Smart Health 32 (2024) 100463

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

APC: Contactless healthy sitting posture monitoring with microphone array

Kaiyuan Ma [a], Shunan Song [a], Lingling An [b], Shiwen Mao [c], Xuyu Wang [d,*]

[a] Guangzhou Institute of Technology, Xidian University, Guangzhou, China
[b] School of Computer Science and Technology, Xidian University, Xi'an, China
[c] Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849, USA
[d] Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, FL 33199, USA

ARTICLE INFO

Keywords:
Acoustic sensing
Posture detection
Microphone array
Smart sensing

ABSTRACT

The prevalence of poor sitting posture in daily work has become a growing concern among office workers and students due to the associated health problems. To address this issue, we design an acoustic sitting posture care system (termed, APC) based on a circular microphone array. Compared with classic posture recognition technologies such as visual perception and sensors, acoustic sensing naturally possesses advantages such as privacy protection and contactless capabilities. Concretely, our system leverages a customized and inaudible sound signal sent from a speaker to a user's body, and an echo signal preprocessing method to sense the body posture. Our system comprises three modules: signal generation and collection, signal preprocessing, and posture classification. The signal generation and collection module is designed to create an appropriate signal waveform for transmitting the sound signal. We also develop a unique alignment method for received signals to implement background interference cancellation. In the signal preprocessing module, we propose a body profile extraction method based on the phase difference between received signals. In the posture classification module, we design an attention mechanism based classification network that can map the output of the previous module to different sitting posture categories. The experimental results show that our proposed method achieves an average accuracy of 98.4% for five common sitting postures. Furthermore, case studies conducted under different practical conditions have validated the robustness of our system.

1. Introduction

Long-term poor posture can lead to serious cervical and lumbar vertebrae diseases (channal, 2015), which are difficult to be noticed early and treated. Moreover, bad sitting posture for a long time may also result in constipation and other non-spinal diseases (Health, 2001). Unfortunately, a long time sitting during work often makes people ignore the posture change, and people always have maintained harmful sitting postures for a period before they realize the problem of their sitting posture. Therefore, a contact-free, low-cost, and real-time sitting posture recognition system would be highly appealing, which will be helpful to remind users to correct their sitting posture, thus avoiding potential disease risks.

Currently, as a controller of household appliances and music players, smart speakers are rapidly becoming popular in families. Beyond the basic functions of smart speakers, smart speakers have been also developed for several Internet of Things (IoT)

* Corresponding author.
E-mail addresses: 21181214038@stu.xidian.edu.cn (K. Ma), 21181214417@stu.xidian.edu.cn (S. Song), all@mail.xidian.edu.cn (L. An), smao@ieee.org (S. Mao), xuywang@fiu.edu (X. Wang).

https://doi.org/10.1016/j.smhl.2024.100463
Received 16 March 2024; Accepted 18 March 2024
Available online 18 April 2024

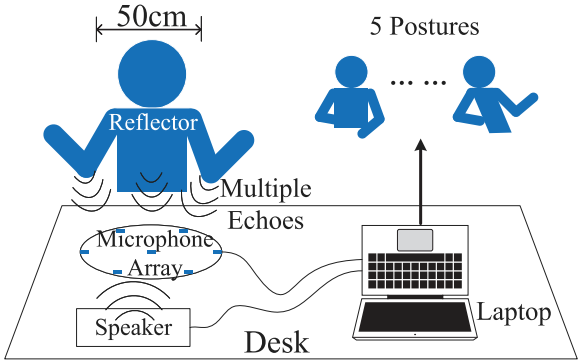


Fig. 1. System scenario.

applications (e.g., gesture recognition, vital monitoring, facial expression monitoring) based on the microphone array (Cai, Zheng, & Luo, 2022; Chara, Zhao, Wang, & Mao, 2023; Shan, Liao, Wang, An, & Mao, 2023; Wang & Mao, 2022). For example, RTrack (Mao et al., 2019) uses a chirp to detect gestures and capture hand locations from a 2D-MUSIC spectrum which is derived by a chirp echo. To further extract echo features, Sonicface (Gao et al., 2021) uses the Doppler effect to extract frequency shifts from facial expressions' tiny differences. Also, UltraGesture (Ling et al., 2020) uses Channel Impulse Response (CIR) to achieve minor finger motion recognition from micro-channel differences. Moreover, acoustic sensing is used for vital sign monitoring (e.g., Sonarbeat Wang, Huang, and Mao (2017), Wang, Huang, Yang, and Mao (2021), and RespTracker (Wan, Shi, Cao, Wang, & Chen, 2021)). In addition to acoustic-based sensing, computer vision and sensors are also used for posture recognition. For example, Sun, Zhu, Cui, and Wang (2021) collect data through a Kinect and classify images by a Convolutional Neural Network (CNN), which can obtain a good performance. Maereg, Lou, Secco, and King (2019) use near-infrared and a sensing wristband to collect hand motion data. Farnan, Dolezalek, and Min (2021) concentrate on sitting postures detection and use a magnetic sensor, which is implemented on users' shirts to collect users' posture data.

Although many existing techniques can achieve a preferable system performance on posture recognition, the methods based on acoustic sensing focus on small targets at close distances or big targets at far distances. Thus, the potential of the sound signal remains to be limited. Methods based on computer vision can be seriously affected by weak light conditions, scene variation, and differences in human wearing, which will result in unstable system performance. Moreover, wearable devices or sensors require expensive equipment with poor versatility and poor user experience. Therefore, a contactless, low-cost, and high-accuracy posture recognition system is still needed.

In this paper, we present a contactless healthy sitting posture monitoring system utilizing a microphone array. The development of our system entails addressing three key challenges. First, commercial microphone arrays often receive signals with uncontrolled deviations on time, leading to nonuniform signal latency and significant errors. Thus, our first challenge involves devising an alignment scheme for the signals. Second, extracting body profile features from the echo presents another challenge. While conventional signal direction finding algorithms prove effective for acoustic hand tracking in narrowband signals, their limitations arise when dealing with a larger number of signals than the available microphones. Moreover, the performance of such algorithms deteriorates as the number of signals increases. In our system, the echo signal from the body will be composed of signals from many different directions and distances, which requires system to analyze multiple distance, altitude, and azimuth information from the echo to identify posture, and this process will inevitably introduce noise. Therefore, constructing a pose classification network for preprocessing results is the third challenge.

To address the aforementioned challenges, we have developed the Acoustic sitting Posture Care system (termed, APC), a contactless, real-time, and highly accurate sitting posture recognition system (see Fig. 1). Our system incorporates several key innovations. First, we utilize a unique sound signal segment as the reference point for signal alignment. Through cross-correlation within a limited sensing range, this approach ensures accurate alignment of the transmission and received signals. Second, we address the second challenge by normalizing the phase of every sample in received signals to zero at a specified time, effectively mitigating interference from undesired positions. Third, we have designed a more lightweight classification network compared to classic Transformer named Posture Echoes Transformer (PET) based on Vision Transformers (ViT) specifically tailored for recognizing common postures. Our comprehensive experimental results demonstrate the better performance of the APC system across various usage scenarios. The primary contributions of this paper are summarized as follows.

- We develop a prototype for the proposed APC system, which is a contactless, low-cost, and adaptable acoustic sitting posture recognition system with a microphone array.
- We leverage the slow propagation property of the sound and the cross-correlation to align echo signals with significant waveform variations, which will not interfere the normal sensing.
- According to the acoustic waveform features, we design a novel body profile extraction method using a circular microphone array, and we propose a new classification network (i.e., PET) including a encoder–decoder feature extraction based on ViT.

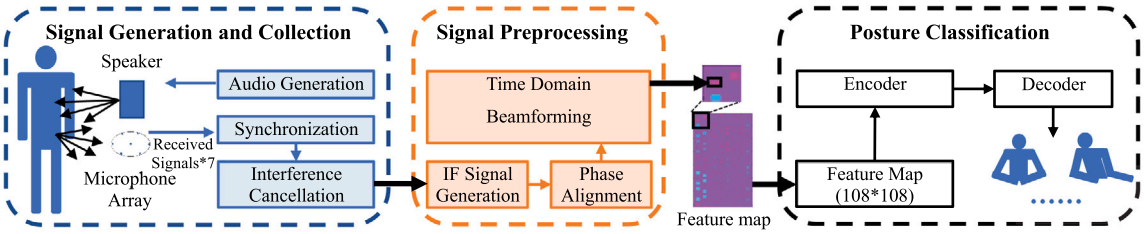


Fig. 2. System overview.

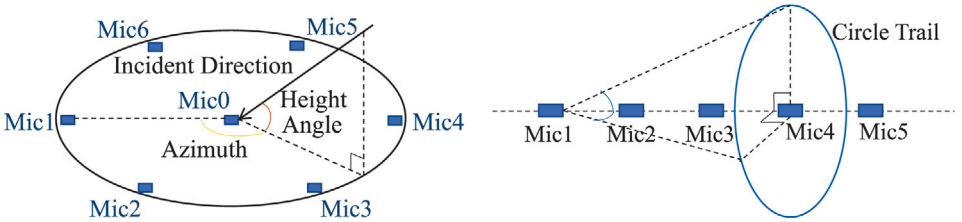


Fig. 3. Microphone array formats.

- We evaluate the performance of the system in different experimental scenarios. The results show the APC system can reach 98.4% for five common sitting postures. Also, case studies under different practical conditions validate the robustness and versatility of our system.

The rest of this paper is arranged as follows: In Section 2, the system design is described. In Section 3, the performance of the system is evaluated. Finally, we conclude this paper in Section 4.

2. System design

In this section, we present three modules of our system as illustrated in Fig. 2. We begin by introducing circular microphone array and the signal generation and collection module, followed by the signal preprocessing module, where we present a human-body ultrasonic reflection model based on a customized sound waveform and then preprocess the signals using this model. Finally, we describe the design of our classification network and provide a detailed explanation of the design method used.

2.1. Circular microphone array

Most acoustic applications utilize pure tone signals and chirp signals. The pure tone signal is employed to capture the phase difference resulting from small movements of the target reflector, making it suitable for motion detection (e.g., fitness movement detection) and heartbeat detection. However, in the case of a stationary sitting position, the chirp signal is required to detect the distances associated with body profiling. However, a single microphone cannot complete the detection of the echo arrival angle, which will cause serious errors in recognition for different postures at the same distance. To find the directions of the whole body, the reflection signal is used to estimate its arrival angle using a microphone array (e.g., a linear microphone array or a circular microphone array). We need to consider which microphone array is better in our usage scenario. Fig. 3(b) shows a linear microphone array. Any reflector on a constant circular trail that is perpendicular to the microphone array could reflect the echo with the same incident angle to a single microphone in the microphone array. On the other hand, the circular microphone array is shown in Fig. 3(a). We assume that the source of the echo wave is at infinity, such that the echoes come from a same distance are parallel, and then the arrival height angle and direction angle can be determined based on the time difference between the echoes arriving at the microphone array. Therefore, the circular form is better than the linear one for our APC system, where the angle from the center microphone to any two adjacent edge microphones in our work is 60°, and the radius of the microphone array is 4.3 cm.

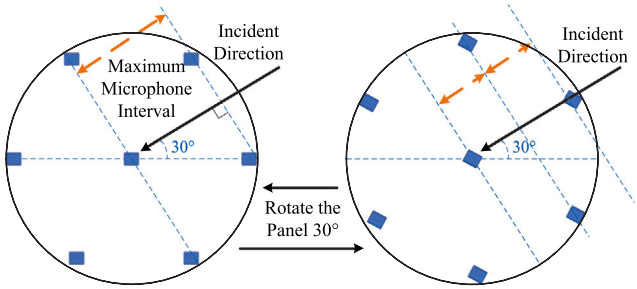


Fig. 4. Maximum microphone interval along the incident direction of echo.

2.2. Signal generation and collection

In this module, our system employs a novel enhancement scheme to the conventional Frequency Modulated Continuous Wave (FMCW) method for signal preprocessing. This enhances the robustness of the signal preprocessing results by increasing the base frequency of the Intermediate Frequency (IF) signal. We also introduce the specific frequency band of the IF signal. Next, we describe the use of a customized sound waveform for the speaker and align every received signal based on this waveform. Finally, we further discuss the resolution of our system based on the improved FMCW and customized waveform. We provide a detailed explanation of the design mentioned above in the remain of this section.

High frequency IF signal: Our method differs from the conventional FMCW method in that the frequency of the IF signal plays a crucial role in the robustness of our APC system. Specifically, we use a higher frequency for the IF signal. We will introduce an example to illustrate the reason. If the IF signal frequency raises up from 0 Hz and the microphone array receives a reflection signal from 30 cm away, we can only obtain an IF signal at 338 Hz approximately. In fact, the wavelength of the IF signal will reach up to 1 m approximately, which will make 0.043 as the largest phase difference between received signals. Furthermore, the preprocession result will become bad. On the other hand, a higher frequency has a smaller wavelength, and this will lead to ambiguity when the microphones interval along the incident direction is larger than half of the wavelength because when we preprocess signals in this case, two signals at the same distance but with phase difference as π will be considered same. Thus, we consider the half of wavelength of the IF signal to be larger than the microphones interval of the microphone array. As shown in Fig. 4, the microphone interval along the incident direction is changing with the change of incident direction, and its variation is also determined by the layout of microphone array. Then we have the maximum interval from the azimuth, which is shown in Fig. 4. Since we have the radius of the microphone array as 4.3 cm, then the maximum interval is 3.72 cm approximately. Thus, the minimum wavelength is 7.44 cm, and then the highest frequency of the IF signal will be set as 4600 Hz. This implies that the frequency of the chirp signal, which is multiplied by the received signal, starts at 11400 Hz.

Signal composition: The transmission signal is combined of two parts, which are positioning segment and chirp segment. To ensure that the echoes from the same distance have the same FMCW frequency characteristics, the echo signal and the transmission signal are required to be aligned. However, most of commercial microphone arrays cannot satisfy this requirement, where each echo signal will appear with a tiny deviation from the transmission signal on the start time. In previous work (Gao et al., 2021), the alignment scheme commonly is to execute cross-correlation for two signals, which can compensate a offset for the one of the signals. However, the scheme is not proper for our work. The result of the alignment scheme above are shown in Fig. 5(b), and the static interference result based on it is shown in Fig. 5(f). As we can see, even if the user is far enough from the microphone to ensure that theoretically the positioning segments from the transmitted and reflected signals do not overlap, after the static interference cancellation, there is still a significant signal strength before the positioning segment of the reflected signal. In the previous work, the sensing process concentrates on a tiny target such as face or hand, however, our work is sensing on a body which is much larger than face or hands. The difference between different usage scenarios leads to that in our case, the difference between static interference and sensing signal is significant, and thus the alignment scheme based on cross-correlation method on complete signals is no longer available. Therefore, we introduce *a novel alignment scheme* that requires to add a segment (i.e., a positioning segment) containing two piece of pure tone signal (i.e., a positioning signal) before the chirp signal, and then we will align the received signal and the transmission signal by shifting the received signal based on the cross-correlation between positioning segments.

Aligning an echo signal with a transmission signal by cross-correlation is intuitive and effective, but there are still two problems as follows. First, the echo will overlap with the transmission signal if a long positioning signal is used, while the echo also will become too weak to implement the cross-correlation if a short positioning signal is used. The above two conditions will cause the signal to have different shifts. To address the problem, we use two combination of a short pure tone signal and a blank signal, which will replace the complete transmission signal in cross-correlation, which is shown in Fig. 6(a). More importantly, the blank part will denoise the positioning signal from echo in cross-correlation procedure, and the two short pure tone signals will also result into a great cross-correlation effect, which is illustrated in Fig. 5(d). The only limitation of this method is that the detectable range should ensure that the echo caused by the first positioning signal will not overlap itself or the second positioning signal.

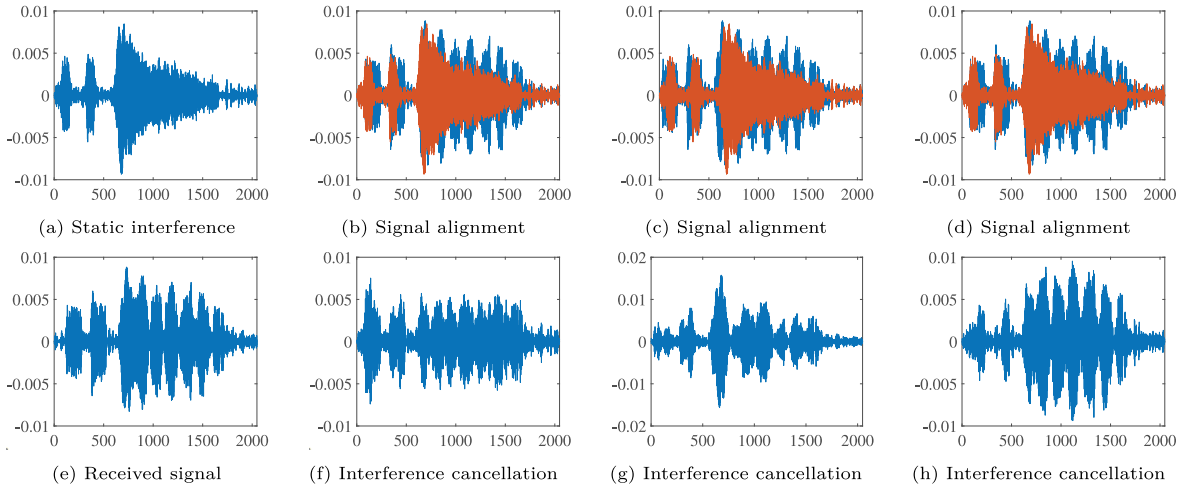


Fig. 5. Performance of two positioning segments in static interference cancellation. (b), (c) and (d) respectively show the signal alignment results with 0, 1 and 2 positioning segments; (f), (g) and (h) respectively show the performance of static interference cancellation according to the signal alignment results of (a), (b) and (c).

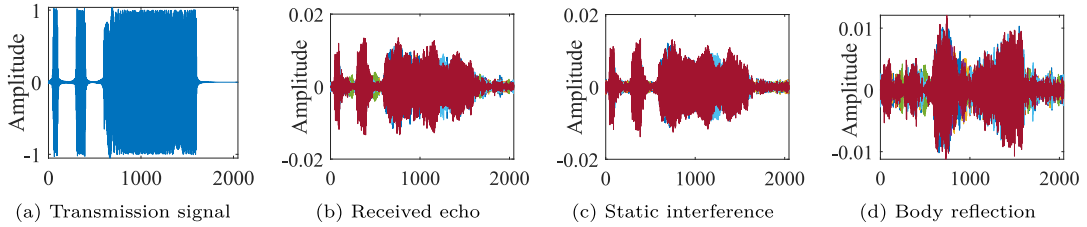


Fig. 6. Static interference cancellation.

To obtain a high-range resolution, we need to make the bandwidth of the chirp signal as wide as possible in our system, but we also need to avoid the low sound frequency response of most commercial speakers. Also, the transmission signal needs to meet the requirement that its highest frequency is less than the maximum working frequency of most speakers and its lowest frequency should not be within the audible range of people, as shown in Fig. 6(a). Thus, we use the chirp bandwidth B with 4 kHz, the length of the chirp signal S_{chirp} with 1000 samples, and the sampling rate f_s with 48 kHz. Specifically, our final transmission signal is shown in Fig. 6(a), where the chirp signal segment is joined by a long blank sample segment with $N_{tail} = 448$ as the length. This means the valid sensing distance is within $D_{valid} = \frac{cN_{tail}}{2f_s} = 1.586$ m of our system, which is sufficient for most sitting posture recognition tasks.

2.3. Signal preprocessing

After acoustic signals are aligned with the waveform discussed, it is ready to obtain the preprocessed data for classification. Next, we will perform interference cancellation and generate the feature matrix based on time domain beamforming.

In order to effectively adapt the system to various environments, it becomes imperative to mitigate the impact of static responses induced by the surroundings. This is particularly crucial as real-life usage scenarios often involve energy reflections originating from the environment that surpass those originating from the human body. Failing to eliminate these environmental reflections would significantly impair the system's accuracy and robustness. To address this problem, we have developed a two-step method for static interference cancellation, we pre-record the static reflection and multiply the power coefficient Ac for the signal, where Ac is the energy ratio of the sum of the absolute values of the pre-recorded signal's position segment and the received signal's position segment. Then, we subtract the pre-recorded static interference from the received signal and generate IF signals from that as the result of the first step. In the second step of our approach, we capture the peak of the IF signal spectrum in the inaugural sensing phase. Subsequently, we define an IF frequency range centered around the peak above with a 250 Hz radius and meticulously track alterations in this pinnacle within the designated frequency span in every successive sensing cycle. If the peak is within the sensing range we have set, it is sufficient for the system to sense a height of 40–48 cm above the chest, which is sufficient for most people. This iterative process ensures the refinement of the peak's representation over time. As a concluding step, we apply a filtering mechanism to eliminate signals that fall outside of this frequency span. This strategic approach culminates in the achievement of effective background noise cancellation.

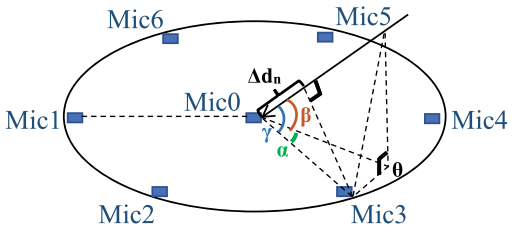


Fig. 7. Triangular pyramid model.

Besides, we derive the formula for generating the feature matrix, which will serve as input for the classification module. We start with a simple model of the reflector as shown in Fig. 3(a). Subsequently, we introduce constraints to the model, leading to the derivation of the ultimate formula for generating a single value of a feature matrix. When a reflector is fixed at a certain distance, azimuth, and height angle from the central microphone (the microphone number 0 in Fig. 3(a)), the IF signal received by the central microphone can be expressed as following,

$$R_{if0} = A_0 \cos(2\pi f_{if0}(t - t_0)), \tag{1}$$

where R_{ifn} is the IF signal from received signal on n_{th} microphone, A_0 is the attenuation coefficient, f_{if0} is the frequency of R_{if0}, t is the time, and t_0 is the time delay of signal flight. After that, we need to calculate the propagation distance deviations between the signal received by the central microphone and signals received by other microphones so that we can further get expressions of signals received by other microphones.

To determine one of the distance deviations, it is necessary to first calculate the angle between the signal direction and the line connecting the central microphone to a surrounding microphone. This angle can be obtained using a triangular pyramid model. In this model, the vertex represents the position of the central microphone, and the three edges correspond to the signal direction, the projection of the signal directly on the microphone array, and the line connecting the central microphone to the surrounding microphone, respectively. Additionally, the bottom surface of the pyramid passes through the surrounding microphone and is perpendicular to the projection of the signal directly on the microphone array. As shown in Fig. 7, we have the relation of γ, α, β, and θ based on (Wang et al., 2023),

$$\cos(\gamma) = cos(\alpha)cos(\beta) + sin(\alpha)sin(\beta)cos(\theta). \tag{2}$$

Due to the angle $\theta = \pi/2$, we have that $\cos(\gamma) = cos(\alpha)cos(\beta)$. Then, the wave path difference between R_{if0} and R_{ifn} can be represented as,

$$\Delta d_n = rcos(\gamma), \tag{3}$$

where r is the radius of the microphone array, and n is the index of n_{th} surrounding microphones. For the case that the reflector is remote, all microphones are considered to receive the same frequency of the IF signal of the reflector, and the IF signal of the number n microphone can be expressed by

$$R_{ifn} = A_n \cos(2\pi f_{if0}(t - t_0) + P_{ifn}), \tag{4}$$

$$P_{ifn} = \frac{2\pi \Delta \mathrm{d}_n f_{if0}}{c}, \tag{5}$$

where c is the sound speed, P_{ifn} is the phase difference between R_{if0} and R_{ifn} caused by Δd_n. In the real usage scenario, we change the reflector mentioned above to a point on a user, and this will result in differences in the frequency of IF signals received by different microphones. According to FMCW method, the frequency of IF signals can be deduced from known values and Δd_n we mentioned above. After the frequency correction, R_{ifn} and P_{ifn} can be further expressed by

$$R_{ifn} = A_n \cos(2\pi(f_{if0} + \frac{B\Delta \mathrm{d}_n}{S_{chirp}c} f_s)(t - t_0) + P_{ifn}), \tag{6}$$

$$P_{ifn} = \frac{2\pi \Delta \mathrm{d}_n \left(f_{if0} + \frac{B\Delta \mathrm{d}_n}{S_{chirp}c} f_s \right)}{c}. \tag{7}$$

We can perform the reflector detection on several fixed distances and different discrete directions. Then, we infer the presence of a reflector by detecting whether the echo generated at this location has caused an IF signal combination of the corresponding frequency on the microphone array. After that, we will introduce how we detect a reflector on a certain location. Since the start frame and the phase of an IF signal which at a specified frequency can be calculated, we add phase for the start frame of every IF signal, and then the following frames can obtain their imaginary parts according to the start frame. Then, we can transform real signals into complex signals coarsely. Because in real cases, echoes come from a large part of the user's body, resulting in many IF

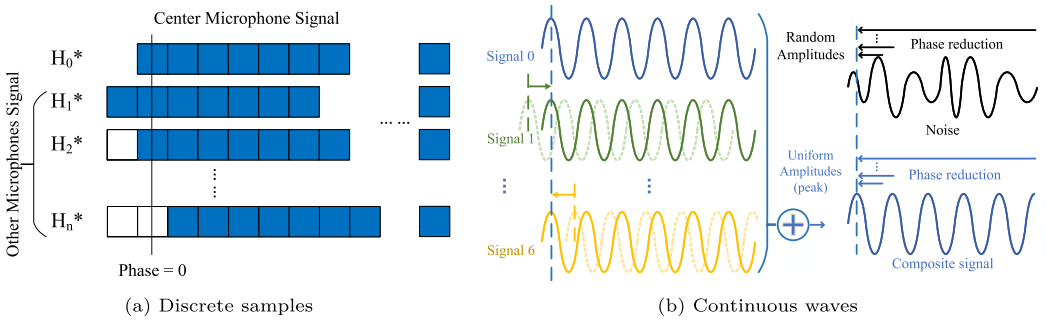


Fig. 8. Time domain beamforming.

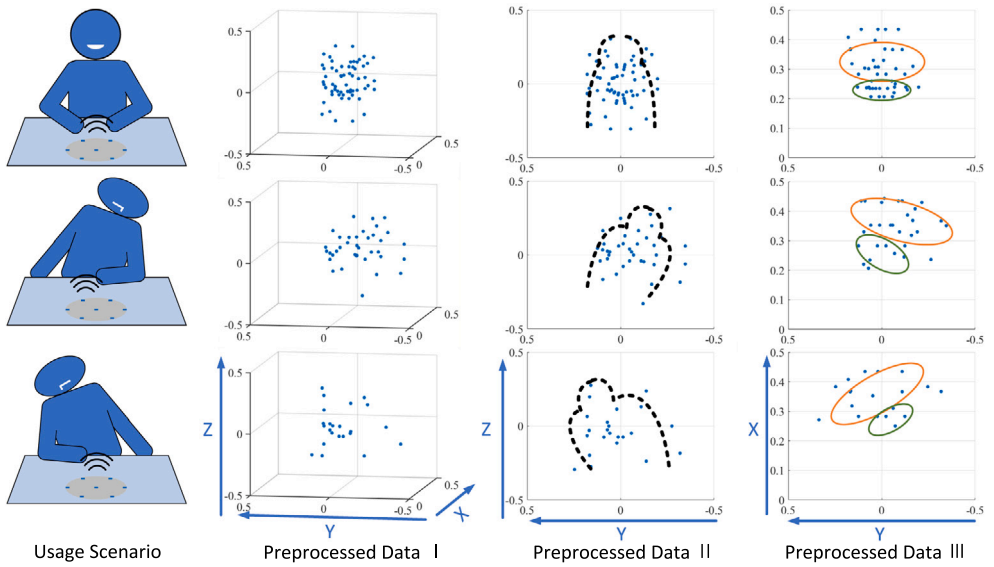


Fig. 9. Preprocessed data. The coordinate origin of the scatter plots is located at the center microphone, and the x axis, y axis and z axis represent the sensing distance, azimuth, and height angle, respectively, and I/II/III represent different viewing directions, with each column having the same viewing direction. Besides, the shown points are the preprocessing results which are stronger than 1/10 average of the preprocessing result.

signals at different frequencies in a signal which can be processed by FMCW method. Also, in each signal we generated by FMCW method, only the IF signal generated by the echo reflected from the desired direction and distance is added with the accurate phase, and other parts of the real signal are added with biased phase. We can leverage the inaccurate complex IF signal to weaken signals from undesired positions.

Next, we implement time domain beamforming to obtain a single value of the feature matrix which implies the intensity of the echo reflected by the desired location, as shown in Fig. 8. Specifically, to attach phase to signals which produced by the FMCW method, we multiple different weights H_n with every sample, and H_n can be expressed by $[e^0, e^{1j\omega_n T_s}, e^{2j\omega_n T_s}, \ldots, e^{Lj\omega_n T_s}]$, where n is the index of n_{th} microphone, ω_n is the angular frequency of the IF signal generated by the echo reflected by the desired location and received by n_{th} microphone, L is the number of samples of the overlapped signal in FMCW procession, T_s is sample interval (i.e., 1/48000 s). We will restore each sample to a phase of zero based on the additional phase and frequency mentioned above, and then we add all the samples up to a single value. Finally, the noise is canceled and the signals from the neighboring area are also weakened.

The frequency of the center microphone IF signal is the same when the signal comes from the same distance. Thus, we first generate a series of body profile values from the same distance but a different azimuth and height angle, and then we only need to implement the calculation for different distances so that we can obtain the result of signal preprocessing. Because the sound reflection and diffusion pattern are complex on various reflectors (Shtrepi, Astolfi, Puglisi, & Masoero, 2017), the intensity of a pixel on a preprocessing result may be weaker, as shown in Fig. 9. The preprocessing cannot capture all the reflection points on the body, and the reflection after twice bounce may wrongly captured. To solve the problem above, we build a machine-learning model for the classification of the preprocessing results.

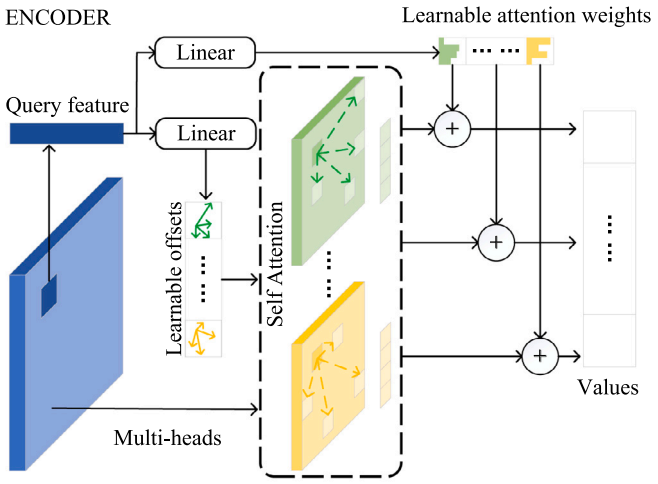


Fig. 10. Encoder in classification network.

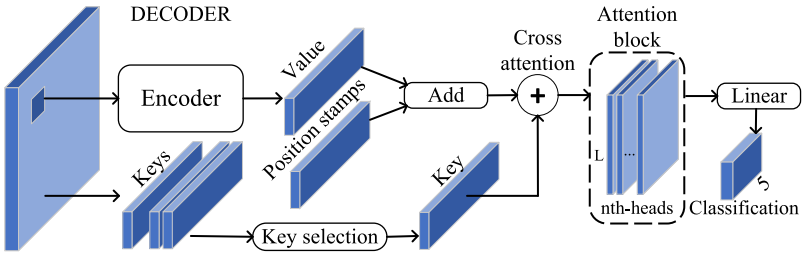


Fig. 11. Decoder in classification network.

2.4. Posture classification

In the context of posture classification, we leverage data derived from an array of matrices, produced by our preceding signal processing technique at varying distances. This data exhibits a strict spatial hierarchy, mirroring the nature of human postures which extend from far to near. Consequently, posture information can be extracted from these diverse spatial distances. Moreover, as human sitting postures are typically confined within a limited spatial domain, the information within this range also presents continuous characteristics, a reflection of the human body's spatial continuity.

Generally, the ViT model can be aptly applied to our data structure for posture classification (Dosovitskiy et al., 2020; Vaswani et al., 2017). In fact, due to the discrepancy between auditory and visual data, ViT tends to overemphasize noise in the attention mechanism, impeding the model's rapid convergence. To address this issue, we introduce a refined attention model (i.e., PET) that meticulously filters the Query (Q) and Key (K) elements, thereby accelerating the model's convergence.

The PET model, as depicted in Fig. 10 and Fig. 11, comprises three main components. The initial component is a feature extraction backbone built upon a shallow-layer CNN. This backbone is designed to capture coarse information via a small receptive field. In this stage, we opt for a shallow variant of ResNets (He, Zhang, Ren, & Sun, 2016) as our network backbone. To obtain global information, we employ a unique classification token (CLS) that is concatenated onto the backbone.

The second component of the PET model is a filterable self-attention transformer layer. We feed the coarse information processed from the previous stage into this transformer layer, following which we apply position embedding to the input sequence. Subsequently, we construct a transformer encoder by stacking a series of alternating multi-head filterable self-attention blocks and Multilayer Perceptron (MLP) blocks. Each block is accompanied by a residual connection and is succeeded by layer normalization. Upon completion of processing within this second component, we acquire coarse-to-fine characteristic information. We iterate this step multiple times to extract progressively precise hidden layer features. Ultimately, this fixed-size vector undergoes a noise-filtering decoder layer to yield the predicted pose class. We delve into further details of the model (i.e., model input, filterable self-attention block, multi-head attention, and model output) as the following.

Model input: Our transformer-based model primarily ingests a multi-channel, compact data matrix ($36 \times 18 \times 18$, with the final dimension, 18, representing channels). This data matrix is first fed into shallow ResNets to extract coarse information denoted as $x_i = x_1, \ldots, x_c$, where $x_i \in \mathbb{R}^{H \times W}$, $i = 1, \ldots, 256$. We then expand the original single-dimensional feature into 256 hidden layer features. Following this expansion, we flatten the data matrix for the transformer, resulting in an input represented as $x_i \in \mathbb{R}^{H \times W}$.

To extract coarse-to-fine characteristic information and incorporate global feature information, we employ a CLS. We generate a layered matrix of equivalent size through a fully connected layer and concatenate it onto the matrix from the previous step. This operation leaves the channel count of our new matrix unchanged, but increments its dimension by one, resulting in $x_i \in \mathbb{R}^{H\times W+1}$. CLS subsequently serves as the final classification layer, facilitating precise classification.

Filterable self-attention block: The primary goal of filterable self-attention blocks is to comprehensively account for the semantic, spatial, and structural relationships among distinct posture patches across all elements in the input data matrix. As a result, the posture patches calculated in this manner take into account the interrelationships among contexts. We introduce x_i into three trainable linear layers to obtain the matrix $x_i W^{Q,K,V} \in \mathbb{R}^{H\times W+1}$, which maintains the input structure. Subsequently, x_i is multiplied with the weight matrix W^Q to procure the Query vector associated with the posture patches. Ultimately, a Query vector, a Key vector, and a Value vector are generated for each patch in the input matrix.

Besides, we compute the attention score. Considering the computation of self-attention for the first posture patch, we need to calculate a score for each posture patches across the entire matrix relative to this patch. These scores serve a crucial function in the context of encoding the initial patch, specifically highlighting the significance of position embedding. These scores are derived by calculating the dot product of the Query vector of the posture patch and the Key vector of the patch. Each score is then divided by $\sqrt{d_i}$ to stabilize the gradient during backpropagation. Following this, softmax operations are performed on these scores to normalize them. Each softmax score is then multiplied by its corresponding Value vector to obtain $z = z_1, \ldots, z_c$, where $z_i \in \mathbb{R}^{H\times W+1}$, $i = 1, \ldots, 756$. For positions with high scores, the resulting multiplied values will be larger, warranting increased attention; conversely, positions with low scores yield smaller multiplied values, indicating less relevance and thus less attention required. We then sum the weighted Value vectors obtained in the previous step to yield the output of the self-attention layer at this position. This completes the self-attention computation. The resulting vector will be input into a feedforward network. In practice, however, this computation is performed matrix-wise for accelerated processing. Our calculation formula is defined as follows,

$$z_i = \sum_{j=1}^{n} \mathrm{softmax}\left(\frac{\left(x_i W^Q\right)\left(x_j W^K\right)^T}{\sqrt{d_i}}\right)\left(x_j W^V\right). \tag{8}$$

Building upon this, we address an important issue, i.e., the fundamental challenge of applying Transformer attention to a compact data matrix is its exhaustive attention to all possible spatial locations. To solve this, we introduce a filterable attention module. This module selectively attends to a limited set of key sampling points around a reference point, independent of the spatial size of the feature maps, as illustrated in Fig. 10. By assigning only a small, fixed number of keys for each query, we can alleviate the convergence issue.

Given a data matrix $x_i \in \mathbb{R}^{H\times W}$, and let q index a query element with a reference point st_q. The filterable attention feature is computed by,

$$Filterable(st_q, x_i) = \sum_{k=1}^{K} A_{qk} \bullet W x_i(\mathrm{st}_q + \Delta st_{qk}), \tag{9}$$

where k indexes the sampled keys and K represents the total number of sampled keys ($K \ll HW$). Δst_{qk} and A_{qk} denote the sampling offset and attention weight of the most significant sampling point in $H\times W$, respectively. The scalar attention weight A_{qk} falls within the range $[0, 1]$, which is normalized such that $\sum_{k=1}^{K} A_{qk} = 1$. As $st_q + \Delta st_{qk}$ is fractional, a bilinear interpolation is applied.

Multi-head attention: We further enhance the self-attention layer by incorporating multi-head attention. Each attention set maps the input to various sub-representation spaces, enabling the model to focus on different positions within these subspaces. The entire computation process can be expressed by

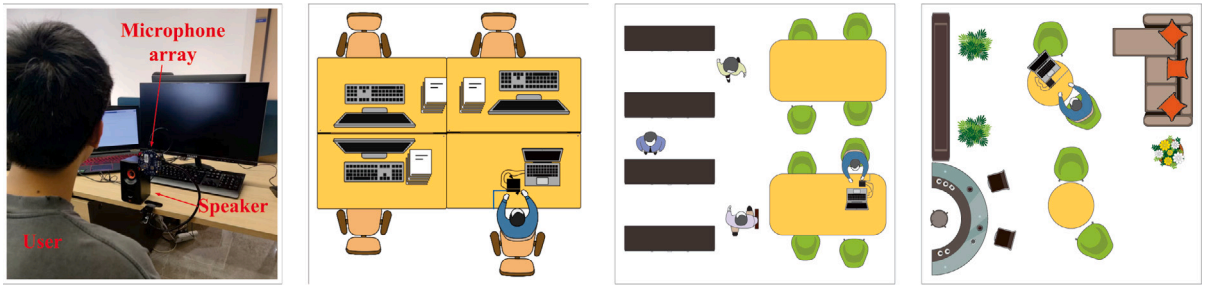
$$MultiHead\,(Q, K, V) = Concat\left(\mathrm{head}_1, \ldots, \mathrm{head}_h\right) W^O, \tag{10}$$

$$\mathrm{head}_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{11}$$

where $W_i^Q \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^K \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^V \in \mathbb{R}^{d_{model}\times d_k}$ and $W_i^O \in \mathbb{R}^{hd_v\times d_{model}}$. In our experiment, we set $h = 8$ and $d_k = d_v = \frac{d_{model}}{h} = 32$.

Following the above self-attention methodology, we can obtain 12 distinct matrices z_i^* by applying 12 multi-head attention calculations with differing weight matrices. Ultimately, we concatenate these 12 matrices z_i^* to form z_i and multiply it by a weight matrix W_i^O after flattening. This yields the final matrix z_i, which encapsulates all the information from the attention heads. This matrix is then input into the MLP layer.

Model outputs: After the sequence passes through the Transformer encoder, a set of high-level features, x^{end}, can be derived. In the decoder layer, to extract more effective features and eliminate interfering noise, we propose a new strategy to isolate the top K effective features to enhance classification accuracy. As illustrated in Fig. 11, we select the top M keys with the highest scores through a scoring mechanism and then incorporate position encoding and Q into the decoder attention layer. Finally, we prepend a CLS token at the start of the input sequence as a representation of the entire input sequence. The features that have just been precisely classified by the decoder layer are fused with the CLS token to generate the final CLS classification head. This CLS classification head is then utilized for the ultimate classification of the postures.



(a) Collection (b) Office (c) Library (d) Cafe

Fig. 12. Experimental scenarios.

3. Performance evaluation

In this section, we present a prototype implementation of our system and collect data from various scenarios for performance evaluation. We begin with the basic evaluation of the system, followed by a detailed analysis of its performance in different aspects, including the classification network, micro-benchmarks, and experimental conditions. Regarding the evaluation of the classification network, we report basic performance metrics and compare our network's accuracy with those of ResNets and ViT. We next evaluate two adjustments to classic methods that enhance system robustness. To evaluate the robustness of the system across different experimental conditions, we use the data from volunteer participants who were not included in the initial training data collection. Specifically, we test the classification accuracy of our system under same environmental conditions but unseen volunteers. Furthermore, we compare the energy variation of signals under static environment, standard posture, and standard posture with different dynamic interference. Finally, we evaluate the classification accuracy of the same volunteer at different distances from experimental device, and different experimental device heights.

3.1. System implementation

Our system prototype uses a UMA-8 circular microphone array to receive the echo at 48 kHz and a JBL PS2200 speaker as a signal generator. The microphone array and speaker are controlled by a laptop through a USB interface. Also, the code implementation of data collection and preprocessing are based on Matlab 2021a. As shown in Fig. 12(a), we set the microphone array panel facing to users and vertically to the desk, and the speaker is behind the microphone array for a tiny interval. The above layout is helpful for the synchronization of signals. After the echo signals are preprocessed, we train the proposed deep learning model on an NVIDIA GeForce GTX 3090 GPU with 24 GB of memory.

Considering the microphone array is located between the person and the screen, and the rapid ultrasound energy loss during transmission, we set the maximum sensing range as 50 cm from the microphone array. In addition, according to the constraint mentioned in Section 2-1 and relevant studies (Pynt, Higgs, & Mackey, 2001), we also set the minimum sensing range as 20 cm from the microphone array. Besides, to balance the calculation complexity and the resolution of the body profile, the range of height angle is set from $\pi/36$ to $\pi/2$ with $\pi/36$ as the step length, and the azimuth range is set from $\pi/18$ to 2π with $\pi/18$ as the step length. Therefore, the preprocessing result obtained from a single echo signal will be a 18 channels matrix with 36×18 as its size.

We also recruited four volunteers to participate in testing our system. For the proposed network training, we collected data from only one volunteer. Additionally, all volunteers were instructed to wear shirts during the basic experiments. To ensure relevance to daily life, we collected raw sound data from three different environments, as shown in Fig. 12, for evaluation purposes. Moreover, we categorized sitting postures into five main types: standard sitting posture, left-leaning, right-leaning, humpback, and back-leaning, based on existing research findings (Villanueva et al., 1997) and common usage scenarios. During the data collection process, we used one volunteer for training data and the remaining three volunteers for test data, in a singular pose within a specific usage scenario, we gather and curate a dataset of 2000 samples for the purpose of model training, additionally, an extra set of 800 samples is allocated exclusively for model testing. The volunteers were positioned approximately 50 cm from the screen, resulting in a horizontal distance of about 35±3cm from the microphone array. Volunteers were instructed to sit in random positions within this distance range. Although the sensing height angle range was set as $\pi/36$ to $\pi/2$, due to the directive nature of ultrasounds, we asked volunteers to sit within a height angle range from $\pi/3$ to $\pi/2$ of the microphone array panel.

3.2. Basic performance

To show the basic performance of our system, we firstly evaluate the general accuracy of APC which means that our training data and testing data come from the same volunteer but at different times and in different scenarios. As shown in Fig. 13, we can obtain a average confusion matrix tested in three experimental scenarios shown in Fig. 12. The results reveal an impressive classification accuracy of 98.41% for our system.

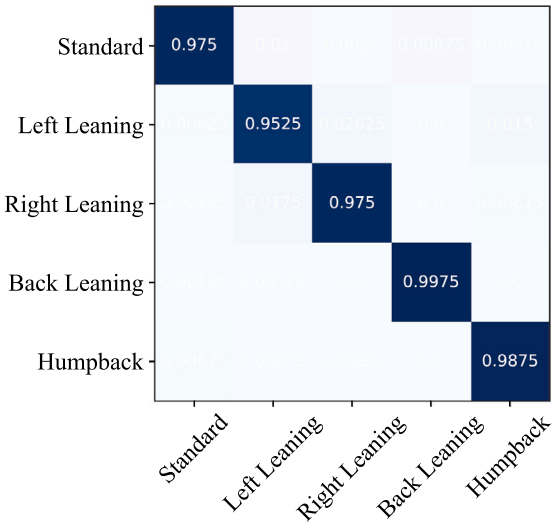


Fig. 13. Basic performance.

Table 1
Comparisons of different methods.

Classes	Recall	Accuracy	F1-score	Parameter (10^7)
ResNets18	0.8012	0.8455	0.8227	1.170
Transformer	0.8322	0.8834	0.8570	8.524
MLP	0.7253	0.7583	0.7414	0.672
LSTM	0.7601	0.8013	0.7801	4.323
ResNets101	0.8521	0.8697	0.8608	4.460
Our model	0.9840	0.9841	0.9840	6.741

Next, we introduce initial parameters, training strategies, and evaluation method of our neural network before the evaluation. Our model training strategies are as follows. Firstly, we conduct a warm-up training phase for 10 epochs at a reduced learning rate of 0.0001. Subsequently, we restore the learning rate to 0.001 and train for additional 500 epochs, and the strategy for learning rate reduction that we adopt in this session involves a linear decrease in the learning rate by a factor of 0.5 every 50 epochs, and our training batch size is set at 256. Furthermore, the neural networks are trained to maximize the square root of the F1-score (Balanced F Score), and the accuracy coefficient is utilized to evaluate our models and optimize the hyperparameters. In this context, the models are evaluated using accuracy, precision, recall, and F1-score.

As a result, our training dataset comprises 10,000 matrices each of dimension $18 \times 36 \times 18$, and the total number of validation sets is 4,000. Through the aforementioned model training strategy, we ultimately achieve the highest accuracy at the 310th epoch, which stands at 0.984, and the F1 score on the validation set is equally high at 0.984. Moreover, the loss value significantly drops from 1.45 to 0.09. The accuracy, recall, and F1 scores of the validation set have also markedly increased from 0.34, 0.34, and 0.24 to above 0.98, respectively.

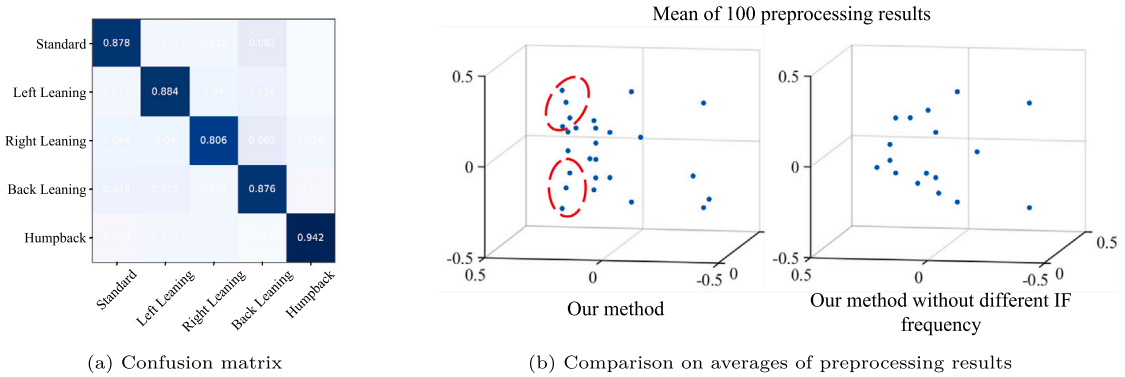
Given that our model is an ensemble model, we also conduct both model comparison experiments and ablation experiments. First, employing the same training strategy on shallow ResNets, the conventional Transformer, and MLP, we achieved the highest accuracies of 0.84, 0.88, and 0.75, respectively. After separately removing the shallow ResNets and Transformer architectures from our model, the highest accuracies stand at 0.89 and 0.91. Additionally, we compared our model with deep ResNets (ResNets101) and LSTM (Tan, Santos, Xiang, & Zhou, 2015). Under the same strategy, the highest accuracies for deep ResNets and LSTM are 0.86 and 0.80, respectively. These results underscore the superior performance of our proposed model Tables 1 and 2.

3.3. Evaluation of system performance with an increased number of classes

Recognizing that a mere five posture classes might not adequately address the requirements of daily usage scenarios, we have expanded the posture categories to eight. Alongside the original five classes, we have introduced three additional posture classes: “Leaning on Table” (LT), “Left Leaning on Table” (LTl), and “Right Leaning on Table” (LTr). For training purposes within an office scenario, we have meticulously collected 2000 fresh samples for each posture class. An additional 200 samples have been reserved for evaluation. The results, reveal an average accuracy of 92.25%. Notably, we observe a decline in the accuracy of classifying the three “on table” postures, with a reduction of approximately 10% when contrasted with other postures. This accuracy drop can be attributed to a common characteristic of the three “on table” postures. These postures involve the user’s hair facing the device

Table 2
Evaluation metrics on test set.

Classes	Precision	Recall	F1-score	Support
Standard	0.9848	0.9700	0.9773	1000
Left Leaning	0.9700	0.9700	0.9700	1000
Right Leaning	0.9657	0.9850	0.9752	1000
Back Leaning	1.0000	1.0000	1.0000	1000
Humpback	1.0000	0.9950	0.9975	1000
Accuracy			0.9840	5000
Macro Avg	0.9841	0.9840	0.9840	5000
Weighted Avg	0.9841	0.9840	0.9840	5000



(a) Confusion matrix (b) Comparison on averages of preprocessing results

Fig. 14. Classification performance of data generated by using the same IF frequency on the detection of single spatial location.

directly. Due to the fluffy texture of hair, it acts as a proficient sound-absorbing material. Nonetheless, our system can still achieve satisfactory recognition by capitalizing on reflections from other body parts, compensating for the challenges posed by these specific postures.

3.4. Adjustments on classic methods

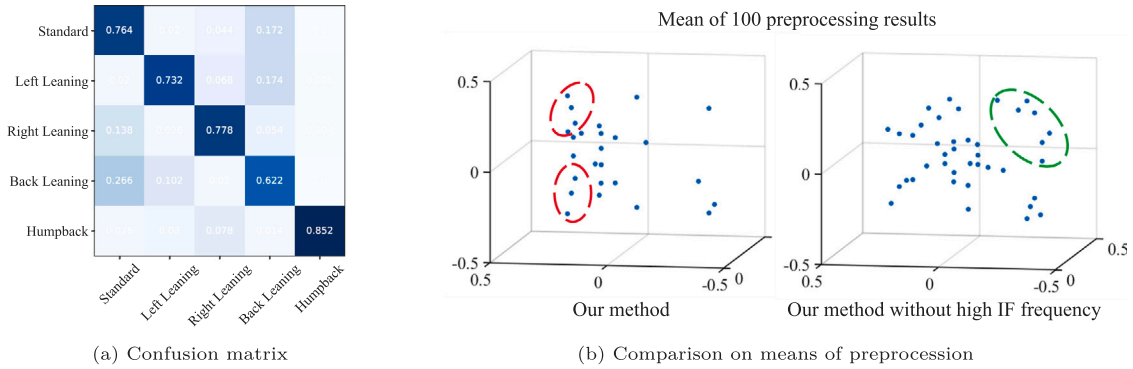
Impact of different IF frequencies: In contrast to the conventional microphone array method, we have adopted a novel approach by assigning different IF frequencies to IF signals on the detection of single spatial location. To evaluate the effectiveness of this scheme, we compared the accuracy of our system with the conventional signal processing method. As shown in Fig. 14(a), we achieved 87.72% as the best accuracy. Also, we compared the average of 100 preprocessing results generated using our method and the unimproved method from the right leaning posture in Fig. 14(b). and the display of the results is the same as in Fig. 9. It is evident that the data generated by the unimproved method lacks information from minor reflectors such as the head or arms.

Impact of high base frequency IF signal: In this part, we evaluate the high IF frequency scheme in the FMCW method, where we compare the accuracy between the system with the conventional FMCW method and our system. As shown in Fig. 15(a), the average accuracy significantly decreases to 74.96%. We also compared the average of 100 preprocessing results generated using our method and the unimproved method from the right leaning posture in Fig. 15(b), and the display of the results is the same as in Fig. 9, due to the long wavelength of IF signals. Preprocessing results from different distances exhibit varying amounts of noise points.

Performance of space domain beamforming: In our approach, we adopt a fundamental concept that involves a fusion of space domain beamforming and time domain beamforming. This choice is made to enhance the inherent sensitivity of wave path differentiation, rather than solely relying on space domain beamforming. The results depicted in Fig. 16(a) substantiate this approach. Comparing the accuracy performance, space domain beamforming exhibits a significantly lower average accuracy of 54.4% when contrasted with our approach. Furthermore, as indicated in Fig. 16(b), the preprocessing results for a right-leaning posture underscore a substantial issue of ambiguity, in contrast to our approach, discerning the underlying pattern of the preprocessing results is notably challenging.

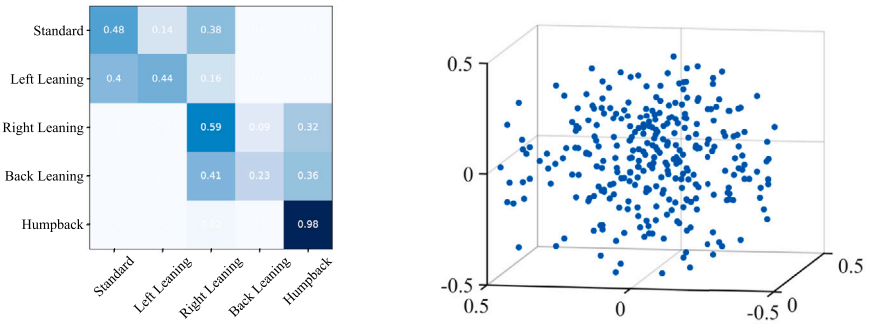
3.5. Experimental conditions

Impact of new users: We collect data from users who did not attend the training process to test the system's accuracy, where we invite three new volunteers for the evaluation. As shown in Fig. 17, the accuracy is 67.9% on average for 5 sitting postures without any extra training. To further evaluate the performance of the network under the condition of training with a small amount of data from new users, we conducted experiments using different sample sizes. Specifically, we set the sample sizes as 10, 20, 30,



(a) Confusion matrix (b) Comparison on means of preprocession

Fig. 15. Classification performance of data generated by using the IF frequency starting from 0 Hz.



(a) Confusion matrix (b) Preprocession result

Fig. 16. Performance of space domain beamforming.

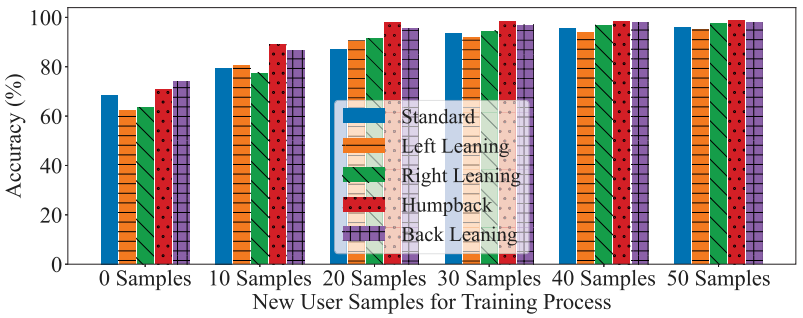


Fig. 17. Accuracy over different numbers of new user samples.

40, and 50 for the training process. Fig. 17 illustrates the accuracy trends observed for distinct convergence patterns as the sample size of new users' data increases. Notably, when the sample size exceeds 30, the accuracy reaches its highest level. The average accuracy achieved in this case is 95.1%. These findings demonstrate that our network can be quickly implemented with acceptable performance after a simple sampling procedure for new users.

Impact of different sensing distances: Because we collect data for training on microphone at about 35 cm away from users body, and the recommended optimal sensing distance is 35 cm away from the microphone array. However, our system still exhibits tolerance for slight deviations from this distance. To demonstrate this, we conducted tests using data sampled at different sensing distances. A volunteer was instructed to assume various postures while maintaining distances of 30 cm, 32.5 cm, 35 cm, 37.5 cm, and 40 cm from the microphone array. As depicted in Fig. 18, our system maintains a stable accuracy even with a small deviation in the sensing distance, approximately 5 cm, and it aligns with the expectation that the accuracy will be better when the sensing distance is closer to 35 cm.

Impact of dynamic interference: The static interference is subtracted from the received signals, but dynamic interference by human walking still exists. Therefore, we need to know how dynamic interference impacts the system, and we compare between

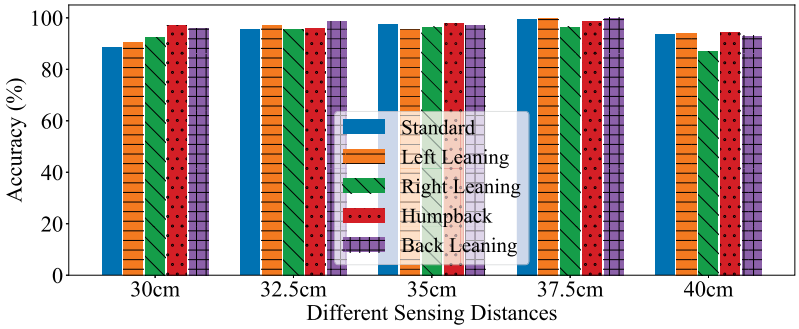


Fig. 18. Accuracy over different sensing distances.

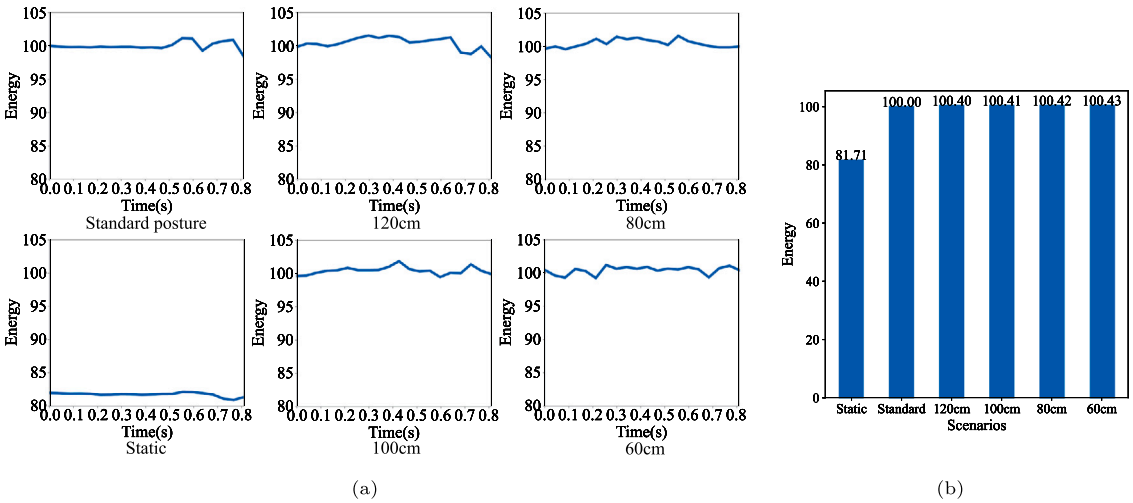


Fig. 19. Normalized signal energy under dynamic interference: (a) shows the normalized energy variation of echos from static experimental scenario, standard posture without dynamic interference, and dynamic interference at different distances. (b) shows the means of normalized energy.

echo energies under different dynamic interference. As shown in Fig. 19, we record echoes and arrange a volunteer walking for 1 m in 0.8s at different distances. As a result, the contribution to total energy of a walking people is minor. We can make a explanation from two aspects. First, the user is very close to the microphone array, thus occupying most of its perception angle, making the reflector behind the person undetectable. Second, because the reflected lobe of the sound is opposite to the normal of the reflector in the direction of sound incidence, dynamic interference in the low altitude angle direction cannot be detected.

Impact of different environments: Our APC system maintains a stable accuracy in different environments, due to its static interference cancellation scheme for near interference and the rapid power attenuation of far interference. To assess the system performance across different environments, we collected data from a single user in various usage scenarios, including an office, a library, and a cafe, as depicted in Fig. 12. The office environment features static interference caused by a crowd but has minimal dynamic interference. In the library, dynamic interference arises from people walking around. The cafe environment, on the other hand, exhibits rare interference due to its clean layout. Fig. 20 presents the accuracy results obtained from these different scenarios. We can see that there is not a large difference in accuracy among the various scenarios. This validates the robustness of our system across different environments.

Impact of different clothes textures: The clothing textures may influence the echo energy. Clothing with rough textures are more likely to absorb more echo energy to varying degrees, but most casual clothes textures do not change the diffusion direction of the sound. Because of the reasons above, it is obvious that different clothes textures have an impact on our system especially when the surface of clothing is very coarse. Therefore, we collect the testing data from a single user but wearing different clothes to evaluate our system. As shown in Fig. 21, thin clothes with neat and smooth surfaces like silk can obtain a good accuracy of 97.5% on average, but thick clothes with coarse surface only obtain a poor accuracy of 41.25% on average. However, because our usage scenarios are mainly indoor environments, it would not be a serious problem for our system.

Impact of different vertical height of experimental device: Due to the variable intensity of body reflection and diffusion at different heights, the feature and echo quality may vary depending on the position of experimental device. In previous experiments we set the experimental device at a distance of 45 cm from the top of a user to the desktop. To evaluate the impact on the accuracy

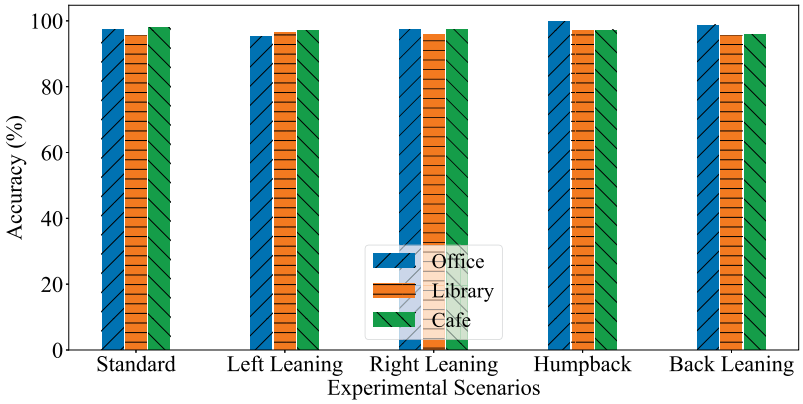


Fig. 20. Accuracy over different experimental environments.

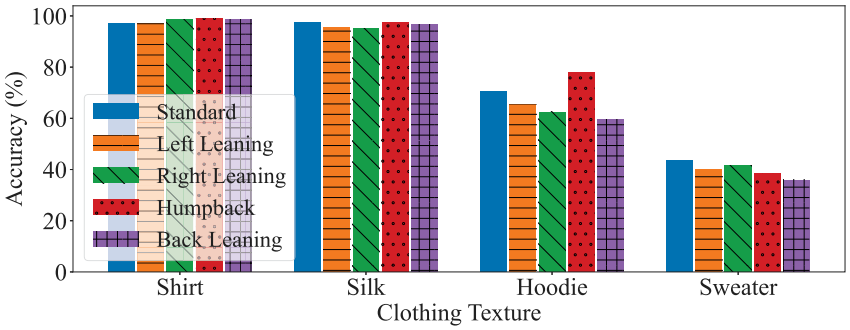


Fig. 21. Accuracy over different clothes textures.

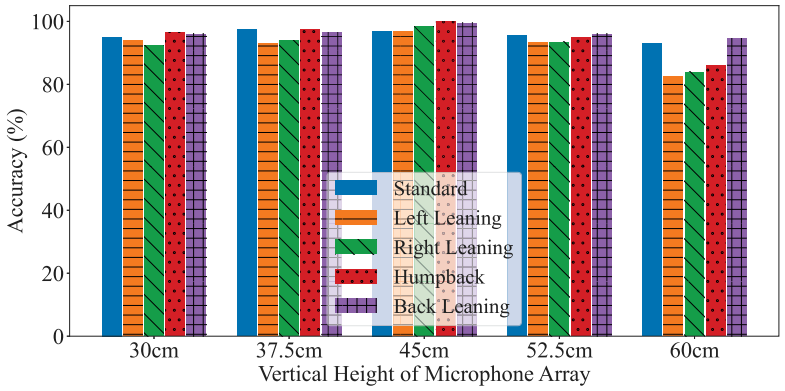


Fig. 22. Experiments under different vertical heights of microphone array.

of different height positions of experimental device, we utilize experimental device to collect data from different height positions to the upper human body. As shown in Fig. 22, the system obtains the best performance at 45 cm height where is the same height of chest, and the system is stable during the height from 30 cm to 52.5 cm. This observation provides a clear evidence that our system maintains a stable performance even when there are small changes in the device height.

Impact of different deflection angle of experimental device: In previous experimental setups, the device was positioned directly in front of the user. However, in real-world scenarios, the device's placement can significantly impact the user experience. To assess the system's resilience to deflection angles deviating from the user's front direction, without necessitating extra training, we gathered data from varying deflection angles for evaluation purposes. As illustrated in Fig. 23, we have 98.41% and 94.5% as the performance on accuracy for 0 and 5 device deflection degree respectively, our system also maintains an acceptable performance on accuracy within a deflection angle from 10 to 25 degrees which is 83%. Beyond this threshold, the accuracy of the system

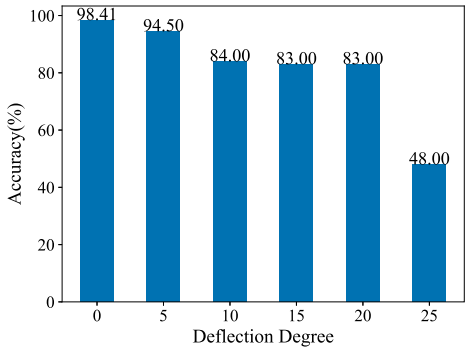


Fig. 23. Experiments under different deflection angle of experimental device.

experiences a sharp decline to 48%. This decrease in accuracy can be attributed to two main factors. Firstly, data stemming from larger deflection angles tend to perplex the classification model. Secondly, when the device deviates from its original position, a substantial portion of the reflection direction and diffusion lobe orientation shifts inversely to the device's deflection.

4. Conclusion

In this paper, we proposed APC, an acoustic sitting posture recognition system. Specifically, we formulated the received signals on a circular microphone array and designed a signal preprocessing method based on a modified FMCW method. Then, we preprocessed signals to obtain feature matrices based on time domain beamforming. Besides, we constructed a new classification network based on ViT for sitting posture classification. Finally, we evaluated this system in many aspects which include different classification methods, different modifications of conventional methods, and different experimental conditions. The experimental results demonstrated our APC system can obtain a great performance with 98.41% classification accuracy.

CRediT authorship contribution statement

Kaiyuan Ma: Conceptualization, Data curation, Formal analysis, Software, Writing – original draft. **Shunan Song:** Software, Validation. **Lingling An:** Supervision. **Shiwen Mao:** Writing – review & editing. **Xuyu Wang:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Cai, C., Zheng, R., & Luo, J. (2022). Ubiquitous acoustic sensing on commodity IoT devices: A survey. *IEEE Communications Surveys & Tutorials*, *24*(1), 432–454.
channal, B. H. (2015). Posture. URL https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/posture.
Chara, A., Zhao, T., Wang, X., & Mao, S. (2023). Respiratory biofeedback using acoustic sensing with smartphones. *Smart Health*, *28*, Article 100387.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
Farnan, M., Dolezalek, E., & Min, C.-H. (2021). Magnet integrated shirt for upper body posture detection using wearable magnetic sensors. In *2021 IEEE international IoT, electronics and mechatronics conference* (pp. 1–5). IEEE.
Gao, Y., Jin, Y., Choi, S., Li, J., Pan, J., Shu, L., et al. (2021). SonicFace: Tracking facial expressions using a commodity microphone array. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *5*(4), 1–33.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
Health, H. (2001). 3 surprising risks of poor posture. https://www.health.harvard.edu/staying-healthy/3-surprising-risks-of-poor-posture.
Ling, K., Dai, H., Liu, Y., Liu, A. X., Wang, W., & Gu, Q. (2020). Ultragesture: Fine-grained gesture sensing and recognition. *IEEE Transactions on Mobile Computing*, *21*(7), 2620–2636.
Maereg, A., Lou, Y., Secco, E. L., & King, R. (2019). Hand gesture recognition based on near-infrared sensing wristband.
Mao, W., Wang, M., Sun, W., Qiu, L., Pradhan, S., & Chen, Y.-C. (2019). RNN-based room scale hand motion tracking. In *The 25th annual international conference on mobile computing and networking* (pp. 1–16).
Pynt, J., Higgs, J., & Mackey, M. (2001). Seeking the optimal posture of the seated lumbar spine. *Physiotherapy Theory and Practice*, *17*(1), 5–21.

Shan, Y., Liao, P., Wang, X., An, L., & Mao, S. (2023). MAA: Modulation-adaptive acoustic gesture recognition. In *2023 IEEE 20th international conference on mobile ad hoc and smart systems* (pp. 62–70). IEEE.

Shtrepi, L., Astolfi, A., Puglisi, G. E., & Masoero, M. C. (2017). Effects of the distance from a diffusive surface on the objective and perceptual evaluation of the sound field in a small simulated variable-acoustics hall. *Applied Sciences*, *7*(3), 224.

Sun, H., Zhu, G.-a., Cui, X., & Wang, J.-X. (2021). Kinect-based intelligent monitoring and warning of students' sitting posture. In *2021 6th international conference on automation, control and robotics engineering* (pp. 338–342). IEEE.

Tan, M., Santos, C. d., Xiang, B., & Zhou, B. (2015). LSTM-based deep learning models for non-factoid answer selection. arXiv preprint arXiv:1511.04108.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems*: *vol. 30*.

Villanueva, M. B. G., Jonai, H., Sotoyama, M., Hisanaga, N., Takeuchi, Y., & Saito, S. (1997). Sitting posture and neck and shoulder muscle activities at different screen height settings of the visual display terminal. *Industrial Health*, *35*(3), 330–336.

Wan, H., Shi, S., Cao, W., Wang, W., & Chen, G. (2021). Resptracker: multi-user room-scale respiration tracking with commercial acoustic devices. In *IEEE INFOCOm 2021-IEEE conference on computer communications* (pp. 1–10). IEEE.

Wang, X., Huang, R., & Mao, S. (2017). SonarBeat: Sonar phase for breathing beat monitoring with smartphones. In *2017 26th international conference on computer communication and networks* (pp. 1–8). IEEE.

Wang, X., Huang, R., Yang, C., & Mao, S. (2021). Smartphone sonar-based contact-free respiration rate monitoring. *ACM Transactions on Computing for Healthcare*, *2*(2), 1–26.

Wang, R., Liu, C., Mou, X., Gao, K., Guo, X., Liu, P., et al. (2023). Deep contrastive one-class time series anomaly detection. In *Proceedings of the 2023 SIAM international conference on data mining* (pp. 694–702). SIAM.

Wang, X., & Mao, S. (2022). Acoustic-based vital signs monitoring. In *Contactless vital signs monitoring* (pp. 281–301). Elsevier.