SpeeChin: A Smart Necklace for Silent Speech Recognition

RUIDONG ZHANG, Cornell University, USA
MINGYANG CHEN, University of California San Diego, USA
BENJAMIN STEEPER, Cornell University, USA
YAXUAN LI, McGill University, Canada
ZIHAN YAN, Zhejiang University, China
YIZHUO CHEN, Zhejiang University, China
SONGYUN TAO, Cornell University and Dartmouth College, USA
TUOCHAO CHEN, Cornell University and University of Washington, USA
HYUNCHUL LIM, Cornell University, USA
CHENG ZHANG, Cornell University, USA

This paper presents SpeeChin, a smart necklace that can recognize 54 English and 44 Chinese silent speech commands. A customized infrared (IR) imaging system is mounted on a necklace to capture images of the neck and face from under the chin. These images are first pre-processed and then deep learned by an end-to-end deep convolutional-recurrent-neural-network (CRNN) model to infer different silent speech commands. A user study with 20 participants (10 participants for each language) showed that SpeeChin could recognize 54 English and 44 Chinese silent speech commands with average cross-session accuracies of 90.5% and 91.6%, respectively. To further investigate the potential of SpeeChin in recognizing other silent speech commands, we conducted another study with 10 participants distinguishing between 72 one-syllable nonwords. Based on the results from the user studies, we further discuss the challenges and opportunities of deploying SpeeChin in real-world applications.

CCS Concepts: • **Human-centered computing** → **Ubiquitous computing**; *Gestural input*.

Additional Key Words and Phrases: Silent Speech recognition, Deep learning, Computer vision

ACM Reference Format:
Ruidong Zhang, Mingyang Chen, Benjamin Steeper, Yaxuan Li, Zihan Yan, Yizhuo Chen, Songyun Tao, Tuochao Chen, Hyunchul Lim, and Cheng Zhang. 2021. SpeeChin: A Smart Necklace for Silent Speech Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 192 (December 2021), 23 pages. https://doi.org/10.1145/3494987

Authors' addresses: Ruidong Zhang, rz379@cornell.edu, Cornell University, USA; Mingyang Chen, mic016@ucsd.edu, University of California San Diego, USA; Benjamin Steeper, bds238@cornell.edu, Cornell University, USA; Yaxuan Li, yaxuan.li@mail.mcgill.ca, McGill University, Canada; Zihan Yan, zihanyan@zju.edu.cn, Zhejiang University, China; Yizhuo Chen, 3170105441@zju.edu.cn, Zhejiang University, China; Songyun Tao, st938@cornell.edu, Cornell University and Dartmouth College, USA; Tuochao Chen, 1600012713@pku.edu.cn, Cornell University and University of Washington, USA; Hyunchul Lim, hl2365@cornell.edu, Cornell University, USA; Cheng Zhang, chengzhang@cornell.edu, Cornell University, USA.

2474-9567/2021/12-ART192 $15.00
https://doi.org/10.1145/3494987

1 INTRODUCTION

Speech input has become one of the most popular input methods on commodity computing devices. Most of the existing speech recognition technologies recognize speech using sound. However, such technologies require users to voice the speech, which may not work well if the users are in public as it can introduce privacy concerns and is less socially appropriate. Thus, a more privacy-aware and discreet speech input technology is needed.

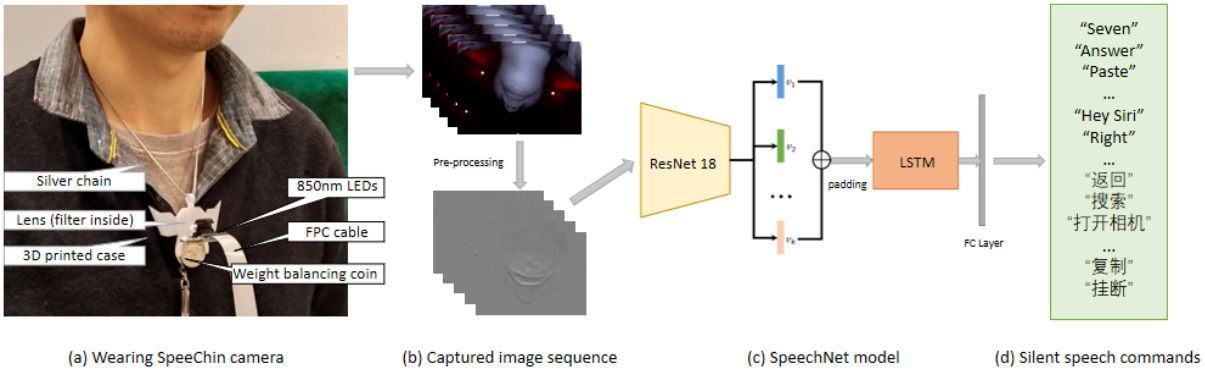


Fig. 1. Overview of SpeeChin

Silent speech recognition (SSR) was invented as an alternative input method for scenarios where vocalizing speech is inappropriate, or speech recognition is compromised by high background noise or for people who have challenges in vocalizing speech [19]. It recognizes speech without requiring the user to voice any sounds into the environment.

Researchers have developed a variety of SSR technologies. Based on whether the sensing technology is worn on the body, existing SSR techniques can be classified into two categories: wearable and non-wearable approaches. Most non-wearable SSR techniques use frontal cameras to capture the face and lip movements to distinguish silent speech commands. However, these systems require the user to be in front of a camera without occlusion, which is neither sustainable nor socially appropriate in daily activities. Wearable-based SSR techniques usually attach different sensors (e.g., EEG [7, 55], sEMG [33, 34, 58, 70], magnetic [4, 9, 29, 36, 37, 57], ultrasonic [18, 32, 40], acoustic [27, 50, 51], capacitive [38, 43]), camera [39] on or around the head to detect the movements of articulators (e.g., skin, ear, tongue, muscles) involved in speech. However, compared to non-wearable-based SSR methods, wearable-based approaches have limitations. Some require the user to attach electrodes on the face or mouth, which is uncomfortable in real-life scenarios. Others do not work well after remounting the sensor, or can only recognize a small set of commands[39]. Thus, there is an apparent need for a minimally obtrusive wearable sensing device that can recognize a rich set of silent speech commands without compromising comfort or performance.

To address this challenge, we present SpeeChin, the first necklace-based SSR technology that can recognize 54 English and 44 Chinese silent speech commands. It uses a customized neck-mounted IR camera to capture the images of the neck and face from under the chin. These images are pre-possessed and sent to an end-to-end convolutional-recurrent-neural-network (CRNN) model for silent speech recognition. An overview of SpeeChin is illustrated in Figure 1. A user study with 20 participants shows that it can recognize 54 English commands and 44 Chinese commands with an average accuracy of 90.5% and 91.6%, respectively.

Unlike standard voice recognition systems, current wearable SSR devices can only recognize a limited set of pre-trained words. However, if they could recognize sound on a phoneme level, they could instead map phoneme patterns to words, allowing them to recognize a much larger dataset without the need for word-level training. To

investigate the potential of SpeeChin on distinguishing between more silent speech commands, we conducted a third study with 10 participants to recognize 72 one-syllable nonwords. The results are presented and analyzed in the discussion section.

To the best of our knowledge, SpeeChin is the first smart necklace that can recognize silent speech phrases in two languages (54 English and 44 Chinese commands) with over 90% accuracy after remounting the device (session-independently) without using frontal cameras. The contributions of SpeeChin are:

- We present the first smart necklace with a built-in IR camera that can distinguish a rich set of silent speech commands in two languages (English and Chinese) by deep learning images of the neck and face captured from under the chin.
- We conducted a user study with 20 participants to evaluate SpeeChin on distinguishing 54 English and 44 Chinese silent speech commands.
- We conducted another user study with 10 participants to understand how SpeeChin can distinguish between phonemes and conducted a corresponding linguistic analysis.
- Based on the study results, we further discuss the opportunity and challenges of deploying SpeeChin in real-world applications.

2 RELATED WORK

Silent Speech Recognition (SSR) has been an active research topic in the research community for decades. This section discusses related works in two categories: 1) non-wearable-based SSR technologies or 2) wearable-based SSR technologies.

Table 1. Summary of wearable-based SSR Techniques. UD: user-dependent, SD: session-dependent, UI: user-independent, SI: session-independent, NM: not mentioned

Reference	Sensing method	Sensor location	Wordlist size	Classification accuracy	Session/user dependency
Sahni et al. [57]	Magnet field	Tongue and glasses	11	90.5%	UD, SD
Hofe et al. [29]	Magnet field	Tongue and head	57	98.8%	SD
TongueBoard [43]	Capacitive	Tongue	21	91.0%	UD, SI
Schultz [58]	sEMG	Face	101	68.5%	UI
AlterEgo [33]	sEMG	Face	10	92.0%	UD, SD
SottoVoce [40]	Ultrasonic	Tightly attached to jaw	40	66.4%	NM
TieLent [39]	Camera	Necklace	15	94%	UD
C-Face [11]	Two cameras	Earphones	8	84.7%	UD, SD
SpeeChin	**One camera**	**Necklace**	**54 English 44 Chinese**	**90.5% 91.6%**	**UD, SI**

2.1 Non-wearable-based SSR Technologies

Humans can read speech by looking at lip movement, also known as lip-reading. Sharing a similar spirit, the most popular SSR techniques use a camera in front of the user's face to capture images of lip and face movements to infer the speech content. Prior works have demonstrated the ability to classify different lists of letters/words/phrases [22, 24, 56, 62, 68, 75], recognize sentences [2, 13, 14, 31, 41, 64, 72], and reconstruct the sound of the speech [1, 5, 15, 20, 48, 67] from these images. Recently, researchers also explored using a built-in camera[62] and acoustic sensors[23, 74] on the smartphone to capture lip movement for SSR.

Although these non-wearable-based SSR techniques have demonstrated promising performance, they require placing the hardware (e.g., cameras, phones) in front of the user's face, which may not always be socially appropriate or convenient. For instance, most previous work uses a camera on a laptop or placed in the environment to capture the user's face. Such systems limit the users' movement and would not work when the environment does not allow setting up a camera (e.g., when the user is in public). Although the user can wear the camera with a chest mount or head-mount or hold it in hand, these settings can be inconvenient and socially inappropriate in many scenarios. A more discreet and flexible SSR system is in need.

2.2 Wearable-based SSR Technologies

Wearables are worn directly on the user's body. Thus, wearable-based systems usually offer more flexibility in terms of range of movement and can work well when frontal cameras can not be set up in the environment. Researchers have explored a variety of wearable-based SSR techniques by capturing the movement of different articulators (e.g., skin, ear, tongue, lips, muscles) using different sensing methods.

To capture the movements of the muscles, tissue, or skin on the head, researchers developed SSR techniques that placed EMG electrodes on the face [33, 34, 58, 70], the head [7], and/or the neck [47]. Other technologies such as RF-based movement tracking [69] and MRI [54] have also been explored to analyse the facial movements or vocal tract shape dynamics while the user is speaking.

Besides movement on the surface of the head, vocalizing speech also involves movements of different articulators (e.g., mouth, tongue) inside the mouth. To capture the tongue and mouth movements for SSR, researchers placed magnets on the tongue and head [4, 9, 29, 36, 37, 57], and capacitive sensors in the mouth [38, 43]. Tongue movement can also be inferred by placing medical-purpose ultrasonic probes tightly under the chin for SSR [16, 18, 32, 40, 66]. The latest research project C-Face [11] shows feasibility in distinguishing silent speech commands by learning the change of facial contours using two ear-mounted cameras. However, it can only recognize eight commands with an accuracy of around 85%. A recent project, TieLent[39] demonstrates the feasibility to recognize 15 silent speech phrases with a necklace-mounted RGB camera, which was evaluated with 4 participants in a controlled setting. It generates synthetic sound of the phrase first which is recognized by the speech recognition engine. Furthermore, it is unclear how it would perform across different sessions (e.g., Remounting) with a larger vocabulary and more participants.

We summarized the closest wearable-based SSR technologies in Table 1. As the table shows, many of these systems require wearing electrodes on the face [33, 40, 58] or putting sensors inside the mouth [29, 43, 57]. Most of these systems are session-dependent, which do not work well after remounting the electrodes or sensors. Although some have shown promising performance [43], heavy instrumentation on the tongue may not be acceptable by many users in daily activities. Furthermore, technologies [11] that are considered non-obtrusive can only recognize a small set of silent speech commands.

Compared to the previous wearable-based SSR technologies, SpeeChin is a minimally-obtrusive wearable-based SSR technology using a smart necklace. It is the first to demonstrate the feasibility of using a neck camera to capture neck and face images for recognizing silent speech commands. Furthermore, it is also the first SSR technology that demonstrates promising session-independent SSR performance in multiple languages (Chinese and English).

3 THEORY OF OPERATION

Although humans use nearly 7,000 distinct languages to communicate with each other across the world, the anatomy for speech is the same: people make sounds using their lips, tongue, and facial muscles (jaw). Speech, whether vocalized or not, requires the speaker to alter the shape of their chin, lips, cheeks, etc. from the perspective of an outside vantage point.

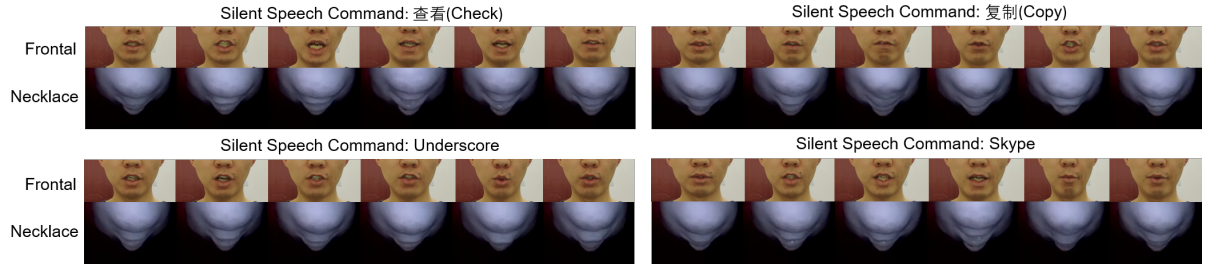


Fig. 2. Change in chin and face shape while uttering different phrases. First line: images captured by frontal camera, second line: images captured by necklace camera

The traditional computer-vision based SSR approach directly captures lip movement using a camera in front of the face. However, as we have discussed previously, placing cameras in front of the face is not always a feasible option. Fortunately, each speaking gesture consists of a series of movements involving multiple articulators (e.g., lips, jaw, tongue) on the face. Capturing the movement of the mouth/lips may not always be convenient or even necessary. However, observing the movements on other corresponding articulators of a speaking gesture may be more convenient in some cases, while still being highly informative for speech recognition. For instance, the latest research [11] has demonstrated using the contours of the face observed from the ears to estimate multiple articulators' movements, including the mouth, eyes, and eyebrows.

However, the main focus of C-Face [11] is not to recognize silent speech commands, and it could only recognize eight commands with around 85% accuracy. We suspect one of the reasons C-Face does not perform well on SSR is that it captures little information on the tongue's movement, which is critical for performing silent speech.

To capture the movement of the tongue, previous research placed sensors inside the mouth, which would be unacceptable for many users. However, based on our observations, we found that tongue movement inside the mouth also leads to changes in the shape of the neck and lower face. Based on this observation, we hypothesize that if the system can observe neck, chin, and lower facial movement, it can accurately distinguish between silent speech commands.

We placed an IR camera below the chin to verify our hypothesis, by taking pictures of the neck and face. We used this system to record changes in the shape of the neck and face while one researcher uttered various silent speech commands. As shown in Figure 2, the shapes and images of the chin change as the commands change. This initial observation was very encouraging, leading us to design SpeeChin, a smart necklace that recognizes silent speech commands using images of the neck and face.

4 HARDWARE DESIGN

Based on the encouraging results from the preliminary experiment, we first implemented the hardware prototype of SpeeChin, which is a camera module housed in a 3D printed necklace case. The necklace case is then hung around the neck with a silver chain, as illustrated in Figure 1(a).

4.1 Infra-red Imaging System

In selecting a suitable camera type for SpeeChin, we first tried using a standard RGB camera. However, we found it difficult to segment the user's head from different backgrounds, as shown in Figure 3(a). However, reliably segmenting the head from backgrounds is critical to the success of SpeeChin, as information is encoded in the shapes of the neck, chin, and face. To better segment the background, we considered using a thermal camera or a depth camera. Unfortunately, these state-of-art thermal cameras are too large to be considered

(a) RGB camera (b) IR camera with 850nm filter and independent lighting (c) IR camera with 650+850nm filter and independent lighting

Fig. 3. Comparison of images captured with different camera and filters. With 850nm filter and independent lighting, the brightness between skin and background is different; with 650+850nm filter and independent lighting, both brightness and color are different.

"minimally-obtrusive" or the image resolution is too low. We settled on an IR camera due to its convenient filter system for background segmentation, small size, and relatively high image resolutions.

Our IR imaging system consists of lighting, optical filter modules, and an image sensor. The lighting module has two customized PCBs, each equipped with a 3W 850nm IR LED (TY-850nm3W, light angle 120 degrees) to project IR light on the skin from the neck. Near-infrared technology has been used in wrinkle and scar treatment [63, 71] and proven to be safe [3, 6]. The power of our IR LEDs is less than 4mW/cm^2, much lower than those used for medical purposes.

The IR light reflected on the skin will be captured by the image sensor, an OV5647 module (320x240@60fps) with a 130-degree FoV lens and adjustable focus. To make segmentation easier, we adopted a 650+850nm dual peak narrow-band optical filter, which allows the 650 and 850nm components of the lights into the imaging system. The skin are mostly reflected with 850nm component using our lighting system. The lights from background contains a wide spectrum of light components, including both 850 and 650nm components. As a result, the color and brightness of the skin in the image, which mostly contains 850nm components, looks visually very distinct from the backgrounds which are mixed with 850 and 650nm components in lights, as seen in Figure 3. Thus, it is very easy to segment the skin from the background based on brightness and color difference. A detailed explanation of our segmentation method is specified in the section 5.1 and Figure 4.

4.2 Form Factor Design

We 3D printed a holder for the IR imaging system. In order to make it stable, we designed a wing on each side and place a coin on the bottom, as Figure 1(a) shows. The imaging system is attached at the center of the holder. A light-weight silver chain was used to hang the form factor around the user's neck. The chain slides smoothly through a hole in the case when the user rotates the head, minimizing device shifting.

5 DATA PROCESSING PIPELINE

The necklace camera images alter over time due to slight shifts in the device's angle and positioning. Pre-processing is therefore necessary to rectify this shifting by affine transformation. Moreover, clothes in the frame or lights in the background may introduce noise. In the captured image, we are only interested in the human body ("foreground"). Thus, we need to narrow down the most significant information within each frame and eliminate noise before feeding the images into the SpeechNet machine learning model. We also need to provide the start frame and the end frame for each utterance. In this section, we offer a detailed pipeline of our algorithm implementation.

5.1 Image Pre-processing

As shown in Figure 4, we first calculate the angle and center for performing affine transformation. We then use differential images to enlarge changes by uttering silent commands and conducting transformation to align the images. Lastly, we crop out the most significant part of the images. The parameters used in image pre-processing were pre-determined based on the results in the preliminary pilot studies. To avoid over-fitting to the training data collected in the user study, we did not change or fine-tune the parameters when evaluating the system on the user study data.

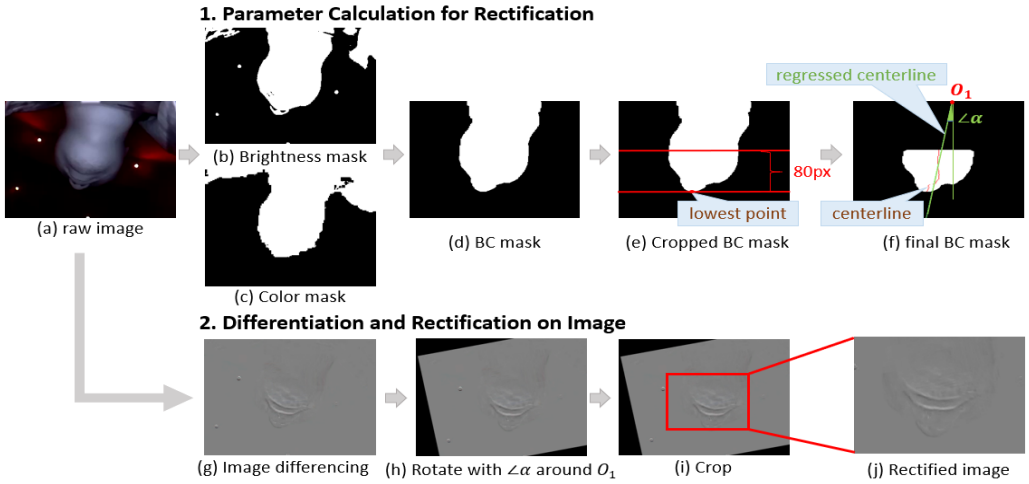


Fig. 4. Pre-processing for images captured by SpeeChin camera

5.1.1 Parameter Calculation for Rectification. Firstly, we need to compute center and angle of rotation in order to rectify the participant's head movement. The first step is to segment out the skin (foreground) from the background in the image. As we utilize a 650+850nm dual peak filter and 850nm LEDs, the foreground and the background are clearly distinguishable by their brightness and color, as seen in Figure 4(a). We first apply Otsu binarization [53] to identify foreground pixels, which corresponds to the user's chin as shown in Figure 4(b). Then, we remove bright spots or stripes in the background caused by artificial light with a color mask. More specifically, we compute the histogram of the H channel with values ranging from 0 to 179 and find the peak value in the histogram (excluding the peak at 0). This peak point represents the color with the largest proportion, which belongs to the foreground. As a result, pixels with H value within -30 to +25 from the peak are considered the foreground. Morphological operations are then applied to fill holes and remove small isolated components to generate a brightness+color (BC) mask as shown in Figure 4(d). Since parts of the user's clothing are often included in corners and top of the frame, we crop out 80 pixels upwards from the foreground mask's lowest position. Finally, we locate the midpoint for each row of pixels in the foreground portion on the cropped BC mask (centerline in Figure 4(f)). Then we regress the centerline with linear regression and set O_1 as the intersection point of the regressed centerline on the first row of the whole mask image. The angle $\angle\alpha$ is the angle between the regressed centerline and vertical line, which indicates the relative rotation angle of the participant's head. Since the device position may shift over time due to body movement, we compute $\angle\alpha$ once for every chunk of frames composing one utterance.

For each chunk of frames corresponding to one utterance, we select a single static frame before the user starts or after the user finishes uttering the phrase to compute $\angle\alpha$. The static frame refers to a frame at which the user's

mouth is not moving. It occurs when the user rests between uttering phrases. We select one static frame for calculation rather than computing $\angle\alpha$ on each frame in order to avoid introducing noise in the time domain. Since each utterance only lasts 1-2 seconds, the possibility of device shifting during this short time period is negligible.

5.1.2 Static Frame Selection. To get this static frame, for each chunk of images composing an utterance, we first select a large window LW of about 160 frames (2.67s). LW ranges from 20 frames (about 33ms) before the recorded start frame to 20 frames after the recorded end time. Secondly, we use a 19-frame sliding window win ($stride$ = 10 frames) over each LW. We resize the raw image to a smaller size of 64x48 pixels and stack the 19 frames in the win together. Thirdly, we calculate the variance on each pixel position in win. The top 20 pixels with the greatest variances are added up and recorded as the "static index" for this win. The center frame of the win with the smallest "static index" is regarded as the static frame.

5.1.3 Differentiation and Rectification of Images. The overarching goal of preprocessing is to improve the signal-to-noise ratio (SNR) to help the model grasp principal features for silent speech command classification. In order to enlarge subtle changes of facial movements in the image, we generate differential images by $I_i - I_{i-2}$, where I_i means the ith frame recorded. We use the B channel of the raw image when calculating difference. Next, we rotate the differential image $\angle\alpha$ to rectify the head rotation, as shown in Figure 4(h)). As mentioned above, a chunk of frames share the same $\angle\alpha$. We rotate this chunk of frames by $\angle\alpha$. At last, we crop the image with a 192x144 rectangle centered at the chin's position on the image as illustrated in Figure 4(i), which are the final images fed into the machine learning model.

5.2 Utterance Detection

In order to split commands and detect the starting and ending frames of each utterance from the image series, we develop an utterance detection algorithm.

5.2.1 Vigorousity Measurement. Detecting the exact start time and end time is based on the degree of mouth movement, which can be reflected by chin movement in the captured video. The vigorousity v_i of such movement in frame i is measured by the standard deviation of all pixels in a smoothed second-order differential image. Let I_i denote the ith frame in the dataset, v_i is measured by

$$v_i = \mathrm{std}\left(\frac{1}{3}\sum_{k=-1}^{1} I_{i+3+k} + \frac{1}{3}\sum_{k=-1}^{1} I_{i-3+k} - \frac{2}{3}\sum_{k=-1}^{1} I_{i+k}\right)$$

The normalized vigorousity index v_i^n is then obtained by averaging v_i in 4 seconds to remove the influence of potential vigorousity shift over time.

In this way, a vigorousity graph is generated as shown in Figure 5(a). v will appear to have lower values when the participant is not speaking.

5.2.2 Utterance Detection. In the generated vigorousity graph, speaking periods visually appear to have a higher vigorousity (or a peak), while silent periods having lower vigorousity (or a valley), as demonstrated in Figure 5(a). Based on this feature, we segment the speaking periods out before performing classification. We first find out all "raw peaks" where $v_i^n \geq v_{th}$, then we merge consecutive "raw peaks" if they are less than t_m from each other. We then remove merged peaks that are shorter than t_{w1} or longer than t_{w2} based on the typical lengths of the phrases. This process is illustrated in Figure 5(a). In practice, we set $v_{th} = 0.92$, $t_m = 0.83s$ (50 frames), $t_{w1} = 0.1s$ (6 frames), $t_{w2} = 2.17s$ (130 frames), based on the typical length of utterances during our user study.

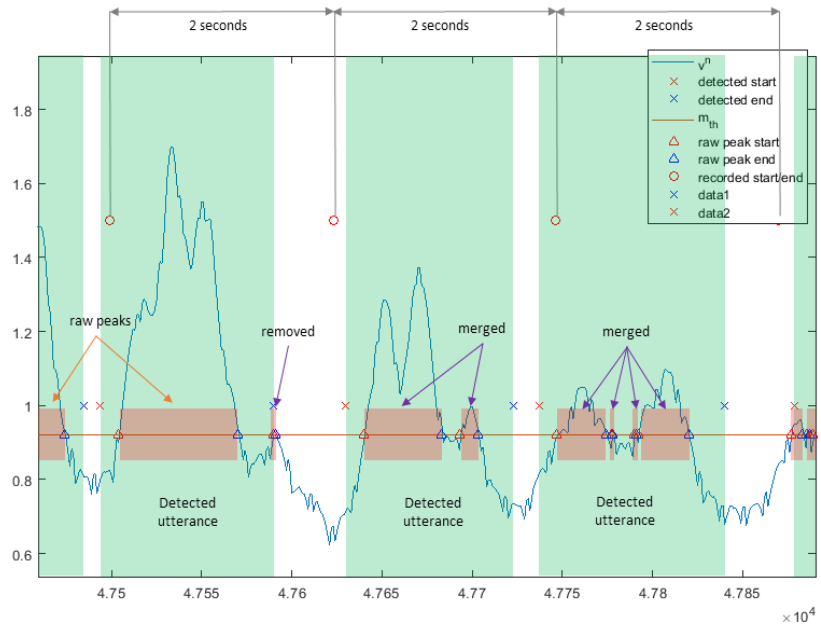


Fig. 5. Vigorousity index calculation for the utterance detection algorithm.

Fig. 6. Evaluation metric for utterance detection. TP: true positives, FP: false positives, FN: false negatives. When the center of a detected segment falls within a recorded segment, it is counted as a TP. If more than one detected segment centers fall within the same recorded segment, only one is counted as TP, other are counted as FP.

5.3 SpeechNet

The pre-processed images on segmented utterances are fed into an end-to-end deep learning model: SpeechNet, for recognizing silent speech commands.

5.3.1 Model Selection. To recognize the silent speeches, we propose SpeechNet, an end-to-end CRNN model for vision-based silent speech recognition. Our choice of model was made by comparing different models and the characteristics of our dataset (spatial and temporal time-series). CNN has shown promising abilities in extracting key spatial features from images [45] in recent years. Therefore, CNN is widely used in various vision tasks including image classification [26, 42, 60], object detection [44] and image segmentation [49]. However, when faced with variable-length sequential data, RNN often fits better by virtue of its flexibility and advantages in extracting temporal features [73]. Combining the strengths of CNN and RNN, researchers have used CRNN models in various vision-based temporal classification tasks including text recognition [59], human activity recognition [46], video-based emotion recognition [21], and silent speech recognition [62]. Given our data set is

image series with variable length, we decide to use CRNN as our model. However, what we provide in this paper is a starting point. Further fine-tuning the model can potentially lead to even stronger performance.

5.3.2 Model Structure. The structure of the model is illustrated in Figure 7. The first block of SpeechNet is a CNN module based on Residual Network (ResNet18) [25] as CNN has demonstrated strong ability in capturing high-level features from images for various tasks in the computer vision field. Each convolutional block in ResNet18 consists of a convolutional layer followed by a batch normalization (BN) [30] and rectified linear unit (ReLU) layer. At the end of the CNN backbone, a global average pooling is applied. The output of Resnet18 is then padded to a fixed size. Subsequently, we apply a Long Short-Term Memory (LSTM) block [28] due to its strength in integrating time serialized information, which has been validated in the natural language processing field. More specifically, we use a single layer, unidirectional LSTM with 128 hidden units, followed by a subsequent dropout (probability=0.5) layer to avoid overfitting. A subsequent linear layer gives the final prediction for the speech command in the end.

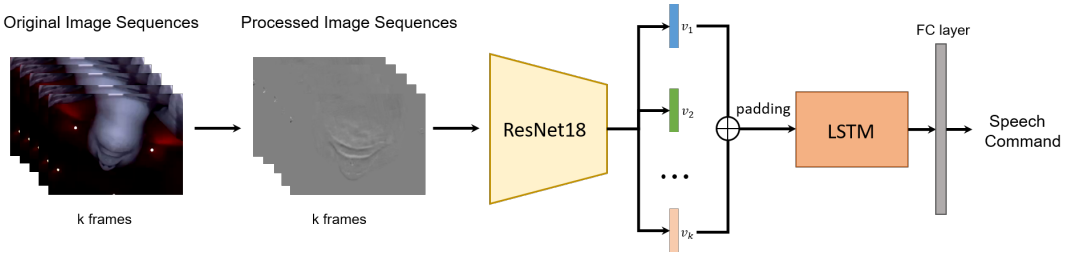


Fig. 7. Architecture of SpeechNet. k represents that there are k frames composing this command. v_i represents the visual feature of the i^{th} frame. Padding means that we pad variable-length sequences to form a fixed-size input.

5.3.3 Data Augmentation. To further improve the robustness of the model against spatial and temporal noises with limited training data , we apply both temporal and spatial data augmentation. The former increases temporal variance by splitting odd and even frames and randomly shifting the starting and ending frame. The latter increases variance on images by applying random affine transformation.

- Temporal augmentation: The original frame rate from our experiment was 60 fps. We group frames with odd frame numbers and even frame numbers to form two new 30 fps data samples, thus doubling the size of our data. A pilot experiment showed that this approach has significantly improved the performance across different settings. The position of starting and ending frame of each utterance is randomly shifted within 10 frames from the detected position in order to increase temporal variance and compensate for errors in utterance detection.
- Spatial augmentation: The position of the camera frequently shifts, which leads to displacements of the images. To enrich the variance of the images without the need of collecting more data, we apply affine transformation. When the training process fetch the set of images of a silent speech command from the dataset, there is an 80% chance that the system would conduct affine transformation on images. If selected, we apply the random affine transformation of translation (range from -20 to 20 pixels) , rotation (range from -8° to $+8^\circ$), scaling (range from 0.9 to 1.1) of any combinations. This approach helps increase the data diversity and avoid overfitting, making our deep model more resilient to spatial noise.

5.3.4 Training Process. We choose the cross-entropy loss as the loss function and Adam as the optimizer during the training process. The learning rate is set to be 0.0001 at the beginning and decreases over time. The batch size is set to 16. We train the model for 400 epochs.

6 USER STUDY

To evaluate the performance of SpeeChin, we conducted two studies with 20 participants in total, both approved by the Institutional Review Board (IRB) of our institution. The first study with 10 participants tested 54 English silent speech commands, and the second study with another 10 participants tested 44 Chinese silent speech commands. The purpose of the two studies was to explore how our system would function cross-linguistically.

6.1 Command Sets

6.1.1 English Commands Dataset. We selected 54 English utterances (listed in Table 2) composed of digits, interactive commands, voice assistant commands, punctuation commands, and navigation commands. We chose interactive commands such as 'Answer', 'Call', and 'Camera'[74] as they are highly applicable for cell phone use. We also selected frequently used voice assistant commands such as 'OK Google', 'Hey Siri', and 'Alexa'. To assist in typing or texting, we included ten punctuation commands. When users type using smartphones, they are typically required to switch to the secondary keyboard to input punctuation. Silently uttering punctuation commands instead could prove to be more convenient.

Table 2. Commands set: English command set with 54 frequently used phrases.

Category	Commands
Digits	Zero, One, Two, Three, Four, Five, Six, Seven, Eight, Nine
Interactive Commands	Answer, Call, Check, Copy, Cut, Hang up, Mute, Paste, Pause, Play, Redial, Screenshot, Search, Skip, Share, Undo, Previous, Next, Open, Close, Volume, Keyboard, Camera, Home, Help, Skype
Voice Assistant	OK Google, Hey Siri, Alexa
Punctuation	Question mark, Exclamation point, Comma, Dot, Semicolon, Colon, Quotation mark, Parentheses, Dash, Slash, Underscore
Navigation	Left, Right, Up, Down

6.1.2 Chinese Command Dataset. We also chose 44 commands in Standard Chinese (Mandarin) to test our system's performance cross-linguistically. Phonetically, Chinese is quite different from English, rendering it a suitable test case. Based on our knowledge, this is the first wearable technology that evaluates the cross-linguistically SSR performance.

The command set is derived from Lip-interact[62], which uses the smartphone's camera to recognize Chinese silent speech commands. It includes five categories: system, home screen, WeChat, Notepad, and pop-ups. As depicted in Table 3, these commands were selected based on their applicability in various smartphone interactions. The command set includes functionality commands related to two popular apps: WeChat[1] and Notepad. With SpeeChin implemented, users could potentially access in-app functions directly without navigating through a hierarchical menu structure.

6.2 Apparatus

We used the SpeeChin necklace described in section 4 to collect facial movements from underneath the chin while users silently uttered commands. As presented in Figure 8(a), the camera in the necklace was connected to a Raspberry Pi through an FPC cable. The Raspberry Pi was connected to a monitor and a control button. The screen was used to display a GUI for participants to follow during data collection. The control button was used

[1]https://www.wechat.com/en/

Table 3. Commands set: Chinese commands set with 44 frequently used phrases. The PinYin (indicating pronunciation) and meaning in English are explained after each command. Alipay, Taobao and Weibo are popular mobile apps in China.

Category	Commands
System	返回 (FanHui, Back), 桌面 (ZhuoMian, Home), 截屏 (JiePing, Screenshot), WiFi, 静音(Jingyin, Mute), 手电筒(ShouDianTong, Flashlight), 通知栏(TongZhiLan, Notification), 最近应用(ZuiJinYingYong, Recent Apps), 蓝牙(LanYa, Bluetooth), 锁屏(SuoPing, Lock)
Home Screen	打开微信(DaKaiWeiXin, Open WeChat), 打开浏览器(DaKaiLiuLanQi, Open Browser), 打开相机(DaKaiXiangJi, Open Camera), 打开支付宝(DaKaiZhiFuBao, Open Alipay), 打开音乐(DaKaiYinYue, Open Music), 打开淘宝(DaKaiTaoBao, Open Taobao), 打开邮箱(DaKaiYouXiang, Open Mailbox), 打开微博(DaKaiWeiBo, Open Weibo), 打开闹钟(DaKaiNaoZhong, Open Alarm), 打开记事本(DaKaiJiShiBen, Open Notepad)
WeChat	朋友圈(PengYouQuan, Moments), 搜索(SouSuo, Search), 添加(TianJia, Add), 发状态(FaZhuangTai, Post), 扫码(SaoMao, Scan QR Code), 点赞(DianZan, Like), 更换头像(GengHuanTouXiang, Change Profile), 二维码(ErWeiMa, Show QR Code), 发送(FaSong, Send)
Notepad	复制(FuZhi, Copy), 剪切(JianQie, Cut), 粘贴(ZhanTie, Paste), 撤销(CheXiao, Undo), 重做(ChongZuo, Redo), 加粗(JiaCu, Bold), 高亮(GaoLiang, Highlight), 向左(XiangZuo, To Left), 向右(XiangYou, To Right)
Pop-ups	删除(ShanChu, Delete), 查看(ChaKan, Check), 接听(JieTing, Answer), 挂断(GuaDuan, Hang up), 是(Shi, Yes), 否(Fou, No)

to start, pause, or resume the data collection process. Lastly, we captured each participant's face with a frontal camera using a mobile phone to compare SpeeChin results with SSR using the images of the user's face.

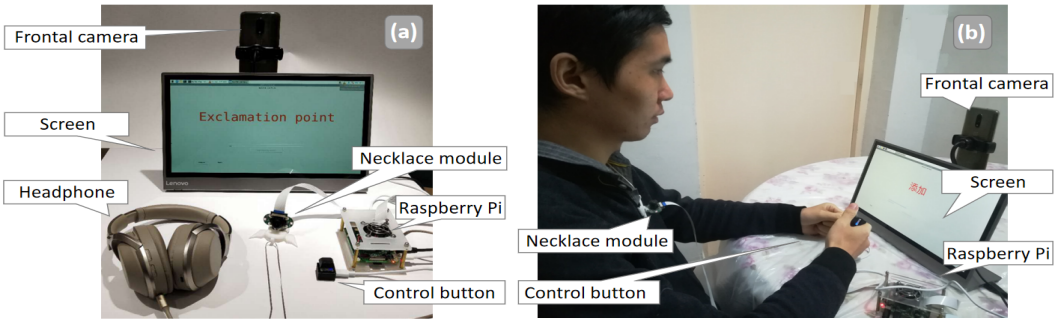


Fig. 8. Apparatus and user study setup. (a) Apparatus for user study. (b) User study setup

6.3 User Study Procedure

We recruited 10 participants for the English Silent Speech Commands study (M_{age} = 20.6; male = 3, female = 7) and 10 participants for the Chinese Silent Speech Commands study (M_{age} = 20.6; male = 8, female = 2). All participants in the studies self-reported being fluent in English and Chinese Mandarin, respectively.

Both user studies shared the same overall procedure. The researcher first introduced the procedures and helped participants set up the necklace and front-facing cameras. The participant was then handed a single button to hold (shown in Figure 8(b)). The participant could decide when to start, pause, or resume a session to take breaks

using the button. Once the study started, the researcher did not interrupt or provide any assistance until the end of the experiment. The user study setup is shown in Figure 8(b).

Each study contained 9 sessions, including a short practice session at the beginning where the participant repeated each phrase twice to get familiarized with the system. The following 8 sessions were used for training/testing, where each phrase was repeated 4 times every session in a random order. As the necklace would likely be taken on and off frequently in real life scenarios, we asked the participants to remount the device by themselves after each session. The participants were free to move around or take a break between sessions.

During each session, a random command would appear on the screen and remain for 2 seconds. The participant was asked to mouth the utterances shown on the screen quietly without voicing a sound. Two progress bars appeared on the screen: one tracked the current utterance progress (2 seconds), and the other tracked the current session progress. The exact time that the command appeared and disappeared on the screen was recorded as ground truth. The participant was instructed to remain relatively still and avoid activities such as touching the face, coughing, swallowing, drinking water, or licking the lips, while uttering commands. However, they were free to move and do these activities after pausing the system.

In order to ensure data quality, we only kept images with a frame rate of 60 ± 2.4 fps. In the event of mistakenly speaking a wrong command, the participant was instructed to press the button to pause the system and then resume it. If there was a disturbance in frame rate (out of range of 60 ± 2.4 fps) or the participant pressed the button to pause, all images recorded for the current and previous utterances were discarded. The corresponding utterances were then added to the end of the current session. In all the user studies combined, 820 out of 55470 utterances (1.48%) were discarded due to participant errors, and 250 (0.45%) were discarded due to a disturbance in frame rate.

7 EVALUATION

In this section, we present the results of our user studies. We first evaluate our utterance detection algorithm. We use precision, recall and F-1 score as the evaluation metrics. We then use accuracy as our evaluation metric for silent speech phrases classification, which is defined as,

$$\text{accuracy} = \frac{\text{\# utterances correctly detected and correctly classified}}{\text{\# utterances correctly detected}} \tag{1}$$

We evaluated our system using user-dependent (UD), user-independent (UI), and user-adaptive (UA) approaches. Specifically, a UD model is trained and tested on the same participant, which means a user needs to provide sufficient amount of training data before being able to use the system; a UI model is trained on some participants, while tested on a different participant, which means a user does not need to provide any training data; a UA model is training on some participants, and fine-tuned with data from the testing participant, which means a user needs to provide a small amount of training data. For UD experiments, the first 6 sessions (excluding the practice session) were used for training and the last two for testing. For UI experiments, we conducted a leave-one-participant-out cross-validation. For UA experiments, we used the first 2 sessions (excluding the practice session) to fine-tune the model trained in UI experiments, and evaluated on the last 2 sessions. Since we asked participants to remount the device after every session, these results are all session-independent.

7.1 Utterance Detection

During the experiment, the phrase stimuli were presented every 2 seconds. This time window was marked as the ground truth of utterances. We used the metrics specified in Figure 6(b) to calculate the precision, recall, and F-1 score of utterance detection.

Results show that the F1 score, precision and recall for all participants in the English study is 98.0% (ranging from 94.7% to 99.3%, SD = 1.34%), 98.3% and 97.8%, respectively. In the Chinese study they are 98.1% (95.6% to

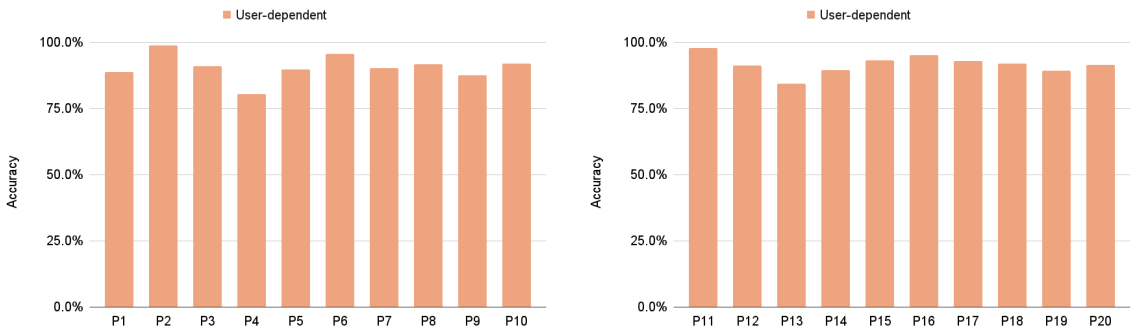
99.4%, *SD* = 1.13%), 97.8%, and 98.4%, respectively. Results demonstrate that the utterance detection algorithm remain highly stable across all participants.

7.2 English Study Silent Speech Recognition Results

We conducted a UD experiment on each participant. Results show that SpeeChin achieves an average accuracy of 90.5% (SD = 4.86%) in classifying 54 English commands, ranging from 80.3% for P4 to 98.7% for P2, as shown in Figure 9(a). The top five confused commands are: misclassifying "Dot" as "Cut" (11% of "Dot" misclassified as "Cut"), "Copy" as "Comma" (9.9%), "Answer" as "Undo" (9.4%), "Home" as "Call" (6.9%), and "Comma" as "Copy"(6.2%). Confused pairs often contain the same number of syllables and share similar families of sounds (such as "Comma" and "Copy" confusing labial phonemes 'm' and 'p'). We observed that glass frame might be captured by our necklace camera, thus polluting the images. To make sure that wearing glasses will not have an impact on system performance, we compared the results of participant with and without glasses. The average accuracy of participants with glasses (P1, P3, P4, P5, P7, and P10) is 88.0%. One-way ANOVA test shows no significant difference on accuracies between participants with or without glasses ($F(1,8) = 3.55$, $p = 0.10$).

7.3 Chinese Study Silent Speech Recognition Results

As for SpeeChin's UD performance on distinguishing 44 Chinese commands, the average accuracy is 91.6% (SD = 3.6%), ranging from 84.2% for P13 to 97.7% for P11, as shown in Figure 9(b). The top 5 confusions are, misclassifying "添加(Add, TianJia)" as "点赞(Like, DianZan)" (12%), "桌面(Home, ZhuoMian)" as "锁屏(Lock, SuoPing)" (12%), "锁屏(Lock, SuoPing)" as "桌面(Home, ZhuoMian)" (10%), "加粗(Bold, JiaCu)" as "删除(Delete, ShanChu)" (6.9%), and "重做(Redo, ChongZuo)" as "搜索(Search, SouSuo)" (6.9%). We observed that 7 out of 10 participants wore glasses. Similarly, one-way ANOVA test shows no significant difference between participants with and without glasses($F(1,8) = 1.53$, $p = 0.25$).



(a) Results on 54 English commands

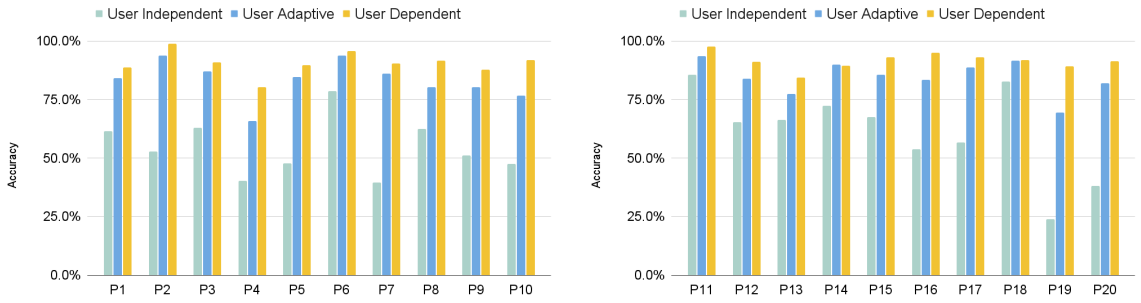
(b) Results on 44 Chinese commands

Fig. 9. Results for all participants on 54 English phrases and 44 Chinese phrases

7.4 User-independent and User-adaptive Experiments

To further discuss the possibility of making our system user-independent, we conducted a leave-one-participant-out (LOPO) experiment. Results are demonstrated in Figure 10. Classification accuracy on 54 English commands across 10 participants ranges from 40.3% (P4) to 78.6% (P6), with an average accuracy of 54.4% and standard

deviation of 12.0%. Classification accuracy on 44 Chinese commands across 10 participants ranges from 23.8% (P19) to 85.5% (P11), with an average accuracy of 61.2% and standard deviation of 19.0%. The high variance across different users indicates that user-independent performance is not very stable, but also demonstrates the potential for future user-independence.



(a) User dependency experiment results for English study

(b) User dependency experiment results for Chinese study

Fig. 10. User dependency experiment results

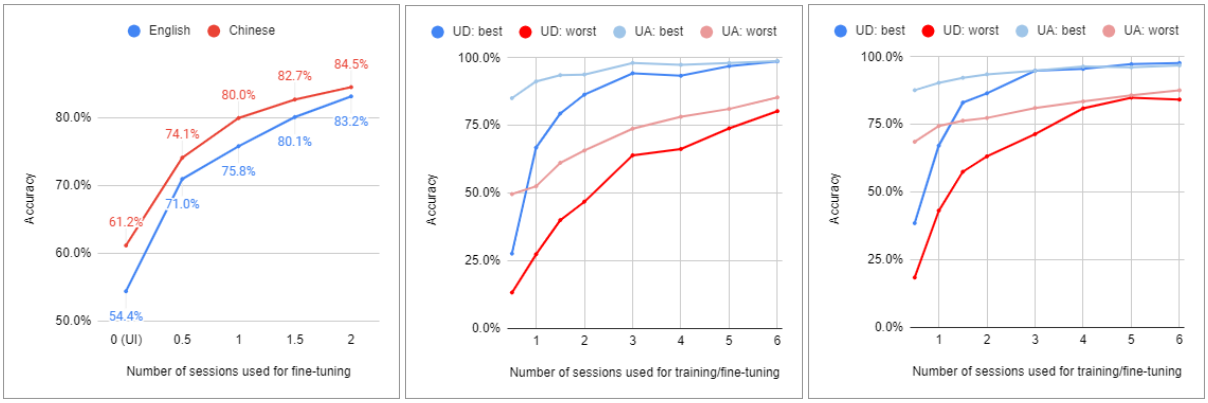
The unsatisfactory UI performance can be compensated by providing a small amount of data from the participant to train a user-adaptive model. We first trained the user-independent model, then we used 2 sessions from the testing participant to fine-tune the model by training 20 epochs. Results show that classification accuracy increases significantly from 54.5% to 83.2% on 54 English commands and from 61.2% to 84.5% on 44 Chinese commands. We also explored the relationship between the amount of data used for fine-tuning and classification performance to see how much fine-tune data is needed. Using only 0.5 session (2 utterances for each command) of training data, the accuracy increases by about 13-16% compared with UI models. Using more data to fine-tune generally leads to better performance, as shown in Figure 11(a). This shows that while there is not enough data to train a user-independent system, the user can provide a small amount of training data (1-2 sessions, 5-15min) and still achieve decent performance.

To remove the possibility that UA model might be over-fitted to a specific participant, we also drew a learning curve with different amount of training data in a UD approach. To achieve this, we gradually increased the amount of sessions used for training from 0.5 to 6, while the testing data unchanged (the last two sessions). All other training parameters are the same as in the UD experiment. To better understand the impact of the amount of training data on different participants, we drew learning curves on participants with best and worst performance from both user studies. We compared the UA and UD learning curves in Figure 11(b)(c). Results show accuracy rises when the amount of training/fine-tuning data increases. Participants with better performance converge faster. When training data is limited (<2 sessions), the UA approach achieves significantly better performance than UD approach by about 7%-60%, showing that the UA approach is not strongly over-fitted to a certain participant.

8 DISCUSSION

8.1 Linguistic Analysis

Modern voiced speech recognition systems are usually able to recognize speech at a phoneme [12] or character [8] level. The ideal silent speech recognition system would also need to distinguish between individual phonemes in order to recognize a much larger set of words in the future. To test the limits of our system and gain a deeper



(a) UA learning curve (b) UD/UA learning curve for English study (c) UD/UA learning curve for Chinese study

Fig. 11. Learning curves. Results acquired by changing the amount of sessions used for training (in UD curves) or fine-tuning (in UA curves). In (a) results are calculated by averaging all participants in English and Chinese study, respectively. In (b) and (c), results for participants with the best (P2 for English, P11 for Chinese) and worst (P4 for English, P13 for Chinese) performance are compared.

understanding of its ability to distinguish between phonemes, we collected data from 10 participants (M_{age} = 20.7; male = 2, female = 8) asking them to silently utter a specially designed list of 72 one-syllable "nonwords".

Each nonword in the list contained two phonemes: a consonant followed by a vowel. This included 18 consonants (p,b,m,f,v,w,r,sh,th,t,d,s,z,n,l,k,g,y) and 4 vowels (i,u,ei,oh) for a total of 72 nonwords. We excluded consonant phonemes that are not allowed at the beginning of words in English such as "ng". The participants were given headphones for this study, which provided audio stimuli to aid with pronunciation. Each nonword appeared on the screen for 1.5 seconds. Apart from the headphones and screen duration, the user study procedure was identical to the procedure described in section 6.3.

We wrote a Python script that analyzed the confusion matrix data from all 10 participants to convert nonword-level confusion into phoneme-level confusion between all consonant pairs and all vowel pairs separately. The confusion matrix in Figure 12 illustrates the average confusion rate for each phoneme pair across all 10 participants. Figure 12 groups the phonemes by certain linguistic features. Namely, the phonemes are organized by their "place of articulation" - a linguistic feature denoting the part of the mouth that constricts to make a particular sound. The main phoneme groupings include labials (full lip closure), dental (lip or tongue touching teeth), lip protrusion, tongue tip, and tongue body. Notice how Figure 12 forms confusion clusters around these groupings. For comparison, were the phonemes to be grouped in a random order, they would show no such clusters around the diagonal confusion line.

According to the confusion clusters, as we move backward from the lips towards the tongue body (down the diagonal line), we see more variance in confusion across groupings. This makes sense since the lips are visible from the POV of an outside camera, while the tongue is not. Still though, the apparent clusters suggest that SpeeChin is able to distinguish between families of phonemes sharing particular linguistic features, even if the articulators responsible for those phonemes are hidden from view. This shows promise for implementing a phoneme-level camera-based SSR system in the future.

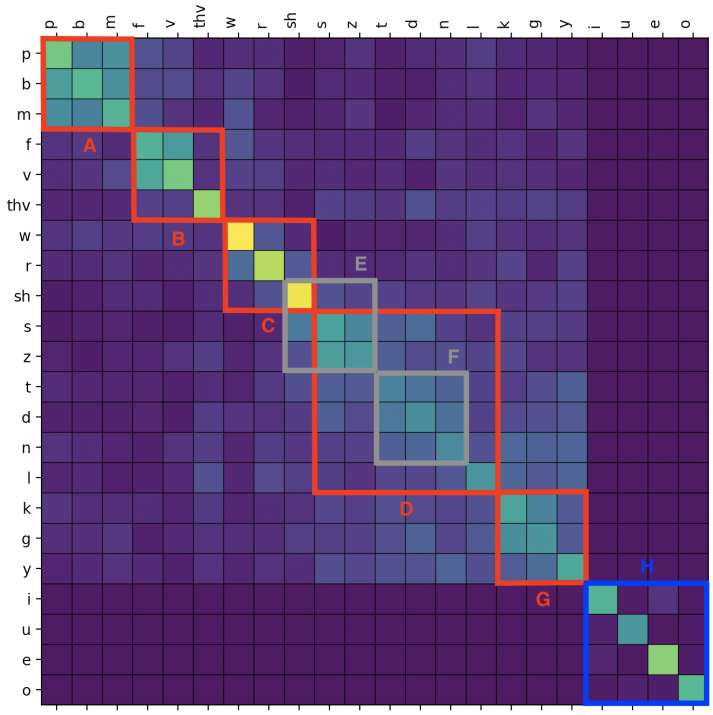


Fig. 12. Phoneme-level confusion pair results grouped by shared linguistic features (A: labials, B: dentals, C: lip protrusion, D: alveolar, E: sibilants, F: alveolar non-sibilants, G: tongue body, H: vowels)

8.2 Performance When the User Is in Motion

In order to evaluate the practicability of applying the system in a mobile setting (e.g., when users are walking), we conducted a small scale mobile setting study. We recruited 6 participants (3 male, 3 female, average age 21.3). We selected 10 Chinese phrases (接听 (answer, JieTing), 挂断 (hang up, GuaDuan), 查看 (check, ChaKan), 扫码 (scan QR code, SaoMa), 打开音乐 (open music, DaKaiYinYue), 手电筒 (flashlight, ShouDianTong), 发送 (send, FaSong), 打开相机 (open camera, DaKaiXiangJi), 是 (yes, Shi), 否 (no, Fou)) and 10 English phrases (Answer, Call, Redial, Play, Previous, Skip, Volume, Next, Up, Down) that can potentially be used while the users are walking. The procedures of the study are the same as static settings study, except that participants were instructed to walk around while they are speaking and that monitor was not used. Participant were asked to wear a bag on the shoulder during the mobile study, in which we put a pair of speakers. The instructions were given as audio through the speakers. In this way, the participants were free to look around instead of staring at a monitor. We used this setup in order to mimic real-life walking scenarios better as well as ensuring safety.

Results show that classification accuracies for 10 English phrases and 10 Chinese phrases are 72.3% (ranging from 40.6% to 91.9%, SD = 17.0%) and 65.5% (ranging from 34.4% to 91.2%, SD = 24.9%), respectively. There is significant variance between different participants. This is because we did not limit the participants' walking style. Some participants rotated their head frequently while walking, while some remained relatively stable. It appears that head movement plays an important role in the final results. However, this issue can potentially be addressed with 1) improving the stability of the device by redesigning the form factor (e.g., the necklace can be combined with a tie clip that can be attached to the clothes), 2) post-processing the images to remove the motion

noise and align the body position, 3) collecting large-scale training data which covers various head movements, and 4) introducing an activation gesture to activate the system. Although the overall accuracy drops significantly compared with static studies, the high variance and good performance on certain participants (up to 91.9%) still demonstrate the potential of making the system practical in mobile settings. We will further investigate this issue in the future.

Most previous wearable-based silent speech recognition technologies were evaluated in a controlled lab-setting, which can only recognize a much smaller set of phrases. TongueBoard [43] and SilentSpeller [38] are among the few that include mobile evaluations. However, they are more obtrusive and less comfortable to wear than SpeeChin. Compared to prior work, we have made significant progress on the wearable-based SSR performance with a less-obtrusive form factor, which is designed towards a SSR technology that has the potential to be deployed in the wild.

8.3 Performance under Specific Application Scenarios

Most applications do not need to distinguish all 54 English commands or all 44 Chinese commands. Depending on the specific application scenario, a subset of the full command list may suffice. For example, texting digits and punctuation on a cellphone can be tedious. For such a scenario, we can extract the 10 digits and 11 punctuation commands from our English dataset and run a separate experiment. Results show that the average accuracy for 10 participants using just these 21 commands is 93.5% (higher than 90.5% with a full command list). Similarly, we can consider a scenario using WeChat and select 10 phrases from the System Group and 9 phrases from the WeChat Group as described in section 6.1.2. Classification accuracy on these 19 phrases is 94.8% (higher than 91.6% with a full command list). This shows that if applying SpeeChin on an application with smaller sets of silent speech commands, its performance can be further improved.

8.4 Activating the SpeeChin System in Daily Scenarios

We envision SpeeChin will serve as a complementary and alternative input method in the future ecosystem of miscellaneous computing devices, where traditional input methods may not best satisfy all the needs from the user (e.g., privacy). In real-life use cases, the system will need to distinguish from intended interaction with the device and various other daily activities, including speaking out loud. To save battery and avoid false-triggering, SpeeChin can operate only when activated. An "activation phrase" that contains a special command (or series of commands) can be selected. Activation phrases have been widely used in real-life interactive systems [2]. For example, "Hey Siri" is usually used to activate the voice assistant "Siri". The activation phrase should 1) have appropriate number of syllables, 2) be easy to memorize and pronounce, 3) be distinct from common daily conversations. An example of such phrases is "Hi SpeeChin". Specifically, the syllable "hi" involves significant chin movement, which is relatively easy to be captured by SpeeChin. Since SpeeChin involves novel hardware, it is also possible to define a special non-verbal "activation gesture", e.g., waving at the camera or touching certain areas of the device. Specifically, users sometimes speak out multiple commands in one utterance (e.g., *Alexa, volume up*, which contains 3 commands), or pause after the activation or between commands (but not long enough to put the system to sleep). To allow for such freedom while using SpeeChin, our utterance detection algorithm (based on scenarios where participants were asked to continuously mouth the phrases in sequences, as described in Section 5.2) will find the exact period of time where the user is mouthing the commands. Further processing and recognition will follow to generate the prediction.

[2]https://www.apple.com/siri/, https://assistant.google.com/

8.5 Using Synthetic Data to Enlarge Training Set

Collecting training data from the user can be expensive and time-consuming in the real-world applications. To address this issue, we plan to use a synthetic non-word method to extend the training set. This approach requires a system to continuously generate 3D human head movements to simulate virtual environment lighting and place virtual cameras on a necklace to collect synthesized data. A series of work in natural language processing and computer vision have demonstrated the possibility of converting text/audio to 2D/3D videos [35, 52]. Specifically, OpenFACS [17] allows generating real-time dynamic facial expressions; Audio2Face platform can generate animated 3D human head movements from just audios [65]. With these studies, we believe it is possible to synthesize datasets on sentence, word, and syllable levels to facilitate training in the future.

8.6 Power Consumption

Powering wearables is a long-standing challenge for wearable technology [61]. The prototype of SpeeChin developed in this paper is intended to show the proof-of-concept, rather than immediately deploying the system in the wild. To provide guidance for future optimization, we evaluated the power consumption of the current prototype. Apart from the data processing pipeline which is run on a remote machine, the sensing system's total power consumption is 5.4W (1.48W for 2 LEDs, 0.92W for camera module, 3W for Raspberry Pi 4B). We plan to further reduce the power consumption by using low power micro-controllers (e.g., ESP32 typically consumes only 0.79W even with wireless module on[3]), using low-power cameras (e.g., OV9755 consumes 100mW at 1280x720@60fps[4]), and reducing the duty cycle of the LEDs.

8.7 Privacy Concerns

Privacy is a concern with many wearable camera-based systems, as they can capture sensitive information in different scenarios. Our device's camera points straight up from the bottom of the neck. In most cases, it would only capture parts of the ceiling or the sky in the background. Even if sensitive backgrounds were to be captured, quality would still below due to our special lighting and filters. For instance, the environment in the background are mostly dark unless there is an near-infrared (NIR) light source. Furthermore, the captured chin and face from the neck arguably have less privacy information of the user compared to the images of the face captured by frontal cameras. For these reasons, humans and background objects captured by SpeeChin camera are arguably less privacy-sensitive than normal RGB cameras. In addition, we expect SpeeChin to be applied in an activate-to-use manner, which means that the system does not need to be always on to avoid capturing sensitive information.

On the other hand, camera-based systems admittedly capture more environmental information than non-camera wearable-based systems. However, SpeeChin is posed to have strong performance on SSR with the user's privacy in mind. In the future, we plan to further explore how to remove the privacy-related information from the IR images and how and where to store the data (e.g., feature extraction on the fly).

8.8 Limitations

Apparently, SpeeChin also has limitations, just like every research prototype/innovation. This section discusses some of the important limitations of the current prototype and possible solutions for the future.

8.8.1 Influence of Strong Sunlight. One limitation of the current system is that segmenting human skin from the backgrounds can be challenging if the camera is directly exposed to strong sunlight in outdoor environments. Figure 13 (b) shows the image captured if there is strong sunlight. To address this issue, there are two possible solutions. One is to train a dedicated machine learning model which segment the skin from the backgrounds in

[3] 3.3V, 240mA, https://www.espressif.com/sites/default/files/documentation/esp32_datasheet_en.pdf
[4] https://www.ovt.com/sensors/OV9755

Fig. 13. Influence of strong sunlight

the image. The other solution is to fine-tune the optical sensing system. For instance, a commodity IR camera (Leap Motion [5]) could be worn on the neck and data can also be collected in strong sunlight, outdoor environment. As shown in Figure 13 (c), the skin is visually distinct from the backgrounds. It shows that using more dedicated filters and cameras can potentially address this issue. A recent work with a similar setup that tested Leap Motion under strong sunlight conditions also demonstrates that performance can be significantly improved compared to IR cameras [10].

8.8.2 Influence of Clothes and Hair. Another apparent limitation of the system is that the camera can not be blocked. As a result, if the user has long hair that may cover the lens or a cloth that block or pushing the camera, the system would not perform well.

8.8.3 Influence of Fluency Level of the Participants. In the user study of SpeeChin, 10 participants were recruited for each language. All participants self-reported as fluent in the language used in their experiment. Specifically, all participants in the Chinese study were native Chinese speakers while all participants in the English study were not native English speakers. We acknowledge that this can introduce a bias in the results. If the participants were not native speakers, it could impact the results, both negatively and positively. Non-native speaker can potentially speak slower than the native speaker, which can positively impact the results. On the other hand, non-native speakers also face more challenges pronouncing certain words, which makes it more challenging to be distinguished. Native speakers tend to speak more consistently. We will leave this part to the future when more data is available.

9 CONCLUSION

In this paper, we present SpeeChin, the first smart necklace that can recognize 54 English and 44 Chinese silent speech commands from images of the neck and face captured by a necklace-mounted IR camera. These images are sent to a customized pre-processing pipeline and an end-to-end CRNN model to infer silent speech commands. A user study with 20 participants shows that SpeeChin can recognize 54 English and 44 Chinese commands with accuracies of 90.5% and 91.6%, respectively. A third study with 10 participants was conducted to analyze how SpeeChin would distinguish between different phonemes. Based on the above results, we discuss the opportunities and challenges in applying SpeeChin in real-world applications in the future.

ACKNOWLEDGMENTS

This work is supported by the Department of Information Science at Cornell University and partially by the SJTU-Cornell seed grant from Cornell China Center. We would like to thank all participants for their participation in the user study during such special times. We would also like to thank all reviewers and lab mates at Cornell SciFiLab for their valuable comments and feedback on the manuscript.

[5] https://www.ultraleap.com/product/leap-motion-controller/

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)*. IEEE, 2516–2520.

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

[3] Daniel Barolet, François Christiaens, and Michael R Hamblin. 2016. Infrared and skin: Friend or foe. *Journal of Photochemistry and Photobiology B: Biology* 155 (2016), 78–85.

[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.

[5] Linnar Billman and Johan Hullberg. 2018. Speech Reading with Deep Neural Networks.

[6] Alper Bozkurt and Banu Onaral. 2004. Safety assessment of near infrared light emitting diodes for diffuse optical measurements. *biomedical engineering online* 3, 1 (2004), 1–10.

[7] Jonathan S Brumberg, Alfonso Nieto-Castanon, Philip R Kennedy, and Frank H Guenther. 2010. Brain–computer interfaces for speech communication. *Speech communication* 52, 4 (2010), 367–379.

[8] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. 2015. Listen, attend and spell. *arXiv preprint arXiv:1508.01211* (2015).

[9] 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 *International Conference on Biomedical Electronics Devices*.

[10] Tuochao Chen, Yaxuan Li, Songyun Tao, Hyunchul Lim, Mose Sakashita, Ruidong Zhang, Francois Guimbretiere, and Cheng Zhang. 2021. NeckFace: Continuously Tracking Full Facial Expressions on Neck-mounted Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–31.

[11] 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

[12] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *arXiv preprint arXiv:1506.07503* (2015).

[13] Joon Son Chung and Andrew Zisserman. 2016. Lip reading in the wild. In *Asian Conference on Computer Vision*. Springer, 87–103.

[14] Joon Son Chung and AP Zisserman. 2017. Lip reading in profile. (2017).

[15] 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.*

[16] 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 *Interspeech 2017*.

[17] Vittorio Cuculo and Alessandro D'Amelio. 2019. OpenFACS: an open source FACS-based 3D face animation system. In *International Conference on Image and Graphics*. Springer, 232–242.

[18] Bruce Denby, Yacine Oussar, Gérard Dreyfus, and Maureen 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. IEEE, I–I.

[19] Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, Jim M Gilbert, and Jonathan S Brumberg. 2010. Silent speech interfaces. *Speech Communication* 52, 4 (2010), 270–287.

[20] Ariel Ephrat and Shmuel Peleg. 2017. Vid2speech: speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5095–5099.

[21] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction*. 445–450.

[22] 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)*. IEEE, 2511–2515.

[23] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-based Silent Speech Interface for Smartphone Users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–27.

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

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[27] Tatsuya Hirahara, Makoto Otani, Shota Shimizu, Tomoki Toda, Keigo Nakamura, Yoshitaka Nakajima, and Kiyohiro Shikano. 2010. Silent-speech enhancement using body-conducted vocal-tract resonance signals. *Speech Communication* 52, 4 (2010), 301–313.
[28] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[29] 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.
[30] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *CoRR* abs/1502.03167 (2015). arXiv:1502.03167 http://arxiv.org/abs/1502.03167
[31] 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.
[32] 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.
[33] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. Alterego: A personalized wearable silent speech interface. In *23rd International Conference on Intelligent User Interfaces*. 43–53.
[34] 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 *Machine Learning for Health Workshop*. 25–38.
[35] Muhammad Zeeshan Khan, Saira Jabeen, Muhammad Usman Ghani Khan, Tanzila Saba, Asim Rehmat, Amjad Rehman, and Usman Tariq. 2020. A realistic image generation of face from text description using the fully trained generative adversarial networks. *IEEE Access* 9 (2020), 1250–1260.
[36] M. Kim, B. Cao, T. Mau, and J. 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.
[37] 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)*.
[38] 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*. 1–5.
[39] 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*. 1–8.
[40] 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*. 1–11.
[41] Alexandros Koumparoulis, Gerasimos Potamianos, Youssef Mroueh, and Steven J Rennie. 2017. Exploring ROI size in deep learning based lipreading.. In *AVSP*. 64–69.
[42] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.
[43] 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*. 1–9.
[44] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. 2020. Deep learning for generic object detection: A survey. *International journal of computer vision* 128, 2 (2020), 261–318.
[45] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi. 2017. A survey of deep neural network architectures and their applications. *Neurocomputing* 234 (2017), 11–26.
[46] Sarfaraz Masood, Adhyan Srivastava, Harish Chandra Thuwal, and Musheer Ahmad. 2018. Real-time sign language gesture (word) recognition from video sequences using CNN and RNN. In *Intelligent Engineering Informatics*. Springer, 623–632.
[47] 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 (2018), 046031.
[48] Daniel Michelsanti, Olga Slizovskaia, Gloria Haro, Emilia Gómez, Zheng-Hua Tan, and Jesper Jensen. 2020. Vocoder-Based Speech Synthesis from Silent Videos. *arXiv preprint arXiv:2004.02541* (2020).
[49] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. 2021. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
[50] Yoshitaka Nakajima, Hideki Kashioka, Nick Campbell, and Kiyohiro Shikano. 2006. Non-audible murmur (NAM) recognition. *IEICE TRANSACTIONS on Information and Systems* 89, 1 (2006), 1–8.
[51] Yoshitaka Nakajima, Hideki Kashioka, Kiyohiro Shikano, and Nick Campbell. 2003. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, Vol. 5. IEEE, V–708.
[52] Osaid Rehman Nasir, Shailesh Kumar Jha, Manraj Singh Grover, Yi Yu, Ajit Kumar, and Rajiv Ratn Shah. 2019. Text2facegan: Face generation from fine grained textual descriptions. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE,

58–67.
[53] N. Otsu. 1979. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics* 9, 1 (1979), 62–66. https://doi.org/10.1109/TSMC.1979.4310076
[54] Laxmi Pandey and Ahmed Sabbir Arif. 2021. Silent Speech and Emotion Recognition from Vocal Tract Shape Dynamics in Real-Time MRI. *arXiv preprint arXiv:2106.08706* (2021).
[55] Anne Porbadnigk, Marek Wester, and Tanja Schultz Jan-p Calliess. 2009. EEG-based speech recognition impact of temporal effects. (2009).
[56] Ahmed Rekik, Achraf Ben-Hamadou, and Walid Mahdi. 2014. A new visual speech recognition approach for RGB-D cameras. In *International conference image analysis and recognition*. Springer, 21–28.
[57] 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*. 47–54.
[58] Tanja Schultz. 2010. ICCHP keynote: Recognizing silent and weak speech based on electromyography. In *International Conference on Computers for Handicapped Persons*. Springer, 595–604.
[59] B. Shi, X. Bai, and C. Yao. 2017. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *IEEE Transactions on Pattern Analysis Machine Intelligence* 39, 11 (nov 2017), 2298–2304. https://doi.org/10.1109/TPAMI.2016.2646371
[60] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[61] Thad Starner. 2001. The challenges of wearable computing: Part 1. *Ieee Micro* 21, 4 (2001), 44–52.
[62] 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
[63] Yohei Tanaka, Kiyoshi Matsuo, and Shunsuke Yuzuriha. 2010. Long-term histological comparison between near-infrared irradiated skin and scar tissues. *Clinical, cosmetic and investigational dermatology: CCID* 3 (2010), 143.
[64] Abhinav Thanda and Shankar M Venkatesan. 2016. Audio visual speech recognition using deep recurrent neural networks. In *IAPR workshop on multimodal pattern recognition of social signals in human-computer interaction*. Springer, 98–109.
[65] Guanzhong Tian, Yi Yuan, and Yong Liu. 2019. Audio2face: Generating speech/face animation from single audio with attention-based bidirectional lstm networks. In *2019 IEEE international conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 366–371.
[66] 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 *Interspeech 2018*.
[67] Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2019. Video-driven speech reconstruction using generative adversarial networks. *arXiv preprint arXiv:1906.06301* (2019).
[68] 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)*. IEEE, 6115–6119.
[69] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I Hong, Carmel Majidi, and Swarun Kumar. 2019. Rfid tattoo: A wireless platform for speech recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–24.
[70] 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), 442.
[71] Alexander Wunsch and Karsten Matuschka. 2014. A controlled trial to determine the efficacy of red and near-infrared light treatment in patient satisfaction, reduction of fine lines, wrinkles, skin roughness, and intradermal collagen density increase. *Photomedicine and laser surgery* 32, 2 (2014), 93–100.
[72] 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)*. IEEE, 548–555.
[73] Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. 2017. Comparative study of CNN and RNN for natural language processing. *arXiv preprint arXiv:1702.01923* (2017).
[74] ZhangYongzhao, HuangWei-Hsiang, YangChih-Yun, WangWen-Ping, ChenYi-Chao, YouChuang-Wen, HuangDa-Yuan, XueGuangtao, and YuJiadi. 2020. Endophasia: Utilizing Acoustic-Based Imaging for Issuing Contact-Free Silent Speech Commands. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (2020).
[75] Guoying Zhao, Mark Barnard, and Matti Pietikainen. 2009. Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia* 11, 7 (2009), 1254–1265.