THE UNIVERSITY
of EDINBURGH

Intelligent Ultrasound Hand Gesture Recognition System

Yinghao Li

A thesis submitted for the degree of Doctor of Philosophy.

The University of Edinburgh.

June 2023

Declaration

I hereby declare that the work presented in this thesis is my own unless otherwise acknowledged. This thesis has not been submitted for any other degree or professional qualification except to The University of Edinburgh for the degree of Doctor of Philosophy.

Yinghao Li

Date: 30/06/2023

Table of Contents

Acknowledgement

First and foremost, I wish to convey my profound gratitude to my primary supervisor, Dr. Jiabin Jia, whose invaluable guidance and unwavering support have been pivotal throughout my PhD journey. Over the course of the past three and a half years, Dr. Jia has embodied the roles of a mentor, confidant, and friend, consistently offering guidance and support. His fervor for research, dedication to academia, and ceaseless pursuit of knowledge have deeply influenced my academic trajectory. The strides I've made in this journey would have been unattainable without his patience, wisdom, and valuable counsel. His academic brilliance and unique personal charm have made our collaboration throughout this challenging doctoral period immensely rewarding.

Moreover, I am grateful to my current and former colleagues in the Agile Tomography Group for their academic assistance and companionship. My heartfelt thanks go out to Dr. Yuan Chen, Dr. Hancong Wu, Dr. Yong Bao, Dr.Zhixi Zhang, Wei Han, Hao Yu, Zhe liu, Dr. Zhou chen, Rui Zhang, Jialei Fu, Jiangnan Xia. Collaborating with these individuals within an inspiring academic environment has immensely enriched my academic journey.

Lastly, and most importantly, my deepest gratitude goes to my family. It is challenging to articulate my immense fortune of having their unwavering support while I pursued my academic dreams in the beautiful city of Edinburgh. Their selfless love and support have been a source of strength during the difficult times and have enabled me to remain focused on my research. Their faith in me has not only facilitated my pursuit of academic dreams but also helped me discover my worth.

Lay summary

Ultrasound gesture recognition system is an existing development in the field of human computer interaction that allows devices to identify and interpret human movement. This technique utilized ultrasound waves to recognize users' input based on unique gestures, creating a new mode of interaction with digital environment. Compared with other techniques, ultrasound remote sensing has the advantage of low cost, illumination irrelevant and easy deploy. Therefore, it has been widely applied in various areas, such as rear parking senser, ultrasound room rangefinder and underwater object detection.

This thesis aims to verify the feasibility of ultrasound gesture recognition under airborne conditions. Airborne gesture recognition can be deployed on smartphone devices, laptop devices and specific designed sensor. One of the key challenges of gesture recognition is the lack of accurate feature extraction methods, robust classification methods and real-time latency problem. Focusing on those problems, this thesis demonstrates a comprehensive investigation process of achieving gesture recognition system based on multiple smart devices system. The illustration process includes an acoustic simulation of smartphone, a gesture recognition system based on smartphone devices, a fast gesture tracking system and a angle-velocity based gesture recognition system.

Abstract

With the booming development of technology, hand gesture recognition has become a hotspot in Human-Computer Interaction (HCI) systems. Ultrasound hand gesture recognition is an innovative method that has attracted ample interest due to its strong real-time performance, low cost, large field of view, and illumination independence. Well-investigated HCI applications include external digital pens, game controllers on smart mobile devices, and web browser control on laptops. This thesis probes gesture recognition systems on multiple platforms to study the behavior of system performance with various gesture features. Focused on this topic, the contributions of this thesis can be summarized from the perspectives of smartphone acoustic field and hand model simulation, real-time gesture recognition on smart devices with speed categorization algorithm, fast reaction gesture recognition based on temporal neural networks, and angle of arrival-based gesture recognition system.

Firstly, a novel pressure-acoustic simulation model is developed to examine its potential for use in acoustic gesture recognition. The simulation model is creating a new system for acoustic verification, which uses simulations mimicking real-world sound elements to replicate a sound pressure environment as authentically as possible. This system is fine-tuned through sensitivity tests within the simulation and validate with real-world measurements. Following this, the study constructs novel simulations for acoustic applications, informed by the verified acoustic field distribution, to assess their effectiveness in specific devices. Furthermore, a simulation focused on understanding the effects of the placement of sound devices and hand-reflected sound waves is properly designed. Moreover, a feasibility test on phase control modification is conducted, revealing the practical applications and boundaries of this model.

Mobility and system accuracy are two significant factors that determine gesture recognition performance. As smartphones have high-quality acoustic devices for developing gesture recognition, to achieve a portable gesture recognition system with high accuracy, novel algorithms were developed to distinguish gestures using

smartphone built-in speakers and microphones. The proposed system adopts Short-Time-Fourier-Transform (STFT) and machine learning to capture hand movement and determine gestures by the pretrained neural network. To differentiate gesture speeds, a specific neural network was designed and set as part of the classification algorithm. The final accuracy rate achieves 96% among nine gestures and three speed levels. The proposed algorithms were evaluated comparatively through algorithm comparison, and the accuracy outperformed state-of-the-art systems.

Furthermore, a fast reaction gesture recognition based on temporal neural networks was designed. Traditional ultrasound gesture recognition adopts convolutional neural networks that have flaws in terms of response time and discontinuous operation. Besides, overlap intervals in network processing cause cross-frame failures that greatly reduce system performance. To mitigate these problems, a novel fast reaction gesture recognition system that slices signals in short time intervals was designed. The proposed system adopted a novel convolutional recurrent neural network (CRNN) that calculates gesture features in a short time and combines features over time. The results showed the reaction time significantly reduced from 1s to 0.2s, and accuracy improved to 100% for six gestures.

Lastly, an acoustic sensor array was built to investigate the angle information of performed gestures. The direction of a gesture is a significant feature for gesture classification, which enables the same gesture in different directions to represent different actions. Previous studies mainly focused on types of gestures and analyzing approaches (e.g., Doppler Effect and channel impulse response, etc.), while the direction of gestures was not extensively studied. An acoustic gesture recognition system based on both speed information and gesture direction was developed. The system achieved 94.9% accuracy among ten different gestures from two directions. The proposed system was evaluated comparatively through numerical neural network structures, and the results confirmed that incorporating additional angle information improved the system's performance.

In summary, the work presented in this thesis validates the feasibility of recognizing hand gestures using remote ultrasonic sensing across multiple platforms. The acoustic simulation explores the smartphone acoustic field distribution and response results in the context of hand gesture recognition applications. The smartphone gesture recognition system demonstrates the accuracy of recognition through ultrasound signals and conducts an analysis of classification speed. The fast reaction system proposes a more optimized solution to address the cross-frame issue using temporal neural networks, reducing the response latency to 0.2s. The speed and angle-based system provides an additional feature for gesture recognition. The established work will accelerate the development of intelligent hand gesture recognition, enrich the available gesture features, and contribute to further research in various gestures and application scenarios.

Abbreviations

RGB - Red, Greed, Blue

LSTM - Long Short-Term Memory

CNN - Convolutional Neural Network

MUSIC - Multiple Signal Classification algorithm

STFT - Short Time Fourier Transfer

SVM - Support Vector Machines

SPL - Sound Pressure Level

DTFT - Discrete-Time Fourier Transform

PWM - Pulse Width Modulation

EMD - Empirical Mode Decomposition.

1D - CNN - One dimension- Convolutional Neural Network

T2B - Gesture Top To Bottom

B2T - Gesture Bottom To Top

BW - Backward

FW - Gesture Forward

SC - Gesture Single Click

DC - Gesture Double Click

ZOOMIN - Finger Gesture Zoom in

ZOOMOUT - Finger Gesture Zoom out

DAQ – Data Acquisition

RNN - Recurrent Neural Network

CRNN – Convolutional-Recurrent Neural Network

BiLSTM - Bidirectional Long Short-Term Memory

RGU - Recurrent Unit

ConvLSTM - Convolutional Long Short-Term Memory

ADC - Analog-To-Digital Converter

Chapter 1 Introduction

1.1 Background and motivation

Gesture is a meaningful body movement with the intent of transforming information and interacting with the environment. As one of the most crucial subjects in human-computer interaction, hand gestures recognition has attracted huge attention from researchers.

So far, most of the relevant research in the gesture recognition area has focused on optical methods, specifically those acquired by RGB cameras [1]–[6] or depth cameras [7], [8], such as infrared cameras [9]–[11]. These optical methods offer high accuracy and versatility, making them popular choices for gesture recognition. However, they have limitations when it comes to hostile environments, such as poor illuminating conditions or lack of external power supply, where their effectiveness diminishes. For example, in a dark room or under direct sunlight, the detection accuracy of optical methods significantly decreases [12].

To address the challenges posed by illumination, alternative approaches have been explored, such as gesture recognition using electrical gloves [13][14] or other types of external devices. These methods aim to mitigate the impact of illumination on gesture recognition accuracy. However, the use of wearable devices introduces certain limitations, including the necessity for users to wear specific equipment, which can restrict their flexibility and increase energy consumption.

In contrast, ultrasound-based gesture recognition offers a promising solution that overcomes the limitations of optical methods in hostile environments. Ultrasound technology operates independently of ambient lighting conditions, making it suitable for low-light or even no-light environments. Additionally, ultrasound sensors can be battery-powered, eliminating the dependence on external power supply.

The use of ultrasound-based gesture recognition eliminates the need for wearable devices and provides users with more freedom and flexibility in their interactions. It

enables gesture recognition in diverse scenarios, including those with poor lighting or where the use of optical devices is impractical. Moreover, ultrasound-based systems offer enhanced privacy as they do not capture visual information, addressing concerns related to surveillance and data security.

The growing interest in ultrasound-based gesture recognition is evident from the increasing number of research studies exploring its potential. These studies focus on improving signal processing algorithms, developing robust feature extraction techniques, and leveraging machine learning and deep learning approaches for accurate gesture recognition.

This thesis primarily focuses on the implementation of ultrasound-based gesture recognition systems on smart devices, specifically targeting smartphones. The research encompasses various aspects, including smartphone device acoustic simulation, ultrasound gesture recognition on smartphones, ultrasound gesture recognition using specific sensors, and the exploration of angle information for gesture recognition.

Feature extraction plays a crucial role in gesture recognition, as it significantly influences the accuracy of gesture classification. Traditional methods such as thresholding detection and distance-dependent feature calculations suffer from drawbacks such as poor environmental robustness, limited adaptability to multiple users, and variations in gesture performance. These limitations pose critical challenges that are not easily overcome.

To address these challenges, this thesis proposes the utilization of machine learning techniques. Compared to the aforementioned traditional approaches, machine learning offers several advantages. It exhibits strong adaptability to different environmental conditions, enabling reliable performance even in complex or challenging settings. Machine learning techniques excel in feature extraction, allowing for the identification and extraction of meaningful features from ultrasound signals, resulting in robust and accurate gesture recognition. Additionally, machine learning approaches offer high resolution, facilitating the capture of intricate details and subtle nuances in hand movements.

1.2 Main Contribution

The main contribution is summarized as follows:

1) A novel pressure-acoustics simulation model that verifying the feasibility of potential application (acoustic gesture recognition) was designed and discussed. The system firstly developed a novel acoustic verification system that adopting simulation of true-to-life sound original components to reproduce a realistic sound pressure environment as closely as possible. The system was calibrated by introducing sensitivity test into simulation and was verified by practical measurement result. Secondly, novel acoustic application simulations based on the verified acoustic field distribution was built to explore the performance for target devices. The system verifies the influence of sound devices' distribution and the reflection wave from hand. Lastly, a feasibility test of phase control modification was carried out and demonstrated the applicable scope and limitations.

2) An ultrasound gesture recognition system based on smartphone devices with speed level classification system was designed. The system adopted a novel short machine learning algorithm with Short Time Fourier Transform and achieved a 95.56% classification accuracy for 9 novel gestures. Two speed classification algorithms were proposed and compared, and the system achieves a 95.6% classification accuracy for 9 gestures with 9 types of gestures with three levels of speed each. The system was designed on a common smartphone device and operated in real time.

3) A fast reaction real time ultrasound gesture recognition system was developed based on sensors and data acquisition devices. This system adopted a novel temporal based convolutional neural network as the classification method and achieved a fast reaction time of 0.2s. The system achieved 100% accuracy over 6 gestures and performed a strong robustness under an interference test. A multi-user system sharing the same neural network was designed and a comprehensive user interface was designed.

4) A multi-feature gesture recognition system was developed to investigate the feasibility of detecting gestures from different directions. A gesture recognition system

adopting novel multi-feature input neural network was developed. This system employed both velocity and the direction of the performed gesture as the input and achieved a 94.9% accuracy over 10 gestures from 2 different directions. The proposed method shows a 9.5% (velocity only) and 19.5% (direction only) improvement over the results of the single-feature approach.

All those works have made step progress in promoting ultrasound gesture recognition as an effective real-time human machine interaction system.

1.3 Overview of the thesis

The thesis is composed of eight chapters and the remaining part is structured as follows.

Chapter 2 presents a brief review of the mathematical theory of ultrasound detection, the development of gesture recognition in various approaches, the existing ultrasound gesture sensing approaches, and the state of the art in ultrasound gesture recognition systems.

Chapter 3 proposes a sound field finite element simulation of two off-the-shelf devices. It is the first time that the smartphone was simulated as a combined entity for conducting gesture recognition application. This simulation included a comprehensive calibration, encompassing both sensor factory sensitivity tests and real model calibrations. Three applications scenarios were simulated to investigate the acoustic performance of conducting a gesture recognition and finally a phase control acoustic phase experiment was carried out.

Chapter 4 proposed a real time ultrasound gesture recognition system based on smartphones. This system adopted Short Time Fourier Transform and Convolutional neural network as the feature extractor and classifier. 9 gestures were designed and classified as a trigger to interact with the smartphone. Two speed evaluation methods were proposed and compared, and three speed levels were designed for each gesture.

Chapter 5 developed a fast reaction gesture recognition system based on data acquisition platform. A novel LSTM-CNN based neural network was designed to

classify the gesture features into gesture type. Two normalization techniques were adopted to reduce the effect of background noise. A multiuser function was conducted, and user interfaces were developed based on computer with mouse and keyboard control.

Chapter 6 conducted an acoustic gesture recognition system using angle information and velocity to improve accuracy. A new sensor array and data collection framework were developed, alongside a feature extraction system. A novel 1D-convolutional neural network classifier was introduced. The multi-feature system outperformed traditional single-feature networks, boasting a 94.9% recognition accuracy.

Lastly, Chapter 7 summarizes the research contributions of this thesis and discusses possible future work. This future work is especially focused on high frequency ultrasound sensing of human gestures and human activity detection, using the developed ultrasound sensor and the whole recognition system from this study.

Chapter 2 Literature review

2.1 Introduction

Following the background introduction, this chapter provides a brief review of the currently study current studies in the area of ultrasound gesture recognition area. This section will be organized in the following order: 1. A background literature review in gesture recognition and acoustic gesture recognition aspect. 2. The disadvantages and shortcomings of existing methodology and areas for improvement and the methodologies. 3. A brief introduction to existing algorithm.

2.2 Review of Ultrasound-Based Gesture Recognition System

Gesture, a form of meaningful body movement, is intended to facilitate the exchange of information between individuals and assists in interacting with the environment. A diverse range of sensing methods has been explored for gesture recognition, including optical, electromagnetic, and acoustic approaches. The optical method, particularly using RGB and infrared cameras, has achieved remarkable success over the past decade. This approach is favored for its high recognition accuracy and has become the predominant system in the field. However, it is not without drawbacks, as it can be negatively impacted by hostile environmental backgrounds, such as inadequate lighting or ambient infrared interference. For instance, gesture detection can be highly unreliable in dimly lit areas or under intense sunlight. Moreover, the continuous operation of these cameras raises privacy concerns among some users, wary of constant recording. Conversely, Ultrasonic sensing has gained popularity in recent years as an innovative method for gesture recognition. It boasts advantages like insensitivity to lighting conditions and a broader field of view. Ultrasonic sensors are typically cost-effective and widely available, with several projects successfully implementing built-in microphones and speakers on common electronic devices, including phones and laptops. A diverse range of sensing methods has been explored for gesture recognition,

including optical[12], [15], [16], electromagnetic[17]–[21], and acoustic approaches[22]–[26]. The optical method, particularly using RGB[1]–[6] and infrared cameras[9]–[11], [27], has achieved remarkable success over the past decade. This approach is favored for its high recognition accuracy and has become the predominant system in the field. However, it is not without drawbacks, as it can be negatively impacted by hostile environmental backgrounds, such as inadequate lighting[28], [29] or ambient infrared [25] interference. For instance, gesture detection can be highly unreliable in dimly lit areas or under intense sunlight[30]. Moreover, the continuous operation of these cameras raises privacy concerns among some users, wary of constant recording. Conversely, Ultrasonic sensing has gained popularity in recent years as an innovative method for gesture recognition. It boasts advantages like insensitivity to lighting conditions and a broader field of view. Ultrasonic sensors are typically cost-effective and widely available, with several projects successfully implementing built-in microphones and speakers on common electronic devices, including phones and laptops[22], [31].

Ultrasound-based gesture recognition has witnessed significant development in the academic domain over the years. Early research in this field focused on exploring the feasibility of using ultrasound technology for gesture recognition. One notable study by Jones et al. (2004) [32]demonstrated the potential of ultrasound sensors in capturing hand movements and recognizing basic gestures.

As the field advanced, researchers started investigating different signal processing and feature extraction techniques for ultrasound-based gesture recognition. In 2010, Liang et al.In 2010, Liang [33]et al. proposed a method that employed time-frequency analysis to extract relevant features from ultrasound signals, achieving improved gesture recognition accuracy.

Machine learning algorithms also found their way into ultrasound-based gesture recognition. In 2013, Nguyen et al.[34] et al. presented a gesture recognition system based on support vector machines (SVM), which demonstrated promising results in recognizing a wide range of hand gestures using ultrasound data.

With the advent of deep learning techniques, researchers began exploring their

application in ultrasound-based gesture recognition. In 2017, Zhang et al. [35]introduced a deep convolutional neural network (CNN) architecture specifically designed for ultrasound data analysis, achieving state-of-the-art performance in gesture recognition tasks.

However, the existing ultrasonic gesture recognition systems face significant challenges in real-world applications [36]. Firstly, there's a noticeable shortfall in research that conducting a simulation test to verify the system feasibility, leading to potential inadequacies in their performance in diverse settings. Secondly, the current systems lack the ability to adjust recognition accuracy based on the speed of the gestures, resulting in either missed or inaccurately interpreted gestures. Another major issue that is the current study's slow response to gestures [30], which usually requires at least 1 second to respond, hindering seamless interactions more seamless.. Lastly, the systems struggle with recognizing gestures from different angles, forcing users to adhere to restrictive and sometimes unnatural gesture execution. These limitations highlight the need for more robust development and research to make ultrasonic gesture recognition systems more reliable and user-friendly in everyday use.

Recent advancements in acoustic gesture recognition systems have shown promising results, but also highlight several areas for improvement. In 2012, a team from Microsoft developed a pioneering gesture recognition program utilizing a laptop's built-in acoustic devices [22]. They employed fast Fourier transform (FFT) to identify frequency changes and a threshold-based peak searching method, achieving a classification accuracy of 94% for five one-dimensional gestures. However, the system's reliance on laptop hardware could limit its adaptability to other devices.

Dolphin [37], introduced in 2014, reached a 93% recognition accuracy among 24 different gestures using FFT. The system classified gestures into groups and used a combination of manual recognition methods and machine learning. Although innovative, Dolphin's classification process could benefit from automated, more sophisticated algorithms to enhance its versatility and accuracy.

In 2016, W. Ruan et al. proposed AudioGest[38], a gesture sensing system based on smartphones. They introduced a novel feature extraction process that decoded the

time-frequency spectrum into a velocity vector using the Doppler effect, achieving 94.15% accuracy with six gestures. A notable drawback of AudioGest is its requirement for 3600 seconds of background signal to build an effective noise filter, which may not be feasible in dynamic real-world environments.

Y. Sang et al. [7] and Q. Zeng et al. [5] presented two projects using 300kHz ultrasound signals to recognize finger movements within a 20cm range. Employing the micro-Doppler effect, they achieved accuracies of 96.32% and 96.57%, respectively. However, these systems' need for high sampling rates (400kHz and 2MHz) and their limited operational range (0.15m and 0.175m) restrict their practical application.Y. Sang et al. [39] and Q. Zeng et al. [40] presented two projects using 300kHz ultrasound signals to recognize finger movements within a 20cm range. Employing the micro-Doppler effect, they achieved accuracies of 96.32% and 96.57%, respectively. However, these systems' need for high sampling rates (400kHz and 2MHz) and their limited operational range (0.15m and 0.175m) restrict their practical application.

These projects indicate significant progress in the field of acoustic gesture recognition. Nonetheless, they also reveal critical areas for improvement, such as the need for systems that can adapt to diverse hardware, operate efficiently in various environmental conditions, and offer broader recognition capabilities without sacrificing accuracy or response time.

To address those problem, this thesis provides a few solutions to address the key issues like slow response time and limited gesture angles. We developed better models and algorithms for smartphones, made gesture recognition faster and more accurate, and included the ability to recognize gestures from different directions. These advancements make gesture control more practical and user-friendly for various technologies and everyday use.

The following section will illustrate a comprehensive literature review regarding specific methodologies and the importance for each adopted method.

2.3 Review of Methodologies

2.3.1 Motion detection

A number of methods have been proposed for gesture motion detection including Time-Of-Arrival/ Time-Difference-Of-Arrival (TOA/TDOA)[25], [41], OFDM,[42], [43] coherent signal, and CIR[30], [31],[44]. However, in terms of ultrasound gesture recognition application, many exhibit limitations that impede their overall performance. For instance, traditional TOA and TDOA system necessitate the emission of burst-like acoustic signals, such as pulse or chirps. These signals can frequently be heard by humans as these signals shift suddenly. Moreover, their accuracy in measuring distance is typically on the scale of centimeters with the exception of the recent phase-based approach using OFDM. OFDM[45] is a digital transmission technique that splits a signal across multiple closely spaced frequencies to enhance data transmission efficiency. However, it has key limitations include sensitivity to frequency synchronization errors, vulnerability to multipath fading, and a high Peak-to-Average Power Ratio (PAPR) requiring complex RF amplifiers, which is not suitable for gesture recognition scenarios. Coherent signal, and CIR approach requires external processing time which is not suitable for gesture recognition application.

The principle utilized by our gesture recognition system to detect motion is to capture alteration in a sound wave's frequency due to the object, which is known as the Doppler effect. Comparing to the aforementioned method, doppler effect has the advantage of non-invasiveness, sensitivity to motion and velocity, simplicity and reliability. Unlike TOA/TDOA systems, which rely on burst-like signals that can be audible to humans, the Doppler effect in ultrasound systems is generally non-invasive and silent, making it more comfortable and less intrusive for users. The Doppler effect is inherently sensitive to motion and velocity changes, which can provide more dynamic and real-time data for gesture recognition compared to methods like CIR that are more static in nature. Compared to OFDM, which is sensitive to frequency synchronization errors and requires complex RF amplifiers, the Doppler effect can be utilized with

simpler hardware and is generally more robust to environmental factors. Hence, Doppler effect is For a brief introduction: Doppler effect is a phenomenon that signals frequency changes when the target has relevant movement regarding the source. The principle utilized by our gesture recognition system to detect motion is to capture alteration in a sound wave's frequency due to the object, which is known as the Doppler effect[46]. Comparing to the aforementioned method, doppler effect has the advantage of non-invasiveness, sensitivity to motion and velocity, simplicity and reliability. Unlike TOA/TDOA systems, which rely on burst-like signals that can be audible to humans, the Doppler effect in ultrasound systems is generally non-invasive and silent, making it more comfortable and less intrusive for users[31], [39], [40], [47], [48]. The Doppler effect is inherently sensitive to motion and velocity changes, which can provide more dynamic and real-time data for gesture recognition compared to methods like CIR that are more static in nature. Compared to OFDM, which is sensitive to frequency synchronization errors and requires complex RF amplifiers, the Doppler effect can be utilized with simpler hardware and is generally more robust to environmental factors. Hence, Doppler effect is utilized in this thesis. For a brief introduction: Doppler effect is a phenomenon that signals frequency changes when the target has relevant movement regarding the source.

$$f = \left(\frac{c}{c + v_s} \right) f_c \tag{1}$$

$$\nabla f = \left(\frac{v_s}{c} \right) f_c \tag{2}$$

where f and f_c represent the received and transmitted signal frequency, v_s is the target velocity, ∇f is the frequency change. A positive frequency change indicates a closer moving to the receiver, and vice versa.

The velocity of the target can be determined as:

$$v_s = \left(\frac{\nabla f \times c}{f_c} \right) \tag{3}$$

Our approach isn't the first to incorporate sound-based methods or the Doppler effect in detecting gestures and movements. As an example, Tarzia and colleagues

analyzed the strength of echo signals captured by a microphone to identify human movement and attention. Paradiso and team utilized a steady 2.4 GHz frequency signal to operate specialized patch antennas, interpreting the modified Doppler signal for understanding human movement and upper body dynamics in interactive areas. Kalgaonkar and associates more recently created a tool capable of recognizing single-handed gestures in three-dimensional space using budget-friendly ultrasonic transducers that generate a 40 kHz sound. They strategically positioned a transmitter and two receivers in a triangular layout to detect these gestures effectively. While these innovations demonstrate the feasibility of economical sonic gesture detection, they rely on either specialized hardware or limited sensors and locations, which poses a significant hurdle for broad-scale adoption. In our research, we concentrate on developing a methodology that is compatible with a broad array of existing devices and array sensors, thereby promoting immediate application development and widespread acceptance. Our approach isn't the first to incorporate sound-based methods or the Doppler effect in detecting gestures and movements [31], [32], [39], [40], [47], [49]. As an example, Tarzia and colleagues[49] analyzed the strength of echo signals captured by a microphone to identify human movement and attention. Paradiso and team [50] utilized a steady high frequency signal to operate specialized patch antennas, interpreting the modified Doppler signal for understanding human movement and upper body dynamics in interactive areas. Kalgaonkar and associates [48] more recently created a tool capable of recognizing single-handed gestures in three-dimensional space using budget-friendly ultrasonic transducers that generate a 40 kHz sound. They strategically positioned a transmitter and two receivers in a triangular layout to detect these gestures effectively. While these innovations demonstrate the feasibility of economical sonic gesture detection, they rely on either specialized hardware or limited sensors and locations, which poses a significant hurdle for broad-scale adoption. In our research, we concentrate on developing a methodology that is compatible with a broad array of existing devices and array sensors, thereby promoting immediate application development and widespread acceptance.

2.3.2 Direction detection

The incoming direction of the gesture is a valuable indicator to distinguish gesture types. In this way, the location and status of user's hand can be better described. In the scenario of limited gesture types, the ability of distinguish directions can enrich the recognizable gestures, by defining different gesture from different directions. As far as we have learned, few studies have been carried out to investigate the directionality in acoustic gesture recognition area. Chen [18][51] illustrated a 2d ultrasound gesture recognition system, and achieved an accuracy of over 94% in the experiment of representing numbers. However, their system adopted a DFT(Discrete Fourier Transform) based appoach whose resolution is not precise enough for determing direction of fast moving hand in a real time recogion system. (Multiple Signal Classification algorithm) MUSIC [1][2][52] is a well-known algorithm in Blind Source Separation [53] (BSS) and Sound Source Localization [54] area (SSL). It is an ingenious algorithm which separate the signal into sub-space, whose directional resolution is outperformed than the traditional angle of arrival algorithm, e.g. DTF, TDOA etc., which is implemented in this study.

The basic idea of MUSIC[52] is to separate the received signal into signal subspace and noise subspace by processing the array output's covariance matrix, then determining the signal direction via the angle scanning.

Assume a Uniform Linear Array (ULA) [55] involved M sensors with equal space $D=\frac{\lambda}{2}$ (λ is the signal wavelength) and there are D source signals from different directions from far field. [56]The signal received from the i^{th} sensor can be expressed as:

$$X_i(t)=\sum_{k=1}^{D}S(k)\exp\left(-j\frac{2\pi}{\lambda}D\sin\theta_k(i-1)\right)+n_i(t) \qquad (4)$$

where $X_i(t)$ is the received signal of the ith sensor, $S(k)$ is the kth signal vector from the source, θ_k is the direction of the kth signal. $n_i(t)$ is the noise signal vector received from the ith sensor. The overall receiver array can be expressed as

$$X_i(t) = \sum_{k=1}^{D} S(k) \exp\left(-j \frac{2\pi}{\lambda} D \sin\theta_k (i-1)\right) + n_i(t) \tag{5}$$

where $X_i(t)$ is the received signal of the ith sensor, $S(k)$ is the k[th] signal vector from the source, θ_k is the direction of the k[th] signal. $n_i(t)$ is the noise signal vector received from the i[th] sensor. Hence, the overall receiver array can be expressed as

$$X(t) = A(\theta)S(t) + N(t) \tag{6}$$

where $A(\theta) = [0, a(\theta_1), a(\theta_2), \ldots, a(\theta_D)]$ is the phase difference vector, $S(t) = [S(1), S(2), \ldots, S(D)]$ is the source signal vector, $N(t)$ is the noise vector.

$$\begin{aligned} R_x &= E\left[X(t)X(t)^H\right] \\ &= E\left(A(\theta)S(t) + N(t)\right)\left(A(\theta)S(t) + N(t)\right)^H) \\ &= A(\theta)\mathrm{E}\left[S(t)S(t)^H\right]A(\theta)^H + E\left[N(t)N(t)^H\right] \\ &= A(\theta)R_s A(\theta)^H + R_N \end{aligned} \tag{7}$$

$$R_N = \sigma_n^2 I \tag{8}$$

where R_s is the signal autocorrelation matrix and R_N is the noise autocorrelation matrix, σ_n^2 is the noise variance.

Equation **(9)** is established upon an ideal condition where $X(t)$ contains infinite samples. In practice, the following estimation is implemented.

$$\hat{R} = \frac{1}{L}\sum_{i=1}^{L} X(t)X(t)^H \tag{9}$$

The eigen decomposition of the received signal can be expressed as :

$$R_x = U\Sigma U^H = U_S \Sigma_S {U_S}^H + U_N \Sigma_N {U_N}^H \tag{10}$$

where U and Σ represent eigenvalue and eigenvector respectively, U_S and U_N represent signal and noise subspace. Then the eigenvalue is sorted from largest to smallest:

$$\Sigma = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_M \end{bmatrix} \tag{11}$$

$$\lambda_1 \geq \lambda_2 \geq \ldots \lambda_D \geq \lambda_{D+1} = \ldots = \lambda_M \tag{12}$$

$$U = \begin{bmatrix} U_S & \\ & U_N \end{bmatrix} \tag{13}$$

The first D eigenvalue corresponds to the source signal and the rest M-D corresponds to noise. Since noise subspace is orthogonal with source signal angle, the equation (8) can get:

$$A(\theta)^H U_N = 0 \tag{14}$$

The MUSIC spectrum is defined as:

$$P(\theta) = \frac{1}{a(\theta)^H U_N U_N{}^H a(\theta)} \tag{15}$$

where the peak represents the angle of the source signal.

2.3.3 ClassificationClassification approach

The proposed methodology, designed to function across varied scenarios of differing complexities, will inherently encounter noise interference in its feature extraction process, particularly in spectrum and angle measurements. This noise predominantly originates from two sources: the multipath issue related to the background environment, and unintended reflections caused by movements of other body parts, such as the arms, chest, and abdomen. Besides, as a wider angle mask was adopted when calculating hand direction, it contained redundant information from noise. There were a few methods proposed to solve the noise issue using frequency-hopping, FMCW, OFDM, frequency selected method [13], [57], [58] with multiple frequencies signals and achieved decent progress. However, all those proposed methods try to eliminate the noise from input features to restore the desired features. It suffers from some problems, for example: 1). When a signal is switching between two frequencies, an undesired and audible transmission noise could be generated. 2). The additional algorithms increase computational cost and slow down operation speed. 3). The algorithm performance may deteriorate when the application environment changes. On the other hand, if a method could properly extract useful features from all input features, the noise will be ignored automatically. Based on this idea, a CNN-based classifier was adopted to extract useful features automatically and reduce the noise simultaneously. CNN is one type of neural network with successful applications such as image recognition, computer vision and natural language processing [59]. By adopting a

convolutional layer, the input feature will be extracted independently by kernels. Through training, those kernels can eventually extract features that represent input features and ignore undesired noise.

2.4 Introduction of Existing algorithm

2.4.1The governing equation

In COMSOL Multiphysics,[60], [61], the governing equations [62]–[64]form the backbone of simulations across various modules, including the Acoustics Module. It represents the fundamental physical principles that underlie the acoustic phenomena being simulated. For example, in acoustics, these could be the Helmholtz equation for harmonic sound fields, the linearized Navier-Stokes equations for fluid flow, or the elasto-dynamic[65] equations for structural vibrations. COMSOL allows for customization of these equations to fit specific scenarios. Users can modify or add source terms, boundary conditions, and material properties to accurately represent the physical situation under investigating. By utilizing the governing equations, the Acoustics Module in COMSOL facilitates a comprehensive and versatile approach to simulating a wide range of acoustic problems in research and industry.

Standard acoustic problem involves solving small acoustic pressure variation p on the top of stationary background acoustic pressure p_0. This represents a small parameter variation around the large stationary values.

The governing equation of fluid flow explains most of the situation for the sound propagation. In compressible lossless fluid flow problem, it can be derived from two equations: (1) the momentum equation and (2) continuity equation as shown.

$$\frac{\partial u}{\partial t}+\left(u\cdot\nabla\right)u=-\frac{1}{\rho}\nabla p$$

$$\frac{\partial \rho}{\partial t}+\nabla\cdot\left(\rho u\right)=0 \qquad (16)$$

where ρ is the total density, p is the total pressure, and u is the velocity field. In classical pressure acoustics all thermodynamic processes are assumed reversible and adiabatic, known as an isentropic process.

The small parameter expansion is performed on a stationary fluid of density ρ0 (SI unit: kg/m3) and at pressure p0 (SI unit: Pa) such that:

$$p = p_0 + p'$$

$$u = u'$$

$$\text{Where } p' \ll p_0 \tag{17}$$

where the primed variables represent the small acoustic variations. Inserting these into the governing equations and only retaining terms linear in the primed variables yields One of the dependent variables, the density, is removed by expressing it in terms of the pressure using a Taylor expansion (linearization)

$$\rho' = \left.\frac{\partial \rho_0}{\partial p}\right|_s p' = \frac{1}{C^2} p' \tag{18}$$

where C is recognized as the (isentropic) speed of sound (SI unit: m/s) at constant entropy s.

Finally, rearranging the equations (divergence of momentum equation inserted into the continuity equation) and dropping the primes yields the wave equation for sound waves in a lossless medium.

$$\frac{1}{\rho_0 C^2}\frac{\partial^2 P}{\partial t^2} + \nabla \cdot \left(-\frac{1}{\rho_0} \nabla_P \right) = 0 \tag{19}$$

Impedance

The formula for acoustic reflection at the interface shown in the figure 3 can be represented as:

$$\left| \frac{E_1 r}{E_1 \mathrm{i}} \right| = \frac{z_z - z_1}{z_2 + z_1} \tag{20}$$

Where $E_1 r$ indicates the sound pressure reflection on the surface. z is the impedance of

the obstacle surface.

This is an acoustic field reflection coefficient which equals the ratio of reflected wave to incident wave in terms of wave amplitude. It determines the reflection rate from the incident energy.

Figure. 1 Transmission coefficient

2.4.2 Signal extraction and filtering

Signal is transmitted omnidirectionally from the speaker, and it reflect by all the obstacles within the operation area.

The emitted signal can be represented as.

$$S_{Tran} = A \times cos 2\pi ft \tag{21}$$

where A is the amplitude and f is the sound frequency.

Considering signal propagates along different path and the signal phase changes along time, it can be represented as:

$$\frac{1}{2}\sum_{k=1}^{P} A_k \left(\cos\left(-\frac{2\pi f d_k(t)}{c} - \theta_k \right) \right) \tag{22}$$

When the signal reaches at an obstacles, a echo reflects back. The echos from N stationary obstacles and P moving obstacles are recorded by the receiver. The received signal can be expressed as:

$$\begin{aligned} S_{rec}(t) &= \sum_{i=1}^{N} A_i \cos(2\pi ft - \theta_i) + \\ &\sum_{k=1}^{P} A_k \cos\left(2\pi ft - \frac{2\pi f d_k(t)}{c} - \theta_k \right) + N(t) \end{aligned} \tag{23}$$

where A_i and A_k represent the amplitudes of echo reflections from stationary obstacles

and moving obstacles, respectively. θ_i and θ_k represent stationary phase changes during propagation, c is the sound speed, and $N(t)$ represents noise. To mitigate noise component $N(t)$, a high pass filter is applied once the signal is received and the rest signal contains reflection signal from both stationary obstacles and the moving hand only. An multiplcation is carred out to derive the moving hand echo, using the remaining signal and the original signal $cos2\pi ft$:

$$
\begin{aligned}
S_I(t) &= S_{rec}(t) \times cos2\pi ft \\
&= \frac{1}{2}\sum_{i=1}^{N} A_i\left(\cos\left(4\pi ft-\theta_i\right)+\cos\left(\theta_i\right)\right)+ \\
&\frac{1}{2}\sum_{k=1}^{\mathrm{P}} A_k\left(\begin{array}{l}\cos\left(4\pi ft-\dfrac{2\pi fd_k(t)}{c}-\theta_i\right)+ \\ \cos\left(-\dfrac{2\pi fd_k(t)}{c}-\theta_k\right)\end{array}\right) \\
&= \frac{1}{2}\sum_{k=1}^{\mathrm{P}} A_k\left(\cos\left(-\frac{2\pi fd_k(t)}{c}-\theta_k\right)\right) \\
&\quad +C_I+\frac{1}{2}\sum_{i=1}^{N} A_i\left(\cos\left(4\pi ft-\theta_i\right)\right) \\
&+\frac{1}{2}\sum_{k=1}^{\mathrm{P}} A_k\left(\cos\left(4\pi ft-\frac{2\pi fd_k(t)}{c}-\theta_i\right)\right)
\end{aligned}
\tag{24}
$$

where $C_I = \frac{1}{2}\sum_{i=1}^{N} A_i\left(\cos\left(\theta_i\right)\right)$ remains as a constant. Note that the maximum hand moving speed is less than 5m/s normally. According to the Doppler effect, both frequency of the first two terms: $= \frac{1}{2}\sum_{k=1}^{\mathrm{P}} A_k\left(\cos\left(-\frac{2\pi fd_k(t)}{c}-\theta_k\right)\right)$ and C_I in **equation (24)** are less than 300 Hz. The frequency of the last two terms, $\frac{1}{2}\sum_{k=1}^{\mathrm{P}} A_k\left(\cos\left(4\pi ft-\frac{2\pi fd_k(t)}{c}-\theta_i\right)\right)$ and $\frac{1}{2}\sum_{i=1}^{N} A_i\left(\cos\left(4\pi ft-\theta_i\right)\right)$, are approximately 40kHz. Those signals are much higher than the desired signal and they are eliminated using a low pass filter. By applying a normalization to remove the constant bias C_I, the signal reflected from the hand can be represented as:

$$S_{lowI} = \frac{1}{2}\sum_{k=1}^{\mathrm{P}} A_k \left(\cos\left(-\frac{2\pi f d_k(t)}{c} - \theta_k \right) \right) \tag{25}$$

In the same way, the signal Quadrature component Q can be obtained by a similar step but multiplying an orignal sin wave signal $sin2\pi ft$:

$$S_I(t) = S_{rec}(t) \times \sin(-2\pi ft) \tag{26}$$

$$= \frac{1}{2}\sum_{k=1}^{\mathrm{P}} A_k \left(\sin\left(-\frac{2\pi f d_k(t)}{c} - \theta_k \right) \right) + C_Q$$
$$+ \frac{1}{2}\sum_{i=1}^{N} A_i \left(\sin(4\pi ft - \theta_i) \right)$$
$$- \frac{1}{2}\sum_{k=1}^{\mathrm{P}} A_k \left(\sin\left(4\pi ft - \frac{2\pi f d_k(t)}{c} - \theta_i \right) \right)$$

$$S_{lowQ} = \frac{1}{2}\sum_{k=1}^{\mathrm{P}} A_k \left(\sin\left(-\frac{2\pi f d_k(t)}{c} - \theta_k \right) \right) \tag{27}$$

Combine both I and Q components, the complex signal can be expressed as:

$$S = \sum_{k=1}^{\mathrm{P}} A_k' e^{-j\left(\frac{2\pi f d_k(t)}{c} + \theta_k \right)} \tag{28}$$

where $\mathrm{j}^2 = 1$.

Chapter 3 Smartphone Sound Field Simulation and Gesture Recognition Application Simulation

3.1 Introduction

Regarding with the developing efficient for effective gesture recognition systems, it is imperative to gain system development, a thorough deep understanding of the acoustic sound field dynamics around the intended devices. This aspect becomes particularly significant considering the diversity in the distribution is necessary and developed. The diverse array of acoustic devices across various device different models, each with their own unique sensor variants sensors and internal structures. Designs requires to be verified before implemented into reality. Such diversity invariably leads to variations in acoustic sound field characteristics, including sound pressure level, frequency response, and radiation direction. These variations in acoustic performance are crucial, as they have a direct bearing on the efficacy of gesture recognition systems. This chapter introduces an in-depth acoustic sound field simulation, meticulously designed to evaluate the feasibility and practicality of implementing gesture recognition on smartphone platforms, a critical step in advancing this field of research. Firstly, two acoustic models of smartphone were recreated based on real dimensional parameters. The sounding components were meticulously designed to simulate sound reproduction and capture process. The acoustic sound field of two smartphones was restored and the acoustic parameters were calculated and analyzed. Secondly, a sensitivity test model was developed. The simulated model was optimized, and parameters were calibrated. The design of the models followed the scenario of real component factory testing, and the acoustic parameters were adjusted based on the specific parameters provided by the manufacturer.

Thirdly, a sensitivity test based on a real smartphone was conducted to further improve the authenticity of the model. The same model of real smartphone was programmed to restore the factory test. A comprehensive evaluation was conducted in

terms of frequency response, Sound Pressure Level (SPL), SPL difference and gain. Fourthly, three application scenario simulations were conducted to simulate an operation scene in the real world. The hand was placed at a different location, in a different direction to the smart devices, and the acoustic reflection from a hand was simulated. A phase control was carried out to verify the potential radiation control.

This simulation has undergone a detailed analysis, providing convenient options for designing real-time gesture recognition in terms of selecting phones, operational distance, and more.

3.2 Smartphone modeling and calibration

3.2.1 Smartphone Modeling

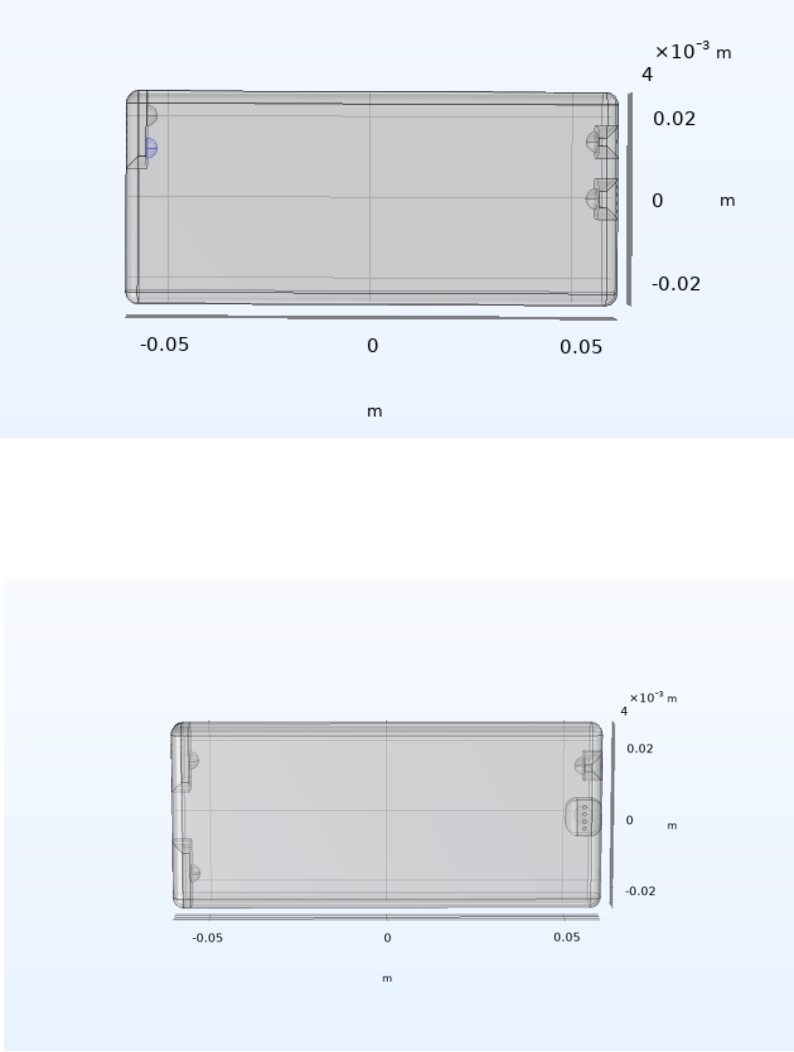


Figure. 2 Mate40 model and P20 pro model

At the beginning of smartphone acoustic field simulation, two common smartphone models were meticulously designed and simulated as shown in figure 3.2 and figure 3.3: 1. Huawei P20; 2. Huawei Mate40

The simulation is built based on the actual shape in the reality, and the model parameters in the simulation were based on the dimensions of real mobile phones.

The dimensions for the two models were set as: 158.6 mm x H 72.5 mm x W 8.8 mm

((6.24 in x 2.85 in x 0.35 in) for Huawei Mate 40, 149.1 mm x 70.8 mm x 7.7 mm (5.87 in x 2.79 in x 0.30 in) for Huawei P20. The acoustic devices (speakers and microphones) were set at the same location with the same size as the real smartphone.

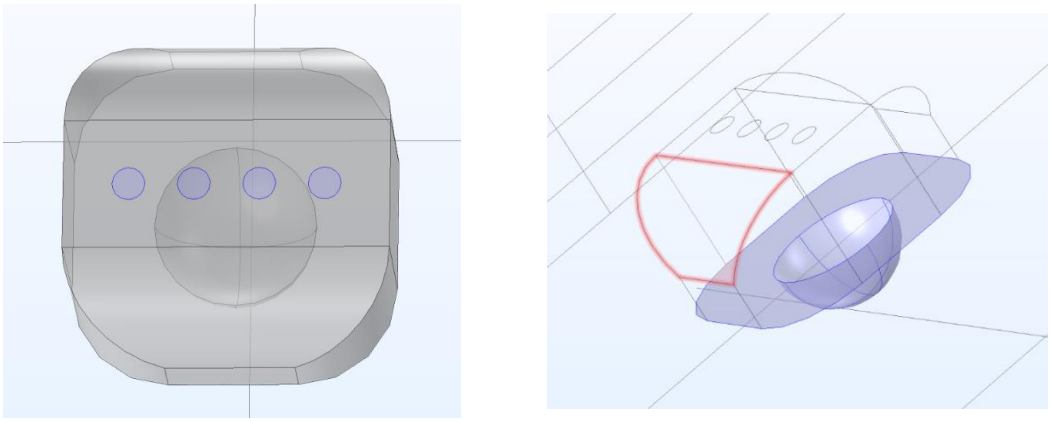


Figure 3 Top speaker transmission hole and Top speaker vibration plate

Acoustic devices have unique designs in smartphone devices, and they were also designed in simulations based on real acoustic devices.

As shown in the figures, the figure shows an example for earpiece on smartphone. It consists of two main components: the vibration plate and the sound transmission holes. When a sound wave arrives, it will firstly pass through the transmission holes, then propagate through the air in the cavity and finally reach the vibration chip at the end. The chip will record the SPL of the sound wave by recording the vibration speed. Similar to the earpiece, the bottom speakers have a similar design in terms of structure. As shown in figure 3, bottom speaker has more transmission holes and a lager vibration plate.

The size of the transmission hole is determined by the real smartphone structure, which is set at 2mm.

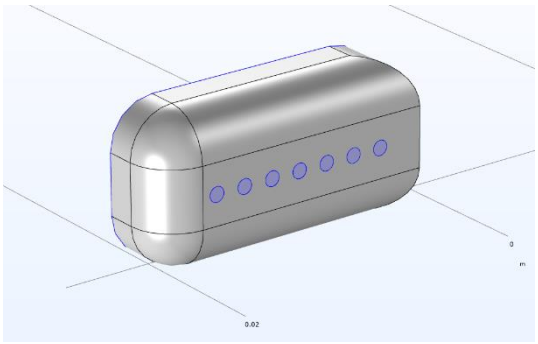


Figure. 4 Bottom speaker transmission hole

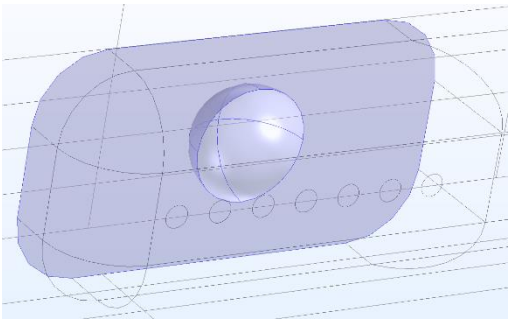


Figure. 3 Bottom speaker vibration plate

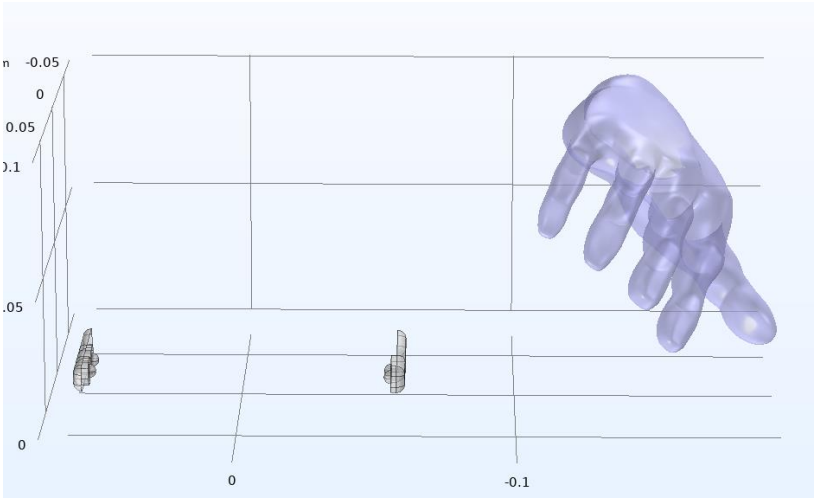


Figure. 5 Model of hand

To simulate the reflection of hand, a hand model has also been simulated as shown in the figure. The dimension of the hand was set as the average size of human's hand, which is 18.4cm length (7.6 inches).

3.2.2 System Calibration

Sensitivity Simulation Calibration

Acoustic sensitivity testing is a process to evaluate the responsiveness of an acoustic sensor, typically a microphone or a transducer, to sound pressure levels. The core objective is to quantify how effectively the sensor converts acoustic energy into

an electrical signal.

The sensitivity test [66] for smartphone microphones and speakers is a meticulous process designed to ensure these devices perform optimally in real-world conditions. This process begins in a controlled environment, often an anechoic chamber, which is crucial for eliminating external noise and echo, thereby ensuring accurate results. For microphones, a calibrated sound source emits sound at known frequencies and sound pressure levels (SPLs), and the microphone's output is recorded and analyzed. The sensitivity is calculated by comparing the electrical output of the microphone to the SPL across a range of frequencies. This helps in determining the microphone's ability to accurately capture sound.

Similarly, for speakers, a known electrical signal is fed into the speaker, and

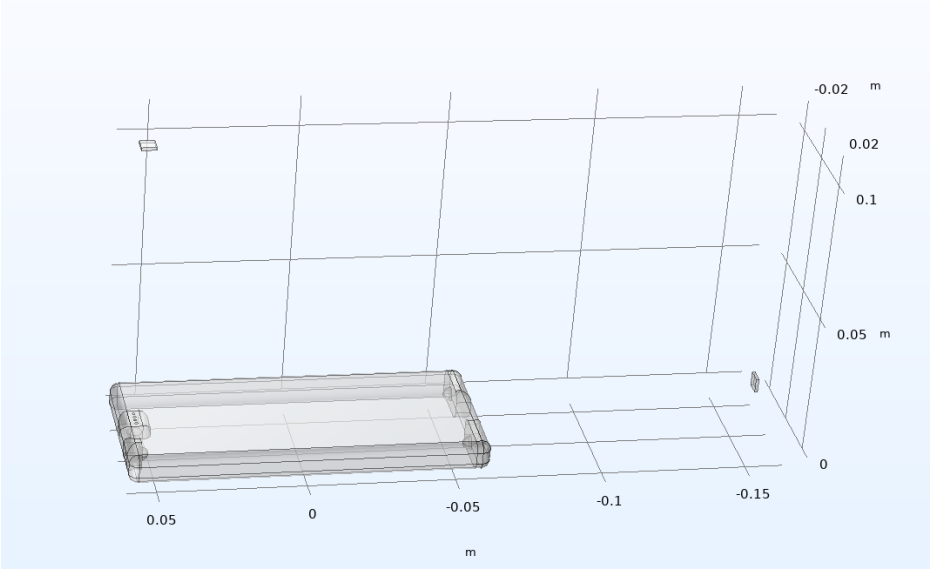


Figure. 6 Calibration scene

precision microphones positioned at a standard distance measure the SPL produced. The speaker's sensitivity is assessed by comparing the SPL output against the input power, often at various frequencies and input levels. This test is crucial in evaluating how effectively the speaker converts electrical signals into audible sound.

To bring the simulation closer to reality, a calibration test was conducted to measure the microphone's sensitivity based on manufacture datasheet. Following the standard sensitivity test procedure, a receiver is placed 10 cm away from the speaker, as illustrated in the figure, and it continuously receives signal. The test summarizes the

sound pressure received over time and calculated an average value. Similarly, the top speaker is tested and adjusted using the same method as the bottom speaker

Modeling Result

The procedure for fine-tuning the sensitivity of the right speaker involved assessing the SPL at a predetermined test position. This was achieved by adjusting the vibration speed of the speaker plate to vary the energy of the emitted sound waves. The average sound pressure on the testing surface was then computed and compared against reference parameters In case of discrepancies, further adjustments to the vibratio

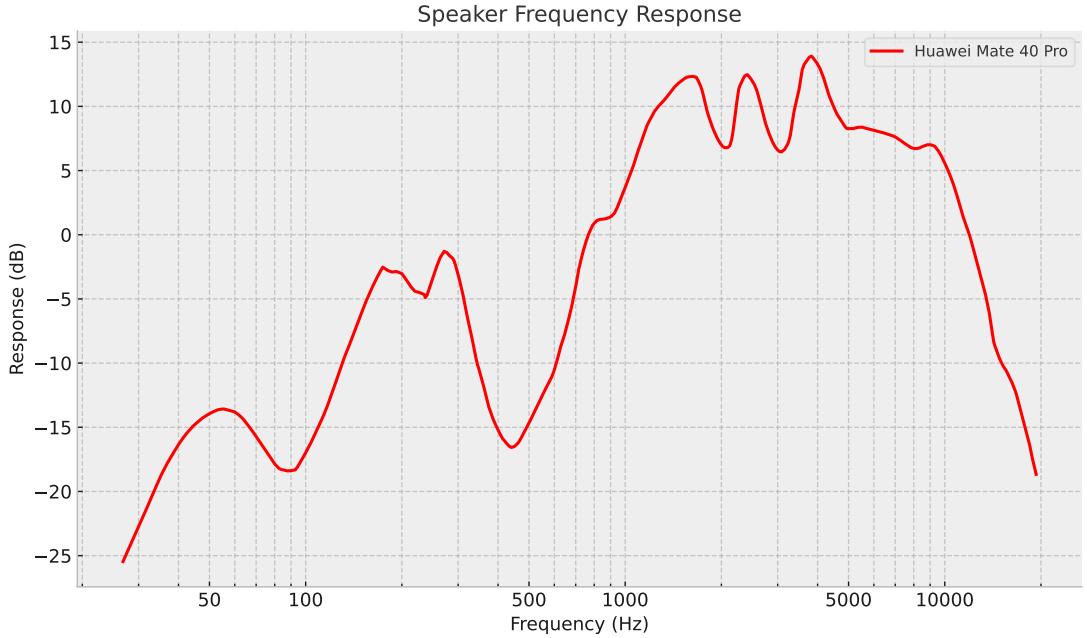


Figure. 7 Speaker frequency response

n speeds are implemented until the SPL aligns with the desired standard.

In addition, speaker frequency response was taken into consideration. As shown in Figure 8, the speaker has different vibration's ability according to different frequencies. This indicates that, with the same output voltage, the speaker produces different transmission energies at various frequencies.

The speaker exhibits its best response at 4.5kHz, producing an SPL that is 14dB higher at the test surface than at other frequencies. The curve is evaluated below 850Hz, and any decimal values shown are relative to 850Hz as the baseline. The calibration is shown as below.

Table. 1 P20 bottom speaker parameters

Top velocity	Bot velocity	F0(Hz)	Freq(Hz)	SPL(dB)
0.019156	0	18000	18000	76.982
0.005006	0	18500	18500	65.335
0.004277	0	19000	19000	64.246
0.003682	0	19500	19500	63.188
0.003124	0	20000	20000	61.952

Table. 2 P20pro top speaker

Top velocity	Bot velocity	F0(Hz)	Freq(Hz)	SPL(dB)
0	0.045793	18000	18000	79.986
0	0.011546	18500	18500	94.399
0	0.010354	19000	19000	73.094
0	0.008671	19500	19500	67.313
0	0.007399	20000	20000	63.835

Table. 3 Mate40 bottom speaker

Top velocity	Bot velocity	F0(Hz)	Freq(Hz)	SPL(dB)
0.096783	0	18000	18000	79.988
0.025174	0	18500	18500	68.344
0.021459	0	19000	19000	67.067
0.018493	0	19500	19500	64.514
0.015675	0	20000	20000	61.952

Table. 4 Mate40 top speaker

Top velocity	Bot velocity	F0(Hz)	Freq(Hz)	SPL(dB)
0	0.10068	18000	18000	78.986
0	0.02940	18500	18500	67.988
0	0.025119	19000	19000	65.535
0	0.058662	19500	19500	65.614
0	0.009633	20000	20000	64.234

Real Scenario Calibration Test

To achieve the most accurate sound pressure level that the transmitter received, a verification test was carried out to acquire the actual value, as shown in the figure below . The speaker has different frequency responds regarding to different frequencies;

Figure. 8 Calibration real scene

hence the test was designed for frequency range in following experiments.

The calibration result is shown in the Table 5, it presents the findings of the sensitivity analysis conducted on two smartphone devices, specifically the Huawei Mate 40 and P20 Pro, focusing on their audio output characteristics across a specified frequency range. The table provides quantitative data across several parameters including frequency, response, difference, gain, and sound pressure level (SPL). The "Frequency (Hz)" column represents the various frequencies tested in the study, ranging from 2kHz to 20kHz. This covers a significant portion of the human auditory range, enabling the assessment of the devices' performance over a broad spectrum of sound frequencies.

The SPL level at the sensitivity point was readjust to 0.05,0.0413,0.035,0.03,0.026 and 0.04579, 0.01154, 0.01035, 0.008671,0.007399 respectively in terms of frequency.

Table. 5 Real scenario calibration result

Frequency(Hz)	2k	18k	18.5k	19k	19.5k	20k
Response(dB)	12.142	-13.872	-15.554	-16.906	-18.217	-19.624
Difference(dB)	0	26.014	27.691	29.048	30.359	31.766
Gain Mate40	1	0.05	0.0413	0.035	0.03	0.026
SPL(dB)Mate40	93	76.986	65.309	63.952	62.641	61.234
SPL(dB)P20pro	93	76.986	68.359	66.371	65.221	64.333
Gain P20pro	1	0.04579	0.01154	0.01035	0.008671	0.007399

Determination of Hand Reflection.

According to the formula in mathematical principle, the transmission coefficient can be written as $T = \frac{Z_{hand} - Z_{air}}{Z_{hand} + Z_{air}}$ which is mainly determined by the acoustic impedance Z_{hand}. The impedance of hand through the tissue medium (m/s) which represented as $z_{hand} = \sqrt{\rho c}$. According to relative research, the density of human's skin is around 1615 kg/m3, the speed of sound in human body is around 1090 m/s. The impedance of hand is calculated: $z_{hand} = \sqrt{\rho c} = 1326.78(Pa \bullet s / m^3)$.

3.3 Application Scenarios Analysis

3.3.1 Smartphone's Acoustic Test

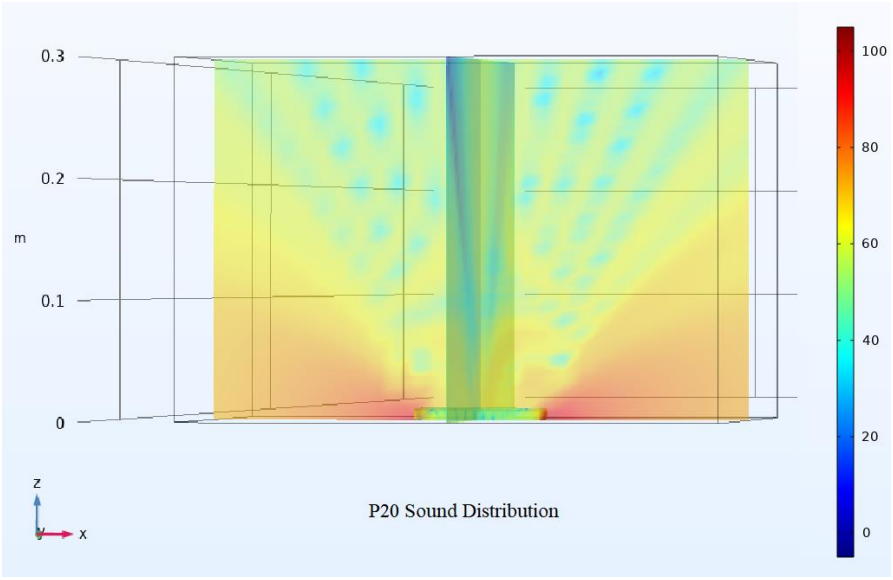


Figure. 9 P20 sound field distribution

The sound field distribution in near field is shown in the figure below in terms of P20 and Mate40

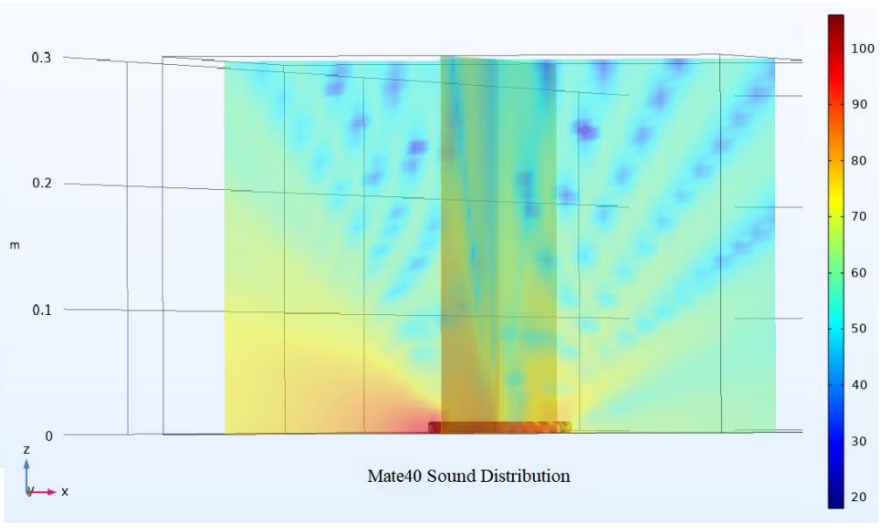


Figure. 10 Mate40 sound field distribution

Due to the different locations of the speakers, the sound field varies across the entire area. The P20 has two speakers: one located at the bottom of the phone, facing

in the -X direction, and the other at the top, facing in the +Z direction. In contrast, the Mate40's speakers are located at the bottom facing in the -X direction and at the top facing in the +X direction. As evident in Figures 23 and 24, the Mate40 exhibits a lower Sound Pressure Level (SPL) compared to the P20 in the +Z direction and significantly lower SPL on the Y-Z surface.

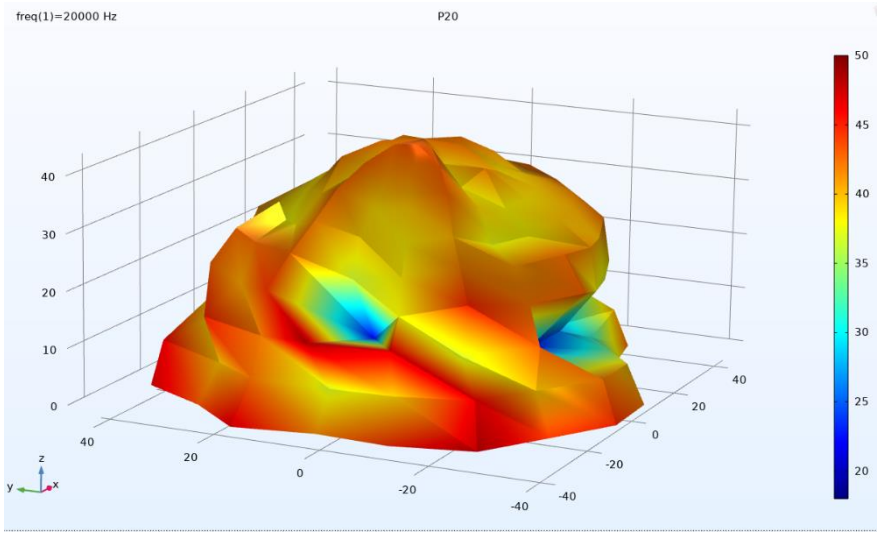


Figure. 12 Radiation pattern P20pro side view(a)

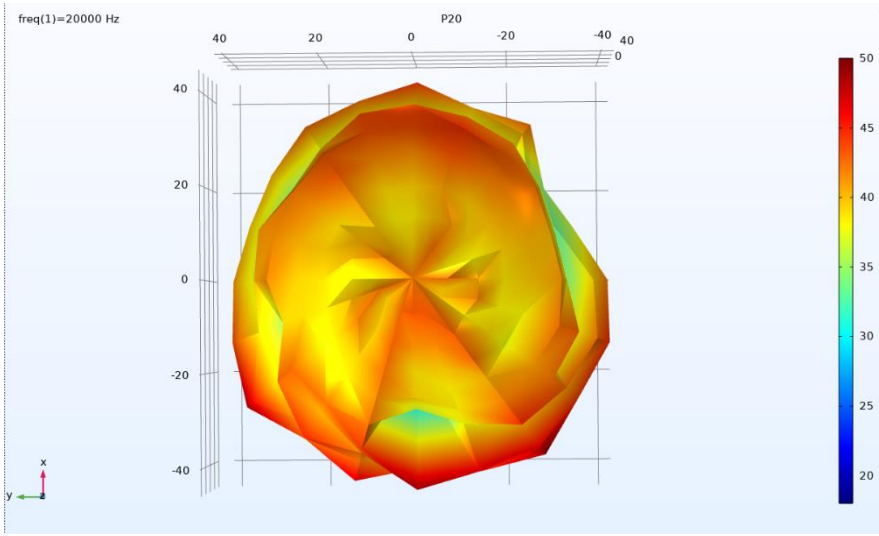


Figure. 13 Radiation pattern P20pro Top view(a)

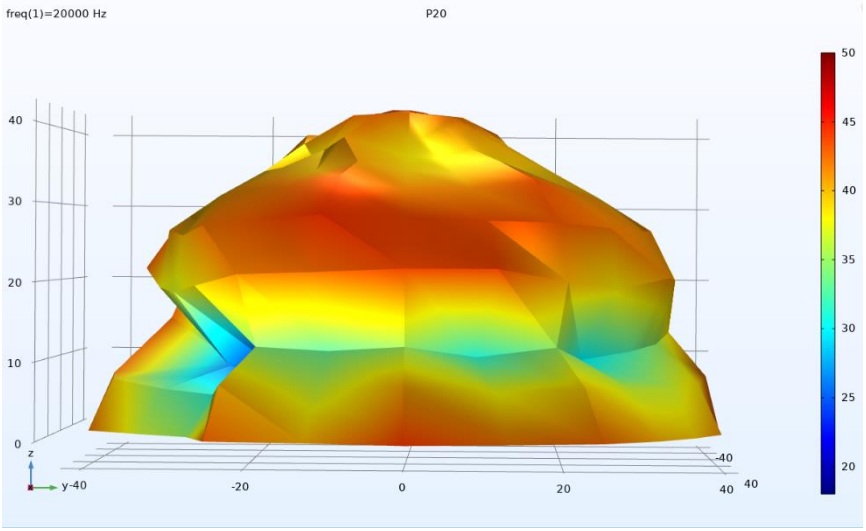


Figure. 11 Radiation pattern P20pro side view(b)

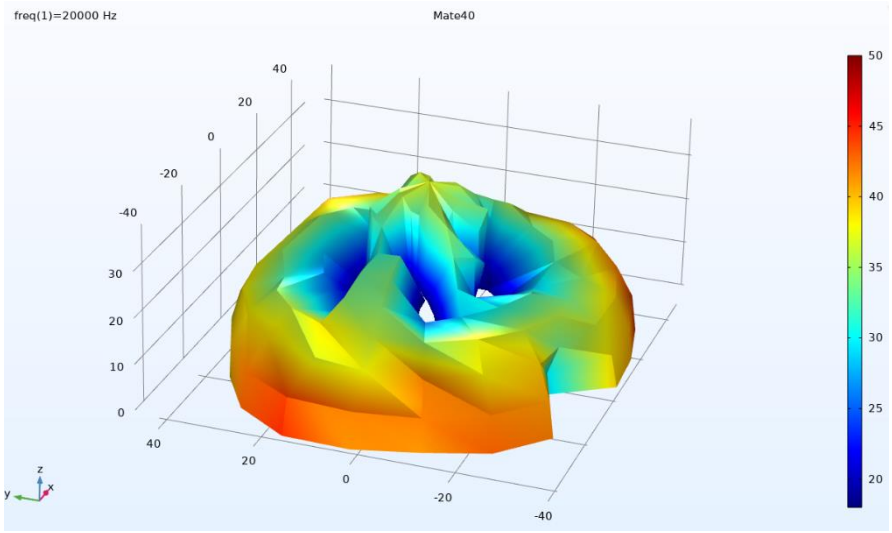


Figure. 16 Radiation pattern Mate40 side view(a)

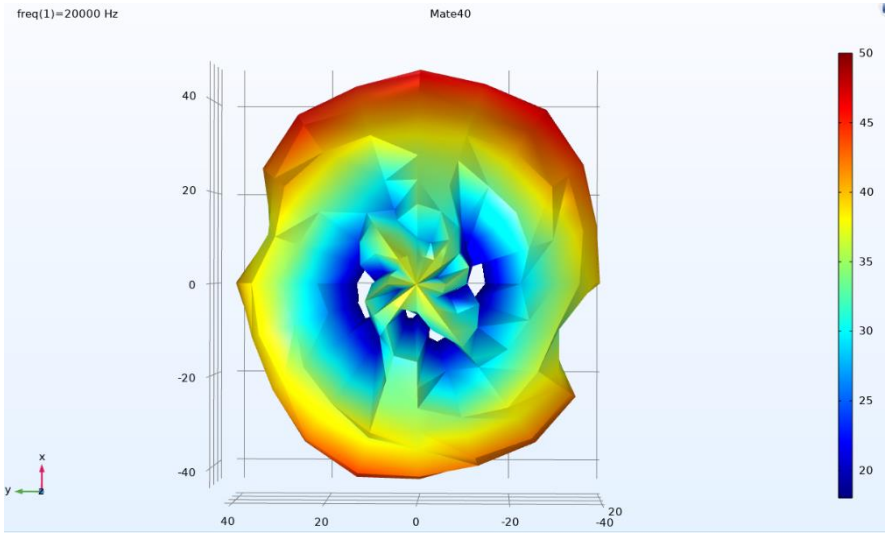


Figure. 15 Radiation pattern Mate40 top view(a)

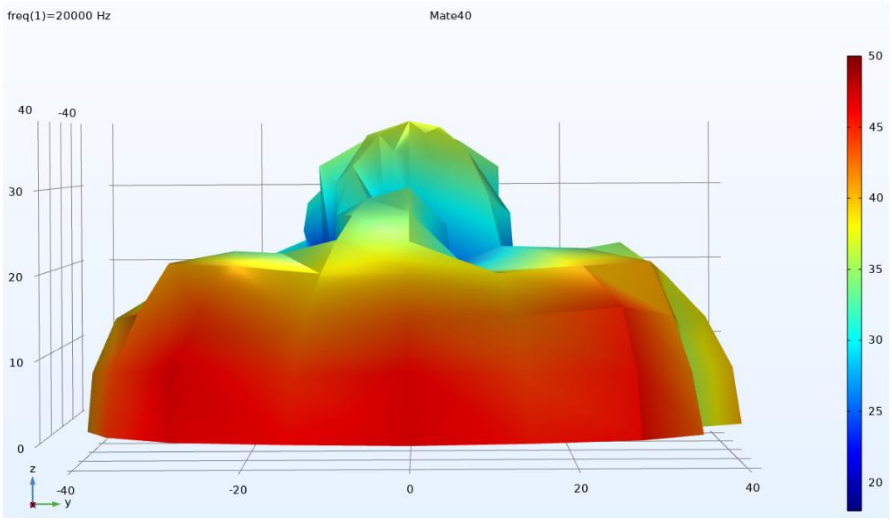


Figure. 14 Radiation pattern Mate40 side view(b)

3.3.2 Hand's Reflection from The Acoustic Field

The section aiming to investigate the sound performance of smart phone when a hand is engaged.

An application scene is carried out when the hand is set at 10 cm above the

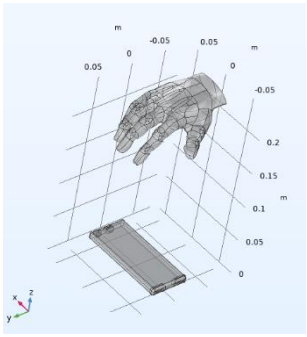


Figure. 17 Hand reflection coordinate

smartphone. Figure below shows P20 acoustic field and the Mate40 acoustic field. The maximum value on the hand surface is shown in Table 6.

Table. 6 P20 hand maximum SPL

Top velocity	Bot velocity	F0(Hz)	Freq(Hz)	SPL(dB)
0.01915	0.04579	18000	18000	69.171

Table. 7 Mate 40 hand maximum SPL

Top velocity	Bot velocity	F0(Hz)	Freq(Hz)	SPL(dB)
0.01915	0.04579	18000	18000	52.145

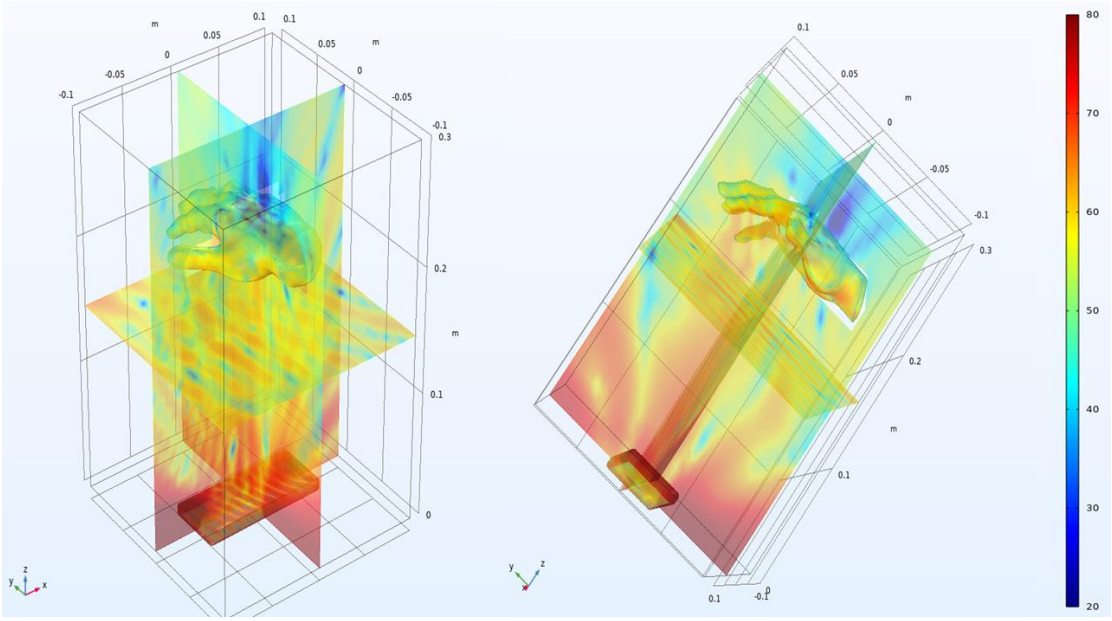


Figure. 18 Acoustic field - hand above situation P20pro

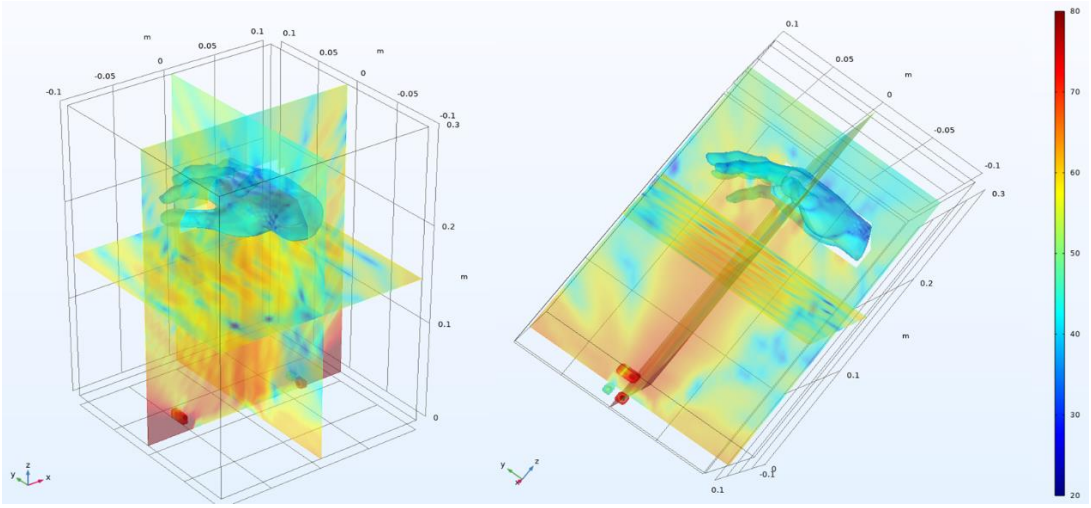


Figure. 19 Acoustic field -hand above situation Mate40

This section is aimed at exploring the influence on hand location.

Microphone Received SPL from Hand.

The line chart shown in the figure below represents the sound pressure level (SPL) received by the microphones of the Huawei P20pro. In the chart, the blue line indicates the SPL from the top microphone, while the green line represents the SPL from the bottom microphone. The general noise background is measured at 22.6dB, determining the maximum operational range for the Huawei P20pro to be 0.15m..

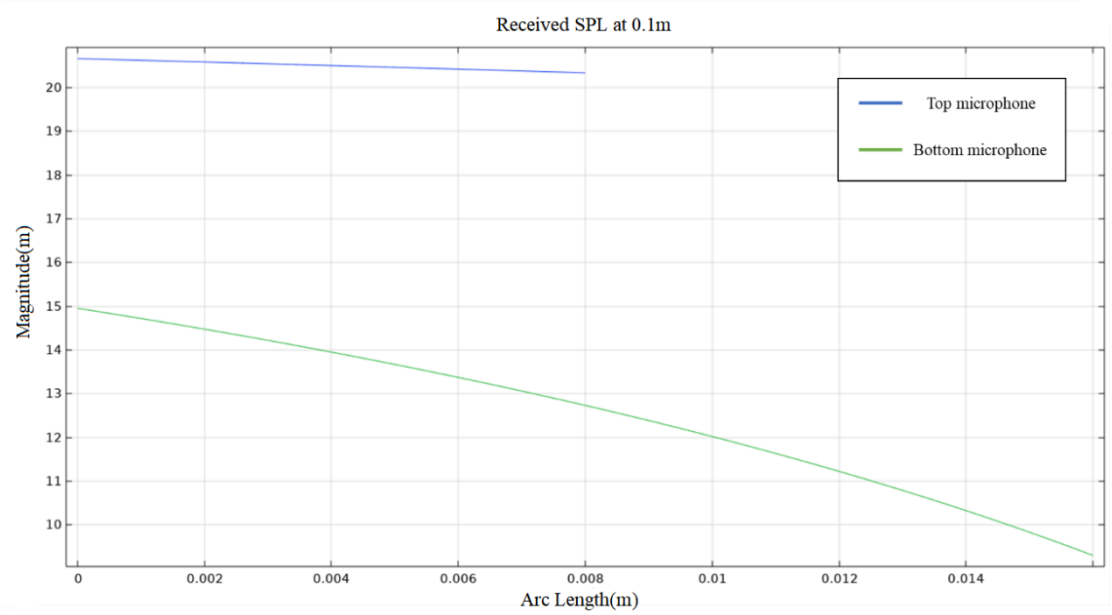


Figure. 20 Received SPL at 0.1m above

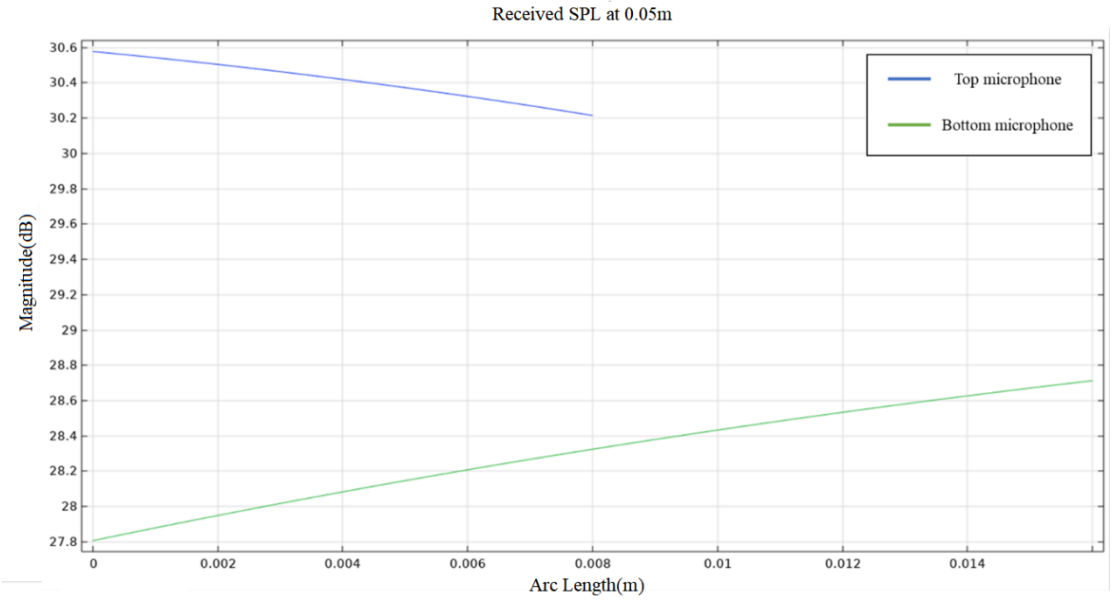


Figure. 21 Received SPL at 0.05m above

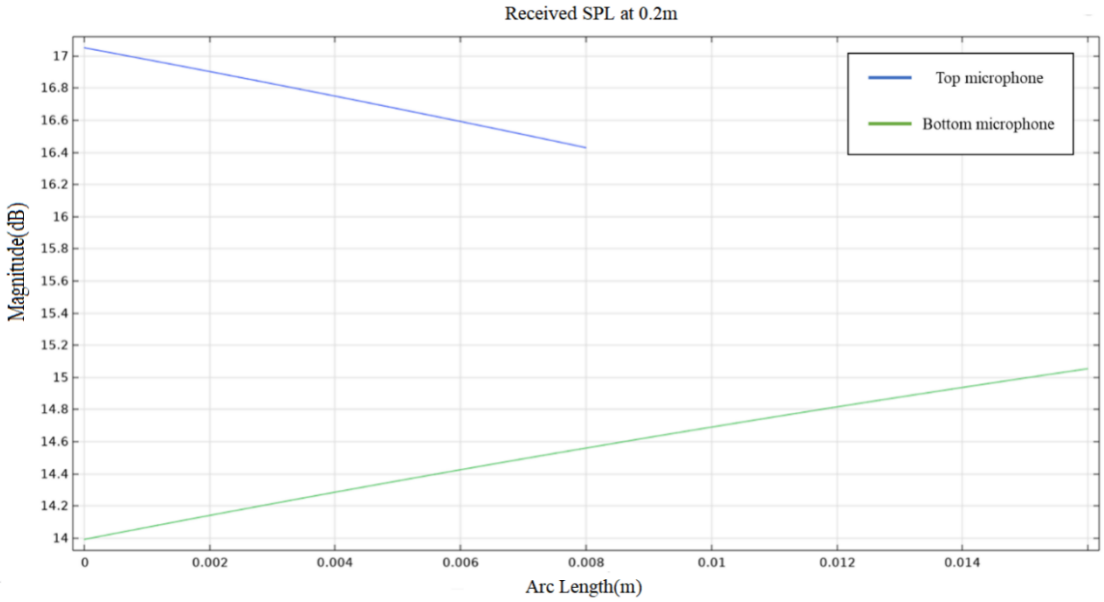


Figure. 22 Received SPL at 0.2m above

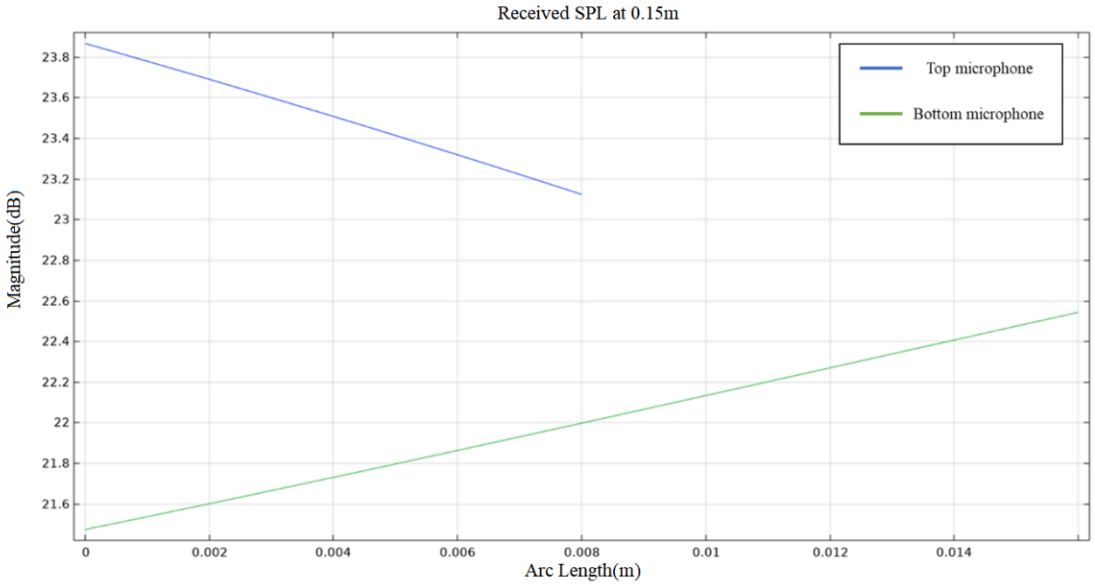


Figure. 23 Received SPL at 0.15m above.

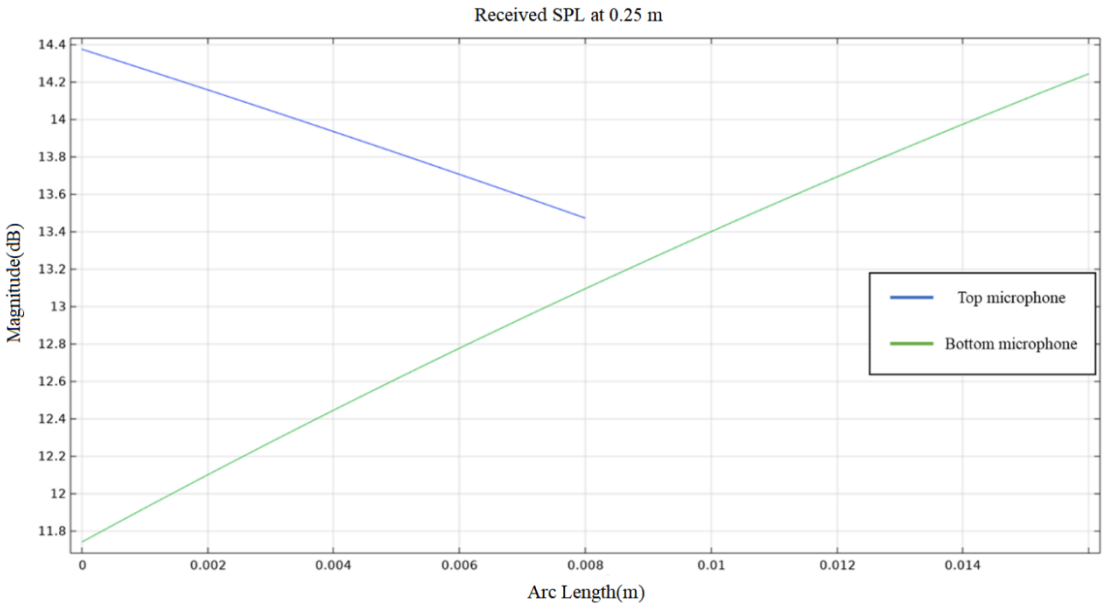


Figure. 24 Received SPL at 0.25m above

The result exhibit that the operation range is limited less than 15cm.

3.4.4 Hand's Direction Experiment

The objective of this experiment is to test the influence of hand placement in different directions. In the simulation, the hand was positioned in five distinct directions (-40°, -20°, 0°, 20°, 40°), covering the majority of common gesture presentation angles. We assessed both the sound pressure level (SPL) on the hand and the signal received by the smartphone's microphone. The distance between the microphone and the hand was consistently maintained at 10 cm, with the results displayed in the figures below. The result is shown below:

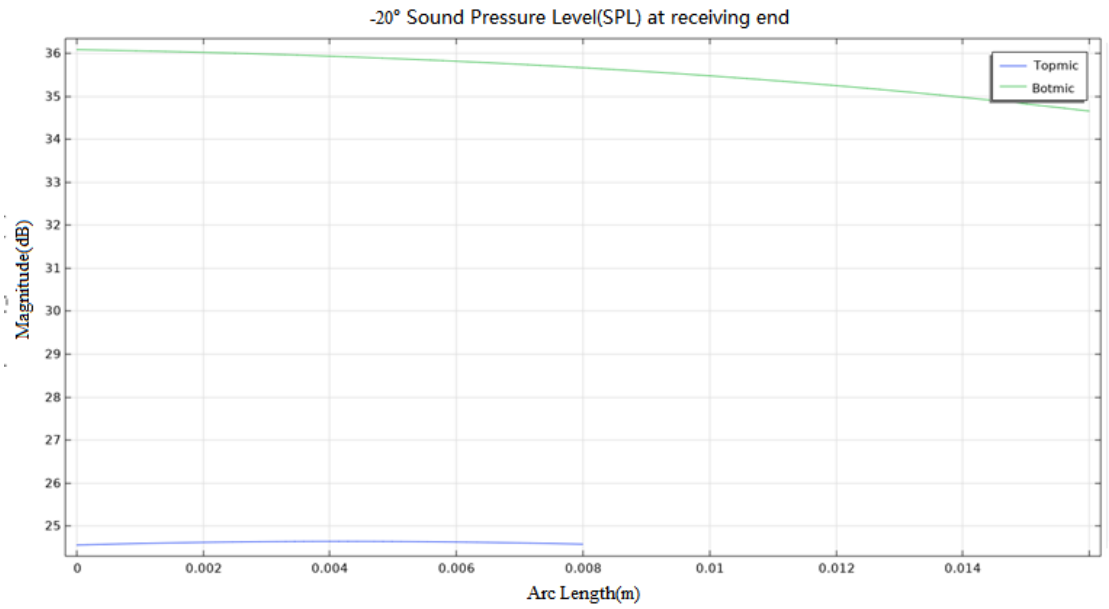


Figure. 26 Received SPL at -20 degree

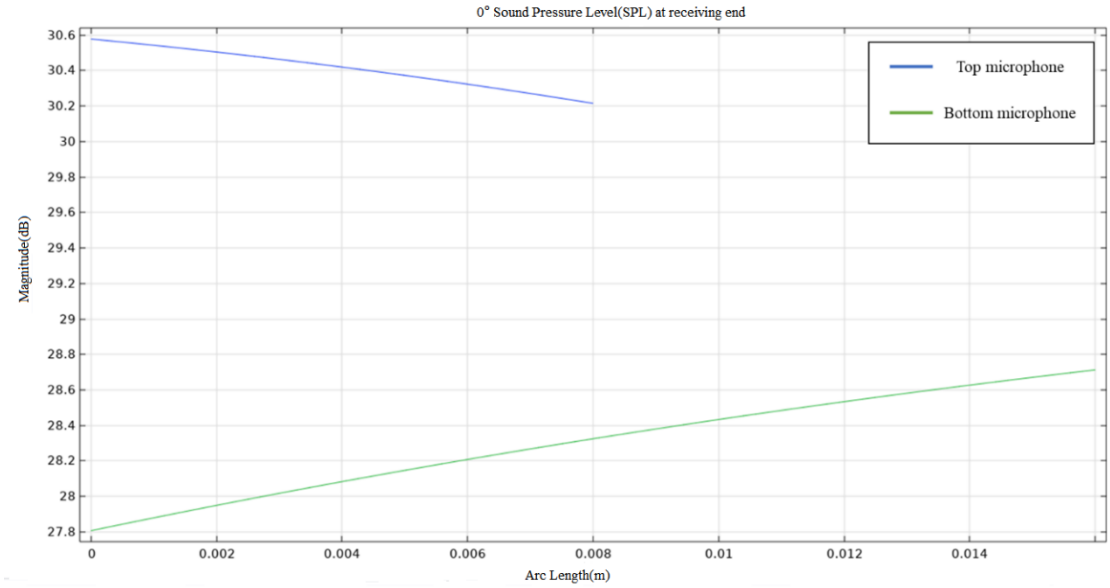


Figure. 27 Received SPL at 0 degree.

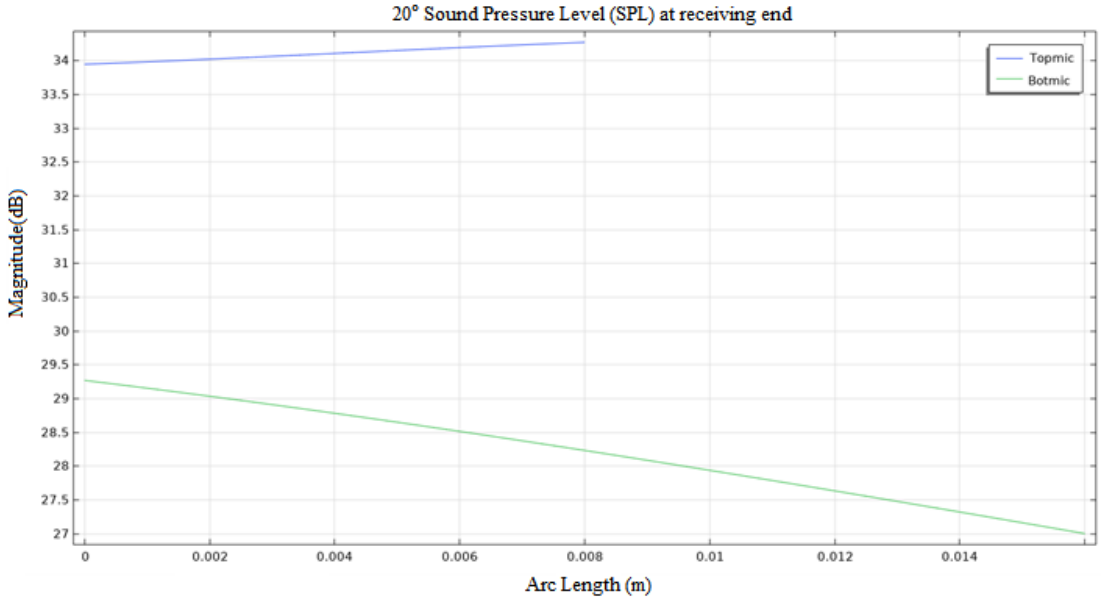


Figure. 28 Received SPL at 20 degree

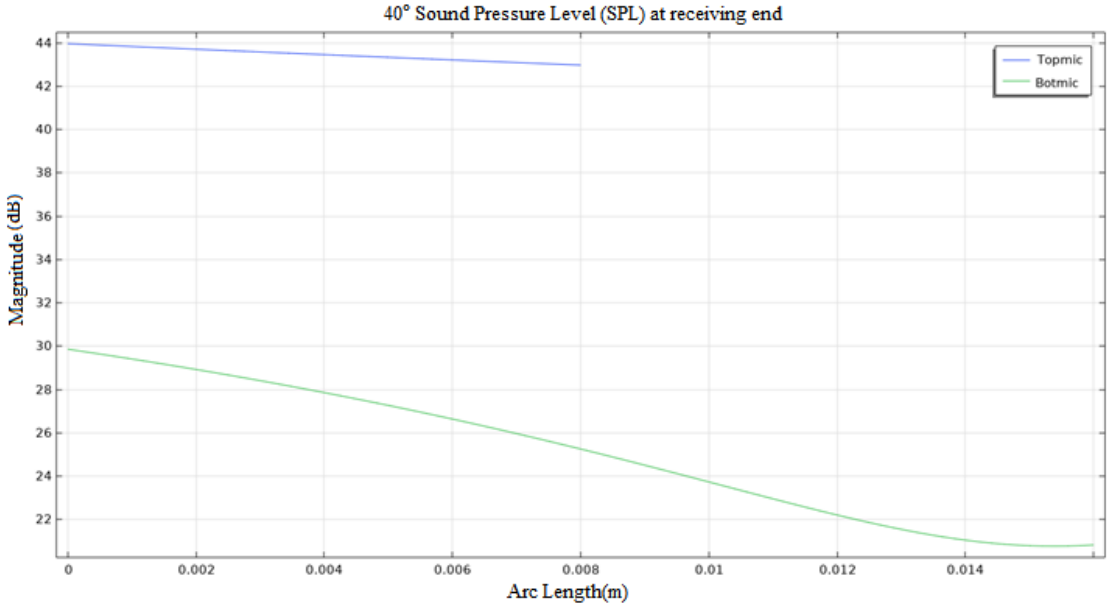


Figure. 29 Received SPL at 20 degree.

The average SPL for 5 kinds of degrees are 42 dB, 35.5 dB, 30.4 dB, 34.3 dB and 41.5dB, it proves that the hand rotation from hand provides a stronger SPL reflection from different direction than 0 degrees. Gestures from different angles can be received and have a high sound reflection pressure than 0 degrees.

3.3.5 Frequency Enhancement Experiment

2D Phase Change Test

This section aims to illustrate a frequency modulation system that take the frequency difference into consideration and seek any potential enhancement by adjusting frequency and shift.

A 2D simulation experiment, adjusting phase difference and observing the result.

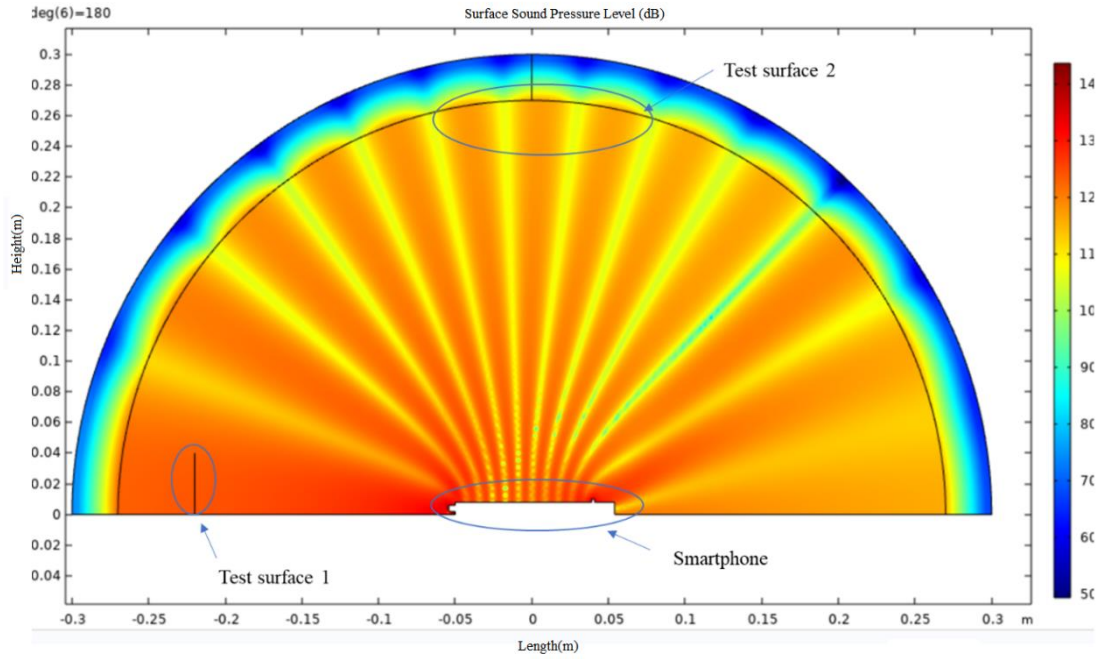


Figure. 30 2D Simulation test scene

By changing the phase on the transmission signal, we evaluate sound pressure level on the two test surfaces to estimate the importance.

In this study, a comprehensive simulation was conducted on six distinct phase conditions, separated by intervals of 30 degrees, to scrutinize the system's performance under various phase shifts. This was coupled with a methodical placement of two test surfaces at -0.2 cm and 0.26 cm, likely along the vertical axis, aiming to evaluate the performance gradient of the device or system across different vertical positions. The

systematic design of the experiment could provide substantial insights into device performance under diverse phase and positional parameters.The result is shown below:

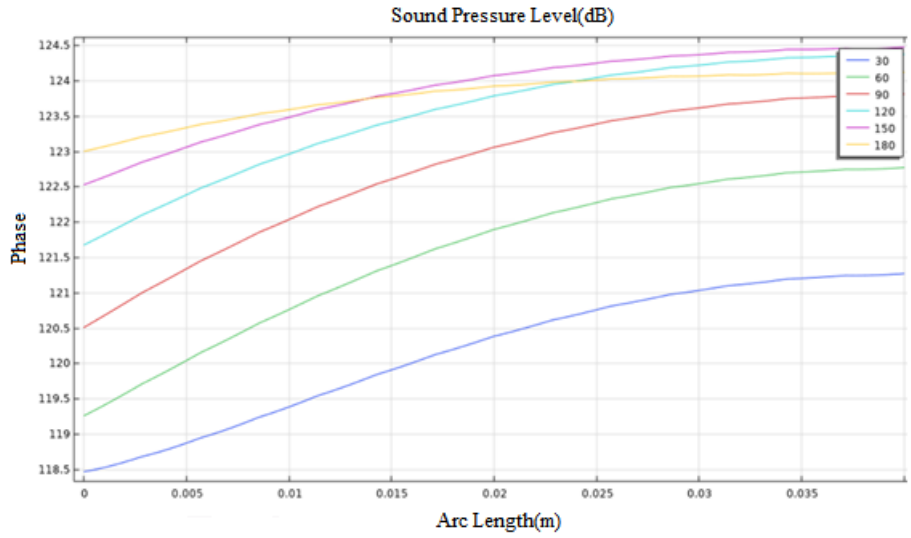


Figure. 31 SPL change on test surface 1 regarding to different phases.

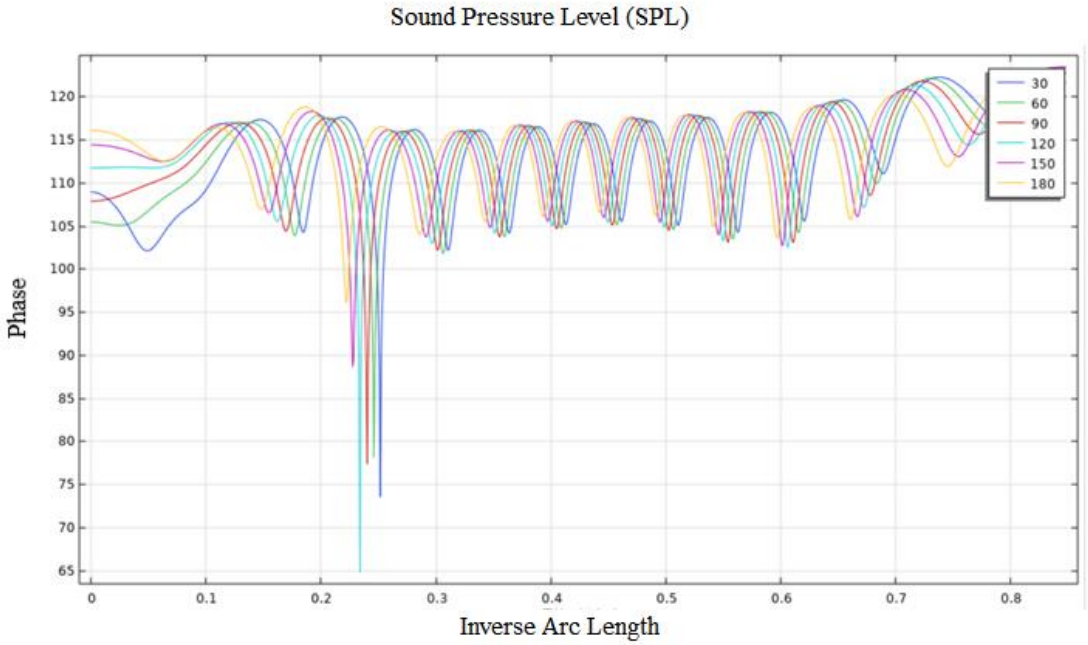


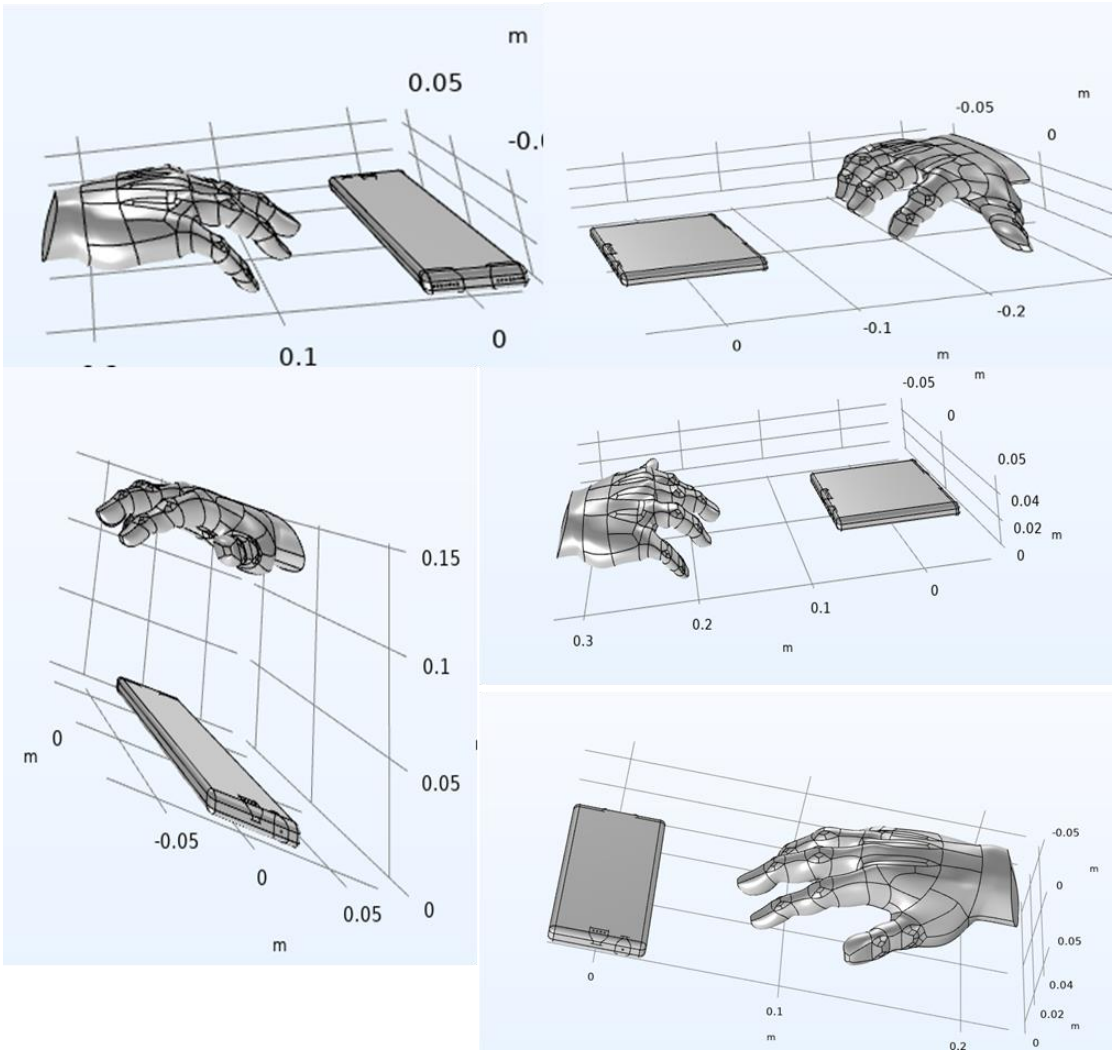
Figure. 32 SPL change on test surface 2 regarding to different phases.

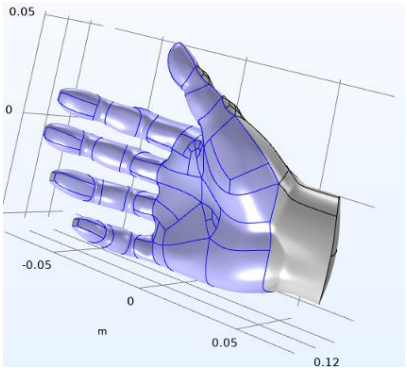
From the observations, it is apparent that the discrepancy between the maximum and minimum Sound Pressure Level (SPL) across the range of tested phases is under 3dB. This relatively small variation suggests a fairly stable performance across the different phase conditions. Additionally, the results reveal that adjacent peaks in the data set are spaced by a distance of 0.48m, equivalent to an angle of 5 degrees. Despite

these changes being noticeable, they are relatively minimal, and as such, may not provide substantial benefit for enhancement, in terms of both magnitude and direction.

Phase Change Regarding Five Locations In 3D.

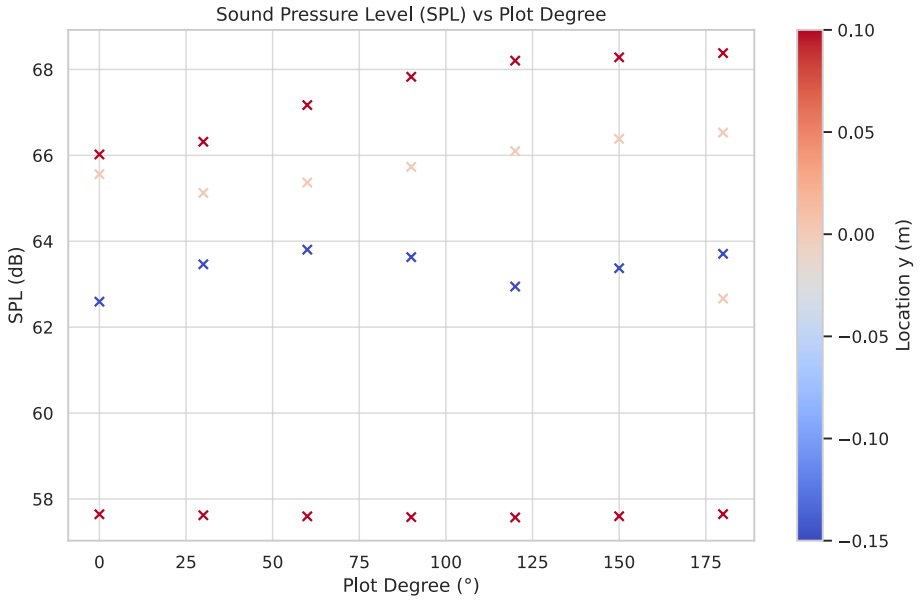
To comprehensively estimate the improvement of phase change, we involved a hand to test the result. The hand is set at 5 different locations, with different heights, different direction with different phase changes to seek any significant difference make by phase. The different situation is shown in the figure.





We adopted the SPL at the inside of palm to test our system behavior.

And the result is evaluated in terms of the average and the largest value of SPL



The findings indicate that there isn't a significant difference in sound pressure levels across all tested locations when different phases are applied. It was observed that at a consistent location, alterations in phase could lead to a change in the sound pressure level of less than 3dB. This observation is consistent with the results obtained from the 2D surface analysis. The average palm result shows less variance, with the difference remaining at around 1dB.

In the context of selecting a phone selection for specific functionalities, the P20 proves to be emerges as a more apt suitable choice for gesture recognition, given its enhanced sound distribution at the top of the speaker. Conversely, the Mate40 is a better fit suited for gesture tracking, owing due to its higher Sound Pressure Level (SPL) at the surface. This scenario highlights underscores the utility value of simulations in facilitating aiding the selection of an appropriate phone model for distinct different purposes.

3.4 Summary

This chapter illustrated a testing of an acoustic sound field simulation for gesture recognition on smartphones. Two smartphone models were built using accurate dimensions and detailed sound components to reproduce the acoustic field. A sensitivity test model was created, optimized, and calibrated, aligning with real factory testing scenarios, using manufacturer-specific parameters. A real-world sensitivity test was performed, followed by evaluations on frequency response, Sound Pressure Level (SPL), SPL difference, and gain. Additionally, three real-world application scenarios were simulated to assess the impact of hand positioning and direction on acoustic reflection. A phase control experiment verified radiation control possibilities. The study offers insights for designing real-time gesture recognition by guiding the selection of phones, operational distance, among other factors.

Chapter 4 Velocity-based Acoustic Gesture Recognition System Based on Smartphones on Smartphones.

4.1 Introduction

In Chapter 3 of the dissertation, the feasibility of gesture recognition applications on smart devices is demonstrated, with the effective area of operation being determined through simulations. This chapter aims to extend this foundation by implementing a real-time gesture recognition system and developing a system for the classification of gesture speeds. Hand gesture recognition system is structured into three parts: (1) Data acquisition. (2) Gesture recognition models. (3) Signal demodulation and gesture speed recognition model.

The data acquisition and signal processing section describe the hardware system's working methodology, including data collection, information extraction from bulk signals, and the detailed procedure of signal design. The gesture recognition models section details how the system achieves recognition functionality with certain accuracy and how interference is reduced during processing. The gesture speed recognition model section explains the methodology for classifying gesture speeds.

A comprehensive performance examination was conducted based on the proposed method. The results show that the system achieves an accuracy of 95.56% in both complex and clean environments.

4.2 Hardware and Software Platform Introduction

The figures explain the main acoustic components used on the smartphone which were acquired from the hardware supplier. The first microphone was used as the receiver, and the second super balanced speaker was used as the transmitter. The amount the microphones and speakers were varied between different smartphone models.

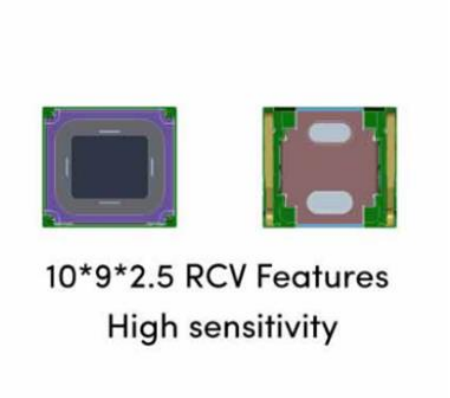


Figure. 37 Hardware RCV

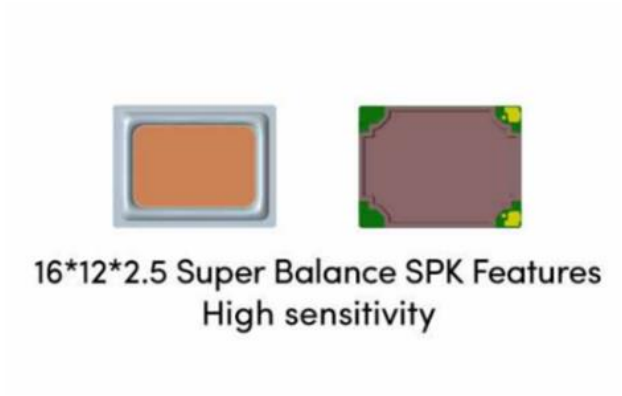


Figure. 36 hardware SPK

Table. 8 Receiver specification

Specification	*Details*
RCV Mode Impedance	(13.2±15%) Ω @2kHz/0.81Vrms
RCV Mode F0	(500±90)Hz@0.81Vrms, in free air
RCV Mode Sensitivity	(122.5±3)dB @0.81Vrms. Type3.2
RCV Mode Listening Test	0.81Vrms, 200~2kHz sweep in 1s
RCV Mode Rated Power	50mW
RCV Mode Max short term power	100mW
SPK Mode Sensitivity	93±3dB@3kHz2.81V, 5cm with baffle
SPK Mode Rated Power	0.6W(2.81Vrms)900~20kHz
Water proof	IPX8

Table. 9 Speaker specification

Specification	*Details*
Impedance	(7±10%) Ω @2kHz, 2.65Vrms
F0	(950±10%)Hz @ 2.65Vrms, in 0.8cc
Sensitivity	(96±3)dB @2kHz, 2.65Vrms, in 0.8cc Baffle
Listening Test	3.5Vrms, in 0.8cc 100-1000Hz sweep in 2s
Rated Power	1W
Max short term Power	1.5W

A couple of key parameters were adopted during the simulation:

(1) Sensitivity: This indicates the amount of voltage outputted when the microphone receives a given sound pressure level in decibels, compared to a reference level.

(2) Impedance, Rated Power, Working Voltage: These parameters indicate the electrical conditions during operation. Any one of these can be derived from the other two.

Additionally, the maximum short-term power represents the highest output power the devices can achieve within a short period. This can determine the maximum voltage in a short time and the maximum sound pressure level via sensitivity.

The reference level for sensitivity is typically evaluated under a standard condition with a transmitter/receiver placed 10cm apart. The sensitivity value is calculated by dividing the received sound pressure by the minimum sound pressure detectable by humans (1dB). In our case, the speaker's sensitivity is 96 dB, meaning the test receiver will receive a sound pressure of 96 dB during operation.

The hardware platform consists of two main parts: acoustic experiment hardware devices and computation devices. For the experiment, we used the main hardware for transmitting signals, receiving signals, processing algorithms, and displaying classification results. The HUAWEI P20 Pro was chosen as the experimental device for this project. As for the computing devices, we utilized a station equipped with an AMD Ryzen 7 5800x and GTX 3070ti, primarily used for machine learning training..

4.3 Data Acquisition and Signal Processing

4.3.1 Purpose and Specification

This section mainly discusses the data acquisition process and the following signal processing process which extracts the hand movement from the background noise. In more detail, the system needs to transmit and receive the signal constantly to capture the movement from hand. When a gesture is performed, the receiver will record the signal during the moving period and save for further process. When the signal is received, the signal processing algorithm will process the signal and analyze the signal into a form that can be processed during the following classification parts. In this section, STFT(Short-Time-Fourier-Transform) and signal demodulation were adopted as the main analysis algorithms.

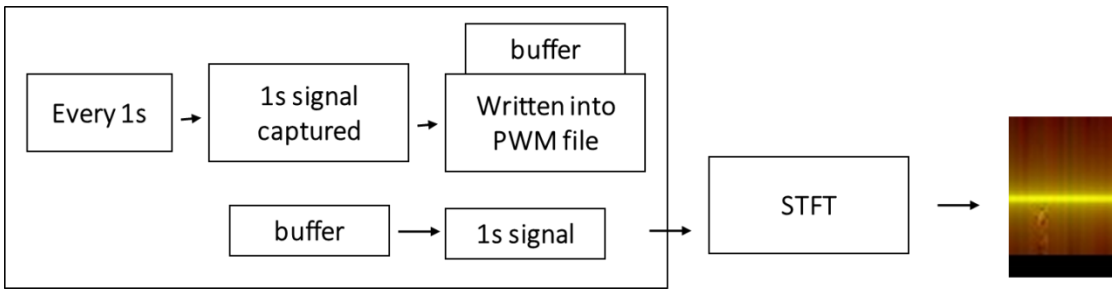


Figure. 38 Flow chart of data acquisition and signal processing

Data acquisition and data processing cannot work at the same time. In this way, firstly the signal was saved into a PWM file and extracted from PWM when processing. The sampling rate was set as 44100 and 50 extras were left for redundancy. Then the STFT algorithm was processed to get the time frequency spectrum. The input of the system was 1 sec of received signal, which was 44100 samples of PWM file, and the output of the system was a bitmap format STFT file.

4.3.2 Choice of Selection

The Purpose of the Short time Fourier transform is to evaluate the signal's frequency composition in terms of short time aspect. The Short-time Fourier transform (STFT)[23] is a Fourier-related transform used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time. In practice, the procedure for computing STFTs is to divide a longer time signal into shorter segments of equal length and then compute the Fourier transform separately on each shorter segment. This reveals the Fourier spectrum on each shorter segment. By adopting STFT, the frequency change, influenced by moving hand can be captured at the certain time intervals.

There two states of art time frequency analysis methods were mainly used in academia: short time Fourier transform, Continuous waveform transform[67]. Short time Fourier transforms uses equivalent sampling length according to the time interval and do Fourier transform along the time axis, which makes the same sampling rate at different frequency. The Continuous waveform transform uses adaptive sampling length along frequency axis, which means signal with higher frequency will lead to

longer sampling interval and lower frequency will have a longer time interval. In our cases, only the high frequency of the spectrum will be adopted for gesture recognition, which means the long-time analysis along lower frequencies are introducing extra time consumption. Therefore, Short-time-Fourier-transform is adopted for our method.

4.3.3 Movement under STFT spectrum

In these cases, the movement of hand will induce a frequency change to the transmit signal, the speed of the gesture will be presented as the magnitude of frequency change.

$$\Delta f = \frac{v_h}{c} f_c \tag{29}$$

Where Δf is the changed frequency, v_h is the speed of hand, c is the speed of sound, f_c is the frequency of the transmitted signal.

STFT: The usual mathematical definition of the STFT [23] is

$$x_m(\omega) = \sum_{n=-\infty}^{\infty} x(n)\omega(n - mR)\mathrm{e}^{-j}\omega n = DTFT_{\omega}\left(x \times SHIFr_m R(\omega)\right) \tag{30}$$

Where $x(n)$ is input signal at time n, $\omega(n)$ is length m window function, $x_m(\omega)$ is the

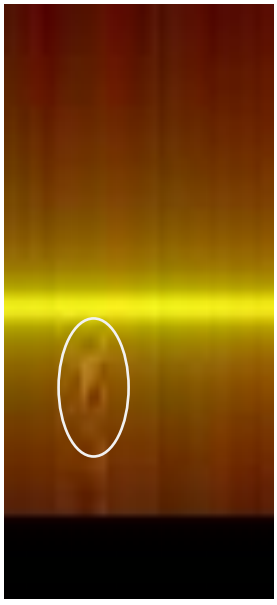


Figure. 39 STFT of Backward gestures

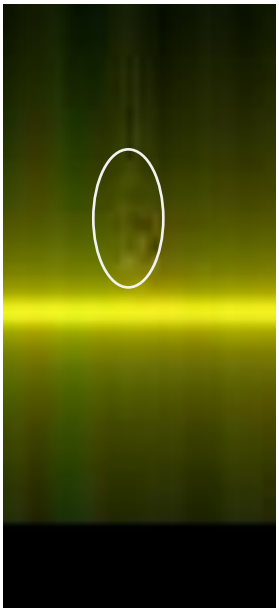


Figure. 40 STFT of Forward gestures

DTFT windowed data centered about time mR, and R is the hop size, in sample, which

between successive DTFTs.

The figure shows the result of a STFT spectrum of “Forward” and “Backward” gesture. The vertical axis indicates the frequency range, and the x axis indicates the time intervals. As shown in the Figure, if a forward gesture is performed, the frequency of received signal will be increased, and the whole spectrum will have a peak in positive direction. If a backward gesture is performed, the frequency of signal will show a decrease trend in frequency axis. The magnitude of the peak indicates the speed of the gesture, which means a faster speed will lead to a higher peak. For specific set up, a 18k sine wave continuously was transmitted and 2048 was adopted as the windows length for STFT. Frequency range was set from17000 Hz to 19000 Hz which cover all the possible movement range.

4.4 Gesture Recognition Module

The purpose of this module is to design a classifier which can take use of the time spectrum obtained from STFT and classify different signal into gestures. The spectrum represents the frequency change obtained from the received signal. When a movement approaching the receiver, the frequency will increase as the effect of doppler effect, and on the contrary, the frequency will decrease when a hand is moving away from the phone. The classifier was designed to extract the frequency changes automatically and classify the frequency into corresponding gestures. In this project, CNN (Convolutional neural network) is used as the classification approach. CNN is one type of neural network with successful applications such as image recognition, computer vision and natural language processing. By adopting a convolutional layer, the input feature will be extracted independently by kernels. Through training, those kernels can eventually extract features that represent input features and ignore undesired noise.

Training dataset sampling:

To restore as many possible scenarios as possible, three volunteers were invited to perform gestures for our training dataset. To enhance the diversity and robustness of the dataset, we selected three volunteers, comprising two males and one female.

Furthermore, these volunteers were instructed to execute gestures in various environments and to perform these gestures randomly. This approach was adopted to ensure maximum variation in the gestures. Volunteers were asked to perform the gesture standing at different locations and using different speeds. For our cases, 1200 samples per gesture were captured for further training.

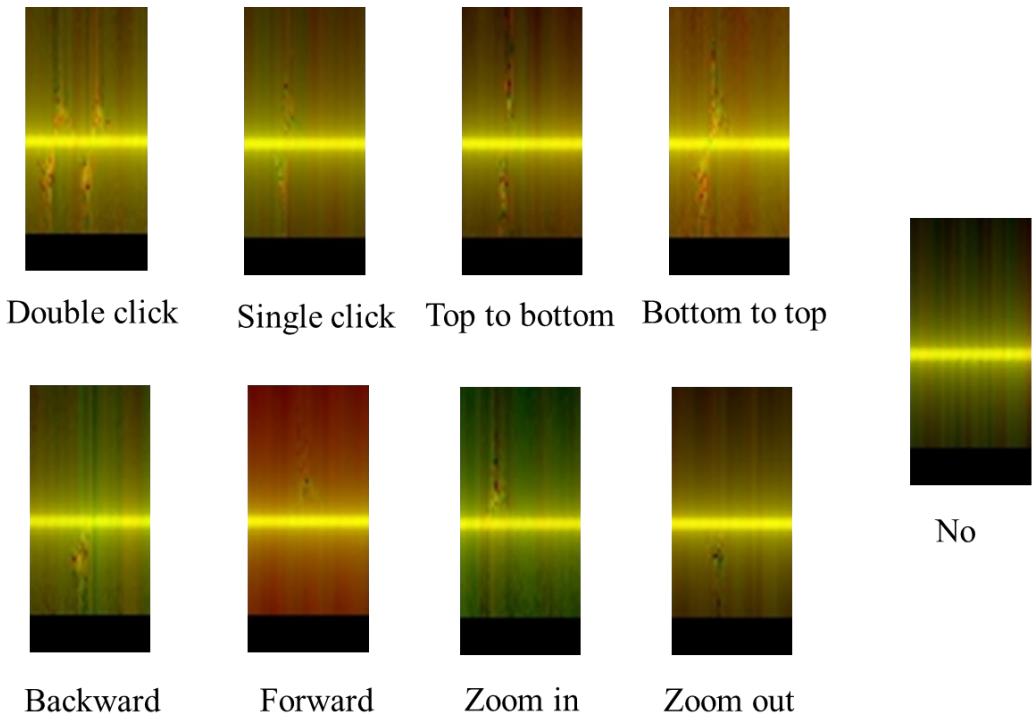


Figure. 41 STFT spectrum examples

The figure shows below are a set of STFT spectrum from smartphone.

The below figure shows the structure that adopted in recognition system. When the STFT is captured by previous step, a Resnet 18 CNN classifier is adopted to classify

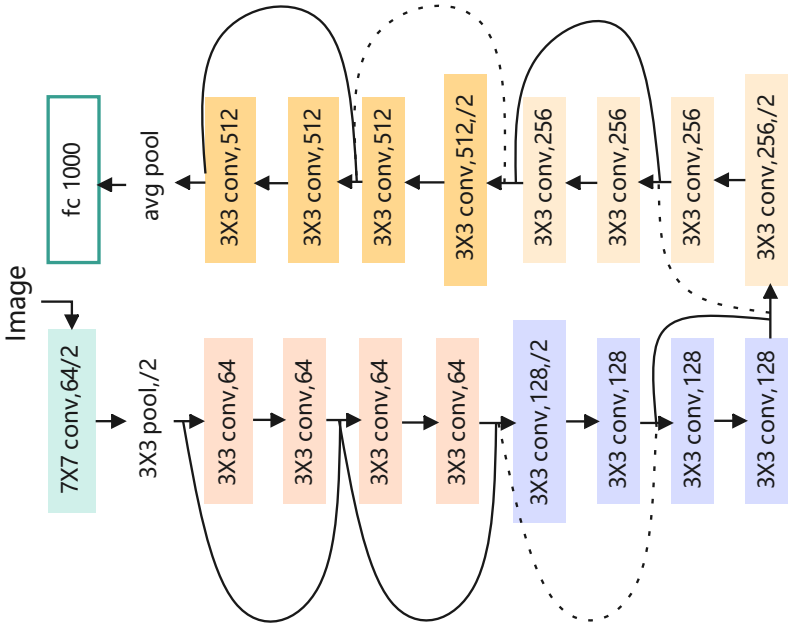


Figure. 42 Flow chart of ResNet 18

our images. The size of the STFT bitmap is set at 44*77 for the purpose of reducing the computational complexity purpose. To implement on smartphone, Torchscript was adopted to transfer Pytorch CNN structure into Java file which can be processed on android devices. The input of this system is the bitmap format STFT spectrum, and the output is classification score which can be transferred into classes.

The figure above shows the 8-fold classification result from the fine-tuned network.

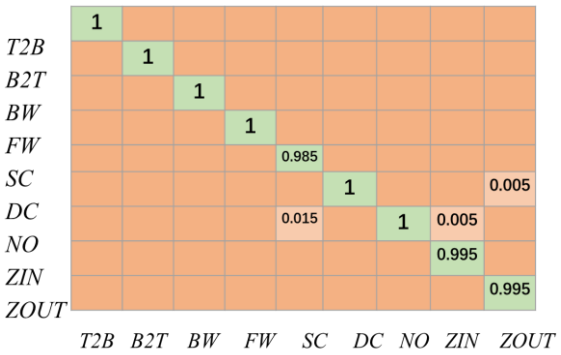


Figure. 43 Classification result from module (Normalized in vertical direction)

The overall accuracy is over 98%.

4.5 Signal Demodulation.

This section was developed to extract the speed information of the gesture performed from the received signal. The signal demodulation was to extract the movement speed from the background noise signal and the gesture speed recognition model was to classify different speed into three levels.

Assuming the received signal is represented as $s_{rec}(t)$:

$$
\begin{aligned}
S_I(t) &= S_{rec}(t) \times \sin(-2\pi ft) \\
&= \frac{1}{2}\sum_{k=1}^{P} A_k \left(\sin\left(-\frac{2\pi f d_k(t)}{c} - \theta_k \right) \right) + C_Q \\
&+ \frac{1}{2}\sum_{i=1}^{N} A_i \left(\sin(4\pi ft - \theta_i) \right) \\
&- \frac{1}{2}\sum_{k=1}^{P} A_k \left(\sin\left(4\pi ft - \frac{2\pi f d_k(t)}{c} - \theta_i \right) \right) \\
S_{lowQ} &= \frac{1}{2}\sum_{k=1}^{P} A_k \left(\sin\left(-\frac{2\pi f d_k(t)}{c} - \theta_k \right) \right)
\end{aligned}
\tag{31}
$$

$S_I(t)$ represents the signal after demodulation $d_k(t)$ represents the sound path's change due to the hand movement, f indicates the central frequency which is set as 18000 in our case. s_{LowQ} represents the Q component of the demodulated signal processed by a low pass filter. For the equation above, a same frequency sine signal is multiplied to the received signal. In this way, the original signal can be split into multiple signals with different frequencies, which are $\left(-\frac{f\, d_k(t)\prime}{c}\right)$,$2f$, $2f - \left(\frac{f\, d_k(t)\prime}{c}\right)$. The average frequency change caused by hand movement is below 300Hz, which means $\left(-\frac{f\, d_k(t)\prime}{c}\right)$ is lower 300Hz, For the terms $2f$and $2f - \left(\frac{f\, d_k(t)\prime}{c}\right)$, The values are around two times of the central frequency, which is 36kHz. By adopting a low pass filter, the s_{LowQ} can be extracted from the received signal which only includes doppler frequency caused by hand movement.

In the similar way, the I component can be obtain by multiply a cosine wave:

$$S_I(t) = S_{rec}(t) \times cos2\pi ft$$

$$S_{lowI} = \frac{1}{2}\sum_{k=1}^{P} A_k \left(\cos\left(-\frac{2\pi f d_k(t)}{c} - \theta_k \right) \right) \tag{32}$$

By combining the two components together, the sound path's change caused by hand can be obtained, which indicates the movement of hand.

$$d_k(t) = -\frac{c}{2\pi f} * \left(\arctan\left(\frac{\sum_{k=1}^{P} A_k \left(\sin\left(-\frac{2\pi f d_k(t)}{c} - \theta_k \right) \right)}{\sum_{k=1}^{P} A_k \left(\cos\left(-\frac{2\pi f d_k(t)}{c} - \theta_k \right) \right)} \right) + \theta_k \right) \tag{33}$$

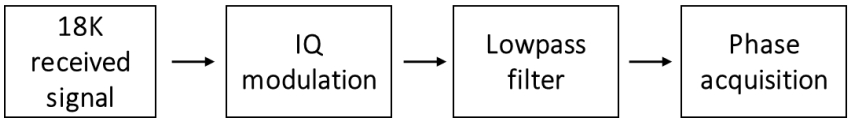


Figure. 44 Signal demodulation flow chart

The above figure shows the process of signal demodulation. The whole process of signal demodulation processes the signal into a phase change which caused by hand movement. IQ modulation is used to split the signal with different frequencies which including the changed phase. The process of lowpass filter is used to split the signal apart. An arctan function is used to transform signal into phase change. The whole process is process in android system where the input is the raw signal from PWM file, and the output is a series with phase change.

The three figures show the result of the whole algorithm when a single click is presented. The first one shows the received data directly from the microphone. It exhibits an erratic fluctuation due to background noise which cannot observe any useful information. The second figure shows the IQ demodulation result from the raw data, it clearly shows the interval of the movement and the phase change along with time axis. The third figure exhibits the phase result extracted from previous IQ demodulation; the movement distance is expressed at the figure.

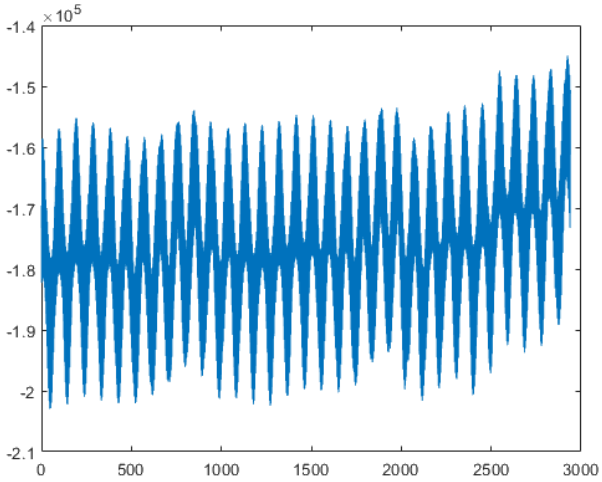


Figure. 46 Raw data of received signal.

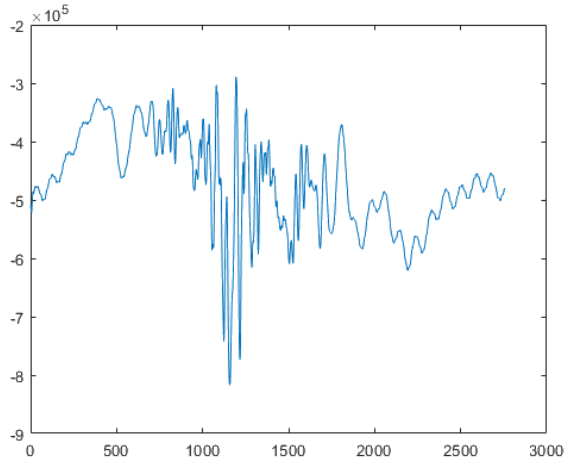


Figure. 45 Signal after IQ modulation

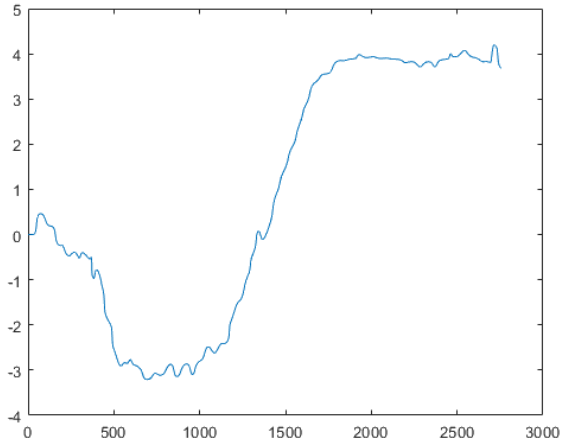


Figure. 47 Signal phase change

4.6 Extreme-Variance Noise Removal Algorithm

The purpose of the extreme-variance noise removal algorithm is to eliminate background noise from stationary objects. In the phase calculation from the demodulated signal, the constant term introduces a consistent phase offset. This offset can lead to continuous calculation errors that may increase or decrease over time. Additionally, background noise, reflections from stationary objects, and multipath reflections from other objects also contribute to unavoidable noise that results in calculation errors.

The concept behind the extreme-variance noise reduction algorithm is based on Empirical Mode Decomposition (EMD). EMD is a data adaptive method that can decompose a signal into physically meaningful components by continuously comparing the envelope of extreme values. In this case, extreme values are used to estimate the peak of the phase wave, and variance values serve as a threshold to estimate background noise.

The extreme-variance noise removal algorithm consists of three parts: 1. Variance calculation, 2. Extreme finding, and 3. Background noise vector update.

Variance Calculation Part:

We use the percentage of variance value as a threshold to determine if a local peak exists. If the difference between two peaks is greater than three times the signal's variance, it is considered a local peak. The value and location of this local peak are saved into a vector.

Extreme Finding:

The method for finding local extremes is straightforward: a value larger or smaller than the preceding and following samples is considered a temporal peak. Once all temporal peaks are identified, a determining algorithm based on the variance value is used to identify local extreme values.

Background noise vector updated:

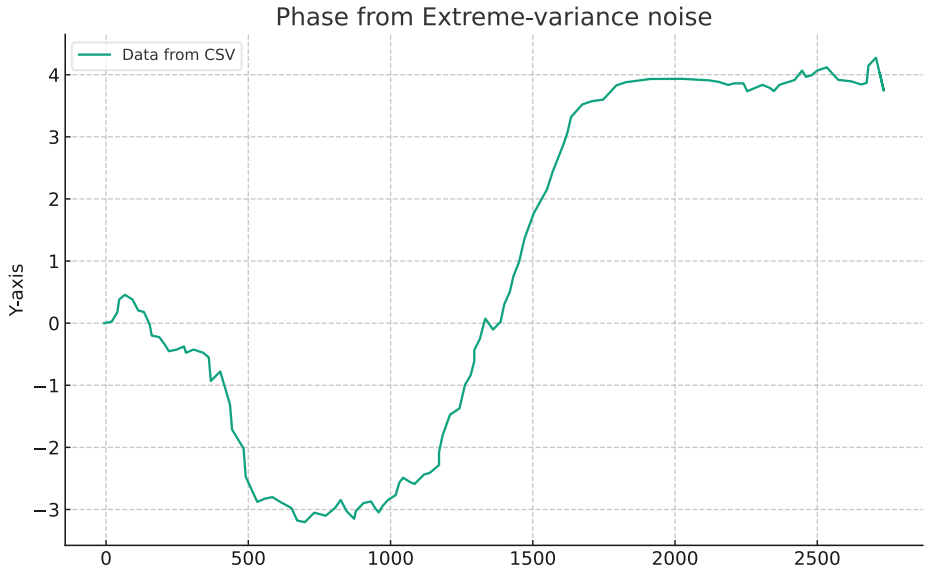


Figure. 48 Phase from Extreme-variance noise

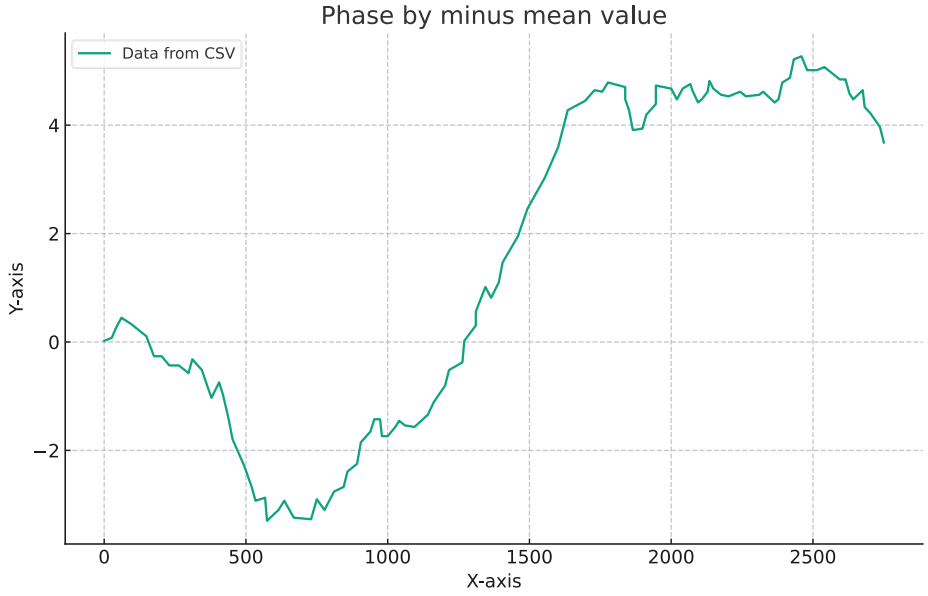


Figure. 49 Phase by minus mean value

When all local extremes are collected, an average summary algorithm was adopted to acquire the moving background noise. As the phase signal is supposed to be varied around zero, the value between two successive peaks should be the background noise plus zero, which is zero in our cases. Therefore, the background noise was extracted by taking the mean of successive peaks and eliminated from the original signal.

4.7 Gesture Speed Recognition Model

The purpose of the speed recognition module was to convert the hand's movement distance into speed. Ideally, the hand's speed should be directly calculated from the movement distance by dividing it by the time taken. However, in practical scenarios, the movement distance (represented as distance vectors) is affected by uncertain noise, leading to disordered fluctuations in the distance vector. As a result, the distance vector cannot be directly transformed into velocity. The primary function of the speed recognition module is to extract as much useful information from the distance vector as possible and to minimize the impact of noise. In our project, we employed the segmented sum average method and utilized signal compression to reduce the influence of noise.

The velocity extraction algorithm comprises two parts: 1. Signal Segmentation Sum Average and 2. Signal Compression.

The equation below shows the first part of the algorithm:

$$V(p) = \sum_{i=100*P}^{n+100*P} \left(s\left(i+n\right) - s\left(i\right) \right) \tag{34}$$

$$P = \frac{L}{n} \tag{35}$$

Where V is speed vector, v_P is the pth element of the vector. L is the whole length of the received signal, n is the size of each element, s represents the distance vector. The maximum value of P is shown above.

For the signal compression, equal-spaced compression is adopted:

$$V_C\left(q\right) = \frac{\sum_{i=1}^{p/N} \left(V\left(i\right)\right)}{N} \tag{36}$$

Where V_C represents the compressed speed vector, N is the length of each compression segments. In this way, the signal will be shortened and compressed to avoid noise.

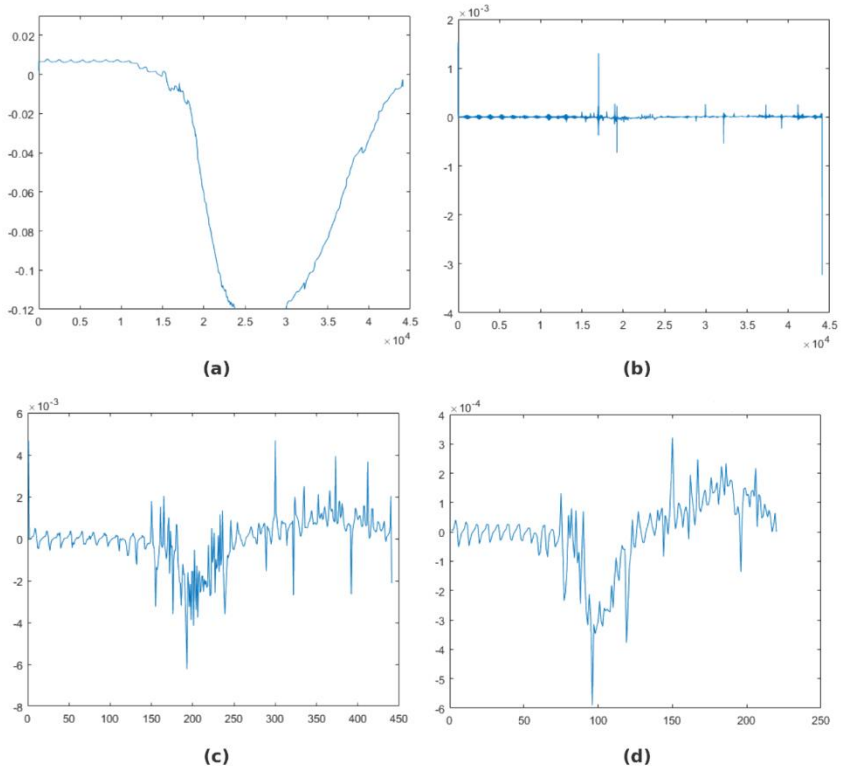


Figure. 50 Segmentation sum average comparing to other algorithms.

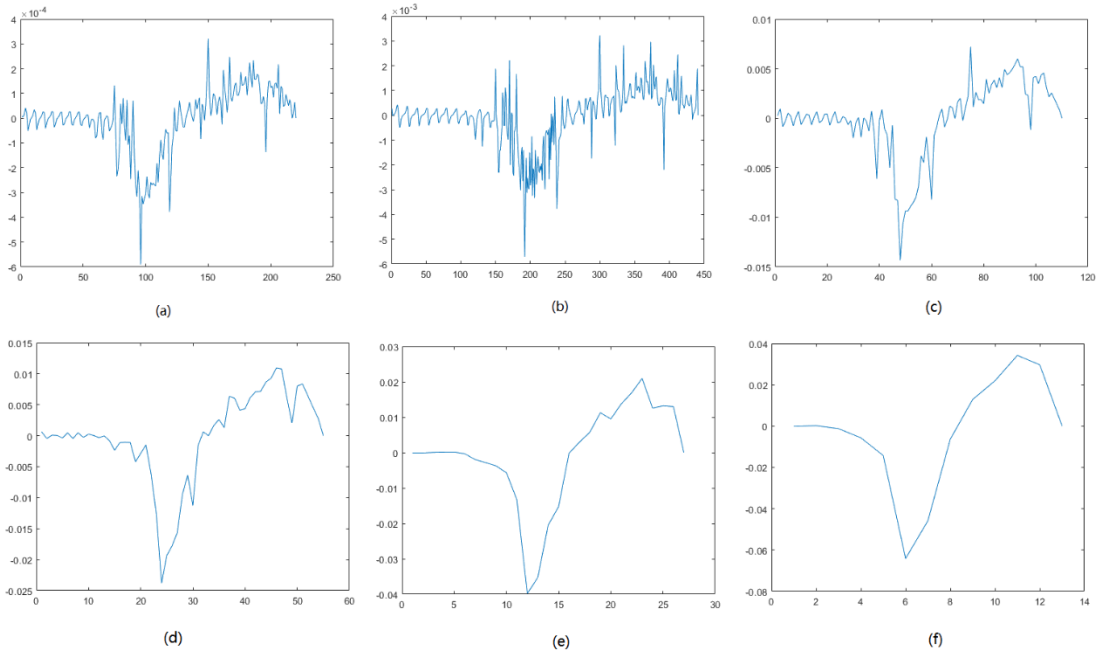


Figure. 51 Speed vector with different compression interval

4.8 Speed Levels Classification

In this section, we discuss the speed levels classification algorithm, which is primarily designed to categorize the speed of performed gestures. By employing this algorithm, the number of distinguishable gestures can effectively be tripled, as each gesture can be classified into three distinct speed levels. For instance, a 'Forward' gesture could be categorized as 'Slow Forward,' 'Mid Forward,' and 'Fast Forward.' We assessed our system's performance using three different algorithms: the Threshold method, the Speed-only CNN method, and the Speed with Gesture 1D-CNN method. Ultimately, the Speed with Gesture CNN method was chosen for speed recognition and is the focus of this section.

A 1D CNN (one-dimensional Convolutional Neural Network) is our main approach for speed classification. A 1D CNN is a specific type of CNN where one of the filter sizes match the corresponding size of the input data. The convolution in the Convolutional Layer operates in only one direction. In our case, the speed vector is a 1-dimensional time-series signal that varies over time. This characteristic makes the 1D-CNN particularly adept at extracting features from speed changes along the time axis. The first part of speed recognition is to transfer phase information into a compressed velocity vector. The phase information is calculated using the signal demodulation process, which has been already stated above. Speed extraction algorithm to reveal the movement speed of the hand afterwards as shown in the flow chart.

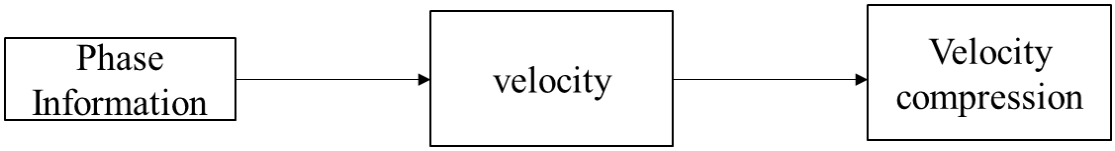


Figure. 52 velocity compression flow chart

The second part of the speed recognition module is the classification module,

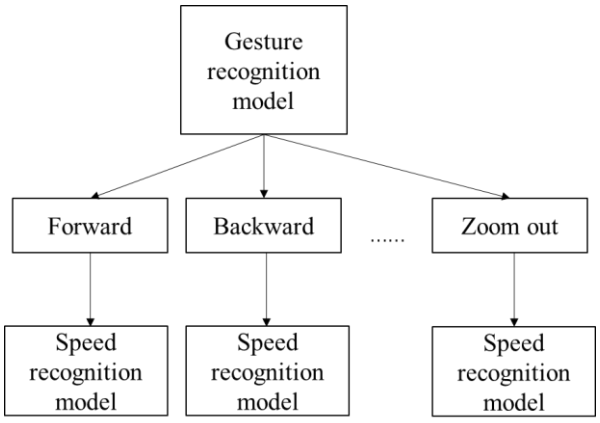


Figure. 53 Flow chart of speed recognition

which is shown in the flow chart. We developed 8 speed classification modules to suit speed difference from different gestures. We implement our speed module after the gesture model get the classification result. The corresponding speed model will be implemented to classify the right speed. During operating, all the module is loaded when the program starts. The overall process time for speed classification module is around 220ms.

4.8.1 Modeling result

Regarding to different gestures, we developed a corresponding 1d CNN. And the Specific structure configuration and accuracy is shown in the figure below.

Table. 10 Neural network structure configuration and result

	T2B	B2T	BW	FW	SC	DC	NO	Zoom-In	Zoom-Out
1D-CNN	2CNN+3FC	3CNN+3FC	3CNN+3FC	3CNN+3FC	3CNN+3FC	2CNN+3FC	N/A	2CNN+3FC	2CNN+3FC
Accuracy	0.99	1	1	0.99	1	1	N/A	1	1

The table presented above displays illustrates eight distinct neural networks, each tailored for recognizing specifically designed to recognize one of eight specific unique

gestures to ensure, thereby ensuring efficient classification speed. The results indicate demonstrate that each network is well-designed crafted and functions effectively functional for its respective designated gesture. The initial step in speed recognition involves converting phase information into a compressed velocity vector.

4.9 Full Functional Test Result and Summary

4.9.1 Gesture Recognition Results

To assess the comprehensive performance of our system, we conducted evaluations in multiple environments with a variety of users. The participants were randomly selected and had no prior exposure to the program. Regarding the environments, we tested in both complex and clean settings. In the complex environment, various obstacles were deliberately placed around the phone to simulate realistic conditions.
Two volunteers were invited for our evaluation test. The volunteers were firstly given

instruction on how the gestures are performed and 5 mins were given for given for them to practice. They were asked to perform each gesture 5 times in one environment.
The results are shown in the following table.

A "clean environment" is described as a space where no obstacles are found within the detection area, which extends up to 1 meter. In this setup, the target devices are positioned horizontally on a surface. On the other hand, a "complex environment" involves the presence of various obstacles located around the area, including both hard obstacles like bottles and walls, as well as soft obstacles such as curtains.

Table 11 Experiment Result

	T2B	B2T	BW	FW	SC	DC	NO	ZOOMIN	ZOOMOUT
Complex environment	0.95	0.95	0.9	0.95	0.95	1	1	0.9	1
Clean environment	0.9	0.85	1	0.9	0.95	1	1	1	1

The accuracy for complex environment achieves 95.56% in both complex environment and clean environment.

Performance summary

Comparing with the state-of-arts, firstly, our system has a higher accuracy to distinguish the similar gestures. The Microsoft team developed a pioneering gestures recognition program using laptop's built-in acoustic devices[22] in 2012. They used fast Fourier transform (FFT) to determine frequency change and derived a thresholding based peak searching method that achieved five one-dimensional gestures classification with 94% accuracy. Dolphin [37] achieved a 93% recognition accuracy among 6 similar gestures using FFT in 2014. In 2016, W. Ruan et al. proposed AudioGest, a gesture sensing system based on smartphone[38]. They introduced an innovative feature extraction process to decode time-frequency spectrum into velocity vector based on Doppler effect. 94.15% accuracy with 6 gestures was achieved. We achieved a 95.5% accuracy gesture classification system which is 1.35% higher than the state of art accuracy.

Secondly, our system has a lower computational cost which can properly operate without any computational help with other devices. Yanwen Wang [28]proposed a

gesture classification system that can detect fist movement. It requires high computation ability and the whole system is working on a cloud computation platform. In our project, we only require 600ms for each 1 second signal and it is able to work on off-the-shelf smartphone devices. For further computation improvement of phone, our system can achieve a better performance.

4.9.2 Speed recognition Results

Table. 11 Test results for two volunteers

Volunteer 1	B2T_Slow	B2T_Mid	B2T_Quick	BW_Slow	BW_Mid	BW_Quick	T2B_Slow	T2B_Mid	T2B_Quick	ZoomIn slow	ZoomIn mid	ZoomIn quick	NO
Accuracy	0.9	0.8	0.8	1	1	1	0.8	0.9	1	1	1	1	1
	FW_Slow	FW_Mid	FW_Quick	SC_Slow	SC_Mid	SC_Quick	DC_Slow	DC_Mid	DC_Quick	ZoomOut slow	ZoomOut mid	Zoomout quick	Ave
	0.9	1	1	1	1	1	1	1	1	1	1	1	0.956

Volunteer 2	B2T_Slow	B2T_Mid	B2T_Quick	BW_Slow	BW_Mid	BW_Quick	T2B_Slow	T2B_Mid	T2B_Quick	ZoomIn slow	ZoomIn mid	ZoomIn quick	NO
Accuracy	1	1	0.8	1	0.9	1	1	0.9	1	1	1	1	1
	FW_Slow	FW_Mid	FW_Quick	SC_Slow	SC_Mid	SC_Quick	DC_Slow	DC_Mid	DC_Quick	ZoomOut slow	ZoomOut mid	Zoomout quick	Ave
	0.9	1	1	0.9	1	0.8	1	0.8	1	1	1	1	0.96

Similar with the result in gesture recognition, we invited two volunteers in our experiment to test the system accuracy, each volunteer was asked to perform 10 gestures for each speed and accuracy is shown above.

As far as we know, there is few research was carried out to defining hand movement's speed in gesture recognition. To evaluate our system performance, we adopted two traditional speed classification approaches for comparison.

Gesture classification using STFT with Threshold method.

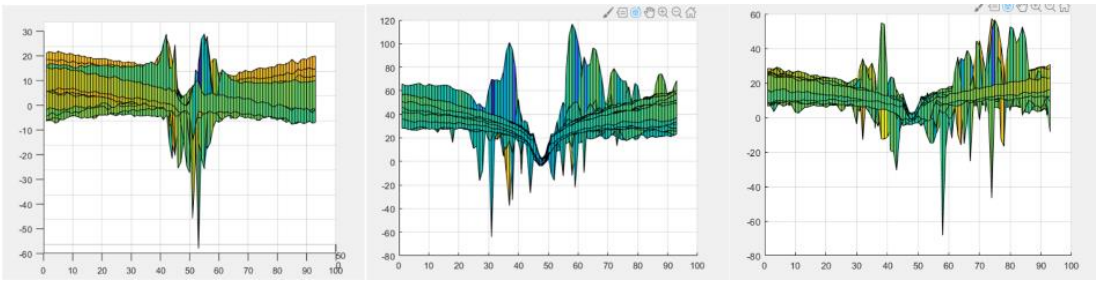


Figure. 56 Normalized STFT for slow a), middle, b) fast c) speed.

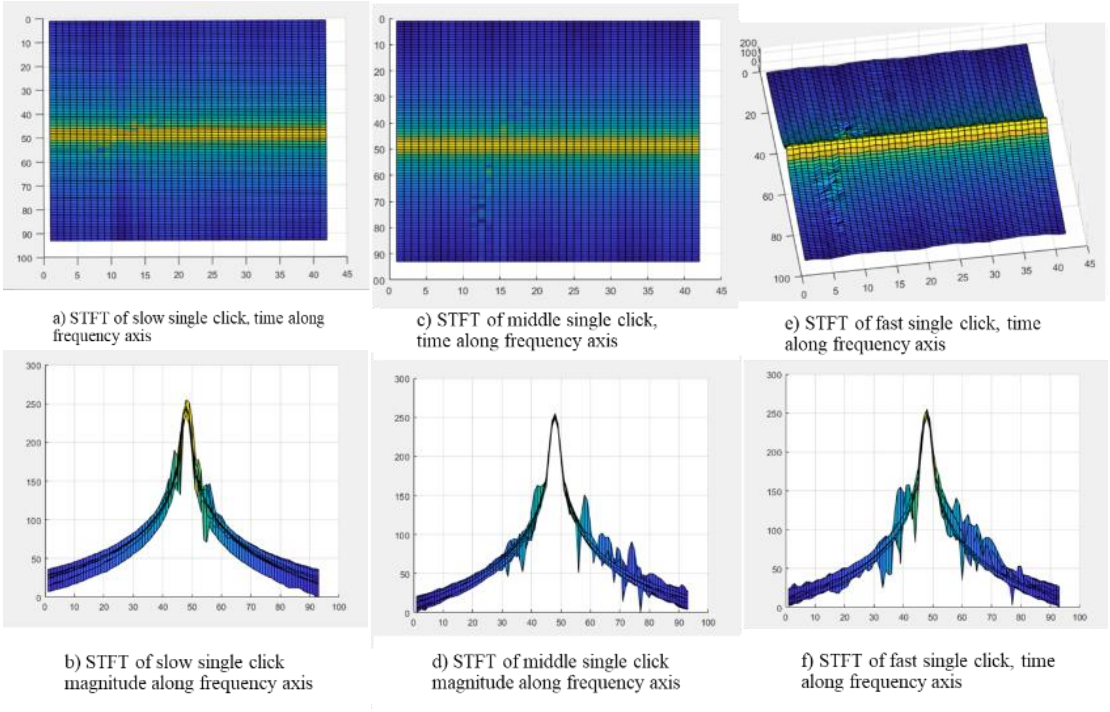


Figure. 57 STFT result for different speed gesture.

As mentioned in the previous section, hand movement introduces an additional frequency shift to the received signal due to the Doppler effect. The greater the speed of movement, the larger the frequency shift. Consequently, an intuitive method that evaluates the shifted frequency should theoretically be able to determine the movement speed. To validate this hypothesis, we conducted an experiment to analyze the STFT spectrum of a moving hand, as depicted in the figure.

The results clearly indicate differences in the frequency peaks for slow and medium-speed single clicks, as shown in parts b) and d) of the figure. For instance, in d), more than three peaks (marked in red) exceed 60 frequency bins, while in b), no peak surpasses 60 bins. This suggests that a threshold method could effectively distinguish speed levels based on the location of frequency peaks in the STFT values Comparing the STFT of middle speed and fast speed, the peak difference is not distinguishable by eyes as they have a similar number of high peaks. To solve this problem, To mitigate noise, the spectrum data was normalized, resulting in more pronounced peak differences and a greater number of unique peaks exceeding 80.

A threshold determining algorithm was developed to differentiate the number of peaks. In this algorithm, a certain percentage of the highest value in the frequency domain is defined as the threshold. Values exceeding this threshold are identified as peaks. Two frequency bin thresholds were established to distinguish between speeds, as detailed in the table. The first row of the table represents the threshold factor, which determines if the STFT contains a peak, expressed as a percentage of the highest value. The second row indicates the determining frequency bins, which are used to ascertain the gesture's speed type. For example, if an STFT spectrum of a 'Top to Bottom' gesture shows a peak at 45 bins and no peak higher than 65, then the gesture is classified as a middle speed. A validation test is carried out and the system achieves an overall 75% accuracy.

Table. 12 Threshold setting for different gestures.

	T2B	B2T	BW	FW	SC	DC	NO	ZI	ZO
Threshold factor	0.55	0.38	0.47	0.38	0.35	0.38	N/A	0.43	0.25
Frequency bins	38/65	35/65	20/70	35/65	40/65	35/65	N/A	27/65	36/65

Table. 13 Validation result

	T2B	B2T	BW	FW	SC	DC	NO	ZI	ZO
Fast	8	10	10	9	9	10	N/A	9	9
Mid	7	8	9	9	8	7	N/A	6	8
Slow	6	5	6	6	7	4	N/A	4	5
	22/30	23/30	25/30	24/30	24/30	21/30		19/30	22/30

The average accuracy is lower than our result.

Speed only classification approach

The second idea designed a classifier to classify both gesture and velocity at the same time. After fine tune the algorithm, the training accuracy reaches 98%. However, when test the performance in practice, the performance decreased dramatically as

shown in the figure. For comparison. Our Speed with gesture 1d-CNN method achieve higher accuracy and more robust in unknow environment.

Table. 14 System performance in different environments

Know environment

	B2T_Slow	B2T_Mid	B2T_Quick	BW_Slow	BW_Mid	BW_Quick	T2B_Slow	T2B_Mid	T2B_Quick	NO
Accuracy	0.8	0.8	1	0.9	0.7	0.8	0.8	0.9	0.7	1
	FW_Slow	FW_Mid	FW_Quick	SC_Slow	SC_Mid	SC_Quick	DC_Slow	DC_Mid	DC_Quick	AVE
	0.9	0.8	0.8	0.8	0.9	0.9	1	1	1	0.86

Unknow environment

	B2T_Slow	B2T_Mid	B2T_Quick	BW_Slow	BW_Mid	BW_Quick	T2B_Slow	T2B_Mid	T2B_Quick	NO
Accuracy	0.6	0.8	1	0.6	0.5	0.8	0.7	0.7	0.9	0.9
	FW_Slow	FW_Mid	FW_Quick	SC_Slow	SC_Mid	SC_Quick	DC_Slow	DC_Mid	DC_Quick	AVE
	0.8	0.6	0.7	0.6	0.6	0.9	0.8	0.6	1	0.78

4.10 Summary

The section aimed to develop a real-time gesture recognition and speed classification system, implemented in three stages: data acquisition, gesture recognition models, and signal demodulation for speed recognition. The data acquisition phase covered the hardware functioning, signal procurement, information extraction, and signal design. The gesture recognition models focused on achieving precise recognition while minimizing processing interference. The gesture speed recognition model elucidated the speed classification process. Comprehensive performance tests showed a 95.56% accuracy in recognizing gestures, both in complex and clean environments.

Chapter 5 Fast Gesture Recognition System with Temporal Neural Network

5.1 Introduction

Problems of high latency and cross-frame recognition failure are encountered in gesture recognition systems based on convolutional neural networks (CNNs). These issues are largely inevitable due to the nature of gesture features and the continuity of multi-gesture sequences. For real-time functionality, systems utilizing CNNs need to take an overlap interval from incoming signals for each processing cycle. However, when processed by STFT (Short-Time Fourier Transform), two consecutive frames may contain identical information. Consequently, if a gesture sequence involves multiple 'Forward' or 'Backward' movements in succession, the gesture feature from the previous frame might be erroneously recognized as a feature for the new gesture. This redundancy in feature consideration can influence the output results of the neural network, leading to cross-frame recognition failure.

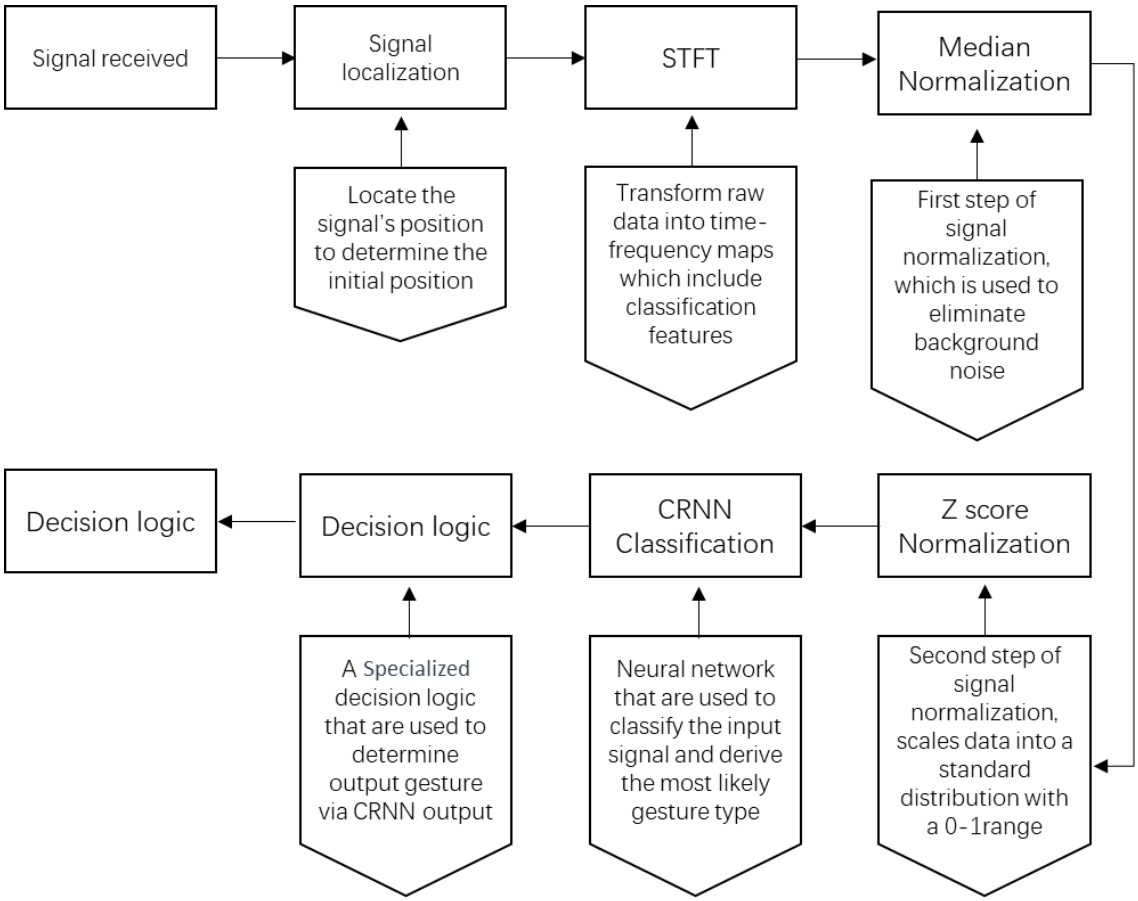
Typically, it takes a user between 0.7 to 1 second to perform a gesture. Conventional CNN-based systems must wait for the entire necessary time period before beginning processing, which means a wait time of 1 second is required for each processing cycle. As a result, the whole system needs to wait at least 1 second before yielding a classification result. Even though a gesture might be completed in a shorter time, such as 0.3 seconds, the system is compelled to wait for the full second to ensure all relevant information is captured. This delay results in a latency of 0.7 seconds in the cited case.

In Chapter 5, a fast gesture recognition system with temporal neural network is presented. To keep the effective feature extraction function, convolutional neural network is adopted as the feature extraction part of the neural network. To introduce temporal functionality, LSTM[68] is utilized to connect the before and after information. In this chapter, as a follow-up study of the work in Chapter 4, a fast reaction algorithm

with short period of reaction time is designed. The reaction time reduces to 0.2s by using the proposed algorithm and the problem of cross frame failure has been mitigated. A normalization algorithm is designed to tackle the background noise issue. The performance is evaluated by numerals parameter setting and the result is compared with other temporal algorithms.

5.2 System Overview

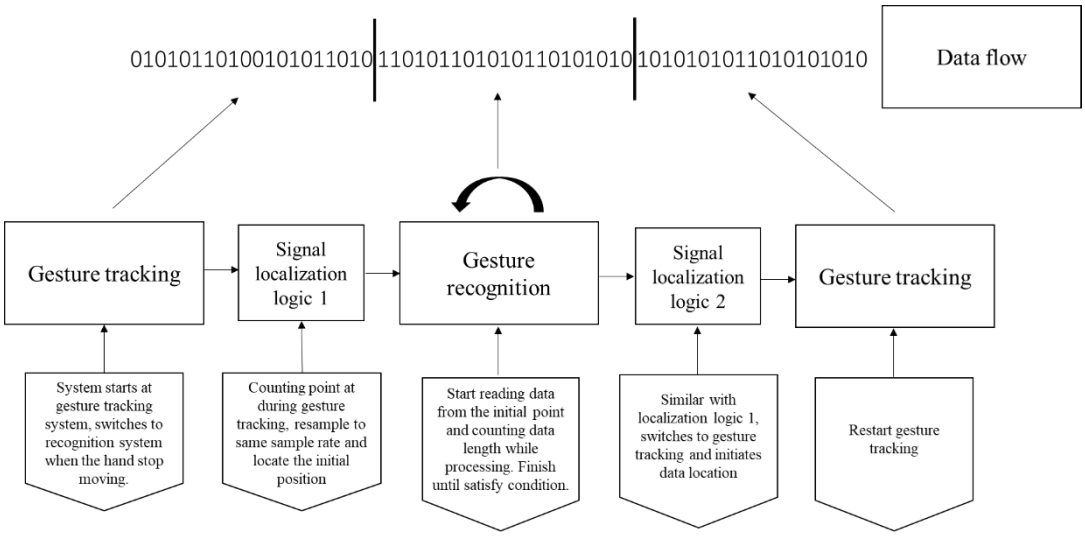
The entire system can be seen in Figure 59. The whole system is developed under python platform. Data acquisition cards were used as the signal collection platform and independent sensors were adopted as transmitter and receiver. Firstly, when the receivers receive signal from users' hand, the DAQ acquisition card captures the signal



and stores the data into a data buffer. The signal was continuously transmitted to the processing PC. Secondly, the signal was localized and identified by a signal localization

algorithm. The initial position of each gesture recognition period was determined by the algorithm and then forwarded to the next stage. Thirdly, a modified STFT algorithm is developed to capture the feature of each gesture, the signal is sliced into short period and a feature map is determined as the input vector.

Fourthly, normalization processes are designed to avoid irrelevant periods and reduce background noise. Two algorithms are developed: a median normalization algorithm to reduce background noise and unneeded frequency bins and a Z-score normalization algorithm to centralize and standardize the input feature map. A separate section below will provide more details on this aspect. Fifthly, a temporal convolutional neural network is designed as the classifier. The network adopts CNN as the main feature extractor and RNN as the temporal information extractor which are combined as a CRNN model. It generates the recognition result from the input feature map and outputs a probability table. Finally, a decision logic is designed to optimize output result. Chapter 5.3 illustrates the real-time processing system and the function of signal localization algorithm. Chapter 5.4 illustrates the STFT processing algorithm and the process of frame selection. Chapter 5.5 illustrates the normalization algorithms and impact of each algorithm. Chapter 5.6 illustrates the temporal convolutional neural



network. Chapter 5.7 illustrates the decision logic and the system performance. Chapter

5.8 illustrates the discussion of system robustness and experiments of system performance under interference.

5.3 Real-time Processing System and Signal Localization.

The real time system combines gesture tracking and gesture recognition. In this thesis, only the gesture recognition system will be discussed. It is good to know how the two systems are combined and here is a brief introduction of the whole system.

Gesture tracking and gesture recognition are shared with the same data flow collected from the sensor. When the system starts, gesture tracking logic will start operating. When the hand stops moving, a decision logic will be triggered, and the gesture recognition system will be activated. A data location logic is designed to initiate the starting point when the transition logic is activated. The two systems adopt different sample rates: gesture tracking adopts 44100 sample/second and gesture recognition adopts 96000 sample/second. When the transition logic starts, the start data point is recalculated. It is used to ensure the location of data is at the same point when the gesture recognition period starts. Figure 60 illustrates the process of signal localization.

In the realm of real-time applications, there has been limited discussion about the impact and significance of system responsiveness. Nonetheless, this is a topic worthy of exploration. In practical scenarios, a typical gesture execution ranges from 0.5 to 1 second. Therefore, it's crucial for a system to respond within 1 second to effectively capture all features and provide a timely response. A reaction time longer than this can hinder the system's efficiency, causing it to lag. Conversely, a shorter reaction time demands a more advanced classification method capable of swiftly capturing all relevant features. Our proposed system boasts a reaction time of just 0.2 seconds, significantly surpassing the standard 1-second benchmark. Additionally, our temporal classification method successfully captures all features comprehensively.

5.4 Convolutional and Temporal LSTM Neural Network.

Convolutional neural network has been demonstrated to possess effective feature extraction capabilities. However, when dealing with temporal problems, it is unable to consider the continuity of information.

In chapter 5, it has been demonstrated that CNN networks have strong feature recognition capabilities. If the feature recognition capability can be retained while adding a temporal aspect to the neural network, then the issue of inability to connect preceding and succeeding information can be resolved. Taking the above considerations into account, a network based on CNN and LSTM has been designed.

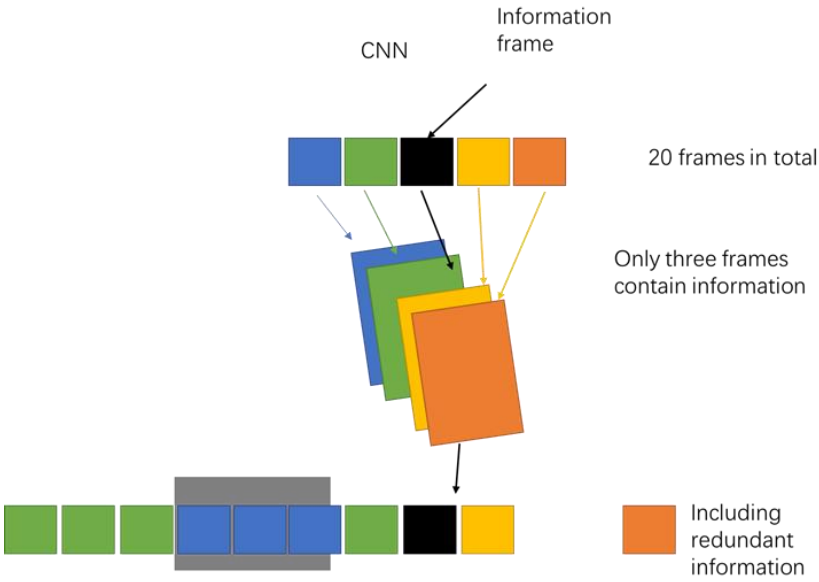
For detailed explanation, the minimal time interval length must be longer than the slowest gesture's time. It means that each interval must have the ability to contain all information. Any information outside the interval will be ignored. A gesture usually took 0.2-0.5s to perform and 0.7s will be the longest time for each gesture. And considering the overlap problem which the gesture might not start from the beginning of every internal, 1second is a proper time for a decent accuracy. However, if there is another approach which can slice every time interval into a short piece and connected all result will largely alleviate this problem.

RNN is famous for connecting the signal context, in our cases, it can take all slices' information and calculate gesture depending on the information from multiple frames.

Figure 61 shows a demo logic for our system. The signal will be sliced into a short period of time (0.2s). Every time new data is generated (0.2s), the neural network will be activated. It takes the new data (STFT map) (squares at the first row in figure) to the CNN first, which will get a feature map for 0.2s (rectangle at second row). Every time a feature map is generated, it will be passed into the RNN network for classification, and also it will be stored for later frame's use. When the RNN network receives the feature map, it will take three frames from previous frames and one upcoming frame to

form a combination and calculate the connection insides via LSTM. In this way, the neural network will obtain one output result for every frame (0.2s) but containing information for 1s.

5.5 Introduction for LSTM



5.5.1 Purpose and Specification

LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) that is commonly used in deep learning applications. Unlike traditional RNNs, LSTM uses a unique architecture that includes memory cells controlled by input, forget, and output gates, allowing it to store and retrieve information over long periods of time without suffering from the vanishing gradient problem. Through the use of gate mechanisms, LSTM can selectively store or discard information, making it well-suited for tasks that require learning and predicting long-term dependencies. LSTM has been successfully applied to various tasks, including natural language processing, speech recognition, and time series analysis, particularly those that involve modeling long-term dependencies or processing sequential data.

5.5.2 State-of-arts and choice of selection

Recent state-of-the-art LSTM models have incorporated various improvements, such as attention mechanisms, multi-layer structures, and sophisticated gating mechanisms, to enhance the performance of the model. One popular variant is the Bidirectional LSTM (BiLSTM)[69], [70], which processes input data in both forward and backward directions to capture dependencies from both directions. Other variants include the Gated Recurrent Unit (GRU) and the Convolutional LSTM (ConvLSTM)[71]–[74], which incorporate convolutional layers into the LSTM architecture for better processing of spatial-temporal data. LSTM-based models have been successfully applied to various tasks, including natural language processing, speech recognition, image captioning, and video analysis.

5.5.3 Mathematical principle

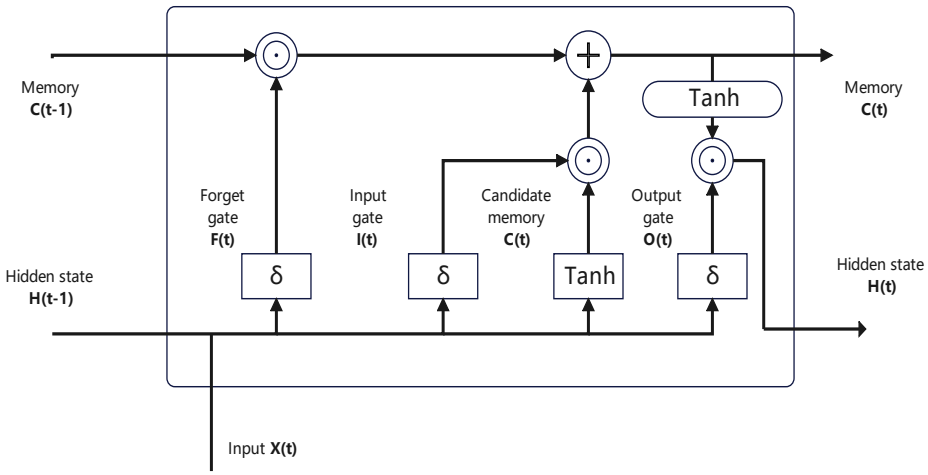


Figure. 61 LSTM principle

LSTM: Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that can process and predict time series data by selectively retaining or forgetting information over long periods of time. LSTM is adopted as our RNN model.

The mathematical principle behind LSTM involves using memory cells and gating

mechanisms to selectively store or discard information over time.

At each time step t, the input x_t is first processed by the input gate $r_t$and a sigmoid activation function to determine how much of the new input to let through. Then, the forget gate $z_t$determines how much of the previous cell state to forget. The new cell state candidate is then computed by applying a hyperbolic tangent activation function to a combination of the input x_t and the previous hidden state h_{t-1}. Finally, the output gate x_t determines how much of the updated cell state to output as the current hidden state h_t.

By using these gating mechanisms, LSTM can selectively store or discard information from previous time steps and thus avoid the vanishing gradient problem that occurs in traditional RNNs. This allows LSTM to capture long-term dependencies and perform well on tasks such as natural language processing and speech recognition.

CRNNs [75]–[77] have been widely used and studied in various applications and have achieved state-of-the-art performance in several tasks. Here are a few examples:

Music analysis: CRNNs have also been used to analyze music, particularly for tasks such as music transcription and genre classification. For example, in the recent MIREX Music Transcription task, CRNN-based approaches achieved the highest performance, beating other methods such as CNNs[78], [79] and RNNs.

Overall, CRNNs have demonstrated their effectiveness and versatility in handling sequential data and have achieved state-of-the-art performance in several applications.

5.6 Comparison with CNN Based Network.

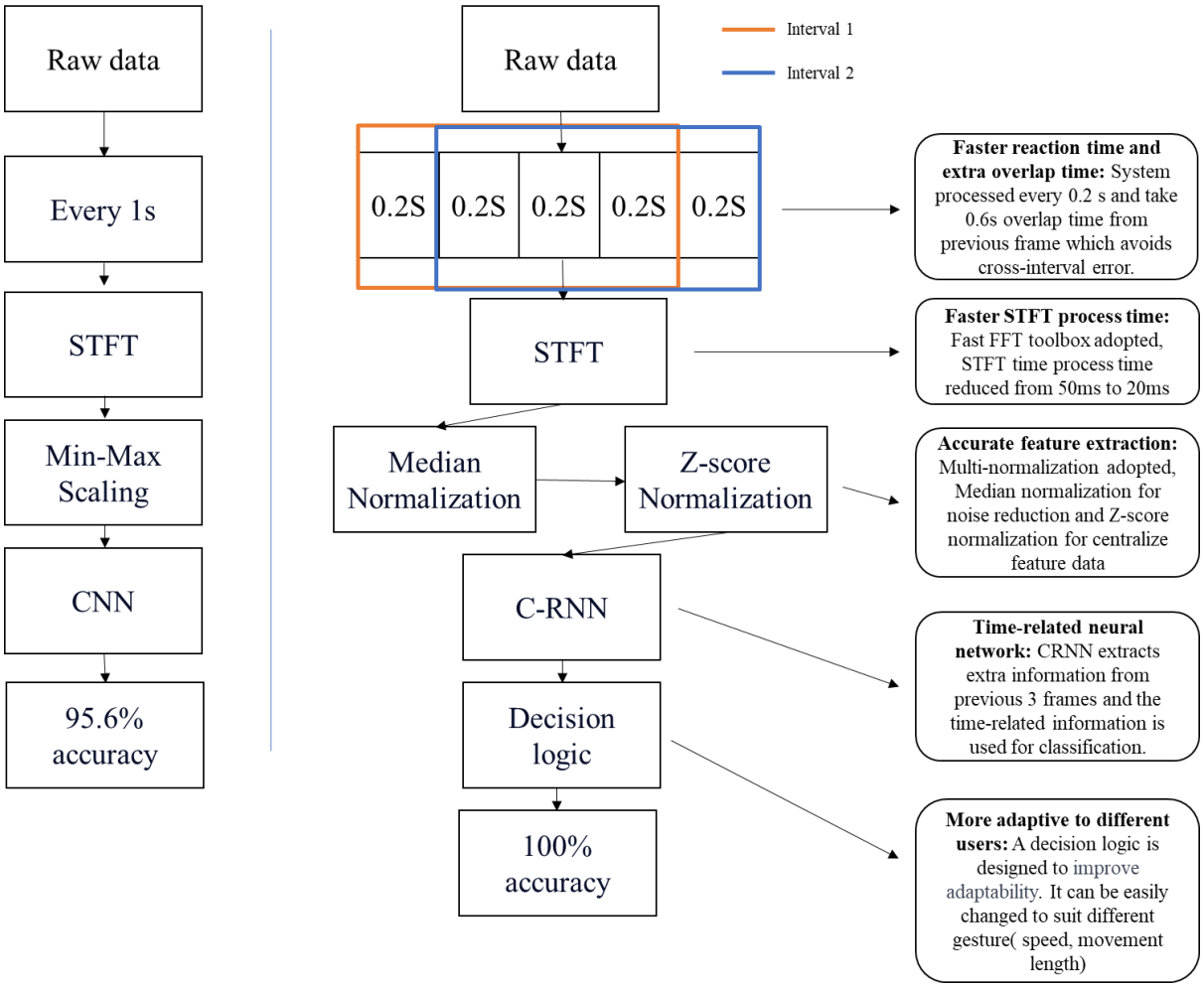


Figure 63 illustrates a logic diagram for CNN project and CRNN project. Comparing with CNN project. CRNN project has a faster reaction time, more overlap interval in every process circle (0.6s's data). The fast reaction time is achieved by shorter operation circle and the system processes every 0.2s. For STFT, CRNN project uses a faster STFT algorithm, which reduces processing time from 50ms to 20ms. Moreover, CRNN project uses a more comprehensive normalization method and a time related CRNN neural work, which improves the accuracy rate to 100% percent and reduces the reaction time to 0.2s. The decision logic is designed to suit different gestures (speed, movement length), it can be easily adjusted for different gestures performed (different

user).

5.7 Modeling Result

Table. 15 CRNN paramete

Neural network	
CNN	Resnet 18
FC1	256
FC2	128
RNN	2 Layers LSTM
FC	256

The neural network parameter is shown in the table.

The dataset includes 6 types of data types, and each includes 400 samples. The classification result is shown in the figure. The average accuracy for gestures reaches 100%. The training process is validated through 4-fold cross validation and the training best model is selected as the real-time operation model.

A performance review is being carried out to validate the accuracy of our system. The overall system keeps 100% accuracy in both no interference and wind, walking interference.

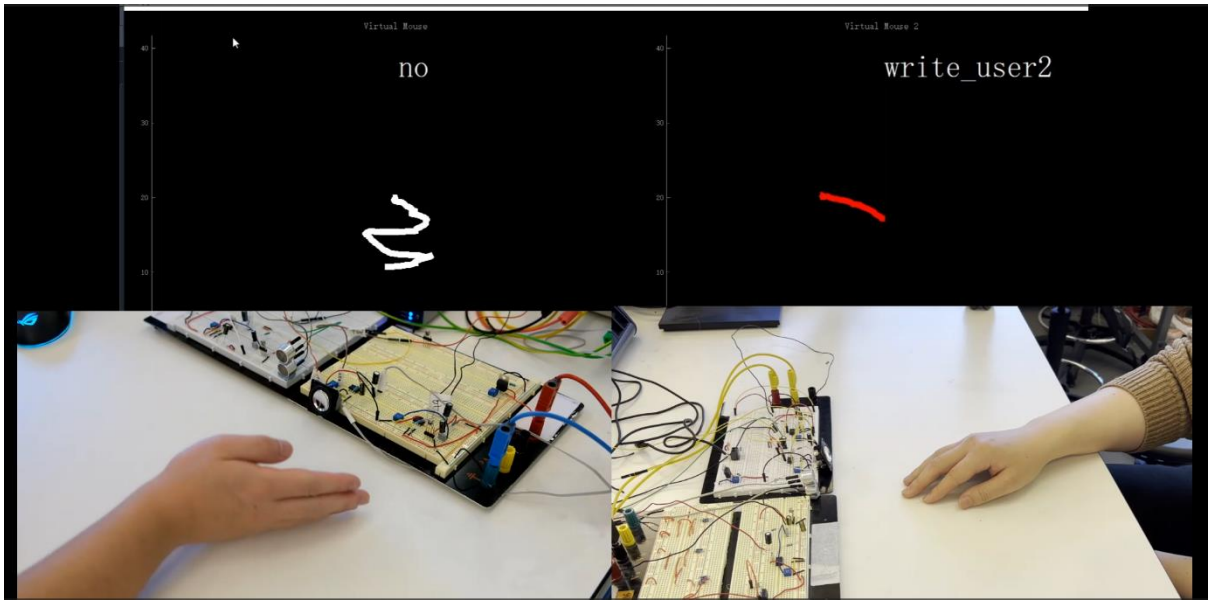
Table. 16 Performance evaluation with interference

	Forward	Backward	Forward backward	Wave hand	Click	No	All
Data collection environment (No interference)	29/30	30/30	29/30	29/30	29/30	30/30	176/180 97.8%
Wind interference	29/30	30/30	28/30	29/30	30/30	30/30	176/180 97.8%
Pedestrian interference	28/30	30/30	29/30	29/30	29/30	30/30	175/180 97.22%
Accuracy in Interference Environment	95.0 %	100%	95%	96.7%	98.3%	100%	97.55%

By comparing with existing system, our system outperforms in both standard environment and interference environment. In standard environment, the second-year system achieved 97.8 % accuracy which outperform Dolphin and AudioGest 44.8% and 2.88 % respectively. In interference environment, second year system outperform the first-year system and AudioGest by 4.55% and 22.5%. In addition, the second-year system can operate under real-time condition with 0.2s reaction time.

Table. 17 Performance evaluation between the start of art

	Fast reaction recognition	Multi-wave[80]	Dolphin [37]	AudioGest [38]
Average accuracy In stable environment	97.8%	96%	93%	95%
Average accuracy In interference environment	97.5%	93%	Not given	94%
Reaction time	0.2s	Not given	Can not working in real time	Not given
Platform	Individual sensor system	Smartphone	Smartphone	Smartphone



5.8 Multi-User Gesture Interaction System

The system developed a multi-user function to allow two users to perform gestures and hand tracking at the same time. Each user has their own recognition system and tracking system, which includes hardware and software system. For hardware, every user has one set of ultrasound sensors for gesture recognition and one set of acoustic sensors for hand tracking. For software, every user has their individual software that is adjusted accordingly with background and hardware.

5.8.1 System Breakdown and Analysis

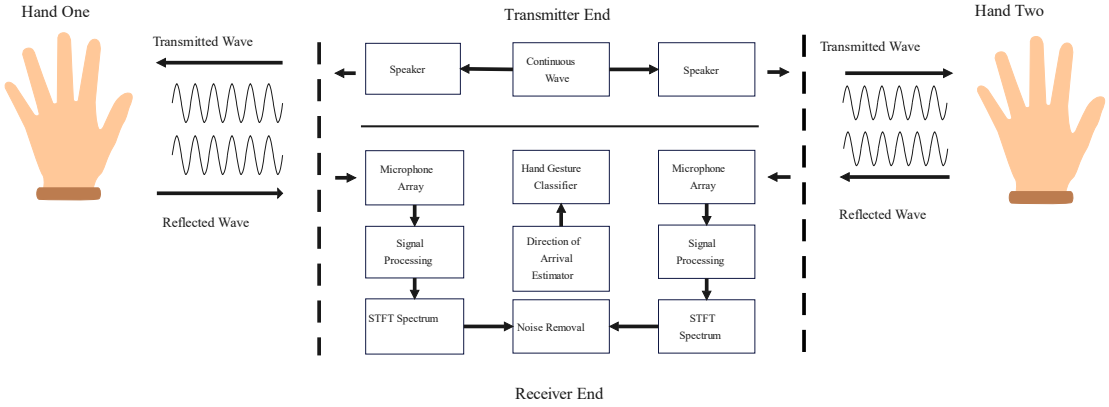
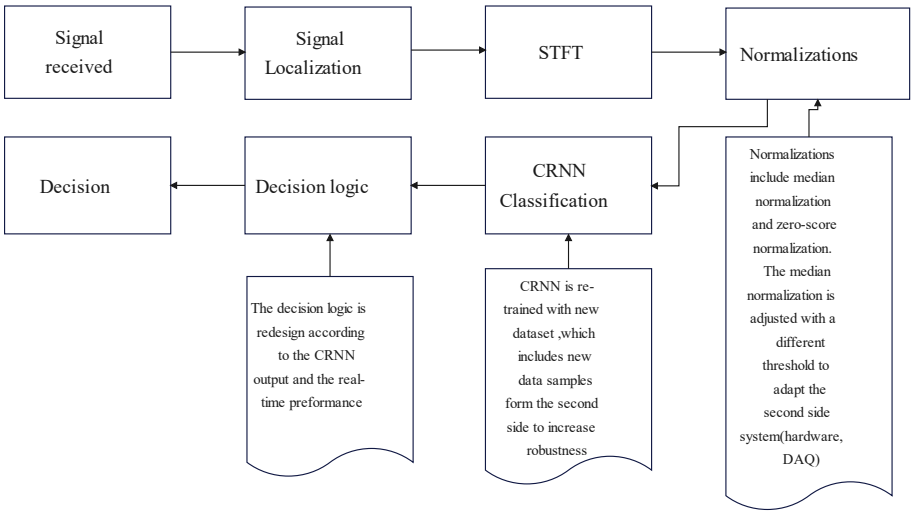


Figure. 65 Overall system flow chart

Multi-user system shares a similar technique with the single user but having a different classification network at end. On one side, an inaudible 20kHz acoustic continuous wave as carrier signal is transmitted via a speaker and the hand is performing gestures. The acoustic wave propagates in air and reflects from the hand, and it is captured by a transmitter. The received signal is evaluated via signal processing, spectrum extraction, irrelevant signal removal, CRNN classifier and a final hand decision logic. The gesture type will be derived from the logic decision algorithm. On the other hand, the signal is running through the same process, but a different classifier and decision logic. It is mainly because of the hardware difference which leads to different STFT performance.



As shown in Figure 66 above, the multi-user system process is similar to the single side's process. Both systems share the same STFT and signal localization technique. However, the normalization, neural network and decision logic are redesigned. In median normalization, the threshold for noise confirmation is reset to adjust with the new hardware that keeps all useful information and removes unnecessary noise. A new dataset is rebuilt by combining the newly collected dataset with old dataset and the new CRNN is built based on the new dataset for better performance. Finally, the decision logic depends on the output CRNN is created.

55.8.2 Performance

The performance is evaluated for both sides separately.

Table. 18 Multi-user system performance

	Forward	Backward	Forward backward	Wave hand	Click	No	All
Multi-user Side one	28/30	28/30	29/30	30/30	30/30	30/30	180/180 97.2%
Multi-user Side two	30/30	22/30	21/30	28/10	30/30	30/30	53/60 88%

Side one retains the similar recognition accuracy as the single side system which around 97.2%. Side two have an overall accuracy of 89.4%, which Backward gesture and Forward gesture is more likely to fail.

5.9 Summary

In this chapter, a novel acoustic gesture recognition system based on temporal convolutional neural network was present. By adopting Short Time Fourier Transform (STFT) as input feature, the convolutional neural network and recurrent neural network were combined to classify the performed gesture. For this goal, the temporal information and feature information of gestures have been combined to solve the recognition latency and cross frame problem. The whole system achieves a fast reaction time with 0.2 second and a 100 percent recognition accuracy among 6 gestures. In comparison with CNN only neural network, it was concluded that CRNN network was a more effective algorithm for real-time continuous gesture input system. A Multi-user system user interference was achieved for multiple user's input.

In chapter 6, The idea of enriching gesture recognition with new feature will be demonstrated below.

Chapter 6 Acoustic Gesture Recognition System Based on Angle of Arrival and Doppler Effect

6.1 Introduction

This section investigates the feasibility of incorporating angle information into acoustic gesture recognition. Angle information, serving as a novel gesture feature, enables the system to accurately determine the direction of the performed gesture. By incorporating angle information, the capturing accuracy was enhanced, and a more comprehensive understanding of hand gestures was obtained.

This chapter presented a comprehensive study on acoustic gesture recognition systems, incorporating both angle and Doppler effect. The study covered various aspects including sensor design, mathematical proof, neural network design, and performance evaluation.

Firstly, a novel acoustic sensor array consisting of one transmitter speaker and four receiver microphones was proposed. This sensor array serves as the foundation for the gesture recognition system.

Secondly, a framework for the acoustic gesture recognition system was developed, utilizing a data acquisition card and Doppler effect. This framework enabled the acquisition and processing of gesture data effectively.

Thirdly, a feature extraction system was designed, which employed the modified Multiple Signal Classification (MUSIC)[52], [81]–[83] algorithm for extracting angle information, and the Short-Time Fourier Transform for extracting velocity information.

Lastly, a novel 1D-convolutional neural network [84]–[86]classifier was proposed, which integrates both velocity and angle information. The structure and modelling of the neural network were thoroughly analysed.

The performance of the proposed multi-feature gesture recognition system was compared with traditional single-feature networks. The final system achieved an impressive recognition accuracy of 94.9% across 10 different gestures and was capable

of distinguishing gestures from two different directions.

6.2 Design and Modeling of Hardware and System Overview

The structure of the active sensing system is depicted in Figure 67. A continuous wave, operating at an inaudible frequency of 20kHz, was emitted from a speaker during the performance of hand gestures. This acoustic wave propagated through the air, interacting with the hand and leading to a reflected wave. This reflected wave was then captured by a uniform linear array of sensors. Detailed information about the hardware configuration of this system is discussed in Section 6.2.1.

In order to enhance the quality of the captured signals, the removal of irrelevant signals and subsequent signal processing techniques were described in Section 6.2.2. The extraction of the velocity spectrum was illustrated in Section 6.2.3.

The algorithm for estimating the direction of arrival and addressing the angle inversion problem was presented in Section 6.2.3. This algorithm allows for the determination of the angle at which the gesture was performed.

Finally, the gesture classifier was explained in Section 6.3. This classifier utilizes the extracted features and the information regarding the direction of arrival to accurately classify the performed gestures.

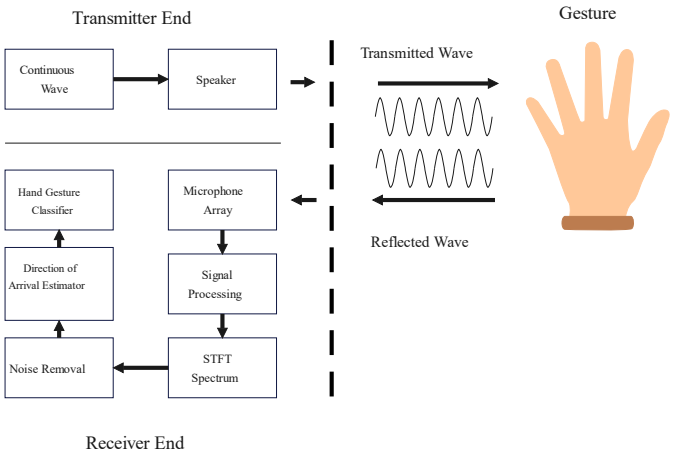


Figure. 67 System schematic

6.2.1 Hardware Configuration

In active sensing system, designed to capture the spatial information of hand gestures, was constructed using a uniform linear array. This array included four microphones and a speaker. The system was configured to acquire the acoustic echo signal, as depicted in the figure. A continuous acoustic signal with a frequency of 20kHz, commonly supported by most off-the-shelf devices, was generated by a digital-to-analog converter (DAC). After amplification, this signal was transmitted through an omni-directional speaker. During gesture performance, the microphone array captured the reflected signal, which was then amplified and transferred to an analog-to-digital converter (ADC). The ADC operated at a sampling rate of 160kHz, recording each gesture for a duration of 2 seconds.

The experimental hardware setup was depicted in Figure 68, providing a visual representation of the system components.

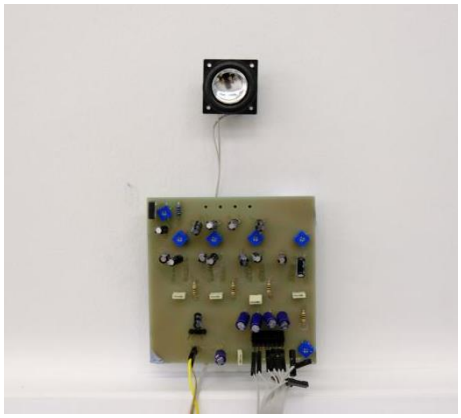


Figure. 68 Linear array hardware

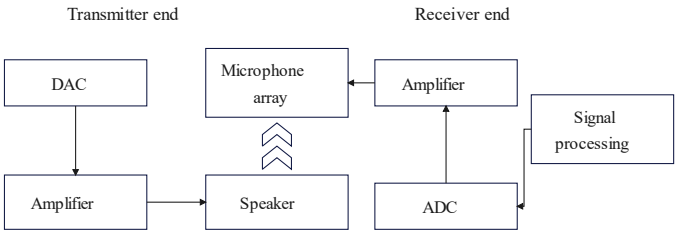


Figure. 69 Hardware schematic

6.2.2 Moving interval.

Focusing on STFT feature map, the frequency change mainly lied in a frequency range from 19.5kHz to 20.5kHz (speed less than 8.5m/s), to avoid the influence of low frequency noise on recognition, only a part of the spectrum sub-band was saved as a

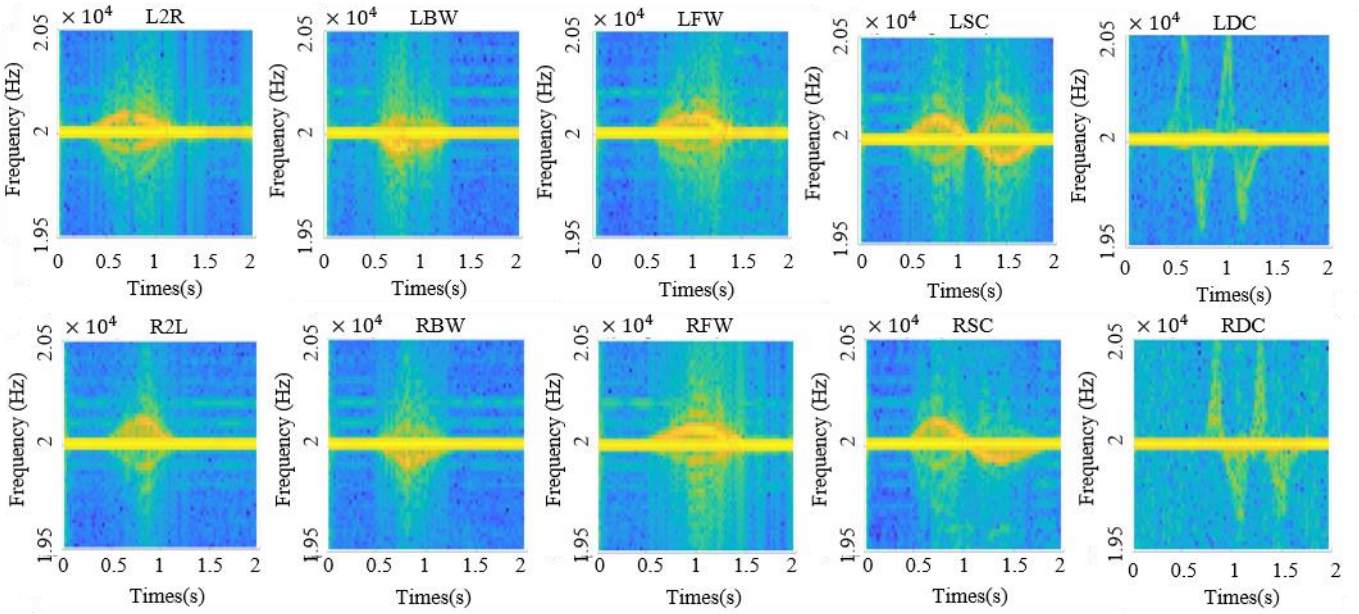


Figure. 70 Time-frequency spectrums of ten different gestures. (1. L2R: Left to right, 2. LBW: Left backward, 3. LFW: Left forward, 4. LSC: Left single click, 5. LDC: Left double click, 6. R2L: Right to left, 7. RBW: Right backward, 8: RFW: Right forward, 9. RSC:

matrix behind each image shown in . The size of the matrix was 52×77, which contained 52 frequency bins and 77 time bins.

6.2.3 Angel of Arrival (AoA) detection

Angle mask

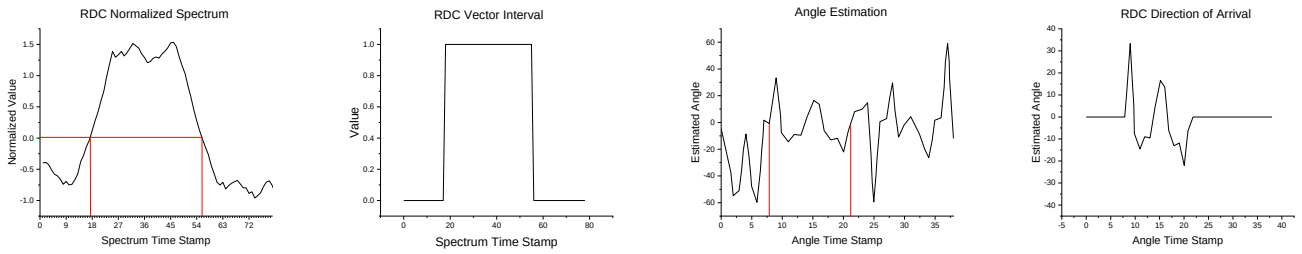
In our system, only the period of hand moving is meaningful,irrelevant time intervals may lead to redundant information. To address this, an angle mask is designed to select the appropriate time period. Initially, an angle spectrum is calculated using the MUSIC algorithm based on the noise source when no signal is present. It is observed that random background noise contributes to angle measurement errors, and this undesired noise cannot be completely eliminated. To mitigate this issue, the filtering process is employed to capture only the periods when the hand is in motion.

The angle mask algorithm, based on the time-frequency spectrum, is adopted for this purpose. When the hand is moving, the amplitude of the moving signal significantly exceeds that of the noise signal, which greatly reduces the angle error. An energy-based signal selection algorithm is developed to aid in this process. During hand movement, variations on the time-frequency plot are observed only within the moving interval. This interval is identified by calculating the energy of the frequency bin over time.

To determine the specific interval, the power spectral density from 20.06kHz to 20.27kHz and 19.73kHz to 19.94kHz is summed, and a threshold is set to identify movement. An example of a normalized power spectral map is shown in Figure 71 a). For broader system applicability, a max-min normalization technique is employed. This approach maintains similar power magnitude ratios between moving and stationary intervals across various environments and volume levels. Additionally, a simple moving average algorithm is utilized to smooth out noise points, given that angle changes continuously.To ensure the adaptability of the system to different environments, a wide-angle mask is chosen to ensure all the useful angles are preserved. For the power spectral map value, which is higher than the threshold, the vector is set to 1, everywhere

else is set to zero. The angle mask vector extracted from figure 71 a) is shown in figure 71 b).

Angle of gesture



(a) (b) (c) (d)

Figure. 71 a) Normalized power spectral. b) Angle mask internal. c) Angle vector with angle masks. d) Angle vector without angle masks.

To achieve AoA detection, MUSIC algorithm in a short time frame was applied. The peak value of MUSIC spectrum in each frame was recorded as the desired angle. To ensure a sufficient number of samples in each frame and maintain the angle's robustness, the signal was splited into 39 pieces and MUSIC is applied in every two consecutive intervals, introducing one overlap interval between two consecutive segments as shown below.

$$S_{seg} = R_{low}\begin{pmatrix} \left((seg-1)\times\left(\frac{length(R_{low})}{39}\right)+1\right) \\ : \\ \left((seg-1)\times\left(\frac{length(R_{low})}{39}\right)+1\right) \end{pmatrix} \quad (37)$$

where seg represents the index of signal segments and the length of angle vector equals to 38. Figure 71 c) and Figure 71 d) shows an example of the angle vector change after the angle mask is applied.

Figure below displays examples of ten distinct gestures under normal and error conditions. In each sub-figure, solid lines represent the normal conditions, depicting the expected outputs. Dash lines indicate one-sided and opposite conditions where a portion of the angle information is estimated in reverse. Dot lines illustrate the extra interval

condition, where a part of an unrelated interval is inadvertently included.

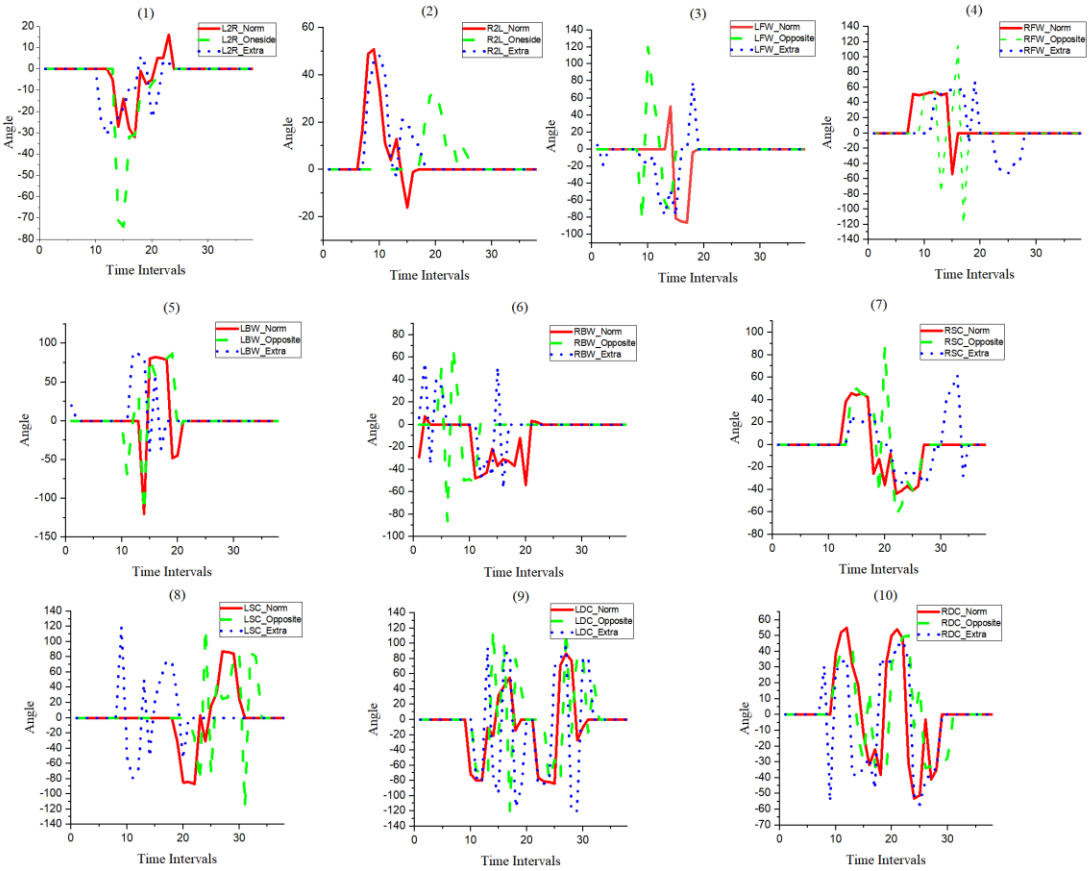


Figure. 72 Angle vector for ten gestures

Inverse angle for backward movement

Intuitively, frequency spectrum and angle information cannot solely distinguish all the gestures. For the frequency spectrum, gestures moving forward and backward from different directions have a similar pattern. For the angle vector, the gesture moving forward and backwards have the same angle pattern in a different direction. For example, the forward movement is detected as a positive angle at the right side of the receiver array and the negative angle at the left side. However, when the backward movement is detected as a positive angle at the left sides and negative at the right side. Considering one reflection point from hand in (4):

$$S = A^{t} e^{-j\left(\frac{2\pi f d(t)}{c}\right)} \tag{38}$$

The received signal from receiver array can be expressed as:

$$X(t) = A(\theta) S \tag{39}$$

$$= \left[e^{-j\left(\frac{2\pi D(0)}{\lambda}\sin(\theta)\right)}, e^{-j\left(\frac{2\pi D(1)}{\lambda}\sin(\theta)\right)}, \ldots, e^{-j\left(\frac{2\pi D(M-1)}{\lambda}\sin(\theta)\right)} \right] S$$

$$= A \left[e^{-j\left(\frac{2\pi D\times(0)}{\lambda}\sin(\theta)+\frac{2\pi fd(t)}{c}\right)}, e^{-j\left(\frac{2\pi D\times(1)}{\lambda}\sin(\theta)+\frac{2\pi fd(t)}{c}\right)}, \ldots, e^{-j\left(\frac{2\pi D\times(M-1)}{\lambda}\sin(\theta)+\frac{2\pi fd(t)}{c}\right)} \right] \quad (40)$$

Note that when backward movement is performed：

$$\begin{aligned} d'(t) &= -d(t) \\ X(t)_i &= A' e^{-j\left(\frac{2\pi D\times(i-1)}{\lambda}\sin(\theta)+\frac{2\pi fd'(t)}{c}\right)} \\ &= A' e^{-j\left(\frac{2\pi D\times(i-1)}{\lambda}\sin(\pi-\theta)+\frac{2\pi f(-d(t))}{c}\right)} \end{aligned} \quad (41)$$

which indicates that the angle vector could not distinguish a movement at direction θ and the opposite movement at the direction $\pi - \theta$. To solve this problem, angle information and frequency spectrum are combined together as the spectrum can distinguish the gesture trendency, the angle can distinguish the movement direction with respect to the sensor array.

6.3 Classification Algorithm

6.3.1 Algorithm overview

Compared with CNN, existing classification methods in the acoustic gestures recognition area adopts machine learning such as K-Nearest Neighbor (KNN) [87]–[90], Decision trees (DT) [91][92][36][93] and Support Vector Machine(SVM) [34], [83], [94]. They achieved a satisfactory classification result. However, most of them require threshold-features extraction as input which is hard to eliminate noise. A low-quality features extraction step might lead to poor classification result.

Generally, a CNN model is composed of three different layers: Convolutional Layer, Pooling Layer and Fully Connected Layer. The Convolutional Layer is used to extract features from the input data automatically. It applies convolutions using multiple filters among the input data and derives a feature map regarding each filter. The Pooling Layer

compresses the feature map to reduce the number of elements transferred to the next layer and it could be replaced by a larger stride for dimension reduction purposes. The Fully Connected Layer flattens all the elements in one layer and connects them to all the next layer elements. It would take the result of the convolutional layer/pooling layer to classify input to labels.

1D-CNN is a special case of CNN where one filter dimension equals one and the others equal to corresponding size of the input data. The convolution in the Convolutional Layer only operates in one direction as shown in figure 73. Detailed classifier architecture and explanation is illustrated in section 2).

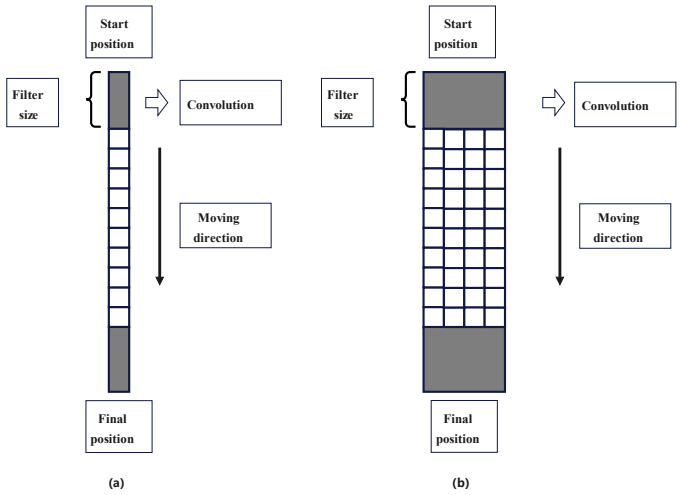


Figure. 73 1D-CNN Convolution Layer of (a) 1D data (b) 2D data

6.3.2 Classifier architecture

The classifier architecture in this chapter is shown in figure below. The convolutional neural network has the property of translational invariance, which means the translation of the input pattern along the kernel scanning axis will be detected as the same output. In our cases, the same gesture performed at any time within the 2 seconds interval would have a similar pattern in terms of angle and spectrum, which means it could be detected as the same classes. The convolutional operation scanning along the time axis can properly ignore the time difference between different samples for both inputs. However, translation invariance could also lead to a misidentification if the same pattern represents a different meaning. For example. The frequency change in the

spectrum indicates the movement speed as well as the movement orientation. A higher frequency speed as well as the movement orientation. A higher frequency change represents a fast hand movement, and a lower frequency change represents a slow movement, frequency above the central frequency line indicates forward movement and frequency below means backward movement. Different frequencies do not represent the same meaning and different patterns along the frequency bin might not represent the same gestures. Hence, we implement 1D-CNN to extract time-frequency spectrum features as it will only scan along the time axis. Two separate initial convolutional layers was adopted at the beginning on both spectrum and angle respectively scanning along the time axis, then stack the outputs into one feature map.

We applied a few subsequent convolutional layers and fully connected layers to extract features further. Batch normalization is used for all the convolutional layers to remove the internal covariate shift [95]. Rectified linear unit (ReLU) is adopted as the activation function for both convolutional layers and fully connected layer, Softmax layer is set at the end of the structure to output labels, Adam [96] and Cross Entropy are used as the optimizer and loss function respectively for all layers. Detailed structure comparison is discussed in section 6.4.2.

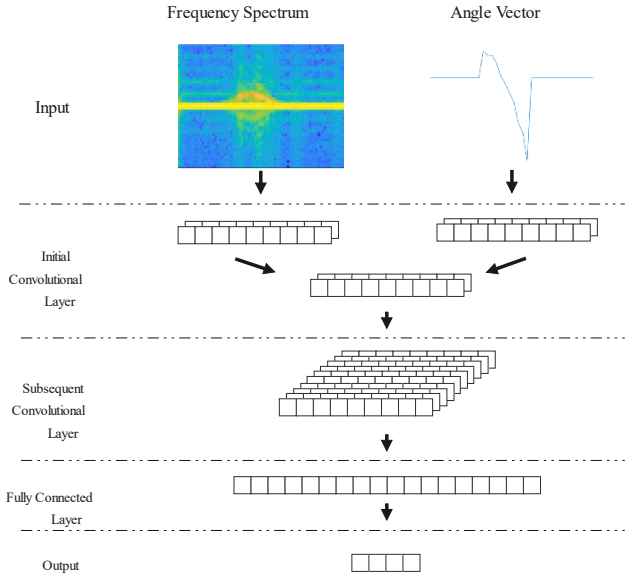


Figure. 74 Network architecture

6.4Experiments Results

6.4.1 Data acquisition

Our system is assumed to work in the situation while the user is standing (e.g., when a teacher is giving a lecture or presenting). According to empirical information, the height of a person's hand is about 70% of the height when the person is performing a gesture, and the average human height in the UK is around 168.5cm. Therefore, the sensor was set at 118 cm above the ground to deliver a satisfactory user experience. Two volunteers performed 10 different gestures (shown in figure below) within two seconds: (1) Left Backward; (2) Left Forward (3) Right Backward; (4) Right Forward; (5) Right Single Click; (6) Left Single Click; (7) Right to Left; (8) Left to Right; (9) Right Double Click; (10) Left Double Click. Each gesture was performed 150 times in both open space and complex space by each volunteer and overall $2 \times 10 \times 150 \times 2 = 6000$ samples were collected.

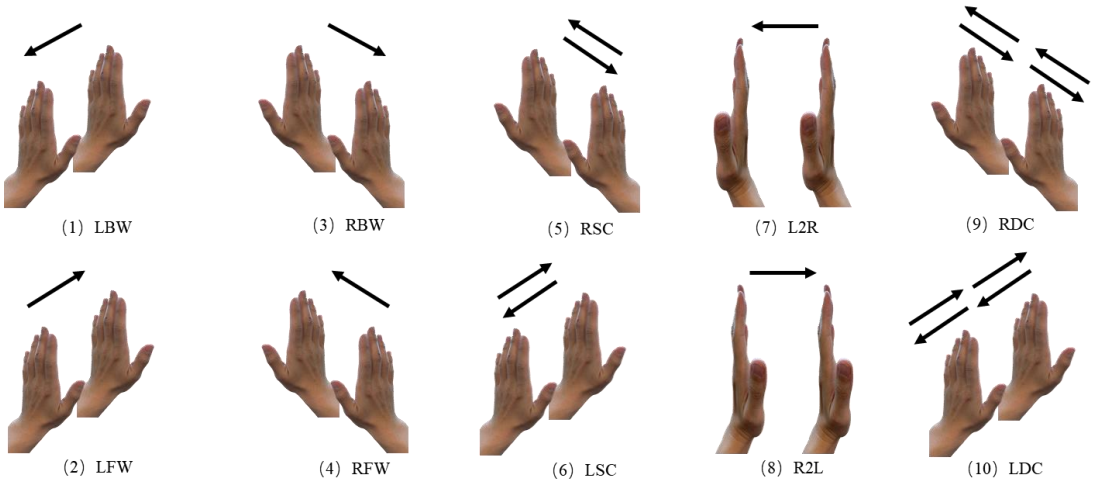


Figure. 75 Gesture set examples

To capture as many scenarios as possible, volunteers were instructed to perform gestures while standing in different locations, at varying speeds, and wearing different types of clothing, all in diverse environments. For gestures (1)-(8), the volunteers were asked to stand on both the left and right sides of the sensor. They performed gestures using the hand on the same side, and additionally, they executed cross-side gestures, such as standing on the left side and using the right hand to perform gestures on the right side. For gestures (9)-(10), the volunteers stood at the sensor's center, waving their hands across the sensor's center with both hands. They were also asked to perform gestures in two speed modes, fast and slow, alternating between them.

To consider the impact of clothing on reflection, volunteers wore two different types of garments: a furry sweater and a smooth jacket, each with distinct reflection rates. Half of the total samples were collected while the volunteers wore each type of clothing. Furthermore, a loop method was employed for gesture collection. After performing one gesture in one direction, the next gesture would be a different gesture from the opposite direction. This approach helped avoid repetitive patterns that could arise from performing the same gestures continuously.

6.4.2 Classifier structure discussion

Regarding our dataset, an investigation was carried out in terms of different network structures in the following aspects: (1)The number of convolutional layers, (2)

Kernel size and number. Four-folds cross-validation is applied to derive an accurate performance evaluation. The dataset is shuffled before being split into the folders to ensure that various gestures' ratios remain approximately consistent. The batch size and epoch are initialized as 50 and 100 respectively. The initial learning rate is set as 0.001. The whole system operates using Pytorch framework.

Table. 19 Comparison result for different convolutional layers

No. of Conv layers	None	1	2	3	4	5	6	7
Accuracy	0.87	0.926	0.933	0.938	0.935	0.945	0.945	0.943

Table. 20 Accuracy with different kernel size

Size of kernel	$1\times3\times16$	$1\times5\times16$	$1\times7\times16$
	$1\times3\times32$	$1\times5\times32$	$1\times7\times32$
	$1\times3\times64$	$1\times5\times64$	$1\times7\times64$
	$1\times3\times128$	$1\times5\times128$	$1\times7\times128$
	$1\times3\times128$	$1\times5\times128$	$1\times7\times128$
Accuracy	0.949	0.937	0.94

Table. 21 Accuracy with different kernel number

No of kernel	$1\times3\times16$	$1\times3\times16$	$1\times3\times32$	$1\times3\times64$
	$1\times3\times16$	$1\times3\times32$	$1\times3\times64$	$1\times3\times128$
	$1\times3\times32$	$1\times3\times64$	$1\times3\times128$	$1\times3\times256$
	$1\times3\times64$	$1\times3\times128$	$1\times3\times256$	$1\times3\times512$
	$1\times3\times64$	$1\times3\times128$	$1\times3\times256$	$1\times3\times512$
Accuracy	0.920	0.949	0.94	0.942

Convolutional layers

Our structure adopts two 1D-convolutional layers for angle vectors and spectrum images respectively at the beginning and then followed by convolutional layers and

fully connected layers. To achieve the best performance according to different convolutional layers, 0 - 7 subsequent convolutional layers are adopted for comparison as shown in Table 19. The classifier accuracy increases with the number of convolutional layers and remains 94.5% when five more layers are adopted. Hence, five subsequent convolutional layers are chosen for our system.

Kernels

Size and number of kernels in convolutional layers are compared. In terms of kernel size, we attempted three, five, and seven for comparison. The system accuracy decreases as the kernel size increases, shown in Table 20. For kernel numbers, three groups of different numbers were tested for comparison, shown in Table 21. The classifier achieves the highest accuracy when the kernel number is set as 16, 32, 64, 128, 128 for five convolutional layers as shown in Table 22.

6.4.3 Performance evaluation

To evaluate system performance, three questions are addressed: (1) How does the angle information influence system performance? (2) How does the standing location influence the classification result? (3) How does a different environment influence the result?

The influence of angle information on system performance

In this section, we implemented both angle information and the time-frequency spectrum as input features to enhance recognition accuracy for multi-directional gestures. To assess the impact of angle information, an experiment was conducted comparing the performance of three different input configurations: (a) spectrum only; (b) spectrum and angle; (c) angle only. To ensure that the experiment's results were not influenced by classifier structures, the same classifier structure was consistently used

across all tests. For the spectrum-only input, the angle vector was set to zero. During the convolution operation, these zero vectors is ignored, ensuring it does not affect the variables containing spectrum information, and vice versa.

Table 23 presents an 8-fold cross-validation result, illustrating the differences between the three input configurations. The average accuracy improved from 85.4% with only the spectrum as input to 94.9% when angle information was added. Notably, the system demonstrated less fluctuation with the angle vector, maintaining classification accuracy around 95% across the cross-validation results. In contrast, accuracy using only the spectrum as input varied from 81.0% to 90.2%.

Table. 22 Neural network parameters

	Angle Vector (1×38)	***Frequency Spectrum*** (77×52)
Initial Convolutional layer	**Angle_Conv_0:** Kernel $1\times3\times16$ Stride=1 **Angle_Conv_1:** Kernel $1\times3\times16$ Stride=1	**Spectrum_Conv_1:** Kernel $7\times52\times16$ Stride=2
Convolutional layer	**Merge_Conv_0:** Kernel $1\times3\times16$ Stride=1 **Merge_Conv_1:** Kernel $1\times3\times32$ Stride=1 **Merge_Conv_2:** Kernel $1\times3\times64$ Stride=1 **Merge_Conv_3:** Kernel $1\times3\times128$ Stride=1 **Merge_Conv_4:** Kernel $1\times3\times128$ Stride=1	
Fully Connect layer	**FC_0:** Output =1024 **FC_1:** Output = 256 **FC_2:** Output = 10	
	Output	

The confusion matrices shown in Figure 76 a) and b) illustrate classification results among 10 different gestures with two types of inputs. It shows that it is more likely to fail with the same gesture from the other direction for the system with the only spectrum as input. As shown in Figure 76 a), 29% RBW is recognized as LBW, 18% of LFW is recognized as RFW, 18% LBW is recognized as RBW, 14 % of RSC is recognized as LSC, 11% of L2R is recognized as R2L, and 10% of LDC is recognized as RDC. It also contains a few errors less than 10%. However, when both spectrum and angle are used as input, gestures from different directions have a higher probability of being distinguished. The misrecognition rate for all gestures is below 6% and some minor

errors in spectrum only cases are also well corrected. The significance of angle information is demonstrated.

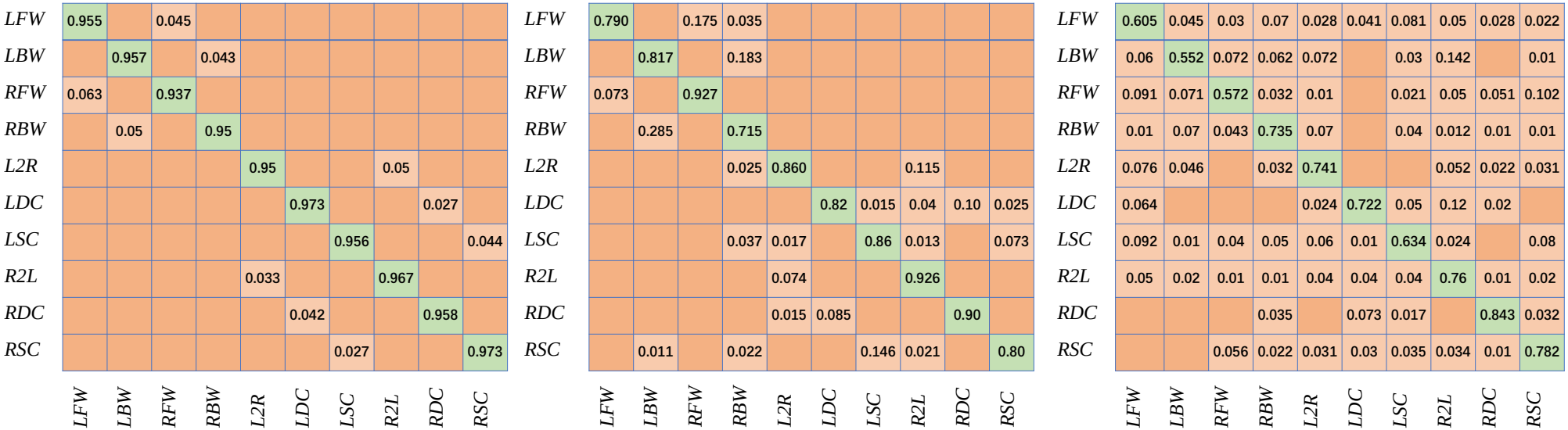


Figure. 76 Confusion matrices of classification results with two types of input a) Spectrum only b) Spectrum and angle c) Using angle only. (Normalized in horizontal direction)

Table. 23 Classification result for two different inputs (a) Spectrum only (b) Spectrum and angle

K-fold index	Spectrum	Spectrum&Angle
1	0.902	0.951
2	0.842	0.948
3	0.870	0.949
4	0.854	0.946
5	0.846	0.961
6	0.813	0.948
7	0.897	0.945
8	0.810	0.945
Ave	0.854	0.949

In addition, angle vector only is tested. The classification result is shown in Figure 76 c). It shows that using angle information as the only input is not sufficient to classify the gestures properly with a classification accuracy of 75.4%. 8-fold classification result is shown in Table 23. It is mainly caused by the angle inverse problem and redundant error. After combining the time-frequency graph, the performance can be greatly improved.

The influence of standing location on classification result

Considering the situation when the user may stand at the different directions to the sensor when performing gestures, a cross-side gestures may be performed. A cross-side gesture means the user is standing on the opposite side when waving their hand, for example, when the user is standing on the left side of the sensor and waving forward using right hand at the right side. In this situation, our system might be affected by extra noise introduced by the arm movement and shoulder movement. To verify this idea, 80 cross-side samples and 80 same-side samples were taken for comparison. The overall accuracy achieves 93.8% and 95% respectively, and the accuracy rate changes within the acceptable range. It indicates that our system is robust to cross-side gestures.

The influence of different environments on system performance.

The complexity of the surrounding environment influences recognition accuracy. More objects around the sensor lead to more reflective noises than an open environment. To investigate system robustness to the environment, 130 samples were taken in an open environment, 130 samples were taken in a complex environment (sensor surrounded by furniture) to compare. The classification accuracy of the complex environment and open environment is 92.2% and 96.1% respectively, which indicates complex environment deteriorates accuracy, but not significantly.

6.5 Discussion

Privacy: As we are using the off-the-shelf microphone array[97], [98], users may be concerned that their private conversation within the operating range will be recorded and might lead to a privacy issue. As a matter of fact, we are aiming at the signal around the carrier frequency, and only the signal around 20kHz will be recorded. A high pass filter is adopted to remove all the irrelevant signals, which includes human voice. Therefore, the private conversation will not be recorded.

Time consumption: We evaluate our system time consumption by processing 100 samples and calculating the average processing time. For the two-second audio signal, our system requires 1.37s for feature extraction, 0.16s for gestures classification. For further time reduction, the data could be sent to a cloud server to process and send the classification result back to the system. The cloud server with a higher computational ability could speed up the calculation process.

Operation distance: Our system is set to operate at round 1m, which is the distance between the sensor to the point where the user stands. For a longer operation distance, the hardware can be adjusted by increasing transmitting energy to mitigate the signal attenuation and we can collect more data to optimize the classifier accuracy.

Real-time: currently, our system is operating in an off-line mode. However, it could be modified into a real-time system in the future. The system could take every two seconds as one frame, extracting features and classifying gestures within every frame. Alternatively, setting every two seconds as a frame and taking every three seconds as a processing interval. The processing interval will contain one second's data from the previous interval, which prevents useful information from being distorted by cutting the interval.

Gestures from omnidirectional: since our system uses a uniform linear array as the receiving sensor, the angle of arrival can only be determined if the gesture is parallel to a sensor array. In the future, our system can be upgraded to detect omnidirectional gestures by adopting more acoustic sensors with a more complex layout. Angle and frequency spectrum extraction processes can be similarly adopted to obtain a gesture's direction from both horizontal and vertical. However, extra data collection and classifier fine-tuning might be necessary.

Chapter 7 Conclusion and Future Work

7.1 Conclusion

With the booming development of technology, hand gesture recognition has become a hotspot in Human-Computer Interaction (HCI) systems. Ultrasound hand gesture recognition is an innovative method that has attracted ample interest due to its strong real-time performance, low cost, large field of view, and illumination independence. Well-investigated HCI applications include external digital pens, game controllers on smart mobile devices, and web browser control on laptops. This thesis probes gesture recognition systems on multiple platforms to study the behavior of system performance with various gesture features. Focused on this topic, the contributions of this thesis can be summarized from the perspectives of smartphone acoustic field and hand model simulation, real-time gesture recognition on smart devices with speed categorization algorithm, fast reaction gesture recognition based on

temporal neural networks, and angle of arrival-based gesture recognition system.

Firstly, in order to facilitate gesture recognition systems on off-the-shelf smartphone devices, the design and evaluation of smartphone acoustic field simulation were carried out to investigate the acoustic field performance. 2D and 3D models were investigated to probe the sound pressure in a slice-wise distribution and an overall sound field distribution. Five application scenarios were evaluated: (1) Smartphone's acoustic experiment, (2) Hand's reflection from the acoustic field, (3) Hand's location experiment, (4) Hand's direction experiment, (5) Frequency enhancement experiment. The simulation was systematically tested and evaluated on the basis of sound radiation magnitude, radiation direction, the reflection of obstacles, and sound pressure in the far field of the selected hardware. The simulation results show the feasibility of ultrasound hand gesture recognition based on smartphones.

Mobility and system accuracy are two significant factors that determine gesture recognition performance. As smartphones have high-quality acoustic devices for developing gesture recognition, to achieve a portable gesture recognition system with high accuracy, novel algorithms were developed to distinguish gestures using smartphone built-in speakers and microphones. The proposed system adopts Short-Time-Fourier-Transform (STFT) and machine learning to capture hand movement and determine gestures by the pretrained neural network. To differentiate gesture speeds, a specific neural network was designed and set as part of the classification algorithm. The final accuracy rate achieves 96% among nine gestures and three speed levels. The proposed algorithms were evaluated comparatively through algorithm comparison, and the accuracy outperformed state-of-the-art systems.

Furthermore, a fast reaction gesture recognition based on temporal neural networks was designed. Traditional ultrasound gesture recognition adopts convolutional neural networks that have flaws in terms of response time and discontinuous operation. Besides, overlap intervals in network processing cause cross-frame failures that greatly reduce system performance. To mitigate these problems, a novel fast reaction gesture recognition system that slices signals in short time intervals was designed. The

proposed system adopted a novel convolutional recurrent neural network (CRNN) that calculates gesture features in a short time and combines features over time. The results showed the reaction time significantly reduced from 1s to 0.2s, and accuracy improved to 100% for six gestures.

Lastly, an acoustic sensor array was built to investigate the angle information of performed gestures. The direction of a gesture is a significant feature for gesture classification, which enables the same gesture in different directions to represent different actions. Previous studies mainly focused on types of gestures and analyzing approaches (e.g., Doppler Effect and channel impulse response, etc.), while the direction of gestures was not extensively studied. An acoustic gesture recognition system based on both speed information and gesture direction was developed. The system achieved 94.9% accuracy among ten different gestures from two directions. The proposed system was evaluated comparatively through numerical neural network structures, and the results confirmed that incorporating additional angle information improved the system's performance.

In summary, the work presented in this thesis validates the feasibility of recognizing hand gestures using remote ultrasonic sensing across multiple platforms. The acoustic simulation explores the smartphone acoustic field distribution and response results in the context of hand gesture recognition applications. The smartphone gesture recognition system demonstrates the accuracy of recognition through ultrasound signals and conducts an analysis of classification speed. The fast reaction system proposes a more optimized solution to address the cross-frame issue using temporal neural networks, reducing the response latency to 0.2s. The speed and angle-based system provides an additional feature for gesture recognition. The established work will accelerate the development of intelligent hand gesture recognition, enrich the available gesture features, and contribute to further research in various gestures and application scenarios.

7.2 Future Work

This thesis has significantly advanced acoustic gesture recognition area using the developed techniques; however, ongoing research is crucial to further progress in the following areas.

Regarding the ultrasound hardware array, a more meticulously designed sensor array could be considered as the hardware platform to further investigate the incoming direction of the recognition system. The current recognition hardware utilizes a linear 4-way microphone array for input. This array is limited to detecting signals in the horizontal direction.[99]–[102] . Improving the quantity of microphones would increase the number of received signal channels, leading to a better signal-to-noise ratio (SNR) from processing. Additionally, enhancing the distribution (or orientation) of the microphone array could expand the detectable dimensions, thereby enriching the area that can be detected.

Regarding the gesture complexity and diversity, future research should delve into the recognition of more complex and diverse hand gestures. For instance, the system could be trained to recognize intricate finger-level movements like sign language alphabets or musical instrument manipulation gestures. This would allow for more detailed and nuanced interactions in HCI applications, such as enabling musicians to control music software with hand gestures or allowing for non-verbal communication in virtual meetings. Additionally, exploring the recognition of dynamic gestures, such as a series of hand movements used in dance or sports coaching, would greatly expand the system's applicability. This could include tracking complex gesture sequences in real-time, enabling interactive dance or fitness applications. Regarding the gesture complexity and diversity, future research should delve into the recognition of more complex and diverse hand gestures. For instance, the system could be trained to recognize intricate finger-level movements [30], [103], [104]like sign language alphabets or musical instrument manipulation gestures. This would allow for more detailed and nuanced interactions in HCI applications, such as enabling musicians to

control music software with hand gestures or allowing for non-verbal communication in virtual meetings. Additionally, exploring the recognition of dynamic gestures, such as a series of hand movements used in dance or sports coaching, would greatly expand the system's applicability. This could include tracking complex gesture sequences in real-time, enabling interactive dance or fitness applications.

Regarding the Cross-Platform [28]compatibility and integration, the development of algorithms should be further studied. For instance, adapting the algorithms to work seamlessly with both Android and iOS systems in smartphones, as well as with wearable devices like smartwatches or fitness bands, which could use gesture control for quick responses or health tracking. Integration with smart home devices, such as gesture-controlled lighting systems or security systems that recognize specific hand signals for operation, is another area of potential development. Ensuring these algorithms can operate efficiently across different processing powers and hardware capabilities of these devices is crucial for universal applicability.

Reference

[1] Z. Ren, J. Meng, and J. Yuan, “Depth camera based hand gesture recognition and its applications in human-computer-interaction,” in *Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on*, IEEE, 2011, pp. 1–5.

[2] S. S. Rautaray and A. Agrawal, “Vision based hand gesture recognition for human computer interaction: a survey,” *AI Review*, vol. 43, no. 1, pp. 1–54, 2015, doi: 10.1007/s10462-012-9356-9.

[3] F. S. Chen, C. M. Fu, and C. L. Huang, “Hand gesture recognition using a real-time tracking method and hidden Markov models,” *Image Vis Comput*, vol. 21, no. 8, pp. 745–758, 2003, doi: 10.1016/S0262-8856(03)00070-2.

[4] D. Xu, Y.-L. Chen, C. Lin, X. Kong, and X. Wu, “Real-time dynamic gesture recognition system based on depth perception for robot navigation,” in *2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, IEEE, 2012, pp. 689–

694.

[5] Z. Wang *et al.*, “Hand Gesture Recognition Based on Active Ultrasonic Sensing of Smartphone: A Survey,” *IEEE Access*, vol. 7, pp. 111897–111922, 2019.

[6] J. Weissmann and R. Salomon, “Gesture recognition for virtual reality applications using data gloves and neural networks,” in *IJCNN’99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)*, IEEE, 1999, pp. 2043–2046.

[7] J.-H. Kim, N. D. Thang, and T.-S. Kim, “3-d hand motion tracking and gesture recognition using a data glove,” in *2009 IEEE International Symposium on Industrial Electronics*, IEEE, 2009, pp. 1013–1018.

[8] R. O. Schmidt, “A signal subspace approach to multiple emitter location and spectral estimation.,” 1982.

[9] G. Bienvenu and L. Kopp, “Principle de la goniometrie passive adaptive,” in *Proc. 7’eme Colloque GRESIT*, 1979, pp. 101–106.

[10] H.-B. Zhang *et al.*, “A comprehensive survey of vision-based human action recognition methods,” *Sensors*, vol. 19, no. 5, p. 1005, 2019.

[11] D. T. Sidhant Gupta1, 2, Dan Morris1, Shwetak N Patel1, 2, “SoundWave: Using the Doppler Effect to Sense Gestures,” pp. 1911–1914, 2012.

[12] S. G. Jin, C. Gao, and J. H. Li, “Atmospheric Sounding from Fengyun-3C GPS Radio Occultation Observations: First Results and Validation,” *ADVANCES IN METEOROLOGY*, vol. 2019, 2019, doi: 10.1155/2019/4780143.

[13] I. Nirmal, A. Khamis, M. Hassan, W. Hu, and X. Q. Zhu, “Deep Learning for Radio-Based Human Sensing: Recent Advances and Future Directions,” *IEEE COMMUNICATIONS SURVEYS AND TUTORIALS*, vol. 23, no. 2, pp. 995–1019, 2021, doi: 10.1109/COMST.2021.3058333.

[14] R. N. Zhao, X. L. Ma, X. H. Liu, and J. Liu, “An End-to-End Network for Continuous Human Motion Recognition via Radar Radios,” *IEEE Sens J*, vol. 21, no. 5, pp. 6487–6496, 2021, doi: 10.1109/JSEN.2020.3040865.

[15] Z. Ju and H. Liu, “Human hand motion analysis with multisensory information,” *IEEE/AsMe Transactions on Mechatronics*, vol. 19, no. 2, pp. 456–466, 2013.

[16] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, “Speech Recognition Using Deep Neural Networks: A Systematic Review,” *IEEE ACCESS*, vol. 7, pp. 19143–19165, 2019, doi: 10.1109/ACCESS.2019.2896880.

[17] M. Liu, “English speech emotion recognition method based on speech recognition,” *Int J Speech Technol*, 2022, doi: 10.1007/s10772-021-09955-4.

[18] S. Ishimitsu, K. Oda, and M. Nakayama, “BODY-CONDUCTED SPEECH RECOGNITION IN SPEECH SUPPORT SYSTEM FOR DISORDERS,” *INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING INFORMATION AND CONTROL*, vol. 7, no. 8, pp. 4929–4940, 2011.

[19] J. Kratt, F. Metze, R. Stiefelhagen, and A. Waibel, “Large vocabulary audio-visual speech recognition using the Janus speech recognition toolkit,” in *PATTERN RECOGNITION*, C. E. Rasmussen, H. H. Bulthoff, M. A. Giese, and B. Scholkopf, Eds., 2004, pp. 488–495.

[20] S. Kimura, “Advances in speech recognition technologies,” *FUJITSU SCIENTIFIC & TECHNICAL JOURNAL*, vol. 35, no. 2, pp. 202–211, 1999.

[21] K. K. Paliwal and K. S. Yao, *Robust Speech Recognition Under Noisy Ambient Conditions*. 2010. doi: 10.1016/B978-0-12-374708-2.00006-1.

[22] H. J. Yoo, S. Seo, S. W. Im, and G. Y. Gim, “The Performance Evaluation of Continuous Speech Recognition Based on Korean Phonological Rules of Cloud-Based Speech Recognition Open API,” *INTERNATIONAL JOURNAL OF NETWORKED AND DISTRIBUTED COMPUTING*, vol. 9, no. 1, pp. 10–18, 2021, doi: 10.2991/ijndc.k.201218.005.

[23] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, “Speech Recognition Using Deep Neural Networks: A Systematic Review,” *IEEE ACCESS*, vol. 7, pp. 19143–19165, 2019, doi: 10.1109/ACCESS.2019.2896880.

[24] N. Wang, X. H. Zhang, and A. Sharma, “A Research on HMM based Speech Recognition in Spoken English,” *RECENT ADVANCES IN ELECTRICAL & ELECTRONIC ENGINEERING*, vol. 14, no. 6, pp. 617–626, 2021, doi: 10.2174/2352096514666210413122517.

[25] S. Jeong and M. Hahn, “Speech quality and recognition rate improvement in care noise environments,” *Electron Lett*, vol. 37, no. 12, pp. 800–802, 2001, doi: 10.1049/el:20010513.

[26] Y. Wu and G. C. Li, “Intelligent Robot English Speech Recognition Method Based on Online Database,” *JOURNAL OF INFORMATION & KNOWLEDGE MANAGEMENT*, vol. 21, no. SUPP02, 2022, doi: 10.1142/S0219649222400123.

[27] X. L. Lu and M. A. Shah, “Implementation of Embedded Unspecific Continuous English Speech Recognition Based on HMM,” *RECENT ADVANCES IN ELECTRICAL & ELECTRONIC ENGINEERING*, vol. 14, no. 6, pp. 649–659, 2021, doi: 10.2174/2352096514666210715144717.

[28] D. Griffin and J. Lim, “Signal estimation from modified short-time Fourier transform,” *IEEE Trans Acoust*, vol. 32, no. 2, pp. 236–243, 1984.

[29] W. Ding and L. He, “Adaptive multi-scale detection of acoustic events,” *IEEE/ACM Trans Audio Speech Lang Process*, vol. 28, pp. 294–306, 2020, doi: 10.1109/TASLP.2019.2953350.

[30] A. Das, I. Tashev, and S. Mohammed, “Ultrasound based gesture recognition,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 406–410. doi: 10.1109/ICASSP.2017.7952187.

[31] C. Yiallourides and P. P. Parada, “Low power ultrasonic gesture recognition for mobile handsets,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 2697–2701.

[32] B. Y. Fu, F. Kirchbuchner, A. Kuijper, A. Braun, and D. V Gangatharan, “Fitness Activity Recognition on Smartphones Using Doppler Measurements,” *INFORMATICS-BASEL*, vol. 5, no. 2, 2018, doi: 10.3390/informatics5020024.

[33] D. P. Tao, Y. G. Wen, and R. C. Hong, “Multicolumn Bidirectional Long Short-Term Memory for Mobile Devices-Based Human Activity Recognition,” *IEEE Internet Things J*, vol. 3, no. 6, pp. 1124–1134, 2016, doi: 10.1109/JIOT.2016.2561962.

[34] J. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, “Deep convolutional neural networks on multichannel time series for human activity recognition,” in *Twenty-*

Fourth International Joint Conference on Artificial Intelligence, 2015.

[35] Y. Qifan, T. Hao, Z. Xuebing, L. Yin, and Z. Sanfeng, "Dolphin: Ultrasonic-based gesture recognition on smartphone platform," in *2014 IEEE 17th International Conference on Computational Science and Engineering*, IEEE, 2014, pp. 1461–1468.

[36] Z. Z. Du Jiang *et al.*, "Gesture recognition based on binocular vision," *Cluster Comput*, vol. 22, no. suppl 6, pp. 13261–13271, 2019.

[37] S. Z. Gurbuz *et al.*, "American Sign Language Recognition Using RF Sensing," *IEEE Sens J*, vol. 21, no. 3, pp. 3763–3775, 2021, doi: 10.1109/JSEN.2020.3022376.

[38] G. Okeyo, L. M. Chen, and H. Wang, "An Agent-mediated Ontology-based Approach for Composite Activity Recognition in Smart Homes," *JOURNAL OF UNIVERSAL COMPUTER SCIENCE*, vol. 19, no. 17, pp. 2577–2597, 2013.

[39] H. Ponce, M. D. Martinez-Villasenor, and L. Miralles-Pechuan, "A Novel Wearable Sensor-Based Human Activity Recognition Approach Using Artificial Hydrocarbon Networks," *SENSORS*, vol. 16, no. 7, 2016, doi: 10.3390/s16071033.

[40] M. Z. Hou, S. Liu, J. L. Zhou, Y. Zhang, and Z. L. Feng, "Extreme Low-Resolution Activity Recognition Using a Super-Resolution-Oriented Generative Adversarial Network," *Micromachines (Basel)*, vol. 12, no. 6, 2021, doi: 10.3390/mi12060670.

[41] G. Okeyo, L. M. Chen, and H. Wang, "Combining ontological and temporal formalisms for composite activity modelling and recognition in smart homes," *FUTURE GENERATION COMPUTER SYSTEMS-THE INTERNATIONAL JOURNAL OF ESCIENCE*, vol. 39, pp. 29–43, 2014, doi: 10.1016/j.future.2014.02.014.

[42] M. Smolen, "Consistency of Outputs of the Selected Motion Acquisition Methods for Human Activity Recognition," *J Healthc Eng*, vol. 2019, 2019, doi: 10.1155/2019/9873430.

[43] A. Shrestha, H. B. Li, J. Le Kernec, and F. Fioranelli, "Continuous Human Activity Classification From FMCW Radar With Bi-LSTM Networks," *IEEE Sens J*, vol. 20, no. 22, pp. 13607–13619, 2020, doi: 10.1109/JSEN.2020.3006386.

[44] A. Bulling, U. Blanke, and B. Schiele, "A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors," *ACM Comput Surv*, vol. 46, no. 3, 2014, doi:

10.1145/2499621.

[45] M. Z. Uddin, J. J. Lee, and T. S. Kim, “Independent shape component-based human activity recognition via Hidden Markov Model,” *APPLIED INTELLIGENCE*, vol. 33, no. 2, pp. 193–206, 2010, doi: 10.1007/s10489-008-0159-2.

[46] C. W. Han, S. J. Kang, and N. S. Kim, “Implementation of HMM-Based Human Activity Recognition Using Single Triaxial Accelerometer,” *IEICE TRANSACTIONS ON FUNDAMENTALS OF ELECTRONICS COMMUNICATIONS AND COMPUTER SCIENCES*, vol. E93A, no. 7, pp. 1379–1383, 2010, doi: 10.1587/transfun.E93.A.1379.

[47] L. M. Chen, C. D. Nugent, and H. Wang, “A Knowledge-Driven Approach to Activity Recognition in Smart Homes,” *IEEE Trans Knowl Data Eng*, vol. 24, no. 6, pp. 961–974, 2012, doi: 10.1109/TKDE.2011.51.

[48] C. Z. Li, “Research on Contactless Identification and Evaluation of Unarmed Fitness Activity,” *Sci Program*, vol. 2022, 2022, doi: 10.1155/2022/4282569.

[49] Y. Shavit and I. Klein, “Boosting Inertial-Based Human Activity Recognition With Transformers,” *IEEE ACCESS*, vol. 9, pp. 53540–53547, 2021, doi: 10.1109/ACCESS.2021.3070646.

[50] C. Meissner, J. Meixensberger, A. Pretschner, and T. Neumuth, “Sensor-based surgical activity recognition in unconstrained environments,” *MINIMALLY INVASIVE THERAPY & ALLIED TECHNOLOGIES*, vol. 23, no. 3–4, pp. 198–205, 2014, doi: 10.3109/13645706.2013.878363.

[51] Z. S. Abdallah, M. M. Gaber, B. Srinivasan, and S. Krishnaswamy, “Adaptive mobile activity recognition system with evolving data streams,” *Neurocomputing*, vol. 150, pp. 304–317, 2015, doi: 10.1016/j.neucom.2014.09.074.

[52] S. Ansah and D. L. Chen, “Wearable-Gait-Analysis-Based Activity Recognition: A Review,” *INTERNATIONAL JOURNAL ON SMART SENSING AND INTELLIGENT SYSTEMS*, vol. 15, no. 1, 2022, doi: 10.2478/ijssis-2022-0021.

[53] B. Y. Fu, F. Kirchbuchner, A. Kuijper, A. Braun, and D. V Gangatharan, “Fitness Activity Recognition on Smartphones Using Doppler Measurements,” *INFORMATICS-BASEL*, vol. 5, no. 2, 2018, doi: 10.3390/informatics5020024.

[54] B. Y. Fu, F. Kirchbuchner, A. Kuijper, A. Braun, and D. V Gangatharan, “Fitness Activity Recognition on Smartphones Using Doppler Measurements,” *INFORMATICS-BASEL*, vol. 5, no. 2, 2018, doi: 10.3390/informatics5020024.

[55] D. Gordon, J. H. Hanne, M. Berchtold, A. A. N. Shirehjini, and M. Beigl, “Towards Collaborative Group Activity Recognition Using Mobile Devices,” *MOBILE NETWORKS & APPLICATIONS*, vol. 18, no. 3, pp. 326–340, 2013, doi: 10.1007/s11036-012-0415-x.

[56] Z. S. Abdallah, M. M. Gaber, B. Srinivasan, and S. Krishnaswamy, “Activity Recognition with Evolving Data Streams: A Review,” *ACM Comput Surv*, vol. 51, no. 4, 2018, doi: 10.1145/3158645.

[57] G. Lavee, M. Rudzsky, and E. Rivlin, “Propagating Certainty in Petri Nets for Activity Recognition,” *IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY*, vol. 23, no. 2, pp. 337–348, 2013, doi: 10.1109/TCSVT.2012.2203742.

[58] S. Q. Wang and G. Zhou, “A review on radio based activity recognition,” *DIGITAL COMMUNICATIONS AND NETWORKS*, vol. 1, no. 1, pp. 20–29, 2015, doi: 10.1016/j.dcan.2015.02.006.

[59] D. Cook, K. D. Feuz, and N. C. Krishnan, “Transfer learning for activity recognition: a survey,” *Knowl Inf Syst*, vol. 36, no. 3, pp. 537–556, 2013, doi: 10.1007/s10115-013-0665-3.

[60] H. Wang, W. A. Xu, G. D. Saxton, and A. Singhal, “Social media fandom for health promotion? Insights from East Los High, a transmedia edutainment initiative,” *SEARCH-JOURNAL OF MEDIA AND COMMUNICATION RESEARCH*, vol. 11, no. 1, pp. 1–15, 2019.

[61] P. Rzeszucinski, M. Orman, C. T. Pinto, A. Tkaczyk, and M. Sulowicz, “Bearing Health Diagnosed with a Mobile Phone ACOUSTIC SIGNAL MEASUREMENTS CAN BE USED TO TEST FOR STRUCTURAL FAULTS IN MOTORS,” *IEEE INDUSTRY APPLICATIONS MAGAZINE*, vol. 24, no. 4, pp. 17–23, 2018, doi: 10.1109/MIAS.2017.2740463.

[62] Y. M. Xiao, C. Z. Li, and J. S. Lin, “A portable noncontact heartbeat and respiration

monitoring system using 5-GHz radar," *IEEE Sens J*, vol. 7, no. 7–8, pp. 1042–1043, 2007, doi: 10.1109/JSEN.2007.895979.

[63] T. B. Wang *et al.*, "Contactless Respiration Monitoring Using Ultrasound Signal With Off-the-Shelf Audio Devices," *IEEE Internet Things J*, vol. 6, no. 2, pp. 2959–2973, 2019, doi: 10.1109/JIOT.2018.2877607.

[64] C. W. Ding *et al.*, "Continuous Human Motion Recognition With a Dynamic Range-Doppler Trajectory Method Based on FMCW Radar," *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, vol. 57, no. 9, pp. 6821–6831, 2019, doi: 10.1109/TGRS.2019.2908758.

[65] S. D. Min, J. K. Kim, H. S. Shin, Y. H. Yun, C. K. Lee, and M. Lee, "Noncontact Respiration Rate Measurement System Using an Ultrasonic Proximity Sensor," *IEEE Sens J*, vol. 10, no. 11, pp. 1732–1739, 2010, doi: 10.1109/JSEN.2010.2044239.

[66] P. A. Braveman, "Monitoring equity in health and healthcare: A conceptual framework," *J Health Popul Nutr*, vol. 21, no. 3, pp. 181–192, 2003.

[67] F. Wu *et al.*, "A lightweight and robust two-factor authentication scheme for personalized healthcare systems using wireless medical sensor networks," *FUTURE GENERATION COMPUTER SYSTEMS-THE INTERNATIONAL JOURNAL OF ESCIENCE*, vol. 82, pp. 727–737, 2018, doi: 10.1016/j.future.2017.08.042.

[68] X. Y. Jin, "Design of bridge health monitoring system based on B/S mode and SOA architecture," *Int J Biom*, vol. 14, no. 2, pp. 199–207, 2022.

[69] H. Wang, W. A. Xu, G. D. Saxton, and A. Singhal, "Social media fandom for health promotion? Insights from East Los High, a transmedia edutainment initiative," *SEARCH-JOURNAL OF MEDIA AND COMMUNICATION RESEARCH*, vol. 11, no. 1, pp. 1–15, 2019.

[70] H. Wang, W. A. Xu, G. D. Saxton, and A. Singhal, "Social media fandom for health promotion? Insights from East Los High, a transmedia edutainment initiative," *SEARCH-JOURNAL OF MEDIA AND COMMUNICATION RESEARCH*, vol. 11, no. 1, pp. 1–15, 2019.

[71] C. V Anikwe *et al.*, "Mobile and wearable sensors for data-driven health monitoring

system: State-of-the-art and future prospect," *Expert Syst Appl*, vol. 202, 2022, doi: 10.1016/j.eswa.2022.117362.

[72] P. Rzeszucinski, M. Orman, C. T. Pinto, A. Tkaczyk, and M. Sulowicz, "Bearing Health Diagnosed with a Mobile Phone ACOUSTIC SIGNAL MEASUREMENTS CAN BE USED TO TEST FOR STRUCTURAL FAULTS IN MOTORS," *IEEE INDUSTRY APPLICATIONS MAGAZINE*, vol. 24, no. 4, pp. 17–23, 2018, doi: 10.1109/MIAS.2017.2740463.

[73] P. Rzeszucinski, M. Orman, C. T. Pinto, A. Tkaczyk, and M. Sulowicz, "Bearing Health Diagnosed with a Mobile Phone ACOUSTIC SIGNAL MEASUREMENTS CAN BE USED TO TEST FOR STRUCTURAL FAULTS IN MOTORS," *IEEE INDUSTRY APPLICATIONS MAGAZINE*, vol. 24, no. 4, pp. 17–23, 2018, doi: 10.1109/MIAS.2017.2740463.

[74] K. Fagher, M. Badenhorst, L. Kunorozva, W. Derman, and J. Lexell, "'It gives me a wake up call'-It is time to implement athlete health monitoring within the Para sport context," *Scand J Med Sci Sports*, vol. 33, no. 5, pp. 776–786, 2023, doi: 10.1111/sms.14281.

[75] B. B. Gupta, A. Gaurav, C. H. Hsu, and B. Jiao, "Identity-Based Authentication Mechanism for Secure Information Sharing in the Maritime Transport System," *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, vol. 24, no. 2, pp. 2422–2430, 2023, doi: 10.1109/TITS.2021.3125402.

[76] S. Venkatraman and S. Parvin, "Developing an IoT Identity Management System Using Blockchain," *SYSTEMS*, vol. 10, no. 2, 2022, doi: 10.3390/systems10020039.

[77] B. Zhao, P. Y. Zhao, and P. R. Fan, "ePUF: A Lightweight Double Identity Verification in IoT," *Tsinghua Sci Technol*, vol. 25, no. 5, pp. 625–635, 2020, doi: 10.26599/TST.2019.9010072.

[78] S. Pal, M. Hitchens, and V. Varadharajan, "IoT for wearable devices: access control and identity management," in *WEARABLE SENSORS: APPLICATIONS, DESIGN AND IMPLEMENTATION*, S. C. Mukhopadhyay and T. Islam, Eds., 2017. doi: 10.1088/978-0-7503-1505-0ch6.

[79] X. Cheng *et al.*, “Secure Identity Authentication of Community Medical Internet of Things,” *IEEE ACCESS*, vol. 7, pp. 115966–115977, 2019, doi: 10.1109/ACCESS.2019.2935782.

[80] J. Long and X. Su, “Anonymous chaotic-based identity authentication protocol in IoT,” *Int J Embed Syst*, vol. 14, no. 2, pp. 194–200, 2021, doi: 10.1504/IJES.2021.113813.

[81] M. M. Zhang *et al.*, “A CPK-Based Identity Authentication Scheme for IoT,” *COMPUTER SYSTEMS SCIENCE AND ENGINEERING*, vol. 40, no. 3, pp. 1217–1231, 2022, doi: 10.32604/csse.2022.017657.

[82] K. M. Sadique, R. Rahmani, and P. Johannesson, “IMSC-EIoTD: Identity Management and Secure Communication for Edge IoT Devices,” *SENSORS*, vol. 20, no. 22, 2020, doi: 10.3390/s20226546.

[83] H. P. Huang, L. K. Hu, F. Xiao, A. M. Du, N. Ye, and F. He, “An EEG-Based Identity Authentication System with Audiovisual Paradigm in IoT,” *SENSORS*, vol. 19, no. 7, 2019, doi: 10.3390/s19071664.

[84] A. G. Reddy, D. Suresh, K. Phaneendra, J. S. Shin, and V. Odelu, “Provably secure pseudo-identity based device authentication for smart cities environment,” *Sustain Cities Soc*, vol. 41, pp. 878–885, 2018, doi: 10.1016/j.scs.2018.06.004.

[85] L. Q. Gong, D. M. Alghazzawi, and L. Cheng, “BCoT Sentry: A Blockchain-Based Identity Authentication Framework for IoT Devices,” *INFORMATION*, vol. 12, no. 5, 2021, doi: 10.3390/info12050203.

[86] J. J. Sun, P. Zhang, and X. H. Kong, “Identity Authentication Protocol of Smart Home IoT based on Chebyshev Chaotic Mapping,” *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, vol. 14, no. 4, pp. 557–565, 2023.

[87] Z. H. Cui *et al.*, “A Hybrid BlockChain-Based Identity Authentication Scheme for Multi-WSN,” *IEEE Trans Serv Comput*, vol. 13, no. 2, pp. 241–251, 2020, doi: 10.1109/TSC.2020.2964537.

[88] M. Hossain and R. Hasan, “P-HIP: A Lightweight and Privacy-Aware Host Identity Protocol for Internet of Things,” *IEEE Internet Things J*, vol. 8, no. 1, pp. 555–571,

2021, doi: 10.1109/JIOT.2020.3009024.

[89] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, “FingerIO : Using Active Sonar for Fine-Grained Finger Tracking,” 2016.

[90] X. Y. Wang, L. J. Gao, and S. W. Mao, “CSI Phase Fingerprinting for Indoor Localization With a Deep Learning Approach,” *IEEE Internet Things J*, vol. 3, no. 6, pp. 1113–1123, 2016, doi: 10.1109/JIOT.2016.2558659.

[91] Q. L. Zeng, Z. Kuang, S. B. Wu, and J. Yang, “A Method of Ultrasonic Finger Gesture Recognition Based on the Micro-Doppler Effect,” *APPLIED SCIENCES-BASEL*, vol. 9, no. 11, 2019, doi: 10.3390/app9112314.

[92] Q. L. Zeng, Z. Kuang, S. B. Wu, and J. Yang, “A Method of Ultrasonic Finger Gesture Recognition Based on the Micro-Doppler Effect,” *APPLIED SCIENCES-BASEL*, vol. 9, no. 11, 2019, doi: 10.3390/app9112314.

[93] Q. L. Zeng, Z. Kuang, S. B. Wu, and J. Yang, “A Method of Ultrasonic Finger Gesture Recognition Based on the Micro-Doppler Effect,” *APPLIED SCIENCES-BASEL*, vol. 9, no. 11, 2019, doi: 10.3390/app9112314.

[94] Q. Zeng, Z. Kuang, S. Wu, and J. Yang, “A method of ultrasonic finger gesture recognition based on the micro-doppler effect,” *Applied Sciences*, vol. 9, no. 11, p. 2314, 2019.

[95] K. K. Liu, X. X. Liu, and X. L. Li, “Guoguo: Enabling Fine-Grained Smartphone Localization via Acoustic Anchors,” *IEEE Trans Mob Comput*, vol. 15, no. 5, pp. 1144–1156, 2016, doi: 10.1109/TMC.2015.2451628.

[96] K. K. Liu, X. X. Liu, and X. L. Li, “Guoguo: Enabling Fine-Grained Smartphone Localization via Acoustic Anchors,” *IEEE Trans Mob Comput*, vol. 15, no. 5, pp. 1144–1156, 2016, doi: 10.1109/TMC.2015.2451628.

[97] K. K. Liu, X. X. Liu, and X. L. Li, “Guoguo: Enabling Fine-Grained Smartphone Localization via Acoustic Anchors,” *IEEE Trans Mob Comput*, vol. 15, no. 5, pp. 1144–1156, 2016, doi: 10.1109/TMC.2015.2451628.

[98] W. C. Huang *et al.*, “Swadloon: Direction Finding and Indoor Localization Using Acoustic Signal by Shaking Smartphones,” *IEEE Trans Mob Comput*, vol. 14, no. 10,

2015, doi: 10.1109/TMC.2014.2377717.

[99] W. C. Huang *et al.*, “Swadloon: Direction Finding and Indoor Localization Using Acoustic Signal by Shaking Smartphones,” *IEEE Trans Mob Comput*, vol. 14, no. 10, 2015, doi: 10.1109/TMC.2014.2377717.

[100] W. C. Huang *et al.*, “Swadloon: Direction Finding and Indoor Localization Using Acoustic Signal by Shaking Smartphones,” *IEEE Trans Mob Comput*, vol. 14, no. 10, 2015, doi: 10.1109/TMC.2014.2377717.

[101] Harikesh, S. S. Chauhan, A. Basu, M. P. Abegaonkar, and S. K. Koul, “Through the Wall Human Subject Localization and Respiration Rate Detection Using Multichannel Doppler Radar,” *IEEE Sens J*, vol. 21, no. 2, pp. 1510–1518, 2021, doi: 10.1109/JSEN.2020.3016755.

[102] X. J. Cai, P. H. Wang, L. Du, Z. H. Cui, W. S. Zhang, and J. J. Chen, “Multi-Objective Three-Dimensional DV-Hop Localization Algorithm With NSGA-II,” *IEEE Sens J*, vol. 19, no. 21, pp. 10003–10015, 2019, doi: 10.1109/JSEN.2019.2927733.

[103] Y. P. Ding, Y. H. Sun, G. W. Huang, R. J. Liu, X. L. Yu, and X. M. Xu, “Human Target Localization Using Doppler Through-Wall Radar Based on Micro-Doppler Frequency Estimation,” *IEEE Sens J*, vol. 20, no. 15, pp. 8778–8788, 2020, doi: 10.1109/JSEN.2020.2983104.

[104] C. Y. Peng, G. B. Shen, and Y. G. Zhang, “BeepBeep: A High-Accuracy Acoustic-Based System for Ranging and Localization Using COTS Devices,” *ACM TRANSACTIONS ON EMBEDDED COMPUTING SYSTEMS*, vol. 11, no. 1, 2012, doi: 10.1145/2146417.2146421.

[105] C. Y. Peng, G. B. Shen, and Y. G. Zhang, “BeepBeep: A High-Accuracy Acoustic-Based System for Ranging and Localization Using COTS Devices,” *ACM TRANSACTIONS ON EMBEDDED COMPUTING SYSTEMS*, vol. 11, no. 1, 2012, doi: 10.1145/2146417.2146421.

[106] C. Y. Peng, G. B. Shen, and Y. G. Zhang, “BeepBeep: A High-Accuracy Acoustic-Based System for Ranging and Localization Using COTS Devices,” *ACM TRANSACTIONS ON EMBEDDED COMPUTING SYSTEMS*, vol. 11, no. 1, 2012, doi:

10.1145/2146417.2146421.

[107] D. Pastina, F. Colone, T. Martelli, and P. Falcone, “Parasitic Exploitation of Wi-Fi Signals for Indoor Radar Surveillance,” *IEEE Trans Veh Technol*, vol. 64, no. 4, pp. 1401–1415, 2015, doi: 10.1109/TVT.2015.2392936.

[108] C. Liu, S. N. Jiang, S. Zhao, and Z. W. Guo, “Infrastructure-Free Indoor Pedestrian Tracking with Smartphone Acoustic-Based Enhancement,” *SENSORS*, vol. 19, no. 11, 2019, doi: 10.3390/s19112458.

[109] C. Liu, S. N. Jiang, S. Zhao, and Z. W. Guo, “Infrastructure-Free Indoor Pedestrian Tracking with Smartphone Acoustic-Based Enhancement,” *SENSORS*, vol. 19, no. 11, 2019, doi: 10.3390/s19112458.

[110] C. Liu, S. N. Jiang, S. Zhao, and Z. W. Guo, “Infrastructure-Free Indoor Pedestrian Tracking with Smartphone Acoustic-Based Enhancement,” *SENSORS*, vol. 19, no. 11, 2019, doi: 10.3390/s19112458.

[111] M. Kawato and K. Fujinami, “Acoustic-sensing-based Gesture Recognition Hierarchical Classifier,” *SENSORS AND MATERIALS*, vol. 32, no. 9, pp. 2981–2998, 2020, doi: 10.18494/SAM.2020.2878.

[112] Y. C. Jin *et al.*, “SonicASL: An Acoustic-based Sign Language Gesture Recognizer Using Earphones,” *PROCEEDINGS OF THE ACM ON INTERACTIVE MOBILE WEARABLE AND UBIQUITOUS TECHNOLOGIES-IMWUT*, vol. 5, no. 2, 2021, doi: 10.1145/3463519.

[113] L. Wang *et al.*, “Watching Your Phone’s Back: Gesture Recognition by Sensing Acoustical Structure-borne Propagation,” *PROCEEDINGS OF THE ACM ON INTERACTIVE MOBILE WEARABLE AND UBIQUITOUS TECHNOLOGIES-IMWUT*, vol. 5, no. 2, 2021, doi: 10.1145/3463522.

[114] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman, and L. Goldstein, “Recognizing articulatory gestures from speech for robust speech recognition,” *JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA*, vol. 131, no. 3, pp. 2270–2287, 2012, doi: 10.1121/1.3682038.

[115] Y. W. Wang, J. X. Shen, and Y. Q. Zheng, “Push the Limit of Acoustic Gesture

Recognition," *IEEE Trans Mob Comput*, vol. 21, no. 5, pp. 1798–1811, 2022, doi: 10.1109/TMC.2020.3032278.

[116] Y. Wang, J. Shen, and Y. Zheng, "Push the Limit of Acoustic Gesture Recognition," *IEEE Trans Mob Comput*, vol. 1233, no. c, pp. 1–1, 2020, doi: 10.1109/tmc.2020.3032278.

[117] T. Amesaka, H. Watanabe, M. Sugimoto, and B. Shizuki, "Gesture Recognition Method Using Acoustic Sensing on Usual Garment," *PROCEEDINGS OF THE ACM ON INTERACTIVE MOBILE WEARABLE AND UBIQUITOUS TECHNOLOGIES-IMWUT*, vol. 6, no. 2, 2022, doi: 10.1145/3534579.

[118] N. Siddiqui and R. H. M. Chan, "Hand Gesture Recognition Using Multiple Acoustic Measurements at Wrist," *IEEE Trans Hum Mach Syst*, vol. 51, no. 1, pp. 56–62, 2021, doi: 10.1109/THMS.2020.3041201.

[119] N. Siddiqui and R. H. M. Chan, "Multimodal hand gesture recognition using single IMU and acoustic measurements at wrist," *PLoS One*, vol. 15, no. 1, 2020, doi: 10.1371/journal.pone.0227039.

[120] H. J. Ai, K. F. Tang, L. L. Han, Y. F. Wang, and S. Zhang, "DuG: Dual speaker-based acoustic gesture recognition for humanoid robot control," *Inf Sci (N Y)*, vol. 504, pp. 84–94, 2019, doi: 10.1016/j.ins.2019.06.065.

[121] W. Wang and A. X. Liu, "Device-Free Gesture Tracking Using Acoustic Signals Limitations of Prior Art," pp. 82–94, 2016.

[122] B. S. Moreira, A. Perkusich, and S. O. D. Luiz, "An Acoustic Sensing Gesture Recognition System Design Based on a Hidden Markov Model," *SENSORS*, vol. 20, no. 17, 2020, doi: 10.3390/s20174803.

[123] G. Luo, P. L. Yang, M. S. Chen, and P. Li, "HCI on the Table: Robust Gesture Recognition Using Acoustic Sensing in Your Hand," *IEEE ACCESS*, vol. 8, pp. 31481–31498, 2020, doi: 10.1109/ACCESS.2020.2973305.

[124] W. H. Jiang, S. Li, Y. C. Zhao, H. W. Tu, and C. Y. Liu, "Fine-grained hand gesture recognition based on active acoustic signal for VR systems," *CCF TRANSACTIONS ON PERVASIVE COMPUTING AND INTERACTION*, vol. 2, no. 4, pp. 329–339, 2020,

doi: 10.1007/s42486-020-00048-w.

[125] L. Kessous, G. Castellano, and G. Caridakis, “Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis,” *JOURNAL ON MULTIMODAL USER INTERFACES*, vol. 3, no. 1–2, pp. 33–48, 2010, doi: 10.1007/s12193-009-0025-5.

[126] M. Zhang, Q. Dai, P. Yang, J. Xiong, C. Tian, and C. Xiang, “idial: Enabling a virtual dial plate on the hand back for around-device interaction,” *Proc ACM Interact Mob Wearable Ubiquitous Technol*, vol. 2, no. 1, pp. 1–20, 2018.

[127] M. Chen, P. Yang, S. Cao, M. Zhang, and P. Li, “WritePad: Consecutive number writing on your hand with smart acoustic sensing,” *IEEE Access*, vol. 6, pp. 77240–77249, 2018.

[128] S. Cao, P. Yang, X. Li, M. Chen, and P. Zhu, “ipand: Accurate gesture input with smart acoustic sensing on hand,” in *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, IEEE, 2018, pp. 1–3.

[129] Y. Iravantchi, M. Goel, and C. Harrison, “BeamBand: Hand gesture sensing with ultrasonic beamforming,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–10.

[130] Y. Sang, L. Shi, and Y. Liu, “Micro hand gesture recognition system using ultrasonic active sensing,” *IEEE Access*, vol. 6, pp. 49339–49347, 2018.

[131] G. M. Zhang, D. R. Braden, D. M. Harvey, and D. R. Burton, “Acoustic time-frequency domain imaging,” *JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA*, vol. 128, no. 5, pp. EL323–EL328, 2010, doi: 10.1121/1.3505760.

[132] T.-N. Nguyen, H.-H. Huynh, and J. Meunier, “Static hand gesture recognition using artificial neural network,” *Journal of Image and Graphics*, vol. 1, no. 1, pp. 34–38, 2013.

[133] C. Zhang *et al.*, “FingerSound: Recognizing unistroke thumb gestures using a ring,” *Proc ACM Interact Mob Wearable Ubiquitous Technol*, vol. 1, no. 3, pp. 1–19, 2017.

[134] Y. Wang, J. Shen, and Y. Zheng, “Push the limit of acoustic gesture recognition,” *IEEE Trans Mob Comput*, vol. 21, no. 5, pp. 1798–1811, 2020.

[135] W. Ruan, Q. Z. Sheng, L. Yang, T. Gu, P. Xu, and L. Shangguan, “AudioGest: enabling

fine-grained hand gesture detection by decoding echo signal," in *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*, 2016, pp. 474–485.

[136] Y. Sang, L. X. Shi, and Y. M. Liu, "Micro Hand Gesture Recognition System Using Ultrasonic Active Sensing," *IEEE ACCESS*, vol. 6, pp. 49339–49347, 2018, doi: 10.1109/ACCESS.2018.2868268.

[137] C. R. Pittman and J. J. LaViola Jr, "Multiwave: Complex hand gesture recognition using the doppler effect," in *Proceedings of the 43rd Graphics Interface Conference*, 2017, pp. 97–106.

[138] Y. Zou, Q. Yang, Y. Han, D. Wang, J. Cao, and K. Wu, "Acoudigits: Enabling users to input digits in the air," in *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom*, IEEE, 2019, pp. 1–9.

[139] Y. Sang, L. Shi, and Y. Liu, "Micro hand gesture recognition system using ultrasonic active sensing," *IEEE Access*, vol. 6, pp. 49339–49347, 2018, doi: 10.1109/ACCESS.2018.2868268.

[140] N. Kim and J. Lee, "Towards grip sensing for commodity smartphones through acoustic signature," in *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, 2017, pp. 105–108.

[141] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[142] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[1] B. Xie, X. He, and Y. Li, "RGB-D static gesture recognition based on convolutional neural network," *J. Eng.*, vol. 2018, no. 16, pp. 1515–1520, 2018.

[2] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li, "Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition," in *Proceedings of the IEEE conference on computer vision and*

pattern recognition workshops, 2016, pp. 56–64.

[3] J. Wan, G. Guo, and S. Z. Li, "Explore efficient local features from RGB-D data for one-shot learning gesture recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1626–1639, 2015.

[4] Y. Li *et al.*, "Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model," in *2016 23rd international conference on pattern recognition (ICPR)*, 2016, pp. 25–30.

[5] R. G. Crespo, E. Verdú, M. Khari, and A. K. Garg, "Gesture recognition of RGB and RGB-D static images using convolutional neural networks," *IJIMAI*, vol. 5, no. 7, pp. 22–27, 2019.

[6] J. Wan, Q. Ruan, W. Li, and S. Deng, "One-shot learning gesture recognition from RGB-D data using bag of features," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 2549–2582, 2013.

[7] Z. Ren, J. Meng, and J. Yuan, "Depth camera based hand gesture recognition and its applications in human-computer-interaction," in *Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on*, 2011, pp. 1–5.

[8] D. Xu, Y.-L. Chen, C. Lin, X. Kong, and X. Wu, "Real-time dynamic gesture recognition system based on depth perception for robot navigation," in *2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2012, pp. 689–694.

[9] T. Mantecón, C. R. Del-Blanco, F. Jaureguizar, and N. García, "A real-time gesture recognition system using near-infrared imagery," *PLoS One*, vol. 14, no. 10, p. e0223320, 2019.

[10] F. Erden and A. E. Cetin, "Hand gesture based remote control system using infrared sensors and a camera," *IEEE Trans. Consum. Electron.*, vol. 60, no. 4, pp. 675–680, 2014.

[11] K. Geng and G. Yin, "Using deep learning in infrared images to enable human gesture recognition for autonomous vehicles," *IEEE Access*, vol. 8, pp. 88227–

88240, 2020.

[12] R. Cutler and M. Turk, “View-based interpretation of real-time optical flow for gesture recognition,” in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 416–421.

[13] J.-H. Kim, N. D. Thang, and T.-S. Kim, “3-d hand motion tracking and gesture recognition using a data glove,” in *2009 IEEE International Symposium on Industrial Electronics*, 2009, pp. 1013–1018.

[14] J. Weissmann and R. Salomon, “Gesture recognition for virtual reality applications using data gloves and neural networks,” in *IJCNN’99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)*, 1999, vol. 3, pp. 2043–2046.

[15] M. B. Holte, T. B. Moeslund, and P. Fihl, “View-invariant gesture recognition using 3D optical flow and harmonic motion context,” *Comput. Vis. Image Underst.*, vol. 114, no. 12, pp. 1353–1361, 2010.

[16] K. Czuszyński, J. Rumiński, and A. Kwaśniewska, “Gesture recognition with the linear optical sensor and recurrent neural networks,” *IEEE Sens. J.*, vol. 18, no. 13, pp. 5429–5438, 2018.

[17] J. Li, H. Liu, and H. Sun, “On a gesture-computing technique using electromagnetic waves,” *arXiv Prepr. arXiv1708.02848*, 2017.

[18] M. Tan, J. Zhou, K. Xu, Z. Peng, and Z. Ma, “Static hand gesture recognition with electromagnetic scattered field via complex attention convolutional neural network,” *IEEE Antennas Wirel. Propag. Lett.*, vol. 19, no. 4, pp. 705–709, 2020.

[19] J. M. Garcia and A. L. Topa, “Gesture recognition by electromagnetic-wave reflection,” *IEEE, April*, 2016.

[20] A. Sluÿters, S. Lambot, and J. Vanderdonckt, “Hand gesture recognition for an off-the-shelf radar by electromagnetic modeling and inversion,” in *27th International Conference on Intelligent User Interfaces*, 2022, pp. 506–522.

[21] H.-Y. Li *et al.*, “Intelligent electromagnetic sensing with learnable data

acquisition and processing," *Patterns*, vol. 1, no. 1, 2020.

[22] A. Das, I. Tashev, and S. Mohammed, "Ultrasound based gesture recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 406–410.

[23] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust.*, vol. 32, no. 2, pp. 236–243, 1984.

[24] C. Yiallourides and P. P. Parada, "Low power ultrasonic gesture recognition for mobile handsets," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2697–2701.

[25] W. Ding and L. He, "Adaptive multi-scale detection of acoustic events," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 294–306, 2020.

[26] W. Wang and A. X. Liu, "Device-Free Gesture Tracking Using Acoustic Signals Limitations of Prior Art," pp. 82–94, 2016.

[27] L. Yu, H. Abuella, M. Z. Islam, J. F. O'Hara, C. Crick, and S. Ekin, "Gesture recognition using reflected visible and infrared lightwave signals," *IEEE Trans. Human-Machine Syst.*, vol. 51, no. 1, pp. 44–55, 2021.

[28] Y. Wang, J. Shen, and Y. Zheng, "Push the Limit of Acoustic Gesture Recognition," *IEEE Trans. Mob. Comput.*, vol. 1233, no. c, pp. 1–1, 2020.

[29] N. Kim and J. Lee, "Towards grip sensing for commodity smartphones through acoustic signature," in *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, 2017, pp. 105–108.

[30] S. Yun, Y. Chen, H. Zheng, L. Qiu, and W. Mao, "Strata : Fine-Grained Acoustic-based Device-Free Tracking," in *MobiSys '17*, 2017, pp. 15–28.

[31] K. Ling, H. Dai, Y. Liu, and A. X. Liu, "Ultragesture: Fine-grained gesture sensing and recognition," in *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking, SECON 2018*, 2018.

[32] S. A. Jones, "Fundamental sources of error and spectral broadening in Doppler

ultrasound signals," *Crit. Rev. Biomed. Eng.*, vol. 21, p. 399, 1993.

[33] W.-Q. Zhang, Y. Deng, L. He, and J. Liu, "Variant time-frequency cepstral features for speaker recognition," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[34] T.-N. Nguyen, D.-H. Vo, H.-H. Huynh, and J. Meunier, "Geometry-based static hand gesture recognition using support vector machine," in *2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*, 2014, pp. 769–774.

[35] X. Zhang, Z. Yang, T. Chen, D. Chen, and M.-C. Huang, "Cooperative sensing and wearable computing for sequential hand gesture recognition," *IEEE Sens. J.*, vol. 19, no. 14, pp. 5775–5783, 2019.

[36] Z. Wang *et al.*, "Hand Gesture Recognition Based on Active Ultrasonic Sensing of Smartphone: A Survey," *IEEE Access*, vol. 7, pp. 111897–111922, 2019.

[37] Y. Qifan, T. Hao, Z. Xuebing, L. Yin, and Z. Sanfeng, "Dolphin: Ultrasonic-based gesture recognition on smartphone platform," in *2014 IEEE 17th International Conference on Computational Science and Engineering*, 2014, pp. 1461–1468.

[38] W. Ruan, Q. Z. Sheng, L. Yang, T. Gu, P. Xu, and L. Shangguan, "AudioGest: enabling fine-grained hand gesture detection by decoding echo signal," in *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*, 2016, pp. 474–485.

[39] Y. Sang, L. Shi, and Y. Liu, "Micro hand gesture recognition system using ultrasonic active sensing," *IEEE Access*, vol. 6, pp. 49339–49347, 2018.

[40] Q. Zeng, Z. Kuang, S. Wu, and J. Yang, "A method of ultrasonic finger gesture recognition based on the micro-doppler effect," *Appl. Sci.*, vol. 9, no. 11, p. 2314, 2019.

[41] C. Peng, G. Shen, Y. Zhang, Y. Li, and K. Tan, "Beepbeep: a high accuracy acoustic ranging system using cots mobile devices," in *Proceedings of the 5th*

international conference on Embedded networked sensor systems, 2007, pp. 1–14.

[42] L. Wang *et al.*, "WiTrace: Centimeter-level passive gesture tracking using OFDM signals," *IEEE Trans. Mob. Comput.*, vol. 20, no. 4, pp. 1730–1745, 2019.

[43] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proceedings of the 19th annual international conference on Mobile computing & networking*, 2013, pp. 27–38.

[44] R. Nandakumar, V. Iyer, D. Tan, and S. Gollakota, "FingerIO : Using Active Sonar for Fine-Grained Finger Tracking," 2016.

[45] J. Armstrong, "OFDM for optical communications," *J. Light. Technol.*, vol. 27, no. 3, pp. 189–204, 2009.

[46] V. C. Chen, *The micro-Doppler effect in radar*. Artech house, 2019.

[47] D. T. Sidhant Gupta1, 2, Dan Morris1, Shwetak N Patel1, 2, "SoundWave: Using the Doppler Effect to Sense Gestures," pp. 1911–1914, 2012.

[48] K. Kalgaonkar and B. Raj, "One-handed gesture recognition using ultrasonic Doppler sonar," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 1889–1892.

[49] S. P. Tarzia, R. P. Dick, P. A. Dinda, and G. Memik, "Sonar-based measurement of user presence and attention," in *Proceedings of the 11th international conference on Ubiquitous computing*, 2009, pp. 89–92.

[50] D. O. Olguin, J. A. Paradiso, and A. Pentland, "Wearable communicator badge: Designing a new platform for revealing organizational dynamics," in *Proceedings of the 10th international symposium on wearable computers (student colloquium)*, 2006, pp. 4–6.

[51] H. Chen, T. Ballal, M. Saad, and T. Y. Al-Naffouri, "Angle-of-arrival-based gesture recognition using ultrasonic multi-frequency signals," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 16–20.

[52] P. Gupta and S. P. Kar, "MUSIC and improved MUSIC algorithm to estimate

direction of arrival," in *2015 International Conference on Communications and Signal Processing (ICCSP)*, 2015, pp. 757–761.

[53] G. R. Naik and W. Wang, "Blind source separation," *Berlin: Springer*, vol. 10, pp. 973–978, 2014.

[54] A. N. Popper, R. R. Fay, and A. N. Popper, *Sound source localization*, vol. 25. Springer, 2005.

[55] W. Shi, S. A. Vorobyov, and Y. Li, "ULA fitting for sparse array design," *IEEE Trans. Signal Process.*, vol. 69, pp. 6431–6447, 2021.

[56] R. L. Haupt, *Antenna arrays: a computational approach*. John Wiley & Sons, 2010.

[57] Y. Wu and G. Li, "Intelligent Robot English Speech Recognition Method Based on Online Database," *J. Inf. Knowl. Manag.*, vol. 21, no. Supp02, p. 2240012, 2022.

[58] H.-B. Zhang *et al.*, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, 2019.

[59] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[60] O. Khrystoslavenko and R. Grubliauskas, "Simulation of Room Acoustics Using Comsol Multiphysics," in *Proceedings of the 20th Conference for Junior Researchers „Science–Future of Lithuania*, 2017.

[61] C. Multiphysics, "Introduction to COMSOL multiphysics®," *COMSOL Multiphysics, Burlington, MA, accessed Feb*, vol. 9, no. 2018, p. 32, 1998.

[62] J. Gu and Y. Jing, "Modeling of wave propagation for medical ultrasound: a review," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 62, no. 11, pp. 1979–1992, 2015.

[63] A. R. N. Meidani and M. Hasan, "Mathematical and physical modelling of bubble growth due to ultrasound," *Appl. Math. Model.*, vol. 28, no. 4, pp. 333–351, 2004.

[64] S. Acosta, G. Uhlmann, and J. Zhai, “Nonlinear ultrasound imaging modeled by a Westervelt equation,” *SIAM J. Appl. Math.*, vol. 82, no. 2, pp. 408–426, 2022.

[65] A. C. Eringen, “Elasto-dynamic problem concerning the spherical cavity,” *Q. J. Mech. Appl. Math.*, vol. 10, no. 3, pp. 257–270, 1957.

[66] J. G. Švec and S. Granqvist, “Tutorial and guidelines on measurement of sound pressure level in voice and speech,” *J. Speech, Lang. Hear. Res.*, vol. 61, no. 3, pp. 441–461, 2018.

[67] L. Aguiar-Conraria and M. J. Soares, “The continuous wavelet transform: A primer,” NIPE-Universidade do Minho, 2011.

[68] Y. Yu, X. Si, C. Hu, and J. Zhang, “A review of recurrent neural networks: LSTM cells and network architectures,” *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, 2019.

[69] S. Siami-Namini, N. Tavakoli, and A. S. Namin, “The performance of LSTM and BiLSTM in forecasting time series,” in *2019 IEEE International conference on big data (Big Data)*, 2019, pp. 3285–3292.

[70] W. Lu, J. Li, J. Wang, and L. Qin, “A CNN-BiLSTM-AM method for stock price prediction,” *Neural Comput. Appl.*, vol. 33, pp. 4741–4753, 2021.

[71] F. Kong, J. Deng, and Z. Fan, “Gesture recognition system based on ultrasonic FMCW and ConvLSTM model,” *Measurement*, vol. 190, p. 110743, 2022.

[72] S. Kim, S. Hong, M. Joh, and S. Song, “Deeprain: Convlstm network for precipitation prediction using multichannel radar data,” *arXiv Prepr. arXiv1711.02316*, 2017.

[73] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, “Bi-directional ConvLSTM U-Net with densley connected convolutions,” in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, p. 0.

[74] Z. Lin, M. Li, Z. Zheng, Y. Cheng, and C. Yuan, “Self-attention convlstm for spatiotemporal prediction,” in *Proceedings of the AAAI conference on artificial*

intelligence, 2020, vol. 34, no. 07, pp. 11531–11538.

[75] P.-C. Kotsias, J. Arús-Pous, H. Chen, O. Engkvist, C. Tyrchan, and E. J. Bjerrum, "Direct steering of de novo molecular generation using descriptor conditional recurrent neural networks (cRNNs)," 2019.

[76] A. Arnault and N. Riche, "CRNNs for Urban Sound Tagging with spatiotemporal context," *arXiv Prepr. arXiv2008.10413*, 2020.

[77] B. Liebl and M. Burghardt, "On the accuracy of CRNNs for line-based OCR: A multi-parameter evaluation," *arXiv Prepr. arXiv2008.02777*, 2020.

[78] M. Odusami, R. Maskeliūnas, R. Damaševičius, and T. Krilavičius, "Analysis of features of Alzheimer's disease: Detection of early stage from functional brain changes in magnetic resonance images using a finetuned ResNet18 network," *Diagnostics*, vol. 11, no. 6, p. 1071, 2021.

[79] A. Ullah, H. Elahi, Z. Sun, A. Khatoon, and I. Ahmad, "Comparative analysis of AlexNet, ResNet18 and SqueezeNet with diverse modification and arduous implementation," *Arab. J. Sci. Eng.*, pp. 1–21, 2022.

[80] C. Pittman, P. Wisniewski, C. Brooks, and J. J. LaViola Jr, "Multiwave: Doppler effect based gesture recognition in multiple dimensions," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2016, pp. 1729–1736.

[81] D. Kundu, "Modified MUSIC algorithm for estimating DOA of signals," *Signal Processing*, vol. 48, no. 1, pp. 85–90, 1996.

[82] B. Friedlander, "A sensitivity analysis of the MUSIC algorithm," *IEEE Trans. Acoust.*, vol. 38, no. 10, pp. 1740–1751, 1990.

[83] A. El Gonnouni, M. Martinez-Ramon, J. L. Rojo-Álvarez, G. Camps-Valls, A. R. Figueiras-Vidal, and C. G. Christodoulou, "A support vector machine MUSIC algorithm," *IEEE Trans. Antennas Propag.*, vol. 60, no. 10, pp. 4901–4910, 2012.

[84] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *arXiv Prepr.*

arXiv1905.03554, 2019.

[85] W. Tang, G. Long, L. Liu, T. Zhou, J. Jiang, and M. Blumenstein, "Rethinking 1d-cnn for time series classification: A stronger baseline," *arXiv Prepr. arXiv2002.10061*, pp. 1–7, 2020.

[86] L. Eren, T. Ince, and S. Kiranyaz, "A generic intelligent bearing fault diagnosis system using compact adaptive 1D CNN classifier," *J. Signal Process. Syst.*, vol. 91, pp. 179–189, 2019.

[87] Y. Zou, Q. Yang, Y. Han, D. Wang, J. Cao, and K. Wu, "Acoudigits: Enabling users to input digits in the air," in *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom*, 2019, pp. 1–9.

[88] G. Luo, P. Yang, M. Chen, and P. Li, "HCI on the table: robust gesture recognition using acoustic sensing in your hand," *IEEE Access*, vol. 8, pp. 31481–31498, 2020.

[89] A. Ferrone *et al.*, "Wearable band for hand gesture recognition based on strain sensors," in *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2016, pp. 1319–1322.

[90] G. Luo, M. Chen, P. Li, M. Zhang, and P. Yang, "SoundWrite II: Ambient acoustic sensing for noise tolerant device-free gesture recognition," in *2017 IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS)*, 2017, pp. 121–126.

[91] S. Gutta, I. F. Imam, and H. Wechsler, "Hand gesture recognition using ensembles of radial basis function (RBF) networks and decision trees," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 11, no. 06, pp. 845–872, 1997.

[92] K. Srinivas and M. K. Rajagopal, "Study of hand gesture recognition and classification," *Asian J. Pharm. Clin. Res.*, pp. 25–30, 2017.

[93] H. Watanabe, T. Terada, and M. Tsukamoto, "Gesture recognition method utilizing ultrasonic active acoustic sensing," *J. Inf. Process.*, vol. 25, pp. 331–340, 2017.

[94] J. Huang, F. Di Troia, and M. Stamp, "Acoustic Gait Analysis using Support

Vector Machines.," in *ICISSP*, 2018, pp. 545–552.

[95] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv Prepr. arXiv1502.03167*, 2015.

[96] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv Prepr. arXiv1412.6980*, 2014.

[97] S. Xiao, X. Ji, C. Yan, Z. Zheng, and W. Xu, "MicPro: Microphone-based Voice Privacy Protection," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 1302–1316.

[98] S. Gray, "Always on: privacy implications of microphone-enabled devices," in *Future of privacy forum*, 2016, pp. 1–10.

[99] A. Nehorai and E. Paldi, "Acoustic vector-sensor array processing," *IEEE Trans. signal Process.*, vol. 42, no. 9, pp. 2481–2491, 1994.

[100] J.-A. Luo, X.-P. Zhang, Z. Wang, and X.-P. Lai, "On the accuracy of passive source localization using acoustic sensor array networks," *IEEE Sens. J.*, vol. 17, no. 6, pp. 1795–1809, 2017.

[101] H. Chen and J. Zhao, "On locating low altitude moving targets using a planar acoustic sensor array," *Appl. Acoust.*, vol. 64, no. 11, pp. 1087–1101, 2003.

[102] L. Hang, C. He, and B. Wu, "Novel distributed optical fiber acoustic sensor array for leak detection," *Opt. Eng.*, vol. 47, no. 5, p. 54401, 2008.

[103] C. Zhang *et al.*, "FingerPing: Recognizing fine-grained hand poses using active acoustic on-body sensing," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–10.

[104] Q. Yang, H. Fu, Y. Zou, and K. Wu, "A novel finger-assisted touch-free text input system without training," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, 2018, p. 533.