# Word-Level ASL Recognition and Trigger Sign Detection with RF Sensors

M. Mahbubur Rahman[1], Emre Kurtoglu[1], Robiulhossain Mdrafi[2], Ali C. Gurbuz[2],
Evie Malaia[3], Chris Crawford[4], Darrin Griffin[5], Sevgi Z. Gurbuz[1]
[1]*Dept. of Electrical and Computer Engineering, The University of Alabama, Tuscaloosa, AL, USA*
[2]*Dept. of Electrical and Computer Engineering, Mississippi State University, Starkville, MS, USA*
[3]*Dept. of Communication Disorders, The University of Alabama, Tuscaloosa, AL, USA*
[4]*Dept of Computer Science, The University of Alabama, Tuscaloosa, AL, USA*
[5]*Dept. of Communication Studies, The University of Alabama, AL, USA*
mrahman17@crimson.ua.edu, szgurbuz@ua.edu

***Abstract*—Current research in the recognition of American Sign Language (ASL) has focused on perception using video or wearable gloves. However, deaf ASL users have expressed concern about the invasion of privacy with video, as well as the interference with daily activity and restrictions on movement presented by wearable gloves. In contrast, RF sensors can mitigate these issues as it is a non-contact ambient sensor that is effective in the dark and can penetrate clothes, while only recording speed and distance. Thus, this paper investigates RF sensing as an alternative sensing modality for ASL recognition to facilitate interactive devices and smart environments for the deaf and hard-of-hearing. In particular, the recognition of up to 20 ASL signs, sequential classification of signing mixed with daily activity, and detection of a trigger sign to initiate human-computer interaction (HCI) via RF sensors is presented. Results yield %91.3 ASL word-level classification accuracy, %92.3 sequential recognition accuracy, 0.93 trigger recognition rate.**

***Index Terms*—ASL, sign language, RF sensing, micro-Doppler, deep learning**

## I. Introduction

Many technologies for human-computer interaction (HCI) have been designed for hearing individuals and depend upon vocalized speech, precluding users of American Sign Language (ASL) in the Deaf community from benefiting from these advancements. Although there has been much research related to technologies for the deaf or hard of hearing (HoH) over the past three decades, much of this work has focused on the translation of sign language into voice or text using camera-based or wearable devices. Although sensor augmented gloves [1]–[3] have been reported to typically yield higher gesture recognition rates than camera-based systems [4]–[6], they cannot capture the intricacies of sign languages presented through head and body movements. In contrast, video can capture facial expressions; but require adequate light and a direct line-of-sight to be effective.

Equally significant to these sensing limitations is the perspectives the Deaf community towards these sensing technologies. In a focus group we conducted with 7 Deaf participants [7], we found that deaf ASL users felt frustrated by wearable gloves, which they described as inaccurate and invasive, while concerned with video-based surveillance in the home. They also expressed excitement towards the potential to one day have Deaf-friendly personal assistants or non-invasive technology-augmented smart environments.

In these regards, RF sensors bring important advantages. RF sensors are non-contact and completely private, fully operational in the dark, and not affected by the color, fabric, or texture of clothes and presence of accessories, such as jewelry or watches. Not only can they provide a sensing capability when video or wearables are not effective, but they can also provide direct measurement of signing kinematics via the micro-Doppler (mD) signature [8] and range-Doppler maps. In previous work [7], we have found that machine learning can be used on RF sensor data to distinguish native ASL signing from copysigning by hearing individuals. This is also known as "imitation" signing, to emphasize the linguistic and kinematic differences that are observable in the RF data [9]. It has been reported that it can take learners of sign language *at least 3 years* to produce signs in a manner that is perceived as fluent by native signers [10]. Because it is much easier to recruit hearing participants than deaf participants, many studies on ASL recognition (e.g. [6], [11]–[13]) have used imitation signing data, despite its differences from native ASL data.

This paper offers four key contributions to ASL recognition and RF signal classification:

1) The gap between imitation and native signing is bridged using domain adaptation, and shown to boost word-level recognition accuracy of 20 signs from 46.2% to 88%.
2) Design of a physics-aware, multi-branch generative adversarial network (MB-GAN) [14] for RF mD signature synthesis is shown to surpass imitation-to-native adaptation results, yielding a word-level recognition accuracy of 91.3%.
3) Sequential classification of a continuous RF data stream, which includes daily activities mixed in with 4-word length signing sequences is accomplished with 92.3% accuracy.
4) Trigger sign recognition for "waking" an RF-enabled ASL-sensitive HCI is demonstrated with 0.93 trigger

recognition rate.

## II. Experimental Datasets and Pre-processing

### A. RF ASL Datasets

The RF data of ASL (ASL-R) used in this study were acquired by a TI AWR1642BOOST 77 GHz frequency modulated continuous wave (FMCW) transceiver. Measurements were made of both native signing from deaf or child-of-deaf-adult (CODA) participants fluent in ASL, and imitation signing from hearing participants mimicking copysigning videos of fluent signers.

- *20-word Native ASL-R Dataset:* 980 samples (49 per class) from 5 deaf/CODA participants were acquired at bandwidth settings of 750 MHz and 4 GHz.
- *20-word Imitation ASL-R Dataset:* 1550 samples from 10 hearing copysigners were acquired at bandwidth settings of 1.5 GHz and 4 GHz.
- *Sequential Mixed Motion Dataset:* Sequences of four ASL signs (YOU (Y), HELLO (H), CAR (C), PUSH (P)) in different orders mixed with different daily activities (walking, sitting, standing, folding laundry, and ironing) were continuously recorded for a duration of 30 seconds. While imitation signing gestures were enacted about 1.5 m from the RF sensor, daily activities were performed at varying distances. A total of 195 samples were recorded of 65 iterations of 3 different sequences were enacted by 4 hearing participants.

Note that participants were presented with a random ordering of single-word signs (see Table 1) to foster independence in each repetition of the signs. Figure 1 shows samples of the word-level and sequential ASL RF mD signatures.

### B. RF Data Representations

The received signal of an RF sensor is a complex I/Q time series, from which line-of-sight distance and radial velocity maybe computed. The amplitude and phase of those complex data are related to the electromagnetic scattering and kinematics of the motion being observed. The mD signature is the time-frequency transform of the I/Q data, and is often computed as the square modulus of the Short-Time Fourier Transform (STFT) of the continuous-time input signal. It reveals the distinct patterns caused by micro-motions, small rotations or vibrations, such as generated during gesturing and daily activity. The STFT itself is computed using Hanning windows with 50% overlap to reduce sidelobes in the frequency domain and convert the 1D complex time stream into a 2D $\mu$D signature. Reflection from static objects can be removed using moving target indicator (MTI) filters whereas sensor noise and artifacts can be mitigated using thresholding.

Range-Doppler (RD) maps are generated by applying 2D Discrete-Time Fourier Transform (DTFT) on the raw data matrix over the coherent processing interval (CPI). Thus, for each pulse, a RD map can be computed, resulting in a slow-time sampling rate of 3200 Hz for a pulse repetition interval (PRI) of 0.3 ms.

## III. Word-Level ASL Recognition

Compilation of large datasets for training state-of-the-art deep neural networks is difficult when human subjects are involved, not only because of the time involved in measuring numerous iterations of each class, but also because it can be difficult to recruit participants, especially if from a minority population, such as the Deaf community. In previous work [7], 20 native ASL signs were classified with an accuracy of 72.5% using minimum-redundancy maximum-relevance (mRMR) selection of 150 handcrafted features extracted from a five node multi-frequency RF sensor network and a random forest classifier. To surpass this performance with just a single sensor, recent advances in deep learning, which have yielded great advances in related fields [15], can be applied. However, deep neural networks (DNNs) rely on large amounts of training data to learn the underlying representations of each class.

One possible approach could be to try to address the data scarcity problem by training the DNN on imitation data, while testing on native ASL data. Unfortunately, this approach is not effective because imitation signing does have kinematic flaws that render them distinguishable from native ASL signing. This is evidenced not only by their differentiability using a support vector machine classifier [7], but also by the poor classification approach attained when deep learning is applied. When a convolutional neural network (CNN) is pre-trained on imitation data and fine tuned with 80% of the native ASL-R dataset, only 46.15% accuracy was attained when testing on the remaining 20% of native ASL-R data.

Consequently, in this study we compare two alternative

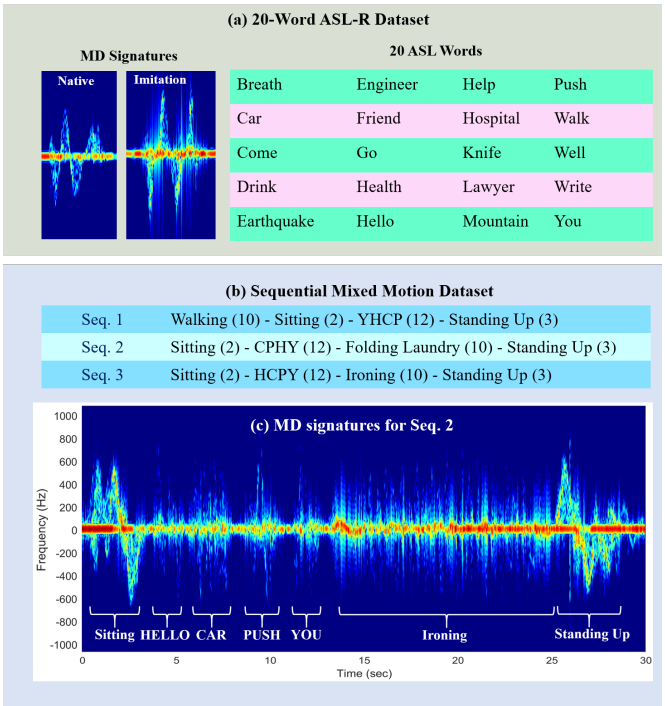


Fig. 1. RF datasets: (a) 20-word ASL-R dataset, (b) table of sequential signing gestures and (c) sequential mixed motion data sample.

approaches to training under low sample support: 1) adversarial domain adaptation to transform imitation signing data to resemble true native samples, and 2) synthetic training sample generation using generative adversarial networks (GANs).

### A. *Transformation of Imitation to "Fake Native" Data*

There are various approaches for image-to-image translation in the literature. We found that the CycleGAN [16] architecture offered better performance when compared to alternatives, such as TravelGAN [17]. CycleGAN translates an image from a source domain A to a target domain B by forming a series connection between two GANs to form a "cycle": the first GAN tries to synthesize "fake native" from the imitation data, while the second GAN works to reconstruct the original sample, synthesizing "fake imitation" samples. Thus, the network tries to minimize the cycle consistency loss, i.e. the difference between the input of the first GAN and the output of second GAN. An example of the input imitation signature and resulting fake native signature is shown in Figure 2. In this study, 25% of the imitation ASL-R data and 40% of the native ASL-R data was reserved for testing, while the remaining was used during training.

The resulting fake native signatures were then used to pre-train a three-layer convolutional autoencoder (CAE), which has been shown to surpass transfer learning in efficacy on small RF datasets [18], [19]. In each layer, a filter concatenation technique [20] is employed in which a filter size of $3 \times 3$ and $9 \times 9$ were concatenated to take advantage of multilevel feature extraction. After training the CAE model, the decoder was removed and two fully connected layers with 128 neurons followed by a dropout of 0.55 were added after flattening the output of the encoder. At the end, a softmax layer with 20 nodes is employed for classification. As shown in Figure 2, when 80% of the native ASL-R data is utilized an accuracy of 88% is achieved.

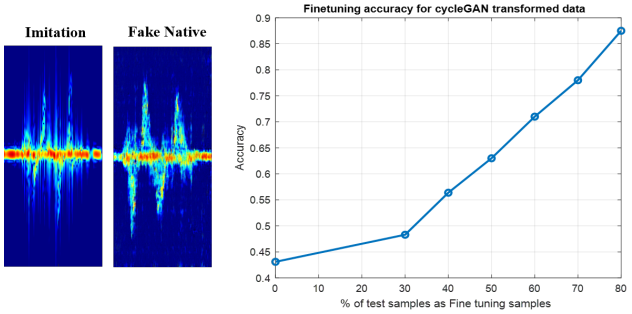


Fig. 2. CycleGAN Results.

### B. *Synthetic Training Data Generation*

Although GANs have resulted in amazing results for computer vision applications, GAN-synthesized RF micro-Doppler signatures have been shown to exhibit significant kinematic errors [21]. This causes the generation of synthetic samples that do not correspond to any physically realizable motion.

One way to reduce such kinematically impossible samples is to amplify adherance of synthetic samples to the envelope of the signature. The envelope represents the fastest moving points on the body, and, in the case of signing and hand gestures, this corresponds to the maximum velocities attained by the hand in motion. Envelope features can be emphasized by incorporating the envelope as an additional, second branch in the discriminator of the GAN, resulting in a multi-branch GAN (MB-GAN) architecture [14]. This has been shown to result in closer correspondence between the synthetic data and the real RF measurements.

In this study, an MB-GAN with 8 convolutional layers is constructed. Each layer is followed by batch normalization with 0.9 momentum and a ReLU activation function. The main branch of the discriminator is a 6-layer CNN where each layer followed by a Leaky-ReLU activation function. In the secondary branch, three 1D-convolutional layers are with the envelope as input. Afterwards, the outputs of the dense layer is concatenated with the flattened output of the main discriminator. This architecture is shown in Figure 3.

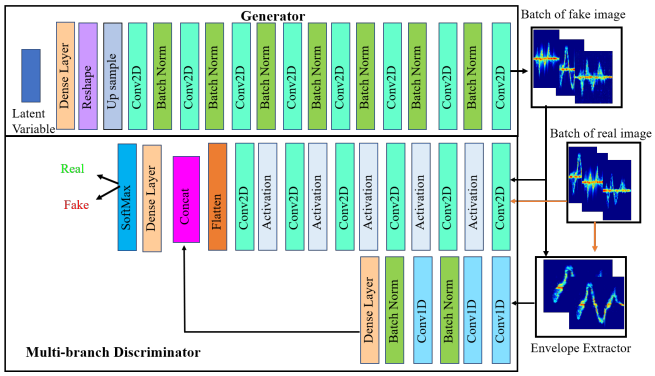


Fig. 3. MB-GAN Architecture.

A total of 10,000 synthetic samples for 20 ASL signs were generated using MB-GAN trained on 75% of the native ASL-R samples. Using these MB-GAN synthesized samples to pre-train the same CAE architecture mentioned earlier, a classification accuracy of 91.3% was achieved with only 30% of the native ASL-R samples used in fine-tuning. This is a 3% increase in accuracy relative to pre-training on CycleGAN-transformed "fake native" data. Moreover, this result was attained with much fewer native ASL-R samples in fine-tuning: just 30% rather than 80%.

## IV. Sequential Classification

Although there has been much work on sequential classification [22] and trigger word spotting [23], [24], [25] in the speech recognition literature, there have been few studies of sequential classification of continuous RF data time streams [26], [27]. Moreover, these studies focus exclusively on gross body motion classification. This study is to the best of our knowledge the first to consider heterogeneous motion sequences, with fine-scale signing gestures mixed with daily activities.

Our approach consists of a two stage procedure. First, we decide whether there is a moving target in the line-of-sight of the radar using range-weighted energy (RWE) plots, which are obtained from RD maps and a short time average over long time average (STA/LTA) based motion detector. Only the samples identified as movements are then classified. Thus, this approach minimizes computational load by eliminating irrelevant data with no motion - a common occurrence in daily living scenarios.

The RWE value of a RD map can be calculated by dividing each pixel intensity value by its range and summing all of the range weighted pixel values up. In this way, when there is a moving target close to the RF sensor, there will be a peak in the RWE plot and it can be detected by the motion detector.

### A. STA/LTA Based Motion Detector

STA/LTA is one of the most broadly used algorithms in motion detection applications. The algorithm continuously keeps track of the average value changes in leading (short) and lagging (long) windows, and the system gets triggered if the STA/LTA ratio goes below a pre-defined threshold. The algorithm has four parameters: short window length, long window length, detection window length, and threshold value. Although the selection of these parameters depends on the application (i.e. desired false alarm rate (FAR), false rejection rate (FRR)), in this study, we empirically optimized them. The detection window is used to select samples to send to the classifier. Figure 4 shows how these windows are positioned. The FRR and FAR are defined as

$$FRR = \frac{T - D}{T}, \qquad FAR = \frac{F}{T} \tag{1}$$

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.

### B. Sequential Classification of Motion Segments

Classification of time series data is done using RD maps. After motion detection with the detector, RD map samples in the detection window goes into a time-distributed (TD) 2D convolutional neural network (CNN) which is followed by a bidirectional LSTM (Bi-LSTM) layer and a TD *softmax* layer. The TD wrapper allows us to apply a layer to each temporal slice. As a result, we can obtain a prediction for each frame using softmax layer with TD wrapper. 80% of each participant's data are used for training and 20% for testing. This results in an overall accuracy of 92.3%.

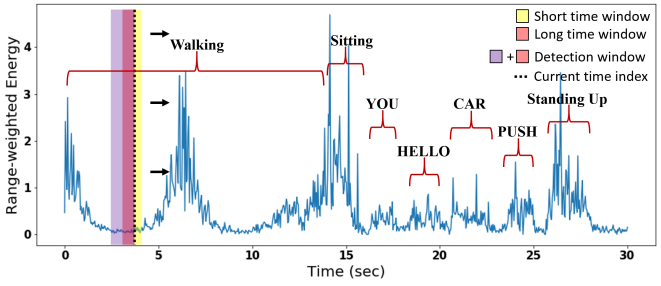


Fig. 4. *STA/LTA* based motion detector.

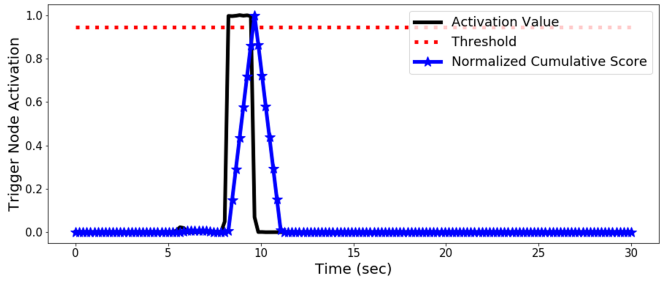


Fig. 5. Trigger detection of the word HELLO.

## V. Trigger Sign Detection

The problem of trigger sign recognition has not yet been adequately addressed in the literature relating to RF sensing. For speech triggered devices, Apple Siri [28] employs a cumulative score aggregation (CSA) algorithm, which ensures the system is triggered only when the phrase is fully completed. In this study, we track the signing process by recording the activation value of the output node of the target class - in this case, HELLO. If activation values are directly used to trigger the system, the system gets triggered at the very beginning of the signing, potentially causing high FA rates since different signs may have a resemblance in initial pattern. This problem is mitigated by CSA, as illustrated in Figure 5. Scores of the last 1.5 seconds of data are accumulated to calculate the cumulative score. Notice that CSA triggers upon completion of the sign, as desired. When the detection threshold is set to 90% of the maximum achievable score, using the activation values as scores has FRR of 0.18 and FAR of 0.03, while CSA method's FRR stays as low as 0.07 and FAR becomes 0. Thus, trigger recognition rate for the word HELLO can be calculated as $1 - 0.07 - 0 = 0.93$.

## VI. Conclusion

This paper has presented word-level ASL recognition and trigger sign detection results using a 77 GHz FMCW RF sensor. In particular, the recognition of up to 20 ASL signs, sequential classification of signing mixed with daily activity, and detection of a trigger sign to wake a device via RF sensors has been demonstrated. A physics-aware, multi-branch GAN was designed to synthesize samples for training a three-layer CAE for classification of native ASL RF signatures with an accuracy of 91.3%. Sequential classification of continuous data streams of signing interwoven with daily activities was accomplished with 92.3% accuracy, while a trigger sign was detected at a rate of 0.93.

## VII. Acknowledgments

This work was funded in part by the National Science Foundation (NSF) Awards #1932547, #1931861, and #1734938. Human studies research was conducted under UA Institutional Review Board (IRB) Protocol #18-06-1271.

## REFERENCES


[1] 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. PP, 03 2015.

[2] Y. Li, X. Chen, X. Zhang, K. Wang, and Z. J. Wang, "A sign-component-based framework for chinese sign language recognition using accelerometer and semg data," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 10, pp. 2695–2704, 2012.

[3] B. G. Lee and S. M. Lee, "Smart wearable hand device for sign language interpretation system with sensors fusion," *IEEE Sensors Journal*, vol. 18, no. 3, pp. 1224–1232, 2018.

[4] G. García-Bautista, F. Trujillo-Romero, and S. O. Caballero-Morales, "Mexican sign language recognition using kinect and data time warping algorithm," in *2017 International Conference on Electronics, Communications and Computers (CONIELECOMP)*, 2017, pp. 1–5.

[5] V. Bheda and D. Radpour, "Using deep convolutional networks for gesture recognition in american sign language," *CoRR*, vol. abs/1710.06836, 2017. [Online]. Available: http://arxiv.org/abs/1710.06836

[6] L. Pigou, M. Van Herreweghe, and J. Dambre, "Gesture and sign language recognition with temporal residual networks," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 3086–3093.

[7] 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.

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

[9] 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.

[10] J. S. Beal and K. Faniel, "Hearing l2 sign language learners: How do they perform on asl phonological fluency?" *Sign Language Studies*, vol. 19, no. 2, pp. 204–224, 2018.

[11] 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. on Embedded Network Sensor Systems*, 2017.

[12] J. Shang and J. Wu, "A robust sign language recognition system with sparsely labeled instances using wi-fi signals," in *Proc. IEEE 14th Int. Conf. on Mobile Ad Hoc and Sensor Systems*, 2017, pp. 99–107.

[13] 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.

[14] B. Erol, S. Z. Gurbuz, and M. G. Amin, "Synthesis of micro-doppler signatures for abnormal gait using multi-branch discriminator with embedded kinematics," in *2020 IEEE International Radar Conference (RADAR)*, 2020, pp. 175–179.

[15] 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.

[16] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[17] M. Amodio and S. Krishnaswamy, "Travelgan: Image-to-image translation by transformation vector learning," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8975–8984, 2019.

[18] M. S. Seyfioğlu and S. Z. Gürbüz, "Deep neural network initialization methods for micro-doppler classification with low training sample support," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2462–2466, 2017.

[19] 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.

[20] X. Bai, Y. Hui, L. Wang, and F. Zhou, "Radar-based human gait recognition using dual-channel deep convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 9767–9778, 2019.

[21] B. Erol, S. Z. Gurbuz, and M. G. Amin, "Motion classification using kinematically sifted acgan-synthesized radar micro-doppler signatures," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 4, pp. 3197–3213, 2020.

[22] 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

[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] P. D. S. A. N. L. S. V. D. N. A. S. Y. M. S. P. J. D. Williams, "Lattice-based improvements for voice triggering using graph neural networks," 2020. [Online]. Available: https://arxiv.org/pdf/2001.10822.pdf

[25] S. Sigtia, P. Clark, R. Haynes, H. Richards, and J. Bridle, "Multi-task learning for voice trigger detection," 2020. [Online]. Available: https://arxiv.org/pdf/2001.09519v1.pdf

[26] 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.

[27] A. Shrestha, H. Li, J. Le Kernec, and F. Fioranelli, "Continuous human activity classification from fmcw radar with bi-lstm networks," *IEEE Sensors Journal*, vol. 20, no. 22, pp. 13 607–13 619, 2020.

[28] "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.