SIN-CHN
SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2024, 21(4), 793.
https://doi.org/10.62617/mcb793

Article

A Chinese sign language recognition system combining attention mechanism and acoustic sensing

Yuepeng Shi*, Yansheng Wu, Qian Li, Junyi Zhang

School of Energy and Intelligent Engineering, Henan University of Animal Husbandry and Economy, Zhengzhou 450011, China
*** Corresponding author:** Yuepeng Shi, YPShi2024@163.com

CITATION

Shi Y, Wu Y, Li Q, Zhang J. A Chinese sign language recognition system combining attention mechanism and acoustic sensing. Molecular & Cellular Biomechanics. 2024; 21(4): 793. https://doi.org/10.62617/mcb793

ARTICLE INFO

Received: 12 June 2024
Accepted: 14 November 2024
Available online: 11 December 2024

Abstract: In recent years, with the widespread popularity of smart devices and the rapid development of communication and artificial intelligence technologies, sign language gestures that can break the communication barriers between ordinary people and those with speech and hearing impairments have received much attention. However, existing human gesture recognition methods include wearable device-based, computer vision-based and Radio Frequency (RF) signal-based. These methods have problems of being difficult to deploy, violating user privacy, and being susceptible to ambient light. Compared with the above methods, using ultrasonic signals to sense sign language gestures has the advantages of not violating user privacy and not being affected by ambient light. For that purpose, we use the built-in speaker and microphone of a smartphone to send and receive ultrasonic signals to recognize sign language gestures. In order to recognize fine-grained sign language gestures, we calculate the Channel Impulse Response (CIR) induced by the sign language action as a sign language gesture special. After that, we compute first-order differences along the time dimension of the Channel Impulse Response matrix to eliminate static path interference. Finally, a convolutional neural network containing convolutional layers, spatial attention, and channel attention is passed in order to recognize sign language gestures. The experimental results show that the scheme has a recognition accuracy of 95.2% for 12 sign language interaction gestures.

Keywords: channel impulse response; sign language gesture recognition; attention mechanism; acoustic sensing

1. Introduction

Sign language is a communication language based on hand body movements, which mainly facilitates people with speech dysfunction and hearing impairment to communicate with others. According to the World Health Organization's Thematic Report on Deafness and Hearing Loss 2023 [1], approximately 430 million people worldwide have disabling hearing impairment, and hearing impairment increases with age, with more than 25% of people over the age of 60 affected by disabling hearing impairment. Similar to other forms of language, learning a language often requires a great deal of effort, which has resulted in sign language being less popular among the general population. In order to help people with speech and hearing impairments to better integrate into society, a recognition system that can effectively recognize sign language gestures is of great significance.

Existing sign language gesture recognition schemes include wearable device based, computer vision based, RF and acoustic Doppler effect based. However, these schemes suffer from invasion of user privacy, need for additional hardware devices, and inability to recognize slower sign language gestures.

In order to solve the above problems, we propose to utilize the built-in speaker and microphone of a smartphone to send and receive ultrasonic signals simultaneously to sense sign language gestures, and later to represent fine-grained sign language gesture features by calculating the channel impulse response of the received signals. Finally, in order to suppress the influence of irrelevant features, we design a classification model with convolutional blocks combining the channel attention mechanism and the spatial attention mechanism to extract the essential features of sign language gestures to improve the recognition accuracy. The main contributions of this work are as follows:

1) We propose to recognize the user's sign language gestures using channel impulse response, which does not depend on the execution speed of the user's sign language gestures and can recognize slower sign language gestures.

2) We design a convolutional block combining channel attention and spatial attention in a classification model that suppresses the influence of irrelevant features to improve the recognition accuracy of sign language gestures.

3) We have implemented a Chinese sign language gesture recognition system, and the experimental results show that the algorithm can accurately recognize 12 kinds of Chinese sign language gestures with an average accuracy rate of 95.2%.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 analyzes the limitations of Doppler effect for recognizing gestures. Section 4 describes the design of sign language gesture recognition system. Section 5 presents the experimental results and analysis. Finally, Section 6 summarizes the paper.

2. Related work

Existing gesture recognition efforts can be categorized into the following four classes based on the different acquisition devices and emitted signals: wearable device based, computer vision based, RF based, and acoustic based.

2.1. Wearable device based

There are a number of traditional gesture recognition efforts that are based on wearable devices, and these types of approaches generally require some customized additional equipment. These devices are usually equipped with a variety of sensors such as accelerometers, gyroscopes, magnetometers, etc. These sensors are capable of sensing information such as acceleration, rotation and orientation of the device to capture subtle changes in gestures. Piskozub et al. [2] designed a data glove equipped with 10 piezoresistive flexible sensors, which were fastened to the textile surface of the glove with Velcro, and finally used a decision tree algorithm to recognize 16 Polish Sign Language gestures. L-Sign [3] utilized a bracelet with built-in sensors to recognize electromyogram (EMG) signals and inertial sensor signals generated by Chinese universal sign language, and the results show that the recognition rate of this scheme for Chinese sign language reaches 90%. WearSign [4] captured the user's sign language gesture data using a smartwatch and an armband with an 8-channel EMG sensor, and then fed the captured data into a network of codecs to recognize sign language gestures.

However, wearing such physical devices can impose additional physical burdens on the user. Some wearable devices may cause comfort issues due to their design or choice of wearing position, thus limiting the practical application of gesture recognition systems.

2.2. Computer vision based

Computer vision-based gesture recognition approaches generally work by capturing images or video streams using a camera and utilizing image processing and computer vision techniques to recognize and understand user gesture movements. This type of technology has been applied in various fields, such as virtual reality, gaming, smart home, healthcare, etc., to provide users with a more natural and intuitive interaction. Pruthvi et al. [5] utilized the YOLOv8 algorithm for real-time recognition of 7 sign language gestures in a video call application. The experimental results show that the recognition accuracy of the method is 98% on 7 sign language gestures. Zhu et al. [6] used the depth image of a hand gesture captured by the Kinect sensor and segmented it, and after obtaining the surface of the hand shape create the corresponding vectors and built a histogram by vector segmentation, then characterized the gesture with a 3D shape background descriptor containing rich 3D information, and finally applied a dynamic time warping algorithm for gesture recognition. Saboo et al. [7] proposed a video gesture detection and tracking algorithm to mitigate the effect of lighting variations or background occlusion on gesture recognition. The experimental results show an accuracy of 98.4% in the presence of light.

2.3. RF based

RF signal recognition is independent of ambient light and does not require the wearing of a physical device. Compared with some near-field sensing techniques, RF signal-aware gestures are usually able to achieve gesture detection at relatively long distances, and thus have received much attention. RFree-GR [8] recognizes 16 commonly used American Sign Language gestures by an array of Radio Frequency Identification (RFID) tags and removes domain-specific information using an adversarial model to overcome the effects of new environments, and the experimental results show that RFree-GR has an accuracy of 88.38% at a new location. Cai et al. [9] used Frequency Modulated Continuous Wave (FMCW) radar to sense four gestures and input the gesture features into a Convolutional Neural Networks (CNN) with residual blocks. The recognition accuracy of the four gestures is 98.7%. Liu H et al. [10] utilized millimeter-wave radios for gesture recognition and utilized transfer learning for the impact of environmental changes on gesture recognition accuracy to reduce the workload of sample collection for new environment adaptation.

However, this approach requires the deployment of additional transceiver devices and the location of the transceiver devices cannot be changed arbitrarily.

2.4. Acoustic based

Compared with the previous ways of recognizing gestures, using sound waves to sense gestures does not violate users' privacy, and existing commercial devices such

as smartphones and smartwatches have microphones and speakers embedded in them, eliminating the need to deploy additional devices. Therefore, acoustic wave-based gesture recognition and hand motion tracking have received a lot of attention. WordRecorder [11] extracts acoustic signals generated by pen rubbing on paper to recognize English letters, then converts the one-dimensional acoustic signals into spectrograms, and inputs the spectrograms into a deep neural network to realize the recognition of English letters. SonicASL [12] integrated a loudspeaker on an existing commercial headset to simulate the way radar works to sense sign language gesture movements. Experimental results show that the recognition accuracy for American Sign Language gestures at the word level is 93.8%. SignID [13] used ultrasound signals to extract the Doppler effect features of user gestures, in order to realize gestures and user identification at the same time, multi-task learning is used to realize gesture and identity recognition, the experimental results show that the recognition accuracy is 96.5% for 7 kinds of customized gestures, and the recognition accuracy is 95.5% for 8 users.

Although current acoustic perception-based gesture recognition can achieve high recognition accuracy, its principle is generally based on the Doppler effect. However, Doppler effect-based recognition methods are limited by the speed of gesture execution, and thus cannot effectively recognize fine-grained sign language gestures.

3. Limitations of the doppler effect

The Doppler effect is the change in the frequency of the signal at the receiving end caused by the motion of an object. Currently, many studies have used the short-time Fourier (STFT) algorithm for feature extraction of Doppler frequency shift characteristics [14]. For the frequency shift resulting from the Doppler effect can be calculated from Equation (1).

$$f = \frac{F_c \times V}{c} \tag{1}$$

where F_c is the frequency of the emitted signal, V is the speed of motion of the object, c is the speed of sound. The minimum resolution of the STFT can be calculated by Equation (2).

$$\Delta f = \frac{F_s}{w} \tag{2}$$

where Fs is the sampling frequency, usually 48 kHz, and w is the window length of the STFT, usually 2048. When F_c = 20 kHz, the resulting speed resolution of the Doppler effect is $\Delta v = (c \times \Delta f)/F_c$=0.4 m/s. Even if zero fill is not effective in improving resolution [15]. This means that the speed resolution of the Doppler effect is very low and therefore requires the user to perform sign language gestures quickly, which is not friendly to slow-moving elderly people. We model sign language gesture actions as a distance-time-signal strength two-dimensional feature matrix by calculating the channel impulse response to overcome the limitation of the Doppler effect that makes it difficult to recognize slower gestures.

4. System design

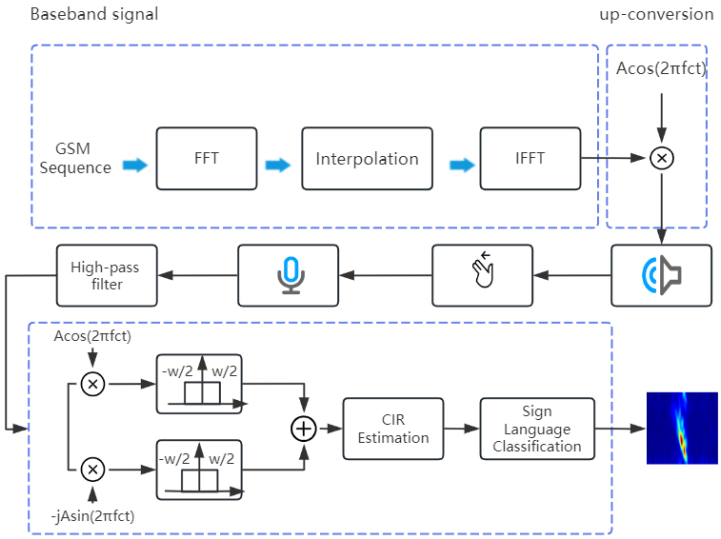


Figure 1. Overall system architecture.

The sign language gesture recognition system consists of three parts: transceiver, CIR estimation, and sign language classification. As shown in **Figure 1**, the 26-bit Global System for Mobile Communications (GSM) training sequence is first modulated into an ultrasonic signal with a center frequency of 20 kHz and a bandwidth of 4 kHz, and a single cycle of the GSM sequence is stored in order to prepare for the subsequent channel estimation stage. After that, the modulated ultrasound signal is stored as a lossless audio file format and played back by a smartphone while receiving the ultrasound signal reflected from the sign language gesture. After receiving the ultrasound signals reflected from the sign language gestures, the channel impulse response is computed using the least squares method for channel estimation, and the one-dimensional ultrasound signals are transformed into a two-dimensional matrix consisting of time-distance-amplitude, which is first-order differenced in order to remove the effects of the reflections of the static objects and the propagation of the signals along the line-of-sight paths. Since the sign language actions are continuous, it is necessary to segment the sign language actions and remove the invalid data in order to reduce the computation amount. Finally, the processed 2D matrix is fed into a convolutional neural network combining residual connectivity, spatial attention mechanism, and channel attention mechanism to extract valid sign language gesture features and recognize the sign language.

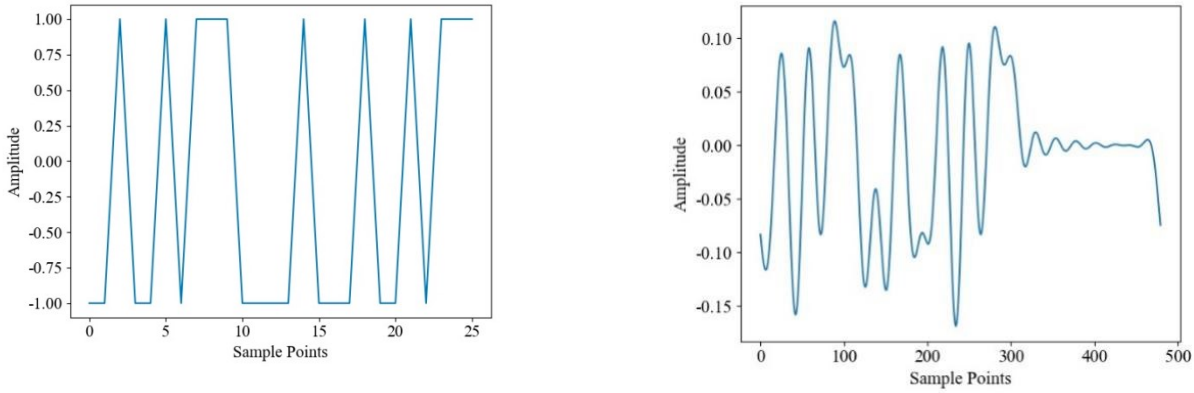
4.1. Transceiver design

Sign language gestures involve a variety of upper limb movements, including fingers, palms and arms. Therefore, the received ultrasonic signals are the superposition of the reflected signals from multiple targets, which has a high complexity. In order to realize the accurate recognition of sign language movements, the received ultrasound signals need to be deconstructed to reveal the patterns and characteristics of different limb movements. To address this challenge, we have adopted concepts from the field of wireless communication and designed a transceiver

system consisting of a speaker and a microphone to measure the channel impulse response using a known training sequence. This approach helps to extract the complex changes that occur in the signal during transmission, leading to a better understanding of the movement patterns of the different limbs in sign language movements. First of all, in order to estimate the channel efficiently we need to select the training sequences with good autocorrelation, such as GSM, Zadoff-Chu, Barker, etc. Since the GSM sequence is widely used in single carrier communication, it has the ideal characteristics of synchronization and channel estimation [16]. Therefore, we choose the 26-bit GSM training sequence, and the GSM sequence can be defined as.

$$N_g = \{g_1, g_2 \cdots g_{n-1}, g_n\} \tag{3}$$

where n is the number of symbol bits. Afterwards, N_g is modulated into binary phase-shift keying (BPSK) symbols, where bit 0 and bit 1 are mapped to baseband symbols 1 and −1, respectively.



(a) **(b)**

Figure 2. GMS sequences and GSM sequences after interpolation. **(a)** GMS sequences; **(b)** GSM sequences after interpolation.

In addition, we add 14 zeros at the end of the GSM sequence as protection bits to avoid superposition between the signals. Currently, most smartphones support a sampling frequency of 48 kHz. According to Nyquist's sampling law, in order to accurately reconstruct a signal, the sampling frequency of that signal is at least twice the highest frequency of the signal. Therefore, the highest frequency of the ultrasound signal sent by a smartphone cannot exceed 24 kHz. In addition, the hearing range of the human ear is mostly below 18 kHz. Due to the above two conditions, the bandwidth of the GSM sequence is 4 kHz at the center frequency fc = 20 kHz. To limit the bandwidth of the GSM sequences, a scheme of interpolating the training sequences is used. The interpolation of sequences can be categorized into time-domain and frequency-domain interpolation, and the autocorrelation of sequences after frequency-domain interpolation is generally higher than that of sequences after time-domain interpolation compared to time-domain interpolation [17]. Therefore, we choose to interpolate the sequence in the frequency domain of the training sequence. Specifically, the zero-padded GSM training sequence (40 bits) is first transformed to the frequency domain by Fourier transform, and interpolated by symmetrically padding the zeros after the positive frequency component and before the negative frequency component

until the length of the sequence is 480 bits, and then transformed to the time domain by Fourier inverse transform to obtain the frequency-domain interpolated single-cycle GSM signal. The frequency-domain interpolated GSM single-period signal is shown in **Figure 2**. The baseband signal can be obtained by expanding the period of single cycle. Since, the frequency band of the baseband signal is lower than 20 kHz, we cannot use the baseband signal directly. The baseband signal can be upconverted to the specified frequency band by multiplying the baseband signal with the carrier signal $A\cos(2\pi f_c t)$. where A is the amplitude and f_c is the carrier signal frequency. Afterwards, a 6th order Butterwolf high pass filter is utilized to filter out the low frequency components and saved as a playable file in wav format. The frequency domain representations of the baseband signal after up-conversion are shown in **Figure 3**.

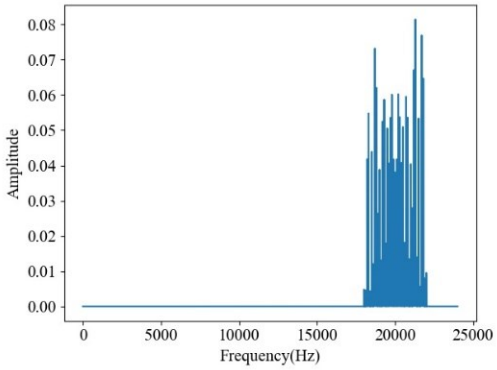


Figure 3. Frequency-domain representations of baseband signals after up-conversion.

4.2. CIR estimation

When the reflected signal $y(t)$ containing the sign language action is received, a filter is first utilized to remove the low frequency noise generated by the surrounding environment such as: talking sound, walking sound, and so on. Next, $y(t)$ is down converted to a baseband signal *r(t)* using IQ demodulation [18]. We need to multiply the reflected signal $y(t)$ by $\sqrt{2}cos(2\pi f_c t)$ and $-\sqrt{2}cos(2\pi f_c t)$, respectively, to decompose the signal into the in-phase component (I) and the quadrature component (Q), which represent the real and imaginary parts of the signal, respectively, and then finally the baseband signal can be obtained by summing up these two parts. The calculation formula is as follows.

$$r(t) = \sqrt{2}\cos(2\pi f_c t)y(t) - j\sqrt{2}\cos(2\pi f_c t)y(t) = \sqrt{2}e^{-j2\pi f_c t}y(t) \quad (4)$$

After obtaining the baseband signal of the received signal, it is necessary to align each frame of the received echo signal to further compute the CIR. In order to segment each signal frame, the starting point of the signal frame can be found by calculating the cross-correlation between the transmitted baseband signal $x(t)$ and the previous frames of the received baseband signal to find the maximum correlation coefficient and thus aligning the received baseband signal with $x(t)$. As frame-by-frame alignment is more computationally intensive, to reduce the amount of computation. In this paper, we choose the first 10 frames for alignment. Since the interval between each frame is

fixed, once a frame is detected, subsequent frames can be determined by adding a constant frame interval. Next, the CIR is computed using the baseband signal at the transmitter and the baseband signal at the receiver. According to wireless communication theory, the channel can be modeled as:

$$y = Wh + n \tag{5}$$

where y is the received signal, h is the channel impulse response, and n is the noise obeying a Gaussian distribution. In contrast to previous work [19] the CIR is computed using cyclic inter-correlation between the transmitted and received baseband signals. We use a least squares implementation to compute the CIR at a low computational cost. According to the least squares algorithm, we need to choose two important hyperparameters L and P, where we need to satisfy $L + P = d$, and d is the length of the effective data portion of the transmitted signal. We set L to 120, so we can distinguish 121 propagation paths. Let the transmit signal sequence $w = [\mathrm{w}_0, \mathrm{w}_1, \ldots, \mathrm{w}_{P+L-1}]^T$,channel impulse response $h = [h_0, h_1, \ldots, , h_L]^T$,then the loop training sequence matrix W can be modeled as:

$$W = \begin{bmatrix} w_L & \cdots & w_1 & w_0 \\ w_{L+1} & \cdots & w_2 & w_1 \\ \vdots & \ddots & \vdots & \vdots \\ w_{L+P-1} & \cdots & w_P & w_{\mathrm{P\text{-}1}} \end{bmatrix} \tag{6}$$

Obviously, the CIR can be estimated by $\hat{h} = argmin_h \|y - Wh\|^2$. Since the time complexity of the algorithm for directly calculating h is high and the loop training matrix W is known, the time complexity of the algorithm can be simplified by pre-computing $(W^H W)^{-1} W^H$ via Equation (7).

$$\hat{h} = (W^H W)^{-1} W^H y \tag{7}$$

The results of the CIR by least squares are shown in **Figure 4**.The ripple representation in **Figure 4a** indicates the movement of the sign language gesture, but the interference between the signal propagation along the line-of-sight path with multiple static path reflections makes the sign language gesture features submerged in these signals, which results in an inconspicuous gesture feature. Consider that gesture interaction gestures produce less channel variation when they involve finger movements. Therefore, there is a need to remove the interference of signals reflecting signals along the line-of-sight path and the static path. In this paper, the channel impulse response result matrix is first order differenced along the time dimension to remove the above interference. As can be seen in **Figure 4b**, the channel variations induced by the gesture motion are characterized more significantly after the first-order differencing operation.

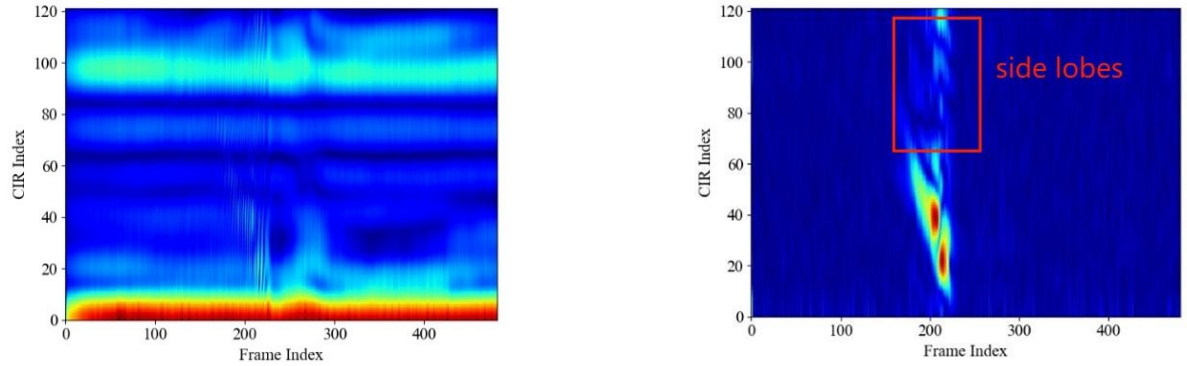


Figure 4. Original channel impulse response and channel impulse response after interference removal. **(a)** Raw channel impulse response; **(b)** Channel impulse response after de-interference.

4.3. Sign language segmentation

For the purpose of segmenting the gestures, we chose to detect the rate of change of each frame in the channel estimation matrix (dCIR) after eliminating static interference in order to find the start and end points of the sign language gestures. The reason for doing so is that the time period in which a sign language gesture exists will show a significant change in the channel estimation results, and this change is more obvious in the channel estimation matrix after eliminating static interference. Specifically, we first calculate the variance of dCIR at time t in the channel impulse response matrix $D_{t\times n}$ after eliminating static interference, to obtain the fluctuation of dCIR at time *t*. The variance $\widehat{D}$ is calculated as follows:

$$\overline{D}_j = \frac{1}{t}\sum_{i=1}^{t} D_{ij} \tag{8}$$

$$\widehat{D} = \frac{1}{t}\sum_{i=1}^{t}(D_{ij} - \overline{D}_j)^2 \tag{9}$$

The execution of a sign language gesture may contain not only movement pauses, but also other small movements that are not sign language gestures. In order to mitigate the effects of the above and make the segmentation algorithm more robust, we utilize a Gaussian filter to smooth the sequence of variances at each moment. Then, the variance at each moment is normalized:

$$norm(\widehat{D}) = \frac{\widehat{D} - \min(\widehat{D})}{\max(\widehat{D}) - \min(\widehat{D})} \tag{10}$$

Finally, when the value of a signal frame after smoothing is larger than a uniform empirical threshold $\sigma = 0.05$, the signal frame is considered to contain sign language features. The "cold" sign language segmentation results are shown in **Figure 5**.

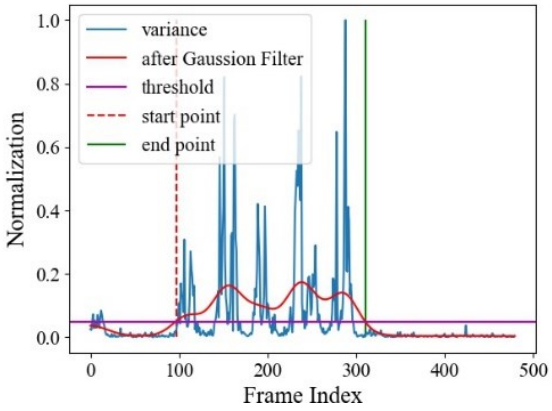


Figure 5. Sign language segmentation results.

4.4. Classification model design

Although we eliminate most of the signal interference along the line-of-sight path with static path reflections. However, in **Figure 4b** we can still see some residual static path reflection signals with interference from the side lobes gain [20], which will affect the recognition accuracy of the model. In order to overcome these interferences, we hope that the classifier model can simulate the human visual cerebral cortex, focusing the attention on regions with important information and reducing or discarding irrelevant information, thus improving the recognition accuracy of the model.

To achieve the above purpose, we introduce the Convolutional Block Attention Module (CBAM) proposed by Woo [21] and others into the CNN model, which combines the channel attention module and the spatial attention module. CBAM combines the channel attention module and spatial attention module, which can tell the CNN model "what" the useful information in the image is and "where" the useful information is located. In addition, CBAM is a lightweight module, so adding this module will not significantly increase the number of parameters and computation of the model. The structure of our classification model is shown in **Figure 6**.

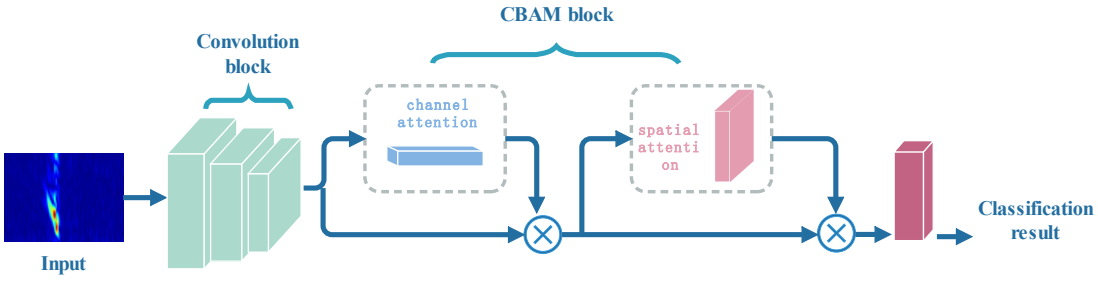


Figure 6. Classification model.

Specifically, the model mainly consists of a convolutional module, a CBAM module, and a fully connected layer. The convolutional module contains 4 convolutional layers, each of which is followed by a pooling layer to aggregate features of sign language gestures. Based on experience, the number of convolution kernels of the four convolutional layers are set to 64, 128, 256, and 256, respectively, where the first convolutional layer has a convolution kernel size of 7 × 7, a step size

of 1, a padding of 3, and a pooling layer behind it with maximal pooling; and the remaining three convolutional layers have convolution kernel sizes of 3 × 3, step sizes of 1, and paddings of 1, and the pooling layers are all average pooled. In addition, the distribution of the output data of each layer is normalized using Batch Normalization (BN) after each convolutional layer, which helps to prevent the model from falling into local minima during the training process, thus enabling the model to achieve faster and more stable training. Finally, in order to increase the expressive power of the model, we use a non-saturated activation function, ReLU, which enables the model to learn complex nonlinear relationships, thus increasing the expressive power of the model while mitigating the problem of gradient vanishing that occurs when the model is backpropagated. The core operations of the convolutional layer are as follows.

$$Z = ReLU(BN(WX + b)) \tag{11}$$

where W denotes the convolution kernel parameters, b is the bias, and X is the input. After the convolution module captures the local features of the input data, the result obtained is input to the CBAM module to better capture the important features in the input data and improve the performance of the model. The CBAM module consists of channel attention and spatial attention connected in series before and after, and the feature map F output from the convolution module first enters the channel attention module. The channel attention module is shown in **Figure 7**.

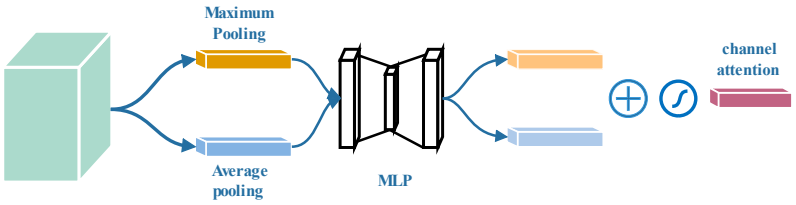


Figure 7. Channel attention module.

In the channel attention module, the spatial features of the input feature map F are first aggregated using maximum pooling and average pooling. After that, the results of global maximum pooling and average pooling are input into a multilayer perceptron to learn the importance of each channel, and finally the two result vectors after the multilayer perceptron are summed and mapped by a sigmoid activation function to obtain the channel attention. After that the channel attention is multiplied with the feature map F to get the output F' of the channel attention module. In short, the channel attention mechanism A_c is computed as follows.

$$F' = A_c(F) \otimes F \tag{12}$$

$$A_c(F) = \sigma(MLP(AvgPool(F) + MLP(MaxPool(F)) \tag{13}$$

where ⊗ denotes the element product, σ is the sigmoid function, and AvgPool() and MaxPool() represent the global average pooling and maximum pooling, respectively. After obtaining the output A_c of the channel attention module, it is fed into the spatial attention module. The spatial attention module is shown in **Figure 8**.

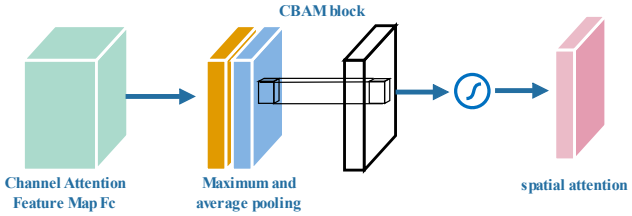


Figure 8. Spatial attention module.

In the spatial attention module, global average pooling and maximum pooling are first performed along the channel dimension of F' to compress the channel information, and then the results of global average pooling and maximum pooling are spliced along the channel dimension, and finally the spliced results are computed using convolution to obtain the final spatial attention A_s. The spatial attention A_s is multiplied by the output of the channel attention module F'' to obtain the output of the spatial attention output F''. The calculation of spatial attention is as follows:

$$F'' = A_s(F') \otimes F' \tag{14}$$

$$A_s = \sigma(Conve^{7\times7}(Cat[AvgPool(F'); MaxPool(F')]) \tag{15}$$

where, $Conve^{7\times7}$ represents the convolution operation with a convolution kernel of 7×7. Finally, the results are passed through a fully connected layer and the raw scores output from the neural network are converted to probability distributions using the Softmax activation function for sign language gesture classification.

5. Experiments and evaluation

5.1. Experimental dataset

For the sign language gesture experiments, we mainly conducted the experiments on a Samsung Galaxy C8 cell phone. For the training and deployment of the classification network, a host computer with model Xeon(R) Silver 4214R, 90 GB of RAM, and GPU model RTX 3080 Ti/12 GB is used. The environment in which the classification model was written and run was python 3.8, PyTorch 2.0.0, and cuda11.8.

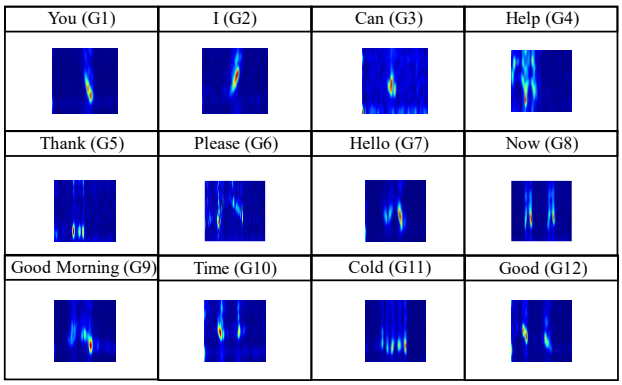


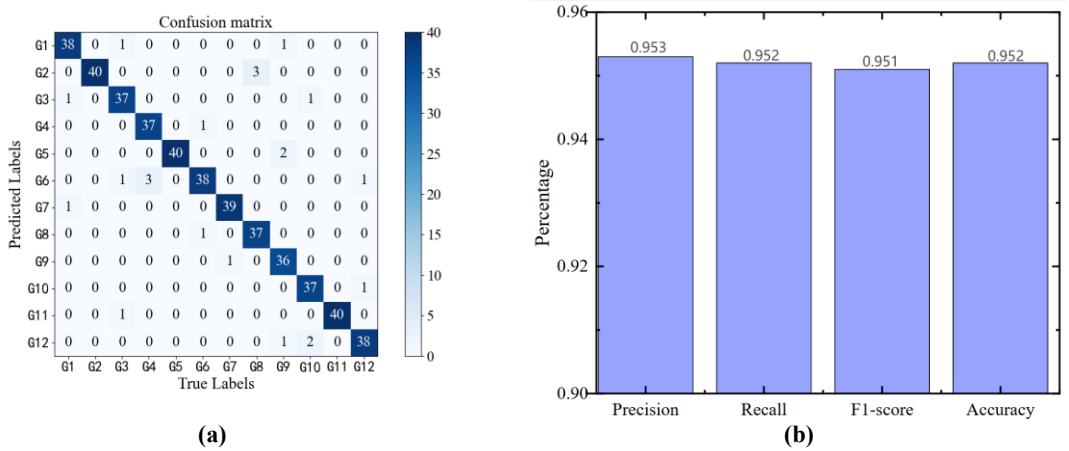
Figure 9. 12 Sign language gestures.

In order to evaluate the sign language gesture recognition system, we invited four volunteers to participate in the experiment; the volunteers' ages ranged from 20 to 60

years old. Between experiments we informed the volunteers of the purpose of the experiment. During the data collection process, the smartphone is placed on the desktop with the speaker facing the volunteers. Each volunteer performs 12 commonly used Chinese sign language gestures, with each gesture repeated 50 times; This process lasted for a month and a half, collecting a total of 2400 data. The collected data were divided into a training set and a test in the ratio of 8:2 for the training and evaluation of the classification model. The feature maps of the 12 commonly used Chinese sign language gestures are shown in **Figure 9**.

5.2. System performance and classification model evaluation

In this experiment, the learning rate for training the classification model is 10^{-5}, the optimizer aspect RMSprop, and the data batch during training is 128. Where the optimizer parameter α: 0.99 and γ is: 10^{-5}. With the above parameters, the classification model was trained for 106 rounds. The recognition accuracy is 95.2% for 12 sign language gestures. In addition to the accuracy rate, in order to better analyze the performance of the classification model, we evaluate the performance of the model by confusion matrix with precision, recall, and F1 score. The results are shown in **Figure 10**. The results are shown in **Figure 10**. Among them, the recognition accuracy of "good morning" is the lowest. This is because the sign language gesture of "good morning" only involves small changes in the fingers, and when the system does not capture certain finger movements, "good morning" is easily mistaken for "you", "thank you", etc. As can be seen in **Figure 10b**, the Precision, Recall and F1-score of the 12 sign language gestures are all higher than 94%, which indicates that our proposed sign language recognition system has good recognition performance.



(a) **(b)**

Figure 10. System performance. **(a)** Sign language sign confusion matrix; **(b)** Precision, recall, and f1 score.

In order to evaluate the effect of CBAM block on the classification model, in addition to the ablation experiments on the sign language gesture classification model, VGG16 [22] and MobileNetV2 [23] are also selected to compare with the classification model proposed in this paper, and the results are shown in **Table 1**.

Table 1. Classification model comparison.

Model	VGG16	MobileNetV2	Our model	Without CBAM block
Accuracy (%)	90.2	85.2	95.2	94.5
Parameters (M)	134.3	2.53	11.89	9.36

The experimental results show that the accuracy of the sign language classification model in this paper is improved by 0.7% compared to the removal of the CBAM block recognition accuracy, and the number of parameters increases by only 2.53 M. This is due to the fact that the CBAM block is a lightweight way to compute the attention mechanism of the image. Compared to MobileNetV2, the recognition accuracy of the sign language classification model proposed in this paper reaches 95.2% compared to MobileNetV2 with a 10% increase in recognition accuracy, but the number of MobileNetV2 parameters is only 7.66 M smaller than the model in this paper.The lower recognition accuracy of MobileNetV2 is due to the fact that MobileNetV2 adopts the depth separable Separable Convolution and other lightweight structures, which can reduce the number of parameters and computational complexity of the model, but may also limit the expressive ability of the model, resulting in difficulty in capturing complex image features. In addition, techniques such as feature mapping refinement and linear bottlenecks in MobileNetV2, although helpful in improving the feature representation ability, may still not be able to achieve the level of advanced feature representation that can be achieved by some of the more complex models. The recognition accuracy of VGG16 is 90.2%, but its parameter count of 134.3 M is much higher than that of other models. Model comparison experiments show that the sign language gesture classification model proposed in this paper achieves the best results in terms of recognition accuracy and does not require more model parameters.

5.3. Impact assessment of noise, distance

A The sign language recognition system we study uses acoustic sensing to perceive sign language actions. Therefore, the effect of environmental noise in daily life on the system needs to be considered. The characteristics of acoustic signal propagation in air are also considered, i.e., when the acoustic signal propagates through air, the acoustic signal will undergo path loss, reflection, refraction and diffraction. Therefore, the energy of the acoustic signal decreases with the increase of the propagation distance. For this reason, it is necessary to evaluate the effects of noise and distance in sign language recognition systems.

For noise evaluation, we evaluated the system from three different noise levels, 50–60db, 60–70db and 70–80db. Due to the unstable noise level generated by other people's activities or electrical equipment in the experimental environment, in order to evaluate accurately, we play music around the experimenters to control the noise level and also monitor the noise level by using the software that detects the noise level. The experimental results of the noise evaluation are shown in **Figure 11**, from which it can be seen that with the increase of the noise level, the recognition accuracy, recall, F1-score, accuracy and other indexes of the sign language recognition system remain

stable, which proves that the sign language recognition system is able to work effectively in a high noise environment.

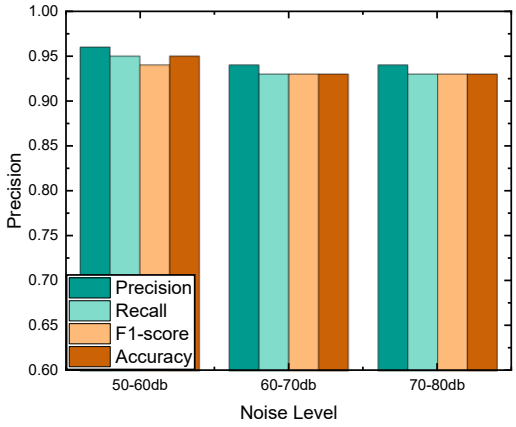


Figure 11. Evaluation of different levels of noise.

For distance evaluation, given that sign language gestures involve arm movements, they have a large range of movements in actual execution. Therefore, in order to comprehensively evaluate the recognition performance of the system, this experiment was conducted in the distance range of 10 cm to 40 cm, and the evaluation experiment was conducted every 10 cm, and the results are shown in **Figure 12**. The experimental results show that the performance of the sign language recognition system tends to decrease as the distance increases. When the sign language action distance exceeds 30 cm, the average recognition accuracy for 12 sign languages drops to about 83% or so. This is mainly due to the fact that the echo signals carrying the sign language gesture features gradually decay as the distance increases, leading to a decrease in the recognition performance of some sign language gestures involving subtle finger movements. Therefore, the optimal performance of our proposed sign language gesture recognition system is within 30 cm.

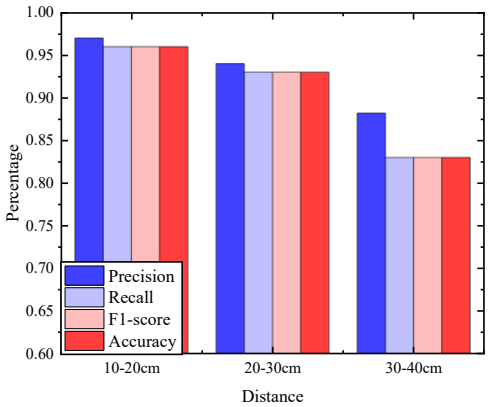


Figure 12. Evaluation at different distances.

5.4. Comparison of different methods

As described in Chapter 2, recognition solutions for sign language interaction gestures can be divided into four main categories: computer vision-based, RF signal-based, wearable device-based and acoustic signal-based. Our proposed sign language

gesture recognition system aims to provide a low-cost, non-contact, easy-to-deploy, low-risk, and widely-applied solution, and therefore compares with several existing representative non-contact sign language gesture recognition methods mainly in terms of four aspects: the recognition method, the type of sign language gesture, the required equipment, and the recognition accuracy. The results are shown in **Table 2**. Compared with existing sign language gesture recognition methods, our method has the following advantages:

(1) It only requires the built-in sensors of a smartphone without additional hardware devices. L-sign recognizes Chinese sign language gestures through a bracelet with built-in motion sensors. In contrast, our sign language recognition system can recognize Chinese sign language gestures without wearing any electronic device, which effectively improves the user experience.

(2) Recognizing slow sign language gestures: Both SonicASL and our proposed system use ultrasonic signals to sense sign language gestures. The difference is that SonicASL computes the Doppler shift of the echo signal as a sign language gesture feature, and this method is limited by the speed of the sign language performer, so if the sign language gesture is done slowly, the sign language gesture cannot be recognized effectively. In this paper, the sign language gesture features obtained from the channel impulse response can effectively characterize the sign language movements with slower execution speed, which is more friendly to the elderly group with slower execution speed.

(3) Easy to deploy without violating user privacy. Since our approach only utilizes the built-in speaker and microphone of existing commercial smartphones to recognize sign language actions. Therefore, our system is easier to deploy than RFree-GR, which is an additional hardware device. Meanwhile, compared to computer vision-based solutions that utilize a camera to capture gesture images, our solution that uses ultrasound signals to sense sign language gestures can effectively protect user privacy.

Table 2. Comparison of different methods.

Work	Method	Sign Language Category	Whether additional hardware is needed	Accuracy
L-sign [3]	Wearable device based	Chinese Sign Language	Yes	90%
SonicASL [12]	Based on acoustic signal	American Sign Language	Yes	93.8%
RFree-GR [8]	RF-based	American Sign Language	Yes	88.38%
Piskozub [2]	Wearable device based	Polish Sign Language gestures	Yes	90%
UltrasonicGS [24]	Based on acoustic signal	Chinese Sign Language	Yes	86.3%
Alyami et [25]	Computer Vision Based	Arabic Sign Language	Yes	98.25%
Our	Based on acoustic signal	Chinese Sign Language	No	95.2%

6. Conclusion

This paper presents a system for recognizing Chinese sign language gestures using acoustic signals sent from a smartphone. The system sends ultrasonic signals through the built-in speaker of a smartphone to sense sign language gesture movements. Afterwards, the fine-grained sign language gestures are characterized by calculating the channel impulse response of the received signals. To achieve real-time

recognition of sign language gestures, we compute the variance of the channel impulse response to segment the effective gesture actions. Finally, we recognize Chinese sign language gestures by combining a convolutional layer with a classification model of channel attention and spatial attention. The experimental results show that our sign language recognition system has a recognition accuracy of 95.2 for 12 commonly used Chinese sign language gestures.

Author contributions: Conceptualization, YS and YW; methodology, YS; software, YS and QL; validation, QL and JZ; data curation, YS and JZ; writing—original draft preparation, YS; writing—review and editing, YS and YW; visualization, QL. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Natural Science Foundation of Henan (202300410190), Science and Technology Research Project of Henan Province (212102110227), Starting Fund for Doctoral Research Project (2020HNUAHEDF007).

Acknowledgments: Thank you to the mentor and classmates for their contributions to this article. The authors wish to express their appreciation to the reviewers for their helpful suggestions which greatly improved the presentation of this paper.

Ethical approval: Not applicable.

Conflict of interest: The authors declare no conflict of interest.

References

1. WHO. Deafness and hearing loss[EB/OL]. https://www.who.int/,2024-04-01.
2. Piskozub J, Strumillo P. Reducing the number of sensors in the data glove for recognition of static hand gestures[J]. Applied Sciences, 2022, 12(15): 7388.
3. Zheng Z, Wang Q, Yang D, et al. L-sign: Large-vocabulary sign gestures recognition system[J]. IEEE Transactions on Human-Machine Systems, 2022, 52(2): 290-301.
4. Zhang Q, Jing J Z, Wang D, et al. Wearsign: Pushing the limit of sign language translation using inertial and emg wearables[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2022, 6(1): 1-27
5. SS P D, Patil S B, Patil B S. Gesture Recognition Machine Vision Video Calling Application Using YOLOv8[C]//2023 22nd International Symposium on Communications and Information Technologies (ISCIT). IEEE, 2023: 105-109.
6. Zhu C, Yang J, Shao Z, et al. Vision based hand gesture recognition using 3D shape context[J]. IEEE/CAA Journal of Automatica Sinica, 2019, 8(9): 1600-1613.
7. Saboo S, Singha J. Vision based two-level hand tracking system for dynamic hand gestures in indoor environment[J]. Multimedia Tools and Applications, 2021, 80(13): 20579-20598.
8. Dian C, Wang D, Zhang Q, et al. Towards domain-independent complex and fine-grained gesture recognition with RFID[J]. Proceedings of the ACM on Human-Computer Interaction, 2020, 4(ISS): 1-22.
9. Cai X, Ma J, Liu W, et al. Efficient convolutional neural network for fmcw radar based hand gesture recognition[C]//Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers. 2019: 17-20.
10. Liu H, Cui K, Hu K, et al. MTransSee: Enabling environment-independent mmWave sensing based gesture recognition via transfer learning[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2022, 6(1): 1-28.
11. Du H, Li P, Zhou H, et al. Wordrecorder: Accurate acoustic-based handwriting recognition using deep learning[C]//IEEE INFOCOM 2018-IEEE Conference on Computer Communications. IEEE, 2018: 1448-1456.

12. Jin Y, Gao Y, Zhu Y, et al. Sonicasl: An acoustic-based sign language gesture recognizer using earphones[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2021, 5(2): 1-30.
13. Xu H, Wang T, Wang X, et al. SignID: Acoustic-based Identification with Single Sign Gesture[C]//2021 7th International Conference on Big Data Computing and Communications (BigCom). IEEE, 2021: 98-105
14. Ling K, Dai H, Liu Y, et al. Ultragesture: Fine-grained gesture sensing and recognition[J]. IEEE Transactions on Mobile Computing, 2020, 21(7): 2620-2636.
15. Zhang Q, Wang D, Zhao R, et al. Soundlip: Enabling word and sentence-level lip interaction for smart devices[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2021, 5(1): 1-28.
16. Yun S, Chen Y C, Zheng H, et al. Strata: Fine-grained acoustic-based device-free tracking[C]//Proceedings of the 15th annual international conference on mobile systems, applications, and services. 2017: 15-28.
17. Zhao R, Wang D, Zhang Q, et al. Smartphone-based handwritten signature verification using acoustic signals[J]. Proceedings of the ACM on human-computer interaction, 2021, 5(ISS): 1-26.
18. Chen Y, Ni T, Xu W, et al. SwipePass: Acoustic-based second-factor user authentication for smartphones[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2022, 6(3): 1-25.
19. Wang Z, Wang Y, Tian M, et al. HearFire: Indoor Fire Detection via Inaudible Acoustic Sensing[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2023, 6(4): 1-25
20. Zhang Y, Huang W H, Yang C Y, et al. Endophasia: Utilizing acoustic-based imaging for issuing contact-free silent speech commands[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2020, 4(1): 1-26
21. Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.
22. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
23. Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks. ArXiv. Org[J]. 1801.
24. Wang Y, Hao Z, Dang X, et al. UltrasonicGS: A highly robust gesture and sign language recognition method based on ultrasonic signals[J]. Sensors, 2023, 23(4): 1790.
25. Alyami S, Luqman H, Hammoudeh M. Isolated arabic sign language recognition using a transformer-based model and landmark keypoints[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2024, 23(1): 1-19.