AI enable Wireless Sensing for Remote Speech Recognition

Zikang Chen

Submitted in fulfilment of the requirements for the
Master of Science(R)

School of Engineering
College of Science and Engineering
University of Glasgow

October 2023

Abstract

Contactless health monitoring is becoming an area of significant attention, especially after the impact of the COVID-19 pandemic. With the development of RF sensing technology, its application prospects in healthcare have garnered significant attention. Radio Frequency (RF) sensing techniques such as ultra-wideband (UWB) radar and Frequency Modulated Continuous Wave (FMCW) radar are used in many contactless monitoring scenarios. Compared to contact-based health monitoring methods, RF sensing technology offers users a non-intrusive experience, which enhances patients' quality of life. Additionally, when compared to traditional non-contact monitoring technologies like imaging, RF sensing provides superior privacy protection, which effectively addresses users' concerns. Artificial intelligence technology is also advancing rapidly and has gained significant attention due to its outstanding performance across various application scenarios. The integration of artificial intelligence with RF sensing technology can offer excellent and convenient solutions for future healthcare. This thesis proposed a multimodal speech recognition system of UWB radar data, acoustic information and visual information. The proposed multimodal approach achieves 96.89% accuracy in the word classification task, which indicates the performance improvement of the incorporation of the UWB for the multimodal system compared to the single-modal system.

Contents

List of Publications

- **Zikang Chen**, C. Tang, Y. Ge, M. Imran and Q. H. Abbasi, "Integrating RF-Visual Technologies for Improved Speech Recognition in Hearing Aids," 2023 IEEE MTT-S International Microwave Biomedical Conference (IMBioC)
- Yao Ge, Chong Tang, Haobo Li, **Zikang Chen**, Jingyan Wang, Wenda Li, Jonathan Cooper, Kevin Chetty, Daniele Faccio, Muhammad Imran, Qammer H. Abbasi "A comprehensive multimodal dataset for contactless lip reading and acoustic analysis" Scientific Data volume 10, Article number: 895 (2023)
- Yao Ge, Wenda Li, Muhammad Farooq, Adnan Qayyum, Jingyan Wang, **Zikang Chen** et al."LoGait: LoRa Sensing System of Human Gait Recognition Using Dynamic Time Warping," in IEEE Sensors Journal, vol. 23, no. 18, Sept.15, 2023

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to express my sincerest gratitude to my supervisor, Prof. Qammer H. Abbasi, for his invaluable guidance and support throughout this academic journey. The guidance from him has guided the direction of my research and provided countless valuable help for my studies during this period.

Secondly, I would like to thank my colleagues who shared valuable experiences and provided assistance to me during this research time. Especially Yao Ge who guided me a lot when I first came to the University of Glasgow.

Last but not least, I am immensely grateful to my parents who encouraged me and supported me in my life and the MSc study.

Declaration

Zikang Chen declare that, with the exception of chapters 1 and 2, which contain introductory material, all work in this thesis was carried out by the author unless otherwise explicitly stated.

Chapter 1

Introduction

Healthcare applications refer to technologies and systems used to monitor, diagnose and manage health conditions. Healthcare plays a crucial role in improving the quality of patients' lives. The importance of advanced healthcare applications has received much attention in recent years, especially after the impact of the Coronavirus disease 2019 (COVID-19) pandemic. The COVID-19 pandemic highlighted the need for advanced healthcare technologies, especially remote monitoring and diagnosis techniques as COVID-19 is highly contagious and can cause severe symptoms, remote monitoring and early detection of symptoms becoming more important to prevent the spread of COVID-19.

In traditional healthcare systems, contact-based techniques are employed for the monitoring of vital signs such as the airflow measurement for respiratory rate monitoring and electrodes to record electrocardiogram (ECG) for heart rate monitoring. These conventional systems provide accurate and reliable monitoring for patients. However, traditional contact-based methods also come with certain limitations. Wearable devices need to be worn constantly, which can be uncomfortable and inconvenient. In particular, the elderly may find it challenging to wear or manage these devices consistently. However, healthcare is more crucial for the elderly, which indicates the importance of the improved system. Privacy is also one of the concerns while video monitoring methods are employed, such as body temperature monitor using infrared cameras and fall detection using optical cameras. Moreover, recharging and maintenance can further reduce the usability of these devices in long-term healthcare.

Radio frequency (RF) sensing provides a promising alternative to address these limitations. RF sensing techniques use the transmitter and the receiver that emit radio waves and detect the reflected radio waves which are affected by human activities in the region of interest (ROI). Common RF sensing techniques include ultra-wideband (UWB) radar, Frequency Modulated Continuous Wave (FMCW) radar and WiFi. Through analysing the RF signal affected by human activities, RF sensing based system provides continuous and non-invasive monitoring, which can

address the concerns of privacy and comfort.

The rapid development of artificial intelligence (AI) has also attracted widespread attention. AI is now being employed in various fields, especially in image recognition and natural language processing areas. The application of artificial intelligence is increasing rapidly in the healthcare domain, such as the analysis and diagnosis of medical images [1]. One of the most significant advantages of AI is its capacity to process a large number of characteristics simultaneously, which makes it an effective approach for data processing and analysis in different systems. By integrating AI to process RF sensing data, the system becomes more capable of extracting comprehensive information and improving computing time consumption. Therefore, integrating AI as the feature extraction model with RF sensing provides a more capable solution for the RF-based healthcare system.

Speech recognition also plays an important role in healthcare. It can not only support the patient with hearing impairment to understand others, but also improve the experience for both patients and doctors. It benefits medical conversations by accurately recording dialogues during consultations, allowing for comprehensive documentation of patients' symptoms. Additionally, it enables physicians to efficiently transcribe their diagnoses and notes, significantly saving time on administrative tasks. This efficiency simplifies procedures and improves the quality of patient care. Furthermore, speech recognition provides a hands-free method of interacting with various medical devices, which is particularly beneficial for patients with disabilities, ensuring they engage with healthcare services more independently and effectively.

This thesis aims to explore the potential of RF sensing in healthcare applications with the assistance of AI in data processing and evaluate its effectiveness compared to traditional methods. Specifically, this thesis proposed a multimodal speech recognition method composed of video, audio and UWB radar data. By incorporating the UWB data, this model aims to provide a more robust and accurate approach for the speech recognition system.

The remainder of this paper is organized as follows: Section 2 reviews the state of the arts of Healthcare applications, RF sensing systems and the common AI algorithm. Section 3 presents the proposed multi-modal speech recognition system, and Section 4 gives a summary of the thesis and states the direction of future experiments.

Chapter 2

Literature review

This chapter provides a comprehensive review of the state of the arts in the field of healthcare monitoring, including traditional contact methods and contactless methods. To explore the applications and limitations of these technologies in various healthcare projects, including vital signs detection, human activity recognition and speech recognition.

Vital signs are key indicators of the body's essential functions, typically including heart rate, respiratory rate, blood pressure, body temperature and blood oxygen saturation. Vital signs are important in indicating a person's health information and can be used to detect early signs of diseases [2], [3].

Heart rate is an indicator which reflects the daily physiological status most. A high resting heart rate is associated with increased cardiovascular and coronary mortality [4]. Heart Rate Variability (HRV) is the time difference between two beats, which indicates the balance between the sympathetic and parasympathetic nervous systems [5], [6].

HRV is also an important indicator of the state of the heart and the autonomic nervous system, and even a sign of sudden death [7]. Heart rate can be extracted from the electrocardiogram (ECG) and photoplethysmography (PPG). ECG signal is commonly used in medical diagnoses, which are recorded by electrodes attached to the body. The bioelectric generated by the heart is recorded with the electrodes and is then reconstructed as the ECG signal. However, the limitation of the ECG method is obvious. In order to collect accurate ECG signals, patients should stay still as electrodes are sensitive to any bioelectric that may caused by any movement. [8] Furthermore, the attached electrodes make patients inconvenient in daily life. For this reason, monitoring Heart rate through ECG is not an ideal solution in daily healthcare scenarios. A PPG sensor emits red or green light to the skin and records the changes in the intensity of the light reflected by or transmitted through the skin. As blood absorbs more light compared with surrounding tissue, the intensity of light becomes a great indicator of the blood flow, which

represents the cardiac cycle [9], [10]. The PPG sensor can be integrated into a small wearable device which makes it a convenient heart monitoring method [11].

Respiratory rate is a sensitive indicator in critical illnesses, especially in cases of illness that affect the respiratory system or the nervous system. Recording respiratory rate is very important for diagnosing obstructive sleep apnea (OSA). OSA can affect the blood oxygen level and disrupt sleep which can impact a person's mental state in the daytime and even increase the risk of stroke [12], [13]. However, the recording of respiratory rate gets little attention in routine healthcare monitoring compared to heart rate [14], which may lead to missed early diagnosis of these diseases. One of the factors that prevent daily and continuous monitoring of respiratory rate is that traditional methods of monitoring respiratory rate tend to be contact-based and need to be worn on the face or chest, which affects the user's daily life severely. The contact-based respiratory monitoring method includes capturing the breath sound, measuring the Chest and abdominal wall movements, measuring the airflow, and monitoring CO_2 or Blood oxygen saturation [15]. In paper [16], the breath sound capturing method is used to detect the sleep-disordered breath. Paper [2] also concludes several wearable techniques such as elastomeric plethysmography (EP), impedance plethysmography (IP) and respiratory inductive plethysmography (RIP), both measure the movement of the Chest and abdominal wall movements to monitor respiratory rate.

2.1 RF sensing techniques applications

RF sensing techniques provide a useful alternative in vital signs monitoring. Paper [17] uses a 77GHz FMCW radar to monitor heart rate and respiratory rate. This paper first unwrapped the signal to eliminate the phase jump caused by the phase change that is greater than π. Then they applied FFT over samples of chirps to detect the distance between the radar and the patient and the second FFT in the range of the patient to reveal the vibration caused by breathing and heartbeats. Through these methods, the authors achieved 94% correlation for breathing rate and 80% correlation for heart rate between the radar estimates and the reference wearable device. These signal processing procedures are the common method while processing radar data in common healthcare tasks. A WiFi-based vital sign monitoring method is proposed in [18]. The channel state information(CSI) is recorded using an off-the-shelf WiFi device. The result shows the WiFi-based system achieves 95% accuracy in heart rate estimation and 98% accuracy in respiratory rate estimation. Work in [19] proposed an in-vehicle vital sign monitoring system. This system is composed of a single-chip computer and an X4M05 impulse radio with 7.3GHz center frequency and 1.4GHz bandwidth. A Multi-Sequence VMD (MS-VMD) algorithm is designed to extract vital signs accurately for the in-vehicle environment. This system achieves a median error of 0.06 rpm for the respiratory rate monitoring and a median error of 0.6 bpm

for the heart rate monitoring. This work demonstrated a practical application scenario for the RF sensing based vital signs system. The accuracy of respiratory rate detection is assessed on hospitalized patients in paper [20]. The respiratory rate is monitored by 2.4 GHz Doppler radar and compared with both thoracic impedance measurement and inductive plethysmographic measurement of respiratory effort as references. The result shows the standard deviation and the root mean square of the difference between radar measurement and references are both below 2 breaths per minute, suggesting the feasibility of radar for respiratory rate monitoring.

2.1.1 Human Activity detection

Human activity detection tasks have also received attention in health care. One of the most popular detection tasks using wearable devices is fall detection. The Accelerometer is often used in contact-based fall detection systems. Papers [21] evaluated the performance of 13 detection algorithms using recorded fall data in real situations. Results show that there are sensitivity of 83% and a specificity of 53% on average using real recorded data, which were much lower than in simulation. Gyroscopes are also used with accelerometers in paper [22] to reduce false positives and false negatives. However, results show there are still 60% false positives when the subject lying down fast. As the accelerometer is often installed in smartphones, there are fall detection systems based on smartphones [23], [24]. Paper [24] detect the fall events by calculating the vertical acceleration and speed recorded by the accelerometer within a time window and defining the threshold through experiments. The average test false negative is 2.67% and the false positive is 8.7%. Paper [23] This paper proposes a sensing process that can be described as a finite state machine, which reduces the recording of irrelevant activities by judging the acceleration threshold. And records the time when a suspected fall occurs and passes the data to the classification system. The classification system extracts 8 features and passes the features to a two-layer feed-forward network. The result of this system is stated with 100% accuracy.

Non-contacted methods are well-suitable for fall detection tasks as the devices can be deployed in necessary places and can monitor continuously. Paper [25] presented a camera-based fall detection system. The system uses a Kalman filter to reduce the noise and track the presented human, and a KNN algorithm is used to determine whether a fall event happens or not. The system achieves a sensitivity of 96% and a Specificity of 97.6%. Paper [26] also demonstrated a vision-based fall detection system but using a Kinect infrared camera to obtain the depth information. The depth information is processed into a 3D bounding box of the subject to track and estimate the motion. Systems using RF sensing for fall detection are also widely discussed. Paper [27] demonstrated a fall detection algorithm based on the radar Doppler time-frequency analysis and used Sparse Bayesian learning for classification. The system proposed in paper [28] uses an fmcw radio to collect activity data and decomposes the reflected RF signal into vertical and horizontal heat maps to obtain spatial information. CNN is then used to extract features

from the heat maps and detect falls. This paper also evaluated the performance of this system in through-wall scenarios, with a 93% through-wall accuracy and 96% Line of Sight accuracy, which shows the advantage of RF sensing based systems. A deep learning framework for RF sensing applications is proposed in [29]. The authors evaluated the proposed framework for different human activity recognition tasks using CSI. A deep residual-learning-based system for indoor localization with median errors at about 0.86m, and a Long Short-Term Memory (LSTM) based system for five types of activity recognition with an overall 90.37% accuracy.

2.1.2 Speech Recognition

Speech recognition has been researched since the last century. Speech recognition can be described as a maximum likelihood estimation problem. The earliest method proposed to solve the problem of speech recognition was the maximum likelihood estimation method based on the Hidden Markov Model(HMM) [30], [31]. The speech recognition processes are modelled into the Language Model and the Acoustic Channel Model. The language model represents each word as a state and the transitions between states represent the relationship between words. The acoustic channel model consists of two subsources, the phonetic model and the acoustic model. Phonemes are the smallest unit of sound in a language. By modelling the phonemes, the parameter of the model can be reduced instead of modelling words directly. Due to realistic factors such as differences among speakers or background noise, the acoustic characteristics of the same word can be different. By learning the output of the acoustic features from the Acoustic processor, the acoustic model enhances the robustness and accuracy of the speech recognition system. This model demonstrates the basic idea of the earliest speech recognition model. The feature extraction method is crucial in speech recognition. The Mel-frequency cepstral coefficient (MFCC) is one of the most popular feature extraction methods. Since the human ear has different sensitivities at different frequencies, it is more sensitive to low-frequency sounds and less sensitive to high-frequency changes. The MFCC simulates the human ear's perception of different frequencies by applying the Mel filter bank to the spectrum. The research on the comparison of different parametric representations shows the great performance of MFCC in feature representation. [32].

Since the development of AI and its excellent performance in various tasks, AI-based speech recognition systems have also been developed. Paper [33] reviews the development of the deep neural network instead of the Gaussian mixture model(GMM). The GMM is used to model the characteristics of each state generated by a mixture of multiple Gaussian distributions and uses the Expectation Maximization (EM) algorithm to estimate the parameters. A DNN is also used to compute the HMM state observation probability similar to the GMM. The DNN-HMM model outperforms the GMM-HMM model on several speech recognition tasks, indicating that DNN is more capable for capturing complex speech than GMM. The DNN in this paper refers to a feed-

forward neural network, the authors further discussed the use of the CNN structure instead of the feed-forward neural networks. By adding convolutional layers and pooling layers, CNN can better capture the characteristics of speech signals, thereby improving the error rate by 6-10% compared to feed-forward neural networks. The development of deep neural networks has also led to the development of lip reading recognition tasks. A multi-modal lip reading system is proposed in the paper [34]. The model in this paper consists of an image encoder, an audio encoder and a character decoder. The image encoder uses a CNN for the image feature extraction and the LSTM network to process the CNN output. The audio encoder uses the LSTM network to process the audio signal directly. The LSTM decoder predicts the sequence of characters, using a dual attention mechanism to focus on relevant parts of both the audio and video inputs. The system is tested on a large scale Lip Reading sentences dataset based on BBC broadcasts and the model error rate outperforms a professional lip reader. Work [35] proposed an attention-based pooling mechanism to track lip movement. A spatio-temporal residual CNN is used to extract the feature map and then processed by a Visual Transformer Pooling (VTP) block. The prediction sequence is generated by an encoder-decoder transformer model, which captures spatial and temporal information from the video frames through the use of temporal positional encodings. The model achieves significant improvements over previous methods. This model is evaluated on the LRS2 and LSR3 datasets, the word error rate (WER) is 22.6% on LSR2 and 30.7% on the LRS3 dataset, showing a great performance of this lip reading system.

2.2 Artificial intelligence

2.2.1 Machine learning

As reviewed before, AI is now widely used in various healthcare applications. This section will review the principle of common AI algorithms including traditional machine learning algorithms and deep learning models. The concept of Artificial intelligence was first introduced in the 1950s, the idea of AI was focusing on making computers handle tasks as humans. After decades of development, AI has become one of the most popular research perspectives, and also one of the most useful tools in various areas. Machine learning has been the main research direction in AI for a long time, and has developed many excellent and practical algorithms.

- Supported vector machine (SVM) was first proposed in [36], SVM is a supervised learning algorithm used to solve classification problems. The principle of SVM is to calculate a Hyperplane that separates different classes, and at the same time has the largest margin for both classes. Though SVM can only work on linear classification tasks, using a kernel function to map the non-linear data to a high-dimension sample space makes the SVM able to work with non-linear tasks [37], [38]. SVM is one of the most popular ML algorithms as It is supported by mathematical theory, and does not rely on statistical methods, thus

simplifying common classification and regression problems. The limitation of SVM is that it takes a long time to train and is difficult to use for large-scale data sets.

- K-nearest neighbour(KNN) is also an algorithm for classification. The principle of the KNN algorithm is to determine the classification of a data point by the category of the k data points that are closest to it. [39] The advantage of KNN is that the implementation is simple and effective, while the disadvantage is that if the sample size of a certain class is much larger than the others, it may be more likely to cause classification errors. Moreover, KNN needs to calculate the distance between the data point to all samples, which requires a large amount of calculation.

- Random forest achieves higher accuracy by integrating more decision trees. The decision tree is a supervised learning method. It can summarize decision rules from a series of data with features and labels, and present these rules in a tree diagram structure to solve classification and regression problems. The decision tree algorithm is popular as it is easy to understand, applicable to various data, and has good performance in solving various problems. [40]

- K-means clustering is suitable for unsupervised classification algorithms. For the number k that needs to be classified, k samples are randomly selected as cluster centers, and then each sample is divided into the cluster closest to it. The center value of the cluster is updated and the process is repeated until the termination condition is reached. The disadvantage of k-means is that it is sensitive to the value of k, the initial center point and abnormal data points. [41] Density-based spatial clustering of applications with noise (DBSCAN) [42] is one of the most commonly used cluster analysis algorithms. Compared with the K-means algorithm, it does not require a given number of k in advance but requires two parameters, the range ε and the minimum number of points $minPTS$ required to form a high-density area. [42] It starts with an arbitrary unvisited point and then explores the ε-area of this point. A new cluster is established if there are enough points in the ε-area. If a point is located in a dense area of a cluster, the points in its ε-area also belong to the cluster. When these new points are added to the cluster, if it is also in a dense area, the points in its ε-area will also be added to the cluster. This process will be repeated until no more points can be added so that a complete cluster is found. The advantage of DBSCAN is that it can find clusters of any shape and can exclude the interference of noise points.

2.2.2 Deep learning

Deep learning is the most popular topic in the field of AI at present. The rapid development of deep learning in recent years has provided more effective methods for processing tasks such

as image recognition and natural language processing(NLP). Lenet as one of the first proposed CNN, attracted attention due to its high performance in text recognition tasks. It proposed the model consists of convolutional layers, pooling layers and the activation function [43]. The Convolutional layers with convolutional cores have local receptive fields to extract features from different input parts. The activation function allows the network to fit with non-linear models. Without an activation function, a neural network would be a linear model regardless of the number of layers, as each layer would just be a linear combination of the previous one. Non-linear activation functions allow the network to model more complex relationships and capture patterns that are not linearly separable.

- AlexNet [44] was the first to introduce the use of ReLU activation functions and dropout regularisation. ReLu relieves the vanishing gradient problem, therefore significantly improving model performance and generalisation. The architecture of AlexNet consists of five convolutional layers, three max-pooling layers and two fully connected layers, which laid the foundation for deeper and more efficient neural networks.

- The VGG Network [45] further increased the depth of the network by replacing the large convolution kernels by stacking multiple small convolution kernels, which increased the non-linearity and reduced the number of parameters while keeping the same receptive field. This architecture enabled the extraction of more complex features and contributed to an improvement in accuracy. The simplicity and effectiveness of the VGG architecture have made it a popular choice for many applications.

- ResNet [46] is one of the most widely used convolutional neural network (CNN) models, primarily because it effectively solves the vanishing gradient problem. This issue, common in deep networks, occurs when gradients become too small during training, causing the model's weights to stop updating, especially in very deep networks. The residual block makes information and gradients propagate more easily in deep networks by introducing skip connections, thus avoiding the rapid decay of gradients. The residual connection enables the training of a very deep network and improves the performance significantly. The architecture of residual connection laid the fundamentals for other deep neural networks.

- LSTM (Long Short-Term Memory) is a specialized architecture of Recurrent Neural Networks (RNN) introduced by Hochreiter and Schmidhuber in 1997 [47]. The core of LSTM is the memory unit. The LSTM processes the information through the three gate structures, which are the input gate, the forget gate and the output gate. This structure effectively preserves the long-term transmission of information and enables LSTM to perform well in processing and predicting sequence data.

- Transformer is proposed as an innovative attention-based model. [48] The Transformer model is composed of multiple stacked encoders and decoders. The Encoder is composed

of several layers with 2 sublayers, the multi-head self-attention mechanism and a feed-forward network. The multi-head self-attention takes each input token and generates three vectors: Query (Q), Key (K), and Value (V). The model calculates attention scores by taking the dot product of the Query and Key vectors. These scores are scaled and passed through a softmax function to get attention weights as the encoder output. The decoder architecture is similar to the encoder, with one more multi-head self-attention sublayer to process the output from the encoder. The input for the encoder is embedded with the positional encoding to represent the connection in context.

These DNN architectures have not only advanced the state of the art in image recognition but have also inspired new directions in neural network design, emphasizing the importance of depth, connectivity, and efficient feature extraction.

2.3 Summary

In summary, this section reviews the recent technologies in healthcare, including RF sensing in human activity recognition and speech recognition, along with the applied machine learning and deep learning algorithms. For vital sign monitoring, the traditional ECG and PPG methods are reviewed, both methods provide accurate vital sign monitoring with multiple attached sensors. In comparison, different RF devices based vital signs monitoring systems are also reviewed, including the fmcw radar based system, the WiFi-based system and the Doppler radar based system.

- **Heart Rate Monitoring:** RF-based methods demonstrate promising accuracy, with WiFi CSI-based approaches achieving up to 95% accuracy, while FMCW radar-based methods report a correlation exceeding 80% with wearable ECG sensors. In-vehicle RF sensing for heart rate monitoring shows a low error margin of approximately 0.6 bpm, making it suitable for real-world applications.
- **Respiratory Rate Monitoring:** RF-based respiratory rate monitoring exhibits high reliability, with Doppler radar-based methods achieving 98% accuracy and an error below 2 breaths per minute (bpm). FMCW-based sensing reports comparable results, while in-vehicle RF monitoring further refines precision, achieving an error of only 0.06 respirations per minute (rpm).
- **Human Activity Recognition:** Traditional accelerometer-based methods report an 83% sensitivity for fall detection but suffer from high false positives of 53% . Camera-based approaches, leveraging Kalman filtering and KNN classifiers, achieve 96% sensitivity and 97.6% specificity, showing strong performance in controlled environments. RF-based

methods using FMCW radar and deep learning achieve 96% accuracy in line-of-sight scenarios and maintain 93% accuracy even through obstacles, highlighting their robustness in real-world applications.

- **Speech and Lip-Reading Recognition:** Traditional HMM-based speech recognition remains widely used but struggles with noise robustness. DNN-HMM hybrid approaches reduce word error rates by 6-10% compared to GMM-HMM systems. RF-assisted lip-reading models demonstrate a 22.6% WER on LRS2 datasets and a 30.7% WER on LRS3 datasets, showing promise in multimodal speech recognition but still requiring improvement for real-world applications.

The results of these different RF-based systems all show a high accuracy in vital signs detection tasks. Moreover, the RF-based systems have the advantage of the ability to perform through the wall monitor and are easy to deploy in the necessary environment. Then fall detection which is the major task of human activity detection in healthcare is reviewed. Including the wearable accelerator and Gyroscopes based systems, the visual-based systems and the RF-based system. The wearable systems are capable of detecting the fall at any place, however, the detection threshold should be strictly examined to reduce the wrong detections. The visual-based system is accurate but it can lead to privacy concerns. For the speech recognition and lip reading tasks, the research on RF-based systems is very limited. The traditional speech recognition algorithm and the feature representation models are reviewed and compared with AI-based systems, the AI-based systems outperform the conventional GMM-HMM systems. Furthermore, the AI-based lip reading systems also show great performance on the visual dataset. These results suggest the ability of the AI models in such context-related tasks. Finally, the most common machine learning algorithms and deep learning models are reviewed to specify the principles of deep neural networks. Through this review, the limitations of traditional technologies and some traditional algorithms are summarised, which indicates the importance of developing the RF sensing healthcare system and the benefits of integrating the AI model.

Chapter 3

Multi-modal speech recognition system

3.1 Introduction

According to a report from the World Health Organization[49], the global population is expected to reach 10 billion by 2050, and the number of people suffering from hearing loss is predicted to reach 2.5 billion. This highlights the need for effective and robust hearing aid systems. Moreover, the COVID-19 pandemic shows a negative impact on people with hearing loss in understanding others' speech. Since masks obstruct the lip reading and affect the performance of audio-based hearing aids[50]. Currently, audio-based and video-based hearing aids are the dominant research aspects, but they face practical issues and are vulnerable to various environmental factors since audio-based systems may not perform optimally in very noisy environments. In recent years, the emergence of RF-based lip-reading systems [51] provides a promising alternative solution for the future of hearing aids. Unlike audio-based systems, RF signals are not affected by acoustic conditions, and their penetrability enables them to work through obstacles and walls. Consequently, hearing aid systems that incorporate RF signals can better resist the variations of environmental factors than traditional solutions. Furthermore, RF-based sensing is a privacy-preserving method as it only detects slight movements of the target's lips without capturing sensitive information. However, the robustness and accuracy of RF-based speech recognition often require the support of vast amounts of measurement data, which can be time-consuming and laborious. Developing effective signal processing and denoising algorithms is also necessary to mitigate interference from other movements during lip-reading, which could be quite challenging.

To address these limitations, the fusion of audio, visual and RF information could provide a more robust and accurate solution than relying on just a single modality. In this way, the strengths of each modality can complement each other, allowing hearing aids to address a variety of challenging sensing scenarios. As a result, we propose a novel multi-modal SR system for hearing aids in this research that can gather information across audio, visual, and RF modalities and

achieve considerable recognition improvements over single-modal SR systems. To simultaneously extract features from different modalities and perform effective analysis, a multi-input convolutional neural network (CNN) is implemented. The network consists of three separate inputting models for initial feature extractions, a concatenation layer for concatenating extracted features along channel dimension, and a ResNet18[46]-based classification model. Meanwhile, our work has comprehensively investigated different sensing modalities individually and also different fusion schemes, including audio-visual, audio-RF, visual-RF and audio-visual-RF fusions. Especially, the RF sensor used in this work is impulse radio ultra-wideband (UWB) radar, which has been extensively deployed in various applications, such as activity recognition[52], [53], localization[54], and vital sign detection[55]. It features low power consumption, high accuracy, high data acquisition rate and high noise resistance, which plays an important role in this work. In the end, we have conducted a series of experiments based on our multi-modal dataset[56] to validate the performance of the proposed system, where the experimental results have demonstrated the effectiveness of the proposed fusion system.

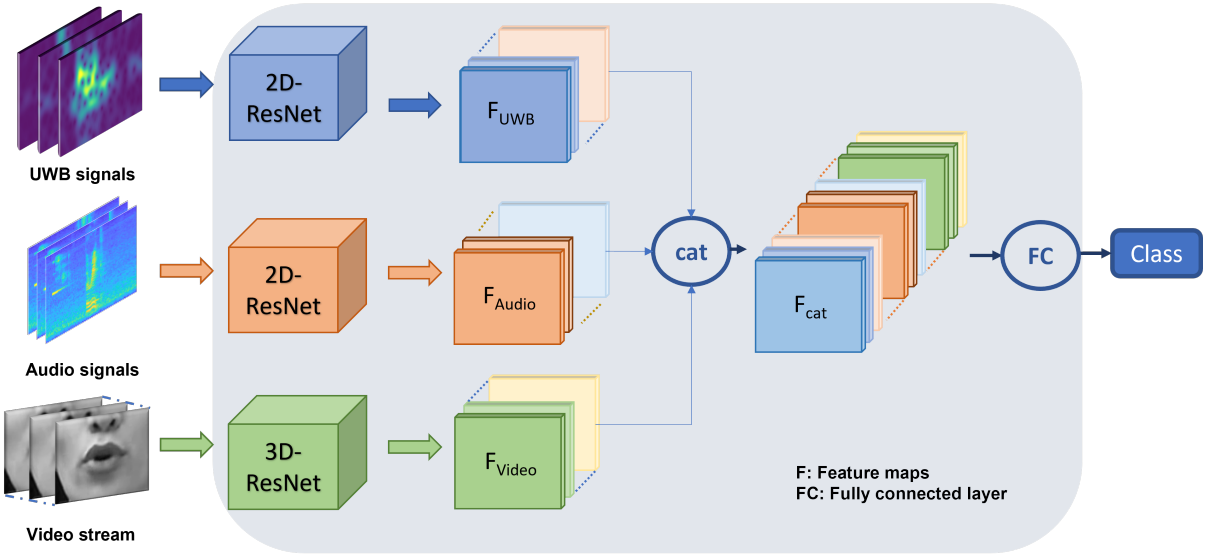


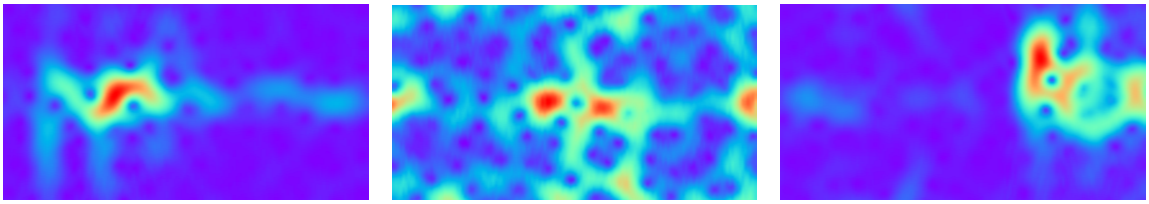
Figure 3.1: Architecture of the multi-modal network

3.2 Methodology

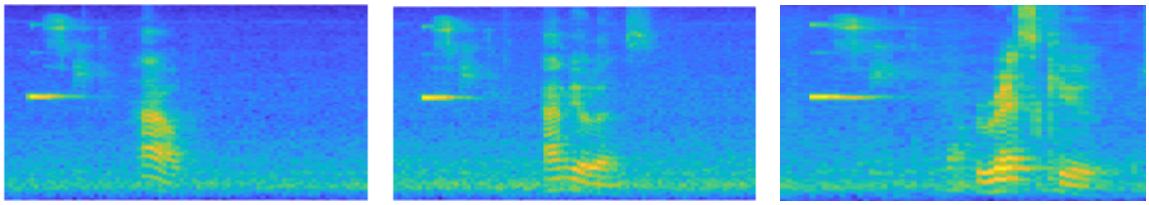
The procedure of the whole system is shown in Fig. 3.1 This section will introduce the principle of the sensing methods, pre-processing method and the feature extraction method involved.

3.2.1 Data pre-processing

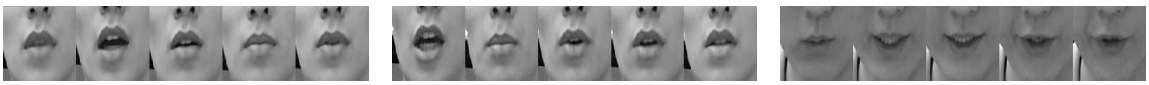
UWB radar is a type of impulse-based radar that uses very short and low-power pulses to transmit the radar signal. UWB is defined as where the bandwidth is larger than 500 MHz which



(a) UWB data of word 'help' (b) UWB data of word 'Ambulance' (c) UWB data of word 'Rescue'



(d) Audio data of word 'help' (e) Audio data of word 'Ambulance' (f) Audio data of word 'Rescue'



(g) Video frames of word 'help' (h) Video frames of word 'Ambulance' (i) Video frames of word 'Rescue'

Figure 3.2: Processed data of UWB, Audio and Video respectively

allows UWB radar to detect objects with high resolution and low power[57]. Therefore UWB radar can be used in multiple detecting applications such as through-wall object detection and tracking, and human motion analysis. The distance of the object is gathered by measuring the Time of Flight (ToF) of the reflected signal $d = \frac{c \times \tau}{2}$, where d is the distance, c is the speed of light and τ is the time delay of the received signal. The distance distinguish resolution is given by $R_\Delta = \frac{c}{2B}$, where R_Δ is the resolution, B is the bandwidth of the impulse signal. Larger bandwidth provides more precise distance detection.

For the UWB data, the in-phase and quadrature (IQ) data of each pulse is extracted from the radar signal. The response of reflected radar signal depends on impulse delay from detected object distance to radar, which can be represented as the Equation 3.1:

$$s(\tau,t) = \sum_{i=1} a_i(\tau,t) e^{-j2\pi \frac{d(t)+d_i(\tau)}{\lambda}} \tag{3.1}$$

where t represents frame time, τ and i represents ToF delay and multipath index, λ represents the wavelength of the UWB signal. After raw data collection, the moving target indication (MTI) technique is applied to filter out static targets, leaving only the moving targets. We eliminate the signal of a specific time interval at a particular weight to significantly suppress static fake peaks that are not related to human motion in this case. The Short-Time Fourier Transform (STFT) is then used to create a time-frequency representation of the signal in the form of a spectrogram. By applying this pre-process method, the UWB data is transformed into a spectrogram for fea-

ture extraction using deep learning models.

The audio data is converted to Mel-Spectrogram to train the network. The Mel-spectrogram offers a more accurate representation of how humans perceive sound compared to a standard spectrogram. This is because human perception of frequency is not linear, it is more sensitive to lower frequencies than to higher ones. By emphasizing this sensitivity, the Mel-spectrogram provides a closer approximation to how we interpret sound in the real world. The formula for calculating the Mel frequency is given in equation 3.2

$$Mel(f) = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \tag{3.2}$$

For the concern of the data size and the computation cost, the video data is first processed using The dlib face recognition to label the landmark of each participant and crop the video frames into lip area as shown in Fig.3.2. The video frames are then resized to the same size and converted into greyscale for training.

3.2.2 Multi-modal feature extraction

The architecture of the multi-input CNN is shown in the block in Fig.3.1 UWB and audio data are processed to the spectrogram, which is represented as a 2D matrix. On the other hand, video data has an additional dimension to indicate the sequence of each frame. Hence, the feature extraction network is divided into two parts. A 2D Resnet architecture is used to extract features from the UWB and audio spectrogram, while a 3D Resnet architecture is used for the video feature extraction as 3D CNN also incorporates temporal information by analyzing the sequence of frames over time. This makes it well-suited for video-based SR tasks. The feature maps obtained from each input source are then concatenated to represent the overall feature of the speech.

3.3 Evaluation and discussion

3.3.1 Experimental Setup

The data collection setup is shown in Fig.3.3 In this research, we utilized three different sensors to collect a multi-modal dataset. The Kinect v2 is used to collect the video and audio data, and the XeThru X4M03 to collect UWB data. The speech data consists of 15 different words that are commonly used in healthcare-related scenarios. Each word was repeated 10 times by the participants. The list of words is as follows: order, assist, help, ambulance, bleed, fall, shock, medical, sanitize, doctor, accident, rescue, emergency, heart and break. Details of the dataset are referred from [56]. The UWB radar setup is shown in Table.3.1.

Figure 3.3: Setup of data collection

The Kinect v2 has a 1080p resolution RGB camera and a 512 $\times$ 424 resolution depth camera. It captures 1920 $\times$ 1080 video at 30 fps. And a 4-microphone array to collect the audio data in 256 kbps and 16-bit depth.

3.3.2 Network Training

The hyperparameter for training is listed in table 3.2. The training of this multimodal model was run on an Nvidia RTX 3080 graphics card with 12GB memory. Considering the data volume of the multimodal dataset and memory limitations, the size of each batch was set to 6. The loss function used is the cross-entropy loss function. The cross-entropy loss is one of the most commonly used loss functions in neural network training. The expression of the cross-entropy loss function is listed below. Where $\mathbf{y}$ is the one-hot encoding real label and $\hat{\mathbf{y}}$ is the probability distribution of the predicted label, n denotes the total classes.

$$l(\mathbf{y},\hat{\mathbf{y}}) = -\sum_{j=1}^{n} y_j \log \hat{y}_j. \tag{3.3}$$

The Adam optimizer was employed instead of the common stochastic gradient descent(SGD) optimizer. Adam is also one of the most widely used optimizers. The Adam optimizer employs

Parameter	UWB Impulse radar
Center Frequency	8.745 GHz
Sampling Frequency	23.328 GHz
Frame Rate	300 Hz
Bandwidth	1.5 GHz
Antennas	1 Tx / 1 Rx

Table 3.1: Parameters of UWB impulse radar

Hyparameter	Value
loss function	Cross-entropy Loss
Optimizer	Adam
β_1	0.9
β_2	0.999
Batch Size	6
Learning rate	0.001

Table 3.2: Hyperparameters in multi-modal network training

two hyperparameters β_1 and β_2, which regulate the gradient descent process, thereby accelerating convergence and reducing sensitivity to the learning rate.[58] Considering the significant volume of data and the relatively modest batch size, each training epoch requires approximately 150 seconds. The SGD optimizer requires more epochs and is also harder to converge in our test. Therefore the Adam optimizer is chosen though it may cause some fluctuation of the loss in the training progress.

3.3.3 Result and Discussion

The confusion matrix in Fig.3.5 shows the performance of three modalities and three fusion models, respectively. The results show that multi-modal SR outperforms any uni-modal detection method. Specifically, the accuracy of video incorporating UWB data SR has achieved 87.55%, improving uni-modal SR only using UWB data or video by around 1.5% to 6.5%. As mentioned earlier, RF and visual information fusion can greatly enhance system performance and robustness in a variety of challenging environments, such as poor lighting conditions, wearing-mask targets, etc., where RF sensors can provide complementary information to fill in the missing (high-quality) visual information. On the other hand, the fusion of audio and UWB information can also improve the corresponding uni-modal SR accuracy by around 2% to 3%. This is of great significance for enhancing SR performance and traditional hearing aids systems in noisy real-world environments. Finally, we also tested the fusion of three modalities that in-

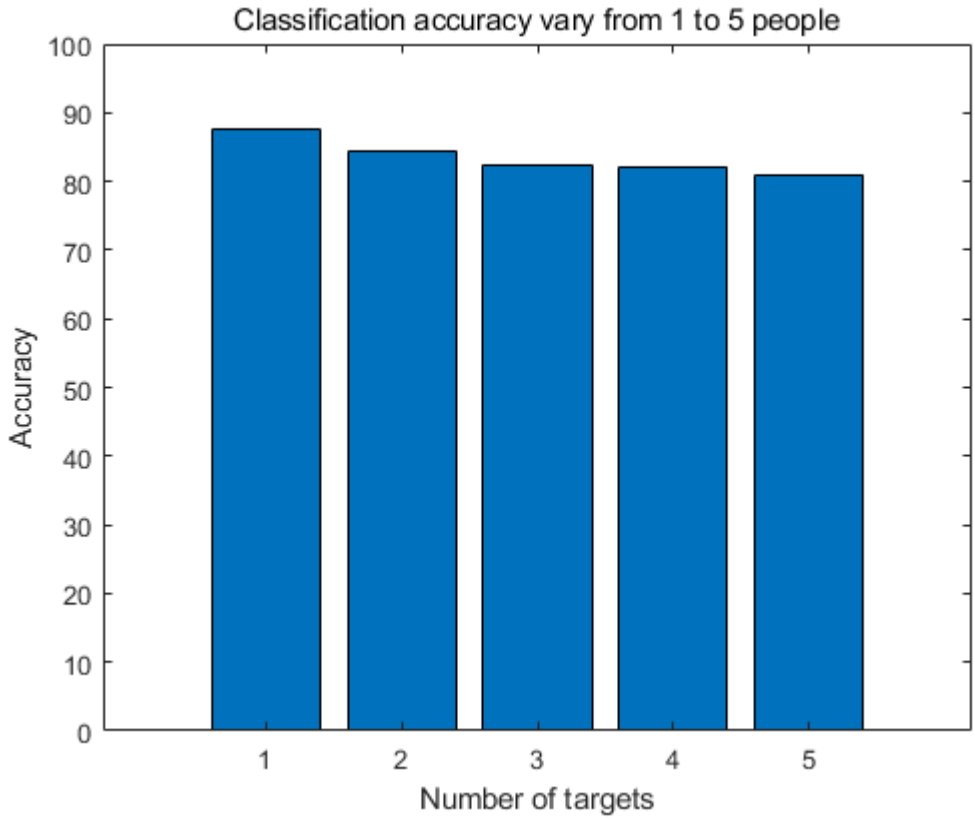


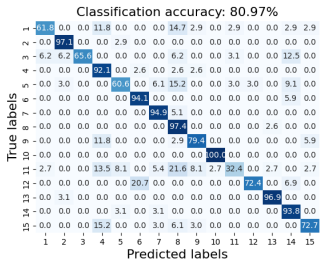
Figure 3.4: UWB result vary from 1-5 people

clude RF, visual and audio data and have demonstrated a high SR accuracy of 96.89%. It can be seen that the proposed system has shown sufficient capability in SR scenarios and has great potential to be implemented in future hearing aids technologies.

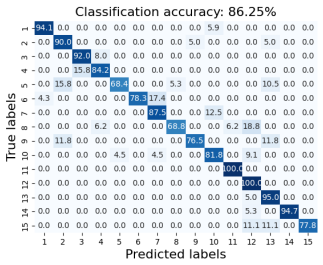
From the bar chart Fig.3.4, it can be seen that the recognition accuracy of the UWB radar decreases from 87.5% to 80.97% as the number of people increases. This may be due to the differences in individual such as different lip shapes, speaking speeds, and speaking habits, leading to the decade of performance in multi-user situations. However, the fusion model of UWB and other sensing sources shows improvement, indicating that the multi-modal SR supplements the information for each modality.

3.4 Conclusion

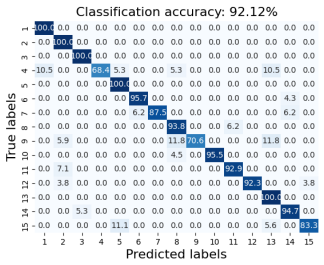
In summary, this research proposed a multi-modal SR system and compared the performance between the multi-modal system and each uni-modal system. Through experimental comparison and analysis, this research has comprehensively demonstrated that the fusion of multiple modalities can effectively improve SR accuracy compared to uni-modal systems. Further, the mutual complementary of different information can greatly improve the robustness of the system



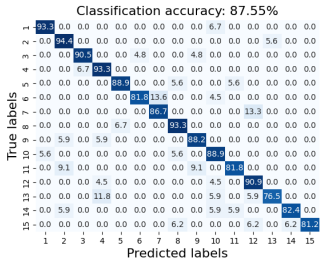
(a) Confusion matrix of word classification based on UWB radar.



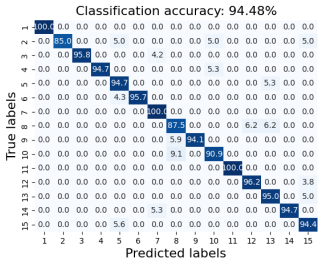
(b) Confusion matrix of word classification based on video stream.



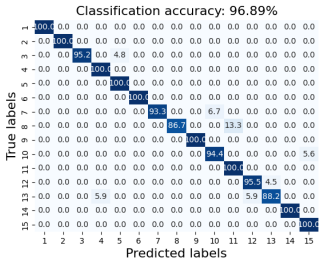
(c) Confusion matrix of word classification based on audio signal.



(d) Confusion matrix of word classification based on video and UWB data.



(e) Confusion matrix of word classification based on audio and UWB data



(f) Confusion matrix of word classification based on video, audio and UWB data

Figure 3.5: Results of human SR task among 15 words with UWB radar, audio signal, video stream, the fusion of two modalities of video and UWB, and fusion of video, audio and UWB, respectively.

in various complicated environments, which is crucial for real-world hearing aids. Especially, the incorporation of RF sensing demonstrated promising outcomes, which indicate the potential of the RF sensing fusion model in SR applications. Overall, the proposed system provides an effective and feasible framework for the development of future hearing aid systems. Meanwhile, there are several main challenges of this model:

- The accuracy of SR based on UWB radar needs improvement to achieve a precise SR.
- The generalization ability of UWB radar for different recognition objects is not adequate. We need to improve the UWB recognition model to enhance its generalization for different targets.
- The use of a Deep Neural Network requires computing power, and the model and the data pre-process procedure need to be improved to achieve real-time recognition.

This limitation indicates that the current dataset may not be sufficiently large or diverse to capture the full range of inter-subject variation. Another source of error comes from the resolution constraints of the IR-UWB radar. Although it is capable of capturing small lip movements theoretically, signal interference and noise from different characteristics of each participant unrelated to speech can introduce additional complexity. Hence, the accuracy of the UWB result is limited, especially in multi-person scenarios. Furthermore, the audio modality remains sensitive to background noise, and fusion strategies currently do not dynamically adapt to changing noise levels. These limitations highlight future improvement directions in dataset expansion, model enhancement, and adaptive fusion mechanisms to mitigate modality-specific weaknesses.

Chapter 4

Conclusions and Future Expectations

This thesis first reviews the commonly used healthcare technologies in various application scenarios. Including vital signs monitoring, motion detection and speech recognition, which are widely concerned areas. By reviewing the state of the arts, some limitations of traditional contact-based techniques are concluded, indicating the advantages of RF sensing based techniques in overcoming these shortcomings.

Then, a multimodal speech recognition model that integrates UWB radar information assistance is proposed. A large-scale multimodal data set is collected and used in verification. This is an innovative multi-modal speech recognition system incorporating the UWB data for lip reading. The system achieves high accuracy in classification among 15 English words commonly used in healthcare scenarios. Compared with the results of the single source model, the incorporation of RF data improves the recognition ability of the system, proving the feasibility of using RF data in the voice recognition system. Compared with traditional speech recognition models, this multimodal speech recognition model is more robust under harsh conditions, such as noisy environments, providing a possibility for future hearing aid design. Even though the resolution of UWB radar is capable of capturing the lip movement of speaking in theory, the classification accuracy of a single RF source is still lower than that of traditional audio and video models, and the classification accuracy decreases to a certain extent as the number of participants providing data increases.

The result shows that there are still certain shortcomings in using UWB as a single lip reading data source. The reason for this phenomenon may be that there are still certain differences in facial features when reading between different individuals, and the amount of data and model size of the existing data set are not enough to learn enough generalized information to eliminate the impact of these differences. This limitation also points out the direction for future work. First, a larger dataset with more participants will help the model learn more generalized and robust representations of facial features. This could help reduce the individual variation

error in the SR system. Secondly, the more advanced deep neural network architectures such as Transformer may contribute to a better multimodal modelling capability. Lastly, more advanced fusion strategies can be explored instead of simple concatenation that can better balance the weights of different modalities under different environments. With the improved system, potential applications could include integration into smart hearing aids that can operate effectively in noisy and dynamic environments, communication systems for patients with speech or hearing impairments. Furthermore, due to the privacy preserving properties of RF sensing, this system may also find applications in elder care and smart home healthcare monitoring, where discreet and continuous interaction support is required without compromising user privacy.

Bibliography

[1] F. Jiang, Y. Jiang, H. Zhi, *et al.*, "Artificial intelligence in healthcare: Past, present and future," *Stroke and vascular neurology*, vol. 2, no. 4, 2017.

[2] D. Dias and J. Paulo Silva Cunha, "Wearable health devices—vital sign monitoring, systems and technologies," *Sensors*, vol. 18, no. 8, p. 2414, 2018.

[3] R. Ross, S. N. Blair, R. Arena, *et al.*, "Importance of assessing cardiorespiratory fitness in clinical practice: A case for fitness as a clinical vital sign: A scientific statement from the american heart association," *Circulation*, vol. 134, no. 24, e653–e699, 2016. DOI: 10.1161/CIR.0000000000000461.

[4] W. B. Kannel, C. Kannel, R. S. Paffenbarger Jr, and L. A. Cupples, "Heart rate and cardiovascular mortality: The framingham study," *American heart journal*, vol. 113, no. 6, pp. 1489–1494, 1987.

[5] J. Achten and A. E. Jeukendrup, "Heart rate monitoring: Applications and limitations," *Sports medicine*, vol. 33, pp. 517–538, 2003.

[6] U. Rajendra Acharya, K. Paul Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, "Heart rate variability: A review," *Medical and biological engineering and computing*, vol. 44, pp. 1031–1051, 2006.

[7] M. T. La Rovere, G. D. Pinna, R. Maestri, *et al.*, "Short-term heart rate variability strongly predicts sudden cardiac death in chronic heart failure patients," *circulation*, vol. 107, no. 4, pp. 565–570, 2003.

[8] S. K. Berkaya, A. K. Uysal, E. S. Gunal, S. Ergin, S. Gunal, and M. B. Gulmezoglu, "A survey on ecg analysis," *Biomedical Signal Processing and Control*, vol. 43, pp. 216–235, 2018.

[9] G. Lu, F. Yang, J. A. Taylor, and J. F. Stein, "A comparison of photoplethysmography and ecg recording to analyse heart rate variability in healthy subjects," *Journal of medical engineering & technology*, vol. 33, no. 8, pp. 634–641, 2009.

[10] A. Temko, "Accurate heart rate monitoring during physical exercises using ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2016–2024, 2017. DOI: 10.1109/TBME.2017.2676243.

[11] T. Tamura, Y. Maeda, M. Sekine, and M. Yoshida, "Wearable photoplethysmographic sensors—past and present," *Electronics*, vol. 3, no. 2, pp. 282–302, 2014.

[12] H. K. Yaggi, J. Concato, W. N. Kernan, J. H. Lichtman, L. M. Brass, and V. Mohsenin, "Obstructive sleep apnea as a risk factor for stroke and death," *New England Journal of Medicine*, vol. 353, no. 19, pp. 2034–2041, 2005.

[13] T. Young, J. Skatrud, and P. E. Peppard, "Risk factors for obstructive sleep apnea in adults," *Jama*, vol. 291, no. 16, pp. 2013–2016, 2004.

[14] M. A. Cretikos, R. Bellomo, K. Hillman, J. Chen, S. Finfer, and A. Flabouris, "Respiratory rate: The neglected vital sign," *Medical Journal of Australia*, vol. 188, no. 11, pp. 657–659, 2008.

[15] F. Q. AL-Khalidi, R. Saatchi, D. Burke, H. Elphick, and S. Tan, "Respiration rate monitoring methods: A review," *Pediatric pulmonology*, vol. 46, no. 6, pp. 523–529, 2011.

[16] H. Alshaer, G. R. Fernie, E. Maki, and T. D. Bradley, "Validation of an automated algorithm for detecting apneas and hypopneas by acoustic analysis of breath sounds," *Sleep medicine*, vol. 14, no. 6, pp. 562–571, 2013.

[17] M. Alizadeh, G. Shaker, J. C. M. De Almeida, P. P. Morita, and S. Safavi-Naeini, "Remote monitoring of human vital signs using mm-wave fmcw radar," *IEEE Access*, vol. 7, pp. 54 958–54 968, 2019.

[18] X. Wang, C. Yang, and S. Mao, "Phasebeat: Exploiting csi phase data for vital sign monitoring with commodity wifi devices," in *2017 IEEE 37th international conference on distributed computing systems (ICDCS)*, IEEE, 2017, pp. 1230–1239.

[19] T. Zheng, Z. Chen, C. Cai, J. Luo, and X. Zhang, "V2ifi: In-vehicle vital sign monitoring via compact rf sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–27, 2020.

[20] A. D. Droitcour, T. B. Seto, B.-K. Park, *et al.*, "Non-contact respiratory rate measurement validation for hospitalized patients," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2009, pp. 4812–4815.

[21] F. Bagala, C. Becker, A. Cappello, *et al.*, "Evaluation of accelerometer-based fall detection algorithms on real-world falls," *PloS one*, vol. 7, no. 5, e37062, 2012.

[22] Q. Li, J. A. Stankovic, M. A. Hanson, A. T. Barth, J. Lach, and G. Zhou, "Accurate, fast fall detection using gyroscopes and accelerometer-derived posture information," in *2009 sixth international workshop on wearable and implantable body sensor networks*, IEEE, 2009, pp. 138–143.

[23] S. Abbate, M. Avvenuti, F. Bonatesta, G. Cola, P. Corsini, and A. Vecchio, "A smartphone-based fall detection system," *Pervasive and Mobile Computing*, vol. 8, no. 6, pp. 883–899, 2012.

[24] J. Dai, X. Bai, Z. Yang, Z. Shen, and D. Xuan, "Perfalld: A pervasive fall detection system using mobile phones," in *2010 8th IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, IEEE, 2010, pp. 292–297.

[25] K. De Miguel, A. Brunete, M. Hernando, and E. Gambao, "Home camera-based fall detection system for the elderly," *Sensors*, vol. 17, no. 12, p. 2864, 2017.

[26] G. Mastorakis and D. Makris, "Fall detection system using kinect's infrared sensor," *Journal of Real-Time Image Processing*, vol. 9, pp. 635–646, 2014.

[27] Q. Wu, Y. D. Zhang, W. Tao, and M. G. Amin, "Radar-based fall detection based on doppler time–frequency signatures for assisted living," *IET Radar, Sonar & Navigation*, vol. 9, no. 2, pp. 164–172, 2015.

[28] Y. Tian, G.-H. Lee, H. He, C.-Y. Hsu, and D. Katabi, "Rf-based fall monitoring using convolutional neural networks," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–24, 2018.

[29] X. Wang, X. Wang, and S. Mao, "Rf sensing in the internet of things: A general deep learning framework," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 62–67, 2018.

[30] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 179–190, 1983.

[31] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[32] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[33] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[34] J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6447–6456.

[35] K. Prajwal, T. Afouras, and A. Zisserman, "Sub-word level lip reading with visual attention," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2022, pp. 5162–5172.

[36] C. Cortes, "Support-vector networks," *Machine Learning*, 1995.

[37] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[38] C. Leslie, E. Eskin, and W. S. Noble, "The spectrum kernel: A string kernel for svm protein classification," in *Biocomputing 2002*, World Scientific, 2001, pp. 564–575.

[39] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[40] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[41] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967.

[42] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, 1996, pp. 226–231.

[43] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[47] S. Hochreiter, "Long short-term memory," *Neural Computation MIT-Press*, 1997.

[48] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[49] W. H. Organization, *World report on hearing*. World Health Organization, 2021, xiv, 252 p.

[50] N. C. Homans and J. L. Vroegop, "The impact of face masks on the communication of adults with hearing loss during covid-19 in a clinical setting," *International Journal of Audiology*, vol. 61, no. 5, pp. 365–370, 2022, PMID: 34319825. DOI: 10.1080/14992027.2021.1952490.

[51] H. Hameed, M. Usman, A. Tahir, *et al.*, "Pushing the limits of remote rf sensing by reading lips under the face mask," *Nature Communications*, vol. 13, no. 1, p. 5168, 2022.

[52] K. Bouchard, J. Maitre, C. Bertuglia, and S. Gaboury, "Activity recognition in smart homes using uwb radars," *Procedia Computer Science*, vol. 170, pp. 10–17, 2020.

[53] J. Maitre, K. Bouchard, C. Bertuglia, and S. Gaboury, "Recognizing activities of daily living from uwb radars and deep learning," *Expert Systems with Applications*, vol. 164, p. 113 994, 2021.

[54] L. Cheng, A. Zhao, K. Wang, H. Li, Y. Wang, and R. Chang, "Activity recognition and localization based on uwb indoor positioning system and machine learning," in *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, IEEE, 2020, pp. 0528–0533.

[55] Z. Duan and J. Liang, "Non-contact detection of vital signs using a uwb radar sensor," *IEEE Access*, vol. 7, pp. 36 888–36 895, 2018.

[56] Y. Ge, C. Tang, H. Li, *et al.*, "A large-scale multimodal dataset of human speech recognition," 2023. DOI: 10.48550/ARXIV.2303.08295. [Online]. Available: https://arxiv.org/abs/2303.08295.

[57] V. Niemelä, J. Haapola, M. Hämäläinen, and J. Iinatti, "An ultra wideband survey: Global regulations and impulse radio research based on standards," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 874–890, 2017. DOI: 10.1109/COMST.2016.2634593.

[58] P. K. Diederik, "Adam: A method for stochastic optimization," *(No Title)*, 2014.