EchoSpeech: Continuous Silent Speech Recognition on Minimally-obtrusive Eyewear Powered by Acoustic Sensing

Ruidong Zhang
rz379@cornell.edu
Cornell University
Ithaca, NY, USA

Ke Li
kl975@cornell.edu
Cornell University
Ithaca, NY, USA

Yihong Hao
yh826@cornell.edu
Cornell University
Ithaca, NY, USA

Yufan Wang
yw583@cornell.edu
Cornell University
Ithaca, NY, USA

Zhengnan Lai
zl345@cornell.edu
Cornell University
Ithaca, NY, USA

François Guimbretière
fvg3@cornell.edu
Cornell University
Ithaca, NY, USA

Cheng Zhang
chengzhang@cornell.edu
Cornell University
Ithaca, NY, USA

ABSTRACT

We present EchoSpeech, a minimally-obtrusive silent speech interface (SSI) powered by low-power active acoustic sensing. EchoSpeech uses speakers and microphones mounted on a glass-frame and emits inaudible sound waves towards the skin. By analyzing echos from multiple paths, EchoSpeech captures subtle skin deformations caused by silent utterances and uses them to infer silent speech. With a user study of 12 participants, we demonstrate that EchoSpeech can recognize 31 isolated commands and 3-6 figure connected digits with 4.5% (std 3.5%) and 6.1% (std 4.2%) Word Error Rate (WER), respectively. We further evaluated EchoSpeech under scenarios including walking and noise injection to test its robustness. We then demonstrated using EchoSpeech in demo applications in real-time operating at 73.3mW, where the real-time pipeline was implemented on a smartphone with only 1-6 minutes of training data. We believe that EchoSpeech takes a solid step towards minimally-obtrusive wearable SSI for real-life deployment.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**; *Gestural input*; • **Computing methodologies** → Speech recognition.

KEYWORDS

Silent Speech Recognition, Acoustic Sensing, Smart Glasses

ACM Reference Format:
Ruidong Zhang, Ke Li, Yihong Hao, Yufan Wang, Zhengnan Lai, François Guimbretière, and Cheng Zhang. 2023. EchoSpeech: Continuous Silent Speech Recognition on Minimally-obtrusive Eyewear Powered by Acoustic Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3580801

1 INTRODUCTION

Silent speech interface (SSI) has drawn increasing attention lately. Compared with voiced speech, silent speech does not require the users to vocalize sounds, which expands its application scenarios to where voiced speech is limited. For instance, SSI can be used in noisy environments where voiced speech may suffer from severe interference or in quiet places and other scenarios where it is socially inappropriate to speak out loud. A recent study founds out that SSI are more socially acceptable than voiced speech, and that users are willing to tolerate more errors [46]. Studies also found that social awkwardness and privacy concerns are important factors affecting user's perception of and willingness to use voice assistants [51, 65]. By removing the need to speak out loud, SSI better preserves privacy. These advantages make SSI promising in expanding the use case of voice assistant with a silent voice assistant. In addition, SSI opens up brand new opportunities where voiced speech has not touched. For instance, SSI can be used to input password without leaking out sounds to the environment. Collaborators in a shared workspace can use SSI to instruct AI agents without disturbing each other.

Despite these promising benefits, there exists substantial challenges preventing existing SSI technologies from being widely used. The most popular SSIs use cameras to capture lip movements. However, these methods require the presence of a camera without severe occlusion, which limits its availability. The wearable community comes up with various solutions to address this limitation. However, most of them require placing skin-contacting sensors inside the mouth [4, 6, 18, 21, 30–33, 39, 52] or on the frontal face [28, 29, 41, 42, 54, 62, 63], which may not be physically or

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3580801

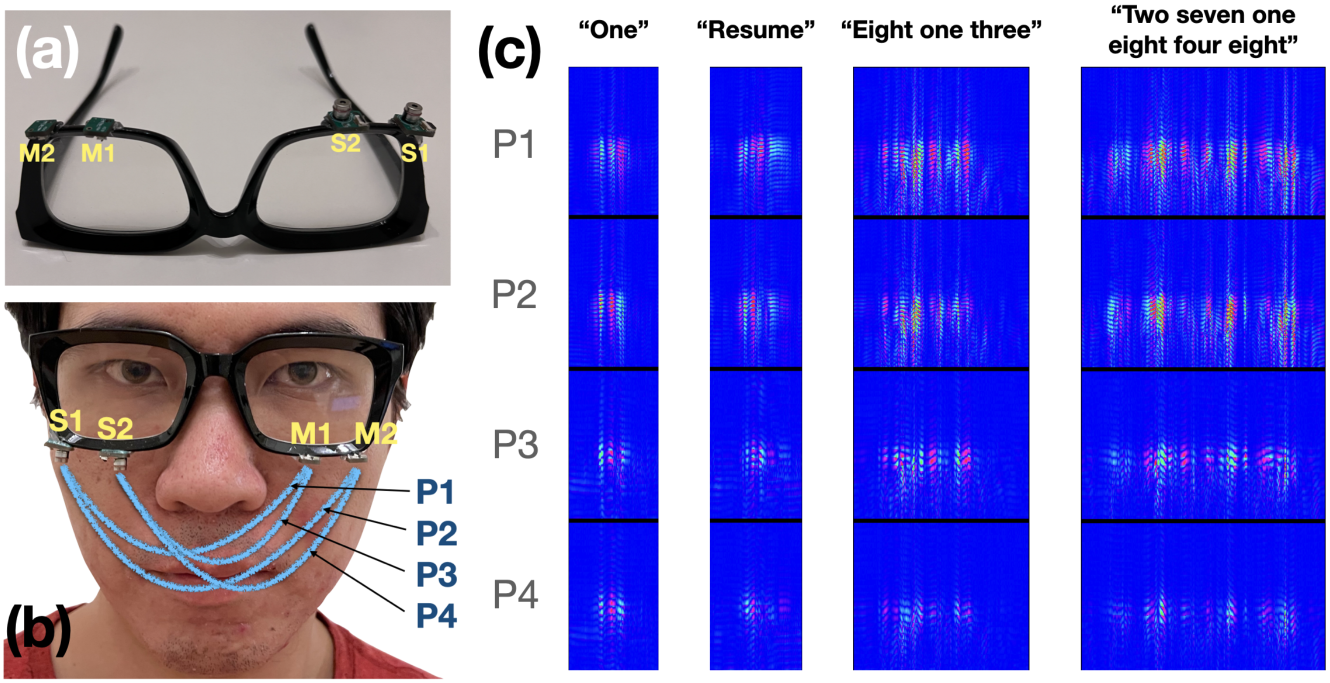


Figure 1: System layout and echo profiles. (a-b)Final sensor position and signal paths. S1, S2: speakers; M1, M2: microphones. P1-P4: Paths. Note that each path consists of multiple path reflection and diffraction that originates from the source speaker and ends at the target microphone. The lines in the figure only illustrate the sources and targets. (c) Echo profiles for different utterances.

socially comfortable. Recent research tries to place less-obtrusive sensors at less-visible positions such as behind the ear [55] or under the chin [50]. However, such positions can only provide limited information thus requiring extra effort such as speaking slowly [55] to ensure performance. Additionally, wearing such devices for an extended period of time may still be uncomfortable [55]. Contact-free SSIs do not need sensors to be tightly coupled with the skin and have drawn recent attentions. Promising results are seen on necklace-mounted camera [71] and in-ear acoustic sensing based SSIs [25]. However, camera-based methods often suffer from high power consumption and privacy concerns [71], while in-ear systems may still be uncomfortable for long-term wearing.

To make things worse, lack of a reliable, comfortable and minimally-obtrusive form factor is not the only obstacle faced by wearable SSIs. Performance is another key challenge. The ability to recognize speech with natural speaking speed and style (e.g., continuously speaking out multiple words together without pausing) is key towards a natural and user-friendly SSI. However, most wearable SSIs are only able to recognize a pre-defined set of discrete commands [7, 28, 34, 71, 72]. Some have extra restrictions such as speaking slowly [25, 55], remaining still [71], exaggerating speech [23], or were only evaluated in-session [23, 28]. In addition, the ability to recognize speech at a sentence level is still extremely limited in wearable SSIs. The past year witnessed most of such advancements in works such as MuteIt [55] and EarCommand [25]. However, their abilities to recognize continuous and connected speech are still limited (discussed in Section 2.3).

To address these challenges, we propose EchoSpeech, a minimally-obtrusive contact-free SSI that is able to recognize both discrete and continuous speech. EchoSpeech is powered by active acoustic sensing using miniature speakers and microphones mounted on the lower edge of a commercial off-the-shelf (COTS) glass-frame to track lip and skin movements from multiple paths. We designed a customized deep learning pipeline with connectionist temporal classification (CTC) loss that enables EchoSpeech to recognize both discrete and continuous speech without segmentation needed. We evaluated EchoSpeech with a study of 12 participants and demonstrate that EchoSpeech achieves a WER of 4.5% (std 3.5%) and 6.1% (std 4.2%) in recognizing 31 isolated commands and 3-6 figure connected digits spoken at a speed of 101 words per minute (wpm). To minimize training effort from new users and improve performance, we designed a two-step (pre-training + fine-tuning) training scheme. We demonstrate that with only 6-7 minutes of training data, EchoSpeech achieves 9.5% and 14.4% WER on recognizing isolated and connected speech. We further demonstratedEchoSpeech's robustness in scenarios such as walking and noise injection. To demonstrate the use case and effectiveness, we applied EchoSpeech in four real-time demos on a low-power variant operating at 73.3mW.

We summarize our contributions as follows:

- We propose EchoSpeech, a minimally-obtrusive, contact-free SSI powered by active acoustic sensing on a glass-frame that

recognizes both isolated and connected speech with around 5% WER.
- To our knowledge, EchoSpeech is the first SSI on a single glass-frame.
- We propose a CNN-based segmentation-free silent speech recognition pipeline for acoustic sensing.
- We evaluated EchoSpeech under multiple scenarios and demonstrate its use case with a real-time demo implemented on a smartphone.

2 RELATED WORK

In this section, we summarize existing silent speech interfaces and discuss their links to this work. We start by briefly summarizing non-wearable SSIs and then discuss wearable systems in two categories: contacting and contact-free silent speech interfaces depending on whether the sensors need physical contact with the skin.

2.1 Non-wearable SSI

Silent speech recognition, as well as similar related tasks known as lip reading or video captioning has been extensively studies in the computer vision community. Large datasets with videos captured by frontal cameras under various scenarios have been established, from controlled lab settings [2, 10, 19, 48, 49, 74] to unconstrained free-living scenarios [8, 9, 69]. SSIs using cameras cover various granularities ranging from isolated commands [15, 17, 49, 56, 61, 74] and sentences [3, 8, 9, 24, 36, 57, 67], to sound restoration [1, 5, 11, 14, 43, 44, 60]. Despite their impressive performance, the drawbacks are also evident: they require users to be present in front of a camera without severe occlusion, which may not be portable and may raise privacy concerns.

A workaround to the portability issue is deploying the system on mobile devices, which has drawn increasing attention lately. Researchers explored using the built-in camera [45, 56] or speaker and microphone [16, 40, 70, 73] of the smartphone to capture lips movements and infer silent speech from them. These systems did an excellent job in utilizing existing resources. However, they still require holding the phone in the hand. For a real hands-free and eyes-free system, fully wearable solutions are needed.

2.2 Contacting SSI on Wearables

Due to the difficulty in scaling up data collection, the wearable community strives to capture as much high quality information as possible. Directly placing sensors on the articulators in the mouth is an efficient way to capture such information. For instance, magnetometers have been placed on the tongue and/or lips to directly capture tongue/lip movements[4, 6, 18, 21, 30, 31, 52]. For similar purposes, capacitive sensors were also used inside of the mouth[32, 33, 39]. This approach has demonstrated promising results in recognize a large set of words [33] or even sentences such as connected digits [21]. However, these systems are highly obtrusive and many users might find putting artifacts inside the mouth uncomfortable.

For improved comfort, researchers also explored putting sensors externally to capture signals that reflect internal movements. In this category, ultrasonic imaging uses skin-contacting probes usually tightly under the chin to obtain direct imaging of the internal structures to infer silent speech [12, 13, 27, 35, 58, 68] or even synthesize voices [35]. Another well-explored direction uses electromyography (EMG) to infer silent speech from muscle movements represented by EMG signals. This approach requires attaching multiple electrodes on the skin, mostly on the frontal face and the chin [28, 29, 41, 42, 54, 63]. Similar approaches with different sensing principles include placing RFID tags around the mouth to capture lip and cheek movements [62], using electrodes around the head to capture electroencephalography (EEG) signals [59] and analyzing vocal tract shape using MRI signals [47].

However, wearing multiple sensors on the frontal face may not be physically or socially comfortable. More recent work tries to mitigate this issue by exploring less obtrusive sensor locations such as motion sensors behind the ear [55] or under the chin [50]. Specifically, MuteIt [55] achieves impressive performance especially in recognizing unseen words. However, it requires users to speak slowly. In addition, these approaches still require users to wear skin-contacting devices which may not be comfortable for long-term deployment (e.g., most users thought that MuteIt was confortable to wear for less than 2 hours [55]). Such limitations in contacting SSIs inspire researchers to explore contact-free solutions.

2.3 Contact-free SSI on Wearables

Compared with putting sensors contacting the skin, contact-free SSIs are usually more comfortable and user-friendly. However, they face more challenges in obtaining high quality signals because sensors need to be put relatively far away from the articulators. This area has not been thoroughly explored but is drawing increasing attention lately. Research in this area usually targets minimal obtrusiveness, trying to deploy the system on COTS form factors. Recent advances include putting camera(s) on earphones/headphones [7] or on a necklace [34, 71], acoustic sensors on a VR headset [72] and infrared distance sensors on eyewear with an extended pole [23]. However, most of them [7, 23, 34, 72] were only evaluated at a limited scale with a small command set or in-session [23]. In addition, cameras usually consume a lot of energy (e.g., SpeeChin's sensing unit operates at 2.4W [71]) and have significant privacy concerns.

Another line of work falls at the boundary of contacting and contact-free systems, where sensors themselves are contact-free but the form factor needs to be tightly attached to the skin. For instance, motion sensors [20] and strain sensors [37] have been deployed on a mask. However, these systems only achieve limited performance, likely due to lack of a reliable representation of articulator movements. Another promising direction is earables, where recent work EarCommand [25] demonstrates encouraging results using in-ear acoustic sensing. However, EarCommand experiences significant performance drop while tested across sessions. Additionally, compared with fully contact-free systems, these systems still have disadvantages during long-term wearing.

In both contacting and contact-free SSIs, continuous recognition ability is still extremely limited. Most recent advances come from MuteIt [55] and EarCommand [25]. The former breaks down words into phonemes to gain ability to adapt to unseen words. However, it requires users to speak slowly and did not evaluate continuous speech with normal speed. The latter demonstrates promising results in recognizing 27 pre-defined sentences. However, these sentences were short (2-5 words) and uttered slowly (5-13s per sentence) with long pauses between sentences. It is unclear how it generalizes to unseen sentences spoken at a normal speed.

Compared with previous work, EchoSpeech provides a low-power, minimally-obtrusive contact-free silent speech interface powered by active acoustic sensing. To our knowledge, EchoSpeech is the first contact-free SSI deployed on a single glass-frame. It deploys miniature speakers and microphones on the lower edge of a COTS glass-frame and achieves around 5% cross-session WER in recognizing 31 isolated commands and 3-6 figure connected digits that are spoken at 101 wpm.

3 THEORY OF OPERATION

In this section, we first explain the rationale and demonstrate the principles behind EchoSpeech. We then explain the rationale behind the current design by explaining our design goals.

When people speak, whether with or without vocalizing voices, muscles on the face drive different parts of the face to move. Among these parts, lip movements are especially useful in inferring speech. As demonstrated in previous work such as EarIO [38], active acoustic sensing works reliably in tracking subtle skin deformation when placed behind the ear. EchoSpeech uses a similar sensing principle. We mount speakers and microphones on the glass-frame close to the face. The speakers emit encoded sound waves, which are reflected and diffracted by various facial parts including the lips and captured by microphones. With our form factor setup in Figure 1(a-b), speakers and microphones are mounted on different sides of the face. Signals emitted by speakers travel across the face through different paths and are captured by microphones on the other side of the face. To demonstrate how the system captures facial movements during speech, we recorded a few silent utterances (defined as a period of silent speech such as a word, phrase or short sentence separated by pauses) while wearing our system and visualized the echo profiles as illustrated in Figure 1(c). In the echo profiles, different silent utterances appear as strong yet distinct patterns, indicating that EchoSpeech is able to capture the movements caused by silent speech. With a customized machine learning pipeline, such patterns can be used to infer speech.

We reach the current design through an iterative process centered around our design goals. As discussed in Section 1 and Section 2, existing SSI systems are faced with two key challenges: 1) lack of a reliable, and physically and socially comfortable form factor, and 2) lack of the ability to recognize speech in a natural and continuous way. We strive to address these challenges with our system. To achieve this, we identify several goals in our system iterations:

(1) The form factor should be minimally-obtrusive and comfortable to wear.
(2) The system should be evaluated in a way that is as natural as possible.
(3) The system should be low-power, privacy aware and require as little training effort as possible.

Bearing these goals in mind, we explored and experimented on different options of hardware, form factor, and algorithms. Finally we evaluated the system with a setup to reflect our goals. We detail this process and its outcome in sections to come.

4 HARDWARE ITERATIONS

We strive to align our system with design goal (1) by exploring and experimenting our options in sensing methods, form factor and sensor configurations. We present these iterations in this section.

4.1 The Choice of Sensing Method

To fulfill design goal (1), we need a contact-free sensing approach that can be easily deployed on wearables. Encouraged by recent successes such as EarIO [38] and EarCommand [25], we chose acoustic sensing as the sensing method. Compared with other contact-free sensing methods such as cameras, acoustic sensing is much more power-efficient and privacy-aware. Compared with methods such as capacitive or distance sensing, acoustic provides better sensing range and resolution. In addition, acoustic sensors are cheap and widely available on wearable devices. To minimize privacy concerns as well as avoiding annoying users with noises, we only use audio signals over 18kHz and apply band-pass filter to remove low frequency components where most sensitive sounds are distributed.

4.2 The Choice of Form Factor

With acoustic sensing, we use reflected and diffracted sound signals to recover the pattern of movements of the articulators and their connecting tissues which naturally occurs while speaking. To be compliant with design goal (1), we focus on COTS form factors. We explored placing acoustic sensors at different positions around the head. A lot of recent work focus on earables and have demonstrated promising results in authentication [64], facial expression tracking [38] and silent speech recognition [25]. Inspired by these successes, we started with a form factor similar to that of EarIO [38] by putting acoustic sensors behind the ear. Note that we did not try in-ear form factors specifically because it may not be comfortable to wear over an extended period of time. Our early results were promising. However, scaling up the system to more participants led to unstable performance on participants with different head shapes.

Behind-the-ear form factor only captures limited information mostly related to jaw movements. To obtain more information to for better performance, we turned to the front side of the face. We considered necklace such as used in SpeeChin [71]. However, necklaces meet additional challenge when the participants are walking as revealed by SpeeChin. This is because necklaces are not attached to the head, where majority of movements occur during speech.

We then experimented on glass-frames. Glass-frame has several benefits that other form factors do not have. It is comfortable for long-term wearing as many people wear it all day long. It is also stably mounted on the head, which allows users to be able to speak naturally without need to keep still. In addition, glass-frame expands from behind the ear to above the nose in the front, which gives us more flexibility in placing sensors at different locations without significant hardware modifications. Many of these locations are close to skins and muscles that have significant deformation during speech, which could result in better performance. These advantages make glass-frame well aligned with our design goals. We experimented on various setups on a glass-frame and gradually improved performances until they are acceptable, which we detail in the next section.

4.3 Identifying the Optimal Sensor Setup

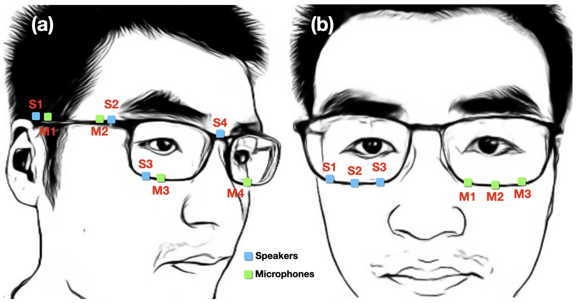


Figure 2: Iterations on sensor positions. (a) Early experiments on sensor positions. (b) Experiments on the lower edge of the glass-frame

In order to find the optimal form factor setup that best balances our design goals, we conducted experiments on various sensor positions, orientations, and quantities. To quickly quantify these explorations, we compared different setups with a small-scale standard test to compare their performances. In the test, one researcher wore the glasses and used a 10-word command set (10 digits, zero to nine) and collected 40 repetitions for each word. We used a simple CNN model to classify these 10 words. We chose a simple CNN to quickly obtain horizontal comparison between different configurations, not to achieve best performance in this step.

To better preserve design goal (1), we started with the most unobtrusive setup by placing the sensors on the leg of the glass-frame. We experimented with S1+M1/M2, S2+M1/M2 as illustrated in Figure 2(a). This setup does not put any sensor in the front. Instead, it captures skin deformations on the side of the cheeks. However, such deformations are very weak and not informative enough for inferring activities as subtle as silent speech with our setup. Accuracy on the standard test was under 50%.

We then moved the sensor to the front side, placing a speaker near the nose bridge while two microphones on either side of the frame to get a symmetric setup (S4+M3+M4 in Figure 2(a)). With this setup, the system was able to capture quite strong patterns during speaking. However, such patterns are not distinguishable enough to tell over 10 different commands apart, achieving around 50% accuracy on the standard test. We believe that this is because the paths that signals travel only cover areas that have limited freedom in deformation, mostly around the eyes and the nose bridge. We then moved both the speaker and the microphones to the lower edges of the frame, using S3+M3 as illustrated in Figure 2(a), hoping to observe the frontal face from a closer-up position. This setup delivered a much better performance than previous trials, around 90% on the standard test. However, when we scaled up the command set size, performance decreased. This is still not enough to reach design goal (2).

Looking into this setup, the signal travels from the speaker and reaches mostly the face and partly the lips before reaching back to the microphone. We hypothesized that having the signal paths over the lips could improve performance by capturing movements of the lip. Therefore, we updated the design and placed the speaker and microphone on different sides of the frame (S3+M4 in Figure 2(a)). With this setup, we achieved a significant performance boost on the standard test from 90% to 98%. More importantly, this performance persisted after scaling up the command set.

We then moved on to further optimize this setup. We first experimented on the microphone location. We placed a speaker near the center and three microphones at the left, center and right of the lower frame as illustrated in Figure 2(b) (S2 + M1 + M2 + M3). Since all setups achieved over 97% on the standard test, we reduced the amount of training data to magnify difference between different setups. Results showed that M2 and M3 achieves best results (around 90% accuracy with less training data) while M1 which was closest to the nose worked significantly worse than the other positions (47% accuracy). Similarly, we experimented on the 3 speaker positions in Figure 2(b) (S1, S2, S3). Results showed that speaker nearest to the nose (S3) had significant worse performance than the other two.

Based on the preliminary findings, we chose to use two speakers (S1+S2) and two microphones (M2+M3) on either sides of the frame, as illustrated in Figure 1(a-b). In this way, signals can travel through different paths to capture more information. We verify in Section 7.6 that this setup indeed yields better performance than using fewer sensors.

5 IMPLEMENTATION

In this section, we describe the hardware and software implementation of EchoSpeech.

5.1 Hardware and Form Factor

As illustrated in Figure 1(a-b), we placed two speakers on the left edge (seen from the front) and two microphones on the right edge. We do not anticipate any difference if the sides of the speakers and microphones are reversed. The speakers and microphones are common commercial products (speaker: OWR-06049T-38D, microphone: ICS-43434).

The speakers and microphones were connected to a micro-controller module (Teensy 4.1) via flexible printed circuits (FPC) cables. We designed a separate add-on board to house the audio amplifier (SGTL5000) and FPC headers. Data were stored in an on-board micro-SD card on the micro-controller. Hardware boards and their dimensions are specified in Figure 4(a-d).

5.2 Echo Profile Calculation

We used active acoustic as the sensing approach. We chose frequency-modulated-continuous-wave (FMCW) as the transmitted signal similar to EarIO [38]. To take advantage of two speaker positions, we used different frequency ranges for the two speakers (18-21kHz for S1, 21.5-24.5kHz for S2). Both frequency ranges are inaudible to most people. The micro-controller was configured to sample at 50kHz.

We applied different band-pass filters to separate signals from the two speakers. In this way, four major paths were possible, as illustrated in Figure 1(b). Each frequency sweep lasted 12ms (corresponding to echo frame length: 600 samples). We experimented on different echo frame lengths and found out that 12ms yielded best performance. After receiving the signals, we calculated echo profiles as specified in EarIO [38] as the representation of patterns.

With this approach, the vertical axis of the echo profiles represents distance, with each pixel representing $\frac{1}{f_s} \times c$, where $f_s = 50$kHz denotes the sampling rate while $c = 343$m/s denotes the speed of sound. Bright strip on the echo profiles represent strong reflection at that certain distance.

In order to remove constant echo reflections from the environment and only focus on the deformations on the skin caused by silent speech, we calculated differential echo profiles by subtracting the previous echo frame from the current one. We stacked the four paths combinations as four channels. The differential echo profiles were used as the representation of facial movement patterns and fed into the following deep learning pipeline. An example of such representation can be found in Figure 1(c), where facial movements during different silent utterances are represented by different patterns in the differential echo profiles.

5.3 Deep Learning Model

We design a customized deep learning pipeline to decipher speech from facial movement patterns represented by echo profiles.

After echo profile calculation, facial movement patterns are already represented by four-channel images. Given its wide application and success in image processing, we use convolutional neural network (CNN) to decode silent speech from echo profiles. We experimented on adding temporal recurrent neural network (RNN) layers including long-short-term memory (LSTM) and gated recurrent unit (GRU) layers. However, they did not improve performance. We detail discussion on the network structure in Section 8.2. We use ResNet-18 as the backbone. The convolutional layers are followed by a one-dimensional average pooling - instead of performing pooling on both axes, we only perform pooling on the spatial axis. In this way, the temporal information is preserved. After this pooling step, the dimensions of the feature vectors become $\lceil T/16 \rceil \times 512$, where T is the original dimension of the time axis before going through the convolutional encoder. It is reduced to $\lceil T/16 \rceil$ during downsampling steps in the encoder. In this way, every 512-dimensional feature vector corresponds to a 16-frame block in the echo profile.

To adapt to variable sequence lengths, we adopt CTC loss. To achieve this, for each of the 512-dimensional feature vector from the encoder, we use a fully-connected decoder network with output dimension of $W+1$ (W distinct labels plus blank) to predict the label of the corresponding position. W is the number of distinct words in the command set. Note that W may not be equal to the number of commands since commands like "Hang up" are represented by two labels "hang" and "up". In the discrete speech recognition task, $W = 32$ while in continuous recognition $W = 10$.

5.4 Sliding-window Implementation

We hope that EchoSpeech can be used in a natural way without the need to segment silent utterances manually. Therefore, we used sliding window during evaluation. In this way, users can speak at different speeds and paces and speak or pause anytime as they wish. To achieve this, we adopted a sliding-window evaluation pattern. In this manner, the system does not rely on pre-existing segmentation that splits different silent utterances apart. Instead, the system automatically generates a prediction where there is an silent utterance detected and gives blank prediction when none detected.

During training, sliding-window was not applied to avoid confusing the model with incomplete silent utterances. We used single silent utterance and consecutive utterances that lasts no more than 800 echo frames (9.6s) to train the model to increase training samples as well as to improve the model's generalizability to variable utterance lengths.

During testing, a sliding window of size 192 echo frames (2.3s) with a stride of 16 echo frames was applied. We experimented on the window size during evaluation and found that windows of size from around 160 to around 800 yielded almost the same performance. We chose 192 for lighter computational cost. Every window of sample went through the same network as training. A prediction label was given for every 16-frame block in the window. We considered the label to represent the prediction at the corresponding location. Since the stride size 16 was smaller than window size 192, every 16-frame block will be covered by multiple windows. We performed a majority voting and assigned it with the label that appeared the most times among the windows covering that block. We then merged consecutive predictions with the same label and removed blank labels to generate the text prediction continuously.

5.5 Data Augmentation

During algorithm iterations, we analyzed and identified challenges that EchoSpeech faced with and tried to address them with data augmentation. Each data augmentation approach was proposed to address a specific challenge, which we detail below.

5.5.1 Merging Consecutive Silent Utterances. As specified in Section 5.4, sliding window was used during testing. With this approach, there is no guarantee that there is an silent utterance at the center of each window. To let the model see the utterances as well as the transitions between utterances, we merge consecutive silent utterances to form longer utterances with pauses in between. For instance, when a user said "One""Pause""Alexa" consecutively, we not only used samples such as "One""Pause""Alexa" for training, samples "One Pause", "Pause Alexa" and "One Pause Alexa" were also added to the training set. Such operation also improves EchoSpeech's ability to adapt to different speaking speeds, as the same window may cover different number or portion of silent utterances for different users. During training, consecutive silent utterances that last no more than 800 echo frames (9.6s) were added. This means that samples with 1 to 4-6 silent utterances were all included during training. Further increasing this window size leads to marginally improved performance, but significantly increased training time.

5.5.2 Random Noise. During training, all pixels were multiplied by a random factor between 0.95 and 1.05. This operation was adopted to increase variance in the samples and avoid over-fitting.

5.5.3 Random Padding. Random padding was applied for two purposes: 1) shifting the position of the samples on the time axis and 2) adapting the model to variable lengths. In order to increase efficiency, samples in a batch were expected to have the same lengths. Due to the vast variance of sample lengths (from less than 100 to 800), we further applied random padding to adapt the model to variable sample lengths. We first sorted all samples according to their lengths. We then took consecutive samples after sorting to form batches. In this way, samples in the same batch have similar lengths. During training, for each batch, in 50% cases, we simply pad all samples to the longest sample in that batch. For the other 50% cases, we pad all samples to a random length between the longest length and 800.

5.5.4 Scenario-specific Noise Addition. A robust SSI should be able to recognize speeches across different scenarios such as remounting the devices, walking, in noisy environments, etc. One method is to collect training data from participants in all these scenarios. However, it is not feasible nor practical. In order to make our system robust across various real-world scenarios, we synthesize training samples by adding scenario-specific noises.

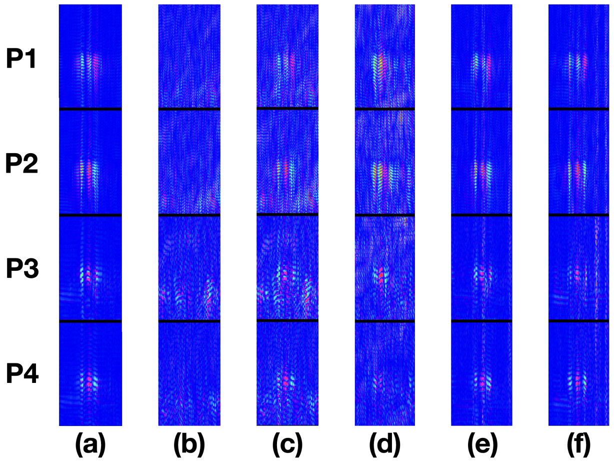


Figure 3: Synthesizing samples for walking and noisy scenarios. (a) Echo profiles of silent utterance "One" collected while sitting. (b) Echo profiles collected while walking but no speaking. (c) Linearly adding (a) and (b). (d) Echo profiles of silent utterance "One" collected while walking. (e) Echo profiles of "One" with injected restaurant noise. (f) Echo profiles of "One" after data augmentation.

We found EchoSpeech can capture clear echo reflection from the environments. If the user is static, since echos from the environment are constant, they can be easily removed while calculating the differential echo profiles. However, if the user is in motion (mobile setting), since the device itself is constantly moving, such echos will leave scenario-specific noises in the echo profiles, as demonstrated in Figure 3. We noticed that such noises were mostly linearly added to static echo profiles. Therefore, we applied data augmentation by randomly adding such noises to echo profiles in the static setting to synthesize echo profiles in the mobile setting.

To collect these noises during walking, researchers wore our device while walking without moving the lips. Using these data, we created noise profiles. During training, a random slice of noise profiles is multiplied by a random factor between 0 and 1 and then linearly added to training samples.

5.5.5 Acoustic Noise Addition. Similar to scenario-specific noises, acoustic noise in the environment can also pollute the signal. Please note that we did apply band-pass filtering. This removed most of environment noises, which mostly only occupies lower frequencies. However, certain noises can still extend beyond audible ranges and mix with our signals, such as silverware touching each other, items dropping on the ground, clapping, etc. Figure 3(e) demonstrates how such noises pollute the data.

The signals that we used are FMCW. While it is possible to decode them frame by frame to improve signal-noise-ratio. However, that will inevitably sacrifice spatial resolution. We adopted an approach similar to scenario-specific noises by recording common noises and linearly adding them into training samples to synthesize noisy data. A researcher recorded noises using EchoSpeech devices that includes the following scenarios: people talking and background music playing in a restaurant, vehicles passing by near a road, home appliances running (washer, dryer, fridge, air conditioner), tap water running. During training, 1-3 random slices of noises of random lengths were multiplied by random factors between 0 and 1 and then mixed and added to training samples at random position. After mixing, a synthesized sample is shown as Figure 3(f).

5.6 Two-step Training Scheme

We propose a two-step training scheme to minimize training effort for new participants as well as to improve performance. The system is still user-dependent, but for each new participant, instead of training a customized model from scratch, we only need to fine-tune the model trained with other people's data. In this way, the entire training process is divided in two steps: 1) pre-train a model using data provided by other participants. The model in this step is denoted as the user-independent (UI) model. And 2) fine-tune the UI model with the new participant's data. In practice, we found that this scheme improves performance and significantly reduces training time for new participants at the same time.

In order to evaluate our system with this approach, we first pre-trained a UI model for each participant using a leave-one-participant-out scheme. We then fine-tuned the UI model using different number of training sessions for each participant.

In both steps, we used an Adam optimizer with cosine scheduler and an initial learning rate of 0.0002. The batch size was set to 5. For the pre-training step, the model was trained for 100 epochs. For the fine-tuning step, the entire model was fine-tuned for 15 epochs.

6 USER STUDY

According to our design goal (2), we want to evaluate EchoSpeech with a setup that is natural and close to real-life applications. To achieve this, we first designed two sets of commands to examine EchoSpeech's ability in recognizing discrete and continuous speech. We also considered two most common use cases, choosing static (sitting at a desk) and mobile (walking) as the evaluation scenarios. We would like to first evaluate how well EchoSpeech works under these scenarios. Then we want to explore more on the practical implications, especially on how much data a user needs to provide before being able to use EchoSpeech in both scenarios. To better encourage natural way of speaking, we did not require users to speak slowly. Instead, we instructed users to speak at their normal speed and control the pace of the study themselves. We elaborate on these considerations in this section.

6.1 The Design of Silent Speech Vocabulary

An ideal silent speech recognition system should be able to recognize any words without limitations, similar to the current speech recognition based on voice. However, training such a system requires resources that are beyond the scope of a research paper involving new sensing hardware, as we need to collect all training data ourselves. Thus, our goal of designing the vocabulary is to strike a balance between usability and training practicality by evaluating EchoSpeech in designed application scenarios and demonstrating its use cases with a real-time demo. We decided to design the command sets used in the popular speech interaction scenarios.

6.1.1 Discrete Command Recognition. We chose our recognition commands for the following popular speech interaction scenarios, many of which have been used in previous silent speech research [71, 73], including 1) hands-free music player control; 2) interacting with smart devices; 3) digits input; 4) activation commands for voice assistants. In total, we have 31 commands for discrete silent speech recognition, as illustrated in Table 1.

6.1.2 Continuous Silent Speech Recognition (Connected Digits). On top of commands, we explored using SSI for continuous input. As discussed previously, continuous recognition is a challenging yet critical step towards adopting SSI in real-world applications. Instead of recognizing a set of pre-defined sentences [25], we are specifically interested in recognizing unseen combinations from existing vocabulary. We combine this task with voiceless passcode input/authentication. In this use case, users can silently utter a three- to six-figure passcode quickly without pauses in between. In total, there are 1,111,000 different possible combinations (from "000" to "999999"), which is impossible to be iterated and learned as a whole and can only be learned through breaking down silent utterances into words.

6.2 Main Study

The main study mainly examines how EchoSpeech works in the static environment (sitting at a desk), as well as in the mobile environment (walking) if no training data from the mobile environment is provided. The study was approved by our institute's Institutional Review Board (IRB). The main study was split into discrete and continuous *sections*, with the former focusing on the isolated commands while the later on connected digits. The study was conducted in a large room on campus. Each participant came in twice to finish the two *sections*, each lasting 70-90 minutes. For each *section*, 18 *sessions* of data were collected. Participants were instructed to remount the device (took off the device and put it back on) after each session. During data collection, instructions were presented on a laptop screen, which showed the participant the command they need to perform. The laptop's webcam was used to record videos of the session with a clear view of the participant's face for reference. Participants were instructed to "mouth the word silently with lip movements similar or slightly larger than how you would have moved your lips when speaking out loud".

The hardware used in the study are illustrated in Figure 4(a-e). Participants wore the glass-frame with sensors installed as specified in Figure 1(b). The sensors were connected to a Teensy 4.1 micro-controller with a customized audio amplifier add-on board and communicates with a laptop (Macbook Pro 2021) via a USB cable, as illustrated in Figure 4(e). At the start of each session, the laptop initiated a signal that started recording on Teensy. Meanwhile participants were instructed to clap their hands. We later manually found the clap in the video and audio to synchronize the audio with the ground truth. Clapping hands was required at the end of the session in case the one at the beginning was not captured. During data collection, instructions on the laptop screen included the silent utterance itself (in large font to make sure participants saw them clearly), progress bar of the current utterance (to let the participant know how much time was left before the system jumped to the next utterance), progress and estimated time left of the current session as shown in Figure 4(f). In the discrete *section*, the maximum duration for each silent utterance was 3 seconds. The system would jump to the next utterance after 3 seconds passed. In the continuous *section*, the participant had 4 seconds to finish each utterance. It was longer because the length of sequence varied with a maximum length of 6 digits long. However, actual utterances were much faster because the participants were told that they should use the space bar (or the right arrow) to jump to the next utterance once the they finished the current one. This was applied to encourage the participants to speak at their natural pace and speed instead of waiting and pausing. The participants were instructed to use the "x" key (or left arrow) on the keyboard, if they made a mistake in the utterance, wanted to adjust the device or wanted a pause (by constantly pressing). In these cases, the current utterance was repeated.

Silent utterances were given in random order. In the discrete *section*, each command were repeated 4 times in each session. In the continuous *section*, each session had 60 connected digits with sequence lengths ranging from three to six. These combination of digits were generated randomly so that each length (3 to 6 digits) had 15 occurrences and each digit (0-9) had the same amount of occurrences (27 times) in each session. In both *sections*, for sessions 1 through 13, participants sat naturally at a desk. For sessions 14 through 18, participants walked in the room. They were instructed to walk in the room in the way, path and speed as they wished. Sessions 1 and 14 were used for participants to get familiarized with the system and not used during training nor testing.

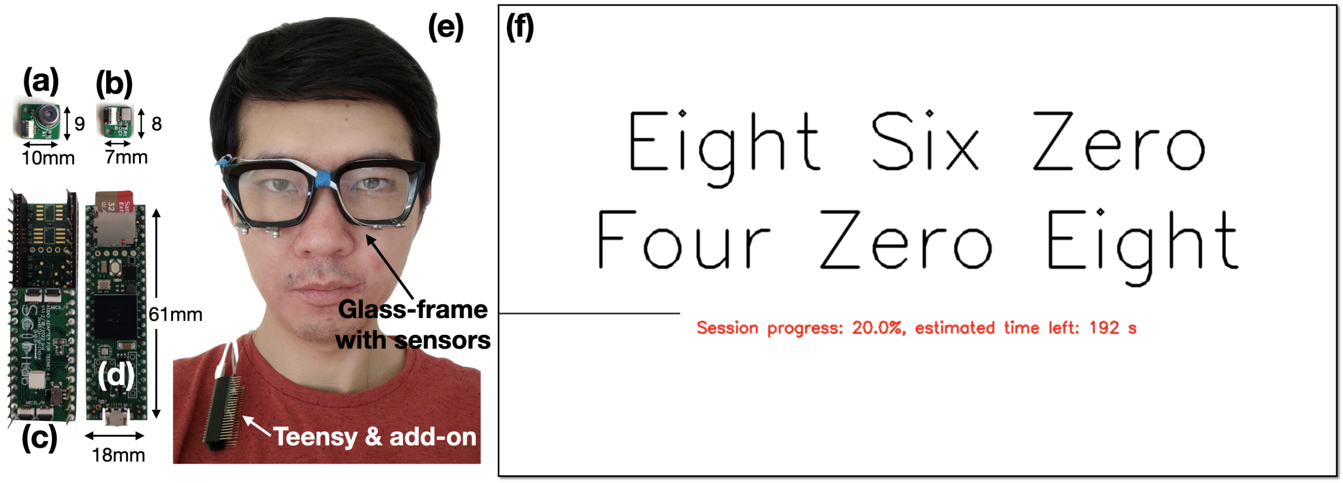
Each participant finished 2 *sections* (continuous and discrete). Therefore, there were 24 *sections* in total. In 12 of the 24 *sections*, participants were asked to hold the laptop in their arms while walking. In the other 12, the laptop was placed on a moving table so that the participant could push it around while walking. This change was adopted because some participants reflected that the laptop was too heavy as well as to increase variance in the way of walking. No significant difference in performance was observed between these two walking styles.

6.3 Followup Study

The followup study was conducted after the main study by 12 users that did not participate in the main study. The followup study

Table 1: Command set for discrete command recognition

Scenario	Commands
Hands-free music player control	Play, Stop, Resume, Pause, Previous, Next, Volume
Interacting with smart devices	Left, Right, Up, Down, OK, Cancel, Menu, Dial, Hang up, Open, Close
Digits input	Zero, One, Two, Three, Four, Five, Six, Seven, Eight, Nine
Activating voice assistant	Hey Google, Hey Siri, Alexa



Eight Six Zero Four Zero Eight

Session progress: 20.0%, estimated time left: 192 s

Figure 4: User study setup. (a) Speaker board. (b) Microphone board. (c) Teensy & add-on board (Add-on board's side) (d) Teensy and add-on board (Teensy's side). (e) Participants wearing the glass-frame with sensors and Teensy. (f) Screenshot of the instructions during the study.

mainly focused on the mobile environment. The main purpose was 3 fold: 1) provide more data to conduct thorough evaluation on the mobile environment, 2) improve performance on the main study with new data and analysis, and 3) explore directions for future optimization on the mobile environment.

The followup study shared almost the same configuration and procedures as the main study except that there were 17 sessions. Participants walked in the room for sessions 1 through 13 and sat at a desk for sessions 14 through 17. Session 1 was used as practicing session. Since participants finished the walking sessions first, they did not need to practice again for sitting sessions. In addition, the laptop was always placed on a moving table. It is worth noting that the followup study and the main study were conducted in different rooms. The room for the main study was quiet and had carpeted floors. The room for the followup study had a noisy ventilation system and hard concrete floors.

6.4 Dataset Characteristic

In the main study, 12 participants (all college students, 5 self-identified as male, 7 female, average age 23.5: from 18 to 32, std 4.4) were recruited. Hardware malfunction happened twice (broken cable on P3, SD card full on P11) but the participants returned to redo the lost sessions. On average, each session lasted 3.3 minutes in the discrete section, and 3.0 minutes in the continuous section. After removing the practice sessions, 23808 valid silent utterances were collected for the discrete section, in which 17856 were collected when participants were static, while 5952 were were collected when participants were in motion. 11520 valid silent utterances (51840 digits uttered) were collected for the continuous section, in which 8640 (38880 digits) were collected when participants were static, and 2880 (12960 digits) were collected when they were in motion. Data collected in the main study is denoted as the main dataset in later text.

Allowing the participants to finish each silent utterance early using keyboard reduced study time. On average, participants spent 1.51 seconds on each discrete silent utterance and 2.67 seconds on connected ones. Variance is also observed: the slowest participants spent 1.83s while the fastest one spent 1.22s on each discrete silent utterance. On continuous silent utterance, the difference is 2.27s (fastest) and 3.09s (slowest).

The followup study only included the discrete section. 12 participants that did not participate in the main study were recruited (5 self-identified as male, 7 female, average age 25.4: from 19 to 35, std 5.0). Each session lasted 2.6 minutes on average. Participants spoke generally faster. The average silent utterance duration was 1.20 seconds (fastest: 0.93s, slowest: 1.53s). Data collected in this study forms the followup dataset.

7 RESULTS

In this section, we describe the experiments conducted to evaluate EchoSpeech. We start by presenting the evaluation metric. We then present the experiments conducted on the main dataset and the followup dataset respectively. After that, we present further experiments and analysis to improve performance and reduce training effort.

7.1 Evaluation Metric

Word Error Rate is commonly used in speech recognition-related tasks. Compared with accuracy, word error rate works better at gauging continuous predictions. For instance, for sequence " Volume up", if the prediction is "Volume", using accuracy as metric will treat the prediction as wrong while the WER will be 0.5, better reflecting that the model finishes half the job. This is especially useful in longer sequences. WER of around 5% is usually considered human performance and is acceptable in conversations [53, 66].

We calculate the metric in the unit of each silent utterance as recorded during the user study. For each silent utterance, we compare the text prediction generated using the sliding window approach as described in Section 5.4 with the ground truth and calculate WER as

$$WER = \frac{S + D + I}{S + D + C}$$

, where S, D, I and C are the numbers of substitutions, deletions, insertions and corrected words, respectively.

7.2 Main Study

Experiments on the main study mainly examines how EchoSpeech works in recognizing discrete and continuous silent speech in the static environment. In addition, we also utilize the mobile sessions in the main study to evaluate how EchoSpeech works while walking without providing training data from the mobile environment.

7.2.1 Discrete Speech Recognition. We evaluated EchoSpeech's capability to recognize discrete speech using the algorithm pipeline as illustrated in Section 5. We adopted the two-step training scheme, training a leave-one-participant-out (LOPO) UI model for each participant first, and fine-tuning the UI model using the same participant's data. To remove random factors, we performed a 6-fold cross-validation on the 12 static sessions. To achieve this, we chose 2 sessions as testing (sessions 2,3, sessions 4,5, ..., sessions 12,13) and used the remaining 10 sessions to fine-tune the model. We used the model after the last epoch to evaluate. Results across participants are illustrated in Figure 5(a). The average WER across 12 participants was 4.5%, ranging from 1.0% (P12) to 13.7% (P3), std=3.5%. We specifically looked into the two participants with worst performance (P3, P11), we found out that they both pushed the glass-frame multiple times as the glass frame frequently slipped down their nose during the study. Since the sensors of EchoSpeech were pointing downwards, pushing the glass-frame introduces significant noise in the signals. We acknowledge this limitation in Section 8.7.

7.2.2 Continuous Speech Recognition. We evaluated EchoSpeech's capability in recognizing continuous speech using similar scheme to that of discrete speech. Results indicate that the average WER across 12 participants is 6.1%, ranging from 2.1% (P12) to 16.3% (P11), std=4.2%. Results for each participant are shown in Figure 5(b). Similar to that of isolated digits, pushing glass-frame causes most issues in participants with worst performances (P3 and P11). For P11, an additional observation was that the participant made some mistakes but did not remove them from the samples. This was observed by watching the recorded study videos. It was difficult to remove all such bad samples, since the utterances were silent and difficult to make out from the video due to limited lip movement and fast speaking speed. If P11 is removed from the study, average WER was improved to 5.2%.

7.2.3 Reducing Training Effort. The results presented in previous sections were all achieved with 10 sessions (around 30 minutes) of training data collected from the same user. Compared with previous work with comparable level of performance [25], this is already significant advancement. For instance, EarCommand requires over 100 minutes of training data from new user to achieve around 10% WER on 32 commands [25], SpeeChin requires 40-50 minutes to reach 10% WER on 54 commands and only works when the user is sitting still [71]. However, we wish to further minimize training effort from new participants, pushing towards higher practicability. Therefore, we conducted several experiments to demonstrate how to minimize training effort from new participants.

We demonstrate that users can provide as little as 2 sessions (6-8 minutes) and still achieve decent performance. We experimented with different number of fine-tuning sessions. When no data was used for fine-tuning, the system works user-independently. In this way, WER for recognizing discrete and continuous speech is around 40%. Results in Figure 6(a) shows that performance improves with more training sessions applied, while flattens after about 4 sessions of data applied. With only 2 sessions of training data, EchoSpeech is already able to recognize 31 isolated commands or 3-6-figure connected digits with 9.5% and 14.4% WER, respectively.

7.2.4 Mobile Performance with No Training Data from the Mobile Environment. Using the 4 mobile sessions from each user, we evaluated how EchoSpeech performs when the user was in motion without providing training data while walking. This can minimize training effort. We applied data augmentation as specified in Section 5.5.4 by adding motion noises to static data and evaluated EchoSpeech without any training data collected in mobile settings. The motion noises were collected by researchers at different locations from the study. We trained a model using data collected from all participants in the static setting, applied data augmentation, and evaluated on all participants' data collected in the mobile setting. Results show that average WER across 12 participants is 16.8% for both discrete and continuous speech(std: 10.3%, 11.0%, respectively), as demonstrated in Figure 5(b). This result is significantly worse than the static setting. We discuss directions and methods to improve this result in the following sections.

7.3 Followup Study

As discussed earlier, the followup study was conducted to provide more data for thorough evaluation on the mobile environment, improve the mobile performance in the main study, and explore directions for further optimization.

7.3.1 Discrete Speech Recognition When User is in Motion. The richness of new data allowed us to train a user-dependent model similar to that in the main study (Section 7.2.1 and 7.2.2) to explore the limit of EchoSpeech when the user was in motion. Adopting the same two-step training scheme as described in Section 7.2.1,

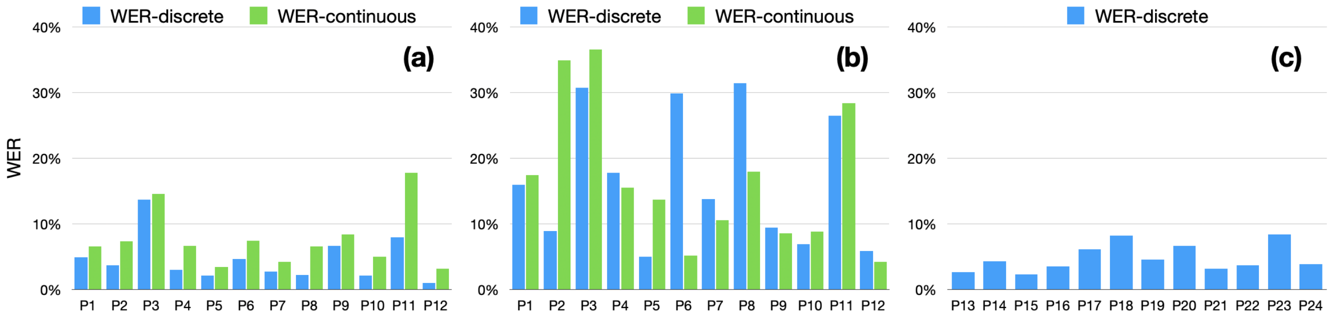


Figure 5: Discrete and continuous speech recognition performance. (a) Performance in the static setting. (b) Performance in the mobile setting. (c) Performance in the mobile setting on new participants from the followup study.

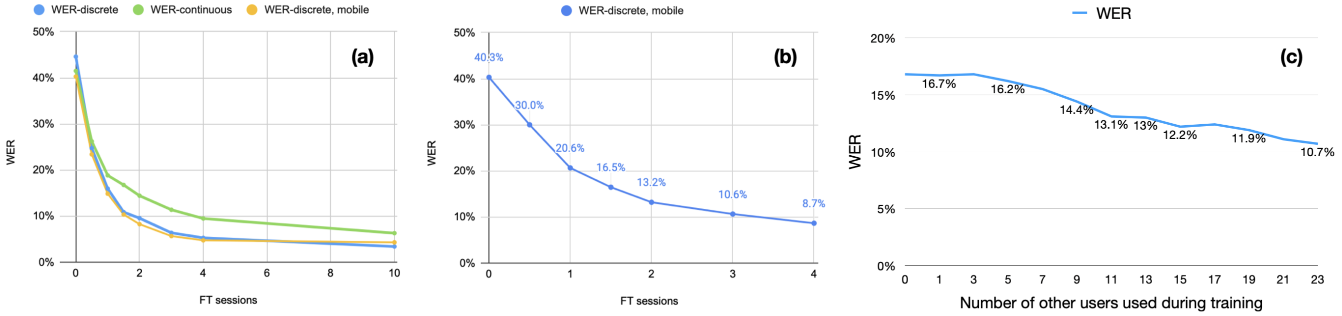


Figure 6: Impact of the amount of data used for fine-tuning. (a) Using data from the same user collected in the same setup to fine-tune the model. (b) Using data from the same user collected in the static setup to fine-tune the model for the mobile setup. (c) Adding other people’s data to training to improve performance.

we trained a LOPO UI model for each user first, then fine-tuned it with 10 mobile sessions from the same user. Results in Figure 5(c) demonstrate consistent performance compared with when the user was static: average WER is 4.7% (std 2.1%), ranging from 2.3% (P15) to 8.4% (P23). This result demonstrate that with enough training data, EchoSpeech can adapt to different scenarios with robust performance.

To reduce training effort, we experimented with different number of fine-tuning sessions similar to the main study. The results are also similar. Figure 6(a) shows that performance improves with more training sessions applied, while flattens after about 4 sessions of data applied. With only 2 sessions of training data (about 6-8 minutes), EchoSpeech is already able to recognize 31 isolated commands in motion with 8.2% WER (std 2.5%).

7.3.2 Improving Mobile Performance in the Main Study with Other Users’ Data. In Section 7.2.4, performance on the mobile environment was significantly worse even after applying data augmentation. We demonstrate that even without new data from the same user, this performance can be improved by incorporating other users’ *walking* data into training. Adding 11 other users’ walking data into training from the original study, performance on discrete and continuous speech recognition improved from 16.8% to 13.1% and 13.2%, respectively. Continue to add the 12 new users’ data from the followup study, performance on discrete speech recognition further improved to 10.7%. We visualize the relationship of other people’s walking data used and performance in Figure 6(c). From the figure, performance steadily improved with more other users’ data added. This also points a future direction for further optimizing performance and reducing training effort with scaled up data collection.

7.4 Further Improving Mobile Performance with Minimal Training Effort

Inspired by the promising improvement from Section 7.3.2, we explored further reducing training effort. We conducted evaluation on the followup study, utilizing all available data from other users. With this approach, for each user, with 4 static sessions as training and no mobile sessions, EchoSpeech achieves 8.7% (std 2.8%) in recognizing 31 discrete commands. Adjusting the number of static sessions yields the curve in Figure 6(b). With only 2 static sessions, performance can still reach 13.2%.

Combined with results in Section 7.2.4, this means that a new user only needs to provide 6-8 minutes of static training data to use the system in both static and mobile environments with decent performance. Additionally, with potential large-scale deployment in the future, this performance can be further improved and the training effort can be further reduced. We believe that this is a solid step towards a practical SSI in daily life.

7.5 Recognizing Silent Utterances with Different Lengths and Speed

While recognizing continuous speech, we are interested in how the errors distribute among sequences with different lengths. We calculated the average WER for 3-6 digit silent utterances separately. We plot the recognition performance for silent utterances with different lengths as well as their average duration in Figure 7(a). Results demonstrate that although longer sequences take more time and have more syllables, no significant performance discrepancy is observed in the recognition performance.

In addition, we also examined the impact of speaking speed on the performance. We plot the average silent utterance duration for each session and its corresponding performance as scatter points in Figure 7(b)(c). Results show that for discrete speech, performance decreases if the duration is shorter than 1.4s and stays steady if participants spoke slower. For continuous speech, no significant trend is observed.

7.6 Impact of Sensor Positions

As specified in Section 4, EchoSpeech utilizes 2 pairs of speakers and microphones, enabling four major signal paths. In order to examine how each path contributes to the performance, we conducted an experiment isolating each path by applying different band-pass filters. We adopted one-step training to save time. Sessions 12-13 of all participants were used as testing and sessions 2-11 of all participants were used for training. Performance in Figure 8 indicate that performance degrades from 5.8% to 12.0% on average if only one path (one speaker and one microphone) is used. Path 1 and path 2 are more useful than path 3 and 4. This confirms that four paths all contribute positively to the system. However, considering the device size, cost, computation load and power consumption, such degrade can be acceptable in certain scenarios. We demonstrate in Section 8.1 that a low-power variation of EchoSpeech with only one speaker and two microphones can operate at as low as 73.3mW and run in real-time on a smartphone.

7.7 Noise Injection

In order to examine how EchoSpeech works in noisy environments, we experimented with noise injection by mixing noises into the data we collected. A researcher used the same device as used in the user study to record two types of noises: 1) street noise. The research walked along a busy street for 5 minutes. Cars passed by frequently. Using the NIOSH Sound Level Meter App [1] on an iPhone 12, the noise level was 64dB(A). 2) restaurant noise. The researcher went into a noisy restaurant and recorded for 5 minutes. Background music and crowd chatting could be heard. The same app measured the noise level at 76dB(A). A spectrum analysis of the injected noises and our signals are visualized in Figure 9. It is clear that most of the noise components are below the range of our signals. For those that do overlap, the amplitute of our signal is much stronger than the noise.

We mixed the two types of noises into each session collected during the user study. We used models trained and fine-tuned on clean data and directly tested them on noisy data. Results in Figure 10 show that performance in static setting slightly decreases around 2% for street noise and around 3% for restaurant noise. However, there was almost no change in performance in the mobile setting. We hypothesize that noisy patterns on the echo profiles from walking makes the model robust to different noises on the echo profiles, thus providing extra resilience against environmental noises.

We then applied the data augmentation as described in Section 5.5.5. One researcher collected the noises from different places. The data augmentation was applied during the fine-tuning stage without the need to re-train the model. After applying data augmentation, EchoSpeech becomes even more resilient against acoustic noises, yielding almost the same performance as when no noise was present.

8 DISCUSSION

In this section, we discuss additional findings and analysis as well as potential opportunities and challenges.

8.1 Applications and Real-time Demo

EchoSpeech can be used as an alternative hands-free and eyes-free input method in a variety of applications. We implemented our system in several sample applications and demonstrate them in our real-time demo video. Please note that these functions can all be achieved by using traditional speech interfaces. However, traditional speech recognition requires the user to speak aloud which is frequently inconvenient or socially inappropriate. EchoSpeech provides a new input form, that can be integrated with other existing interactive technologies or used alone. We discuss these two cases with two promising applications that could be immediately available, respectively. In addition, we demonstrate several further possible use cases of EchoSpeech that could lead to improved user experiences in the future.

Using CAD software: Software like CAD usually involves many options, configurations and dimensions. It can be challenging to incorporate them with the natural way of design - directly drawing on canvas because users need to switch between options, specifications and the graph itself constantly. Furthermore, users often need to work at quiet places (e.g. library, lab) while working on these design tasks. Thus, speech recognition is not feasible. EchoSpeech provides an option to use silent utterance as an extra interface, without disturbing others with voice. We demonstrate this use case by drawing basic shapes on CAD. Here EchoSpeech provides an additional input modality that can be used together with existing stylus input smoothly. The user first chose the types of shapes with silent commands. Then he naturally drew the shape with a pencil and used silent commands to specify the dimensions.

Silent utterance in mobile use: While in motion, it is usually inconvenient and even dangerous to engage in interactions with hands and/or eyes. In such cases, uttering the interaction intention can be a good alternative. EchoSpeech provides a solution where users have access to a hands-free and eyes-free interface without needing to speak out loud, which can be particularly suitable for mobile use.

[1] https://www.cdc.gov/niosh/topics/noise/app.html

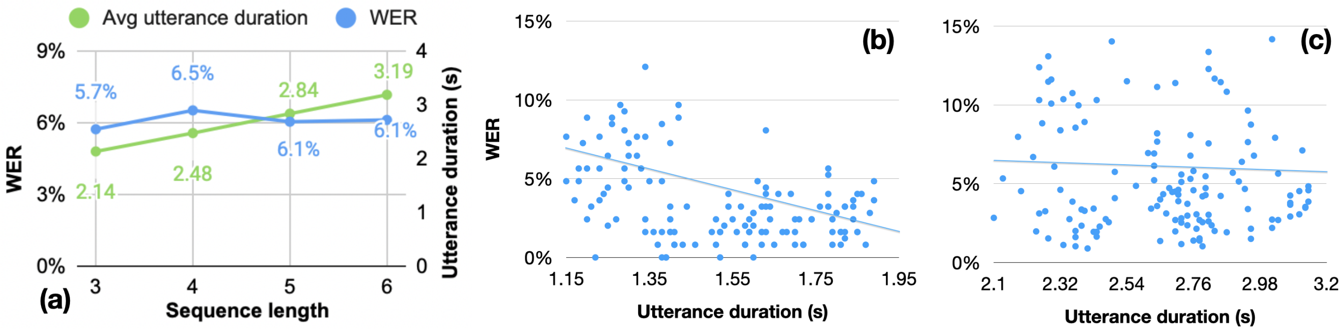


Figure 7: Recognizing speech with various lengths and speed. (a) Performance on sequences with different lengths. (b-c) Performance on sessions with different speaking speed: (b) discrete speech, (c) continuous speech

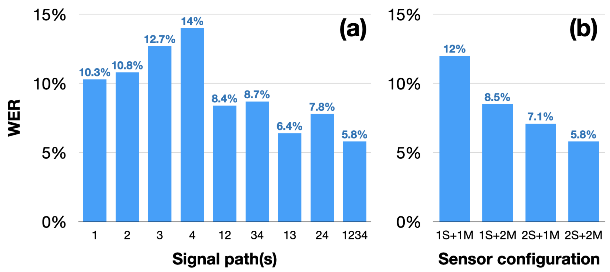


Figure 8: Impact of different signal paths and sensor configurations. Signal paths illustrated in Figure 1(b). 1S+1M: using one speaker and one microphone, etc.

In addition to the applications discussed above, we also demonstrate that EchoSpeech can be used in certain scenarios as a replacement or improvement to existing interfaces. These scenarios particularly focus on cases where the users have occupied hands or inaccessible devices. We acknowledge that the following use cases may still be limited at this stage, we present them as a basic illustration of EchoSpeech's potentials. In addition, we believe that with future engineering effort, EchoSpeech system can be fully integrated into real glasses so that the sensors are invisible from appearance. In this way, EchoSpeech can be used with minimal level of social awkwardness, thus making the following use cases possible.

Controlling music player: EchoSpeech can be integrated with earphone/headsets to control music players. Voice has already been used to control music players. EchoSpeech provides an alternative approach without making sound, which could expand the use cases of voice music player control. We demonstrate this use case in our demo video. To clearly demonstrate how EchoSpeech worked live, the smart phone was placed at a desk where it is playing music out loud so that the action on the screen and the sound can all be recorded.

Assisting text input on mobile phones: Inputting punctuation and symbols using the keyboard on a smartphone is not very convenient, as it requires users to switch to secondary keyboards. In such cases, if these keys can be silently mouthed, then users can keep their focus on the main input without switching between keyboards. We demonstrate this use case with inputting a short equation that involves numbers and special marks. In the future, with a larger vocabulary, it is also possible to integrate more words and functions to realize a dictation-style input interface that is even more natural and smooth. We leave the development for future work.

We developed these real-time demos using a low-power variant of EchoSpeech and deploy the processing pipeline on a smartphone. We employed the wireless module with nRF52840 micro-controller for Bluetooth Low Energy (BLE) data transmission. Since the module only supports one-channel audio transmission, we used the one speaker (S1) and two microphones (M1 + M2) setup. With this setup, we measured the power consumption of the entire module while transmitting data via BLE using a Current Ranger [2]. Results show that the system operates at 73.3mW (3.96V, 18.5mA).

We implemented the data processing and deep learning pipeline on an Android phone (Xiaomi Redmi K40) with the help of PyTorch Mobile [3]. For each demo application, one researcher collected a small amount of training data (1-6 minutes) with a command set including all the desired commands. The phone handled all processing and prediction and transmitted results to an ESP32 [4] that registered itself as a Bluetooth keyboard. The ESP32 analyzed the predictions and sent corresponding action keys to the device that it paired with. For instance, when controlling a music player, the ESP32 paired with the phone that is playing music. When it received the command "Next", it sent a "Next song" key to the phone.

8.2 Selection of the Deep Learning Model

EchoSpeech uses a CNN-based model for both discrete and continuous silent speech recognition. Recent work on similar tasks [25, 26, 71] also attached Recurrent Neural Networks (RNN) such as LSTM or GRU layers at the end of CNN layers to extract temporal patterns. We actually experimented on such CRNN models including attaching LSTM and GRU after the CNN encoder in our exploration. Our early results indicated that GRU works better than LSTM. In most cases, CNN and CRNN networks work similarly. However, CNN converges faster than CRNN with same number of epochs. In

[2] https://lowpowerlab.com/guide/currentranger/
[3] https://pytorch.org/mobile/home/
[4] https://www.espressif.com/en/products/socs/esp32

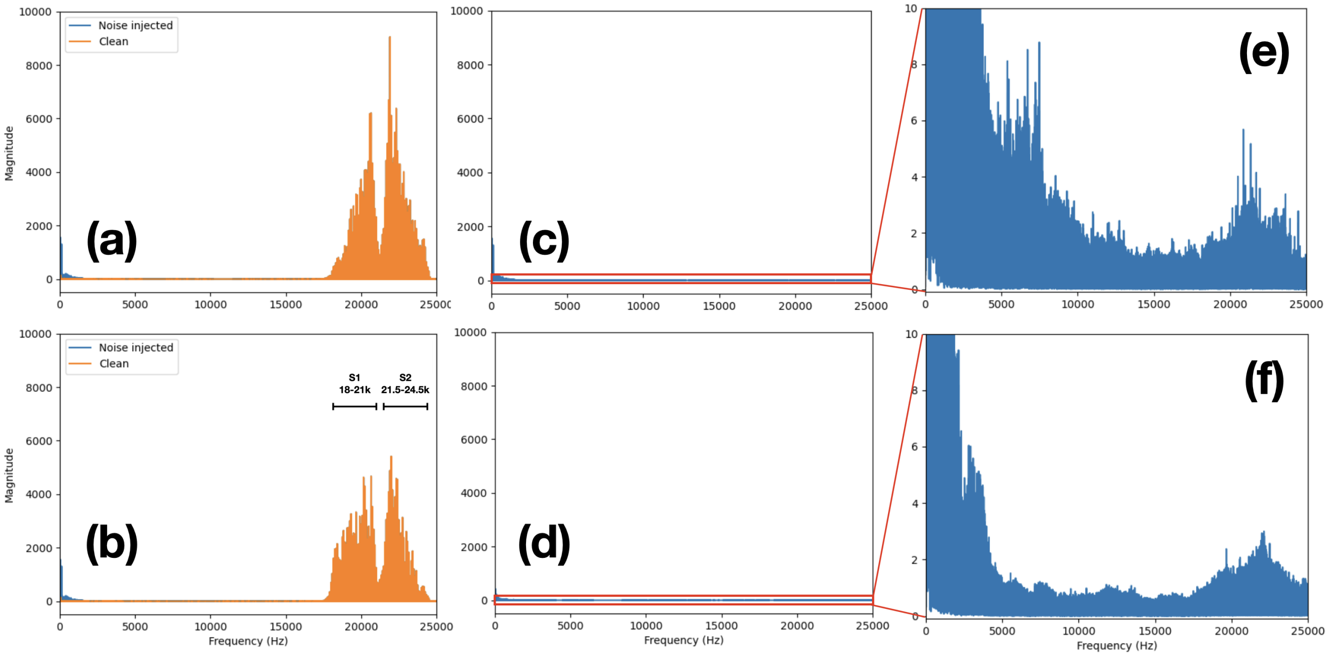


Figure 9: Spectrum of the signals and noises. (a-b) Signals recorded using microphone M1 and M2. M1 is physically closer to the speakers and thus having slightly stronger echos. Both the original signal and the signal after injecting restaurant noise are plotted. (c-d) Spectrum of the restaurant noise and street noise. (e-f) Zooming in on the spectrum of the restaurant noise and street noise.

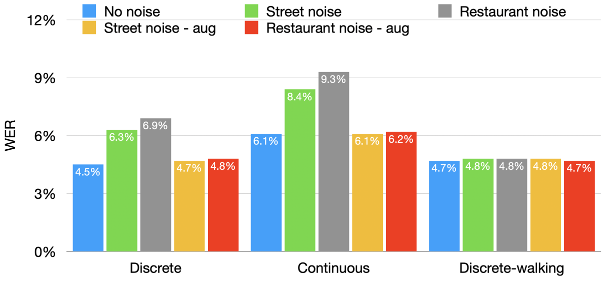


Figure 10: Impact of noise injection. Two types of noises were added to the data

certain cases, the latter had trouble converging. Also, CNN without RNN layers needs less computational resources and runs faster. For these reasons, we used the CNN model without RNN layers. We attribute the success of CNN to our feature representation and customized pooling strategy. With echo profile calculation, we converted all temporal features into spatial ones. For instance, speech at different speed will be reflected in the echo profiles as having different lengths. Our one-dimension average pooling preserves temporal information and enables the network to cope with silent utterances with variable lengths without loosing information.

8.3 Adjusting Speaker Power

In order to minimize power consumption as well as to reduce the potential impact on the user and the surrounding environment, we conducted an experiment on speaker power. One researcher collected data with 10 different signal amplitude configurations, ranging from 0.67% to 100% (using the amplitude used during the user study as 100%). For each amplitude configuration, 8 sessions of data of a smaller commandset (10 digits) were collected. For each amplitude, 8-fold cross-validation were performed to minimize randomness. All 80 sessions were collected in random order to make sure that results are maximally directly comparable. The data was collected in a quiet environment while we injected noises later. Noise augmentation was also performed. Results in Figure 11 show that in a quiet environment, even with very low amplitude, the system still worked reliably. However, when the restaurant noise was injected, performance significantly degraded. The larger the amplitude, the less degradation happened. When noise augmentation was applied, performance significantly improved except for very low amplitude. The flattening point was between 10% and 20%. For speaker amplitude greater than 20%, the system performance was basically not impacted by noises when data augmentation was applied. At 20% amplitude, each speaker roughly consumes 1.2mW, compared with 28mW at 100%. This further reduced the system's power signature to around 50mW with both speakers, compared with 73.3mW with 1 speaker.

In the real-time demo, EchoSpeech operates at 73.3mW, which can last for over a day on AR glasses such as Google Glass [5], Espon Moverio [6], and Microsoft HoloLens [7] which all have a battery size of over 800mAh. Adjusting the amplitude further reduced power comsumption to around 50mW. Further reduction of power consumption can be achieved by adjusting the duty cycle. For instance, it is possible to adjust EchoSpeech to operate at low sweeping rate when not actively used. Once an activation is detected, the system can be turned on at full speed.

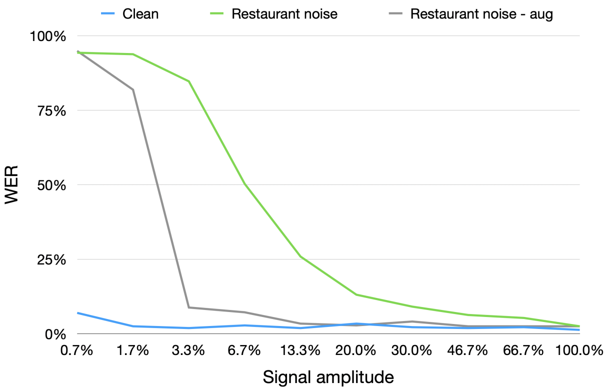


Figure 11: The performance of EchoSpeech with different speaker power strength. Please note that the horizontal axis is not presented in a linear scale.

8.4 Impact of Native/Non-native Speakers

Some studies found out that native speakers tend to perform better in silent speech tasks [39]. In our study, all participants were fluent with English, but only P7 and P12 were native speakers. While their performance was indeed better than average (average WER 1.9% vs. 4.6% on isolated commands), the sample size is too small to draw any conclusion. We believe that just like voiced speech, SSI should be developed for all fluent speakers regardless of whether they are native speakers or not. We leave further analysis on the impact of native/non-native speakers for future exploration.

8.5 Health Implications on Ultrasound Exposure

EchoSpeech uses near ultrasound as the sensing medium. NIOSH recommendation of noise exposure mainly focuses on noises below 16kHz, which we did not use. A review focusing on airborne ultrasound exposure recommends 75-85dB SPL as the limit for long-term exposure for frequencies near 20kHz [22]. To investigate the sound pressure level of our system, one researcher wore the device and placed a microphone near the edge of the left ear canal - the one closer to the speakers. At 100% amplitude, the RMS value of the recorded sound is 1107 (-26.4 dB FS). Taking the sensitivity (-26 dB FS @ 1kHz, 94 dB SPL) and frequency response (+15dB at 20kHz) data[8] into account, the estimated intensity at the ear canal is 78.6 dB SPL, near the edge of the strict versions of recommendations. However, as discussed in Section 8.3, reducing the amplitude to 20% has little impact on performance. At 20% amplitude, the RMS value of recorded sound is 167 (-42.8 dB FS), and the estimated intensity is 62.2 dB SPL, well below recommended levels.

Our signal ranges from 18 to 24.5kHz. To most adults, this range is not audible. However, it may still be audible to kids and certain animals. To minimize impact on the environment, EchoSpeech can be used in an activate-to-speech way. For instance, it is possible to define an activation gesture such as nodding and integrate an IMU module on the glass-frame to detect system activation. In this way, both power and computational resources can be saved, meanwhile bringing less disturbance into the environment.

8.6 Impact of Form Factor and Shape of Face

Different sizes of glass-frame may have some impact on the performance in that the signal paths will be different. Intuitively, a lager glass-frame usually has lower edges, which makes the sensors closer to the mouth. However, based on our limited experiments, we did not observe discrepancy in different glass-frames. During early stages of the exploration, researchers experimented on three different glasses: a small one that tightly fits the face, a light yet large one with lower edges, and the one used in the user study, as shown in Figure 12. All glass-frames were commercial products purchased online or at a local store. We chose the current one because its size is easier to fit different head shapes: the small one was too small for some people while the large one slid down easily.

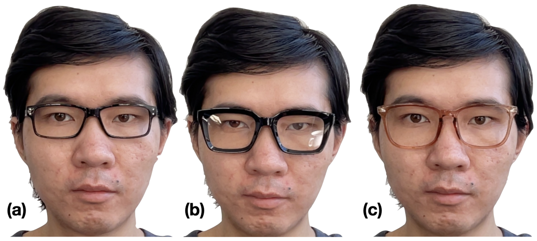


Figure 12: Glassframes used in pilot studies. (a) Small frame. The frame is light and small. However, for some people the lower edge is too close to the skin. (b) Medium frame. This frame was used in the user study. (c) Large frame. The frame is light yet big. For some people it is too loose and slides down easily.

We also examine possible impacts of different shapes of face. We measure the height and width of participants' face from the video and compared them against the size of the glass-frame to obtain the actual sizes. We draw the relationship between the height/width and performance of the participant in Figure 13. From the figure it is not evident that face shape has any correlation with performance. However, the current sample size (12) is very small to draw any conclusion.

[5] https://www.google.com/glass/tech-specs/

[6] https://mediaserver.goepson.com/ImConvServlet/imconv/b1cac7eaccf8017600cf8e0ac112f5403b86e4de/original?assetDescr=Moverio_BT-35ES_Glasses_and_Intelligent_Controller_Specification_Sheet_CPD-60652R1.pdf

[7] https://www.niora.net/en/p/microsoft_hololens

[8] https://invensense.tdk.com/wp-content/uploads/2016/02/DS-000069-ICS-43434-v1.2.pdf

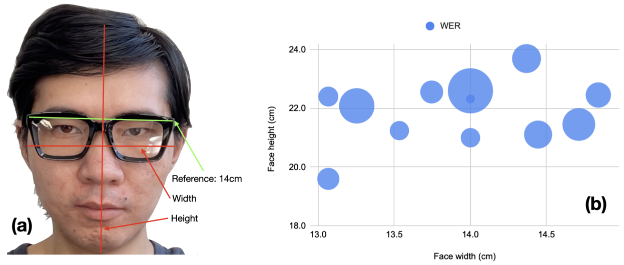


Figure 13: Impact of face shape. (a) Illustration of measurement. The width of the glass-frame served as the reference. (b) Distribution of face size and WER. WER value represented in size of the bubble.

8.7 Limitations and Future Work

Like any other systems, EchoSpeech has its own limitations. We list the limitations we have identified as well as future directions of optimizing the system.

8.7.1 Pushing Glasses. A major limitation is that EchoSpeech does not work well when objects get too close to the sensors, such as pushing glasses with fingers. We believe that this issue might be improved by applying data augmentation. However, we also believe that such a limitation might be acceptable. Research already show that users are willing to tolerate more errors for silent speech system [46]. Pushing glass as to SSI can be comparable to coughing or sneezing as to voiced speech interfaces.

8.7.2 Device Stability. Several participants reflected that the glass-frame was not particularly stable during the study (P7, P8, P11). They all have relatively small faces. This issue may negatively impacts performance. We believe that more glass-frames size options or personalized glass-frame can mitigate this issue.

8.7.3 Activating the System. Activating the system before use can save significant power and computational resources. Although EchoSpeech uses a segmentation-free pipeline that automatically detects the start and end of speech, it was not evaluated specifically for activation purpose. To be used with activation, the system needs to tell certain activation gestures apart from various other activities even including speech. We leave this exploration for future work.

9 CONCLUSION

We present EchoSpeech, a minimally-obtrusive contact-free silent speech interface on a glass-frame that can recognize both discrete and continuous speech. EchoSpeech strives to address the key challenges faced by wearable SSIs by placing two pairs of speakers and microphones on either sides of a glass-frame. Such configuration allows EchoSpeech to capture subtle yet highly-informative skin deformations with acoustic sensing at a close-up yet comfortable position. A customized deep learning pipeline enables EchoSpeech to recognize discrete and continuous speech without segmentation. Our user study with 12 participants shows that EchoSpeech achieves a WER of (std 3.5%) and 6.1% (std 4.2%) on recognizing 31 isolated commands and 3-6 figure connected digits, respectively. Further evaluation demonstrates EchoSpeech's robustness across different scenarios such as walking and injected noises. Finally, we demonstrate with a real-time demo that operates at 73.3mW with pipelines running on a smart phone to demonstrate EchoSpeech's use cases in demo applications.

ACKNOWLEDGMENTS

The authors would like to thank all participants for their participation in the user study. We would also like to thank all reviewers for their valuable comments. In addition, we would like to thank Cornell Information Science Department for supporting this project.

REFERENCES

[1] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani. 2018. Lip2Audspec: Speech Reconstruction from Silent Lip Movements Video. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2516–2520. https://doi.org/10.1109/ICASSP.2018.8461856
[2] Najwa Alghamdi, Steve Maddock, Ricard Marxer, Jon Barker, and Guy J. Brown. 2018. A corpus of audio-visual Lombard speech with frontal and profile views. *The Journal of the Acoustical Society of America* 143, 6 (2018), EL523–EL529. https://doi.org/10.1121/1.5042758 arXiv:https://doi.org/10.1121/1.5042758
[3] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. 2016. LipNet: Sentence-level Lipreading. *CoRR* abs/1611.01599 (2016). arXiv:1611.01599 http://arxiv.org/abs/1611.01599
[4] Abdelkareem Bedri, Himanshu Sahni, Pavleen Thukral, Thad Starner, David Byrd, Peter Presti, Gabriel Reyes, Maysam Ghovanloo, and Zehua Guo. 2015. Toward Silent-Speech Control of Consumer Wearables. *Computer* 48, 10 (2015), 54–62. https://doi.org/10.1109/MC.2015.310
[5] Linnar Billman and Johan Hullberg. 2018. Speech Reading with Deep Neural Networks.
[6] Lam A. Cheah., James M. Gilbert., Jose A. Gonzalez., Phil D. Green., Stephen R. Ell., Roger K. Moore., and Ed Holdsworth. 2018. A Wearable Silent Speech Interface based on Magnetic Sensors with Motion-Artefact Removal. In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies - BIODEVICES,*. INSTICC, SciTePress, 56–62. https://doi.org/10.5220/0006573200560062
[7] Tuochao Chen, Benjamin Steeper, Kinan Alsheikh, Songyun Tao, François Guimbretière, and Cheng Zhang. 2020. C-Face: Continuously Reconstructing Facial Expressions by Deep Learning Contours of the Face with Ear-Mounted Miniature Cameras. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 112–125. https://doi.org/10.1145/3379337.3415879
[8] J Chung and A Zisserman. 2017. Lip reading in profile. *British Machine Vision Conference, 2017* (2017).
[9] Joon Son Chung and Andrew Zisserman. 2017. Lip Reading in the Wild. In *Computer Vision – ACCV 2016*, Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato (Eds.). Springer International Publishing, Cham, 87–103.
[10] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America* 120, 5 (2006), 2421–2424. https://doi.org/10.1121/1.2229005 arXiv:https://doi.org/10.1121/1.2229005
[11] Thomas Le Cornu and Ben Milner. 2015. Reconstructing intelligible audio speech from visual speech features. In *sixteenth annual conference of the international speech communication association.*
[12] Tamás Gábor Csapó, Tamás Grósz, Gábor Gosztolya, László Tóth, and Alexandra Markó. 2017. DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface. In *Proc. Interspeech 2017*. 3672–3676. https://doi.org/10.21437/Interspeech.2017-939
[13] B. Denby, Y. Oussar, G. Dreyfus, and M. Stone. 2006. Prospects for a Silent Speech Interface using Ultrasound Imaging. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1. I–I. https://doi.org/10.1109/ICASSP.2006.1660033
[14] Ariel Ephrat and Shmuel Peleg. 2017. Vid2speech: Speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5095–5099. https://doi.org/10.1109/ICASSP.2017.7953127
[15] Ivan Fung and Brian Mak. 2018. End-To-End Low-Resource Lip-Reading with Maxout Cnn and Lstm. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2511–2515. https://doi.org/10.1109/ICASSP.2018.8462280
[16] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-Based Silent Speech Interface for Smartphone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 80 (sep 2020), 27 pages. https://doi.org/10.1145/3411830

[17] Amit Garg, Jonathan Noyola, and Sameep Bagadia. 2016. Lip reading using CNN and LSTM. *Technical report, Stanford University, CS231 n project report* (2016).

[18] Jose A. Gonzalez, Lam A. Cheah, Angel M. Gomez, Phil D. Green, James M. Gilbert, Stephen R. Ell, Roger K. Moore, and Ed Holdsworth. 2017. Direct Speech Reconstruction From Articulatory Sensor Data by Machine Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 12 (2017), 2362–2374. https://doi.org/10.1109/TASLP.2017.2757263

[19] Naomi Harte and Eoin Gillen. 2015. TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech. *IEEE Transactions on Multimedia* 17, 5 (2015), 603–615. https://doi.org/10.1109/TMM.2015.2407694

[20] Hirotaka Hiraki and Jun Rekimoto. 2021. SilentMask: Mask-Type Silent Speech Interface with Measurement of Mouth Movement. In *Augmented Humans Conference 2021* (Rovaniemi, Finland) *(AHs'21)*. Association for Computing Machinery, New York, NY, USA, 86–90. https://doi.org/10.1145/3458709.3458985

[21] Robin Hofe, Stephen R. Ell, Michael J. Fagan, James M. Gilbert, Phil D. Green, Roger K. Moore, and Sergey I. Rybchenko. 2013. Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. *Speech Communication* 55, 1 (2013), 22–32. https://doi.org/10.1016/j.specom.2012.02.001

[22] Carl Q Howard, Colin H Hansen, and Anthony C Zander. 2005. A review of current ultrasound exposure limits. *The Journal of Occupational Health and Safety of Australia and New Zealand* 21, 3 (2005), 253–257.

[23] Yuya Igarashi, Kyosuke Futami, and Kazuya Murao. 2022. Silent Speech Eyewear Interface: Silent Speech Recognition Method Using Eyewear with Infrared Distance Sensors. (2022), 33–38. https://doi.org/10.1145/3544794.3558458

[24] Yan Ji, Licheng Liu, Hongcui Wang, Zhilei Liu, Zhibin Niu, and Bruce Denby. 2018. Updating the Silent Speech Challenge benchmark with deep learning. *Speech Communication* 98 (2018), 42–50. https://doi.org/10.1016/j.specom.2018.02.002

[25] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 57 (jul 2022), 28 pages. https://doi.org/10.1145/3534613

[26] 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 (jun 2021), 30 pages. https://doi.org/10.1145/3463519

[27] Eloi Moliner Juanpere and Tamás Gábor Csapó. 2019. Ultrasound-Based Silent Speech Interface Using Convolutional and Recurrent Neural Networks. *Acta Acustica united with Acustica* 105, 4 (2019), 587–590.

[28] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. AlterEgo: A Personalized Wearable Silent Speech Interface. In *23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 43–53. https://doi.org/10.1145/3172944.3172977

[29] Arnav Kapur, Utkarsh Sarawgi, Eric Wadkins, Matthew Wu, Nora Hollenstein, and Pattie Maes. 2020. Non-Invasive Silent Speech Recognition in Multiple Sclerosis with Dysphonia. In *Proceedings of the Machine Learning for Health NeurIPS Workshop (Proceedings of Machine Learning Research, Vol. 116)*, Adrian V. Dalca, Matthew B.A. McDermott, Emily Alsentzer, Samuel G. Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones (Eds.). PMLR, 25–38. https://proceedings.mlr.press/v116/kapur20a.html

[30] Myungjong Kim, Beiming Cao, Ted Mau, and Jun Wang. 2017. Speaker-Independent Silent Speech Recognition From Flesh-Point Articulatory Movements Using an LSTM Neural Network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 12 (2017), 2323–2336. https://doi.org/10.1109/TASLP.2017.2758999

[31] Myungjong Kim, Nordine Sebkhi, Beiming Cao, Maysam Ghovanloo, and Jun Wang. 2018. Preliminary Test of a Wireless Magnetic Tongue Tracking System for Silent Speech Interface. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. 1–4. https://doi.org/10.1109/BIOCAS.2018.8584786

[32] Naoki Kimura, Tan Gemicioglu, Jonathan Womack, Richard Li, Yuhui Zhao, Abdelkareem Bedri, Alex Olwal, Jun Rekimoto, and Thad Starner. 2021. Mobile, Hands-Free, Silent Speech Texting Using SilentSpeller. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 178, 5 pages. https://doi.org/10.1145/3411763.3451552

[33] Naoki Kimura, Tan Gemicioglu, Jonathan Womack, Richard Li, Yuhui Zhao, Abdelkareem Bedri, Zixiong Su, Alex Olwal, Jun Rekimoto, and Thad Starner. 2022. SilentSpeller: Towards Mobile, Hands-Free, Silent Speech Text Entry Using Electropalatography. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 288, 19 pages. https://doi.org/10.1145/3491102.3502015

[34] Naoki Kimura, Kentaro Hayashi, and Jun Rekimoto. 2020. TieLent: A Casual Neck-Mounted Mouth Capturing Device for Silent Speech Interaction. In *Proceedings of the International Conference on Advanced Visual Interfaces* (Salerno, Italy) *(AVI '20)*. Association for Computing Machinery, New York, NY, USA, Article 33, 8 pages. https://doi.org/10.1145/3399715.3399852

[35] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300376

[36] Alexandros Koumparoulis, Gerasimos Potamianos, Youssef Mroueh, and Steven J. Rennie. 2017. Exploring ROI size in deep learning based lipreading. In *Proc. The 14th International Conference on Auditory-Visual Speech Processing*. 64–69. https://doi.org/10.21437/AVSP.2017-13

[37] Yusuke Kunimi, Masa Ogata, Hirotaka Hiraki, Motoshi Itagaki, Shusuke Kanazawa, and Masaaki Mochimaru. 2022. E-MASK: A Mask-Shaped Interface for Silent Speech Interaction with Flexible Strain Sensors. In *Augmented Humans 2022*. Association for Computing Machinery, New York, NY, USA, 26–34. https://doi.org/10.1145/3519391.3519399

[38] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. EarIO: A Low-Power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 62 (jul 2022), 24 pages. https://doi.org/10.1145/3534621

[39] Richard Li, Jason Wu, and Thad Starner. 2019. TongueBoard: An Oral Interface for Subtle Input. In *Proceedings of the 10th Augmented Human International Conference 2019* (Reims, France) *(AH2019)*. Association for Computing Machinery, New York, NY, USA, Article 1, 9 pages. https://doi.org/10.1145/3311823.3311831

[40] Jian Luo, Jianzong Wang, Ning Cheng, Guilin Jiang, and Jing Xiao. 2021. End-To-End Silent Speech Recognition with Acoustic Sensing. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. 606–612. https://doi.org/10.1109/SLT48900.2021.9383622

[41] Hiroyuki Manabe, Akira Hiraiwa, and Toshiaki Sugimura. 2003. "Unvoiced Speech Recognition Using EMG - Mime Speech Recognition". In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI EA '03)*. Association for Computing Machinery, New York, NY, USA, 794–795. https://doi.org/10.1145/765891.765996

[42] Geoffrey S Meltzner, James T Heaton, Yunbin Deng, Gianluca De Luca, Serge H Roy, and Joshua C Kline. 2018. Development of sEMG sensors and algorithms for silent speech recognition. *Journal of Neural Engineering* 15, 4 (jun 2018), 046031. https://doi.org/10.1088/1741-2552/aac965

[43] Daniel Michelsanti, Olga Slizovskaia, Gloria Haro, Emilia Gómez, Zheng-Hua Tan, and Jesper Jensen. 2020. Vocoder-Based Speech Synthesis from Silent Videos. (2020). https://doi.org/10.48550/ARXIV.2004.02541

[44] Rodrigo Mira, Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, Björn W. Schuller, and Maja Pantic. 2022. End-to-End Video-to-Speech Synthesis Using Generative Adversarial Networks. *IEEE Transactions on Cybernetics* (2022), 1–13. https://doi.org/10.1109/TCYB.2022.3162495

[45] Laxmi Pandey and Ahmed Sabbir Arif. 2021. LipType: A Silent Speech Recognizer Augmented with an Independent Repair Model. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 1, 19 pages. https://doi.org/10.1145/3411764.3445565

[46] Laxmi Pandey, Khalad Hasan, and Ahmed Sabbir Arif. 2021. Acceptability of Speech and Silent Speech Input Methods in Private and Public. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 251, 13 pages. https://doi.org/10.1145/3411764.3445430

[47] Laxmi Pandey and Ahmed Sabbir Arif. 2021. Silent Speech and Emotion Recognition from Vocal Tract Shape Dynamics in Real-Time MRI. In *ACM SIGGRAPH 2021 Posters*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3450618.3469176

[48] E.K. Patterson, S. Gurbuz, Z. Tufekci, and J.N. Gowdy. 2002. CUAVE: A new audio-visual database for multimodal human-computer interface research. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2. II–2017–II–2020. https://doi.org/10.1109/ICASSP.2002.5745028

[49] Ahmed Rekik, Achraf Ben-Hamadou, and Walid Mahdi. 2014. A New Visual Speech Recognition Approach for RGB-D Cameras. In *Image Analysis and Recognition*, Aurélio Campilho and Mohamed Kamel (Eds.). Springer International Publishing, Cham, 21–28.

[50] Jun Rekimoto and Yu Nishimura. 2021. Derma: Silent Speech Interaction Using Transcutaneous Motion Sensing. In *Augmented Humans Conference 2021* (Rovaniemi, Finland) *(AHs'21)*. Association for Computing Machinery, New York, NY, USA, 91–100. https://doi.org/10.1145/3458709.3458941

[51] Christine Rzepka. 2019. Examining the use of voice assistants: A value-focused thinking approach. (2019).

[52] Himanshu Sahni, Abdelkareem Bedri, Gabriel Reyes, Pavleen Thukral, Zehua Guo, Thad Starner, and Maysam Ghovanloo. 2014. The Tongue and Ear Interface: A Wearable System for Silent Speech Recognition. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers* (Seattle, Washington) *(ISWC '14)*. Association for Computing Machinery, New York, NY, USA, 47–54. https://doi.org/10.1145/2634317.2634322

[53] George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny,

Lynn-Li Lim, Bergul Roomi, and Phil Hall. 2017. English Conversational Telephone Speech Recognition by Humans and Machines. (2017). https://doi.org/10.48550/ARXIV.1703.02136

[54] Tanja Schultz. 2010. ICCHP Keynote: Recognizing Silent and Weak Speech Based on Electromyography. In *Computers Helping People with Special Needs*, Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 595–604.

[55] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. MuteIt: Jaw Motion Based Unvoiced Command Recognition Using Earable. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 3, Article 140 (sep 2022), 26 pages. https://doi.org/10.1145/3550281

[56] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. 2018. Lip-Interact: Improving Mobile Device Interaction with Silent Speech Commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 581–593. https://doi.org/10.1145/3242587.3242599

[57] Abhinav Thanda and Shankar M. Venkatesan. 2017. Audio Visual Speech Recognition Using Deep Recurrent Neural Networks. In *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*, Friedhelm Schwenker and Stefan Scherer (Eds.). Springer International Publishing, Cham, 98–109.

[58] László Tóth, Gábor Gosztolya, Tamás Grósz, Alexandra Markó, and Tamás Gábor Csapó. 2018. Multi-Task Learning of Speech Recognition and Speech Synthesis Parameters for Ultrasound-based Silent Speech Interfaces. In *Proc. Interspeech 2018*. 3172–3176. https://doi.org/10.21437/Interspeech.2018-1078

[59] Darya Vorontsova, Ivan Menshikov, Aleksandr Zubov, Kirill Orlov, Peter Rikunov, Ekaterina Zvereva, Lev Flitman, Anton Lanikin, Anna Sokolova, Sergey Markov, and Alexandra Bernadotte. 2021. Silent EEG-Speech Recognition Using Convolutional and Recurrent Neural Network with 85% Accuracy of 9 Words Classification. *Sensors* 21, 20 (2021). https://doi.org/10.3390/s21206744

[60] Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2019. Video-Driven Speech Reconstruction using Generative Adversarial Networks. (2019). https://doi.org/10.48550/ARXIV.1906.06301

[61] Michael Wand, Jan Koutník, and Jürgen Schmidhuber. 2016. Lipreading with long short-term memory. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6115–6119. https://doi.org/10.1109/ICASSP.2016.7472852

[62] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I. Hong, Carmel Majidi, and Swarun Kumar. 2020. RFID Tattoo: A Wireless Platform for Speech Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 155 (sep 2020), 24 pages. https://doi.org/10.1145/3369812

[63] You Wang, Ming Zhang, RuMeng Wu, Han Gao, Meng Yang, Zhiyuan Luo, and Guang Li. 2020. Silent Speech Decoding Using Spectrogram Features Based on Neuromuscular Activities. *Brain Sciences* 10, 7 (2020). https://doi.org/10.3390/brainsci10070442

[64] Zi Wang, Yili Ren, Yingying Chen, and Jie Yang. 2022. ToothSonic: Earable Authentication via Acoustic Toothprint. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 78 (jul 2022), 24 pages. https://doi.org/10.1145/3534606

[65] Philip Weber and Thomas Ludwig. 2020. (Non-)Interacting with Conversational Agents: Perceptions and Motivations of Using Chatbots and Voice Assistants. In *Proceedings of Mensch Und Computer 2020* (Magdeburg, Germany) *(MuC '20)*. Association for Computing Machinery, New York, NY, USA, 321–331. https://doi.org/10.1145/3404983.3405513

[66] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. 2016. Achieving Human Parity in Conversational Speech Recognition. (2016). https://doi.org/10.48550/ARXIV.1610.05256

[67] Kai Xu, Dawei Li, Nick Cassimatis, and Xiaolong Wang. 2018. LCANet: End-to-End Lipreading with Cascaded Attention-CTC. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 548–555. https://doi.org/10.1109/FG.2018.00088

[68] Kele Xu, Yuxiang Wu, and Zhifeng Gao. 2019. Ultrasound-Based Silent Speech Interface Using Sequential Convolutional Auto-Encoder. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 2194–2195. https://doi.org/10.1145/3343031.3350596

[69] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. 2019. LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. 1–8. https://doi.org/10.1109/FG.2019.8756582

[70] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2021. SoundLip: Enabling Word and Sentence-Level Lip Interaction for Smart Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 43 (mar 2021), 28 pages. https://doi.org/10.1145/3448087

[71] Ruidong Zhang, Mingyang Chen, Benjamin Steeper, Yaxuan Li, Zihan Yan, Yizhuo Chen, Songyun Tao, Tuochao Chen, Hyunchul Lim, and Cheng Zhang. 2022. SpeeChin: A Smart Necklace for Silent Speech Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 192 (dec 2022), 23 pages. https://doi.org/10.1145/3494987

[72] Yongzhao Zhang, Yi-Chao Chen, Haonan Wang, and Xingyu Jin. 2021. CELIP: Ultrasonic-Based Lip Reading with Channel Estimation Approach for Virtual Reality Systems. 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* (Virtual, USA) *(UbiComp '21)*. Association for Computing Machinery, New York, NY, USA, 580–585. https://doi.org/10.1145/3460418.3480163

[73] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 37 (mar 2020), 26 pages. https://doi.org/10.1145/3381008

[74] Guoying Zhao, Mark Barnard, and Matti Pietikainen. 2009. Lipreading With Local Spatiotemporal Descriptors. *IEEE Transactions on Multimedia* 11, 7 (2009), 1254–1265. https://doi.org/10.1109/TMM.2009.2030637