Endophasia: Utilizing Acoustic-Based Imaging for Issuing Contact-Free Silent Speech Commands

YONGZHAO ZHANG, Shanghai Jiao Tong University, China
WEI-HSIANG HUANG, National Chiao Tung University, Taiwan
CHIH-YUN YANG, National Chiao Tung University, Taiwan
WEN-PING WANG, National Chiao Tung University, Taiwan
YI-CHAO CHEN*, Shanghai Jiao Tong University, China
CHUANG-WEN YOU, National Taiwan University, Taiwan
DA-YUAN HUANG, National Chiao Tung University, Taiwan
GUANGTAO XUE*, Shanghai Jiao Tong University, China
JIADI YU, Shanghai Jiao Tong University, China

Using silent speech to issue commands has received growing attention, as users can utilize existing command sets from voice-based interfaces without attracting other people's attention. Such interaction maintains privacy and social acceptance from others. However, current solutions for recognizing silent speech mainly rely on camera-based data or attaching sensors to the throat. Camera-based solutions require 5.82 times larger power consumption or have potential privacy issues; attaching sensors to the throat is not practical for commercial-off-the-shell (COTS) devices because additional sensors are required. In this paper, we propose a sensing technique that only needs a microphone and a speaker on COTS devices, which not only consumes little power but also has fewer privacy concerns. By deconstructing the received acoustic signals, a 2D motion profile can be generated. We propose a classifier based on convolutional neural networks (CNN) to identify the corresponding silent command from the 2D motion profiles. The proposed classifier can adapt to users and is robust when tested by environmental factors. Our evaluation shows that the system achieves 92.5% accuracy in classifying 20 commands.

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

Additional Key Words and Phrases: silent command, acoustic-based imaging, mobile devices

ACM Reference Format:
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

*Corresponding authors.

Authors' addresses: Yongzhao Zhang, Shanghai Jiao Tong University, China, e-mail:zhangyongzhao@sjtu.edu.cn; Wei-Hsiang Huang, National Chiao Tung University, Taiwan, e-mail:wei.hsiang.tw@gmail.com; Chih-Yun Yang, National Chiao Tung University, Taiwan, e-mail:jean123456789@kimo.com; Wen-Ping Wang, National Chiao Tung University, Taiwan, e-mail:emilywang@nctu.edu.tw; Yi-Chao Chen, Shanghai Jiao Tong University, China, e-mail:yichao@sjtu.edu.cn; Chuang-Wen You, National Taiwan University, Taiwan, e-mail:cwyou2004@gmail.com; Da-Yuan Huang, National Chiao Tung University, Taiwan, e-mail:dayuan.huang.tw@gmail.com; Guangtao Xue, Shanghai Jiao Tong University, China, e-mail:xue-gt@cs.sjtu.edu.cn; Jiadi Yu, Shanghai Jiao Tong University, China, e-mail: jdyu@cs.sjtu.edu.cn.

2474-9567/2020/3-ART37 $15.00
https://doi.org/10.1145/3381008

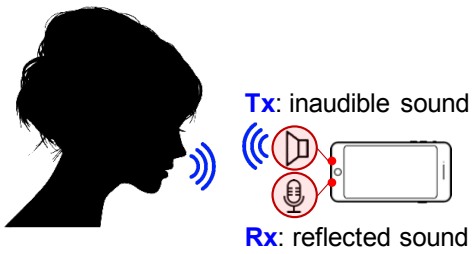


Fig. 1. In Endophasia, inaudible sounds are generated from a mobile device, radiated toward all directions, reflected by the face, and received by the mobile device.

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

1 INTRODUCTION

The advancement and proliferation of mobile and wearable devices have enabled a broad spectrum of applications. To interact naturally with those devices, people design various natural user interfaces (NUI), including gesture-based, touch, and voice inputs, which allow people to issue commands at any point in their daily lives. Among these, voice input is one of the most natural input options for interactions with those devices using everyday speech; however, this method suffers from fundamental limitations, e.g., privacy concerns or inference from background noise or the voices from others, that prevent this method from being adopted widely. To maintain the simplicity of natural human speech interaction while preserving a user's privacy, we present an acoustic-based solution to identify command words, i.e., an intuitive word such as "Pause", 'mouthed' (i.e., spoken without voicing) by users through measuring the movements and shape changes of a user's mouth as they silently 'speak' the words.

In this work, we propose Endophasia, which utilizes acoustic-based face imaging to detect silent speech commands. This method relies on a built-in speaker and a microphone in COTS mobile devices to transmit and receive inaudible audio signals. The received audio is deconstructed to extract the signals reflected by various parts of the face and a deep learning classifier is adopted to identify the commands.

There are several challenges implementing Endophasia: i) The process of acoustic-based imaging requires a microphone array [5, 8, 23] or requires users to move their phone following a specific trajectory [25, 30, 40]. However, most COTS phones have 1 to 3 microphones facing different angles, so forming a microphone array on the phone is not feasible or creates a factor of poor quality. It is also impractical to ask users to move their phones or other devices precisely while issuing a silent speech command. ii) For the same command, the reflected acoustic signals could be different among users due to the subtle differences in their lip movements or the signals reflected by the faces; therefore, it's difficult to find a universal model that works well for everyone. However, training a model for an individual user is time-consuming and imposes significant effort on new users.

To address the first challenge, we propose to send a known acoustic training sequence and use the received signals to compute the channel impulse response (CIR). As a result, the received signals are deconstructed into impulse at various channel taps. By analyzing the phase shift in each tap, we can profile the motion patterns from face components with various distances to the device. To address the second challenge, we propose a three-step training scheme. In the cold start step, we fine-tune a standard *ResNet18* CNN network on a large training data-set. To reduce the burden for a new user, in the warm start step, we use a transfer learning scheme to reduce the

training data required for the new user. In the online learning step, we allow users to utilize the unlabeled samples collected while using Endophasia to improve system performance continuously.

We demonstrate the technical feasibility of Endophasia by implementing it on a mobile phone. We tested the system using 20 commands: “Search”, “Home”, “ScreenShot”, “Skype”, “Camera”, “Play”, “Skip”, “Pause”, “Previous”, “Mute”, “Answer”, “Call”, “Check”, “Copy”, “Cut”, “HangUp”, “Help”, “Undo”, “Paste”, and “Redial”. The results from 14 participants show 88.83% within-user accuracy. In cross-user experiments, although leaving-one-person-out shows low cross-user accuracy, by adding a small number of a new user’s samples, we can achieve 92.5% accuracy.

Our contributions include:

- We propose an approach to identify silent speech commands using a built-in speaker and a microphone using COTS mobile devices.
- We propose to use channel impulse response to deconstruct the reflected acoustic signals and profile the motion patterns of various face components.
- We propose to use the transfer learning scheme to reduce the data collection time for a new user and use Few-Shot Adversarial Domain Adaptation scheme to address the issue of unbalanced data.
- A semi-supervised learning scheme is adopted to utilize the unlabeled motion profiles, which are collected when the user uses the system over time to continuously update the model and improve the classification accuracy.
- We perform an extensive evaluation to show the accuracy and robustness of Endophasia.

This paper is organized as follows. Sec. 3 details the system design. In Sec. 4, we extensively evaluate the performance of Endophasia. Related works are introduced in Sec. 2. Lastly, In Sec. 6, we conclude the work.

2 RELATED WORK

2.1 Silent Speech Interface

The idea of Silent Speech Interface (SSI) [11] is raised because of the increasing possibility of speech processing but without an intelligible acoustic signal. The SSI is mainly applied in two typical scenarios: i) As an aid for the speech-handicapped. ii) Operating in silence-required or high-background-noise environments as part of the communication system. Several SSIs have been proposed to recognize inaudible speech with either bulky/invasive sensor deployment or non-invasive wearable devices. By invasively implanting/placing sensors within human bodies, researchers proposed solutions to recognize the brain activities in the speech motor cortex [6], or capture tongue and jaw movements with in-mouth magnetic beads [21] or capacitive touch sensors [26]. The inconveniences caused by these invasive solutions impede their widespread use. To offer more practical or affordable solutions, other studies have designed schemes to identify speech content by using alternative sensors (e.g., EEG [37], sEMG [32], ultrasound imaging [24]) to detect tongue, facial, throat movements, and microphones attached to the skin to hear non-audible murmurs (NAM) [20, 35, 36] or put close to the front of the mouth to capture whisper-like tiny voice while ingressive breathing [17]. Although these solutions are non-invasive, they require attaching specialized sensors on the human body. Attaching sensors for the speech-handicapped is an acceptable solution to help them to ‘speak’ again. However, in general cases, SSIs are mainly used to deal with the occasional challenging scenarios, e.g., silence-required environments, for most users. Therefore, more feasible and convenient solutions are proposed.

2.2 Camera-Base Solutions

A natural idea is to utilize the cameras on COTS devices. Hence no additional sensors are required. Therefore, some recent works are proposed to use camera-based techniques for lip reading. LipNet [2] proposed an end-to-end sentence-level lip reading method, with a high accuracy of up to 95.2% in GRID Corpus [10]. However, the GRID Corpus is very different from the daily conversation scenarios because of the limited command set and the

same grammar in all sentences. Chung etc. conducted experiments on the Oxford-BBC Lip Reading in the Wild (LRW) dataset [9], which consists of up to 1000 utterances of 500 different words spoken by hundreds of different speakers in the wild. They only achieve low recognition accuracy in general sentence lip reading with a word error rate of 50.2%, which is unusable for general purpose applications. This is mainly caused by two reasons: i) In general speech, some phonemes often result in similar lip movements. ii) In uncontrolled environments, or in the daily conversation, the amount of information carried by lip movements is not sufficient for general speech recognition, because sometimes people may do little mouth movements when speaking. To make the camera-based silent speech truly usable on COTS devices in daily life, Sun, etc. proposed Lip-Interact [46], which supports 44 pre-defined commonly used Chinese commands for mobile phones, instead of recognizing general speech. They use the front camera in a mobile phone to capture users' mouth movements, achieving an average accuracy of 95.464%. Nevertheless, using a camera on mobile devices has many limitations, such as (1) unstable recognition accuracy (dim light decreases accuracy). (2) Currently, many mobile devices do not have a camera, like smartwatches and Google glass. (3) Tiny wearable devices cannot support long periods of video recording because of power consumption. (4) Taking videos for lip reading may occupy the full capability of a camera, implying that this camera can not execute other concurrent tasks, like taking a video call. To address these issues, Endophasia utilizes an acoustic-based imaging solution to use inaudible sound produced by COTS mobile devices.

2.3 Acoustic-based Tracking and Imaging

Many acoustic-based tracking technologies using COTS mobile devices have been proposed because of the wide availability of microphones and speakers. These works transmitted inaudible acoustic signals using speakers in various waveform including sinusoidal waves [27, 48, 52, 53], FMCW [29, 31, 51, 55], PN sequence [38], GSM sequence [54], Zadoff–Chu sequence [47], etc. When the reflected signals are captured, according to the type of waveform used, different methods are proposed to estimate the distances which include the Doppler frequency shift [3, 27, 28, 53], phase change [51, 52], channel state [47, 54], time-of-flight [38, 55], envelope difference [48], or the frequency of the mixed signals [29, 31, 55]. The methods proposed in these works can potentially be used to track facial movements for identifying silent speech commands; however, these methods either require more than one microphone or speaker to track multiple objects or have limited resolutions to distinguish objects which are as close as lips. Unlike these works, Endophasia deconstructs the signals reflected from various face components and generates corresponding 2D motion profiles. The 2D motion profiles can distinguish reflectors which are more than $0.7cm$ apart and provide information with finer granularity for silent command classification.

Recently, acoustic-based imaging technologies are also proposed to use one speaker and one microphone on COTS mobile phones to reconstruct the 2D structure of the space or objects. AIM [30] used a speaker and a microphone on a mobile phone to send acoustic signals and apply Synthetic Aperture Radar technology to reconstruct the 2D image of a remote object. SAMS [40] used a speaker and a microphone on a mobile phone to send/receive FMCW signals and estimate the distance from the phone to a wall. By combining the distances with the walking trajectory estimated from the accelerometer, SAMS can reconstruct the 2D map of the building. Although these works can obtain 2D images of the target object, there are two issues that prevent them from being applicable to silent command recognition. First, these works require the phone to move in a specific pattern to image an object; however, that is impractical to ask users to move their phones precisely while issuing a silent speech command. Second, these works assume the targeted object is stationary during the imaging process; however, the face components are changing while issuing a command.

3 SYSTEM DESIGN

Audio input and output quality of mobile devices are becoming better and better due to the evolvement of hardware and software of speakers/mics. However, the high frequency band ($17 - 24$KHz) may be easily ignored

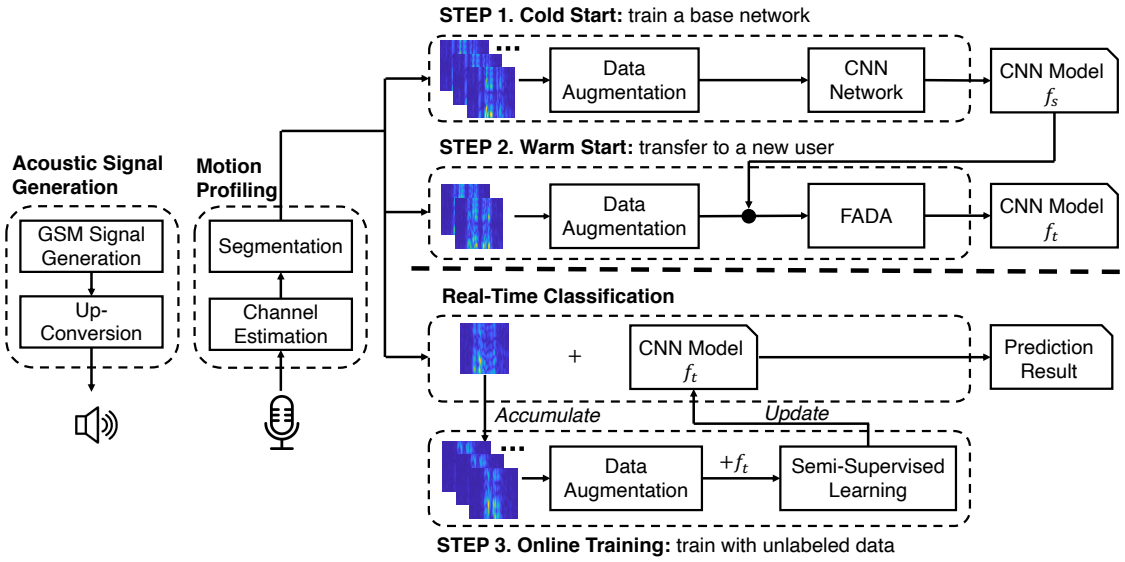


Fig. 2. System overview.

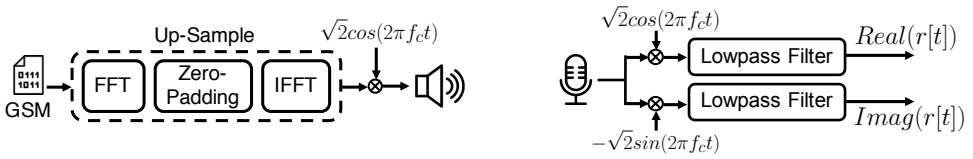
by researchers and engineers, because acoustic signals in this frequency range are inaudible to most adults. Endophasia fully utilizes acoustic signals in the inaudible band to capture the user's facial movements. To use Endophasia, a user holds a phone and puts it close to the mouth, pressing the volume button to start recording. The acoustic signal is emitted and received by speaker/mic on the lower side. Then the user silently issues a command by mouthing the verbal command but not vocalizing the sound, pressing the volume button again to stop recording. The inaudible acoustic signal will be reflected by the moving facial components and received by the microphone. Finally, Endophasia will predict the command and trigger the corresponding functionalities.

3.1 System Overview

Endophasia uses inaudible sound signals actively transmitted and received by mobile devices to capture changes in mouth shape while giving silent commands. As shown in Fig. 1, inaudible sounds generated from a built-in speaker on a mobile device travel through a straight line, are then reflected by the face and received by a built-in microphone. Because most commonly available speakers are omnidirectional, a single transmitted signal radiates in all directions in free space and will reach the microphone via multiple paths (e.g., paths going through different reflectors). Therefore, the received signal is a superposition of multiple signals with various delays.

The intuition behind Endophasia is that, while giving a silent speech command, the user's face, especially the mouth, changes accordingly over time. If we can deconstruct audio signals reflected by the various parts of the face and monitor the changes, we may infer the corresponding silent command.

Fig. 2 shows the system flow of Endophasia. To identify a silent command, Endophasia generates and upsamples a Global System for Mobile (GSM) training sequence to produce inaudible sound (Sec. 3.2). The reflected signals are captured by a microphone and used to estimate channel states and segmented to produce a motion profile corresponding to facial movements (Sec. 3.3). In the training phase, we first collect a large amount of data from various users to train a CNN base network in the cold start step (Sec. 3.4.1). In the warm start step, in order to transfer the knowledge for a new user, a transfer learning technique is adopted to adapt the model to unseen



(a) Transmitting inaudible audio.

(b) Receiving reflected audio.

Fig. 3. Transmitter and receiver design.

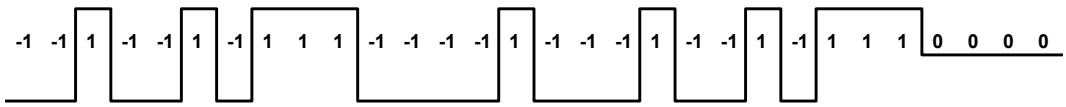


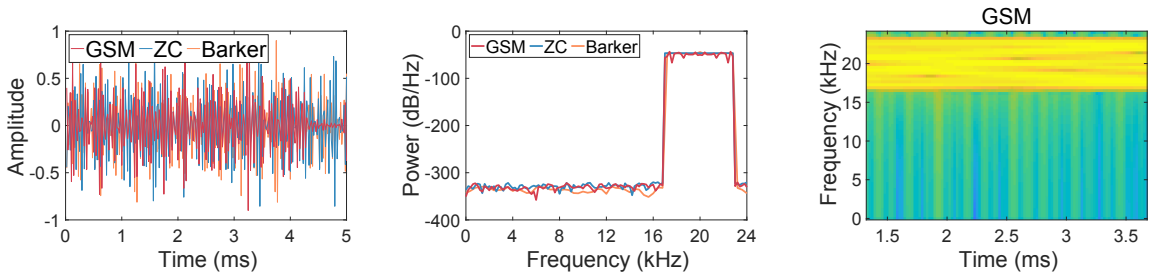
Fig. 4. 26bit GSM training sequence with 4 guard bits.

users (Sec. 3.4.2). In the real-time prediction phase, the motion profiles are fed into the customized model and return prediction results. At the meanwhile, these unlabeled motion profiles are also collected for the online learning step where a semi-supervised learning scheme is adopted to utilize these unlabeled motion profiles to continuously update the model and improve the classification accuracy (Sec. 3.4.3). We detail each system component hereafter.

3.2 Transmitting Inaudible Audio

Issuing silent speech commands involves the movements of various face components, including the upper lip, lower lip, cheeks, etc. The received signals, therefore, are the superposition of signals reflected from each of these components. In order to accurately identify a silent command, we need to deconstruct the received signals to reveal the movement patterns of these face components. To achieve that, we borrow the idea from wireless communication, where we send a known acoustic training sequence and use the received signals to compute the channel impulse response (CIR). CIR is a characterization of all signal traversal paths with different delays and magnitudes [42]. Specifically, it is a vector of channel taps where each channel tap corresponds to multi-path effects within a specific delay range. Because the reflected signals from various face components travel through different paths with different lengths, by focusing on the CIR changing patterns of certain channel taps whose delay ranges are close to the distances to these face components, we can effectively profile their movements. Fig. 3(a) further details the signal generation and transmission process.

GSM Signal Generation: As described prior, a transmitter sends a known acoustic training sequence for channel estimation. Let $S = \{s_1, ..., s_K\}$ denotes the training sequence, where K is the length of the sequence. It can be any random bits. We choose the 30-bit GSM sequence [15, 41], where 4 guard bits are zero-padded. As shown in Fig. 4, we modulate S to the Binary Phase Shift Keying (BPSK) symbols, where bits 0 and 1 are mapped to baseband symbols 1 and -1, respectively. GSM sequence has the property of high autocorrelation [1], which will be utilized for channel estimation (see Sec. 3.3). There are some commonly-used training sequences having the same property, such as the Barker sequence [18] and chirp-like Zadoff-Chu (ZC) sequence [39], etc. These three sequences provide similar performance for Endophasia (see Sec. 4.2.3). We chose the GSM training sequence because it has been widely used in single carrier communication and is known to have desirable properties for synchronization and channel estimation.



(a) Time-domain sequences. (b) Power spectrum. (c) Spectrogram.

Fig. 5. The representations of the up-sampled source trinaing sound of GSM, ZC, and Barker sequence in time and frequency domain.

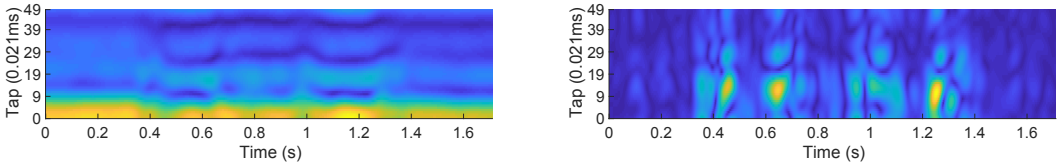
Up-Sampling: To transmit a modulated symbol over the inaudible frequency band, we first need to up-sample the signal to reduce its bandwidth so that it does not exceed the maximum allowed bandwidth of the inaudible band. Let f_S and BW denote the sampling rate and the channel bandwidth, respectively. To limit the bandwidth of the transmitted symbol, we use the fast Fourier Transform (FFT) to convert the time domain audio signal to the frequency domain and perform zero-padding to limit the bandwidth of the audio signal. Then we perform the inverse fast Fourier Transform (IFFT) to convert it back to the time domain.

After the up-sampling, the length of the audio signal becomes $30 \times f_S/BW$. In Endophasia, we set $f_S = 48000$, which is a common audio sampling rate and supported by most of the available mics, speakers, and mobile devices. We chose $BW = 6000$ (i.e., set the bandwidth to $6KHz$), so the length of the up-sampled signals is 240. That is, each GSM signal can be transmitted within $240/f_S = 5ms$. The advantage of using such a short sequence is that it can better capture the quick movements of the face and the mouth while giving silent commands. The downside of using a short signal is the inter-symbol interference (ISI) [16]. ISI is a form of distortion in which one symbol interferes with subsequent symbols. In our case, if the reflected path is longer than $240/f_S \times v_s/2 = 84cm$ ($v_s = 343m/s$ denotes the sound speed in the air), the reflected signal overlaps the next signal resulting in ambiguous reflectors. Fortunately, Endophasia requires users to move their mobile device close to their face (e.g., a few centimeters) and all the face components related to silent commands are with $84cm$ from the mobile device; then, after having the up-sampled audio signals, we up-convert the signal to transmit it over the inaudible band. Let f_c denote the central frequency of the passband. We change the frequency of the signal by multiplying $\sqrt{2}cos(2\pi f_c t)$ to the baseband signal: $x(t) = \sqrt{2}cos(2\pi f_c t)s(t)$, where $s(t)$ and $x(t)$ are up-sampled baseband and passband signals, respectively. After up-conversion, the sequence is normalized before being played by a speaker. Fig. 5 shows the acoustic sound generated from GSM, ZC, and Barker training sequences in time and frequency domain, respectively. We omit the spectrogram of ZC and Barker sequences because they look similar to that of GSM as shown in Fig. 5(c). As expected, the bandwidth usage of the source sound manifests from 17KHz to 23KHz for all of the three training sequences.

3.3 Receiving Audio

Fig. 3(b) illustrates the signal reception and baseband conversion process.

Channel Estimation: The received passband signal $y(t)$ arriving at the microphone is converted into a baseband symbol $r[n]$ using the following down-conversion process: $y(t)$ is multiplied by $2cos(2\pi f_c t)$ and $-2sin(2\pi f_c t)$ to get the real and imaginary parts of the received baseband symbol, respectively. We then perform low-pass filtering to remove background noise. This gives us the following baseband signal:



(a) CIR. (b) Differential CIR.

Fig. 6. The CIR and differential CIR estimation of the command "ScreenShot".

$$\begin{aligned} r(t) &= \sqrt{2}cos(2\pi f_c t)y(t) - j\sqrt{2}sin(2\pi f_c t)y(t) \\ &= 2e^{-j2\pi f_c t}y(t) \end{aligned}$$

The received signal via multi-path is traditionally modeled as the Linear Time-Invariant (LTI) system: suppose the free space has L paths and the received signal from the path i has delay τ_i and amplitude a_i which is determined by the propagation distance between the paths and reflectors. Then the received signal $rx(t)$ can be modeled as the summation of L signals:

$$rx(t) = \sum_{i=1}^{L} a_i tx(t - \tau_i) = h(t) \cdot tx(t)$$

where $tx(t)$ is the transmitted passband audio at time t and $h(t)$ is the channel impulse response (CIR). In the analog world, $h(\cdot)$ is a continuous function and modeled using Dirac's delta function [14]. However, because a microphone captures baseband audio symbols as a discrete output of $h(\cdot)$ sampled at T_s, CIR is regarded as the discrete-time filter in the LTI system:

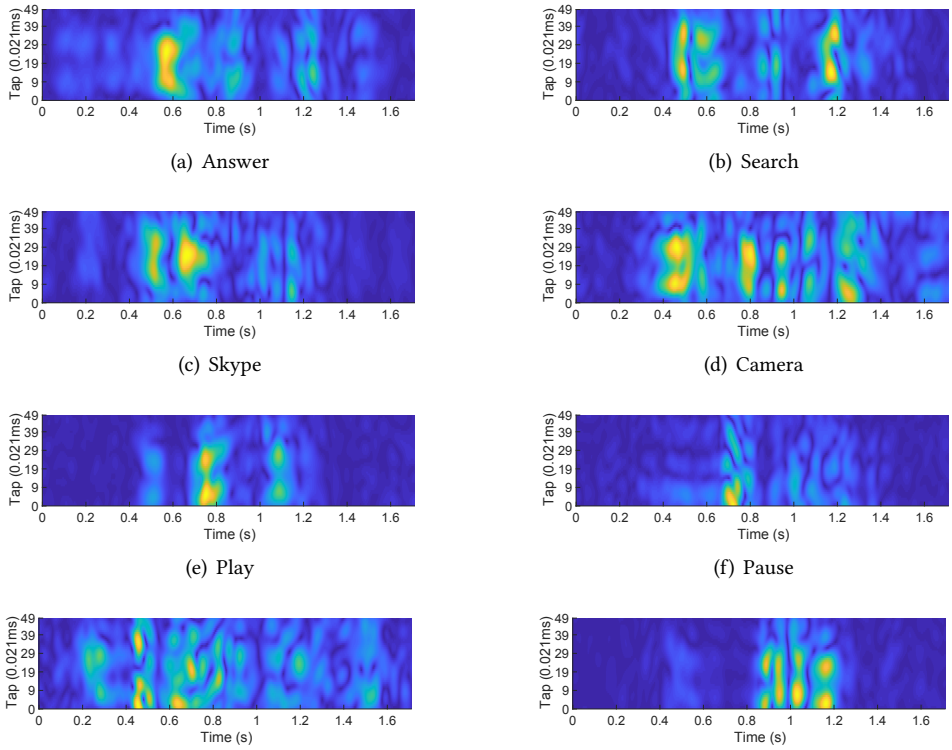
$$h[n] = \sum_{i=1}^{L} a_i e^{-j2\pi f_c \tau_i} sinc(n - \tau_i W)$$

where $h[n]$ is called the n^{th} channel tap, $sinc(t) = \frac{sin(\pi t)}{\pi t}$, and f_c represents the central frequency of the transmitted signals. $sinc$ function which is also called sampling function has a high and narrow peak at the point where the delay τ matches the delay of the corresponding n^{th} tap (i.e., $n = \tau_i W$). Because the auto-correlation of a GSM sequence is close to zero with any delay within one period, we can use the auto-correlation as an approximation for the CIR $h[n]$:

$$h[n] \approx h'[n] = gsm^*_{rx}(-n) * gsm_{tx}(n)$$

where $gsm_{rx}(t)$ and $gsm_{tx}(t)$ represent the received and transmitted GSM signals at time t, respectively; the operation * represents the Hermitian transpose.

In Endophasia, $h'[n]$ is sampled with an interval of $T_s = 1/f_S = 0.021ms$ where $f_S = 48000$ representing the audio sampling rate. It implies the propagation distance is $0.7cm$ ($343m/s \times 0.021ms$) per tap. In other words, at any time t, we can monitor the reflected signals in each of 240 taps where audio signals in the tap n correspond to those traveling through a distance of $0.7 \times n$ cm. Therefore, Endophasia can distinguish two reflectors, which are more than $0.7cm$ apart. The resolution is enough for us to capture the changes occurring in the posture of the lips.



(a) Answer (b) Search (c) Skype (d) Camera (e) Play (f) Pause (g) Previous (h) Mute

Fig. 7. Differential CIR estimation of various silent speech commands.

Facial movements while giving silent commands incur both magnitude and phase changes $h'[n]$. Fig. 6(a) shows the magnitude of the CIR when the user gave the command "Search". The x-axis represents time t while the y-axis represents taps n. Regions in red indicate a strong reflection at the corresponding tap (i.e., distance) in the CIR estimation. While we can observe there are several reflectors in the CIR estimation and they change while giving the command, it is difficult to distinguish the reflected signals as they are much weaker than those from the line-of-sight path or self-interference. To remove these static paths and amplify the changes, we take the difference of the CIR estimation by subtracting it from the same tap in the previous time snapshot. The result is shown in Fig. 6(b). The value (i.e., color) of each point represents the difference of the tap n at the time t. Therefore, we can use $P(n, t)$ to represent a differential CIR estimation and treat it as an image.

Fig. 7 shows the differential CIR estimation from various commands. The brighter areas indicate stronger movements at the corresponding distance and the time. We can make two observations. First, the brightest areas usually are located at tap 5-20, which corresponds to 1.8-7.2cm. Since the phone was placed at around 1 to 3cm from the mouth while the data was being collected, it implies that the brightest areas capture the reflection from the lip and cheek movements as we would expect. Second, different commands exhibit very different patterns due to their various facial movements. The results provide preliminary support that the differential CIR estimation *profiles facial movements* and can be used for silent speech command classification.

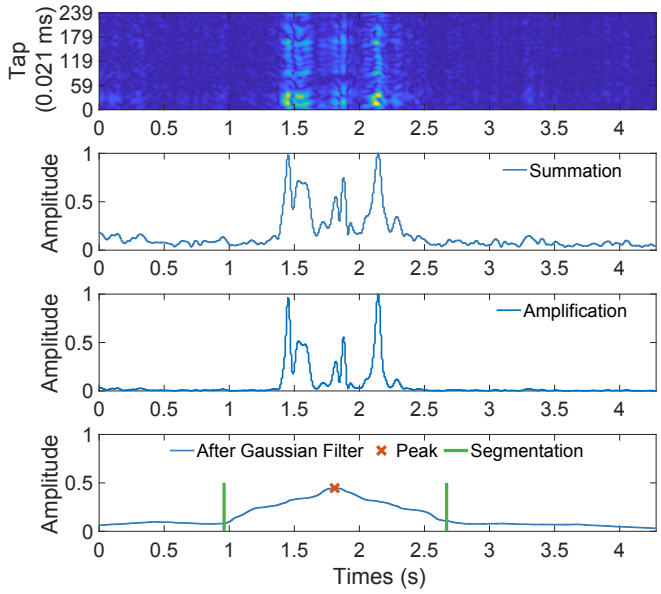


Fig. 8. Example of segmentation flow: The figures from top to bottom represent the results of original differential CIR estimation of "Search" command, after the summation across taps, after peak amplification, and after applying a Gaussian filter.

Segmentation: Differential CIR estimation $P(n,t)$ can be viewed as an n-factor time-series, which is continuously retrieved from the audio streaming captured by the microphone. The timestamp of pressing the button can be used to locate a coarse region containing a command. Then, to extract the desired command, we need to segment the time-series of the differential CIR estimation. The segmentation algorithm should satisfy three criteria: i) It should be able to cover a whole command. ii) Segments should have a fixed window length to avoid the distortion in the shape of the brighter areas. iii) The command should locate at the center of each segment to avoid misclassification caused by the location of brighter areas. As a result, we propose a segmentation scheme, as shown in Fig. 8. For time t, we compute the summation of CIR changes across taps ($\hat{P}(t) = \sum_{i=1}^{n} P(n,t)$), before normalization. After that, we amplify the difference between peaks and noise ($\hat{P}'(t) = \hat{P}(t)^2$). Then we apply the Gaussian filter to remove random noise and merge nearby peaks. Finally, for each peak, whose maximal value is located at time t_1, we segment the time-series using the time window $[t_1 - T/2, t_1 + T/2]$. T represents the size of each segment. For our experiments, the average time duration μ of commands in our dataset is 1.08s and a standard deviation σ is 0.21s. The ground truth is measured by VICON [50] (see Sec. 4.1). Therefore we set T to $T = \mu + 3\sigma = 1.71s$ to ensure that, statistically, 99.87% segments include a complete command. The accuracy change caused by varying length of T is evaluated in Sec. 4.2.4.

With the proposed segmentation scheme, we get a 2D array of differential CIR estimation $P(n,t)$ where $t = t_1 - T/2 \ldots t_1 + T/2, n = 0 \ldots 239$. After that, the estimation $P(n,t)$ is resized to a 224×224 image to reduce the computational load in the training phase, which represents a 2D motion profile of a silent speech command. Hereafter, we show how to train a model and use the model to identify the command of a given 2D motion profile.

3.4 CNN Classifier

To classify the 2D motion profiles of various commands, we adopt CNN [43] as a classifier since CNN is widely used for image classification and shows promising accuracy. Our CNN model is designed to be in accord with three principles: i) The model needs to achieve a high accuracy to offer a high-quality user experience. ii) A new user should be able to use the system with little effort, i.e., only needing to collect a little data before using the model. iii) The model can be further improved upon while the new user uses the model over time.

Therefore, the training process for our CNN model includes three steps. In the **cold start step**, we collect a large amount of data from various users to train a base network that can extract representative features to identify commands from these users accurately. In the **warm start step**, a transfer learning technique is adopted to adapt the model from the cold start phase for new users. The transfer learning technique only requires the new user to collect a small amount of data. In the **online training step**, a semi-supervised learning scheme is adopted to utilize the unlabeled motion profiles, which are collected when the user uses the system over time in order to continuously update the model and improve its classification accuracy.

3.4.1 STEP 1. Cold Start: Train a Feature Extractor. In the cold start step, we collect a large amount of data from various users to train a feature extractor which can extract representative features to identify commands from these users accurately. Assume that we have a large amount of labeled data $\mathcal{D}_s$. We call this source domain, whose input space is denoted as $\mathcal{X}^s$ and label space is denoted as $\mathcal{Y}^s$ (i.e., the commands). The goal of the cold start step is to tune a CNN network on $\mathcal{D}_s$. To achieve this, we adopt a pre-trained *ResNet18* [19] as our base network. By conducting experiments on $\mathcal{D}_s$, we found that *ResNet18* converges quickly and requires less time compared with other frequently used CNN networks, such as *VGGNet*, *InceptionNet*, *DenseNet*, and *SqueezeNet* (see Sec. 4.2.1). We denote the mapping function of our CNN network as f, which is composed of two functions, i.e., $f = h \circ g$. Here $g : \mathcal{X} \rightarrow \mathcal{Z}$, which is called a feature extractor, represents an inference from input space $\mathcal{X}$ to feature space $\mathcal{Z}$, while $h : \mathcal{Z} \rightarrow \mathcal{Y}$, which is called fully connected layers, represents an inference from feature space to output space $\mathcal{Y}$. With this notation, we represent the CNN network trained in source domain as $f_s = h_s \circ g_s$, where g_s is composed of 4 *ResNet* layers and h_s contains a single fully connected layer. Fig. 9(a) shows the process of training such a network, which is a typical training process of CNN networks. The total loss in one epoch is computed by the average loss of 2D motion profiles with a standard *Cross-Entropy Loss* ℓ:

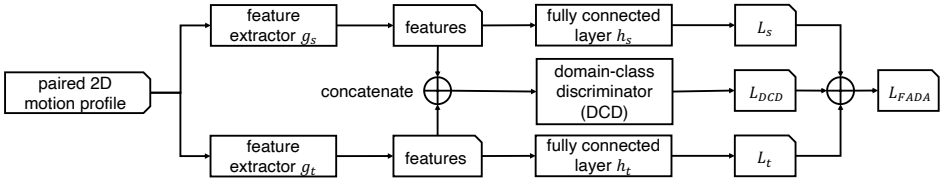
$$L_s = E[\ell(f_s(\mathcal{X}^s), \mathcal{Y}^s)] \tag{1}$$

Data Augmentation: One thing to note is that, instead of feeding the 2D motion profiles to the CNN model directly, we first apply a data augmentation scheme to improve the generalization ability of the model and alleviate the overfitting problem. Specifically, given a 2D motion profile, the data augmentation scheme randomly crops the 2D motion profile and flips the cropped portion horizontally. The size of the cropped portion is randomly selected, which ranges from 8%-100% [7] of the original 2D motion profile size. The scheme increases the dataset size by 200 times.

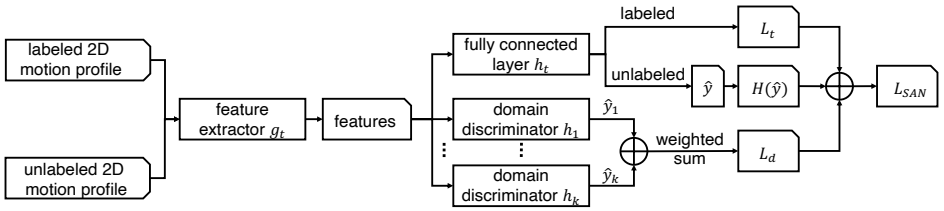
3.4.2 STEP 2. Warm Start: Transfer to a New User. Although f_s performs well in $\mathcal{D}_s$, it may not be well generalized for a new user, because of the distinct characteristics in the 2D motion profiles for different users.

Assume that we have a small amount of labeled data $\mathcal{D}_t$. We call this target domain, whose input space is denoted as $\mathcal{X}^t$ and label space is denoted as $\mathcal{Y}^t$. The goal of the warm start step is to transfer the knowledge learned from $\mathcal{D}_s$ to $\mathcal{D}_t$. One intuitive method to tackle this problem is to pre-train a model in $\mathcal{D}_s$ and then re-train it in $\mathcal{D}_t$. This process is known as fine-tuning. Although fine-tuning is wildly used for its simplicity, when there is only a small amount of labeled data in $\mathcal{D}_t$ and a deep neural network like *ResNet18* is used, the performance may not be satisfactory due to the underfitting problem in the target domain. In our case, f_s contains the knowledge learned from $\mathcal{D}_s$ and has 11.2 million trainable parameters. Since $\mathcal{D}_t$ contains only a small amount of labeled

(a) **Cold Start:** train a base network.



(b) **Warm Start:** transfer to a new user.



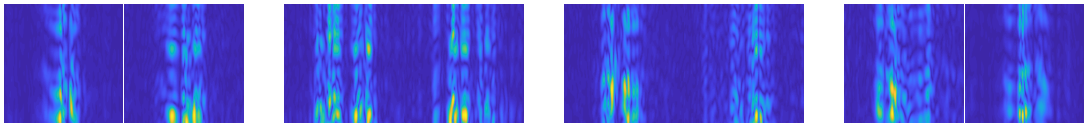
(c) **Online Training:** train with unlabeled data.

Fig. 9. The training scheme for Endophasia is demonstrated by showing the way to compute the loss at each step. After obtaining the loss in each epoch, the model can be updated by back-propagating the loss. The training scheme is composed of three main steps: **cold start**, **warm start** and **online training**. (a) **cold start:** A typical training process for a CNN network. Train a base network ($f_s = h_s \circ g_s$) with a large number of 2D motion profiles from source domain. (b) **warm start:** Update f_s and f_t by freezing the parameters in the DCD so that the DCD can no longer distinguish the paired 2D motion profiles. By applying the FADA scheme, the knowledge learned in f_s can be transferred to f_t with considerably less data from a new user. (c) **online training:** The features of each 2D motion profile will be fed into h_t, while only unlabeled 2D motion profiles are fed into domain discriminators ($h_1, ..., h_k$). During the daily use, many unlabeled 2D motion profiles will accumulate in the data pool. Endophasia utilizes the unlabeled 2D motion profiles to continuously improve the performance of the personalized model f_t, while introducing no extra burden to the user.

data, f_s can not be fully tuned and the transferable knowledge will not be sufficiently utilized (as we show in Sec. 4.3.3).

To address the imbalance issue, we adopted the Few-Shot Adversarial Domain Adaptation (FADA) scheme [33]. FADA divides $\mathcal{D}_s$ and $\mathcal{D}_t$ into 4 groups, denoted as $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3 \cup \mathcal{G}_4$. $\mathcal{G}_1$ is composed of pairs of 2D motion profiles of the same command, which are randomly selected from $\mathcal{D}_s$. $\mathcal{G}_2$ is composed of pairs of 2D motion profiles of the same command but, in each pair, one is from $\mathcal{D}_s$ while the other one is from $\mathcal{D}_t$. $\mathcal{G}_3$ contains pairs of 2D motion profiles of different commands randomly selected from $\mathcal{D}_s$. $\mathcal{G}_4$ contains pairs of 2D motion profiles of different commands where one is from $\mathcal{D}_s$ and the other one is from $\mathcal{D}_t$. The 4 groups contain the same amount of sample pairs so that they are balanced. Moreover, the dataset is enlarged by pairing 2D motion profiles from $\mathcal{D}_s$ and $\mathcal{D}_t$ because each sample pair in $\mathcal{G}$ is different. Fig. 10 presents examples of paired 2D motion profiles.

The training process contains two iterative steps. First, we need learn a domain-class discriminator (DCD) [33] to distinguish the 2D motion profile pairs from $\mathcal{G}$. The DCD is composed of two fully connected layers with



(a) $\mathcal{G}_1$ (b) $\mathcal{G}_2$ (c) $\mathcal{G}_3$ (d) $\mathcal{G}_4$

Fig. 10. Example of paired 2D motion profile in Few-Shot Adversarial Domain Adaptation (FADA) scheme. (a) In $\mathcal{G}_1$, the command of the two profiles is "Pause" from the source domain. (b) In $\mathcal{G}_2$, the command of the two profiles is "ScreenShot". The one on the left is from $\mathcal{D}_s$ and the one on the right is from $\mathcal{D}_t$. (c) In $\mathcal{G}_3$, the commands of the left and right profiles are "Copy" and "Skip", respectively. Both are from $\mathcal{D}_s$. (d) In $\mathcal{G}_4$, the command of the left profile is "Undo" from $\mathcal{D}_s$ and the command of the right profile is "Play" from $\mathcal{D}_t$.

ReLU and *DropOut*. *ReLU* is used to ensure the non-linear mapping from a feature space to a domain output space [34] and *DropOut* is used to prevent overfitting [44]. To initialize the DCD, first we make a copy of f_s and call it $f_t = h_t \circ g_t$. g_s and g_t deals with the first and the second 2D motion profiles of each sample from $\mathcal{G}$, respectively. Then, outputs of g_s and g_t are concatenated and fed into the DCD. In each epoch, we compute the standard *Cross-Entropy Loss* and backpropagate errors to update the DCD while keeping the parameters in g_s and g_t frozen. Second, we update f_s and f_t while keeping DCD unchanged by minimizing the loss function defined below:

$$L_{FADA} = L_s + L_t - \lambda E[\mathcal{Y}_{\mathcal{G}_1} log(DCD \circ g(\mathcal{G}_2)) - \mathcal{Y}_{\mathcal{G}_3} log(DCD \circ g(\mathcal{G}_4))] \tag{2}$$

where λ strikes a balance between classification and confusion, $\mathcal{Y}_{\mathcal{G}_x}$ represents the label of a sample from group $\mathcal{G}_x$ ($x = 1, ..., 4$). $DCD \circ g(\cdot)$ represents the function of the composition of DCD and two feature extractors. L_t is the loss in $\mathcal{D}_t$, computed in the same manner with L_s. $L_s + L_t$ is designed to maintain high classification accuracy. The term $-\lambda E[y_{\mathcal{G}_1} log(DCD \circ g(\mathcal{G}_2)) - y_{\mathcal{G}_3} log(DCD \circ g(\mathcal{G}_4))]$ is denoted as L_{DCD} in Fig. 9(b) and is used to confuse the DCD such that it can no longer distinguish between $\mathcal{G}_1$ and $\mathcal{G}_2$ as well as between $\mathcal{G}_3$ and $\mathcal{G}_4$. Lastly, we repeat the above two steps to iteratively update f_s, f_t and the DCD until f_t converges in our dataset. As a result, the feature extractor g_t can be updated to extract features that are only sensitive to commands but not domains. The feature space $\mathcal{Z}_t$ will have the same distribution with $\mathcal{Z}_s$ such that the DCD can not distinguish samples from $\mathcal{G}$ anymore. Therefore, the knowledge learned in $\mathcal{D}_s$ by f_s will be fully utilized when training f_t in $\mathcal{D}_t$.

3.4.3 STEP 3. Online Training: Improving the Model with Unlabeled Data. Since we have a customized model f_t for a new user trained with few training 2D motion profiles, we are interested in making the system more robust and accurate with daily use. To achieve that, we adopt a semi-supervised learning scheme to utilize the unlabeled 2D motion profiles, which are collected when the user uses the system over time. We modified a transfer learning scheme of domain adaptation with Selective Adversarial Network (SAN) proposed by Cao *et al.* [7]. The network of the scheme is shown in Fig. 9(c). We assign a domain discriminator to each class of the commands, which are denoted as h_k, where $k = 1, 2, ..., |\mathcal{Y}|$. The scheme is modified as a semi-supervised network in our system. The intuition behind the modification is that we use the labeled data $\mathcal{X}^t$ (collected in the warm start step) to update f_t and ensure that the labeled data can be accurately distinguished. In the meanwhile, we use the unlabeled data $\mathcal{X}^r$ (collected via the daily use of Endophasia) to update each domain discriminator h_k as well as f_t. We call the union of $\mathcal{X}^t$ and $\mathcal{X}^r$ as $\mathcal{X}^u$. As a result, f_t is iteratively improved and can better map the labeled and unlabeled data to a feature space where they have the same distributions.

Specifically, the loss function of the entire network is designed below:

$$L_{SAN} = L_t + E[H(f_t(\mathcal{X}^r))] - \sum_{k=1}^{|\mathcal{Y}|} E[f_t(\mathcal{X}^r)]E[f_t(\mathcal{X}^u)\ell(h_k \circ g_t(\mathcal{X}^u), d)] \quad (3)$$

where d represents the domain of each 2D motion profile. In our case, it denotes the 2D motion profile is from $\mathcal{X}^t$ or $\mathcal{X}^r$. L_t is used to maintain the high classification accuracy of f_t in $\mathcal{D}_t$. The term $-\sum_{k=1}^{|\mathcal{Y}|} E[f_t(\mathcal{X}^r)]E[f_t(\mathcal{X}^u)\ell(h_k \circ g_t(\mathcal{X}^u), d)]$ is to compute the sum of weighted domain loss, which aims at confusing domain discriminators. Note that this loss is weighted in proportional to $\hat{y} = f_t(\mathcal{X}^r)$, which represents the probability of the predicted labels for the unlabeled data. Moreover, to speed up this process, a conditional-entropy loss is added: $H(\hat{y}) = -\hat{y}^T log\hat{y}$, which narrows down the probability density of $\hat{y}$ and boost the performance of the domain adversarial mechanism.

Note that the aforementioned semi-supervised scheme does not assume the number of unlabeled data of various labels is balanced. It implies that a user does not need to 'speak' all commands before the online learning scheme can be applied. Therefore, Endophasia can optimize the accuracy of the frequently used commands and have little impact on those unused or less-used commands.

4 EVALUATION

4.1 Experiment Configuration

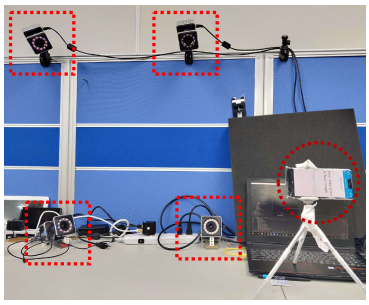
System Implementation: We developed an Android Application to generate inaudible signals and collect reflected signals. The recorded signals are sent to a server for processing and prediction in real-time.

Data Collection: We spent 6 weeks to collect 10100 silent command samples (includes 17 hours of audio signals) from 14 participants. The participants include 12 males and 2 females, ranging in age from 22 to 26 years old. All of the participants are well-educated, university students, capable of speaking English. Before the experiment, we explained the goal of Endophasia and presented the way to issue silent speech commands, then gave participants 5 minutes to get familiar with the system. After that, the participants were instructed to sit in front of a desk where a mobile phone was fixed by a holder; then, they silently issue commands at 3cm away from the phone. They were required to issue 20 silent commands and repeat 30 times. We used VICON vero [50] to localize the phone and track lip movements, as shown in Fig. 11(a). VICON is a camera-based tracking tool that can track the markers (5mm diameter spheres with special coating) with a 1mm error. The trajectories collected from VICON are used as the ground truth to compute the distances and angles between participants and the phone and to determine the start and end time when issuing commands. Considering that the participants would likely become more familiar with the mechanism of the system during the experiment, a Latin Square [4] was used to determine the collection order of each command. If the commands were collected in the same order, the first few commands might have a relatively bad classification accuracy. With the Latin Square, this error would be distributed fairly into each command. Moreover, the participant could have a short pause to re-adjust his or her posture between each command to simulate the real application scenarios. Each participant spent around four and a half hours to finish the experiments.

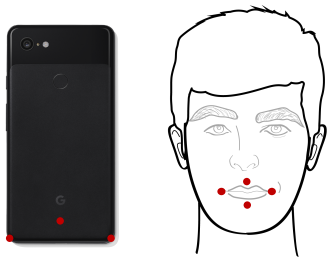
4.2 Micro-benchmark

In this section, we evaluate the impact of system parameters and motivate the values selected in the following evaluation.

4.2.1 Impact of Different CNN networks. The structure of the CNN network has a significant impact on performance. Thus, we fine-tune 7 pre-trained CNN networks using our dataset and the accuracy, rate of convergence, and training time are shown in Fig. 12. Each network is tuned for 300 epochs to ensure convergence. We can see that the *ResNet* family, *InceptionNet*, and *DenseNet* outperform *VGGNet* and *SqueezeNet* in terms of the accuracy

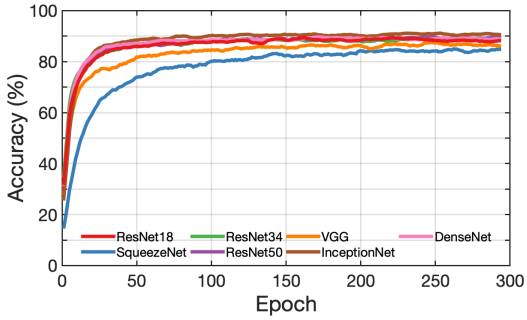


(a) Data collection configuration.

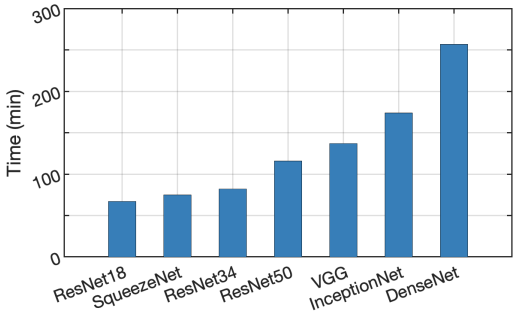


(b) Markers.

Fig. 11. (a) The mobile phone is held by a holder and 4 VICON cameras are used to track the location of markers. (b) 3 markers are attached to the back of the phone, which are used to determine the plane of the phone. 4 markers are attached around the user's mouth.



(a) Accuracy and convergence rate.



(b) Running time.

Fig. 12. The performance of 7 CNN networks.

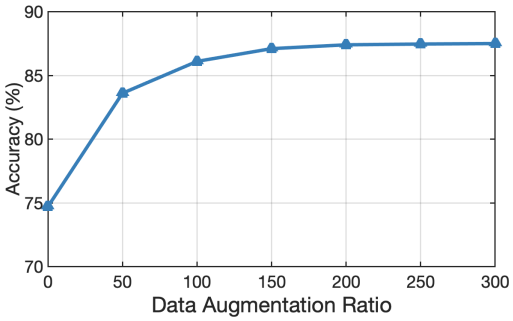


Fig. 13. The accuracy under various augmentation ratios. The x-axis represents the ratio of the training data size w/ data augmentation to that w/o augmentation.

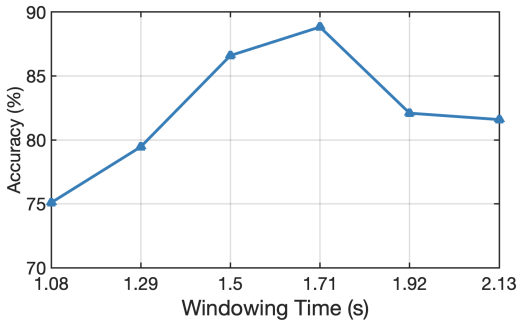
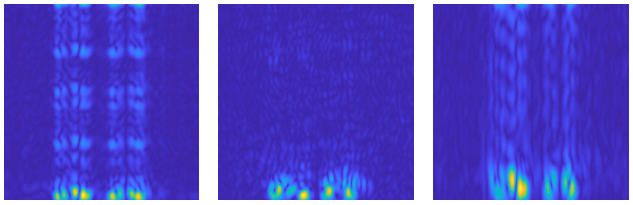


Fig. 14. The impact of windowing time T in the segmentation algorithm on recognition accuracy.



(a) GSM. (b) ZC. (c) Barker.

Fig. 15. The 2D motion profiles generated by different training sequences.

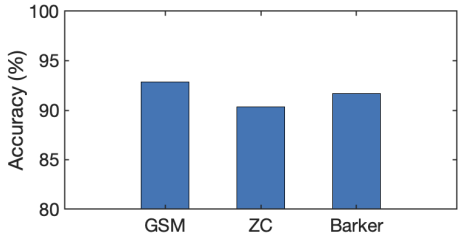


Fig. 16. The impact of different training sequences on recognition accuracy.

and the rate of convergence. Although the accuracy of *InceptionNet* is slightly better (0.9% better than *ResNet18*), the training time of *ResNet18* is 61.27% less than that of *InceptionNet*. Therefore, we choose a pre-trained *ResNet18* as our CNN network.

4.2.2 Impact of Data Augmentation. We evaluate the impact of the data augmentation scheme. As described in Sec. 3.4.1, the data augmentation scheme randomly crops 2D motion profiles to augment the number of inputs. We vary the number of augmented motion profiles and the results are shown in Fig. 13. The x-axis represents the ratio by which the training dataset size is increased. We can see that the accuracy increases by 13.3% when the number of augmented motion profiles increases from 0 to 200 times. After that, increasing the augmented profiles does not improve the accuracy markedly. The results suggest that the data augmentation scheme can effectively increase the generalization ability of our system and alleviate the overfitting problem.

4.2.3 Impact of Training Sequences. To evaluate the impact of different training sequences, we collected additional 2D motion profiles generated by a 13-bit Barker sequence and a 30-bit ZC sequence from one participant. Similarly, this participant collected 30 samples for each command by using both the Barker sequence and the ZC sequence. Example 2D motion profiles of the command "ScreenShot" generated by these three training sequences are shown in Fig. 15. The brighter areas of GSM and Barker seem to appear in all taps, while those of ZC only appear in lower taps. This phenomenon is due to the fact that the autocorrelation result of the ZC sequence has a smaller side lobe gain compared with that of the GSM sequence and the Barker sequence. However, the side lobes may bring more features for the 2D motion profiles, which would likely contribute to the classification accuracy. Our experiments demonstrate that these three sequences result in a similar performance in Endophasia, but the accuracy for the GSM sequence is slightly greater than that of the ZC sequence and the Barker sequence, as shown in Fig. 16. The result implies that no significant difference appears among these three training sequences. Thus, we choose the GSM sequence as our training sequence mainly because of the wide adoption in single carrier communication and the well-known properties for synchronization and channel estimation.

4.2.4 Impact of Windowing Time. The segmentation algorithm is proposed in Sec. 3.3. To evaluate the impact of the fixed windowing time, we segment the 2D motion profiles with varying time windows in the range from 1.08s to 2.13s with a resolution of the standard deviation, which is 0.21s. Then we use a pre-trained *ResNet18* CNN network with a data augmentation ratio of 200 to test the accuracy. As shown in Fig. 14, the accuracy reaches the maximum 88.83% when the windowing time equals 1.71s and drops to 82.09% beyond this value. It implies that if the windowing time is extended, additional noise is likely to be included, resulting in the squeeze of the brighter areas of the commands. However, if the windowing time is short, some segments cannot cut off the full commands. Therefore, we set the windowing time T to 1.71s.

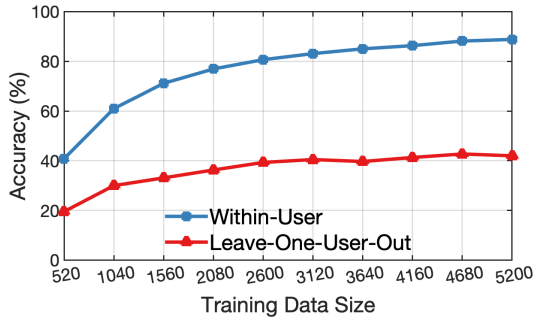


Fig. 17. The accuracy of within-user test and leave-one-user-out test in the cases of various training data sizes.

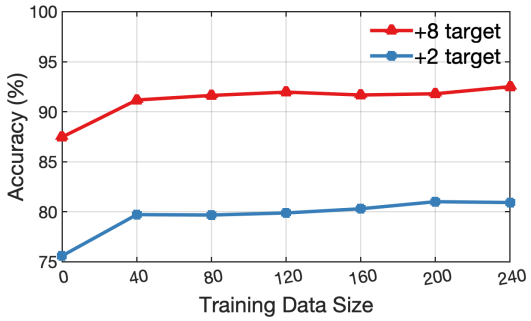


Fig. 18. The performance of the online learning scheme. The x-axis represents the number of unlabeled samples.

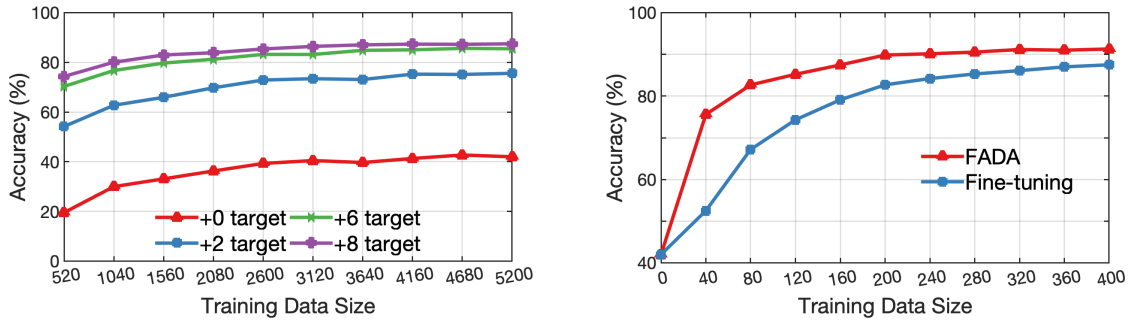
4.3 System Performance

In this section, we evaluate the overall system performance and detail the performance of each component.

4.3.1 Within-User Performance. We fine-tune a pre-trained *ResNet18* in $\mathcal{D}_s$ with 10-fold cross validation and incrementally increase the training data size from 560 samples (i.e., 2 samples × 20 commands × 14 users) to 5600 samples (i.e., 20 samples × 20 commands × 14 users), while the remaining 2D motion profiles are used for evaluation. Note that the following sections have the same data separation scheme. The results are highlighted by the blue line seen in Fig. 17. The accuracy in the validation set increases from 40.81% to 88.83% when the training data size equals 5200. The results suggest that the dataset is highly separable, which further prove that the reflected acoustic signals indeed carry the information of users' facial movements and Endophasia can effectively extract this information.

4.3.2 Leave-One-User-Out Performance. We evaluate the leave-one-user-out performance to test the accuracy of applying the CNN network to a new user directly. We train a CNN model by using the 2D motion profiles of 13 users and test the model with the 2D motion profiles of the last user. This process is repeated 14 times and the average accuracy is reported. The results are highlighted by the red curve shown in Fig. 17. The accuracy increases from 19.54% to 39.32% when the training size increases from 520 to 2600. After that, the accuracy fluctuates at around 40% and does not improve while further increasing the training size. It implies that when different users issue the same command, their facial movements indeed share some similarities because the accuracy for a new user achieves 40% for 20 commands. However, the accuracy converging at 40% suggests that the distinct features of different users become the major bottleneck for improving the performance. Different users have unique features, even when they issue the same command [28, 48].

4.3.3 Warm Start Performance. The effectiveness of the FADA scheme is evaluated. Similar to the process described in Sec. 4.3.2, but instead of directly testing the model with 2D motion profiles from a new user, we apply FADA to transfer the knowledge obtained from the source domain (i.e., data from the 13 users) to the target domain (i.e., data from the remaining user). Also, the average accuracy after repeating 14 times is reported. The results are shown in Fig. 19(a). The x-axis represents the training data size in the source domain. "+0 target" means no extra training data from the target domain is added, which is the same as the data plotted by the red curve in Fig. 17. "+2", "+6" and "+8 target" represent that 2, 6 and 8 samples, respectively, for each command from the target domain are added for the training process of the FADA scheme.



(a) FADA accuracy.

(b) FADA v.s. Fine-Tuning.

Fig. 19. (a) The performance of the FADA scheme while varying the training data size from the source domain (520-5200) and the target domain (+0, +2, +6, and +8). (b) Compare the performance of FADA and fine-tuning by incrementally adding the training data from target domain (0 – 400). +0, +2, +6, and +8 target are equivalent to add 0, 40, 120 and 160 training data size in target domain.

We can see that there is a significant improvement even if we collect 2 more samples per command from a new user. The accuracy is improved by 33.61% and converges at about 75.6%. When we collect 8 more samples per command from the new user, the accuracy converges at 87.47%, which is only 1.36% lower than that of the source domain. The results suggest that FADA can effectively reduce the effort for a new user to use Endophasia.

The performance of FADA and fine-tuning is also compared and the results are shown in Fig. 19(b), where the x-axis represents the training data size in target domain. We first train a CNN model by using all the 2D motion profiles from the source domain. Then we apply FADA and fine-tuning by incrementally increasing 2D motion profiles from the target domain, from 0 to 400. We can see that FADA converges faster and outperforms fine-tuning by 3.8%-23.1% in cases of various training data sizes. FADA almost converges when the training data size is larger than 200, while the accuracy of fine-tuning seems to keep increasing even beyond the 400 training data size. Note that FADA outdistances fine-tuning when data size is extremely small, e.g., 40 and 80, due to the benefits of implementing the idea of re-grouping the 2D motion profiles into pairs. On the one hand, it solves the imbalance of samples between the source domain and the target domain. On the other hand, it increases the size of training samples, because each 2D motion profile pair in the re-grouped dataset is distinct. The improvement brought about by this benefit is more significant when the data size in the target domain is very small. It implies that Endophasia significantly reduces the requirements of data size for a new user by adopting a well-designed transfer learning scheme.

4.3.4 Online Learning Scheme Performance. After applying the FADA scheme, a customized model for a new user can be initialized with little effort. To improve the performance of Endophasia with daily use, we adopt an online learning scheme. The models obtained in Sec. 4.3.3 denoted by "+2 target" and "+8 target" with the source domain training data size of 5200 are used to evaluate the performance of the online learning scheme. In this case, 8 samples for each command in the target domain are used for the training of the FADA scheme. Therefore, the remaining samples are considered as unlabeled 2D motion profiles. We vary the number of unlabeled 2D motion profiles and apply the proposed online learning scheme to update the models iteratively. This process is repeated 14 times and we report the average accuracy. Fig. 18 shows the results. The red curve plots data of the average accuracy for the "+8 target" models. When we increase the unlabeled 2D motion profiles from 0 to 240,

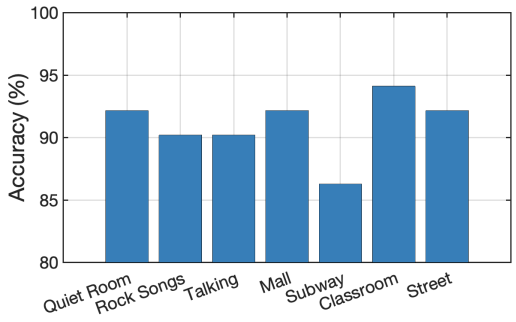


Fig. 20. The accuracy within various environments.

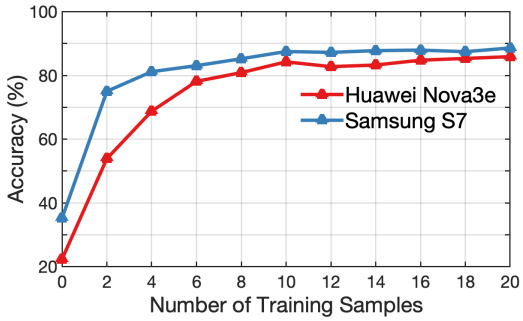


Fig. 21. The performance of transferring a customized model to different phones. The x-axis represents the number of training samples per command.

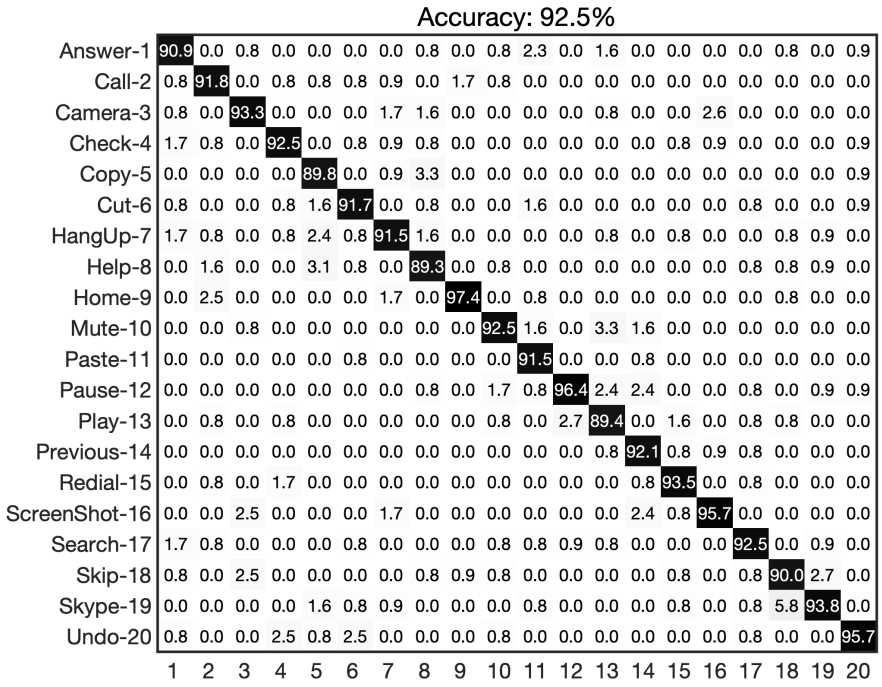


Fig. 22. The confusion matrix of the model trained with 8 labeled samples per command and 240 unlabeled data.

the accuracy increases by 4.39%. Similarly, for the "+2 target" models, the online learning scheme improves the accuracy by 5.03% with 240 unlabeled 2D motion profiles.

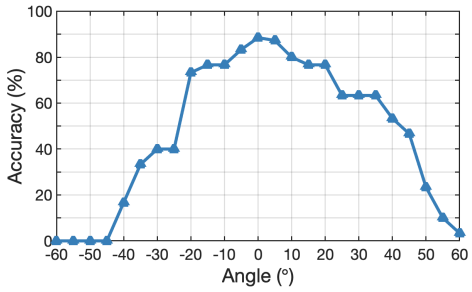


Fig. 23. When the classifier is trained at a fixed angle, we evaluate the accuracy at varying angles between the phone and the user in the test dataset.

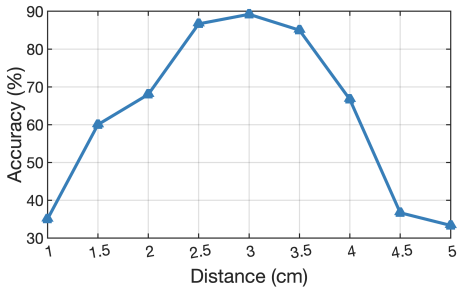


Fig. 24. When the classifier is trained at a fixed distance, we evaluate the accuracy while varying distances between the phone and the user in the test dataset.

Fig. 22 shows the confusion matrix of "+8 target" models, which are further improved by 240 unlabeled 2D motion profiles. We present the average accuracy among 14 users. The average accuracy is 92.5%. We can observe that, among all the commands, "Help" is the most difficult to identify, but we still achieve 89.3% accuracy. In addition, "Skype" is likely to be confused with "Skip" (5.8%) because these two commands have a similar pronunciation, which results in similar facial movements.

4.4 Environmental Dynamics

In this section, we evaluate the impact of various environmental factors that show the robustness of our system.

4.4.1 Impact of Noise. We evaluate the robustness of Endophasia in noisy environments. We first conducted experiments in controlled environments where users issue silent commands in i) a quiet room, ii) a room with rock songs being played, and iii) a room with 5 people talking loudly. We then test Endophasia in uncontrolled and noisy environments, including i) a classroom while a lecturer was teaching, ii) a shopping mall, iii) a subway, and iv) a street. Fig. 20 shows the results. We can see that the accuracy remains similar in all environments ranging from 86.3% to 94.1%. The results are expected because Endophasia operates at 17-23KHz band while music, the human voice, and other audible sounds are usually below 10KHz [22]. The only exception is that the accuracy in the subway is slightly lower (drops by 5.6%). This is likely due to the high frequency noise caused by the friction between the train cars and the rails.

4.4.2 Impact of Angle. We evaluate the impact of the angles between the phone and the user's mouth. In the training data, we fix the angle to 0°. While in the testing data, we vary the angle from −60° to 60°. The results are shown in Fig. 23. We can see that Endophasia's performance vis-à-vis the angles. When the angle offset is within 20°, the accuracy remains above 73.3%.

4.4.3 Impact of Distance. We then evaluate the impact of the distances between the phone and the user's mouth. In the training data, we fix the distance to 3cm. While in the testing data, we vary the distance from 1cm to 5cm. The results are shown in Fig. 24. We can see that when the distances in the training data and the testing data are close (e.g., 2.5cm-3.5cm), the accuracy remains similar. However, when the distance offset becomes greater, the accuracy drops significantly. When the distance offset is 1cm, the accuracy drops by 21.2%-22.53%.

Although the distance offset has a significant impact on the accuracy, we observed from extensive experiments that users can quickly adapt and adjust the phone's position. Specifically, we gave users an Android phone with Endophasia installed. After issuing silent speech commands, Endophasia gives feedback with the detected command in real-time. If users found the identified command is incorrect, they adjusted the phone position

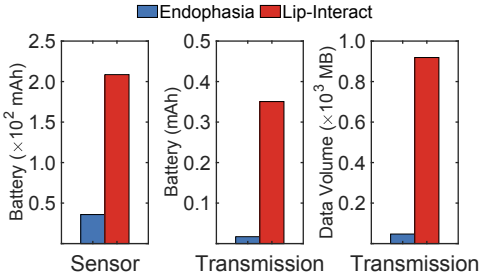


Fig. 25. Battery consumption and transferring data volume comparison between Endophasia and camera-based solution, e.g., Lip-Interact.

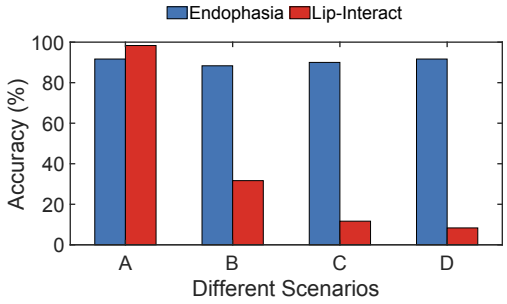


Fig. 26. The impact of ambient light intensity on recognition accuracy in 4 different scenarios.

the next time when they issued a command. After a 5-minute practice session, users can hold the phone at the designated position every time.

4.4.4 Impact of Phones. To evaluate the impact of different devices, we train a model with the 2D motion profiles collected by Pixel 3L and then test the model with another 2 different mobile phones (Huawei Nova3e and Samsung S7). Because the location and type of mics and speakers on different phones vary, a model trained for one phone may not be directly applicable to the other phone. Fortunately, the signals received by different phones still share some similarities because they are produced by the same user whose face motion profiles remain unchanged. Therefore, we can also apply FADA in the same manner to help the user to transfer the model from one phone to another quickly. The results are shown in Fig. 21. It suggests that, for a registered user, he/she can transfer the model trained using one phone to another by collecting 6 labeled samples per command to reach an accuracy of 80%.

4.5 Compare with Camera-based Solution

We implement a camera-based solution with our command set, as described in Lip-Interact [46], achieving an accuracy of 97.47%. We believe that our implementation is equivalent to Lip-Interact. Battery consumption and accuracy in varying light intensity conditions are compared.

4.5.1 Battery Consumption. The data processing and inference are finished via a remote server for both Endophasia and Lip-Interact. Therefore we take the battery consumed by sensors (including CPU) and network transmission into account and compare the data transfer volume of both methods. The experiment was conducted on a Google Pixel 3L. We measure the battery consumption via Android Debug Bridge (ADB) wireless debugging mode over WI-FI connection. We issue 100 commands within 10 minutes to measure the total battery consumed by Endophasia and Lip-Interact. The results are shown in Fig. 25. For Endophasia, 35.82mAh battery is consumed by mic and speaker. In the meantime, 47.15MB data is transferred. However, for Lip-Interact, a total of 208.60mAh power is consumed by the camera, which is 5.82 times greater than the battery consumed by Endophasia. Correspondingly, a total of 918.70MB of data is transferred by Lip-Interact, which is 19.48 times greater than that of Endophasia.

4.5.2 Performance in Varying Light Conditions. One of the main advantages of acoustic methods is that it is independent of light conditions, while a camera-based solution may be sensitive to the ambient light intensity. Therefore, a camera-based solution may not maintain consistent performance in varying light conditions. To

evaluate this, we measure the recognition accuracy of Endophasia and Lip-Interact in 4 scenarios with decreasing ambient light intensities, as shown in Fig. 26. Scenario A is during a lecture in a classroom with adequate supplementary lighting. Scenario B is watching a concert with relatively weak supplementary lighting from the stage. Scenario C represents driving at night with little supplementary lighting along the street. Scenario D is in an indoor environment at night with all lights off, except for the light from the phone's screen. We issue the 20 commands 3 times per each and compute the average accuracy. Fig. 26 shows that the performance of the camera-based solution decreases quickly with a decrease in light intensity, while the number of Endophasia remains unaffected and shows promising results under all conditions.

5 DISCUSSION

Based on the evaluation results reported in the previous section, we have demonstrated the effectiveness of the Endophasia system, which can correctly identify users' silent speech commands in a contact-free way with their mobile phones. Also, the command recognition algorithm of Endophasia can be used to complement existing phone-based solutions or be integrated with wearables in commonly-used form factors to enable a wide spectrum of applications. In this section, we outline potential application scenarios of Endophasia to enrich existing interactions in the following directions, including (1) complementing with existing solutions and (2) integrating with commonly-used wearables.

5.1 Complementing with Existing Solutions Recognizing Silent Speech Commands

The acoustic-imaging-based solution proposed in this paper transmits/receives acoustic signals in the frequency inaudible to humans to enable silent speech command recognition, which facilitates users to issue the commands without worrying about eavesdropping. This frequency band used by the speaker (or the microphone) to transmit (or receive) is idle in voice applications of current COTS devices, e.g., smartphones. So, Endophasia offers good extensibility to be implemented on these devices without affecting normal functionalities of the devices. Moreover, based on the measurement results of power consumption in Section 4.5.1, Endophasia consumed an acceptable amount of battery power when recognizing each command, which is 5.82 times less than that consumed by a camera-based solution, i.e., Lip-Interact. These characteristics facilitate the Endophasia system to complement existing silent speech recognition solutions in the following two ways.

First, we can introduce Endophasia to complement current solutions only when they could not perform well under some challenging scenarios. Given the fact that each individual sensing algorithm cannot be 100% accurate, existing solutions might fail to correctly differentiate commands when the sensors were confused by some environmental noises, such as poor light conditions for camera sensors. When detecting the environmental changes, the system can switch to the Endophasia recognition function so that users can still interact with the system in an accurate and privacy-preserving way through silent speech commands. Moreover, the existing solutions consumed more battery power, comparing with the Endophasia approach. When the power of the device is going to run out, the users who have privacy concerns can save power by switching to use the Endophasia recognition function. Second, Endophasia can be integrated with existing solutions to jointly generate inferences to achieve better recognition results. Since the Endophasia recognition function consumed a small amount of power, it could be used concurrently with a silent speech recognition algorithm to detect some keywords which can be fed into the inference pipeline of that silent speech recognition solution for further improve the performance of the whole system. In the future, we would explore a better complementary way to improve the interaction of existing silent speech command recognition with the help of Endophasia.

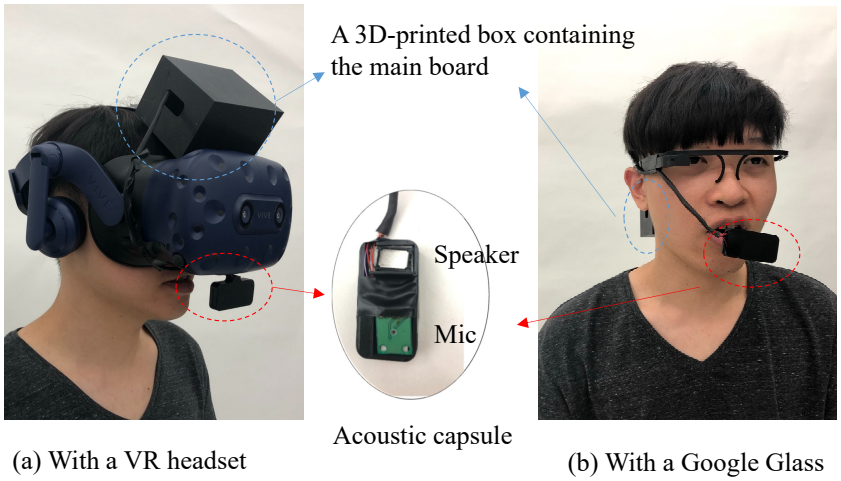


Fig. 27. The prototyped headset integrated with (a) a VR headset and (b) a Google glass. The main board placed (a) on the VR headset or (b) behind the person's neck is in a 3D-printed box. The acoustic capsule contains both a speaker and a microphone to transmit the GMS signals and receive the acoustic signals reflected from the user's face.

5.2 Integrating with Commonly-used Wearables

In this paper, we leveraged the prevalent mobile phones as an exemplar vehicle to realize the interaction of issuing contact-free silent commands. However, people might not be able to use mobile phones to detect the commands in some everyday application scenarios, e.g., when they do not carry their phones, or their hands are occupied. Under these application scenarios, they might still wear devices, such as headset, glasses, and watches, used in their daily lives. For example, when people drive their cars, they might wear their headset so that they can answer the phone and talk on the headset. Moreover, the low level of battery power consumption, as described in Section 4.5.1, made Endophasia be easily integrated into power-limited wearable devices. Therefore, enabling this contact-free silent speech command recognition on everyday-use wearables would offer an alternative interface.

To demonstrate the feasibility of recognizing silent speech commands with modified everyday-use wearables, we took the headset form factor which could be easily integrated into head-mounted devices used in AR/VR scenarios (Figure 27(a)) and Google Glasses (Figure 27(b)) as examples. The goal of this prototype is to explore the feasibility of enabling the silent speech command recognition in other everyday-use wearables. To prototype an Endophasia-enabled headset as shown in Figure 27, we added in an extra low-cost speaker [12] into the acoustic capsule, containing a microphone [13], positioned in a place near the user's mouth. To control the microphone and speaker in the acoustic capsule near the user's mouth, a main board [49] equipped with an STM32 microcontroller is placed in the 3D-printed box. The headset can emit the GSM acoustic signals through the speaker after users initiate the command recognition function. When users want to issue a silent speech command, users wear the device as the way they usually did. After issuing a command, the main board [49] samples the acoustic signals sensed by the microphone with a DFSDM peripheral [45] and transmitted the sampled acoustic signals back to a Notebook through a USB wire link so that the Notebook can analyze the received acoustic signals to generate the corresponding acoustic images for recognizing the silent speech commands dropped by users. We tested the Endophasia-enabled headset by asking a graduate student to issue 2 silent speech commands. The student repeatedly issued each command for five times. The resulting acoustic images corresponding to those commands issued by the student have similar patterns, which can be easily be recognized by the Endophasia algorithm as

described in Section 3. In the future, we plan to redesign the circuit board to further shrink the size of the device, e.g., the main board. Also, we plan to use the improved version of the headset device to collect more training and testing samples and conduct a field study to verify the performance of this device when used in daily lives.

6 CONCLUSION

In this paper, we present Endophasia, a sensing technique enabling issuing contact-free silent speech commands by utilizing acoustic-based face imaging. Endophasia perceives users' facial movements by detecting the phase change of acoustic signals and then it extracts users' 2D motion profiles by deconstructing the received signal into 240 channel taps. As generating acoustic-based images requires only a speaker and a microphone, Endophasia offers a non-invasive, power-efficient silent speech interface. Twenty silent speech commands were tested on an Android mobile phone powered by Endophasia. To verify the feasibility of the proposed solution, we collected data from 14 participants who tested the system using the 20 speech commands. Results from the 14 participants show an 88.83% within-user accuracy. We adopted transfer learning techniques to reduce the effort to customize the model for a new user. By collecting only 8 labeled samples per command, Endophasia achieves an average accuracy of 87.47% for new users. In addition, we designed an online learning scheme to incrementally improve the performance of a target user while the users use the system over time. With 240 unlabeled samples, our online learning scheme further improves the accuracy to 92.5%.

ACKNOWLEDGMENTS

We thank the anonymous reviewers whose suggestions helped improve and clarify the work. This work is supported by the Joint Key Project of the NSFC (U1736207), National Key R&D Program of China (2018YFB2101102), Startup Fund for Youngman Research at SJTU, and in part by the Ministry of Science and Technology of Taiwan (MOST108-2636-E-009-011- and 108-2633-E-002-001-), National Chiao Tung University.

REFERENCES

[1] Enrique Alameda-Hernandez, Des C McLernon, Aldo G Orozco-Lugo, M Mauricio Lara, and Mounir Ghogho. 2007. Frame/training sequence synchronization and DC-offset removal for (data-dependent) superimposed training based channel estimation. *IEEE Transactions on Signal Processing* 55, 6 (2007), 2557–2569.

[2] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. 2017. LipNet: End-to-End Sentence-level Lipreading. *arXiv: Learning* (2017).

[3] Adeola Bannis, Shijia Pan, and Pei Zhang. 2014. Adding directional context to gestures using doppler effect. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication - UbiComp '14 Adjunct* (2014), 5–8. https://doi.org/10.1145/2638728.2638774

[4] James V Bradley. 1958. Complete counterbalancing of immediate sequential effects in a Latin square design. *J. Amer. Statist. Assoc.* 53, 282 (1958), 525–528.

[5] Michael Brandstein and Darren Ward. 2013. *Microphone arrays: signal processing techniques and applications.* Springer Science & Business Media.

[6] Jonathan S. Brumberg, Alfonso Nieto-Castanon, Philip R. Kennedy, and Frank H. Guenther. 2010. Brain-computer Interfaces for Speech Communication. *Speech Commun.* 52, 4 (April 2010), 367–379. https://doi.org/10.1016/j.specom.2010.01.001

[7] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I Jordan. 2018. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2724–2732.

[8] Huawei Chen, Wee Ser, and Zhu Liang Yu. 2007. Optimal design of nearfield wideband beamformers robust against errors in microphone array characteristics. *IEEE Transactions on Circuits and Systems I: Regular Papers* 54, 9 (2007), 1950–1959.

[9] Joon Son Chung and Andrew Zisserman. 2016. Lip Reading in the Wild. (2016), 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.

[11] B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, and J.S. Brumberg. 2010. Silent speech interfaces. *Speech Communication* 52, 4 (2010), 270 – 287. https://doi.org/10.1016/j.specom.2009.08.002 Silent Speech Interfaces.

[12] digikey.com. 2020. 8 Ohms General Purpose Speaker 700mW 100Hz 20kHz Top Rectangular. Retrieved January 20, 2020 from https://www.digikey.com/product-detail/en/cui-inc/CMS-15113-078SP/102-5644-ND/8581915

[13] digikey.com. 2020. Knowles SPH0641LU4H-1 Microphone. Retrieved January 20, 2020 from https://www.digikey.com/product-detail/en/knowles/SPH0641LU4H-1/423-1402-1-ND/5332430

[14] Björn Engquist, Anna-Karin Tornberg, and Richard Tsai. 2005. Discretization of Dirac delta functions in level set methods. *J. Comput. Phys.* 207, 1 (2005), 28–51.

[15] EN ETSI. [n.d.]. 300 908 (GSM 05.02), Digital Cellular Telecommunications System. *Multiplexing and Multiple Access on the Radio Path* ([n. d.]).

[16] GDJR Forney. 1972. Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference. *IEEE Transactions on Information theory* 18, 3 (1972), 363–378.

[17] Masaaki Fukumoto. 2018. SilentVoice: Unnoticeable Voice Input by Ingressive Speech. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18)*. ACM, New York, NY, USA, 237–246. https://doi.org/10.1145/3242587.3242603

[18] S Golomb and R Scholtz. 1965. Generalized barker sequences. *IEEE Transactions on Information theory* 11, 4 (1965), 533–537.

[19] 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.

[20] 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. https://doi.org/10.1016/j.specom.2009.12.001 Silent Speech Interfaces.

[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] Harry Hollien, Donald Dew, and Patricia Philips. 1971. Phonational frequency ranges of adults. *Journal of Speech and Hearing research* 14, 4 (1971), 755–760.

[23] Walter L Kellermann. 2001. Acoustic echo cancellation for beamforming microphone arrays. In *Microphone Arrays*. Springer, 281–306.

[24] 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 (CHI '19)*. ACM, New York, NY, USA.

[25] Swarun Kumar, Stephanie Gil, Dina Katabi, and Daniela Rus. 2014. Accurate indoor localization with zero start-up cost. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 483–494.

[26] 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 (AH2019)*. ACM, New York, NY, USA, Article 1, 9 pages. https://doi.org/10.1145/3311823.3311831

[27] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, L. Kong, and M. Li. 2019. Lip Reading-Based User Authentication Through Acoustic Sensing on Smartphones. *IEEE/ACM Transactions on Networking* 27, 1 (Feb 2019), 447–460. https://doi.org/10.1109/TNET.2019.2891733

[28] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Yunfei Liu, and Minglu Li. 2018. Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 1466–1474.

[29] Wenguang Mao, Jian He, and Lili Qiu. 2016. CAT: High-Precision Acoustic Motion Tracking. *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking - MobiCom '16* (2016), 69–81. https://doi.org/10.1145/2973750.2973755

[30] Wenguang Mao, Mei Wang, and Lili Qiu. 2018. Aim: acoustic imaging on a mobile. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 468–481.

[31] Wenguang Mao, Zaiwei Zhang, Lili Qiu, Jian He, Yuchen Cui, and Sangki Yun. 2017. Indoor Follow Me Drone. *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys '17* (2017), 345–358. https://doi.org/10.1145/3081333.3081362

[32] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline. 2017. Silent Speech Recognition as an Alternative Communication Device for Persons With Laryngectomy. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 12 (Dec 2017), 2386–2398. https://doi.org/10.1109/TASLP.2017.2740000

[33] Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. 2017. Few-shot adversarial domain adaptation. In *Advances in Neural Information Processing Systems*. 6670–6680.

[34] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.

[35] Yoshitaka Nakajima, Hideki Kashioka, Nick Campbell, and Kiyohiro Shikano. 2006. Non-Audible Murmur (NAM) Recognition. *IEICE - Trans. Inf. Syst.* E89-D, 1 (Jan. 2006), 1–4. https://doi.org/10.1093/ietisy/e89-d.1.1

[36] Y. Nakajima, H. Kashioka, K. Shikano, and N. 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. V–708. https://doi.org/10.1109/ICASSP.2003.1200069

[37] Chuong H Nguyen, George K Karavas, and Panagiotis Artemiadis. 2017. Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features. *Journal of Neural Engineering* 15, 1 (nov 2017), 016002. https://doi.org/10.1088/1741-2552/aa8235

[38] Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan. 2007. BeepBeep: A High Accuracy Acoustic Ranging System Using COTS Mobile Devices. In *Proceedings of the 5th International Conference on Embedded Networked Sensor Systems (SenSys '07)*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/1322263.1322265

[39] Branislav M Popovic. 1992. Generalized chirp-like polyphase sequences with optimum correlation properties. *IEEE Transactions on Information Theory* 38, 4 (1992), 1406–1409.

[40] Swadhin Pradhan, Ghufran Baig, Wenguang Mao, Lili Qiu, Guohai Chen, and Bo Yang. 2018. Smartphone-based Acoustic Indoor Space Mapping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 75.

[41] Markku Pukkila. 2000. Channel estimation modeling. *Nokia Research Center* 17 (2000), 66.

[42] Theodore S Rappaport et al. 1996. *Wireless communications: principles and practice*. Vol. 2. prentice hall PTR New Jersey.

[43] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 806–813.

[44] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[45] st.com. 2020. Getting started with sigma-delta digital interface on applicable STM32 microcontrollers. Retrieved January 5, 2020 from https://www.st.com/content/ccc/resource/technical/document/application_note/group0/b2/44/42/9d/46/b4/4d/34/DM00354333/files/DM00354333.pdf/jcr:content/translations/en.DM00354333.pdf

[46] 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 (UIST '18)*. ACM, New York, NY, USA, 581–593. https://doi.org/10.1145/3242587.3242599

[47] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie. 2018. Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 591–605.

[48] Jiayao Tan, Xiaoliang Wang, Cam-Tu Nguyen, and Yu Shi. 2018. SilentKey: A new authentication framework through ultrasonic-based lip reading. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 36.

[49] taobao.com. 2020. STM43L476 Mini Development Board. Retrieved January 20, 2020 from https://item.taobao.com/item.htm?spm=a230r.1.14.298.499c2265yYV2qF&id=582824201272&ns=1&abbucket=20#detail

[50] VICON vero. 2019. Vero X, large field of view. Retrieved August 10, 2019 from https://www.vicon.com/products/camera-systems/vero

[51] Anran Wang and Shyamnath Gollakota. 2019. MilliSonic: Pushing the Limits of Acoustic Motion Tracking. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 18, 11 pages. https://doi.org/10.1145/3290605.3300248

[52] Wei Wang, Alex X Liu, and Ke Sun. 2016. Device-free gesture tracking using acoustic signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 82–94.

[53] Sangki Yun, Yi-Chao Chen, and Lili Qiu. 2015. Turning a mobile device into a mouse in the air. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 15–29.

[54] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-grained acoustic-based device-free tracking. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 15–28.

[55] Bing Zhou, Jay Lohokare, Ruipeng Gao, and Fan Ye. 2018. EchoPrint: Two-factor Authentication Using Acoustics and Vision on Smartphones. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom '18)*. ACM, New York, NY, USA, 321–336. https://doi.org/10.1145/3241539.3241575