ASL Trigger Recognition in Mixed Activity/Signing Sequences for RF Sensor-Based User Interfaces

Emre Kurtoğlu, Ali C. Gurbuz, *Senior Member, IEEE*, Evie A. Malaia, Darrin Griffin, Chris Crawford, and Sevgi Z. Gurbuz, *Senior Member, IEEE*

***Abstract*—The past decade has seen great advancements in speech recognition for control of interactive devices, personal assistants, and computer interfaces. However, deaf and hard-of-hearing (HoH) individuals, 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. RF sensors have been recently proposed as a new modality that can be effective under the circumstances where video is not. This article considers the problem of recognizing a trigger sign (wake word) in the context of daily living, where gross motor activities are interwoven with signing sequences. The proposed approach exploits multiple RF data domain representations (time-frequency, range-Doppler, and range-angle) for sequential classification of mixed motion data streams. The recognition accuracy of signs with varying kinematic properties is compared and used to make recommendations on appropriate trigger sign selection for RF-sensor-based user interfaces. The proposed approach achieves a trigger sign detection rate of 98.9% and a classification accuracy of 92% for 15 ASL words and three gross motor activities.**

***Index Terms*—American sign language (ASL), gesture recognition, human-computer interaction, sign language, trigger detection, wake word.**

I. Introduction

THE past decade has seen great advancements in sensing for ambient intelligence, including speech recognition for control of interactive devices, personal assistants, and human–computer interfaces. However, deaf and hard-of-hearing (HoH) individuals, whose primary mode of communication is sign language, cannot benefit from voice-controlled interfaces. Research in recognition of American sign language (ASL) has focused primarily on wearable sensors [1]–[3], optical cameras [3]–[5], and infrared depth sensors [6], [7]. Wearables, such as "signing" gloves embedded with inertial measurement units (IMUs), or surface elecromyography (sEMG) sensors have yielded relatively higher recognition accuracy, but inhibit natural motion, and are thus not highly preferred by members of the deaf community [8].

Video cameras are perhaps the most often used device by deaf/HoH individuals for interpersonal communications. RGB-D cameras, which add depth measurements to video recordings for the purposes of skeleton tracking, such as Kinect or leap motion sensors, have improved recognition accuracy relative to video-only approaches. However, RGB-D cameras are ineffective under low illumination and may invade privacy through the acquisition of personal imagery of the face and environment.

RF sensors have been recently proposed [9]–[12] as a new modality for ASL recognition that has the capability of measurning human kinematics through fine range, angle, and velocity measurements. RF sensors are also effective in the dark and do not make any visual recordings of the people or environment. RF sensors, also known as radar, which is short for *radio detection and ranging*, acquire independent measurements of distance, velocity, and angle. Using a technique known as *stretch processing* [13], the frequency difference between the transmitted and received frequency modulated continuous wave (FMCW) signals can be used to compute the round-trip travel time of the signal, and, hence, distance to an object. The Doppler shift, on the other hand, relates radial velocity to the frequency shift in the received signal. Rotations or vibrations result in additional Doppler frequency modulations, known as micro-Doppler (μD) frequencies, centered around the main Doppler shift due to translational motion [14]. The μD signature is a 2D time-frequency representation of the RF data, which reveals the unique kinematics of the observed motion. Thus, μD has been exploited as a biometric [15] for recognizing individuals [16], activities [17], aided/unaided walking [18], falls [19], [20], and even different gaits [21].

Although RF sensors cannot effectively perceive facial expressions or hand shape, radar does provide data that is complementary to that of video: While video is effective in capturing spatial parameters, radar is more adept at capturing temporal or dynamic parameters. This is because radar measurements of distance and velocity are based on independent physics-based

Manuscript received July 15, 2021; revised October 12, 2021; accepted November 6, 2021. Date of publication December 22, 2021; date of current version July 14, 2022. This work was supported in part by the National Science Foundation Awards 1932547, 1931861, and 1734938. This article was recommended by Associate Editor B. Guo. *(Corresponding author: Sevgi Z. Gurbuz.)*

Emre Kurtoğlu and Sevgi Z. Gurbuz are with the Department of Electrical and Computer Engineering, University of Alabama, Tuscaloosa, AL 35487 USA (e-mail: ekurtoglu@crimson.ua.edu; szgurbuz@ua.edu).

Evie A. Malaia is with the Department of Communication Disorders, University of Alabama, Tuscaloosa, AL 35487 USA (e-mail: eamalaia@ua.edu).

Ali C. Gurbuz is with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762 USA (e-mail: gurbuz@ece.msstate.edu).

Darrin Griffin is with the Department of Communication Studies, University of Alabama, Tuscaloosa, AL 35487 USA (e-mail: djgriffin1@ua.edu).

Chris Crawford is with the Department of Computer Science, University of Alabama, Tuscaloosa, AL 35487 USA (e-mail: crawford@cs.ua.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/THMS.2021.3131675.

Digital Object Identifier 10.1109/THMS.2021.3131675

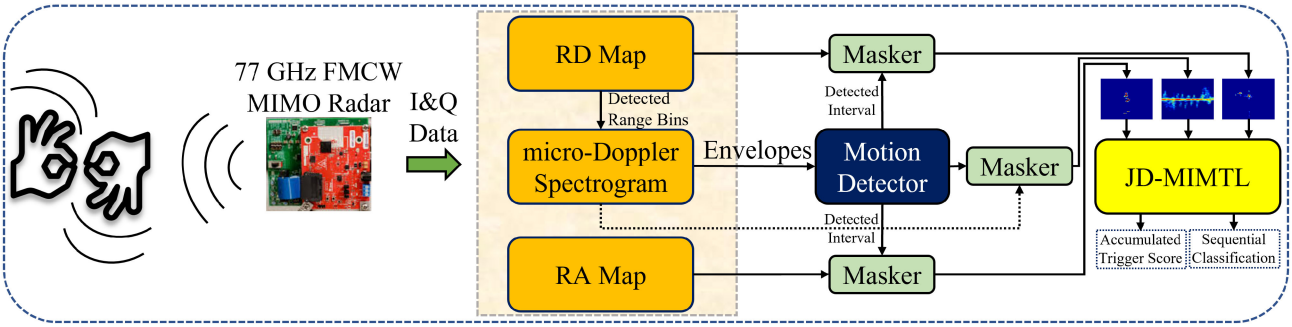


Fig. 1. Flowchart for the proposed approach.

measurements: Distance is computed from round-trip travel time, while velocity is computed from the Doppler shift, which has greater accuracy than the computation of displacement over time. Thus, although RF sensors cannot be used to reconstruct facial expressions or hand shape, radar can provide a new way of recognizing signs in a noncontact, ambient fashion based primarily on signing kinematics and range profiles.

As a result, there has been much research on the use of RF sensing for hand gesture recognition [22], [23], especially since the development of low-cost, low-power, high resolution, integrated millimeter wave RF transceivers [24]. 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 work represents the first to consider triggering and command recognition of RF-sensor enabled devices under more realistic conditions, where the RF data are 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 joint-domain, multiinput–multitask learning (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. Fig. 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 recognize a sequence of three activities and 15 ASL signs with 92% accuracy, while detecting trigger signs with rates as high as 98.9%.

In Section II, an overview of the current state-of-the art and work related to radar-based gesture and sign language recognition is given. In Section III, we describe the RF sensor utilized, experiments conducted, and preprocessing algorithms applied to the data. In Section IV, kinematic and replicability considerations are applied to select 15 out of a total of 110 measured ASL signs as example trigger signs. Next, in Section V, a motion detection method is presented and its efficacy on the acquired datasets demonstrated for temporal segmentation. In Section VI, the proposed JD-MIMTL framework for sequential classification is detailed. Results demonstrating the performance of the proposed approach for trigger word detection and sequential recognition are discussed in Section VII. Discussions of key conclusions and future work is given in Section VIII.

II. Related Work

A. Radar-Based Activity and Gesture Recognition

A variety of deep learning approaches [25] have been leveraged for human motion recognition with RF sensors. Most approaches consider either daily activities (e.g., walking, sitting, running) or hand gestures (e.g., left/right, up/down swiping, push buttons). Recurrent neural networks (RNNs) have been proposed in many works, but results have been primarily demonstrated on fixed-duration snapshots that include just one class of motion. For example, the author in [26] applied stacked gated RNNs to 2D micro-Doppler signatures, while [27] constructs a 3D data representation from shifted windows of the micro-Doppler signature, applying both a 3D convolutional neural network (CNN) and long-short-term memory (LSTM). A more common approach is to use a time series of range-Doppler maps as input [28], [29] to the 3D CNN-LSTM network, while also using connectionist temporal classification (CTC) [27] and triplet loss [30].

Studies considering recognition performance in real-world conditions are limited; most of the aforementioned works involve experiments conducted in controlled environments with participants positioned at a fixed distance from the RF sensor. Work involving real-world use cases have considered the effects of sensor positioning and environment. For example, [23], investigates the dependence of performance on sensor position at different heights from the ground and distance from the user. Dynamic time warping is used on RF data acquired from a dual-Doppler radar to accurately recognize 12 different gestures with at least 80% accuracy, depending on positioning. These results are consistent with expectations based on the radar range equation, which show that the received power decreases with distance (R) as $1/R^4$. Indoor environments also tend to have comparable stationary clutter properties, and, hence do not

result in significant variations in performance. The environment-independence of RF sensing of ASL was verified in a recent study [31], which found the recognition accuracy across several different rooms to be comparable. Another important factor in gesture recognition is the upper body movements, which can change the received RF signal from gestures. In [32], a single transmitter, dual receiver RF transceiver was proposed to decouple hand gestures from random body movements, and thereby improve gesture recognition accuracy.

An important real-world use case that is gaining increasing attention is the challenge of sequential motion recognition. Human movement is inherently dynamic, greatly varied, and sequential in nature. The continuous data streams acquired by RF sensors in real-world environments will consist of an intertwining of gross body activities and finer movements, such as gestures. But, most works on human activity recognition consider either daily activities or fine-grain gestures, while the approaches proposed reflect a similar technique to that applied over fixed-duration snapshots. For example, the authors in [33] applies bi-LSTM networks to continuous sequences of micro-Doppler signatures, while [34] applies RNNs to continuous streams of 3D inputs formed from the time series of range-Doppler maps.

B. Radar-Based ASL Recognition

Radar-based ASL recognition to-date has primarily focused on the recognition of snapshots of specific words or phrases. In [35], ten (10) different ASL phrases that would be relevant to emergency response were recognized with an accuracy of 95% using transfer learning from VGG-16 to classify X-band micro-Doppler signatures. In [12], feature-level fusion of RF sensors operating at three different transmit frequencies (10, 24, and 77 GHz) were used together with a random forest classifier trained only an measured micro-Doppler signatures from fluent ASL signers yielded a classification accuracy of 72.5% for 20 ASL signs. Moreover, using a support vector machine (SVM) classifier, it was shown that the ASL articulations of hearing imitation signers were distinguishable from that of fluent ASL users. With the use of a multimodal DNN for fusion [36], the classification accuracy for 20 signs was improved to 95.5% and shown to surpass by 22% and 19% the accuracy given from use of a single RF sensor classified using transfer learning from VGG-16 and unsupervised pretraining with convolutional autoencoders (CAEs), respectively.

In [31], micro-Doppler signatures of 50 ASL signs are classified with an average accuracy of 87%, while the sign KNOCK is specified as a wake word and detected at a rate of 94% using a fixed-window binary DNN classifier. Word-level ASL recognition with RF sensors is shown to be tolerant to the presence of other interfering users, different user positions and different environments. These results were achieved by collecting over 12-k samples from 15 different participants.

Because of the differences in fine-grained temporal dynamics and linguistic parameters, such as prosody and grammatical structure, the RF data acquired from hearing imitation signers versus fluent ASL users are actually quite different. In [36], it is shown that imitation signing cannot be used to train classifiers of fluent signers. To overcome this challenge, adversarial learning has been proposed [37] to 1) adapt imitation signing data to resemble that of fluent signers, and 2) synthesize kinematically accurate samples for training DNNs. This approach has yielded over 77% top-1 and over 93% top-5 accuracy for recognition of 100 ASL signs using micro-Doppler signatures acquired from a 77-GHz RF sensor.

Note that all of the above works classify fixed-duration snapshots of micro-Doppler signatures of ASL. Thus, this work fills an important gap in current literature by addressing the challenge of trigger sign detection and sequential ASL recognition in continuous RF data streams of mixed motion sequences that are typical of daily living.

III. RF Data Acquisition and Preprocessing

A. 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 frequency modulated continuous wave (FMCW) short-range automotive radar that has two transmit (TX) channels and four receive (RX) channels, 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 m from the ground.

B. Participants

Although ASL has been used as example motions in some gesture recognition studies [38], [39], sign language greatly differs from gesturing in that it possesses a much greater degree of physical complexity and Shannon information [40]–[42]. Like other complex system-generated signals, raw physical signal from signing data contains information at multiple timescales, spanning phonological, semantic, syntactic, and prosodic cues ([43], [44]).

While some studies [45], [46] have utilized imitation signers—hearing participants who mimic signs observed in video—it has been shown [47] 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 [12], [36], we have found that imitation signing is distinguishable from native signing using classification of RF μD signatures.

Thus, in this work, 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-m 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 adults (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

HOT	BOOK	SOON	WEEK	TODAY	MAYBE	FRIEND	TOMORROW
BED	THIS	SHOP	LONG	READY	MONEY	SCHOOL	LET ME SEE
WHY	HOME	LIKE	PUSH	DRINK	NIGHT	COFFEE	OH, I SEE
PET	MORE	YOUR	COME	WHERE	TABLE	CHANGE	I IOVE YOU
ONE	GOOD	HELP	WATER	BRING	RIGHT	BETTER	SOMETHING
GAS	MUST	HAVE	SLEEP	MONTH	THERE	FATHER	THANK YOU
CAN	WHAT	WALK	TEACH	THREE	KNIFE	PLEASE	BREAKFAST
WANT	HOLD	FINE	TIRED	WRONG	EARTH	SHOULD	DON'T LIKE
DEAF	SHOES	READ	CITY	HELLO	THRILLED	ALWAYS	NOTHING
TIME	WORK	LICENSE	WRITE	KITCHEN	HOSPITAL	TEACHER	ENGINEER
TIE UP	PEOPLE	WINTER	FAMILY	TOILET	LAWYER	HEALTH	TECHNOLOGY
COOK	PAPER	AGAIN	ME	SUMMON	EVENING	GO AHEAD	EXPLANATION
SEE	CAR	EAT	YES	EXCITED	MOTHER	MOUNTAIN	DOESN'T MATTER
	OK	HE	MY	GO	YOU	MORNING	

Fig. 2. Experimental setup (left) for acquisition of ASL signs (listing on the right).

TABLE I
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

CODAs and 2 ASL learners, who were not used in acquisition of training samples.

C. RF Datasets

A total of two different datasets were acquired as follows.

1) *Single ASL Signs*: 110 of the more frequently used ASL signs were selected from the ASL-LEX database [48], including nouns, verbs, and adjectives. A complete listing of the signs acquired is given in Fig. 2. 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 (see Section IV). 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 I. 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 three 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 have been made available for download.[1]

D. Transmit Waveform Parameters

The raw data provided by each receive channel of the RF sensor is a time stream of complex in-phase (I) and quadrature (Q) data. The presence of multiple receive channels enables not only the extraction of range and velocity but also the direction (or angle) of arrival of the received signals. The 77-GHz TI transceiver has 2 TX and 4 RX channels, which forms a uniform linear array (ULA). If, instead, the transmitters send two different chirp combinations, binary phase modulation (BPM) can be used to form a virtual array that behaves like a single TX, but 8 RX channel transceiver. To accomplish this, in the first chirp, C_a, we send exactly the same chirps from both transmitters, TX1 and TX2, with phase of $\Phi = 0°$, while in the second chirp, C_b, TX2 transmits with phase of $\Phi = 180°$. The chirps of TX1 and TX2 then can be retrieved by $C_1 = (C_a + C_b)/2$ and $C_2 = (C_a - C_b)/2$, respectively. In this way, we obtain a phase-shift for our second transmitter as well and synthesize a virtual array. An illustration of the transmitted chirp waveforms are provided in Fig. 3(a), while the resulting actual and virtual ULAs are shown in Fig. 3(b).

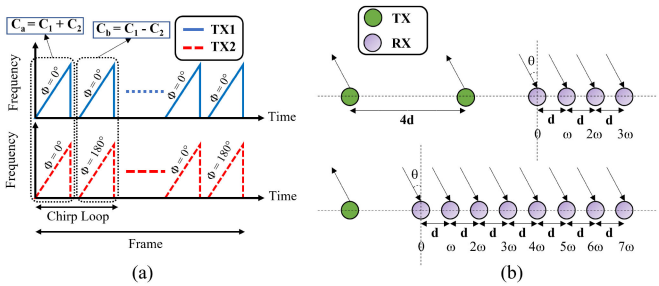


Fig. 3. (a) BPM chirp configuration and (b) virtual array synthesis.

For a target at angle θ, the phase difference between receiver channels will be as follows:

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

The angular resolution, θ_{res}, is given by

$$\theta_{\text{res}} = \frac{\lambda}{M \times d \times \cos(\theta)} \tag{2}$$

where M is the number of channels, so the doubling of M from 4 (real) to 8 (virtual) using BPM improves the angular resolution by a factor of 2. Thus, the 77-GHz TI transceiver was set to operate in BPM mode with a bandwidth of 4 GHz, pulse

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

repetition frequency (PRF) of 6.4 kHz with 256 samples per pulse and a coherent processing interval (CPI) of 40 ms.

E. RF Data Representations

Typically, the received I/Q data stream from each channel is reshaped into a 3D array, known as the *radar data cube*, with dimensions of fast-time (the number of ADC samples) $\times$ slow-time (the number of pulses) $\times$ channels. From the radar data cube, several different ways of representing the information acquired by the radar may be formed. The fast Fourier transform (FFT) across fast-time can be used to find the frequency difference, f_b, between the transmitted and received signals at any instant of time. If the chirp rate of the frequency modulated waveform is γ, then the distance, R, between the radar and scatterer can be found as $R = cf_b/2\gamma$, where c is the speed of light. An FFT across slow-time reveals the velocity of moving scatterers, $v = cf_d/2f_t$, where f_d is the Doppler shift and f_t is the transmit frequency.

Thus, several different 2D data representations may be computed from the radar data cube as follows.

1) *Range-Doppler Map*: The 2D FFT of the slow-time/fast-time data matrix for a single channel can be computed to find a range-Doppler image for each CPI. Because an RD map is computed from all the received returns acquired over a CPI, some researchers have adapted terminology from video processing and refer to the RD map as a *frame* and the CPI as the *frame* duration. Time series of RD maps can be formed to form RD videos. With a CPI comprised of 256 pulses, the resulting video as a frame rate of 25 fps (1/40 ms).
2) *Time-Frequency (Micro-Doppler) Map*: While there are many time-frequency transforms that yield the μD frequency versus time, the most often used is the spectrogram [14], which is the square modulus of the short-time Fourier transform (STFT) across slow-time. In order to generate μD spectrograms independent of the subjects' range, cell averaging constant false alarm rate (CA-CFAR) is applied on RD maps for detection of range bins with motion. Detected range bins are then used to generate the spectrograms.
3) *Range-Angle Map*: Angle can be computed from multiple-channel data using a beamforming method, e.g., multiple signal classification (MUSIC), to determine the angle of arrival of the received returns at a specific range and Doppler. Repeating this process for each CPI yields a time-series of RA maps, i.e., RA videos.

The visibility of target-related motion in the RA maps 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 [49] and take its elementwise 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 Fig. 4.

IV. 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 fingerspelling 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 subculture in the South [50]. 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.

Thus, this article 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 handshape 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 fingerspelled 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 nonnative ASL users, such as interpreters, who could also be using the interface. However, as noted in Section III-B, there can be noticeable differences in the articulation of signs based on fluency. Thus, the replicability of the 110 signs listed in Fig. 2 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 [51], [52]. Next, both the discrete Fréchet distance (DFD) [53] and dynamic time warping (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 [54] 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,

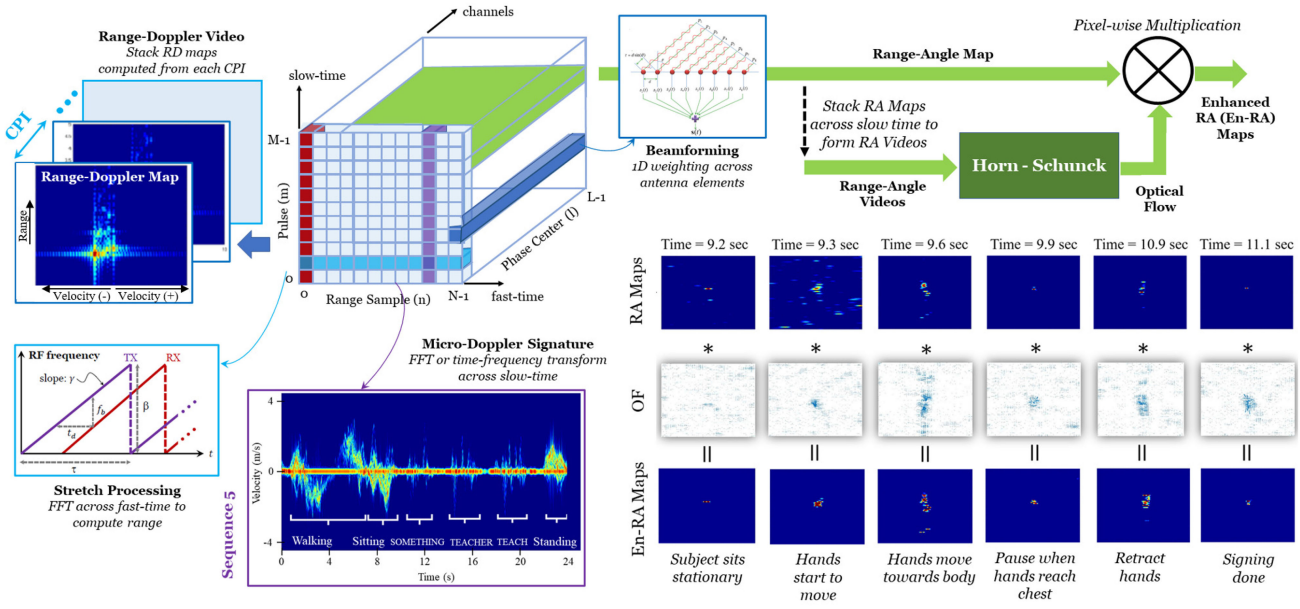
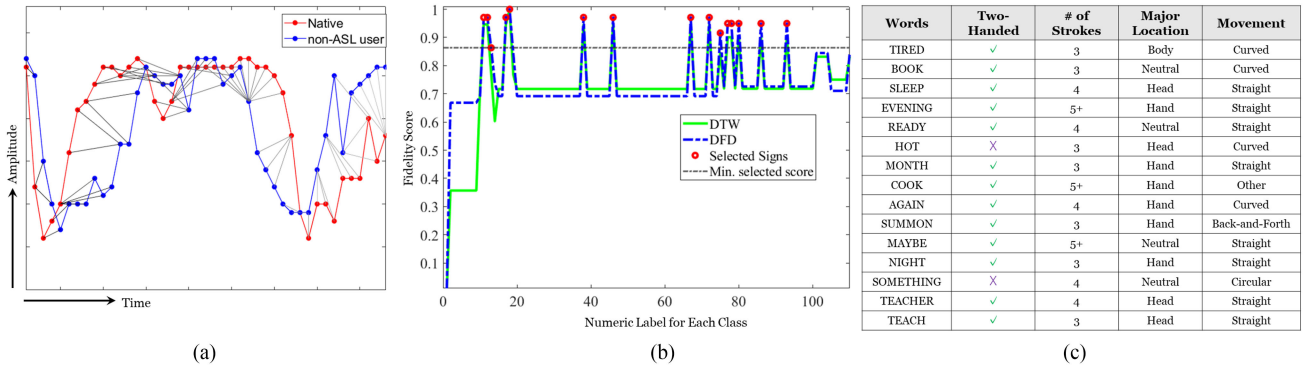


Fig. 4. Signal processing diagram for computation of various RF data representations.



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

(a) (b) (c)

Fig. 5. Selection of replicable ASL signs using DFD and DTW. (a) DFD of upper envelopes for the sign WANT. (b) Fidelity analysis between native and copy signing. (c) Selected replicable ASL signs.

consider the comparison of the upper envelopes of the μD signatures for imitation signing and native signing for the sign WANT, shown in Fig. 5(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 rescaled 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 Fig. 5(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 Fig. 5(c) along with their kinematic properties, as given by ASL-LEX.

V. 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 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 *short-time*

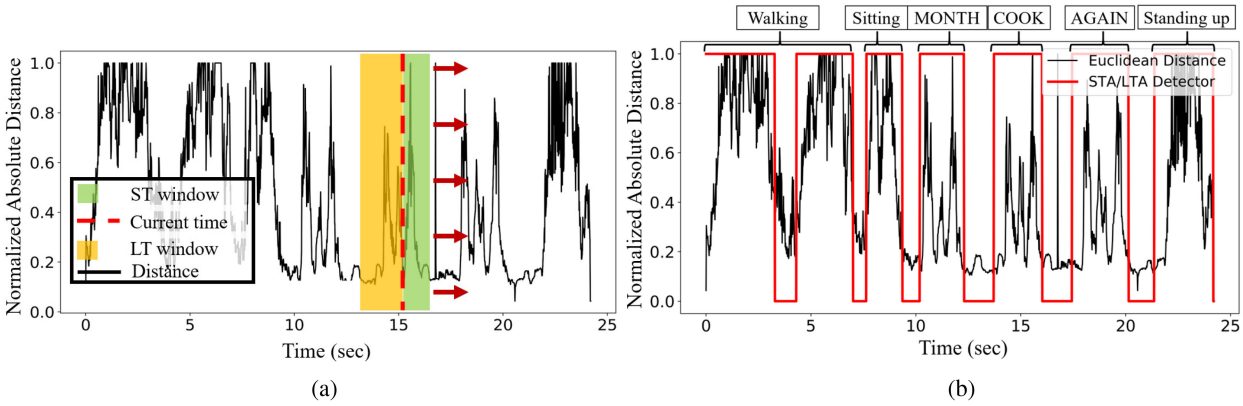


Fig. 6. Illustration of the operation of STA/LTA based motion detector on SEQUENCE 3. (a) Operation of detector on absolute distance vectors. (b) Intervals with motion detected by STA/LTA detector.

average over long-time average (STA/LTA) [55], [56], dynamic boundary detection (DBD) [57], or power burst curve [58] (PBC) 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 [56] 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 noninformative 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 [12] 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 Fig. 6(a). 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, $\mathrm{STA}(t)$ and $\mathrm{LTA}(t)$ can be defined as the leading and lagging windows at time t as

$$\mathrm{STA}(t) = \frac{1}{T_1}\sum_{k=t+1}^{t+T_1} v_i(k),\ \mathrm{LTA}(t) = \frac{1}{T_2}\sum_{k=t-T_2+1}^{t} v_i(k) \quad (3)$$

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:

$$\mathrm{STA}(t) > \sigma_1 \quad \text{and} \quad \frac{\mathrm{STA}(t)}{\mathrm{LTA}(t)} > \sigma_2 \quad (4)$$

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

$$\mathrm{STA}(t) < \sigma_3 \quad \text{and} \quad \frac{\mathrm{STA}(t)}{\mathrm{LTA}(t)} < \sigma_2 \quad (5)$$

where σ_3 is the detection threshold for the stopping point.

Note that in order to locate the starting point, according to (4), $\mathrm{STA}(t)$ 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 Fig. 6(b).

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 Fig. 7. Segmentation accuracy is computed by comparing segmentation mask with the ground truth generated by a human analyst for each time step. Note that

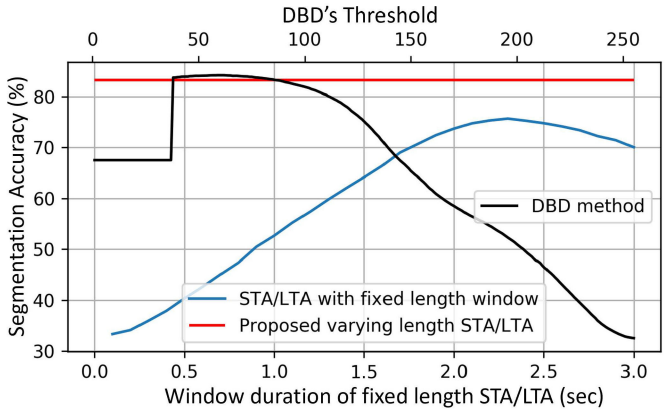


Fig. 7. Comparison of the segmentation accuracy of DBD, fixed-window STA/LTA and the proposed variable-window STA/LTA.

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 s. 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 (see Section VI-C). 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.

VI. Multiinput–Multitask Learning Framework

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

A. 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 and 2) making prediction for each time step. The former is achieved using a motion detector described in Section V, while the latter will be discussed in this section. In real-world scenarios, training a model with the entire stream of data sequences (24 s 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 [59]. 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.

TABLE II
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%

B. 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 three time-distributed (TD) 2D convolutional blocks with kernel sizes of 3, followed by max pooling layers and a bidirectional LSTM (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 [60], and speech recognition [61]. 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-s *nonoverlapping* windows to be used as time steps. In RD and RA map classification, the interval between each RD/RA map or frame is 40 ms, so to obtain a data structure corresponding to the same (0.2 s) duration, five RD/RA frames were stacked (5×40 ms $=$ 0.2 s). For both inputs, 80% of the data are 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 have 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.

1) Original Sequential Data: Table II 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 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.

TABLE III
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
Detected Intervals	61.4 sec	69.3	67.8

TABLE IV
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%**

The performance using μDS changes around 17% while that using RD maps and RA maps change around 20% from 1-s sequences to 24-s 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.

2) 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 s for μD, and 1 s for RD/RA maps, while the length of detected segments vary between 0.6 and 10 s. Moreover, using MDI rather than fixed length windows significantly reduces the computation time for prediction by masking out the intervals that do not contain any motion. Table III presents the total computation time of an NVIDIA Titan V GPU to make predictions for data durations of 1 and 2 s. The total computation time is reduced by 45% on average for different input representations when compared with 2-s 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.

C. 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 IV 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.

D. Proposed Approach: JD-MIMTL

To improve the classification accuracy obtained with just one input representation, this article proposes utilizing fusion of multiple input representations in a multiple-task learning [62] framework with CTC [63]. Although MTL has been implemented successfully in computer vision [64] and natural language processing [65], 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 Fig. 5(c). Five auxiliary tasks are defined as follows.

1) *Task 1*: One versus two handedness.
2) *Task 2*: Major location of hands.
3) *Task 3*: Movement type.
4) *Task 4*: Daily activity versus ASL sign.
5) *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

$$L_{\text{total}} = \lambda_{\text{ctc}} L_{\text{ctc}} + \sum_{i}^{I} \lambda_i L_i \tag{6}$$

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 [66] were compared, namely, the uniform combination of losses (i.e., equal weights across all tasks), the uncertainty-based weighing method [67], 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 $\lambda_{\text{ctc}} = 1$ and $\lambda_i = 0.2$ were used. The overall proposed JD-MIMTL approach is depicted in Fig. 8. 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.

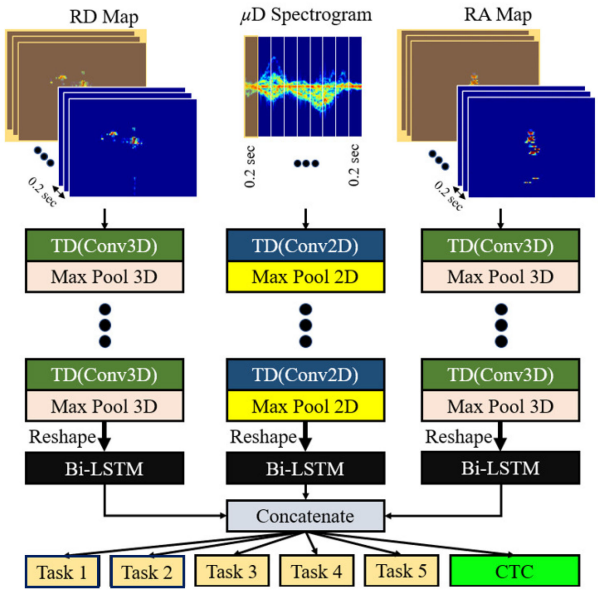


Fig. 8. Proposed multiinput–multitask learning network.

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.

VII. Results and Discussion

A. 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) [68], 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 $w/2$ s and the motion is classified as the trigger sign.

TABLE V
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%**

In trigger word detection, effect of using single versus double thresholding can been seen from Fig. 9(a), which shows the tradeoff between the 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, $D_r = 1 - \text{FRR} - \text{FAR}$, where FRR and FAR are defined as

$$\text{FRR} = \frac{n_t - n_d}{n_t}, \qquad \text{FAR} = \frac{n_f}{n_t} \tag{7}$$

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

As shown in Fig. 9(b), 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 ([69], [70]), 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 four were prone 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 [71], or, in quantitative terms, several entropy (high information-density) peaks within the continuous signal.

B. 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 V. 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

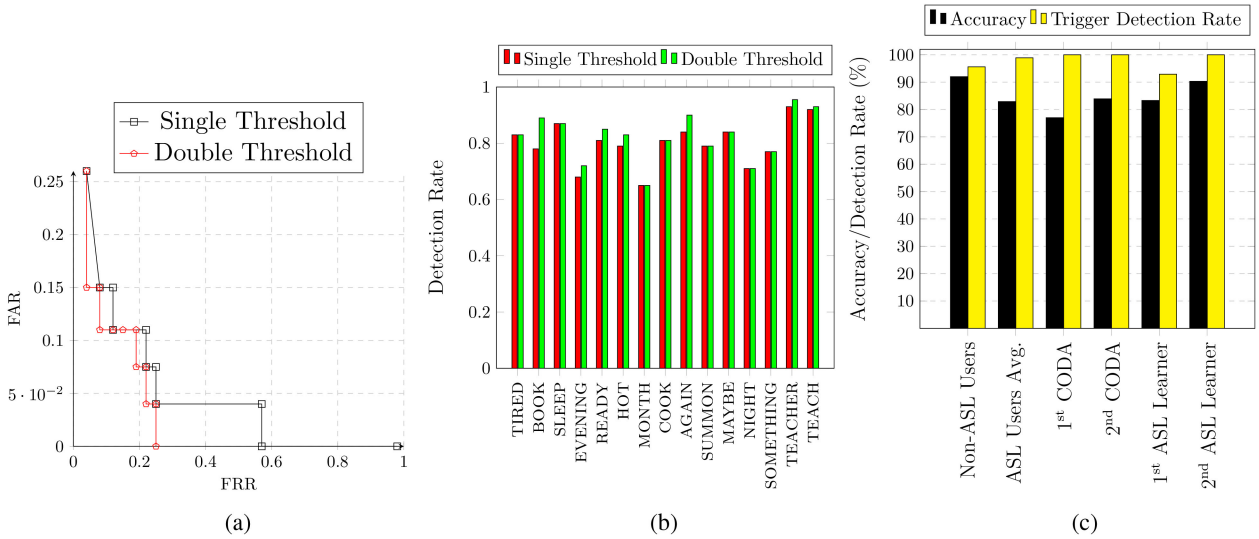


Fig. 9. Trigger word detection results. (a) FAR and FRR of the word AGAIN for single and double threshold. (b) Detection rates of the words for single and double threshold. (c) Performance of the proposed method on different ASL fluency groups.

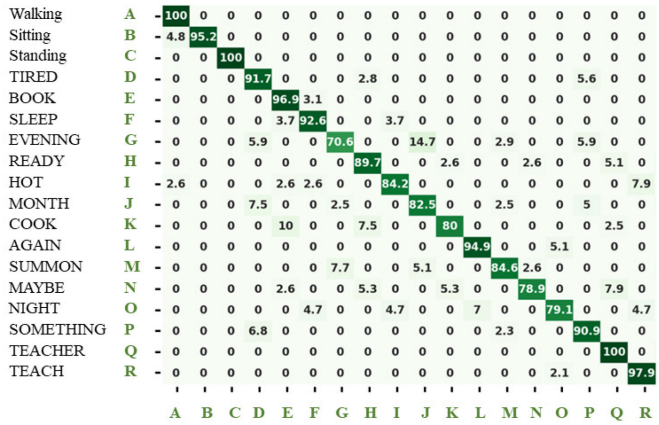


Fig. 10. Confusion matrix of the proposed JD-MIMTL.

baseline established in Section VI-B 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 multiinput 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 slightly 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 multiinput feature level fusion without using MTL.

The confusion matrix for the proposed architecture is provided in Fig. 10. 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.

C. 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 Fig. 9(c), 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 four 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.

D. 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. 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 [56], [72]–[74], 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 [75] 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.

VIII. Conclusion

The proposed techniques in this article enable 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.

Acknowledgment

The human studies research was conducted under UA Institutional Review Board (IRB) Protocol #18-06-1271.

References

[1] J. Wu, L. Sun, and R. Jafari, "A wearable system for recognizing american sign language in real-time using IMU and surface EMG sensors," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 5, pp. 1281–1290, Sep. 2016.

[2] N. Siddiqui and R. H. M. Chan, "Hand gesture recognition using multiple acoustic measurements at wrist," *IEEE Trans. Human-Mach. Syst.*, vol. 51, no. 1, pp. 56–62, Feb. 2021.

[3] M. Mohandes, M. Deriche, and J. Liu, "Image-based and sensor-based approaches to arabic sign language recognition," *IEEE Trans. Human-Mach. Syst.*, vol. 44, no. 4, pp. 551–557, Aug. 2014.

[4] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, "Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2306–2320, Sep. 2020.

[5] R. Cui, H. Liu, and C. Zhang, "A deep neural framework for continuous sign language recognition by iterative training," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1880–1891, Jul. 2019.

[6] C. Sun, T. Zhang, B. Bao, C. Xu, and T. Mei, "Discriminative exemplar coding for sign language recognition with kinect," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1418–1428, Oct. 2013.

[7] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian, and B. B. Chaudhuri, "A modified LSTM model for continuous sign language recognition using leap motion," *IEEE Sensors J.*, vol. 19, no. 16, pp. 7056–7063, Aug. 2019.

[8] D. Bragg *et al.*, "Sign language recognition, generation, and translation: An interdisciplinary perspective," in *Proc. 21st Int. ACM SIGACCESS Conf. Comput. Accessibility*, 2019, pp. 16–31.

[9] S. Gurbuz, A. Gurbuz, C. Crawford, and D. Griffin, "Radar-based methods and apparatus for communication and interpretation of sign languages," U. S. Patent Appl. US2020/0334452, Oct. 22, 2020.

[10] S. Z. Gurbuz *et al.*, "A linguistic perspective on radar micro-Doppler analysis of American sign language," in *Proc. IEEE Int. Radar Conf.*, 2020, pp. 232–237.

[11] S. Z. Gurbuz *et al.*, "ASL recognition based on kinematics derived from a multi-frequency RF sensor network," in *Proc. IEEE Sensors*, 2020, pp. 1–4.

[12] S. Z. Gurbuz *et al.*, "American sign language recognition using RF sensing," *IEEE Sensors J.*, vol. 21, no. 3, pp. 3763–3775, Feb. 2021.

[13] S. Gurbuz, S. Sun, and D. Tahmoush, "Radar systems, signals, and phenomenology" in *Proc. Deep Neural Netw. Des. Radar Appl.*, 2020, pp. 11–40.

[14] V. Chen, *The Micro-Doppler Effect in Radar*. Norwood, MA, USA: Artech House, 2011.

[15] N. Boulgouris, D. Hatzinakos, and K. Plataniotis, "Gait recognition: A challenging signal processing technology for biometric identification," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 78–90, Nov. 2005.

[16] B. Vandersmissen *et al.*, "Indoor person identification using a low-power FMCW radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 3941–3952, Jul. 2018.

[17] Y. Kim and H. Ling, "Human activity classification based on micro-Doppler signatures using a support vector machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 5, pp. 1328–1337, May 2009.

[18] S. Gurbuz *et al.*, "Micro-doppler based in-home aided and unaided walking recognition with multiple radar and sonar systems," *IET Radar, Sonar Navigation*, vol. 11, no. 1, pp. 107–115, 2016.

[19] M. G. Amin, Y. D. Zhang, F. Ahmad, and K. D. Ho, "Radar signal processing for elderly fall detection: The future for in-home monitoring," *IEEE Signal Process. Mag.*, vol. 33, no. 2, pp. 71–80, Mar. 2016.

[20] B. Y. Su, K. C. Ho, M. J. Rantz, and M. Skubic, "Doppler radar fall activity detection using the wavelet transform," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 3, pp. 865–875, Mar. 2015.

[21] A.-K. Seifert, M. G. Amin, and A. M. Zoubir, "Toward unobtrusive in-home gait analysis based on radar micro-doppler signatures," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 9, pp. 2629–2640, Sep. 2019.

[22] C. Gu, J. Wang, and J. Lien, "Motion sensing using radar: Gesture interaction and beyond," *IEEE Microw. Mag.*, vol. 20, no. 8, pp. 44–57, Aug. 2019.

[23] 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 Trans. Human-Mach. Syst.*, vol. 51, no. 1, pp. 32–43, Feb. 2021.

[24] A. Arbabian, S. Kang, S. Callender, J. C. Chien, B. Afshar, and A. Niknejad, "A 94GHz mm-wave to baseband pulsed-radar for imaging and gesture recognition," in *Proc. Symp. Very Large Scale Integration Cir.*, 2012, pp. 56–57.

[25] S. Z. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring," *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 16–28, Jul. 2019.

[26] M. Wang, G. Cui, X. Yang, and L. Kong, "Human body and limb motion recognition via stacked gated recurrent units network," *IET Radar, Sonar Navigation*, vol. 12, no. 9, pp. 1046–1051, 2018.

[27] Z. Zhang, Z. Tian, and M. Zhou, "Latern: Dynamic continuous hand gesture recognition using FMCW radar sensor," *IEEE Sensors J.*, vol. 18, no. 8, pp. 3278–3289, Apr. 2018.

[28] J. Lien *et al.*, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Trans. Graph.*, vol. 35, no. 4, Jul. 2016, Art. no. 142.

[29] S. Hazra and A. Santra, "Robust gesture recognition using millimetric-wave radar system," *IEEE Sens. Lett.*, vol. 2, no. 4, Dec. 2018, Art. no. 7001804.

[30] S. Hazra and A. Santra, "Short-range radar-based gesture recognition system using 3D CNN with triplet loss," *IEEE Access*, vol. 7, pp. 125623–125633, 2019.

[31] P. S. Santhalingam, A. A. Hosain, D. Zhang, P. Pathak, H. Rangwala, and R. Kushalnagar, "MmASL: Environment-independent ASL gesture recognition using 60 GHz millimeter-wave signals," in *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 1, Mar. 2020, Art. no. 26.

[32] Z. Gu *et al.*, "Blind separation of doppler human gesture signals based on continuous-wave radar sensors," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 7, pp. 2659–2661, Jul. 2019.
[33] H. Li, A. Mehul, J. Le Kernec, S. Z. Gurbuz, and F. Fioranelli, "Sequential human gait classification with distributed radar sensor fusion," *IEEE Sensors J.*, vol. 21, no. 6, pp. 7590–7603, Mar. 2021.
[34] C. Ding *et al.*, "Continuous human motion recognition with a dynamic range-doppler trajectory method based on FMCW radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6821–6831, Sep. 2019.
[35] H. Kulhandjian, P. Sharma, M. Kulhandjian, and C. D'Amours, "Sign language gesture recognition using Doppler radar and deep learning," in *Proc. IEEE Globecom Workshops*, 2019, pp. 1–6.
[36] S. Z. Gurbuz *et al.*, "Multi-frequency RF sensor fusion for word-level fluent ASL recognition," *IEEE Sensors J.*, to be published, doi: 10.1109/JSEN.2021.3078339.
[37] M. M. Rahman *et al.*, , "Word-level sign language recognition using linguistic adaptation of 77 GHz FMCW radar data," in *Proc. IEEE Radar Conf.*, 2021, pp. 1–6, doi: 10.1109/RadarConf2147009.2021.9455190.
[38] P. Melgarejo, X. Zhang, P. Ramanathan, and D. Chu, "Leveraging directional antenna capabilities for fine-grained gesture recognition," in *Proc. ACM UbiComp.*, 2014, pp. 541–551.
[39] H. Li, W. Yang, J. Wang, Y. Xu, and L. Huang, "Wifinger: Talk to your smart devices with finger-grained gesture," in *Proc. ACM UbiComp*, 2016, pp. 250–261.
[40] E. Malaia, J. D. Borneman, and R. B. Wilbur, "Assessment of information content in visual signal: Analysis of optical flow fractal complexity," *Vis. Cognition*, vol. 24, no. 3, pp. 246–251, 2016.
[41] J. D. Borneman, E. Malaia, and R. B. Wilbur, "Motion characterization using optical flow and fractal complexity," *J. Electron. Imag.*, vol. 27, no. 5, 2018, Art. no. 051229.
[42] E. A. Malaia and R. B. Wilbur, "Syllable as a unit of information transfer in linguistic communication: The entropy syllable parsing model," *Wiley Interdiscipl. Reviews: Cogn. Sci.*, vol. 11, no. 1, 2020, Art. no. e1518.
[43] A. Blumenthal-Dramé and E. Malaia, "Shared neural and cognitive mechanisms in action and language: The multiscale information transfer framework," *Wiley Interdiscipl. Reviews: Cogn. Sci.*, vol. 10, no. 2, 2019, Art. no. e1484.
[44] R. B. Wilbur and E. Malaia, "Contributions of sign language research to gesture understanding: What can multimodal computational systems learn from sign language research," *Int. J. Semantic Comput.*, vol. 2, no. 1, pp. 5–19, 2008.
[45] B. Fang, J. Co, and M. Zhang, "Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation," in *Proc. 15th ACM Conf. Embedded Netw. Sensor Syst.*, 2017, pp. 1–13.
[46] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "Signfi: Sign language recognition using WiFi," in *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, Mar. 2018, Art. no. 23.
[47] J. S. Beal and K. Faniel, "Hearing l2 sign language learners," *Sign Lang. Stud.*, vol. 19, no. 2, pp. 204–224, 2019.
[48] N. K. Caselli, Z. S. Sehyr, A. M. Cohen-Goldberg, and K. Emmorey, "ASL-LEX: A lexical database of American sign language," *Behav. Res. Methods*, vol. 49, no. 2, pp. 784–801, Apr. 2017.
[49] B. K. Horn and B. G. Schunck, "Determining optical flow," vol. 17, nos. 1/3, pp. 185–203, 1981.
[50] J. Hill, "Black ASL," *J. Amer. Sign Lang. Literatures*, 2012.
[51] P. V. Dorp and F. C. A. Groen, "Feature-based human motion parameter estimation with radar," *IET Radar, Sonar Navigation*, vol. 2, no. 2, pp. 135–145, 2008.
[52] C. Karabacak, S. Z. Gurbuz, A. C. Gurbuz, M. B. Guldogan, G. Hendeby, and F. Gustafsson, "Knowledge exploitation for human micro-Doppler classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 10, pp. 2125–2129, Oct. 2015.
[53] T. Eiter and H. Mannila, "Computing discrete Frechet distance," Christian Doppler Lab., Vienna Univ. of Technology, Tech. Rep. 94/64, 1994.
[54] A. Abbott and A. Tsay, "Sequence analysis and optimal matching methods in sociology: Review and prospect," *Sociol. Methods Res.*, vol. 29, no. 1, pp. 3–33, 2000.
[55] Y. Vaezi and M. Van der Baan, "Comparison of the STA/LTA and power spectral density methods for microseismic event detection," *Geophysical J. Int.*, vol. 203, no. 3, pp. 1896–1908, 2015.
[56] Y. Sun, T. Fei, X. Li, A. Warnecke, E. Warsitz, and N. Pohl, "Real-time radar-based gesture detection and recognition built in an edge-computing platform," *IEEE Sensors J.*, vol. 20, no. 18, pp. 10706–10716, Sep. 2020.
[57] Y. Sun *et al.*, "Moving target localization and activity/gesture recognition for indoor radio frequency sensing applications," *IEEE Sensors J.*, vol. 21, no. 21, pp. 24318–24326, Nov. 2021.
[58] Z. Zeng, M. G. Amin, and T. Shan, "ARM motion classification using time-series analysis of the spectrogram frequency envelopes," *Remote Sens.*, vol. 12, no. 3, 2020, Art. no. 454.
[59] F. Jafariakinabad, S. Tarnpradab, and K. Hua, "Syntactic recurrent neural network for authorship attribution," 2019, *arXiv:1902.09723*.
[60] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018, doi: 10.1109/MCI.2018.2840738.
[61] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 6645–6649.
[62] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.
[63] A. Graves, S. Fernández, and F. Gomez, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 369–376.
[64] R. Girshick, "Fast R-CNN," in *Proc. IEEE /CVF Int. Conf. Comput. Vision*, 2015, pp. 1440–1448.
[65] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning" in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 160–167.
[66] T. Gong *et al.*, "A comparison of loss weighting strategies for multi task learning in deep neural networks," *IEEE Access*, vol. 7, pp. 141627–141632, 2019.
[67] 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. Comp. Vis. Pattern Recognit.*, 2018, pp. 7482–7491.
[68] "Hey siri: An on-device DNN-powered voice trigger for Apple's personal assistant." *Apple Mach. Learn. Res.*, 2017.
[69] E. Malaia, "Current and future methodologies for quantitative analysis of information transfer in sign language and gesture data," *Behav. Brain Sci.*, vol. 40, 2017, Art. no. 063.
[70] L. K. Ford, J. D. Borneman, J. Krebs, E. A. Malaia, and B. P. Ames, "Classification of visual comprehension based on EEG data using sparse optimal scoring," *J. Neural Eng.*, vol. 18, no. 2, 2021, Art. no. 026025.
[71] T. Tsai and P. Hao, "Customized wake-up word with key word spotting using convolutional neural network," in *Proc. Int. SoC Des. Conf.*, 2019, pp. 136–137.
[72] H. Liu *et al.*, "M-gesture: Person-independent real-time in-air gesture recognition using commodity millimeter wave radar," *IEEE Internet Things J.*, to be published, doi: 10.1109/JIOT.2021.3098338.
[73] M. Chmurski, M. Zubert, K. Bierzynski, and A. Santra, "Analysis of edge-optimized deep learning classifiers for radar-based gesture recognition," *IEEE Access*, vol. 9, pp. 74406–74421, 2021.
[74] A. Ninos, J. Hasch, and T. Zwick, "Real-time macro gesture recognition using efficient empirical feature extraction with millimeter-wave technology," *IEEE Sensors J.*, vol. 21, no. 13, pp. 15161–15170, Jul. 2021.
[75] O. O. Adeoluwa, S. J. Kearney, E. Kurtoglu, C. J. Connors, and S. Z. Gurbuz, "Near real-time ASL recognition using a millimeter wave radar," in *Proc. Radar SensorTechnology XXV*, K. I. Ranney and A. M. Raynal, Eds., vol. 11742, 2021, pp. 173–184.

Emre Kurtoğlu received the B.S. degree in electrical and electronics engineering from Koc University, Istanbul, Turkey, in 2018. He is currently working toward the Ph.D. degree with the Lab for Computational Intelligence in Radar, Department of Electrical and Computer Engineering, the University of Alabama, Tuscaloosa, AL, USA.

From August to September 2017, he was an Intern for Honeywell, Istanbul, Turkey, and, from June to July 2018, for Aselsan, Ankara, Turkey. His research interests include machine learning, human activity recognition, radar signal processing, and human–computer interaction.

Kurtoglu was the recipient of a UA Graduate Council Fellowship, in September 2020.

Ali C. Gurbuz (Senior Member, IEEE) received B.S. degree in electrical engineering from Bilkent University, Ankara, Turkey, in 2003, and the M.S. and Ph.D. degrees in electrical and computer engineering from Georgia Institute of Technology (Georgia Tech), Atlanta, GA, USA, in 2005 and 2008.

From 2003 to 2009, he researched compressive sensing-based computational imaging problems with Georgia Tech. Between 2009 and 2017, he held faculty positions with TOBB University, Ankara, Turkey, and University of Alabama, Tuscaloosa, AL, USA, where he pursued an active research program on the development of sparse signal representations, compressive sensing theory and applications, radar and sensor array signal processing, and machine learning. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Mississippi State University, Mississippi State, MS, USA, where he is Co-Director with Information Processing and Sensing Lab.

Dr. Gurbuz was the recipient of The Best Paper Award for *Signal Processing* Journal in 2013, and the Turkish Academy of Sciences Best Young Scholar Award in Electrical Engineering in 2014. He was an Associate Editor for several journals such as *Digital Signal Processing, EURASIP Journal on Advances in Signal Processing, and Physical Communications*.

Evie A. Malaia received the Ph.D. degree in computational linguistics from Purdue University, West Lafayette, IN, USA, in 2004.

She was formerly a Research Scientist with Indiana University, Bloomington, IN, USA, and Purdue University, and an Assistant Professor with the University of Texas at Arlington, Arlington, TX, USA. She is currently an Associate Professor with the Department of Communicative Disorders, University of Alabama at Tuscaloosa, Tuscaloosa, AL, USA. Her research interests include neural and physical bases of sign language communication, classification of higher cognitive states, and neural bases of autism spectrum disorders.

Dr. Malaia was the recipient of the Ralph E. Powe Award from DOE/ORAU, the EurIAS Research Fellowship, the EU Marie Curie Senior Research Fellowship, and the APS Award for Teaching and Public Understanding of Psychological Science.

Darrin Griffin received the B.S. degree in communication sciences and disorders with a focus on deaf education and the M.A. degree in communication studies from The University of Texas at Austin, Austin, TX, USA, in 2004 and 2007, respectively, and the Ph.D. degree in communication with a focus on deceptive communication from The University at Buffalo, SUNY, Buffalo, NY, USA, in 2010.

From August 2010, he was a Faculty Member with the Department of Communication Studies, University of Alabama, Tuscaloosa, AL, USA, where he currently teaches and conducts research as an Associate Professor. He is fluent in American Sign Language and participates in various forms of community engagement with the deaf community. His research interests include nonverbal communication, deceptive communication, and deafness.

Dr. Griffin was the recipient of the 2020 College of Communication and Information Sciences Board of Visitors Research Excellence Award, the 2018 President's Faculty Research Award at The University of Alabama, and the 2018 Premiere Award from The University of Alabama Council on Community-Based Partnerships for research that raised weather awareness and preparedness for the deaf & hard of hearing community.

Chris Crawford received the Ph.D. degree in human-centered computing from the University of Florida, Gainesville, FL, USA.

He is currently an Assistant Professor with the Department of Computer Science, University of Alabama, Tuscaloosa, AL, USA. He directs the Human-Technology Interaction Lab. He has investigated multiple systems that provide computer applications and robots with information about a user's cognitive state. In 2016, he led the development of a BCI application that was featured in the world's first multiparty brain–drone racing event. His research interests include computer science education, human–robot interaction, and brain–computer interfaces.

Sevgi Z. Gurbuz (Senior Member, IEEE) received the B.S. degree in electrical engineering with minor in mechanical engineering and the M.Eng. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1998 and 2000, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2009.

From February 2000 to January 2004, she was a Radar Signal Processing Research Engineer with the U.S. Air Force Research Laboratory, Sensors Directorate, Rome, NY, USA. She was an Assistant Professor with the Department of Electrical-Electronics Engineering, TOBB University, Ankara, Turkey, and a Senior Research Scientist with the TUBITAK Space Technologies Research Institute, Ankara, Turkey. She is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Alabama at Tuscaloosa, Tuscaloosa, AL, USA. Her research interests include physics-aware machine learning, RF sensor-enabled cyber-physical systems, radar signal processing, sensor networks, human motion recognition for biomedical, automotive autonomy, and human–computer interaction applications.

Dr. Gurbuzwas the recipient of the IEEE Harry Rowe Mimno Award for 2019, the 2020 SPIE Rising Researcher Award, the EU Marie Curie Research Fellowship, and the 2010 IEEE Radar Conference Best Student Paper Award.