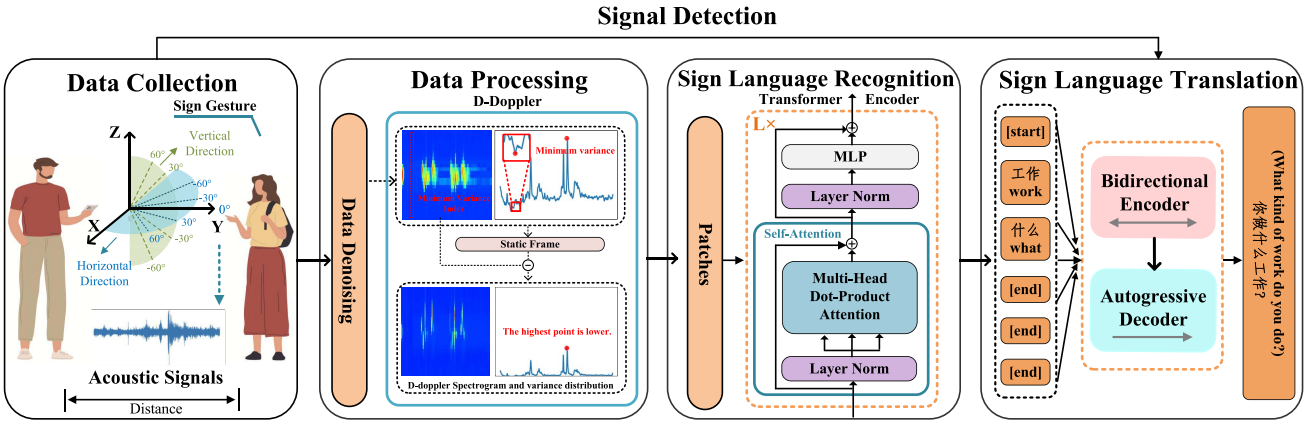


Fig. 2. EarHear pipeline overview.

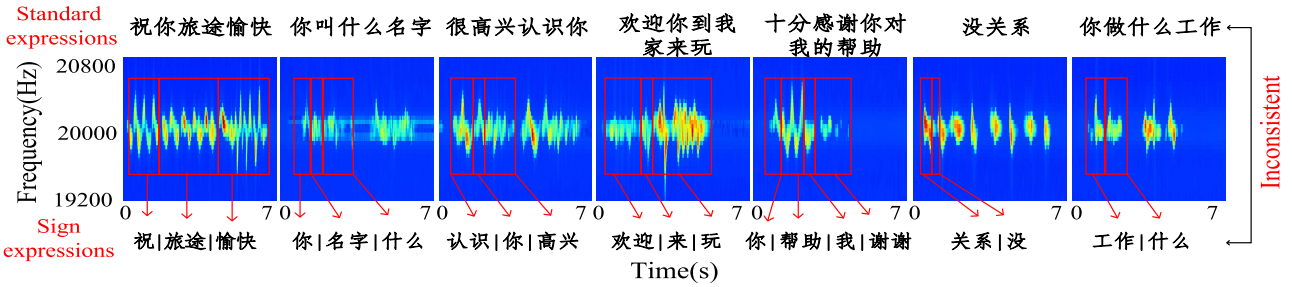


Fig. 3. Schematic diagram of the Doppler effect of inconsistency between seven sign language gestures and natural language norms, namely, “祝你旅途愉快”(Enjoy your trip!), “你叫什么名字” (What’s your name?), “很高兴认识你” (Nice to meet you!), “欢迎你到我家来玩” (Welcome to my home!), “十分感谢你对我的帮助” (Thank you very much for your help!), “没关系” (No matter.), “你做什么工作” (What do you do?).

III. System Design

A. Overview

Fig. 2 illustrates the overview of EarHear. The pipeline of the system consists of four parts: data collection, data processing, sign language recognition, and sign language translation. During the data collection phase, the smartphone’s speaker acts as a transmitter, sending a single audio signal at 20 kHz, while the microphone acts as a receiver, recording and storing the original echo signal.

In the data processing phase, the clutter is first removed using a Butterworth bandpass filter and a bandstop filter. Then, a short-time Fourier transform (STFT) is applied to convert the time-domain signal into a frequency-domain signal. Next, a Gaussian filter is employed to smooth the resulting image. Finally, the D-Doppler strategy is implemented to eliminate the interference of static environment and other moving objects. In the sign language recognition phase, ViT4SLR is proposed for sign language recognition, which uses an attention mechanism instead of the commonly used CTC loss to deal with the prediction problem of variable length sequences. In the sign language translation phase, the gloss sequences obtained from ViT4SLR are translated into natural language sequences matching the verbal descriptions by fine-tuning the MuCGEC pretraining model on the acoustic sign language data set. Next, we will describe each section in detail.

B. Data Collection and D-Doppler-Based Signal Processing

Data Collection: The frequency of ambient noise typically ranges from 1000 to 4000 Hz [32]. To mitigate the impact of ambient noise on the experimental signal, the speaker emits a single audio signal of 20 kHz. The single audio signal offers significant advantages in terms of low complexity and high resolution for Doppler shift analysis [33]. Therefore, we leverage the Doppler shift to conduct a comprehensive analysis of the received signal. Figs. II-B and 4 show the schematic diagrams of the Doppler effect for seven sign language gestures that are inconsistent with the natural language norm and eight sign language gestures that are consistent with the natural language norm, respectively. By analyzing the Doppler spectrogram, distinct patterns corresponding to different sign language gestures can be observed.

Signal Processing: First, to eliminate noise interference, we implement a Butterworth bandpass filter with a frequency range between 19 and 21 kHz. Second, considering that the frequencies around 20 kHz are affected by the audio source signal, we use a band-stop filter with a frequency range from 19 985 to 20 015 Hz. Subsequently, the STFT is employed to extract Doppler frequency shift information caused by gestural movements. STFT has demonstrated satisfactory performance in analyzing acoustic signals in recent studies [34], [35], [36], [37], [38]. This technique involves dividing the input signal into segments of equal length and calculating the FFT

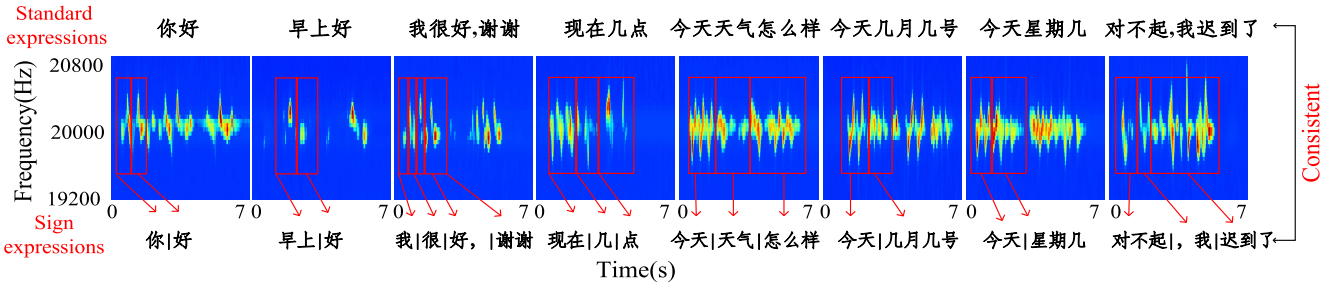


Fig. 4. Schematic diagram of the Doppler effect of consistency between eight sign language gestures and natural language norms, namely “你好” (Hello!), “早上好” (Good Morning!), “我很好谢谢” (I’m fine, thanks.),“现在几点” (What time is it now?),“今天天气怎么样” (What’s the weather like today?), “今天几月几号” (What month is it today?), “今天星期几” (What day is today?),“对不起,“我迟到了” (Sorry I’m late.).

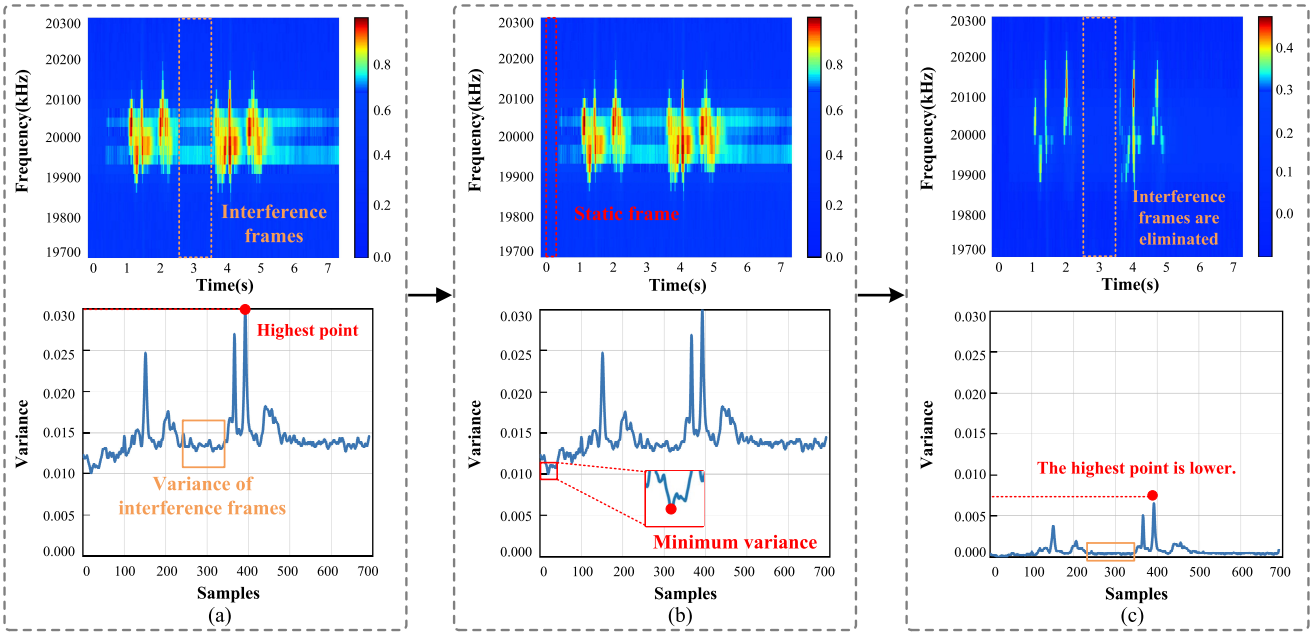


Fig. 5. Differential-Doppler process diagram. Each subplot contains a spectrogram and its corresponding variance variation diagram. (a) Example of interference frames in orange. (b) Process of selecting static frame. (c) Visualizes the effect of interference elimination by D-Doppler.

coefficients for each segment individually. To strike a balance between temporal and frequency resolution, the frame length is set to 8192, and the window step is set to 2048. Next, the Doppler synthetic frequency shift is calculated by (1) to estimate the frequency interval of gesture activity after signal reflection

$$\Delta f = f_0 \times |1 - \frac{v_s \pm v_f}{v_s \mp v_f}| \tag{1}$$

where f_0 is the transmit signal frequency (20 kHz), v_s is the speed of sound (340 m/s), v_f is the gesture movement speed (maximum movement speed 4 m/s), so the synthesized frequency shift is about 470.6 Hz, and the effective frequency range should be narrowed between 19 530 and 20 470 Hz.

Finally, to suppress noise and achieve a smoother output, a Gaussian filter is used. The filter’s weight distribution approximates a bell-shaped curve, with the center pixel receiving the highest weight and the weights of surrounding pixels gradually decreasing, making it well suited for image smoothing. For a 2-D spectrogram, the Gaussian function in (2) is used to obtain the spectrogram shown in Fig. 5(a)

$$G(x, y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \tag{2}$$

where x is the distance of the horizontal axis from the origin, y is the distance of the vertical axis from the origin, and σ is the standard deviation of the Gaussian distribution.

To accurately extract dynamic sign language feature components from the received signal and eliminate the interference of static environment and other moving objects, we propose the D-Doppler strategy based on the Doppler spectrogram. Specifically, we first compute the variance for each column of the Doppler matrix in Fig. 5(a) using (3) to obtain the corresponding variance variation plot. It is widely recognized that smaller variances indicate less energy fluctuation within a frame, implying lower disturbance from static environments and other moving objects. Therefore, we consider the frame with the smallest variance as the static frame and find its corresponding subscript in the Doppler spectrogram shown in

Fig. 5(b) using

$$\mathrm{var}_j = \frac{1}{n}\sum_{i=1}^{n}\left(D_{i,j} - \bar{D}_j\right)^2 \tag{3}$$

$$k = \mathrm{argminvar}_j \tag{4}$$

where D represents the matrix of the Doppler spectrogram, m and n represent the number of samples of the matrix, and the dimension of each sample, respectively. $\bar{D}_j$ represents the mean of the jth column of the matrix, k represents the index of the feature with the minimum variance, and $i = \{1, 2, \ldots, n\}$ and $j = \{1, 2, \ldots, m\}$.

Next, we apply (5) to make the difference between each frame in the spectrogram and the static frame. This process effectively eliminates interference from the static environment and other moving objects while minimizing the impact of multipath effects. The resulting spectrogram based on the D-Doppler strategy is depicted in Fig. 5(c). It can be observed that only the Doppler shift caused by the gestural movements is preserved in the figure. Meanwhile, the overall trend of the variance variation plot is smoother, and the Doppler shift caused by the sign language gesture becomes more pronounced. This observation reconfirms the effectiveness of our proposed D-Doppler strategy in mitigating static environment and other moving object interference, leading to improved accuracy in extracting dynamic sign language feature components. Finally, we use (6) to normalize all features to eliminate the scale differences between different features

$$X_{i,j} = D_{i,j} - D_{i,k} \tag{5}$$

$$X_{i,j} = \frac{X_{i,j} - \mathrm{minX}_{\cdot,j}}{\mathrm{maxX}_{\cdot,j} - \mathrm{minX}_{\cdot,j}} \tag{6}$$

where X represents the matrix after performing the D-Doppler operation.

C. Sign Language Recognition

Currently, the majority of research on acoustic-based sign language recognition [8], [31] relies on the LSTM series of algorithms that model current, past, and future time frames but struggle to effectively capture long-range features. Additionally, these methods commonly rely on time-consuming CTC algorithms to predict variable-length sign language sequences, limiting their deployment. To balance accuracy and inference speed,we propose ViT4SLR which utilizes a straightforward one-stage model built on a ViT. Fig. 6 shows the model architecture of ViT4SLR in detail.

The architecture of ViT4SLR is similar to the original transformer introduced by Vaswani et al. [39]. The main difference is that we only use the encoder part of the transformer, and the final output is an indeterminate length sequence. In particular, each input spectrogram $x \in R^{H\times W\times C}$ is reshape into a sequence of flattened 2-D patches, each patch is noted as $x_p \in R^{N\times P^2C}$. The image dimension is $H \times W$ with C channels while the patch dimension is $P \times P$. The resulting patch sequence length is N. Then, to further extract the features embedded in patches to satisfy the transformer encoder input, linear projection is executed to convert patches into N 1-D tokens $z_i \in R^D$, where D is the dimensions of the

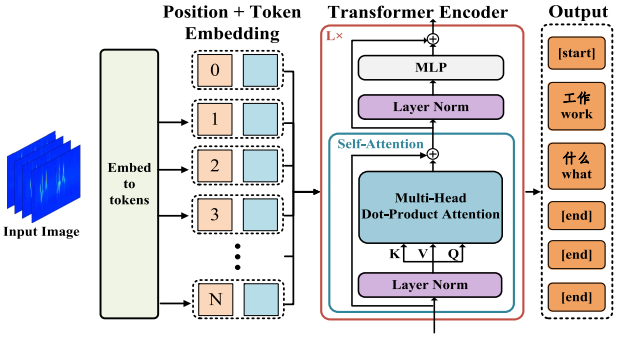


Fig. 6. Model architecture of ViT4SLR.

feature embedding. In addition, a learned positional embedding, $p \in R^{N\times D}$, is added to the tokens to retain positional information, as the subsequent self-attention operations in the transformer are permutation invariant. Then, these tokens are fed into a transformer encoder consisting of L transformer layers.

In the encoder layer, each input undergoes layer normalization (LN) to normalize the feature vectors. The multi-head self-attention layer (MSA) is then applied to determine the relationships between the feature vectors. It has been shown that employing multiple heads, as opposed to a single head, enables the model to jointly attend to information from different representation subspaces at different positions. Vaswani et al. [39] found that the number of heads, denoted as H, can significantly impact the performance of the model. Following MSA, the multilayer perceptron (MLP) is utilized to perform feature extraction. The input to the MLP is also normalized with LN. It consists of two layers, each with GELU activation [40]. Finally, a residual connection is inserted between the output of LN and MSA/MLP to facilitate information flow through the block. This process is shown in the (7) (8)

$$y^l = \mathrm{MSA}\left(\mathrm{LN}\left(z^l\right)\right) + z^l \tag{7}$$

$$z^{l+1} = \mathrm{MLP}\left(\mathrm{LN}\left(y^l\right)\right) + y^l \tag{8}$$

where z^l indicates the input of the transformer block of the l_{th} layer, y^l indicates the result of the l_{th} layer after self-attention operation, as well as z^l indicates the output of the transformer block of the l_{th} layer, which is also the input of the transformer block of the $l + 1_{th}$ layer.

During the output stage, the original ViT employs learnable class embeddings to predict object categories based on the corresponding output vectors. However, the result of sign language recognition is an uncertain length gloss sequence. Therefore, ViT4SLR needs to be capable of variable length sequence output. To achieve this, we extract multiple feature vectors from the encoder instead of just one output vector. In the training stage, we set the labels uniformly to the maximum text length in the data set plus 2 for [start] and [end] tokens. The [start] token denotes the beginning of the text, while the [end] token signifies the end or a space. To indicate that there is no further content after the text characters, we repeat the [end] token until the maximum sequence length

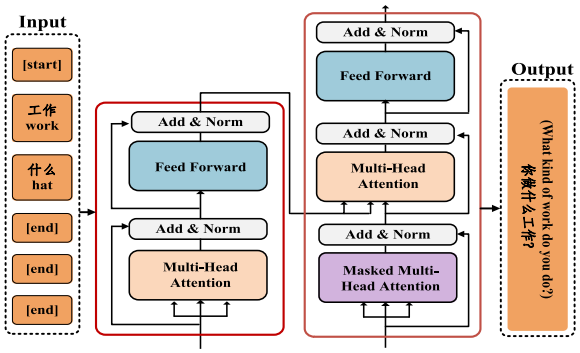


Fig. 7. Model architecture of CSLT.

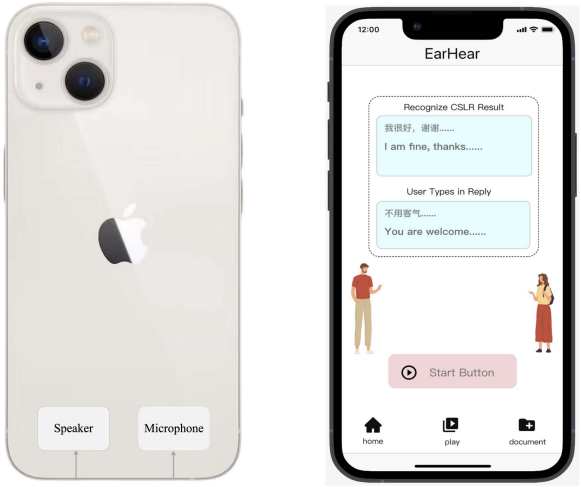
is reached. This serves as a marker to indicate the end of the predicted text. In the inference stage, we can restore the real prediction results based on the special token in the model output, thus realizing the variable-length sequence prediction. Fig. 6 provides an example of the output sequence structure.

D. Sign Language Translation

After the sign language recognition stage, the gloss sequence representing the sign language gestures can be obtained. Glosses are considered the fundamental units of sign language expression, representing single or multiple gestures performed by the signer. In the task of generating sentences from spectrograms through continuous sign language recognition, the resulting sentences composed of glosses should ideally maintain the same order as the gestures in the spectrograms. However, the gloss sequences obtained by continuous sign language recognition alone may not only differ from the grammar in natural language but also differ significantly in order. To obtain natural language sequences matches with verbal descriptions, we add the CSLT model after ViT4SLR.

Sign language translation is a text generation task, and BART [41] absorbs the respective features of BERT's [42] bidirectional encoder and GPT's [43] left-to-right decoder and builds on the standard seq2seq transformer model, which makes it more suitable for text generation scenarios. Therefore, we choose BART as our baseline model. As shown in Fig. 7, the encoder in the BART model uses a bidirectional transformer structure that represents each word in the input sequence z^{l+1} as a vector, and this vector not only depends on the current input word but is also influenced by the entire input sequence. The decoder in the BART model uses an autoregressive transformer structure. When generating each word, the model takes into account the words already generated earlier, which are entered as the context in the next time step. Hence,the features of BART allow it to better understand the context and generate coherent sign language translation results.

Achieving an exceptional text translation model typically necessitates a sufficiently large data set to support it. However, in the field of acoustic sign language translation, there is no data set of sufficient size. To address this limitation, we use a well-trained model on the multireference multisource evaluation data set for the MuCGEC [44] for fine-tuning. The MuCGEC data set consists of 7063 sentences collected from three Chinese-as-a-Second-Language learner sources. It contains five major and 14 minor error types. Each sentence is corrected by three annotators, and their corrections are carefully reviewed by a senior annotator, resulting in 2.3 references. Specifically, we initialize the model weights with the pretraining weights of the MuCGEC model. The gloss sequences serve as the input for the sign language translation model, while the ground truth is a Chinese sentence that adheres to natural language specifications. The model output is optimized to be as close to the ground truth as possible while maintaining good generalization capabilities. Following the fine-tuning principle of the BART model, we train the source encoder in two stages. In the first stage, we only update the encoder, the positional embeddings, and the self-attention input projection matrix of the first layer of the BART encoder. In the second stage, we perform a small number of iterations to train all model parameters. Once the training is complete, we obtain a sign language translation model that performs well in acoustic sign language recognition.



(a) (b)

Fig. 8. (a) iPhone 13 smartphone. (b) WeChat application UI design diagram.

IV. Experimentation and Evaluation

A. Experiment Setting

Hardware Configuration: In the experimental phase, an iPhone 13 smartphone (with IOS 16.2 OS, an A15 CPU, and 4-GB RAM) with built-in dual microphones and dual speakers is chosen as the data collection device. As shown in Fig. 8(a), two speakers act as signal transmitters (Tx), and two microphones act as signal receivers (Rx). We set speakers to send a continuous single audio signal at 20 kHz, set the sampling rate of the microphones to 48 kHz, and train the ViT4SLR and CSLT models on a server equipped with an Intel(R) Xeon(R) Gold 6226R, 64-GB RAM, and an NVIDIA Tesla T4 graphics card.

Software Implementation: We design a WeChat application to demonstrate the effect of sign language recognition in real-time. By clicking the start button, the speaker sends inaudible acoustic signals, and the microphone receives the reflected

echo signals and transmits the data to a remote server. On the remote server, we use the trained model to predict the data and return the results to the user. To identify the start and end of sign language, we calculate the percentage of gesture regions (nonblue regions in the spectrogram) in each 1-s interval of the data set. The minimum percentage value obtained is 3.26%, which we define as the wake-up threshold δ. During real-time sign language recognition, the system calculates the percentage of gesture regions in the spectrogram per second and considers the start of the sign language once it exceeds the threshold δ. Until the percentage of gesture regions is less than δ for 2 s consecutively, the sign language is considered to be finished. After that, we feed the spectrogram from the start-to-end interval into the sign language recognition and sign language translation model. Finally, the recognition results are displayed on the WeChat application, Fig. 8(b) shows our user interface.

Data Collection: We invite four male volunteers and four female volunteers totally to participate in this study as sign language performers (aged between 25 and 40 years old) in four scenarios: laboratory, apartment, corridor, and outdoors. The experimental scenario is shown in Fig. 10(a). None of these volunteers is deaf, and CSL is not their main language for daily communication, but they all perform an hour-long Chinese sign language practice. In addition, the sign language performers performed each sign language gesture 20 times under four practical influencing factors, including distance, speed, noise, and angle, and a total of 1200 data are collected for the experiment. In the training phase, we augment the data by a factor of 20 using two DA methods based on traditional image transformations and GAN [45].

B. Overall Performance Evaluation

1) Training Data Set Size: It is well known that different sizes of training data sets can lead to different learning results of machine learning models [46]. To investigate the impact of the number of training samples on model accuracy, we conduct two sets of experiments:1) using 60%, 70%, 80%, and 90% of the entire collected samples for each experimenter as the training data set and keeping the remaining samples as the test data set and 2) choosing the same training data set as in experiment 1 and the same 8% samples as the test data set.

In Fig. 9, the red and blue parts present the results of experiment 1, while the orange and green parts present those of experiment 2. Experiment 1 shows increasing the number of training samples significantly improves EarHear's accuracy in word and sentence recognition. At the same time, experiment 2 confirms the change in accuracy is not attributable to variations in the size of the training data set (or test data set), but rather an improvement in the model's capabilities. To balance the performance and generalization ability of the model, we chose 80% of the samples for training the final model, achieving a word-level accuracy of 96.69% and a sentence-level accuracy of 92.45%. However, it is worth noting that the accuracy of sentence-level recognition is slightly lower than that of word-level recognition. There are various levels of uncertainty with the word transitions (e.g., time intervals),

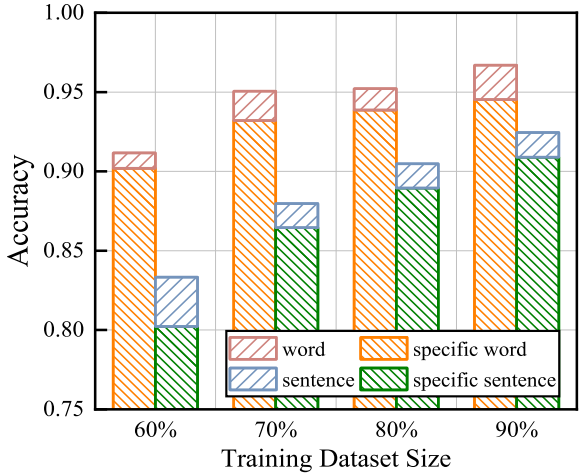


Fig. 9. User-dependent recognition accuracy of word-level and sentence-level CSL with increasing training data set size.

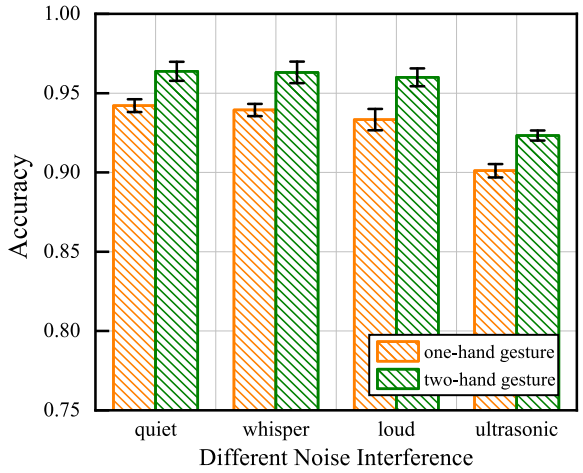


Fig. 10. User-dependent recognition accuracy of one-hand gesture and two-hand gesture with the different noise.

which will affect the performance. Therefore, it is inevitably more challenging to achieve satisfactory sentence-level CSL recognition than word-level recognition. To provide a truly useful assistive communication tool for CSL signers, CSL recognition systems must consider and evaluate sentence-level recognition performance. This part was largely missed in prior studies reported in the literature.

2) One-Hand Gesture and Two-Hand Gesture: In order to investigate the recognition performance of one-hand and two-hand gestures in different noise environments, we asked experimenters to perform one-hand and two-hand sign language gestures in a laboratory scenario with no noise, low-frequency noise, high-frequency noise, and 19-kHz ultrasonic noise environment at a distance of 40 cm from the smartphone, and the experimental results are shown in Fig. 10.

Analyzed from the perspective of noise, the recognition accuracy of one-hand sign language gestures is about 92.9%, while that of two-hand sign language gestures is about 95.3% under four different sound frequencies: 1) no noise; 2) low-frequency noise; 3) high-frequency noise; and 4) 19-kHz ultrasonic noise. This shows that noise has minimal

interference with the experiment and further verifies the effectiveness of the proposed D-Doppler-based data processing method in noise removal. In addition, in an environment with the same sound frequency, the recognition accuracy of two-hand sign language movements is 2.4% higher than that of the one-hand sign language movements. It is due to the larger amplitude of two-hand gestures, which involve the movement of body parts such as arms, thus causing a stronger Doppler shift and enhancing the distinctness of the gesture features in the spectrogram.

3) Different Environments: Multipath effects caused by reflections from unrelated objects have proven to be a major challenge in the field of acoustic sensing. When the smartphone's speaker sends an acoustic signal and the microphone acts as the receiver to record and store the original echo signal, the echo signal contains reflections from surrounding static objects and dynamic pedestrian activity. To evaluate the performance of EarHear in recognizing gestures in real environments, we set up four environmental scenarios commonly used for daily communication, including three indoor environments, such as an apartment, laboratory, and corridor, and one outdoor environment. The experimental scenarios and results are shown in Fig. 11(a) and (b).

In Fig. 11(a), we chose a laboratory with an area of 8 m $\times$ 10 m, an apartment of 3 m $\times$ 6 m, a corridor of 10 m $\times$ 20 m, and an open outdoor area to test the anti-interference capability of our proposed sign language recognition system, respectively. We recruited two experimenters, one performing sign language gestures and the other holding a smartphone, and the sign language performer performed 15 sentences containing 32 words, each repeated 20 times.

In the training phase, we used the data recorded in the laboratory scenario, while in the testing phase, we used the data recorded in the four environments. Fig. 11(b) illustrates the performance of the test set, showing the EarHear system achieves an average accuracy of 96.56% at the word level and 94.33% at the sentence level. Although there are regularly distributed equipment with tables and chairs in the lab, beds and tables in the apartment, walls in the hallway, and high-frequency noise and active people outdoors, four environments have minimal impact on the accuracy of sign language recognition, with differences within 0.38%. This further verifies that our proposed D-Doppler-based data processing method can improve the system's anti-interference capability.

C. Robustness Evaluation

To investigate the robustness of EarHear in different application scenarios, we evaluated it comprehensively in terms of perceived distance, perceived speed, interaction angle, similar gestures, DA, and individual differences in behavior.

1) Perceived Distance: Since the intensity of sound waves is inversely proportional to the square of the distance between two objects, the farther the distance between objects, the more likely it is to affect the accuracy of gesture recognition. In order to balance recognition accuracy with the comfort of communication and to find the optimal perceptual distance, we asked six experimenters to perform sign language gestures at 20, 40, 60, 80, and 120 cm from the device. Three of the experimenters were sign language performers and the other three were hand-held devices. The experiment required all sign language performers to execute 15 sign language sentences, which could use different hand shapes, different movements, and different orientations. The experimental scenario is shown in Fig. 12(a).

The experimental results in Fig. 12(b) show that when the interaction distance between two experimenters is in the range of 20–40 cm, the average accuracy of gesture recognition is the highest, reaching 94.8%. As the interaction distance increases, the recognition accuracy gradually decreases. At distances of 80–120 cm, the recognition accuracy fluctuates around 88%. The reason for this phenomenon is that the interaction distance is too small and the signal reflected from the hand cannot be fully received by the microphone, resulting in a 1.41% lower recognition accuracy than at the optimal distance. When the distance is too large, the acoustic intensity decays sharply, which significantly weakens the system's ability to recognize CSL. Therefore, we recommend users choose a relatively close distance for sign language communication to obtain better recognition accuracy.

It has been widely proven that the propagation distance of sound signals increases with higher power. Based on our analysis, we conjectured that increasing the power of the speaker can effectively improve the accuracy of the system in recognizing sign language gestures at a distance. To validate the conjecture, we chose a smartphone with a higher powered speaker and conducted the experiment with the same settings as before. The results showed that the recognition accuracy was 90.45% in the distance range of 80–120 cm, which is 2.45% better than before. It demonstrates the applicability of our approach to different communication distances. Therefore, when the distance between experimenters is large, we can improve the recognition accuracy by using higher powered hardware devices. However, considering the limitation of battery capacity in smart devices, we suggest that the distance between users should not be too far to ensure satisfactory recognition accuracy.

2) Perceived Speed: We conduct experiments to evaluate the impact of gesture speed. Due to the difficulty in directly measuring gesture speed, we use the gesture execution time as a proxy. Two subjects are recruited to perform the same sign language sentence at five different speeds. From Fig. 12(c), the highest average accuracy of 95.1238% is achieved when the gesture duration is between 3 and 4 s. When the gesture duration is too short, although the Doppler shift is enhanced, the microphone struggles to capture the complete signal. On the other hand, when the gesture duration is too long, the signal change caused by the Doppler shift is slight. The experimental results demonstrate that the best performance can be achieved in a range of 3–4 s.

3) Interaction Angle: In practical applications, it is not always possible to maintain a strictly fixed angle in a conversation. To investigate the effects of horizontal and vertical angles on recognition performance, we invite six experimenters, with

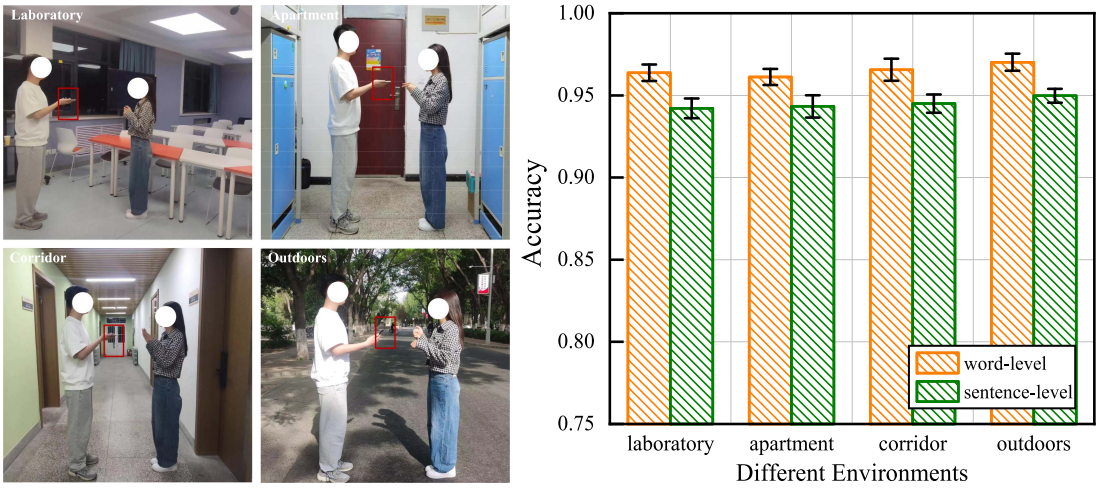


Fig. 11. (a) User study in the four different environments. (b) Recognition performance.

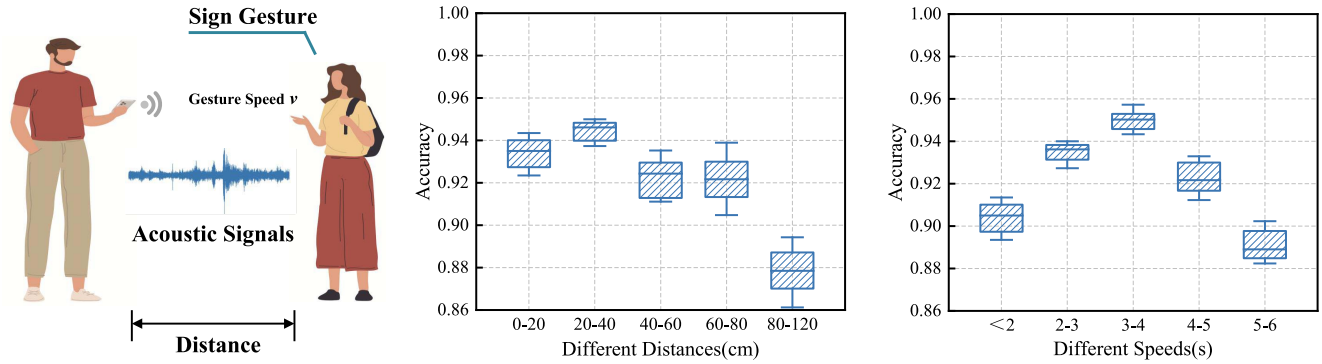


Fig. 12. (a) Explanation of distance and speed. (b) Resulting in increasing distances between the CSL signer and the smartphone user. (c) Resulting in different speeds between the CSL signer and the smartphone user.

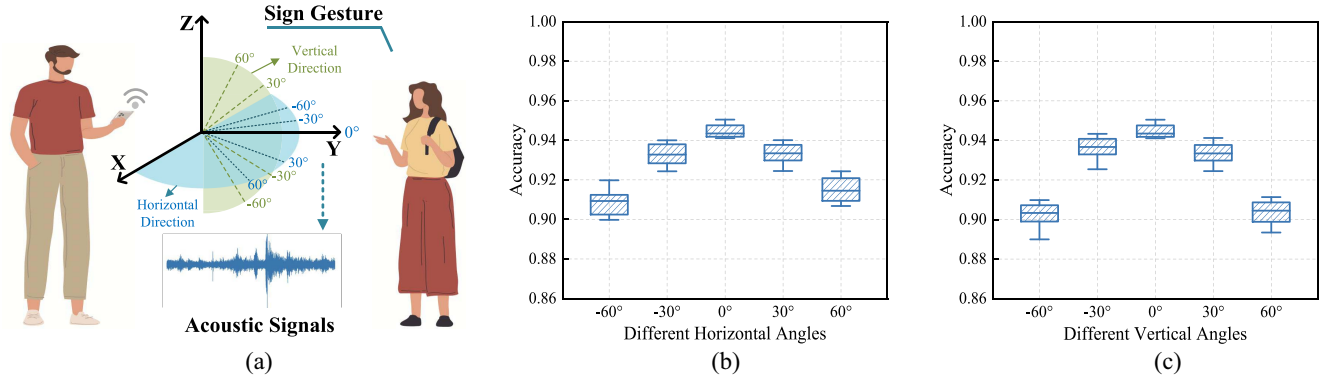


Fig. 13. (a) Explanation of aiming orientations. (b) Resulting in different horizontal angles between the CSL signer and the smartphone user. (c) Resulting in different vertical angles between the CSL signer and the smartphone user.

each group of two experimenters, to set the angle of the smartphone at −60°, −30°, 0°, −30°, and 60°, respectively.

Considering user habits, we collect the experimental data from five angles at horizontal and vertical orientations, respectively. The blue part in Fig. 13(a) represents the horizontal angle, while the green part represents the vertical angle. Correspondingly, the experimental results are shown in Fig. 13(b) and (c). From Fig. 13(b), when the interaction angle is 0°, the highest accuracy of gesture recognition is 94.34%. However, the accuracy decreases to 93.42% and 91.74% when the interaction angle shifts to 30° and 60°, respectively. Similarly, from Fig. 13(c), the recognition accuracy remains highest at 0° and gradually decreases as angular shift increases. This is because when the interaction angle is 0°, the extent of hand motion is vertical to the direction of signal emission, which has a greater impact on the signal. When the

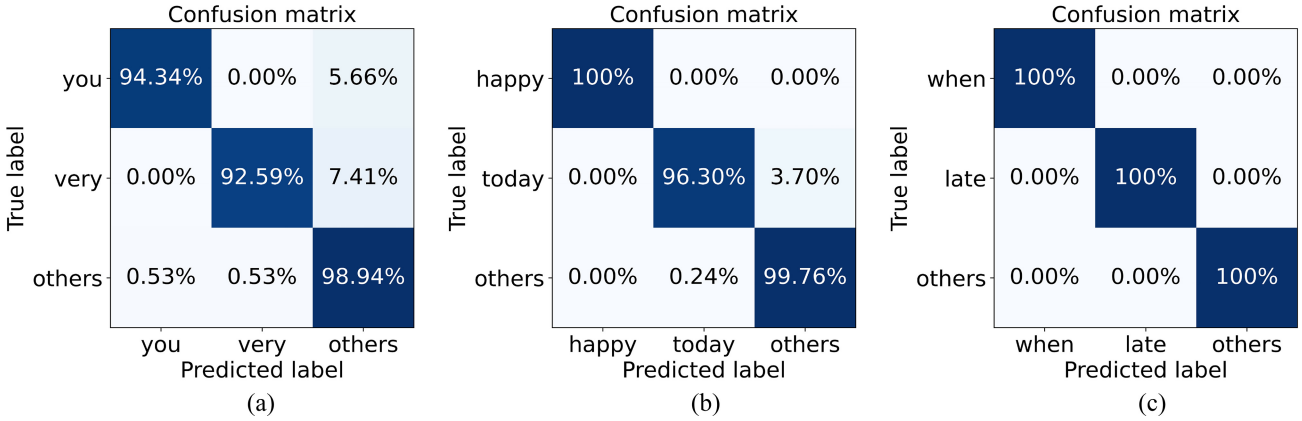


Fig. 14. (a) Performance of "你 and "很 sign language gestures. (b) Performance of "高兴 and "今天 sign language gestures. (c) Performance of "几点 and "迟到 sign language gestures.

interaction angle changes, the effect of the gesture movement on the signal is reduced, leading to a decrease in recognition accuracy. Overall, EarHear demonstrates consistently high performance across various interaction angles.

4) Similar Gesture: Each Chinese sign language word is represented by a unique gesture, and this uniqueness generates different sound reflection signals. However, the presence of a fraction with similar gestures leads to similar gesture features, which affects the recognition performance of the system. To verify EarHear's ability in processing similar sign language gestures, we explored a total of three sets of similar gestures, namely "你 (you)" and "很 (very)," "高兴 (happy)" and "今天 (today)," "几点 (when)" and "迟到 (late)."

Fig. 14 shows the confusion matrix for three sets of similar sign language gestures, and the overall accuracy ranges from 92.59% to 100%. It indicates that our proposed data processing method and model recognition method can distinguish similar gestures well. In Fig. 14(a), the gesture "you" achieves a recognition accuracy of 94.34%, with 5.66% of "you" gestures being misclassified as other gestures. Similarly, the gesture "very" achieves a recognition accuracy of 92.59%, with 7.14% of "very" gestures being recognized as other gestures. Both "you" and "very" have similar gesture characteristics, as they involve a top-to-bottom movement of the index finger. In Fig. 14(b), the gesture "happy" achieves a perfect recognition accuracy of 100%, while the gesture "today" achieves a slightly lower accuracy of 96.30%. In Fig. 14(c), the recognition accuracy of "when" and "late" are both 100%, and these two words are two-hand sign language gestures. Therefore, we can conclude that EarHear can detect small finger movements well and recognize similar sign language gestures.

5) Data Augmentation: To evaluate the impact of the DA technique on the robustness of the sign language recognition system, we analyze the recognition results from two perspectives: word level and sentence level. Fig. 15 shows the cumulative distribution functions (CDFs) curves of the error rates with and without DA for these two levels.

In the figure, the x-axis represents the recognition error rate and y-axis represents the CDF percentage. From the perspective of recognition performance, when the CDF value is 0.8,

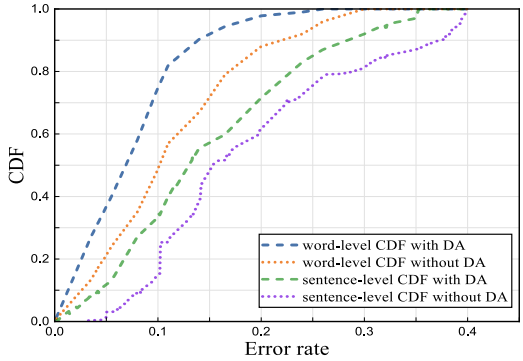


Fig. 15. Impact of DA on the robustness of sign language recognition systems.

the word-level error rate with DA is 0.08, the word-level error rate without DA is 0.17, the sentence-level error rate with DA is 0.225, compared with the sentence-level error rate without DA is 0.293. These results show that word-level recognition accuracy is higher than sentence-level sign language recognition accuracy. This is because multiple uncertainties between words (e.g., time intervals) affect performance. Therefore, achieving high-performance sentence-level sign language recognition is more challenging than word-level recognition. From the perspective of DA, we found that the recognition error rate was significantly reduced and the system performance was significantly optimized after adding DA techniques. This is due to our use of DA methods, such as random rotation, random flip, random scaling, and random occlusion, which allow the model to better adapt to different scenes and environments, thus improving the robustness and reliability of the system. We will continue to employ DA techniques in a series of subsequent experiments.

6) Individual Differences in User Behavior: We also conduct experiments to evaluate the impact of user behavior. Ten subjects are recruited to evaluate the performance under different user behavior. It involves the user's hand movement habits, hand size, finger length, etc. We train the ViT4SLR model on the data set of one subject and test it on the data sets of all other subjects. Fig. 16 illustrates the average sign language recognition accuracy for each subject. The experimental

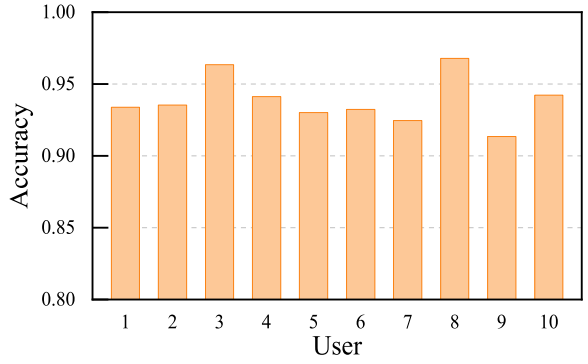


Fig. 16. Impact of different user behavior on the robustness of sign language recognition systems.

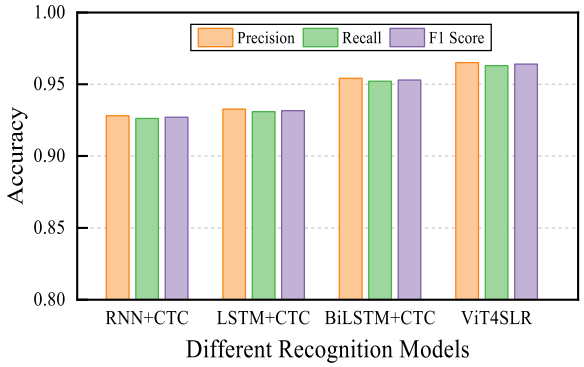


Fig. 18. User-dependent recognition accuracy of different models.

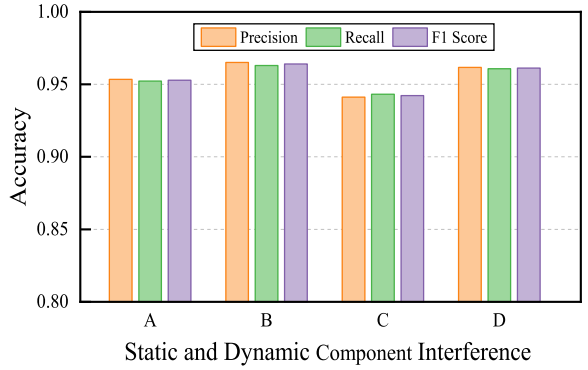


Fig. 17. Performance of ViT4SLR with different interference factors and signal preprocessing methods.

results indicate that individual differences in user behavior have a smaller impact on the system's performance, and the overall average accuracy exceeds 91.34%. It further demonstrates that EarHear has consistently high performance across various user behavior.

D. Ablation Experiments

1) Evaluation of the D-Doppler Method: Static environment and object movement are two main interference sources. To demonstrate the effectiveness of the D-Doppler method in eliminating the two types of interference during signal processing, we set up two experimental scenarios: 1) two experimenters perform sign language gestures in an empty hall and in a classroom with a regular distribution of equipment, respectively and 2) two experimenters perform sign language gestures in the scenarios with fan interference and people walking around, respectively.

Fig. 17 shows the experimental results under different interference factors and signal preprocessing methods, where A, B, C, and D represent the results without D-Doppler under static interference, with D-Doppler under static interference, without D-Doppler under dynamic interference, and with D-Doppler under dynamic interference, respectively. With static environmental interference, the D-Doppler method leads to notable improvements in the evaluation indexes of precision, recall, and $F1$-score by 1.17%, 1.07%, and 1.112%, respectively, compared to the absence of the D-Doppler method. Similarly, with dynamic interference, the D-Doppler method results in significant improvements in the evaluation indexes of precision, recall, and $F1$-score by 2.05%, 1.76%, and 1.9%, respectively. These experimental results provide compelling evidence that our proposed D-Doppler method can effectively eliminate static environment and object movement. As a result, it delivers high-quality data inputs for subsequent neural network processing.

2) Evaluation of Sign Language Recognition Models: To demonstrate the good performance of our proposed sign language recognition model, we evaluate the model in terms of both recognition accuracy and computational speed of sign language gestures.

Accuracy Evaluation of Sign Language Recognition Models: Considering the differences in morphology, speed, and direction of sign language actions in real scenes, how to make the recognition model with good long-range context modeling capability and make correct judgments on gesture behaviors is one of the challenges we want to solve. In order to verify the effectiveness of ViT4SLR in solving this problem, we selected several commonly used models as comparisons, including RNN+CTC, LSTM+CTC, and Bi-LSTM+CTC. We compared them in terms of precision, recall and $F1$-score, and the experimental results are as follows:

Fig. 18 shows that ViT4SLR has the highest scores on the three evaluation metrics of precision, recall, and $F1$-score with 96.51%, 96.30%, and 96.40%, respectively. While the models using RNN+CTC, LSTM+CTC, and Bi-LSTM+CTC have slightly lower accuracy rates. This is because traditional sign language recognition models based on the LSTM series of algorithms only consider information from current, past, and future time frames. They cannot capture long-range global contextual information, thus failing to recognize some weak gestures and context-dependent gestures. In contrast, our proposed ViT4SLR model takes advantage of the transformer to better overcome the above problems and has the ability to better adapt to real scenarios.

Efficient Evaluation of Sign Language Recognition Models: To evaluate the performance of our proposed ViT4SLR in terms of accuracy and speed, we conducted a comparative experiment using several state-of-the-art models as baselines. The chosen models include RNN+CTC, LSTM+CTC, Bi-LSTM+CTC, and ResNet+CTC. We input the same

TABLE I
PERFORMANCE EVALUATION RESULTS OF FOUR TRANSLATION MODELS

	DEV				TEST			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ViT4SLR	71.12	62.32	51.74	41.80	68.05	59.93	50.76	41.72
ViT4SLR+CSLT (Acoustic Dataset)	73.25	64.04	57.66	48.70	68.89	61.08	53.76	45.17
ViT4SLR+CSLT (MuCGEC)	76.22	69.67	62.36	55.15	69.71	63.50	57.23	51.00
ViT4SLR+CSLT (MuCGEC Finetune)	80.73	75.67	70.14	65.08	75.83	72.09	68.66	65.52

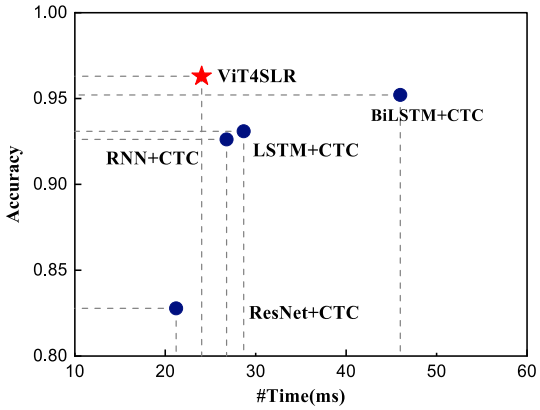


Fig. 19. Performance of ViT4SLR in terms of accuracy and speed.

acoustic spectrogram samples into each of the five models under identical hardware conditions.

Fig. 19 shows the experimental results, where the horizontal axis represents the single inference time for the selected sample and the vertical axis represents the recognition accuracy. The closer to the upper left corner of the figure indicates the higher inference speed and accuracy, reflecting better model performance. It can be clearly observed that our proposed ViT4SLR outperforms the other models in terms of both recognition accuracy and inference speed. Specifically, the ViT4SLR achieves a recognition accuracy of 0.965 while maintaining an inference time of only 24.02 ms. This impressive balance between accuracy and computational efficiency is evident when compared to other models, such as Bi-LSTM+CTC, which reaches a slightly lower accuracy of 0.954 but at a significantly higher inference time of 45.98 ms. The primary reason for this demonstration is that CTC is a very time-consuming algorithm, while our ViT4SLR avoids this operation through an attention mechanism and significantly improves the inference time. Therefore, ViT4SLR is more suitable to be deployed on mobile devices, enabling real-time sign language recognition.

3) Evaluation of Sign Language Translation Models: Next, we evaluate the sign language translation model in terms of both the sign language translation performance and the advantages of the two-stage model.

Performance Evaluation of Sign Language Translation Models: To solve the problem of sequence inconsistency between sign language gestures and natural language, we added a CSLT model after ViT4SLR. The purpose of this model is to translate the lexical sequences obtained by ViT4SLR into natural language sequences matching the verbal description. To evaluate the performance of the translation model, we used the BLEU as a metric. BLEU is a widely used machine translation evaluation metric that assesses the quality of machine translation by comparing the similarity between machine translation results and human translation results. Its calculation formula is as follows:

$$\text{BLEU}-n=\frac{\sum_{n-\text{gram}\in\hat{y}}\text{count}_{\text{clip}}(n-\text{gram})}{\sum_{n-\text{gram}\in\hat{y}}\text{count}(n-\text{gram})} \tag{9}$$

where count is the result of model translation, $\text{count}_{\text{clip}}$ is the result of human translation, $n-$ gram is the set of n words, and $\hat{y}$ is the candidate word.

Table I shows the performance evaluation results of the four translation models. Among them, ViT4SLR in the first row refers to the direct use of the sign language recognition model to calculate BLEU scores without using the sign language translation model. ViT4SLR+CSLT (Acoustic Data set) in the second row refers to the use of our acoustic sign language recognition data set to train the translation model. ViT4SLR+CSLT (MuCGEC) in the third row refers to the direct evaluation using the MuCGEC pretrained model. ViT4SLR+CSLT (MuCGEC Finetune) in the fourth row is our final approach, i.e., fine-tuning the MuCGEC pretrained model on our acoustic sign language recognition data set by migration learning. As can be seen from Table I, our adopted sign language translation method achieves BLEU-1 scores of 80.73/75.83, BLEU-2 scores of 75.67/72.09, BLEU-3 scores of 70.14/68.66, and BLEU-4 scores of 65.08/65.52 on the development set and test set, respectively. ViT4SLR+CSLT obviously outperformed the other models. This result indicates that the migration learning strategy not only alleviates the problem of insufficient acoustic sign language translation data set but also significantly improves the performance of the CSLT model.

Superiority Evaluation of the Two-Stage Model: The two-stage model refers to the need to train a sign language recognition model and a sign language translation model separately, and the one-stage model refers to simultaneous sign language recognition and sign language translation. To compare the advantages and disadvantages of the two-stage and one-stage models in terms of translation performance, we trained an additional one-stage model on the acoustic sign language recognition data set and the ViT4SLR structure. And

TABLE II
PERFORMANCE EVALUATION RESULTS OF ONE-STAGE AND TWO-STAGE MODELS

	DEV				TEST			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ViT4SLR (Recognition and Translation)	76.28	70.04	64.66	59.70	73.88	68.98	62.76	57.47
ViT4SLR+CSLT (MuCGEC Finetune)	80.73	75.67	70.14	65.08	75.83	72.09	68.66	65.52

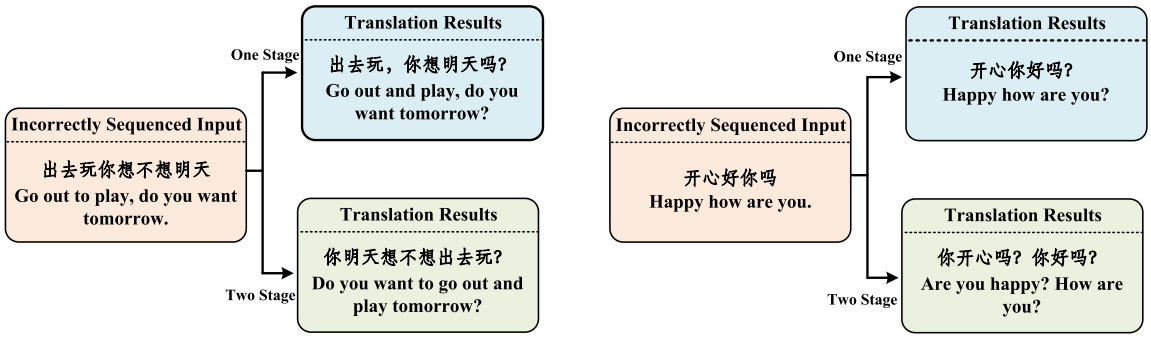


Fig. 20. Two examples of sentences with incorrect word order.

TABLE III
COMPARISON WITH EXISTING WIRELESS-BASED GESTURE RECOGNITION INTERFACES

Interfaces	Air-CSL [48]	mmASL [49]	RobuCIR [22]	UltrasonicG [23]	SonicASL [8]	UltrasonicGS [9]	HearASL [32]	**EarHear(ours)**
Device	Two Laptops	Radar	Smartphone	ASDP	Smartphone	ASDP	Smartphone	Smartphone
Signal	Wi-Fi	Millimeter Wave	Ultrasound	Ultrasound	Ultrasound	Ultrasound	Ultrasound	Ultrasound
Feature	Doppler	Doppler	CIR	Doppler	Doppler	Doppler	CIR	Doppler
Algorithm	DBM+GRU	Deep Learning	CNN+LSTM	ResNet+Bi-LSTM	VGG+LSTM+CTC	ResNet+Bi-LSTM+CTC	CNN+GRU+CTC	ViT4SLR+CSLT(MuCGEC Finetune)
Gesture Type	30	15	15	15	42	27	50	32
Sentence Type	N/A	N/A	N/A	N/A	30	6	30	15
Single Gesture	91.77%	87%	98.40%	98.80%	93.80%	98.80%	97.2%	98.89%
ContinuousSign Gesture	N/A	N/A	N/A	N/A	90.60%	86.30%	N/A	93.38%
Sign Language Translation	N/A	N/A	N/A	N/A	N/A	N/A	N/A	80.73%

later, we evaluated the one-stage and two-stage models using the same test set.

The results in Table II show that using the two-stage model performs better on the development set and test set compared to using the one-stage model. Specifically, on the development set, the two-stage model improved the BLEU-1, BLEU-2, BLEU-3, and BLEU-4 metrics by 4.45, 5.63, 5.48, and 5.38 percentage points, respectively, compared to the one-stage model. On the test set, the improvement was 1.95, 3.11, 5.9, and 8.05 percentage points, respectively. It demonstrates that employing a two-stage model can effectively improve the performance of downstream tasks with the help of a large-scale pretrained language model. In addition, as shown in Fig. 20, we input two sentences with the wrong order into the one-stage model and the two-stage model, respectively. The translation results show that the translation of the two-stage model is more consistent with the daily expression habits, which further indicates that the two-stage model has a stronger generalization ability. In real sign language communication, the two-stage model can make relatively correct predictions even in the case of incorrect sequences of sign language actions and irregular sign language expressions.

E. Comparison With Other State-of-the-Art Methods

To evaluate the performance of our proposed method in sign language recognition and sign language translation tasks, we compared it with recent state-of-the-art gesture recognition methods based on Wi-Fi, millimeter waves, and acoustic sensing. Specifically, we compared these eight methods in terms of device, signal, feature extraction, classification algorithm, gesture category, sentence category, and performance of single gestures, continuous sign language gestures, and sign language translation. Where single and continuous sign language gestures were evaluated using accuracy and sign language translation was evaluated using the BLEU-1 metric, and the details are shown in Table III.

Air-CSL uses two laptops with Intel 5300 NIC for data collection, while mmASL uses millimeter-wave radar. Both methods require specialized equipment and are not easily portable. Additionally, they are limited to recognizing single sign language gestures and lack generalization capabilities. RobuCIR and UltrasonicG are based on acoustic sensing methods using ultrasonic waves. RobuCIR collects data using smartphones and extracts fine-grained features using CIR, while UltrasonicG employs a dedicated acoustic development board

called ASDP and extracts features using Doppler frequency shift. However, these methods can only recognize a small number of gestures. SonicASL, UltrasonicGS, and HearASL are capable of recognizing both single sign gestures and continuous sign sentences. UltrasonicGS and HearASL achieve higher accuracy rates for single sign gestures, with rates of 98.80% and 97.2%, respectively. SonicASL and UltrasonicGS attain accuracy rates of 90.60% and 86.30% for continuous sign sentences, respectively. However, it is worth noting that HearASL only considers the comparison between predicted words and true words without accounting for the specific positions and orders of the words, rendering its accuracy rate incomparable. In comparison, our proposed method achieves accuracy rates of 98.89% for single sign gestures and 93.38% for continuous sign sentences, surpassing all other methods. Notably, our method is the only one that incorporates sign language translation, achieving a BLEU-1 score of 80.73%. These results highlight the strong competitiveness of our method, reaching the state-of-the-art level.

V. Discussion and Future Work

Two-Way Communications Between Sign Language Users and Nonsign Language Users: EarHear system is capable of simple two-way communications. We add a text input window, allowing enter text and respond to sign language users. This design not only helps nonsign language users understand sign language but also enables timely positive feedback for sign language users. To further expand the functionality of EarHear, we will add text-to-speech conversion and sign language animation to the system. With the text-to-speech conversion feature, the system will be able to convert recognized sign language into spoken feedback for nonsign language users. Additionally, providing sign language animations will provide visual representations of sign language expressions, helping nonsign language users better understand and learn sign language.

Expand the CSL Vocabulary Data Set: We recognize 32 sign language words and 15 sentences totally in our system, but it only covers a small portion of the CSL vocabulary. To make our system cover more sign language, we will introduce our mobile application to the deaf community and special education schools, teaching them how to collect data. In the future, we will expand the CSL vocabulary data set to ensure a more comprehensive sign language recognition capability in the system, filling the gaps in the existing acoustic sign language data set.

Nonmanual Markers: For CSL communications, nonmanual markers comprised of nonaffective facial expressions, head positions, lip motions, and body positions often provide crucial grammatical context to the manual signs. However, it is challenging to achieve comprehensive gesture recognition that takes into account both finger movements and nonmanual markers. In future work, we will further explore and optimize our algorithms to establish a practical and reliable CSL recognition system.

Violent Dynamic Interference: Our innovative D-Doppler-based data processing strategy effectively mitigates static and mild dynamic interference (e.g., someone walking past the user or mild body movement of users). However, violent dynamic interference can lead to significant changes in spectrograms or signals, which may invalidate the D-Doppler strategy and limit the performance of subsequent recognition models. In the future, we will make further attempts to remove violent dynamic interference.

CSL Grammar: CSL has a unique grammar system. When working with Chinese, we must take into account sentence structures, grammar rules, and aesthetic factors. In this work, we utilize the MuCGEC data set to train our models, aiming to achieve stable and flexible output. To enhance the practicality of the SonicASL system, we will continue exploring novel sign language translation models including ongoing improvements and optimizations of our algorithms. Furthermore, we will also prioritize user feedback and demands to enrich system functionality and provide an improved sign language translation experience.

VI. Conclusion

In this article, we introduce EarHear, a groundbreaking CSLR and translation system that excels in performance by leveraging acoustic signals for wireless sensing. EarHear pioneers the realm of sign language recognition by accurately translating recognized word sequences into natural language. Our innovative D-Doppler-based data processing strategy effectively mitigates interference caused by environmental factors, providing a feasible solution for data noise reduction in wireless sensing. Furthermore, the proposed sign language recognition model, ViT4SLR, and the sign language translation model, CSLT, not only significantly improve recognition performance and speed up model inference but also achieve more stable and flexible outputs with the assistance of large-scale language models. Experimental results show that EarHear surpasses the state-of-the-art in terms of accuracy and robustness under various practical influences. We anticipate that our work will make a substantial technical contribution to the research on wireless CSLR.

References

[1] N. Aloysius and M. Geetha, "Understanding vision-based continuous sign language recognition," *Multimedia Tools Appl.*, vol. 79, nos. 31–32, pp. 22177–22209, 2020.
[2] Y. Du, P. Xie, M. Wang, X. Hu, Z. Zhao, and J. Liu, "Full transformer network with masking future for word-level sign language recognition," *Neurocomputing*, vol. 500, pp. 115–123, Aug. 2022.
[3] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra, "A multimodal framework for sensor based sign language recognition," *Neurocomputing*, vol. 259, pp. 21–38, Oct. 2017.
[4] F. Wang, C. Li, C.-W. Liu, Z. Zeng, K. Xu, and J.-X. Wu, "An approach based on 1D fully convolutional network for continuous sign language recognition and labeling," *Neural Comput. Appl.*, vol. 34, no. 20, pp. 17921–17935, 2022.
[5] N. Zhang, J. Zhang, Y. Ying, C. Luo, and J. Li, "Wi-Phrase: Deep residual-multihead model for WiFi sign language phrase recognition," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 18015–18027, Sep. 2022.
[6] G. Lin et al., "Human activity recognition using smartphones with WiFi signals," *IEEE Trans. Human–Mach. Syst.*, vol. 53, no. 1, pp. 142–153, Feb. 2023.
[7] S. Z. Gurbuz et al., "American sign language recognition using RF sensing," *IEEE Sensors J.*, vol. 21, no. 3, pp. 3763–3775, Feb. 2021.
[8] Y. Jin et al., "SonicASL: An acoustic-based sign language gesture recognizer using earphones," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 5, no. 2, pp. 1–30, 2021.

[9] Y. Wang, Z. Hao, X. Dang, Z. Zhang, and M. Li, "UltrasonicGS: A highly robust gesture and sign language recognition method based on ultrasonic signals," *Sensors*, vol. 23, no. 4, p. 1790, 2023.
[10] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, 2016, pp. 82–94.
[11] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "FingerIO: Using active sonar for fine-grained finger tracking," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2016, pp. 1515–1525.
[12] R. Nandakumar, A. Takakuwa, T. Kohno, and S. Gollakota, "CovertBand: Activity information leakage using music," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–24, 2017.
[13] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao, "STRATA: Fine-grained acoustic-based device-free tracking," in *Proc. 15th Annu. Int. Conf. Mobile Syst. Appl. Services*, 2017, pp. 15–28.
[14] H. Jiang, M. Wang, D. Liu, and S. Zhou, "CTrack: Acoustic device-free and collaborative hands motion tracking on smartphones," *IEEE Internet Things J.*, vol. 8, no. 19, pp. 14658–14671, Oct. 2021.
[15] W. Liu, W. Shen, B. Li, and L. Wang, "Toward device-free micro-gesture tracking via accurate acoustic doppler-shift detection," *IEEE Access*, vol. 7, pp. 1084–1094, 2018.
[16] X. Xu, J. Yu, Y. Chen, Y. Zhu, L. Kong, and M. Li, "BreathListener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals," in *Proc. 17th Annu. Int. Conf. Mobile Syst. Appl. Services*, 2019, pp. 54–66.
[17] T. Wang et al., "OmniResMonitor: Omnimonitoring of human respiration using acoustic multipath reflection," *IEEE Trans. Mobile Comput.*, early access, Jun. 1, 2023, doi: 10.1109/TMC.2023.3281928.
[18] T. Wang et al., "MultiResp: Robust respiration monitoring for multiple users using acoustic signal," *IEEE Trans. Mobile Comput.*, early access, May 25, 2023, doi: 10.1109/TMC.2023.3279976.
[19] W. Xie, Q. Zhang, and J. Zhang, "Acoustic-based upper facial action recognition for smart eyewear," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 5, no. 2, pp. 1–28, 2021.
[20] Z. Wang, Y. Wang, M. Tian, and J. Shen, "HearFire: Indoor fire detection via inaudible acoustic sensing," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 6, no. 4, pp. 1–25, 2023.
[21] Y. Yang, Y. Wang, J. Cao, and J. Chen, "HearLiquid: Nonintrusive liquid fraud detection using commodity acoustic devices," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13582–13597, Aug. 2022.
[22] Y. Wang, J. Shen, and Y. Zheng, "Push the limit of acoustic gesture recognition," *IEEE Trans. Mobile Comput.*, vol. 21, no. 5, pp. 1798–1811, Jul. 2020.
[23] Z. Hao, Y. Wang, D. Zhang, and X. Dang, "UltraSonicg: Highly robust gesture recognition on ultrasonic devices," in *Proc. 17th Int. Conf. Wireless Algorithms Syst. Appl. (WASA)*, Nov. 2022, pp. 267–278.
[24] Y. Zou, Z. Xiao, S. Hong, Z. Guo, and K. Wu, "Echowrite 2.0: A lightweight zero-shot text-entry system based on acoustics," *IEEE Trans. Human–Mach. Syst.*, vol. 52, no. 6, pp. 1313–1326, Dec. 2022.
[25] I. Papastratis, K. Dimitropoulos, and P. Daras, "Continuous sign language recognition through a context-aware generative adversarial network," *Sensors*, vol. 21, no. 7, p. 2437, 2021.
[26] H. Luqman and E.-S. M. El-Alfy, "Towards hybrid multimodal manual and non-manual arabic sign language recognition: Marsl database and pilot study," *Electronics*, vol. 10, no. 14, p. 1739, 2021.
[27] L. Kraljević, M. Russo, M. Pauković, and M. Šarić, "A dynamic gesture recognition interface for smart home control based on croatian sign language," *Appl. Sci.*, vol. 10, no. 7, p. 2300, 2020.
[28] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li, "Improving sign language translation with monolingual data by sign back-translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1316–1325.
[29] M. Lee and J. Bae, "Real-time gesture recognition in the view of repeating characteristics of sign languages," *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 8818–8828, Dec. 2022.
[30] A. Qaroush, S. Yassin, A. Al-Nubani, and A. Alqam, "Smart, comfortable wearable system for recognizing arabic sign language in real-time using IMUs and features-based fusion," *Exp. Syst. Appl.*, vol. 184, Dec. 2021, Art. no. 115448.
[31] Y. Wang, F. Li, Y. Xie, C. Duan, and Y. Wang, "HearASL: Your smartphone can hear American sign language," *IEEE Internet Things J.*, vol. 10, no. 10, pp. 8839–8852, May 2023.
[32] M. Basner et al., "Auditory and non-auditory effects of noise on health," *Lancet*, vol. 383, no. 9925, pp. 1325–1332, 2014.
[33] C. Cai, H. Pu, M. Hu, R. Zheng, and J. Luo, "Acoustic software defined platform: A versatile sensing and general benchmarking platform," *IEEE Trans. Mobile Comput.*, vol. 22, no. 2, pp. 647–660, Feb. 2023.
[34] Q. Lin, Z. An, and L. Yang, "Rebooting ultrasonic positioning systems for ultrasound-incapable smart devices," in *Proc. 25th Annu. Int. Conf. Mobile Comput. Netw.*, 2019, pp. 1–16.
[35] W. Mao, J. He, and L. Qiu, "CAT: high-precision acoustic motion tracking," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, 2016, pp. 69–81.
[36] W. Ruan, Q. Z. Sheng, L. Yang, T. Gu, P. Xu, and L. Shangguan, "AudioGEST: Enabling fine-grained hand gesture detection by decoding echo signal," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 474–485.
[37] S. Yun, Y.-C. Chen, and L. Qiu, "Turning a mobile device into a mouse in the air," in *Proc. 13th Annu. Int. Conf. Mobile Syst. Appl. Services*, 2015, pp. 15–29.
[38] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "BatTracker: High precision infrastructure-free mobile device tracking in indoor environments," in *Proc. 15th ACM Conf. Embedded Netw. Sensor Syst.*, 2017, pp. 1–14.
[39] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–8.
[40] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUS)," 2016, *arXiv:1606.08415*.
[41] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.
[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[43] A. Radford et al., "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
[44] Y. Zhang et al., "MUCGEC: A multi-reference multi-source evaluation dataset for Chinese grammatical error correction," 2022, *arXiv:2204.10994*, 2022.
[45] I. J. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[46] D. R. Stockwell and A. T. Peterson, "Effects of sample size on accuracy of species distribution models," *Ecol. Model.*, vol. 148, no. 1, pp. 1–13, 2002.
[47] H. Chen, D. Feng, Z. Hao, X. Dang, J. Niu, and Z. Qiao, "AIR-CSL: Chinese sign language recognition based on the commercial WiFi devices," *Wireless Commun. Mobile Comput.*, vol. 2022, Sep. 2022, Art. no. 5885475.
[48] P. S. Santhalingam, A. A. Hosain, D. Zhang, P. Pathak, H. Rangwala, and R. Kushalnagar, "MMASL: Environment-independent ASL gesture recognition using 60 GHZ millimeter-wave signals," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 1–30, 2020.

Zhanjun Hao (Member, IEEE) received the B.S. degree from the School of Computer Science, Xi'an University of Electronic Science and Technology, Xi'an, China, in 2003, the M.S. degree from the School of Mathematics and Information Technology, Northwest Normal University, Lanzhou, China, in 2011, and the Ph.D. degree from the School of Electronic Information Engineering, Lanzhou Jiaotong University, Lanzhou, in 2021.

He is the Vice President of the Academy of Sciences, Northwest Normal University, Lanzhou, a Professor, and a Doctor. His research interests include Internet of Things, sensor networks, and wireless sensing technology.

Prof. Hao is the Chairman of the CCF YOCSEF Lanzhou Academic Committee, a member of the IoT Special Committee and Embedded System Special Committee of the Chinese Computer Society, and a Senior Member of CCF.

Yuejiao Wang received the B.E. degree from the School of Software, Hebei Normal University, Shijiazhuang, China, in 2021. She is currently pursuing the master's degree with the School of Computer Science and Engineering, Northwest Normal University, Lanzhou, China.

Her research interests include mobile computing, human–computer interaction, and acoustic sensing.

Zhenyi Zhang received the B.E. degree from the School of Computer and Technology, Anhui University, Hefei, China, in 2022. He is currently pursuing the master's degree with the School of Computer Science and Engineering, Northwest Normal University, Lanzhou, China.

His research interests include deep learning, computer vision, and wireless perception, particularly acoustic wireless passive perception.

Xiaochao Dang (Member, IEEE) received the B.S. degree from the School of Physics, Northwest Normal University, Lanzhou, China, in 1985, and the M.S. degree from the School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China, in 1995.

He is the Vice President of the School of Computer Science and Engineering, Northwest Normal University, Lanzhou, a Professor, a Doctor, the Director of the Gansu IoT Engineering Research Center, Lanzhou, and a Senior Visiting Scholar with the University of Old Territories in the United States. He has led five national projects and published more than 150 articles. His research interests include Internet of Things, sensor networks, and wireless sensing technology.

Prof. Dang is a Senior Member of CCF.