Capturing Motion: Using Radar to Build Better Sign Language Corpora

Evie Malaia, Joshua Borneman, Sevgi Gurbuz
University of Alabama, Purdue University, University of Alabama
Tuscaloosa, AL; West Lafayette, IN; Tuscaloosa, AL
eamalaia@ua.edu, jdbornem@purdue.edu, szgurbuz@eng.ua.edu

Abstract

Sign language conveys information using dynamic visual signal. Proficient signers rely on the skill in processing and predictive motion information during sign language comprehension. Much current work in sign language corpora development relies on video data. However, from the perspective of information transfer in communication, video recordings are limited in capturing spatial and temporal frequencies of sign language signal in sufficient resolution. In contrast, radar can capture 3D motion data at high temporal and spatial resolution, preserving depth articulations lost in 2D video. Radar's recording parameters can also be adapted in real time to optimize temporal resolution for rapid signing motions. Thus, radar recordings provide higher-fidelity corpora for analyzing linguistic features of sign languages and creating smart environments that respond to signed input. Crucially, radar recordings uphold user privacy, only capturing kinematic parameters of communicative signal, as opposed to signer identity. Radar resolution in capturing dynamic data from sign language production, and privacy advantages it provides to users, make it uniquely suited for advancing sign language research through corpora development.

Keywords: sign language, production, radar

1. Signed Communication

Sign languages convey linguistic information dynamically through articulator motion. Although linguistic analyses of signs only identifies motion as a component of sign phonology, on par with handshape, hand orientation, and place of articulation, research in visual perception and sign comprehension has long been clear on relevance of dynamic motion, as opposed to static components of articulation, to proficient signers (Malaia et al., 2023). Lifelong exposure to visual complexity inherent in sign motion affects both perceptual and cognitive processing in sign language users compared to non-signers, and enhances signers' perceptual tuning to the information density in motion signals, allowing them to parse continuous signal, identifying discrete signs and their grammatical modifications (Klima et al., 1999; Bavelier et al., 2006). Linguistic distinctions in meaning and grammar are reflected in the movement dynamics of the signed signal. These distinctions can be captured in a manner parallel to acoustic and phonetic analysis of spoken signals (Borneman et al., 2018).

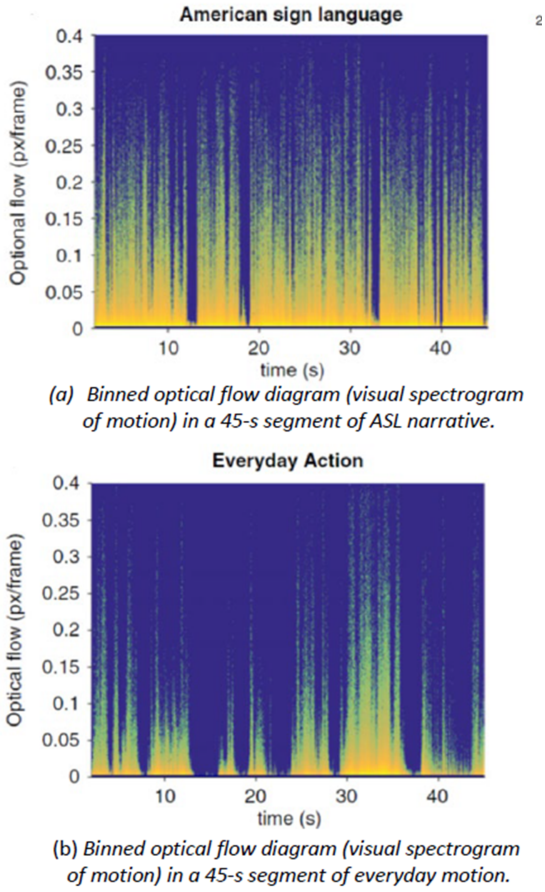
Fully visible articulator motion in sign language carries all communicated information. At the same time, sign language motion carries more information defined as visual signal entropy than everyday human motion (cf. Fig. 1). The parameters that are critical to capturing information-dense features of the continuous signal are the temporal resolution and the amount of change present in the signal within the given time window. When signs are produced fluently in sentences, there are almost always transitional movements between them, for example, when one sign ends with the hand(s) located in one place and the next sign starts with them located somewhere else, there must be a movement of the hands to that next location before the next sign can start its lexical movement. These transitions are clearly differentiable to signers, and ignored when they are asked to count/tap to syllables (Klima et al., 1999).

The variability of motion between sign-syllables and transitions forms the basis of the quantitative distinction between non- informative, biological motion, and the sign language signal (Malaia et al., 2018). Mathematically quantified amount of information (i.e. variability) in the motion of the articulators in sign language forms the basis of sign syllables (Malaia and Wilbur, 2020). Experimental approaches, including video analysis using optical flow and motion capture data analysis, indicate that information transfer in sign language critically relies on the entropy of the articulator signal, making it critical to capture dynamic changes in it with sufficient spatial and temporal resolution.

2. Information Transfer in Sign Language Signal

When evaluating and comparing modalities for capturing sign language motion, and for analyzing languages in general, a key factor is the fidelity and dimensionality with which each modality can capture the information content of the original dynamic signal over time (Malaia et al., 2022). It is first useful

Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages, pages 213–218



(a) Binned optical flow diagram (visual spectrogram of motion) in a 45-s segment of ASL narrative.

(b) Binned optical flow diagram (visual spectrogram of motion) in a 45-s segment of everyday motion.

Figure 1: ASL and action: comparative variability optical flow spectrograms (a - American Sign Language; b - everyday motion).

to explain the common framework on which different language types, and capture methods, may be compared.

Starting with a simple example, a spoken language is a 1-dimensional time series signal, carrying information in amplitude as a function of temporal frequency $[f_t]$, written here as $[S_0(f_t)]$. Recording of this spoken signal, usually limited in amplitude and frequency by electronics/sampling method, may then be treated as a series of transfer functions. For instance we may have recording/electronics/sampling effects, $[T_r(f_t)]$, and effects on the data due to preprocessing $[T_p(f_t)]$. Importantly, $T(f_t) < 1$, i.e. no recording or capture method is perfect. This means that the final recorded language signal is not a pure recorded sample of the original spoken language, but is rather a modified signal $[S_1(f_t)]$, where $S_1(f_t) = S_0(f_t) \cdot T_r(f_t) \cdot T_p(f_t)$. Therefore, final analysis of the spoken language is always done on a reduced fidelity recording. Knowing this, a spoken language recording method should be selected which preserves the overall information density within the temporal component, $[f_t]$. This would require high temporal sampling and analysis frequencies, and most acoustic recordings may contain a minimum of 20k samples per second. This characterization may seem trivial for a 1-dimensional spoken language, but the framework becomes useful when dealing with a multi-dimensional signal, such as sign language. Compared to linear sound recordings, capture of sign language presents a significant difficulty. Sign language conveys information over spatial frequencies in 3 space dimensions (f_x, f_y, f_z), as well as in temporal frequency (f_t), and therefore any analysis of sign languages will depend on the accuracy and dimensionality with which the original signal can be recorded, as well as potential dimensionality reduction and fidelity loss during further analysis. Each recording and processing step acts as a filtering function, potentially reducing the fidelity of the data. Therefore, it is important to select measurement and analysis methods which preserve, or at least are intentional about, how dimensionality and fidelity are addressed. Sign language, as a 3-dimensional spatial signal also varying in time, $S_0(f_x, f_y, f_z, f_t)$, is filtered in both spatial frequencies, f_x, f_y, f_z, and temporal frequencies, f_t, depending on how it is recorded $[T_r]$ and how it is preprocessed $[T_p]$. $S_1(f_x, f_y, f_z, f_t) = S_0(f_x, f_y, f_z, f_t) \cdot T_r(f_x, f_y, f_z, f_t) \cdot T_p(f_x, f_y, f_z, f_t)$ Although sign languages use relatively lower temporal frequencies as compared to spoken language, sign language also transfers information in additional spatial dimensions. These spatial dimensions must also be recorded in order to preserve the overall information density. This description may now be used to describe various methods of language capture in a common framework. For example, video capture recordings of sign language are, in essence, a 2D spatial frequency filter, which removes depth information $[T_r(f_z) = 0]$, and in which the x, y spatial plane is downsampled to s, t by the camera distance and resolution $[(f_u, f_v) \approx (f_x, f_y)]$, and filtered such that $[T_r(f_u, f_v) < 1]$. The camera resolution and position should ideally be placed such that all hand/arm articulators in the signing space are resolved, that is, that the articulator frequencies are in the camera band-pass. Further, f_t is subsampled by the frame rate of the video recording $[f_T \approx f_t)]$, resulting in $T_r(f_T) < 1$. Therefore, our pure real-world sign language information signal $[S_0(f_x, f_y, f_z, f_t)]$ is now recorded by video and subsampled to only two spatial dimensions and time $[S_{1,video}(f_s, f_t, f_T)]$.

In contrast, radar is capable of capturing 3D motion data over time, with adaptive temporal resolution based on user-configurable recording parameters. Radar signal processing algorithms may extract range-Doppler (RD) maps (2D images of range versus Doppler frequency) or micro-Doppler

signature (Doppler frequency versus time). Therefore, radar records motion along the depth axis z, subsampled to w resolution$[f_w \approx f_z]$, such that $[T_r(f_w) < 1]$ through the line of sight distance. The remaining spatial dimensions f_x, f_y are convoluted into a radial velocity and angle of arrival such that $[(f_r, f_a) \propto (f_x, f_y)]$ and therefore $[T_r(f_r, f_a) < 1]$. Temporal resolution is adjustable based on the pulse repetition frequency (PRF), and can be set to match sign language motion bandwidths, and is generally faster than video frame rates, $[T_r(f_T) < 1]$. For Frequency Modulated Continuous Wave (FMCW) radar, the PRF also determines the maximum measurable radial velocity ($v_{max} = PRF \times \lambda/2$) and the velocity resolution $\Delta v = PRF/N$, where N is the total number of pulses transmitted. With higher transmit frequencies, the Doppler frequency shift incurred due to even slower motions is greater and hence more easily measurable; however, this also reduces the maximum velocity limit. Thus, selecting a high PRF is advantageous both from the perspective of ensuring unaliased velocity measurements and high temporal sampling of motion during signing. In prior work comparing the resulting radar micro-Doppler signatures of lower bandwidth (β), lower PRF signal with low duty cycle (d) (β = 750 MHz, PRF = 3.2 kHz, d = 51.2%) versus one of high bandwidth, PRF and duty cycle (β = 4 GHz, PRF = 6.4 kHz, d = 96%), we found that the latter enabled crisp and pristine micro-Doppler signatures of sign language (Gurbuz et al., 2022a). Spatial depth resolution depends on the transmitted waveform's bandwidth as $\Delta r = c\beta/2$, where c is the speed of light. Ideally, an FMCW radar with high bandwidth and high PRF is best suited for sign language measurements, as this enables both high spatial and temporal resolution measurements. Automotive radars are well-suited for this aim as they typically have bandwidths of 4 GHz and are designed with PRFs so high as to measure vehicular speed. As the commercial applications of low-cost, low-power radar sensors are ever expanding, it is now possible to find sensors having bandwidths of 5 or even 7 GHz. The main disadvantage of operating the sensor at such high bandwidth and PRF is the high volume of data that results from high spatiotemporal sampling. However, such considerations can be mitigated by interactively adapting the transmission parameters of the waveform so that a low spatiotemporal resolution waveform is transmitted when no human presence is detected, or if a person is simply engaging in daily activities, while a high spatiotemporal resolution waveform is transmitted once a device is triggered and sign language recognition is needed (Kurtoglu, 2024).

Therefore, our pure real-world sign language information signal $[S_0(f_x, f_y, f_z, f_t)]$ is sub-sampled with radar to two convoluted spatial dimensions, one pure spatial dimension, and time $[S_{1,radar}(f_v, f_a, f_z, f_t)]$. This dimensional analysis is useful to evaluate not just hardware capture, but signal processing (Malaia et al., 2022). However it is seen here that compared to 2D video, radar provides crucial depth information about sign articulations in 3D space. Radar's recording parameters can also be selected to maximize temporal resolution appropriate for capturing the rapid motions of signing - a PRF of 6.4 kHz, as utilized in our earlier example, offers much higher temporal sampling than that of a high-speed webcam, which can have a frame rate of about 200-300 frames per second (fps). Thus, the micro-Doppler signature offers a novel representation of sign language corpora that can capture sign language kinematics in a unique fashion, while also doing so in an ambient fashion without recording private imagery. Consider, for example, the micro-Doppler signature for the sign KNIFE, shown in Figure 2. Not only can the maximum and minimum velocity in both directions be measured, but also the timing of the repetitive motion and the number of times the fingers moved back and forth. Notice also that from the radar image we cannot infer any information about the location or environment the recording was made or even who was signing.

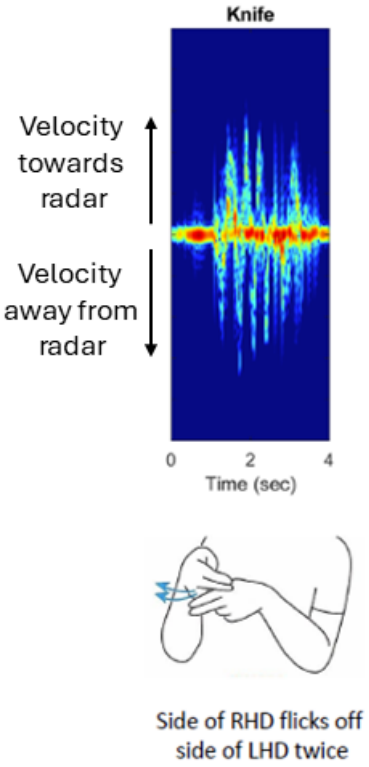


Figure 2: Sample radar micro-Doppler signature for the sign KNIFE.

In addition to manual articulations, sign languages also involve facial expressions, mouth shapes, contact between the fingers and body, as well eye movements, which hold linguistic significance. These are areas of ongoing, active research in radar technology, which may one day make radar-based sign language studies beyond manual articulations possible. For example, lip reading under face masks using radar has been proposed (Hameed et al., 2022) to enable speech recognition when camera-based techniques are not possible due to the obstruction by the mask. Emotion recognition (Dang et al., 2022) has also become a topic of interest, as such facial movements during expressions is coupled with vital signs recorded by the radar. Moreover, through the use of a high number antenna array elements in both the azimuth and elevation, newly developed high-resolution imaging radars (Bräunig et al., 2023) have been developed that can provide a distinct image of hand shape, which can thus enable recognition of fingerspelling. The principle downside of this current technology, however, is that such imaging radars are not able to dynamically acquire images and require the hand to be stationary. However, as automotive radars are commercialized with an increasing number of array elements, so is the azimuth and elevation angular resolution increasing so that potentially new AI/ML algorithms operating directly on the raw radar data can be developed to enable such functions that require high spatial resolution and localization (such as detection of finger-body contact).

For these reasons, radar provides a uniquely informative way for capturing sign language corpora, which we have only yet begun to explore. Radar's higher-fidelity 3D motion data over time offers potential for more detailed analysis of linguistic and kinematic features of sign languages. This advantage highlights radar's promise for advancing sign language research through improved corpora.

3. Radar-based Sign Corpora and Machine Learning

Unlike video, radar measurements are not inherently an image, but are actually a time-stream of complex I/Q data from which line-of-sight distance, radial velocity, and angle of arrival may be computed. The radar measurements may be visualized via a variety of 2D and 3D data tensors. The most widely used representation is the *micro-Doppler signature* (Chen, 2019), which is computed using the short-time Fourier transform and reveals the micro-Doppler frequencies - or radial velocity - due to rotational motion centered about the Doppler shift due to translational motion. Thus, the micro-Doppler signature is a rich source of kinematic information relating to signing. In our prior work (Rahman et al., 2022), we have shown that using micro-Doppler signatures alone, snapshots of over 100 word-level signs can be classified at over 77% top-1 and 92% top-5 accuracy. Moreover, we found that RF micro-Doppler frequencies also captured significant linguistic properties of the signer, such as co-articulation (Gurbuz et al., 2020), whether the signer was fluent in ASL versus being a hearing imitation signer (Gurbuz et al., 2021, 2022b), and whether or not the signer was being directed to articulate a sign versus doing a natural articulation as part of freely playing a game (Kurtoglu et al., 2024).

Thus, the linguistic characteristics of a signer have a significant impact on model training: deep neural networks (DNNs) for recognition of natural, fluent signing cannot be effectively trained using imitation signing or signing acquired via controlled experiments. Integration of kinematic constraints into the DNN architecture itself is also greatly beneficial. For example, the envelopes of the micro-Doppler signature measure the peak radial velocity incurred during signing and can be provided as a dual input to the discriminator of a Generative Adversarial Network (GAN) and used to compute a physics-based loss function, which combine enable to GAN to synthesize kinematically more accurate data for model training (Rahman et al., 2022, 2023). The utilization of multi-task learning where loss functions for each task are defined based on kinematic properties is also beneficial for recognition performance (Kurtoğlu et al., 2022). For example, we showed that a trigger sign (or wake word) could be more effectively recognized if the DNN optimized for five distinct tasks: 1) one versus two handedness, 2) major location of hands, 3) movement type, 4) daily activity versus ASL, and 5) number of strokes comprising the sign. Linguistic metrics, such as fractal complexity, were also found to be indicative of whether a person was signing versus doing an everyday activity at home (Gurbuz et al., 2020).

In addition to the micro-Doppler signature, 3D RF data tensors can be used to provide an enriched input to DNNs and achieve greater accuracy when trying to recognize sign language in a real-world environment, such as would occur if a user were to use sign language to interact with an electronic personal assistant, such as Alexa or Siri. Using radar signal processing, the raw radar data stream can be converted into a time series images of range-velocity and range-angle. Joint utilization of multiple radar data representations has been used to design a Joint Domain Multi-Input Multi-Task Learning (JD-MIMTL) network (Kurtoğlu et al., 2022) that can automatically segment and extract signing sequences from continuous recordings of daily life, detect whether a trigger sign has been articulated, and recognize subsequent signs as device com-

mands. In fact, estimation of the angle at which a person is located relative to the location of the radar can be used to generate an angular projection of the RF data tensor for the left and right hands (Kurtoğlu et al., 2023). A multi-view DNN was designed to leverage the separate projections of the left and right hand for increased sign recognition performance.

A major challenge to deep learning based ASL recognition with both video and radar remains the limited availability of data that truly captures the nuanced variations of natural signing. To overcome this challenge, an interactive game (Kurtoglu et al., 2024), ChessSIGN, was developed that acquires both video and radar data as a user articulates ASL to move the pieces of the chess game. When the user clicks on a piece, different ASL words corresponding to valid chess moves appear on the screen. The piece moves its position based on recognition of the user's articulation of the sign. We have shown that for both video and radar data, machine learning models trained under data collected via controlled experiments is not effective in recognizing signing in such an unconstrained, natural setting. However, as the system acquires more and more natural signing data during the course of the game, recognition accuracy increases. Moreover, the signs recorded are natural language recordings, which more accurately reflect 1) variations in ASL due to person-specific traits, regional dialects, and fluency; and 2) natural effects such as coarticulation, which occur due to the variation in the position with which a sign can begin or end, as typical of daily life. ChessSIGN thus provides an entertaining way to minimize the burden on the Deaf community to acquire ASL data, while also continually building improved models. Also, because the system captures simultaneous recordings of video with radar, this unique dataset can enable the exploration of new ASL recognition algorithms that jointly exploit the strengths of radar and video together.

Ultimately, our work has shown that RF sensing can capture the kinematics of the rapid progression of dynamic sign sequences that is characteristic of ASL usage. We not only bring to bear, for the first time, a linguistic perspective to RF-based motion recognition, but also a physics-based machine learning approach achieved through the integration of kinematics with deep learning. These advances have enabled the development of RF-sensing based ASL-sensitive human computer interaction (HCI) and as a tool for linguistic analysis of ASL.

4. Ethical Consideration for Sign Language Corpora

Collecting sign language data with radar sensors also offers important privacy advantages over video recording. Video cameras capture detailed visual information about a person's appearance, clothing, surroundings, and any visible actions. This raises significant personal privacy concerns, especially when recording in homes or private spaces. In contrast, radar does not actually record images or videos. Radar sensors operate by transmitting electromagnetic waves and analyzing the reflected signals. The sensors only measure the time-varying position and velocity of body parts as they move through space. No identifying visual features are recorded. The raw radar data itself reveals nothing about a person's identity, gender, attire, or environment. While video provides full visual details, this level of information is unnecessary for analyzing sign language gestures. The intricate motions of signing are characterized by the changing spatial relationships and dynamics of the hands, arms, and face over time. Radar captures exactly these articulatory parameters relevant to sign language, without any personal identifying visuals. Participants are also more comfortable being recorded by radar since their privacy is protected. No video footage exists that could be leaked or exploited. Radar enables collecting natural, unrestrained sign language data even in private real-world environments. Radar recordings capture information-bearing motion from sign language signal with fidelity sufficient for both linguistic or ML-based analysis, while upholding signers' privacy. The ability to gather realistic sign language data in a completely private manner makes radar systems uniquely suited for building sign language corpora and recognition datasets in an ethical, non-invasive way.

5. Bibliographical References

Daphne Bavelier, Matthew WG Dye, and Peter C Hauser. 2006. Do deaf individuals see better? *Trends in cognitive sciences*, 10(11):512–518.

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

Johanna Bräunig, Vanessa Wirth, Christoph Kammel, Christian Schüßler, Ingrid Ullmann, Marc Stamminger, and Martin Vossiek. 2023. An ultra-efficient approach for high-resolution mimo radar imaging of human hand poses. *IEEE Transactions on Radar Systems*, 1:468–480.

Victor C Chen. 2019. *The micro-Doppler effect in radar*. Artech house.

Xiaochao Dang, Zetong Chen, and Zhanjun Hao. 2022. Emotion recognition method using millimetre wave radar based on deep learning. *IET Radar, Sonar & Navigation*, 16(11):1796–1808.

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

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

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

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

Hira Hameed, Muhammad Usman, Ahsen Tahir, Amir Hussain, Hasan Abbas, Tie Jun Cui, Muhammad Ali Imran, and Qammer H. Abbasi. 2022. Pushing the limits of remote RF sensing by reading lips under the face mask. *Nature Communications*, 13(1):1–9.

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

E. Kurtoglu. 2024. *Fully-Adaptive RF Sensing for Non-INtrustive ASL Recognition via Interactive Smart Environments*. Ph.D. thesis, Department of Electrical and Computer Engineering, The University of Alabama, Tuscaloosa, AL.

Emre Kurtoglu, Kenneth DeHaan, Caroline Kobek Pezzarossi, Darrin J. Griffin, Chris Crawford, and Sevgi Z Gurbuz. 2024. Interactive learning of natural sign language with radar. *IET Radar Sonar and Navigation*.

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

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

Evie A Malaia, Joshua D Borneman, Emre Kurtoglu, Sevgi Z Gurbuz, Darrin Griffin, Chris Crawford, and Ali C Gurbuz. 2022. Complexity in sign languages. *Linguistics Vanguard*, 9(s1):121–131.

Evie A Malaia, Joshua D Borneman, and Ronnie B Wilbur. 2018. Information transfer capacity of articulators in american sign language. *Language and Speech*, 61(1):97–112.

Evie A Malaia, Sean C Borneman, Joshua D Borneman, Julia Krebs, and Ronnie B Wilbur. 2023. Prediction underlying comprehension of human motion: an analysis of deaf signer and non-signer eeg in response to visual stimuli. *Frontiers in Neuroscience*, 17:1218510.

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

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

Mohammed Mahbubur Rahman, Sevgi Z Gurbuz, and Moeness G Amin. 2023. Physics-aware generative adversarial networks for radar-based human activity recognition. *IEEE Transactions on Aerospace and Electronic Systems*, 59(3):2994–3008.