SonicASL: An Acoustic-based Sign Language Gesture Recognizer Using Earphones

YINCHENG JIN*, University at Buffalo, State University of New York, USA
YANG GAO*, University at Buffalo, State University of New York, USA
YANJUN ZHU, University at Buffalo, State University of New York, USA
WEI WANG, University at Buffalo, State University of New York, USA
JIYANG LI, University at Buffalo, State University of New York, USA
SEOKMIN CHOI, University at Buffalo, State University of New York, USA
ZHANGYU LI, University at Buffalo, State University of New York, USA
JAGMOHAN CHAUHAN, University of Southampton, UK
ANIND K. DEY, University of Washington, USA
ZHANPENG JIN†, University at Buffalo, State University of New York, USA

We propose SonicASL, a real-time gesture recognition system that can recognize sign language gestures on the fly, leveraging front-facing microphones and speakers added to commodity earphones worn by someone facing the person making the gestures. In a user study (N=8), we evaluate the recognition performance of various sign language gestures at both the word and sentence levels. Given 42 frequently used individual words and 30 meaningful sentences, SonicASL can achieve an accuracy of 93.8% and 90.6% for word-level and sentence-level recognition, respectively. The proposed system is tested in two real-world scenarios: indoor (apartment, office, and corridor) and outdoor (sidewalk) environments with pedestrians walking nearby. The results show that our system can provide users with an effective gesture recognition tool with high reliability against environmental factors such as ambient noises and nearby pedestrians.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; *Ubiquitous and mobile computing systems and tools.*

Additional Key Words and Phrases: Acoustic sensing, sign language gesture recognition, earphones

*The first two authors contributed equally.
†This is the corresponding author.

Authors' addresses: Yincheng Jin, University at Buffalo, State University of New York, Department of Computer Science and Engineering, Buffalo, NY, 14260, USA; Yang Gao, University at Buffalo, State University of New York, Department of Computer Science and Engineering, USA; Yanjun Zhu, University at Buffalo, State University of New York, Department of Computer Science and Engineering, USA; Wei Wang, University at Buffalo, State University of New York, Department of Computer Science and Engineering, USA; Jiyang Li, University at Buffalo, State University of New York, Department of Computer Science and Engineering, USA; Seokmin Choi, University at Buffalo, State University of New York, Department of Computer Science and Engineering, USA; Zhangyu Li, University at Buffalo, State University of New York, Department of Electrical Engineering, USA; Jagmohan Chauhan, University of Southampton, Department of Electronics and Computer Science, Southampton, SO17 1BJ, UK; Anind K. Dey, University of Washington, Information School, Seattle, Washington, 98195, USA; Zhanpeng Jin, University at Buffalo, State University of New York, Department of Computer Science and Engineering, USA, zjin@buffalo.edu.

2474-9567/2021/6-ART67 $15.00
https://doi.org/10.1145/3463519

ACM Reference Format:
Yincheng Jin, Yang Gao, Yanjun Zhu, Wei Wang, Jiyang Li, Seokmin Choi, Zhangyu Li, Jagmohan Chauhan, Anind K. Dey, and Zhanpeng Jin. 2021. SonicASL: An Acoustic-based Sign Language Gesture Recognizer Using Earphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 67 (June 2021), 30 pages. https://doi.org/10.1145/3463519

1 INTRODUCTION

Approximately 38.2 million people (14.3%) in the United States suffer from some forms of hearing loss, and this rate becomes one in five people worldwide [13]. Sign languages have been widely used by people who are deaf or are hard of hearing and people with communication disorders as their daily communication form. In particular, American Sign Language (ASL) is arguably one of the most popular and predominant sign languages globally, not only in North America but also in many countries around the world. Many Deaf people live successful and productive lives using sign languages. However, it is still quite challenging for the Deaf group to communicate with the hearing population conveniently. ASL and Deaf customs are not widely understood within the hearing community [22]. Common traditional options for alternative communications include written notes and interpreters. Firstly, the ambiguity and slowness of handwriting are the two major barriers to make exchanging notes not very "interactive" in a lively sense [22]. Secondly, interpreter-based communications are commonly used in the Deaf community, including person-to-person services and free-of-charge mobile apps [47]. However, these type of solutions are not always available and accessible, especially for some unforeseen scenarios (e.g., traveling in foreign countries or in an emergency such as the car accident where a doctor needs to get all the information of the Deaf patient; however, a personal interpreter has longer turnaround time.) and sensitive topics (e.g., face-to-face lawsuit consultation to the lawyer). In recent years, it has been brought to the attention of the entire society that it is necessary and crucial to break the communication barriers between Deaf sign language users and hearing people and seek pervasive, easy-to-use approaches to allow the general population to understand sign languages easily.

Past research in human-computer interaction and ubiquitous computing communities has explored different ways to facilitate communications between sign language users and non-signers. While there has been much recent work on captioning for the Deaf and hard of hearing [6, 27, 42] and producing ASL avatars [67], we focus on a proof-of-concept method of using acoustics to recognize gestures such as those used in ASL. We feed the sentence training dataset into SonicASL, and recognize all words with a large vocabulary and unique grammatical structures that enable both speed and subtlety of expression. As in the early years of speech recognition efforts which addressed small sections of the language in limited domains, ASL recognition efforts [7, 54] attempt to find meaningful but limited applications that can be addressed with current technology. While SonicASL has a small vocabulary, it can recognize phrases. DARPA's early One-Way Phraselator speech recognition effort [24] had similar technical limitations but articulated appropriate applications.

One can imagine a similar scenario where a signer wears SonicASL to interact with a host at a restaurant or a person at a railway ticket counter. They sign a phrase in ASL, and the system recognizes it. The system displays the closest matching phrase in English on the signer's smartwatch, and the signer acknowledges it as being correct. The smartwatch then speaks the phrase for the hearing person [38]. The phrase that is spoken is designed to elicit responses in the form of nodding the head Yes or No, holding up fingers to indicate amount, or pointing. For example, signing "Is there a train to New York" would correspond to a spoken phrase of "Please nod your head up and down if there is a train to New York." The conversation might continue: "Using your fingers, show how many hours until the next train" and "Please point to the track where the train will arrive." Different sign vocabulary can be loaded for recognition for different scenarios. In some senses, the system acts as a foreign language phrasebook. The user finds the right section of the book for "Train travel," "Restaurants" or "Hotel" and then finds a list of the most commonly used phrases for those situations. Like most people working in their non-native language, the Deaf often find it easier to recognize an appropriate English phrase than generating

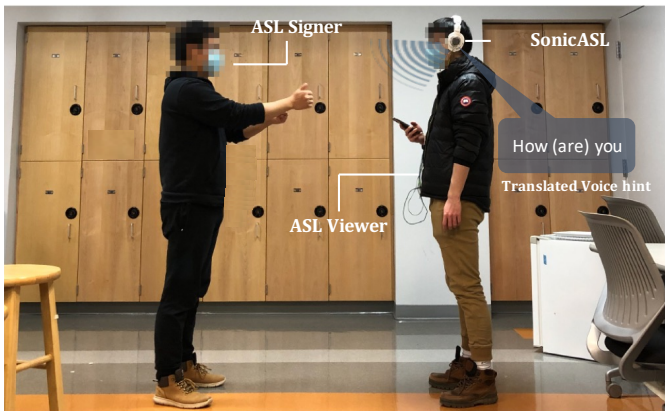


Fig. 1. SonicASL: an acoustic-based sign language gesture recognition that leverages dual speakers and microphones to capture the sonic feedback for hand gestures recognition of the ASL signer.

it themselves. Having a recognition system that allows signing a phrase to discover an appropriate English equivalent instead of attempting to generate it is preferred by the Deaf [53]. In this manner, a small vocabulary phrase-level recognizer may prove useful to the Deaf community.

For the purpose of experimentation and impact of our interactive demonstration, we choose to have the hearing partner in the conversation wear SonicASL. This perspective, that of SonicASL being mounted on the viewer of the sign as opposed to the signer themselves, may be more practical in other scenarios. For example, one can imagine SonicASL being mounted in an information kiosk at a hospital to allow a signer to interact in ASL.

Our work takes advantage of the recently rapid-growing headphones and earphones as the on-body sensor platform for sign language gesture recognition. Over the past few years, as earphones have become increasingly popular among all age groups and a new *de facto* common norm in our daily lives, they hold the potential to become a new integrated sensory platform for providing personal-scale, heterogeneous sensing capabilities. Many earphones now use a pair of outward-facing microphones for Active Noise Cancellation (ANC) [32]. Specifically, an outward-facing microphone picks up the environmental noise, and the circuitry rapidly produces an anti-sound (a soundwave with the same amplitude but with an inverted phase of the surrounding noise) to cancel the external noise. So, we ask a question: "if we add an additional speaker next to the outward-facing microphone, can the earphone sense moving objects similar to a radar, with a rather high accuracy to capture the hand gestures in sign language performed by another person?"

Intrigued by this question, we propose the SonicASL, a new acoustic-based wearable system that leverages acoustic sensing to recognize sign language gestures on the fly, as shown in Figure 1. Augmenting commercial earphones with SonicASL will provide earphone users with on-the-fly *recognition* of an ASL signer's gesture.

We conducted a pilot study with eight non-Deaf subjects using American Sign Language (ASL) gestures, consisting of 42 individual words and 30 sentences. The result shows that SonicASL can achieve 93.8%, 90.6% recognition accuracy for word-level and sentence-level sign language gestures, respectively. Finally, we evaluated our system in the wild with two subjects, including indoor and outdoor environments with ambient noises and nearby pedestrians.

Specifically, we make the following contributions in this work:

- We develop the SonicASL, an accessible, ubiquitous, easy-to-use sign language gesture recognition system on commodity earphones. The proposed approach does not rely on wearable sensors and specialized gadgets, various types of cameras, or a pre-defined environment.

- We address the challenge of distinguishing sign language gestures from subtle variations by introducing the dynamic-threshold-based spectrum enhancement of reflected echoes and propose a CRNN-based framework to recognize sentence-level sign language gestures.
- We conduct rigorous performance evaluations of the proposed approach in diverse real-world scenarios, including surrounding environments, talking distances, aiming angles, and head motions. Our results show that SonicASL possesses strong robustness and reliability.

2 RELATED WORK

We describe research works done in the areas of sign language gesture recognition and acoustic-based human activity recognition. To provide a better background of earphone-based sensing, we also introduce some recent research on earphone interaction and sensing applications.

2.1 Sign Language Recognition

Existing sign language recognition solutions could be generally categorized into three major mechanisms: vision-based, remote sensing-based, and wearable sensor-based. Vision-based approaches usually rely on the videos of sign language hand gestures recorded from a camera (e.g., RGB color cameras [8, 15, 66], Kinect depth cameras, and time-of-flight (ToF) cameras). Early research used hand-crafted features (e.g., the edge orientation histogram [43], upper body joints [21], skeleton information [71], hand shape features [58, 68, 70]) from images to build the ASL recognition systems. Zafrulla *et al.* [69] developed an ASL recognition system using multimodal Kinect system. Camgoz *et al.* [10] proposed an end-to-end deep learning approach with SubNet and Connectionist Temporal Classification (CTC) to recognize continuous sign language gestures from video frames. Later on, Camgoz *et al.* also proposed a new deep learning method "Sign language transformers" [11] by utilizing the encoder-to-decoder to bind the recognition and translation problems into a single unified architecture. Similarly, Pu *et al.* [46] developed a weak-supervised sign language recognition system based on an alignment network that consists of a 3D convolutional residual network and an encoder-decoder network with iterative optimization.

Remote sensing techniques (e.g., WiFi, mmWave) provide less intrusive and privacy-preserving approaches for sign language recognition tasks than vision-based solutions. Ma *et al.* [36] designed SignFi, a sign language recognizer by extracting Channel State Information (CSI) measured by WiFi packets to recognize over 200 sign language gestures. mmASL [50] is an environment-independent ASL gesture recognition system using 60 GHz millimeter-wave radio platform. Nevertheless, those wireless sensing approaches need the supports of available network infrastructures (such as the router and the radio platform) or specialized radar sensors, which is not suitable for mobile application scenarios. Different from aforementioned non-contact solutions, wearable sensor-based systems require ASL signers to wear sensory devices (e.g., data gloves [18, 49, 72], wristbands [72, 73, 75, 76], and smartwatches [25]). Glove-based approaches utilize inertial measurement unit (IMU) sensors to capture hand sign motions[17]. However, the form factor of gloves brings extra burden and inconvenience to the user [31], limiting the usability of those glove-shape gadgets in many application scenarios. Wristbands and smartwatches are increasingly popular these days and are used for monitoring a wide variety of physical activities and gestures. Nevertheless, given the precision limitations of those built-in inertial measurement units (IMUs), wearable sensors can usually recognize only standalone, custom-designed, or coarse-grained hand gestures, which is insufficient and inapplicable for fine-grained, continuous sign language gestures involving both hands.

Unlike existing solutions mentioned above that either require additional hardware or demand special environmental setup, not to mention the potential concerns about privacy and usability in daily-life use, SonicASL recognizes sign gestures using earphones, which can provide a ubiquitous, accessible, and privacy-preserving interface.

2.2 Acoustic-Based Sensing Applications

Acoustic sensing has been increasingly adopted for a wide range of tasks, from gesture recognition to respiratory monitoring, from lip-reading to authentication [5, 12, 19, 35, 48, 52, 60, 63]. As speakers and microphones have become prevalent and accessible in people's everyday lives, acoustic sensing has also found traction in being utilized as a ubiquitous human-computer interface. Gao *et al.* [19] proposed a silent speech interface by leveraging the dual speakers and microphones in the smartphone to capture the user's mouth and tongue movements with the light-weighted MobileNet. Iravantchi *et al.* [26] developed a wrist-worn system that used an array of small transducers to emit ultrasonic beamforming for hand gesture recognition. MilliSonic [59] was proposed for the acoustic-based motion tracking and localization on the VR headset with sub-mm 1D tracking accuracy.

As the closest work to ours, Mao *et al.* [37] proposed a hand motion tracking system by utilizing a 4-element microphones array and dual speakers to measure the propagation distance and angle-of-arrival (AoA) of reflected signals. This technique, however, requires high power (> 50 watts) speakers and a linear microphone array, which is not suitable for wearable and mobile applications. Wang *et al.* [61] designed a smartphone-based gesture recognition system, which can recognize 15 gestures. To the best of our knowledge, SonicASL is the first of its kind to recognize fine-grained, heterogeneous sign language gestures by leveraging acoustic sensing on commodity earphones and ordinary consumer-grade speakers and microphones.

2.3 Earphone-Related Interaction and Sensing

Recent advances in hearable devices such as earphones have bought great opportunities for providing new human-computer interfaces and ubiquitous sensing applications. Researchers have explored various modalities [33, 40, 57] to investigate the potentials of earphone-related applications. Xu *et al.* [65] developed an on-face interaction through wireless earphones by recognizing different sounds of on-face gestures (e.g., single tap, slide on the cheek, and complex slide on pinch). Kikuchi *et al.* [28] proposed EarTouch, an ear-based input interface by utilizing an optical sensor to capture the slight device deformation caused by pulling the ear. Besides, as a human-computer interface, earphones can also turn into a sensory platform for in-ear physiological sensing. EarEcho [20] utilized the built-in speaker and microphone of the earbud to capture the acoustic profile obtained from the user's ear canal for authentication. Similarly, Amesaka *et al.* [1] developed a facial expression recognition system utilizing the user's ear canal transfer functions corresponding to different facial muscle movements. Bui *et al.* [9] proposed a device called "eBP" to monitor the blood pressure inside the user's ear canal, which brings a positive impact on users' daily health monitoring.

3 SONICASL DESIGN

This work aims to recognize the gestures used in sign language in a user-friendly and privacy-preserving manner. Leveraging an acoustic sensing technique, SonicASL can recognize the ASL hand gestures in a typical face-to-face conversation setting, without requiring extra handheld or hand-attached devices, fixed camera setup, or specialized infrastructure support. The system, consisting of a pair of ordinary microphone and speaker, can be added on or built into existing commodity earphones. When an ASL signer performs sign language gestures, an inaudible sonic wave is emitted from the earphone-speaker (worn by the viewer) towards the upper body of the ASL signer. The sonic wave is partially reflected by the moving hands of the ASL signers while they are performing the fine-grained hand gestures (sign language). We leverage the long-term recurrent convolutional networks with connectionist temporal classification (CTC) to continuously recognize subtle pattern changes of the reflected sonic wave and convert them into text. To provide a seamless user experience, we generate the speech audio from the recognized text using a state-of-the-art text-to-speech (TTS) engine and then play the speech audio in the user's earphone. In this section, we introduce the processing flow and implementation methods of the SonicASL system.

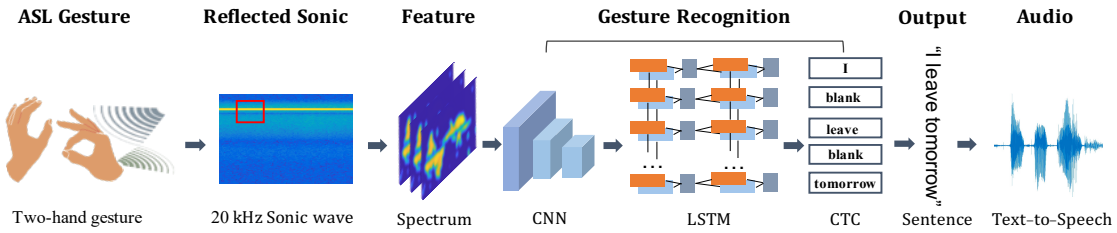


Fig. 2. Diagram of SonicASL processing flow. The inaudible sonic wave is reflected by the moving hands and captured by the microphone. The system extracts spectrums from the reflected echoes based on the Doppler effect. Afterward, SonicASL utilizes the deep learning technique (CNN+LSTM+CTC) to convert the spectrums into text and also generates speech as the feed source to the earphone.

3.1 System Processing Flow

SonicASL recognizes the sign language gestures in four steps, as shown in Fig. 2. Firstly, the earphone user needs to wake up the system by double-tapping near the microphone. Then an inaudible sonic wave is emitted from the earphone's outward-facing speaker, and the outward-facing microphone captures the reflected echoes from hand motions of the ASL signer. Afterward, the system analyzes the echoes and generates a Doppler-effect-based spectrogram as a feature map, which is passed into a sign language gesture recognition model. Finally, the predicted output sentence from the recognition model is converted into speech and delivered to the user via their earphones.

3.1.1 Wake-up/Activation and Disabling Operations. In SonicASL, we design an activation operation to avoid recognizing irrelevant, non-ASL hand movements. We ask the user to register the double-tapping in front of the SonicASL at a distance of 2–5 cm from the microphone as a wake-up operation. Similarly, we also define a disabling function for which the user performs hand-grasping in front of the microphone to stop the recognition. This is useful when the conversation with the ASL signer ends and the ASL viewer switches to the normal usage mode of earphones without physically touching or even looking at their earphone-paired smartphones.

3.1.2 Acoustic Sensing. This section introduces our approach for performing signal enhancement and then presents our spectrogram extraction and data augmentation strategies.

Signal Enhancement. The frequency band for urban noise (which is generated predominantly by traffic) typically ranges from 1 kHz to 4 kHz [45], and the operating frequency of the continuous wave (CW) signal in our design is 20 kHz. Hence, the signals used in our work do not conflict with the frequency bandwidth of natural ambient noise. To remove background noise from the sensed signal, we apply a Butterworth band-pass filter with the pass frequency from 18 kHz to 21 kHz.

Feature Extraction. Based on the Doppler effect, sign language gestures, including both hands and arms, will cause phase and frequency changes of the reflected sonic wave. Some existing works[61, 74] utilized channel impulse response (CIR) to extract fine-grained features, however, it requires audio signal encoding and decoding scheme, which costs extra computational resource and processing time. As a real-time communication tool, we make the trade-off between the performance and the latency by using spectrograms. After signal enhancement, we utilize the short-time Fourier transform (STFT) to calculate a spectrogram as the feature representation of the reflected near-ultrasound waves. The spectrogram contains information in both frequency and time domains. The 2D patterns are highly correlated with the hand and arm movements [64]. The spectrogram is also defined

as the Power Spectral Density of the function:

$$\text{spectrogram}\{x(t)\}(\tau,\omega) \equiv |X(\tau,\omega)|^2 = \left|\sum_{n=-\infty}^{\infty} x[n]\omega[n-m]e^{-j\omega n}\right|^2 \tag{1}$$

where $x[n]$ is input signal, and $\omega[n-m]$ represents the overlapping Kaiser window function with an adjustable shape factor β that improves the resolution and reduces the spectral leakage close to the sidelobes of the signal. The coefficients of the Kaiser window are computed as:

$$\omega[n] = \frac{I_0\left(\beta\sqrt{1-\left(\frac{n-N/2}{N/2}\right)^2}\right)}{I_0(\beta)}, 0 \le n \le N \tag{2}$$

Spectrogram Enhancement. Given the fact that sign language gestures often involves swift, subtle hand movements, the raw spectrograms may not demonstrate apparent patterns corresponding to those hand motion. For instance, as shown in Figure 3(a), it can be seen that the shifts with the 20 kHz pure tone are unclear and ambiguous. To highlight and spotlight the frequency changes along with the time, we propose a series of steps to enhance the spectrogram. First, we used a Butterworth band-pass filter to obtain the target frequency bins between 19 kHz and 21 kHz, because the frequency shift will lie within this range and most of the irrelevant noise would be removed. Second, a band-stop filter was adopted to remove the bins near the dominant frequency (19,985 Hz - 20,015 Hz). Then, we performed STFT with the same segmented frame length of 8,192 points and an overlap of $0.95 \times 8,192$, which can provide sufficient frequency resolution to recognize the ASL. After that, we calculated the sharpest change of STFT values and set that point as the threshold. Any STFT values less than the dynamic and spectrogram-dependent threshold will be set to zero. Lastly, a 2-dimensional Gaussian low-pass filter and Winnier filter were applied to improve the quality of spectrogram pictures by reducing some isolated noise points. Based on these steps, we can clearly recognize and distinguish the subtle changes in every spectrogram. To demonstrate the effects of the proposed enhancement strategy, we plot the acquired spectrograms for the hand gestures of the ASL word "LOVE," with and without enhancement. Fig. 3 (a) shows the original spectrogram around 20 kHz, 3 (b) shows the spectrogram with bandstop filter but without removing noise based on the threshold (b), and 3 (c) shows the final enhanced spectrogram with clearly distinguishable patterns corresponding for the ASL word “LOVE”.

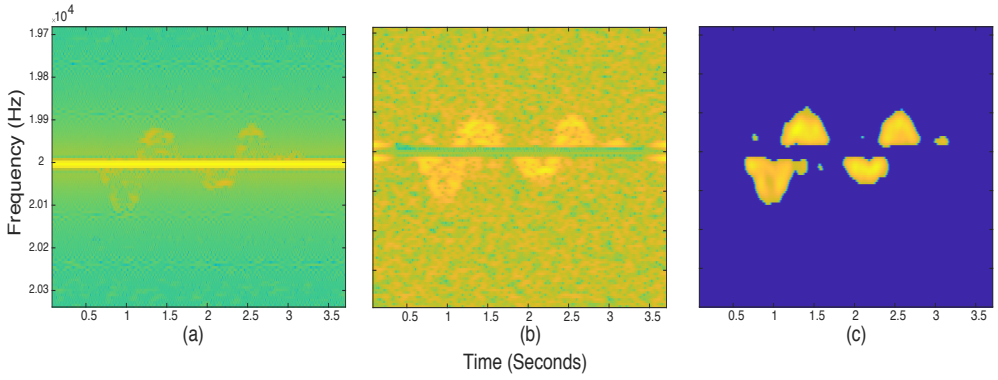


Fig. 3. (a) The raw spectrogram figure without bandstop around 20 kHz;(b) The spectrogram figure without threshold noise remover which means a lot of noise; (c) The final enhanced spectrogram;

To further verify our hypothesis that different gestures would generate unique spectrogram patterns, as a pilot feasibility analysis, we illustrate the spectrogram patterns resulting from several hand gestures corresponding to distinct words and sentences, as shown in Fig. 4. Specifically, the brighter areas (yellow) indicate more distinct and distinguishable hand movements in the corresponding time and frequency domain.

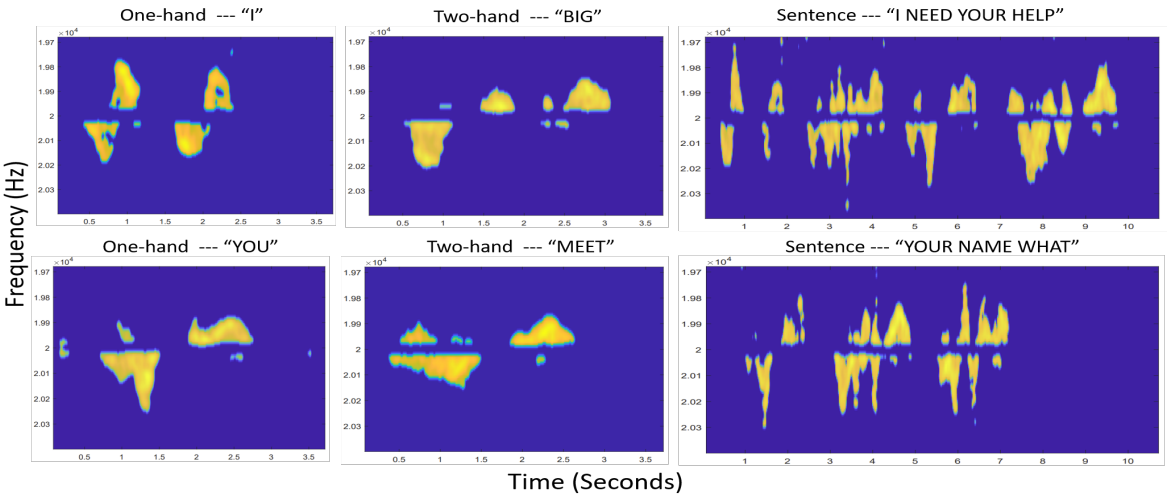


Fig. 4. Example plots of extracted features for different words and sentences. The brighter blocks (yellow) contain unique features from the beginning of the gestures to the end of the gestures for both words and sentences.

Augmentation. To achieve better robustness against noise which is caused by variations of earphone wearing position, we aim to generate additional noisy training data for our deep neural network model. Inspired by the recent success of data augmentation in the speech recognition domain, we utilized the augmentation policy proposed by Daniel Park *et al.* [44], which consisted of warping the features along with time steps, masking blocks of frequency channels, and masking blocks of time steps. Given a Mel-spectrogram with τ time steps, time warping is applied by fixing anchor points on the boundary - four corners and two mid-points of the vertical edges, and warping the random points along with the horizontal line within the time step $(W, \tau - W)$ to either left or right by a certain distance $w \in [0, W]$, where W is the time-warp parameter. Time-Frequency masking is applied by masking consecutive frequency channels $[f_0, f_0 + \Delta f)$ and time steps $[t_0, t_0 + \Delta t)$, where Δf is chosen from a uniform distribution from 0 to the frequency mask parameter F, and $f_0 \in [0, v - \Delta f]$, and Δt is chosen from a uniform distribution from 0 to the time mask parameter T, and $t_0 \in [0, \tau - \Delta t)$. In this paper, we use both time warping and time-frequency masking with customized augmentation parameters.

3.1.3 Sign Language Gesture Recognition. The augmented spectrogram features with the size of $224 \times 224 \times 3$ that hold the information of hand gesture motions are fed into the CRNN (CNN-LSTM) - CTC based deep learning model (CRNN: convolutional recurrent neural network; CNN: convolutional neural network; LSTM: long-short-term-memory; CTC: connectionist temporal classification), inspired by the analogy with optical character recognition and speech recognition tasks [2, 51]. Compared with conventional CNN models, the CRNN model is more suitable for image-based sequence recognition tasks and possesses multiple advantages, such as the ability to learn directly from sequence labels (e.g., words, sentences) and the flexibility in input sequence lengths (i.e., number of frames and words). As shown in Fig. 2, this model contains three major components: the convolutional layers, the recurrent layers, and a transcription layer, from bottom to top. Specifically, in the

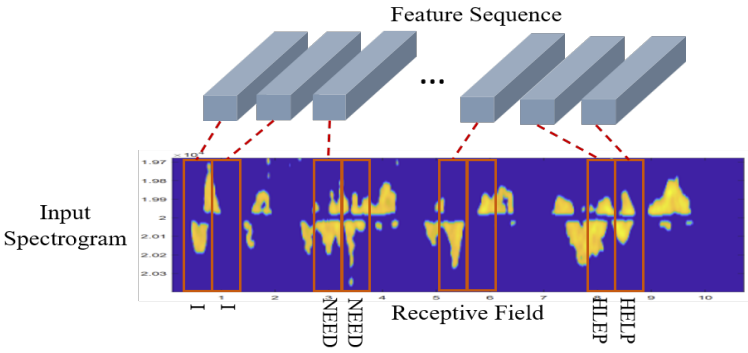


Fig. 5. The receptive field in the input spectrogram. Each vector in the feature sequence in the extracted from the receptive field in the image.

bottom part, a VGG16-based convolutional network extracts the feature sequence from each input spectrogram. Each vector of the feature sequence is associated with a receptive field in the spectrogram. As shown in Fig. 5, some complicated sign language gestures (e.g., help, cold, space) might need multiple successive frames to be properly recognized. Afterward, a BiLSTM-based recurrent network is built on the top of convolutional layers to learn and predict the sequence labeling. On top of the CRNN model, a transcription layer aligns the per-frame predictions from the recurrent layers and converts them into the sentence-level label sequence.

3.1.4 Recognition Result Feedback. After predicting the sign language content, we design a recognition result feedback module such that the content can be naturally delivered to the user without any interruption during the conversation. In this module, every predicted sentence is fed through txt files in the plain text form and passed into the app with standard I/O. To convert texts to speech audio for the user, we apply the *speech.tts* package provided in Android API. The ultrasound signal is played via android's media player library in a parallel manner.

4 SONICASL IMPLEMENTATION

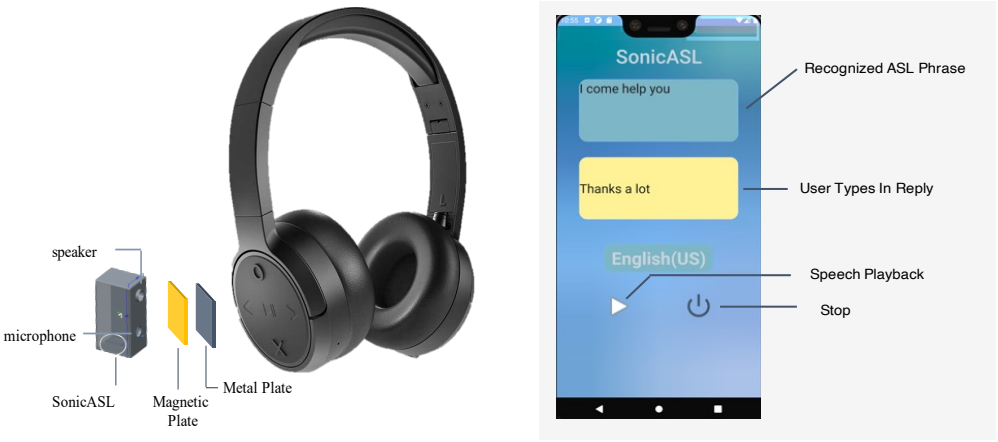
4.1 Hardware

SonicASL contains a microphone (Sound Professionals SP-TFB-2) and a speaker with 1 cm apart from each other vertically in a 3D printed case, as shown in Fig. 6(a). The dimension of the entire device is 25.4 mm × 21.05 mm × 30.00 mm. The front side of device has two holes, the upper one is designed for the sound outlet from the inside speaker, and the lower one is used for placing the microphone. Both the speaker and microphone are connected to the smartphone via a 3.5mm audio splitter adapter for data transmission. To enable a flexible adjustment of the aiming angle of the device towards the upper body region of the ASL signer, we utilize a magnetic attachment from a magnetic plate and metal plate attached to the SonicASL and the earphone.

4.2 Mobile Software

We designed a mobile app to support SonicASL. This app provides the user with two forms of recognition feedback, including on-screen text and audio speech generated through the TTS engine (Figure 6(b)). The app receives the reflected echo signals from the microphone in SonicASL and transmits the data to a remote server. Once the server sends back the recognition result, the user can either check the ASL recognized text on the smartphone screen or play the synthesized speech audio of the recognized text in an automated manner.

In this work, we assume that the earphone user is not familiar with sign language. To facilitate a two-way communication between the sign language users and non-sign language users, we will also intend to support



(a) (b)

Fig. 6. (a) SonicASL Layout. (b) Application UI design.

state-of-the-art voice input/typing functions in SonicASL and convert regular speech of the non-ASL users to text or sign language animations on the smartphone screen. These details are discussed in Section 6.5.

4.3 Sign Language Gestures Selection

4.3.1 Word Selection. As shown in Table 1, we tested the different designs with a selection of 42 ASL words from the most frequently used words in the ASL vocabulary in terms of five categories: noun, verb, adjective, adverb and pronoun [14]. Amongst these 42 words, 22 words are associated with two-handed gestures and the rest (20 words) are performed with the dominant hand (the right hand for our participants).

Table 1. The ASL word selection (words with two-hand gestures are indicated with underlines)

Category	Words
noun	name, friends, time, family, space, camera, boy, woman, mirror, uncle, aunt
verb	want, don't want, live,meet, love, help, choose, need, like, no thanks
adjective	big, small, cold, hot, nice, bad, sad, many, no
adverb	but, how, too/same, and, please, hello
pronoun	what, I, you, who, other, he/she

4.3.2 Sentence Selection. ASL has its only nature and structure compared with English language. In this work, as a preliminary study, we followed the simplified selection strategy in sentence-level gesture and lipreading recognition tasks [3, 16], we designed the sentence template by "$subject^{(3)}$ + ($predicate^{(6)}$) + $object^{(8)}$", where the superscript represents the number of word choices for each term. Specifically, *subject* could be { I, you, he/she }, *predicate* is designed to be { need, love, meet, don't want, help }, and *object* is chosen from { name, camera, time, family, mirror, nice, cold, help }. Theoretically, there are over 140 combinations, and we chose 30 meaningful sentences as our sentence-level dataset such as "Your name what", "How (are) you" ,"Nice (to) meet you".

4.4 Participants and Apparatus

We recruited 8 subjects (7 male, 1 female, average age 28 years) to be the ASL signers for this study. None of the participants were Deaf (i.e., they do not use ASL as their primary language), but all practiced the selected ASL words and sentences for a period of 30-60 minutes. The ASL viewer wore a commercial wireless headphone with the attached SonicASL system. The SonicASL was connected to a smartphone via a 3.5 mm audio cable for data transmission.

4.5 Procedure

We implemented the within-subject study [23] with a 2×42 factorial design, with *Session* and *Gestures* being the factors. Specifically, we asked participants to perform 42 different word-level gestures and 30 different sentence-level (i.e., consisting of 2-4 words) gestures for 6-10 times in two separate sessions in front of the earphone user (i.e., the ASL viewer). In total, we collected 1,200 samples from each participant (i.e., (42 word gestures × 10 times + 30 sentence gestures × 6 times) × 2 sessions). To provide better time alignment for annotations, participants were instructed to perform the gestures followed by a 1-second audio cue for a certain period (i.e., word-level: 5 seconds, sentence-level: 13 seconds). After finishing each session, the participants were asked to walk away and have a 10-minute break. The experiment for each participant lasted for about 5-6 hours.

5 PERFORMANCE EVALUATION

In this section, we conducted a comprehensive evaluation with 8 participants to evaluate the SonicASL's gesture recognition performance. A quiet environment (a 400 ft^2 noise-controlled room) with few people walking around was chosen to simulate a real-world conversation setting. During the experiments, the participants were seated in the room with an ambient noise averaged at 40 dB. The experiments were approved by the Internal Review Board (IRB) of the University at Buffalo, State University of New York for human subjects.

We calculate the word accuracy W.Acc for word-level and sentence-level recognition according to Equation 3,

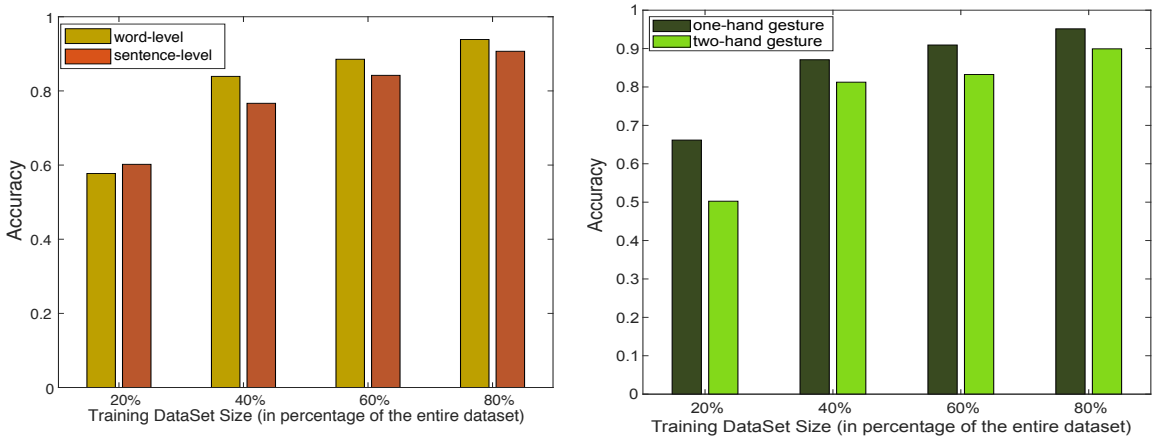
$$\text{W.Acc} = 1 - \text{WER} = 1 - \frac{D + S + I}{D + S + C} \tag{3}$$

where D, S, I, and C represent the number of deletions, substitutions, insertions, and correct words, respectively. The WER (word error rate) denotes a similar metric of performance as the one used in speech recognition systems [29]. A higher $W.Acc$ indicates a higher accuracy of the sign language gesture recognition by SonicASL.

5.1 Recognition Accuracy

5.1.1 Word-Level. Because of the significance of individual ASL words in sign language gesture recognition, SonicASL first evaluated the user dependent recognition accuracy for the word-level ASL signs and presented the average accuracy of eight subjects. We collected data from eight subjects for all 42 representative ASL words, and repeated every ASL word for 20 times. The evaluation metric was defined as the percentage of gesture samples for all 42 ASL words that can be correctly detected and recognized by SonicASL.

It is well known that different sizes of training data would affect and result in different learning outcomes for machine learning-based models [55]. To investigate the influence of the number of training samples on model accuracy, we assessed the recognition accuracy by using 20%, 40%, 60%, and 80% of the entire collected samples for each subject as the training dataset, and keeping the remaining samples as the testing dataset. Unsurprisingly, Fig. 7(a) shows that the increasing number of training samples can significantly improve the recognition accuracy of ASL words. When 80% of the collected samples are used for training, the average accuracy of eight subjects based on 42 ASL words is 93.8%. Thus, it can be concluded that SonicASL can precisely recognize individual ASL words.



(a) (b)

Fig. 7. (a) User dependent Recognition accuracy of word-level and sentence-level ASL with increasing training dataset size. (b) User dependent Recognition accuracy of one-hand and two-hand ASL gestures with increasing training dataset size. We tested the training dataset with four different sizes, which were 20%, 40%, 60%, and 80% of the entire dataset for each subject respectively.

Table 2. Accuracy of isolated signs

Words	Acc	Words	Acc	Words	Acc	Words	Acc	Words	Acc
"AND"	1	"I"	0.86	"AUNT"	0.90	"BAD"	0.96	"NO,THANKS"	1
"YOU"	0.97	"LIKE"	0.93	"COLD"	1	"FAMILY"	0.96	"CHOOSE"	0.83
"HELP"	0.93	"BOY"	0.89	"HOT"	0.96	"HOW"	0.96	"FRIEND"	1
"LOVE"	0.69	"MANY"	0.96	"MEET"	0.93	"MIRROR"	1	"NAME"	1
"NICE"	0.93	"BIG"	1	"NO"	0.93	"OTHER"	0.90	"DON'T WANT"	0.89
"SAD"	0.94	"SHE/HE"	1	"SMALL"	0.83	"SPACE"	0.93	"TOO/SAME"	1
"UNCLE"	0.93	"WANT"	1	"WHAT"	0.93	"WOMAN"	0.93	"CAMERA"	0.93
"BUT"	0.90	"WHO"	0.82	"LIVE"	0.97	"NEED"	1	"PLEASE"	0.86
"TIME"	1	"HELLO"	0.9						

In order to assess the performance of every single gesture, we randomly chose two subjects and used 80% of data for training and the rest 20% for testing for each user. Table 2 lists the two subjects' average performance for every single word, which shows that most of the words can be recognized accurately.

5.1.2 Sentence-Level. We evaluated the sentence-level recognition performance of SonicASL by collecting gesture data of 30 meaningful ASL sentences and repeating 12 times for each of the eight subjects. As the user-dependent evaluation, we trained individual models for each subject, and utilized 80% (9 samples for each sentence) ,60% (7 samples for each sentence), 40% (5 samples for each sentence) and 20% (3 samples for each sentence) of whole sentences from the same subject as training set, and used the remaining sentences as testing set. To evaluate the recognition performance, we calculate the recognition accuracy of individual words in testing sentences (i.e., similar as the Word Error Rate). In Fig. 7(a), the ASL recognition accuracy is 90.6% when the training dataset is 80% of the entire dataset. The results showed that SonicASL performed well in sentence-level recognition.

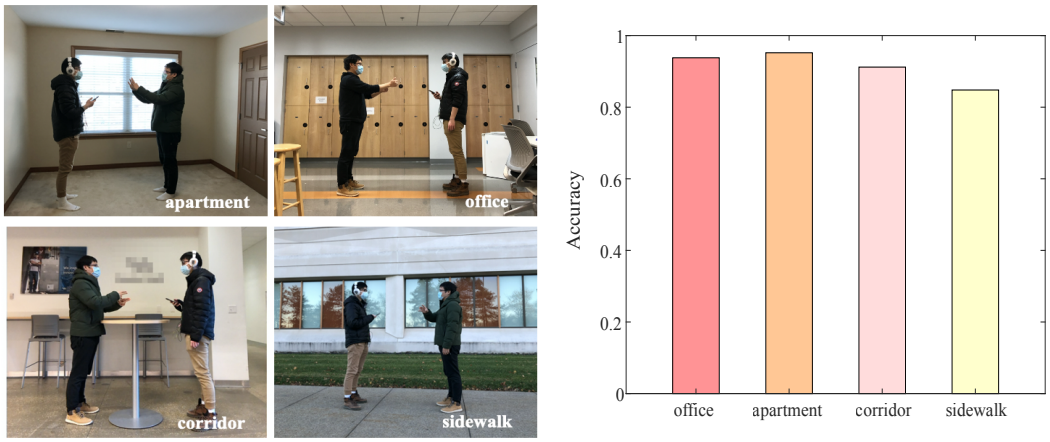
However, it is also observed that the sentence-level recognition accuracy is slightly lower than the word-level recognition accuracy, because there are various levels of uncertainty with the word transitions (e.g., time-intervals), which will affect the performance. Thus, it is inevitably more challenging to achieve satisfactory sentence-level ASL recognition than the word-level recognition. However, in order to provide a truly useful assistive communication tool for ASL signers, it is imperative for ASL recognition systems to consider and assess the performance of sentence-level recognition. This part was largely missed in prior studies [34, 61, 62] reported in the literature.

5.1.3 Different Types of Gestures. To better understand the performance of the proposed approach on different types of gestures, we examined the performance based on both common types of ASL gestures, i.e., one-hand and two-hand gestures. Thus, we divided these gestures into two groups: the first group contains 20 one-hand gestures and the second group contains 22 two-hand gestures. It is well known that these two types of gestures may cause significantly distinct reflected patterns of the sonic wave. Fig. 7(b) also shows that SonicASL can achieve a relatively higher accuracy for recognition of one-hand gestures, than the accuracy for two-hand gestures recognition. A possible reason is that, two-hand gestures have more significant reflection of sonic waves and thus more complex acoustic patterns. Therefore, if there are some subtle variations when performing the same gesture for more times, two-hand gestures would be more vulnerable to various levels of gesture variations, which will result in a slightly higher chance of being labelled incorrectly.

5.2 In-the-wild Evaluation

As the signal's bandwidth (20 kHz) in SonicASL is far beyond the bandwidth of urban noises (e.g., loud radios, background music, vehicle alarms, and roadway traffic) that range from 1 - 4 kHz, our acoustic sensing is resistant against direct ambient noise. However, multi-path interference caused by the reflections from unrelated objects is still a major challenge for acoustic sensing. For example, when the sonic beam aims at a target ASL signer, nearby walking pedestrians around the ASL user might also reflect echoes and introduce additional noises. Therefore, in this section, we evaluate the effectiveness of SonicASL in recognizing sign gestures in a real-world environment. We experimented with four common environmental scenarios for daily conversations, including three indoor environments (apartment, office, corridor) and an outdoor environment (sidewalk) to test the reliability of our proposed sign language gesture recognition system.

In order to prove this, we provide the ASL recognition accuracy in different settings including an office with the area of 8m × 10m, an apartment with the area of 3m × 6m, a corridor with the area of 10m × 20m, and a sidewalk. We recruited two participants (two male with an average age of 30 years) to act as an ASL signer and an ASL viewer wearing the SonicASL earphone prototype (and swapping the roles later). The ASL signer performed a selected subset of 15 sign language gestures (e.g., "I", "YOU", "COME", "HELP", "LOVE", "MEET", "MIRROR", "NAME", "NICE", "DON'T WANT", "WHAT", "SAD", "HELLO", "NO THANKS", "SMALL") with a repetition of 20 times. We used the data recorded from the office in the training phase and used other data recorded from other environments in the testing phase. In addition to the indoor environments, we also selected the sidewalk by the campus road as the outdoor testing environment (ambient noise averaged at 60 dB) shown in Fig. 8(a). The ASL signer and viewer were standing face to face on one side of the sidewalk, while the pedestrians were walking back and forth on the other side. As shown in Fig. 8(b), the performance is quite consistent among all three indoor environments, while showing a lower performance outside the building given the complexity of the environmental setting. Another finding is that SonicASL achieves the best recognition performance in the apartment setting, as it is a quiet environment with the least noise interference.



(a) (b)

Fig. 8. (a) User study in the four different environments. (b) Recognition Performance.

Table 3. The average power consumption and time latency on different smartphones.

	Pixel 3 XL	Pixel 3
Power (mAh/min)[1]	6.86	5.83
Avg. Latency (s) [2]	1.03	1.07

[1] The battery size for Pixel 3 XL and Pixel 3, is 3430 and 2915.

[2] Results are tested under the same data sample (single word gesture) and averaged for 10 runs.

5.3 Usability Analysis

To enable natural and smooth communications between the ASL signer and the viewer, SonicASL should be capable of recognizing ASL signs in real-time and on the fly. Thus, in this section, we seek to evaluate the power consumption for transmitting and recording ultrasound signals, as well as the latency for recognizing the ASL signs with SonicASL implemented on different smartphones. To measure power consumption, we keep the speaker and microphone in our system in the active status (emitting and recording) for over 20 minutes and measure the power consumption during the usage. As shown in Table 3, it is observed that the power consumption of the front-end module is quite low (consuming 2% for every 10 minutes) and could well satisfy the long-term conversation requirements. Even though our system needs about one second to process and predict a single gesture on mobile ends. Considering the speed of sign language gestures during the conversation is much slower compared with vocalized speech, the 1-second latency is still acceptable.

5.4 Robustness Quantification

In order to investigate the robustness of SonicASL in different application scenarios, we conducted a comprehensive evaluation with various real-world situations and device configurations, including sensing distances, aiming orientations of the earphone, data augmentation, and head motions.

5.4.1 Impacts of Sensing Distances. As the sound wave intensity decreases in inverse proportion to the squared distance between the two objects, as well as due to channel fading, the gesture recognition accuracy will drop along

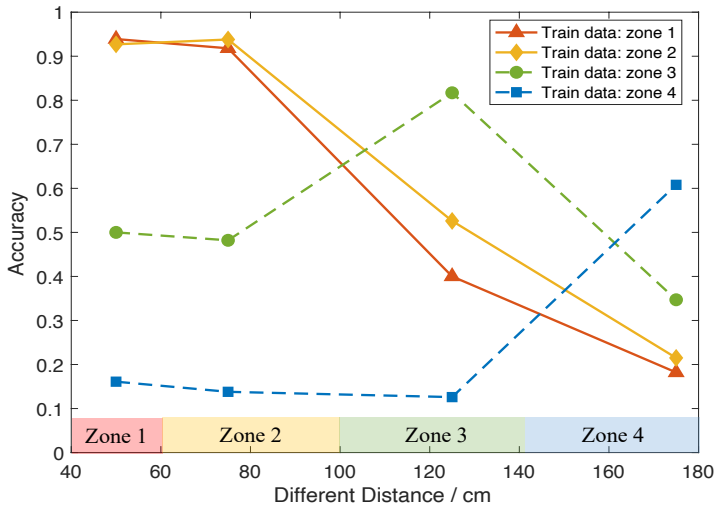


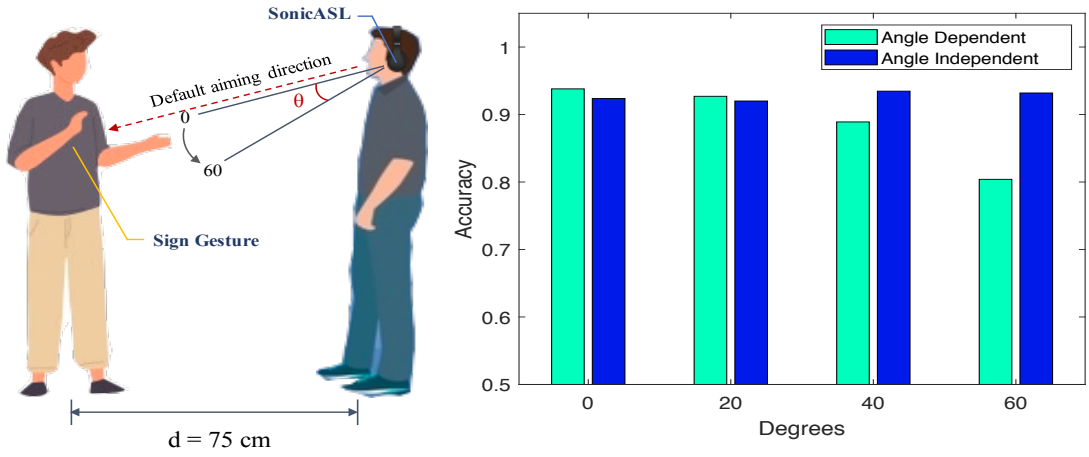
Fig. 9. Result with different distances between the ASL signer and the earphone user (viewer). Different lines represent training data from different zones, (e.g., "Train data: zone 1": Training data from distance 50 cm, "Train data: zone 2": Training data from distance 75 cm, "Train data: zone 3": Training data from distance 125 cm, "Train data zone 4": Training data from distance 175 cm). We label the three comfort zones based on distances d. (e.g., intimate: d < 80 cm; normal: 80 cm < d < 150 cm; stranger: d > 150 cm).

with the distance between the ASL signer and viewer. Hence, to understand the trade-off between recognition accuracy and comfortable distance in conversations, we chose to experiment with four different physical distances (i.e., 50 cm, 75 cm, 125 cm, and 175 cm) between the ASL signer and the earphone user (ASL viewer). The reason to choose these four values is twofold: (a) based on our empirical analysis, the sensing distance limit for a low-power speaker (0.5 Watt, 50 dB) is around 200 cm; and (b) the normal social conversation distance between two people ranges from 45 cm to 200 cm [41].

In this experiment, we recruited two participants (two male with an average age of 30 years) to act as an ASL signer and an ASL viewer worn the SonicASL earphone prototype (and swap the roles later) in a noise-controlled room (40 dB ambient noise). The ASL signer performed a selected subset of 15 sign language gestures (e.g. the same subset with the section 5.2) with a repetition of 20 times. These 15 words are one-hand or two-hand gestures that contain different handshapes, different movements, different orientations, and different locations. We recorded the data at four distances and utilized the dataset of one distance for training and the datasets of the other three distances for testing, respectively. As shown in Fig. 9, when the training and testing datasets are both from the intimate zones, the recognition accuracy is high enough to accurately recognize the ASL signs. However, when both the training and testing datasets come from the normal or stranger zone, the ASL recognition for the datasets from other zones is relatively low.

To address this issue and improve the applicability in real-world scenarios, we can train a relatively robust model based on a wide range of distances. Hence, we make use of the data from all these four distances and randomly choose 80% of data for training and the remaining for testing. An average accuracy of 82.0% is obtained. The overall performance degrades because a too long distance between the signer and viewer (e.g., over 140 cm) would significantly weaken the capability of accurately recognizing the ASL signs. Thus, we suggested that the user should chose a relative near distance to communicate.

It is well acknowledged that the acoustic signals can propagate a longer distance along with the increased power. Thus, we consider a speaker with higher power will be more effective to improve the ASL recognition



(a) (b)

Fig. 10. (a) An explanation of aiming orientations (θ); (b) Result with increasing angles between the ASL signer and the earphone user (ASL viewer).

accuracy for a rather large distance. In order to validate this hypothesis, we chose a speaker with a higher power (1 Watt, 70 dB) with the same microphone in the SonicASL prototype, and recorded the data using the same four distances as the settings above. We utilized the recorded data at the distances of 50 cm, 75 cm, and 125 cm as the training dataset, and used the dataset recorded at the distance of 175 cm as the testing data. The recognition performance jumps to 72%, which is much higher than the performance we obtained (23%) using the original speaker (0.5 Watt, 50 dB).

Based on all the evaluations above, it is shown that our approach is applicable for different communication distance scenarios. However, we will see a drastic performance drop when the distance between the ASL signer and viewer is relatively large. This performance limitation could be partially overcome by using a higher-power hardware, whereas we should make an optimal trade-off between the desired sensing distance and the acceptable speaker power. Given the typical constraints of battery capacity in hearable devices, we are aiming at a low-power design and it is thus recommended that the SonicASL users don't stay too far away from each other to ensure a satisfactory recognition accuracy.

5.4.2 Impacts of Aiming Orientations. Within a natural face-to-face conversation, the user's earphone wearing angle (i.e., aiming direction relative to the hand gestures of the ASL signer) is sometimes inconsistent and easy to change depending on the user's attention (e.g., looking at the ASL signer's face or hands) or body postures (e.g., sitting or standing). As discussed in Section 5.2, the reflected echo is sensitive to the propagation path, which is dominated by the surrounding obstacles and propagation angles. In order to examine the impacts of aiming orientations of the SonicASL device on recognition performance, we conducted the experiment with various aiming angles towards the hand gestures of the ASL signer. As shown in Fig.10(a), we designate the angle (θ) between the SonicASL System and the hands which perform the sign language. During the training phase, the earphones faced the hands directly ($\theta = 0$). We measured the reflected echoes with different angles θ ranging from 0 to 60 to simulate the varying wearing behaviors in daily conversation scenarios.

In this experiment, we also set the same environment as Section 5.4.1 and recorded the same 15 representative ASL words which were repeated 20 times for every specific angle. To explore the effect of varying aiming angles, We performed an angle-dependent experiment with limited training angles, and another angle-independent

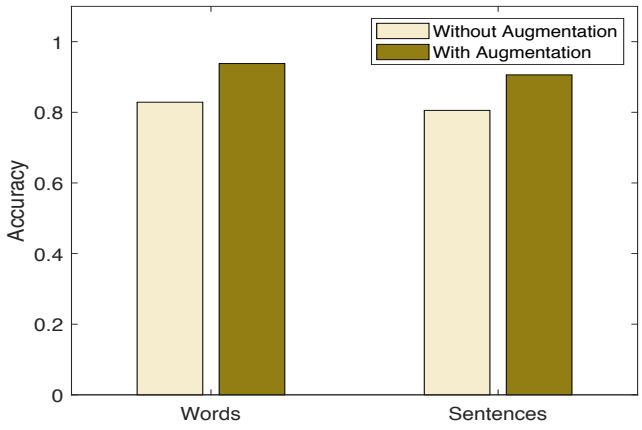


Fig. 11. The recognition accuracy with and without data augmentation

experiment with more training angles. For the angle-dependent experiment, we utilized the data recorded at angle 0 ($\theta = 0$) for training and the data recorded with changing angles ($\theta = 20, 40, 60$) for testing. The ASL viewer stands in front of the ASL signer at the default distance of 75 cm. Fig. 10 (b) shows the recognition accuracy (in light blue) based on different aiming angles. When the aiming angle increases, the recognition accuracy gradually drops. Accordingly, the recognition accuracy has shown a clear drop for the $\theta = 60$ case. Thus, we will further explore the scenario when including more angles data in the training.

For the angle-independent experiment, we utilized any three out of the four angle datasets for training and the remaining one for testing, and obtained the performance (in dark blue) as shown in Fig. 10 (b). This result is consistent with our expectations because the hands would go beyond the coverage of transmitted sounds with the increasing angles. It is observed that, only when the aiming angle is larger than 40 degrees, the signer's hands were partially out of range of the ultrasound signal regions which will reduce the quality of reflected echoes. Thus, based on the angle-independent training method, the performance drop is not observed for any angle dataset and we have achieved a satisfactory level of resistance to the aiming angle variations.

At last, we picked all the data from four angles, and randomly chose 80% of them for training and the rest for testing. We achieved an accuracy of 95% which means that the ASL recognition accuracy can be achieved at a very high level when using enough training data from different angles.

5.4.3 Impacts of Data Augmentation. As discussed in Section 3.1.2, to generate more training data with similar characteristics and traits and increase the system generalizability, we applied the data augmentation strategy of time warping and time-frequency masking. To validate the performance of data augmentation, we tested all the 42 ASL words and 30 sentences from 8 subjects with augmentation and without augmentation. As shown in Fig. 11, we can observe a clear accuracy increase by using the augmentation in both word-level and sentence-level datasets.

5.4.4 Impacts of Similar ASL Signs. Each ASL word has a unique sign gesture, which could generate distinct sound reflection patterns. however, there are some words that are represented with very similar gestures. For instance, we list several groups of words in Table 4 and the words in one group all have very similar sign gestures with only subtle differences. In order to distinguish them in SonicASL, we recruited two subjects to perform these words for 20 times and then utilized 80% of these words as the training data. Fig. 12 presents the confusion matrix for 20 sign words based on 6 similar groups. The overall accuracy ranges from 87% to 99%, except for the group "Safe, Free, Independent" with 65%. As indicated in Fig. 12, the sign "INDEPENDENT" is wrongly recognized as

"FREE" or "SAFE", which come from the same similar group. These three signs have the same hand gesture but with slightly different finger shapes (please refer to the Appendix for detailed sign gestures), which increases the chance of misclassification. Thus, we can conclude that SonicASL has great potentials of distinguishing similar ASL signs with hand-level gesture variations. We aim to improve the recognition accuracy for those fine-grained, finger-level similar ASL signs.

Table 4. The averaged accuracy for each group of similar ASL signs

	Multiple Groups of Similar ASL Signs				
Words	But, Opposite	Group, Family Class, Team	Safe, Free Independent	Name, Brief, Sit Chair, Train	Paper, School, College University, Clean-up
Accuracy	99%	87%	65%	87%	89%

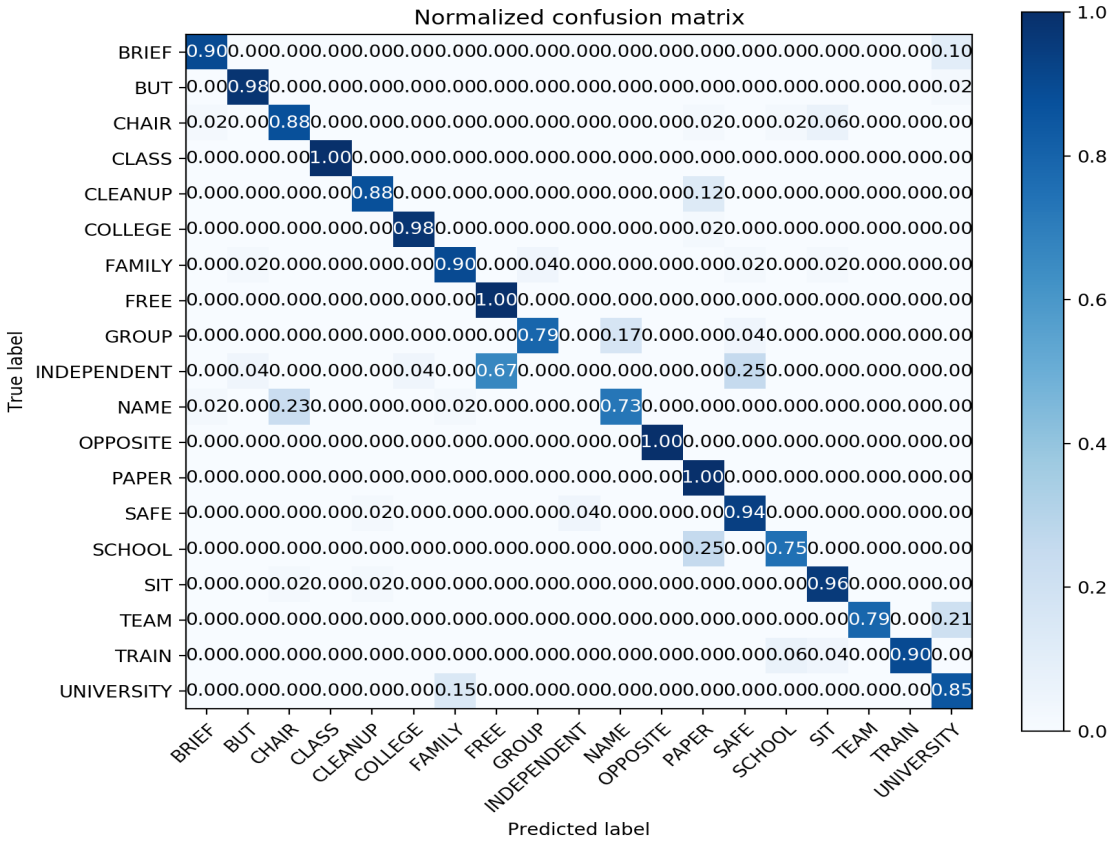


Fig. 12. The Confusion Matrix of Similar Words Recognition.

5.4.5 Impacts of Different Apparel. It is well acknowledged that the reflected signals resulting from different reflectors would lead to distinct recognition patterns. In this section, we seek to examine the effect of apparel on ASL recognition. We recorded the data at a distance of 75 cm with a subset of 15 sign gestures (e.g. the same

subset with the section 5.2). When the ASL signer wears the latex gloves, the ASL recognition accuracy is 93% which is almost the same as the case without any glove. With a pair of thick snow gloves, the performance drops to 87%. We can thus conclude that wearing gloves will not affect the ASL recognition performance drastically because our model learns to recognize the ASL signs based on the user's hand and arm movements, instead of the hand shapes or apparel. It is noted that the performance of ASL recognition has dropped slightly because some gestures with fine-grained finger movements would be affected by the thick snow gloves.

6 DISCUSSION

In this section, we discuss the comparisons between SonicASL and existing wireless gesture recognition systems, performance over new users, the adjustment strategy for sensing aiming directions, the potential generalizability on multiple hardware devices, and limitations.

6.1 Comparison with Existing Contact-Free Gesture Recognition Solutions

As discussed in Section 2.1, there have been a wide variety of solutions for sign language gesture recognition, including vision-based, wearable-sensor-based, and remote-sensing-based approaches. Given the fact that remote sensing techniques provide less intrusive and privacy-preserving experiences for sign language users, they have gained increasing attention and popularity recently. Because SonicASL also aims to provide a non-intrusive, contact-free, daily-use solution, we hence present a comparison with several existing representative methods in contact-free sign language and gesture recognition (the research work in other categories will not be discussed and compared here due to the space limit). As shown in Table. 5, most existing acoustic-based studies only recognized a limited number (less than 15) of user-defined, isolated, and coarse-grained hand gestures and other wireless modalities with high-frequency band (e.g., mm Wave and WiFi) lack the portability, which can not be easily deployed in daily usage scenarios and is also not suitable for outdoor environments. Thus, different from those existing solutions in the literature, we have to address several unique challenges in SonicASL to achieve accurate and robust recognition of general, continuous, fine-grained ASL hand gestures.

- The first challenge is that some sign language gestures have a high degree of similarity with only subtle differences at the finger level. Thus it is hard to distinguish those gestures from a large candidate pool. In this work, to recognize the fine-grained gestures, inspired by Xu *et al.* [64], we improve the spectrogram enhancement method by introducing customized image filters and the dynamic threshold mechanism.
- The second challenge is that, different from the recognition of standalone individual gestures in the existing works, ASL phrases/sentences inevitably contain natural transitions between each word (gesture) to construct meaningful sentences. To address this challenge, we utilized a deep learning model (CRNN) with the architecture of CNN+LSTM+CTC.
- The third challenge is the body motions. During the sign language communications, the user would involuntarily and unintentionally introduce some motion artifacts such as head and body movements. To overcome those subtle variations caused by undesired motions, it is necessary to train a robust deep learning model with the ability to recognize sufficient and fine-grained gestures. Therefore, to expand our dataset, we utilized the data augmentation policy [44], which consisted of warping the features along with time steps, masking blocks of frequency channels, and masking blocks of time steps.

6.2 Performance on New Users

Since individuals have unique preferences for performing particular ASL sign gestures, those preferences would introduce variations on the feature spectrograms and further affect the recognition performance. To evaluate the generalizability performance on new users based on different scales of training datasets, we conducted three groups of experiments based on single-word signs: the first experiment used two subjects for training and one

Table 5. Comparison with existing wireless-based hand gesture recognition interfaces.

Interfaces	**SonicASL**	mmASL [50]	SignFi [36]	RobuCIR [61, 62]	UltraGesture [34]
Features	Acoustic	60 GHz mmWave	WiFi	Acoustic	Acoustic
No. Words	42	50	276	15	12
No. Sentences	30	N/A	N/A	N/A	N/A
Gesture Type	Sign Language	Sign Language	Sign Language	Self-defined	Self-defined
Devices	Earphone	Radio platform	WiFi AP	Smartphone	Smartphone
Portable	Yes	No	No	Yes	Yes
Accuracy	93.8%	87%	86.6% - 98.01%	96.9%	97%

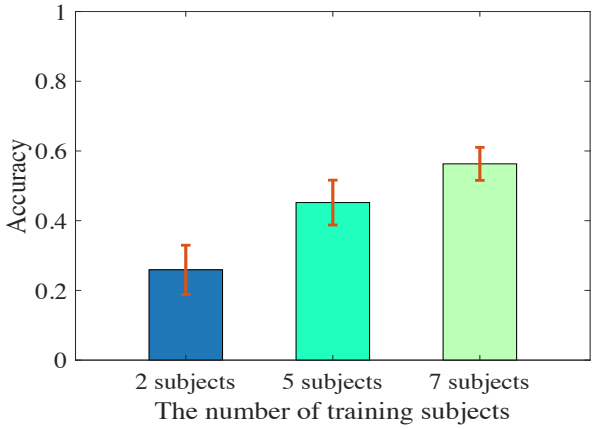


Fig. 13. Recognition accuracy with standard deviation for different training subject number in leave-one-user-out testing

unknown subject for testing, the second experiment used five subjects for training and one unknown subject for testing, and the third experiment used seven subjects for training and one unknown subject for testing. For both the first and second experiments, we randomly selected subjects from the 8-subject pool for 50 different combinations. For the third experiment, eight leave-one-user-out combinations were attempted, and the average accuracy and standard deviation were calculated in Fig. 13. The average accuracy of leave-one-user-out (56.3%) for word-level recognition was much lower than the average accuracy across each subject (93.8%). However, those variances introduced from individuals would be reduced by a generalized model with more subjects' training data. As shown in Fig. 13, given an unknown subject's data, the accuracy using the model trained with seven known subjects samples is much higher than the accuracy using the model trained with only two subjects' data. This result indicates that involving more subjects' gesture data into the training will efficiently improve the system's robustness against individuals' variations. Due to the large variations introduced from individuals when performing sign language gestures (e.g., user preferences, degree of familiarity) and limited participants (N=8) in our current experiment, we can only achieve 56.3% leave-one-user-out accuracy. In our future work, we are planning to enroll more participants to further improve the system's generalizability.

6.3 Adaptive Adjustments for Aiming Direction

Besides the unique behavioral preferences possessed by each individual, the different heights of people would also cause the aiming deviations of the sonic waves. For example, considering a reference height of 175 cm for both the ASL signer and viewer standing with a distance of 75 cm as the baseline, a 10 cm height difference will cause ± 5° - 6° aiming deviation. In our design, SonicASL utilizes a pair of magnetic and metal plate to provide

the attachment on the earphone that not only provides an easy-to-use interface for installment but also enables the flexible rotation for the user to adjust the aiming angle. During the system setup, we will first ask the ASL viewer to look straight towards the signer's hands in front of the chest, as the initial aiming direction. For a better calibration, we can then ask the signer to perform the same sign (e.g., "I") multiple times while the viewer keeps slightly adjust the sensor's aiming orientation to get the best recognition results.

6.4 Generalizability on Hardware

In this study, we design an earphone-based (Audio-Technica M50 headphone) ASL recognition system with magnetic-attached SonicASL. However, some commercial earbuds with small form factors (e.g., Apple AirPods, Samsung Galaxy Buds) are difficult to be compatible with our current design. A miniaturized version with fine components (MEMS microphone [30] and micro-speaker [56]) will be included in the future design to improve the generalizability. SonicASL also has the potential of being deployed on other head-worn devices such as VR headsets and smart glasses as long as there is a symmetrical space on both sides for magnetic attachment of SonicASL without impairing the user's visual sight.

6.5 Limitations and Future Work

As the first attempt of its kind to provide an accessible, non-intrusive, easy-to-use sign language gesture recognition system that can be used in diverse daily conversation scenarios, SonicASL is still in its early stages with many limitations and imperfections.

6.5.1 Two-way Communications between Sign Language Users and Non-Sign-Language Speakers. Our current design of SonicASL enables the one-way communication for helping the non-sign-language speaker understand the ASL signer. To increase the system usability and realize the two-way communication that the earphone-user can talk to the Deaf, we currently design a text window in the app where the user can type-in the reply, as shown in Fig. 6(b). In future work, we would like to extend the functionality of SonicASL by supporting more user-friendly two-way communications. Specifically, we will plan to support state-of-the-art voice input/typing functions in SonicASL and convert regular speech of the non-ASL users to text or sign language animations on the smartphone screen to be shown to the ASL signer.

6.5.2 Expanded ASL Vocabulary Dataset. In this work, as discussed in Section 4.3, we experimented with 42 individual words and 30 sentences, which only cover a small portion of ASL vocabulary pool. To provide the SonicASL with better recognition ability of more words, we will make our app and data collection process publicly available to the Deaf or hard-of-hearing communities. By collecting more data from sign language users, we aim to expand the training basis of SonicASL and improve its generalisability for more words and sentences. It is anticipated that our work could potentially benefit the ASL-related research and facilitate hassle-free communications for sign-language users.

6.5.3 Non-Manual Markers. For ASL communications, non-manual markers comprised of non-affective facial expressions, head positions, lip motions, and body positions often provide crucial grammatical context to the manual signs [39]. Unfortunately, it is challenging to recognize the reflections of acoustic waves based on those non-manual markers. Thus, in future work, we would further explore and optimize our algorithms to recognize the fine-grained motion movements to explore a real practical ASL recognition system.

6.5.4 ASL Grammar. ASL has its own grammar system (i.e., its own rules of phonology, morphology, syntax, and pragmatics), separate from that of English [4]. In this work, as a preliminary study of exploring ASL recognition through acoustic sensing, we evaluate the sentence-level ASL signs by combining the isolated ASL words and English-style sentence structure for easier understanding of the hearing population. In the future work, to improve

the practicability of SonicASL, we will replace the English-style sentence structure with the true ASL grammar and sentence structure.

7 CONCLUSION

In this study, we propose SonicASL, an earphone-based sign language gesture recognition system that leverages inaudible sonic waves to capture and recognize the fine-grained gestures of the ASL signer. We conducted a performance study with an ASL vocabulary dataset consisting of 42 individual words and 30 sentences, based on 8 participants. Our system can achieve a 93.8% word-level accuracy and 90.6% sentence-level accuracy. To evaluate the robustness of SonicASL in real-world scenarios, experiments with various simulated configurations (e.g., head movements, aiming angles, and distances) and environmental factors (e.g., indoor and outdoor with pedestrians walking nearby) have shown the effectiveness and robustness of SonicASL to recognize ASL words and sentences. It is expected that this work could provide an alternative, technological contribution towards technology for the ASL community and ASL gesture recognition research.

ACKNOWLEDGMENTS

We thank anonymous reviewers for their insightful comments and constructive suggestions for the improvement of the paper, especially about the scope, structure, and wording of the manuscript. We also want to express our sincere appreciation to Dr. Henry Adler (Center for Hearing and Deafness, University at Buffalo) for his valuable time in reviewing the manuscript and providing very helpful feedback and suggestions on this work.

REFERENCES

[1] Takashi Amesaka, Hiroki Watanabe, and Masanori Sugimoto. 2019. Facial Expression Recognition Using Ear Canal Transfer Function. In *Proceedings of the 23rd International Symposium on Wearable Computers (ISWC'19)*. ACM, New York, NY, USA, 1–9.

[2] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in English and Mandarin. In *International Conference on Machine Learning (ICML'16)*. 173–182.

[3] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. 2016. LipNet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599* (2016).

[4] Rochelle Barlow. 2018. *ASL Grammar: The Workbook* (1 ed.). CreateSpace Independent Publishing Platform.

[5] Vincent Becker, Linus Fessler, and Gábor Sörös. 2019. GestEar: combining audio and motion sensing for gesture recognition on smartwatches. In *Proceedings of the 23rd International Symposium on Wearable Computers (ISWC'19)*. 10–19.

[6] Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 16–31.

[7] Helene Brashear, Valerie Henderson, Kwang-Hyun Park, Harley Hamilton, Seungyon Lee, and Thad Starner. 2006. American sign language recognition in game development for deaf children. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*. 79–86.

[8] Helene Brashear, Thad Starner, Paul Lukowicz, and Holger Junker. 2003. Using multiple sensors for mobile sign language recognition. Georgia Institute of Technology.

[9] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, Taeho Kim, Nicholas Farrow, Anh Nguyen, Jianliang Xiao, Robin Deterding, Thang Dinh, and Tam Vu. 2019. eBP: A Wearable System For Frequent and Comfortable Blood Pressure Monitoring From User's Ear. In *The 25th Annual International Conference on Mobile Computing and Networking (MobiCom '19)*. ACM, New York, NY, USA, Article 53, 17 pages.

[10] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. SubUNets: End-to-end hand shape and continuous sign language recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 3075–3084.

[11] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10023–10033.

[12] Mingshi Chen, Panlong Yang, Jie Xiong, Maotian Zhang, Youngki Lee, Chaocan Xiang, and Chang Tian. 2019. Your Table Can Be an Input Panel: Acoustic-based Device-Free Interaction Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 3.

[13] Debbie Clason. 2019. Hearing loss statistics at a glance. https://www.healthyhearing.com/report/52814-Hearing-loss-statistics-at-a-glance. [Online; accessed 30-August-2020].
[14] ASL Dictionary. 2021. SIGN LANGUAGE. https://www.handspeak.com/. ASL Dictionary.
[15] Philippe Dreuw, Carol Neidle, Vassilis Athitsos, Stan Sclaroff, and Hermann Ney. 2008. Benchmark Databases for Video-Based Automatic Sign Language Recognition.. In *LREC*.
[16] Biyi Fang, Jillian Co, and Mi Zhang. 2017. DeepASL: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems (SenSys'17)*. Article 5, 13 pages.
[17] Gaolin Fang, Wen Gao, and Debin Zhao. 2006. Large-vocabulary continuous sign language recognition based on transition-movement models. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 37, 1 (2006), 1–9.
[18] Jakub Gałka, Mariusz Mąsior, Mateusz Zaborski, and Katarzyna Barczewska. 2016. Inertial motion sensing glove for sign language gesture acquisition and recognition. *IEEE Sensors Journal* 16, 16 (2016), 6310–6316.
[19] Yang Gao, Yingcheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-based Silent Speech Interface for Smartphone Users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3, Article 80 (2020), 27 pages.
[20] Yang Gao, Wei Wang, Vir V. Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using Ear Canal Echo for Wearable Authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3, Article 81 (Sept. 2019), 24 pages.
[21] Srujana Gattupalli, Amir Ghaderi, and Vassilis Athitsos. 2016. Evaluation of deep learning based pose estimation for sign language recognition. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 1–7.
[22] Blaine Goss. 2003. Hearing from the Deaf Culture. *Intercultural Communication Studies* 12, 2 (2003), 9–24.
[23] Anthony G. Greenwald. 1976. Within-subjects designs: To use or not to use? *Psychological Bulletin* 83, 2 (1976), 314–320.
[24] Matthew D Hickman. 2010. Translation device eases communication problems. https://www.army.mil/article/32679/translation_device_eases_communication_problems. [Online; accessed 23-April-2021].
[25] Jiahui Hou, Xiang-Yang Li, Peide Zhu, Zefan Wang, Yu Wang, Jianwei Qian, and Panlong Yang. 2019. SignSpeaker: A real-time, high-precision smartwatch-based sign language translator. In *The 25th Annual International Conference on Mobile Computing and Networking (MobiCom'19)*. Article 24, 15 pages.
[26] Yasha Iravantchi, Mayank Goel, and Chris Harrison. 2019. BeamBand: Hand gesture sensing with ultrasonic beamforming. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–10.
[27] Dhruv Jain, Rachel Franz, Leah Findlater, Jackson Cannon, Raja Kushalnagar, and Jon Froehlich. 2018. Towards Accessible Conversations in a Mobile Context for People who are Deaf and Hard of Hearing. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. 81–92.
[28] Takashi Kikuchi, Yuta Sugiura, Katsutoshi Masai, Maki Sugimoto, and Bruce H. Thomas. 2017. EarTouch: Turning the Ear into an Input Surface *(MobileHCI'17)*. ACM, New York, NY, USA, Article 27, 6 pages.
[29] Dietrich Klakow and Jochen Peters. 2002. Testing the correlation of word error rate and perplexity. *Speech Communication* 38, 1-2 (2002), 19–28.
[30] Knowles. 2020. Surface Mount MEMS Microphoness. https://www.knowles.com/subdepartment/dpt-microphones/subdpt-sisonic-surface-mount-mems. [Online; accessed 30-August-2020].
[31] Karly Kudrinko, Emile Flavin, Xiaodan Zhu, and Qingguo Li. 2020. Wearable Sensor-Based Sign Language Recognition: A Comprehensive Review. *IEEE Reviews in Biomedical Engineering* Early Access (2020), 1–15.
[32] Sen M. Kuo and Dennis R. Morgan. 1999. Active Noise Control: A Tutorial Review. *Proc. IEEE* 87, 6 (1999), 943–973.
[33] Haouyu Li, Hunglin Hsu, Liying Hu, Shufen Guo, and Rongjian Huang. 2017. Gesture control earphone. US Patent App. 15/125,002.
[34] Kang Ling, Haipeng Dai, Yuntang Liu, and Alex X Liu. 2018. Ultragesture: Fine-grained gesture sensing and recognition. In *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.
[35] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Yunfei Liu, and Minglu Li. 2018. LipPass: Lip reading-based user authentication on smartphones leveraging acoustic signals. In *Proceedings of the 2018 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 1466–1474.
[36] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. 2018. SignFi: Sign language recognition using WiFi. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1, Article 23 (2018), 21 pages.
[37] Wenguang Mao, Mei Wang, Wei Sun, Lili Qiu, Swadhin Pradhan, and Yi-Chao Chen. 2019. RNN-Based Room Scale Hand Motion Tracking. In *The 25th Annual International Conference on Mobile Computing and Networking (MobiCom'19)*. ACM, New York, NY, USA, Article 38, 16 pages.
[38] R Martin McGuire, Jose Hernandez-Rebollar, Thad Starner, Valerie Henderson, Helene Brashear, and Danielle S Ross. 2004. Towards a one-way American sign language translator. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 620–625.
[39] Dimitris Metaxas, Bo Liu, Fei Yang, Peng Yang, Nicholas Michael, and Carol Neidle. 2012. Recognition of Nonmanual Markers in American Sign Language (ASL) Using Non-Parametric Adaptive 2D-3D Face Tracking. In *Proceedings of the Eighth International Conference on*

Language Resources and Evaluation (LREC'12). 2414–2420.
[40] Christian Metzger, Matt Anderson, and Thad Starner. 2004. FreeDigiter: A contact-free device for gesture control. In *Eighth International Symposium on Wearable Computers*, Vol. 1. IEEE, 18–21.
[41] Riall Nolan. 1999. *Communicating and adapting across cultures: Living and working in the global village*. ABC-CLIO.
[42] Alex Olwal, Kevin Balke, Dmitrii Votintcev, Thad Starner, Paula Conn, Bonnie Chinh, and Benoit Corda. 2020. Wearable Subtitles: Augmenting Spoken Communication with Lightweight Eyewear for All-day Captioning. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20)*. 1108–1120.
[43] Jayshree R Pansare and Maya Ingle. 2016. Vision-based approach for American sign language recognition using edge orientation histogram. In *2016 International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 86–90.
[44] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Proc. Interspeech 2019* (Sep 2019).
[45] Dominique A Potvin, Kirsten M Parris, and Raoul A Mulder. 2011. Geographically pervasive effects of urban noise on frequency and syllable rate of songs and calls in silvereyes (Zosterops lateralis). *Proceedings of the Royal Society B: Biological Sciences* 278, 1717 (2011), 2464–2469.
[46] Junfu Pu, Wengang Zhou, and Houqiang Li. 2019. Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4165–4174.
[47] Purple. 2020. Purple Communications - On-site ASL Interpreting and VRI. https://signlanguage.com//. [Online; accessed 01-Feb-2021].
[48] Kun Qian, Chenshu Wu, Fu Xiao, Yue Zheng, Yi Zhang, Zheng Yang, and Yunhao Liu. 2018. Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices. In *Proceedings of the 2018 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 1574–1582.
[49] Mina I Sadek, Michael N Mikhael, and Hala A Mansour. 2017. A new approach for designing a smart glove for Arabic Sign Language Recognition system based on the statistical analysis of the Sign Language. In *34th National Radio Science Conference (NRSC)*. 380–388.
[50] Panneer Selvam Santhalingam, Al Amin Hosain, Ding Zhang, Parth Pathak, Huzefa Rangwala, and Raja Kushalnagar. 2020. mmASL: Environment-Independent ASL Gesture Recognition Using 60 GHz Millimeter-wave Signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1, Article 26 (2020), 30 pages.
[51] Baoguang Shi, Xiang Bai, and Cong Yao. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 11 (2016), 2298–2304.
[52] Xingzhe Song, Boyuan Yang, Ge Yang, Ruirong Chen, Erick Forno, Wei Chen, and Wei Gao. 2020. SpiroSonic: monitoring human lung function via acoustic sensing on commodity smartphones. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking (MobiCom'20)*. ACM, Article 52, 14 pages.
[53] Thad Starner. 2009. *Telesign: Towards a one-way American sign language translator*. Technical Report. Georgia Institute of Technology.
[54] Thad Starner, Joshua Weaver, and Alex Pentland. 1998. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 12 (1998), 1371–1375.
[55] David RB Stockwell and A Townsend Peterson. 2002. Effects of sample size on accuracy of species distribution models. *Ecological modelling* 148, 1 (2002), 1–13.
[56] USound. 2020. MEMS Microspeakers. https://www.usound.com/product/ganymede/. [Online; accessed 30-August-2020].
[57] B. Venema, J. Schiefer, V. Blazek, N. Blanik, and S. Leonhardt. 2013. Evaluating Innovative In-Ear Pulse Oximetry for Unobtrusive Cardiovascular and Pulmonary Monitoring During Sleep. *IEEE Journal of Translational Engineering in Health and Medicine* 1 (2013), 8.
[58] Christian Vogler and Dimitris Metaxas. 2003. Handshapes and movements: Multiple-channel american sign language recognition. In *International Gesture Workshop*. Springer, 247–258.
[59] Anran Wang and Shyamnath Gollakota. 2019. Millisonic: Pushing the limits of acoustic motion tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. 1–11.
[60] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW based contactless respiration detection using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 170.
[61] Yanwen Wang, Jiaxing Shen, and Yuanqing Zheng. 2020. Push the Limit of Acoustic Gesture Recognition. In *IEEE Conference on Computer Communications (INFOCOM'20)*. IEEE, 566–575.
[62] Yanwen Wang, Jiaxing Shen, and Yuanqing Zheng. 2020. Push the Limit of Acoustic Gesture Recognition. *IEEE Transactions on Mobile Computing* Early Access (2020), 1–14.
[63] Zhengjie Wang, Yushan Hou, Kangkang Jiang, Wenwen Dou, Chengming Zhang, Zehua Huang, and Yinjing Guo. 2019. Hand Gesture Recognition Based on Active Ultrasonic Sensing of Smartphone: A Survey. *IEEE Access* 7 (2019), 111897–111922.
[64] Wei Xu, ZhiWen Yu, Zhu Wang, Bin Guo, and Qi Han. 2019. Acousticid: gait-based human identification using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–25.
[65] Xuhai Xu, Haitian Shi, Xin Yi, WenJia Liu, Yukang Yan, Yuanchun Shi, Alex Mariakakis, Jennifer Mankoff, and Anind K. Dey. 2020. EarBuddy: Enabling On-Face Interaction via Wireless Earbuds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing*

Systems (CHI'20). 1–14.
[66] Hee-Deok Yang, Stan Sclaroff, and Seong-Whan Lee. 2008. Sign language spotting with a threshold model based on conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 7 (2008), 1264–1277.
[67] Yuancheng Ye, Yingli Tian, Matt Huenerfauth, and Jingya Liu. 2018. Recognizing american sign language gestures from within continuous videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2064–2073.
[68] Zahoor Zafrulla, Helene Brashear, Harley Hamilton, and Thad Starner. 2010. A novel approach to american sign language (asl) phrase verification using reversed signing. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 48–55.
[69] Zahoor Zafrulla, Helene Brashear, Thad Starner, Harley Hamilton, and Peter Presti. 2011. American sign language recognition with the kinect. In *Proceedings of the 13th International Conference on Multimodal Interfaces*. 279–286.
[70] Zahoor Zafrulla, Helene Brashear, Pei Yin, Peter Presti, Thad Starner, and Harley Hamilton. 2010. American sign language phrase verification in an educational game for deaf children. In *2010 20th International Conference on Pattern Recognition*. IEEE, 3846–3849.
[71] Zahoor Zafrulla, Himanshu Sahni, Abdelkareem Bedri, Pavleen Thukral, and Thad Starner. 2015. Hand detection in American Sign Language depth data using domain-driven random forest regression. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 1. IEEE, 1–7.
[72] Cheng Zhang, Qiuyue Xue, Anandghan Waghmare, Ruichen Meng, Sumeet Jain, Yizeng Han, Xinyu Li, Kenneth Cunefare, Thomas Ploetz, Thad Starner, Omer Inan, and Gregory D. Abowd. 2018. FingerPing: Recognizing Fine-Grained Hand Poses Using Active Acoustic On-Body Sensing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. 1–10.
[73] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2019. MyoSign: enabling end-to-end sign language recognition with wearables. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 650–660.
[74] Yongzhao Zhang, Wei-Hsiang Huang, Chih-Yun Yang, Wen-Ping Wang, Yi-Chao Chen, Chuang-Wen You, Da-Yuan Huang, Guangtao Xue, and Jiadi Yu. 2020. Endophasia: Utilizing Acoustic-Based Imaging for Issuing Contact-Free Silent Speech Commands. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–26.
[75] Tianming Zhao, Jian Liu, Yan Wang, Hongbo Liu, and Yingying Chen. 2018. PPG-based Finger-level Gesture Recognition Leveraging Wearables. In *International Conference on Computer Communications (INFOCOM)*. IEEE, 1457–1465.
[76] Tianming Zhao, Jian Liu, Yan Wang, Hongbo Liu, and Yingying Chen. 2020. Towards Low-cost Sign Language Gesture Recognition Leveraging Wearables. *IEEE Transactions on Mobile Computing* Early Access (2020), 1–16.

A APPENDIX

A.1 Gestures and Related Spectrograms

We provide all gestures instructions [14] and the corresponding spectrogram features of all 42 selected ASL signs, 16 similar signs, and two start and end gestures.

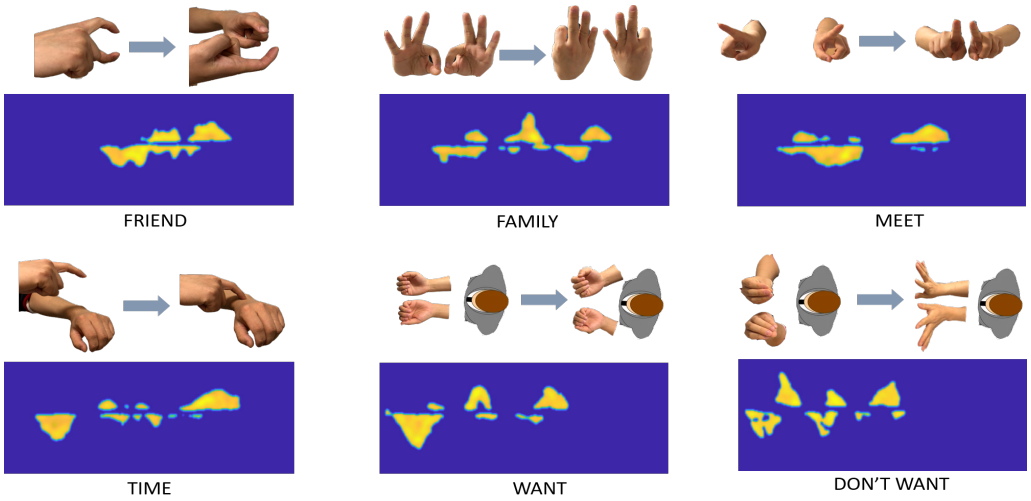


Fig. 14. The 1st part gestures and corresponding spectrograms

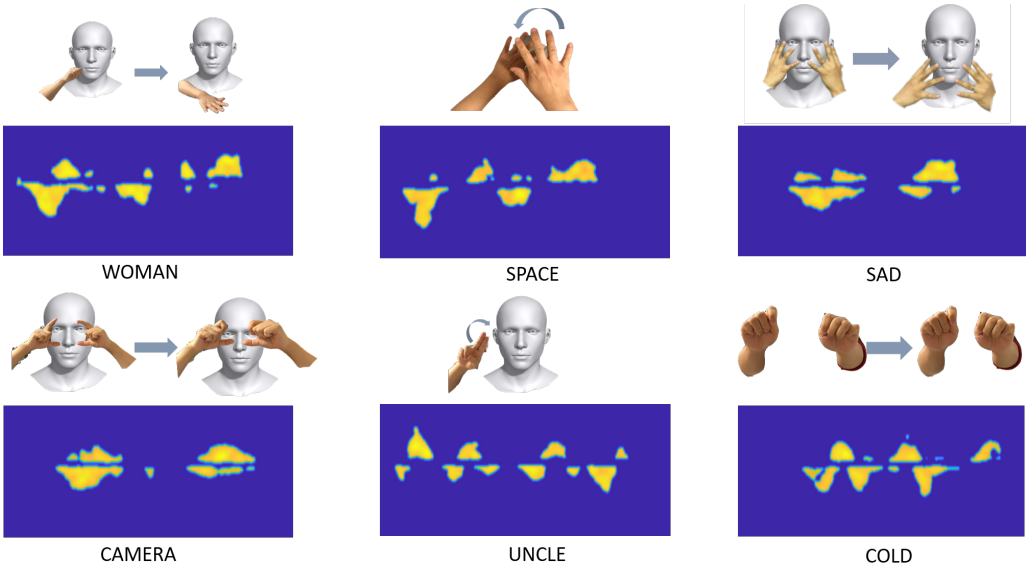


Fig. 15. The 2nd part gestures and corresponding spectrograms

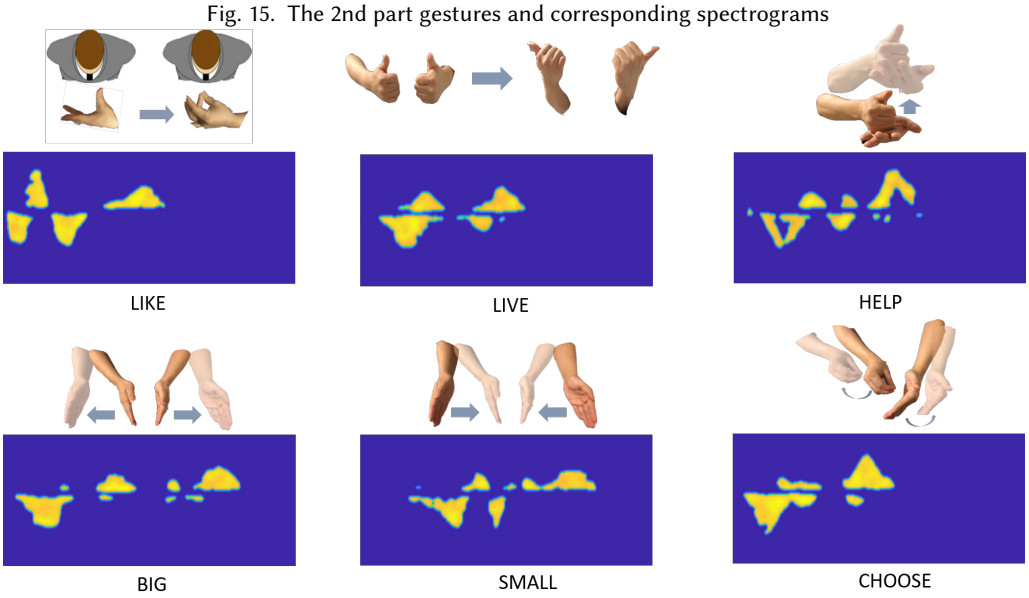


Fig. 16. The 3rd part gestures and corresponding spectrograms

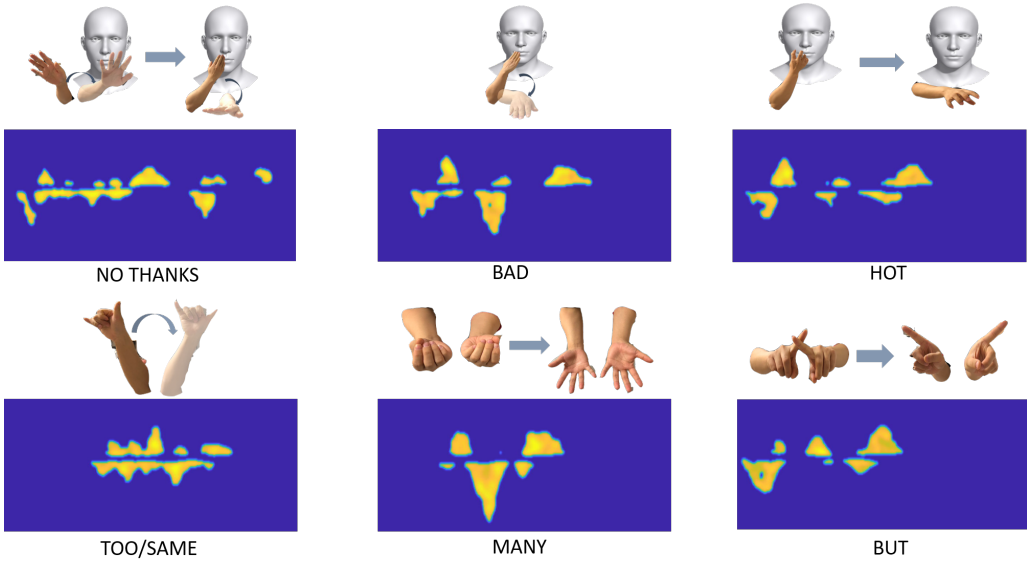


Fig. 17. The 4th part gestures and corresponding spectrograms

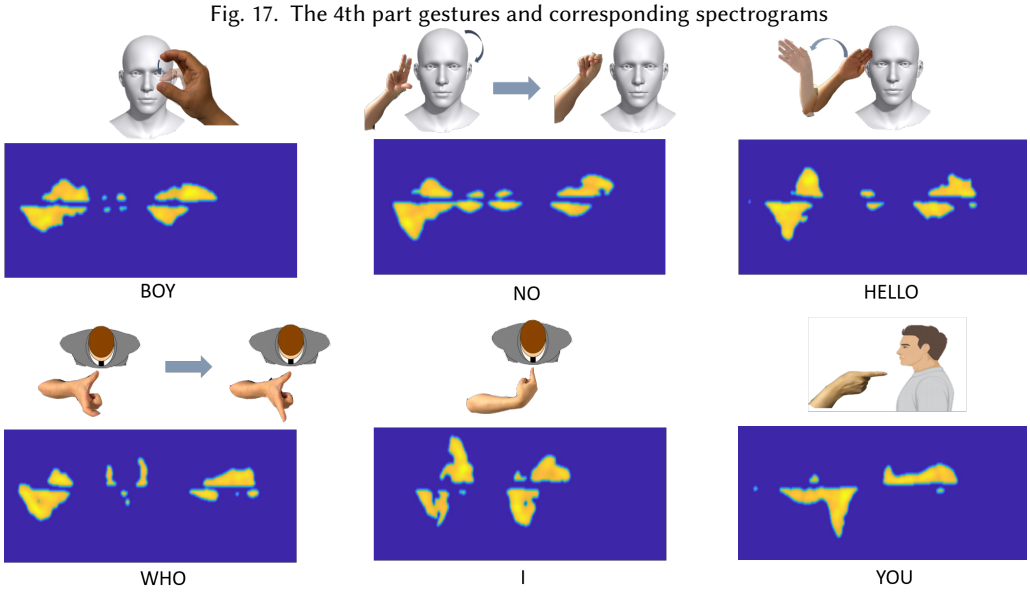


Fig. 18. The 5th part gestures and corresponding spectrograms

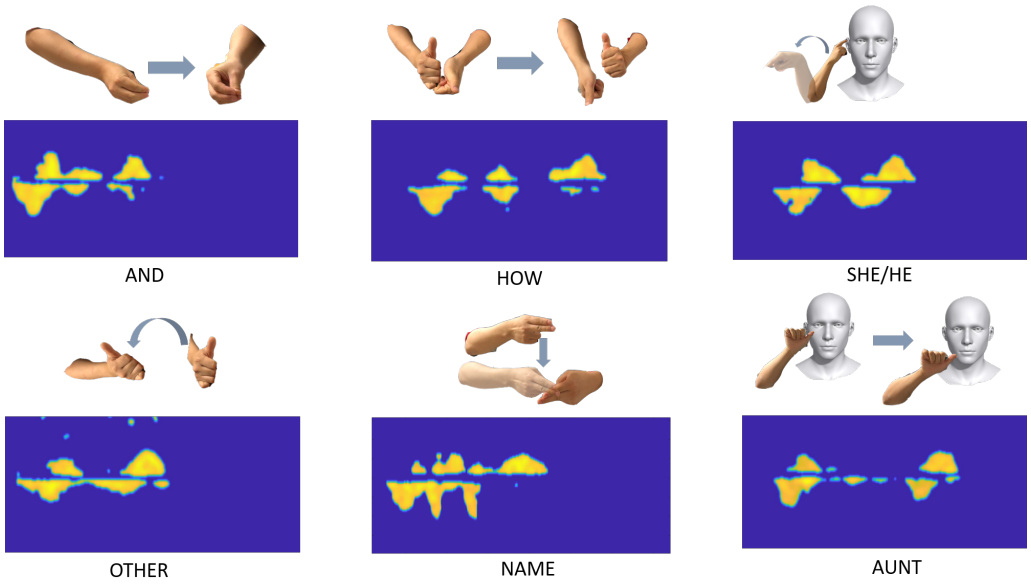


Fig. 19. The 6th part gestures and corresponding spectrograms

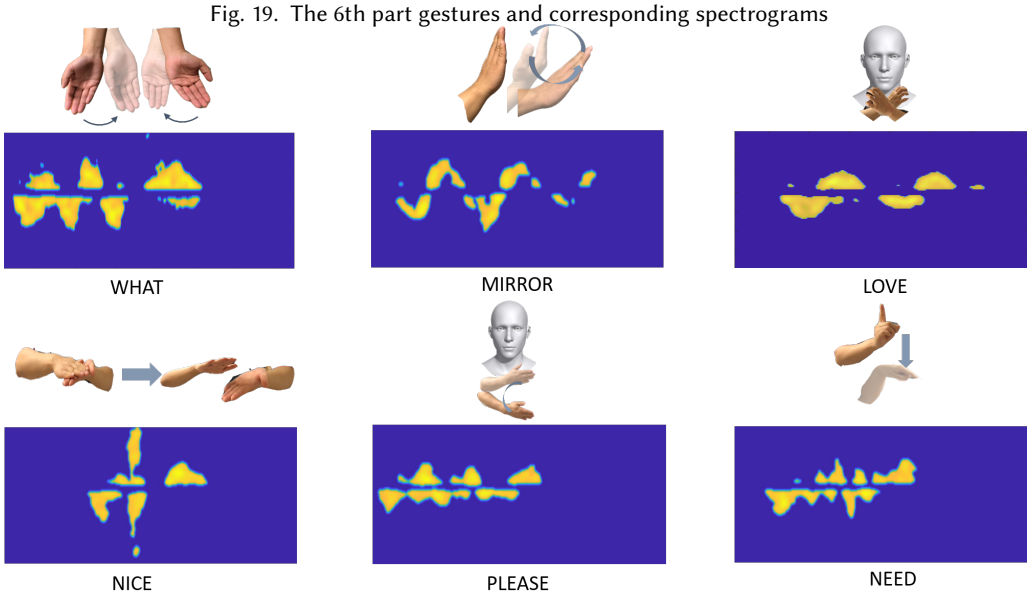


Fig. 20. The 7th part gestures and corresponding spectrograms

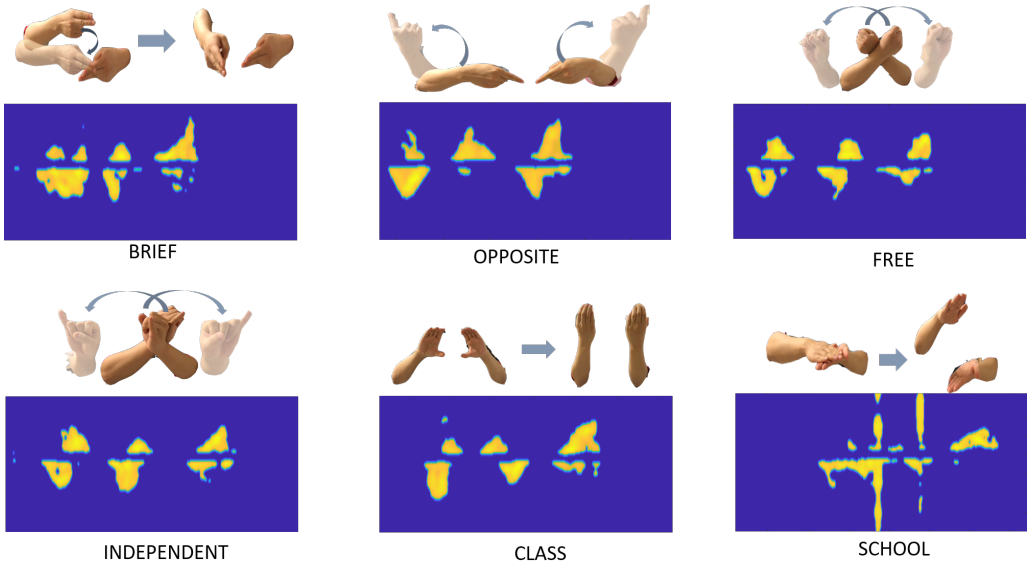


Fig. 21. The 1st part gestures for similar words and corresponding spectrograms

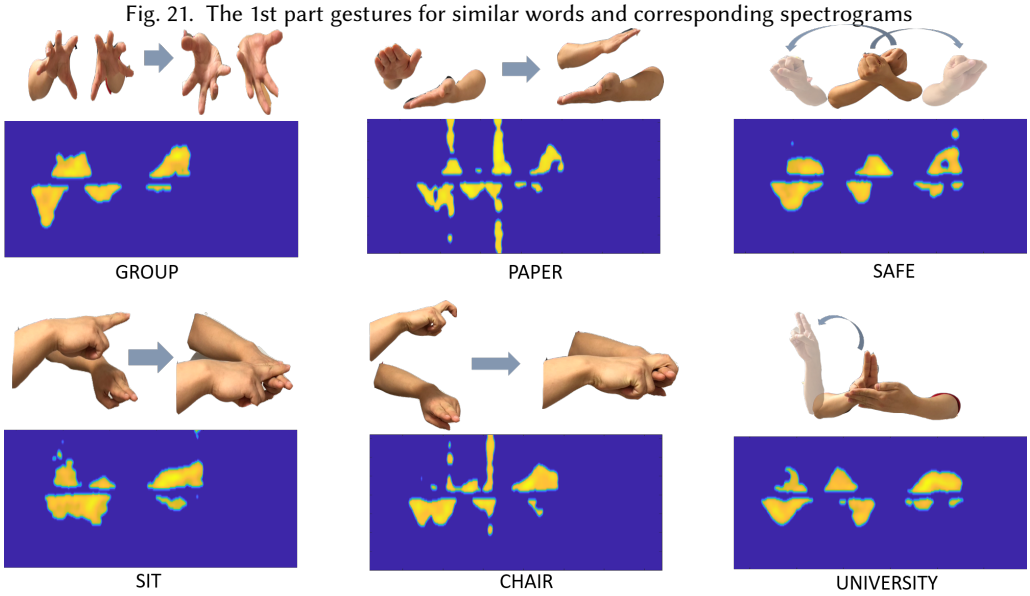


Fig. 22. The 2nd part gestures for similar words and corresponding spectrograms

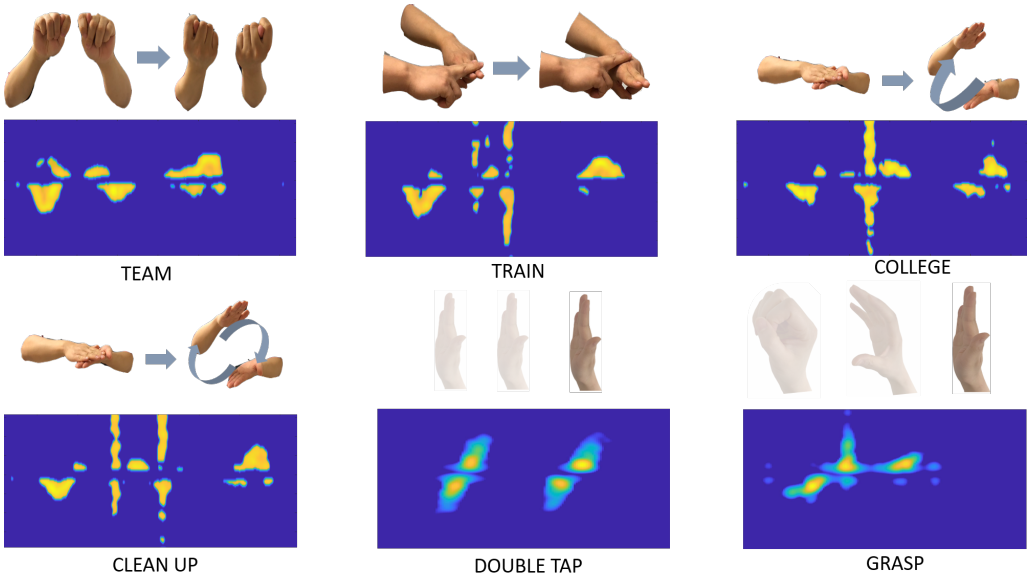


Fig. 23. The 3rd part for similar words and start (double tap) - end (grasp) gestures and corresponding spectrograms