SmartASL: "Point-of-Care" Comprehensive ASL Interpreter Using Wearables

YINCHENG JIN, University at Buffalo, USA
SHIBO ZHANG, HP Inc., USA
YANG GAO, East China Normal University, China
XUHAI XU, University of Washington, USA
SEOKMIN CHOI, University at Buffalo, USA
ZHENGXIONG LI, University of Colorado Denver, USA
HENRY J. ADLER, University at Buffalo, USA
ZHANPENG JIN*, South China University of Technology, China and University at Buffalo, USA

Sign language builds up an important bridge between the d/Deaf and hard-of-hearing (DHH) and hearing people. Regrettably, most hearing people face challenges in comprehending sign language, necessitating sign language translation. However, state-of-the-art wearable-based techniques mainly concentrate on recognizing manual markers (e.g., hand gestures), while frequently overlooking non-manual markers, such as negative head shaking, question markers, and mouthing. This oversight results in the loss of substantial grammatical and semantic information in sign language. To address this limitation, we introduce SmartASL, a novel proof-of-concept system that can 1) recognize both manual and non-manual markers simultaneously using a combination of earbuds and a wrist-worn IMU, and 2) translate the recognized American Sign Language (ASL) glosses into spoken language. Our experiments demonstrate the SmartASL system's significant potential to accurately recognize the manual and non-manual markers in ASL, effectively bridging the communication gaps between ASL signers and hearing people using commercially available devices.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: Comprehensive ASL Recognition, Smartwatch, Earbuds, Manual Markers, Non-manual Markers

ACM Reference Format:
Yincheng Jin, Shibo Zhang, Yang Gao, Xuhai Xu, Seokmin Choi, Zhengxiong Li, Henry J. Adler, and Zhanpeng Jin. 2023. SmartASL: "Point-of-Care" Comprehensive ASL Interpreter Using Wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 60 (June 2023), 21 pages. https://doi.org/10.1145/3596255

*This is the corresponding author.

Authors' addresses: Yincheng Jin, University at Buffalo, Department of Computer Science and Engineering, Buffalo, NY, 14260, USA; Shibo Zhang, HP Inc., USA; Yang Gao, East China Normal University, Department of Computer Science, China; Xuhai Xu, University of Washington, Information School, Seattle, Washington, USA; Seokmin Choi, University at Buffalo, Department of Computer Science and Engineering, USA; Zhengxiong Li, University of Colorado Denver, USA; Henry J. Adler, University at Buffalo, Department of Communicative Disorders and Sciences, USA; Zhanpeng Jin, South China University of Technology, School of Future Technology, China and University at Buffalo, Department of Computer Science and Engineering, USA, zjin@scut.edu.cn.

2474-9567/2023/6-ART60 $15.00
https://doi.org/10.1145/3596255

1 INTRODUCTION

With the aging population and the shifting lifestyles in modern society, it is reported that more than 400 million people suffer from hearing loss [56]. Sign languages provide effective solutions for communication among d/Deaf and hard-of-hearing (DHH) people [11], becoming the primary language in the DHH community. In particular, American Sign Language (ASL) is considered one of the most widely used and predominant sign languages across the globe [25]. As a comprehensive and structured visual language, ASL is articulated through manual markers, such as hand shapes and movements to convey basic information, along with non-manual markers, which include non-affective facial expressions and head/body postures to form grammatical structure, adjectival or adverbial content [36, 54]. Without non-manual markers, ASL signers can barely create a comprehensible ASL construction, such as expressing negation, forming question expressions, and employing topicalization [12]. For example, comparing the ASL word "late" and "Not-Yet", the only difference is that "Not-Yet" is produced with the required non-manual marker - "TH" [48].

Professional ASL interpreters can help communication between d/Deaf and hearing people because hearing people rarely understand ASL. However, professional ASL interpreters are costly and limited in many regions. Also, professional interpreters are rarely available and accessible for impromptu applications. A straightforward solution is to utilize portable ASL translation devices to help d/Deaf people communicate with hearing people. Existing ASL recognition systems via different sensing technologies can be generally divided into three categories — camera-based methods [32, 43], wireless signal-based methods [25, 44, 46], and mobile/wearable-based methods [21, 25, 31, 33, 44, 57, 62]. Among them, only camera-based solutions can provide the recognition of manual and non-manual markers at the same time [32, 43]. However, they are often restricted from being deployed in real-world scenarios, due to severe privacy concerns of the DHH community when facing and recording the facial/body videos of the ASL signer, as well as the challenges of addressing the influence of noises such as diverse backgrounds and camera viewpoints. On the other side, signal-based and mobile/wearable-based ASL recognition systems have recently gained wide attention due to their portability and privacy-preserving nature [21, 25, 60]. However, these solutions have mainly focused on manual markers, neglecting non-manual markers. As we mentioned above, non-manual markers are an essential aspect of ASL, and neglecting them would lose a substantial portion of essential linguistic information. Hence, it is imperative to seek an effective approach to recognize both the manual markers and non-manual markers simultaneously leveraging the signal-based and mobile/wearable-based platforms, during the daily communications between d/Deaf and hearing people.

Besides, ASL has its own linguistic features and grammatical structures, which are different from spoken/written languages. In addition, ASL has its own independent lexicon, grammar, and syntax, which present significant obstacles for hearing people in understanding ASL glosses [60]. Currently, existing works translate the sensor data into spoken English texts directly [60]. However, the high complexity of the neural network model proposed in [60] renders the training process very difficult to converge, undermining the practicability of the proposed method especially when an expansion of the dataset is demanded. Thus, it is better to adopt the cascaded mechanism to translate recognized glosses into spoken/text English by fine-tuning the current successful natural language processing (NLP) model.

Motivated by the pros and cons of existing ASL recognition methods, in this study, we propose and develop SmartASL, an affordable, off-the-shelf, user-friendly proof-of-concept system to achieve end-to-end ASL recognition which shows great potential to facilitate communications between ASL signers and hearing people. As shown in Fig. 1 (a), we adopt the smartwatch as a vehicle to recognize manual ASL components using built-in IMU sensors data, and utilize the built-in IMU sensors on the earbuds to detect head/facial movements to recognize non-manual markers in ASL. Combining these two modalities, we can obtain ASL glosses from the signer's signs (i.e., ASL sign gesture word labels). In addition, because ASL has its own linguistic features and grammatical structures [60], we propose to fine-tune a pre-trained natural-language-processing (NLP) model to translate ASL

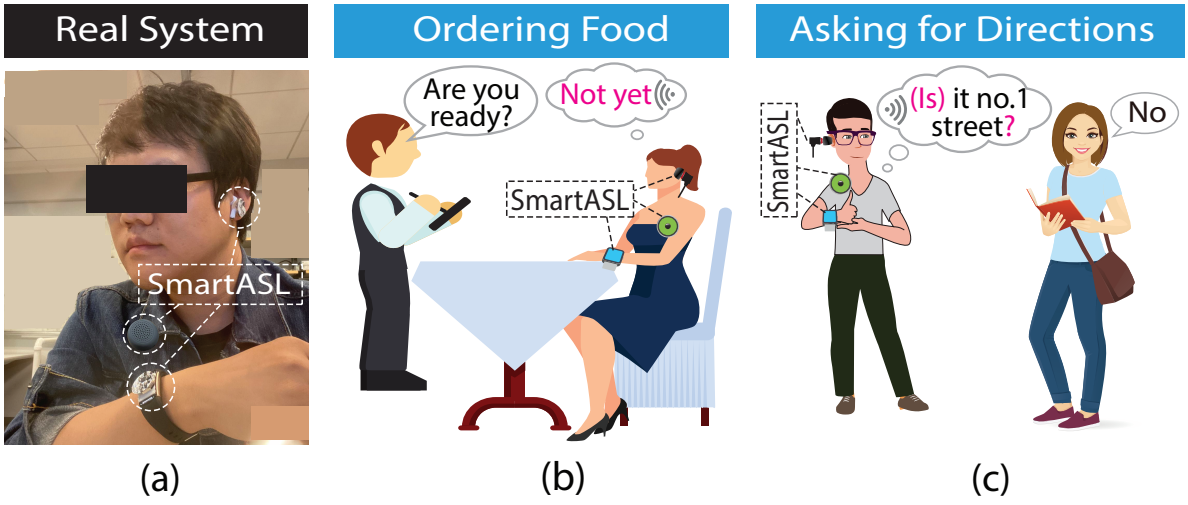

Fig. 1. (a) A user wears the real SmartASL system, a smartwatch for manual markers recognition, a pair of earphones for non-manual markers recognition, and a wearable speaker for playing these recognized words, (b) One potential application scenario, the user gives the response 'Not-Yet', which includes the required non-manual marker - "TH", (c) Another potential application scenario, the user asks a question, which includes the required non-manual marker - "raised brows, widened eyes and forward-tilted head".

glosses into comprehensible spoken texts. As illustrated in Fig. 1(b) and (c), SmartASL could be a convenient and accessible approach assisting with daily causal communication between the signer and hearing people, when the ASL interpreter is not available and exchanging written notes is discouraged. The contributions of our work can be summarized as follows:

- We propose that SmartASL, the first proof-of-concept wearable ASL system, can recognize both manual markers (e.g., hand movements) using a smartwatch, and non-manual markers (e.g., question markers, negative meanings, and non-affective facial expressions) using a pair of earbuds, simultaneously. Our proposed SmartASL system demonstrates promising recognition performance.
- We develop and fine-tune a well-pre-trained transformer-based T5-small model to translate the recognized ASL glosses into spoken English. It can bridge the gap between the recognized ASL glosses and spoken English.
- We develop and implement a portable prototype to showcase the end-to-end pipeline of our system. Our system consists of three IMU sensors, a Raspberry Pi board, and a wearable speaker that audibly delivers the translated English. Additionally, we investigate the user experience by conducting experiments with our portable SmartASL prototype.

2 RELATED WORK

In this section, we begin by introducing the recognition of manual markers, followed by an explanation of the recognition of non-manual markers. Lastly, we discuss the related work in the field of earable computing.

2.1 Manual Markers Recognition in ASL

It has been a long-standing challenge in ASL research to accurately capture and recognize manual markers using various computing techniques. Existing hand gesture recognition solutions can be generally divided into

three groups, camera-based recognition [43], wireless signal-based recognition [25, 44], and wearable sensor-based recognition [21, 60]. Along with the recent advancements in deep learning and exponential growth of computing power, camera-based hand gesture recognition solutions [13, 22, 59], which utilize the technologies including Hierarchical Attention Network or Transformer, have been well explored in the recent decade. However, rapidly growing privacy and ethical concerns about camera video recording have restricted the wide adoption of such vision-based solutions in real-world scenarios. Furthermore, wireless signal-based ASL recognition (e.g., acoustic signal [25], WiFi signal [46], and millimeter wave [44]) has been widely investigated because of their unique advantages in terms of privacy protection and non-contact remote sensing. However, the applicability of wireless signal-based recognition techniques is limited by environmental dependency and interferences. Besides, wearable sensor-based hand gesture recognition (e.g., IMU sensors [30], EMG sensors [4], or multi-modal sensors [21, 57, 60]), is another well-known approach widely adopted in daily life because of its accessibility and portability. In particular, among all existing approaches, SignSpeaker [21] utilizes a CRNN model to detect 103 word-level and 73 sentence-level ASL expressions using IMU data collected from a smartwatch. This SignSpeaker system shares a similar goal with our SmartASL in recognizing manual markers using the smartwatch. In addition, some prior works using smartwatches (e.g., WearSign [60], MyoSign [61]) successfully proved the feasibility of recognizing ASL hand gestures using only a smartwatch with IMU sensors. Thus, to make SmartASL more affordable and accessible for users in daily-life scenarios, we adopted a similar method that utilizes a smartwatch with built-in IMU sensors to recognize the manual markers in ASL.

2.2 Non-Manual Markers Recognition in ASL

As introduced in Section 1, non-manual markers consist of head shaking, head tilt, facial gestures, and mouth morphemes. In this section, we introduce related work focusing on recognizing non-manual markers in ASL. Previously, camera-based non-manual marker recognition systems in ASL have been widely explored. For example, Michael *et al.* [58] proposed a camera-based ASL recognition system by utilizing the Hidden Markov Support Vector Machine (HMSVM) to recognize non-manual markers. However, it is well acknowledged that camera-based non-manual marker recognition raises privacy and ethical concerns for users. Therefore, several non-contact wireless signal-based facial expression sensing technologies have been proposed. For example, Gao *et al.* [16] proposed an acoustic-based sensing technology to recognize emotional facial expressions which provides the possibility of non-camera-based facial gesture detection adopting a multi-view CNN structure. Unfortunately, although providing a privacy-preserving detection mechanism, such approaches suffer from the limitations of environmental and position dependency, which means they can hardly be used in a different environment or position. To enhance portability, wearable systems have also been examined to sense facial expressions. For example, Verma *et al.* [52] proposed the ExpressEar system to capture facial gestures by using IMU sensors embedded in the earbuds utilizing a basic CNN structure. Inspired by those prior scientific explorations, in this work, we propose to utilize commercial earbuds with an embedded IMU sensor to recognize different non-manual markers including head motions, mouth morphemes, and facial gestures.

2.3 Earable Computing

In the past decade, the market of earbuds and earphones has been rapidly expanding and is expected to grow at a compound annual growth rate (CAGR) of 20.3% from 2020 to 2027 [41]. Multiple earable sensing-based applications span from authentication to human sensing[42], such as head gesture recognition [5], authentication [14, 17, 53], blood flow [29], ASL recognition [25], blood pressure [7], heart rate [8], navigation [2], hearing loss detection [23], and facial gestures detection [3, 10, 47, 52]. For example, Bui *et al.* proposed a customized device to measure blood pressure from inside the user's ear by leveraging the light-based pulse sensor. However, such systems rely on high-cost, customized devices that are hard to deploy in daily life. Another example is an

earphone-based ASL recognition system — SonicASL [25], which leverages the acoustic signals to recognize 42 word-level hand gestures and 30 sentence-level hand gesture groups by deploying a CRNN neural network structure. In addition, Jin *et al.* [24] and Wang *et al.* [53] proposed their own systems that utilize a pair of commercial inward-facing microphone and speaker to detect facial gestures and mouth movements based on the structural changes of the ear canal. Besides, for the DHH community, the market of hearing aids is growing at a considerable level [19, 28], by amplifying the target sound. Except for this fundamental hearing aid function [19], some novel systems have been deployed in earable systems to help the DHH community, such as for navigation [35], sign language recognition [25], or speech recognition [18]. Since earable devices have been widely accepted in the DHH community, we will deploy IMU sensors on earable devices to recognize the non-manual markers in our study.

3 DESIGN CONSIDERATIONS

In this section, we first elaborate on the rationale and necessity of recognizing both manual and non-manual markers in ASL. Then, we present the rationale that the smartwatch is capable of capturing manual markers, and the earphone is able to recognize non-manual markers. Furthermore, given the complexity of ASL, we will provide the principles of ASL understanding and speech translation.

3.1 Combination and Interaction of Manual Markers and Non-manual Markers in ASL

As discussed in Section 1, a complete sign must be described using both manual markers (e.g., hand shape, movement, palm orientation, and hand location) and non-manual markers (e.g., non-affective facial expressions, head tilting, and body shifting) simultaneously. It has been widely shown in literature [9, 26, 49, 55, 61] that, without non-manual markers, it is impossible to create a truly comprehensible construction in ASL by manual-only signs [54]. As a consequence, the goal of our proposed SmartASL system is not only to recognize different manual markers, but also to distinguish similar manual markers with distinct non-manual markers with convenient wearable devices. For instance, the ASL word "Late" and "Not-Yet" have similar manual markers, but different non-manual markers. As shown in Fig. 2 [48], SmartASL targets recognizing them without any contextual environment.

3.2 Rationale of Manual and Non-manual Markers Recognition in ASL

According to the National Institute of Health (NIH), 28.8 million U.S. adults utilize hearing aids in their daily life [37], demonstrating the widespread acceptance of these devices within the DHH community. Additionally, earbud-type hearing devices with built-in IMU sensors offer a privacy-preserving, unobtrusive, user-friendly method for capturing non-affective facial gestures and head motions [6, 52], which are crucial components in constructing non-manual markers in ASL [36, 54]. As a result, our goal is to utilize earbud-type earable devices to detect non-manual markers by integrating low-cost IMU motion sensors. Besides, the smartwatch has been investigated and utilized for manual markers recognition in recent years [21, 60]. Following a similar approach, we also employ a smartwatch with IMU sensors to recognize and identify manual markers in our work. In this section, we will discuss the rationale and significance of recognizing non-manual and manual markers using earbuds and smartwatches, respectively.

3.2.1 Rationale of Non-Manual Markers Using Earbuds with IMU Sensors. It is well acknowledged that IMU sensors attached to the head can be employed to recognize head/body movements. Moreover, a prior study [52] showed that a pair of earbuds with IMU sensors could be effective in capturing diverse facial expressions by sensing ear canal deformation. Behavior scientists (e.g., Carl-Herman *et al.* [20], Friesen and Ekman *et al.* [15]) have developed the Facial Action Coding System (FACS), which is based on the anatomy of the human face, to encode the slight movements of individual facial muscles. Action units (AUs) are the fundamental actions of

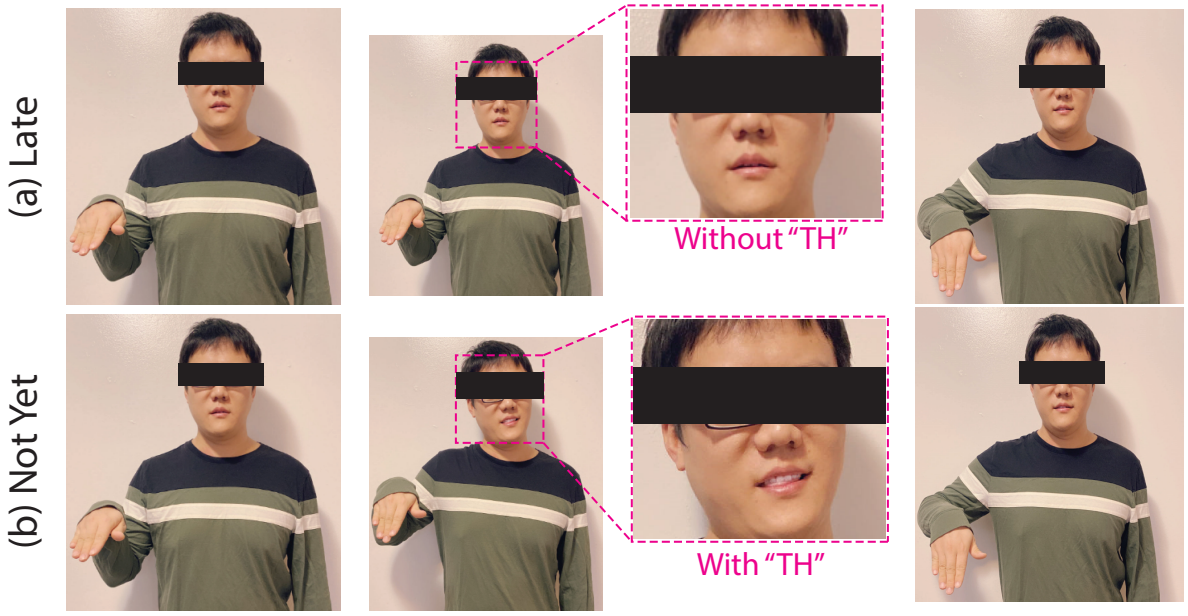

Fig. 2. Examples showing that manual markers and non-manual markers together construct comprehensible ASL. (a) ASL word of "Late". (b) ASL word of "Not-Yet". The only difference is that "Not-Yet" is produced with the required non-manual marker - "TH". (The pictures are excerpted from the online ASL dictionary [48].)

individual facial muscles or groups of muscles. In total, there are 30 AUs related to the contractions of specific facial muscles, 12 for the upper face and 18 for the lower face. Thus different facial expressions can be represented by single or multiple AUs together. As shown in Fig. 3, head movements (e.g., platysma, sternocleidomastoid), facial expressions (e.g., Masseter, Buccinator), and lip motions (e.g., Zygomaticus Minor, Zygomaticus Major) are connected to TMJ. When people are executing non-manual markers, the facial muscle and bone movements will lead to the movements of TMJ, further leading to ear canal deformation. Thus, we deployed the earbuds to capture the richness of non-affective facial expressions by sensing the ear canal deformation. In a nutshell, the earbuds with attached IMU sensors in the SmartASL system can be deployed to recognize the non-manual markers, encompassing head and body movements and facial expressions.

3.2.2 Rationale of Manual Markers Using Smartwatches with IMU Sensors. The IMU data associated with each ASL expression exhibits its own movements in spatial environments. Thus, it is essential to utilize a sensor that can track the 3D space to distinguish unique ASL expressions. Given that the acceleration data's x, y, and z-axis correspond to 3D space and are widely used for hand gesture (i.e., manual markers in ASL) recognition [21, 60, 61], we also utilize the accelerometer on the smartwatch to sense hand gestures to track the ASL manual markers.

3.3 Conversion from ASL Glosses to Spoken English

American Sign Language has its own linguistic rules. For example, "What's your name" in English should be expressed as "Your name what" in ASL, which means different word sequences and some missing words (e.g., "be verbs," "articles," "auxiliary verb"). It is rather difficult for hearing people to understand ASL correctly based on some basic ASL glosses, as shown in Fig. 4. Therefore, to meet the need for an end-to-end ASL translation solution,

we proposed a new approach to translating ASL glosses into spoken English. Over the past few years, given the success of transfer learning, a variety of natural language processing (NLP) tasks can achieve state-of-the-art results. The basic idea of Text-To-Text Transfer Transformer (T5) architecture [40] is based on a two-stage training pipeline comprising (a) pre-training a model on a large dataset and (b) fine-tuning the model on a smaller and specialized dataset. This methodology is well aligned with the goal of ASL translation and thus adopted to translate the ASL glosses into spoken English by fine-tuning this model with our created text data.

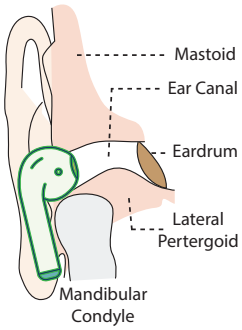

Fig. 3. Rationale of recognizing ear canal deformation using earbuds.

Fig. 4. Conversion from recognized ASL gloss to spoken English

4 SYSTEM OVERVIEW

SmartASL is a novel system that can recognize manual and non-manual markers at the same time, and then translate the recognized ASL glosses into spoken English. As shown in Fig. 5, SmartASL consists of four steps to generate spoken English. Firstly, the user wears a pair of earbuds and a smartwatch, which collects IMU data while the user is performing ASL expressions; the data will be transmitted to terminal devices, such as smartphones, laptops or Raspberry Pi, for further processing and analysis. Secondly, recognition modules for both non-manual and manual markers will execute the signal processing, feature extraction, and classification separately to obtain basic ASL glosses. Afterward, SmartASL concatenates and feeds the recognized ASL glosses into a fine-tuned T5 model to translate the ASL glosses into spoken English text. Finally, the terminal device will either display the spoken English text or play out the speech via a wearable speaker to allow hearing people to properly comprehend the signer's ASL expressions.

4.1 Non-manual Markers Recognition

As introduced in Section 3.2.1, non-manual markers could involve subtle, fine-grained facial expressions and/or significant, coarse-grained head movements. To this end, our approach is motivated by the rationale that fine-grained facial expressions in ASL (e.g., "MM", "CHA") would result in small-scale deformations of the ear canal geometry, while coarse-grained head motions (e.g., "NOT", "Question") lead to more considerable changes of the head position.

Challenge: The main challenge of sensing non-manual markers in ASL recognition is that the morphological characteristics of non-manual markers reflected in IMU signals data are hidden in the data of head motions.

Solutions: To address this challenge, we adopt an optimized signal pre-processing solution to eliminate the side effect of head/body motions, then utilize a CNN model to sense and recognize non-manual markers through the ear canal deformation.

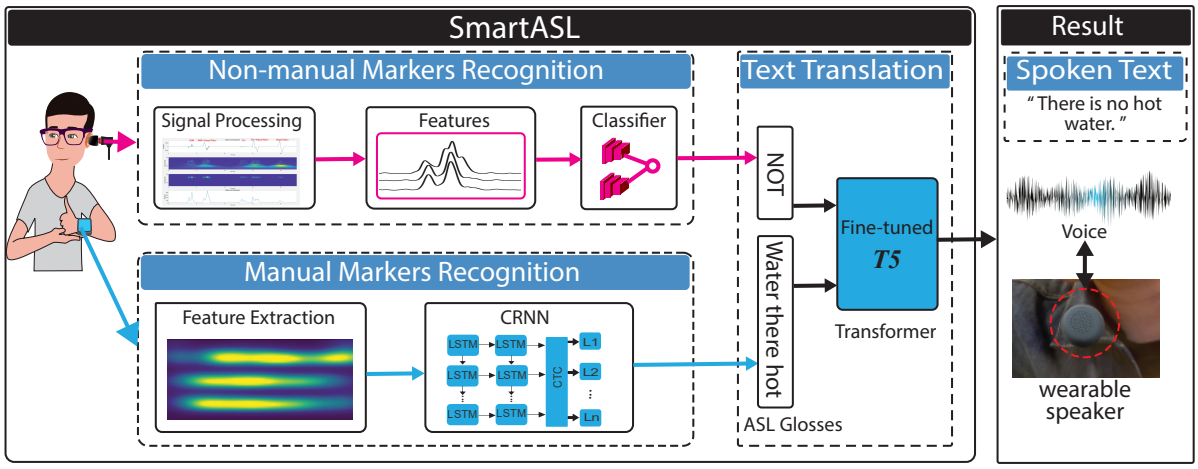

Fig. 5. A flow diagram shows the processing mechanism of SmartASL, which consists of a non-manual markers recognition block, and a manual markers recognition block, followed by a gloss-to-text translation block. Then, the recognized texts can be played by a wearable speaker.

4.1.1 Signal Processing for Facial Gestures. Because the displacements caused by ear canal deformations are orders of magnitude smaller than position variations resulting from unintentional head/body movements, the IMU signals are thus dominated by intentional/unintentional head or body movements (e.g., negative statements, question statements, and random head/body motion artifacts). However, through our observation, it is witnessed that the acceleration of unintentional head movements is smaller than facial expressions when signing ASL. This is because when exhibiting facial expressions, the head is usually moving slowly and steadily. As shown in Fig. 6, the frequency component corresponding to unintentional head motions has a lower frequency than that of facial gestures. Thus, we can magnify the elements of facial expressions and remove the components of unwanted head motions by using a low-pass filter. Since facial expressions and facial movements are highly dynamic, we need to achieve a high level of resolution in both the time and frequency domains to accurately capture corresponding information. This essentially requires SmartASL to properly mitigate the head movements from facial expressions and facial movements, as shown in Fig. 6. Thus we first utilized CWT to plot the time-frequency domain features. Then, we set a CWT-dependent threshold to remove unintentional head movements. Lastly, we utilized i-CWT to reconstruct the signal. The figure shows that our proposed method can effectively mitigate noisy head movements to magnify the facial expression signals. It can be observed that the facial expression - "MM" is different from the combined signal of head movements and the same facial expression -"MM" in Fig. 6. Besides, the reconstructed signal (after removing the component of noisy head movements) shows similar morphological patterns as the original one. As shown in Fig.6, some unique signal patterns continue after the coarse head movement signals. Thus, we utilize the same process to obtain final patterns for non-manual based head/body movements (e.g., negative expression with a head shake).

4.1.2 Feature Structure. After signal processing steps, we can obtain six channels of the optimized IMU data (i.e., acc-x, acc-y, acc-z, gyro-x, gyro-y, gyro-z) from either ear (i.e., left and right ear) at iCWT phase as shown in Fig. 6. After obtaining the optimized signal, we need to feed it to the neural network system. However, since the input length of each individual ASL word or sentence is variable, and the neural network requires a consistent input size, we resize each channel of the input signal to 512 bits. To recognize the various patterns of different

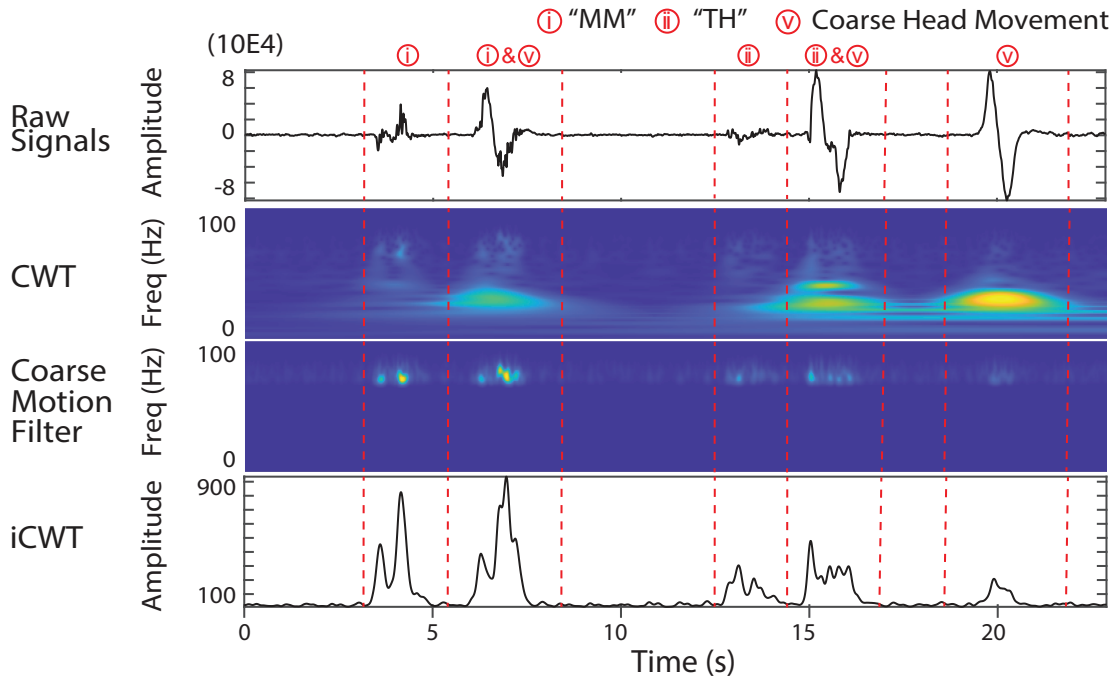

Fig. 6. An example shows the elimination of unintentional coarse-grained head movements from facial expressions. 'MM' and 'TH' represent two different facial expressions. Head movement means turning the head left and back without facial expressions.

facial gestures, we extract the iCWT data as the features and form a 6-dimensional data structure for each ear (i.e., 512×6 for each ear). Then, this data is fed into the following neural network.

4.1.3 Classifier. The extracted and preprocessed IMU data were fed into a specially designed classifier to recognize and distinguish different facial gestures. Motivated by the superior capability of convolutional neural networks (CNN) in extracting, representing, and capturing subtle characteristics, we utilized the CNN classifier for classifying the features. Besides, to reduce interference from different ears, we utilize a multi-view-based classifier that consists of two same CNN structures (i.e., each CNN was fed into an input with structure of 512×6). And then we combined these two output vectors and linked them to a softmax layer to classify seven different non-manual markers. Both CNN structures consist of three 2D convolutional layers with 128 units, 256 units, and 128 units, respectively, and each layer includes the ReLU activation function.

4.2 Manual Markers Recognition

4.2.1 Feature Extraction. Hand gesture-related IMU data is collected by the smartwatch from the dominant hand and then upsampled to 256 Hz (the actual sampling frequency varies from 100 Hz to 150 Hz). Time-variant features can provide time-series-related features, and frequency-domain features are more reliable to be recognized. To represent different hand gestures using three channels (i.e., x, y, and z-axis), we leverage Short-Time Fourier Transform (STFT) spectrogram as a feature to recognize different hand gestures, which the Power Spectral

Density of the function can be expressed as follows:

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

where $x[n]$ is the 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:

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

Considering the size of images and the complexity of neural network models, we concatenate three spectrograms along with the same time axis to form a single feature representation input as shown in the bottom half of Figure 12(a). It is observed that there are three high-energy areas that represent the motions along with different spatial dimensions. Thus, the concatenation is effective and efficient in representing different motions in any direction of space. The concatenation order is d_x, d_y, d_z, where d represents the hand data.

4.2.2 Manual Markers Recognition Classifier. For an ASL translation system, it is necessary to support isolated ASL words and sentences. As convolutional neural networks (CNN) exhibit superior representation capability in image classification tasks, we propose to employ a CNN to serve as the front end to embed the spectrogram images into vectors for word or sentence recognition. In our work, we designed a 2D-CNN network composed of six 2D convolutional layers (i.e., $64 \times 128 \times 256 \times 256 \times 512 \times 512$) with the ReLU activation function.

Besides, LSTM has been proven to be effective in handling time-series data by learning temporal dependencies of meaningful features. As introduced above, STFT is a kind of time-series data and can be used to determine the sequential dependencies between words and phrases in both the forward and backward directions. Therefore, two LSTMs are placed following the previous front-end CNN model in our work. Then, a Connectionist Temporal Classification (CTC) loss is stacked on top of the LSTM layers [27] to align the words. Combining these multiple pieces, we build a CRNN-based structure with three components, including CNN, LSTM, and CTC to recognize the manual markers.

4.3 Translate ASL Glosses to Spoken English

Since the majority of hearing people cannot understand ASL and ASL possesses its own unique syntax and grammar, there still exists a communication barrier between d/Deaf and hearing people, even though those ASL glosses can be recognized separately. Therefore, it is necessary to translate the recognized ASL gloss into spoken English (i.e., text or speech) to truly break down the communication barrier between d/Deaf and hearing people. Sign language translation from recognized sign glosses to spoken language usually requires a large amount of data to cover the diversity (e.g., lemmatization, word reordering, and dropping or adding words). However, it is well acknowledged that collecting sufficient labeled data with broad variations is often expensive and difficult [39]. For example, the WearSign system [60] achieves an excellent performance of sequence learning based on 250 common sentences, which means that the WearSign system is prone to overfitting. This issue is particularly true in the medical domain, due to the intrinsic complexity and high costs of data collection [45, 51].

To improve the generalizability of sign language gloss translation, we fine-tuned a well-pre-trained model (i.e., T5) to translate the recognized glosses into spoken language text. Thus, we construct our own training dataset from several previous works [25, 60], a small portion of dataset ASLG-PC12 [38] with a length of fewer than seven words in one meaningful ASL glosses, and augment our own dataset by adding variance to the existing

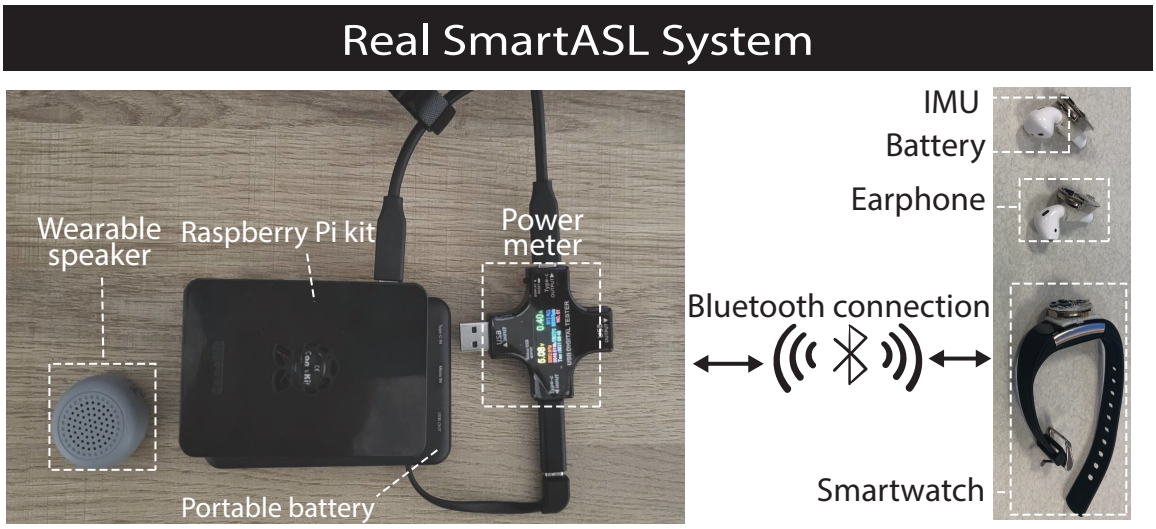

Fig. 7. The hardware block diagram for the real-time system.

sentences (i.e., replace a noun, adjective, and verb with a new word to form new sentences). And then, we utilize our own dataset as the test dataset. During training, we choose the small-size version of T5 (i.e., the model name is “t5-small”) as our pre-trained model. We set the batch_size as 8, epochs as 32, max_input_length as 10, and max_output_length as 10, respectively. During testing, we utilize the generate_text to translate the ASL glosses as spoken English. The test result shows that the re-trained model can translate all the ASL glosses into spoken English accurately. For example, the re-trained model can translate the ASL glosses - “I sad why coffee cold” into spoken English - “I am sad. Why? The coffee is cold.”

5 IMPLEMENTATION

5.1 Hardware

In recent years, IMU sensors have become widely accessible in various commercial earbuds and almost any brand of smartwatch/fitness tracker. However, we need to synchronize the data from earbuds with the data from smartwatches for our recognition simultaneously. Thus, for the development and prototyping purposes of our study, we attached three lightweight wireless IMU sensors on a pair of earbuds and a smartwatch to collect simultaneous data, as shown on the right side of Fig. 7. In this part, these three sensors will be connected to the MacBook Pro laptop with an Intel quad-core processor of 2.9 GHz and a memory of 8 GB. The IMU sensor adopted in this study is STMicroelectronics’ STEVAL-BCN002V1B Bluetooth LE enabled sensor node development kit, which has a circular shape of 2.2 cm in diameter and 14.9 grams in weight and can be powered by a button battery (i.e., CR2032). The small form factor makes it suitable to be attached to earphones. Thus, we utilized the IMU sensors to detect ear canal deformation which corresponds to different non-manual markers.

5.2 Set of ASL Words and Sentences in Experimental Evaluations

Selection of ASL Words In our work, we proposed the idea of utilizing a smartwatch and a pair of earbuds to recognize the manual and non-manual markers in ASL translation. To validate this idea, we selected a set of representative ASL words either with different manual markers or with similar manual markers but with diverse non-manual markers, organized under the following four groups [34, 48]:

Table 1. Selected set of ASL words in experimental evaluation

Category	Words
Noun	name, friend, time, space, uncle, aunt, mirror, camera, plate, family
Verb	bicycle, sign, run, discuss, open, drive, write, read, study, carry, meet
Adjective	big, small, cold, hot, nice, bad, happy, sad, clean, dirty, arrogant, tall, late, not-yet
Pronoun	I, you, he/she, what, who, where, when, how, which, somebody, why, here, there
Negative Word	(Not)drive, (Not)write, (Not)read, (Not)study, (Not)CARRY, (Not)here, (Not)there
Question Word	(?)you, (?)here, (?)there, (?)she/he, (?)I
Mouth Morpheme (CHA)	(CHA)space, (CHA)big, (CHA)tall, (CHA)arrogant, (CHA)plate
Mouth Morpheme (MM)	(MM) drive, (MM)write, (MM)read, (MM)Study, (MM)CARRY
Mouth Morpheme (TH)	(TH) drive, (TH)write, (TH)read, (TH)Study, (TH)CARRY
Mouth Morpheme (CS)	(CS) drive, (CS)write, (CS)read, (CS)Study, (CS)CARRY

- **Base Group**: Similar to previous works [25, 60, 61], we first include four basic groups of words in languages (i.e., noun, verb, adjective, and pronoun).
- **Positive/Negative Group**: It is critical and also challenging to differentiate the same statement under positive or negative expressions. So we intend to explore some negative word expressions compared with their positive counterparts.
- **Question-type Group**: Another essential category is the question-type expressions, compared with the corresponding plain statements.
- **Mouth-morphemes Group**: Linguistic facial expressions have a clear onset and offset. We would like to explore different mouth morphemes and choose the following four representative expressions as examples: "MM" (i.e., the lips press together and protrude) indicates an action done effortlessly, "TH" (i.e., the tongue protrudes slightly) means to do something carelessly, "CS" (i.e., pronouncing the word 'CS' silently) is used to describes a short distance, "CHA" (i.e., pronouncing the word 'CHA' silently) is used to describe the severe levels.

Selection of ASL Sentences ASL grammar is also reflected in sentences, we thus designed the sentence templates with two or three different meaningful patterns — Statement sentence, Negative sentence, and Question sentence. All sentences (with three patterns) have the same hand gestures with different non-manual markers. Specifically, we first created a regular sentence as the statement; then, we added a question marker ("?") to create the question sentence or a negative marker ("not") to create the negative sentence. For example, "Somebody (is) here" is the statement, "Somebody (is) here?" is the corresponding question sentence, and "Somebody (is) not here" has the opposite meaning, as compared to the statement. In total, we created 40 comprehensive sentences.

5.3 Participants & Data Collection

We conducted a data collection study with 10 participants (6 females and 4 males) to make different ASL words or sentences. Three of them are native ASL signers, who are d/Deaf and use ASL as their primary language for more than 10 years. The remaining seven participants are non-d/Deaf ASL users who learned ASL from courses or other d/Deaf individuals. To ensure the accuracy of the communications and experiments, professional ASL interpreters are involved in facilitating the communications between the d/Deaf participants and the experiment coordinator.

During data collection, we divided the process into two stages: experiment preparation and data collection. During the experiment preparation stage, we worked with native ASL signers to review and finalize the list of ASL words and sentences to be examined in the experiment. During the data collection stage, we first asked participants to sit in front of the desktop in a comfortable position. To reduce the artifacts induced by random

physical motions and fatigue, we asked participants to perform our chosen ASL words and sentences with their own preferable styles. Each participant repeated every ASL word and sentence listed in Section 5.2 10 times with an interval of 5 seconds. During the transition time between every two ASL words or sentences, participants were allowed to take a break (e.g., stand up, leave the seat, re-sit on the chair, and take off their smartwatch and earbuds). The study lasted for around 4.5 hours. In total, we collected 10 samples per ASL word and sentence for each subject. The experiments were approved by the Internal Review Board (IRB) of [the university name is hidden for the double-blind review], and the participants were compensated $20 per hour for their participation.

6 EVALUATION

In this section, we conduct a couple of pilot experiments to assess SmartASL's performance on comprehensive ASL recognition. We will report the performance of non-manual markers recognition in Section 6.1, the comprehensive ASL recognition performance based on both non-manual and manual markers in Section 6.2, and the portable prototype system evaluation in Section 6.3.

6.1 Non-manual Markers Recognition Evaluation

To validate the effectiveness of our proposed method in recognizing the non-manual markers in ASL, SmartASL first evaluated the user-dependent recognition accuracy. Then, we evaluated the influence of different numbers of earbuds. Since there are significant differences for each participant in expressing non-manual markers when performing ASL, we also explored the performance of the user-independent model (i.e., leave-one-user-out evaluation) with the calibration to investigate the generalizability of recognizing non-manual markers.

6.1.1 User-Dependent Performance. We first evaluate the recognition accuracy for different non-manual markers (i.e., negative, question marker, "CHA", "CS", "MM", "TH") based on user-dependent training, which means that both training and testing data (80/20 split) are from the same user. As shown in Fig. 8, most of the non-manual markers can be properly distinguished for both native ASL signers and ASL learners. However, it is worth noting that, the non-manual class "MM" has the lowest recognition accuracy for both the learner ASL participants and native ASL participants. The predictions of this class are easy to be mixed with non-manual markers, "CHA" and "TH." The possible reason is that "TH" and "MM" have minimal motions to generate the ear canal deformation.

6.1.2 The Plurality of Earbuds. As discussed above, d/Deaf and hard-of-hearing people show positive attitudes toward using earphone-type devices if they can accurately capture non-manual elements of sign language, according to our pilot user study (Section 5.1). There is another option to embed IMU motion sensors into the hearing aids. Since numerous DHH people may wear a hearing aid on only one ear, instead of hearing aids on both ears, we investigated the recognition accuracy of non-manual markers using one single IMU from only one ear. As shown in Fig. 9, using the data from two IMU sensors can get the highest accuracy due to obtaining more information following training the model. We also observed that the left ear could provide better accuracy than the right ear. One possible reason for this phenomenon is that the left facial side may play a more dominant role in facial activities in response to non-manual markers.

6.1.3 User-Independent Performance With Calibration. Due to different face geometries [16, 52]) and unique behaviors in expressing non-manual markers, even for the same non-manual marker, different individuals will exhibit unique patterns. The inter-individual differences, even for the same non-manual class, will affect the discrimination capability of different non-manual classes. One popular method is to utilize a pre-trained model with a small set of the user's own data samples to calibrate the performance. As a consequence, to investigate the generalizability of our system, we evaluated the leave-one-user-out performance with calibration by adopting a

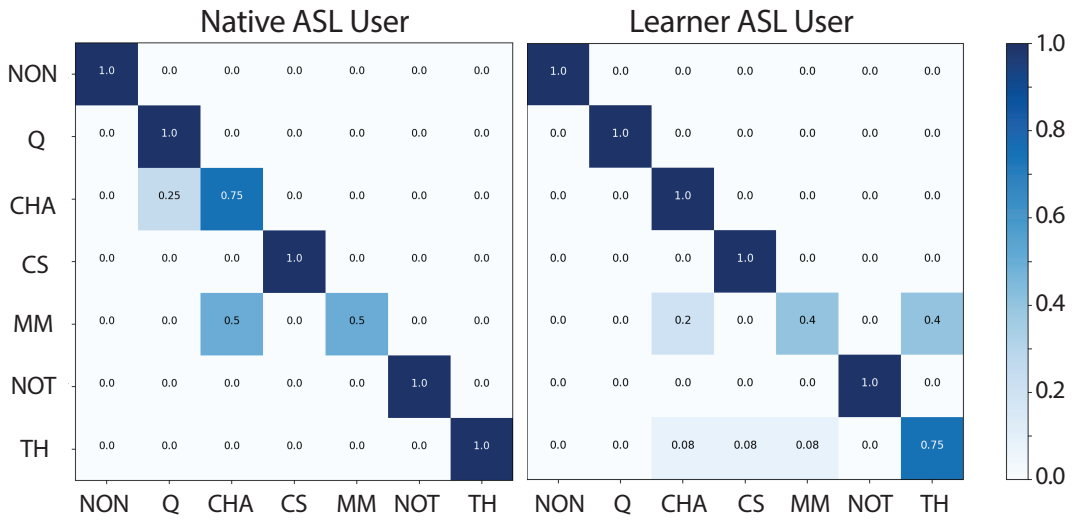

Fig. 8. Confusion matrix for the recognition of seven non-manual classes, 'NON' - natural facial expressions, 'Q' - question marker. (a) Native ASL participants and (b) Learner ASL participants.

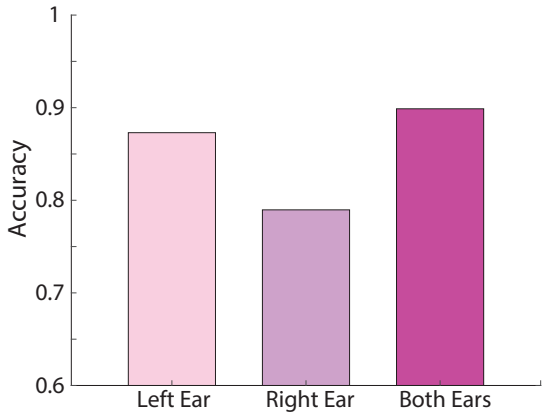

Fig. 9. The accuracy across three IMU positions, mounted in the left ear, right ear, and both ears.

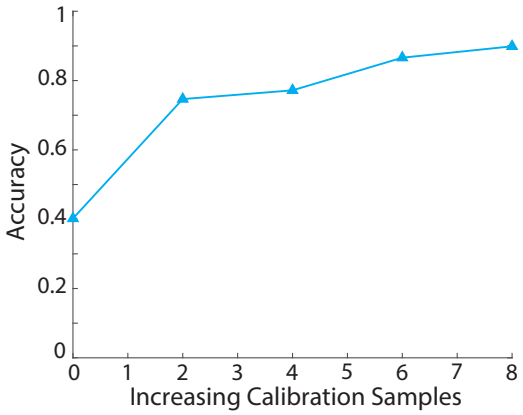

Fig. 10. The accuracy over increasing samples for calibration. "0" means leave-one-user-out accuracy.

small amount of the new user's data to fine-tune the pre-trained model. As shown in Fig. 10, the leave-one-user-out recognition accuracy is only about 40%. With a small amount of two samples for each class, the accuracy increases to more than 77%. Thus we recommend adding several samples to achieve a satisfactory accuracy level.

6.2 ASL Recognition Evaluation Using Both Manual and Non-manual Markers

As discussed before, we have explored the feasibility of recognizing non-manual markers, which were largely neglected in prior studies, using a pair of earbuds equipped with IMU sensors. In this section, we will demonstrate the overall performance of the SmartASL system for recognizing both manual markers and non-manual markers at the same time.

Table 2. The WER in our recognition with different word length

	one word	two words	three words	more words
WER	6.1%	8.7%	7.4%	9.0%

6.2.1 Performance of Comprehensive ASL Recognition. Since Single-word or multi-word ASL expressions usually are widely used in daily communications (e.g., the single-word "Yes" to express the affirmative meaning, the multi-word "Your name what" to ask " What is your name?"), thus we evaluated the performance of SmartASL in recognizing 120 widely used ASL expressions based on user-dependent accuracy, i.e., both training and testing data (80/20 split) were from the same user. To validate the hypothesis that our SmartASL system can effectively recognize the non-manual and manual markers at the same time, we conducted two different experiments for our chosen dataset including words and sentences as listed in Section 5.2. One was to only utilize manual markers to recognize our chosen dataset, and the other one was to utilize manual and non-manual markers to recognize our chosen dataset. As shown in Fig. 11, the average WER for manual and non-manual recognition is 7.7%. As a comparison, it can be observed that an average WER for ASL recognition when considering only manual markers is 20.9%. Without non-manual markers, WER increases by 13.2% compared with the case considering both manual and non-manual markers. This result indicates that our SmartASL system is applicable and effective in improving the capability of recognizing the non-manual and manual markers in ASL simultaneously. Fig.11 visualizes the performance for each subject respectively. We can see that all subjects demonstrate different WERs due to individual differences in signing the same ASL expressions. Those subjects with lower WERs could perform the same manual and non-manual markers in a more stable manner while those subjects with higher WERs were shown to sign ASL in a much less stable manner.

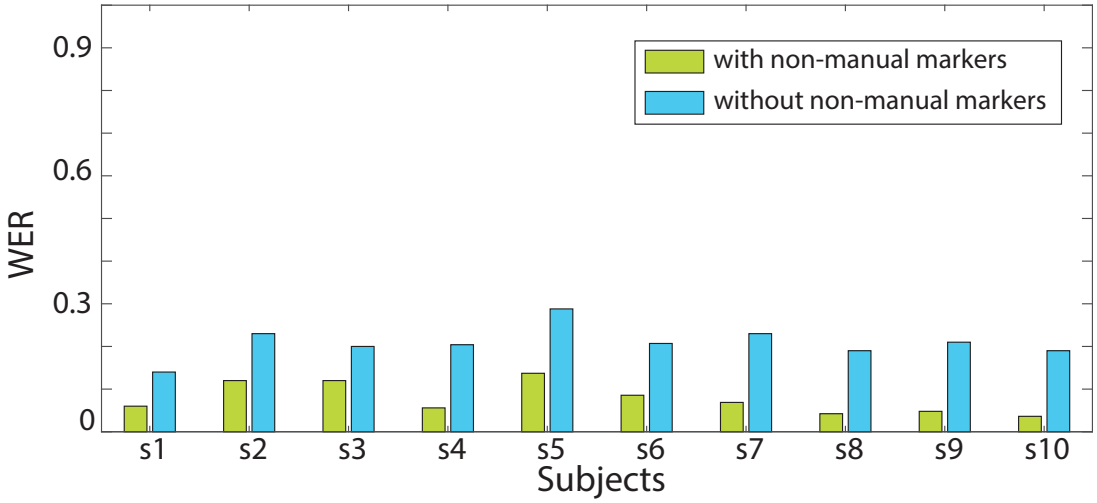

Fig. 11. WER of ASL Recognition based on two settings, (1) utilizing only manual markers for ASL recognition (i.e., without non-manual markers), (2) utilizing both manual and non-manual markers for ASL recognition (i.e., with non-manual markers).

Impact of different word lengths. As shown in Table 2, SmartASL has been observed to achieve lower WERs with one word in ASL, as compared to the average WER when an ASL expression involves more than just words. It is possible because non-manual marker patterns in long-time ASL expressions could be much more vulnerable than those in short-time ASL expressions.

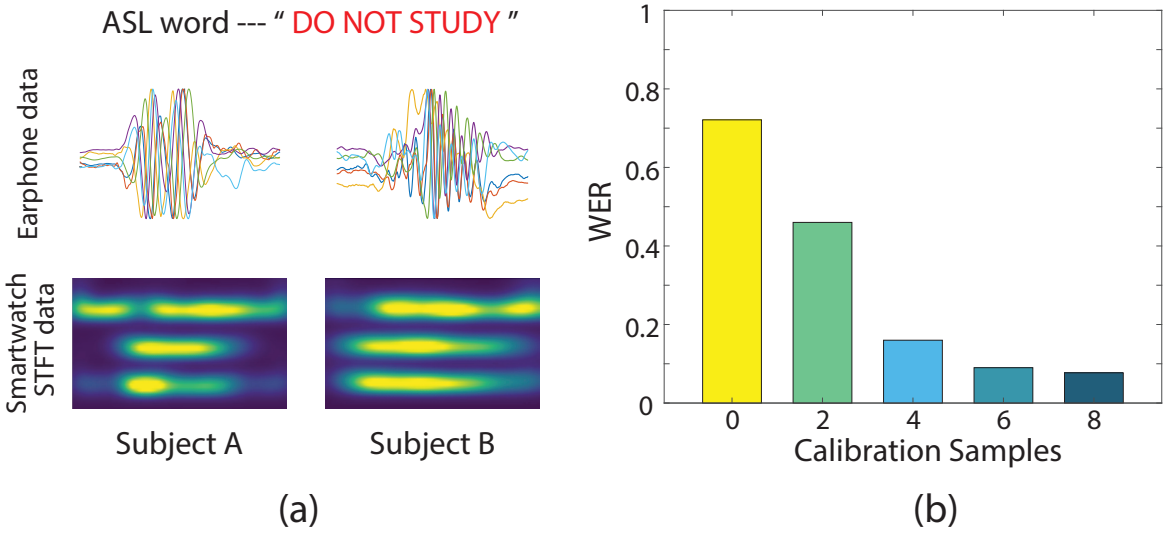

Fig. 12. (a) Different channel response patterns for the same word (i.e., "DO NOT STUDY") from different subjects. (b) The WER over increasing training samples. 0 represents the leave-one-user-out performance without any new user data.

6.2.2 User-independent Performance with Calibration. Since our dual-channel ASL recognition is based on manual and non-manual markers recognition, vast differences in physical and physiological properties, preferences, facial geometries, and hand gesture behaviors among individuals will generate different patterns, as shown in Fig. 12(a). It means that our system may complicate any recognition of different ASL expressions from a new user. To validate our hypothesis, we first calculate the leave-one-user-out performance by implementing the 10-fold cross-validation. As shown in Fig. 12(b), zero calibration samples represent leave-one-user-out accuracy, which can only achieve the WER of 74%. This result is consistent with our expectation that we cannot deploy a well-trained model to a new user without any changes.

An effective and practical solution to adopt a well-trained model to a new user is to include a small set of personal samples to fine-tune the trained model [24, 25, 50] in order to achieve satisfactory performance. Therefore, adopting a small amount of the new user's data to fine-tune the leave-one-user-out model may improve recognition accuracy. Thus, we measure the relationship between the calibration samples and recognition accuracy to evaluate the adaptability of the SmartASL system. We obtain the accuracy by increasing the training samples gradually. Fig. 12 (b) presents the results when adding 2, 4, 6, and 8 samples per class from the new user. These results demonstrate that the WER will decrease to 15.4% when adding just 4 samples. When the number of training samples increases to 6, the WER will decrease to 9.1%, which is comparable to the user-dependent accuracy. Thus, in real-life deployment, a small but increasing number of new user samples are required to improve the accuracy, just like the calibration process in most commercially available electronic systems.

6.3 Portable Prototype Evaluation

To enable long-time seamless communication between hearing people and ASL signers, our SmartASL system should be capable of recognizing ASL signs in the real world with low power consumption. In addition, given the connection distance of Bluetooth, our SmartASL system should be connected to a portable terminal device (i.e., Raspberry Pi). In this section, we seek to implement a portable approximate real-time SmartASL prototype to

Table 3. Latency and power consumption based on our real-time SmartASL system.

	Avg. Power of Raspberry Pi (Watt)	Avg. Latency for System Execution (ms)
Value	4.85	430

achieve the goal of ASL recognition and provide the parameters of latency and power. Then we adopt the System Usability Scale (SUS) to measure this system.

6.3.1 Hardware Implementation and Configuration. As shown in Fig. 7, the proof-of-concept portable SmartASL system consists of an integrated Raspberry Pi 4B development kit, three IMU sensors mounted on a smartwatch, and a pair of earbuds (the right part in Fig. 7), a card size portable battery pack charger with a capacity of 10000 mAh (under the Raspberry Pi part in Fig. 7), a power meter (i.e., testing the power in this evaluation, nor necessary for real application system), and a quarter coin size wearable speaker. With the exception of the portable battery and power meter, all devices are connected to Raspberry Pi 4B through Bluetooth.

To install the proposed SmartASL system on Raspberry Pi 4B, we first trained a robust model and then downloaded this trained model to the Raspberry Pi. The system collected the data streaming from the IMU sensors on the smartwatch and earbuds all the time, and the Raspberry Pi 4B system caught the data when detecting there is a hand event. Then, the SmartASL system extracted the features and fed them into the neural network to predict the ASL expressions. To evaluate the power consumption, we ran our system continuously and obtained the power reading from the power meter. In order to ensure the power measurement accuracy, we calculated the average value by repeating for 5 times with an interval of more than 5 seconds. As shown in Table 3, Table 3, we observed the power consumption to be 4.85 Watts, which means that it is possible to use the SmartASL system for 7.6 hours. Since our SmartASL system is in its early stages without any available correction system, we may need a waiting period to ensure that our system must receive a complete ASL expression to avoid recognition errors. This waiting time is non-negligible to avoid irrational partition and separation of all motion activities corresponding to a complete ASL expression. In summary, the latency of the SmartASL system includes the waiting time and the system execution time, thus we only evaluate the system execution time. During this experiment, we executed the data of 169 word-level ASL gestures and obtained a total processing time of 73 seconds. Table 3 shows that the SmartASL system has a latency of 430 ms for system execution, suggesting its potential capability of achieving real-time ASL recognition and translation in real-world scenarios.

6.3.2 User Study. We conducted a user study to evaluate the proof-of-concept portable SmartASL system. Due to COVID restrictions, we could only host a small-scale in-person study with five participants, which were a subset of the participants in Section 5.3. We adopted the System Usability Scale (SUS) to measure the portable approximate real-time system.

After participants put on the SmartASL, they started to sign the chosen 20 ASL words from our word dataset in a random order, with an interval of at least four seconds to make sure we can receive a whole ASL word. Then, all predicted words were played via the wearable speaker. Afterwards, participants filled out the SUS scale and then went through a brief interview. During the experiment, we recorded the times of wrong predictions and right predictions. The study lasted 20-30 minutes.

SmartASL achieved a robust WER of 13% in the portable stream mode, which is comparable to the performance reported in Section 6.2. Moreover, our survey results showed that SmartASL achieved an average SUS score of 79, indicating good usability. Participants commented that SmartASL was *"easy-to-use"* (P1,P2,P4,P5) and *"portable and easy to adapt in daily life"* (P1,P4,P5).

Table 4. Methodology comparison with existing hand gesture recognition interfaces (complex words & sentences mean the words or sentences are composed of hand motions and facial expressions).

Interfaces	**SmartASL** (Ours)	mmASL [44]	SignFi [33]	WearSign [60]	SonicASL [25]
Non-manual Marker Support	Yes	No	No	No	No
Devices	Earphone + Smartwatch	Radio platform	WiFi AP	Smartwatch + Armband	Headphone by listener
Glosses to Natural Language	Yes	No	No	Yes	No
Features	IMU	60 GHz mmWave	WiFi	IMU + EMG	Acoustic
Support Complex Words & Sentences	Yes	No	No	No	No
Portable	Yes	No	No	Yes	Yes

7 DISCUSSION

7.1 Methodology Comparison with Prior Results

To demonstrate that our prototype is competitive and promising among non-camera-based ASL systems, we compared SmartASL with several representative non-camera-based ASL recognition systems, including mmASL [44], SignFi [33], WearSign [60], and SonicASL [25] with respect to the accuracy, features, the inclusion of non-manual markers, and ASL gloss-to-text translation. As shown in Table 4, the most distinguishing characteristic of the proposed SmartASL is that it is the first system focusing on the recognition of both manual and critical non-manual markers simultaneously by utilizing consumer-grade wearable devices (i.e., earbuds and smartwatches). Among those prior works, WearSign [60] also relied on wearable devices and trained a transformer structure to translate sensor-based ASL signs to spoken language. However, there were only 250 sentences which meant that it was prone to overfitting in practical usage and needed to re-train the model from scratch while expanding the dataset. In our work, we trained both non-manual marker and manual marker recognition models, and fine-tuned the well-trained NLP (i.e., T5-small). We then combined them together for fine-tuning the end-to-end model, which means easy expansion of the dataset of ASL glosses to reduce the training efforts. In summary, our system — SmartASL — is a promising solution and opens a door for user-friendly, low-cost, and accessible technology to bridge the communication gaps between sign language users and spoken language users through recognizing the manual and non-manual markers and adjusting the ASL grammar.

7.2 Limitations and Future Work

Even though we have demonstrated the efficacy and efficiency of the proposed SmartASL system through a variety of evaluations and user studies, several limitations still exist and call for further research.

First, as the first of its kind, our work introduces a unified, end-to-end solution that is capable of handling and recognizing both manual and non-manual components in ASL by demonstrating the potential feasibility through a proof-of-concept prototype. Given the complexity of ASL grammar and syntax [1] and the limited dataset for conducting the experiment, SmartASL only scratches the surface of the complexity of ASL translation and is still in its early stages. Thus, in order to make it truly applicable in daily life, it is mandatory to conduct a more thorough study from the perspective of the linguistic relativity of sign languages. Accordingly, more diverse manual and non-manual markers based on ASL expressions are necessary to improve the robustness of

our system. In addition, considering the vast diversity of sign languages, future work will be conducted on a larger group of d/Deaf people, especially with a wide range of demographic characteristics and experiences.

The second limitation of our study is that the SmartASL system cannot indicate whether or not the translated language is wrong. This will generate misunderstandings to interrupt communication. Thus, it is necessary to seek a method that will facilitate the recognition of incorrect words that SmartASL has generated during translation. A possible solution for this is to provide a glass (like Google Glass [18]) to observe their own recognized transcription. This may help correct errors by re-signing the ASL expressions.

Third, our current design of SmartASL enables one-way communication for helping ASL signers to translate comprehensive ASL into spoken English, which may help hearing people comprehend ASL more accurately. Besides, Google AR glasses were recently introduced to the public, which could be a promising method to transcribe spoken language and display it on the AR glasses in real time [18]. In the future, it may be necessary to combine our SmartASL system with Google Glass to achieve the goal of two-way real-time communications.

8 CONCLUSION

In this work, we propose SmartASL, a wearable-based sign language translation system that can translate sensory signals into spoken texts in an end-to-end way. This system first achieves the goal of manual and non-manual marker recognition through IMU data obtained from a pair of earphones and a smartwatch and forming the recognized glosses. To make hearing people understand the recognized glosses more accurately, we then translate the recognized glosses into a trained model by fine-tuning a well-pre-trained T5 model. We conduct the performance study by recognizing 80 word-level and 40 sentence-level ASL expressions which consist of both manual and non-manual markers based on 10 participants and achieved the WER of 7.7%. Thus, we propose a promising method to improve the communication ability for several impromptu application scenarios between hearing people and d/Deaf people.

ACKNOWLEDGMENTS

This work was supported in part by the Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004) and the Endowed Professorship from the Shenzhen Holdfound Foundation.

REFERENCES

[1] Accessibility.com, LLC. 2022. Is American Sign Language (ASL) a language? https://www.accessibility.com/blog/is-american-sign-language-asl-a-language/.

[2] Ashwin Ahuja, Andrea Ferlini, and Cecilia Mascolo. 2021. PilotEar: Enabling In-ear Inertial Navigation. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*. 139–145.

[3] 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*. 1–9.

[4] F Berzin and CRH Fortinguerra. 1993. EMG study of the anterior, superior and posterior auricular muscles in man. *Annals of Anatomy-Anatomischer Anzeiger* 175, 2 (1993), 195–197.

[5] Hongliang Bi and Jiajia Liu. 2022. CSEar: Meta-learning for Head Gesture Recognition Using Earphones in Internet of Healthcare Things. *IEEE Internet of Things Journal* (2022).

[6] Eric Branda and Tobias Wurzbacher. 2021. Motion Sensors in Automatic Steering of Hearing Aids. In *Seminars in Hearing*, Vol. 42. Thieme Medical Publishers, Inc., 237–247.

[7] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, Taeho Kim, Nicholas Farrow, Anh Nguyen, Jianliang Xiao, et al. 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*. 1–17.

[8] Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, and Cecilia Mascolo. 2021. Motion-resilient Heart Rate Monitoring with In-ear Microphones. *arXiv preprint arXiv:2108.09393* (2021).

[9] George Caridakis, Stylianos Asteriadis, and Kostas Karpouzis. 2014. Non-manual cues in automatic sign language recognition. *Personal and ubiquitous computing* 18, 1 (2014), 37–46.

[10] Seokmin Choi, Yang Gao, Yincheng Jin, Se jun Kim, Jiyang Li, Wenyao Xu, and Zhanpeng Jin. 2022. PPGface: Like What You Are Watching? Earphones Can" Feel" Your Facial Expressions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–32.
[11] Deaf Community. 2021. Deaf Culture. https://www.startasl.com/what-does-d-d-and-d-deaf-mean-in-the-deaf-community/. [Updated May 13, 2021].
[12] ASLLRP Continuous Signing Corpora. 2022. American Sign Language Linguistic Research Project. https://dai.cs.rutgers.edu/dai/s/dai. [Online].
[13] 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*. 1–13.
[14] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. 2021. EarGate: gait-based user identification with in-ear microphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 337–349.
[15] E Friesen and Paul Ekman. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto* 3, 2 (1978), 5.
[16] Yang Gao, Yincheng Jin, Seokmin Choi, Jiyang Li, Junjie Pan, Lin Shu, Chi Zhou, and Zhanpeng Jin. 2021. SonicFace: Tracking Facial Expressions Using a Commodity Microphone Array. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–33.
[17] 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.
[18] Google. 2022. AR Glass. https://nerdist.com/article/google-ar-glasses-live-translation-real-time-transcription/.
[19] Audien Hearing. 2023. Atom Pro. https://audienhearing.com/products/audien-atom-pro-pair?variant=39511193255999.
[20] Carl-Herman Hjortsjö. 1969. *Man's face and mimic language*. Studentlitteratur.
[21] 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.
[22] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2018. Video-based sign language recognition without temporal segmentation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
[23] Yincheng Jin, Yang Gao, Xiaotao Guo, Jun Wen, Zhengxiong Li, and Zhanpeng Jin. 2022. EarHealth: an earphone-based acoustic otoscope for detection of multiple ear diseases in daily life. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 397–408.
[24] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: " Hearing" Your Silent Speech Commands In Ear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–28.
[25] 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. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–30.
[26] Sara Askari Khomami and Sina Shamekhi. 2021. Persian sign language recognition using IMU and surface EMG sensors. *Measurement* 168 (2021), 108471.
[27] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4835–4839.
[28] Nicolas Le Goff, Jesper Jensen, Michael Syskind Pedersen, and Susanna Løve Callaway. 2016. An introduction to OpenSound Navigator™. *Oticon A/S* (2016).
[29] Steven F LeBoeuf, Michael E Aumer, William E Kraus, Johanna L Johnson, and Brian Duscha. 2014. Earbud-based sensor for the assessment of energy expenditure, heart rate, and VO2max. *Medicine and Science in Sports and Exercise* 46, 5 (2014), 1046.
[30] Yilin Liu, Fengyang Jiang, and Mahanth Gowda. 2020. Finger Gesture Tracking for Interactive Applications: A Pilot Study with Sign Languages. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–21.
[31] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. When Video Meets Inertial Sensors: Zero-Shot Domain Adaptation for Finger Motion Analytics with Inertial Sensors. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation* (Charlottesvle, VA, USA) *(IoTDI '21)*. ACM, New York, NY, USA, 182–194.
[32] Hamzah Luqman and El-Sayed M El-Alfy. 2021. Towards hybrid multimodal manual and non-manual Arabic sign language recognition: MArSL database and pilot study. *Electronics* 10, 14 (2021), 1739.
[33] 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.
[34] Stephen McCullough, Karen Emmorey, and Martin Sereno. 2005. Neural organization for recognition of grammatical and emotional facial expressions in deaf ASL signers and hearing nonsigners. *Cognitive Brain Research* 22, 2 (2005), 193–203.
[35] Meta. 2016. Binaural Audio for Narrative AR. https://www.oculus.com/story-studio/blog/binaural-audio-for-narrative-vr/.

[36] Nicholas Michael, Peng Yang, Qingshan Liu, Dimitris N Metaxas, Carol Neidle, and CBIM Center. 2011. A Framework for the Recognition of Nonmanual Markers in Segmented Sequences of American Sign Language.. In *BMVC*. 1–12.
[37] NIH. 2008. Hearing Loss and Hearing Aid Use. https://www.nidcd.nih.gov/news/multimedia/hearing-loss-and-hearing-aid-use-text-version. [Updated July 17, 2017].
[38] Achraf Othman and Mohamed Jemni. 2012. English-ASL gloss parallel corpus 2012: ASLG-PC12. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon LREC*.
[39] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2009), 1345–1359.
[40] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67.
[41] Grand Review Research. 2023. Grand Review Research. https://www.grandviewresearch.com/industry-analysis/earphone-and-headphone-market. [Online].
[42] Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. 2022. Sensing with Earables: A Systematic Literature Review and Taxonomy of Phenomena. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–57.
[43] Arman Sabyrov, Medet Mukushev, and Vadim Kimmelman. 2019. Towards Real-time Sign Language Interpreting Robot: Evaluation of Non-manual Components on Recognition Accuracy.. In *CVPR Workshops*.
[44] 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.
[45] Torgyn Shaikhina and Natalia A. Khovanova. 2017. Handling limited datasets with neural networks in medical applications: A small-data approach. *Artificial Intelligence in Medicine* 75 (2017), 51–63.
[46] Jiacheng Shang and Jie Wu. 2017. A robust sign language recognition system with multiple Wi-Fi devices. In *Proceedings of the Workshop on Mobility in the Evolving Internet Architecture*. 19–24.
[47] Xingzhe Song, Kai Huang, and Wei Gao. 2022. FaceListener: Recognizing Human Facial Expressions via Acoustic Sensing on Commodity Headphones. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 145–157.
[48] StartASL. 2020. ASL Dictionary – Learn Essential Vocabulary. https://www.handspeak.com/word/. [Updated April 28, 2020].
[49] Karush Suri and Rinki Gupta. 2019. Continuous sign language recognition from wearable IMUs using deep capsule networks and game theory. *Computers & Electrical Engineering* 78 (2019), 493–503.
[50] Noeru Suzuki, Yuki Watanabe, and Atsushi Nakazawa. 2020. Gan-based style transformation to improve gesture-recognition accuracy. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 4, 4 (2020), 1–20.
[51] Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J. Casson. 2019. Machine learning algorithm validation with a limited sample size. *PLoS ONE* 14, 11 (2019), 1–20.
[52] Dhruv Verma, Sejal Bhalla, Dhruv Sahnan, Jainendra Shukla, and Aman Parnami. 2021. ExpressEar: Sensing Fine-Grained Facial Expressions with Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–28.
[53] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. EarDynamic: An Ear Canal Deformation Based Continuous User Authentication Using In-Ear Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–27.
[54] Katharine L Watson. 2010. *WH-questions in American Sign Language: Contributions of non-manual marking to structure and meaning*. Purdue University.
[55] Traci Patricia Weast. 2008. *Questions in American Sign Language: A quantitative analysis of raised and lowered eyebrows*. The University of Texas at Arlington.
[56] WHO. 2022. Deafness and hearing loss. https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss. [Online].
[57] Jian Wu, Lu Sun, and Roozbeh Jafari. 2016. A Wearable System for Recognizing American Sign Language in Real-Time Using IMU and Surface EMG Sensors. *IEEE Journal of Biomedical and Health Informatics* 20, 5 (2016), 1281–1290.
[58] Kayo Yin. 2020. Sign language translation with transformers. *arXiv preprint arXiv:2004.00588* 2 (2020).
[59] 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.
[60] Qian Zhang, JiaZhen Jing, Dong Wang, and Run Zhao. 2022. WearSign: Pushing the Limit of Sign Language Translation Using Inertial and EMG Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.
[61] 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.
[62] Zhihao Zhou, Kyle Chen, Xiaoshi Li, Songlin Zhang, Yufen Wu, Yihao Zhou, Keyu Meng, Chenchen Sun, Qiang He, Wenjing Fan, Endong Fan, Zhiwei Lin, Xulong Tan, Weili Deng, Jin Yang, and Jun Chen. 2020. Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays. *Nature Electronics* 3 (2020), 571–578.