EchoWhisper: Exploring an Acoustic-based Silent Speech Interface for Smartphone Users

YANG GAO, University at Buffalo, State University of New York, USA
YINCHENG JIN, University at Buffalo, State University of New York, USA
JIYANG LI, University at Buffalo, State University of New York, USA
SEOKMIN CHOI, University at Buffalo, State University of New York, USA
ZHANPENG JIN, University at Buffalo, State University of New York, USA

With the rapid growth of artificial intelligence and mobile computing, intelligent speech interface has recently become one of the prevalent trends and has already presented huge potentials to the public. To address the privacy leakage issue during the speech interaction or accommodate some special demands, silent speech interfaces have been proposed to enable people's communication without vocalizing their sound (e.g., lip reading, tongue tracking). However, most existing silent speech mechanisms require either background illuminations or additional wearable devices. In this study, we propose the EchoWhisper as a novel user-friendly, smartphone-based silent speech interface. The proposed technique takes advantage of the micro-Doppler effect of the acoustic wave resulting from mouth and tongue movements and assesses the acoustic features of beamformed reflected echoes captured by the dual microphones in the smartphone. Using human subjects who perform a daily conversation task with over 45 different words, our system can achieve a WER (word error rate) of 8.33%, which shows the effectiveness of inferring silent speech content. Moreover, EchoWhisper has also demonstrated its reliability and robustness to a variety of configuration settings and environmental factors, such as smartphone orientations and distances, ambient noises, body motions, and so on.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **Mobile devices**; • **Computer systems organization** → *Sensors and actuators.*

Additional Key Words and Phrases: Acoustic, echo, silent speech, smartphone

ACM Reference Format:
Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-based Silent Speech Interface for Smartphone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 80 (September 2020), 27 pages. https://doi.org/10.1145/3411830

1 INTRODUCTION

With the prevalence of smart and mobile devices and rapid advances of deep learning techniques, voice/speech interaction is becoming more popular and even pivotal in human-computer interaction. For decades, speech is

Authors' addresses: Yang Gao, University at Buffalo, State University of New York, Department of Computer Science and Engineering, Buffalo, NY, 14260, USA, ygao36@buffalo.edu; Yincheng Jin, University at Buffalo, State University of New York, Department of Computer Science and Engineering, Buffalo, NY, 14260, USA; Jiyang Li, University at Buffalo, State University of New York, Department of Computer Science and Engineering, Buffalo, NY, 14260, USA; Seokmin Choi, University at Buffalo, State University of New York, Department of Computer Science and Engineering, Buffalo, NY, 14260, USA; Zhanpeng Jin, University at Buffalo, State University of New York, Department of Computer Science and Engineering, Buffalo, NY, 14260, USA.

2474-9567/2020/9-ART80 $15.00
https://doi.org/10.1145/3411830

the most efficient way for people to interact with each other. Thus, speech interaction has natural advantages over typing, gestures, and visual interaction, in terms of better accessibility and higher efficiency. Nowadays, voice/speech interaction plays an irreplaceable role in many daily tasks such as communication (e.g., voice messages, phone calls), information search (e.g., "Hi, Siri"), car-navigation (e.g., voice assistant in Google Maps), and home-automation. It is reported that, the speech and voice interaction market is experiencing great high potential and expected to reach $21.5 billion by 2024 [33].

With the great convenience enabled by the speech interaction, the application scenarios (e.g., noisy street, crowded restaurant, and quiet library) are popularizing and expanding. However, to provide a robust voice-user interface in diverse environments, two challenges must be addressed. The primary information carrier of existing speech interaction is the user's voice, and it is very tough to clearly filter out the ambient noise, especially on some specific occasions (e.g., crowded streets, noisy trains). Also, due to the omnidirectional propagation of the sound, voice is not only received by the target microphone but is also exposed to the public. Disclosing personal information or sensitive content to the public is always risky and uncomfortable for users themselves.

Mitigating the noise from speech has been studied for decades, prior research has proposed many useful noise cancellation techniques from both the active and passive perspectives. One classic solution is the active noise reduction (ANR) that has already been deployed in some off-the-shelf earphones. It utilizes the inside speaker to emit a sound wave with the same amplitude but the inverted phase of the ambient noise captured by the external microphone. The emitted sound is merged into the original collected sound by the earphone as a destructive interference with the unwanted noise [19, 29]. However, although existing ANR systems are effective for steady noises within the low-frequency range [28] (<1000Hz), they can do little for transient sounds like a door slam or car roar/horn. The other common solution used in many smart speakers is microphone-array based noise cancellation. This technique requires at least two microphones with a differential topology and utilizes the spatial information such as phase difference of the signal that arrives at different microphones to suppress the noise [11].

Even though those noise cancellation approaches can largely overcome the influences of ambient noises, they still cannot avoid the privacy leakage during the speech interaction. Therefore, the concept of "silent speech" has been proposed. It generalizes the communication interface that does not need to utilize the sound when the user is vocalizing the speech. Lip reading is one of the most well-known solutions that can identify the speech content based on the user's mouth movement captured by a camera. Given the fact that the pronunciations of different vowels or words might have similar mouth shapes (e.g., "mike" and "bike"), the performance of vision-based lip reading could be restricted by imperfect illumination conditions which makes it hard or even impossible to capture the tongue's movement. Recent studies also explored the feasibility of inferring the speech through non-acoustic sensing. One solution called "non-audible murmur" leveraged the attached microphone to capture the skin vibration near the throat or neck during the speech [20, 38]. Magnetic based tongue tracking system was also proposed for silent speech recognition [45]. Nevertheless, those sensors require continuous contact with the skin, which would affect the user's daily activities and may cause skin irritations.

In this paper, without using any additional wearable sensors, we propose a novel smartphone-based, silent speech interface, named "*EchoWhisper*", utilizing near-ultrasound signals to interpret a user's speech as shown in Fig 1. Compared with existing mobile and wearable silent speech recognition (e.g., lip reading, throat vibration, airflow), *EchoWhisper* possesses three advantages in real-world application scenarios:

1) **Ubiquitous:** We make use of the microphones and speakers of off-the-shelf smartphones and the light-weight built-in app without introducing extra wearable sensors or sensor arrays.
2) **Secure:** Due to the inaudible transmission and its high resistance to noises and multi-device interferences, *EchoWhisper* provides the user with comprehensive security and privacy protection.
3) **User-friendly:** EchoWhisper allows the user to collect training data under the typical vocalized speech scenarios without tedious data preparation for silent speech behaviors.

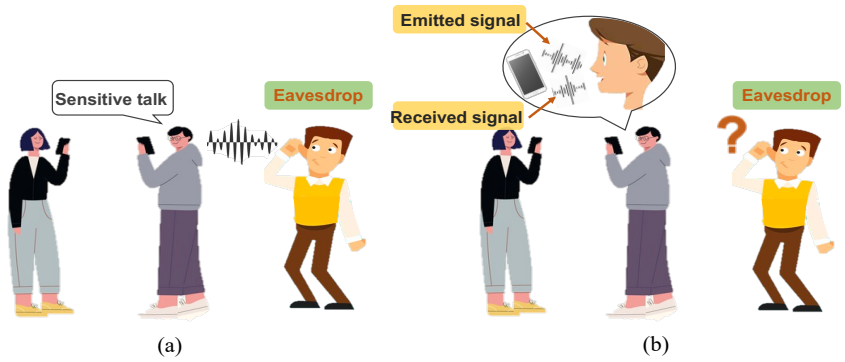


Fig. 1. Example of an application scenario for the *EchoWhisper* silent speech interface. (a) The malicious user can eavesdrop on the target user's speech with the sensitive and private content. (b) *EchoWhisper* leverages the reflected signals from the mouth motions to realize the mobile silent speech interaction without the information leakage.

Specifically, we make the following contributions in this work:

- We develop the *EchoWhisper*, an ease-of-use silent speech interface that leverages the Doppler shift of reflection of near-ultrasound sound waves caused by the mouth and tongue movements to interpret the speech.
- We customize a mobile application that is capable of emitting a pre-designed Continuous Wave (CW) signal and receiving the reflected echo through the dual microphones. Then we utilize the beamforming mechanism to localize the user's mouth and enhance the reflected echoes. A deep learning network *EchoNet* is proposed to recognize the speech content from the reflected echoes.
- We perform extensive experimental evaluations about the system's performance under diverse application scenarios, including smartphone orientations and distances, ambient noises, body motions, and multi-device interferences. Our results indicate that *EchoWhisper* possesses strong robustness and reliability and could be a promising mobile silent speech interface in people's daily life.

To the best of our knowledge, *EchoWhisper* is the first of its kind to leverage the smartphone's dual speakers and microphones with near-ultrasound signals for mobile silent speech interaction. It has been demonstrated that *EchoWhisper* has robust performance and possesses superior advantages in terms of its ubiquitous, secure, and user-friendly nature.

The rest of this paper is organized as follows. In Section 2, we review the state of the arts of existing silent speech interaction and acoustic sensing techniques. We present the scientific rationale for our system and the feasibility study in Section 3. In Sections 4 and 5, we elaborate on the system design scheme and frontend sensing module of EchoWhisper. Then we describe our signal processing, including data transfer, feature extraction, and customized EchoNet in Section 6. After that, we introduce the implementation details in Section 7 and present the performance evaluation and robustness analysis with human subjects in Section 8. In the end, we discuss the limitations and conclusions in Sections 9 and 10, respectively.

2 RELATED WORK

2.1 Silent Speech Interface

Silence Speech Interfaces (SSI) leverage inaudible information such as vibration [25], EMG [13, 25], lip and facial images [2, 41], Brain-Computer-Interface (BCI) [37, 51], ultrasound imaging [12, 22], and WiFi signals [55] to recognize the speech utterance.

Sun *et al.* [50] developed a camera-based silent speech interface by capturing the user's mouth movements, which was capable of recognizing issued commands. SilentVoice [17] was proposed as an "ingressive speech" method by placing a microphone to the user's mouth. This technique achieved silent speech interaction by recognizing the airflow during the whisper-like voice. Kimura *et al.* [27] designed the SottoVoce, an ultrasound imaging based silent voice interface with a customized ultrasonic prob attached under the user's jaw. Kim *et al.* [26] leveraged a magnetic tracer attached on the tongue's blade to capture the tongue's movement for SSI purpose. V-speech [34] was presented by Maruri *et al.*, which captured the voice signal with a bone-conduction sensor located in the nasal pads in smart glasses.

Tan *et al.* proposed the SilentTalk [52], an ultrasound-based lip-reading system on the mobile phone. Nevertheless, SilentTalk required a rather large set of silent speech data for training, and the phoneme-based model with limited mouth motions constrains the recognition accuracy for actual words and short sentences. Also, the airflow caused by plosives was not considered in the system, which might affect the performance.

2.2 Acoustic Sensing on Wearable and Mobile Devices

Acoustic sensing has been widely explored in the wearable and mobile computing community, mostly as a function of geometry approximation (e.g., floor plans [43, 49], obstacle estimation [57]), and motion detection including hand gestures [6, 8], respiration [56], heart beat [44], mouth movement [31], and object angular speed[30].

As smartphones become more and more prevalent and accessible in modern society, many studies have been focusing on mobile sensing in users' daily lives. Mao *et al.* [32] proposed an acoustic imaging system utilizing the speaker and microphone in the smartphone. They moved the smartphone along with the pre-defined path around the target to mimic the synthetic aperture radar imaging. SilentKey [53] took advantage of the fine-grained effects of selected mouth motions on the reflected signals emitted by the smartphone's speaker to present as an authentication modality. Similarly, Zhou *et al.* [61] developed a mobile facial authentication system combining acoustic and visual information.

Besides the mobile platform, acoustic sensing has also demonstrated great potentials in wearable devices. EarEcho [18] utilized the built-in speaker and microphone of the earbud to capture the acoustic profile in the user's ear canal for authentication. Yatani *et al.* [60] proposed a wearable acoustic sensor called BodyScope to record the sound near the user's throat for activity recognition. Beamband [23] is a wrist-worn sensing system containing an array of small transducers that utilize ultrasonic beamforming to track hand gestures.

3 CONCEPT AND RATIONALE OF SILENT SPEECH SENSING

In this section, we present the background of silent speech utterance and the scientific rationale of the acoustic sensing of mouth and tongue movements. We also conduct the feasibility study for proof of concept.

3.1 The Principle of Silent Speech

The human vocalization system generally consists of three parts: the lungs, vocal folds, and articulators, including the tongue, palate, and lip. The lungs first generate enough airflow to vibrate the vocal folds [36], and then the muscles of the larynx fine-tune the vocal folds to produce a specific pitch and tone. Lastly, the articulators articulate and filter the sound from the larynx and finalize it as the actual voice.

However, different from normal vocalizations, silent speech in our application scenario does not require airflow from the lungs and vibration from the vocal folds, it only replies on the inaudible information provided by articulators (e.g., lip, palate, and tongue).

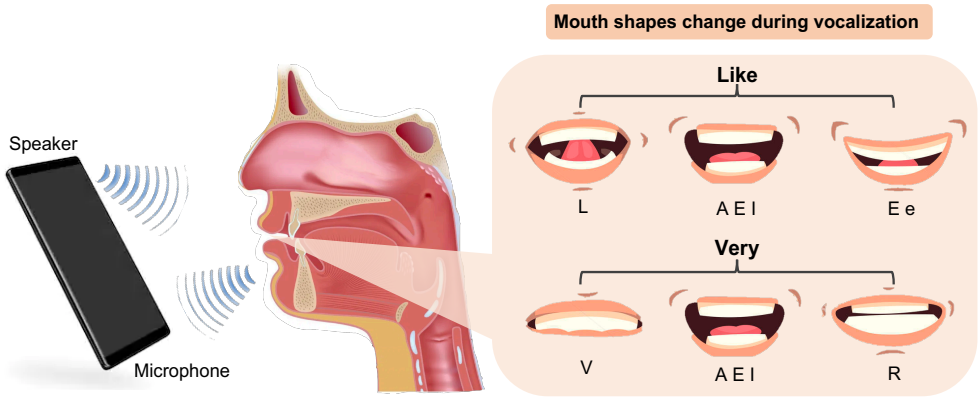


Fig. 2. Conceptual illustration of acoustic sensing of lip and tongue movements by a smartphone

3.2 Acoustic Sensing on Silent Speech

Based on the background of the silent speech process, we make an assumption that the lip and tongue, as well as their movements could be sensed by the near-ultrasound emitted by the speaker and captured by the microphone of smartphones. As illustrated in Fig. 2, the near-ultrasonic wave in the inaudible frequency band is emitted by the speaker of the smartphone to the user's mouth and oral cavity, and then reflected back to the microphone with the stationary path. When the user is vocalizing, the shape of the mouth and the position of the tongue will vary with the corresponding utterance. The mouth movements interfere with the incident and reflection paths of the near-ultrasonic wave, which would be recorded by the microphone probe of the smartphone.

According to the Doppler effect, assuming that the emitting wave comes with the frequency f_o, c denotes the propagation speed of the acoustic wave in the air, and Δv represents the relative moving speed of the lip and the tongue, the frequency variation observed by the microphone can be expressed below:

$$\Delta f = \frac{2\Delta v}{c} f_o \tag{1}$$

Given a typical value of Δv as 5cm/s, we observe from Eq. 1, to obtain an apparent frequency change, a higher frequency wave is needed. Thus, we choose the high-frequency sound at 20 kHz as the emitting wave to reach the resolution of frequency change at around 6 Hz.

3.3 Micro-Doppler Effect

In addition to the constant Doppler frequency shift caused by the excessive motions of the radar target, the micro-motions such as vibration, rotation, and small displacement [9](e.g., blades rotation in the helicopter, swings of arm for walking pedestrians) would also induce unique frequency modulations). Similarly, in our silent speech scenario, mouth and tongue are not the only reflection targets when the user is holding the smartphone near the mouth, the entire face behaves as a stronger reflection source. The echo reflection by the mouth would somehow be disturbed or even shadowed by the Doppler shift caused by the other facial motions such as eye blinking and sniffs. Therefore, an accurate estimation of the mouth location is essential to help increase the sensing resolution, which will be discussed in Section 5.2.

3.4 Feasibility Analysis

Given the hypothesis and the theoretical modeling of acoustic sensing with the Micro-Doppler Effect, to validate our idea of leveraging near-ultrasound at 20 kHz to sense the mouth and tongue motions, we performed and

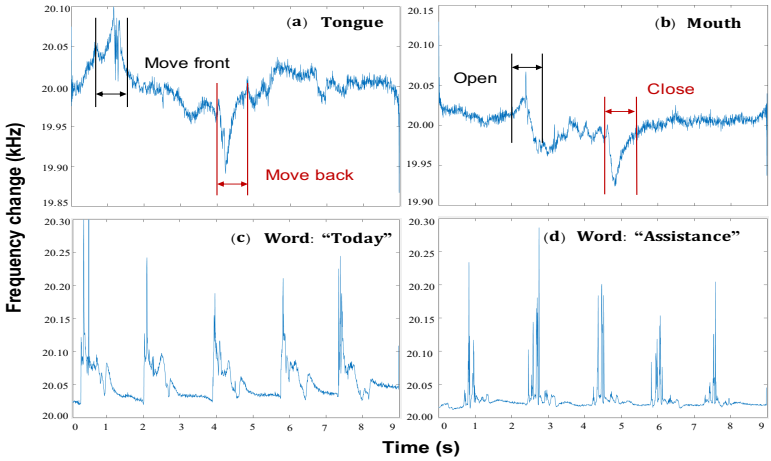


Fig. 3. Examples of the frequency changes of the reflected echoes over time, for different words or different mouth/tongue movements. (a) and (b) show the Doppler shift caused by the tongue and mouth movements. (c) and (d) show the unique and consistent patterns for different words.

examined a preliminary study. In this experiment, we asked the subject to perform the following tasks: 1) moving the tongue back and forth with the mouth open; 2) opening and closing the mouth; 3) speaking the word "Today" and "Assistance" without vocalization. The smartphone was fixed at a distance of 12 cm and faced to the subject's mouth. To better observe the frequency variations caused by the Doppler Effect, we chose the Continuous Wave (CW) with a fixed 20 kHz frequency as an emitted wave. Fig. 3 shows the frequency changes of the reflected signals received by the microphone of the smartphone for the corresponding tasks. We observed that the mouth and tongue movements would cause significant shifting and parabolic curves in the time-frequency domain of the reflected signal. For the same word or mouth shape, the frequency changes of near-ultrasound signals show a high level of similarity, and different words have distinguishable frequency patterns. Based on this observation, we validate our assumption that sensing the user's silent utterance by utilizing near-ultrasound emitted from the smartphone is feasible.

4 SYSTEM OVERVIEW

In this study, we propose the *EchoWhisper*, a smartphone-based silent speech interface, which utilizes the built-in speaker and microphone of the smartphone for echo sensing to interpret the user's silent speech. Based on the Doppler Effect, the movements of the mouth and tongue could be recognized by the acoustic profile of the reflected wave. The end-to-end methodological framework and processing flow are shown in Fig. 4, which consists of two major components: echo sensing and echo processing.

Echo Sensing. We utilize the dual speakers and microphones in the smartphone to emit and capture the reflected echoes. In the calibration stage, the mouth localization is estimated. Given the angle information of the mouth, we leverage the beamforming mechanism to enhance the echoes received by the top and bottom microphones.

Echo Processing. We segment the enhanced echoes through the event detection module. To mitigate the traditional data preparation process, we apply the airflow cancellation module to transfer the data collected under the vocalized speech scenarios to the desirable features for the silent speech recognition. Then, the spectrogram of the transformed data is fed into a deep learning model, EchoNet, to extract the Doppler-shift patterns.

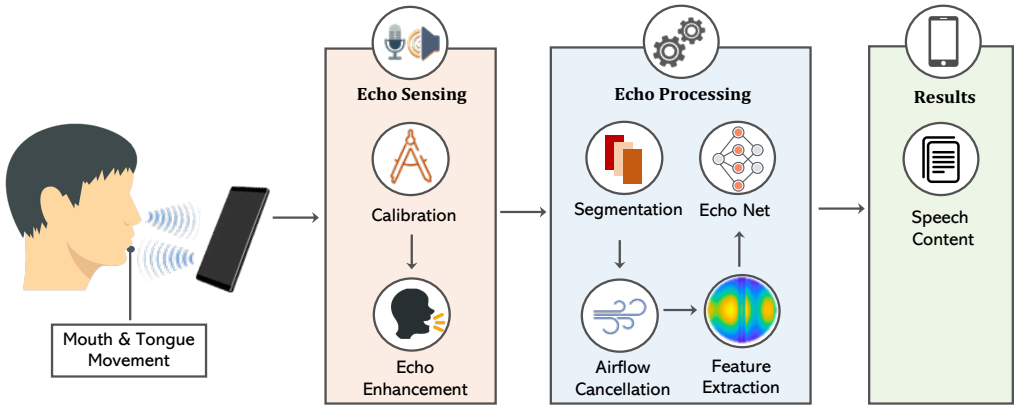


Fig. 4. Diagram of *EchoWhisper* methodological framework, consisting of an Echo Sensing front-end module to enhance the reflected echoes based on mouth localization, and an Echo processing module to recognize the Doppler shift patterns caused by mouth and tongue motions.

5 ECHOWHISPER SENSING

In this section, we will first introduce the dual speaker and microphone setting in the smartphone and then further present the adaptive enhancement module, including the mouth localization and echo enhancement.

5.1 Dual Microphone and Speaker Setting

Typically, most recently released smartphones (e.g., Google Pixel 3, Samsung Galaxy S8, and Apple iPhone X) would have two pairs of microphones and speakers, one pair in the bottom and the other pair next to the top camera, as shown in Fig. 5. Even though the top microphone is designed to record the ambient environment sound for noise cancellation, to capture the mouth location relative to the smartphone, we choose both the bottom and top microphones as the near-ultrasound receivers. We also use two speakers to emit the acoustic signal for better detection coverage.

5.2 Adaptive Enhancement

Although most smartphone speakers are non-directional and have wide spatial radiation range, our *EchoWhipser* utilizes inaudible and high frequency-band signal of 20 kHz, which has a narrower beamwidth of the center lobe, a large displacement of the mouth would cause a non-negligible incident and reflection path deviation. Thus, it is crucial to precisely estimate the mouse location for higher recognition accuracy. Since the frequency of the emitted signal is 20 kHz, the corresponding wavelength is 1.7 cm. The distance between the user's mouth and smartphone is roughly around 3 cm to 10 cm during the silent speech, which is smaller than the Fraunhofer distance of the acoustic signal ($d_F = \frac{2D^2}{\lambda}$, where D is the longitude of the microphone array and λ is the signal wavelength) [5]. However, due to the limited number of microphones N in the smartphone (N=2), to approximately estimate the mouth location with incident angle θ as shown in Fig. 5, we consider the acoustic propagation as a far-field model which only analyzes the phase changes of received signals.

Mouth Localization. Given the two received acoustic signals by the top and bottom microphones, the signals are reflected from both the face and mouth of the user. To better localize the mouth position among other static reflections (e.g., noise, face), we ask the user to open the mouth or speak any word without vocalization.

Beamforming is a widely used wireless sensing technique used in sensor arrays for directional signal processing [16] and estimation of Direction of Arrival (DOA) [58]. Even though the two microphones (i.e., top and bottom)

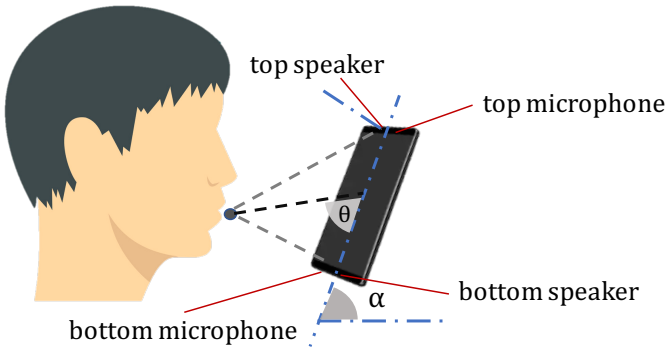


Fig. 5. Layout of the smartphone microphones and speakers. α indicates the holding angle of the smartphone. Azimuth θ represents the angles from the mouth to the center of two-element microphone array. We assume the user faces perpendicularly to the smartphone surface with the zero elevation angle.

might not be exactly identical, after calibration, those two microphones could still be used as a uniform linear dual-microphone array. In this paper, we use the MUSIC (Multiple Signal Classification) algorithm [47], which infers the orthogonality of signal subspace and noise subspace to construct the spatial spectral function, and detects the DOA of signal through the peak search. The algorithm performs an eigenspace analysis of the signal's correlation matrix to estimate the signal's frequency content. The MUSIC spatial spectrum estimate is given by

$$R_x = \frac{1}{N}\sum_{i=1}^{N} X(i)X^H(i) = AR_sA^H + \sigma^2 I \tag{2}$$

where N is the number of receiving signals, $X(i)$ is the signal array collected from microphones, R_x is the corresponding eigendecomposition, and R_s denotes the signal autocorrelation matrix, with the noise variance σ^2 and the Vandermonde matrix A of samples of the signal frequencies. Then we can get the noise subspace matrix E_n with the given dimension of the signal subspace p:

$$E_n = [\mathrm{v}_{p+1}, \mathrm{v}_{p+2}, ..., \mathrm{v}_M], A^H\mathrm{v}_i = 0 \tag{3}$$

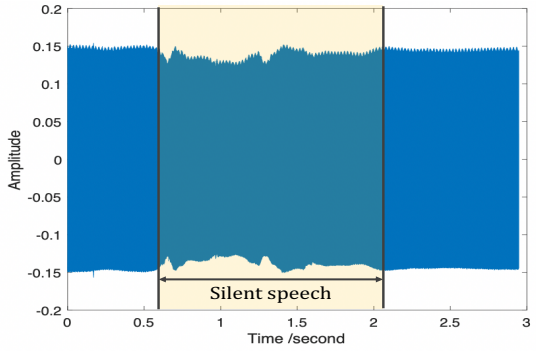
We can get the spatial spectrum shown as below:

$$P_{\mathrm{MUSIC}}(\hat{\theta}) = \frac{1}{\mathrm{a}^H(\hat{\theta})E_nE_n^H\mathrm{a}(\hat{\theta})} = \frac{1}{|E_n^H a(\hat{\theta})|^2} \tag{4}$$

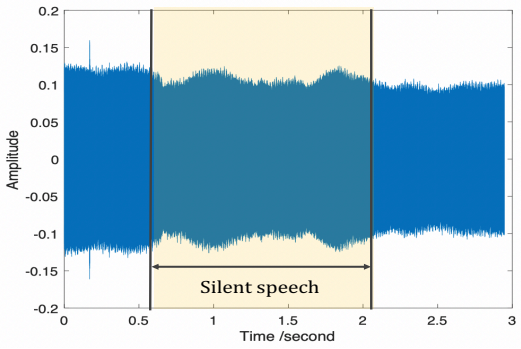
From the orthogonality of the signal and noise subspaces, in this work, we assume the incident direction θ as shown in Fig. 5 is approximately close to the arriving angle $\hat{\theta}$ of microphones which can be estimated by step searching the peak in the estimator function $P_{\mathrm{MUSIC}}(\hat{\theta})$.

Signal Enhancement. After obtaining the information of the incident direction of the mouth towards the smartphone, given the distance of the microphone array d, we can calculate the time delay Δt of the signal between the two microphones. Then the signals from each microphone are time-aligned and summed. Time alignment is applied by transforming the signals into the frequency domain and applying the linear phase shifts corresponding to the time delay. The synthesized signal in the time domain is shown below:

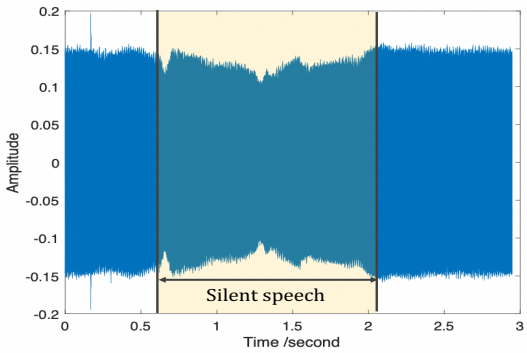
$$y(t) = \frac{1}{N}\sum_{i=1}^{N} \omega_i x_i(t - \Delta t) \tag{5}$$



(a) Received signal from the bottom microphone



(b) Received signal from the top microphone



(c) Beamformed signal

Fig. 6. With 45 degree in azimuth and 0 degree in elevation, the raw signals received in both top and bottom microphone and the beamformed signal during the silent speech of word "apple" (['æpl]).

where y(t) is beamformed output, ω_i is the weight for each microphone, and $x_i(t)$ indicates the receiving signal from each microphone.

With the beamformed signal, we can enhance the reflected echo from the mouth and minimize those undesired echoes reflected from other objects (e.g., eye, nose, background environment). As shown in Fig. 6, we can observe the amplitude changes in raw signals caused by silent speech. Compared with signals from the top and bottom microphones during the 1.5 seconds of silent speech, the beamformed signal would be more explicit corresponding to the mouth movement.

6 ECHOWHISPER PROCESSING

6.1 Segmentation / Event Detection

To eliminate the effects of the low SNR segments and to increase the energy efficiency, especially for the long sentence recognition scenarios, we adopt a Likelihood Radio Test (LRT) and Hidden Markov Model (HMM)-based event detection module to filter out undesired low-power density echo segments [48]. Given that:

$$\begin{aligned} H_0 &: \text{event absence} : X = N \\ H_1 &: \text{event presence} : X = N + S \end{aligned} \tag{6}$$

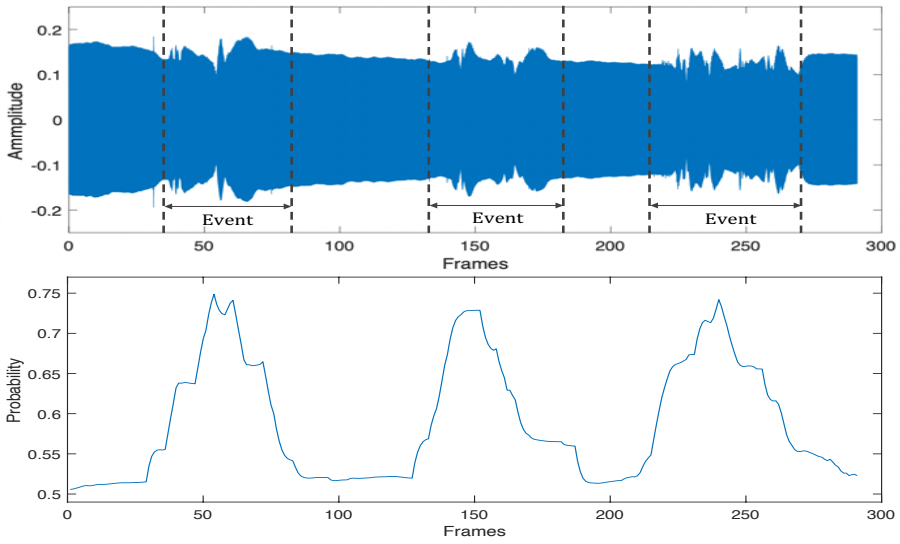


Fig. 7. Example of event detection in the sentence of silent speech and the corresponding event occurrence probability

where S, N, X are Discrete Fourier Transform (DFT) coefficient vectors of echo with motions, noise, and noisy echo, with their kth elements S_k, N_k, and X_k, respectively.

The probability density functions conditioned on H_0 and H_1 are given by:

$$\begin{aligned} p(X|H_0) &= \prod_{k=0}^{L-1} \frac{1}{\pi\lambda_N(k)} \exp\{-\frac{|X_k|^2}{\lambda_N(k)}\} \\ p(X|H_1) &= \prod_{k=1}^{L-1} \frac{1}{\pi[\lambda_N(k)+\lambda_S(k)]} \cdot \exp\{-\frac{|X_k|^2}{\lambda_N(k)+\lambda_S(k)}\} \end{aligned} \tag{7}$$

where $\lambda_N(k)$ and $\lambda_S(k)$ represent the variances of N_k and S_k. The likelihood ratio for the kth frequency band is

$$\Lambda_k \triangleq \frac{p(X_k|H_1)}{p(X_k|H_0)} = \frac{1}{1+\xi_k} \exp\{\frac{\gamma_k \xi_k}{1+\xi_k}\} \tag{8}$$

where ξ_k and γ_k are called *a priori* and *a posteriori* Signal-to-Noise Ratios (SNR's). The decision rule is obtained from the average likelihood ratio for each band, which is given by

$$\log\Lambda = \frac{1}{L}\sum_{k=0}^{L-1} \log\Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{9}$$

As shown in Fig. 7, given a beamformed signal collected by the microphone, we first calculate the event probability for every frame, and then segment frames within the events (i.e., mouth movement) with the high probability.

6.2 Feature Extraction

As discussed in Section 3.2, the silent speech behaviors, including mouth and tongue movements, would cause phase and frequency changes of the reflected signals. After signal segmentation, we utilize the spectrogram as the feature presentation of the reflected near-ultrasound echoes. The spectrogram contains information in both

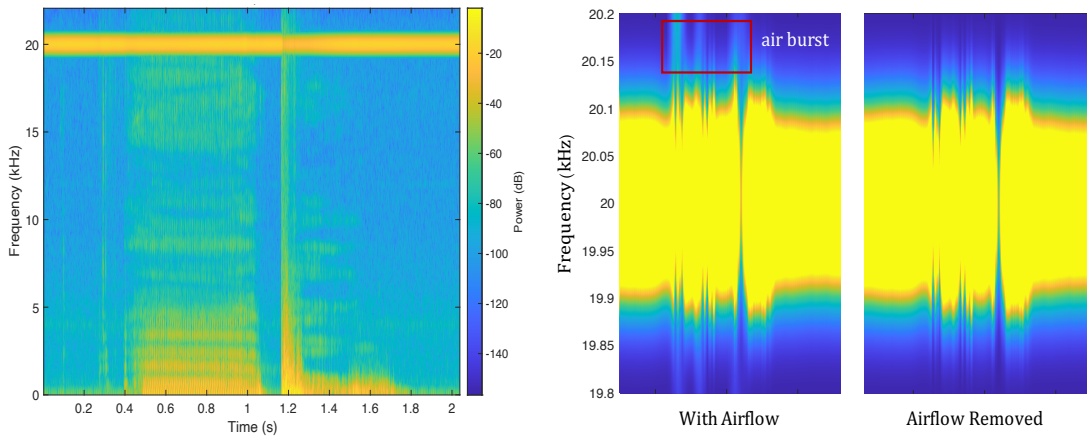


Fig. 8. (a) Spectrum of the word "apple" (['æpl]) with plosive p. (b) Comparison of patterns in frequency domain with airflow burst and airflow removed.

frequency and time domains. The 2D patterns are highly correlated with mouth and tongue movements. The spectrogram is also defined as the Power Spectral Density of the function:

$$\text{spectrogram}\{x(t)\}(\tau,\omega) \equiv |X(\tau,\omega)|^2 = \left|\sum_{n=-\infty}^{\infty} x[n]\omega[n-m]e^{-j\omega n}\right|^2 \tag{10}$$

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

$$\omega[n] = \frac{I_0\left(\beta\sqrt{1-\left(\frac{n-N/2}{N/2}\right)^2}\right)}{I_0(\beta)}, 0 \le n \le N \tag{11}$$

6.3 Data Augmentation

Because the speaking behaviors (e.g., mouth shapes, the distance between the mouth and the smartphone) and smartphone models are user-dependent, to achieve a good performance of the deep learning model, a tremendous amount of training data from the individual is essential. However, it is extremely inconvenient, and perhaps impracticable to ask users to collect hours of silent speech data. Considering the popular usage of normal vocalized speech recognition nowadays (e.g., voice commands, voice messages, phone calls), we propose a data augmentation strategy that uses the microphones to collect the normal sound under typical speech scenarios, which could be then transformed and used for training the silent speech model. Similar to the silent speech scheme, we also enable the speakers to emit 20 kHz CW signals and record both the human voices and the reflected echoes by both top and bottom microphones.

Label Generation. Before the data transformation, we utilize the Google Could Speech-to-Text API [10] to generate the corresponding labels, including words and segmentation frames.

High-Pass Filter. As the classic theorem of the human voice, the frequency covers from 20 Hz to the highest tone C8 at 4,186 Hz [3]. The operating frequency of the CW signal in our design is 20 kHz, which has no

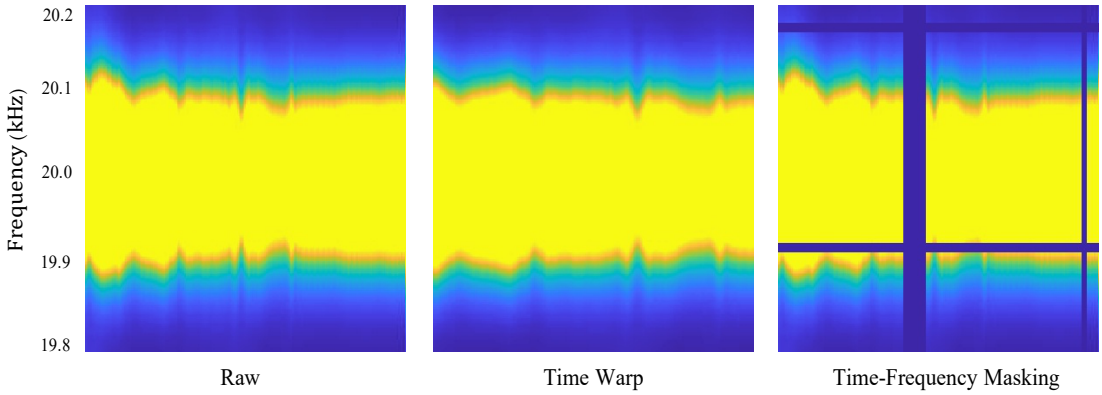


Fig. 9. Augmentation policies applied to the spectrum input, including time warping and time-frequency masking.

conflict with the regular frequency bandwidth of the human voice. To generate the trainable data for silent speech recognition, we first apply a Butterworth high-pass filter with a stop frequency at 15 kHz to remove the interference caused by the undesired human voice,

Airflow Cancellation. As discussed in Section 3.1. The major difference between vocalized speech and silent speech is the vocal cord vibrations caused by air pressure or airflow, especially for plosive phonemes. Plosive, also known as an oral occlusive, is a consonant that blocks the vocal tract to cease the airflow in the mouth. There are many common phonemes related with plosives, including ([t], [d]) with tongue tip or blade occlusion, ([k], [g]) with tongue body occlusion, ([p], [b]) with lips occlusion [21]. In phonetics, it consists of three steps, airflow stop, articulation, and burst release. The release of airflow burst combined with the voice is both recorded by the microphone. The airflow bursts of some plosives would induce noises in a wide frequency range (20 Hz - 10 kHz), or even the full frequency range (20 Hz - 22 kHz) in a short distance to the microphone (around 5 cm). In the training phase, to substantially reduce the users' burden of data collection and transfer the data collected from the vocalized speech data to the silent speech data, we apply an airflow cancellation process. Since the vibration energy of an microphone's membrane caused by airflow bursts hold high consistency within the frequency band from 15 kHz to 22 kHz, we can estimate the noise power of our target zone (19 kHz - 21 kHz) based on the knowledge of the buffer zone (16 kHz - 18 kHz). The filtered spectrogram is calculated as below:

$$\hat{\mathrm{S}}(\tau, \omega_{\mathrm{t}}) = \mathrm{S}(\tau, \omega_{\mathrm{t}}) - \mathrm{S}(\tau, \omega_{\mathrm{b}}) \otimes \mathcal{G} \tag{12}$$

where ω_t and ω_b represents the frequency band of the target zone and the buffer zone, respectively. $\mathcal{G}$ is the Gaussian blur kernel for smoothing.

We illustrate the airflow cancellation mechanism in Fig. 8. The overall spectrogram of the word "apple" (left) shows that the 20 kHz CW signal is interfered with by the airflow of [ae] and plosive [p]. Specifically, in Fig. 8(b), we can observe the comparison of frequency shift patterns with airburst tailing and airflow removed, respectively.

Spectrum Augmentation. To achieve better robustness against noise, we aim to generate additional noisy training data for our deep neural network model. Inspired by the recent success of data augmentation in the speech recognition domain, we utilized the augmentation policy proposed by Dainel Park *et al.* [40], which consisted of warping the features along time steps, masking blocks of frequency channels, and masking blocks of time steps. As shown in Fig. 9, time warping is applied by fixing anchor points on the boundary - four corners and two mid-points of the vertical edges, and warping the random points along with the horizontal line to either left or right by a certain distance. Time-Frequency is applied by masking consecutive frequency channels and

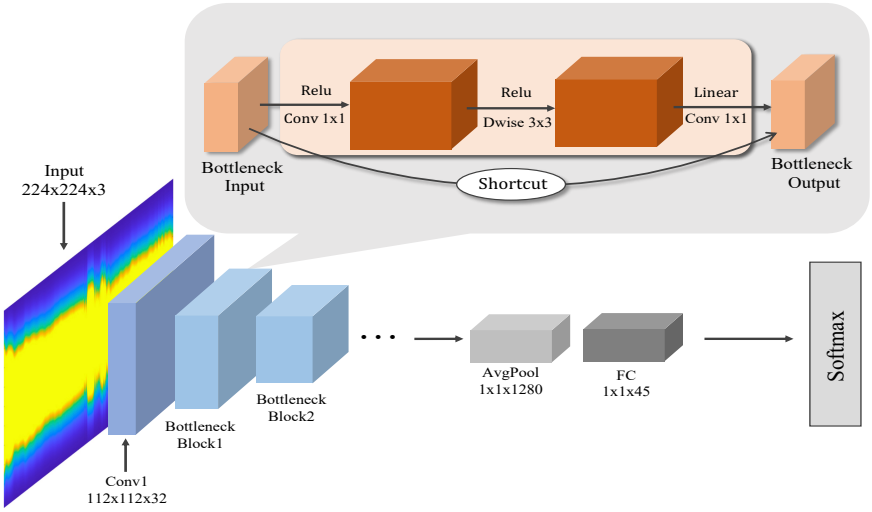


Fig. 10. Architectural diagram of EchoNet for performing silent speech recognition from the spectrograms of the reflected echoes.

time frames. In this paper, we use both time warping and time-frequency masking with customized augmentation parameters.

6.4 *EchoNet* Deep Neural Network Architecture

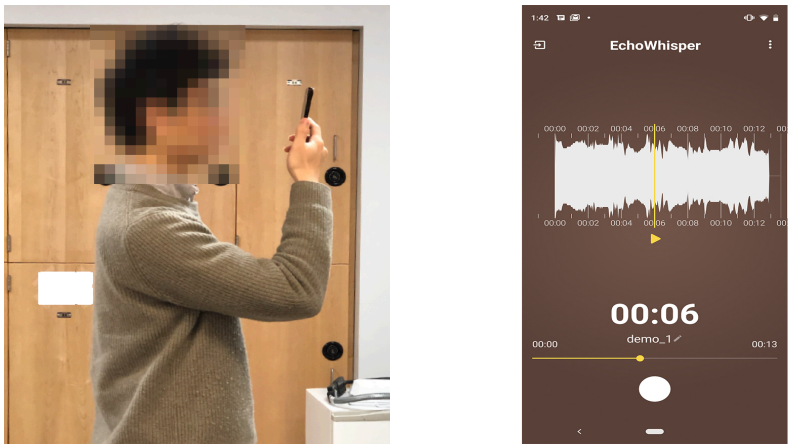
EchoWhisper aims to recognize patterns of different words from the reflected echoes on the 2D spectrograms, which could be eventually performed on personal mobile devices. However, storage space and power consumption are two vital challenges for mobile and embedded computing. To this end, we propose an *EchoNet* deep neural network architecture, based upon and modified from the popular MobileNet V2 [46], which is a lightweight and efficient on-device visual recognition model to extract the rich feature representations. To reduce the computational resources, MobileNet replaces the traditional convolution operation using the combination of depthwise separable convolution and 1×1 pointwise convolution. Also, it introduces two new features to the architecture: 1) linear bottlenecks between the layers, and 2) shortcut connections between the bottlenecks.

As shown in Fig. 10, in *EchoNet*, the input is the $224 \times 224 \times 3$ spectrogram image, and the first layer is the standard 112×112 fully convolution layer with 32 filters followed by batch normalization and clipped ReLU with the ceiling of 6. Then 17 residual bottleneck layers, including two types "stride=1" and "stride=2", are added after the Conv1 layer. Each bottleneck contains three steps: 1) 1×1 Conv2D with ReLU6 for expansion; 2) 3×3 depthwise with ReLU6 for feature filtering; 3) 1×1 Conv2D with linear activation for compression. To improve the gradient propagation towards deep layers, similar with the classical residual connection, shortcuts are added in "stride=1" bottlenecks. After the bottleneck layers, we add a global average pooling layer and a fully connected layer for the output dimension matching. Finally, the probability of 45 different word categories could be calculated from the softmax layer with the cross-entropy loss.

7 SYSTEM IMPLEMENTATION

7.1 Dataset

7.1.1 Preparation. We conducted extensive experiments to validate the effectiveness and robustness of our proposed *EchoWhisper*. Human participants were instructed to sit on the chair in a casual position. We asked the



(a) (b)

Fig. 11. (a) Experimental setup for silent speech interface. (b) An example of the smartphone application interface.

participants to hold the smartphone with about 5 cm away from the mouth and interact with the customized app installed on the smartphone. The recorded echoes were then sent to the desktop computer equipped with an Nvidia GeForce GTX 1080 8GB GPU.

7.1.2 Participants. 5 subjects aged from 24 to 30 years (2 females and 3 males) were recruited in our experiments. These participants included both native and non-native English speakers. All subjects had experiences of using voice-chats in their daily smartphone usages. The experiment was approved by the Internal Review Board (IRB) of the University at Buffalo for human subjects.

7.1.3 Data Collection. To ensure the generality and practicability of our experiment, we choose one daily conversation in the Test of English for International Communication (TOEIC) [15] that has been widely used for evaluating the daily speaking skills, and 20 words covering different phonemes and syllables from the lipreading practice list [35] as our reading materials. In total, our reading material covers 45 different words, including many high-frequency words (e.g., 'going', 'first', 'same', 'now'). To compare the performance against existing vision-based lip reading solutions, some pairs of words such as 'my', 'nine' with similar mouth shapes, but different tongue placements are also included. To increase the robustness of our recognition capability in daily usage, we also introduce the extra unknown class that contains irrelevant but common behaviors such as yawning and coughing. During the experiment, all participants were asked to read the reading material 30 times with vocalization in a controlled lab environment (on average 60 seconds each time). As the testing dataset, all participants read the material in a silent manner for five times. In total, we have 6,750 samples for training and 1,125 samples for testing. Considering the fact that each subject might perform a silent speech in different ways for the same phoneme or word, including mouth shape and speech speed, our EchoWhisper model is user-dependent and is trained separately for each participant. Thus, each user's model has 1,350 samples for training and 225 samples for testing.

7.2 EchoNet Implementation

We implemented the proposed EchoNet in TensorFlow. We first created the base model from the pre-trained MobileNet V2 on the ImageNet database [1] with over 1 million images and various object categories. This base of knowledge helped us better extract features such as edges and verticals. Then, we froze the first 10 layers

and added the classifier head on top of it. To prevent the network from overfitting, we also implemented image augmentation by flipping and translating spectrogram images. We used the Adam optimizer with an initial learning rate of 0.002, a 10% piecewise decay schedule, and the batch size of 16, to train our EchoNet network.

7.3 Smartphone Application

To provide an easy-to-use silent speech recognition interface, we developed an Android application for *EchoWhisper*. This app can emit the pre-defined CW wave and then capture the reflected echoes through both the top and bottom microphones when the user is performing silent speech behavior. Built upon the open-source app, Audio recorder [42], we added the playback function that enabled the ability of playback and record simultaneously. We used Android Studio 3.5.2 to develop the app running on a Google Pixel 3XL. This app is also lightweight, with an overall size of 4 MB. After recording, to prevent information loss during data compression, the app can output the lossless Waveform Audio File (.wav) format at a 44,100 sample rate.

7.4 Evaluation Metrics

Word Error Rate (WER). In an automatic speech recognition system, WER is the most widely used metric for performance evaluation. Instead of the phoneme level, the WER works at the word level based on the Levenshtein distance, which describes the difference between two sequences. Word Error Rate can be calculated as:

$$\mathrm{WER} = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \tag{13}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words, and N is the total number of words in the sentence.

8 PERFORMANCE EVALUATION

In this section, based on extensive experiments, we evaluate our *EchoWhisper* from the following three perspectives: 1) overall accuracy for different words, 2) usability for daily operations, and 3) robustness and reliability against environmental factors.

8.1 Recognition Performance

To better demonstrate the association between the frequency-domain features and the corresponding silent speech activities in the *EchoWhipser*, we illustrate an example of a 5-word sentence in the Doppler-based spectrogram in Fig. 12. From the spectrogram, we can observe clear frequency variation patterns, including edges and vertices caused by lip and tongue movements related to different words. For example, some words such as "my" and "was" that have strong mouth movements would result in sharper edges. Those features could be extracted through multiple convolution layers in our EchoNet network.

8.1.1 Performance over Different Network Structures. We evaluated the performance of *EchoWhisper* based on the metric discussed in Section 7.4. Fig. 13 presents the WERs for different deep learning models, including AlexNet, VGG-16, ResNet-18, ResNet-50, and EchoNet. ResNet-50 and EchoNet outperform other structures. Considering that EchoNet is easier to train (3 million parameters) compared with ResNet-50 (25 million parameters), and higher transportability (13 MB) on mobile devices compared with ResNet-50 (96 MB), we choose the EchoNet as our silent speech classifier.

8.1.2 Performance over Different Word Types. To further analyze our system's performance on the collected words from the phonology perspective, we listed the accuracy over words with different numbers of syllables, as shown in Table 1. It can be seen that two-syllable words present higher accuracy than three-syllable words. This is because, compared with two-syllable words, during vocalization, the three-syllable words need more

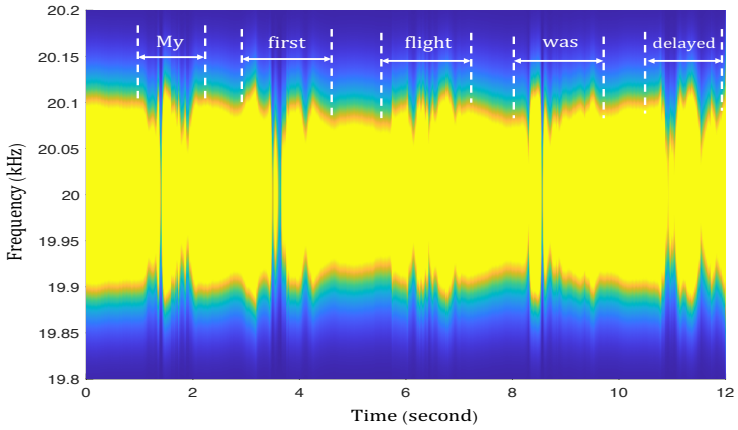


Fig. 12. An example illustrates the spectrogram of sentence "My first flight was delayed" with segmentations.

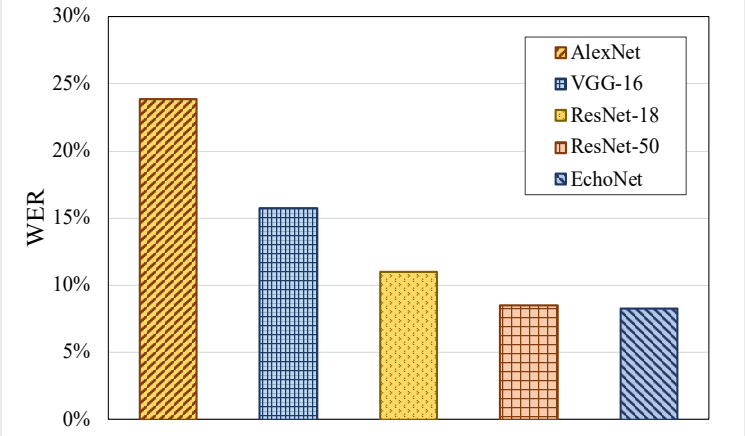


Fig. 13. The average WER of *EchoWhisper* with different deep learning structures

Table 1. The average accuracy among testing words with different number of syllables

	One-syllable	Two-syllable	Three-syllable	Confusing Words[1]	Total
No. words	23	12	10	6	45
Avg. Accuracy	92.08%	94.92%	86.36%	81.82%	91.67%

[1] Confusing words denote the pair of words have similar lip movements and are easy to be incorrectly recognized by visual lip reading. For example, {my, nine}, {take, eight}, {want, was}.

complicated and subtle mouth movements and might introduce small noisy variations, which decreases the consistency. Since we added three pairs of confusing words with one-syllable such as {my, nine}, {take, eight}, {want, was}, the overall accuracy of one-syllable words drops to around 92.08%.

8.2 Usability Analysis

8.2.1 Learnability. Learnability is a major factor in measuring the usability of any human-computer interfaces. It is a well-known fact that the performance of a deep learning model highly relies on the number of training

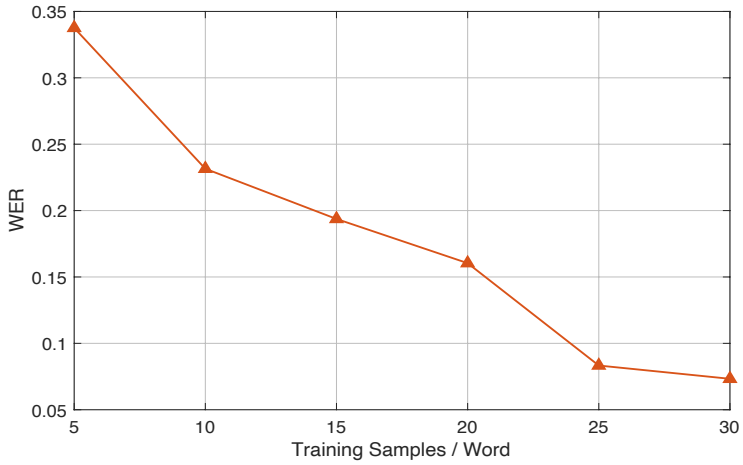


Fig. 14. WER varies under different numbers of training samples per word

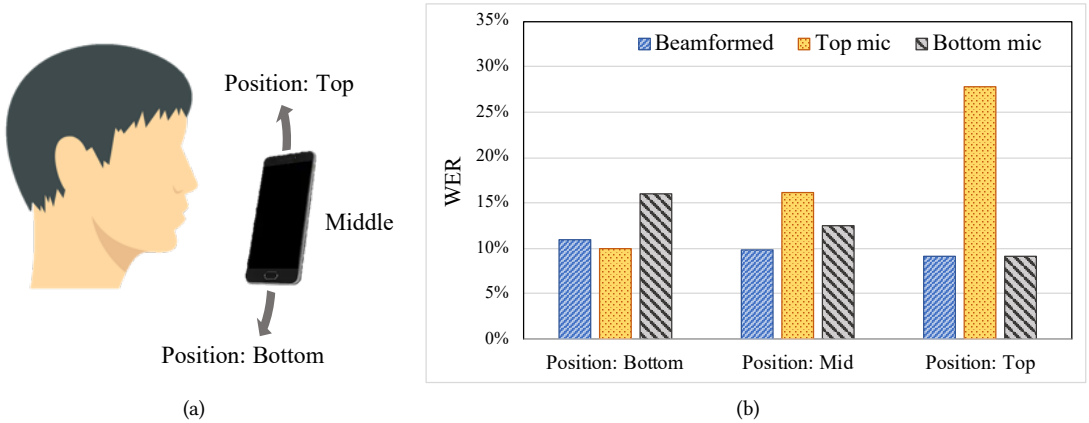
Table 2. The average WER and power consumption on different smartphone platforms.

	Pixel 3 XL	**Pixel 3**	**Galaxy S8**
No. Microphones	Dual	Dual	Dual
Power (mAh/min)[1]	5.88	5.40	3.13
Avg. WER[2]	8.33%	6.31%	9.26%

[1] The average smartphone battery size for Pixel 3 XL, Pixel 3, and Galaxy S8 is 3000 mAh.
[2] The results are tested under the same subject group.

data samples. Even our system can facilitate the learnability by augmenting data from daily usage of typical voice chats; it is still necessary to evaluate how the data collection would affect the system's performance. Fig. 14 shows the *EchoWhisper*'s performance (in terms of word error rate, WER) along with the increasing training samples for each word. It is manifest that, when we only collect a small amount of the user's voice data, the captured features corresponding to mouth and tongue movements are underrepresented. Thus, it results in higher WERs. After we collect over 25 samples per word for training, the WER becomes stable at the level of around 8.33%, which can provide a better user experience.

8.2.2 Smartphones. As a silent speech interface on mobile devices, it is necessary to investigate the potential influences of design and manufacturing variations of speakers and microphones on our system's performance. Thus, we also tested the *EchoWhisper* on different smartphone platforms. We asked two subjects to read the same materials, including all 45 words on multiple smartphones. We evaluated the power consumption of emitting and recording near-ultrasound signals and the recognition accuracy in Table 2. From the table, considering the typical battery size of a smartphone is approximately 3000 mAh, we can see that the power consumption of our system's front-end module is quite low (0.1% - 0.2% per minute) on all the smartphones and could well satisfy the daily-usage requirements. The performance of all three models, including Google Pixel 3XL, Pixel 3, and Samsung Galaxy S8, have low and similar WERs, which demonstrates satisfactory usability and stability of the proposed approach over different smartphone platforms.



(a) (b)

Fig. 15. (a) An example of holding the smartphone with different positions. (b) The average WER comparison among different smartphone holding positions.

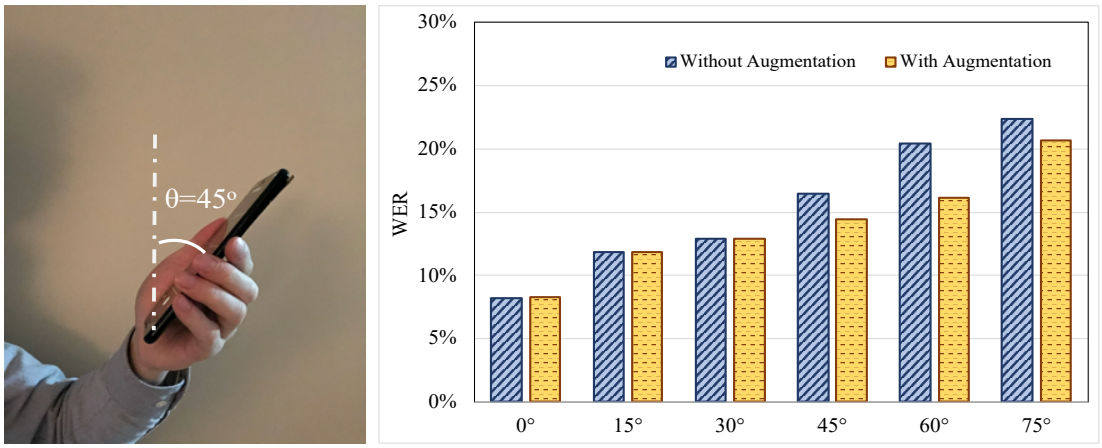
8.3 Robustness Quantification

In this section, we evaluate the robustness of *EchoWhisper* in terms of smartphone positions, smartphone orientations, varying distances between the mouth and the smartphone, background noises, and body motions.

8.3.1 Robustness on Smartphone Positions. Different users may have various preferences on smartphone-holding positions during the silent speech or normal speech interactions. Sometimes, the relative positions of the microphones on the smartphone towards the user's mouth would vary, as shown in Fig. 15(a) that includes three smartphone positions relative to the mouth during the usage. To validate the performance of beamforming discussed in Section 5.2, we recorded silent speech data from three subjects, in all three different smartphone-holding positions for both training and testing. In each position, we evaluated WERs for the top, bottom microphone signals, and the beamformed signal, respectively. It is seen that, as shown in Fig. 15(b), in the scenario of the middle position which is a very common position, the beamformed signal outperforms the top and bottom microphone signals. While in the scenarios of the bottom or top position, either the top or the bottom microphone signal has relatively larger WERs compared with the beamformed one. Therefore, our beamforming mechanism is effective for the robustness of smartphone position variations.

8.3.2 Robustness on Smartphone Orientations. Besides the smartphone position, holding angles are also user-dependent. Even the same user may change the holding orientations in different scenarios. Therefore, we implemented the experiment to examine how the smartphone orientation would affect the performance. During the registration phase for training, the subjects recorded the echoes with the smartphone holding vertically ($\theta = 0°$). As shown in Fig 16(a), we then measured the reflected echoes with different smartphone angles θ from $0°$ to $75°$ to simulate the varying smartphone holding behaviors in daily usage. From Fig. 16(b), unsurprisingly, it is seen that the WER increases along with the accumulating holding angle θ, which is consistent with our expectation, as the angle increases, the mouth would gradually exceed the top microphone's sensing coverage, which reduces the quality of reflected echoes. Specifically, with the help of time-warping augmentation policy, starting from $45°$, the WER gets relative 2% - 5% lower than the result without augmentation, which shows that our augmentation strategy helps increase the system robustness against smartphone holding variations.

8.3.3 Robustness on Smartphone Distances. Besides the orientation, acoustic sensing is also sensitive to the propagation distance. Thus, We are motivated to evaluate the WER under different distances between the



(a) (b)

Fig. 16. (a) An example of holding the smartphone with a specific orientation. (b) The average WER comparison among different smartphone holding orientations with and without the augmentation strategy.

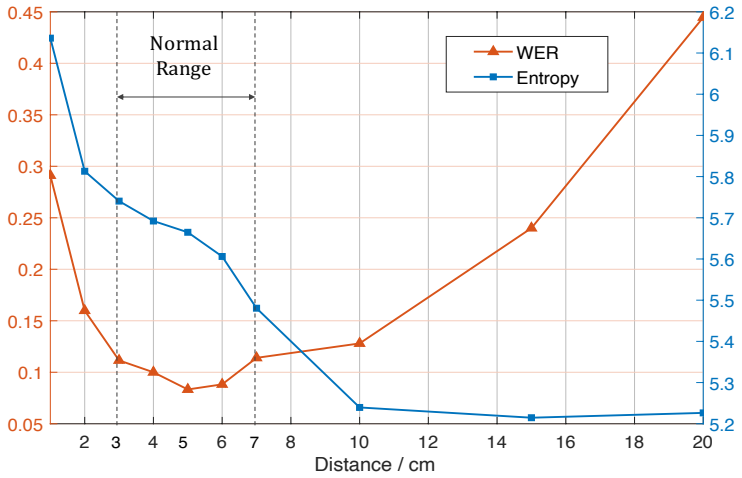


Fig. 17. WER variations under different distances. Red line indicates the varying WER, and blue line represents the spectrogram entropy of the reflected echo. The normal distance between the mouth and the smartphone during voice-based usage ranges from 3cm to 7cm.

smartphone and the mouth. Considering that the typical distance between the smartphone's microphone and the mouth for voice-based usages (e.g., voice-messages) ranges from 3 cm to 7 cm, thus such distance is fixed at 5 cm during the training phase of our experiments. We asked the subjects to place the smartphone away from their mouth at a distance from 1 cm to 20 cm, including some abnormal distances and perform the silent speech task. To describe the information loss of reflected echoes, we measured the entropy of the gray-scaled spectrogram features which are defined as follows:

$$\mathrm{H}(x) = \sum_{i} \mathrm{P}(x_i) \log_2 \mathrm{P}(x_i) \tag{14}$$

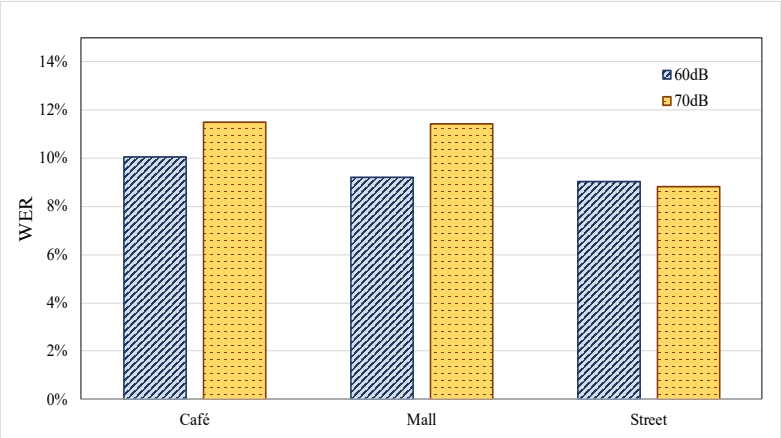


Fig. 18. The average WER comparison among different ambient noises.

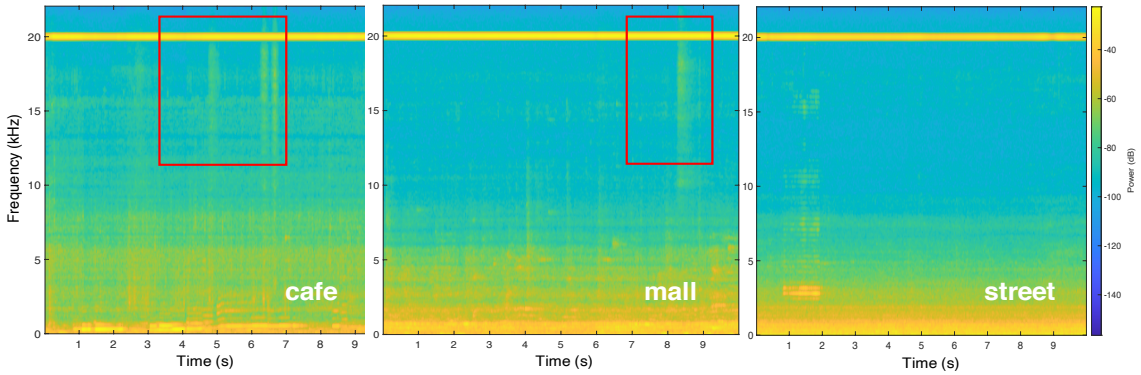


Fig. 19. The spectrogram comparison of recorded echoes among different ambient noises

where x_i represents the pixel in the spectrogram image. The lower entropy means the less information being carried by the image.

As shown in Fig. 17, within the distance of normal range from 3cm to 7cm, the WER remains stable. If the smartphone is placed too close to the mouth, the microphone will no longer be able to capture the whole shape of the mouth, which results in lower accuracy. On the other side, if the distance is too large (> 10 cm), the spectrogram pattern caused by the Doppler shift would be submerged by the echoes reflected from surrounding static multi-paths, which could be observed from the gradually dropped entropy.

8.3.4 Resistance to Ambient Noises. With the popularity of smartphones and the convenience of voice-chatting and voice-messaging, people are more and more willing to use voices in many mobile communication scenarios. Nevertheless, rich and diverse application scenarios also inherently demand a high level of resistance to various background noises. Thus, we examined our system under three different noisy environments: a cafe with background music, a shopping mall, and a noisy street. To ensure the replicability and controllability of the experiments, during the training phase, we used data collected in the quite room and then simulated the testing scenarios by playing background noises at two different sound pressure levels [7, 24, 39]. We used a smart speaker to play the background sound as the noise source with 0.5 meters away from subjects. The results demonstrate

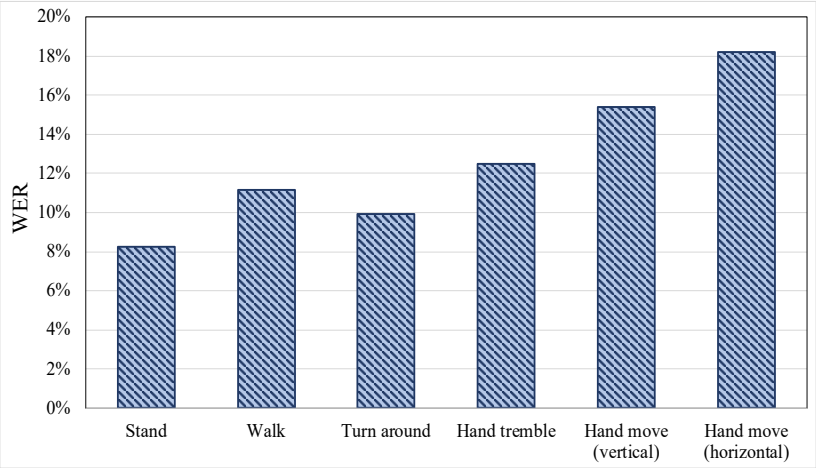


Fig. 20. Performance comparison (average WER) for different body motions

that, as shown in Fig. 18, various background noises have limited impacts on the performance of *EchoWhisper* and only slightly increase WER (<4%) compared with the controlled environment. Specifically, cafe and mall environments with background music would cause higher WER compared with the street with random ambient noise. We further investigated the frequency distribution of the ambient noise as shown in Fig. 19, there are several high-frequency sparkles (e.g., bells, ringing of cymbals, or harmonics) marked by red boxes in the spectrograms of the cafe and mall environments that are mixed with the CW signals, which slightly affect the performance of our system.

8.3.5 Reliability on Body Motions. Besides the background noises, users' body motions during the usage of *EchoWhisper* (e.g., voice messaging while walking, yawning, tuning the head around while making a phone call) would also cause disturbance of received echo signals. To evaluate the robustness of our system, we tested the EchoWhisper for multiple user body motions: standing, walking, turning around, hand movements in the vertical and horizontal directions. To better emulate the real-life scenarios, we asked the participants to perform selected body motions with a high degree of freedom (e.g., random handshakes, walking at a common pace) while holding the smartphone. Fig. 20 represents the overall performance under each selected body motion scenario. Except for the hand movements, behaviors such as standing, walking, turning around, hand tremble would only cause less than 5% WER increment, which indicates the resistance of our approach to those motion artifacts. Compared with the first four motions, hand movements with higher speeds and larger displacements between the mouth and the smartphone would introduce a higher level of undesired frequency shift that should be avoided during the usage.

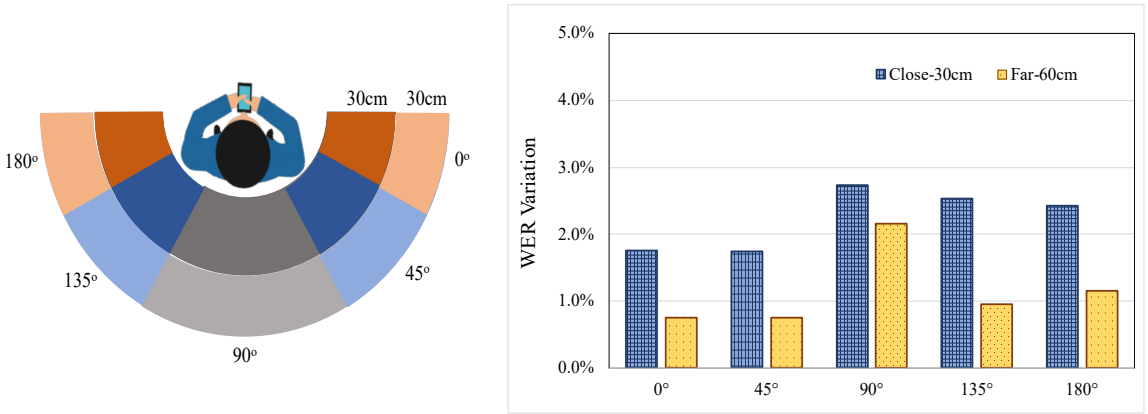
9 DISCUSSIONS

9.1 Advantages over Visual Lip-reading

As aforementioned, one of the major challenges for visual lip-reading is the information loss of the tongue movement, especially under a poor illumination condition. LipNet [4] is a state-of-the-art visual lip-reading solution, which achieves an average 4.8% WER on the GRID corpus dataset. However, the authors also mentioned that LipNet could not perform well to distinguish several vowels such as /æ/ and /ih/. This is because these two vowels hold similar mouth shapes but different tongue displacements (i.e., for /æ/, the tongue stays low and at the front of the mouth; while for /ih/, the front of the tongue slightly arches up towards the roof of the mouth [14]). To evaluate the *EchoWhisper*'s capability of capturing the tongue movements, in addition to the 3 pairs

Table 3. The average accuracy of distinguishing confusing-word pairs

	[æ] / [ih]	my / nine	want / was	take / eight
No. Samples	30	47	47	45
Avg. Accuracy	83.3%	82.9%	76.6%	88.9%



(a) (b)

Fig. 21. Multiple Devices Interference. (a) Interference map with 10 sub-zones. (b) WER variations of all sub-zones

of confusing words, we also tested the two extra vowels, /æ/ and /ih/, the overall accuracy of distinguishing confusing-word pairs is listed in Table 3. The results clearly demonstrate the superior capability and advantage of our proposed approach in terms of sensing the subtle displacement of the tongue.

9.2 Multiple Devices Interference

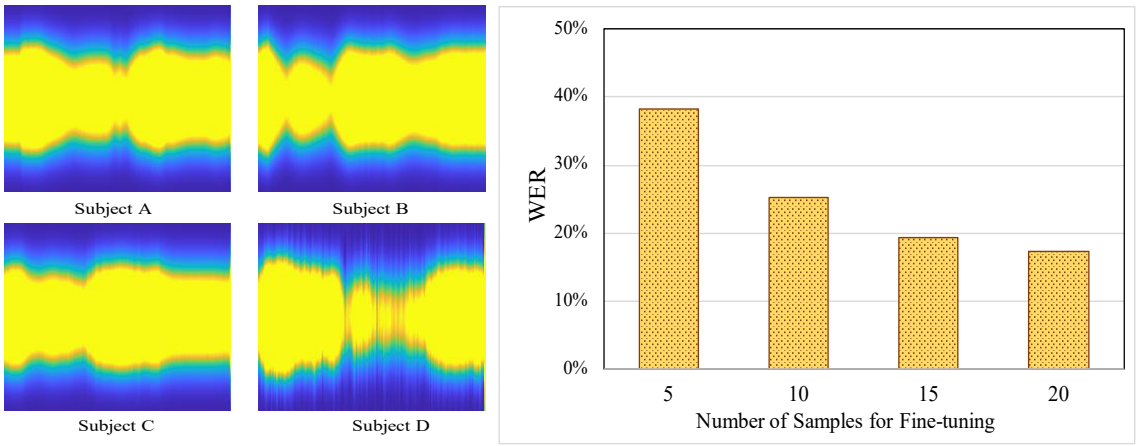
As a mobile application for daily usage, it is possible to meet a situation where other users in the surrounding area may hold their smartphones to use the same app. Those apps are also transmitting the near-ultrasound signals and might cause some interference with each other. To evaluate our system under a real and complex environment, we implemented the experiment to check how our system's performance would be affected by another user's interference. As shown in Fig. 21(a), we divided the target user's space behind into 10 sub-zones covering 0-180 degree and 0-60 cm distance. We asked one subject to perform the silent speech while there was another smartphone emitting a 20 kHz signal from one of the sub-zones towards the target user's direction. As shown in Fig. 21(b), in general, the *EchoWhisper*'s performance is robust to the interference from other devices with less than 3% WER variation. Specifically, the interference from the closer sub-zones (within 30cm) causes slightly higher WER drops.

9.3 Comparison with Inaudible Silent Speech Interfaces

We compared the performance of *EchoWhisper* with other emerging inaudible speech interfaces deployed on mobile and customized devices, including SilentTalk [52], WiHear [55], WaveEar [59], and SottoVoce [27]. As shown in Table 4, generally, sensing modality with higher frequency (i.e., Wifi: 2.4 GHz; mmWave: 24 GHz; Ultrasound: 3.5 MHz) could reach higher accuracy and relatively larger recognition capability. Nonetheless, customized mmWave and USRP devices would be the major limitation and not suitable for mobile application

Table 4. Comparison with existing wireless-based silent speech interfaces

Interfaces	**EchoWhisper**	SilentTalk [52]	WiHear [55]	WaveEar [59]	SottoVoce [27]
Features	Acoustic	Acoustic	WiFi	24 GHz mmWave	3.5 MHz Ultrasound
Capability	45 words	12 motions	32 words	100+ words	100+ words
Devices	Smartphone	Smartphone	USRP	mmWave Probe	Ultrasound Probe
Portable	Yes	Yes	No	No	Yes
Contact Type	Non-contact	Non-contact	Non-contact	Non-contact	Skin-contact
Accuracy	92%	95%/75%	91%	94%	90%



(a) (b)

Fig. 22. Speaker-independent Modelling. (a) Echo patterns of the word "OK" among four different users. (b) Given a generalized model, WERs of the new user with different numbers of samples per word for fine-tuning.

scenarios. SottoVoce utilizes the portable ultrasound probe, but the apparatus needs to be attached to the user's jaw to recognize the silent speech content, which is also less feasible and user-friendly for daily usage. Compared with SilentTalk, our proposed *EchoWhisper* has better sensing capability and robustness by leveraging the beamforming technique to enhance the reflected echoes with dual microphones.

9.4 Speaker-independent Modelling

To explore the potential of the speaker-independent model, we first illustrated the echo patterns of the same word among different users in Fig. 22(a). We can observe a general pattern (i.e., frequency shifts) resulted from the pronunciation rules. To further investigate the speaker-independent model in the warm start scenario, we trained a simplified speaker-independent model based on four subjects silent speech data with a WER of 18%. Then we evaluated a new user's data on the pre-trained, generalized model with different amounts of fine-tuned samples. As seen from the results shown in Fig. 22(b), the WER gradually drops and converges with the increasing numbers of fine-tuned samples per word, which suggests the great potential of the speaker-independent modelling with more testing subjects.

9.5 Limitations and Future Work

In this study, the proposed *EchoWhisper* presents a new promising way of silent speech interface in daily life. However, the proposed technique still exhibits several limitations in its current stage. In order to further enhance the *EchoWhisper*'s accuracy and usability, we discuss the future work from the following aspects:

9.5.1 Small Vocabulary. In this work, we only selected 45 sensitive words in a daily conversation with limited training samples. To avoid the skewed phoneme-level samples, we chose the word-level segmentation and utilized the CNN-based end-to-end model. As the further work, by expanding the dataset, we would segment audio samples into short-time frames and adopt the CNN-LSTM-based architecture to generate the estimated phoneme for each frame, and then infer the speech content with a CTC loss.

9.5.2 Language Modeling. The current recognition system solely relies on the phonic model that recognizes features from acoustic profile caused by mouth and tongue movements with a limited dataset. Given the fact that some words have very similar mouth and tongue movements (e.g., the confusing words in Table 1), to help further correct the prediction grammatically and semantically, a language modeling (e.g., N-gram, Conditional Random Field) needs to be included in the decoding part of EchoNet.

9.5.3 Mobile Deployment of Deep Learning Model. In our current system, we developed the mobile app for the front-end module to emit and capture the reflected signals. But the deep-learning-based model runs on the PC end, which is still not a fully-engineered application. In the future work, to have a better evaluation of system overhead on mobile devices, we are planning to deploy the EchoNet to the Android platform by using TensorFlow Lite [54].

10 CONCLUSION

In this paper, we proposed a smartphone-based silent speech interface named *EchoWhisper*, leveraging the dual microphone and speaker setting to capture and recognize the user's mouth and tongue movements without vocalization. First, we explored the feasibility of sensing the mouth and tongue movements through the near-ultrasound emitted by the smartphone. Then, we proposed a dual-channel mouth localization mechanism to enhance the reflected echoes. After that, the reflected echoes containing the Doppler shift information caused by the movements of mouth and tongue were fed into a customized EchoNet deep learning model for pattern recognition. Extensive experiments under real-world scenarios with various simulated configuration and environmental variations have shown the effectiveness and robustness of EchoWhisper to precisely recognize the silent speech content for up to 45 words.

REFERENCES

[1] 2019. ImageNet. http://www.image-net.org. [Online; accessed 20-Jan-2020].
[2] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani. 2018. Lip2AudSpec: Speech reconstruction from silent lip movements video. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2516–2520.
[3] Abdullah I Al-Shoshan. 2006. Speech and music classification and separation: a review. *Journal of King Saud University-Engineering Sciences* 19, 1 (2006), 95–132.
[4] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. 2016. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599* (2016).
[5] Constantine A Balanis. 2016. *Antenna theory: analysis and design*. John wiley & sons.
[6] Vincent Becker, Linus Fessler, and Gábor Sörös. 2019. GestEar: combining audio and motion sensing for gesture recognition on smartwatches. In *Proceedings of the 23rd International Symposium on Wearable Computers*. 10–19.
[7] Alfredo Calixto, Fabiano B. Diniz, and Paulo H. Zannin. 2003. The statistical modeling of road traffic noise in an urban setting. *Cities* 20, 1 (2003), 23–29.

[8] Mingshi Chen, Panlong Yang, Jie Xiong, Maotian Zhang, Youngki Lee, Chaocan Xiang, and Chang Tian. 2019. Your Table Can Be an Input Panel: Acoustic-based Device-Free Interaction Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 3.
[9] Victor C Chen, Fayin Li, S-S Ho, and Harry Wechsler. 2006. Micro-Doppler effect in radar: phenomenon, model, and simulation study. *IEEE Transactions on Aerospace and electronic systems* 42, 1 (2006), 2–21.
[10] Google Cloud. 2019. speech-to-text. https://cloud.google.com/speech-to-text/docs/apis. [Online; accessed 15-Jan-2020].
[11] Mattias Dahl and Ingvar Claesson. 1999. Acoustic noise and echo cancelling with microphone array. *IEEE transactions on Vehicular Technology* 48, 5 (1999), 1518–1526.
[12] Bruce Denby, Jun Cai, Thomas Hueber, Pierre Roussel, Gérard Dreyfus, Lise Crevier-Buchman, Claire Pillot-Loiseau, Gérard Chollet, Sotiris Manitsaris, and Maureen Stone. 2011. Towards a practical silent speech interface based on vocal tract imaging.
[13] Yunbin Deng, James T Heaton, and Geoffrey S Meltzner. 2014. Towards a practical silent speech recognition system. In *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association.*
[14] Rachel's English. 2020. English: How to Pronounce IH Vowel. https://rachelsenglish.com/english-pronounce-ih-vowel/. [Online; accessed 20-Jan-2020].
[15] ETS. 2019. TOEIC Listening and Reading Test. https://www.ets.org/s/toeic/pdf/toeic-listening-reading-sample-test-updated.pdf. [Online; accessed 15-Jan-2020].
[16] Otis Lamont Frost. 1972. An algorithm for linearly constrained adaptive array processing. *Proc. IEEE* 60, 8 (1972), 926–935.
[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.* ACM, 237–246.
[18] Yang Gao, Wei Wang, Vir V Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using Ear Canal Echo for Wearable Authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–24.
[19] Colin N Hansen. 1999. *Understanding active noise cancellation.* CRC Press.
[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.
[21] Yoshiyuki Horii and Paul A Cooke. 1978. Some airflow, volume, and duration characteristics of oral reading. *Journal of Speech and Hearing Research* 21, 3 (1978), 470–481.
[22] Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone. 2010. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication* 52, 4 (2010), 288–300.
[23] Yasha Iravantchi, Mayank Goel, and Chris Harrison. 2019. BeamBand: Hand Gesture Sensing with Ultrasonic Beamforming. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, 15.
[24] Jian Kang. 2006. *Urban sound environment.* CRC Press.
[25] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. Alterego: A personalized wearable silent speech interface. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces.* ACM, 43–53.
[26] Myungjong Kim, Nordine Sebkhi, Beiming Cao, Maysam Ghovanloo, and Jun Wang. 2018. Preliminary Test of a Wireless Magnetic Tongue Tracking System for Silent Speech Interface. In *Proceedings of the 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS).* IEEE, 1–4.
[27] 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.* ACM, 146.
[28] Sen M Kuo, Sohini Mitra, and Woon-Seng Gan. 2006. Active noise control system for headphone applications. *IEEE Transactions on Control Systems Technology* 14, 2 (2006), 331–335.
[29] Sen M Kuo and Dennis R Morgan. 1999. Active noise control: a tutorial review. *Proc. IEEE* 87, 6 (1999), 943–973.
[30] Zeshui Li, Haipeng Dai, Wei Wang, Alex X Liu, and Guihai Chen. 2018. Pcias: Precise and contactless measurement of instantaneous angular speed using a smartphone. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–24.
[31] Li Lu, Jiadi Yu, Yingying Chen, Hongbo Liu, Yanmin Zhu, Yunfei Liu, and Minglu Li. 2018. Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals. In *Proceedings of the 2018 IEEE Conference on Computer Communications (INFOCOM).* IEEE, 1466–1474.
[32] 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.
[33] Marketsandmarkets. 2019. Speech and Voice Recognition Market by Technology (Speech and Voice Recognition), Vertical (Automotive, Consumer, Government, Enterprise, Healthcare, BFSI), Deployment (On Cloud On-Premises/Embedded), and Geography - Global Forecast to 2024. https://www.marketsandmarkets.com/Market-Reports/speech-voice-recognition-market-202401714.html. [Online; accessed 15-Jan-2020].
[34] Héctor A Cordourier Maruri, Paulo Lopez-Meyer, Jonathan Huang, Willem Marco Beltman, Lama Nachman, and Hong Lu. 2018. V-Speech: Noise-Robust Speech Capturing Glasses Using Vibration Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 180.

[35] Gloria McGregor. 2019. Lipreading Practice 2018. https://lipreadingpractice.co.uk/. [Online; accessed 15-Jan-2020].
[36] Deirdre D. Michael. 2018. About the voice. http://www.lionsvoiceclinic.umn.edu/page2.htm#physiology101. [Online; accessed 19-Jan-2020].
[37] Beomjun Min, Jongin Kim, Hyeong-jun Park, and Boreom Lee. 2016. Vowel imagery decoding toward silent speech BCI using extreme learning machine with electroencephalogram. *BioMed Research International* 2016 (2016).
[38] Yoshitaka Nakajima, Hideki Kashioka, Kiyohiro Shikano, and Nick Campbell. 2003. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing.(ICASSP'03).*, Vol. 5. IEEE, V–708.
[39] Christopher C. Novak, Joseph L. Lopa, and Robert E. Novak. 2010. Effects of sound pressure levels and sensitivity to noise on mood and behavioral intent in a controlled fine dining restaurant environment. *Journal of Culinary Science & Technology* 8, 4 (2010), 191–218.
[40] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779* (2019).
[41] Stavros Petridis, Jie Shen, Doruk Cetin, and Maja Pantic. 2018. Visual-Only Recognition of Normal, Whispered and Silent Speech. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6219–6223.
[42] Dmitriy Ponomarenko. 2019. Audio recorder (Voice recorder, Sound recorder). https://github.com/Dimowner/AudioRecorder. [Online; accessed 15-Jan-2020].
[43] 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.
[44] Kun Qian, Chenshu Wu, Fu Xiao, Yue Zheng, Yi Zhang, Zheng Yang, and Yunhao Liu. 2018. Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices. In *Proceedings of the 2018 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 1574–1582.
[45] Himanshu Sahni, Abdelkareem Bedri, Gabriel Reyes, Pavleen Thukral, Zehua Guo, Thad Starner, and Maysam Ghovanloo. 2014. The tongue and ear interface: a wearable system for silent speech recognition. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers*. ACM, 47–54.
[46] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4510–4520.
[47] Ralph Schmidt. 1986. Multiple emitter location and signal parameter estimation. *IEEE transactions on antennas and propagation* 34, 3 (1986), 276–280.
[48] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. 1999. A statistical model-based voice activity detection. *IEEE signal processing letters* 6, 1 (1999), 1–3.
[49] Qun Song, Chaojie Gu, and Rui Tan. 2018. Deep Room Recognition Using Inaudible Echos. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 135.
[50] 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*. ACM, 581–593.
[51] Yousef Rezaei Tabar and Ugur Halici. 2017. Brain Computer Interfaces for Silent Speech. *European Review* 25, 2 (2017), 208–230.
[52] Jiayao Tan, Cam-Tu Nguyen, and Xiaoliang Wang. 2017. SilentTalk: Lip reading through ultrasonic sensing on mobile phones. In *Proceedings of the 2017 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 1–9.
[53] 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.
[54] TensorFlow. 2020. TensorFlow Lite Guide. https://www.tensorflow.org/lite/guide. [Online; accessed 10-Feb-2020].
[55] Guanhua Wang, Yongpan Zou, Zimu Zhou, Kaishun Wu, and Lionel M Ni. 2016. We can hear you with Wi-Fi! *IEEE Transactions on Mobile Computing* 15, 11 (2016), 2907–2920.
[56] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW based contactless respiration detection using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 170.
[57] Zi Wang, Sheng Tan, Linghan Zhang, and Jie Yang. 2018. ObstacleWatch: Acoustic-based Obstacle Collision Detection for Pedestrian Using Smartphone. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 194.
[58] Darren B Ward, Zhi Ding, and Rodney A Kennedy. 1998. Broadband DOA estimation using frequency invariant beamforming. *IEEE Transactions on Signal Processing* 46, 5 (1998), 1463–1469.
[59] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. 2019. Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 14–26.
[60] Koji Yatani and Khai N Truong. 2012. BodyScope: a wearable acoustic sensor for activity recognition. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 341–350.

[61] 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*. ACM, 321–336.