FULLY-ADAPTIVE RF SENSING FOR NON-INTRUSIVE ASL RECOGNITION VIA INTERACTIVE SMART ENVIRONMENTS

by

EMRE KURTOĞLU

SEVGI ZUBEYDE GURBUZ, COMMITTEE CHAIR
EDWARD SAZONOV
AIJUN SONG
CHRIS S. CRAWFORD
SHUNQIAO SUN
KENNETH DEHAAN

A DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Electrical and Computer Engineering
in the Graduate School of
The University of Alabama

TUSCALOOSA, ALABAMA

2024

ABSTRACT

The past decade has seen great advancements in speech recognition for control of interactive devices, personal assistants, and computer interfaces. However, Deaf people and people with hard-of-hearing, whose primary mode of communication is sign language, cannot use voice-controlled interfaces. Although there has been significant work in video-based sign language recognition, video is not effective in the dark and has raised privacy concerns in the Deaf community when used in the context of human ambient intelligence. Radars have recently been started to be used as a new modality that can be effective under the circumstances where video is not.

This dissertation conducts a thorough exploration of the challenges in RF-enabled sign language recognition systems. Specifically, it proposes an end-to-end framework to acquire, temporally isolate, and recognize individual signs. A trigger sign detection with an adaptive thresholding method is also proposed. An angular subspace projection method is presented to separate multiple targets at raw data level. An interactive sign language-controlled chess game is designed to enhance the user experience and automate the data collection and annotation process for labor-intensive data collection procedure. Finally, a framework is presented to dynamically adjust radar waveform parameters based on human presence and their activity.

DEDICATION

To my missed mother, father, brothers, and my beloved wife, Ayşenur.

LIST OF ABBREVIATIONS AND SYMBOLS

ADC	Analog-to-Digital Converter
AI	Artificial Intelligence
ASL	American Sign Language
ASPS	Angular Subspace Projection-Based Separation
BPM	Binary Phase Modulation
CA-CFAR	Cell Averaging Constant False Alarm Rate
CAE	Convolutional Autoencoder
CNN	Convolutional Neural Network
CODA	Child of Deaf Adult
CPHS	Cyber-Physical & Human System
CPI	Coherent Processing Interval
CPS	Cyber-Physical System
CSA	Cumulative Score Aggregation
CTC	Connectionist Temporal Classification
CW	Continuous Wave
μD	Micro-Doppler
DAM	Doppler-Angle Map
DBD	Dynamic Boundary Detection
DFD	Discrete Fréchet Distance
DFT	Discrete Fourier Transform
DL	Deep Learning
DNN	Deep Neural Network
DoA	Direction of Arrival

DTW	Dynamic Time Warping
FFT	Fast Fourier Transform
FMCW	Frequency Modulated Continuous Wave
FoV	Field of View
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
HAR	Human Activity Recognition
HCI	Human Computer Interaction
HGR	Hand Gesture Recognition
HITL	Human in the Loop
HoH	Hard of Hearing
I/Q	In-Phase and Quadrature
IF	Intermediate Frequency
IoT	Internet-of-Things
IR	Infrared
JD-MIMTL	Joint-Domain Multi-Input Multi-Task Learning
k-NN	k-Nearest Neighbors
LOS	Line of Sight
LSTM	Long Short Term Memory
MIMO	Multiple Input Multiple Output
ML	Machine Learning
MRMR	Minimum Redundancy Maximum Relevancy
MSE	Mean-Squared Error
MTI	Moving Target Indication
NLP	Natural Language Processing
PAC	Perception/Action Cycle
PBC	Power Burst Curve

PhGAN	Physics-Aware Generative Adversarial Network
PRF	Pulse Repetition Frequency
PRI	Pulse Repetition Interval
RADAR	RAdio Detection and Ranging
RAM	Range-Angle Map
RCS	Radar Cross Section
RDM	Range-Doppler Map
ReLU	Rectified Linear Unit
RF	Radio-Frequency
RNN	Recurrent Neural Network
RP	Range Profile
SAE	Stacked Autoencoder
SAR	Synthetic Aperture Radar
SL	Sign Language
SLR	Sign Language Recognition
SNR	Signal-to-Noise Ratio
SOTA	State-of-the-Art
STA/LTA	short-time averaging over long-time averaging
STFT	Short-Time Fourier Transform
SVMs	Support Vector Machines
TDM	Time Division Multiplexing
TF	Time-Frequency
TL	Transfer Learning
t-SNE	t-distributed Stochastic Neighbor Embedding
ULA	Uniform Linear Array
VAE	Variational Auto-encoder
VCO	Voltage Controlled Oscillator

ACKNOWLEDGMENTS

I would like to thank National Science Foundation (NSF) [Awards #1932547 and #2238653] and the American Association of University Women (AAUW) via a Research Publication Grant for Engineering, Medicine and Science for supporting this research.

Foremost, I would like to thank and express my sincere gratitude to my advisor, Dr. Sevgi Zubeyde Gurbuz, for her uninterrupted support throughout my Ph.D, for her thoughtfulness, motivation, patience, and immense knowledge. Her both technical knowledge and guidance had a game changing effect in my research and technical skills. I cannot appreciate her enough for understanding the student psychology so well and giving them the mental support when the times are rough. I will always be grateful for her valuable feedback, patience, understanding and being an exemplary leader.

I also would like to extend my sincere thanks to my dissertation committee members, Dr. Kenneth DeHaan, Dr. Chris S. Crawford, Dr. Shunqiao Sun, Dr. Edward Sazonov and Dr. Aijun Song for their constructive feedback, guidance and suggestions throughout the dissertation process. I would like to express my appreciation to them for taking the time to understand my research and providing their invaluable expertise in the field. Especially, I would like to extend my deepest thanks Dr. Kenneth DeHaan for helping and coordinating the studies with the Gallaudet University students, and Dr. Chris S. Crawford for providing his insights on human-computer interaction designs, experiments and user experience studies. Their technical expertise and insights were fundamental to design and implement my research studies.

I would like to express my sincere gratitude to Dr. Ali Cafer Gurbuz for his collaboration and investing the time to take part in our research studies and provide his fundamental theoretical and technical knowledge to develop novel methods and approaches. I also would like to thank Dr. Ali Cafer Gurbuz's Ph.D. student, Sabyasachi Biswas, for his collaboration in the angular projection project.

I would like to thank Dr. Darrin J. Griffin and Dr. Evie Malaia for their guidance and insights in the ASL project and helping us to get connected with the Deaf/HoH community, build multi-disciplinary relationships, understand the Deaf culture and the linguistic properties of ASL.

I would also like to thank Dr. Moeness G. Amin for his collaboration and constructive feedback in the human activity recognition with ethogram project. His ideas and insights have been instrumental for navigating the project.

I would like to thank my old roommate, Ozgur Satici, for being a brother to me during my Ph.D. journey. He has been a great company with enormous support and lots of memories.

I would also like to thank my lab mate Dr. Mahbubur (Mahbub) Rahman for his mentorship, friendship and guidance in technical works. I would also like to thank Ladi for providing his technical expertise in Linux and embedded systems related topics. I also would like to thank Sean, Josh and Sultan for helping with the experimental setups, data collections and enabling a supportive lab environment.

Finally, I would like to express my deepest gratitude to my parents for their constant love and uninterrupted support and being patient to my absence over the years.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

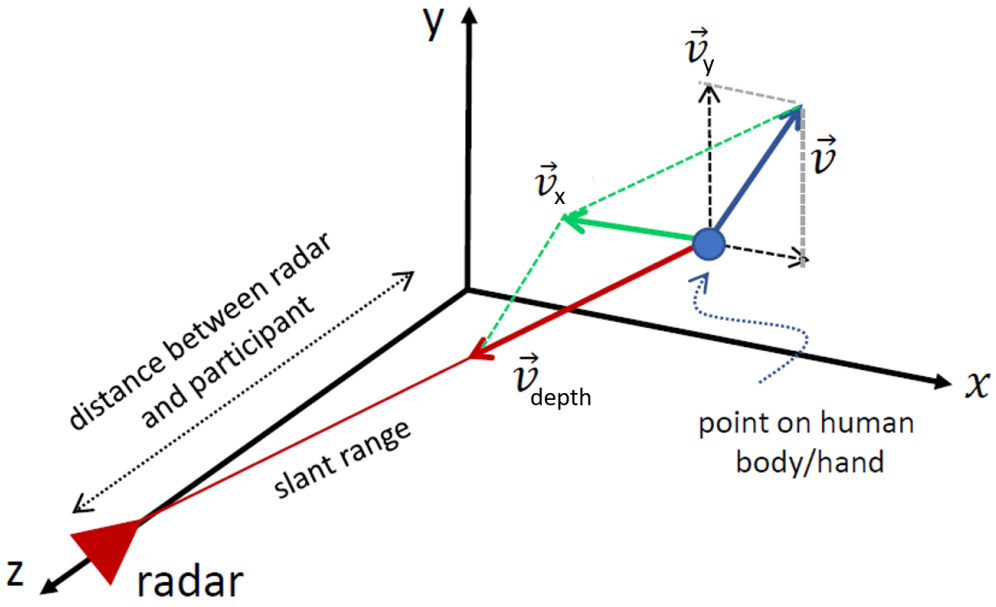
INTRODUCTION

1.1 Motivation & Context

Users of American Sign Language (ASL) make up over 1 million people in the U.S. and Canada, based on statistics provided by Gallaudet University (the world's only university designed to be barrier-free for deaf and hard of hearing students located in Washington, D.C.). People in the Deaf community, who rely on ASL as their primary mode of communication, rely heavily on technology as an assistive device as they navigate communication/language barriers that status quo society often creates. Unfortunately, many technologies are designed for hearing individuals, where vocalized speech is the preferred mode of communication, and has driven a burgeoning market of voice recognition software and voice-controlled devices. This precludes the Deaf community from benefiting from advances in technology, which, if designed to be compatible with ASL, could in fact generate tangible improvements in their quality of life.

Research related to technologies for the deaf or Hard of Hearing (HoH) has been ongoing for the past three decades, but, has primarily focused on camera-based and wearable technologies, such as gloves or wrist bands containing accelerometers and other sensors to translate sign language into voice or text. Among these approaches, sensor-augmented gloves have been reported to typically yield higher gesture recognition rates than camera-based systems. However, such wearable gloves cannot capture intricacies of sign languages offered through

Figure 1.1: Geometry of video and radar velocity measurements.



head and body movements [46, 119]. This issue is addressed by optical sensors, however, video cameras trigger user concerns over privacy and require light to be effective.

In contrast, especially in the context of technologies for the Deaf/HoH users, RF sensors have several advantages over alternative sensing modalities, which make them uniquely desirable. RF sensors are non-contact and protective of privacy, fully operational in the dark, and can even be used for through-the-wall sensing. RF sensors do not acquire personal imagery. Therefore, they are considered to be more privacy protective sensors when compared to video cameras. However, if the system security is compromised, there are still certain information the intruder can infer from the environment such as presence of human [39], number of people in a room [153], or person-specific activity or presence data (when acquired enough model training data) [180]. Most importantly, RF sensors can acquire a *new source of information* that is inaccessible to optical sensors: visual representation of kinematic patterns of motion via the micro-Doppler signature [27], as well as more accurate velocity measurements and range profiles. Figure 1.1 depicts the components of a 3D velocity vector, v. While video sensors are more sensitive to motions in x-y axes, radars have

higher sensitivity towards radial motions. This feature makes them promising complimentary sensors for multi-modal recognition scenarios.

1.1.1 Involvement of Deaf Community

Previous investigations of these existing prototypes often fail to involve participants and investigators fluent in ASL [16, 41]. A major aspect of my research methodology has involved feedback from Deaf users, and advice and collaboration with Deaf researchers at Gallaudet University, in particular Dr. Kenneth DeHaan and Dr. Caroline Kobek Pezzarossi.

In a preliminary focus group we conducted in 2019, Deaf participants reacted negatively to the idea of having to use anything wearable in their daily lives. Indeed, wearable technology limits signer's freedom in conducting daily activities and is not designed with ASL movements and language constraints in mind. In contrast, the Deaf participants reported that while they used on a regular basis some form of video-based technology for communication in their jobs; those video technologies have limitations (such as a narrow field-of-view, privacy issues, and reliance on light). Although video tends to be viewed favorably for interpersonal communication with society-at-large, Deaf participants in our focus group lamented its limitations – video usage was dependent upon being in an office/work environment or with access to cell phones (and battery life). A significant thematic point of discussion that occurred with the Deaf participants was concern over technology enabling invasion of privacy and potential surveillance of their personal and private lives.

There are also concerns about cultural exploitation and monetizing/commercializing products with no involvement or ownership by the Deaf community. As researchers we also have a responsibility with cultural sensitivity and awareness. Therefore, we believe it is important to have a strong collaboration with Deaf community while designing and commercializing Deaf-centric products, and there should be more initiatives towards including Deaf researchers into the product development team so that we can ensure a fair economic outcome.

1.2 Linguistic Significance of Kinematics

In speech communication, quantitative measurements of the temporal dynamics have resulted in fundamental insights into perceptual and mathematical properties of information exchange [160] . Temporal quantification of the properties of signed languages has been, to date, substantially behind that of speech due to the higher dimensionality of visual modality. When sign language linguistics research began in 1960s, signs had been defined based on their static properties: hand shape, place of articulation (i.e. location of the articulator/hand at the beginning and end of the sign), and hand shape orientation [161].

A study on signers' perception of writing in point-light displays [98] has demonstrated that signers viewing the dynamics of hieroglyph writing can tell the difference between 'strokes' (information-bearing portions of point-light movement) and 'transitions' (movement of the point-light from the end of one meaningful portion, to the beginning of another). A 2x2 Latin Square design that assessed the difference in perception between signers and non-signers, and users of Chinese and English, showed that sensitivity to transitions was due entirely to experience with sign language, and not due to experience with hieroglyphic writing systems.

Current neurolinguistic research indicates that dynamic properties of signs (i.e., speed and temporal contour of motion) contribute crucial linguistic information to the meaning of signs [120, 117]. Analysis of information content in speech vs. everyday motion using the visual properties of the signal and optical flow [118, 12] has indicated that signers transmit more information (in the sense of mathematical entropy) than humans carrying out dynamic tasks, and that the intelligibility of a signing stream is crucially dependent on the ability to parse entropy changes in visual information [121, 14]. RF sensors allow for improved measurements of these temporal dynamics in conjunction with shape dynamics, combining information picked up from the moving hands with the information on other articulators (head and body).

1.3 Sign Language-Sensitive Cyber-Physical & Human Systems

There has been extensive research towards technologies for the Deaf and HoH people over the past three decades, and most of these works have focused on the translation of sign language into text or voice. While Sign Language (SL) translation contributes towards facilitating interaction between Deaf/HoH and hearing people, potential of more broad range of applications to make their lives easier is overlooked. In contrast with many works focused on sign language translation, this dissertation is concerned with recognition only to facilitate interaction, which can be accomplished without needing to recognize all the words comprising ASL.

Smart Deaf spaces [9] are environments that can respond to the natural language of the Deaf community for the purposes of remote health, environment control, Human Computer Interaction (HCI), and security. There are several key design considerations when designing Deaf-centric smart spaces:

1. **Culture:** The Deaf culture and what feels more comfortable and natural for them should be prioritized for the device control commands. Having an open space and less visual noise in the room/office are important to facilitate a visually comfortable environment.

2. **Representation:** Deaf people have very diverse backgrounds and the data used by the prediction model (if there are any) should be comprehensive enough to represent data of all the signers. Regional dialects and different versions of the same sign are also other important factors to take into account while designing a sign language recognition system.

3. **Data Authenticity:** The data utilized in the system should be acquired from Deaf/HoH people and not from hearing individuals since SL is a language with an involved grammar and has contextual nuances, and should be articulated by the fluent signers to be able to capture those features. In our earlier study [69], we have found that there

are significant differences in the data distribution of fluent and imitation signers which affect the model performance. In another work [104], we have shown the importance of acquiring data in a natural setting instead of having strict experimental limitations and assumptions which do not represent the real world scenarios.

4. **Ease of Operation:** Usage of an accessible system should be intuitive and easy enough so that it does not require any technical expertise or external assistance. In an RF sensor based system, this would correspond to having a standalone system operating continuously without needing user to connect it to an edge device or computer and set it up every time when the system is being used.

Omitting any of these items can result in unfavorable user experiences. Therefore, a special attention needs to be paid in the design process of Deaf-centric spaces.

A wide range of sensor modalities from cameras to wearable-based devices have been proposed for SL recognition. Although sensor augmented gloves typically yield the most accurate 3D positioning of the hands and fingers, they cannot capture facial expressions and other body movements which play a significant role in SL grammar and can completely change the meaning of a phrase. In addition, they are often cumbersome devices to be used in a daily living scenario, and found intrusive by the Deaf community [47]. Since they need to be worn every time when the user wants to use the system, it interferes with other daily activities and causes uncomfortable experiences.

Video-based solutions, on the other hand, are currently the most effective way to capture facial expressions. High 2D spatial resolution provides sufficient precision for scene understanding and enables further methods like skeleton and body landmark estimation from 2D video frames by utilizing Deep Learning (DL)-based methods. Inclusion of Infrared (IR) sensors in cameras enables depth estimation as well with the cost of increased price and size. While video-based solutions have a lot to offer, they have certain limitations which preclude them to be ubiquitously used in home environments. First, they collect visual imagery which raises serious privacy concerns about data security. Second, they can easily get affected by

the lighting conditions of the environment, and skin and dress color of the people. Finally, they require direct Line of Sight (LOS) within a close distance to be able to understand the human motions accurately. These drawbacks limit the deployment of video-based solutions for Sign Language Recognition (SLR) in indoor environments. Therefore, a more secure, non-intrusive and non-invasive modality is needed.

Devices for interaction and communication with humans in a home environment should utilize sensors with several key characteristics:

1. **Privacy:** Not acquiring any personal data which can compromise identity or personal information of the individuals.

2. **Data Security:** The data should be acquired, processed, transferred and stored securely without allowing intruders to interfere.

3. **Accuracy and Precision:** Perceiving the environment and people with enough resolution and no or minimal error.

4. **Reliability and Robustness:** Environmental conditions and other external factors should not affect the sensor in a degree that it malfunctions or provides misinformation.

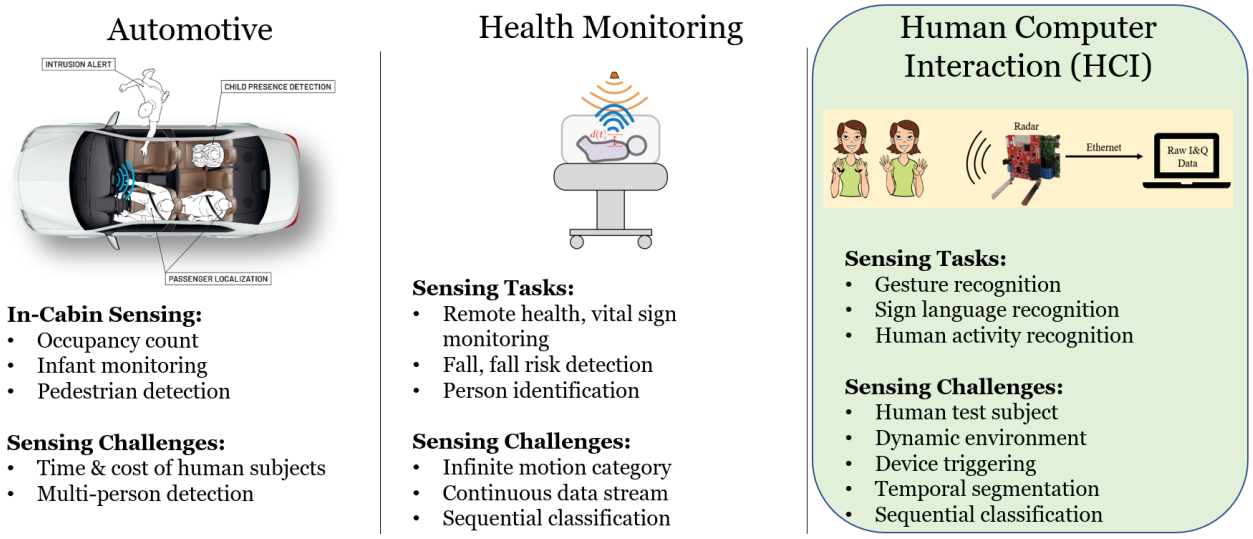
While each sensor has its own pros and cons, RF sensors check all these requirements and present distinct advantages over other sensor types. They can provide range, velocity and angle information with high resolution in various data representation formats. Furthermore, radars provide these information through physical measurements instead of relying on estimation methods as used in video-based solutions. Time-varying radial range information can be attained through round-trip time delay of the transmitted signal. Velocity information can be obtained from the Doppler shifts of the transceived signal. Finally, azimuth and elevation angles can be extracted from the time delay between the receiving antenna elements. The raw RF data is flexible enough to generate range, velocity, angle profiles, Range-Doppler Map (RDM), Range-Angle Map (RAM), Doppler-Angle Map (DAM), point clouds and any

other RF data representations. This flexible nature of RF data becomes especially handy when computational resources or the time allocated for data processing are limited. The generated data representations then can be used to train various DL models which lead to end CPHS applications.

Radars can also operate in adverse weather conditions from long distances which make them suitable sensors for autonomous driving and other surveillance applications. They can even be used for through-the-wall sensing in certain center-frequency bands. The development of low-cost, low-power, high-resolution and small size antenna modules enabled new research topics by allowing the use of RF sensors almost anywhere as a part of Internet-of-Things (IoT) applications, smart home systems, autonomous driving, wearables and even cell phones (e.g., Google Pixel 4). As a result, businesses including Google [112], NXP [143], Aptiv [147], Motional and Ghost have started to build and commercialize radar-based devices.

1.3.1 Radar as an Emerging Modality

Figure 1.2: Radar-based CPHS applications.



Development of commercially available, small package, high frequency radars has enabled numerous applications in automotive, health monitoring, security and IoT fields. Automotive applications include traffic sign detection, object detection and tracking, pedestrian detection,

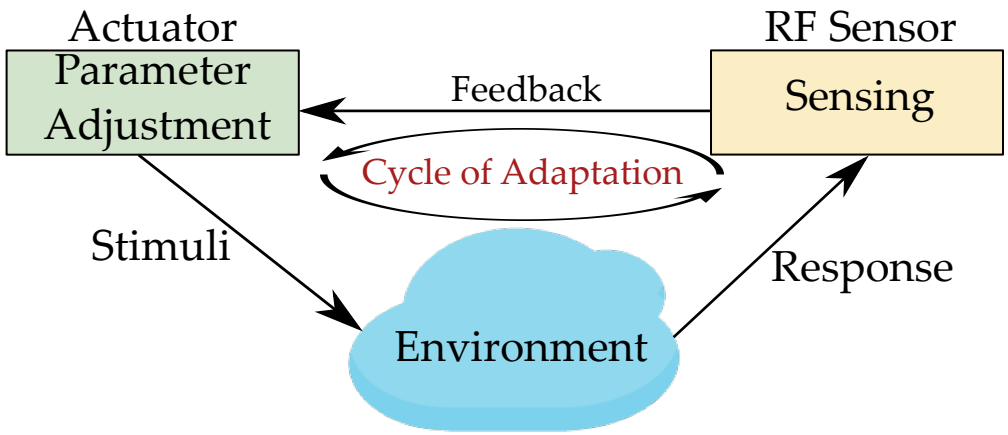
blind spot detection, automated parking systems, and in-cabin sensing applications such as occupancy count, and infant monitoring. Health monitoring applications such as vital sign monitoring, fall detection and gait abnormality detection have also been investigated. Other indoor applications include but not limited to gesture recognition, Human Activity Recognition (HAR), SLR and presence detection.

There are several reasons why radars have such a wide range of applications:

- **Data Reliability:** Radars can provide consistent data in unfavorable conditions such as adverse weather, very high or very low temperatures.
- **Low Cost:** Radars are relatively cheaper sensors when compared to LiDARs which makes them more affordable to integrate into various systems.
- **High Resolution Data:** Radars can provide high range, velocity and angle resolution depending on the RF waveform parameters and the antenna layout.
- **Privacy Protection:** Radars do not acquire personal imagery which alleviates the privacy concerns for indoor applications.
- **Small Package:** Commercially available RF sensors have very small package sizes that makes them easy to deploy in various environments.
- **Color, Texture and Lighting Agnostic:** Radars do not get affected from the color or texture of the object, and having very high or very low lighting conditions in the environment.

These features make radars well-suited sensors for various applications, and draw more attention from the researchers each day. Figure 1.2 depicts the radar-based CPHS applications for various objectives along with their sensing challenges.

Figure 1.3: The Cycle of Adaptation.



1.3.2 Adaptive RF Sensing

RF waveforms used in radar systems are often selected and adjusted based on the application needs. These include maximum unambiguous range and velocity, range, velocity and angle resolution, sampling rate, computational overhead and storage space. While having a fixed set of waveform parameters work well for a particular task, when there exists multiple tasks or a dynamic environment, utilization of different waveform becomes a necessity. For instance the RF sensor might need to adjust its parameters based on existence of a person in the room, or the activities they are performing.

With the development of software-defined RF sensors, it has become possible adjust transmitted signal's waveform parameters such as center frequency, bandwidth, Analog-to-Digital Converter (ADC) sampling rate, Pulse Repetition Frequency (PRF), and number of receiving antenna elements to be used depending on the needs of the application. However, these selections are made manually prior to deployment of the system. In addition to sensing, if the radar can understand the surroundings and takes action to adjust its waveform parameters in an autonomous fashion, it would count towards cognition, and the radar can operate in more optimal modes for different scenarios.

Human brain considered to be the most powerful cognitive dynamic system. Haykin [74] describes the functions of five building blocks of a cognitive dynamic system attributed to Fuster's Paradigm [52]:

1. **Perception/Action Cycle (PAC):** Extracting information about the environment by processing the received signal, where the amount of information gain increases at each cycle.

2. **Memory:** Encoding and storing the information, and recalling it when queried by some cue.

3. **Attention:** Utilizing computational resources in an effective and efficient manner in order to avoid the information overload problem.

4. **Intelligence:** Enabling an algorithmic decision-making process to choose a strategy for the optimal solution of a predefined objective.

5. **Language:** Enabling an effective and efficient communication between people. Therefore, Haykin discards the language form its model in the concept of cognitive radar and it is not considered further.

Cognition mechanism is often depicted with PAC [71, 60]. Similarly, adaptation of radar parameters for the environment can be depicted with Cycle of Adaptation as depicted in Figure 1.3. Here, adaptation cycle first starts with RF sensor's response acquisition from the environment. This collection is realized by transceiving antenna elements. Upon extracting the relevant information from the data, they are passed to the actuator to interpret and take intelligent actions to optimize the resource allocation and radar parameters. When the new actions are realized and used on the environment, the cycle is complete. The main objective of the adaptation cycle is to mimic human operator to adjust the radar parameters and other computational resources. By doing so, the RF sensor will be able to perceive, understand and adapt to the environment intelligently.

In addition to adaptive optimization processes we can perform on the hardware side, there are certain computational stages we can optimize on the signal processing side as well. For instance, generating high resolution optical images/videos from raw RF data using computationally intensive algorithms in order to obtain range, velocity and angle information can be cumbersome, and also increases the overall response delay of the system. Continuously running these algorithms also allocates so much space in the random access memory (RAM) of the processor and can interfere or block concurrently running computational threads used by other modules of the system. This can be processing of other sensor's data or a DL model used for prediction. Such software-related artifacts can result in software crashes and unfavorable user experience. Therefore, it is important to adapt the signal processing pipeline for the environmental feedback to maximize the system efficiency.

Similarly, for the inference part, various DL models with different computational complexities can be present in the system. Light-weight models can be used for initial device trigger (i.e., wake-word) detection and more sophisticated models can be preferred for actual command sign/gesture understanding. This would alleviate the heavy computational load on the Graphics Processing Unit (GPU) and keep its memory more accessible to other units.

1.4 Vision and Objectives of This Dissertation

This dissertation aims to enable accessible, interactive RF technologies controlled via sign language. While the scope of this dissertation is not complete sign language translation, we aim to explore the ways how RF technology can be utilized to build accessible applications for Deaf/HoH people in an indoor environment. We first tackle the word-level ASL recognition task and propose various signal processing and Machine Learning (ML) techniques to improve the recognition performance of the system. Next, we consider the use case of sign language in a daily living scenario where other human activities can also be part of the observed data. Then, we consider the case of presence of multiple people in the radar Field of View (FoV) and how to recognize activities of people individually. Next, we explore the ways

of automating the data collection and annotation procedure via gamification. Finally, we aim to improve the intelligence of the RF system by dynamically adjusting the waveform parameters based on the actions in the environment.

Putting it all together, we examine different components of a sign language-sensitive smart spaces. While our objective is not to replace our modalities in SLR, we rather show how RF sensors can provide a new source of information and how they can be a well-suited complimentary sensors when other sensors become suboptimal or inadequate.

Outcomes of this dissertation can be leveraged in other applications including building STEM applications to teach RF sensing and AI/ML for high school students, or controlling other electronic devices via sign language. Complete translation of ASL (ASL-to-text or text-to-ASL) will only be viable when a large amount of annotated sentence-level data are acquired, perhaps with the utilization of multiple sensors and adequate amount of computational resources.

1.5 Related Work

Radar-based CPHS applications has been expanded to various fields and applications including HAR [48], fall/fall risk detection [4], gait analysis [148, 67], healthcare monitoring [131] and many more. A great deal of these innovative solutions utilize the recent advances in GPU technology - enabling powerful ML and DL models to be used on end user devices. DL methods used in radar-based recognition technology exploit different RF data representations and processing techniques. This is due to flexible nature of the raw radar data in a sense that it can be processed in many different ways to obtain range, velocity or angle information.

In the past, radar-based recognition frameworks often adopted conventional ML methods such as Support Vector Machines (SVMs) [187], decision trees [126], random forest [174], k-Nearest Neighbors (k-NN) [35] and Dynamic Time Warping (DTW) [156]. Different feature extraction methods are used while employing these methods for recognition problems. Although conventional ML methods work well on different radar-based tasks, they have

several drawbacks which prevent the further improvement of robustness and generalization. Firstly, in order to employ these methods domain expertise need to manually and heuristically extract the features. Further feature selection algorithms may also needed to choose the only most important features to maximize the model's performance. Minimum Redundancy Maximum Relevancy (MRMR) [115], neighborhood component analysis [54] and sequential feature selection can be given as example to some of the popular feature selection methods. Secondly, hand-crafted features, in most cases, refer to low-level statistical information such as mean, median, variance, power level and amplitude, which are task specific. When the model is trained with these low-level features and tested on a new dataset, the performance of the model usually degrades. Therefore, conventional ML methods are not competent enough to train a robust model with good generalization capability.

Deep neural networks (DNNs), on the other hand, tend to break these limitations as a new branch of ML. DL approaches are able to extract a wide range of low-level to high-level features without requiring any manual or heuristic effort, through hierarchical architectures. Moreover, thanks to recent advances in GPU and parallel computing technology, they are capable of processing large amounts of data within a very short time interval. However, since DL approaches are data-driven methods, they require a handful amount of quality data to be able to train the model with a good generalization capability. While there exists a broad range of publicly available datasets for many image-based recognition tasks, this is not the case for radar datasets. They are not only limited by the certain application areas, also the number of samples is often not adequate to train very deep networks. Quality of the provided datasets are also questionable. Moreover, since there exists a large number of commercially available RF sensors, research groups use different RF sensors depending on their budget and project requirements. These sensors' operation characteristics and the hardware/software artifacts observed in the data are so different that it is almost impossible to utilize a data collected with a different sensor model even if they are operating at the same

center frequency. Collecting real radar samples is also a labor intensive, time consuming and an expensive task.

In order to overcome these problems, different synthetic data generation methods are proposed such as Variational Auto-encoder (VAE) [97] and Generative Adversarial Network (GAN) [57]. Most of the data augmentation techniques used in computer vision such as flipping, rotating or shearing are also not applicable to radar data since RF data are often presented to the networks in the form of heat maps, and the location of each blob in the heat map has a physical meaning. Therefore, any modification made on the image would correspond to a physical change in the observed environment and hamper the model's performance. As an alternative to synthetic data generation methods, Transfer Learning (TL) techniques [149] are often employed to initialize the network weights with a better starting point. In this way, a network trained on a different task with large amount data can be leveraged to be used in a different task by only fine-tuning the network weights with a small amount of real data. ImageNet [40] weights which are trained with over one million images can be given as example to one of most popular network initialization methods. While the earlier layers of the DNN models extract the low-lewel features such as edges, corners and curves which are common in most computer vision tasks, deeper layers capture higher-level features such as eyes, lips or nose in a facial image. Therefore, keeping weights of the initial layers fixed and fine tuning the latter layers with the task of interest is a common approach while employing TL methods. Although such pre-trained networks are not initially going to be familiar with spatial features of RF data, they can extract primitive features very effectively and can learn high level RF data features as more data acquired over time.

1.5.1 Gesture versus Sign Language Recognition

Machine learning algorithms are data greedy methods that require large amounts of training samples to enable the network to learn complex models. Thus, it is a common practice for researchers to acquire data from non-native signers, who may not know any sign language, since it is an expeditious source of data. Although ASL is often likened to

gesturing, it is important to recognize that ASL is a *language*, and not reduce signing to mechanical hand and arm movements that can be easily imitated. Thus, while gestures can be made using any participant, studies of ASL require participants for whom sign language is their native language, e.g. Deaf/HoH individuals.

In an earlier study [66], it is found that there are also significant differences between fluent ASL signers and imitation signers who do not know ASL but imitates a learned sign for the experimental study, and signing of hearing imitation signers is distinguishable from that of fluent ASL signers, exhibiting greater kinematic variation, more erratic cadence and significant signing errors. Although some studies, e.g. [89, 162, 32, 49, 116], of ASL recognition have employed hearing imitation signers or ASL learners, perhaps due to the greater ease in recruiting a larger number of participants, the intended benefactor of Deaf spaces are fluent ASL signers.

1.5.2 Machine Learning

Various ML methods are used to learn important data statistics and features of moving targets from the radar signal. A handful amount of studies have been published in the literature for radar target recognition using ML approaches [21, 2, 141]. Rathi et al. [141] proposed a method which uses SVM and Naïve Bayes for classifying airborne targets. An airborne radar is employed to obtain the measurements of the moving targets on the ground and the sea surface. The authors in [2] employed a ground penetration radar data and presented an automatic target recognition method based on an ML algorithm. The proposed system is able to detect complex features which are relevant to threading targets. Carrera et al. [21] presents an ML-based approach for target detection employing radar processors, where the performances of random decision forests and Recurrent Neural Network (RNN) are compared. It is shown that although the ML-based approaches are capable of differentiating the targets from clutter with a good precision, they require feature extraction before the final prediction. Gurbuz et al. [66] has benchmarked k-NNs, random forests and SVMs for

different RF sensors operating at 10, 24 and 77 GHz center frequencies for ASL recognition task.

With the recent advances in DL methods due to data availability and better hardware components (e.g., GPUs), ML-based methods are regarded as outdated and inferior in performance. However, they are lighter methods in terms of the computational complexity when compared to DL methods, and high performances can be obtained even with a low amount of training data.

1.5.3 Convolutional Neural Networks

Convolutional neural networks are one of the most popular DL models in many computer vision tasks. They can effectively capture the spatial relationships of the input data with 2D and 3D kernels. Most popular Convolutional Neural Network (CNN) approaches include VGG-Net [154], Alex-Net [99], Google-Net [165], Res-Net [75], Dense-Net [82] and Mobile-Net [80].

In the last few years, different CNN methods are successfully used with different radar system data types for various tasks such as object detection and recognition, HAR [150], Hand Gesture Recognition (HGR) [164], SLR [106] and many other tasks [107, 19, 6, 129, 38, 173] with high accuracy performance. This is due to ability of convolutional kernels to capture the spatial relationships in different RF data representations such as RDM, RAM, μD, range profiles and point cloud data. In [164], a multi-feature encoder is designed to extract 4D range-velocity-azimuth-elevation information, and followed by a CNN model on an edge computing platform for real-time HGR. One weakness of CNNs is that they cannot capture the temporal relationships in sequential data. Although 3D-CNNs are proposed instead of 2D-CNNs for temporal data, the long term dependencies are still not being captured effectively by the convolutional kernels.

1.5.4 Recurrent Neural Networks

Recurrent neural networks (RNNs) [145, 94], on the other hand, proposed as a new approach to capture the long term dependencies, utilizing memory cells. However, RNNs

suffered from the vanishing and exploding gradient problems, especially when the input sequences are long. Next, Long Short Term Memory (LSTM) networks [77] are proposed as an alternative to RNNs, which were able to overcome the vanishing and exploding gradient problems by using three different gates in their structure, namely, input, output and forget gates.

In radar-based HAR problems, it is shown that RNNs can capture temporal and spatial characteristics of the radar signal, which is crucial in HAR [6]. In [106], a short-time averaging over long-time averaging-based motion detector is implemented to extract the motion detected intervals in a time sequence data. A multi-task learning-based multi-branch temporal CNN + LSTM model is proposed for SLR. In [189], hand gestures are decomposed into sub-classes similar to the phonemes in speech recognition methods. The proposed method predicts the class labels from in-progress gestures in unsegmented input streams.

1.5.5 Encoder-Decoder Networks

Encoder-decoder networks are type of networks which are composed of two subnetworks to map the input data to output data. The encoder part tries to create a dense representation of the input data (i.e., latent space), while the decoder part tries to reconstruct the output from the latent space representation. Such models are heavily used in Natural Language Processing (NLP) and image-to-image translation tasks. Some encoder-decoder variances include VAEs, Convolutional Autoencoder (CAE) and stacked autoencoder (SAE). These models are employed in various radar signal processing tasks as well [176]. Taking the advantage of unsupervised pre-training technique, the network weights can be better initialized before the actual task is started to learn when compared to training the network from scratch. In [93], multi-branch sparse autoencoders are stacked to fuse the information obtained from time-range and time-Doppler maps. In [92], a similar stacked autoencoder approach is presented with a decision level fusion of multiple input representations (i.e., time-range, time-Doppler and RDMs. The decision level fusion is implemented on the softmax outputs with a majority voting approach. Seyfioğlu et al. [150] used convolutional autoencoders to

classify human activities and shown that they can outperform CNN models since the model weights has a better initialization when compared to training from scratch.

1.5.6 Training Under Low Sample Support

Although DL methods are proven to be very effective in various radar-based recognition tasks, they are data-greedy methods. Therefore, when the number of samples in a dataset is not sufficient or the distribution of the training data is different than the testing data, their performance degrade and the models are more likely to be overfitting. In order to eliminate this problem, different DL-based data augmentation methods are proposed which are covered in this section.

Variational Autoencoders

VAEs are a type of encoder-decoder networks whose aim is to generate similar synthetic samples to the original input. They achieve this goal by outputting a 2-dimensional vector with mean and variance from a random variable. The created vector is used to sample an encoding which is passed to the decoder. Since latent spaces are generated from a distribution consisting of the same mean and variance, the decoder part learns from the nearby points referred to the same encoded space. The diversity of the generated samples are controlled through KL divergence. Reducing the KL divergence corresponds to optimizing the mean and variance to be similar to that of the target distribution. Optimizing both the encoder and the decoder parts together (reconstruction loss, decoding and KL divergence loss) generates a latent space which preserves the resembling of close encodings on the local scale. VAEs are considered to be effective generative methods since they work seamlessly on various data types including continuous or discrete, temporal or non-temporal and 1D, 2D or 3D data.

Charlish et al. [25] employed VAEs to generate non-linear FM radar waveforms using a custom reconstruction loss. The proposed VAE has capability to synthesize new radar waveform modulations which have required ambiguity function characteristics, even though they were not represented in the training data. In [91], VAE is coupled with an RNN to

compute the anomaly level of the body motion based on the acquired point cloud. The proposed model generates a spike in the anomaly level when an abnormal motion, such as fall, occurs. Stephan et al. [159] proposed a parametrically constrained VAE with residual and skip connections, which can generate the clustered and localized target detections on the RA map. They present domain adaptation strategies whereby the neural network is first trained using ray tracing based model data and then fine-tuned on the real sensor data. This method improves the generalization and scalability of the proposed model even though it is trained with limited real radar data.

Generative Adversarial Networks

GANs are one of the most popular generative methods. They are composed of two subnetworks: a generator and a discriminator. While generator takes a noise vector as input and tries to output a fake sample which resembles to the data distribution, the discriminator takes both fake and a real sample and tries to differentiate them. The generator tries to minimize a joint loss function while the discriminator tries to maximize it. GANs take longer time to train when compared to VAEs, and they more sophisticated architectures for generative modeling. Therefore, the use of GANs is considered and proved a lot more stable.

Labeling of the real data is one of the most labor-intensive tasks in computer vision and radar-based recognition problems. However, utilizing the unsupervised generative models like GANs, a large amount of synthetic data can be generated in an unsupervised manner to be used in the training stage without needing the labor-intensive labeling tasks. Although GANs are powerful methods for generative modeling, they suffer from a critical issue called mode collapse where the model starts to output the same fake after certain number of iterations. More recent GAN architectures propose various techniques to overcome this specific problem.

Lekic et al. [109] proposed a Conditional Multi-Generator Generative Adversarial Network (CMGGAN) which can produce scene images conditioned on the radar sensor measurements.

The proposed model fuses the features from both radar and camera sensors. Rahman et al. [137, 138, 134] tackled the problem of synthetic μD spectrogram generation using Physics-Aware Generative Adversarial Network (PhGAN) architectures for HAR and SLR tasks. They present a modified GAN architecture which makes use of the envelope trajectory of the spectrograms. A custom loss function is proposed which takes the consistency of the envelopes of the generated fake samples into account. This helps loss function to better guide the training process, and generate more kinematically accurate μD samples.

Transfer Learning

TL is a sub-field of ML whose aim is to store the knowledge gained while solving a problem and applying it to another but related problem. It is a popular approach in many computer vision and NLP tasks. TL methods are alternative ways to synthetic data generation techniques under low sample support scenarios. They can be utilized in combination with other generative methods as well. In computer vision tasks, applying TL with ImageNet weights is a common method, however since radar data looks significantly different than other image datasets, the TL may not necessarily be as efficient as in the case of computer vision tasks.

Seyfioglu et al. [150] used CAE to pretrain the network weights in an unsupervised manner for identity mapping. The decoder part of the trained CAE is then removed and the model is augmented with fully connected layers for classification. The modified architecture transforms into a CNN model with pretrained weights instead of random or Gaussian weight initialization. Later added layers are, then, fine-tuned for the task of interest. The pretrained CAE model performed significantly better than other ImageNet-based TL methods. A similar approach is followed in [68] for HAR using multiple radars operating at different center frequencies. It is shown that while CNN models trained at different center frequencies perform poorly on a cross-frequency dataset testing, TL across different frequency bands helps to alleviate this problem.

Huang et al. [85] proposed a method to transfer knowledge gained from a large unlabeled dataset to small amount of labeled dataset. The proposed CNN architecture is composed of stacked CAEs, along with a feedback bypass additionally. First, the reconstruction pathway with stacked CAEs is trained in an unsupervised manner. Then, the pretrained CNN layers are reused to transfer knowledge to the classification task, with feedback bypass introducing the reconstruction loss simultaneously. In [84], fully connected network and U-Net based two TL methods are proposed for classification with only 50 image patches. Chen et al. [26] proposed a modified CNN which incorporates expert knowledge of target scattering mechanism interpretation and polarimetric feature mining that assists the training of the model and increases the classification performance.

Zhang et al. [188] presents a semi-supervised TL method based on GANs. Initially, the GAN is trained with a variety of unlabeled samples in order to learn generic features of radar images. Next, the learned network parameters are reused to initialize the target network weights to transfer the knowledge gained from the unsupervised stage to specific recognition task. Finally, the network is fine-tuned in a semi-supervised manner using both the labeled and unlabeled training samples. It is shown that the proposed TL method outperforms the randomly initialized model by accuracy difference of 23.58%. Zheng et al. [190] proposed a semi-supervised recognition method composed of a GAN and a CNN. A handful amount of unlabeled images are generated using the GAN, and they are fed into the CNN subnetwork as input along with the original labeled images. In order to address the mode collapsing issue faced in GANs, a dynamic adjustable multi-discriminator GAN architecture is introduced. At the same time, label smoothing regularization method is applied to better regularize the semi-supervised recognition model of the CNN. In other studies [50, 72, 96, 152], publicly available pretrained models and radar datasets are utilized for TL.

Other RF Data Augmentation and Model Regularization Techniques

Data augmentation techniques used in computer vision such as mirroring, shearing, random cropping cannot be directly employed in radar-based recognition methods because of the structural differences in the data types. Radar images correspond to physical measurements and such distortions change the underlying physical structure of the observed scene. This section covers the data augmentation and model regularization techniques tailored to RF data.

In order to eliminate the overfitting problem on a small RF dataset, Ding et al. [43] proposed three ways to augment the data, namely, translation, speckle noising and pose synthesis. Pei et al. [133] introduced a multi-view DL framework for limited RF data. They present a novel multi-branch CNN architecture to generate multiple views of the data as input. The features of multiple views are progressively fused in consecutive layers of the network. Song et al. [158] introduced an autoencoder-based cyclic network using adversarial learning to generate synthetic samples at different azimuth angles. Hua et al. [81] proposed a dual-channel CNN model for classifying the dataset with a small number of labeled samples. The proposed method, first, enlarges the labeled sample set using a neighborhood minimum spanning tree, and then extracts the spatial features using the dual-channel CNN.

Zhang et al. [186] proposed a feature augmentation and ensemble learning method. The selected features from the CNN layers are concatenated to obtain an enriched representation for the recognition. The Adaboost rotation forest is proposed instead of using a softmax layer for classification to realize the low sample-based recognition task with merged features. In [181, 151] fully connected layers are replaced with CNN layers and deep memory CNNs are proposed to alleviate the overfitting problem caused by low-sample support. Zhai et al. [185] presents a transferred max-slice CNN with L2-regularization term. The proposed method augments the feature space and enables the recognition of the targets with a greater performance using a small number of samples.

1.6 Challenges

Just like any other modality radars also have certain limitations yet to be tackled. The proposed radar-based recognition methods in this work make use of the dynamic movements of the arms, hands and fingers. Therefore, static hand shapes cannot be recognized since they require 3D imaging of the skeleton joints. Although imaging radars have capability to reconstruct the depth map of the objects and the scene, they require longer observation times and large number of TX-RX antenna elements in both azimuth and elevation dimension. Schuessler et al. [146] proposed a radar-based solution to detect static hand shapes in ASL. However, since it is an imaging radar, the target (hands) should be stationary during radar's observation time. Having longer observation times reduces the frame rate and hinders the applicability of the proposed solution for the recognition of dynamic motions which are crucial for most of the signs. In addition, the designed radar system has 47 RX antenna in both azimuth and elevation directions with a bulky hardware, which makes it hard to deploy in indoor environments with limited spaces.

In this work, we are not concerned with the reconstruction of the object. Instead, our objective is to extract the characteristics of kinematic motions over time and utilize them to recognize different activities and signs. It has been shown that high temporal resolution of radar can compensate for the low spatial resolution for hand gesture recognition applications [112]. Note that motivation of employing RF sensors in CPHS applications is not to replace other sensor modalities with radar. Instead, we show the capabilities of the radar with its distinct advantages like high temporal resolution and high sensitivity towards radial motions, and show that radars have valuable information to offer where other sensors can become suboptimal or inefficient. Therefore, radars can be great complimentary sensors when integrated into a system with other sensor modalities like camera or LiDAR.

RF-based recognition methods in CPHS applications have several other challenges which preclude machine understanding of human activities and gestures with high accuracy and robustness:

1. **Temporal Segmentation:** RF data are usually in the form of time stream of raw In-Phase and Quadrature (I/Q) ADC samples. There is no prior information regarding where a motion is occurring in the time-series data which is crucial to spot when an activity or gesture starts and ends. Missegmentation of the data can result in incomplete actions/gestures which would confuse classifiers and can cause wrong predictions. Therefore, a way to autonomously and accurately segment the data in temporal domain is a necessity.

2. **Open-Set Problem:** The number of possible human motions/activities are almost infinite. However, most of the existing studies limit the number of experimental activities to be very small (10-20). These methods are often not expandable to new classes, and when the RF sensor experiences a data sample which does not resemble to any of the learned classes, it is unconfidently dumped into one of the classes. Usage of an "unknown" class is also a sub-optimal solution since the unknown sample's data distribution might resemble more to one of the valid classes, and the representation and generalization capability of the unknown class is also questionable considering so many possible human movements.

3. **Multi-Person Differentiation:** When multiple people exist in the radar FoV, back-scattered signals from all the targets superimpose on top of each other. This causes generated RF data representations to have signatures from different targets and it becomes challenging for classifiers to identify the correct motion. Resolving targets in range-angle domain and differentiating their signatures from each other is both a crucial and a challenging task, especially when targets are close to each other.

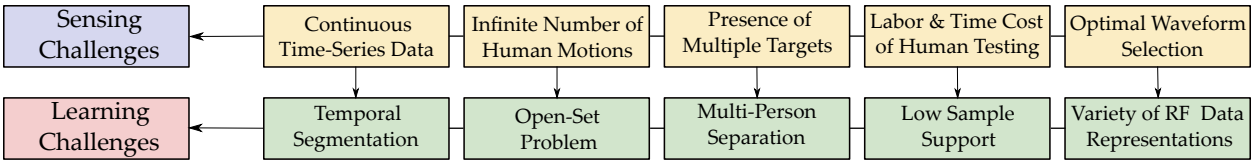
4. **Data Scarcity:** Collecting real RF data with human subjects in a laboratory environment is a time consuming, labor intensive and an expensive task. This becomes even more challenging when the focus group of people are not easily accessible (e.g., Deaf community). In addition, there is no consensus on the radar type, waveform parameter

selection and format of the publicly available datasets, leaving them unusable when any of these items do not match-up.

5. **Robustness Across Multiple Tasks:** Radar waveform parameters affect range, velocity and angle resolutions along with their maximum and minimum unambiguous boundaries. They are often optimized according to application requirements. However, when the applications consists of different tasks, using the same waveform can be suboptimal and can introduce unnecessary computational overhead to the system. Therefore, the waveform parameters need to be adjusted based on dynamic environmental changes and for different tasks in an autonomous fashion.

Sensing challenges in RF data are the root causes of the subsequent learning challenges. Figure 1.4 depicts the aforementioned issues yet to be solved.

Figure 1.4: Sensing and learning challenges of RF data.



1.7 Contribution of This Dissertation

RF-based CPHS applications present both sensing and learning-related challenges. Overlooking or omitting any of these issues can result in suboptimal and unfavorable CPHS applications. This dissertation aims to tackle all of these challenges individually, but in a correlated and compatible fashion. Main contributions of this dissertation can be listed as:

1. **Automated Motion Detector:** RF data are often in the form of a time-series of complex I/Q samples. An automated labeling strategy is needed to locate where a motion is starting and ending. This work presents a short-time averaging over long-time averaging (STA/LTA) based motion detector [106] to spot the starting and ending times of a motion. The method first generates the μD spectrogram of the data. Then, upper and lower envelopes of the spectrogram are extracted, and their Euclidean distance is computed. The resulting 1D vector is passed to STA/LTA-based detector to find the starting and ending points of the motions. A 1D binary masking vector with the same length as the input vector is outputted for subsequent segmentation of individual activities. Details of the proposed method are explained in Section 3.4.

2. **RF Data Fusion Classifier:** The raw complex RF data do not provide any meaningful information without certain pre-processing steps to extract range, velocity and angle measurements through various signal processing techniques. It is common practise to generate μD spectrograms, RDMs, RAMs, range profiles, point clouds and other data representations to visualize and inspect the data. These data representations are further used to train DL classifiers and enabling them to learn certain spatial features of RF data better instead of enforcing the Deep Neural Network (DNN) model to figure out these features by itself. In this dissertation, we propose a Joint-Domain Multi-Input Multi-Task Learning (JD-MIMTL) network [106] to fuse information gathered from multiple RF data representations. Each data representation is processed parallelly with time-distributed 2D and 3D CNN blocks followed by LSTM layers to capture the

temporal dependencies. Auxiliary tasks for SLR such as one versus two handedness, major hand location, hand movement type, daily activity versus SL and number of arm strokes are used to regularize the network training. The proposed method is shown to outperform other State-of-the-Art (SOTA) methods by a large margin. Architectural and implementation details of the JD-MIMTL network are discussed in Section 3.5.4.

3. **Trigger Sign Detection:** In order to activate a SL-enabled CPHS device, a wake sign should be accurately recognized within a stream of RF data, and the device should activate only when the articulation of the wake sign is completed. Trigger sign detection task differs from conventional classification in a sense that trigger sign should be intuitive and making sense for the context. In addition, it should be recognized with high recognition rate to prevent false alarms and false rejections (i.e., missing the actual trigger). Incomplete or overextended trigger attempts should also be taken into account. Therefore, this work studies the design considerations of trigger signs both from device perception and user experience point of views. In this work, an adaptive double-threshold Cumulative Score Aggregation (CSA) approach [106] is proposed to recognize the wake sign in RF data streams. The proposed method is shown to yield better detection rates with lower false rejections while preserving the low false alarm rate. Details of the proposed motion detection approach are provided in Section 3.6.1.

4. **Automation of Natural ASL Data Collection via Gamification:** An important challenge in SL-sensitive CPHS applications is the lack of publicly available RF datasets for model training. Not just the amount of data, but the quality of data is also critical. Just like any other language, SLs have their own grammar, linguistic features, dialectal nuances and diversities across people. Traditional way of collecting SL data in a laboratory environment from people whose primary way of communication is not SL results in pristine datasets which do not well represent the features of actual SL. Training learning models with such datasets unsurprisingly yield overoptimistic

results. When these models are tested on real-world SL data collected free from experimental restrictions and assumptions, they are likely to underperform and cause poor user experience. In this work, we developed an ASL-controlled chess game where the pieces are moved on the board with ASL sign articulations instead of mouse clicks. The proposed method [104] acquires data from both camera and an RF sensor simultaneously during the game play, and processes it, uploads it to a cloud platform, and runs the prediction model and the chess engine in the backend. This CPHS application eliminates the need for an external operator to monitor the data collection procedure. Moreover, the game interactively communicates with the user through a pop-up window to correct the ground truth labels of the mispredicted samples. Finally, the game is designed in a way that more words can be added to its dictionary. Such flexibility paves the way to curate a large and diverse multi-modal SL datasets for SLR tasks. This approach enables acquisition of natural SL, in an enjoyable and sustainable manner in long term. Implementation and gameplay details of the designed SL-controlled chess game are provided in Section 5.2.2.

5. **Multi-Person Separation:** Most RF-based CPHS techniques assume only one target in the radar FoV. However, presence of multiple targets in the scene is a very typical case in real-world environments. Their backscattered radar signal returns superimpose on top of each other, making generated data representations hard to interpret. Consequently, DL models trained with single target data samples will fail to correctly classify the activity of the targets individually. Many of the CPHS works considering multiple targets focus on counting the number of people present and tracking them in a room environment. Moreover, separation techniques applied on the pre-processed data are not scalable to other domain representations. In this work, we propose an Angular Subspace Projection-Based Separation (ASPS) method [103] to resolve the raw data of the targets in the scene. This approach differs from current SOTA target separation methods in a sense that it can output projected raw data for each target. This low-level

separation technique enables to generate any RF data representation using the project raw data for a particular target. Efficacy of the proposed method is demonstrated for a HAR application in an end-to-end framework. In addition, a multi-view DNN is also proposed for very close targets when the angular resolution of the device is not sufficient to completely resolve the targets in angular domain. Algorithmic details and performance for varying number of antenna elements of the ASPS method are explained and discussed in Section 4.2.3.

6. **Human-Centric Adaptive RF Sensing:** It is a common practise to optimize the radar waveform parameters based on the application requirements. However, for RF-controlled CPHS applications, these parameters are often fixed before deploying the system. Continuously running the radar system in its fully functional mode with high sampling rate and a large bandwidth allocates so many resources like high memory and RAM usage to store and process the data, GPU memory to make inference using the DL-based prediction model on the edge device. Therefore, a more strategic approach is needed to intelligently switch between working modes. This work presents an automated framework to switch between different working modes of the RF system based on the activities of the person. To the best of our knowledge, this dissertation is the first study on introducing human-centric adaptive RF system where radar takes action to adjust the waveform parameters and the operation characteristics according to human behavior. The proposed system not only optimize the resource allocation intelligently, but also reduces the computational overhead significantly without compromising the recognition performance. Details of the proposed fully-adaptive RF waveform selection approach are discussed in Section 6.3.

1.8 List of Publications

The section lists all the articles published as the accomplishment of this dissertation.

1.8.1 Journals (7)

[66] Sevgi Z. Gurbuz, Ali Cafer Gurbuz, Evie A. Malaia, Darrin J. Griffin, Chris S. Crawford, Mohammad Mahbubur Rahman, Emre Kurtoglu, Ridvan Aksu, Trevor Macks, and Robiulhossain Mdrafi. American sign language recognition using rf sensing. IEEE Sensors Journal, 21(3):3763–3775, 2021.

[69] Sevgi Z. Gurbuz, M. Mahbubur Rahman, Emre Kurtoglu, Evie Malaia, Ali Cafer Gurbuz, Darrin J. Griffin, and Chris Crawford. Multi-frequency rf sensor fusion for word-level fluent asl recognition. IEEE Sensors Journal, 22(12):11373–11381, 2022.

[106] Emre Kurtoğlu, Ali C. Gurbuz, Evie A. Malaia, Darrin Griffin, Chris Crawford, and Sevgi Z. Gurbuz. Asl trigger recognition in mixed activity/signing sequences for rf sensor-based user interfaces. IEEE Transactions on Human-Machine Systems, 52(4):699–712, 2022.

[67] Sevgi Z. Gurbuz, Emre Kurtoglu, M. Mahbubur Rahman, and Dario Martelli. Gait variability analysis using continuous rf data streams of human activity. Smart Health, 26:100334, 2022.

[124] Evie A. Malaia, Joshua D. Borneman, Emre Kurtoglu, Sevgi Z. Gurbuz, Darrin Griffin, Chris Crawford, and Ali C. Gurbuz. Complexity in sign languages. Linguistics Vanguard, 9(s1):121–131, 2023.

[103] Emre Kurtoğlu, Sabyasachi Biswas, Ali C. Gurbuz, and Sevgi Zubeyde Gurbuz. Boosting multi-target recognition performance with multi-input multi-output radar-based angular subspace projection and multi-view deep neural network. IET Radar, Sonar & Navigation, 17(7):1115–1128, 2023.

[104] Emre Kurtoğlu, Kenneth DeHaan, Caroline Kobek Pezzarossi, Darrin J. Griffin, Chris Crawford, and Sevgi Zubeyde Gurbuz. Interactive learning of natural sign language with radar. IET Radar, Sonar & Navigation (Accepted), 2024.

1.8.2 Conference Proceedings (13)

[65] Sevgi Z. Gurbuz, Ali C. Gurbuz, Evie A. Malaia, Darrin J. Griffin, Chris Crawford, M. Mahbubur Rahman, Ridvan Aksu, Emre Kurtoglu, Robiulhossain Mdrafi, Ajaymehul Anbuselvam, Trevor Macks, and Engin Ozcelik. A linguistic perspective on radar micro-doppler analysis of american sign language. In 2020 IEEE International Radar Conference (RADAR), pages 232–237, 2020.

[68] Sevgi Z. Gurbuz, M. Mahbubur Rahman, Emre Kurtoglu, Trevor Macks, and Francesco Fioranelli. Cross-frequency training with adversarial learning for radar micro-Doppler signature classification (Rising Researcher). In Kenneth I. Ranney and Ann M. Raynal, editors, Radar Sensor Technology XXIV, volume 11408, page 114080A. International Society for Optics and Photonics, SPIE, 2020.

[64] Sevgi Z. Gurbuz, Ali C. Gurbuz, Evie A. Malaia, Darrin J. Griffin, Chris Crawford, Emre Kurtoglu, M. Mahbubur Rahman, Ridvan Aksu, and Robiulhossain Mdrafi. Asl recognition based on kinematics derived from a multi-frequency rf sensor network. In 2020 IEEE SENSORS, pages 1–4, 2020.

[3] Oladipupo O. Adeoluwa, Sean J. Kearney, Emre Kurtoglu, Charles J. Connors, and Sevgi Z. Gurbuz. Near real-time ASL recognition using a millimeter wave radar. In Kenneth I. Ranney and Ann M. Raynal, editors, Radar Sensor Technology XXV, volume 11742, page 1174218. International Society for Optics and Photonics, SPIE, 2021.

[100] Emre Kurtoglu, Ali C. Gurbuz, Evie Malaia, Darrin Griffin, Chris Crawford, and Sevgi Z. Gurbuz. Sequential classification of asl signs in the context of daily living using rf sensing. In 2021 IEEE Radar Conference (RadarConf21), pages 1–6, 2021.

[135] M. Mahbubur Rahman, Emre Kurtoglu, Robiulhossain Mdrafi, Ali C. Gurbuz, Evie Malaia, Chris Crawford, Darrin Griffin, and Sevgi Z. Gurbuz. Word-level asl recognition and

trigger sign detection with rf sensors. In ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8233–8237, 2021.

[105] Emre Kurtoğlu, Ali C. Gurbuz, Evie Malaia, Darrin Griffin, Chris Crawford, and Sevgi Z. Gurbuz. Rf micro-doppler classification with multiple spectrograms from angular subspace projections. In 2022 IEEE Radar Conference (RadarConf22), pages 1–6, 2022.

[136] M. Mahbubur Rahman, Emre Kurtoğlu, Muhammet Taskin, Kudret Esme, Ali C. Gurbuz, Evie Malaia, and Sevgi Z. Gurbuz. Performance comparison of radar and video for American sign language recognition. In 2022 IEEE Radar Conference (RadarConf22), pages 1–6, 2022.

[169] Emin Ucer, Emre Kurtoglu, Mithat Kisacikoglu, Ali C. Gurbuz, and Sevgi Z. Gurbuz. Local detection of oltc operation to support decentralized control of active end-nodes. In 2022 IEEE Power & Energy Society General Meeting (PESGM), pages 1–5, 2022.

[70] Sevgi Z. Gurbuz, M. Mahbubur Rahman, Emre Kurtoglu, and Dario Martelli. Continuous human activity recognition and step-time variability analysis with fmcw radar. In 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), pages 01–04, 2022.

[63] Sevgi Z. Gurbuz, Chris Crawford, Darrin J. Griffin, Emre Kurtoglu, Oladipupo Adeoluwa, and Josh Haeker. Interactive rf game design for deciphering real-world human motion: Activities, gestures, and sign language. In 2023 IEEE Radar Conference (RadarConf23), pages 1–6, 2023.

[101] Emre Kurtoğlu, Moeness G. Amin, and Sevgi Z. Gurbuz. Radar-based joint human activity and agility recognition via multi-input multi-task learning. In 2024 IEEE Radar Conference (Accepted), pages 1–6, 2024.

[102] Emre Kurtoğlu, Kenneth DeHaan, Caroline Kobek Pezzarossi, Darrin J. Griffin, Chris Crawford, and Sevgi Z. Gurbuz. Gamification of rf data acquisition for classification of natural human gestures. In 2024 IEEE Radar Conference (Accepted), pages 1–6, 2024.

CHAPTER 2

RADAR BACKGROUND

2.1 Introduction

RADAR is an acronym for "RAdio Detection and Ranging". It is a technology that uses radio waves to detect, locate, track and identify objects. It is used in a wide range of applications including aviation, military, navigation, weather monitoring and traffic control. Radar systems are often customized to meet the application requirements and for specific objectives. For instance, weather radar is used to monitor precipitation and severe weather conditions. Air traffic control radars are used to track the position of planes and aircrafts in the airspace. Military radars are used to detect missiles, aircrafts and other potential targets. Ground penetrating radars are used for subsurface imaging in archaeological and geological studies. Police radars are used to measure the velocity of the vehicles on the roads.

Until the last decade, radar was mostly associated with military, intelligence and defense related applications with strict regulations. In the past decade, recent developments of low-cost, small package, high frequency radar systems enabled more civilian applications of off-the-shelf radars. Automotive radars are started to be used in semi-autonomous and fully-autonomous driving applications. Smart home devices and even cell phones (e.g., Google Pixel 4) have started to utilize radars as a complimentary sensor. Radar's capability to provide range, velocity and angle information of the objects in the scene with high resolution makes it an indispensable sensor for certain applications.

2.2 Radar Principle

Radar operates based on the principles of broadcasting radio waves, receiving their back-scattered reflections from objects in the environment, and analyzing the received signal to extract certain information about the objects. First, the radar system generates short pulses of radio waves using a transmitter antenna. The generated pulses propagate through the airspace with the speed of light. A portion of the waves is reflected back to radar when they encounter an object or obstacle. Reflected signal strength is related to object's shape, material and reflectivity coefficient. This constant is often referred as Radar Cross Section (RCS). The reflected waves are received by a receiver antenna. The receiver module is typically scheduled to listen for reflected waves during the idle times between transmitted pulses.

Radial distance of the targets from the radar antenna can be computed from the round-trip time delay of the transceived waves. Velocity of the moving targets can be found using the Doppler principle. Moving objects in the radar FoV cause frequency shift between the transmitted the received waves. This phenomenon is called Doppler shift, and it's value is related to object's moving direction, center frequency of the radar system and the radial speed of the object. If there are multiple transmitter or receiver antennas, azimuth and elevation angles of the targets can also be computed by analyzing the phase shift between the antenna elements considering the antenna array geometry. The received signal often undergoes certain signal processing steps to enhance the signal quality, filter out the noise and clutter, and extract useful information. The specifications and capabilities of a radar system can vary depending on the use case.

2.2.1 Radar System and Hardware

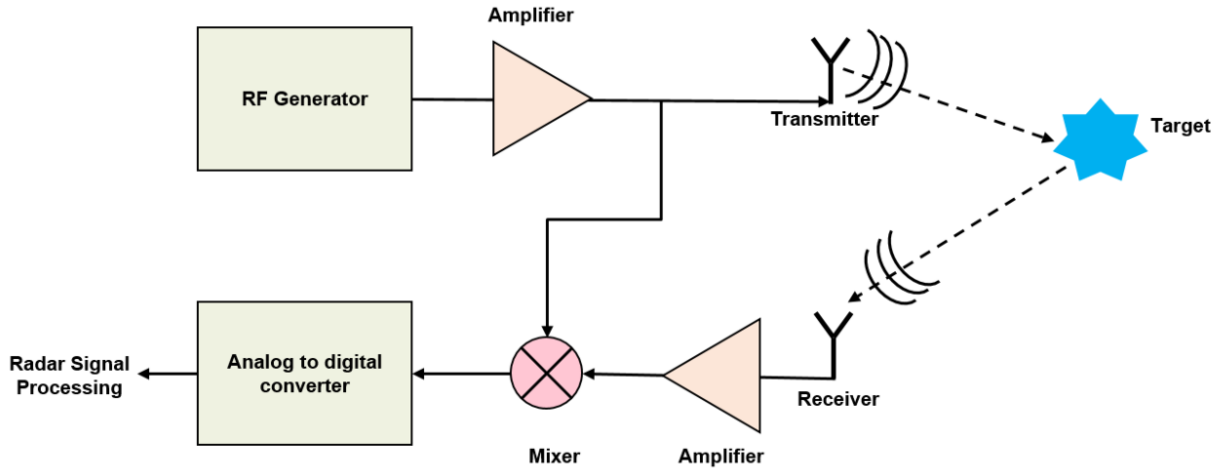
A typical radar system consists of the following components:

- RF Generator
- Amplifiers

- Transmitter
- Receiver
- Mixer
- Analog to Digital Converter

An electromagnetic wave is created, amplified and transmitted through RF Generator, Amplifier and Transmitter, respectively. The back-scattered signal from the objects is collected by the Receiver, amplified and passed to a Mixer which mixes it with the transmitted signal. The output signal from the mixer is called as Intermediate Frequency (IF) signal. IF signal's instantaneous frequency and phase are equal to the difference of instantaneous frequencies and phases of the two input signals, respectively. The IF signal is sampled and digitized by an ADC to be further processed by a computer. Figure 2.1 depicts the overall block diagram of a typical radar system.

Figure 2.1: Radar system block diagram [86].



2.2.2 Radar Types Based on the Waveform

This section discusses and provides information about different types of radar systems. Radar systems can be classified into two types based on the type of signal they are operating with:

- Pulse Radar
- Continuous Wave Radar

Pulse Radar

Pulse radar operates by transceiving a high power signal. It waits for the reflected signal to get received and then transmits the next signal. It employs a single antenna for both transmitting and receiving signals with the help of a duplexer which isolates the receiver from the transmitter modules while permitting them to share a common antenna. The antenna transmits a pulse signal at every clock cycle. The time interval between the two clock pulses should be long enough so that the reflected signal corresponding to the current clock pulse should be collected before the next clock pulse.

A variant of pulse radar is Moving Target Indication (MTI) Radar. It uses the Doppler effect phenomenon to differentiate the moving targets from stationary objects.

Continuous Wave Radar

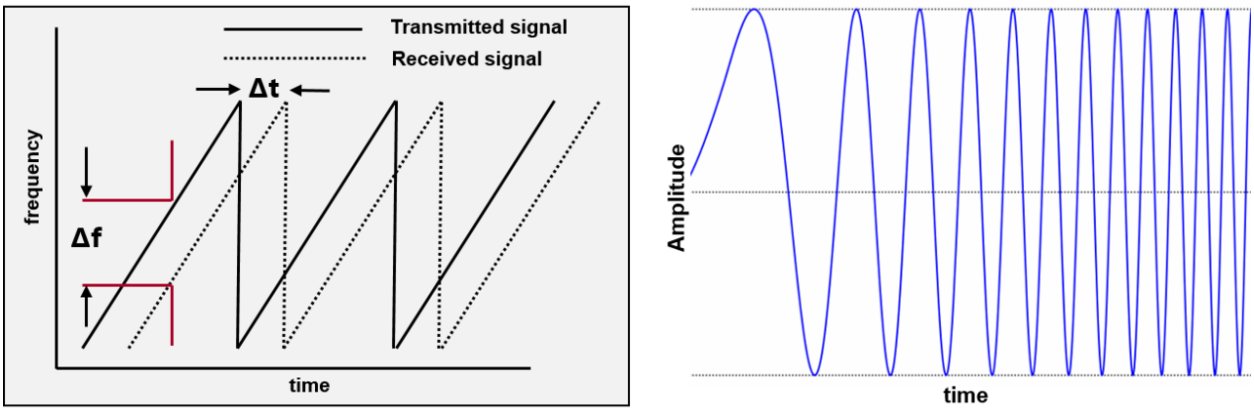
Continuous Wave (CW) radar transmits a continuous signal or wave at a constant frequency at all times. Similarly, they utilize Doppler effect for moving target detection and speed measurement. CW radar systems can further be divided into two categories: Modulated and Unmodulated radar systems.

Unmodulated CW radar systems require two separate antennas for transmitting and receiving the signal. They can only measure the velocity of the targets, but not the radial distance of the targets from the radar. Velocity is extracted from the instantaneous rate of change of the target's radial range by calculating the Doppler shift of the reflected signal.

Frequency Modulated Continuous Wave Radar

A variant of modulated CW radar is called Frequency Modulated Continuous Wave (FMCW) radar whose center frequency linearly increases during each pulse. Each pulse sweeps a

Figure 2.2: Chirp signal representation [86].



certain bandwidth. Such pulse whose frequency increases with a constant rate over time is called a chirp. This radar system requires two separate antennas for transmitting and receiving the radio waves. It can measures not only the velocity of the target, but also the radial distance of the target from the radar system. Figure 2.2 (left) shows the time-frequency representation while the right one shows the time-amplitude plot for a chirp signal. A chirp is typically characterized by its start frequency (f_c), bandwidth (BW) and duration (T_c). The slope (S) of the chirp defines the rate at which the chirp ramps up.

Figure 2.3: FMCW radar system block diagram [86].

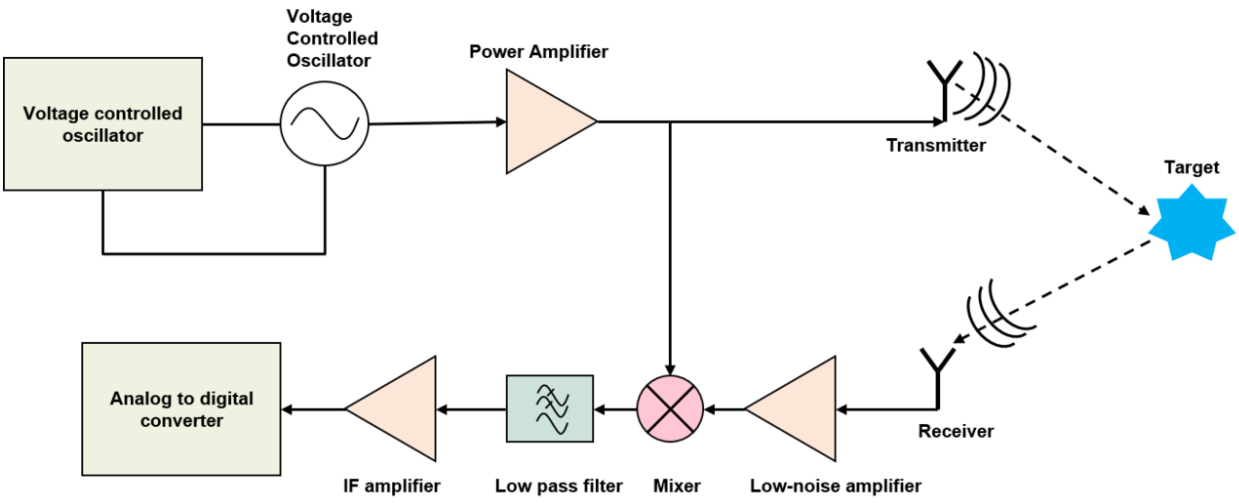


Figure 2.3 depicts the FMCW radar system. Here, the frequency synthesizer generates the FMCW chirp together with the Voltage Controlled Oscillator (VCO). The output of the power amplifier is passed to transmitter antenna for broadcasting, and mixer to down-convert the received and amplified signal. The resulting IF signal from the mixer is then low pass filtered, amplified and finally passed to an ADC. The sampled signal is then transferred to a computing device for further processing.

A frequency mixer is a 3 port device with 2 inputs and 1 output which combines two input signals to generate a new signal with a new frequency [87, 140]. For two sinusoidal input signals x_1 and x_1, the output signal, x_{output}, can be written as:

$$x_1 = \sin(\omega_1 t + \phi_1) \tag{2.1}$$

$$x_2 = \sin(\omega_2 t + \phi_2) \tag{2.2}$$

$$x_{output} = \sin((\omega_1 - \omega_2)t + (\phi_1 - \phi_2)) \tag{2.3}$$

Notice that the instantaneous frequency of x_{output} is equal to the difference of the instantaneous frequencies of x_1 and x_2. The phase of the x_{output} is also equal to the difference of the phases of x_1 and x_2.

The fundamental advantages of an FMCW radar include:

- Ability to measure both range and velocity simultaneously.
- High range and velocity resolution.
- Enabling of performing signal processing at a low frequency range.

2.3 FMCW Radar Metrics and Parameters

The received FMCW signal is a time-delayed and frequency-shifted version of the transmitted signal. The time-of-flight of the transceived signal, τ can be derived as:

$$\tau = \frac{2d}{c} \tag{2.4}$$

where d is the radial distance of the object from the radar platform and c is the speed of light. A single target in the radar FoV creates an IF signal with a constant frequency given by:

$$\text{IF}_{frequency} = \frac{2Sd}{c}, \text{ where } S = \frac{BW}{T_c} \tag{2.5}$$

where S is the slope of the chirp. IF bandwidth is limited by the ADC sampling rate F_s given by:

$$F_s = \frac{S2R_{max}}{c} \tag{2.6}$$

where R_{max} is the maximum unambiguous range. This gives the maximum unambiguous range of the radar as:

$$R_{max} = \frac{F_s c}{2S} = \frac{F_s c T_c}{2BW} \tag{2.7}$$

Range resolution, $R_{resolution}$, of a radar system is the ability to differentiate two or more objects in range domain. It solely depends on the BW and can be computed by:

$$R_{resolution} = \frac{c}{2BW} \tag{2.8}$$

It can be noticed that higher BW yields better range resolution. The relationship between R_{max} and $R_{resolution}$ can also be written as:

$$R_{max} = \frac{R_{resolution}}{2} N_{samples} \tag{2.9}$$

where $N_{samples}$ is the number of ADC samples per chirp. When two targets are closer than $R_{resolution}$, they will look like a single target in the frequency spectrum. The maximum allowed BW of the RF system should comply with Federal Communications Commission (FCC) legislation.

When two targets moving at different speeds are equidistant from the radar, their range spectrum will have a single peak they cannot be resolved in range domain. Phases of the two targets can be used to estimate the velocity of the targets by computing the phase difference measured across two consecutive chirps. The maximum unambiguous velocity, v_{max} that can be measured by consecutive two chirps is given by:

$$v_{max} = \frac{\lambda}{4\text{PRI}},\ \text{PRI} = \frac{1}{\text{PRF}},\ \lambda = \frac{c}{f_c} \tag{2.10}$$

where λ is the wavelength in meters, f_c is the center frequency, PRI is the pulse repetition interval and PRF is the pulse repetition frequency. If any of the targets move faster than v_{max}, aliasing (i.e., wrapping of the peak to the opposite side of the spectrum) in Doppler spectrum would be observed. Estimating over more chirps instead of two would yield better estimation results. Coherent processing of N_c chirps constructs a Coherent Processing Interval (CPI) (i.e., frame). The velocity resolution, $v_{resolution}$, can then be written as:

$$v_{resolution} = \frac{2v_{max}}{N_c} \tag{2.11}$$

The total frame duration, T_{frame}, can also be computed as:

$$T_{frame} = N_c\text{PRI} \tag{2.12}$$

Figure 2.4 depicts the attributes of a typical FMCW radar chirp, while Figure 2.5 shows the frame structure.

In order to locate the object in the 2D or 3D space, azimuth and elevation angles of the object is also needed in addition to the radial distance. DoA of a target can be estimated by

Figure 2.4: Typical FMCW chirp [42].

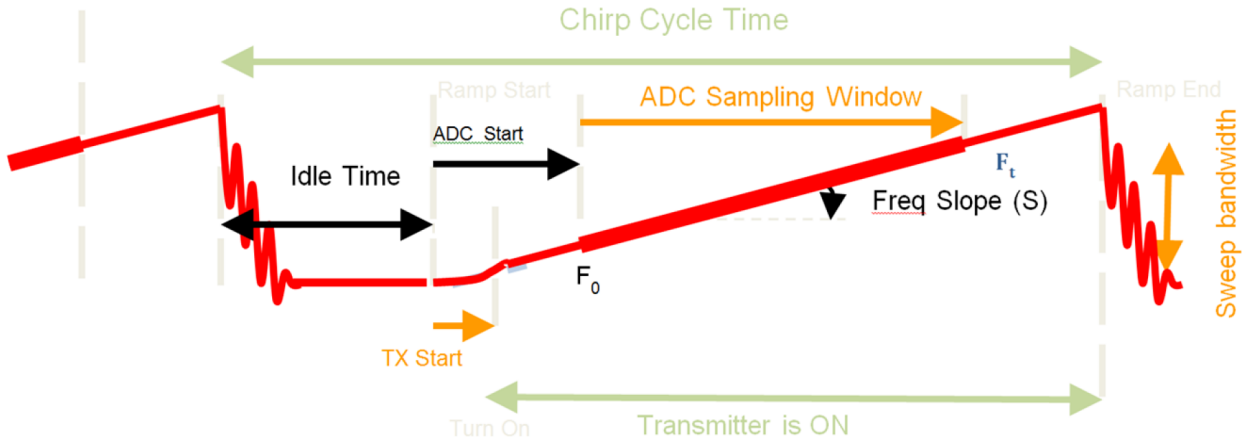
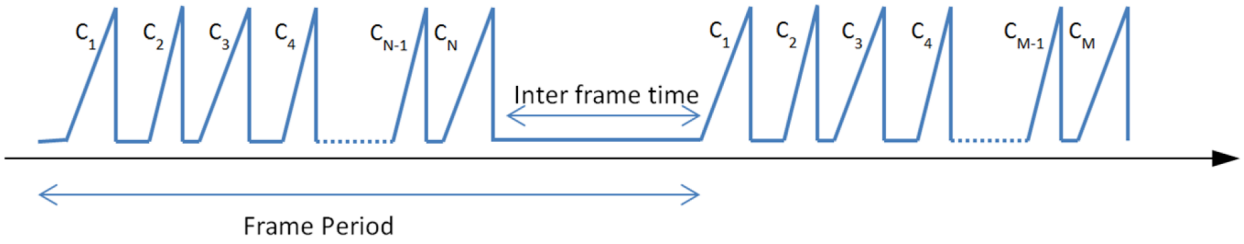


Figure 2.5: Typical FMCW frame structure [42].



utilizing the phase difference of the collected signal from multiple receivers which are spaced apart with a distance, d. Therefore, estimating the DoA of an object requires at least two RX antennas.

Figure 2.6: DoA estimation using two RX antennas [139].

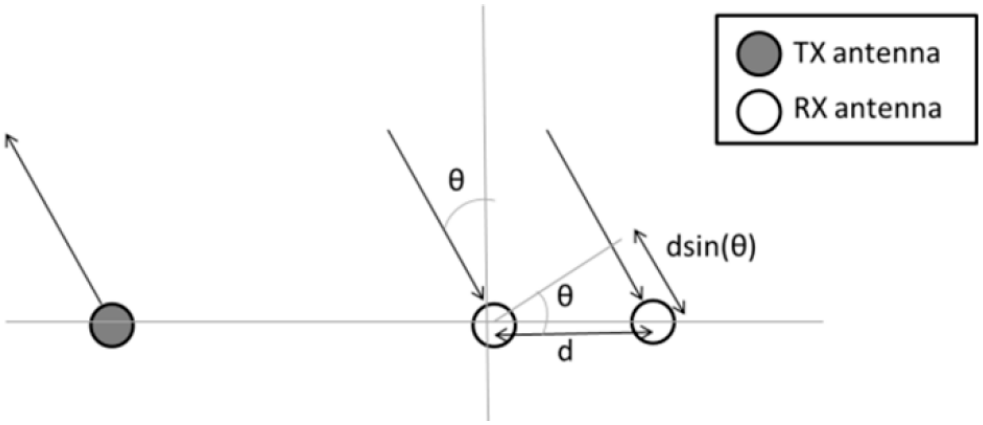


Figure 2.6 depicts the geometry of a radar antenna array with two RX antennas separated by distance d. The reflected signal approaches at an angle of θ with respect to the radar, and is collected by both RX antennas. The signal received at the secondary RX antenna travels an additional distance of $d\sin\theta$ which corresponds to a phase difference of $\omega = (2\pi/\lambda)d\sin\theta$ between the two RX antennas. Therefore, the DoA, θ, can be computed as:

$$\theta = \sin^{-1}\left(\frac{\omega\lambda}{2\pi d}\right) \tag{2.13}$$

Since ω, can be estimated only within the range $(-\pi,\ \pi)$, the equation 2.13 can be re-written with the unambiguous FoV of the radar as:

$$\theta_{FoV} = \pm\sin^{-1}\left(\frac{\lambda}{2d}\right) \tag{2.14}$$

When the inter-antenna distance, $d_{RX} = \lambda/2$, the maximum FoV is achieved, $\theta_{FoV} = \pm 90°$. In general, a MIMO radar system has more than two RX antennas. Angular resolution, $\Delta\theta$, of a radar system can be written as:

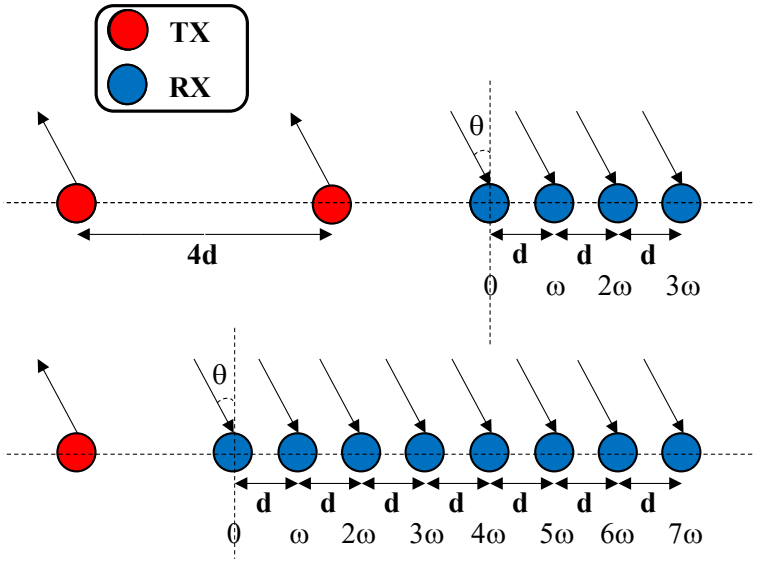
$$\Delta\theta = \frac{\lambda}{N_{RX} d_{RX}\cos\theta} \tag{2.15}$$

where N_{RX} is the number of RX antennas. Notice that $\Delta\theta$ depends on target's DoA, θ, and when the target is located in the bore-sight view (θ=0) and d_{RX}=$\lambda/2$, angular resolution becomes $\Delta\theta = \frac{2}{N_{RX}}$.

2.3.1 Principle of MIMO Radar

Angle resolution of a radar system can be doubled by doubling the number of RX antennas (i.e., N_{RX}). The same enhancement can be achieved by utilizing the MIMO concept where one more TX antenna is added to the array geometry. The radar system depicted in Figure 2.7 has two TX and four RX antennas. The signal emitted from the right TX antenna will have phases of [0, ω, 2ω, 3ω] at the receiver Uniform Linear Array (ULA). Since the left TX antenna is placed at a distance of 4d from the right TX antenna, any signal emitted from

Figure 2.7: Virtual array formation with MIMO concept.

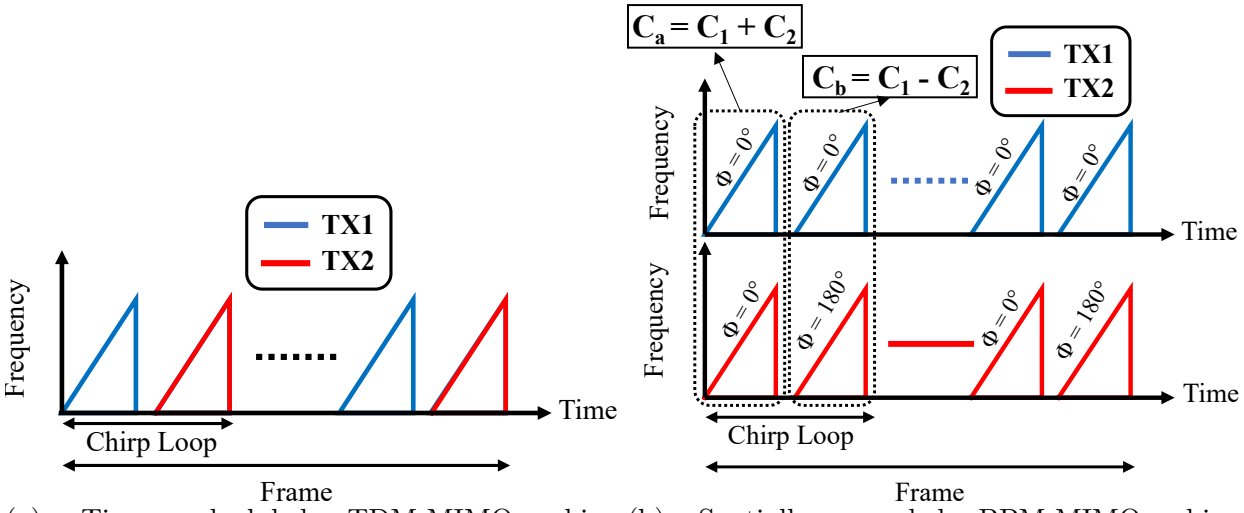


it will travel additional distance of $4d\sin\theta$. Consequently, they will have additional phase shift of 4ω which can be listed as [4ω, 5ω, 6ω, 7ω]. The phase sequence of [0, ω, 2ω, 3ω,k 4ω, 5ω, 6ω, 7ω] can be obtained by concatenating the phase sequences of four receivers for the signals emitted from two TX antennas. Such phase sequence is equal to the one can obtained from one TX and eight RX antennas. This process is called virtual array formation or synthesis, and it can be generalized to arbitrary number of $N_{TX} \times N_{RX}$ virtual arrays elements. Therefore, utilization of MIMO radar principle has the advantage of multiplicative increase in the number of virtual antenna channels, which results in finer angle resolution. The MIMO concept can also be extended to multi-dimensional array geometries.

Multiplexing Strategies for MIMO Radar

In MIMO arrays, RX antennas must be able to separate the signals emitted from each TX antenna. There exists various techniques to achieve this separation, and in this section two of these methods are discussed: Time Division Multiplexing (TDM) and Binary Phase Modulation (BPM).

Figure 2.8: MIMO radar multiplexing methods.



(a) Time scheduled TDM-MIMO chirp sequencing.

(b) Spatially encoded BPM-MIMO chirp sequencing.

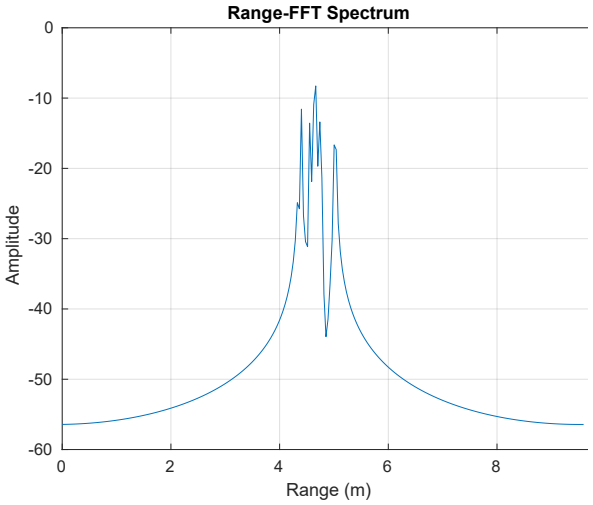
Time Division Multiplexing

In TDM-MIMO, TX antennas emit signals in a scheduled manner. Depending on the chirp sequence in the frame configuration, they can transmit chirps consecutively. Figure 2.8a depicts the TDM-MIMO chirp sequencing for the case of N_{TX}=2. In the collected data, each channel will correspond to one virtual antenna (i.e., TX-RX pair).

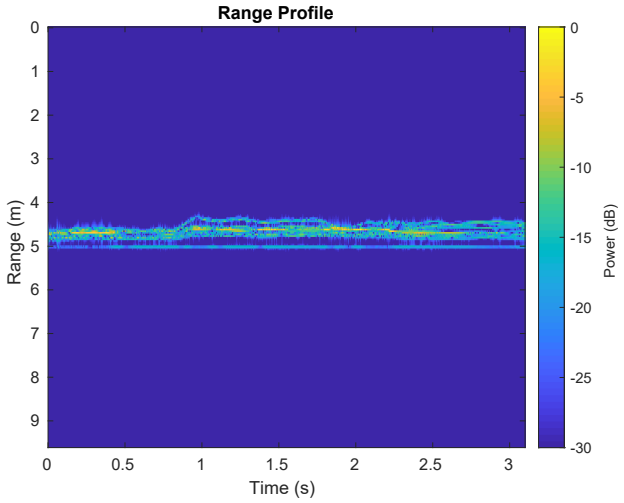
Binary Phase Modulation

TDM-MIMO is a simple method to implement, however it does not utilize full transmission capability of the system since only one transmitter is activated at a time. Different from TDM-MIMO technique, BPM-MIMO activates and transmits signal from both TX antennas simultaneously. While configuring the chirps, the first chirp, C_a, is configured to use both TX antennas with zero phase (Φ=0°), while the second chirp, C_b, is configured to use again both antennas but with phases of 0° and 180°, which is equivalent to multiplying each chirp with +1 and –1. One iteration of transmission of C_a and C_b is called a chirp loop. This process allows received data to be subsequently decoded by each virtual channel. Allowing simultaneous transmission from all TX antennas increases the total transmission power per

Figure 2.9: Range information extraction from raw data.



(a) Range-FFT spectrum.



(b) A range profile sample for arm gesture.

time slot which results in an Signal-to-Noise Ratio (SNR) increase of 10log10(N_{TX}). Figure 2.8b visualizes the chirp configuration for BPM-MIMO multiplexing.

One downside of employing MIMO techniques is the reduction in PRF which results in lower maximum unambiguous velocity.

2.4 Radar Signal Processing

Although radar signal processing is a wide topic to cover, this section primarily discusses the most common ways of processing raw data of FMCW MIMO radars. RF data are often in the form of time-series of complex I/Q samples. After a reshaping operation, the data can be converted to a 3D array with the shape of ($N_{samples}, N_c, N_{TX} \times N_{RX}$). ADC samples in the first and the second dimensions are also referred as fast-time and slow-time samples, respectively.

Fast Fourier Transform (FFT) is one of the most commonly employed radar signal processing technique to extract useful information from the raw data. FFT is more efficient version of Discrete Fourier Transform (DFT). It transforms a data from its original domain

to a representation in the frequency domain. In essence, it decomposes a signal into the frequency components that build it up.

2.4.1 Range Extraction

Applying FFT along the fast-time samples, range spectrum can be obtained. Peak locations along the spectrum indicate the radial distance of the targets, and the magnitude indicates the received signal strength. Figure 2.9a shows an FFT along the fast-time samples for range extraction for human arm gesture movement. Repeating this for each chirp results in a time series of range information in a heatmap matrix called Range Profile (RP). Figure 2.9b shows a sample of RP for human arm gesture for 3 sec.

2.4.2 Micro-Doppler Spectrogram

In human activity observation, the occurrence of kinematic motions in the radar FoV are reflected in the received signal's frequency components. Micro motions generated by hand, fingers and limbs result in μD [27] modulations centered around the main Doppler shift caused by the torso motion. The μD spectrogram, S, (also called μD signature) is a time-frequency analysis technique that can be used to observe these patterns and can

Figure 2.10: μD spectrogram sample for different activities and hand gestures.

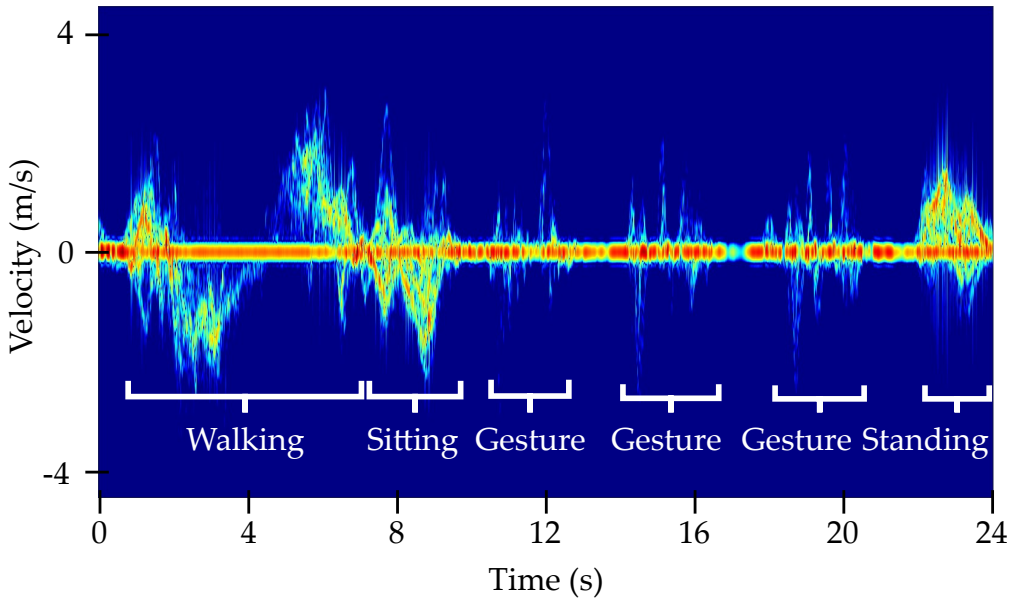
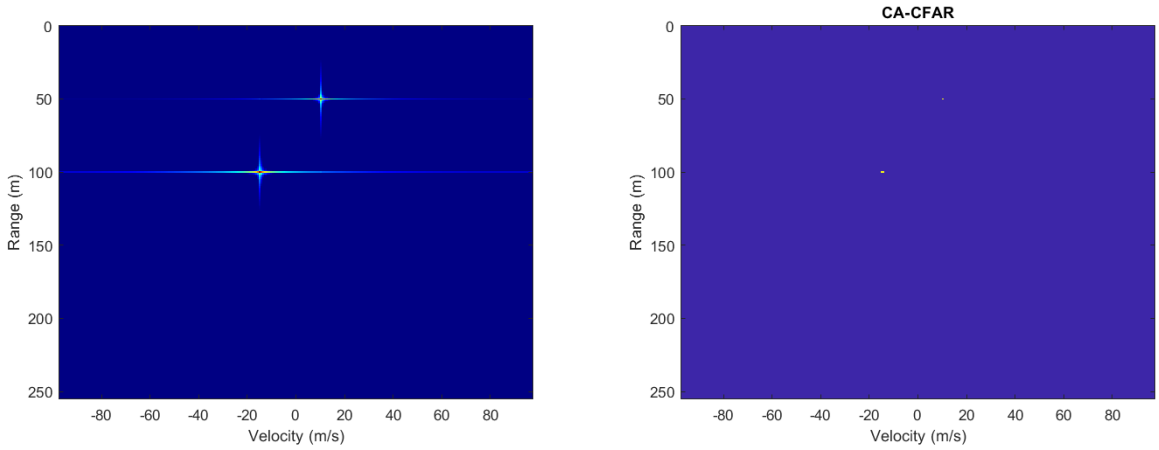


Figure 2.11: Range-Doppler processing and target detection of two targets.



(a) Range-Doppler map.

(b) CA-CFAR target detection result.

be computed as the square modulus of the Short-Time Fourier Transform (STFT) of the discretized input signal:

$$S(t, w) = \left| \int_{-\infty}^{\infty} h(t-u)x(u)e^{-jwt}\, du \right|^2 \tag{2.16}$$

where $h(\cdot)$ is the windowing function, $x(\cdot)$ is the received signal. Reflected signals from the stationary objects such as the walls or furniture, will locate at 0 Hz or 0 m/s in the μD spectrogram. Figure 2.10 shows a sample μD spectrogram for different human activities.

2.4.3 Range-Doppler Processing

Both range and velocity information can be jointly obtained via range-Doppler processing. While a range-FFT along fast-time resolves objects in range, a Doppler-FFT along the slow-time resolves each row (i.e., range bin) in velocity. Doppler FFT should be applied separately for each CPI in a non-overlapping windowing fashion, resulting in time-series of range-Doppler heatmaps with the same frame rate as the radar CPI. Peaks appear at targets' range and velocity bins. The resulting heatmap matrix is called range-Doppler map. Figure 2.11a shows a sample RDM for two targets located at 50 and 100 m away from the radar with radial velocities of 10 and -15 m/s, respectively.

2.4.4 Target Detection

A target detection method is needed in order to determine the number of targets, and range, velocity and angle information of the targets. In essence, most detection methods work by comparing the signal with threshold. The threshold is, in general, a function of both the probability of detection and the probability of false alarm. In this section, we study one of the most widely used adaptive thresholding method: Cell Averaging Constant False Alarm Rate (CA-CFAR). CA-CFAR extracts noise samples from both leading and lagging RDM bins (i.e., training cells) around the cell under test (CUT). The noise power estimate, P_n, can be computed as [142]:

$$P_n = \frac{1}{N_T} \sum_{m=1}^{N_T} x_m \tag{2.17}$$

where N_T is the number of training cells and x_m is the sample in each training cell. Guard cells are chosen adjacent to the CUT, both leading and lagging it. They serve for avoiding signal components from leaking into the training cells, which could adversely affect the noise estimate. The threshold factor, a, can be written as:

$$a = N_T(P_{fa}^{-1/N_T} - 1) \tag{2.18}$$

where P_{fa} is the desired false alarm rate. When the power of CUT exceeds a, detection occurs. Figure 2.11b shows the CA-CFAR detection result for the RDM given in Figure 2.11a.

2.4.5 Angle Estimation

In order to locate a target in 2D or 3D space, estimation of azimuth and elevation angles are needed. After detecting targets in range-Doppler domain for each channel (i.e., TX-RX pair), the measured phase difference across channels can be used to estimate the angle of arrival of the object. An FFT across channels of detected CA-CFAR peaks resolves the objects in angle domain even if they are located in the same range-Doppler bin. Other angle

estimation methods like digital beamforming, MUSIC [73] and ESPRIT [144] can also be applied for enhanced angular resolution.

2.5 Conclusion

Recent advances in FMCW MIMO radar technology enables various ways of processing data and achieving better target detection and tracking results. Software-defined adaptability of radar parameters also paves the way for environment-aware waveform selection. Such functionality leads to more optimal use of the hardware based on the application.

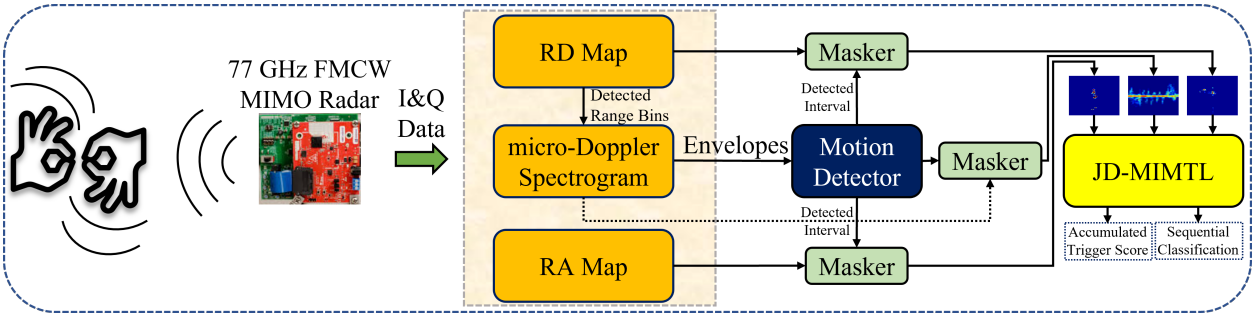
CHAPTER 3

SIGN LANGUAGE RECOGNITION IN A DAILY LIVING

3.1 Introduction

There has been much research on the use of RF sensing for hand gesture recognition [61, 177], especially since the development of low-cost, low-power, high resolution, integrated millimeter wave RF transceivers [7]. However, most current research involves controlled data acquisition with the participant located in a fixed position relative to the radar, articulating only a single gesture or sign. A critical challenge that has not been adequately addressed in the literature, however, is the challenge of ASL recognition in the context of daily living. To the best of our knowledge, this study [106] represents the first to consider triggering and command recognition of RF-sensor enabled devices under more realistic conditions, where the RF data is acquired in a continuous fashion to capture mixed sequences of gross body motion/activity intertwined with ASL signing. In particular, we analyze the design considerations for selection of a trigger sign based on kinematics, replicability, and recognition accuracy. Whereas current approaches rely on just one RF data representation, we propose a JD-MIMTL framework coupled with a motion detector to isolate the intervals over which the user is engaged in meaningful movement, and thus prevent unnecessary expenditure of computation resources when the RF system is not being used. Figure 3.1 shows a flowchart providing an overview of the proposed approach. Our results show that the proposed approach exceeds that offered by approaches common in the literature and can

Figure 3.1: Flowchart for the proposed approach.



recognize a sequence of 3 activities and 15 ASL signs with 92% accuracy, while detecting trigger signs with rates as high as 98.9%.

3.2 Sequential Human Activity and Sign Language Dataset

3.2.1 RF Sensor

In this study, a TI AWR1642BOOST 77 GHz RF transceiver paired with a DCA1000EVM data capture card were used to record data directly to a laptop. The TI 77 GHz transceiver is a FMCW short-range automotive radar that has two TX and four RX antennas, which offer additional sensing capabilities in comparison to other commercially available RF sensors that may have only 1 TX/RX channel. The antenna for the sensor has a roughly $\pm 70^{\circ}$ azimuth and $\pm 15^{\circ}$ elevation beamwidths. The sensor was positioned on a small table at a distance of about 1 meter from the ground.

3.2.2 Participants

Although ASL has been used as example motions in some gesture recognition studies [128, 110], sign language greatly differs from gesturing in that it possesses a much greater degree of physical complexity and Shannon information [123, 13, 125]. Like other complex system-generated signals, raw physical signal from signing data contains information at multiple timescales, spanning phonological, semantic, syntactic, and prosodic cues [11, 178].

Table 3.1: Listing of ASL Signs Acquired

Water	Good	Want	Thank You	Paper	Lawyer	I Love you	He	Book	Yes	Coffee
Teach	You	Time	Mountain	Bed	Knife	Nothing	Father	Car	Bring	Always
Eat	Drink	Tired	Don't Like	Work	Engineer	Oh, I see	See	Tie up	City	Evening
Deaf	My	Breath	Doesn't matter	School	Me	Why	Better	Shoes	Read	Write
Hold	Soon	Teacher	Technology	Sleep	Where	Go Ahead	Hot	Pet	Ready	Like
Shop	Maybe	Help	Earthquake	Mother	Your	Walk	Can	Month	License	Please
Morning	Gas	Hello	Tomorrow	Again	Cook	OK	Should	Go	Again	This
Have	Excited	Week	Let me see	Fine	Friend	Summon	Home	Three	More	Push
Kitchen	What	Wrong	Breakfast	Money	Come	Health	Today	Night	Must	One

Table 3.2: Description of Mixed Activity/Sign Sequences

Seq. #	**Motion Sequence**
1	Walking, sitting, TIRED, BOOK, SLEEP, standing up
2	Walking, sitting, EVENING, READY, HOT, standing up
3	Walking, sitting, MONTH, COOK, AGAIN, standing up
4	Walking, sitting, SUMMON, MAYBE, NIGHT, standing up
5	Walking, sitting, SOMETHING, TEACHER, TEACH, standing up

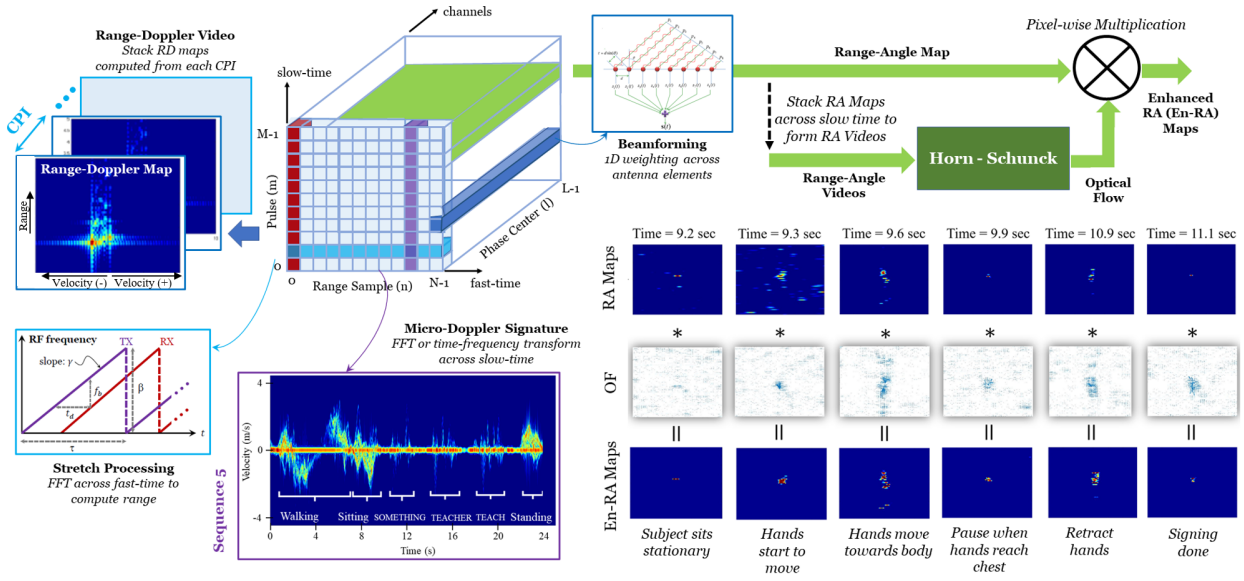
While some studies [49, 116] have utilized imitation signers - i.e., hearing participants who mimic signs observed in video - it has been shown [10] that it takes at least three years before the signing of ASL learners is perceived as fluent by native ASL users. Imitation signers exhibit greater kinematic variations, erratic cadence and signing errors, especially in replicating repetitive signs. Indeed, in our previous works [66, 69], we have found that imitation signing is distinguishable from native signing using classification of RF μD signatures.

Thus, in this study, RF data from both imitation signers and native ASL users were acquired and used for comparative study in trigger sign selection. A total of 110 single ASL signs were recorded from participants sitting 1 meter away from the radar. A total of 19 participants contributed to the database, including 4 native ASL users, who were either Deaf or Child of Deaf Adult (CODA), and 6 hearing individuals. Continuous recordings of mixed activity/signing sequences were recorded from 13 hearing participants, while testing on native users was conducted with 2 CODAs and 2 ASL learners, who were not used in acquisition of training samples.

3.2.3 RF Datasets

A total of two different datasets were acquired:

Figure 3.2: Signal processing diagram for computation of various RF data representations.



1. **Single ASL Signs:** 110 of the more frequently used ASL signs were selected from the ASL-LEX Database [23], including nouns, verbs, and adjectives. A complete listing of the signs acquired is given in Table 3.1. Each participant was asked to repeat the signs 5 times, resulting in 20 native and 30 imitation samples per sign.

2. **Mixed Motion Sequences:** Of these 110 signs, based on kinematics and replicability, a subset of 15 ASL signs are selected. Five different sequences of three ASL signs mixed with three different gross motor activities (walking, sitting, and standing up) were acquired, as shown in Table 3.2. For example, in SEQUENCE 1, the participant first walks for a few seconds, then sits on a chair located in front of the radar and enacts 3 different signs (TIRED, BOOK, SLEEP), and finally stands up. The participants were instructed to perform these activities consecutively in the line-of-sight of the radar. A total of 200 hearing participant samples and 94 native participant samples for each sequence were acquired, and made available for download [1].

[1] https://github.com/ci4r/ASL-Sequential-Dataset

3.2.4 RF Data Representations

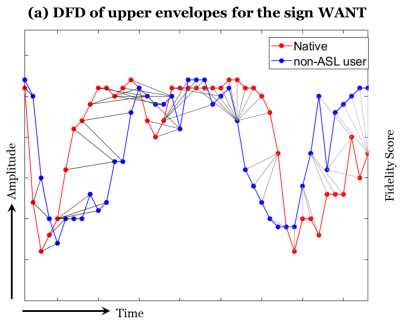
From the radar data cube, several different ways of representing the information acquired by the radar may be formed. In this work, RDMs, μD spectrograms and RAMs are used as RF data representations.

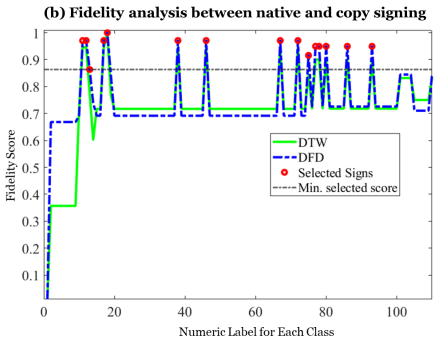
The visibility of target-related motion in the RAMs may be enhanced using optical flow, which indicates the spatial change in the location of pixels from one frame to another in a video. In this work, we compute the optical flow using the Horn-Schunck method [79] and take its element-wise multiplication with the pixels in the RA maps to accentuate motion-related returns. This process puts more weight on pixels where there is a moving target, and suppresses pixels comprised of clutter or minimal motion. Because the MUSIC algorithm is relatively prone to noise, this approach can enable significant visual enhancements in the RA maps. An overview of the radar signal processing steps utilized to compute the stated RF data representations are summarized in Figure 3.2.

3.3 Trigger Sign Fidelity Analysis and Selection

There are many different considerations for the design of a device trigger sign (also known as a wake word). Trigger signs should be distinct, not easily confused with signs frequently used in daily discourse, easy to articulate and culturally appropriate. In Deaf culture, for example, while it is common for finger-spelling to be used to state the names of a hearing individuals, personal *name signs* can only be used if the name sign has been given by a member of the Deaf community. Moreover, ASL does have some differences in dialects used in different geographical regions within the U.S., such as Black ASL, which represents a unique ethnic sub-culture in the South [76]. The cultural context of signs may differ and take on different meanings in different regions. Therefore, the design of culturally-appropriate trigger signs can only be accomplished through partnership with Deaf community organizations, who can provide cultural perspectives and facilitate studies soliciting Deaf community feedback on the design.

Figure 3.3: Selection of replicable ASL signs using DFD and DTW.





(c) Selected replicable ASL signs

Words	Two-Handed	# of Strokes	Major Location	Movement
TIRED	✓	3	Body	Curved
BOOK	✓	3	Neutral	Curved
SLEEP	✓	4	Head	Straight
EVENING	✓	5+	Hand	Straight
READY	✓	4	Neutral	Straight
HOT	X	3	Head	Curved
MONTH	✓	3	Hand	Straight
COOK	✓	5+	Hand	Other
AGAIN	✓	4	Hand	Curved
SUMMON	✓	3	Hand	Back-and-Forth
MAYBE	✓	5+	Neutral	Straight
NIGHT	✓	3	Hand	Straight
SOMETHING	X	4	Neutral	Circular
TEACHER	✓	4	Head	Straight
TEACH	✓	3	Head	Straight

Thus, this study focuses on technical aspects of trigger sign design as a precursor to a subsequent Deaf-centric design study. First, as RF sensors are sensitive to distance and motion, signs that are dynamic, with strong radial velocity components (i.e., include primary arm motion, as well as secondary motion of the hand, such as hand shape or orientation change), or which traverse greater distance and have a longer flight times are better suited as trigger signs for automatic detection. This is in contrast with signs primarily characterized by secondary hand motion, such as finger-spelled words.

Second, the replicability of the trigger sign is important to enable consistent and robust recognition. Although native ASL users are the target population for ASL-sensitive user interfaces, there is a wider community of ASL learners and non-native ASL users, such as interpreters, who could also be using the interface. However, there can be noticeable differences in the articulation of signs based on fluency. Thus, the replicability of the 110 signs listed in Table 3.1 were evaluated using a comparison of the imitation signing and native ASL μD signatures. This was done by first computing the upper and lower envelopes of each sign based on the percentiles of the cumulative amplitude distribution [44, 95]. Next, both the Discrete Fréchet Distance (DFD) [45] and DTW were used to compare the replicability of signs based on fluency.

DTW is a method for measuring the similarity between two time-series and finds the optimal match [1] between sequences that satisfy all restrictions and rules with the minimum

cost. The DFD computes the similarity between two curves by taking into account both ordering of the points and the location along the curves. It is defined as the shortest cord-length required to join a point traveling forward along one curve and one traveling forward along the other curve, and the rate of travel for either point may not necessarily be uniform. As the similarity of two curves increases, DFD gets closer to zero. As an example, consider the comparison of the upper envelopes of the μD signatures for imitation signing and native signing for the sign WANT, shown in Figure 3.3(a), where the grey lines represent the cord-length.

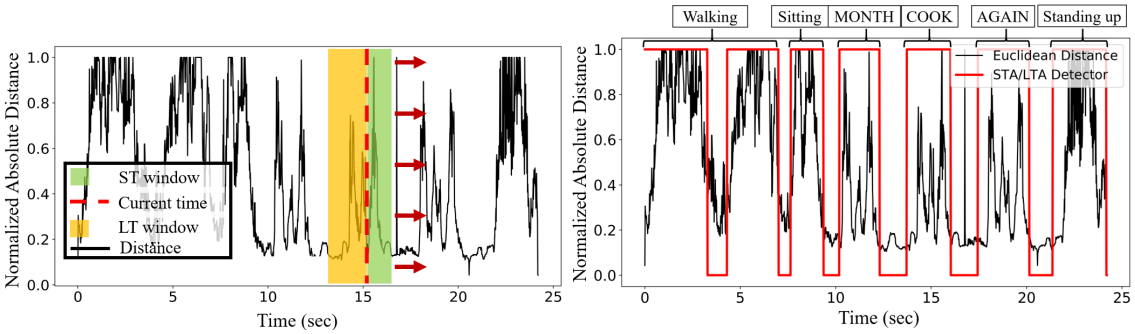
To identify the most easily replicable signs (independent of fluency), the envelopes of the native ASL signatures and those from hearing imitation signers are compared on a sign-by-sign basis. The DTW and DFD metrics are averaged and re-scaled between 0 and 1. Once the distance metrics, dtw and dfd, are normalized, the fidelity scores, s_{dtw} and s_{dfd}, for each class (sign) are found by taking the inverse of the normalized distance (i.e., $s_{dtw} = 1/dtw$, $s_{dfd} = 1/dfd$). The results are shown in Figure 3.3(b). It may be observed that both the DTW and DFD are consistent in their assessment of which signs are consistently articulated across deaf, CODA, and hearing users.

The top 15 signs that have the shortest distance (i.e. highest similarity) between native ASL and imitation signing users were selected as trigger sign candidates, which will next be evaluated based on detection rate and sequential recognition accuracy. The selected signs are listed in Figure 3.3(c) along with their kinematic properties, as given by ASL-LEX.

3.4 Motion Detection and Segmentation

Continuous activities and ASL signing create a time series of sequential activities, for which segmentation is an important initial step in the analysis of sequential data. Utilization of a motion detector can facilitate segmentation, which helps define the length of the input samples to be fed to a learning model. It can also improve the power and computational efficiency of the system by making a prediction only when an activity or sign is detected as

Figure 3.4: Illustration of the operation of STA/LTA based motion detector on SEQUENCE 3.



(a) Operation of detector on absolute distance vectors.

(b) Intervals with motion detected by STA/LTA detector.

opposed to every time step. While motion detection can be done with a human-in-the-loop approach, this is not desirable in automate, stand-alone systems. Instead, a power-based automated segmentation algorithm, such as STA/LTA [170, 164], Dynamic Boundary Detection (DBD) [163] or Power Burst Curve (PBC) [184] may be utilized.

The PBC can be used for motion detection using thresholding. The start and end of the motion is determined by when the input power exceeds or falls below this threshold, respectively. An important drawback of this method, however, is that it is prone to a high rate of false triggering, especially in the presence of noise, because the threshold is not adaptive and unaware of past and future power levels.

STA/LTA-based techniques solve this problem by defining two consecutive windows; namely, short-time and long-time windows. Their relative average power is used to define an adaptive threshold value. The STA/LTA method proposed in [164] has proven to be very successful in detecting the tail (end point) of hand gestures. However, the method uses fixed length detection windows, whose duration is selected based on the duration of the longest gesture in the dataset. This approach is not well suited to sign language, since ASL signs possess great variability in duration. Basing window size on the longest duration sign can result in a long blank period at the beginning of the detected region for short signs, thereby introducing non-informative or redundant input to the feature space.

DBD, on the other hand, requires application of high-pass filtering to the Doppler information, resulting in elimination of the low and zero frequency components of the spectrograms. Prior work [66] has shown, however, that filtering at 77 GHz results in significant loss of low-frequency information in the signal, together with removal of the clutter, thereby degrading classification accuracy.

Thus, this work proposes a variable window STA/LTA-based motion detection algorithm to identify both the starting and ending point of a motion. First, the absolute difference between the upper and lower envelopes at a time index is computed to create absolute distance vectors. An exemplary, normalized absolute distance vector is shown in Figure 3.4a. The absolute distance for each data recording, i, can be computed as $v_i = |u_i - l_i|$, where v_i is the absolute distance vector, u_i and l_i are the upper and lower envelopes, respectively.

Then, $STA(t)$ and $LTA(t)$ can be defined as the leading and lagging windows at time t as:

$$STA(t) = \frac{1}{T_1} \sum_{k=t+1}^{t+T_1} v_i(k),\ LTA(t) = \frac{1}{T_2} \sum_{k=t-T_2+1}^{t} v_i(k) \tag{3.1}$$

where T_1 and T_2 are the lengths of short and long windows respectively. The starting point of a motion is detected when the following conditions are satisfied:

$$STA(t) > \sigma_1 \quad \text{and} \quad \frac{STA(t)}{LTA(t)} > \sigma_2 \tag{3.2}$$

where σ_1 and σ_2 are predefined detection thresholds. Similarly, the ending point is detected if

$$STA(t) < \sigma_3 \quad \text{and} \quad \frac{STA(t)}{LTA(t)} < \sigma_2 \tag{3.3}$$

where σ_3 is the detection threshold for the stopping point.

Note that in order to locate the starting point, according to (3.2), $STA(t)$ needs to exceed the threshold σ_1, implying that the the motion has to appear in the short window. Also, the ratio of average power in the short and the long window should be higher than σ_2. In

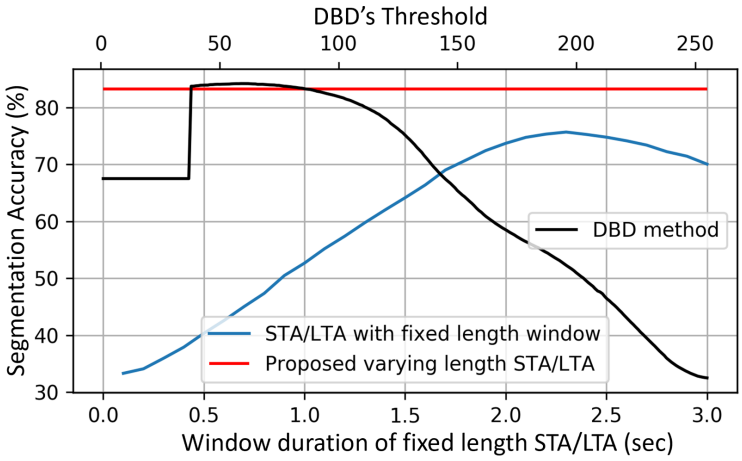
this way, if there is noise, the system will not be triggered unless the ratio exceeds the σ_2. Similar conditions apply to ensure correct detection of the endpoint; i.e., the case when the motion disappears from the proceeding window and the ratio drops below the threshold σ_2. The resulting detection mask found with the proposed *vw*-STA/LTA approach is able to separate the intervals with and without motion, as shown in Figure 3.4b.

While DBD requires the optimal selection of a threshold based on the returned signal strength, fixed length STA/LTA bases selection on the window length. In contrast, the proposed variable length STA/LTA approach adaptively changes its detection window interval irrespective of the returned signal strength. A comparison of the segmentation accuracy for these three methods is presented in Figure 3.5. Segmentation accuracy is computed by comparing segmentation mask with the ground truth generated by a human analyst for each time step. Note that the segmentation accuracy of DBD and fixed-window STA/LTA exhibit great variance in efficacy for different thresholds or window lengths. Fixed-window STA/LTA achieves a peak accuracy of 75.7% when the window length is 2.3 seconds. DBD performs better by comparison, achieving a peak accuracy of 84.2% when the threshold is set to 61, but with the cost of information loss in low frequency components. This peak value is only slightly higher than the 83.5% accuracy achieved by the proposed motion detector, while the propose approach can maintain this accuracy irrespective of any parameter values due to the use of variable, adaptive window lengths.

3.5 Joint-Domain Multi-Input Multi-Task Learning

Conventional approaches to RF signal classification rely on a single data representation, presented as either 2D or 3D inputs. In contrast, to take advantage of all available physics-based information (range, velocity, frequency and angle), we propose a JD-MIMTL-based DNN architecture, where each input representation is processed in parallel and the final feature space is constructed by fusing individual feature spaces. Auxiliary tasks are used to regularize

Figure 3.5: Comparison of the segmentation accuracy of DBD, fixed-window STA/LTA and the proposed variable-window STA/LTA.



and better guide the training loss. The accuracy of the proposed approach surpasses that of conventional single-input models by over 13%.

3.5.1 Mixed Motion Sequential Recognition

Sequential classification of daily activities and ASL signs differs from conventional hand gesture recognition tasks because it is not comprised of just an isolated, short duration, single type of motion. Instead, it consists of a time series of consecutive motions, which might belong to different classes of gross daily activities or ASL signs. A typical approach to classify a continuous time series data includes: 1) temporal segmentation, 2) making prediction for each time step. The former is achieved using a motion detector described in Section 3.4, while the latter will be discussed in this section. In real-world scenarios, training a model with the entire stream of data sequences (24 sec each) is not feasible, because this significantly increases the computation time, rendering outputs only after a long delay, which is undesirable in interactive systems. However, when models are trained with shorter input sequences, performance also tends to drop gradually, because performance of LSTMs are dependent on input sequence lengths [90]. Since LSTM networks have the flexibility to be trained with varying sequence lengths, the data segments isolated by the motion detector

were used as input sequences. These segments will have varying lengths depending on the user's pace and the motion itself.

3.5.2 Training a Spatio-Temporal Model

In this section, the effect of input sequence length on prediction accuracy is examined. For this purpose, we use a DNN consisting of 3 time-distributed (TD) 2-D convolutional blocks with kernel sizes of 3, followed by max pooling layers and a bidirectional long-short-term-memory (BiLSTM) layer. A TD *softmax* layer is employed for temporal classification. While convolutional layers extract the spatial features, the TD wrapper enables application of the same nested layer to each time step. BiLSTM is a kind of recurrent neural network which is used to extract temporal relationships between time steps. They have proven to be very successful in terms of learning long term dependencies in various tasks such as natural language processing [183], and speech recognition [58]. By employing LSTMs in our final encoded feature space, both spatial and temporal features are extracted for classification.

In μD spectrogram (μDS) classification, spectrograms are divided into 0.2 sec *non-overlapping* windows to be used as time steps. In RD and RA map classification, the interval between each RD/RA map or frame is 40 milliseconds, so to obtain a data structure corresponding to the same (0.2s) duration, five RD/RA frames were stacked (5×40ms = 0.2s). For both inputs, 80% of the data is used for training and 20% for testing, with an equal number of samples from each sequence. Adam optimizer and categorical cross entropy is used along with early stopping with patience of 10 epochs to train the model. Hence, the input data has the shape of `(batch size, number of windows, width, height, channels)`. A 2D-CNN+BiLSTM network for μDS and 3D-CNN+BiLSTM network for RD/RA maps are employed. The impact of the motion detector is discussed next.

Original Sequential Data

Table 3.3 shows the classification accuracy for each input data representation as a function of various input durations. It may be observed that the accuracy of the models for all

Table 3.3: Sequential Classification with CNN+BiLSTM

Data	Length of Sequences	μD Spectrogram	RD Map	RA Map
Original Sequences	1/24 (1 sec)	69.2%	72.5%	69.9%
	1/12 (2 sec)	78.6%	76.3%	73.7%
	1/6 (4 sec)	81.3%	82.4%	79%
	1/3 (8 sec)	84.3%	89.9%	85.9%
	Half (12 sec)	84.6%	90%	87%
	Full (24 sec)	86.1%	92.4%	89.7%
MDI	Varying	78.8%	72.8%	67.5%

Table 3.4: Computation Times Spent for Prediction

Length of Sequences	μD Spectrograms	RD Map	RA Map
1 second	201.8 sec	207.5 sec	205.8 sec
2 seconds	111.7 sec	125.7 sec	123.1 sec
MDIs	61.4 sec	69.3 sec	67.8 sec

input domains decreases as the length of input sequences gets shorter. Best performances are obtained using longest sequences with RD maps providing a 92.4% accuracy. The performance using μDS changes around 17% while that using RD maps and RA maps change around 20% from 1 sec. sequences to 24 sec. sequences. While the longer sequences give better performance, they also result in greater prediction delay and higher memory requirement due to increased data size. This situation demonstrates the challenge of deciding an appropriate input length while doing sequential classification and the trade-off between prediction performance and delay.

Motion Detected Intervals (MDI)

The detector extracts data segments containing motion, eliminating periods of no movement. Thus, each MDI is of varying duration, and models are trained using variable length data. The testing accuracies obtained when using μDS, RD and RA maps are 78.8%, 72.8%, 67.5% respectively. These results are comparable to those obtained with fixed length sequences of 2 sec. for μD, and 1 sec. for RD/RA maps, while the length of detected segments vary between 0.6 and 10 sec. Moreover, using MDI rather than fixed length windows significantly

Table 3.5: Classification Accuracy of the Motion Detectors

Motion Detector	μD Spectrogram	RD Map	RA Map
DBD	72.4%	70.9%	63.8%
Fixed STA/LTA	76.8%	71.5%	67.1%
Varying STA/LTA	**78.8%**	**72.8%**	**67.5%**

reduces the computation time for prediction by masking out the intervals that do not contain any motion. Table 3.4 presents the total computation time of an NVIDIA Titan V GPU to make predictions for data durations of 1 sec and 2 sec. The total computation time is reduced by 45% on average for different input representations when compared with 2 sec length sequences. Note that the amount of computational savings obtained using the motion detector does depends on the data, in that as MDI increases so does the time savings. As daily life often involves extended stationary periods, in practical settings the use of MDI can result in significant savings.

3.5.3 Effect of Motion Detector on Classification Accuracy

The performance of DNN models rely heavily on the data presented at the input, which in turn is extracted based upon the starting and ending points of the MDIs as determined by the motion detector. Thus, the ability of a motion detector to accurately extract intervals containing movement impacts the efficacy of classifiers. Table 3.5 compares the classification accuracy attained from different input representations extracted using DBD, fixed-length STA/LTA and the proposed variable-length STA/LTA motion detectors. It may be observed that the proposed variable-length STA/LTA detector yields greater classification accuracy in comparison to other approaches, surpassing fixed-length STA/LTA by 0.4-2% and DBD by 1.3-6.4%. Note that the relatively worse accuracy of DBD is due to information loss incurred during the high-pass filtering, which removes low-frequency signal as well as clutter components, and hence degrades the resulting classification accuracy.

3.5.4 Proposed Approach: JD-MIMTL

To improve the classification accuracy obtained with just one input representation, this paper proposes utilizing fusion of multiple input representations in a multiple-task learning [22] framework with connectionist temporal classification (CTC) [59]. Although MTL has been implemented successfully in computer vision [53] and natural language processing [34], these applications all involve a single data representation (image, text, speech signal). In RF sensing, the various physical variables measurable by radar - namely, range, μD, and angle versus time - are reflected in different data representations, to base recognition decisions on all physical properties, multiple inputs to MTL are advantageous. The joint feature space derived from multiple input representations is enriched by fusing in a concatenation layer.

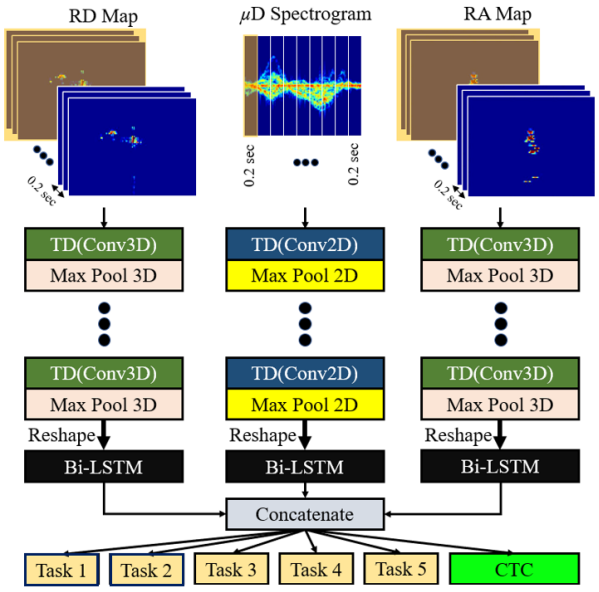
MTL jointly optimizes multiple objectives by exploiting domain-specific information contained in commonalities and differences across tasks. By sharing representations among related (auxiliary) tasks, the generalization capability of the model can be improved on the main task. ASL classification can be aided by basing decisions on consistency with certain physical properties of signing, based on the categorization provided in Figure 3.3(c). Five auxiliary tasks are defined:

- Task 1: One versus two handedness;
- Task 2: Major location of hands;
- Task 3: Movement type;
- Task 4: Daily activity versus ASL sign; and
- Task 5: Number of strokes.

The overall loss function, L_{total}, utilized in the JD-MIMTL framework is the weighted sum of the CTC loss, λ_{ctc}, and the loss L_i specific to each task i:

$$L_{total} = \lambda_{ctc} L_{ctc} + \sum_{i}^{I} \lambda_i L_i \tag{3.4}$$

Figure 3.6: Proposed multi-input multi-task learning network.



where λ are the weights assigned to the various loss terms. Since each task has its own loss function, and, hence, varying convergence times, the weights λ needs to be jointly optimized. Three different loss optimization techniques [56] were compared, namely, the uniform combination of losses (i.e. equal weights across all tasks), the uncertainty based weighing method [33], and grid search. The first two methods minimize L_{total} without taking into account the importance of each individual task. Since we aim to minimize L_{ctc}, which is derived from the prediction layer, the grid search method was preferred. The use of smaller auxiliary task weight values during grid search was found to perform better than that obtained with using the uniform combination of losses or uncertainty-based weighting. Specifically, weight values of $\lambda_{ctc} = 1$ and $\lambda_i = 0.2$ were used. The overall proposed JD-MIMTL approach is depicted in Figure 3.6. After training the model, all of the auxiliary task and CTC output layers are removed and the model is augmented with a $softmax$ layer for classification.

The probability distribution of the classes, which is obtained as the output of the JD-MIMTL, can be decoded two ways in parallel for sequential classification and trigger word detection. Best path decoding is used as the decoding scheme of the CTC outputs for both objectives.

However, the final prediction class is defined as the statistical mode of the time steps of an MDI for sequential classification, and as the prediction scores for the trigger sign accumulated over the time steps of an MDI for trigger word detection.

3.6 Results and Discussion

3.6.1 Trigger Word Detection

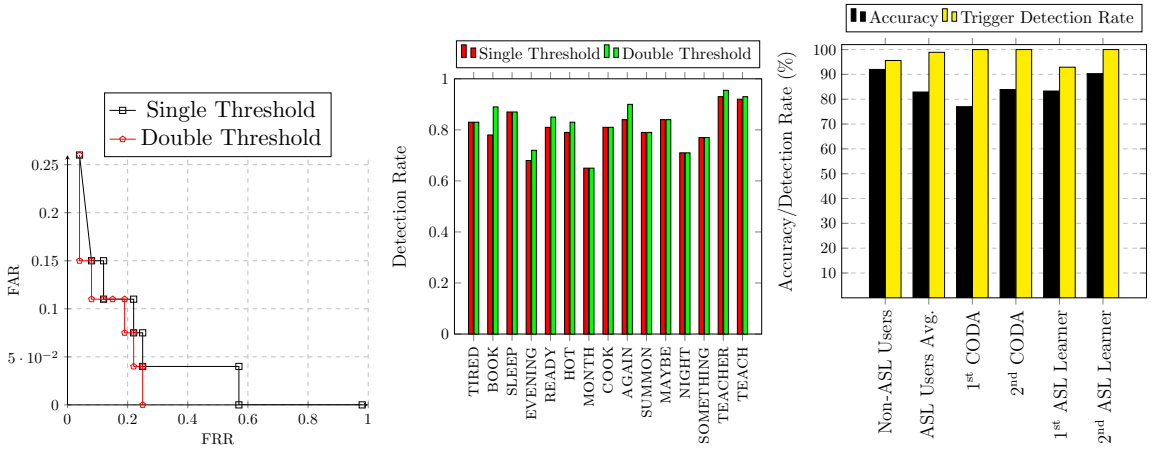
To activate a device, the trigger sign must be correctly recognized from within a stream of data, and the activation should occur when the articulation of the sign is completed. One approach is cumulative score aggregation (CSA) [166], where the scores (i.e., prediction probabilities) of the trigger sign are accumulated over time, and a detection is recorded when the accumulated score, s_a, exceeds a predefined threshold. The threshold can be adjusted to ensure the detection is triggered only when the trigger sign is complete.

In this work, an adaptive, double-threshold CSA approach is proposed for trigger sign detection. Since the MDIs have varying lengths, the value of the threshold, T, is adaptively determined based on the interval length as: $T = w * \gamma$, where w is the length of the MDI and γ is a predefined confidence factor. To mitigate the false rejection rate (FRR) of the detector, a second (lower) threshold, T_{low}, is also defined. When the accumulated score exceeds the T_{low}, but not T, the detector is alerted to the possibility of a trigger and begins recording the duration over which the score stays above T_{low}. The system is triggered if score exceeds T_{low} for more than $w/2$ seconds and the motion is classified as the trigger sign.

In trigger word detection, effect of using single versus double thresholding can been seen from Figure 3.7a, which shows the trade-off between the false alarm rate (FAR) and FRR for $\gamma \in \{0.01 : 0.99\}$ for the word AGAIN. When a single threshold is used, the FRR can climb as high 0.6, while double thresholding limits this value to just over 0.2. This is significant because decreasing the FRR boosts the detection rate, $D_r = 1 - FRR - FAR$, where FRR and FAR are defined as:

$$FRR = \frac{n_t - n_d}{n_t}, \qquad FAR = \frac{n_f}{n_t} \tag{3.5}$$

Figure 3.7: Trigger word detection results.



(a) FAR and FRR of the word AGAIN for single and double threshold.

(b) Detection rates of the words for single and double threshold.

(c) Performance of the proposed method on different ASL fluency groups.

where n_t, n_d and n_f are the number of total, detected and false detected samples respectively.

As shown in Figure 3.7b, when the resulting detection rates for single thresholding versus the proposed double thresholding approach are compared, it may be observed that for each considered trigger sign, the proposed approach yields a same or improved detection rate. The word TEACHER has the highest detection rate for both thresholding methods, achieving a detection rate of 0.93 and 0.96, while the word MONTH (self-occluded) has the lowest score of 0.65 for both cases. Signs with higher classification accuracy tend to have higher detection rates as well, such as TEACHER and TEACH.

The number of strokes (i.e., length) of the sign is an important consideration in trigger sign selection. For the purposes of automatic detection, strokes were defined as components surrounding the sign-initial and sign-final handshapes; thus, both the motion inherent to the sign (i.e., the *stroke* as defined in sign language phonology), and transitional motions preceding and following the sign, were included in the analysis. This approach approximated predictive processing in human sign language recognition [122, 51], while remaining consistent with ecological paradigm of wake sign use. Signs with few strokes defined in this manner (less than 3) were found to have many false alarms, while those with more than 4 were prone

Table 3.6: Comparison of DNNs for MDI Classification

Architecture	μD	RD Map	RA Map	Feature-Level Fusion
CNN + BiLSTM	78.8%	72.8%	67.5%	84.3 %
CNN + BiLSTM + CTC	80.6%	78.4%	71.3%	87.5%
CNN + BiLSTM + CTC + MTL	**83.6%**	**78.6%**	**71.4%**	**JD-MIMTL 92%**

to a high number of false rejections. This is similar to results in speech recognition, which report optimal wake word lengths of 3 to 4 syllables [167] - or, in quantitative terms, several entropy (high information-density) peaks within the continuous signal.

3.6.2 Sequential ASL Recognition

A testing accuracy of **92%** is achieved using the proposed JD-MIMTL approach, and surpasses the results achieved with various state-of-the-art sequential recognition approaches, as shown in Table 3.6. This result is also quite close to the 93.5% accuracy attained using JD-MIMTL when the motion detector is replaced with ground truth segmentation. Moreover, the baseline established in Section 3.5.2 using CNN+BiLSTM on single-input representation MDI data is improved to 84.3% by application of feature-level fusion. Consideration of CTC loss improves the results obtained for both single-input and fusion of multi-input representations.

The accuracy using μDS increased to 80.6%, RD maps to 78.4% and RA maps to 71.3%, thus providing an average improvement of 3.73%. For RD maps and RA maps, MTL only slight improves performance by just 0.1%-0.2%, while the accuracy with μDS increases by 3%. The proposed JD-MIMTL approach yields a performance improvement of 8.4% over μDS as a single-input to MTL, and 4.5% improvement over multi-input feature level fusion without using MTL.

The confusion matrix for the proposed architecture is provided in Figure 3.8. It can be seen JD-MIMTL exhibits the most confusion in signs with low radial motion (EVENING, MAYBE, NIGHT) and self-occlusion (MONTH). The signs with high radial motion (TEACHER,

TEACH) have the highest recognition rates. This is due to higher sensitivity of radars to radial velocity components.

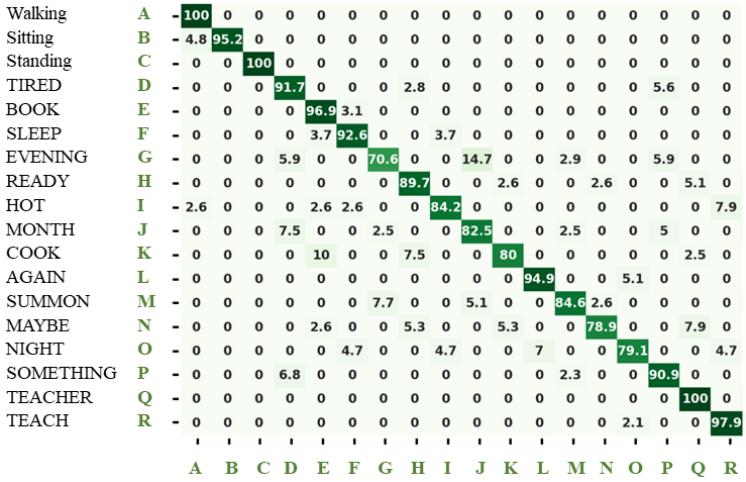
3.6.3 Performance Across Different Fluency Groups

The proposed approach is tested on different fluency groups to evaluate is efficacy across different users. This is done by training the model solely with data from non-ASL users, but testing on ASL users' data. Thus, not only are the participants between training and test sets different, but also their fluency levels. In Figure 3.7c, the overall testing accuracy for all signs, and the trigger detection rate for the selected trigger word, TEACHER, are presented for different fluency groups. While the first two columns report average results, the remaining 4 columns break down the results for specific participants, indicating whether the participant was an ASL learner or CODA. On average, the sequential ASL classification accuracy for ASL users was 10% less than that attained from non-ASL users. But, the trigger detection rates remained above 94% irrespective of fluency. In fact, 3 out of 4 ASL users' trigger word is detected with 100% accuracy.

3.6.4 Discussion

Because RF sensors rely on kinetic properties of signing during recognition, signs that inherently contain greater movement (especially inter-sign movements) are easier to recognize.

Figure 3.8: Confusion matrix of the proposed JD-MIMTL.



For example, the signs TEACHER and TEACH both involve raising the hands to the level of the head, whereas MONTH involves just a short swipe of a finger downward and NIGHT involves a more subtle downward, curved motion of the hand/arm, resulting in a detection rate that is over 20% lower. Effective ASL-based device triggering will require the design of a unique sign for this purpose, as commonly used daily expressions may mistakenly trigger a device. In this regard, it is important to note that it is not necessary for such a trigger sign to have meaning in English; e.g. that KNOCK might be sensible in meaning has little bearing on efficacy in terms of detectability, practical and cultural considerations. In future work, we aim to work with deaf community partners to jointly evaluate usability and efficacy of kinetically unique trigger signs.

Another important consideration for device operation with ASL is real-time implementation on dedicated edge computing platforms. Although there have been some studies of real-time gesture recognition using micro-Doppler signatures [164, 114, 28, 130], these works have considered only a small number of classes (less than 12), and focus on hardware acceleration or reduction of the computational complexity of the model itself. However, our initial work [3] in evaluating computational latency in the processing pipeline has shown that a significant part of the latency is not in the classification stage, but in the computation of the input representations themselves, especially micro-Doppler signatures. Latency depends not just on the duration (length) of the data, but also on short-time Fourier transform parameters, such as window length and overlap, which determine the dimensionality of the resulting spectrogram and impacts classification accuracy. Joint optimization of input representation generation and DNN model will be necessary to maximize real-time recognition performance.

3.7 Conclusion

The proposed techniques in this chapter enables trigger sign detection for device activation and sequential recognition of ASL in the context of daily living. While conventional approaches to RF signal classification utilize just one RF data representation, this work exploits μD

spectrograms, RD maps, and RA maps in a JD-MIMTL framework for sequential classification. By defining tasks in terms of physically-relevant concepts for ASL recognition, sequences involving a mixture of 18 different daily activities and ASL signs was classified with 92% accuracy. The proposed double-thresholding trigger detection method achieves detection rates of 96% and 98.9% for non-ASL and ASL users, respectively, for the sign TEACHER. Potential selections for trigger signs are evaluated based on sequential activity recognition accuracy and replicability across the fluency levels of users. The results demonstrate the potential for RF sensing to be used for ASL-sensitive HCI.

CHAPTER 4

MULTI-PERSON SEPARATION VIA ANGULAR PROJECTION

4.1 Introduction

Most of the current RF-based activity/gesture recognition literature is limited to consideration of just single target scenarios, even though the presence of multiple targets is typical in real-world environments. In multi-target scenarios, the RF μD signature computed from the raw I/Q data will result in the signature for each target super-imposed upon each other. Consequently, DNNs trained with data with just a single target present will not be able to correctly recognize the targets' activity. Many of the works involving multiple targets focus on counting the number of people present [30, 29, 182, 8], while just a few works actually aim at separating the micro-Doppler signatures. Vishwakarma, et al. [172] propose a sparse coding dictionary learning based algorithm to separate the μD returns from multiple targets. However, the approach relies on parametric models for simulating the human and fan returns considered as part of a binary classification problem. The Boulic model used to simulate human returns only approximates walking, thus precluding the approach from being effective when generalized human activities, which are not easily represented by a parametric model, are observed. Even in the limited case presented, the separated μD signatures suffer from losses in comparison to their measured counterparts.

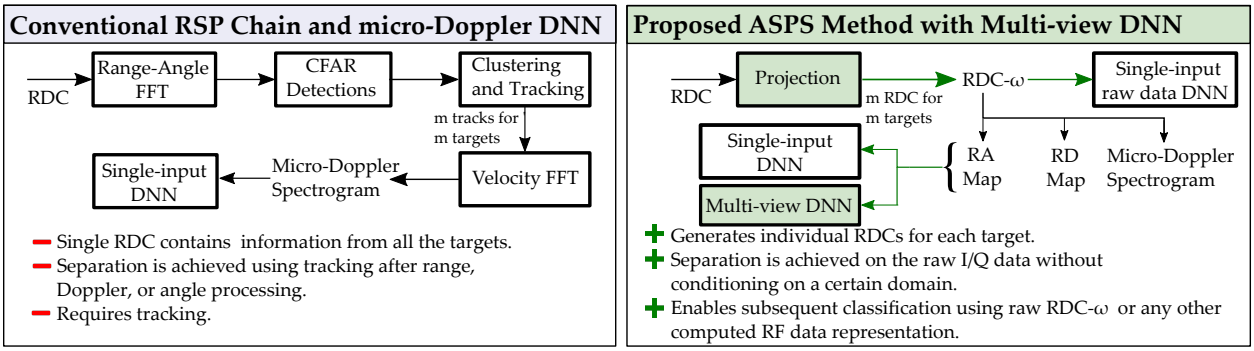
Alternatively, Huang, et al. [83] applies a multi-stage separation scheme in which range gating is first applied for preliminary separation, followed by design of a multi-task learning network for fine signature separation and recognition. However, because multiple targets can

be present at the same range, this approach is essentially relying on the DNN to figure out the signal separation solely based on a learned, data-driven model. This inherently limits the applicability of the proposed approach to only those target signatures for which the model has been trained, and will not generalize easily to previously unseen target μD signatures.

Another common approach is to explicitly track the components of the overall μD signature. Wang, et al. [175] formulates the problem as one of path planning and proposes ant colony optimization to identify the points corresponding to different paths within the overall μD signature. Simulation-based results are presented only for sinusoidal μD curves - in real measured signatures where the backscatter from each target is not comprised of a distinct curve, it is not clear how effective this approach would be. Moreover, the μD signatures of more complicated targets, such as humans, are comprised of multiple trajectories corresponding to the movements of each body part. This approach does not address the subsequent association problem that would ensue for target signatures comprised of multiple trajectories. A similar challenge is faced by the multiple target tracking approach proposed in [24], where the coning targets considered similarly consist of distinct sinusoidal trajectories.

To overcome these limitations, Pegoraro, et al. [132] proposed a density clustering and tracking based technique to separate the μ-D for up to four people walking back and forth along a corridor. A trajectory association algorithm is utilized to match clusters with tracked trajectories, and a CNN that incorporates a reconstruction loss term in the cost function is utilized to correctly identify the person walking in case tracking fails and clusters of some subjects cannot be separated. Although this method in principle can generalize to realistic target signatures, results are only presented for identification of walking people - HAR for multiple people in a scene is not considered. Moreover, it can potentially suffer from trajectory instability due to missed detections, a probable event in scenes where there is significant clutter, and ghost targets resulting from multi-path reflections. Significant effort

Figure 4.1: Conventional vs. angular projection-based radar signal processing (RSP) chain for multi-target scenarios.



is involved for trajectory management, a task that is increasingly complicated as the number of targets present increases.

A common limitation of all of the aforementioned works, however, is that they can perform separation only after tracking results are obtained or after radar images are computed from the raw data cube. The computation of tracking trajectories introduces additional delays in the classification process, which is detrimental to real-time applications. For cases like multiple close targets or gesture recognition for right and left hands, separation of targets in range-angle domain is generally not possible, detrimentally effecting performance of existing approaches. Additionally, decomposition or target separation made at the image level is not scalable to other data domains as the raw I/Q signals of the targets are still superimposed. This precludes the utilization of joint domain classification techniques, which require lower-level signal separation in the raw radar data and hence RDC itself.

The proposed angular subspace projection-based separation (ASPS) technique [103] projects the raw radar data onto an angle subspace and generates multiple low-level RDC-ω representations for the targets at different aspect angles. Figure 4.1 shows the fundamental differences between radar signal processing stages with and without the proposed projection method along with its advantages and disadvantages. In particular, we show that the proposed projection method improves the similarity between a target's original signal and the decomposed multi-target signal after projection. This enables the utilization of a DNN trained for the

single-target case to be utilizable even in multi-target scenarios. For a 9-class μD signature recognition problem, a four-layer CNN achieves 97.8% when three targets are present within the radar field-of-view. In cases where multiple targets are positioned with close proximity to each other, the RDC-ω representation offers multi-view inputs, which highlights the differences at each angle that can be exploited via a multi-view deep neural network to achieve improved classification accuracy. This work expands on preliminary work presented as a conference paper [105] by 1) adding a weighting stage to the projection pipeline, 2) characterizing target separability as a function of the number of antenna elements and aspect angle for an expanded number of targets, which includes human activities and several different gaits of a robotic dog, 3) demonstrating improved classification results for a 9-class HAR scenario using a 4-layer CNN and 10-class gesture/sign language recognition scenario using a multi-view DNN on an expanded dataset that has a greater number of samples per class.

The specific contributions of this study are as follows:

1. We propose the ASPS method to separate and boost the relative SNR of targets at different aspects angles, generating raw multi-view RDC-ω data representations.

2. The effect of the number of antenna channels and target aspect angle on ASPS performance is evaluated.

3. The effectiveness of ASPS is demonstrated for a HAR application in an end-to-end framework.

4. We propose a novel multi-view DNN that utilizes multiple RDC-ω derived micro-Doppler signatures as its input to boost classification performance when targets are in close proximity, such as when separately considering left and right hand movements for sign language recognition.

4.2 Multiple Target Signal Model

In an FMCW radar, the frequency of the transmit signal changes over time, typically as a linear sweep across the bandwidth, so that both range and velocity measurements can be acquired. [113]. In MIMO FMCW radar systems, the presence of multiple channels also enables the estimation of the angle of arrival of the radar backscatter from a target. The transmitted FMCW signal, $S_T(t)$, can be modelled as:

$$S_T(t) = \exp(j2\pi f_c t + j\pi\alpha t^2) \tag{4.1}$$

where, f_c is the carrier frequency, α is the chirp rate defined as the ratio of bandwidth, B, to the sweep duration, T, as $\alpha = \frac{B}{T}$.

Now suppose that once transmitted, the signal reflects back from a set of K targets with corresponding ranges R_i and radial velocities v_i $i = 1, 2, \ldots, K$. The received signal in one channel of the radar system can be expressed as:

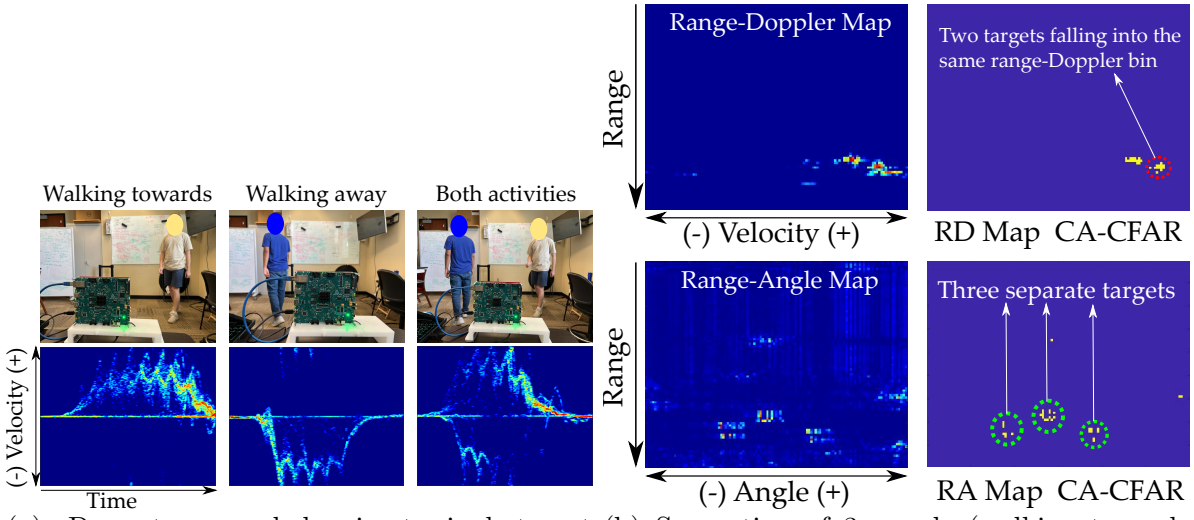
$$S_R(t) = \sum_{i=1}^{K} A_i \exp(j2\pi f_c(t - t_i) + j\pi\alpha(t - t_i)^2) \tag{4.2}$$

where, A_i is a complex constant related to target radar cross section, and t_i is the round trip time delay for the i^{th} target. Each RF sensor channel's raw data is collected as a time stream of in-phase (I) and quadrature (Q) samples. In FMCW transceivers, the received signal is mixed with a copy of the transmitted signal and then low-pass filtered to remove unwanted high-frequency mixing byproducts. The output of the filter as an intermediate frequency (IF) signal $S_{IF}(t)$:

$$S_{IF}(t) = \sum_{i=1}^{K} A_i \exp(j2\pi\alpha t_i t + \phi_i), \tag{4.3}$$

where ϕ_i is a constant phase term over time and is function of time delays, chirp rate and carrier frequency. Sampling the IF return signal results in N fast-time samples for each pulse,

Figure 4.2: Target separation in range, Doppler and angle domain for a multi-target scenario.



(a) μD spectrograms belonging to single-target and multi-target cases.

(b) Separation of 3 people (walking towards radar) in range, Doppler and angle domain.

while computation of its Fourier transform reveals the beat frequencies $F_b = \alpha t_i$, which are directly related to the round-trip travel time and distance between the radar and target. The velocity of a target can be obtained through transmission and coherent processing of P pulses. Because the PRI is typically much longer than the ADC sampling interval, the pulse number is typically referred to as slow-time, while the ADC samples are referred to as fast-time. For a MIMO radar with a ULA consisting of M virtual channels, a RDC with dimensions of $(N \times P \times M)$ can be formed. Fourier processing can be utilized to compute from the RDC various RF data representations, such as RD or RA versus time.

When multiple targets are present in the radar FoV, the received signal is comprised of the backscatter from all the targets in the scene. Thus, the μD spectrogram consists of the superposition of the μD signatures for all targets. Figure 4.2a illustrates the resulting μ-D spectrogram obtained when two people present in the FoV - one person is walking towards radar and the other is walking away from the radar. Comparing the multi-target spectrogram with the one obtained when the same activities are recorded individually in the radar FoV, it can be observed that the μD for each person exhibits the same patterns, but that in the multi-target spectrogram these signatures are overlayed or superimposed so that

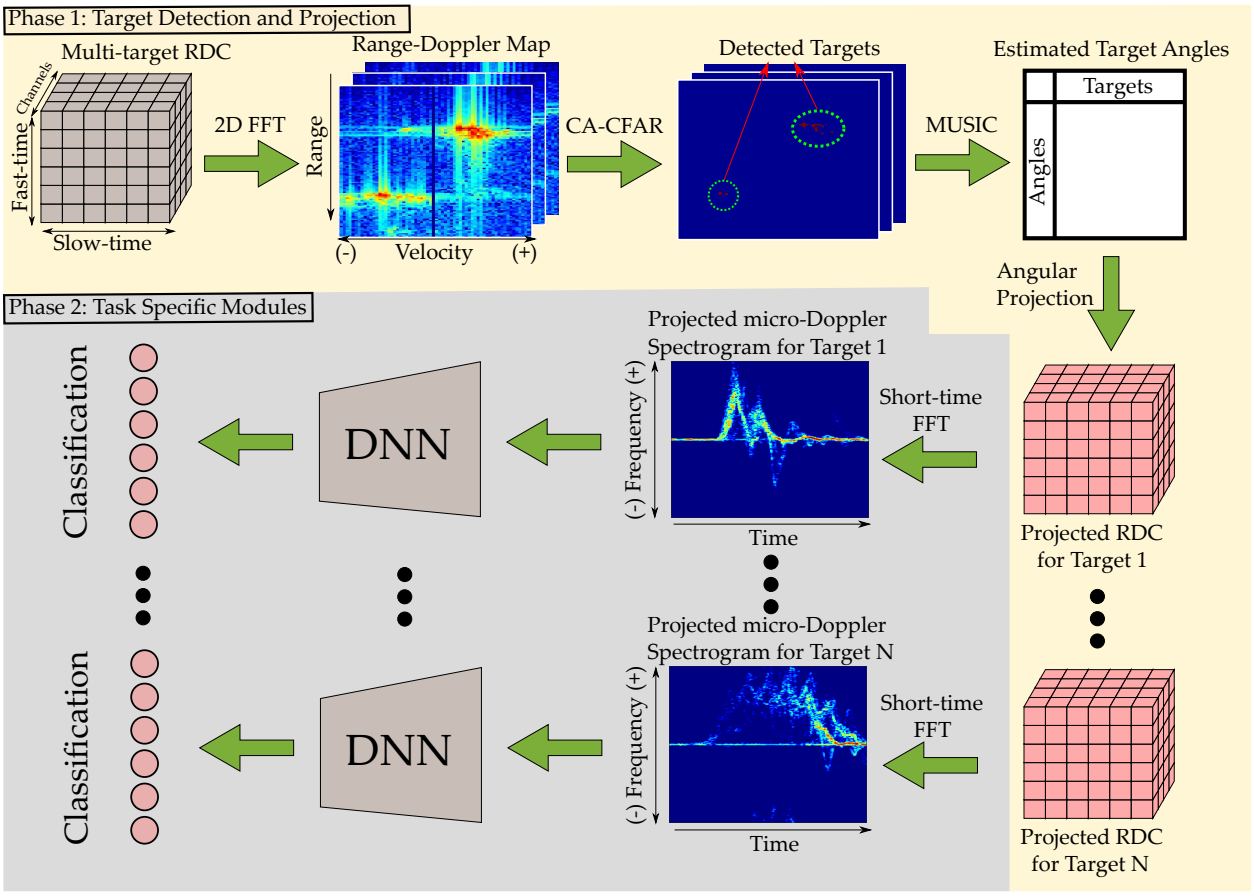
they interfere with each other. As DNNs for HAR are typically trained on μD signatures recorded when just a single person is present, such activities of the subjects in multi-target spectrograms will not be accurately classified. Thus, the signatures for each person need to be separated prior to input to a DNN for classification.

4.2.1 Rationale of Proposed Approach

Rather than immediately compute the μD, the potential separability of targets can be observed through the use of other radar data representations, especially the RD and RA maps. Consider the case of three people walking towards a radar. The RD and RA maps for this scenario is shown in Figure 4.2b. When a CA-CFAR detector is applied to the RD and RA maps, the presence of several distinct targets can be seen. In the RD map, however, despite three targets being present, only two can be detected as two of the people were located at the same distance from the radar. Consequently, in the RD map, their RF returns fall into the same range-Doppler bin and cannot be distinguished. On the other hand, in the CA-CFAR results for the RA map, three separate targets can now be observed. Hence multiple targets live in the range-Doppler or angle domains and might not be always be separated. Our goal is to generate multi-view raw radar data representations that will enhance target classifications even for cases that might not be separated in RA domains.

Towards this goal, we propose an angular subspace projection technique for signal separation that decomposes the raw complex RDC of multiple targets from each other by generating individual RDCs for each angle subspace. This process strengthens the signals received from the projection angle subspace, but weakens the signals received from any other angle. Because the signal separation is accomplished at the RDC-level, subsequently any desired radar data representation for the separated multi-view RDCs, including micro-Doppler signature, RD and RA maps, can be computed. This yields advantages over trajectory tracking-based approaches in the literature because the joint use of multiple input representations has given improved classification performance over just using micro-Doppler signatures. In essence, a

Figure 4.3: Proposed end-to-end framework of the angular projection method for a classification application.



more fundamental form of signal separation is being accomplished at the raw radar signal level.

The proposed framework, shown in Figure 4.3, consists of two main stages. The first stage involves target detection and angular projection. RD maps are obtained through 2D FFT on the fast and slow-time dimensions of each channel of the acquired raw RDC. Then, CA-CFAR adaptive thresholding is applied on the RD maps to detect the targets. Angle-of-arrival (AoA) of the detected targets are estimated using the MUltiple SIgnal Classification (MUSIC) [73] super-resolution algorithm. Next, the proposed projection approach is applied with the detected angles to form new, projected RDCs for each target. The second stage involves the implementation of task-specific processing on each individual target RDC.

For the task of classification, individual μD signatures can be generated by applying any desired time-frequency transformation on the projected RDCs. The μD signatures can then be given as input to a DNN to separately classify each target. Although μD is used as the input representation for classification here, different representations from multi-view RDC can also be computed. Even though angular subspace projection is applied to detected targets here, we also show that the proposed projection idea boosts the classification performance even for targets that were not separated at the initial stage such as right or left-hand gestures. The details of the computations comprising each stage are presented next.

4.2.2 Target Detection and AoA Estimation

The first stage to decompose multiple target RDC into individual RDCs of each target is the detection of targets in range-Doppler-angle domain. Number of targets in the radar FoV along with their angles are utilized to apply the projection algorithm. If the user has the priori knowledge of this information, angles of the targets can directly be fed into the angular projection algorithm without the initial detection stage. Otherwise, a target detection and an angle estimation method is applied.

A 2D FFT can be applied on the fast and slow-time dimension of the multi-target RDC to obtain the RD maps spanning each coherent processing interval, a.k.a. frame. In order to detect the targets in the RD maps, a widely used adaptive thresholding method, CA-CFAR detection, is applied.

Once CA-CFAR is applied on the RD maps, the detected target range-Doppler bins are passed through the MUSIC algorithm for angle estimation. This procedure is applied to all frames. To reduce grid effects, clustering is applied in angle space (i.e., detections with close angles are clustered into one group). If the number of detections for a particular angle cluster exceeds the pre-defined threshold, the center angle of the cluster is added to the list of projection angles, which is then given as input to the projection algorithm.

4.2.3 Projection of RDC

The previous detection stage generates angle estimates of each target. In this part, we project the raw multi-target radar data to the angular subspace of each target. To do this let us first define the angular space. If there is a target at a specific angle with respect to the array, the steering vector defines the signal model received at the array. For the case of a ULA the steering vector, $\mathbf{a}$, can be formulated as follows:

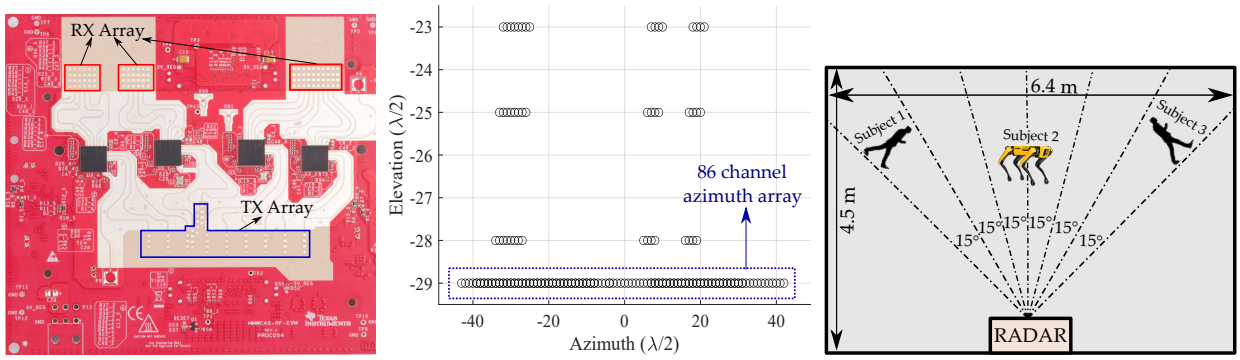
$$\mathbf{a}(\theta) = [1 \; e^{-j(2\pi d \sin(\theta)/\lambda)} \; ... \; e^{-j(2\pi(M-1)d \sin(\theta)/\lambda)}]^T \tag{4.4}$$

where θ is the aspect angle, λ is the signal wavelength, d is the spacing between array elements, and M is the total number of channels in the ULA. Repeating equation 4.4 for each discretized angle θ yields the steering matrix. The projection method takes the lower and upper bounds of the desired projection angular interval (i.e., θ_l and θ_u, respectively) as inputs. For a given angle space $[\theta_l, \theta_u]$, we can discretize this space and create a steering matrix, as $\mathbf{B} = [\mathbf{a}(\theta_1)\, \mathbf{a}(\theta_2), \ldots, \mathbf{a}(\theta_i), \ldots, \mathbf{a}(\theta_S)]$ where each column is a steering vector for the corresponding aspect angle $\theta_i \in [\theta_l, \theta_u]$. Here the column space of $\mathbf{B}$ spans the angular subspace we want to project and slow time index k can be projected onto the column space of $\mathbf{B}$ as follows:

$$\widehat{\mathbf{x}}_{nk} = (\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T)\mathbf{x}_{nk} \tag{4.5}$$

where $\widehat{\mathbf{x}}_{nk}$ is the projected data. Repeating the projection for relevant fast-time and slow-time indexes will construct the projected RDC for the given angle subspace. The angle subspace can be selected as the angular extent of the detected targets. In some cases the angular subspace can also be designed depending on the application. For example, in radar-based sign language recognition, to better represent right and left-hand activities, angular projections for each hand can be done considering an angular subspace that can include all angles at

Figure 4.4: Virtual array generation from MIMO array using TDM (a-b) and the experimental setup (c).



(a) Position of antenna elements.

(b) Created virtual array.

(c) Top-view of the experimental setup.

one hand can be. As the output of this subspace projection multi-view RDCs are obtained and varying feature representations can be computed for each RDC.

4.3 Experimental Setup and Dataset

In this work, Texas Instrument's AWR2243 Cascade MIMO radar is employed as the RF sensor which uses a sawtooth FMCW model. Frequency of the transmitted signal linearly increases as a function of time during sweep repetition period or sweep time, T. The start frequency in the radar system is 77 GHz and a bandwidth of 4 GHz is used. So, the signal is linearly increased up to 81 GHz. The radar system can be configured as a long-range radar (LRR) in the beamforming mode for higher signal-to-noise ratio (SNR) or as a short-range radar (SRR) using the MIMO mode for enhanced angular resolution. In this work, MIMO configuration is utilized.

4.3.1 Virtual Array Generation with MIMO Processing

The experimental MIMO radar contains 4 radar sensor chips cascaded together, where each containing 3 transmitter TX and 4 RX antennas, resulting in a total of 12 TX and 16 RX channels. 9 of the 12 TX antennas are in the same vertical position and remaining 3 of them are at different heights. As for the Rx antennas, all of them are located at the same

Table 4.1: The acquired dataset for different number of targets.

Number of Targets	**Angles**	**Number of Samples**
1	0°, ±45° ±15°, ±30°	476 16
2	0°, ±45°	15
3	0°, ±45°	6

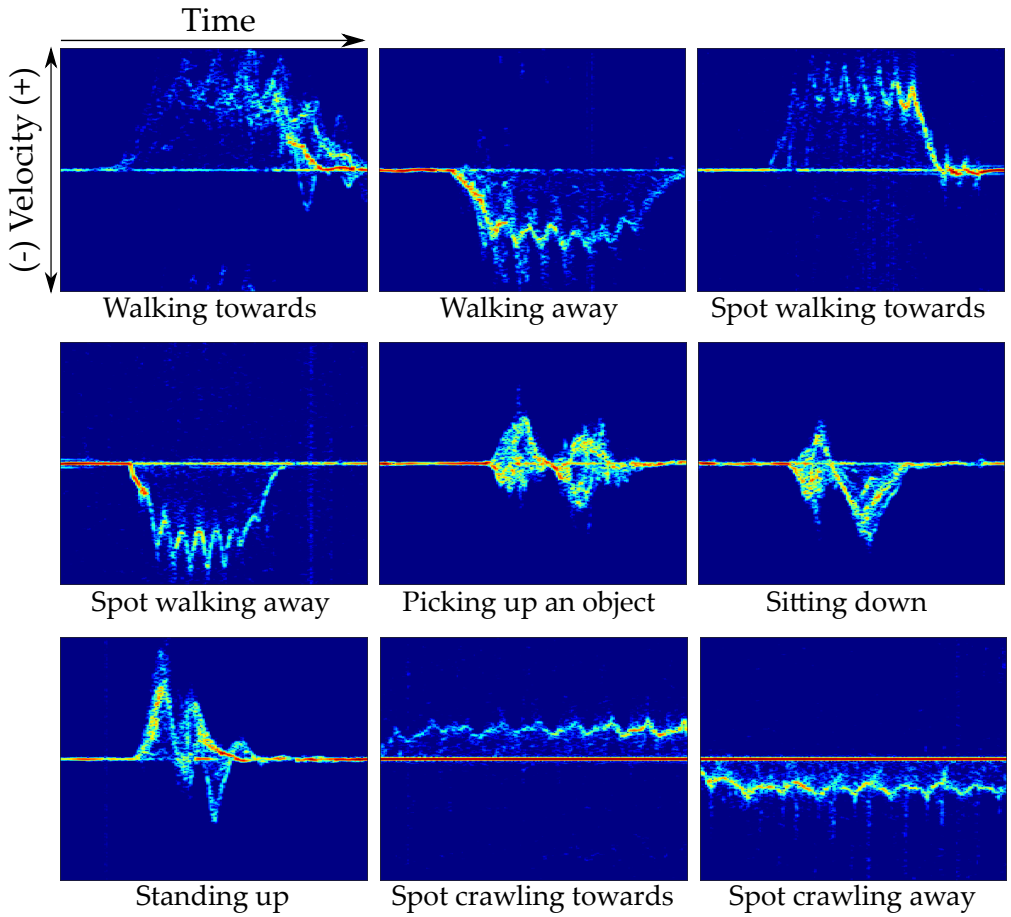
height with a sparse distribution in the horizontal axis as depicted in Figure 4.4a. Utilizing relative positions of different TX-RX pairs in MIMO radars, one can form a virtual array with a larger aperture size than what is provided on the hardware. It can be achieved by employing different modulation techniques such as TDM, BPM, code-division multiplexing (CDM).

BPM works by introducing binary phase difference (i.e., $0 and 180$) between consecutive pulses which is not effectively applicable when there exists more than two TX channels. In this work, TDM is used as the virtual array generation scheme which schedules each TX-RX pair to transmit pulses. Using 12 TX and 16 RX, a virtual array with 192 channels (including the overlapping TX-RX pairs) can be formed as depicted in Figure 4.4b where 86 of these 192 channels can be used to form a ULA in azimuth direction, providing angular resolution of as small as 1.4°.

4.3.2 Data Collection

In order to assess the effectiveness of the proposed method, a HAR dataset with 9 activities are acquired where 5 of them are human activities and 4 of them belong to a robotic dog (i.e., Boston Dynamics's Spot). The acquired classes for human participants can be listed as: walking towards radar, walking away from radar, picking up an object from the ground, sitting down to a chair, standing up from a chair. The classes for Spot are: walking towards radar, walking away from radar, crawling towards radar, crawling away from radar. The experiment is conducted in a 4.5 m × 6.4 m indoor area. Figure 4.4c shows

Figure 4.5: μD spectrogram samples of different classes.



the experimental setup and the layout of the room where the data are acquired. Each of the activities are performed at the angles of 0° and ±45°. Few more samples are also collected at ±15° and ±30° in order assess the limitations of the proposed method. Each recording for an activity is lasted for 7 sec, and walking/crawling activities are started at 4.5m away from the radar. Single and multi-target cases are considered, and Table 4.1 summarizes the acquired dataset for varying number of targets.

In this study, μD spectrogram is used as the RF data representation type for similarity analysis and classification. Sample μD spectrograms belonging to different classes are provided in Figure 4.5. Three different datasets are generated from the acquired samples:

1. The first dataset consists of single activity samples where only one target present in the radar FoV. This dataset is referred as *single activity dataset*.

2. Secondly, merging of raw I/Q signals of the targets located at different aspect angles by summing the raw signals up allows us to create a scene with multiple targets in a synthetic way. This dataset is referred as *merged multi-target dataset* throughout the chapter.

3. Finally, the data of multiple subjects located at different aspect angles are also recorded which are named as *real multi-target dataset*.

Next, we present the performance analysis of the proposed approach and the effect of several radar system parameters. Classification performance for two additional applications and their corresponding datasets will also be presented.

4.4 Performance Analysis of the ASPS Method

4.4.1 Similarity Comparison

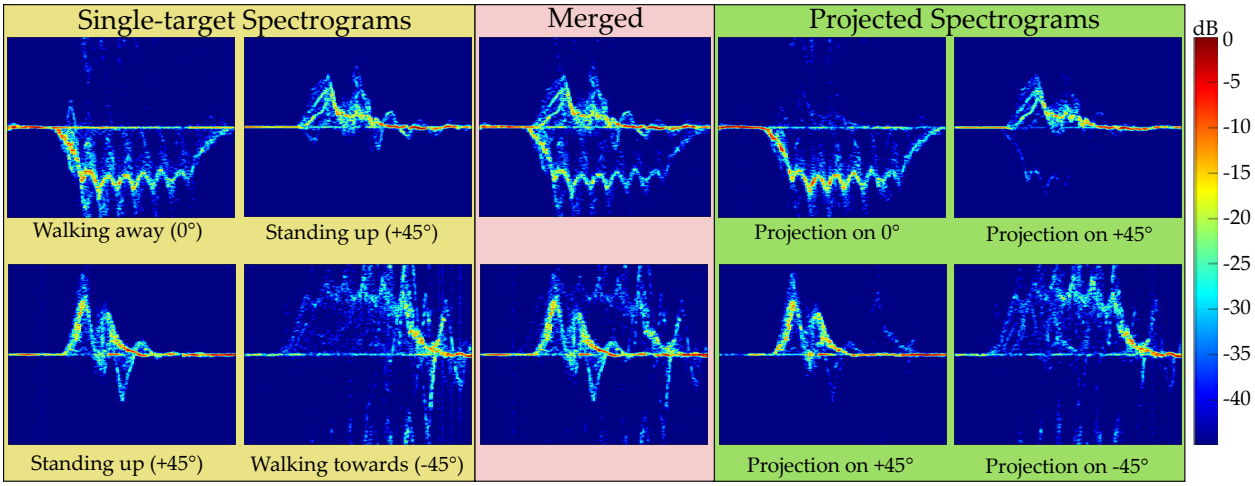
Projection of the raw signal onto the angular subspace spanned by $\mathbf{B}$ keeps the returned signal strength from the targets located at the angular interval spanned by $\mathbf{B}$, while fades out the return signals of the targets located at other angles. In order to quantitatively evaluate the quality of the projected RDCs, their similarity with the original, single target spectrograms are compared by applying the following steps:

- **Step 1:** Raw data with single targets at different aspect angles are recorded, their RDCs are formed and corresponding μD spectrograms are generated.

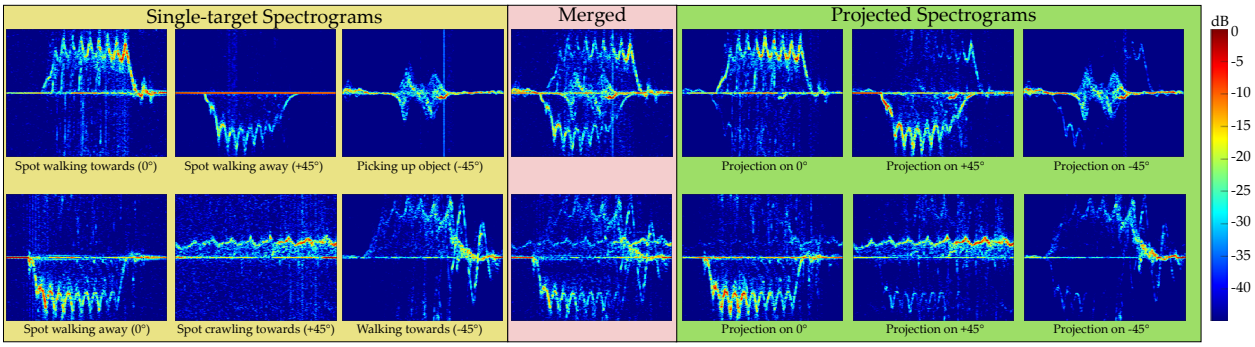
- **Step 2:** Multiple single target RDCs are merged by adding them up to create a combined RDC which contains the raw return signals of multiple targets, and the μ-D spectrogram of the merged RDC is generated.

- **Step 3:** The combined RDC is projected onto the target angular subspace. Resulting projected RDCs are used to generate new, projected μD spectrograms for each projection.

Figure 4.6: Projection results for two and three target cases.



(a) Two target case.



(b) Three target case.

- **Step 4:** Compare the similarity of the projected spectrograms to the original, single target spectrograms with varying metrics.

Figure 4.6 illustrates the aforementioned steps with the μ-D spectrograms for two and three target cases. In the first row of the Figure 4.6a, single target spectrograms for walking away (at 0°) and standing up (at +45°) activities and their merging result is presented. After the projection of the combined RDC onto 0° and +45°, it can be seen that individual targets' signals are recovered almost perfectly. Second row, presents the results for standing up and walking towards activities where both of them have mostly positive Doppler frequencies. It can be seen that the projection method is still able to separate two targets quite well, indicating that the performance of ASPS method is agnostic to the signature of the Doppler

frequency caused by the direction of the target's radial motion. Figure 4.6b presents the projection results for a three-target case performed at -45°, 0° and +45° angles. In the first row, while Spot's *walking towards* and *walking away* activities seem to be well separated than others, projection result for *picking up an object* has some leftover signals of Spot's activities before and after the *picking up an object*'s signature. This is due to the fact that the projection angle subspaces are not fully orthogonal and still some weak projections from other angles are observed.

In order to quantitatively assess the similarity of the projected spectrograms to the original (i.e., single target) ones, three different similarity metrics are considered, namely, structural similarity index (SSI), pixelwise mean-squared error (MSE) and peak signal-to-noise ratio (PSNR). Table 4.2, presents the averaged similarity results across all the samples of *merged multi-target dataset* for two and three target cases where X, Y, Z $\in \{\pm 45, 0\}$. It can be observed that when the target angle and the projection angle matches, resulting spectrograms have higher SSI and PSNR and lower MSE than non-matching case as expected. Although the presented results are the average of all samples, having higher SSI and PSNR with lower MSE when the target and the projection angle matches is consistent for all individual samples.

4.4.2 Effect of Angular Difference of the Targets

So far, only angles of -45°, 0° and +45° are considered for the similarity measures. In order to understand the effect of angular difference, $\Delta\Theta$, of the targets on the projection results, more data samples are collected from -30°, -15°, 0°, +15°, and +30° for the activity of *walking away from radar*. These samples are then merged with a *picking up an object* activity sample recorded at -45° one-by-one, resulting in varying $\Delta\Theta$ between two targets where $\Delta\Theta \in \{15°, 30°, 45°, 60°, 75°$ and $90°\}$. Table 4.3 presents the similarity results for μD spectrograms when RDCs of the targets at varying angles merged with the RDC of the target located at -45°, and projected onto the target's original angle. It can be observed that as $\Delta\Theta$ increases, SSI and PSNR increase as well and MSE decreases, meaning that

Table 4.2: Mean similarity results for the μD spectrograms where X, Y, Z $\in \{\pm 45°, 0°\}$.

Num. of Targets	Target Angle	Projection Angle	SSI	MSE	PSNR
2	X	X	0.59	1.82e3	16.93
		Y	0.46	3.6e3	13.16
	Y	X	0.47	3.56e3	13.2
		Y	0.55	2.17e3	16.46
3	X	X	0.56	2.22e3	16.43
		Y	0.45	3.68e3	13.14
		Z	0.42	4e3	12.8
	Y	X	0.46	3.73e3	13.07
		Y	0.57	2.06e3	16.52
		Z	0.44	3.9e3	12.9
	Z	X	0.44	3.95e3	12.88
		Y	0.44	3.79e3	12.97
		Z	0.53	2.48e3	16.05

the projected spectrogram resembles more to the original single target spectrogram. Figure 4.7 shows the geometry of the multi-target scenario and resulting projected spectrograms belonging to Table 4.3. It can be seen that the projection method has hard time to separate targets when $\Delta\Theta = 15°$ where both activities are well visible in the resulting spectrogram. When $\Delta\Theta = 30°$, although a complete isolation of two targets is not achieved yet, *picking up an object* activity is mostly suppressed and *walking away* seems to be the dominant activity. When $\Delta\Theta = 45°$, μD signature of the *picking up an object* activity is barely noticeable and there exists only a small portion of the leftover weak signatures. When $\Delta\Theta = 60°$ or $75°$ a complete separation can be observed with μD spectrograms containing only one target's signatures. From these results, it can be inferred that as $\Delta\Theta$ between targets increases, a better separation is achieved and all similarity metrics perform better.

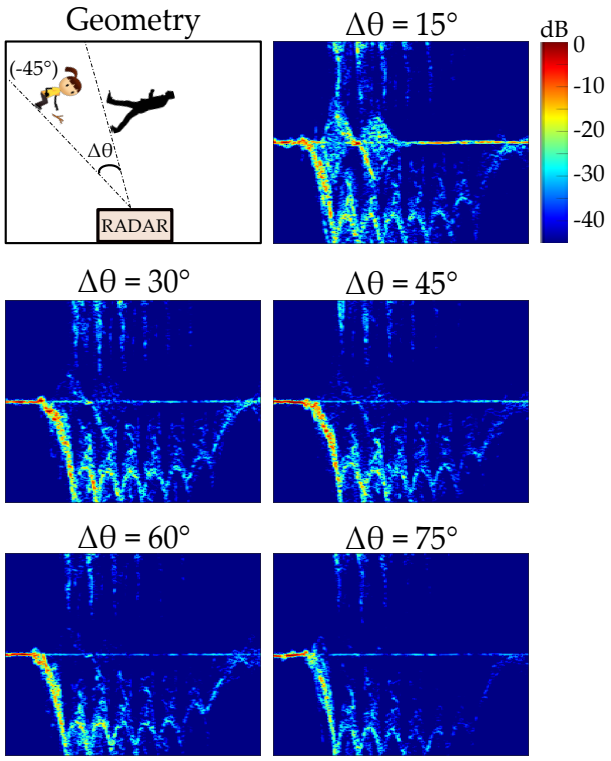
4.4.3 Effect of Number of Antennas

The created virtual array has 86 channels which forms a ULA in the azimuth direction. In any kind of angle estimation method, the angular resolution is proportional to the number of channels in the antenna array. However, it worsens as the aspect angle deviates from

Table 4.3: Similarity results when two activities (i.e., *picking up an object* and *walking away*) with angular difference of $\Delta\Theta$ are merged, and projected onto the targets' original angle.

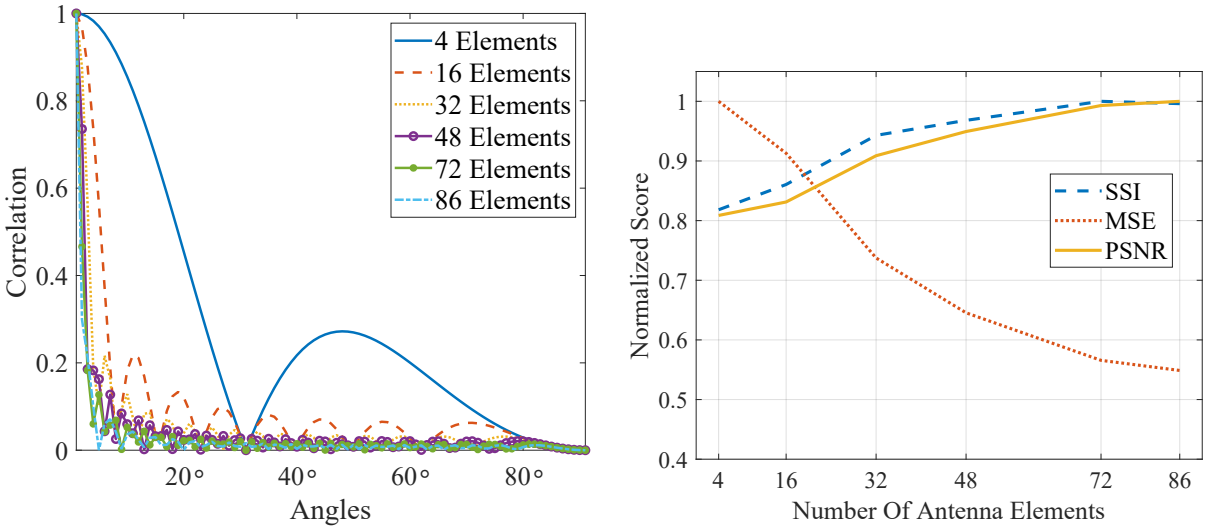
$\Delta\Theta$	15°	30°	45°	60°	75°	90°
SSI	0.67	0.72	0.77	0.83	0.85	0.85
MSE	1.75e3	0.63e3	0.55e3	0.45e3	0.39e3	0.3e3
PSNR	15.7	20.15	20.73	21.63	22.25	23.36

Figure 4.7: Generated μD spectrograms after projecting multi-target (*picking up an object* and *walking away*) RDCs onto original target angles.



the direct line-of-sight of the radar. A similar phenomenon can be observed in the ASPS method as well. In some cases, it can be seen that the projected μD spectrograms still contain signatures of some portion of the activities from other angles. This due to the fact that the steering vector, $a(\theta)$, for an angle, θ, is not fully orthogonal to the other angles. Figure 4.8a shows the correlation between the steering vector of angle 0° and the steering vectors of other angles for different number of antenna elements. It can be stated that although there are some ups and downs, especially noticeable for 4-elements case, the correlation between

Figure 4.8: Correlation of angles and similarity between the original single target and projected μD spectrograms for varying number of antenna elements.



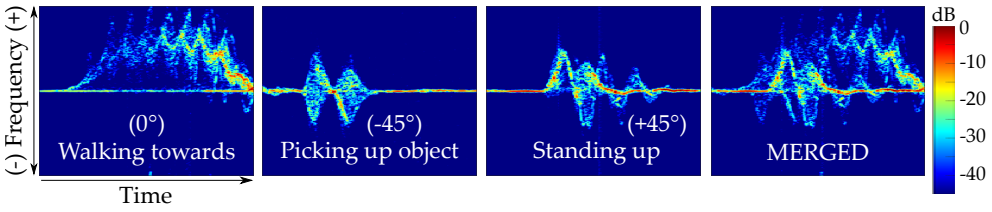
(a) Correlation between steering vector of angle 0° and other angles for different number of array elements.

(b) Similarity between single target and the projected spectrograms for different number of array elements.

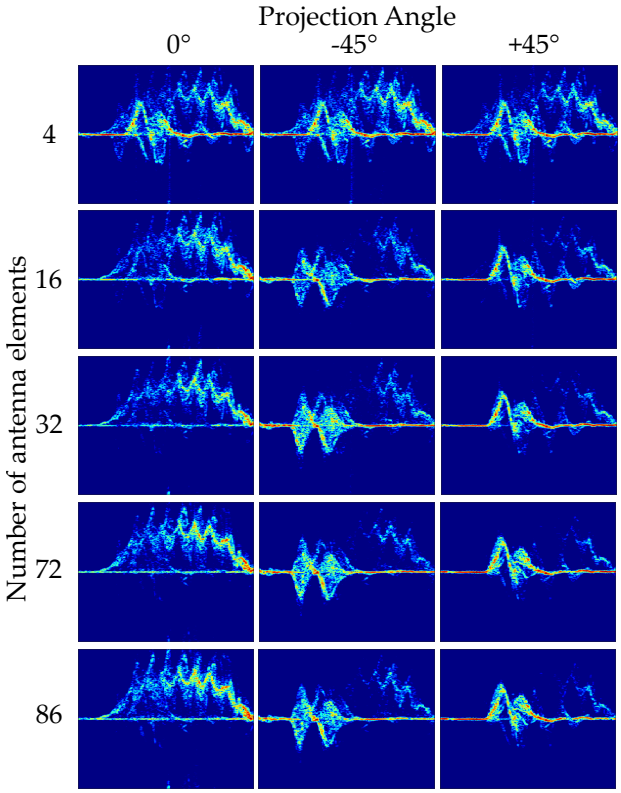
steering vectors gets smaller as the number of antenna elements increases which is in favor of the projection algorithm as lower correlation between two angles will lead to a better separation of the targets located at those angles.

Figure 4.8b presents the normalized SSI, MSE and PSNR values of the projected and the original spectrogram pairs for different number of enabled antenna elements. It can be observed that SSI and PSNR increase as with the number of antenna elements while MSE decreases, indicating that the projections made with a larger antenna aperture size resembles more to the original target signatures. Qualitative results for this observation are presented in Figure 4.9. Figure 4.9a shows the μD spectrograms belonging to 3 different activities at different angles and their merging result. Figure 4.9b shows the resulting μD spectrograms for projections for the target angles with different number of antenna elements. While the separation is barely noticeable and poor for lower number of antenna elements, it gets better as the number of MIMO channels increases. Isolation of individual targets start to become quite clear after 32 channels, and 72 and 86 channels yield very close results.

Figure 4.9: Projection results for varying number of antenna elements.



(a) Individual and merged activities.



(b) Projected spectrograms.

These qualitative results are found to be in-line with the quantitative results obtained in Figure 4.8b.

4.5 Classification with RDC-ω Representation

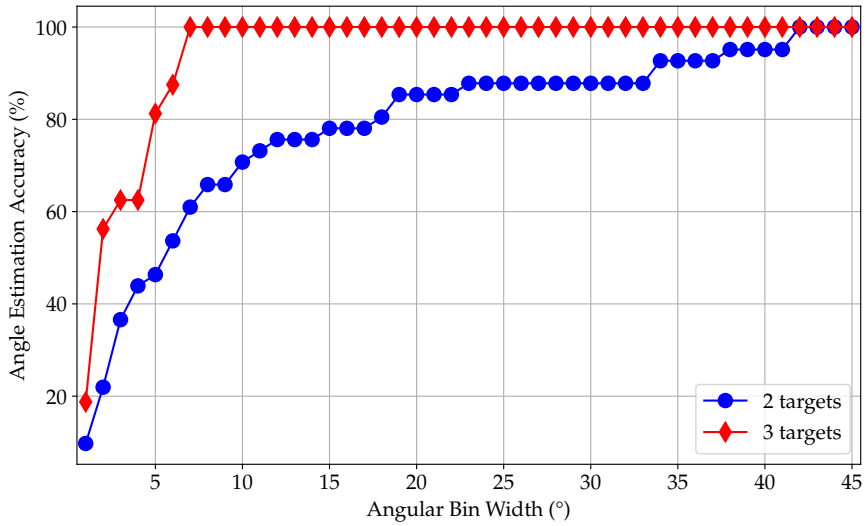
In this section, utilization of ASPS derived RDC-ω representation for multi-target activity classification is presented. First, multi-person HAR where there is sufficient angular separation is considered. The RDC-ω representation for each target is input to a DNN trained with

data from single-target. Next, two cases where targets are in close proximity is considered: 1) the left and right hands during articulation of two-handed gestures/sign language and 2) two-people doing different activities next to each other. In these cases, the proposed multi-view DNN, taking multiple spectrograms generated from the RDC-ωs as inputs, is shown to boost the multi-target classification performance.

4.5.1 Multi-Person Activity Recognition

HAR and indoor monitoring are drawing more attention as with the development of state-of-the-art end-to-end solutions utilizing RF sensing. HAR dataset is used to assess the effectiveness of the method. First, a very basic case, classification of the *single activity dataset* is considered. μD spectrograms of single targets are fed into a 4-layer CNN followed by two fully connected layers and one more fully connected layer having number of nodes equal to the number of classes for classification. There exists a *Reshape* layer before the final *Dense* layer, so that the output of the model will have the shape of $T_{max} \times N_{class}$ where T_{max} is the predefined maximum number of detectable targets and N_{class} is the number of classes. This modification allows model to give prediction for each target. *softmax* activation function is often employed in the last layer of a classification network for multi-class classification problems. It enables normalization of the output of a network to a probability distribution over predicted output classes, based on Luce's choice axiom. In this work, in the final classification layer, *sigmoid* activation function is preferred over *softmax* which is more common for multi-class classification problems. Such unconventionality is needed because output values of the *softmax* should add up to 1, however, when there are less number of targets than T_{max} in the scene, all the output values for non-existent target nodes should be close to 0, but *softmax* cannot provide such output while *sigmoid* can. In this application, T_{max} is set to 3. The trained model achieved the testing accuracy of 98.7% for the *single activity dataset*. However, when the trained model is tested on *merged multi-target* and *real multi-target* datasets where multiple targets present in the scene, accuracies drop down to

Figure 4.10: Angle estimation accuracy for different angular tolerance values.



5.8% and 8.3%, respectively. Such performance drop is expected as the latent space of the multi-target spectrograms may not resemble to any of the individual classes.

Application of the projection idea can be quite useful in this scenario as it has capability to isolate μ-D signatures of the individual targets. As mentioned earlier, estimation of the target angles is a necessary step in this framework, and the ground truth label of the projected spectrogram will depend on the ground truth class of the original target. If the original target angle and the estimated projection angle matches, they will share the same class label. In order to decide whether the estimated angle and the original angle matches, the angular grid needs to be divided into angular bins. If the estimated angle and the original are in the same angular bin, they are said to be matching and will share the same class label, otherwise, the target will be regarded as a false detection, and will not be classified since its ground truth becomes vague. Figure 4.10 shows the angle estimation accuracy for varying angular bin widths for 2 and 3-target cases of the *real multi-target* dataset. Accuracy here is defined as the ratio of the number of correct angle estimations divided by the total number of estimated targets in the dataset. It can be seen that the angle estimation accuracies for the 3-target case is higher than the 2-target case, especially for lower angular bin widths. One reason of this could be that the ground truth of the 3-target case spans a larger angular

Table 4.4: Classification accuracy of the projected spectrograms for the *real multi-target dataset*.

Angular Bin Width	5°	10°	15°	20°	30°	35°	45°
Accuracy (%)	96.9	97.8	93.8	92.2	92.3	88.9	79

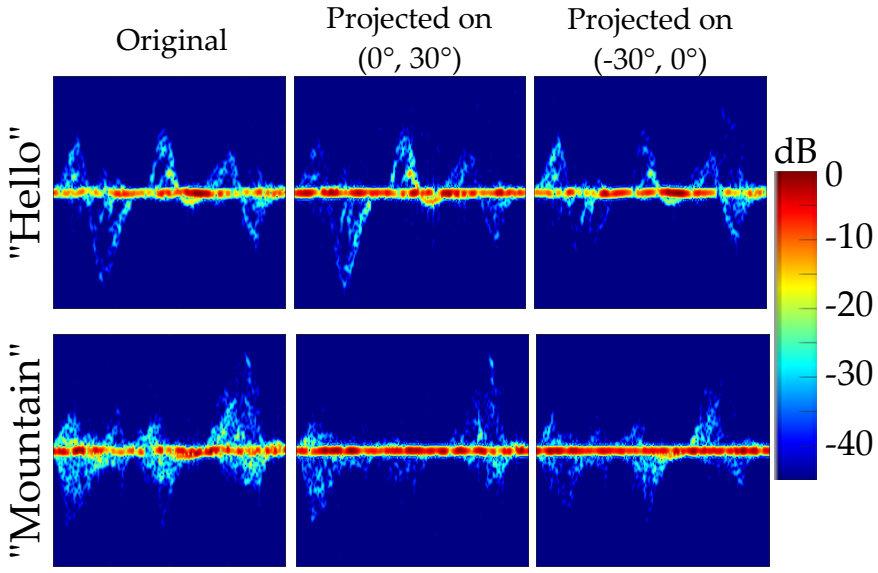
Table 4.5: Classification accuracy of the projected spectrograms for varying number of MIMO channels.

Number of MIMO Channels	4	16	32	72	86
Accuracy (%)	46.8	82	93.1	94.6	94.6

interval than 2-target case, hence it is more likely for an estimated angle to fall into detection interval of a ground truth angle.

After obtaining the ground truth labels for the projected spectrograms, they are given as input to the model trained with *single activity dataset*. Table 4.4 presents the prediction accuracies of the projected spectrograms for the *real multi-target dataset* for varying angular bin widths. It can be seen that the accuracy reaches its maximum with 97.8% at 10° which is only 1% lower than the single target prediction task, although none of the *real multi-target dataset* samples are used in the training stage. This shows the benefit of utilizing the ASPS method in a multi-target scenario by decomposing the multi-target spectrograms into individual spectrograms and enabling flexibility to treat them as single target spectrograms, and classify them in that way. Finally, Table 4.5 presents the testing accuracy results for different number of virtual channels in the MIMO array. It can be observed that the accuracy improves with increasing number of MIMO channels since more number of channels yields better separation of μD signatures and the projected spectrograms start to resemble more to the single target samples.

Figure 4.11: μD spectrograms of the projected ASL signs.



4.5.2 Multi-view DNN for Multiple Targets in Close Proximity

American Sign Language Recognition

ASL recognition using radars has become an emerging research field, especially with the development of small package, commercially available RF sensors. ASL signs are composed of a mixture of various hand movement types (e.g., circular, straight, and back-and-forth). While some signs are articulated with one hand, some are articulated using both hands. Separation of return signals from left and right hands can be quite useful in order to retrieve the individual characteristics of each hand's motion. However, rapid change in the spatial position of the hands and two hands being very close to each other introduce challenging scenarios and classical representations such as RA domain cannot separate the right and left hand as two separate targets.

In order to demonstrate the performance of the proposed ASPS approach, an ASL dataset with 10 different signs (YOU, HELLO, WALK, DRINK, FRIEND, KNIFE, WELL, CAR, ENGINEER, MOUNTAIN) are collected from 6 participants. Moreover, considering not all the commercially available MIMO radars have antenna apertures as large as 86 channels, TI's AWR1642BOOST single-chip radar with 2 TX and 4 RX channels is employed as the RF

Figure 4.12: Proposed multi-view CNN model where W_i denotes the shared weights at the i^{th} layer.

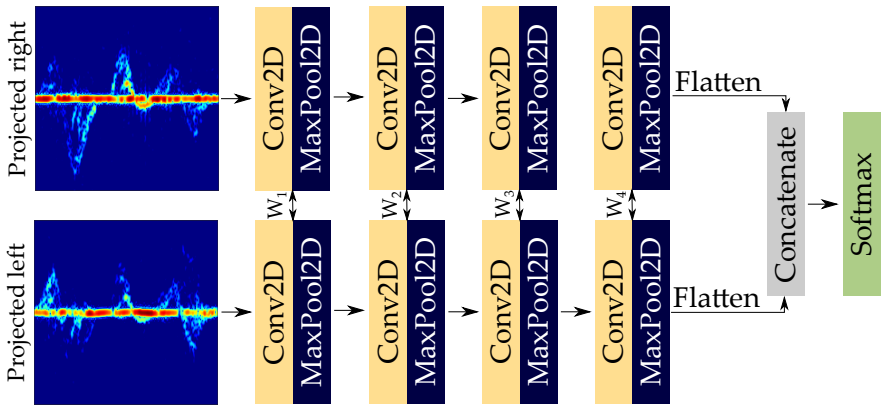


Table 4.6: Classification accuracy (%) comparison results for the ASL recognition task for varying projection angle intervals.

No Projection	**Projection Angle Interval**		
	0 to ±30	±30 to ±60	±60 to ±90
93.3	**96.9**	93.9	93.3

sensor. A virtual array of 8 channels is formed using BPM scheme for MIMO processing. ASL samples are then projected onto left and right angular intervals of (0° to ±30°), (±30° to ±60°) and (±60° to ±90°). Figure 4.11 shows the original and projected samples for the words HELLO and MOUNTAIN. In the first row, it can be seen that while projecting onto (0°, 30°) interval strengthens the power of first negative and second positive peaks, projecting onto (-30°, 0°) strengthens the power of first positive and the last negative peaks. A similar observation can be made for the word MOUNTAIN in the second row. Although ASPS method, in this case, cannot completely separate μD signatures of left and right hands, the resulting spectrograms can be used to enrich the feature space in the learning-based classification algorithms.

For this 10-class ASL recognition problem, a multi-branch DNN model which takes left and right projected spectrograms as inputs is proposed. The network has 2 input layers and 4 CNN blocks. Weights of two branches are shared across corresponding layers to reduce the number of trainable parameters, hence better regularizing the model. Two branches are then

merged in a *concatenation* layer and *softmax* is employed for classification. The proposed network model is presented in Figure 4.12. The performance of this model is then compared with the baseline model which is essentially the same network, but with a single branch that takes the original spectrograms as inputs. Table 4.6 presents the classification results when the projection is employed and not employed. It can be seen that while baseline model's mean accuracy is limited to 93.3%, the classification accuracy of the projected spectrograms with the proposed network can go up to 96.9%. The reason projection performs better for angular subspaces of $[0, \pm 30]$ might be because the articulating ASL signs mostly live in this angular subspace. These results show that although the ASPS method has some limitations for challenging scenarios such as separation of left and right hands signals, it can still be employed to enrich the feature space in the model which yields a better recognition performance.

Two People Performing Activities Side-by-Side

Similar to the ASL case, when the targets are very close to each other or side-by-side, it is a challenging task to individually isolate and extract each target's μD signal as depicted in Figure 4.7. When the generated spectrograms from the ASPS method have major μD components from multiple targets, it is not plausible to feed the resulting projected spectrograms into a DNN model trained with only single target samples as their feature spaces are different from the single target samples. Therefore, a similar approach can be followed as in the ASL case, and the ASPS method can still be benefited in the cases when there is a prior knowledge about the number of targets present in the scene.

In order to demonstrate an alternative use of the ASPS method for closely located targets, a separate HAR dataset is acquired where two targets were performing the same or different activities in the direct line of sight of the radar, very closely on the lateral axis and side-by-side for the stationary activities like sitting down, standing up and picking up an object. In total, 276 samples for 5 different activities are acquired including walking towards

Table 4.7: Classification accuracy (%) comparison results for closely located targets for varying projection angle intervals.

No Projection	**Projection Angle Interval**			
	0 to ±5	0 to ±30	±30 to ±60	±60 to ±90
91.1	92.9	**94.6**	92.9	87.5

and away from the radar. The acquired samples are projected onto different pairs of left and right angular intervals: (0° to ±5°), (0° to ±30°), (±30° to ±60°) and (±60° to ±90°). The projected left and right samples are then used to train the multi-input CNN model presented in Figure 4.12 by only modifying the number of nodes in the softmax layer according to the number of classes. Performance of the multi-input CNN model with projected samples is compared with the single branch model with no projection. Single branch model has the same hyperparameters and the identical architecture to the one branch of the multi-branch CNN model.

Table 4.7 presents the testing results for the single branch baseline model (i.e., no projection) and the proposed multi-branch CNN with the projected spectrograms. It can be seen that all the projected angular intervals except (±60° to ±90°) outperform the the baseline method with no projection. Obtaining lower performance for the (±60° to ±90°) case is an expected result since all the activities were performed in the direct line of sight of the radar, hence not much information is present in the higher angular intervals. On the other hand, projection on (±0° to ±30°) yields the highest performance with 94.6% which is more than 3% performance gain when compared to the baseline method.

4.6 Discussion and Conclusions

This study proposes two techniques to address the challenge of multi-target recognition: 1) an angular subspace projection-based separation (ASPS) method to emphasize at the raw signal-level the data of targets located at different aspect angles with respect to the radar; and 2) a multi-view DNN, which takes as input spectrograms generated from the multiple

RDC-ω representations generated via ASPS. The approach was demonstrated for several use cases: multi-person/robot motion recognition, HAR for closely spaced targets and sign language recognition, where ASPS is used to generate RDC-ω representations for the left and right hands that are then used in a multi-view network for classification. The significance of the new RDC-ω representation is that the angle-depended boosting of target signatures is accomplished at the level of the raw RDC, not after post-processing in images as is done with current approaches. The RDC-ω representation can thus be also used to develop DNNs that directly operate on the raw RDC-ω data or other 2D/3D radar data representations than micro-Doppler, such as range-Doppler, range-Angle maps.

The case studies presented in this paper show that the proposed method is able to generate individual RDCs for each target so that the newly generated RF data representations from the projected RDCs can be treated as samples belonging to a single target in a classification task. For a nine-class activity recognition scenario, the projected multi-target RF data samples were classified with 97.8% accuracy by a CNN model which is trained solely on single target samples, despite multiple targets being within the field of view. We also characterize the effect of the number of MIMO channels on the performance of the projection method in terms of different similarity metrics and classification accuracy. In the case of close proximity between targets, we show that the multiple RDC-ω representations can be used in a multi-input DNN framework to boost classification performance.

In future work, we plan to further investigate the design of DNNs operating on the raw RDC-ω representations to enable real-time recognition applications, such as RF-enabled cyber-physical human systems for explicit and implicit control of personal assistants and autonomous vehicle in-cabin driver/passenger monitoring sub-systems. Future work will also include multi-target activity classification under more challenging scenarios such as moving clutter.

CHAPTER 5

INTERACTIVE LEARNING OF NATURAL SIGN LANGUAGE

5.1 Introduction

An important challenge to the development of RF-sensing based human motion recognition algorithms, and sign language processing technologies more broadly, is the lack of availability of adequate datasets for model training. Not just the *amount* of data, but the *quality* of data is critical. Sign language is comprised of not just physical spatio-temporal articulations, but also non-manual markers (such as eyebrow, eye, cheek, and mouth postures, head and body position) to convey linguistically and emotionally rich messages. Like all languages, sign language is influenced by personal, regional, and cultural traits that can result unique variations in expression, as well as linguistic properties influenced by grammar and prosidy [124]. In prior work, we have shown that RF sensing data also captures these human and linguistic qualities, including co-articulation [65] and degree of fluency [69]. In particular, we showed that a support vector machine (SVM) could be trained on RF micro-Doppler signatures to discriminate between fluent users of American Sign Language (ASL) and imitation signers - hearing participants who strive to replicate ASL signs after watching and practicing from videos of fluent signers. More significantly, we found that using imitation data to train and validate machine learning (ML) algorithms, as done in some works [49, 116], over-optimistically predicts the recognition accuracy of 20-signs by as much as 20% in comparison with that obtained using data from fluent signers - the actual prospective users of ASL recognition technologies.

Given the great variety and complexity of the expression of sign language, existing ASL datasets (for any sensing modality) lack adequate size and diversity to adequately train models that can generalize to signers of all ethnicities, regions and accents. Traditional in-lab data collection typically involves inviting participants to articulate ASL in a directed fashion and in a controlled setting. However, this type of data collection has many disadvantages, from both a sociological and technical perspective. In-lab data collection may attract participants from only a certain demographic [17], unwittingly resulting in datasets that contain inherent biases and that do not adequately represent certain groups of signers. Moreover, the controlled settings of a lab result in pristine data, which may not be representative of real-world environments or natural articulation. For example, in collecting directed datasets, the signer may position the hands on the knees before repeating the directed sign, subsequently returning the hands to the same position. This type of scripting precludes capture of variations due to co-articulation - the variation in the spatio-temporal properties of the sign due to the preceeding or proceeding word or activity. Moreover, it is just human nature to behave differently when we know that we are recorded [127]. In daily settings, when one is not explicitly focusing on what one is saying, the participant may behave and sign differently. Circumstances may also dictate differences in signing with two hands versus one, if the person is signing while holding a cup of coffee, for example.

Finally, directed dataset collection is simply not scalable. The costs in terms of time to collect the data and money to compensate participants for their contributions are often too prohibitive to collect massive amounts of data. This has driven efforts to develop alternative means for acquiring sign language datasets. In 2021, Bragg, et al. [18] proposed using crowdsourcing to record videos with specific content to facilitate automatic labeling and perform quality control with experts to check for consistency. In another work, Bragg, et al. [15] also conducted a user study to explore the data quality that could be obtained by participants playing ASL Sea Battle, a variant of Battleship that uses ASL, and reported favorable user experiences and reliable collection of videos for 20 ASL signs. However, this

work utilized an approach similar to the "Wizard of Oz" [36] procedure. In particular, a researcher was required to interact with the technology alongside participants. Moreover, the quality of the data collected was only visually evaluated by experts - no investigation or demonstration of the data's utilization for model training or ASL recognition was conducted.

In contrast, this study proposes ChessSIGN[1], an interactive chess game autonomously controlled via video to collect natural ASL from both video and radar. To the best of our knowledge, this paper is the first to explore the *learning via interaction* of radar micro-Doppler signatures, considering both pre-deployment batch training and post-deployment model updates. First, Section 5.2 describes the design of the ChessSIGN game, interface design, model training, and real-time recognition accuracy. Next, Section 5.3 shows how directed datasets are not effective even for model pre-training of DNNs to classify natural articulations. Section 5.4 then describes several possible solutions to this challenge, including the use of physics-aware generative adversarial networks (PhGANs) for synthetic training data generation, style transfer and domain adaptation networks for leveraging directed data for model training, and post-deployment training strategies for improving recognition accuracy as an increasing amount of data is acquired. Beyond ASL recognition, the results of this paper provide insights into the real-world challenges in the development and deployment of effective ML models for classification of human RF signatures, and show that post-deployment interaction can be used to improve recognition of natural signing over time. Section 5.5 discusses key conclusions and plans for future work, including the use of ChessSIGN as an interface for evaluating real-time radar-based recognition algorithms and closed-loop sensing paradigms, such as cognitive radar.

[1]We refrained from utilizing ASL in the name to reflect the broader applicability of the proposed approach to all sign languages, not just ASL.

5.2 Interactive ASL-Enabled Chess Game

Chess is a popular strategy game that as drawn great interest from people of all age groups and backgrounds. It is well-suited for our proposed interactive data acquisition approach because it is a slow-paced game, which gives users enough time to decide and select their move. As such, it alleviates real-time processing constraints and allows enough time for locally saving and transferring the data, signal processing, and model prediction. Furthermore, chess is flexible enough to allow for the addition of other features for collecting more complex signing sequences and collecting user feedback using a small pop-up window. This can enable users to effectively self-annotate their data, minimizing subsequent quality control efforts.

When the data collection procedure is transformed into a gaming environment, several concerns emerge that do not exist in controlled experiments, such as designing the game in an enjoyable manner and ensuring that any overhead for self-annotation is not overwhelming or so intrusive that users get bored or frustrated with the interface. Additionally, it is important to minimize computational overhead due to data processing so as to avoid introducing delays in the game, which can then degrade a user's playing experience. Finally, predictions made by the game control model should be accurate enough so that users do not have to often undo their move, or feel like they are doing something wrong or are not skilled enough to play the game.

The proposed interactive ASL-enabled chess game is designed to acquire data from both an RGB camera and an FMCW radar simultaneously. In our initial pilot version, the game itself is controlled using predictions made using video data only. To minimize potential user frustration due to misclassifications, we took advantage of a publicly available video-based ASL dataset to train our initial game control model, as described in the next section.

WATER	YES	BOOK	SLEEP	CAR	HELLO
HOME	READ	TIME	BETTER	DRINK	TOMORROW
SEE	HOT	BED	WHY	WHERE	LIKE
PLEASE	HAVE	MORNING	FINE	GO	NIGHT
CAN	TABLE	THERE	FINISH	HATE	-

Table 5.1: ASL signs utilized in the chess game.

5.2.1 Video Dataset

We used Google's Isolated Sign Language Recognition (GISLR) dataset [31] to train our initial video-based control model. This dataset contains 250 of the first concepts/vocabulary based signs that are taught to infants in any language. Around 100k videos (∼400 samples per class) of isolated signs are articulated by 21 Deaf participants fluent in ASL. The corpus itself is a collection of hand and facial landmarks generated by MediaPipe Holistic pipeline. It integrates separate models for pose, face and hand components, each of which are optimized for their particular domain. This dataset is mainly used in the PopSign mobile game[2] to improve the ability of the game to help relatives of Deaf/Hard-of-Hearing(HoH) children learn basic signs and communicate better with their loved ones. In this work, a subset of 29 signs - the maximum number of different positions the most mobile piece, the Queen, can move - from the GISLR dataset is utilized to control the movement of game pieces.

Command sign selection was done based on the basis of several factors, including the selection of signs that were unique in their articulation (e.g. not signs that had many variants based on regional dialects), were more kinetic in nature (did not rely exclusively on shape for distinction), and which were one of the 100 signs acquired in prior studies [138] conducted with directed data collection using radar. This enabled comparison of our proposed interactive approach with conventional directed data and study of its implications for ML model training and recognition of natural ASL - the core contribution of this paper. A list of signs utilized is given in Table 5.1.

Figure 5.1: Screenshots of the ASL-enabled chess game.

(a) Highlighted possible moves with assigned words for a selected piece (left Bishop).

(b) UNDO button for correcting wrong predictions.

(c) Feedback GUI for correcting misclassifications.

5.2.2 Chess Interface Design and Game Play

The graphical user interface (GUI) of the game is comprised of a central region showing the chess board itself, captioned with an banner at the top to provide information and instructions to the user. The game begins by the user selecting the piece they wish to move by hovering over the piece with a mouse and clicking. This triggers the GUI to reveal to the user all of the possible moves for the selected piece by highlighting those squares on the chess board and displaying text for the English word that has the closest conceptual correspondence an ASL sign, randomly selected from among the 29 command signs. Note that there is no guarantee that the user will respond by articulating exactly the same sign as recorded in the training data, due to regional and cultural variations of ASL. In our initial selection of command signs, we aimed to select signs that had unique articulations and no significantly different variants. A screenshot showing the textual prompts for moving a chess piece using ASL is illustrated in Figure 5.1a.

Once the user decides to which position they want to move the piece, the user clicks the green "CLICK HERE" botton on the top right corner of the screen to trigger the data

[2]https://www.popsign.org/

recording of both the camera and radar sensors. A pop-up appears on the center of the screen and counts down from 3 to 1, after which the word "GO" is displayed to indicate to the user when to begin signing. Both sensors record the users signing for 3 seconds, after which the camera data is processed and input to the video-based model, described in Section 5.2.3, to recognize the user's articulation.

The prediction made is displayed to the user and the chess engine makes the move accordingly. An "UNDO" button is then displayed on the top right corner of the screen, as shown in Figure 5.1b. If the prediction is correct, the user selects another piece and continues to play. Otherwise, the user can click on the "UNDO" button to reverse the move. When a move is undone, the last move is reversed (i.e., the game goes back to the previous board state before the prediction), and it opens-up a small GUI to allow user to select the actual word they signed from a drop-down menu, as shown in Figure 5.1c, and enables correct labeling of the recorded signs as ground truth. The game history, along with a record of the incorrectly predicted samples, is logged into a file to allow further offline analysis of the data. The recorded data samples are transferred to a local hard drive and backed-up to a cloud platform automatically after each recording for storage safety purposes.

The interactive ASL-enabled chess game essentially inherits all the features and preserves the rules of a regular chess game. The main difference from a regular chess game comes from the way it is being played from a user point of view. Instead of clicking on the position users want to move their piece on, they use ASL signs to give the move command to the game. The game, on the other hand, operates sensors, collects user's data and runs the prediction model and the chess engine in the backend. Such operating capability eliminates the need for an operator during the game play and the data collection process and the need for an annotator to label the acquired dataset.

5.2.3 Video Prediction Model

The GISLR dataset was first introduced in an online hackathon organized by Google on Kaggle. The first place was achieved with a network composed of a 1D-CNN and

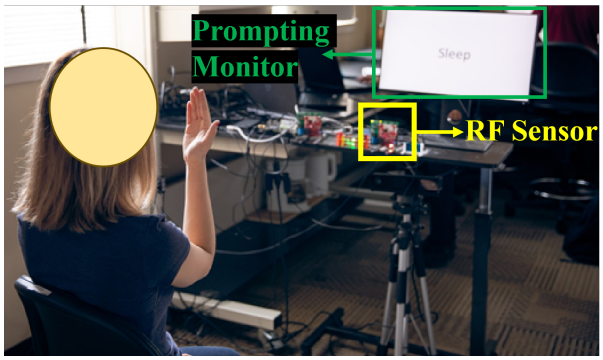
a Transformer subnetworks by Hoyeol Sohn [157]. Combining CNNs with Transformers is a prevalent idea applied in different networks such as CoAtNet [37], Conformer [62], MaxViT [168] and Next-ViT [111]. In the proposed solution, the Transformer is applied with batch normalization and Swish activation function instead of typical layer normalization and Gaussian Error Linear Unit (GELU) activation since the former yielded slightly lower inference time with the same accuracy than the latter one. The proposed model has around 1.85M trainable parameters. Handling variable-length input was achieved with padding and truncation. This approach provided sufficient inference speed and enabled the use of reasonably large models. In this work, we modified the network for 29 output classes and re-trained the model.

The final model consists of 12 convolutional and 4 Transformer blocks stacked in 3+1 fashion for four times. After global average pooling (GAP) for flattening, a multi-layer perceptron (MLP) with *Softmax* activation function is applied for classification. Drop-path [108] - a high rate of dropout (p=0.8) - and Adversarial Weight Perturbation (AWP) [179] is applied for regularization. These methods were very crucial to prevent overfitting when training for long epochs ($\geq$ 300), and removing any one of them resulted in significant performance drops. To improve generalization, temporal and spatial augmentation techniques were applied to the training data, such as random re-sampling (0.5$\times$ - 1.5$\times$ the original length), random masking, horizontal flip of the skeleton, random affine transformations (scale, shift, rotate and shear) and random cutout.

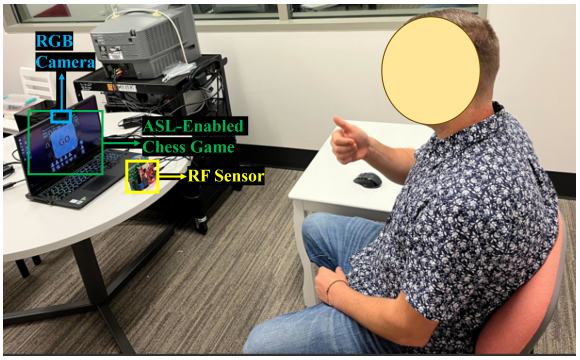
5.2.4 ASL Datasets Acquired

Both directed and interactive data was acquired during this study. Directed data is acquired via a controlled experiment in which the users is specifically directed to articulate a particular sign. Interactive data is acquired via the proposed ASL-enabled chess game, and the data is acquired in free form during game play with limited instructions and no external intervention. IRB approval was obtained prior to the study, and data was collected with informed consent from each participant.

Figure 5.2: Dataset acquisition environments.



(a) Directed RF Dataset Acquisition.



(b) Natural Multi-Modal Dataset Acquisition.

Directed RF ASL Data

This initial dataset is acquired under controlled experimental settings in a laboratory environment. The RF sensor was placed around 1.5m away from the participants and 0.91m elevated from the ground. Participants were seated on a chair directly facing towards a monitor that was placed behind the RF sensor. The monitor is used to prompt the words to be articulated. This ensures the dialectal consistency across participants by displaying a specific articulation of each sign. The experimental setup for the directed RF ASL dataset acquisition is demonstrated in Figure 5.2a. Since there exists different ways of articulating a sign, the signing videos were also displayed to the participants in order to have a consistent way of signing a word across participants.

Data was acquired in 2022 from 19 participants at Gallaudet University, the only university in the U.S. for Deaf/HoH students where ASL is used as the primary language of instruction, and 4 participants at the Lab for Computational Intelligence in Radar (CI4R) at the University of Alabama. Of these participants, twenty-one were Deaf, while two were Child-of-Deaf Adults (CODAs) fluent in ASL. All experiments were conducted using the same RF settings and operators. A total of 110 signs were acquired, based on selection from the ASL-Lex Database [23] including nouns, verbs and adjectives based on their usage frequency and

kinematic variance. A total of 4,455 samples were acquired for 110 signs which corresponds to around 40 samples per class.

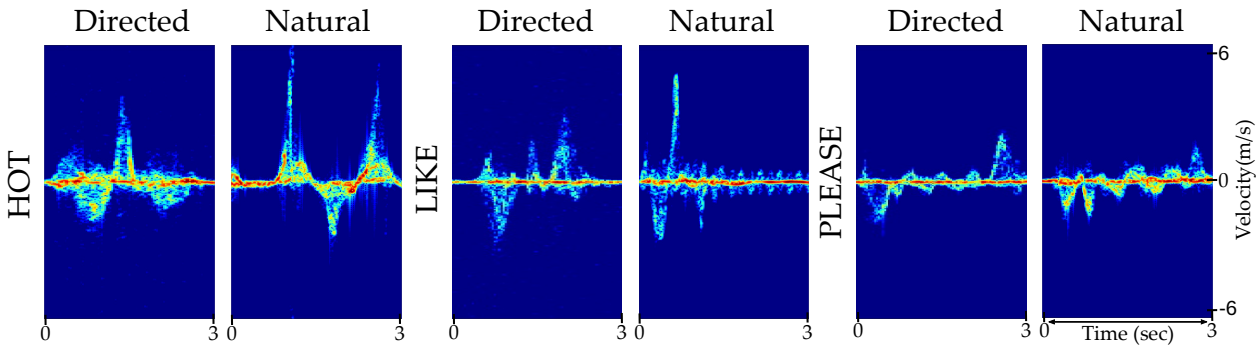
Interactive Multi-Modal ASL Dataset Acquired

This dataset is acquired during the course of interactive ASL-enabled chess game play. An integrated RGB camera is used for recording the videos, while an RF sensor (the same sensor as used in the directed ASL dataset) was placed in front of a laptop that runs the game as shown in Figure 5.2b. This data was acquired in 2023 from 23 Deaf participants at Gallaudet University. Note that the people who participated in the 2022 Directed ASL data collection are not the same individuals as those who played ChessSIGN to provide interactively-acquired data. To assess the difference between the directed and the natural signing, 29 common words between the directed dataset and the GISLR are selected as listed in Table 5.1. The dataset is acquired in a synchronized, multi-modal fashion and in total 1,078 samples were collected for 29 signs ($\sim$37 samples per class). Since the chess game is a slow-paced game, total number of samples acquired during the experiment is relatively lower when compared to the directed case. However, number of samples per class is comparable to that of the directed case.

5.3 Directed versus Natural ASL Data

In this section, we examine the differences between RF ASL signatures acquired through natural interactions versus that of the conventional, directed (controlled) experimental approach. In particular, we first show qualitatively through observation of the RF signatures, the various ways in which directed experiments fail to capture the nuances of natural signing. Then, we show the detrimental impact of using directed ASL data in model training for sign language recognition.

Figure 5.3: μD signatures of directed and natural ASL samples for the signs HOT (left), LIKE (center) and PLEASE (right).



5.3.1 Comparison of μD Signatures

Consider the pairs of directed versus natural μD signatures shown in Figure 5.3 for the signs HOT, LIKE, and PLEASE. For the same sign, the natural signature exhibit significantly greater variance. These variations are not just slight variations in the spatio-temporal artifacts of the signature, but can be significant differences in the shape, speed (μD bandwidth), number and bandwidth of repetitive features and strength of the signatures.

For example, based on viewing video recordings of the articulation corresponding to the μD signatures, it may be observed that for the word HOT, the signer articulating the word in a natural, interactive setting repeats the sign two times, hence two positive peaks can be observed in the μD signature while there is only one repetition and one positive peak in the directed samples. For the word LIKE, in an unconstrained, interactive setting, the signer shakes her hand after finishing the sign. This causes some jittering effect at the lower frequencies of the μD spectrogram. For the word PLEASE, the signer moves her hand towards her chest in two steps instead of one which causes two consecutive negative peaks in the μD spectrogram. Also, the negative and the positive peaks at the beginning and at the ending of the sign when the arms are being moved towards and away from the chest are not as sharp as in the directed case.

Figure 5.4: Maximum and minimum velocity distributions of directed and natural ASL signing.

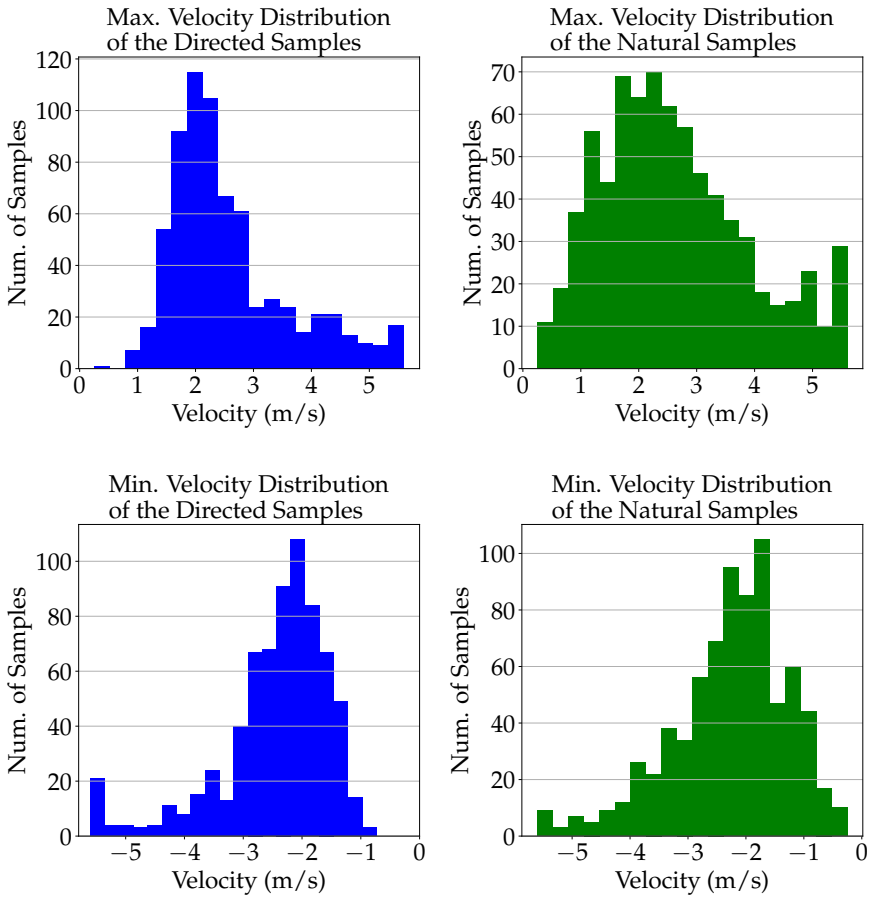


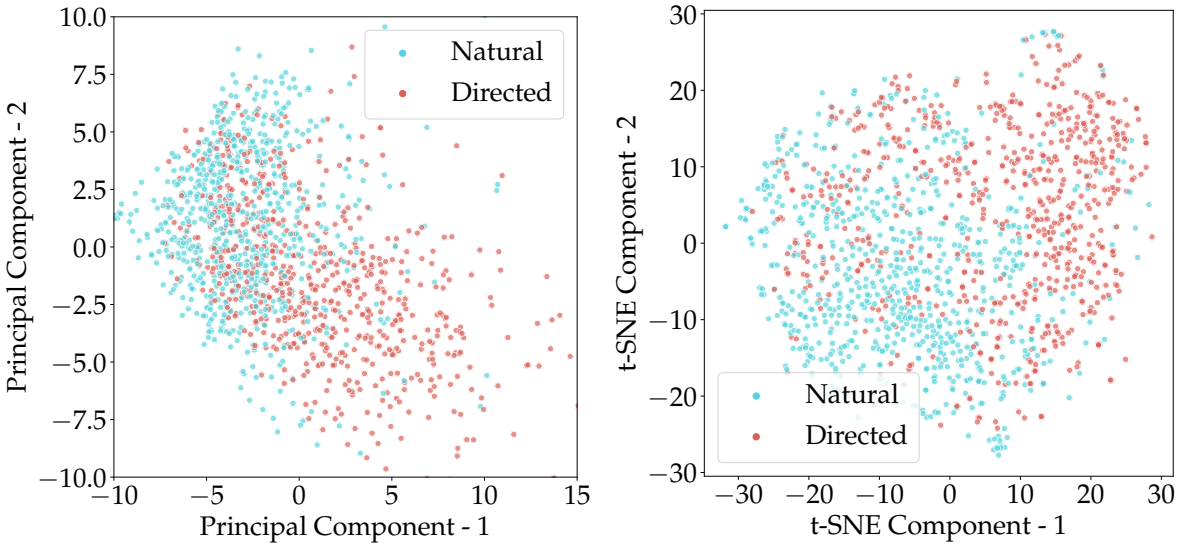
Table 5.2: Statistical comparison of speed in directed and natural.

Data Type	Avg. Max. Velocity	Var. of Max. Velocity	Avg. Min. Velocity	Var. of Min. Velocity
Directed	2.58 m/s	1.06 $(m/s)^2$	-2.46 m/s	0.89 $(m/s)^2$
Natural	2.61 m/s	1.59 $(m/s)^2$	-2.28 m/s	1.03 $(m/s)^2$

5.3.2 Comparison of Velocity and Feature Distribution

Figure 5.4 shows maximum and minimum velocity distributions of directed and natural ASL signing. From the histograms, it can be observed that although directed and natural samples have close mean maximum and minimum velocity values, variance of natural samples are much larger as the histograms bins are more evenly distributed. Table 5.2 summarizes the average and variance values of maximum and minimum velocities of directed and natural ASL samples. It can be seen that while the variance of maximum velocity of the directed

Figure 5.5: Data distribution difference exploration of directed and natural ASL samples via dimension reduction techniques.



(a) PCA analysis of directed and natural ASL samples.

(b) t-SNE analysis of directed and natural ASL samples.

samples are 1.06 $(m/s)^2$, it is 1.59 $(m/s)^2$ for the natural signing samples. Similarly, for the variance of minimum velocity, while it is 0.89 $(m/s)^2$ for directed samples, it is 1.03 $(m/s)^2$ for the natural case. This quantifies the statistical difference between the two datasets.

The impact on the statistical distribution of features can be visualized by utilizing the principal component analysis (PCA) [5] and t-SNE [171] dimension reduction techniques. Figure 5.5a and 5.5b present the 2-dimensional PCA and t-SNE maps for directed and natural ASL samples. From these visualizations, it may be observed that the directed ASL samples exhibit - as expected - some overlap with the natural signing samples. However, there are significant regions overwhich the two distributions do not overlap, indicating that for many samples, the directed data is ***not*** representative of natural signing.

5.3.3 Impact on Model Training

The significance of the difference between the distributions is underscored when its impact on model training is examined. In particular, the distributions of direct data signatures versus natural signing is so significant that models trained with data collected in a directed

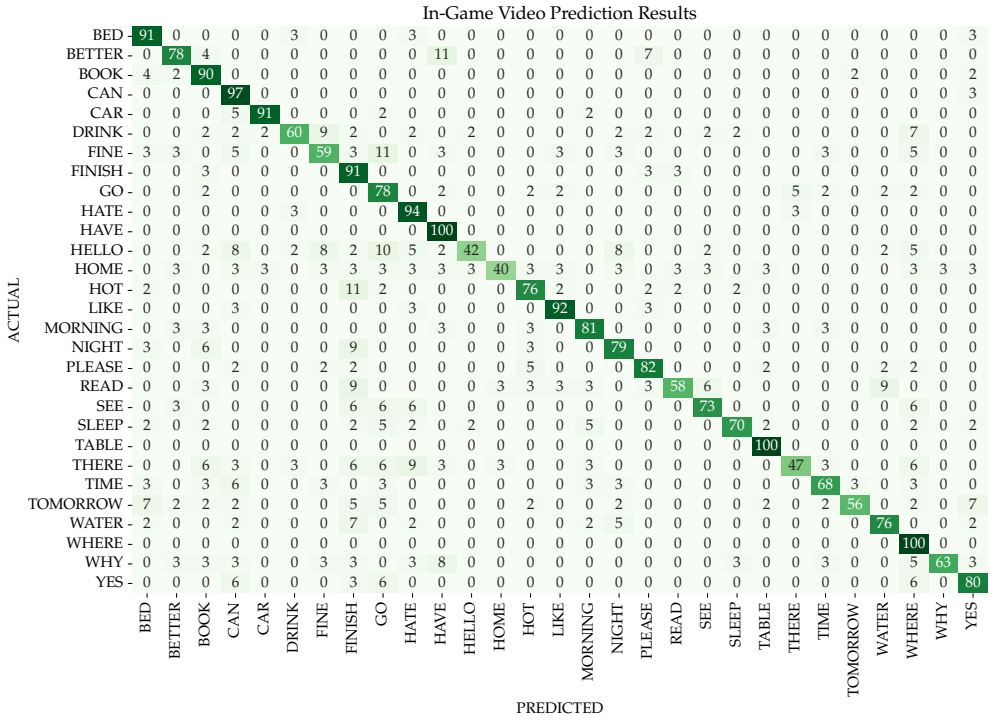
fashion are entirely unable to recognize natural signing at all - a result seen not just with radar data, but with video data as well.

First, it should be noted that there are actually two different ways of evaluating model performance: computation of the "in-game accuracy" and evaluation of 29-sign model accuracy once the game has concluded based on recorded data. Although the complete model is trained for 29 signs, during a game, the selected piece will only be able to move to a much lower number of possible positions. This reduces the classification problem to one of just recognizing the signs for the possible positions. For example, a Pawn can move one square forward if unobstructed (or two on the first move), or one square diagonally forward when making a capture. This results in three possible positions. Or a completely unobstructed Knight may only move to eight different positions. In contrast, once the game is completed, all the data acquired for all 29 command signs can be utilized as test data for the model, resulting in a true assessment of the ability of the interactive data to train a 29-class model. Due to the few number of classes encountered during the game, the in-game accuracy of a model is typically higher than that of the true 29-class accuracy.

Video-Based Model Accuracy

When the video-based model is trained and tested with directed data from the GISLR dataset, an accuracy of 92.3% was obtained for the 29 signs selected to control the movement of chess pieces during the game. However, during the actual chess game, this model performed significantly worse, achieving a 76.62% in-game classification accuracy. The confusion matrix for in-game predictions are shown in Figure 5.6. A number of words appear to be consistently confused over 10% of the time: HOT with FINISH, FINE and HELLO with GO, and BETTER with HAVE. All of these signs are more kinetic in nature, which may be one reason for higher misclassification by video: video tends to be more effective in characterizing spatial variance, rather than temporal variance (a weakness remedied by radar, which is effective in recognizing signing dynamics - not shapes).

Figure 5.6: Confusion matrix of the video-based prediction model with in-game restrictions. (All the values are in terms of percentages.)



If we evaluate the efficacy of the GISLR-trained model to classify 29 signs outside the constraints of the game, however, a natural signing classification accuracy of just 48.24% is attained. This result demonstrates the inherent limitations of existing sign language video databases, which utilize directed data collection and hence cannot capture the natural features of sign language, which emerge in the interactive game environment.

Radar-Based Model Accuracy

A similar effect is observed in radar data as well. First, let us consider the baseline training scenario of training the RF model with directed data, but also testing the model on directed data. The directed RF dataset used in this work has significantly lower number of samples per class when compared to the GISLR video dataset (i.e., 40 vs 400 samples per class). With only real data itself used during training, a classification accuracy of 68.9% is obtained using a 4-layer CNN comparised of 2D convolutional blocks followed by max-pooling layers

and two fully connected layers for classification. The Adam optimizer and cross-entropy loss function are used in the training phase.

This performance can be boosted, however, through the use of synthetic data generation to increase the size of the training sample support. In prior work, physics-aware generative adversarial network (PhGAN) [137] was shown to be effective in improving the classification accuracy of both sign language and, more broadly, human activity datasets. The PhGAN model improves the kinematic fidelity of synthetically generated RF signatures by adding another branch to the discriminator that takes as input not just the μD signature, but its envelope as well. The envelope represents the maximum velocity incurred during the articulation of a certain movement and as such represents an important physical bound on the resulting signatures. Moreover, a physics-based loss was also added into the cost function of the network to quantify how effectively the envelopes of the synthetic versus real signatures matched. The utilization of these two innovations was shown to result in the fewest kinematic errors as comparison to alternative GAN [57] architectures. In this work, a dual-branch PhGAN network was utilized to generate an additional 500 synthetic samples per class for the directed RF dataset. When trained using PhGAN-synthesized signatures, the same 4-layer CNN yielded 100% recognition accuracy on directed radar-based ASL data.

Inasmuch as this is a great result when training and testing on directed ASL data, this model completely fails to recognize any natural ASL samples: only a 9.56% accuracy is obtained when testing on the natural ASL samples acquired via the interactive chess game. This result is substantially worse than that obtained from video, which exhibited a 44% performance drop. Here, the radar-based model exhibits a 90% drop in performance! One possible reason for radar being more effected could be that one of the major ways in which directed ASL data differs from that of natural ASL is that the kinematics - temporal progression and revelation of co-articulation - is much different even though the spatial component of the signal is still similar. As radar is much more sensitive than video to

Table 5.3: Performance comparison of different training methods and datasets. (Note that no natural signing data are used in the training phase of Exp. 5).

Exp. ID	Training Data	Testing Data	Modality	Model	Acc. %
1	GISLR	GISLR	Video	1D-CNN + Transformer	92.3
2	GISLR	Natural ASL	Video	1D-CNN + Transformer	48.2
3	Directed ASL	Directed ASL	RF	2D-CNN + MLP	68.9
4	PhGAN(Dir. ASL)	Directed ASL	RF	2D-CNN + MLP	100
5	PhGAN(Dir. ASL)	Natural ASL	RF	2D-CNN + MLP	9.6

kinematics rather than hand shape or spatial variables, its performance is more negatively effected.

5.4 Interactive Learning of Natural ASL

The results of the prior section clearly show the challenge of recognizing natural ASL and validate the necessity of the proposed interactive ASL-enabled chess game for capturing and learning from natural ASL. But the question then remains of how to best train an RF system prior to deployment so that the prediction accuracy improves as we get an increasing amount of natural ASL data: *interactive learning in-situ*.

5.4.1 Fine-Tuning Model Pre-Trained with Directed ASL

The results given in Table 5.3 do not utilize any natural signing data during the training phase. However, after the ChessSIGN game is deployed, we will be acquiring an increasing number of natural ASL samples that can then be leveraged to fine-tune models initially trained using 1) real RF samples acquired in a directed fashion, or 2) synthetic RF samples generated from directed ASL data. Synthetic RF data generation using GANs has been shown to be an effective method for increasing the sample support during model training, especially when the availability of real data is limited.

Figure 5.7: Accuracy of 4-layer CNN pre-trained with Directed and PhGAN-Directed ASL data, and fine-tuned with natural ASL data.

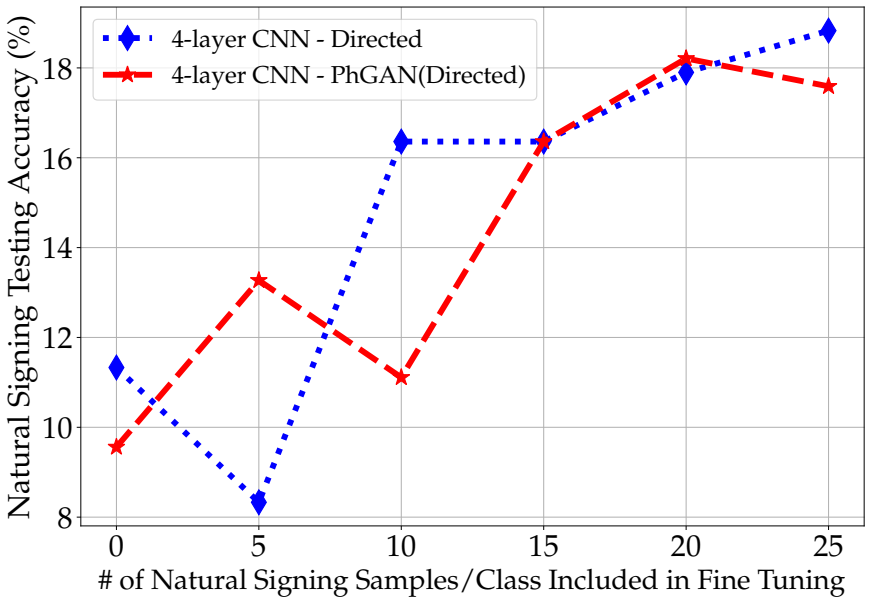
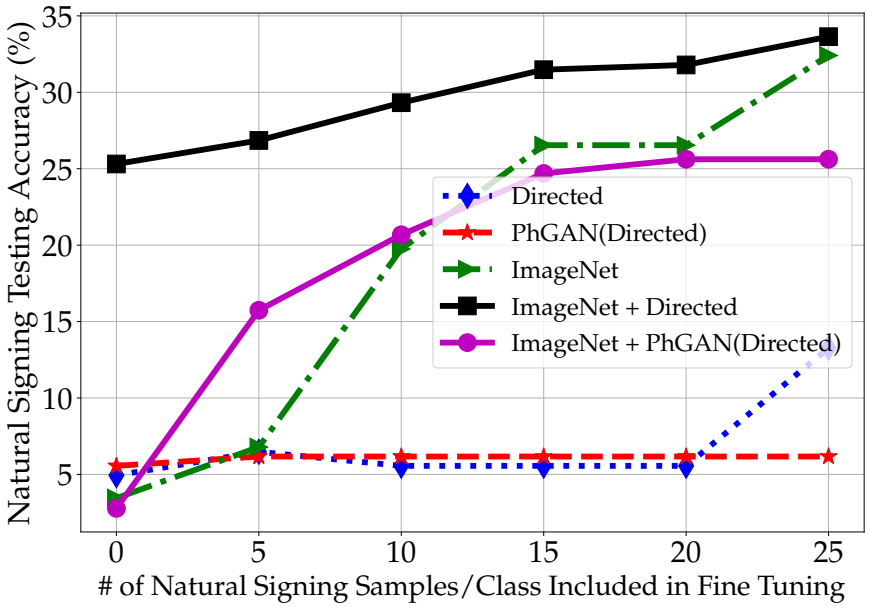


Figure 5.7 shows how utilizing varying amounts of natural signing data for fine-tuning the model improves prediction accuracy. Note that the 4-layer CNN has the identical architecture and hyperparameters as used in Exp. 4 in Table 5.3. It should be noted that while the use of PhGAN-synthesized samples for training offered tremendous performance gains when testing on directed ASL data (accuracy increases from 68.9% to 100%), when testing on natural ASL data, as shown in Figure 5.7, increasing the amount of training data using PhGAN synthesis does not offer an performance benefits. This is because there is a fundamental difference in the distributions of directed versus natural ASL data, and GAN-based synthesis does not bridge this gap - it only generates more samples from the same distribution. As directed ASL does not accurate capture the articulation of signs that occurs during natural signing, the fundamental differences in kinematics severely limits the efficacy of directed data to inform model training of networks intended for classification of natural ASL.

Ultimately, while fine-tuning with natural ASL samples collected *in-situ* increases the testing accuracy, this increase is not sufficient to train a viable model as accuracy remains under 20%.

Figure 5.8: Accuracy of VGG-16 pre-trained with Directed/PhGAN-Directed data only versus initialization with ImageNet.



5.4.2 Fine-Tuning Model Pre-Trained with ImageNet

One potential way to improve the efficacy of pre-training with RF datasets is to initialize the network with optical imagery from a large database, such as ImageNet [40], a database of 1.5 million RGB images. While such pre-trained network is not initially going to be familiar with the spatial features of the RF data except certain primitive image features such as edges and corners, directed dataset can be utilized to introduce the spatial features of RF data to the network. In this section, we examine the impact of using a two-step pre-training process for the 16-layer CNN architecture of VGG-16 [155] by 1) utilize the stored VGG-16 weights obtained from training with ImageNet, and 2) training VGG-16 again using Directed or PhGAN-synthesized Directed ASL data.

Figure 5.8 shows the resulting accuracy as we fine-tune the network with increasing amounts of natural ASL samples. First, it may be observed that using Directed or PhGAN-Directed samples for training VGG-16 results in the worst performance - baseline results consistent with that seen in Figure 5.7. If we utilize the two-step training process with ImageNet-based initialization, we can see a significant performance improvement by as much as 25% when

25 samples per class natural ASL is used in fine-tuning. The best performance is attained when both ImageNet weights and training with Directed ASL samples are used. However, notice that this is only slightly better than the result obtained if would have just fine-tuned from ImageNet only initialization. Using a second round of training with Directed ASL offers greater initial performance when fewer samples of natural ASL are used in fine tuning; however, once we have at least 25 samples per class natural ASL, utilizing a second round of training with Directed ASL does not offer much benefit.

This result is significant because essentially it is showing that when an interactive learning paradigm is utilized, we are better off just initializing with ImageNet, and learning as we go.

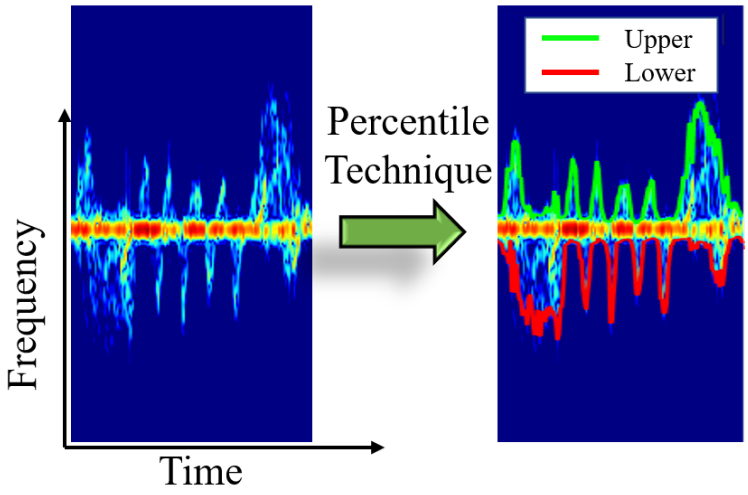
5.4.3 Domain Adaptation of Directed to Natural ASL

Rahman et al. [138] showed how the degree of fluency of the ASL user effected classification accuracy. For example, an often encountered practice is the use of hearing participants to imitate actual sign language based on videos showing ASL articulations - also known as "imitation signing". It is shown that discrepancies in the fluency level of signers in training and test data resulted in significantly degraded classification accuracy, but that domain adaptation techniques could be use to bridge the gap.

In this study, we consider the efficacy of utilizing domain adaptation techniques to bridge the gap between the distributions of directed versus natural ASL data. In particular, rather than directly using the natural ASL samples for fine-tuning, we consider two alternative ways of exploiting the interactively acquired natural ASL samples: 1) utilization for training a PhGAN to general additional synthetic samples from the distribution of natural ASL, and 2) utilization for training a domain adaptation network to learn the mapping from the directed data distribution to the natural ASL distribution. In particular, we consider two domain adaptation networks: CycleGAN [191] and Pix2Pix [88] (abbreviated as P2P in this study).

CycleGAN is a network that aims to learn a mapping from directed (D) to the natural (N) ASL domain, $G : D \rightarrow N$ such that the distribution of μD spectrograms from $G(D)$ is indistinguishable from the distribution N using an adversarial loss. This mapping is coupled

Figure 5.9: Upper and lower envelope extraction.

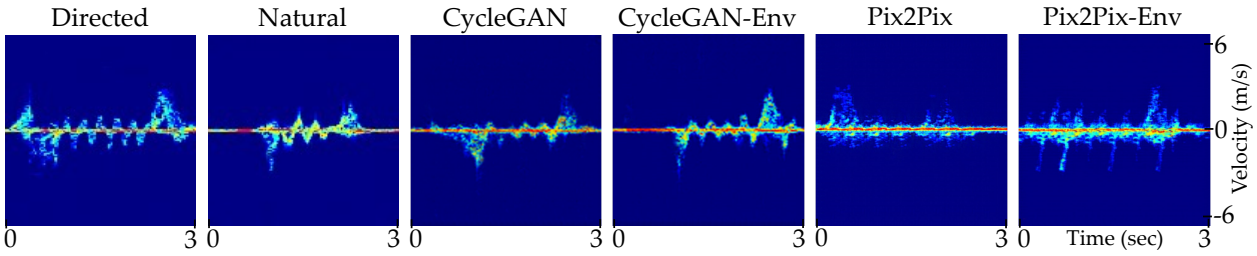


with an inverse mapping $F : N \rightarrow D$, and a cycle consistency loss is introduced to push $F(G(D)) \approx D$ (and vice versa). One advantage of CycleGAN is that it can generate more synthetic samples than the number of images provided during training.

Pix2Pix, on the other hand, is an image-to-image translation method that utilizes conditional adversarial networks. In addition to learning the mapping from input image to output image, this network also learns a loss function to train this mapping. This enables Pix2Pix to apply a consistent process across all datasets without having to explicitly modify the loss function for each case. However, Pix2Pix requires matching input-output image pairs during its training process. This is in contrast to CycleGAN, which can be trained on unpaired samples drawn from each distribution. Because the amount of real data available is limited, we use PhGAN to generate a greater number of synthetic-Directed and synthetic-Natural samples. We then create input-output pairs by matching samples from the same class to train Pix2Pix.

To improve the kinematic fidelity of the synthetic data generated by the two networks, in addition to vanilla CycleGAN and Pix2Pix networks, modified versions that utilizes a physics-based loss term based on consistency of the upper and lower envelopes is also developed. The modified versions, CycleGAN-Env and Pix2Pix-Env, extract the upper and lower envelopes of the μD signatures using the percentile method [44]. Figure 5.9 shows the results of the envelope extraction for a sample μD signature. The mean-squared error

Figure 5.10: μD signatures of the PhGAN-generated directed and natural samples, and benchmarking of transformed samples generated by CycleGAN, CycleGAN-Env, Pix2Pix and Pix2Pix-Env models.



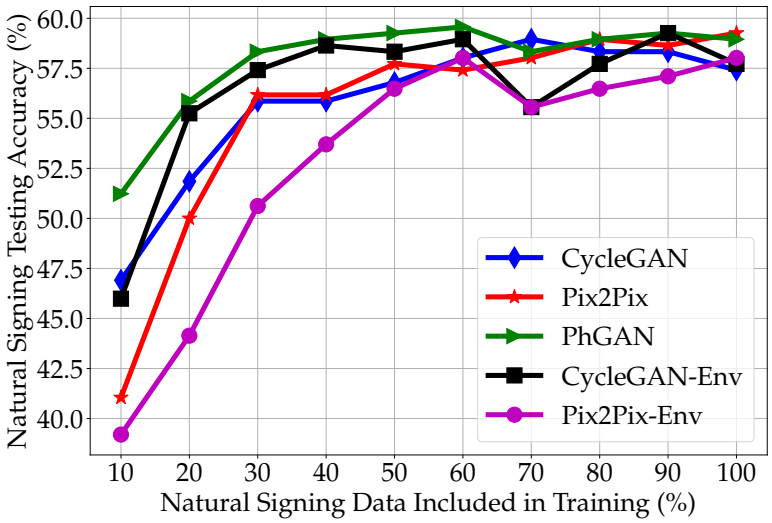
(MSE) between the envelopes of the generated and target signatures is computed as the physics-aware loss. Thus, the total loss of the generator, $\mathcal{L}_{\text{GAN}}$, is computed as

$$\mathcal{L}_{\text{GAN}}(G, D_N, D, N) = \mathbb{E}_n[\log(D_N(n))] + \mathbb{E}_d[\log(1 - D_N(G(d)))] + \lambda \mathcal{L}_{\text{env}}, \qquad (5.1)$$

where G is the generator, D_N is the discriminator for natural domain, λ is the weighting factor and $\mathcal{L}_{\text{env}}$ is the MSE between generated and target envelopes. While the first two terms represent the discriminator and the generator losses, respectively, the last term computes the error rate between generated and target envelopes. λ is empirically selected to be 0.001 to balance multiple loss terms in Equation 5.1.

Figure 5.10 shows examples for visual comparison of PhGAN-generated directed and natural μD signatures, and the transformed samples generated by Cycle-GAN, CycleGAN-Env, Pix2Pix and Pix2Pix-Env methods. It may be observed that while initial and final peaks are well represented by both CycleGAN and CycleGAN-Env, peaks in the center of the signature are not replicated effectively by the vanilla CycleGAN model. The signal power of the CycleGAN-Env sample is also noticeably higher than that of CycleGAN. A similar phenomenon can be observed in the Pix2Pix model. While the vanilla model is performing very poorly and incapable of reconstructing the peaks in the μD signatures, Pix2Pix-Env can replicate periodic peaks with high signal power. However, much of the detail of the signature is lost in the synthetic Pix2Pix-Env samples. While our goal does not necessarily

Figure 5.11: Accuracy of 4-layer CNN fine-tuned with synthetic samples generated from natural ASL or adapted from directed ASL data.



require perfect emulation of natural ASL signatures, we did observe that this loss of detail does degrade the ability of Pix2Pix-Env samples to adequately train classifiers of natural signing, as discussed more in the next section.

5.4.4 Fine-Tuning Model with Synthetic Natural ASL

Synthetic data generated from a small amount of natural ASL or adapted from directed ASL data can be used to augment and improve the training of networks for recognition of natural ASL. Figure 5.11 shows the improvement of the classification accuracy of a 4-layer CNN when incrementally fine-tuned with 500 synthetic samples per class that are generated from 25 natural ASL samples/class (70% of data). Note that 30% of the natural signing data is always preserved for testing.

It may be observed that the best performance is attained when a PhGAN is used to synthesize additional samples from the natural ASL data itself, irrespective of the amount of natural ASL samples available for training. The domain adaptation method that yields the most comparable results - when a larger amount of natural ASL data is available - is

Table 5.4: Final classification results of VGG-16 for RF data of natural ASL.

Method	PhGAN	CycleGAN	CycleGAN with Env.	P2P	P2P-Env
Acc. (%)	**69.14**	62.04	62.35	61.73	60.49

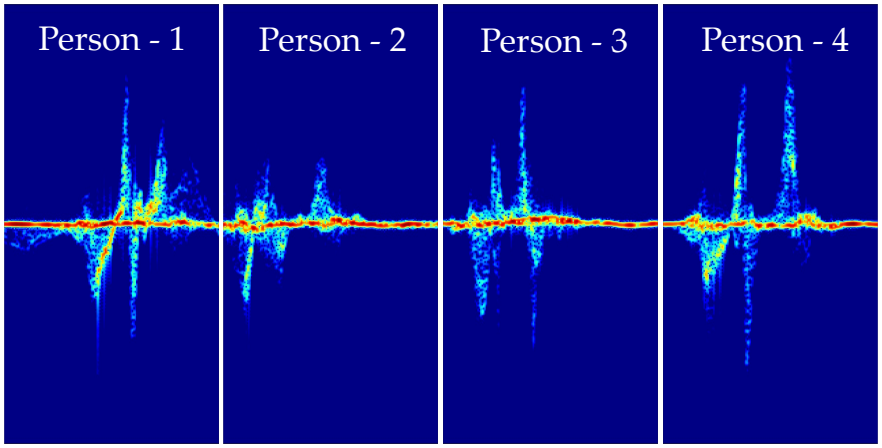
Pix2Pix. However, with only a small about of natural ASL data, CycleGAN-Env provides the best results.

As more data are acquired, all models begin to level off at about 58% accuracy. We believe this is because all 25 samples/class is used to train the PhGAN model that synthesizes the augmented directed and natural ASL data required to train the domain adaptation methods. If even more natural ASL data were acquired interactively, the sample support for the domain adaptation step would also be increased, and consequently increase the fidelity of synthesized samples.

Nevertheless, an important benefit of synthetic data generation, however, is that also a deeper network can be trained to further improve classification accuracy. For each of the different synthesis approaches - PhGAN synthesis from natural ASL, CycleGAN/CycleGAN-Env/Pix2Pix/I adaptation from directed ASL - the VGG-16 model is trained and used to compute the final achieved classification accuracy for the interactively-acquired, natural ASL dataset. Table 5.4 presents the accuracy achieved 70% of the training data (25 samples/class) are used to train the network. It may be observed that utilizing PhGAN to synthesize samples from natural ASL itself provides the best accuracy of 69.1%, while CycleGAN and Pix2Pix-based methods yield between 60-62% accuracy. In general, utilizing training data synthesized via domain adaptation from directed ASL underperformed that of simply synthesizing from natural ASL itself.

Thus, even in combination with domain adaptation, the utilization of directed ASL samples in the training process does not offer tangible benefits that would render worthwhile the time, cost and effort involved with the acquisition of directed ASL data, even if from

Figure 5.12: μD signatures of different participants for the sign FINISH.



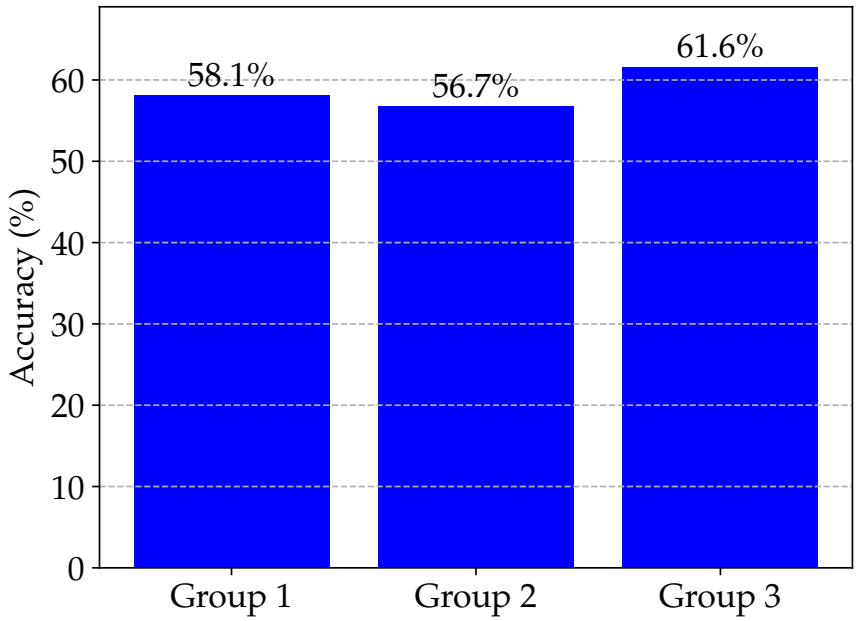
participants who are Deaf and fluent in ASL. In all cases, we achieve better results by simply learning interactively and *in-situ* from natural signing only.

5.4.5 Generalization Across Participants

It is critical for user interfaces to be effective regardless of the user or study participant. This is especially significant for interfaces reliant on sign language recognition as different participants may articulate ASL differently based on regional and cultural differences. For example, consider the variation in sign articulation for four different participants, illustrated in Figure 5.12 for the μD signatures of the sign FINISH. While significant differences in peak velocities may be observed by comparing Person 1 and 2's samples, the Person 3's second positive stroke (i.e., arm motion) is faster than the initial positive stroke; however, this situation is reverse for Person 1, whose second positive peak is much wider - indicating a longer duration motion with less acceleration - than that of Person 3. Person 4, on the other hand, has the longest sign articulation duration by spanning over 2.5 s interval with higher positive and negative peaks, which indicates greater peak speed.

Due to the individualized differences in articulation across participants, in this section, we evaluate the robustness of the proposed system across different group of participants using the leave-one-group-out (LOGO) method: the data of certain group of participants are used for training the model while the remaining participants' data are used for testing only.

Figure 5.13: Recognition performance of different participant groups when leaving-one-group-out for testing.



LOGO cross-validation is repeated for three participant groups, where in each repetition, the data of randomly selected seven participants used for testing and the remaining data of 16 participants are used for training the recognition model. Figure 5.13 presents the recognition performance across different participant groups. It may be observed that the random selection of different participants does result, as expected, in a variation of performance based on the participants, but this variation is only +/- 2.45%. However, the average LOGO cross-validation accuracy is about 10% lower than data for all participants are utilized in training. As data from more and more participants is acquired, we expect this discrepancy to become increasingly smaller. This result corroborates the conventional wisdom that building a very diverse dataset for training effective models is crucial to the performance of real-world systems.

5.4.6 Discussion

Our results show that there is a significant difference in the RF and video recordings of sign articulations that are acquired under a controlled setting (a.k.a. directed data

collection format) versus the natural articulations acquired via the proposed interactive gaming environment. While conventional wisdom may lead us to believe that despite these differences there is still value in acquiring directed ASL data for model pre-training, in fact our results show that this is not the case. Pre-training with ImageNet and Directed ASL yields the best performance irrespective of the amount of natural ASL acquired. However, once 25 samples/class of natural ASL is available, there is no significant difference between pre-training on ImageNet only, versus pre-training on ImageNet and Directed ASL.

It is important to note that video data also exhibits significant performance degradation due to the difference between directed and natural ASL articulations. Fine-tuning the video-based GISLR model with 70% of our acquired natural ASL data improves the prediction accuracy from 48.2% to 88.1%. Note that direct comparison of this result with the radar-based accuracy of 69% achieved via our proposed approach is not a fair comparison of the sensor modalities, as the GISLR model is pre-trained with an enormous amount of ASL data acquired by Google. However, both modalities exhibit massive performance gains when data acquired via the interactive ChessSIGN is utilized to fine-tune models for recognition of natural ASL. We do not view either the 88% video-based accuracy or the 69% radar-based accuracy as the ultimate achievable classification performance for the 29 ASL signs considered in this work as these accuracies will further increase as the interactive ChessSIGN game is continually played and the increasing amount of data is used to further improve model training.

In fact, the proposed interactive game ChessSIGN can be used to expand the dictionary to as many words as desired, since the words used for moving pieces during gameplay are randomly selected among a list of words.

5.5 Conclusion

This work proposes an interactive gaming environment, ChessSIGN, as a new way of acquiring video and radar recordings of natural sign language acquired in an unconstrained,

real-world setting. We show that the conventional way of collecting human RF signatures via directed experiments results in data that does not reflect how sign language is typically articulated in natural settings. The differences in movement result in a shift in distribution if directed data is used to train models for classification of natural ASL. This difference can be observed in both video-based models as well as radar-based models, which are more severely impacted due to its exclusive reliance on kinematic features, rather than spatial features, for ASL recognition.

In particular, radar-based ASL recognition performance for a 29-sign dataset is shown to drop from 100% to just 9% when natural ASL is used as test data rather than directed data. Several possible ways to exploit directed data for data synthesis via generative learning and domain adaptation are explored, but we show that such methods cannot overcome the differences in sign articulation due to participants being directed on when/what to articulate. Ultimately, our work shows that network initialization using transfer learning from ImageNet is sufficient to enable *learning via interaction* with the ChessSIGN game. As an increased amount of natural ASL data are acquired, we show the performance gains of augmenting natural ASL data using a physics-aware generative adversarial network (PhGAN). Fine tuning of the RF model with PhGAN-augmented natural samples yields promising results even when a small amount of data (around 25 samples per class) are acquired. In this way, we achieve a classification accuracy of 69% for a 29-sign natural ASL dataset acquired using the ChessSIGN game.

In future work, we aim to expand the concept of gaming-enabled interaction to the domain of embodiment games coupled with virtual reality to naturally engage participants in a wider range of natural movements and daily activities. We believe that the proposed interactive gaming approach can evolve into a valuable interface for evaluating real-time radar-based recognition algorithms. Moreover, as the software-defined radar systems, such as that used in ChessSIGN, can be controlled via command-line, the proposed interactive

learning framework can contribute to the testing and evaluation (T&E) of closed-loop sensing paradigms, such as cognitive radar.

CHAPTER 6

HUMAN-AWARE FULLY-ADAPTIVE RF SENSING

6.1 Introduction

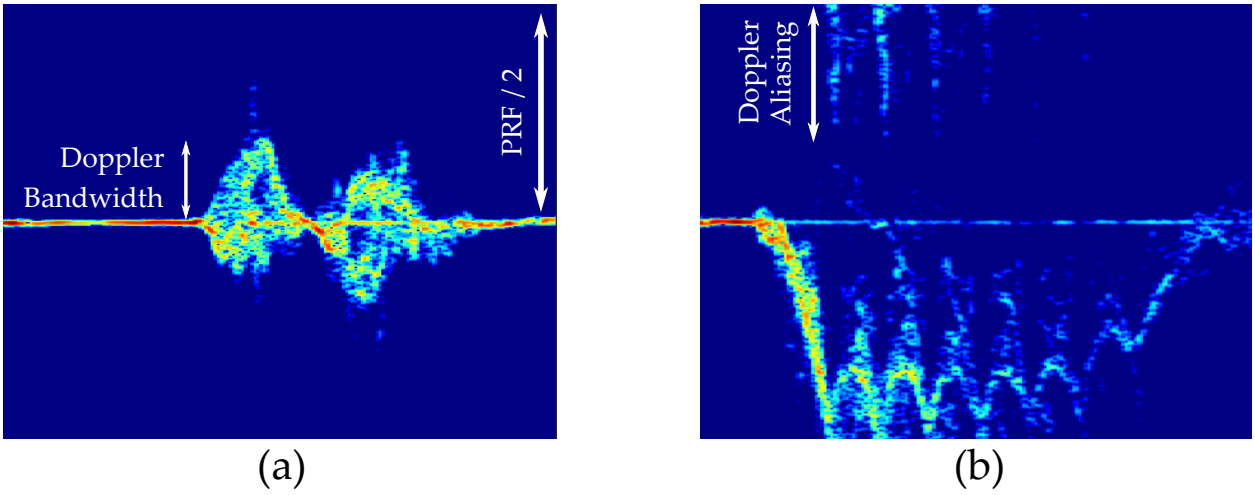
Selection of RF waveform parameters play a crucial role in the quality and the characteristics of the received signal. As discussed in Section 2.2.2, selection of certain parameters affect range, velocity and angle resolution, unambiguous velocity and range, frame rate and other metrics. These parameters are often optimized and selected based on the end-application's needs. For instance, while front-looking automotive radars require higher maximum ranges (e.g., ≥ 150 m) without needing fine range resolution (e.g., ≤ 30 cm), human activity and sign language recognition applications typically require much less maximum range (≤ 10 m) but with a finer range resolution (≤ 5 cm). Therefore, a special attention should be paid while adjusting the RF waveform parameters as the data cannot be recovered or enhanced after the data acquisition if there are certain errors or sub-optimal selections in the parameters.

Most of the RF-based sign language recognition studies select a certain set of fixed parameters and acquire all the data with the same parameter set (i.e., parameter profile). This is a viable option for developing an end-to-end functional system. However, it requires radar to be continuously operational in high data rate mode while occupying all the RF-related and computational resources (e.g., bandwidth, random access memory (RAM), data storage memory, GPUs etc.) even if there is no informative or communicative action occurring in the radar FoV. Constantly allocating a large bandwidth can raise interference problems in the presence of other RF sensors. Time-scheduled allocation of it can enable more

spectrum-efficient solutions. RAM is often utilized by other computational modules of the system as well and unnecessary occupation of it can cause processing delays or out-of-memory issues for certain applications. Therefore it should be occupied only when a computational resource is needed with a reasonable amount. In addition, it is often desired for the system to be able to store the acquired data locally or in a cloud platform. When a daily living scenario is considered where an RF sensor is mounted in a corner or on a wall of the room to observe and recognize sign language, continuous recording of the acquired data can easily result in a very large amount of storage memory without containing much informative data since there is not much communicative interaction between the sign language recognition systems and the people during the day. Therefore, an automated system is needed to understand the presence of the people in a room, temporally isolate the individual activities and differentiate daily, non-communicative activities from the communicative sign language articulations. This way, the informative data can be separated from other activities and can be efficiently stored without needing to manually segment and label the data which is a labor-intensive and an expensive task. GPUs, on the other hand, are the hardware units used to make inference from the trained prediction model. Although modern computers and laptops are equipped with high-end GPU units, smaller edge-computing devices such as NVIDIA-Jetson Nano usually have GPUs with less computational capabilities and memory. Considering GPUs are also used in rendering other graphical displays, they can be in high demand by several modules of the system. Hence, they should be utilized in an effective manner to maximize the system efficiency and mitigate the unwanted computational overhead.

Optimization of RF waveform in the context of data quality is still a pristine area with a very limited amount of study. Hong et al. [78] recently proposed a reinforcement learning-based cognitive Doppler radar approach to optimize the carrier frequency and the sampling rate (PRF is referred as sampling rate throughout the paper). They used Carnegie Mellon University (CMU) Graphics Lab's human activity motion capture data to simulate the RF μD spectrograms and evaluate the proposed method. It is found that for 7 human

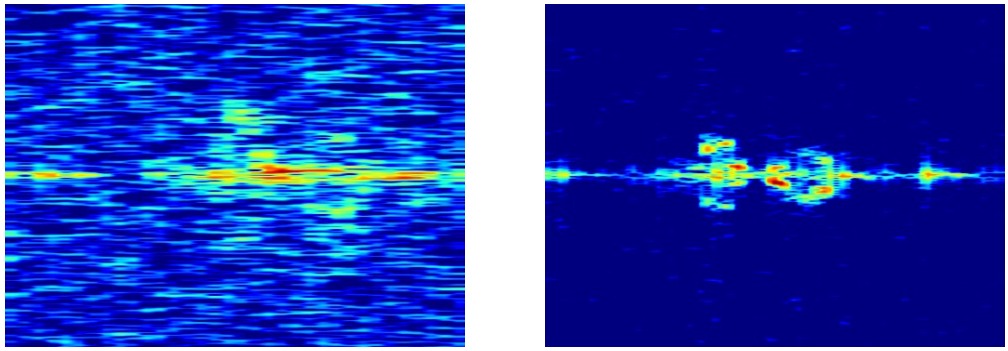
Figure 6.1: Doppler bandwidth and sampling rate on μD spectrogram (a), and Doppler aliasing affect due to low PRF in parameter selection (b).



(a) (b)

activity classes there is an optimal interval for the carrier frequency and going beyond that results in lower classification accuracy. However, when the accuracy of the individual activities are considered, this interval is subject to change. Therefore, there is no clear pattern which can be generalized to all human activities or sign language recognition tasks. Nonetheless, it is a fact that the center frequency is proportional to the Doppler resolution of the signature. For a more expressive and detailed μD spectrograms, higher carrier frequencies are desired. Figure 6.1a shows the Doppler bandwidth and the PRF/2 frequency span on a μD spectrogram. PRF is also another important parameter which determines the maximum unambiguous velocity. If the PRF is not chosen wisely, unambiguous velocity can be lower than application needs, and when targets move faster than the upper limit, Doppler aliasing effect can be observed. Doppler aliasing is basically wrapping of the Doppler components to the other side of the spectrum if they exceed the maximum unambiguous velocity. Figure 6.1b illustrates this effect. Such artifact causes corrupt and kinematically incorrect data representations. When such data are used to train or test the ML/DL models, they result in sub-optimal performances.

Figure 6.2: μD spectrogram for the word HAVE when sampling rate, f_s=1 MHz (a), and when f_s=2 MHz (b).

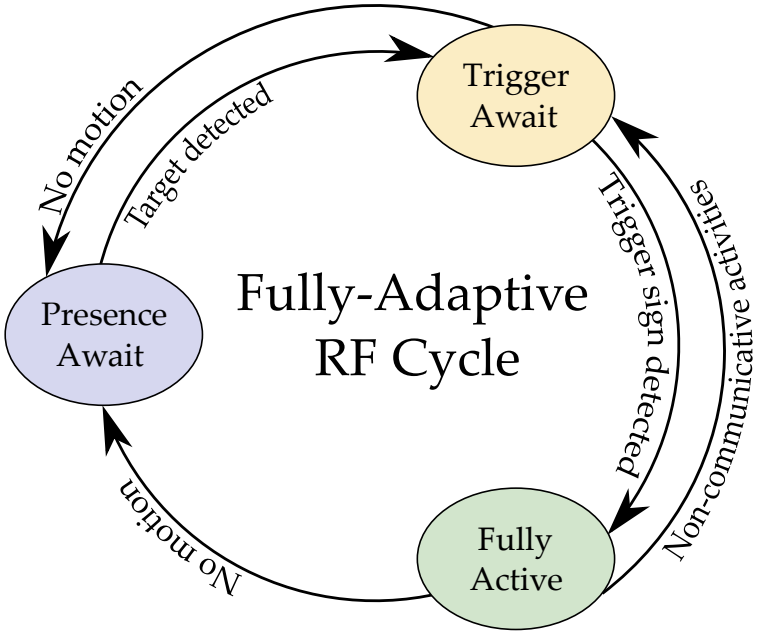


It is also found that there is a clear pattern which suggests higher sampling rates yield higher accuracies. This is an expected result as higher sampling rates result in higher SNRs in the collected signal. Figure 6.2 illustrates this phenomenon for the ASL word HAVE. It can be observed that when the sampling rate is higher, SNR of the μD signature increases drastically.

Gong et al. [55] studied the effect of CPI and PRF on the μD signals of a DJI Phantom 4 drone and a P750 aircraft. They observed that changes in CPI and PRF values have differential effects on Jet Engine Modulation (JEM) spectra and the blade flash patterns of the μD spectrograms.

In this chapter, we propose an adaptive RF waveform parameter adjustment framework according to the observed scene of the radar. The system is able to observe, understand and adapt to the environment by changing its operation mode and waveform characteristics. Specifically, we define three operation modes: presence await (PA), trigger await (TA) and fully-active (FA) mode. The main objective of the radar in PA mode is to determine if there is a moving target/subject in the room or not. The PA mode makes use of waveform parameters with relatively low data rate and bandwidth. Therefore, its computational cost is very low. It utilizes a small fraction of the full bandwidth and do not store any data during acquisition. No prediction model or GPU memory is used since determination of the target

Figure 6.3: Fully-adaptive RF cycle.



presence is based on the received power strength and do not need any learning model to make this decision. Once target presence is detected, the system switches to the TA mode. The TA mode is used to spot the trigger or wake sign from the user. Its computational cost is relatively higher than the PA mode and lower than the FA mode. The TA mode is continuously tries to determine if the trigger sign is articulated or not. When the system is turned on with the trigger sign, the system switches to the FA mode. The FA mode enables the full capabilities of the system by temporally segmenting activities, separating daily activities from ASL signs and recognizing different signs. After switching to TA or FA mode, if there is no moving target in the radar FoV for a certain duration the system goes back to the idle (i.e., PA) mode. In the FA mode, if the performed activities are not communicative ASL signs, but random movements or daily activities, the system goes back to the TA mode. Figure 6.3 depicts the proposed fully-adaptive RF cycle framework. It is found that the proposed method maintains high level ASL sign recognition performance while minimizing the computational costs and allocation of computational units in the system.

6.2 Adaptive RF Dataset and Experimental Setup

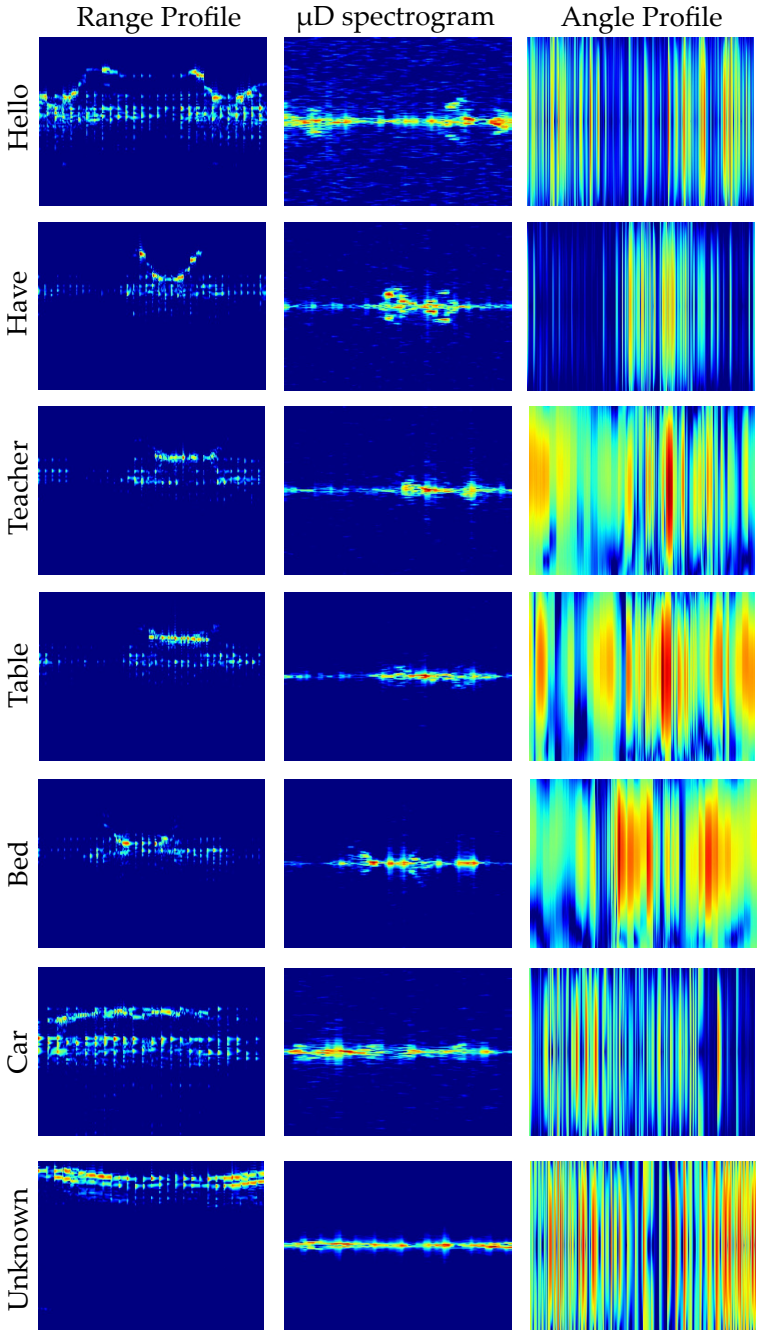
6.2.1 RF Sensor and Dataset Description

In order to evaluate the proposed method, we use Infineon's BGT60TR13C Demo FMCW radar operating at 60 GHz. This device has several advantages over other FMCW sensors for indoor monitoring. First, the device has a very small package size (40.64mm x 25.4mm) enabling it to be deployed in any corner of the room seamlessly. Second, it has a relatively large bandwidth (5.5 GHz) which yields very fine range resolution of 0.027m. The device also has MATLAB and Python software development kit (SDK) support to be operable in real-time without needing to rely on manufacturer's GUI for end-applications. Finally, the RF sensor also has an L-shaped RX antenna array with 3 receivers in a single-input-multiple-output (SIMO) fashion which enables angle estimation in both azimuth and elevation directions.

As discussed in the previous section, the proposed fully-adaptive RF cycle approach has three operation modes: PA, TA and FA mode. Each of these modes have their own respective waveform parameter profiles: PA-RF, TA-RF and FA-RF. The adaptive RF dataset consists of 6 ASL signs (HELLO, HAVE, TEACHER, TABLE, BED, CAR) and 1 UNKNOWN class for random motions and daily activities like walking, sitting and standing up. For the UNKNOWN class, participants were free do whatever they want in the radar FoV. The ASL signs are performed in the direct line-of-sight of the radar in a sitting position. The RF sensor was placed 1.5 away from the subjects and approximately 0.9 above the ground. The experiment is repeated for each parameter profile. In total approximately 1,500 samples were collected for each parameter set with a uniform distribution across classes ($\sim$ 210 samples/class for each parameter set). 5 participants have attended the study with various ages. The acquired data are then split into 70% and 30% portions for training and testing.

Range-Doppler maps, μD spectrograms, range profiles and angle profiles are used as RF data representations in this study. While computation of range-Doppler maps, μD spectrograms and range profiles are described in detail in Section 2.4, for angle profile computation, two different angle estimation approaches are considered: Fast Fourier Transform

Figure 6.4: Range, Doppler and Angle Profiles of fully-adaptive RF dataset samples.



(FFT) and Capon beamforming (i.e., minimum variance distortionless response (MVDR) beamformer) [20]. While MVDR has better resolution with the cost of higher computational complexity, FFT has lower resolution with the advantage of requiring less computation. In this work, both methods are evaluated and it is found that using MVDR increases the

computation time to the extend that radar frame rate drops when operated in real-time and causes buffering issue. Since the proposed method is designed to operate in real-time, we opted for using FFT-based angle estimation. Figure 6.4 shows range, Doppler and angle profile samples for each class of the dataset.

6.2.2 Parameter Profiles

Parameter profiles play a crucial role in the adaptive RF paradigm as they directly affect the operation and waveform characteristics of the system. Therefore, this section describes the parameter selection and the reasoning behind it for each parameter profile.

PA-RF parameters are used in the PA mode to determine the presence of a moving target in the radar FoV. PA-RF parameters are empirically optimized to be both computationally light and sensitive to detect the moving targets. The presence detection algorithm uses range-Doppler maps to make a decision. If the cumulative power level from the moving targets are above certain threshold, detection occurs. Considering an indoor environment (e.g., room, lab, office etc.), presence detection for targets closer than $\sim$5m should suffice. Frame interval of 0.1s (i.e., 10 FPS) is also determined to be temporally satisfying for the application needs. Sampling rate and PRF kept lower than other parameter profiles since this parameter profile is going to be used by the PA mode which will be running during a great portion of the day with the assumption that there is no moving person or object in the environment most of the day time. Keeping sampling rate and PRF reduces the data size drastically and the computational overhead on the computational units like RAM and the CPU. Finally, only one RX antenna is used since there is no need for the angle information in the presence detection algorithm.

TA-RF parameters are used when radar is operating in the TA mode which is activated when presence of a target is detected in the PA mode. TA-RF profile has the same bandwidth of 1 GHz as the PA-RF has. Higher allocation of bandwidth is not needed in TA-RF since only μD spectrograms are used for trigger sign detection and range resolution has no effect on them which is the only motivation for allocating larger bandwidths. Sampling

Figure 6.5: RF waveform selection of different parameter profiles.

RF Parameter \ Parameter Profile	PA-RF	TA-RF	FA-RF
Start Frequency (GHz)	58	58	58
End Frequency (GHz)	59	59	63.5
Sample Rate (MHz)	1	2	2
Number of ADC Samples	128	256	256
Number of RX Antennas	1	1	2
Frame Repetition Time (s)	0.1	0.1	0.1
Pulse Repetition Interval (s)	0.0004	0.0002	0.0002
Max. Unambiguous Range (m)	9.59	19.19	3.49
Max. Unambiguous Velocity (m/s)	3.2	6.41	6.17
Range Resolution (m)	0.15	0.15	0.027
Velocity Resolution (m/s)	0.1	0.1	0.096

rate and the PRF are higher than the PA-RF profile since the TA mode is using μD spectrograms to make predictions on whether the trigger sign is articulated or not, and the spectrograms should have high temporal and frequency resolution. Increasing PRF, also increases the maximum unambiguous velocity which prevents the aforementioned aliasing effect. Maximum unambiguous velocity of 6.41 m/s is obtained for the TA-RF parameter selection which is sufficient for indoor activity/signing monitoring applications considering average signing speed is $\leq$ 3 m/s. TA-RF profile also utilizes single RX antenna since angle information is not used in the trigger sign detection algorithm.

Finally, FA-RF parameters are used in the FA mode upon the trigger sign detection in the TA mode and switching to the FA mode. FA-RF profile utilizes the full available bandwidth of 5.5 GHz yielding 0.027m range resolution. Sampling rate and the PRF are also maximized to the limit allowed by the real-time processing frame rate requirements. Different than the TA-RF profile, FA-RF profile utilizes the all 3 RX antennas for angle estimation. The maximum unambiguous velocity of 6.17 m/s and velocity resolution of 0.096 m/s is obtained with the selected parameters. The maximum unambiguous range is

set to 3.49 m which is the shortest range amongst different parameter profiles, expecting user to be present close to the radar for greater returned signal SNR. Figure 6.5 summarizes the three parameter profiles.

6.3 Fully-Adaptive RF Cycle

6.3.1 Adaptive RF Operation Modes

This section describes the three modes of the adaptive RF operation cycle in detail and compares them.

Presence Await Mode

The radar starts with the PA mode. In this mode, it continuously observes the scene with the purpose of detecting a moving object. For that, we use a power-based presence detection algorithm. The acquired raw data are first reshaped into a 3D array with the shape of (number of ADC samples $\times$ number of chirps $\times$ number of RX channels). After the reshaping operation, a moving target indicator (MTI) filter is applied to suppress the signals reflected back from stationary objects and enhance the SNR of the moving target signals. Then, a 2D-FFT is applied to obtain the range-Doppler map. Total energy of the range-Doppler is computed by summing up the power levels (dB) of all the range-Doppler bins. If the cumulative power level exceeds a predefined threshold, detection occurs and the radar switches to the TA mode.

Trigger Await Mode

In the TA mode, the system utilizes the TA-RF parameter profile to detect the occurrence of the trigger sign. In order to detect the beginning and the ending of a sign and isolate its raw data, the STA/LTA based motion detector presented in an earlier study [106] is utilized. In the original study, the presented motion detector operates on μD spectrogram envelopes. Although the presented method works well when the data is processed offline without

considering real-time processing challenges, it becomes computationally costly and infeasible due to the high computational complexity of generating μD spectrograms in real-time with high resolution and their upper and lower envelope extraction process. Therefore, a more lightweight method is needed to run the motion detector in real-time.

In order to reduce the computational complexity of the μD envelope-based motion detector, in this work, we propose a modified STA/LTA-based motion detector which makes use of the total power, P_T, in the range-Doppler map. The total power in a 2D range-Doppler map (RDM) can be computed by:

$$P_T = \sum_{r=1}^{R} \sum_{d=1}^{D} 20 \log_{10} |RDM(r,d)| \tag{6.1}$$

where R and D are the number of range and Doppler bins respectively. Then, $STA(t)$ and $LTA(t)$ can be defined as the leading and lagging windows at time t as:

$$STA(t) = \frac{1}{T_1} \sum_{k=t+1}^{t+T_1} P_T(k),\ LTA(t) = \frac{1}{T_2} \sum_{k=t-T_2+1}^{t} P_T(k) \tag{6.2}$$

where T_1 and T_2 are the lengths of short and long windows respectively. While greater T_1 and T_2 values are more robust to false alarms in noisy data, they increase the response time of the system proportionally since at least a total time of T_1 should pass after a motion is performed and for it to appear in the lagging window. Therefore, T_1 and T_2 values should be selected based on the application requirements. In this work, we empirically optimized the window sizes of $T_1 = T_1 = 0.5$s. The starting point of a motion is detected when the following conditions are satisfied:

$$STA(t) > \sigma_1 \quad \text{and} \quad \frac{STA(t)}{LTA(t)} > \sigma_2 \tag{6.3}$$

where σ_1 and σ_2 are predefined detection thresholds. Similarly, the ending point is detected if

$$STA(t) < \sigma_3 \quad \text{and} \quad \frac{STA(t)}{LTA(t)} < \sigma_2 \tag{6.4}$$

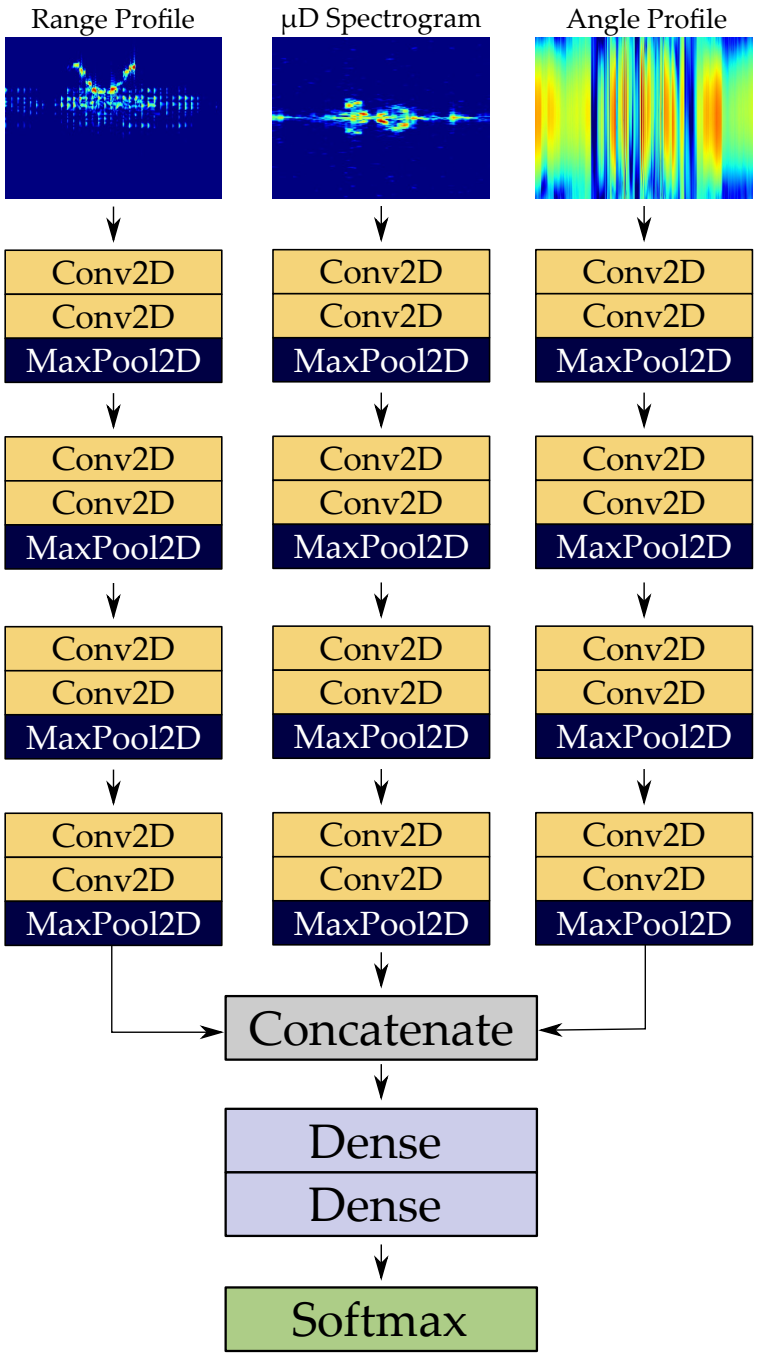
where σ_3 is the detection threshold for the stopping point.

Once the beginning and ending of a motion is detected, the interval corresponding to the beginning and the ending point of the motion is isolated from the raw data. μD spectrogram of the isolated data portion is generated using the STFT method. Since STFT operation is only performed on a small interval of the data where the motion is occurring, its computation is light enough and does not hinder the real-time processing. After μD spectrogram is generated it is passed into a trigger recognition network to determine if the observed motion is the trigger sign or not. The trigger recognition network is a CNN-based binary classification model which inherits 4 convolutional blocks from the VGG-16 model trained with ImageNet weights. The model is fine-tuned with the μD spectrograms of the TRIGGER sign and NON-TRIGGER class samples. NON-TRIGGER class samples include daily activities (e.g., walking, sitting, standing up, arm gestures) and other 5 ASL signs in the acquired dataset. Since the number of samples in the NON-TRIGGER class is significantly larger ($\geq 5\times$) than the TRIGGER class, class-weighting approach based on the number of samples for each class is applied in the loss function. This step is crucial to prevent network to bias its weight optimization process towards NON-TRIGGER class samples. If the observed motion is predicted as TRIGGER, the system switches to the FA mode. Otherwise, the system stays in the TA mode and continues to isolate and predict different motions occurring in the radar FoV. If there is no motion for a certain duration, the system cycles back to the PA mode.

Fully-Active Mode

When the TRIGGER sign is detected, the FA mode gets activated. The FA mode is the operation mode where capabilities of the system are maximized. The system uses the full

Figure 6.6: Multi-input ASL recognition network.



available bandwidth, all the RX channels and high data rate. Computational hardware units such as memory, RAM and GPU are also occupied accordingly. The motivation behind this selection is to maximize the acquired data quality and information and utilize them to make better sign language predictions and increase the user experience by yielding more accurate

results. Similar to the TA mode, the FA mode also utilizes the proposed STA/LTA-based motion detector for temporal segmentation of the individual activities/signs. Upon isolation of the raw data of an individual activity/sign, range profile, μD spectrogram and angle profile of the data are generated. The generated RF data representations are then passed to a multi-input CNN-based network to recognize a particular sign.

The multi-input CNN network inherits 4 convolutional blocks of the VGG-16 network pretrained with ImageNet weights. Range profiles, μD spectrograms and angle profiles are processed in the network with identical but separate CNN layers in a parallel fashion. After the CNN blocks, a global average pooling layer is used to flatten the feature embeddings. Feature spaces of three inputs are fused in a concatenation layer which is followed by two fully-connected layers. The model is finally augmented with a softmax layer with 7 nodes for 6 ASL signs and 1 UNKNOWN class. The overall architecture of the proposed model is presented in Figure 6.6. If the predicted sign is an ASL sign, the system stays in the FA mode and continues to interact with the user. If the predicted sign is the UNKNOWN class which includes random movements or daily activities and gestures, the system goes back to the TA mode since these activities are non-communicative and it is highly likely that the user has finished interacting with the system. Similar to the TA mode, if there is no motion for a certain duration, the system cycles back to the PA mode which lowers and stops the occupation of certain computational resources.

6.4 Non-Adaptive versus Adaptive RF Sensing

In conventional RF-based activity or sign language recognition applications, it is a common practise to optimize the RF waveform parameters and use a fixed set of parameters (i.e., non-adaptive) to collect all the data and make inference on the trained model. Although this approach works well when the radar is coupled with high-end computational units and do not interfere with other modules of the system, when the edge-computing device has limited computational capabilities such as low RAM, GPU or memory, and need to share

Resource \ Mode	**Non-Adaptive RF**			**Adaptive RF**
Parameter Profile	PA-RF	TA-RF	FA-RF	Switching between 3 profiles
Bandwidth (GHz)	1	1	5.5	Varying [1, 5.5]
Storage Memory (MB/s)	0.67	2.5	7.49	7.49 after the system trigger
RAM (MB)	487	591	744	Varying [487, 591, 744]

Table 6.1: Resource allocation comparison of non-adaptive versus adaptive parameter selection approaches.

the computational resources with other modules of the system, allocation of computational units needs to be taken into account and utilized in an effective manner. Therefore, in this section, we compare the non-adaptive and the proposed adaptive RF sensing approach in terms of both resource allocation efficiency and overall recognition capability of the system.

6.4.1 Resource Allocation Efficiency Benchmark

In order to compare the resource allocation efficiency of the non-adaptive and adaptive RF approaches, we evaluate them using three different metrics: bandwidth allocation, storage memory allocation and RAM allocation. These spectral and computational resources are common in almost all radar-based recognition applications. Non-adaptive approach is evaluated by fixed utilization of PA-RF, TA-RF and FA-RF parameter profiles without any transition between them.

Table 6.1 summarizes the resource allocation results for two approaches. It can be seen that non-adaptive parameter profiles always allocate a fixed bandwidth. While lower bandwidths has the advantage of less chance of interference with other RF sensors might be present in the environment, they provide poor range resolution. Higher bandwidths, on the other hand, yield superior range resolution with a higher chance of interfering with other RF sensors. Considering this phenomenon, a fixed selection of PA-RF or TA-RF will be more advantageous when the system is not actively being used as the RF system will have less chance of affecting other sensors or being affected by them. Fixed selection of FA-RF, on the other hand, has the advantage of providing more high quality data with the cost of allocating a larger bandwidth continuously. Adaptive RF approach, basically, takes the

good part of different bandwidth selections by using a lower bandwidth when the system is not active and switching to the full-bandwidth after the system is triggered. This ensures the acquisition of high resolution data during interaction with the user while minimizing the interference when the system is not actively being used.

In terms of storage memory allocation, using lower sampling rate, PRF and less number of RX channels reduces the acquired data size drastically with the cost of low resolution, lower unambiguous velocity and not being able to estimate the target's direction-of-arrival. However, DNN models used in RF data recognition tasks are data driven models and the quality of the training data plays a crucial role in the recognition performance of the model. Therefore, data with rich spatial and temporal features are needed to train a robust learning model. While non-adaptive fixed PA-RF and TA-RF yield small data size, they compromise the data quality and features can be useful during the model training. Fixed FA-RF profile, on the other hand, yields high quality data but the data size is always large even when the system is not actively used. Considering such RF-controlled interactive system will potentially be used less than a few hours a day and the cost of cloud-based storage solutions per GB such as Microsoft's Azure, Google Cloud Platform or AWS Cloud Storage, it is not needed to save data always in high quality when data are not informative. The adaptive RF approach optimizes this problem by reducing the data size during non-communicative actions in PA and TA modes in a daily scenario and increasing the data rate during interaction with the user by switching to the FA-RF profile for a higher data quality and performance.

Finally, RAM allocations of non-adaptive and adaptive RF approaches are compared. RAM is a significant computational resource which stores all the variables and data in its memory during a program's execution. Several modules of a system and the operation system can make use of the RAM simultaneously and continuously. Since it is a shared unit, it should be occupied efficiently in order to mitigate system response delays and out-of-memory errors. In the fixed PA-RF and TA-RF parameter selection, the RF sensor occupies a smaller portion in the memory with 487 MB and 591 MB of space respectively when compared to

the FA-RF profile which allocates 744 MB of space. This is mainly due to the smaller data size per frame and the lighter computational cost on a smaller data chunk. The adaptive RF approach minimizes the RAM allocation by using PA-RF profile when no moving target present in the environment. It starts to occupy a larger space as a result of switching to the TA-RF and FA-RF profiles when a target is detected and the system is triggered respectively. Therefore, it minimizes the unnecessary RAM occupation during idle times.

6.4.2 Classification Results

While optimizing the usage efficiency of computational resources are important from a system point-of-view, we should also ensure and maximize the recognition capability of the system. Compromising the recognition performance of the system for the sake of minimizing computational cost can result in poor user experience due to wrong predictions. Therefore, it is important to balance and maximize the two metrics. The proposed approach listens for the trigger sign in the TA mode before switching to the FA mode for ASL recognition. Once the trigger sign is detected, it starts to predict the ASL signs in the FA mode.

Trigger Sign Recognition Results

It is a common practise to trigger/awake an interactive system before starting to use it such as "Hey Siri" phrase in Apple's products, "Hey Alexa" in Amazon's products or "Okay Google" for Google Home. In this work, we follow a similar approach and evaluate the detection performance of each trigger candidate word separately by training individual models for each word. For the evaluation of trigger sign recognition task, we define two metrics false alarm rate (FAR) and false rejection rate (FRR). They are defined as:

$$FAR = \frac{FP}{TN + FP}, \qquad FRR = \frac{FN}{TP + FN} \tag{6.5}$$

where FP is the number of false positives (i.e., predicting trigger when it is not), TN is the number of true negatives (i.e., predicting non-trigger motions correctly), FN is the number of

Param. Profile	Metric \ Word	HELLO	HAVE	TEACHER	TABLE	BED	CAR
PA-RF	FAR (%)	13.25	7.23	16.87	13.25	10.84	16.87
	FRR (%)	13.24	2.6	5.41	5.33	13.58	2.56
	Detection Rate (%)	73.51	90.17	77.73	81.41	75.58	80.57
TA-RF	FAR (%)	8.14	4.65	8.14	10.47	5.81	9.3
	FRR (%)	2.6	2.9	2.74	1.32	6.78	3.61
	Detection Rate (%)	89.26	92.45	89.12	88.22	87.41	87.08
FA-RF	FAR (%)	21.52	0	12.66	20.25	16.46	18.99
	FRR (%)	4.29	6.06	1.41	2.5	1.19	5.13
	Detection Rate (%)	74.2	93.94	85.93	77.25	82.35	75.88

Table 6.2: Trigger sign recognition results of each word for different parameter profiles.

false negatives (i.e., missed triggers), and TP is the number of true positives (i.e., predicting the trigger sign correctly). Based on the FAR and FRR of a sign, detection rate, D_R, of a sign can be computed as:

$$D_R = 1 - FAR - FRR \tag{6.6}$$

Table 6.2 presents the trigger recognition results of each sign for different parameter profiles. Note that a binary classification model described in Section 6.3.1 is trained for each word and parameter profile pair separately. It can be observed that the TA-RF profile has the highest average detection rate of 88.92% for the six signs while the PA-RF and FA-RF profiles can achieve 79.83% and 81.59%, respectively. The models trained with the TA-RF profile perform significantly better than the other two parameter sets except for the word HAVE where the FA-RF profile performs only 1% better which can be due to various reasons including different number of samples in the training/testing datasets or the way participant articulates the sign. The poor performance of the models trained with data collected in PA-RF mode can be attributed to the low quality data and the lower SNR in μD spectrograms. Although the lower performance for the FA-RF mode is a bit unexpected, we can see that the lower detection rates are mostly due to high FARs when compared to the TA-RF mode even though the FRRs remain low. Other reasons can also include the free

form activities participants perform during the data collection. If the μD signatures of the motions they perform during data collection resemble to the trigger sign, there is a higher chance that they might be confused with the trigger sign. Recall that participants were free to do anything they want during free form activities except the ASL signs. Nonetheless, the overall trend does not change and we can conclude that the TA-RF parameter set is well-suited for the trigger recognition task.

It can be seen that for all the parameter profiles the word HAVE yields the highest detection rates of 90.17%, 92.45% and 93.94% and the lowest FARs of 7.23%, 4.65% and 0%. This might be due to the high radial displacement of the arms while moving both hands towards the chest and retrieve them back. Radar is the most sensitive to the motions in radial direction which makes motions with high kinematic variance in radial axis more perceivable to the radar. The word TABLE consistently has very high FARs for all the parameter profiles with over 13%, 10% and 20% which indicates the higher chance of being confused with other motions and unintentional activation of the system. This is not a desired behavior from a user experience point-of-view. High FAR, on the other hand, seems to reduce the FRR of the word which is a desired behavior, but it comes with the cost of so many false alarms. The word TABLE is articulated by parallelly moving both arms stacked on top of each other up and down. This can cause strong returned signal strengths from the arms which is similar to torso movements observed on daily activities. Therefore, it can be easier for the DNN model to confuse the daily activities as if they are trigger signs. The words HELLO, TEACHER, BED, CAR, on the other hand, have similar detection rates ranging between 87%-89% for the TA-RF mode. Based on these results, we chose the word HAVE as the trigger sign of the system for the TA mode.

Sign Language Recognition Results

When the system is activated with the trigger sign, it starts to recognize the ASL signs articulated by the user. In order to predict the articulated sign, we use the model described in

Model Input(s)	Param. Profile	Accuracy (%)
μD Spectrogram	PA-RF	52.24
	TA-RF	67.11
	FA-RF	82.01
μD Spectrogram Range Profile	PA-RF	49.63
	TA-RF	69.22
	FA-RF	86.17
μD Spectrogram Range Profile Angle Profile	FA-RF	**87.5** (Proposed)

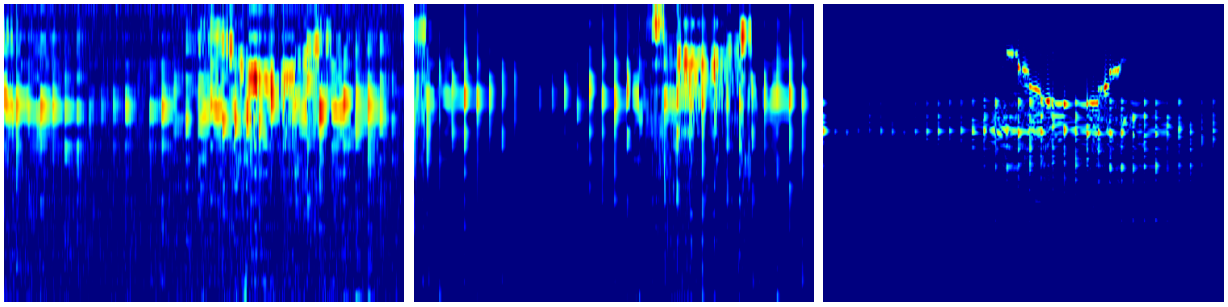
Table 6.3: ASL recognition results of the parameter profiles with various model input(s).

Section 6.3.1 and Figure 6.6. Performance of the proposed network is evaluated by applying certain ablation studies on the network architecture and the RF parameter profile.

First, we compare the performance of different parameter profiles for the classification of μD spectrograms in a single input network. Such CNN-based network is the baseline model which is commonly used μD spectrogram classification tasks. The single input network has the identical architectural structure to the proposed method with the exception of having only one branch since only μD spectrograms are used to train/test the model. Other training settings such as optimizer, number of layers, kernel sizes, learning rate, learning rate scheduler, dropouts also kept the same to have a fair comparison. The models trained with the data acquired in PA-RF, TA-RF and FA-RF profiles yielded the accuracies of 52.24%, 67.11% and 82.01%, respectively for the 7-class (6 ASl sign + 1 UNKNOWN class) ASL recognition task. It can be observed that as the sampling rate, PRF, and SNR increases, the learning models can achieve higher accuracies as a result of having more expressive and high quality data samples.

Next, we expand the network to two branches in order to incorporate the range information into the model. This model takes both μD spectrogram and range profile as separate inputs and processes them parallelly, and fuses their feature spaces after the global average pooling operation. This network resembles to the proposed network illustrated in 6.6 with the

Figure 6.7: Range profile comparison of different RF waveform profiles.



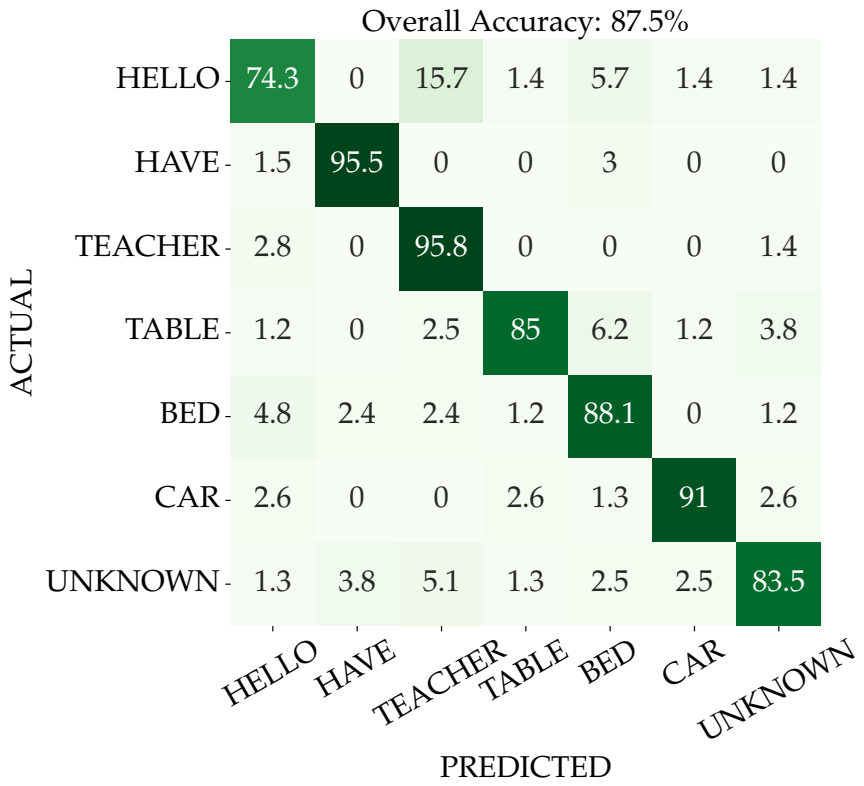
(a) PA-RF. (b) TA-RF. (c) FA-RF.

exception of not having the third branch for the angle profile. After training, the augmented two-branch model yielded the accuracies of 49.63%, 69.22% and 86.17% for PA-RF, TA-RF and FA-RF profiles. It can be observed that while there is not much improvement for the PA-RF and TA-RF profiles, accuracy of the FA-RF profile is increased over 4%. This can be due to the high bandwidth selection in the FA-RF profile which results in finer range resolution as depicted in Figure 6.7. The detrimental effect observed in the PA-RF profile for around 3% can be because of the increased number of trainable parameters while not providing very informative and distinct input samples.

Finally, we evaluate the performance of the proposed three-input network which incorporates range, Doppler and angle information by expanding the network with a third branch which processes angle profiles. This method is evaluated only for the FA-RF profiles since other parameter profiles utilize single RX antenna data and the angle estimation cannot be performed. The proposed model outperformed other method by achieving the testing accuracy of 87.5% on 7-class ASL recognition task. These results show the efficacy of utilizing all the physical information about the environment RF sensors can offer. Table 6.3 summarizes the performances of different waveform profiles along with different RF data representations as DNN inputs.

Figure 6.8 shows the final confusion matrix of the proposed model. The words HAVE and TEACHER yield the highest recognition rate with over 95% accuracy. These results are also in line with the trigger recognition results obtained in Section 6.4.2 where the word HAVE had the highest trigger recognition rate among other signs. The word HELLO, on the other

Figure 6.8: Confusion matrix of the proposed multi-input network for ASL recognition.



hand, has the lowest recognition rate with the accuracy of 74.3%. It is mostly confused with the word TEACHER. This potentially be because of the similarity of the way the beginning of these signs are articulated. While the word HELLO is articulated by raising one hand to the forehead and moving forward towards the person of interest, the word TEACHER is articulated by raising both hands to the forehead and moving them forward for a certain distance, and lowering them parallelly at the end. Confusion of the UNKNOWN activities with the other signs seem to be distributed across the signs.

6.5 Conclusion

This chapter introduces the idea of fully-adaptive RF where the waveform parameter as well as the operation characteristics of the RF-based ASL recognition system changes. Advantage of the proposed method is evaluated both in terms of computational and spectral

resource allocation, and the recognition performance of the overall system. We propose a cyclic multi-state operation diagram for the RF system where each state utilizes a certain set of waveform parameters to optimally use the computational resources without compromising the recognition performance. Resource allocation and recognition performances of different parameter profiles are evaluated for an ASL dataset consisting of 6 ASL signs and daily activities. Incorporation of range and angle information in addition to the Doppler in a multi-input network is shown to be a promising approach for enriching the feature space and achieving better recognition performance than the baseline methods.

CHAPTER 7

CONCLUSION AND DISCUSSION

This dissertation is mainly focused on the utilization of RF sensors for ASL-enabled smart environments. We tackle several challenges in RF-based ASL recognition systems including data variance stemming from background-related and cultural dialects, pre-processing of RF data, presentation of RF data to the deep learning models, temporal segmentation of sequential activities/signs, differentiation of ASL signs from daily activities, limitations in real-time recognition, separation and isolation of RF data of multiple people, automation of data collection and annotation stages, and adaptively changing of RF waveform parameters to optimize the resource allocation and to maximize the recognition performance. While most of these are on-going and not well-explored challenges, we propose intuitive solutions for them collectively. We provide both qualitative and quantitative results for the each proposed method/approach.

7.1 Summary of the Contributions

For the temporal segmentation problem, an STA/LTA-based motion detector is presented to locate the starting and ending point of a motion. The proposed method is shown to outperform other power-based methods and it eliminates the need for relying on fixed length windows. The method utilizes the Euclidean distance between upper and lower envelopes of the μD spectrograms to decide when a motion is starting and ending. Efficacy of the method is shown on a sequential mixed activity/signing data to isolate individual activities and signs in RF data in a daily living scenario.

Next, a joint-domain multi-input multi-task learning network is presented to aggregate information from different RF data representations including μD spectrograms, range-Doppler maps and range-angle maps. Each data representation is processed in parallel branches of the network. In the proposed network, while time-distributed 2D and 3D CNN layers are used to extract spatial features, temporal dependencies are obtained with bidirectional LSTM layers. We define certain auxiliary tasks for the signs such as one versus two handedness, major hand location, hand movement type, daily activity versus SL and number of arm strokes to better regularize the network in the training stage. The proposed technique is shown to outperform other SOTA methods in the sequential sign language recognition task.

When the RF sensors are deployed in an indoor environment to interact with the users via sign language, it is a natural need to turn on/off the system with certain gestures or signs similar to the wake/trigger words in voice-based personal assistant systems. In this study, we explore the design considerations and radar's capability to perceive and recognize these words. Accurate recognition of the trigger sign is crucial to reduce and eliminate the false triggers and false trigger rejections. The system should be robust enough to spot incomplete or overextended trigger attempts as well. In this work, we present an adaptive double-threshold cumulative score aggregation approach to recognize the trigger sign in continuous RF data streams.

Collecting data with RF sensors is often a time consuming and expensive task. A certain number of participants need to be recruited to attend the study and follow the instructions of the researchers to help with the data collection. In sign language recognition tasks, it is even harder to find and recruit the participants since the target community (Deaf/HoH) is narrower. Recruiting hearing participants is shown to be not a viable solution since there are significant differences between the signings of fluent and non-fluent (i.e., imitation) signers, and imitation signing cannot effectively represent the nuances of sign language. In addition, recruiting fluent signers (Deaf/HoH) for data collection in a laboratory environment is also not a sustainable solution since there are cultural differences across Deaf communities and

sign language is also an evolving language with the addition of new words and dialects in everyday life. In this work, we propose an interactive gamification approach to integrate sign language into the chess game where users control the pieces on the chess board with sign language instead of mouse clicks. The game collects, processes and classifies the collected signs. Users also have chance to correct the mispredictions of the game by canceling the last motion. This feature of the game eliminates the need for manual data annotation. The designed game also presents a new way to acquire data without boring the participants. This approach enable to curate a diverse, multi-modal sign language datasets in a sustainable and enjoyable fashion.

Presence of multiple people in the environment presents certain challenges in the RF data since the received signal becomes superimposition of the individual signals from each target. Although there exists certain methods to estimate target ranges and angles, they can be applied only after certain signal processing steps, and unprocessed raw data of the individual targets cannot be recovered. In this work, we present an angular subspace projection-based separation technique to separate the signals of individual targets at a low level. Achieving separation at the raw data level enables further signal processing and learning techniques for individual targets. We show the efficacy of the proposed method on human activity recognition and sign language recognition tasks. For closely spaced targets, we present a multi-view DNN model which incorporates left and right side of the boresight view.

In radar-based recognition tasks, RF waveform parameters are often optimized on the software based on the application needs. Although this is a commonly applied approach, software-defined RF sensors allow users to change the waveform parameters at any given time. There are certain pros and cons, and trade-offs while adjusting the radar parameters. Therefore, ideally, we want to keep the radar parameters optimal and aware of the surroundings so that it can perceive, understand and adapt to the environment. RF sensors are often coupled with other computational units to process and store the data. It is highly likely that these systems are also hosts for other modules of the system where computational resources

are shared across the units. Therefore, RF system should adjust its parameters and operation characteristics so that it does not occupy the spectral and computational resources it does not need especially when users are not interacting with the RF system. In this work, we propose a fully-adaptive cyclic RF sensing paradigm where radar has three operation states and each state is associated with a RF waveform profile. It is shown that the proposed paradigm can understand the presence of a person, whether the trigger sign is articulated and different ASL signs. The proposed method is shown to be effective in reducing the unnecessary occupation of the spectral and computational resources while preserving the high recognition performance.

7.2 Discussion

This work explores and tackles on-going challenges in RF sensor-based end-to-end sign language and human activity recognition systems. The presented studies and experiments show and prove that RF is a promising modality for indoor monitoring and human-computer interaction especially considering they are non-intrusive, non-invasive and robust against lightning conditions. While the purpose and claim of this research are not to replace the existing modalities such as video or wearables, we show what radars can offer and how they can be integrated into existing systems as a compatible modality seamlessly.

Other challenges current RF sensor-based recognition systems face include data scarcity. Although certain synthetic data generation methods have been proposed including GANs, simulated data, transfer learning and other physics-based methods, still each recognition task require at least certain amount of real data to be collected to drive the DNN-based solutions. Also, generalization capability of the synthesized data are often questionable as their variance is limited by the distribution of real data samples.

Low spatial resolution in point cloud data is also another on-going challenge especially for the autonomous driving scenarios. Accurate 3D reconstruction of the hand and finger shapes are still not possible with the existing RF systems. This limits radar's capability to

recognize only dynamic motions with high radial movement. Recognition of static shapes and finger-spelling are still challenging tasks without high resolution 3D hand and finger point clouds. However, this challenge can be alleviated in a near future with the recent advances in high resolution imaging radar technology and AI-based oversampling and RF data-to-skeleton methods.

Overall, radar looks like will be in our lives especially with smart home and human-computer interaction applications as a key player.

REFERENCES

[1] ANDREW ABBOTT and ANGELA TSAY. Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research*, 29(1):3–33, 2000.

[2] Canicious Abeynayake, Vidar Son, Hedayetul Islam Shovon, and Hiroshi Yokohama. Machine learning based automatic target recognition algorithm applicable to ground penetrating radar data. In *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXIV*, volume 11012, pages 1101202–. SPIE, May 2019.

[3] Oladipupo O. Adeoluwa, Sean J. Kearney, Emre Kurtoglu, Charles J. Connors, and Sevgi Z. Gurbuz. Near real-time ASL recognition using a millimeter wave radar. In Kenneth I. Ranney and Ann M. Raynal, editors, *Radar Sensor Technology XXV*, volume 11742, page 1174218. International Society for Optics and Photonics, SPIE, 2021.

[4] M. Amin. *Radar for Indoor Monitoring: Detection, Classification, and Assessment*. CRC Press, 2017.

[5] T. W. Anderson. *An introduction to multivariate statistical analysis*. John Wiley & Sons Inc., 1958.

[6] Aleksandar Angelov, Andrew Robertson, Roderick Murray-Smith, and Francesco Fioranelli. Practical classification of different moving targets using automotive radar and deep neural networks. *IET Radar, Sonar & Navigation*, 12(10):1082–1089, 2018.

[7] A. Arbabian et al. A 94ghz mm-wave to baseband pulsed-radar for imaging and gesture recognition. In *Symp. VLSI Cir.*, pages 56–57, 2012.

[8] Runhan Bao and Zhaocheng Yang. Cnn-based regional people counting algorithm exploiting multi-scale range-time maps with an ir-uwb radar. *IEEE Sensors Journal*, 21(12):13704–13713, 2021.

[9] Hansel Bauman. Gallaudet university deafspace design guidelines. https://app.dcoz.dc.gov/Exhibits/2010/ZC/15-24/Exhibit95.pdf.

[10] Jennifer S Beal and Kia Faniel. Hearing l2 sign language learners. *Sign Language Studies*, 19(2):204–224, 2019.

[11] Alice Blumenthal-Dramé and Evie Malaia. Shared neural and cognitive mechanisms in action and language: The multiscale information transfer framework. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(2):e1484, 2019.

[12] J. D. Borneman, E. A. Malaia, and R. B. Wilbur. Motion characterization using optical flow and fractal complexity. *Journal of Electronic Imaging*, 27(05), July 2018.

[13] Joshua D Borneman, Evie Malaia, and Ronnie B Wilbur. Motion characterization using optical flow and fractal complexity. *Journal of Electronic Imaging*, 27(5):051229, 2018.

[14] R. G. Bosworth, C. E. Wright, and K. R. Dobkins. Analysis of the visual spatiotemporal properties of american sign language. *Vision Research*, 164:34–43, Nov. 2019.

[15] Danielle Bragg, Naomi Caselli, John W. Gallagher, Miriam Goldberg, Courtney J. Oka, and William Thies. Asl sea battle: Gamifying sign language data collection. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[16] Danielle Bragg, Naomi Caselli, Julie A. Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E. Ladner. The fate landscape of sign language ai datasets: An interdisciplinary perspective. *ACM Trans. Access. Comput.*, 14(2), jul 2021.

[17] Danielle Bragg, Naomi Caselli, Julie A. Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E. Ladner. The fate landscape of sign language ai datasets: An interdisciplinary perspective. *ACM Trans. Access. Comput.*, 14(2), jul 2021.

[18] Danielle Bragg, Abraham Glasser, Fyodor Minakov, Naomi Caselli, and William Thies. Exploring collection of sign language videos through crowdsourcing. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2), nov 2022.

[19] Peibei Cao, Weijie Xia, Ming Ye, Jutong Zhang, and Jianjiang Zhou. Radar-id: human identification based on radar micro-doppler signatures using deep convolutional neural networks. *IET Radar, Sonar & Navigation*, 12(7):729–734, 2018.

[20] J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8):1408–1418, 1969.

[21] Enrique V. Carrera, Fernando Lara, Marcelo Ortiz, Alexis Tinoco, and Rubén León. Target detection using radar processors based on machine learning. In *2020 IEEE ANDESCON*, pages 1–5, 2020.

[22] Rich Caruana. Multitask learning. *Machine Learning*, 28, 07 1997.

[23] Naomi K. Caselli, Zed Sevcikova Sehyr, Ariel M. Cohen-Goldberg, and Karen Emmorey. Asl-lex: A lexical database of american sign language. *Behavior Research Methods*, 49(2):784–801, Apr 2017.

[24] Shao Changyu, Du Lan, Han Xun, and Liu Hongwei. Multiple target tracking based separation of micro-doppler signals from coning target. In *2014 IEEE Radar Conference*, pages 0130–0133, 2014.

[25] Alexander Charlish and Carolin Schwalm. Generating nlfm radar waveforms using variational autoencoders. In *2022 IEEE Radar Conference (RadarConf22)*, pages 1–6, 2022.

[26] Si-Wei Chen, Chen-Song Tao, Xue-Song Wang, and Shun-Ping Xiao. Polarimetric sar targets detection and classification with deep convolutional neural network. In *2018 Progress in Electromagnetics Research Symposium (PIERS-Toyama)*, pages 2227–2234, 2018.

[27] V. Chen. *The Micro-Doppler Effect in Radar, Second Edition*. Artech House, 2019.

[28] Mateusz Chmurski, Mariusz Zubert, Kay Bierzynski, and Avik Santra. Analysis of edge-optimized deep learning classifiers for radar-based gesture recognition. *IEEE Access*, 9:74406–74421, 2021.

[29] Jeong Woo Choi, Xuanjun Quan, and Sung Ho Cho. Bi-directional passing people counting system based on ir-uwb radar sensors. *IEEE Internet of Things Journal*, 5(2):512–522, 2018.

[30] Jeong Woo Choi, Dae Hyeon Yim, and Sung Ho Cho. People counting based on an ir-uwb radar sensor. *IEEE Sensors Journal*, 17(17):5717–5727, 2017.

[31] Ashley Chow, Glenn Cameron, Mark Sherwood, Phil Culliton, Sam Sepah, Sohier Dane, and Thad Starner. Google - isolated sign language recognition. *Kaggle*, 2023.

[32] C. Chuan, E. Regina, and C. Guardino. American sign language recognition using leap motion sensor. In *2014 13th International Conference on Machine Learning and Applications*, pages 541–544, 2014.

[33] R. Cipolla, Y. Gal, and A. Kendall. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proc. IEEE/CVF Conf. on Comp. Vis. and Patt. Recog.*, pages 7482–7491, 2018.

[34] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. *Association for Computing Machinery*, page 160–167, 2008.

[35] Angelo Coluccia, Alessio Fascista, and Giuseppe Ricci. A knn-based radar detector for coherent targets in non-gaussian noise. *IEEE Signal Processing Letters*, 28:778–782, 2021.

[36] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. Wizard of oz studies: Why and how. In *Proceedings of the 1st International Conference on Intelligent User Interfaces*, IUI '93, page 193–200, New York, NY, USA, 1993. Association for Computing Machinery.

[37] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes, 2021.

[38] Andreas Danzer, Thomas Griebel, Martin Bach, and Klaus Dietmayer. 2d car detection in radar data with pointnets. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 61–66, 2019.

[39] D. Deiana, E.M. Suijker, R.J. Bolt, A.P.M. Maas, W. J. Vlothuizen, and A.S. Kossen. Real time indoor presence detection with a novel radar on a chip. In *2014 International Radar Conference*, pages 1–4, 2014.

[40] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[41] Aashaka Desai, Maartje De Meulder, Julie A. Hochgesang, Annemarie Kocab, and Alex X. Lu. Systemic biases in sign language ai research: A deaf-led call to reevaluate research agendas, 2024.

[42] Vivek Dham. Programming chirp parameters in ti radar devices. `https://www.ti.com/lit/an/swra553a/swra553a.pdf`, 2020. Texas Instruments.

[43] Jun Ding, Bo Chen, Hongwei Liu, and Mengyuan Huang. Convolutional neural network with data augmentation for sar target recognition. *IEEE Geoscience and Remote Sensing Letters*, 13(3):364–368, 2016.

[44] P. V. Dorp and F. C. A. Groen. Feature-based human motion parameter estimation with radar. *IET Radar, Sonar Navigation*, 2(2):135–145, 2008.

[45] Thomas Eiter and Heikki Mannila. Computing discrete frechet distance. Technical Report 94/64, Christian Doppler Lab., Vienna Univ. of Technology, 1994.

[46] M. Erard. Why sign language gloves don't help deaf people. `https://www.theatlantic.com/technology/archive/2017/11/why-sign-language-gloves-dont-help-deaf-people/545441/`, November 2017.

[47] Michael Erard. Why sign-language gloves don't help deaf people. *The Atlantic*, November 2017.

[48] Baris Erol and Moeness G. Amin. Radar data cube processing for human activity recognition using multisubspace learning. *IEEE Transactions on Aerospace and Electronic Systems*, 55(6):3617–3628, 2019.

[49] Biyi Fang, Jillian Co, and Mi Zhang. Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. Association for Computing Machinery, 2017.

[50] Jeremy Fix, Israel Hinostroza, Chengfang Ren, Giovanni Manfredi, and Thierry Letertre. Transfer learning for human activity classification in multiple radar setups. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1576–1580, 2022.

[51] Linda K Ford, Joshua D Borneman, Julia Krebs, Evguenia A Malaia, and Brendan P Ames. Classification of visual comprehension based on eeg data using sparse optimal scoring. *J. Neural Engineering*, 18(2):026025, 2021.

[52] J. M. Fuster. *Cortex and mind: Unifying cognition*. Oxford University Press, 2003.

[53] R. Girshick. Fast r-cnn. In *Proc. IEEE ICCV*, pages 1440–1448, 2015.

[54] Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. Neighbourhood components analysis. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.

[55] Jiangkun Gong, Jun Yan, Deyong Kong, and Deren Li. Introduction to cognitive micro-doppler radar: Optimization and experiment. In *2023 IEEE International Radar Conference (RADAR)*, pages 1–6, 2023.

[56] T. Gong, T. Lee, C. Stephenson, V. Renduchintala, S. Padhy, A. Ndirango, G. Keskin, and O. H. Elibol. A comparison of loss weighting strategies for multi task learning in deep neural networks. *IEEE Access*, 7:141627–141632, 2019.

[57] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[58] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.

[59] Alex Graves, Santiago Fernández, and Faustino Gomez. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *in Proc. Int. Conf. on Mach. Learn.*, pages 369–376, 2006.

[60] Hugh Griffiths. *Introduction to waveform diversity and cognitive radar*. Radar, Sonar and Navigation. Institution of Engineering and Technology, 2017.

[61] Changzhan Gu, Jian Wang, and Jaime Lien. Motion sensing using radar: Gesture interaction and beyond. *IEEE Microwave Magazine*, 20(8):44–57, 2019.

[62] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition, 2020.

[63] Sevgi Z. Gurbuz, Chris Crawford, Darrin J. Griffin, Emre Kurtoglu, Oladipupo Adeoluwa, and Josh Haeker. Interactive rf game design for deciphering real-world human motion: Activities, gestures, and sign language. In *2023 IEEE Radar Conference (RadarConf23)*, pages 1–6, 2023.

[64] Sevgi Z. Gurbuz, Ali C. Gurbuz, Evie A. Malaia, Darrin J. Griffin, Chris Crawford, Emre Kurtoglu, M. Mahbubur Rahman, Ridvan Aksu, and Robiulhossain Mdrafi. Asl recognition based on kinematics derived from a multi-frequency rf sensor network. In *2020 IEEE SENSORS*, pages 1–4, 2020.

[65] Sevgi Z. Gurbuz, Ali C. Gurbuz, Evie A. Malaia, Darrin J. Griffin, Chris Crawford, M. Mahbubur Rahman, Ridvan Aksu, Emre Kurtoglu, Robiulhossain Mdrafi, Ajaymehul Anbuselvam, Trevor Macks, and Engin Ozcelik. A linguistic perspective on radar micro-doppler analysis of american sign language. In *2020 IEEE International Radar Conference (RADAR)*, pages 232–237, 2020.

[66] Sevgi Z. Gurbuz, Ali Cafer Gurbuz, Evie A. Malaia, Darrin J. Griffin, Chris S. Crawford, Mohammad Mahbubur Rahman, Emre Kurtoglu, Ridvan Aksu, Trevor Macks, and Robiulhossain Mdrafi. American sign language recognition using rf sensing. *IEEE Sensors Journal*, 21(3):3763–3775, 2021.

[67] Sevgi Z. Gurbuz, Emre Kurtoglu, M. Mahbubur Rahman, and Dario Martelli. Gait variability analysis using continuous rf data streams of human activity. *Smart Health*, 26:100334, 2022.

[68] Sevgi Z. Gurbuz, M. Mahbubur Rahman, Emre Kurtoglu, Trevor Macks, and Francesco Fioranelli. Cross-frequency training with adversarial learning for radar micro-Doppler signature classification (Rising Researcher). In Kenneth I. Ranney and Ann M. Raynal, editors, *Radar Sensor Technology XXIV*, volume 11408, page 114080A. International Society for Optics and Photonics, SPIE, 2020.

[69] Sevgi Z. Gurbuz, M. Mahbubur Rahman, Emre Kurtoglu, Evie Malaia, Ali Cafer Gurbuz, Darrin J. Griffin, and Chris Crawford. Multi-frequency rf sensor fusion for word-level fluent asl recognition. *IEEE Sensors Journal*, 22(12):11373–11381, 2022.

[70] Sevgi Z. Gurbuz, M. Mahbubur Rahman, Emre Kurtoglu, and Dario Martelli. Continuous human activity recognition and step-time variability analysis with fmcw radar. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 01–04, 2022.

[71] Sevgi Zubeyde Gurbuz, Hugh D. Griffiths, Alexander Charlish, Muralidhar Rangaswamy, Maria Sabrina Greco, and Kristine Bell. An overview of cognitive radar: Past, present, and future. *IEEE Aerospace and Electronic Systems Magazine*, 34(12):6–18, 2019.

[72] Esra Al Hadhrami, Maha Al Mufti, Bilal Taha, and Naoufel Werghi. Transfer learning with convolutional neural networks for moving target classification with micro-doppler radar spectrograms. In *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 148–154, 2018.

[73] M.H. Hayes. *Statistical Digital Signal Processing and Modeling*. Wiley India Pvt. Limited, 2009.

[74] Simon Haykin. Cognitive dynamic systems: Radar, control, and radio [point of view]. *Proceedings of the IEEE*, 100(7):2095–2103, 2012.

[75] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[76] J. Hill. Black asl. *Journal of American Sign Languages and Literatures*, 2012.

[77] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997.

[78] Amin Hong, Young-Hoon Chun, Sangyeol Oh, and Youngwook Kim. Human activity classification based on cognitive doppler radar to optimize carrier frequency and sampling rate using reinforcement learning. *IEEE Sensors Journal*, 24(2):1696–1705, 2024.

[79] Berthold K.P. Horn and Brian G. Schunck. Determining optical flow. Technical report, Massachusetts Institute of Technology, USA, 1980.

[80] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

[81] Wenqiang Hua, Shuang Wang, Wen Xie, Yanhe Guo, and Xiaomin Jin. Dual-channel convolutional neural network for polarimetric sar images classification. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 3201–3204, 2019.

[82] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2016.

[83] Xuejun Huang, Jinshan Ding, Dongxing Liang, and Liwu Wen. Multi-person recognition using separated micro-doppler signatures. *IEEE Sensors Journal*, 20(12):6605–6611, 2020.

[84] Zhongling Huang, Corneliu Octavian Dumitru, Zongxu Pan, Bin Lei, and Mihai Datcu. Classification of large-scale high-resolution SAR images with deep transfer learning. *IEEE Geoscience and Remote Sensing Letters*, 18(1):107–111, jan 2021.

[85] Zhongling Huang, Zongxu Pan, and Bin Lei. Transfer learning with deep convolutional neural network for sar target classification with limited labeled data. *Remote Sensing*, 9(9), 2017.

[86] Infineon. Radar system. RFS SDK Documentation, 2023.

[87] Cesar Iovescu and Sandeep Rao. The fundamentals of millimeter wave radar sensors. `https://www.tij.co.jp/jp/lit/wp/spyy005a/spyy005a.pdf`, 2020. Texas Instruments.

[88] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

[89] J. Huang, We. Zhou, H. Li, and W. Li. Sign language recognition using 3d convolutional neural networks. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2015.

[90] Fereshteh Jafariakinabad, Sansiri Tarnpradab, and Kien Hua. Syntactic recurrent neural network for authorship attribution. *FLAIRS-33*, 02 2019.

[91] Feng Jin, Arindam Sengupta, and Siyang Cao. mmfall: Fall detection using 4-d mmwave radar and a hybrid variational rnn autoencoder. *IEEE Transactions on Automation Science and Engineering*, 19(2):1245–1257, 2022.

[92] Branka Jokanovic, Moeness Amin, and Baris Erol. Multiple joint-variable domains recognition of human motion. In *2017 IEEE Radar Conference (RadarConf)*, pages 0948–0952, 2017.

[93] Branka Jokanović and Moeness Amin. Fall detection using deep learning in range-doppler radars. *IEEE Transactions on Aerospace and Electronic Systems*, 54(1):180–189, 2018.

[94] Michael I. Jordan. Chapter 25 - serial order: A parallel distributed processing approach. In John W. Donahoe and Vivian Packard Dorsel, editors, *Neural-Network Models of Cognition*, volume 121 of *Advances in Psychology*, pages 471–495. North-Holland, 1997.

[95] C. Karabacak, S. Z. Gurbuz, A. C. Gurbuz, et al. Knowledge exploitation for human micro-doppler classification. *IEEE Geoscience and Remote Sensing Letters*, 12(10):2125–2129, 2015.

[96] Youngwook Kim and Taesup Moon. Human detection and activity classification based on micro-doppler signatures using deep convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1):8–12, 2016.

[97] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[98] Edward S. Klima, Ovid J. L. Tzeng, Y.Y. A. Fok, Ursula Bellugi, David Corina, and Jeffrey G. Bettger. From sign to script: Effects of linguistic experience on perceptual categorization. *Journal of Chinese Linguistics Monograph Series*, pages 96–129, 1999.

[99] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[100] Emre Kurtoglu, Ali C. Gurbuz, Evie Malaia, Darrin Griffin, Chris Crawford, and Sevgi Z. Gurbuz. Sequential classification of asl signs in the context of daily living using rf sensing. In *2021 IEEE Radar Conference (RadarConf21)*, pages 1–6, 2021.

[101] Emre Kurtoğlu, Moeness G. Amin, and Sevgi Z. Gurbuz. Radar-based joint human activity and agility recognition via multi-input multi-task learning. In *2024 IEEE Radar Conference (Accepted)*, pages 1–6, 2024.

[102] Emre Kurtoğlu, Kenneth DeHaan, Caroline Kobek Pezzarossi, Darrin J. Griffin, Chris Crawford, and Sevgi Z. Gurbuz. Gamification of rf data acquisition for classification of natural human gestures. In *2024 IEEE Radar Conference (Accepted)*, pages 1–6, 2024.

[103] Emre Kurtoğlu, Sabyasachi Biswas, Ali C. Gurbuz, and Sevgi Zubeyde Gurbuz. Boosting multi-target recognition performance with multi-input multi-output radar-based angular subspace projection and multi-view deep neural network. *IET Radar, Sonar & Navigation*, 17(7):1115–1128, 2023.

[104] Emre Kurtoğlu, Kenneth DeHaan, Caroline Kobek Pezzarossi, Darrin J. Griffin, Chris Crawford, and Sevgi Zubeyde Gurbuz. Interactive learning of natural sign language with radar. *IET Radar, Sonar & Navigation (Accepted)*, 2024.

[105] Emre Kurtoğlu, Ali C. Gurbuz, Evie Malaia, Darrin Griffin, Chris Crawford, and Sevgi Z. Gurbuz. Rf micro-doppler classification with multiple spectrograms from angular subspace projections. In *2022 IEEE Radar Conference (RadarConf22)*, pages 1–6, 2022.

[106] Emre Kurtoğlu, Ali C. Gurbuz, Evie A. Malaia, Darrin Griffin, Chris Crawford, and Sevgi Z. Gurbuz. Asl trigger recognition in mixed activity/signing sequences for rf sensor-based user interfaces. *IEEE Transactions on Human-Machine Systems*, 52(4):699–712, 2022.

[107] Jihoon Kwon and Nojun Kwak. Human detection by neural networks using a low-cost short-range doppler radar sensor. In *2017 IEEE Radar Conference (RadarConf)*, pages 0755–0760, 2017.

[108] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Fractalnet: Ultra-deep neural networks without residuals. *CoRR*, abs/1605.07648, 2016.

[109] Vladimir Lekic and Zdenka Babic. Automotive radar and camera fusion using generative adversarial networks. *Computer Vision and Image Understanding*, 184:1–8, 2019.

[110] Hong Li et al. Wifinger: Talk to your smart devices with finger-grained gesture. In *Proc. ACM UbiComp*, page 250–261, 2016.

[111] Jiashi Li, Xin Xia, Wei Li, Huixia Li, Xing Wang, Xuefeng Xiao, Rui Wang, Min Zheng, and Xin Pan. Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios, 2022.

[112] Jaime Lien et al. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Trans. Graph.*, 35(4), July 2016.

[113] Jau Jr Lin, Yuan Ping Li, Wei Chiang Hsu, and Ta Sung Lee. Design of an fmcw radar baseband signal processing system for automotive application. *SpringerPlus*, 5:1–16, 12 2016.

[114] Haipeng Liu, Anfu Zhou, Zihe Dong, Yuyang Sun, Jiahe Zhang, Liang Liu, Huadong Ma, Jianhua Liu, and Ning Yang. M-gesture: Person-independent real-time in-air gesture recognition using commodity millimeter wave radar. *IEEE Internet of Things Journal*, pages 1–1, 2021.

[115] F. Long, H. Peng, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(08):1226–1238, aug 2005.

[116] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. Signfi: Sign language recognition using wifi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1), March 2018.

[117] E. Malaia. It still isn't over: Event boundaries in language and perception. *Languistics Compass*, 8(3):89–98, 2014.

[118] E. Malaia, J. D. Borneman, and R. B. Wilbur. Assessment of information content in visual signal: analysis of optical flow fractal complexity. *Visual Cognition*, 24(3):246–251, 2016.

[119] E. Malaia, J.D. Borneman, and R.B. Wilbur. Information transfer capacity of articulators in american sign language. *Language and speech*, 61(1):97–112, 2018.

[120] E. Malaia and R. B. Wilbur. Kinematic signatures of telic and atelic events in asl predicates. *Lang Speech*, 55(3):407–421, 2012.

[121] E. A. Malaia and R. B. Wilbur. Syllable as a unit of information transfer in linguistic communication: The entropy syllable parsing model. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2019.

[122] Evie Malaia. Current and future methodologies for quantitative analysis of information transfer in sign language and gesture data. *Behavioral and Brain Sciences*, 40, 2017.

[123] Evie Malaia, Joshua D Borneman, and Ronnie B Wilbur. Assessment of information content in visual signal: analysis of optical flow fractal complexity. *Visual Cognition*, 24(3):246–251, 2016.

[124] Evie A. Malaia, Joshua D. Borneman, Emre Kurtoglu, Sevgi Z. Gurbuz, Darrin Griffin, Chris Crawford, and Ali C. Gurbuz. Complexity in sign languages. *Linguistics Vanguard*, 9(s1):121–131, 2023.

[125] Evie A Malaia and Ronnie B Wilbur. Syllable as a unit of information transfer in linguistic communication: The entropy syllable parsing model. *Wiley Interdisciplinary Reviews: Cognitive Sci.*, 11(1):e1518, 2020.

[126] Jan Matuszewski. Applying the decision trees to radar targets recognition. In *11-th INTERNATIONAL RADAR SYMPOSIUM*, pages 1–4, 2010.

[127] Jim Mccambridge, John Witton, and Diana Elbourne. Systematic review of the hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology*, 67:267 – 277, 2014.

[128] Pedro Melgarejo et al. Leveraging directional antenna capabilities for fine-grained gesture recognition. In *Proc. ACM UbiComp.*, page 541–551, 2014.

[129] Ramin Nabati and Hairong Qi. Rrpn: Radar region proposal network for object detection in autonomous vehicles. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3093–3097, 2019.

[130] Alexandros Ninos, Jürgen Hasch, and Thomas Zwick. Real-time macro gesture recognition using efficient empirical feature extraction with millimeter-wave technology. *IEEE Sensors Journal*, 21(13):15161–15170, 2021.

[131] Giacomo Paterniani, Daria Sgreccia, Alessandro Davoli, Giorgio Guerzoni, Pasquale Di Viesti, Anna Chiara Valenti, Marco Vitolo, Giorgio M. Vitetta, and Giuseppe Boriani. Radar-based monitoring of vital signs: A tutorial overview. *Proceedings of the IEEE*, 111(3):277–317, 2023.

[132] Jacopo Pegoraro, Francesca Meneghello, and Michele Rossi. Multiperson continuous tracking and identification from mm-wave micro-doppler signatures. *IEEE Transactions on Geoscience and Remote Sensing*, 59(4):2994–3009, 2021.

[133] Jifang Pei, Yulin Huang, Weibo Huo, Yin Zhang, Jianyu Yang, and Tat-Soon Yeo. Sar automatic target recognition based on multiview deep learning framework. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2196–2210, 2018.

[134] M. Mahbubur Rahman, Sevgi Z. Gurbuz, and Moeness G. Amin. Physics-aware design of multi-branch gan for human rf micro-doppler signature synthesis. In *2021 IEEE Radar Conference (RadarConf21)*, pages 1–6, 2021.

[135] M. Mahbubur Rahman, Emre Kurtoglu, Robiulhossain Mdrafi, Ali C. Gurbuz, Evie Malaia, Chris Crawford, Darrin Griffin, and Sevgi Z. Gurbuz. Word-level asl recognition and trigger sign detection with rf sensors. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8233–8237, 2021.

[136] M. Mahbubur Rahman, Emre Kurtoğlu, Muhammet Taskin, Kudret Esme, Ali C. Gurbuz, Evie Malaia, and Sevgi Z. Gurbuz. Performance comparison of radar and video for american sign language recognition. In *2022 IEEE Radar Conference (RadarConf22)*, pages 1–6, 2022.

[137] M.M. Rahman, S.Z. Gurbuz, and M.G. Amin. Physics-aware generative adversarial networks for radar-based human activity recognition. *IEEE Transactions on Aerospace and Electronic Systems*, pages 1–15, 2022.

[138] Mohammad Mahbubur Rahman, Evie A. Malaia, Ali Cafer Gurbuz, Darrin J. Griffin, Chris Crawford, and Sevgi Zubeyde Gurbuz. Effect of kinematics and fluency in adversarial synthetic data generation for asl recognition with rf sensors. *IEEE Transactions on Aerospace and Electronic Systems*, 58(4):2732–2745, 2022.

[139] Sandeep Rao. Mimo radar. `https://www.ti.com/lit/an/swra554a/swra554a.pdf`, 2018. Texas Instruments.

[140] Sandeep Rao. Introduction to mmwave sensing: Fmcw radars. `https://www.ti.com/content/dam/videos/external-videos/2/3816841626001/5415528961001.mp4/subassets/mmwaveSensing-FMCW-offlineviewing_0.pdf`, 2020. Texas Instruments.

[141] Aparna Rathi, Debasish Deb, N. Sarath Babu, and Reena Mamgain. Two-level classification of radar targets using machine learning. In Yu-Dong Zhang, Jyotsna Kumar Mandal, Chakchai So-In, and Nileshsingh V. Thakur, editors, *Smart Trends in Computing and Communications*, pages 231–242, Singapore, 2020. Springer Singapore.

[142] M.A. Richards. *Fundamentals Of Radar Signal Processing*. McGraw-Hill Education (India) Pvt Limited, 2005.

[143] MARC ROBASZKIEWICZ. Saving lives when temperatures rise. `https://www.nxp.com/docs/en/white-paper/RADAR-CHILD-DETECTION-WP.pdf`.

[144] R. Roy and T. Kailath. Esprit-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7):984–995, 1989.

[145] David E. Rumelhart and James L. McClelland. *Learning Internal Representations by Error Propagation*, pages 318–362. MIT Press, 1987.

[146] Christian Schuessler, Wenxuan Zhang, Johanna Bräunig, Marcel Hoffmann, Michael Stelzig, and Martin Vossiek. Radar-based recognition of static hand gestures in american sign language, 2024.

[147] Rick Searcy. Machine learning takes automotive radar further. `https://www.aptiv.com/docs/default-source/white-papers/2020_aptiv_whitepaper_machinelearning_radar.pdf?sfvrsn=dd3a9a3e_6`.

[148] Ann-Kathrin Seifert, Abdelhak M. Zoubir, and Moeness G. Amin. Radar classification of human gait abnormality based on sum-of-harmonics analysis. In *2018 IEEE Radar Conference (RadarConf18)*, pages 0940–0945, 2018.

[149] Mehmet Saygin Seyfioglu, Baris Erol, Sevgi Zubeyde Gurbuz, and Moeness G. Amin. Dnn transfer learning from diversified micro-doppler for motion classification. *IEEE Transactions on Aerospace and Electronic Systems*, 55(5):2164–2180, 2019.

[150] Mehmet Saygın Seyfioğlu, Ahmet Murat Özbayoğlu, and Sevgi Zubeyde Gürbüz. Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities. *IEEE Transactions on Aerospace and Electronic Systems*, 54(4):1709–1723, 2018.

[151] Ronghua Shang, Jiaming Wang, Licheng Jiao, Rustam Stolkin, Biao Hou, and Yangyang Li. Sar targets classification based on deep memory convolution neural networks and transfer parameters. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(8):2834–2846, 2018.

[152] Marcel Sheeny, Andrew Wallace, and Sen Wang. 300 ghz radar object recognition based on deep neural networks and transfer learning, 2019.

[153] Cheng-Che Shih, Xinrui Zhou, Thinh Nguyen, and Khanh Pham. People counting system using mmwave mimo radar with 3d convolutional neural network. In *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, pages 1–5, 2023.

[154] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[155] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[156] Graeme E. Smith, Karl Woodbridge, and Chris J. Baker. Radar micro-doppler signature classification using dynamic time warping. *IEEE Transactions on Aerospace and Electronic Systems*, 46(3):1078–1096, 2010.

[157] Hoyeol Sohn. Google - isolated sign language recognition. `https://www.kaggle.com/code/hoyso48/1st-place-solution-training`, 2023. Accessed: Sep. 13, 2023.

[158] Qian Song, Feng Xu, and Ya-Qiu Jin. Sar image representation learning with adversarial autoencoder networks. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 9498–9501, 2019.

[159] Michael Stephan, Thomas Stadelmayer, Avik Santra, Georg Fischer, Robert Weigel, and Fabian Lurz. Radar image reconstruction from raw adc data using parametric variational autoencoder with domain adaptation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9529–9536, 2021.

[160] C. E. Stilp and K. R. Kluender. Stimulus statistics change sounds from near-indiscriminable to hyperdiscriminable. *PLOS ONE*, 11(8), 2016.

[161] W. C. Stokoe. Studies in linguistics: Occasional papers 8. *Sign Language Structure: An Outline of the Visual Communication System of the American Deaf*, 1960.

[162] C. Sun, T. Zhang, and C. Xu. Latent support vector machine modeling for sign language recognition with kinect. *ACM Trans. Intell. Syst. Technol.*, 6:20:1–20:20, 2015.

[163] Yingxiang Sun, Haoqiu Xiong, Danny Kai Pin Tan, Tony Xiao Han, Rui Du, Xun Yang, and Terry Tao Ye. Moving target localization and activity/gesture recognition for indoor radio frequency sensing applications. *IEEE Sensors Journal*, pages 1–1, 2021.

[164] Yuliang Sun, Tai Fei, Xibo Li, Alexander Warnecke, Ernst Warsitz, and Nils Pohl. Real-time radar-based gesture detection and recognition built in an edge-computing platform. *IEEE Sensors Journal*, 20(18):10706–10716, 2020.

[165] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[166] Siri Team. Hey siri: An on-device dnn-powered voice trigger for apple's personal assistant. Technical report, Apple, October 2017.

[167] T. Tsai and P. Hao. Customized wake-up word with key word spotting using convolutional neural network. In *2019 International SoC Design Conference (ISOCC)*, pages 136–137, 2019.

[168] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer, 2022.

[169] Emin Ucer, Emre Kurtoglu, Mithat Kisacikoglu, Ali C. Gurbuz, and Sevgi Z. Gurbuz. Local detection of oltc operation to support decentralized control of active end-nodes. In *2022 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5, 2022.

[170] Y. Vaezi and M. Van der Baan. Comparison of the sta/lta and power spectral density methods for microseismic event detection. *Geophysical Journal International*, 203(3):1896–1908, 2015.

[171] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

[172] Shelly Vishwakarma and Shobha Sundar Ram. Classification of multiple targets based on disaggregation of micro-doppler signatures. In *2016 Asia-Pacific Microwave Conference (APMC)*, pages 1–4, 2016.

[173] Li Wang, Jun Tang, and Qingmin Liao. A study on radar target detection based on deep neural networks. *IEEE Sensors Letters*, 3(3):1–4, 2019.

[174] Yanhua Wang, Wei Chen, Jia Song, Yang Li, and Xiaopeng Yang. Open set radar hrrp recognition based on random forest and extreme value theory. In *2018 International Conference on Radar (RADAR)*, pages 1–4, 2018.

[175] Yizhe Wang, Yongshun Zhang, Cunqian Feng, Bin Chen, and Qichao Ge. Micro-doppler separation of multi-target based on aco in midcourse. *The Journal of Engineering*, 2019(19):5967–5970, 2019.

[176] Yizhou Wang, Zhongyu Jiang, Xiangyu Gao, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. Rodnet: Object detection under severe conditions using vision-radio cross-modal supervision. *CoRR*, abs/2003.01816, 2020.

[177] Z. Wang, Z. Yu, X. Lou, B. Guo, and L. Chen. Gesture-radar: A dual doppler radar based system for robust recognition and quantitative profiling of human gestures. *IEEE Transactions on Human-Machine Systems*, 51(1):32–43, 2021.

[178] Ronnie B Wilbur and Evguenia Malaia. Contributions of sign language research to gesture understanding: What can multimodal computational systems learn from sign language research. *International journal of semantic computing*, 2(01):5–19, 2008.

[179] Dongxian Wu, Yisen Wang, and Shutao Xia. Revisiting loss landscape for adversarial robustness. *CoRR*, abs/2004.05884, 2020.

[180] Zhaoyang Xia, Genming Ding, Hui Wang, and Feng Xu. Person identification with millimeter-wave radar in realistic smart home scenarios. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[181] Feng Xu, Haipeng Wang, and Y. Jin. Deep learning as applied in sar target recognition and terrain classification. *Journal of Radars*, 6:136–148, 04 2017.

[182] Xiuzhu Yang, Wenfeng Yin, Lei Li, and Lin Zhang. Dense people counting using ir-uwb radar with a hybrid feature extraction method. *IEEE Geoscience and Remote Sensing Letters*, 16(1):30–34, 2019.

[183] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *CoRR*, abs/1708.02709, 2017.

[184] Zhengxin Zeng, Moeness G. Amin, and Tao Shan. Arm motion classification using time-series analysis of the spectrogram frequency envelopes. *Remote Sensing*, 12(3), 2020.

[185] Yikui Zhai, Wenbo Deng, Ying Xu, Qirui Ke, Junying Gan, Bing Sun, Junying Zeng, and Vincenzo Piuri. Robust sar automatic target recognition based on transferred ms-cnn with l 2 -regularization. *Computational Intelligence and Neuroscience*, 2019:1–13, 11 2019.

[186] Fan Zhang, Yunchong Wang, Jun Ni, Yongsheng Zhou, and Wei Hu. Sar target small sample recognition based on cnn cascaded features and adaboost rotation forest. *IEEE Geoscience and Remote Sensing Letters*, 17(6):1008–1012, 2020.

[187] Renyuan Zhang and Siyang Cao. Support vector machines for classification of automotive radar interference. In *2018 IEEE Radar Conference (RadarConf18)*, pages 0366–0371, 2018.

[188] Wei Zhang, Yongfeng Zhu, and Qiang Fu. Semi-supervised deep transfer learning-based on adversarial feature learning for label limited sar target recognition. *IEEE Access*, 7:152412–152420, 2019.

[189] Zhenyuan Zhang, Zengshan Tian, and Mu Zhou. Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor. *IEEE Sensors Journal*, 18(8):3278–3289, 2018.

[190] Ce Zheng, Xue Jiang, and Xingzhao Liu. Semi-supervised sar atr via multi-discriminator generative adversarial network. *IEEE Sensors Journal*, 19(17):7525–7533, 2019.

[191] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017.

APPENDIX A

IRB APPROVAL LETTERS FOR HUMAN SUBJECT TESTING

THE UNIVERSITY OF ALABAMA® | Research & Economic Development
Office for Research Compliance

August 11, 2023

To: Sevgi Zubeyde Gurbuz, PhD
Assistant Professor
Department of Electrical and Computer Engineering
College of Engineering
The University of Alabama
Box 870286

From: Edward M. Shirley, MA, CIP
Interim IRB Team Lead

Re: **Notice of Approval**

IRB Application #:	e-Protocol 23-04-6553
Project Title:	"Radio-Frequency (RF) / Radar-Based Studies on Cyber-Physical Human Systems (CPHS)"
Submission Type:	New
Approval Date:	August 11, 2023
Expiration Date:	August 10, 2024
Funding Source:	NSF /22-1046- 23-0654
Review Category:	Expedited
Approved Documents:	Informed Consent, Recruitment Script/Flyer

Dear Dr. Gurbuz:

The University of Alabama Institutional Review Board has approved your proposed research. Therefore, your application has been approved according to 45 CFR part 46 *as outlined below:*

(7) Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

The approval for your application will lapse, as noted above. If your research will continue beyond this date, please submit the Continuing Review to the IRB as University policy requires before the lapse. Please note any modifications made in research design, methodology, or procedures must be submitted to and approved by the IRB before implementation. Please submit a final report form when the study is complete.

All the best with your research.

166 Rose Administration | Box 870127 | Tuscaloosa, AL 35487-0127 | 205-348-8461
Fax 205-348-7189 | Toll Free 1-877-820-3066 | rscompliance@ua.edu

THE UNIVERSITY OF ALABAMA® | Research & Economic Development
Office for Research Compliance

June 22,2023

To: Sevgi Zubeyde Gurbuz, PhD
Assistant Professor
Department of Electrical and Computer Engineering
College of Engineering
The University of Alabama
Box 870286

From: Carpantato T. Myles, MSM, CIM, CIP
Director & Research Compliance Officer

Re: **Notice of Approval**

IRB Application #:	e-Protocol 18-OR-364-ME-R5-A (18-06-1271)
Project Title:	"Radar-Based Indoor Human Motion Recognition Studies"
Submission Type:	Revision (Amendment, Modification)
Approval Date:	June 22, 2023
Expiration Date:	June 7, 2024
Funding Source:	NSF(OSP#19-0652), AFOSR(OSP#22-0615), ECE and CS Departmental Funding
Review Category:	Expedited
Approved Documents:	Informed Consent, Recruitment Email/Flyer

Dear Dr. Gurbuz:

The University of Alabama Institutional Review Board has reviewed the revision to your previously approved expedited protocol. The board has determined that the change does not affect the expedited status of your protocol.

Should you need to submit any further correspondence regarding this proposal, please include the assigned IRB application number. Changes in this study cannot be initiated without IRB approval, except when necessary to eliminate apparent immediate hazards to participants.

All the best with your research.

166 Rose Administration | Box 870127 | Tuscaloosa, AL 35487-0127 | 205-348-8461
Fax 205-348-7189 | Toll Free 1-877-820-3066 | rscompliance@research.ua.edu