Multi-Frequency RF Sensor Fusion for Word-Level Fluent ASL Recognition

Sevgi Z. Gurbuz – University of Alabama
M. Mahbubur Rahman – University of Alabama
Emri Kurtoglu – University of Alabama
Evie Malaia – University of Alabama
Ali C. Gurbuz – Mississippi State University
Darrin J. Griffin – University of Alabama
Chris Crawford – University of Alabama

Citation of published version:
Guruz, S., Rahman, M., Kurtoglu, E., Malaia, E., Gurbuz, A., Griffin, D., Crawford, C. (2021): Multi-Frequency RF Sensor Fusion for Word-Level Fluent ASL Recognition. *IEEE Sensors Journal*.

DOI: https://doi.org/10.1109/JSEN.2021.3078339

Multi-Frequency RF Sensor Fusion for Word-Level Fluent ASL Recognition

Sevgi Z. Gurbuz, *Senior Member, IEEE*, M. Mahbubur Rahman, Emre Kurtoglu, Evie Malaia, Ali C. Gurbuz, *Senior Member, IEEE*, Darrin J. Griffin, and Chris Crawford

***Abstract*—Deaf spaces are unique indoor environments designed to optimize visual communication and Deaf cultural expression. However, much of the technological research geared towards the deaf involve use of video or wearables for American sign language (ASL) translation, with little consideration for Deaf perspective on privacy and usability of the technology. In contrast to video, RF sensors offer the avenue for ambient ASL recognition while also preserving privacy for Deaf signers. Methods: This paper investigates the RF transmit waveform parameters required for effective measurement of ASL signs and their effect on word-level classification accuracy attained with transfer learning and convolutional autoencoders (CAE). A multi-frequency fusion network is proposed to exploit data from all sensors in an RF sensor network and improve the recognition accuracy of fluent ASL signing. Results: For fluent signers, CAEs yield a 20-sign classification accuracy of %76 at 77 GHz and %73 at 24 GHz, while at X-band (10 Ghz) accuracy drops to 67%. For hearing imitation signers, signs are more separable, resulting in a 96% accuracy with CAEs. Further, fluent ASL recognition accuracy is significantly increased with use of the multi-frequency fusion network, which boosts the 20-sign fluent ASL recognition accuracy to 95%, surpassing conventional feature level fusion by 12%. Implications: Signing involves finer spatiotemporal dynamics than typical hand gestures, and thus requires interrogation with a transmit waveform that has a rapid succession of pulses and high bandwidth. Millimeter wave RF frequencies also yield greater accuracy due to the increased Doppler spread of the radar backscatter. Comparative analysis of articulation dynamics also shows that imitation signing is not representative of fluent signing, and not effective in pre-training networks for fluent ASL classification. Deep neural networks employing multi-frequency fusion capture both shared, as well as sensor-specific features and thus offer significant performance gains in comparison to using a single sensor or feature-level fusion.**

***Index Terms*—American sign language, gesture recognition, radar micro-Doppler, RF sensing, deep learning, autoencoders**

I. Introduction

MOST indoor environments designed by hearing individuals present a variety of challenges to Deaf individuals, who primarily perceive the world through visuo-spatial awareness. American Sign Language (ASL) is widely used in the Deaf community as a visual-kinetic mode of communication, which requires direct visual observation of other signers to be effective. Thus, Deaf spaces [1] - indoor environments designed to optimize Deaf cultural expression - involve modifications, such as a higher number of windows, increased lighting, furniture re-arrangements and even openings in windows that allow clear sightlines, to improve visual accessibility between signers.

Research on sensing technologies for the Deaf has primarily focused on the use of video [2], [3] or wearable devices [4], [5] for ASL translation and facilitating the understanding of hearing individuals of Deaf communications. The objective of this work, however, is to instead focus on machine understanding of ASL as a means for better designing technology to serve the Deaf community. Through the involvement of community partners, such as the Alabama Institute of the Deaf and Blind (AIDB), we aim to reflect a Deaf-centric approach to ASL recognition, which reflects Deaf perspectives. Wearable devices are intrusive, restrict natural hand usage, and burden Deaf individuals based on "hearing" perceptions of deafness as a disability, as opposed to a unique sub-culture of American society. Video-based technologies, such as video-based cell phone communication apps, are often used by the Deaf to great benefit; however, in the context of smart environments, where video would offer constant opportunity for surveillance, cameras elicit significant concern over privacy.

In contrast, radio frequency (RF) sensors can operate even in the dark, in a non-contact fashion. RF sensors only record the range and velocity profiles of the signing motion, and, thus, even if hacked, completely protect the privacy of the individual (e.g. face) and environment (no visual background information). Although RF sensing cannot offer complete perception of sign language, which also involves facial expressions and hand shapes, RF sensors are responsive to the kinematics and position of the hands. Radar point clouds [6] may be extracted from multi-channel RF sensors, however, transmit waveform bandwidth bounds slant range resolution, while the number of channels limits azimuth resolution. This results in spatially and temporally sparse point clouds that are not effective in capturing hand shape or motion dynamics. Alternatively, the short-time Fourier transform (STFT) can be used to compute the spectrogram of the RF sensor returns, which reveals the unique patterns of micro-Doppler [7] frequency resulting from

S.Z. Gurbuz, M.M. Rahman, and E. Kurtoglu are with the Dept. of Electrical and Computer Engineering, University of Alabama, Tuscaloosa, AL, 35487 USA (e-mail: szgurbuz@ua.edu, mrahman17@crimson.ua.edu, ekurtoglu@crimson.ua.edu.

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

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

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

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

signing. In recent prior work [8]–[10], we showed that the RF micro-Doppler signatures are effective in capturing both linguistic and kinematic properties of signing. Handcrafted features were extracted from the micro-Doppler signatures obtained from an RF sensor network and utilized to classify 20 ASL signs with %72.5 accuracy [10].

This paper explores the effect of various RF sensor transmit waveform parameters and the fluency of ASL users on the performance attained by deep learning based approaches to RF-based ASL recognition. Over the past five years, a significant body of work on RF sensor-based gesture recognition [11]–[13] has emerged in the literature. While some researchers have used "sign language gestures" to test gesture recognition approaches, in fact, signing possesses a much greater degree of complexity and nuance [14], [15]. Due to the communicative and linguistic nature of the signal, signing presents additional challenges relating to fine-grained temporal dynamics and linguistic parameters, such as prosody (e.g. pauses and suprasegmental components, such as phrase-final lengthening [16]) and grammatical structure. Moreover, the signing of hearing imitation signers is distinguishable from that of fluent ASL signers [10], exhibiting greater kinematic variation, more erratic cadence and significant signing errors. Thus, both the transmit waveform parameters can affect the extent to which the RF sensor accurately captures motion during signing.

Although some studies, e.g. [17]–[21], of ASL recognition have employed hearing imitation signers or ASL learners, perhaps due to the greater ease in recruiting a larger number of participants, the intended benefactor of Deaf spaces are fluent ASL signers. Thus, in this paper, we investigate the performance of transfer learning and convolutional autoencoders in the classification of fluent ASL signing and show that 1) the accuracy achieved by deep neural networks (DNNs) on imitation signing data is overly optimistic (higher) than that achieved with fluent ASL data; and that 2) imitation signing data is not effective in pre-training networks intended to classify fluent ASL signing data. Furthermore, we compare the performance achieved with RF sensors with different transmit waveform, center frequency, pulse repetition frequency (PRF), and bandwidth. Finally, we propose a multi-frequency DNN for fusing the simultaneous measurements of three RF sensors transmitting at different frequencies, boosting accuracy relative to that achieved with feature-level fusion.

The paper is organized as follows. In Section II, a description of the RF sensor network and acquired datasets is given. Section III examines the variation in DNN classification accuracy across different transmit waveforms for both fluent and imitation signing. In Section IV, the design of the multi-frequency fusion DNN is presented, and results are compared with that obtained from alternative fusion approaches. Section V concludes the paper with a discussion of main implications.

II. RF Measurements of ASL

Three different RF sensors are utilized in this work: 1) the TI IWR1443BOOST 76 GHz - 81 GHz automotive short-range radar (SRR) sensor, which has frequency modulated continuous wave (FMCW) tranmissions; 2) the Ancortek 2400AD transciever, which transmits FMCW with a center frequency of 24 GHz; and 3) the Xethru X4M03 ultra-wide band (UWB) impulse radar with a transmission frequency range of 7.25 - 10.2 GHz. Measurements were acquired with 77 GHz automotive radar at two different bandwidths, namely, 750 MHz and 4 GHz, while the 24 GHz radar was operated with a bandwidth of 1.5 GHz, and the Xethru radar had a bandwidth of 3 GHz. While the bandwidth of the 77 GHz sensor is adjustable, the 24 GHz sensor allows for selection among only three possible bandwidths, 1.5 GHz being the heighest, and the bandwidth of the Xethru sensor is fixed.

Nominal Transmit Frequency	77 GHz FMCW		24 GHz FMCW	10 GHz UWB
Bandwidth (MHz)	750	4000	1500	3000
Pulse Duration (ms)				-
Chirp Rate (MHz/μs)	75	400	1.5	-
# Samples/Class (Native ASL)	20	30	50	50
# Samples /Class (Copy Signing)	100	80	100	80
ASL SIGNS ACQUIRED				
YOU	KNIFE		LAWYER	HELP
HELLO	WELL		HOSPITAL	PUSH
WALK	CAR		HEALTH	GO
DRINK	ENGINEER		EARTHQUAKE	COME
FRIEND	MOUNTAIN		BREATHE	WRITE

Fig. 1. Summary of RF sensor parameters and ASL signs in study.

The three sensors were placed side by side, directly facing the participants, at an elevation of 0.91 m from the ground. Participants sat on a chair directly facing a computer monitor, which was placed immediately behind the radar systems, and used to relay prompts indicating the signs to be articulated. The radar systems were positioned at a distance of 1.2 - 1.5 meters from the participant. The output transmission power of the RF sensors are 4.3 mW, 100 mW, and 40 mW for the Xethru 10 GHz UWB, Ancortek 24 GHz FMCW, and TI 77 GHz FMCW transceivers, respectively. These levels are lower than those incurred during cell phone usage, e.g. 250 mW to 2 W [22], and are further reduced by propagation losses proportional to $1/r^2$, where r is the distance between the sensor and user. This study was approved by the Institutional Review Board (IRB) of the University of Alabama.

A total of 6 fluent ASL users took part in the study of whom 3 were Deaf and 3 were a Child-of-Deaf Adult (CODA). A total of 15 hearing participants, who did not know sign language, also participated. Hearing participants were first tutored for about 10-15 minutes on how to articulate the desired signs. During the experiment, hearing participants were prompted with a copy-signing video were a CODA articulated the sign and afterwards the participant was expected to repeat the same sign. Participants were presented a random ordering of single-word signs to foster independence in the repetition of the signs. The 20 signs considered in this study were selected using the ASL-LEX [23] database (http://asl-lex.org), choosing words that are higher frequency, but not phonologically related to ensure a more diverse dataset. The specific ASL signs used as well as the number of samples acquired for different radar waveform types and transmit parameters are listed in Fig. 1.

III. Influence of RF Transmit Waveform Parameters and Fluency on ASL Recognition

The signal received by a radar is, in general, weighted summation of time-delayed, frequency-shifted versions of the transmitted signal from multiple scatterers. In many practical scenarios, it has been shown that the scattering from the human body can be approximated using the superposition of returns from K points on the body [24]. Thus,

$$x[n] = \sum_{i=1}^{K} a_i exp\left\{ -j\frac{4\pi f_c}{c} R_{n,i} \right\}, \tag{1}$$

where $R_{n,i}$ is the range to the i^{th} body part at time n, f_c is the transmit center frequency, c is the speed of light, and the amplitude a_i is the square root of the power of the received signal as given by the radar range equation [25]. Thus, RF sensors provide a complex-time series of measurements in the form $x[n] = I[n] + jQ[n]$.

Typically, this data stream is re-shaped into a 2D matrix for each RF receive channel, so that the columns represent fast-time, e.g. range samples, and the row represents slow-time, e.g. pulse number. The range between the radar and any scattering point is found from the round-trip travel time, while the radial velocity of motion, v_r, is given by computation of the Doppler shift,

$$f_D = \frac{2v_r f_t}{c} \tag{2}$$

where f_t is the instantaneous transmit frequency. In addition, the range and velocity estimates obtained from RF sensors are independent measurements.

A. RF Data Pre-Processing

The kinematic behavior of the signer is captured by the frequency modulations in the phase of the received signal. Micro-motions [7], e.g. rotations and vibrations, result in micro-Doppler (μD) frequency modulations centered about the main Doppler shift, which is caused by translational motion. Signing results in a time-varying pattern of micro-Doppler frequencies. Each sign generates its own unique patterns, which can be revealed through time-frequency analysis. The *micro-Doppler signature*, or spectrogram, is found from the square modulus of the Short-Time Fourier Transform (STFT) of the continuous-time input signal $x(t)$ and can be expressed in terms of the window function, $h(t)$, as

$$S(t,\omega) = \left| \int_{-\infty}^{\infty} h(t-u)x(u)e^{-j\omega t}du \right|^2. \tag{3}$$

Ground clutter from stationary objects, such as furniture and the walls, will appear in the micro-Doppler signature as a band centered around 0 Hz. Based on earlier studies [10], we found that for the 10 GHz and 24 GHz RF sensors, performance is improved with removal of the ground clutter via high pass filtering. At 77 GHz, however, the elimination of low-speed signal components during clutter filtering results in performance degradation [10]. Thus, a 4$^{\text{th}}$-order Butterworth high pass filter was applied only to the 10 GHz and 24 GHz RF sensor data. Samples of the micro-Doppler signatures for fluent ASL signers as acquired from the different RF sensors are shown in Figure 2.

B. Transmit Frequency

The transmit frequency has a significant impact on the perception of micro-motions by the RF sensor. As revealed by Eq. 2, the higher the transmit frequency, the greater a Doppler shift is observed. For movements such as signing, where the finer-scale motion is involved both temporally and spatially, transmission at higher frequencies has great benefits: even small movements result in observable Doppler shifts, resulting in greater detail in the time-frequency representation, i.e. the micro-Doppler signature of the motion. Both the 77 GHz and 24 GHz FMCW sensors appear to acquire much crisper μD signatures in comparison to the 10 GHz UWB radar. For example, for the sign WALK as illustrated in Fig. 2, the number of times that the hand waves back and forth can only be clearly counted in 77 GHz and 24 GHz data.

C. FMCW Transmit Waveform Parameters

An ideal FMCW waveform may be specified using three parameters: 1) the pulse duration, τ, 2) the bandwidth, β, and 3) the number of pulses transmitted, N. The range resolution, ΔR, is dependent upon the waveform bandwidth as $\Delta R = c\beta/2$, while the velocity resolution, Δv, is a function of the total coherent duration that the radar interrogates the target, i.e. dwell time, as computed from $\Delta v = \lambda/N\tau$, where λ is the wavelength of the transmit waveform. Thus, the greater the bandwidth, the better the range resolution; and the higher the transmit frequency, the shorter the wavelength and better the velocity resolution. Because signing can be quite dynamic with rapid progressions, keeping the pulse duration as short as possible would be an advantage as this also increases the sampling rate across Doppler frequency. Moreover, the maximum unambiguous velocity depends upon the pulse duration (which for FMCW equals the pulse repetition interval): $-1/\tau < f_D < 1/\tau$, where f_D is as defined in Eq. 2. The shorter the pulse duration, the greater is the maximum unambiguous velocity that can be measured.

In real FMCW transmitters, however, additional parameters factor into the specification of the transmit waveform, as shown in Figure 3. For example, due to the finite switching time of the transceiver, there is a short duration between the transmission of each pulse, known as *idle time*, t_{idle}. In the user interface of the TI 77 GHz sensor, not just the idle time, but also a *frame period*, T, can be specified. The term *frame* is borrowed from video processing literature, but, in this case, refers to the 2D range-Doppler map that is computed from returns received from N pulses transmitted over a *coherent processing interval (CPI)*. An *inter-frame period*, t_{if}, can also be specified in the user interface to allow for a time delay between successive CPIs. Thus, the duty cycle, d, of the entire transmission can be defined as

$$d = \frac{N \times t_{chirp}}{T}, \tag{4}$$

where t_{chirp} is the chirp cycle time.

For the purposes of sign language recognition, we recommend that the transmit waveform not only have the minimum possible pulse duration and maximum possible bandwidth, but also a duty cycle as close as possible to 100%; e.g.,

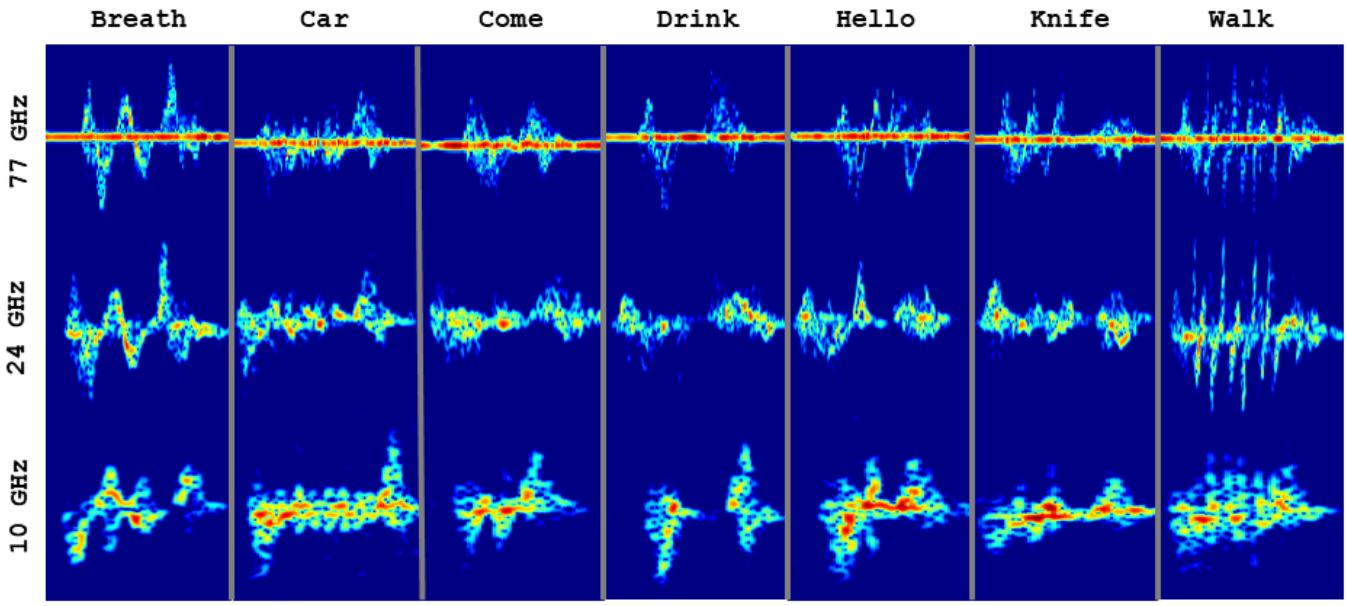


Fig. 2. Comparison of example micro-Doppler signatures for fluent ASL signing as measured by various RF sensors. The 77 GHz signatures are shown for the high PRF of 6400 Hz.

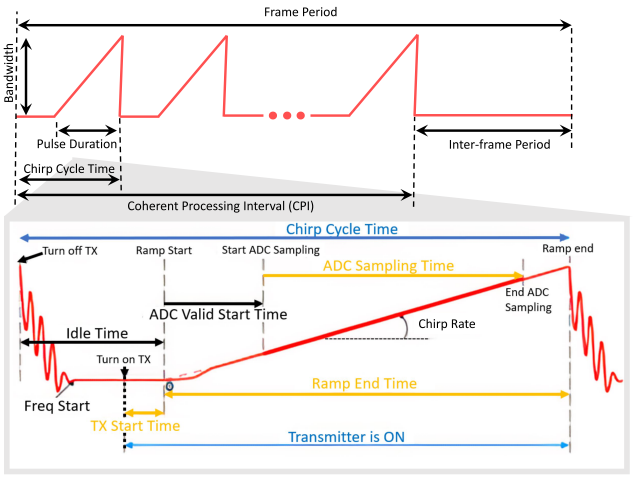
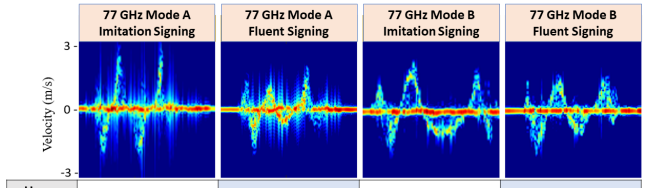


Fig. 3. TI 77 GHz mmWave Studio chirp design parameters.

	Number of chirp loops	t_{idle} (µs)	d (%)	τ (µs)	β (GHz)
Min.	1	2	>0	50	0
Max.	256	5242	100	-	4

Fig. 4. Minimum and maximum values of waveform parameters when other parameters are selected as follows: *# of ADC samples: 256, ADC sampling frequency: 6.25 kbps. (Min/max T depends on N and τ.)*

minimum idle time and inter-frame period. The minimum and maximum values that each parameter may be assigned in TI mmWaveStudio are listed in Fig. 4. To see the effect of the duty cycle, and in particular, the inter-frame period on the acquired ASL data, consider the RF signatures acquired under two different settings for hearing imitation and fluent signer, shown in Fig. 5:



	77 GHz Mode A Imitation Signing	77 GHz Mode A Fluent Signing	77 GHz Mode B Imitation Signing	77 GHz Mode B Fluent Signing
Upper Envelope Mean	0.824	0.709	0.498	0.318
Upper Envelope Max	2.418	2.189	2.079	1.621
Lower Envelope Mean	−0.861	−0.863	−0.773	−0.631
Lower Envelope Max	−2.527	−2.332	−2.219	−1.825

Fig. 5. Comparison of the μD spectrograms for "BREATH" and their envelope statistics for fluent and imitation signers.

- *Mode A:* 77 GHz, $\beta = 750$ MHz, PRF = 3.2 kHz, $\tau = 60\mu s$, $N = 128$, $d = 51.2\%$, $T = 40ms$, $t_{idle} = 100\mu s$, $t_{if} = 18.8ms$.
- *Mode B:* 77 GHz, $\beta = 4$ GHz, PRF = 6.4 kHz, $\tau = 50\mu s$, $N = 256$, $d = 96\%$, $T = 40ms$, $t_{idle} = 100\mu s$, $t_{if} = 0.3ms$.

Notice that when the waveform has a low duty cycle, and, hence, a significant inter-frame period, the signatures are effected by vertical streaking across Doppler, which corrupts the measurement. This is made more evident when the peak and mean of the upper and lower envelopes are compared across the two modes for both imitation and fluent signing. The peak values of the upper envelope and the minimum values of the lower envelope, i.e. the extreme velocities, are greater in the data from the corrupted Mode A waveform versus that from the pristine Mode B data.

Due to the erratic nature of imitation signing, which results

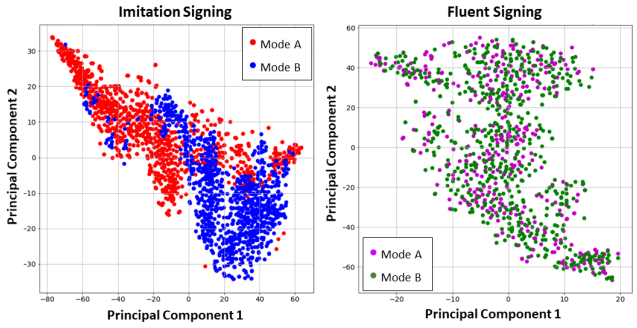


Fig. 6. Comparison of the effect of different RF transmit modes on the feature spaces of imitation and fluent signing data.

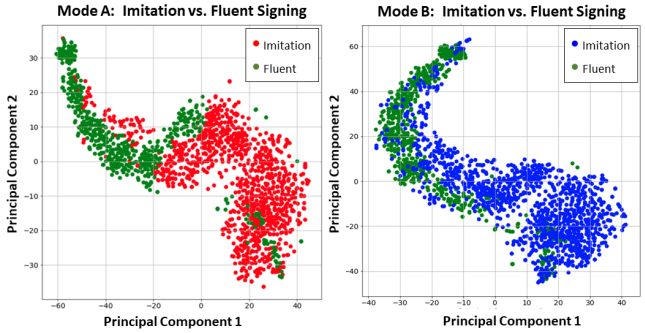


Fig. 7. Comparison of the overlap between the feature spaces of imitation signing and fluent signing data for Mode A and Mode B.

in greater micro-Doppler frequency diversity, the streaking effect appears to impact imitation signing more severely than fluent signing. This can also be seen by comparing the degree of overlap in the feature space spanned by fluent and imitation signing data for the two different transmit waveforms, Mode A and Mode B, as shown in Fig. 6. Principal Component Analysis (PCA) was used for feature extraction [26], while the T-distributed Stochastic Neighbor Embedding (t-SNE) [27] algorithm was used to visualize the feature spaces. The t-SNE visualization reveals that while in imitation signing data the differing transmit waveforms result in a tangible shift in the centroid and extent of the imitation signing feature space, fluent signing is not as affected by mode and the feature space spanned by both modes predominantly overlap.

D. Imitation Signing versus Fluent Signing

Studies of sign language have shown that it can take ASL learners *at least 3 years* to produce signs in a manner that is perceived as fluent by other fluent signers [28]. Visualizations of feature space as given by t-SNE can also be used to compare the extent to which imitation signing statistically resembles fluent signing. Consider Fig. 7, which shows the overlap between the feature spaces of imitation and fluent signing data for Mode A and Mode B. The overlap is greater when the Mode B transmit waveform parameters are utilized; but, in both cases, there is a significant discrepancy between the feature spaces of imitation signing versus fluent signing.

This discrepancy can be quantified by considering the ability of support vector machines (SVM) to classify imitation signing versus fluent signing using PCA. With a Mode A transmit waveform, SVM is able to distinguish imitation signing from fluent signing with an accuracy of %96. With a Mode B transmit waveform, which is optimized for spatiotemporal parameters of signing, the acquired signatures are pristine, and the accuracy to distinguish drops to %76. This level of capability to distinguish between fluent and imitation signers is still a high percentage, and reinforces the main point that imitation signing is not representative of fluent signing.

Thus, ASL recognition algorithms should not be validated using imitation signing data. Even in the context of ASL-sensitive human-computer interfaces (HCI), it should be remembered that the target audience for such technologies is the Deaf community and broader population of ASL users, who are fluent signers. Just as speech recognition systems would never be designed using vocalizations from non-speakers, so should ASL recognition systems not be evaluated using the imitation data of hearing non-signers.

E. Single-Sensor Classification Accuracy

Although DNNs have yielded great improvements in performance in many fields, including RF micro-Doppler signature classification [29], they rely on large amounts of training data to learn the underlying representations. However, RF sensing typically involves much fewer samples than in computer vision due to the cost and time to acquire data from human subjects. Several approaches have been proposed for addressing the training of DNNs when the amount of training data is limited: e.g., transfer learning and unsupervised pre-training. In prior work [30], [31], the efficacy of these methods on micro-Doppler signatures for human activities was investigated. Among pre-trained networks, VGGNet [32] was found to be more effective than GoogleNet [33], while the performance of CAEs surpassed that of VGGNet and convolutional neural networks (CNN) when the amount of training data exceeded 600 samples. Thus, the single-sensor classification accuracy for each sensor, imitation signing and fluent signing were compared for transfer learning using VGGNet and a CAE.

1) VGGNet: VGGNet is a 16-layer convolutional neural network (CNN), which uses 3×3 convolutional filters in each layer. Volume size is reduced using max pooling, with the convolutional layers followed by two fully-connected layers having 4096 nodes per layer and a softmax classifier. A slight modification to VGGNet was made in this work by utilizing global average pooling in the final layer, rather than max pooling. The two dense layers use ReLU activation functions, with each followed by 50% dropout. A batch size of 8, epoch number of 120, learning rate of 10^{-4}, momentum of 0.9, and the ADAM [34] optimizer were utilized.

VGGNet was initially pre-trained using 1.2 million optical images from the ImageNet [35] database. This results in improved initialization of the network weights relative to random initialization, while also reducing the number of RF samples required during training. Real RF data samples are thus only used for fine tuning and testing.

RF Sensor	DNN	Training Data	Testing Data	Accuracy
77 GHz FMCW with 4 GHz bandwidth	VGGNet	Pre-Train on ImageNet Fine-Tune on Imitation Signing	Imitation Signing	94%
	CAE	Imitation Signing	Imitation Signing	96%
	VGGNet	Pre-Train on ImageNet Fine-Tune on SMOTE-Augmented Fluent ASL	Fluent ASL	72%
	CAE	Fluent ASL	Fluent ASL	77.68%
24 GHz FMCW with 1.5 GHz bandwidth	VGGNet	Pre-Train on ImageNet Fine-Tune on Imitation Signing	Imitation Signing	75.90%
	CAE	Imitation Signing	Imitation Signing	77.11%
	VGGNet	Pre-Train on ImageNet Fine-Tune on SMOTE-Augmented Fluent ASL	Fluent ASL	73.22%
	CAE	Fluent ASL	Fluent ASL	74.55%
10 GHz UWB With 3 GHz bandwidth	VGGNet	Pre-Train on ImageNet Fine-Tune on SMOTE-Augmented Fluent ASL	Fluent ASL	68.20%
	CAE	SMOTE-Augmented Fluent ASL	Fluent ASL	68.90%

Fig. 8. Comparison of classification accuracy for imitation signing and fluent signing using various RF sensors.

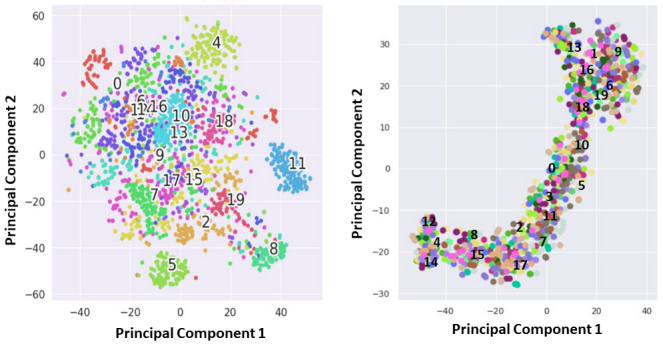


Fig. 9. Illustration of distribution of each ASL class for imitation signing (left) and fluent signing (right).

2) CAE: In this work, a three-layer CAE that employs multilevel feature extraction was utilized. A total of 128 convolutional filters with two different sizes (64 3×3 and 64 9×9) were applied and their outputs concatenated. After use of unsupervised pre-training to initialize network weights, the decoder was removed and replaced with two fully-connected layers having 128 neurons per layer. Dropout of 55% was added after flattening the output of the encoder. A softmax layer with 20 nodes was employed for classification.

3) Results: The recognition accuracy of 20 signs were compared across all RF sensors for imitation signing versus fluent signing. Only Mode B imitation signing data was utilized in these assessments, given the distinct differences demonstrated in Figure 6. In the case of fluent ASL data, both Mode A and Mode B data were utilized with equal proportions in training and test datasets. A ratio of 80% to 20% was used between training and test sets. The Synthetic Minority Over-sampling TEchnique (SMOTE) [36] was applied to equalize the number of real RF samples used for training when comparing imitation and fluent signing recognition accuracy. Results are tabulated in Figure 8.

In all cases, the CAE slightly outperforms transfer learning from ImageNet with VGGnet. At 77 GHz, the imitation signing recognition accuracy (96%) achieved significantly exceeds that of fluent signing (76%) by 20%. While this at first glance may seem surprising, a visualization of the distribution of each sign, illustrated in Figure 9 shows in fact how distinctly group each sign is in 77 GHz imitation signing data. At 24 GHz, the imitation signing recognition accuracy still exceeds that of fluent signing, but with a lesser difference of just 5%. This is primarily because of the greater detail in the signatures of the 77 GHz sensor, which exhibits a greater Doppler shift for a given velocity than the other sensors. As the coarseness of the micro-Doppler signatures increase, the classification accuracy decreases. This indicates that investigation into higher-frequency resolution time-frequency transforms may lead to tangible gains for ASL recognition applications. Moreover, these results reveal that the use of imitation signing to evaluate ASL recognition algorithms can lead to over optimistic results, so that even if the objective were purely for ASL-sensitive user interfaces, as opposed to translation, which encompasses the richness of language, fluent ASL data should always be used for testing.

F. Imitation Signing Data as a Source for Pre-Training

Given the high classification accuracies of imitation signing data, it may be thought, however, that one way of mitigating the burden of acquiring fluent signing data could be by pre-training networks on imitation signing data as opposed to alternative, entirely unrelated sources of data, such as ImageNet. Pre-training with imitation signing data, however, results in significantly poorer network initialization: the bottleneck classification accuracy obtained by pre-training the CAE with the imitation signing samples is just 24%. Imitation signing data misleads the network in its understanding of the kinematic characteristics of each class due to the many signing errors and differences in tempo. A better solution is to instead illuminate the signer with RF sensors transmitting across multiple frequencies, which allows for the extraction of unique features at each frequency. This approach is discussed next.

IV. Multi-Frequency Fluent ASL Recognition

Because there is no overlap in the transmit frequency bands of the three RF sensors compared in this work, all sensors may be simultaneously operated and used to illuminate the participant. Various types of fusion can then be utilized to increase the performance afforded by each sensor individually. In decision fusion, the received return from each sensor is first separately classified and then an overall decision made through majority voting. In feature level fusion, separate networks are used to extract features from each sensor, concatenated, and then supplied a classifier. Cross-modal fusion networks [37], [38] aim to capture both share features in the data, while also separately extracting features specific to each modality. This approach is particularly well-suited for fusion in RF sensor networks because the common observations will result in shared target-specific features, while the phenomenological difference across frequency create sensor-specific differences in the data.

Thus, a multi-frequency fusion network is designed that consists of sensor-specific layers and shared layers. We compare two approaches towards training the network: 1) end-to-end training, and 2) two-step modality tuning. In end-to-end

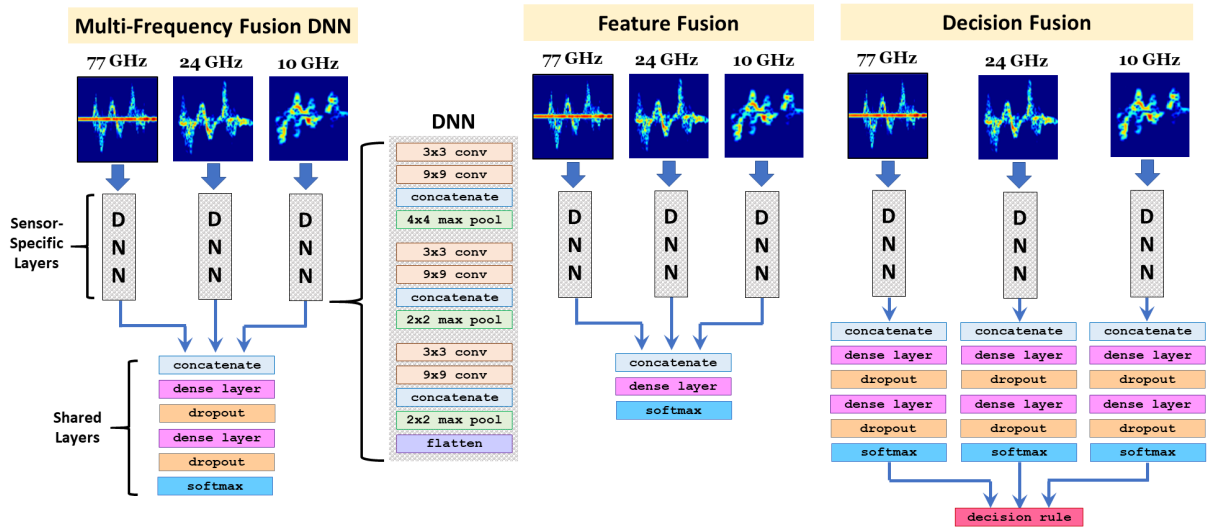


Fig. 10. Block diagrams for proposed and compared fusion approaches.

Fusion Type	Accuracy
Multi-Frequency Fusion DNN	95.53%
w/No Modality Tuning	83.67%
Feature Level Fusion	79.59%
Decision Fusion	75.00%

Fig. 11. Comparison of fusion results.

training, all network weights are optimized in a supervised fashion. In modality tuning, the shared layers are initially frozen while the modality specific layers are trained. By freezing the shared layers of the network, a high-level representation is transferred to the other modalities. Essentially, with this approach, the network is being fine-tuned for a modality as opposed to a task. After training the network for each modality for a fixed number of iterations, the shared layers are unfrozen and the entire network is trained jointly, allowing the incorporation of information from the other modalities without overfitting modality specific representations.

Fig. 11 shows the results for the various fusion methods in comparison to the proposed multi-frequency fusion DNN with and without modality tuning, illustrated in Fig. 10. All results are shown for training and testing with fluent ASL signing data. The highest classification accuracy of 95% is achieved with the multi-frequency fusion DNN trained with modality tuning, which provides a 12% increase in accuracy relative to the same DNN with all layers trained simultaneously, 16% increase relative to feature-level fusion, and 20% increase relative to decision fusion.

V. Conclusions and Future Directions

This work illustrates the potential of RF sensing for recognition of fluent ASL signs at a high (>95%) accuracy. It is significant that these results were obtained using only kinematic information captured by the micro-Doppler signatures of the signs. In future work, we plan to investigate further performance improvements enabled by fusion with spatial information provided by multi-channel radars, namely, slant range and direction-of-arrival, as reflected in the range-Doppler map and range-angle representations of the RF data. Moreover, although the currently possible radar point cloud spatial resolutions are too coarse for hand shape recognition, advancements in commercially available multi-channel radar transceivers could one day make this possible. Indeed, the proposal of RF sensing for silent lip reading and voice recognition [39], [40] suggests another interesting way mouth movements perceived by RF sensors could be exploited for ASL recognition.

While this work has examined the recognition of independently articulated signs, in natural settings, device triggering could be embedded within connected discourse or daily activities resulting in gross body movements, such as walking or picking up an object. Thus, future work should consider not just sensor positioning within a room, but also sequential recognition in continuous, long duration recordings. We are currently working to integrate the RF sensors used in this work with edge computing platforms to develop an indoor test bed for more realistic studies of ASL recognition in smart environments.

A second important conclusion of this work is to underscore the importance of testing algorithms on fluent signing, which has been demonstrated through visualization of the feature space of imitation signing versus fluent signing and comparison of their respective recognition accuracies. The difference in representation of fluent vs. imitation signing by principal components emphasizes the gap in quantitative understanding of sign articulation, and importance of careful calibration of sensors to ensure appropriate spatiotemporal resolution in the data that would allow capture of lingustic features in continuous fluent signing.

We believe that it is essential for research on technologies benefiting the Deaf community to be conducted in partnership with the Deaf community [41], [42]. As essential as the

involvement of Deaf participants and fluent signers is, the development of community partnerships and the conduct of joint research with Deaf researchers are critical to ensure that the developed technology addresses the concerns and problems of the Deaf community as the primary audience/beneficiary. Although the scope of this work is limited to recognition of individual signs, in the future we plan to work with Deaf community partners on the development of non-invasive sign language recognition technologies under more natural settings as a means for opening to door to the design of smart Deaf spaces.

ACKNOWLEDGMENT

The authors would like to thank Dr. Caroline Kobek-Pezzarossi from Gallaudet University, Washington D.C. and Dr. Dennis Gilliam from AIDB for their support of this research. 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. Human studies research was conducted under UA Institutional Review Board (IRB) Protocol #18-06-1271.

REFERENCES

[1] K. Tsymbal, "Deaf space and the visual world - buildings that speak: An elementary school for the deaf," Ph.D. dissertation, 2010.

[2] 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 Journal*, vol. 19, no. 16, pp. 7056–7063, 2019.

[3] 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 Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2306–2320, 2020.

[4] J. Galka, M. Masior, M. Zaborski, and K. Barczewska, "Inertial motion sensing glove for sign language gesture acquisition and recognition," *IEEE Sensors Journal*, vol. 16, no. 16, pp. 6310–6316, 2016.

[5] 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.

[6] K. Qian, Z. He, and X. Zhang, "3d point cloud generation with millimeter-wave radar," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 4, Dec. 2020.

[7] V. Chen, *The Micro-Doppler Effect in Radar, 2nd Ed.* Boston: Artech House, 2019.

[8] 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.

[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] S. Z. Gurbuz, A. C. Gurbuz, E. A. Malaia, D. J. Griffin, C. S. Crawford, M. M. Rahman, E. Kurtoglu, R. Aksu, T. Macks, and R. Mdrafi, "American sign language recognition using rf sensing," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3763–3775, 2021.

[11] 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.

[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. Wang, Z. Yu, X. Lou, B. Guo, and L. Chen, "Gesture-radar: A dual doppler radar based system for robust recognition and quantitative profiling of human gestures," *IEEE Transactions on Human-Machine Systems*, vol. 51, no. 1, pp. 32–43, 2021.

[14] J. D. Borneman, E. Malaia, and R. B. Wilbur, "Motion characterization using optical flow and fractal complexity," *Journal of Electronic Imaging*, vol. 27, no. 5, p. 051229, 2018.

[15] E. Malaia, J. D. Borneman, and R. B. Wilbur, "Assessment of information content in visual signal: analysis of optical flow fractal complexity," *Visual Cognition*, vol. 24, no. 3, pp. 246–251, 2016.

[16] E. Malaia and R. B. Wilbur, "Kinematic signatures of telic and atelic events in asl predicates," *Language and speech*, vol. 55, no. 3, pp. 407–421, 2012.

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

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

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

[20] B. Fang, J. Co, and M. Zhang, "Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation," in *Proc. of the 15th ACM Conf. on Embedded Network Sensor Systems*, 2017.

[21] 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.

[22] P. Joshi, D. Colombi, B. Thors, L. Larsson, and C. Törnevik, "Output power levels of 4g user equipment and implications on realistic rf emf exposure assessments," *IEEE Access*, vol. 5, pp. 4545–4550, 2017.

[23] N. Caselli, Z. Sehyr, A. Cohen-Goldberg, and K. Emmorey, "Asl-lex: A lexical database of american sign language," *Behavior Research Methods*, vol. 49, 05 2016.

[24] P. van Dorp and F. Groen, "Human walking estimation with radar," *IET Radar, Sonar and Navigation*, vol. 150, pp. 356–365(9), October 2003.

[25] M. Richards, *Fundamentals of Radar Signal Processing*. New York: McGraw-Hill Education, 2014.

[26] B. Erol and M. G. Amin, "Radar data cube processing for human activity recognition using multisubspace learning," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 55, no. 6, pp. 3617–3628, 2019.

[27] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, 2008.

[28] J. S. Beal and K. Faniel, "Hearing l2 sign language learners," *Sign Language Studies*, vol. 19, no. 2, pp. 204–224, 2019.

[29] 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.

[30] 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.

[31] M. S. Seyfioglu, A. M. Ozbayoglu, and S. Z. Gurbuz, "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.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[34] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, Dec. 2014.

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[36] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[37] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Computation*, vol. 32, no. 5, pp. 829–864, 2020.

[38] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba, "Cross-modal scene networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 10, pp. 2303–2314, 2018.

[39] Y. H. Shin and J. Seo, "Towards contactless silent speech recognition based on detection of active and visible articulators using ir-uwb radar," *Sensors (Basel, Switzerland)*, vol. 16, 2016.
[40] R. Khanna, D. Oh, and Y. Kim, "Through-wall remote human voice recognition using doppler radar with transfer learning," *IEEE Sensors Journal*, vol. 19, no. 12, pp. 4571–4576, 2019.
[41] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef, C. Vogler, and M. Ringel Morris, "Sign language recognition, generation, and translation: An interdisciplinary perspective," in *The 21st Int. ACM SIGACCESS Conference on Computers and Accessibility*, ser. ASSETS '19, 2019, p. 16–31.
[42] J. Charlton, *Nothing About Us Without Us: Disability Oppression and Empowerment*. University of California at Berkeley, 1998.

Sevgi Z. Gurbuz (S'01–M'10–SM'17) 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 Georgia Institute of Technology, Atlanta, GA, USA, in 2009.

From February 2000 to January 2004, she worked as a Radar Signal Processing Research Engineer with the U.S. Air Force Research Laboratory, Sensors Directorate, Rome, NY, USA. Formerly an Assistant Professor in the Department of Electrical-Electronics Engineering at TOBB University, Ankara, Turkey and Senior Research Scientist with the TUBITAK Space Technologies Research Institute, Ankara, Turkey, she is currently an Assistant Professor in the University of Alabama at Tuscaloosa, Department of Electrical and Computer Engineering and director for the UA Laboratory of Computational Intelligence in Radar (CIR4). Her current research interests include radar signal processing, physics-aware machine learning, human motion recognition for biomedical, vehicular autonomy, and human-computer interaction (HCI) applications, and sensor networks.

Dr. Gurbuz is a recipient of the 2020 SPIE Rising Researcher Award, EU Marie Curie Research Fellowship, and the 2010 IEEE Radar Conference Best Student Paper Award.

M. Mahbubur Rahman received the B.S. degree in Electronics and Communication Engineering from Khulna University of Engineering and Technology (KUET), Bangladesh, in 2016. He is currently a Ph.D. student in Electrical and Computer Engineering at the University of Alabama (UA), Tuscaloosa, AL, USA, and a research assistant in the UA Laboratory of Computational Intelligence for Radar (CI4R). His research interests include radar signal processing, physics-aware machine learning and domain adaptation for human activity recognition using radar.

Emre Kurtoglu received the B.S. degree in electrical and electronics engineering from the Koc University, Istanbul, Turkey, in 2018.

He worked as an intern in Honeywell, Istanbul, Turkey and in Aselsan, Ankara, Turkey, from August 2017 to September 2017 and from June 2018 to July 2018 respectively. He is currently a Ph.D. student in the Department of Electrical and Computer Engineering at the University of Alabama, Tuscaloosa, and research assistant in the UA Laboratory for Computational Intelligence for Radar (CI4R). His current research interests include machine learning, human activity recognition and radar signal processing.

Evie A. Malaia received her Ph.D. degree in Computational Linguistics from Purdue University, West Lafayette, in 2004.

Formerly a Research Scientist at Indiana University and Purdue University, and an Assistant Professor at the University of Texas at Arlington, she is currently an Associate Professor at the University of Alabama at Tuscaloosa, Department of Communicative Disorders. Her current 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 is a recipient of the Ralph E. Powe Award from DOE/ORAU, EurIAS Research Fellowship, EU Marie Curie Senior Research Fellowship, and the APS Award for Teaching and Public Understanding of Psychological Science.

Ali Cafer Gurbuz (M'08–SM'18) received B.S. degree from Bilkent University, Ankara, Turkey, in 2003, in Electrical Engineering, and the M.S. and Ph.D. degrees from Georgia Institute of Technology, Atlanta, GA, USA, in 2005 and 2008, both in Electrical and Computer Engineering. From 2003 to 2009, he researched compressive sensing based computational imaging problems at Georgia Tech. He held faculty positions at TOBB University and University of Alabama between 2009 and 2017 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. Currently, he is an Assistant Professor at Mississippi State University, Department of Electrical and Computer Engineering, where he is co-director of Information Processing and Sensing (IMPRESS) Lab. He is the recipient of The Best Paper Award for Signal Processing Journal in 2013, the Turkish Academy of Sciences Best Young Scholar Award in Electrical Engineering in 2014 and National Science Foundation CAREER Award in 2021.

Darrin J. 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 in 2004 and 2007, respectively, from The University of Texas at Austin. The Ph.D. degree was completed at The University at Buffalo, SUNY in 2010 in communication with a focus on deceptive communication.

From August 2010 to current he has served as a faculty member at The University of Alabama, Department of Communication Studies where he currently teaches and conducts research as an associate professor on topics related to nonverbal communication, deceptive communication, and deafness. Dr. Griffin is fluent in American Sign Language and participates in various forms of community engagement with the Deaf community.

Dr. Griffin is 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 a 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 S. Crawford received the Ph.D. degree in human-centered computing from the University of Florida, Gainesville, FL, USA. He is currently an Assistant Professor at the University of Alabama's Department of Computer Science. He directs the Human-Technology Interaction Lab (HTIL). He has investigated multiple systems that provide computer applications and robots with information about a user's cognitive state. In 2016, he lead the development of a BCI application that was featured in the world's first multiparty brain-drone racing event. His current research focuses on computer science education, human-robot interaction, and brain-computer interfaces.