Sequential Classification of ASL Signs in the Context of Daily Living Using RF Sensing

Emre Kurtoglu[1], Ali C. Gurbuz[2], Evie Malaia[3], Darrin Griffin[4], Chris Crawford[5], Sevgi Z. Gurbuz[1]
[1]*Dept. of Electrical and Computer Engineering, The University of Alabama, Tuscaloosa, USA*
[2]*Dept. of Electrical and Computer Engineering, Mississippi State University, Starkville, USA*
[3]*Dept. of Communication Disorders, The University of Alabama, Tuscaloosa, USA*
[4]*Dept. of Communication Studies, The University of Alabama, Tuscaloosa, USA*
[5]*Dept. of Computer Science, The University of Alabama, Tuscaloosa, USA*
ekurtoglu@crimson.ua,edu, szgurbuz@ua.edu

***Abstract*—RF sensing based human activity and hand gesture recognition (HGR) methods have gained enormous popularity with the development of small package, high frequency radar systems and powerful machine learning tools. However, most HGR experiments in the literature have been conducted on individual gestures and in isolation from preceding and subsequent motions. This paper considers the problem of American sign language (ASL) recognition in the context of daily living, which involves sequential classification of a continuous stream of signing mixed with daily activities. In particular, this paper investigates the efficacy of different RF input representations and fusion techniques for ASL and trigger gesture recognition tasks in a daily living scenario, which can be potentially used for sign language sensitive human-computer interfaces (HCI). The proposed approach involves first detecting and segmenting periods of motion, followed by feature level fusion of the range-Doppler map, micro-Doppler spectrogram, and envelope for classification with a bi-directional long short-term memory (BiLSTM) recurrent neural network. Results show 93.3% accuracy in identification of 6 activities and 4 ASL signs, as well as a trigger sign detection rate of 0.93.**

***Index Terms*—Sequential classification, trigger detection, RF sensing, ASL recognition, gesture recognition**

I. Introduction

Using RF sensing for human activity recognition has become an emerging research area with multiple applications like gait abnormality recognition [1]–[3], non-contact [4] measurement of heart rate [5] and respiration [6], [7], fall detection [8], [9] concussion detection [10], and hand gesture recognition (HGR) [11]–[13], among others. Sign language recognition (SLR) with wi-fi [14] or millimeter wave radar [15] has been recently proposed, and is related to HGR, but possesses a much greater degree of complexity and nuance, due to the communicative/linguistic nature of the signal, which brings additional challenges relating to temporal resolution and complexity of the physical signal, as well as linguistic parameters of the message, such as phonotactic constraints (i.e. linguistically permissible handshapes and their combinations), prosody (pauses and suprasegmental components of articulation, similar to intonation in spoken languages), pragmatic context, and grammatical structure. In prior work [16], [17], we have shown that radar-measurements of ASL can capture linguistic features, such as signal changes due to coarticulation, and signer proficiency in comparison of native vs. imitation signing - when a hearing individual attempts to replicate signing motions based on video stimuli of signs.

Optical cameras [18], sensor-augmented gloves [19] and motion capture (MOCAP) cameras/gloves [20], [21] have also been used for SLR applications. Gloves may have greater accuracy than video, but are not practical in daily life due to the severe restrictions to natural motion they impose on the user. Unlike video, radar does not provide any optical imagery of the scene and therefore does not violate user privacy (even if hacked), and are also effective when cameras are not, such as in the dark and through the wall [22]. These advantages make radar a promising sensor for the design of ASL-sensitive home assistants or Deaf-centric smart environments.

The current popular approaches for personal assistance services such as Amazon's Alexa, Apple's Siri and Google Assistant mainly use speech as the medium for device control; but, deaf or hard-of-hearing individuals cannot take the advantage of any of these products. Radar-based SLR, however, has the potential to enable access of the Deaf community to many new advancements by facilitating the design of ASL-sensitive human-computer interaction (HCI).

Most speech-driven devices function by first detecting a trigger/wake-up word, followed by speech recognition to interpret verbal commands. Similarly, an important practical consideration in developing SLR for HCI with radar is being able to detect a trigger sign in the context of daily living, which involves many other unrelated motions. Although there has been much research on trigger word detection [23], [24] and sequential classification [25] in speech recognition literature, there are very limited number of radar-based sequential classification works [26], [27].

This paper addresses the challenge of sequential classification in daily scenarios for the purpose of trigger sign detection and SLR recognition. Three different input representations are computed from the raw I/Q radar data: range-Doppler (RD) maps, micro-Doppler signatures, and their envelopes. In Section II, the sequential mixed-motion dataset and its preprocessing are described. In Section III, the efficacy of two

TABLE I
DESCRIPTION OF THE SEQUENCES

Sequence No.	Activities and ASL Gestures
1	Walking (10) - Sitting (2) - NM (3) - YHCP (12) - Standing Up (3)
2	Sitting (2) - CPHY (12) - NM (3) - Folding Laundry (10) - Standing Up (3)
3	Sitting (2) - HCPY (12) - NM (3) - Ironing (10) - Standing Up (3)

different sign detection techniques are compared to ascertain whether there is any motion, and extract motion-related data segments. Next, in Section IV, these segments are classified using a fusion of the three input representations and a bi-directional long short-term memory (Bi-LSTM) recurrent neural network. Details of the proposed trigger word detection method are presented and compared with conventional trigger detection approaches. Finally, the paper concludes in Section VI with a discussion of future work.

II. DATA COLLECTION AND PRE-PROCESSING

For this study, data was acquired such that several daily activities and signs are continuously recorded as 30 second samples. A total of 4 participants enacted 6 different daily activities (No/low movement (NM), walking, sitting, standing up, folding laundry, ironing) and 4 ASL signs (YOU (Y), HELLO (H), CAR (C), PUSH (P)). Although the participants were not native signers, they practiced copysigning prior to the recording from videos by native signers. The selected ASL signs utilized familiar/iconic handshapes, and gross hand motion, helping minimize error potential. Each participant performed one of the three sequences listed in Table II. A total 196 samples were acquired, with 80% of each participant's data used for training and 20% for testing.

The RF data was acquired using a Texas Instrument's AWR1642BOOST frequency modulated continuous wave (FMCW) radar with a transmit frequency of 77 GHz and a bandwidth of 4 GHz, which resulst in a range resolution of 3.75 cm. When signing, the participant was located about 1.5 m directly in front of the radar, while daily activities were conducted at varying distances within a 4 meters radius from the radar.

The raw data acquired from RF sensor is a time-stream of complex I/Q data. The data is then reshaped to form a matrix in which the rows are populated with the fast-time analog-to-digital converter (ADC) samples of the return from a single pulse, and the columns are formed by stacking the return from each pulse, thus representing slow-time samples. From the slow-time / fast-time data matrix, we generated three input representations for DNN models, namely range-Doppler (RD) maps, micro-Doppler (μD) spectrograms and spectrogram envelopes.

RD maps are computed by taking the 2D Discrete Fourier Transform (DFT) of the data matrix corresponding to a coherent processing interval (CPI). A CPI of 128 pulses were selected with the frame interval of 40 ms, so that each 30 second sample resulted in 750 RD maps being generated. This results in a 3D time-series of RD maps, which shows how the target's radial velocity and range varies with time.

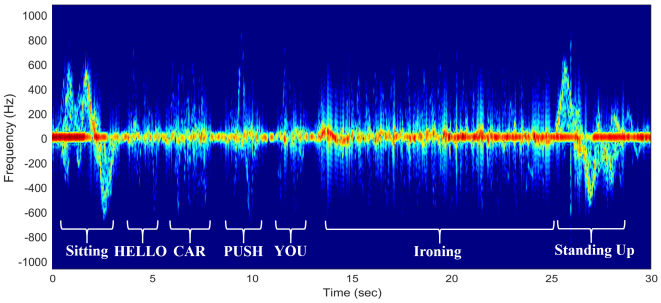

Fig. 1. μD spectrogram of a SEQUENCE 3 sample.

μD spectrograms are a time-frequency representation of RF data. The μD spectrogram can be calculated from the data matrix by taking the square modulus of the Short-Time Fourier Transform (STFT) computed across slow-time. A sample spectrogram is shown in Figure 1, where frequency is proportional to target velocity and the intensity is proportional to the power of the received radar signal.

The upper and lower envelopes of μD spectrograms provide significant information, as they represent the maximum velocities attained during the signing. Envelope features of μD spectrograms have shown to be [28] effective to provide information about different human activities. Upper, central and lower envelopes are extracted using the technique proposed by Van Dorp and Groen [29].

III. MOTION DETECTION

Unlike conventional one stage classification methods, we propose dividing the processing of continuous RF data into two stages: motion detection/segmentation, and segment recognition. Thus, classification will be done only on those periods of time during which motion is detected by the signing detector. This approach allows us to eliminate a great deal of the unnecessary computations, which stems from the classification of large segments of data that do not contain any motion in daily living scenarios.

Motion detection is done using Range-Weighted Energy (RWE) plots obtained from RD maps with two different detection methods: Cell Averaging Constant False Alarm Rate (CA-CFAR) thresholding and the Short Time Average over Long Time Average (STA/LTA) algorithm. RWE plots can be computed by: 1) normalizing the intensity value of each pixel by its range, and then 2) summing all range-weighted pixel values to obtain the total energy of the RD map at each CPI.

A. CA-CFAR Based Trigger Detector

CA-CFAR is one of the most popular CFAR detectors. In most cases, it is used as a baseline comparison method for other CFAR techniques. In this work, we apply CA-CFAR to RWE plots to detect peaks in the data which are caused by movement along the radar line-of-sight. In the CA-CFAR

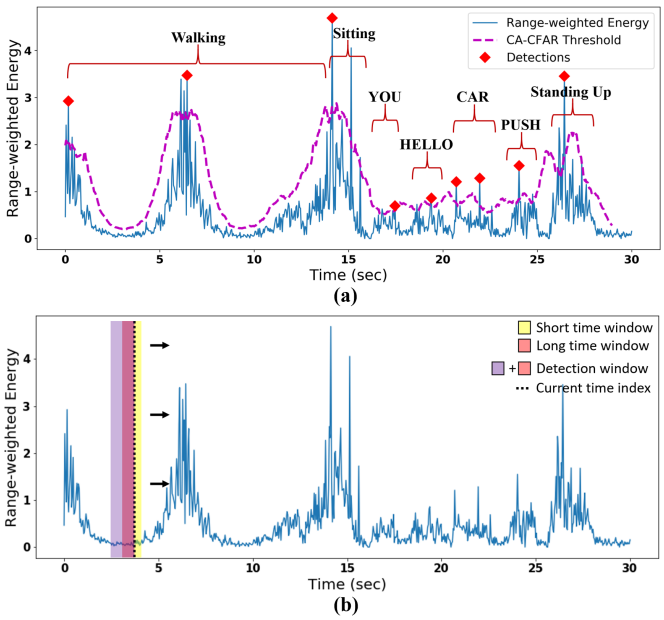

Fig. 2. RWE plot of a SEQUENCE 1 sample with (a) *CA-CFAR* and (b) *STA/LTA*.

detector, the decision for a given cell (also called a *cell under test* (CUT)) is given based on comparison with a threshold, T, given by

$$T = aP_n \tag{1}$$

where a is a scaling factor (also called the threshold factor) and P_n is the noise power estimate. The noise power is estimated from leading and lagging neighboring cells.

From Eq. (1), it may be observed that the adaptive threshold varies according to the data. Moreover, the desired probability of false alarm (P_{fa}) rate can be kept at a constant level using a proper threshold factor, a. The noise power estimate can be computed as

$$P_n = \frac{1}{M}\sum_{m=1}^{M} x_m \tag{2}$$

where M is the number of training cells and x_m is the sample in each training cell. The threshold, a, then can be written as

$$a = M(P_{fa}^{-1/M} - 1) \tag{3}$$

In total, we used 20 training and 10 guard cells for detection and the adaptive threshold was able to detect movement-related peaks successfully, as shown in Figure 2a. After each detection, a fixed 2 second detection window was applied to select motion-related samples for classification.

B. STA/LTA Based Trigger Detector

The STA/LTA algorithm continuously keeps track of the ratio between the average amplitude value in the leading (short) and the lagging (long) window. A third detection window is defined to choose how many preceding samples to be send to the next stage upon detection. When the STA/LTA ratio goes below the pre-defined threshold value the system will go into the trigger mode. In our case, the trigger stage corresponds to sending samples in detection window to the classifier. The algorithm has four parameters: length of short, long and detection windows, and the threshold value, which are illustrated in Figure 2b. As with the CA-CFAR based detector, samples extracted using a 2 second detection window are then sent to the classifier.

TABLE II
CLASSIFICATION RESULTS

Detectors	RD Map	μD Spectrogram	Envelope	Decision level fusion	Feature level fusion
CA-CFAR	91.6%	91.4%	85.9%	92.8%	92.9%
STA/LTA	91.6%	90.8%	88.8%	93.1%	**93.3%**

IV. SEQUENTIAL CLASSIFICATION

Classification of time series data is done using three different input representations, namely RD maps, μD spectrograms and envelopes of μD spectrograms.

In RD map classification we divided videos into 0.2 second windows in order to have a comparable input shape with other input representations for fusion which will be discussed in Section IV-B. Hence, the input data has the shape of `(batch size, time samples, frames per window, width, height, channels)` where each window is a time sample, and similarly the label data has the shape of `(batch size, time samples, number of classes)`. A time-distributed 3D convolutional neural network (CNN) followed by a bidirectional LSTM (Bi-LSTM) layer has been used for classification. While time-distributed layer applies the same nested layer to every time sample, Bi-LSTM layer can capture the long term dependencies amongst the time samples. In order to obtain a prediction for each temporal slice, we returned the sequences from the output of the LSTM layer and employed a time-distributed $softmax$ layer at the output. After movement detection using two detectors, RD maps within the detection window goes into the CNN classifier. Overall, 91.6% testing accuracy is obtained using both the CA-CFAR and the STA/LTA based detectors.

In μD spectrogram and envelope classification, we again divided the time series data into 0.2 second *non-overlapping* windows and those windows are used as time samples in the input data. We employed time-distributed 2D & 1D CNNs with Bi-LSTM layers, again followed by a time-distributed $softmax$ layer for spectrogram and envelope classification. Similar to RD map classification, the input data are classified after detection. While we obtain slightly better testing accuracy for the μD spectrograms using the CA-CFAR based detector, the STA/LTA based detector performs a lot better for envelope classification with $\sim$3% improvement when compared to the CA-CFAR method as can bee seen from Table II.

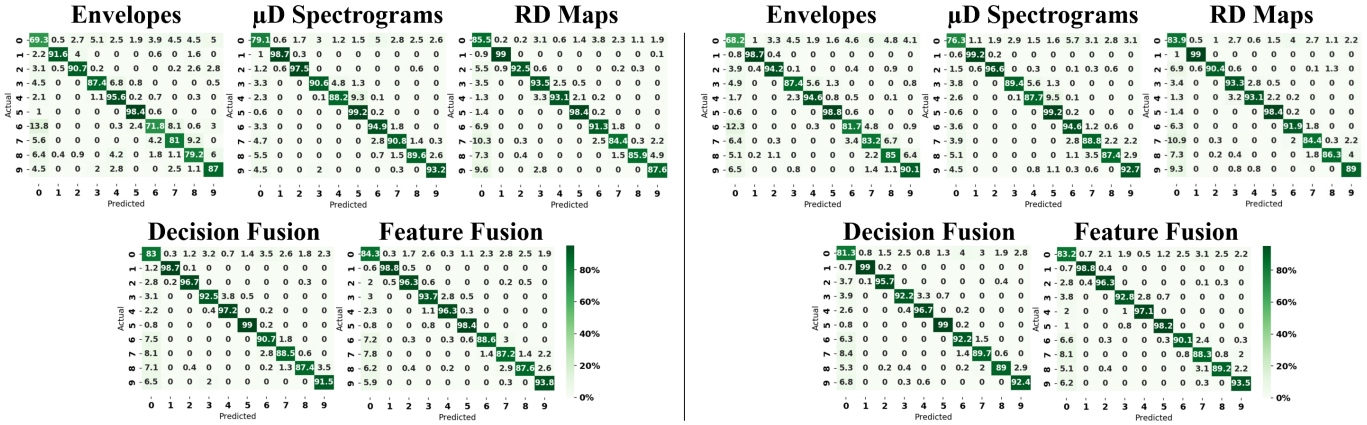

Fig. 3. Confusion matrices of detected samples (windows), where the confusion matrices in **(a)** belongs to the *CA-CFAR* detector and those in **(b)** belongs to the *STA/LTA* detector, where 0: NM, 1: walking, 2: sitting, 3: standing up, 4: folding laundry, 5: ironing, 6: YOU, 7: HELLO, 8: CAR, 9: PUSH.

A. Decision Level Fusion

In order to examine the efficacy of decision level fusion, hard majority voting method is applied. Mode of the predictions of the three aforementioned networks is used as the final prediction for a given time sample. From Table II, it may be observed that decision level fusion of multiple input representations helps to mitigate the misclassification rate with both detectors.

B. Feature Level Fusion

Feature level fusion refers to combining the feature spaces of three input representations using a multi-input network. This is done by taking the output of the LSTM layers of individual networks and feed them into a *concatenation* layer. Concatenation layer is then followed by a time-distributed softmax layer for the final prediction. Weights of all the layers are frozen before the concatenation since they have already been trained and to prevent overfitting that might stem from over-training. Details of the multi-input fusion DNN architecture can be found in Figure 4.

It should also be noted that although envelopes do not seem to perform as good as μD spectrograms and RD maps, when they are fused with other the input representations, they provide additional information about the limits on velocity. In class #4 *(folding laundry)*, envelopes give more accurate predictions than other domains, as shown in Figure 3a-b. This demonstrates the fact that different input representations can contribute to the final decision (prediction) and improve the performance as long as they contain useful information even if they do not perform well by themselves.

V. Trigger Word Detection

In speech recognition literature, trigger words with 3 to 4 syllables have been considered to be the most effective [30]. While trigger words less than 3 syllables will increase false alarm rate (FAR), those more than 5-6 syllables will increase the false rejection rate (FRR). ASL is a primarily monosyllabic language [31]; however, syllable complexity and duration can vary depending on the path complexity (single vs. reduplicated motion), and whether the path dynamics is combined with a change in handshape aperture, or handshape orientation [32]. Given the range of sign types, we chose an average-complexity sign HELLO as the trigger. As a monosyllabic sign with motion component, it consists of three stages: raising one hand to forehead, moving hand towards the conversation partner (assume conversation partner is radar), and finally retracting hand back to the original position.

The spotting of trigger sign with radar is quite similar to wake-up word detection task in speech recognition. The time series signal/input is continuously analysed in order to locate the trigger/wake-up word. One method to do this is using a *cumulative score aggregating (CSA)* approach, which has also been used in Apple's Siri [33]. The output of the network model provides a distribution of scores/probabilities

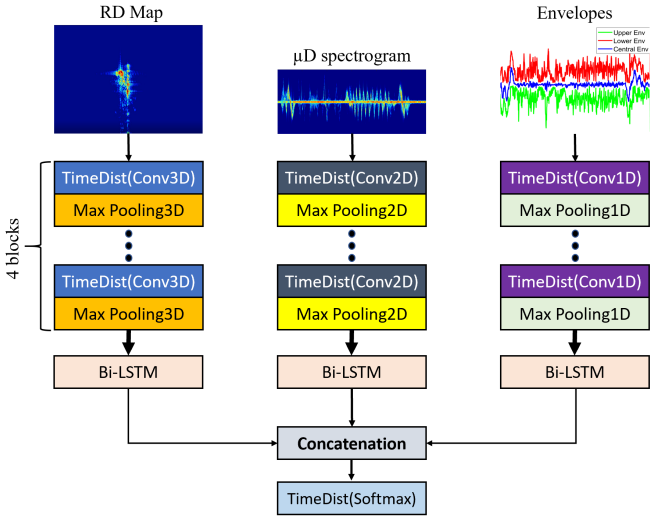

Fig. 4. Feature level fusion network architecture.

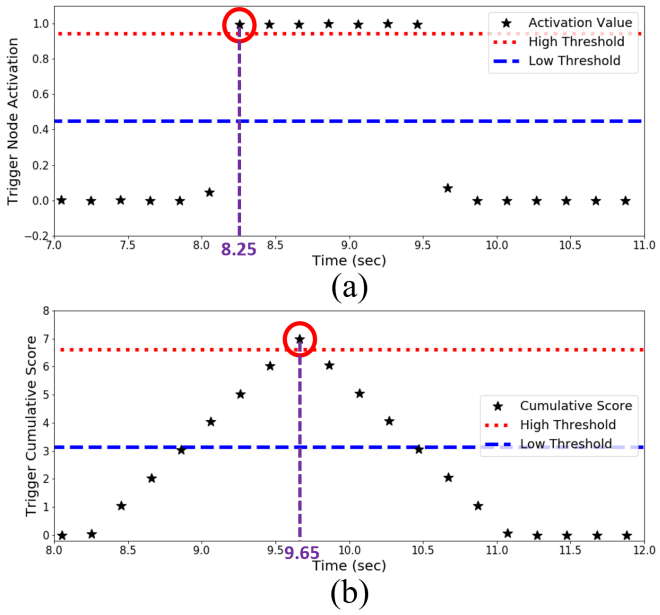

Fig. 5. Trigger sign detection of word HELLO: (a) early trigger with the activation value of the target node, (b) on-time trigger with the accumulated score of the target node.

over classes for each time sample (window). We would like to detect the trigger sign if the prediction confidence of the model is high for the target class, and we would like to go into trigger mode only when the trigger sign is fully complete. For instance, if we consider the sign HELLO, the signing process takes approximately between 1.5-2 seconds depending on the participant's signing rate [34]. If we directly use the activation values of the target softmax node as scores, the system gets triggered at the very beginning of the sign, which can potentially cause high FAR when the dataset contains signs with similar beginning patterns. For example, in the Figure 5a, the system is triggered (exceeds the high threshold) at t=8.25, while the sign finishes at t=9.65. To alleviate this problem, we accumulate the scores of last 1.4 seconds of windows (1.4/0.2 = 7 windows). This way, we ensure that the cumulative trigger score is maximized, hence the system is triggered only when the trigger sign is fully completed (see the Figure 5b).

The threshold value is also crucial to accurate detection and to preventing undesired false rejections and false alarms. While higher threshold values eliminate the false alarms, at the same time, they increase the number of false rejections. In this work, we applied a second threshold (lower) instead of a single threshold value to reduce the FRR. If the cumulative score of the windows exceeds the lower threshold, but not the higher one, the system goes into a sensitive stage for the next 1.4 seconds, and if the cumulative score stays above the lower threshold in that duration, the system gets triggered without any need for a second trial. In this work, we define FRR and FAR as

$$FRR = \frac{T - D}{T}, \qquad FAR = \frac{F}{T} \tag{4}$$

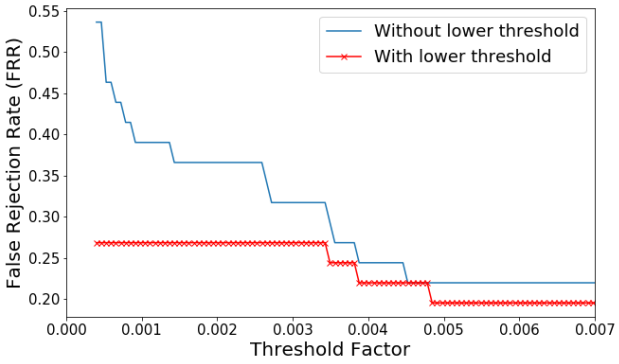

Fig. 6. False rejection rate of the word HELLO with and without the lower threshold.

where T is the total number of triggers in the test dataset, D is the detected number of triggers and F is the number of misdetections. We also define two threshold values, τ_l and τ_h, as

$$\tau_{l,h} = W(1 - (\ln C)^2 \rho_{l,h}) \tag{5}$$

where W is the number of windows whose scores are accumulated which also defines the maximum cumulative score can be achieved since the individual window scores can be maximum of 1, C is the total number of classes and $\rho_{l,h}$ are the lower/higher threshold factors. From Figure 6, it can be seen that applying the second threshold drastically reduces the false rejections, especially with the lower threshold factors. When we set the detection threshold to 90% of the maximum achievable score, while using the activation values as scores suffer from high FRR and FAR (0.18 and 0.03 respectively), double-threshold CSA method's FRR stays as low as 0.07 and FAR becomes 0. Hence, the trigger recognition rate for the word HELLO can be calculated as $1 - 0.07 - 0 = 0.93$.

VI. Conclusion

This paper presents initial work on sequential classification of continuous data streams of daily activities mixed with ASL signs. This can be potentially used for the design of sign language controlled, intelligent personal assistant systems. We demonstrate that while different RF data inputs may yield comparable performances, feature level fusion of them gives the best classification performance with 93.2% testing accuracy for 10 classes using a multiple input fusion DNN.

The issue of trigger sign recognition in a daily living scenario is addressed and a CSA-based approach proposed to mitigate the false alarms, which may stem from early triggering of the system when the sign motion is not completely finished or when the target class has a similar initial pattern as that of another class. In order to reduce the FRR and thereby improve the triggering capability of the system, we defined a second threshold that allows the system to be triggered even though the cumulative score does not exceed the higher

threshold but stays above the lower threshold for a certain duration. One drawback of using a double threshold is that it may increase the FAR if the network has high confidence for the wrong classes. A complete trigger sign detection pipeline is demonstrated for RF sensing and a detection rate of 0.93 is achieved for the word HELLO. These results show the potential of RF sensing to be used for sign language sensitive HCI applications and personal assistance services.

ACKNOWLEDGMENT

This work was funded in part by the National Science Foundation (NSF) Cyber-Physical Systems (CPS) Program Awards #1932547 and #1931861, NSF Integrative Strategies for Understanding Neural and Cognitive Systems (NCS) Program Award #1734938, as well as the University of Alabama ECE Department. Human studies research was conducted under UA Institutional Review Board (IRB) Protocol #18-06-1271.

REFERENCES

[1] A. Seifert, A. M. Zoubir, and M. G. Amin, "Radar classification of human gait abnormality based on sum-of-harmonics analysis," in *2018 IEEE Radar Conference (RadarConf18)*, 2018, pp. 0940–0945.

[2] S. Z. Gurbuz, C. Clemente, A. Balleri, and J. J. Soraghan, "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, 2017.

[3] M. S. Seyfioğlu, A. M. Özbayoğlu, and S. Z. Gürbüz, "Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 4, pp. 1709–1723, 2018.

[4] C. Li, V. M. Lubecke, O. Boric-Lubecke, and J. Lin, "A review on recent advances in doppler radar sensors for noncontact healthcare monitoring," *IEEE Transactions on Microwave Theory and Techniques*, vol. 61, no. 5, pp. 2046–2060, 2013.

[5] W. Massagram, V. M. Lubecke, A. HØst-Madsen, and O. Boric-Lubecke, "Assessment of heart rate variability and respiratory sinus arrhythmia via doppler radar," *IEEE Transactions on Microwave Theory and Techniques*, vol. 57, no. 10, pp. 2542–2549, 2009.

[6] A. Rahman, V. M. Lubecke, O. Boric–Lubecke, J. H. Prins, and T. Sakamoto, "Doppler radar techniques for accurate respiration characterization and subject identification," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 2, pp. 350–359, 2018.

[7] A. Dell'Aversano, A. Natale, A. Buonanno, and R. Solimene, "Through the wall breathing detection by means of a doppler radar and music algorithm," *IEEE Sensors Letters*, vol. 1, no. 3, pp. 1–4, 2017.

[8] M. Amin, "Radar for indoor monitoring: Detection, classification, and assessment," 2017.

[9] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "Dnn transfer learning from diversified micro-doppler for motion classification," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 5, pp. 2164–2180, 2019.

[10] J. W. Palmer, K. F. Bing, A. C. Sharma, and E. F. Greneker, "Detecting concussion impairment with radar using gait analysis techniques," in *2011 IEEE RadarCon (RADAR)*, 2011, pp. 222–225.

[11] J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Trans. Graph.*, vol. 35, no. 4, Jul. 2016. [Online]. Available: https://doi.org/10.1145/2897824.2925953

[12] 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 Journal*, vol. 20, no. 18, pp. 10 706–10 716, 2020.

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

[14] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "Signfi: Sign language recognition using wifi," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, Mar. 2018.

[15] S. Gurbuz, A. Gurbuz, C. Crawford, and D. Griffin, "Radar-based methods and apparatus for communication and interpretation of sign languages," in *U.S. Patent Application No. US2020/0334452 (Invention Disclosure filed Feb. 2018; Provisional Patent App. filed Apr. 2019.)*, October 2020.

[16] S. Z. Gurbuz, A. C. Gurbuz, E. A. Malaia, D. J. Griffin, C. Crawford, M. M. Rahman, R. Aksu, E. Kurtoglu, R. Mdrafi, A. Anbuselvam, T. Macks, and E. Ozcelik, "A linguistic perspective on radar micro-doppler analysis of american sign language," in *2020 IEEE International Radar Conference (RADAR)*, 2020, pp. 232–237.

[17] S. Z. Gurbuz, A. C. Gurbuz, E. A. Malaia, D. J. Griffin, C. Crawford, M. M. Rahman, E. Kurtoglu, R. Aksu, T. Macks, and R. Mdrafi, "American sign language recognition using rf sensing," *IEEE Sensors Journal*, pp. 1–1, 2020.

[18] Guan-Feng He, Sun-Kyung Kang, Won-Chang Song, and Sung-Tae Jung, "Real-time gesture recognition using 3d depth camera," in *2011 IEEE 2nd International Conference on Software Engineering and Service Science*, 2011, pp. 187–190.

[19] N. Tubaiz, T. Shanableh, and K. Assaleh, "Glove-based continuous arabic sign language recognition in user-dependent mode," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 4, pp. 526–533, 2015.

[20] Q. Yang, W. Ding, X. Zhou, D. Zhao, and S. Yan, "Leap motion hand gesture recognition based on deep neural network," in *2020 Chinese Control And Decision Conference (CCDC)*, 2020, pp. 2089–2093.

[21] J. McDonald, R. Wolfe, R. B. Wilbur, R. Moncrief, E. Malaia, S. Fujimoto, S. Baowidan, and J. Stec, "A new tool to facilitate prosodic analysis of motion capture data and a data-driven technique for the improvement of avatar motion," in *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2016, pp. 153–159.

[22] E. Lagunas, M. Amin, and F. Ahmad, "Through-the-wall radar imaging for heterogeneous walls using compressive sensing," 06 2015.

[23] V. Këpuska and G. Bohouta, "Improving wake-up-word and general speech recognition systems," in *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, 2017, pp. 318–321.

[24] S. Zhang, Wen Liu, and Y. Qin, "Wake-up-word spotting using end-to-end deep neural network system," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 2878–2883.

[25] W. Jeon, L. Liu, and H. Mason, "Voice trigger detection from lvcsr hypothesis lattices using bidirectional lattice recurrent neural networks," 2020. [Online]. Available: https://arxiv.org/pdf/2003.00304

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

[27] H. Li, A. Shrestha, H. Heidari, J. Le Kernec, and F. Fioranelli, "Bi-lstm network for multimodal continuous human activity recognition and fall detection," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1191–1201, 2020.

[28] C. Karabacak, S. Z. Gurbuz, A. C. Gurbuz, M. B. Guldogan, G. Hendeby, and F. Gustafsson, "Knowledge exploitation for human micro-doppler classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 10, pp. 2125–2129, 2015.

[29] 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.

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

[31] R. B. Wilbur, "Productive reduplication in a fundamentally monosyllabic language," *Language Sciences*, vol. 31, no. 2-3, pp. 325–342, 2009.

[32] D. Brentari, *A prosodic model of sign language phonology*. Mit Press, 1998.

[33] "Hey siri: An on-device dnn-powered voice trigger for apple's personal assistant." Apple Machine Learning Research, October 2017. [Online]. Available: https://machinelearning.apple.com/research/hey-siri.

[34] 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*, vol. 11, no. 1, p. e1518, 2020.