The Trumpet of the Swan: Decoding Lip Language for Speech-impaired Users through COTS Earbuds

MING GAO, Nanjing University of Posts and Telecommunications, China
YICHEN DAI, Nanjing University of Posts and Telecommunications, China
ZHANBO ZHOU, Nanjing University of Posts and Telecommunications, China
ZHENGXIN GUO, Nanjing University of Posts and Telecommunications, China and Tongda College of Nanjing University of Posts and Telecommunications, China
JINSONG HAN, Zhejiang University, China
FU XIAO*, Nanjing University of Posts and Telecommunications, China

Lip language is a straightforward method of non-verbal communication for people with laryngeal and lingual injuries, dysphonia, or vocal cord lesions. It is convenient for these speech-impaired individuals, especially when their hands are occupied. However, it depends mainly on line-of-sight-required sensors (e.g., cameras for vision-based and speaker/microphone for acoustic-based solutions) or specialized peripherals to collect and understand the lip language. Therefore, the usage of lip language is limited. Recently, inertial measurement units (IMUs) have been increasingly embedded among commercial off-the-shelf (COTS) earbuds (e.g., Apple AirPods). The built-in IMUs have the potential to sense silent articulatory gestures. However, these sensors present low sampling rates (of merely 25 Hz) and signal-to-noise ratios, blocking error-free lip reading. As a countermeasure, we comprehensively analyze the articulations of speech-impaired persons. Accordingly, we map inertial signals to lip language with effective noise cancellation. We propose an inertia-based lip language decoding system, named SwanTrumpet. It supports lip reading, unvoiced interaction, and user identification for speech-impaired users. In addition, we address the issue of user diversity in lip language translation, which is significant among the speech-impaired. We implement the prototype connected with two COTS earbuds. Extensive evaluations on 11 speech-impaired participants demonstrate its capability to improve the interaction experience for people lacking a voice.

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

Additional Key Words and Phrases: Lip Language Decoding, Speech-impaired, IMU Sensing, Earable Sensing

ACM Reference Format:
Ming Gao, Yichen Dai, Zhanbo Zhou, Zhengxin Guo, Jinsong Han, and Fu Xiao. 2025. The Trumpet of the Swan: Decoding Lip Language for Speech-impaired Users through COTS Earbuds. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 4, Article 175 (December 2025), 32 pages. https://doi.org/10.1145/3770648

*Fu Xiao is the corresponding author.

Authors' Contact Information: Ming Gao, Nanjing University of Posts and Telecommunications, Nanjing, China, gaomingppm@njupt.edu.cn; Yichen Dai, Nanjing University of Posts and Telecommunications, Nanjing, China, yichen.dai@njupt.edu.cn; Zhanbo Zhou, Nanjing University of Posts and Telecommunications, Nanjing, China, 1223045238@njupt.edu.cn; Zhengxin Guo, Nanjing University of Posts and Telecommunications, Nanjing, China and Tongda College of Nanjing University of Posts and Telecommunications, Yangzhou, China, guozx@njupt.edu.cn; Jinsong Han, Zhejiang University, Hangzhou, China, hanjinsong@zju.edu.cn; Fu Xiao, Nanjing University of Posts and Telecommunications, Nanjing, China, xiaof@njupt.edu.cn.

ACM 2474-9567/2025/12-ART175
https://doi.org/10.1145/3770648

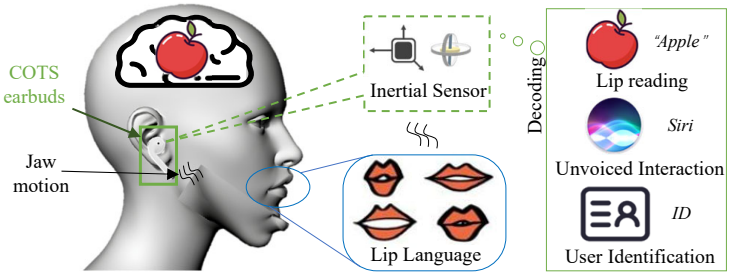


Fig. 1. The illustration of SwanTrumpet, a lip language decoder using IMUs in COTS earbuds for speech-impaired users with the capabilities of lip reading, unvoiced interaction, and user identification.

1 Introduction

There are over 1.3 million individuals suffering from organic anarthria, dysphonia, postlaryngectomy, and voice disorders [40] in China [15]. Even if with intact articulation skills (jaw and lip movements), these users are unfortunately unable to experience the convenience that voice communication and speech interaction offer, which are popular on mobile devices due to the straightforwardness, hands-free accessibility, and swift input capabilities [64]. Lip language provides a user-friendly way of communication requiring no acoustics from users. It reuses the existing human articulatory process, without additional learning cost.

It remains notoriously difficult to understand lip language. Solutions based on computer vision [56, 69] and acoustic sensing [71, 83] have achieved respectable recognition performance. However, they would degrade in non-line-of-sight (NLOS) scenarios [48, 64]. For example, the user's mouth is covered by a mask, which is common in an epidemic of infectious diseases (e.g., COVID-19) [48]. Customized peripherals overcome the issues of NLOS and hand occupation. They achieve this by attaching specialized sensors (e.g., triboelectric sensors [48], RFID [76, 86] and ultrasonic imaging sensors [43]) to the user's tongue [61], mouth [48], face [76], jaw [43], or periauricular skin [64]. However, those attached peripherals are reported to make users feel obtrusive and socially awkward [64].

The emergence of earable sensing provides a new solution for lip language translation. Users wear wireless earbuds naturally in real-world scenes. Recent earbuds (e.g., Apple AirPods Pro 2) are equipped with various sensors, e.g., in-ear microphones [13], photoplethysmography (PPG) sensors [60], and inertial measurement units (IMUs) [46]. Among them, inertial data has been reported to be capable of measuring subtle vibrations [42, 80, 88]. Consequently, IMUs (including accelerometers and gyroscopes) on earbuds have the potential to detect and recognize movements of the jaw and temporomandibular joint caused by lip language [64].

Unfortunately, the applications of inertia-based earable sensing are restricted. Existing schemes are mainly conducted on homemade prototypes, rather than commercial off-the-shelf (COTS) devices [11, 45, 46, 53, 87]. Users should pay an additional cost for these specialized peripherals or hardware modifications. People with speech impairment present a low willingness to purchase specialized devices, according to our semi-structured interview with 28 speech-impaired participants. Instead, we aim at a 'seamless integration' scheme on COTS wireless earbuds for speech-disabled users, as illustrated in Figure 1. It supports lip reading, unvoiced interaction, and user identification for speech-impaired users. However, there are three practical challenges confronting us when decoding lip language on COTS earbuds.

1) *How to leverage COTS earbuds with restricted capability?* Existing approaches [11, 45, 46, 53, 87] on lip reading typically rely on a variety of external sensors or peripherals. They design homemade prototypes and position them carefully. For example, Tanmay Srivastava et. al. [64, 65] proposed a setup involving two peripheral IMUs. One is pressed against the user's jaw to detect articulation-related signals. The other is attached to the temporal

bone as a reference to mitigate motion artifacts from the user's body and head and gravity. Instead, COTS earbuds adhere less tightly to users and are equipped with only a single IMU (even without gravity sensors). Thereby, built-in IMUs capture lip motion signals with a significantly lower signal-to-noise ratio (SNR). Furthermore, the lack of specialized reference peripherals exacerbates the SNR degradation due to motion artifacts and other external disturbances. The potential of COTS earbuds for lip language decoding has not been fully exploited.

2) *How to decode lip language using IMUs with an extremely low sampling rate of 25 Hz?* Existing studies employ IMUs but at a high sampling rate, typically exceeding 250 Hz [11, 64, 65]. However, this assumption is impractical for real-world applications. Due to power consumption constraints, COTS earbuds sample inertial data at much lower rates, such as 25 Hz in Apple AirPods [52]. It results in a low sensing band below 12.5 Hz according to the Nyquist sampling theorem. Consequently, all high-frequency features essential for accurate lip language decoding in prior works [64, 65] are entirely lost. Furthermore, the mixture of articulation-related signals at 25 Hz and motion artifacts, which are distributed below 20 Hz, is hard to separate. This further degrades the SNR in COTS earbuds. Consequently, it is challenging to characterize silent articulatory gestures using IMUs operating at such low sampling rates.

3) *How to achieve generalized application among diverse speech-impaired users?* Existing works [11, 45, 46, 53, 87] are mainly mutely conducted on users without speech disability. We observe that the articulatory gestures of people with a missing voice are individual and significantly different from those of voice-able persons. The differences and diversity are barely explored. Consequently, these methodologies can hardly be adapted for individuals with speech disability without individualized training, which demands abundant personal data. The diversity would degrade the recognition performance, along with an unsettled issue of scalability among lip language users.

To address the above challenges, we propose SwanTrumpet, a lip language decoding system. It allows people lacking a voice to converse, control voice assistants (e.g., Siri), and authenticate their silent 'voiceprint'. SwanTrumpet can be implemented on COTS earbuds, requiring no peripherals that might be unpleasant to users. Built-in accelerometers and gyroscopes on earbuds are employed to measure silent articulatory gestures. These inertial data are fused for lip language translation. For scalable applications, SwanTrumpet merely requires a quick and simple registration, in which a new user performs few-shot articulatory gestures by leveraging an encoder-decoder model. We implement SwanTrumpet on two COTS earbuds (Apple AirPods 3 and Pro 2). Different from prior works evaluated on participants who are able to voice but perform mutely during experiments, our evaluation involves 11 speech-impaired participants. Extensive evaluations demonstrate the practical value of SwanTrumpet for people with speech disorders.

Our work is specifically oriented towards the Chinese lip language. Despite its massive users, Chinese lip language has received relatively little attention in the literature [48], compared to English lip language [13, 25, 46, 60, 71, 83, 85, 89]. We built the first inertial dataset for Chinese lip language. It involves 11 speech-impaired participants, along with 26 volunteers who are able to voice but mouth mutely. The dataset is composed of over 92,000 inertial items. We have released the dataset with necessary data desensitization in accordance with relevant laws, regulations, and the IMWUT Dataset guidelines[1]. We believe that SwanTrumpet will promote the development of disability-friendly applications, especially in China. The contributions of SwanTrumpet are summarized as follows.

- We propose SwanTrumpet, a practical lip language decoder using earable IMUs. It provides a real-time service for users lacking a voice with barrier-free identification, interaction, and communication.
- SwanTrumpet realizes a non-vision solution for lip language recognition. It leverages COTS earbuds, requiring no peripheral. SwanTrumpet fully exploits the limited capability of mobile devices. To our knowledge, it is the first to implement inertia-based sensing on COTS earbuds.

[1]The open-source dataset including 26 volunteers can be found at [70] https://doi.org/10.5281/zenodo.14723896.

Table 1. Participants' attitudes and comments towards lip language via COTS earbuds

"Easy to perform phone and voice calls."
"I feel uncomfortable using specialized auxiliary tools in public. Earbuds would draw less attention from strangers."
"Access in-vehicle voice interaction during driving, e.g., for navigation and message response. Otherwise, I must stop driving to type."
"Comfortable to wear."
"Cheap and convenient."

- We suppress the impact of diversity among users and devices, along with an investigation about articulations of speech-impaired users. This promotes the few-shot scalable applicability of SwanTrumpet in the real world.

2 Background and Related Work

Lip language is a mode of expression that requires no acoustics. It relies on the existing process of human speech, eliminating the requirement to learn a new norm (e.g., sign language [30, 34, 90]). As such, lip language is useful for those who can mouth through lip movements but are unable to pronounce audibly, e.g., who have undergone laryngectomy. To understand lip language, various solutions are proposed but with practical limitations.

Vision-based techniques leverage advanced neural networks for lip reading [16, 56, 69]. However, visual methods are limited by NLOS interference from light conditions, face angles, and blocks. Moreover, the privacy issue during image/video capturing and processing raises public concern [47].

Acoustic-based lip reading methods [25, 71, 83, 85, 89] utilize the ubiquitous speakers and microphones on mobiles. These methods require the users to hold their phones towards their mouths [71, 83]. Acoustic-based solutions do not perform well in NLOS scenes, e.g., when the users' hands are occupied.

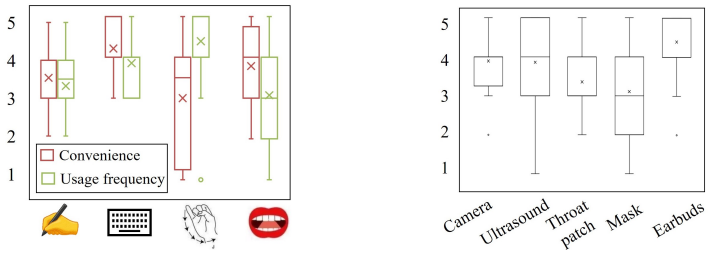
Earable-based means for lip language leverage non-invasive and contact sensors to measure the motion of muscles, e.g., in-ear microphones [13], PPG sensors [60], and IMUs [46, 64, 65]. Although with the advantages of less privacy leakage, NLOS application, and low cost, existing methods depend on specialized and customized peripherals. Few inertia-based translators could be adapted onto COTS devices.

3 Motivation and Formative Study

Before the system design, a formative study approved by the institutional review board (IRB) is performed to motivate our work. We conduct a semi-structured interview to understand practical requirements for lip language translation, particularly from users who cannot pronounce due to laryngeal and lingual injuries, dysphonia, or vocal cord lesions.

We recruit 32 speech-impaired participants (16 male and 16 female, as detailed in Appendix A) ranging in age from 16 to 60 years. These participants are certified as having Level 1 or Level 2 speech disabilities by the China Disabled Persons Federation, according to GB/T26341-2010 [51]. They are with no speech function (certified as Level 1 disability) or nearly speech-disabled with speech clarity below 25% (certified as Level 2 disability). None of the participants has additional impairments, such as hearing or brain disabilities, and all are able to articulate silently using lip movements. Their native language is Mandarin Chinese.

During the semi-structured interviews, we ask participants about their preferences regarding communication and interaction methods with humans and electronics. They are encouraged to share their experiences and requirements. Each interview lasts approximately one hour, and participants receive compensation of 200 CNY for their time.



(a) Interaction modes (b) Lip language translators

Fig. 2. Preference scores of speech-impaired users.

Following the interviews, 11 participants (6 male and 5 female) agree to participate in the data collection (with an additional compensation of 400 CNY). They are instructed to wear a pair of COTS earbuds, Apple AirPods 3. These participants mouth 2,500 common Chinese characters [66], covering the most common pronunciation in Mandarin. During this process, the built-in IMUs in the earbuds record their silent articulatory gestures.

We conclude features of speech-impaired users from the two following perspectives: preference and articulation.

3.1 Preference Analysis of Users with Speech Impairment

In the semi-structured interview, the participants rate the communication methods and expected interaction devices on a 5-point scale ranging from 1 to 5. A high score refers to the high priority in the users' subjective experiences.

Need for lip language: Common interaction methods for speech-impaired users include handwriting, typing (via keyboards or touchscreens), Chinese sign language, and lip language. The participants are asked to score the four methods from two perspectives: 1) the convenience in their opinions and 2) the usage frequency in their daily lives. The subjective preference scores are shown in Figure 2(a).

Among the current methods of communication and interaction, lip language is acknowledged as user-friendly (average convenience score: 3.79) but is used most infrequently. Participants express their desire of being 'heard'. Table 1 lists a few responses regarding the question 'advantages of lip language'. Unfortunately, lip language is not commonly utilized in practice due to its poor understandability towards audiences and machines. While participants identified typing as the most convenient mode of communication (average score: 4.21), followed by lip language, they noted that it is still not frequently employed. The long time-consuming is to blame. Moreover, participants mentioned that lip language is more convenient when their hands are occupied. In contrast, sign language is perceived as the least convenient option (average score: 2.96), because it requires both users and their surroundings to learn a new mode of communication from scratch. On average, users typically spend about 25 months mastering sign language.

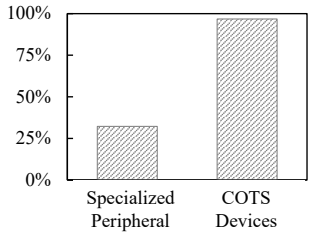


Fig. 3. Purchase willingness of lip reading devices.

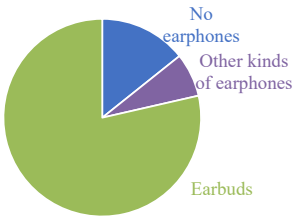


Fig. 4. Earphone ownership ratio.

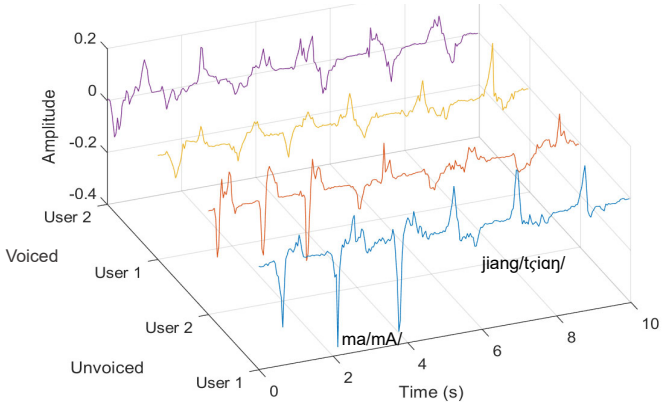


Fig. 5. Articulation-related IMU signals in Apple AirPods 3 (two characters, three repetitions each) from two speech-impaired and voice-able participants. The signal differences illustrate that individuals articulate the same syllable in distinct ways.

Device preference: The participants score five models of lip-reading devices in the literature including cameras [69], speakers [83], throat patches [43, 76], masks [48], and earables, after a detailed introduction with illustrations. Among these, earables are the most acceptable with the highest scores in Figure 2(b). However, as shown in Figure 3, there is a low willingness to purchase specialized peripherals. Less than one-third of participants express interest in the specialized mask in [48] and the homemade earable in [64, 65]). This reluctance stems from concerns regarding various factors, such as cost, convenience, learning curve (particularly for older individuals), wearing comfort, attention-drawing appearance, and compatibility with smartphone applications. Instead, participants prefer to reutilize COTS devices if they would be capable of lip reading, due to the low cost and easy access. Only one participant is concerned that COTS devices might not perform well enough. In particular, earbuds are popular among speech-impaired users. As shown in Figure 4, 87.5% of participants wear earphones frequently, and 81.3% possess wireless earbuds. Thus, lip reading using COTS earbuds is generally acceptable for speech-impaired users.

3.2 Insight about Articulatory Characteristics

We investigate the articulatory features presented by speech-disabled users in terms of stability and differentiation. The collection process and setup are detailed in Section 8.1, and source data are included in our dataset [70].

Stability: Users generally articulate the same syllables in the same manner. To make a comparative analysis of silent articulatory gestures between participants with versus without speech impairments, we recruit 2 voice-able volunteers (22~23 years old, 1 male and 1 female) to articulate the same characters silently. Figure 5 illustrates the inertial signals for two syllables mouthed by two speech-impaired and two voice-able participants. Distinct articulatory gestures produce uniquely identifiable waveforms. On the contrary, the waveforms for the same syllables remain similar, with a dynamic time warping (DTW) distance [29] of 0.442 on average. The consistency in articulating identical words and distinctiveness among different words facilitate lip reading, e.g., through comparison with pre-established templates.

Diversity: Intuitively, each individual possesses unique articulatory gestures due to personal muscle stretching habits. This diversity is even more significant among speech-impaired people. As shown in Figure 5, the variations among these signals from different participants demonstrate the diversity in individual articulation patterns when producing the same syllable. The average DTW correlation coefficient between the two speech-impaired participants is 1.113, well higher than that between voice-able (0.808). Addressing user diversity remains an open issue in lip language translation and decoding.

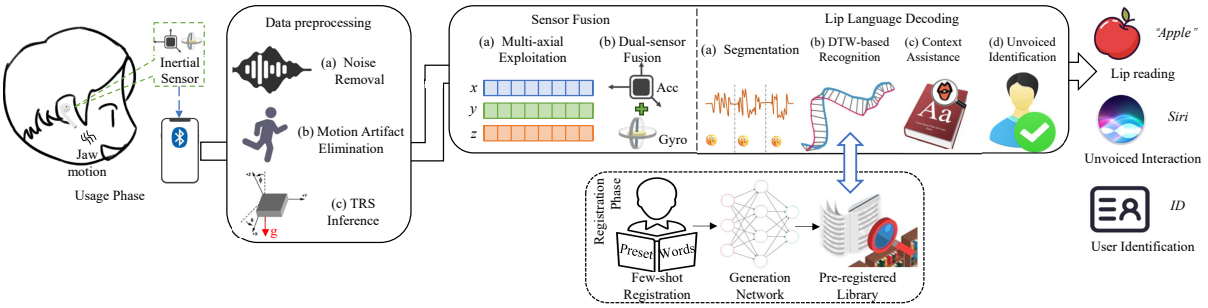


Fig. 6. Workflow of SwanTrumpet. It leverages COTS earables equipped with inertial sensors to decode lip language for speech-impaired users. Inertial data are first preprocessed for SNR improvement, then fused for bandwidth expansion, and finally decoded for various applications, including lip reading, unvoiced interaction, and identification. Besides, in the registration phase, users simply need to read a few preset words, followed by an encoder-based generation network to construct a reference library for DTW-based recognition.

4 System Overview

Motivated by the practical needs of Chinese users with speech impairments, we design a lip language decoding system using COTS earbuds. An overview of the system is illustrated in Figure 6.

To address the challenge of sensor limitations inherent in COTS earbuds, we compare their capabilities with those of homemade prototypes equipped with redundant sensors. Through a comprehensive noise analysis, we propose solutions to improve the SNR in COTS earbuds in Section 5.

To address the challenge of low sampling rates, we fuse multi-axial data from gyroscopes and accelerometers in the COTS earbuds in Section 6. This data fusion technique effectively expands the equivalent bandwidth of the earable inertial sensing system. The fused data then undergoes decoding for various applications, including lip reading, unvoiced interaction, and user identification. Specifically, we recognize segmented lip language using DTW-based approaches and refine character selections according to contextual information, while simultaneously supporting user identification functionality.

To address the challenge of user diversity, we employ an encoder-decoder model to transform articulation-related inertial data from one user to the unique characteristics of another in Section 7. This approach enables us to generate sufficient personalized inertial templates for individual lip language recognition. Therefore, users are only required to perform a few-shot registration before usage.

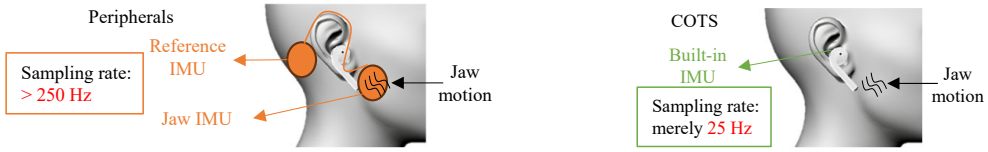
5 Noise-free Exploitation of Earable Inertial Signals

Although preferred by speech-impaired users, COTS earbuds are typically equipped with limited sensors. In this section, we explore the potential of built-in IMUs in COTS devices for noise-free signals.

5.1 Homemade Prototypes vs. COTS: Sensor Limitations

Before introducing our design, we compare COTS earbuds with homemade specialized prototypes used in existing studies about inertial-based lip reading [11, 64, 65]. This reveals a significant limitation: COTS devices' fewer sensors restrict access to information that prior works rely upon. Figure 7 makes an illustrated comparison between the homemade earable prototypes in existing works [64, 65] and COTS earbuds (Apple AirPods) when being worn.

Existing methods with redundant sensors. Previous research has demonstrated that IMUs are able to measure articulatory gestures [11, 64, 65]. However, these studies rely on ideal conditions using homemade



(a) Specialized prototypes with position-given peripherals (b) COTS earbuds with only one IMU

Fig. 7. Homemade earable prototypes in existing works [64, 65] *vs.* COTS earbuds (Apple AirPods).

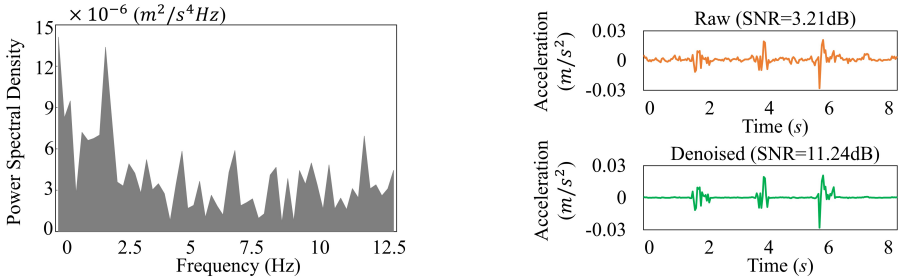
specialized prototypes, e.g., equipped with two pairs of IMUs [64, 65]. In these setups, one IMU is attached directly to the temporomandibular joint for measuring jaw motion. The other is pressed against the user's temporal bone to serve as a reference. It measures motion artifacts from the user's body and head and gravity. This configuration enables noise-free articulation-related signal collection, not to mention the much higher sampling rates than COTS (as detailed in Section 6).

COTS earbuds. In comparison, these additional components are rare in COTS earbuds. As a representative, Apple AirPods carry only one IMU in each earbud. Due to the lack of additional reference sensors, COTS earbuds suffer from motion artifacts. Even worse, COTS earbuds are not equipped with a gravity sensor (in order to reduce cost and size). In this case, we cannot transform inertial data into a general coordinate system, e.g., the terrestrial reference system (TRS), when users wear earbuds under different positions. It seems that the capacity of COTS earbuds presents little capacity for inertial-based lip language under the interference of ambient noises, motion artifacts, and TRS absence.

5.2 Noise Analysis and Removal

IMUs embedded in COTS earbuds are vulnerable to noise interference. There are three main categories of noises, including intrinsic noises in sensors, interference from earbuds' speakers, and motion artifacts. Here, we analyze the former two noises and correspondingly propose solutions. Motion artifacts will be solved in Section 5.3.

5.2.1 Intrinsic Noise. Due to hardware imperfections, IMUs suffer from intrinsic noise, which results in a low SNR in earable sensing. Intrinsic noise typically follows a stationary distribution, i.e., a direct-current (DC) bias that is constant or near-constant and an additive white noise [3]. Based on this insight, we employ a Wiener filter [79] after migrating the DC bias. The Wiener filter allows for effective cancellation of the noise when its distribution is known [79]. We can estimate the intrinsic noise distribution by gathering inertial readings from the earbuds in advance when they are stationary. The data refer to our open-source dataset [70]. We conduct the



(a) Energy distribution (b) Example of denoising

Fig. 8. Intrinsic noises in earable IMUs and an example of our denoising method.

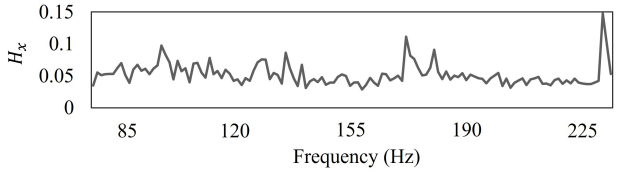


Fig. 9. Frequency response within the human voice's fundamental band in Apple AirPods 3.

Wiener filtering on a pair of Apple AirPods 3, with its noise energy distribution shown in Figure 8(a). There is an increase in SNRs by over 7 dB, with an example of the x-axis in the on-board accelerometer in Figure 8(b).

5.2.2 Interference from Earbuds' Speakers. It has been reported that IMUs, especially accelerometers, are sensitive to audio played by speakers [4, 5, 9, 24, 50, 67, 68, 77, 84]. Due to the physical contact, the vibrations generated by the earphone's speaker during audio playback inevitably impact the inertial measurements.

We perform a quantitative analysis on the interference caused by the earbuds' speakers. Insecure low-pass filters fail to cancel these out-band noises [73]. According to the Nyquist sampling theorem, audio from the earbuds, with sampling rates much higher than that of the inertial sensor (Fs=25 Hz), is downsampled as follows,

$$IES_i(t) = H_i(f)A_0 sin(2\pi||f_0 - n \times Fs|| + \phi_0),\ (n \in \mathbb{N}) \tag{1}$$

where $IES_i(t)$ $(i = x, y, z)$ represents the downsampled noise along the i-axis of the inertial sensor, with the corresponding transfer function $H_i(f)$, and A_0, f_0, and ϕ_0 denote the amplitude, frequency, and phase of the audio, respectively. We observe that the transfer function $H_i(f)$ is consistent for a given sensor, with an example in Figure 9.

Based on this insight, we employ an adaptive noise filter [21] to estimate clean articulation-related signals in real time by removing interference from the speaker. Specifically, we pre-characterize the transfer function $H_i(f)$ of the user's earphones through chirp signal excitation before system deployment. Given the acoustic signals $a(t)$ played by the earphone speakers, which can be accessed and recorded via the `ReplayKit` API, we compute the corresponding downsampled signals $IES_i^a(t)$ according to Equation 1. Next, we apply an adaptive noise filter [21]. For the mixed inertial signal $M^a[n]$ $(n \in N)$, we can obtain the articulation-related signal $C[n] = M^a[n] \ominus IES_i^a[n]$, where $\ominus$ denotes the adaptive noise filtering operation. The filter coefficients $W[n]$ are updated iteratively based on the normalized least mean square (NLMS) algorithm [32], as follows,

$$W[n+1] = W[n] + \mu \times (M^a[n] - C[n]) \times IES_i^a[n], \tag{2}$$

where μ is the learning rate.

In particular, IMUs are sensitive only to sound transmitted through solid media (from earphone speakers here), rather than through air [4]. Thus, even in noisy environments, IMU readings from the jaw remain unaffected, neither by airborne noise nor by noise transmitted via the air-muscle-IMU pathway. Experimental results in Section 8.5.1 confirm that SwanTrumpet is immune to ambient noise interference.

5.3 Motion Artifact Elimination

Following the common sense [11, 45, 46], we recognize the movement of the user's body and head as interference in inertial-based earable sensing. A common approach utilizes a high-pass filter with a cutoff frequency exceeding 20 Hz [9, 11, 24, 65], as motion artifacts predominantly occur in the low-frequency band. However, this approach is ineffective in COTS earbuds. The low sampling rate of the built-in IMU results in a narrow bandwidth of merely 12.5 Hz according to the Nyquist sampling theorem. In this case, the approach of high-pass filtering would also remove the articulation-related signals. To mitigate the impact of user motions, we leverage a deep regression

model for artifact elimination under different conditions. Besides, we take the non-language lip movement into consideration.

To remove motion artifacts, we propose a deep regression model [62]. Its purpose is to separate the articulation-related inertial data from the noisy measurements distorted by human movements. The model consists of two fully connected layers and a regression layer. The model takes as input the spectrogram of inertial signals contaminated by motion artifacts, i.e., data recorded while users articulate during body movements and processed by the short-time Fourier transform (STFT). It outputs the spectrogram of clean articulation-related signals with motion artifacts removed. Subsequently, the Griffin-Lim algorithm [28] is applied to reconstruct the time-domain inertial signals from the predicted spectrogram. In the training phase, we independently collect inertial data from lip movements (with the users' bodies stationary) and common user activities (such as walking, bicycling, driving, head nodding, and head shaking), and then manually mix these signals. This process generates paired datasets of artifact-corrupted and ground-truth signals for training the regression model. This approach enables the efficient and low-cost generation of large-scale training data. In this way, we suppress the influence of the user's movement in an earable-only manner.

In practical applications, earable IMUs are likely to capture lip movements with no semantic information, e.g., mastication, brushing teeth, yawn, and cough. If these movements are mistakenly identified as lip language, the user experience would degrade with a high false positive rate. We explore the distinction between signals that contain semantic information and those that do not. Non-verbal lip movements can be classified into two categories: continuous movements, exemplified by mastication, and sudden movements, represented by yawning and coughing. We remove mastication-represented continuous movements using the above deep regression model. Moreover, users can turn off SwanTrumpet during continuous movement to avoid unnecessary errors. As for sudden movements, they typically occur at longer intervals than lip language. Consequently, we differentiate them based on these intervals.

5.4 TRS Inference without Gravity Sensors

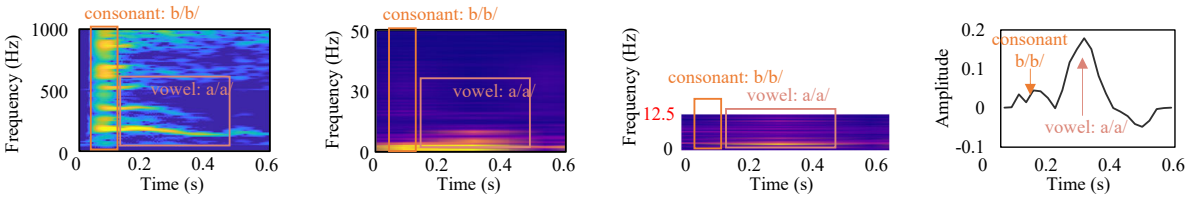
In practice, users are likely to wear earbuds in slightly different positions. To mitigate the influence of varying wearing positions, it is essential to transform the earable inertial data into a general coordinate system, typically the terrestrial reference system (TRS). Traditionally, a gravity sensor is necessary to infer the TRS, while it is absent in COTS earbuds. To solve this issue, we propose a method for TRS inference utilizing IMUs from a pair of earbuds.

In a pair of COTS earbuds, each bud is equipped with an IMU. The pair enable us to leverage their correlation to estimate gravity. We denote a_t^i, ω_t^i, v_t^i, and g_t^i as the acceleration, angular velocity, egovelocity, and gravity at the time point t in the coordinate of each earbud $i(= L, R)$ respectively. The sampling rate is denoted as Fs. We establish the egovelocity interrelationship as follows,

$$v_t^L = f(v_t^R, \omega_t^R) = M^{L \leftarrow R}(v_t^R + \omega_t^R \times V^{L \leftarrow R}), \tag{3}$$

where the rotation matrix $M^{L \leftarrow R}$ and the rotation vector $V^{L \leftarrow R}$ characterize the relative position between the IMU coordinates in the pair of earbuds. The two parameters can be calculated based on the rotation matrix of each earbud to the charging case, which is recorded by the application programming interface (API) `CMHeadphoneMotionManager.DeviceMotion Handler`. The approximate relationship [20] between the egovelocity and gravity is

$$f(v_t^R, \omega_t^R) - f(v_{t-\Delta t}^R, \omega_{t-\Delta t}^R) = \frac{1}{Fs} \sum_{k}^{Fs \cdot \Delta t} (a_{t-\frac{k}{Fs}}^R - g_{t-\frac{k}{Fs}}^R) \approx \frac{1}{Fs} \sum_{k}^{Fs \cdot \Delta t} (a_{t-\frac{k}{Fs}}^R - M_{t-\frac{k}{Fs} \to t}^R g_t^R) \tag{4}$$



(a) Audio spectrogram (b) Inertial spectrogram sampled at 100 Hz (c) Inertial spectrogram sampled at 25 Hz (d) Inertial signal sampled at 25 Hz

Fig. 10. Comparison of (a) audio, (b) earable inertial signals sampled at 100 Hz in prototypes and (c&d) inertial signals at 25 Hz in COTS earbuds.

where $M^R_{t-\frac{k}{Fs}\to t}$ is the rotation matrix from the time stamp $t-\frac{k}{Fs}$ to t during the time interval Δt. Accordingly, we can estimate gravity in the earbud R as follows,

$$g^R_t \approx [\sum_{k}^{Fs\cdot\Delta t} M^R_{t-\frac{k}{Fs}\to t}]^{-1}[\sum_{k}^{Fs\cdot\Delta t} a^R_{t-\frac{k}{Fs}} - Fs\cdot(f(v^R_t,\omega^R_t) - f(v^R_{t-\Delta t},\omega^R_{t-\Delta t}))] \tag{5}$$

Similarly, we can derive the gravity g^L_t for the other earbud. Our experimental results show that this method achieves high accuracy. The mean error in the gravity estimation is 0.028°, with its standard deviation of 0.055°. In this way, we normalize earable inertial data into the TRS, without requiring a dedicated gravity sensor. Therefore, SwanTrumpet is able to be robust to variations in how users wear the device.

Combining the above efforts together, we obtain error-free articulation-related inertial data from COTS earbuds.

6 Vocabulary-unrestricted Lip Language Decoding

In COTS earbuds, IMUs are typically sampled at an extremely low rate. This results in a limited bandwidth for inertial sensing. In this section, we utilize comprehensively built-in 3-axis accelerometers and gyroscopes to expand the equivalent bandwidth. Subsequently, we realize a lip language decoder with capabilities of vocabulary-unrestricted lip reading, mute interaction, and unvoiced identification.

6.1 Narrow Bandwidth of Earable Inertial Sensing

Prior research has demonstrated that earable inertial data can enable voice inference [11, 62] and lip reading [64, 65]. However, these studies rely on a high sampling rate, e.g., over 250 Hz. Thus, these approaches cannot be applied to COTS earbuds represented by Apple AirPods, whose sampling rates are limited at 25 Hz [52]. According to the Nyquist sampling theorem, the sensing bandwidth is restricted within just 12.5 Hz. This results in the loss of high-frequency articulatory gesture characteristics. Figure 10 illustrates how different sampling rates affect characterization. While articulatory gestures contain frequencies up to 50 Hz (e.g., the consonant /b/), these high-frequency components become distorted in COTS earbuds. Nevertheless, the fundamental components remain detectable in COTS earbuds, as shown in Figure 10(d). This observation evidences the feasibility of using COTS earbuds. However, it is unsettled to decode lip language effectively given the distortion caused by the low sampling rate.

6.2 Sensor Fusion for Equivalent Bandwidth Expansion

As a countermeasure, we exploit the built-in 3-axis accelerometers and gyroscopes together, to maximize the potential of COTS earbuds with limited sampling rates. The basic idea is illustrated in Figure 11.

6.2.1 Sensor Selection. In current researches, there is often an exclusive choice between using gyroscopes and accelerometers, rather than employing both simultaneously [9, 11, 35, 41, 62]. The majority favor accelerometers due to the common belief that these vibration-sensitive sensors have a stronger correlation with articulation-related movements [9, 11, 35, 62]. A recent study suggests that gyroscopes may be more effective in disability services due to their lower intrinsic noise levels than accelerometers [41]. Instead, we utilize both sensors. Benefiting from the Wiener filter in Section 5.2.1, we eliminate the intrinsic noise in accelerometers effectively. Moreover, we observe that gyroscopes are also able to capture features associated with articulatory gestures. Our insights present a promising opportunity to fuse sensing data with low sampling rates for improved lip language translation.

6.2.2 Multi-axis Utilization. Three axes in a sensor describe the three-dimensional movement of articulation. In this case, prior processing methods are defective. It would lead to incomplete characterization of articulatory gestures if focusing exclusively on one axis while neglecting the others [5, 9]. The L2-Norm-based approach [24, 82] is also commonly used for dimensionality reduction, while it sacrifices spatial information of the three-dimensional movement.

We propose a multi-axis utilization method to expand the equivalent bandwidth without any loss of spatial information. The multi-axis data is rearranged into a one-dimensional format, following the order illustrated in Figure 11(b). Taking the accelerometer as an example, data from different axes $a_i[n](i = 0, 1, 2)$ are interwoven as

$$a^{\ddagger}[n] = a_{n \ mod \ 3}[\lfloor \frac{n}{3} \rfloor], \ (n \in N) \tag{6}$$

where *mod* is the modulo operation, and $\lfloor \cdot \rfloor$ is the floor function. This method adds equivalent samples to 75 Hz in a lossless manner. In the equivalent bandwidth of 37.5 Hz, the high-frequency components reflect the interrelationship among the multiple axes. Similarly, we can obtain a super-resolution gyroscope data $g^{\ddagger}[n]$.

6.2.3 Coherence-based Fusion. The earable accelerometer and gyroscope serve as separate yet coherent observers of individual silent articulatory gestures. Based on this property, we proposed a coherence-based approach to fuse the two IMUs with different modalities. We opt not to adopt the Kalman filter [39] here, despite its widespread use as a motion fusion algorithm in robotics and kinematics [59], due to its unsatisfactory experimental performance in our application. This limitation is likely attributable to the absence of a magnetometer in COTS earbuds, which blocks effective yaw correction. Without a magnetometer-based heading reference, errors in gyroscope measurements tend to accumulate over time, resulting in drift and degraded state estimation accuracy. As a potential improvement, the unscented Kalman filter (UKF) [12] could be employed to compensate for yaw error accumulation in earable gyroscopes.

Our proposed method emphasizes distinguishing features relevant to silent articulation through point-wise multiplication. Given the coherence between the accelerometer and gyroscope, the articulation-related characteristics captured by both sensors are tightly synchronized, exhibiting high intensity simultaneously. Therefore, a point-wise multiplier could fuse the two sources into a one-dimensional characteristic vector, i.e. $a^{\ddagger}[n] \odot g^{\ddagger}[n]$. Generally speaking, the product's maximum frequency is double the multiplier's [23]. Benefiting from the three-fold bandwidth expansion in Section 6.2.2, the product is free from distortion. The T-F spectrum in Figure 11 demonstrates the effectiveness of our proposed sensor fusion method for equivalent bandwidth expansion. This coherence-based fusion method is computationally efficient, which facilitates real-time processing capabilities.

6.3 Lip Language-based Recognition, Interaction, and Identification

We identify syllables in segmented fused inertial data according to DTW distances from users' pre-registered reference data to translate lip language. We first introduce a language-agnostic decoding framework. Next, we describe Chinese-specific adaptations by leveraging Chinese phonetics.

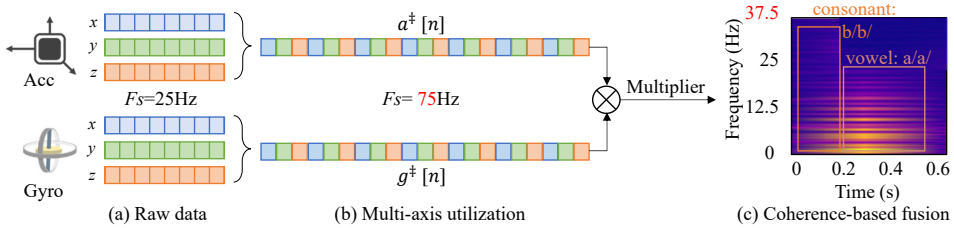


Fig. 11. Illustration of the sensor fusion approach for equivalent bandwidth expansion.

6.3.1 Segmentation. We propose a lightweight approach for segmenting syllables (characters) in inertial data.

First, we detect articulation-related signals using the widely-adopted double-threshold scheme [22]. Continuously streaming and processing data from the earables for lip reading can be resource-intensive. To minimize power consumption and computational costs, it is essential to pinpoint the lip moments when a user is likely performing an articulatory gesture. Our observation shows that during periods of silent articulation, there is a correlation between accelerometer and gyroscope readings, characterized by their simultaneous high intensity, whereas non-speech periods generally exhibit low intensity. We implement the maximum between-class variance (OTSU) algorithm [55] to calculate thresholds for each sensor independently. The regions where both sensor readings exceed their respective thresholds are identified to be relevant to articulations.

Second, we extend each identified region's start and end points by 2 samples forward and backward, respectively. This segmentation strategy ensures coverage of the articulatory gestures associated with the entire syllable.

Third, we segment articulatory gestures into vowels and consonants. From a phonetic perspective [36, 78], consonant and vowel articulation exhibit markedly different jaw movement characteristics. Consonants typically involve more subtle, smaller-range oral movements. Many consonants (such as plosives and fricatives) primarily depend on the intricate coordination of the tongue, lips, and anterior oral structures, with relatively minimal jaw displacement. Consequently, the corresponding inertial data demonstrate low amplitude. In contrast, vowel production necessitates a more substantial jaw opening to generate characteristic vocal tract resonances. When articulating vowels, users' lips perform more pronounced movements [64]. This causes significant inertial data to be captured by earbuds. These larger movements produce peaks in inertial sensor readings, especially in the gyroscope data. To identify vowels, we first determine the axis exhibiting the greatest rotational variation (usually the z-axis). We then detect the peak that exceeds the previously defined OTSU threshold and designate the zero-crossing points immediately before and after this peak as the vowel boundaries. After detecting the vowels, the remaining intervals are classified as consonants.

6.3.2 DTW-based Recognition. Instead of existing methods that demand extensive data for training a classifier [48, 64, 83], we utilize Fast-DTW [54] to align the fused data with a pre-registered library of reference data. This library is built during a registration phase, in which the user articulates either a large vocabulary or just a few words/characters. In the latter case, extensive reference data are generated using the method described in Section 7.

Fast-DTW operates with linear time complexity, making it appropriate for real-time implementation on mobile devices. We rank the potential articulatory gestures based on their DTW distance, where a lower distance indicates a higher similarity between the reference and the gesture. In particular, consonants are more challenging to detect and recognize. Consonant articulation is rapid and tends to have low intensity. Moreover, the articulation is easily affected and distorted by the surrounding vowels. As a result, directly matching consonants with pre-registered templates can be difficult. To make matters worse, several consonants produce similar lip movements. In this case, it is indistinguishable among consonants based solely on inertial readings [64]. These indistinguishable

consonants are named visemes [76]. The DTW distances are calculated sequentially from vowels or visemes to their respective reference signals.

6.3.3 Context-assisted Interaction. We leverage the contextual information to correct wrongly recognized syllables for interaction and communication. We denote the transition probability as $P(I, G_i|F)$, where I represents a consonant under the articulation of $G_i (i = 1, 2, ..., 7)$ and F represents a vowel. We use the Bayesian Equation to calculate the transition probabilities,

$$P(I, G_i|F) = \frac{N(I, G_i|F)}{\sum N(I, G_i|F)} \tag{7}$$

where $N(I, G_i|F)$ is denoted as the numbers of a consonant I under the articulation of G_i followed by the vowel F, according to statistics from Sketch Engine [63].

In this way, we are capable of continuously translating lip language sentences based on the smallest DTW distance with the highest probabilities. Besides, a recent work [65] reports that the Large Language Models (LLMs) can also be utilized for syllable correction. In this way, SwanTrumpet translates lip language to speech or text, which enables voice communication for speech-impaired users and interaction with voice assistance (VA).

6.3.4 Unvoiced Identification. As an essential requirement, identification ensures security during silent interactions and commands [46]. We observe that each individual articulates uniquely when mouthing the same words, as illustrated in Figure 5. Inspired by this, we explore earable inertial data for biometric extraction. When interaction wakewords (e.g., "Siri") are detected, SwanTrumpet utilizes a three-layer convolutional neural network (CNN) for user authentication. To be specific, we first transform the fused data $a^{\ddagger}[n] \odot g^{\ddagger}[n]$ (corresponding to the wakeword) into the time-frequency (T-F) spectrum. Then, the spectrum acts as the input of three convolutional layers, with the rectified linear unit (ReLU) activation. Each convolutional layer is followed by a max pooling layer. The convolutional output passes through a flatten layer, a fully connected layer, and a softmax function to project into identification.

6.4 Chinese-specific Adaptation

In China, over 1.3 million individuals live with speech impairments [15]. Despite this significant population, research on Chinese lip language has received relatively little attention in the literature [48], compared to English lip language [13, 25, 46, 60, 71, 83, 85, 89]. Chinese has distinctive phonetic components and tonal features. Investigating these distinctive linguistic features would benefit lip language recognition techniques specific to Chinese.

6.4.1 Chinese Phonetics. The pronunciation of each Chinese character (namely Pinyin) is single-syllable. It consists of two components: initials and finals [36, 78], as detailed in Appendix B. Mandarin Chinese consists of 21 consonant initials and one zero initial, along with 39 finals. This structure simplifies Chinese lip-reading into a two-step process: identifying the initial consonants and the finals.

The initials and finals exhibit distinct characteristics. Figure 10(a) shows a syllable 'ba/ba/' (gathered from a voice-able volunteer), in which the syllable is composed of the initial (consonant) 'b/b/' and the final (monophthong) 'a/A/'.

In addition, modern Standard Chinese features four distinct lexical tones. Identical phonetic forms differing in tone correspond to different characters. We experimentally observe that tones have little impact on lip movement during articulation. As illustrated in Figure 12, the inertial data for different tones exhibit similar patterns. The average DTW distances between tones are below 0.55. This similarity makes it difficult to distinguish tones solely based on inertial data. As a countermeasure, we utilize contextual information to infer the tone and thus improve lip language recognition.

Based on the distinctive phonetics, we refine the recognition process in Section 6.3 for Chinese as follows.

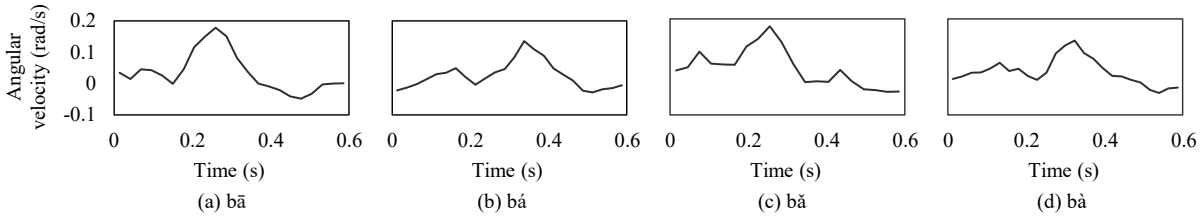


Fig. 12. Chinese phonetics has four tones. They differ insignificantly in inertial data.

6.4.2 Pinyin Segmentation. We adjust the syllable segmentation into initials and finals, instead of vowels and consonants. According to the Chinese linguistic structure, the initial consists solely of consonants, while the finals begin with vowels necessarily [36, 78]. By detecting the most prominent peak, we designate the zero-crossing point preceding it as the cutting point. This point demarcates the initial before it and the final after it.

6.4.3 Pinyin Recognition. In the DTW-based approach, the Pinyin recognition is divided into three steps.

First, we detect specific special Pinyin cases involving certain particular collocations, decreasing the computational load. Several finals only follow the given initials. Specifically, the monophthong '-i/ɻ/' is found only after the initials 'z/ts/', 'c/ts^{h}/', and 's/s/'. The monophthong '-i/ʅ/' appears solely after the initials 'zh/tʂ/', 'ch/tʂ^{h}/', 'sh/ʂ/', and 'r/ʒ/'. The monophthong 'er/ɭ/' is merely used independently, i.e., following the zero initial. Moreover, the monophthong '$\hat{e}$/ɛ/' is exclusively observed in 'ye/iɛ/'. These cases are found out according to reference [38, 78].

Next, we focus on identifying finals that contain vowels. Compared to the initials, the finals are characterized by their long duration and high intensity. We observe that each final possesses a distinct articulation. Consequently, we match the final segmentation against the referred final data (totally 32 templates as listed in Table 5, excluding '-i/ɻ/', '-i/ʅ/', 'er/ɭ/', and '$\hat{e}$/ɛ/'). In particular, the end (empirically the last quarter) of an articulatory gesture is likely to be influenced, distorted, and overlapped by the subsequent syllables [78]. Therefore, we primarily compare the DTW distance between the former in an articulatory gesture with that of the reference signals. Here, we select the three finals with the shortest DTW distances for the following steps.

Lastly, we recognize the initials in terms of visemes [76]. We list the viseme groups in Mandarin Chinese in Table 6 in Appendix B. To address the above issues, we recognize entire syllables using prior knowledge of the identified finals based on the Bayesian Equation (as detailed in Section 6.3.3). Thus, we distinguish initials independently.

Benefiting from the above approaches, the computational workload is significantly reduced from 792(=22×36) DTW distance calculations to just 107(=9+32+22×3). This reduction efficiently enhances real-time performance on mobile devices for lip language translation.

6.4.4 Chinese Character Correction. To describe contextual Chinese linguistics, we refine the Bayesian Equation.

There are three constraints to consider in Chinese linguistic and contextual analysis. First, not all combinations of initials and finals are possible in Chinese. For example, the final 'ia' is exclusively found after a select group of initials such as 'd', 'j', 'l', 'q', and 'x', resulting in a conditional probability where $\sum_{I\in d,j,l,q,x} P(I, G_i|ia)$=1. Second, the frequency of syllable usage in Chinese is unevenly distributed. For instance, the syllable 'dui/duei/' is much more common than 'tui/d^{h}uei/', while the syllables 'nui/nuei/' and 'lui/luei/' are not used in Chinese, though the four articulatory gestures are similar. Third, Mandarin Chinese frequently employs phrases rather than individual characters. Correspondingly, the context would pose an impact on the transmission probabilities. In this case, we rewrite the Bayesian Equation 7 to describe the contextual relationship as follows,

$$P(I, G_i|F, C_{forward,back}) = \frac{N(I, G_i|F, C_{forward,back})}{\sum_I N(I, G_i|F, C_{forward,back})} \tag{8}$$

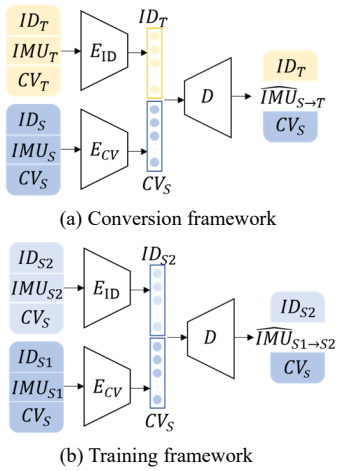


Fig. 13. Illustration of generation frameworks.

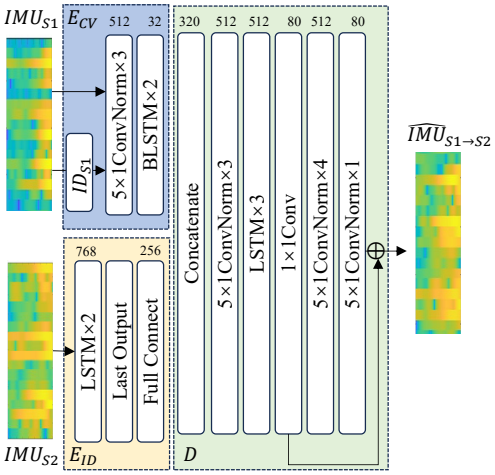


Fig. 14. Encoder-decoder model architecture.

where $N(I, G_i|F, C_{forward,back})$ represents the numbers of an initial I following the character $C_{forward}$ or followed by the character $C_{backward}$. For example, the probability of 'tui/d^{h}uei/' would exceed that of 'dui/duei/' if followed by 'chu/ts^{h}u/' though $P(t, G_7|ui)$<$P(d, G_7|ui)$.

In short, we decode lip language by recognizing, understanding, and identifying silent articulatory gestures just using inertial data on COTS earbuds, despite the limitation of low sampling rates.

7 Few-shot Scalability Study for Diverse Speech-impaired Users

Generalization across various users and devices blocks the practice of applying SwanTrumpet. It requires a substantial dataset comprising a wide range of scenarios. Instead, we aim at a few-shot solution, in which a new user merely performs a simple registration.

To overcome user diversity, an intuitive way is to ask a new user to mouth every syllable (character) as pre-registered templates. However, this method is impractical due to the significant time consumption (approximately dozens of minutes). A straightforward compromise employs average data from other users as templates. Unfortunately, this approach yields poor results because of the unique articulatory characteristics of each individual, especially those with speech impairments. A Chinese-focused approach is to mouth each initial and final separately, then manually combine them as templates. This is infeasible because these artifacts differ quite significantly from the actual lip motion, in which the articulatory gestures of initials are affected (distorted) by the subsequent finals. Their DTW distance could reach up to 1.013. Even worse, in our interview, the speech-impaired participants are reported to have nearly forgotten how to articulate initials and finals independently from characters. To this end, we propose a data generation method based on a limited set of inertial data during the registration phase.

Inspired by the voice conversion techniques [44, 58] in natural language processing, we exploit the encoder-decoder model [72]. The approach is adopted to modify given inertial data related to silent articulations from source users to match the unique characteristics of a target user (with speech disorders) to be registered. Therefore, we can generate sufficient inertial templates for individual lip language.

We consider two sets of independent and identically distributed (iid.) variables (ID_S, CV_S, IMU_S) collected from a source user and (ID_T, CV_T, IMU_T) from a target speech-impaired user. Here, IMU represents articulation-related inertial readings, from which we can extract the user identity characteristic ID and the content vector CV. The

expression (ID, CV, IMU) can be regarded as a series of processes randomly sampled from the inertial data distribution $p_{IMU}(\cdot|ID, CV)$. As shown in Figure 13(a), our objective is to design a converter $\hat{IMU}_{S\to T}$ that maintains the content CV_S while aligning with the user characteristics of the target user ID_T. An ideal converter can be formulated as follows,

$$p_{\hat{IMU}_{S\to T}}(\cdot|ID_T, CV_S) = p_{IMU}(\cdot|ID_T, CV_S) \tag{9}$$

To achieve this objective, we employ the autoencoder [91] framework. It comprises three modules: an ID encoder E_{ID}, a content encoder E_{CV}, and a decoder D. In detail, the ID encoder E_{ID} extracts the identity characteristic ID_T from the inertial signals IMU_T collected from the target user ID_T. The content encoder E_{CV} takes the inertial signals IMU_S collected from the source user and outputs the contained content vector CV_S. The decoder D integrates the identity ID_T and the content CV_S to generate the inertial data $\hat{IMU}_{S\to T}$ in which the target user articulates the source content.

To train this converter, we utilize the datasets we collected as the training set, following the framework illustrated in Figure 13(b). Our training employs parallel data in which two source users (ID_{S1} and ID_{S2}) mouth the same content (CV_S). The model architecture is depicted in Figure 14. The ID encoder E_{ID} is independently trained on the GE2E loss following the design in [75]. The content encoder E_{CV} constructs the information bottleneck from the concatenated features of the spectrogram IMU_{S1} and the ID feature ID_{S1} (extracted by the pre-trained E_{ID}). The decoder follows AutoVC [58] under the loss function,

$$\min_{E_{CV},D} L = \mathbb{E}[||\hat{IMU}_{S1\to S2} - IMU_{S2}||_2^2] - \lambda\mathbb{E}[||E_{CV}(\hat{IMU}_{S1\to S2}) - E_{CV}(IMU_{S2})||_1]. \tag{10}$$

Figure 14 demonstrates the similarity between an authentic spectrogram IMU_{S2} and a generated one $\hat{IMU}_{S1\to S2}$. Then the Griffin-Lim algorithm [28] is leveraged to estimate the inertial signals from spectrograms. Because training data from speech-impaired users is inadequate, voice-able participants also serve as the sources ID_s for training. Empirically, new users are asked to silently 'read' a sequence of 30 characters (as listed in Appendix C) in succession in the registration phase. Such a registration requirement is acceptable for speech-impaired users according to the semi-structured interview. The corresponding inertial data are exploited to fine-tune the trained model, and serve as IMU_T for signal generation. The registration duration lasts around 30 seconds, which is acceptable in practice. We recruit 26 voice-able volunteers (21∼49 years old, 13 males and 13 females) to articulate the 2,500 common Chinese characters [66] silently, as training data for the encoder-decoder model. They are collected on Apple AirPods Pro 2 paired to an iPhone 13 mini (iOS 17.6.1). Our approach enables few-shot generation of inertial templates for lip language, allowing SwanTrumpet to be scalable among speech-impaired users.

Furthermore, we take into account the influence of diversity among devices. For the issue of inertia-based lip reading, the device diversity primarily stems from the distinct intrinsic noise in IMUs [24]. Wiener filters, as detailed in Section 5.2.1, can eliminate such intrinsic noise along with the associated impact of device diversity.

In addition, jaw motion measurements are influenced by variations in muscular exertion across different speech modes. For example, the muscular exertion during lecturing-style articulation (with exaggerated enunciation) significantly exceeds that of natural conversational speech. Our proposed few-shot approach can be utilized here to improve the scalability across different speech modes. We consider the articulation with exaggerated enunciation as the target set (ID_T, CV_T, IMU_T) and repeat the generation procedures. In this way, SwanTrumpet is able to function under diverse speech modes. Furthermore, interviews with speech-impaired participants indicate that their primary need is the recognition of natural conversational speech to facilitate communication and interaction. In typical usage scenarios, users perform jaw movements within habitual speech modes, while lecture-style articulations involving exaggerated enunciation are seldom encountered among speech-impaired individuals.

Table 2. Overall performance of SwanTrumpet

Accuracy	Segmentation	Recognition & Interaction				Identification	
		Vowel	Viseme	Single word	Word in context	Wakeword	Authentication
Top-1	99.8%	70.6%	78.6%	47.1%	87.6%	98.2%	97.8%
Top-2	/	85.6%	90.3%	73.7%	94.3%	/	/
Top-3	/	93.9%	97.1%	82.6%	97.9%	/	/

8 Evaluation

We conducted real-world, IRB-approved experiments to evaluate SwanTrumpet. Different from prior works evaluated on volunteers who can voice but perform mutely, our evaluations are conducted on *11 speech-impaired* participants.

8.1 Experimental Setup

System implementation. We implement SwanTrumpet on two COTS earbuds (Apple AirPods 3 and Pro 2). By calling the API `CMHeadphoneMotionManager.DeviceMotion Handler`, the earbuds sample inertial data at 25 Hz and send them to mobiles (including smartphones, tablets, and laptops) paired via Bluetooth in real time, on which a SwanTrumpet application runs. The lip language will be played in audio, or then transferred into text via ASR.

Implementation of networks. The networks in this paper are trained in a server with Intel(R) Xeon(R) Silver 4210R CPU@2.40GHz and two Nvidia GeForce RTX 3090. In particular, to train the encoder-decoder model, we collect data from 26 volunteers who are able to voice but mouth silently during collection.

Data from speech-impaired participants. We recruit 11 speech-impaired (6 male and 5 female) participants. These participants age between 16 and 60 years. They are familiar with SwanTrumpet since they have taken part in our semi-structured interview. These participants mouth 2,500 common Chinese characters [66], covering the most common pronunciation in Mandarin. The data is collected on Apple AirPods 3 and Pro 2 paired to an iPad 9 (iPadOS 17.5.1). In total, we develop the first inertial dataset on Chinese lip language involving 37 (11 speech-impaired and 26 voice-able) participants. The dataset is composed of over 92,000 inertial items.

8.2 Overall Performance

We evaluate the overall performance of SwanTrumpet in terms of segmentation, translation, and identification. The results are listed in Table **??**, in which the top-k accuracy means how often the ground truth of a sample is within the first k predicted probabilities.

8.2.1 Segmentation. Our proposed method supports an error-free and automatic realization for segmentation. The success rate reaches up to 99.8%. This lays the foundation for the following lip reading. In addition, we estimate the false positive detection in Section 8.6.

8.2.2 Recognition and Interaction. We conduct recognition evaluations at the syllable/word, phrase, and sentence levels, respectively. We employ Top-N accuracy as the metric, which evaluates the proportion of instances in which the true class label is included among the top N predicted classes with the shortest DTW distances.

Syllable/word-level. We evaluate SwanTrumpet's ability to distinguish syllables and words. It achieves a top-1 accuracy of over 70% and a top-3 accuracy exceeding 95% in identifying vowels and visemes. Nevertheless, directly recognizing consonants is difficult. This is because consonants often involve subtle inertial cues, and several consonants produce similar articulatory patterns, as analyzed in Section 6.3. Experimentally, consonant

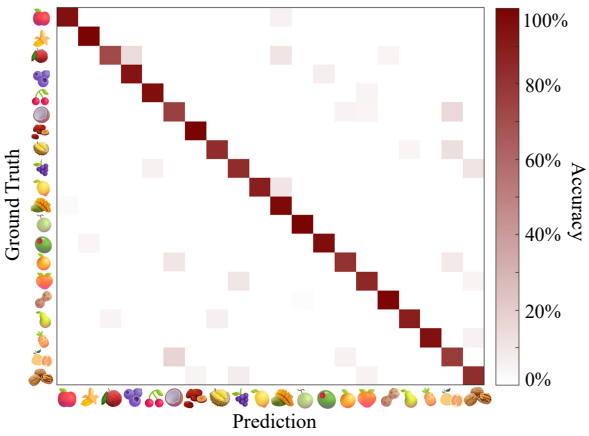


Fig. 15. Confusion matrix for noun recognition.

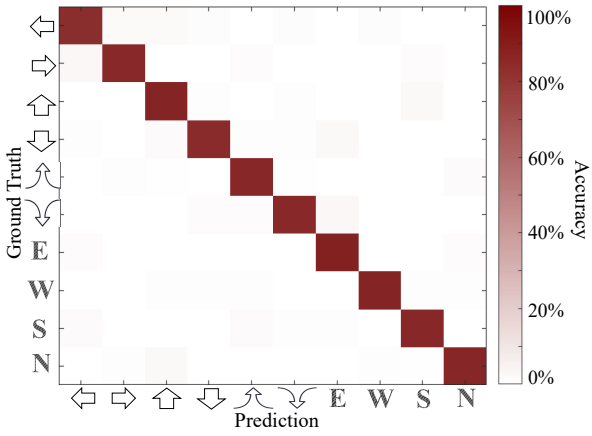
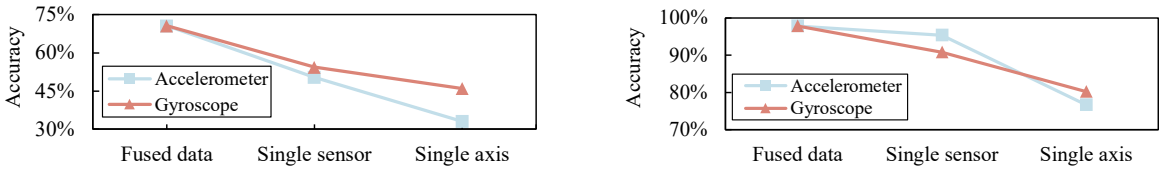


Fig. 16. Confusion matrix for preposition recognition.

recognition achieves only a top-1 accuracy of 24.1% and a top-3 accuracy of 80.2%. Consequently, the subsequent word recognition performance is adversely affected by the poor accuracy in consonant recognition. Among 2,500 words, it attains a top-1 accuracy of 41% and a top-3 accuracy of 82.6%. Given that the dataset covers all common Chinese characters, this accuracy is quite acceptable, though somewhat modest. If with the contextual information, we can infer the word based on the Bayesian Equation. The performance can reach up to 87.6% in the top-1 accuracy and over 97% in the top-3 accuracy.

Phrase-level. The phrase-level recognition is fundamental in effective communication and interaction. Otherwise, the user's intent may become unclear or even distorted. For instance, if nouns or prepositions are misrecognized in sentences like "I want a XXX (noun)" or "Go XXX (directional preposition)," the intended meaning would be ambiguous or lost entirely. Furthermore, if these key words are missing, relying solely on contextual information or the assistance of LLM techniques would be insufficient to fully understand the user's intent. Therefore, we conduct evaluations on the performance of inferring nouns and prepositions. In the noun recognition, we set the identical translation task as the previous literature for Chinese lip language [48]. The task to be recognized includes 20 phrases (fruits), as listed in Appendix C. These phrases consist of two characters. We follow the task of phrase recognition [48] because phrases are more frequently used in daily life than single words in Chinese. SwanTrumpet obtains a high accuracy in phrase recognition. The detailed confusion probabilities for these phrases are shown in the confusion matrix in Figure 15. The average recognition accuracy stands at 90.3%. The performance of SwanTrumpet on COTS earbuds is only 4.2% lower than the state-of-the-art (SOTA) lip language translators [48] (of 94.5%), which rely on specialized sensors/devices. The performance of each phrase in SwanTrumpet is presented along the diagonal. Thirty percent of the phrases achieve an accuracy exceeding 95%, 45% of phrases maintain an accuracy of 90%, 85% of phrases have ≤80% accuracy, and none phrases demonstrate an accuracy below 70%. In the recognition of prepositions, we evaluate SwanTrumpet on distinguishing ten directional prepositions (i.e., left, right, forward, backward, upward, downward, east, west, south, and north). Figure 16 shows its confusion matrix. The average accuracy reaches up to 93.8%. The above results demonstrate that SwanTrumpet can support the accurate translation of lip language.

Sentence-level. We conduct the evaluation on recognizing complete sentences. We select 10 commonly used interaction commands [10] of voice assistants (VAs), represented by Siri. Appendix C lists these commands and their Chinese pronunciations. The overall word-recognition accuracy is 53.7%. Leveraging the contextual information within sentences, the word-recognition accuracy improves significantly to 82.0%. With the advancements of large language models (LLMs), these sentences can be further corrected. Particularly, Siri is now integrated



(a) Vowel recognition (b) Authentication

Fig. 17. Ablation Study.

with ChatGPT [8], which is conducive to accurate command recognition. For instance, the sentence “Gei Ma Ma Da Dian Hua." (meaning “Call Mom.") is initially recognized as “Gai Ma Ma Ba Tian Hua," then corrected to “Gai Ma Ma Da Dian Hua" with contextual cues, and finally accurately recognized as “Gei Ma Ma Da Dian Hua" by GPT-4o. Out of the 10 sentences, 9 are correctly recognized. The only misrecognized sentence is “Guan Deng" (meaning “Turn off the lights"), which was mistaken as “Huan Deng" (meaning “slide show"). This error likely arises because the sentence consists of only two words with common meanings, making it more prone to confusion. We observe a general increasing trend in recognition accuracy as sentence length grows, and plan to validate this assumption on a larger set of sentences in future work. In general, SwanTrumpet is competent in complete sentence recognition.

8.2.3 Identification. SwanTrumpet supports unvoiced authentication among speech-impaired users. It can detect the wakeword (i.e., ‘Siri’) with an accuracy of 98.2%. After the short-term registration of 30 words (also the requirement in the few-shot scalability, as listed in Appendix C), the CNN network achieves an accuracy of 97.8% among 11 speech-impaired participants with an equal error rate (EER) of 2.05%.

8.2.4 Ablation Study on Sensor Fusion. We conduct ablation experiments to quantitatively evaluate the effectiveness of the sensor fusion technique. We test performance using either a single sensor (a three-axis accelerometer or gyroscope) or data from a single axis (defaulting to the x-axis). As shown in Figure 17, the accuracy of both vowel-recognition and identification degrades progressively as data from fewer axes/sensors are involved.

8.3 Latency

We evaluate the real-time performance on COTS devices. The latency is measured on a pair of earbuds (AirPods 3) to the application reaction on a iPad 9. The built-in `Date` function [7] is employed to synchronize the clocks of the earbuds and the mobile device and record time stamps. The average latency measures 181.6 ms across 110 attempts of lip reading. The reaction time fluctuates as articulation, with an experimental maximum of 196 ms. For the task of user identification, the additional calculation delay is 43.2 ms on average. The result indicates that SwanTrumpet operates in near real-time.

8.4 Generalization Performance

8.4.1 Impact of Device Diversity. SwanTrumpet is capable of generalizing across various device models with different mobile operating systems (OSs). To evaluate device compatibility, we invited a speech-impaired participant to test 3 earbuds in total. SwanTrumpet achieves phrase recognition accuracies of 90.6% on AirPods 3, 89.5% on AirPods Pro 2, and 91.0% on Beats Powerbeats Pro 2. In particular, the Powerbeats Pro 2 earbuds are connected to a vivo X200s smartphone running AndroidOS (OriginOS 5). This shows SwanTrumpet’s ability to work beyond Apple. The above results indicate that our proposed process fits diverse devices.

8.4.2 Impact of User Diversity. In this experiment, we select one speech-impaired participant as the target new user, while the rest served as source users for training the encoder-decoder model. We select the registered

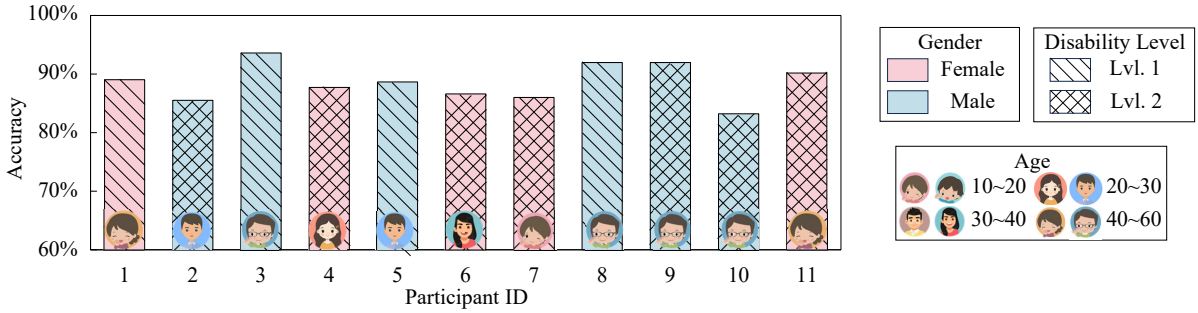


Fig. 18. Scalability among few-shot new users.

characters to fine-tune the trained model for few-shot generation. SwanTrumpet achieves an average accuracy of 88.5% in identifying unregistered nouns and prepositions across speech-impaired users. It just requires a registration of merely 30 characters. As shown in Figure 18, all speech-impaired participants achieve a scalable performance above 79%, regardless of their gender, age, or disability level.

The performance declines slightly when the generated data are used as templates in place of the actual ones. The average DTW distance is 0.790 between the generated and actual inertial data related to the same articulatory gesture. The probable limitation lies in the Griffin-Lim algorithm that compensates the phase information approximately but not fully accurately [9]. Nevertheless, the generated data also exhibit sufficiently unique waveforms. A potential solution is to adopt advanced spectrogram inversion techniques, e.g., WaveNet [74], for the phase compensation.

8.5 Robustness

8.5.1 Impact of Environmental Noise. In our interviews, data were collected based on the willingness of the speech-impaired participants in two settings: a meeting room and the living rooms of the participants' homes. SwanTrumpet demonstrated robustness in these cases, with an accuracy of 90.6% and 89.6% in the noun recognition, respectively.

To simulate real-world noisy environments, we place a loudspeaker one meter away from two speech-impaired participants (who are unwilling to conduct experiments outdoors). The background noise level in the meeting room is approximately 48.1 dB. The loudspeaker plays audio recordings from three real-world noisy scenarios: (1) a laboratory with people talking (low noise, 54.4 dB), (2) a bar with music playing (middle noise, 65.6 dB), and (3) a crowded bus station (high noise, 71.2 dB). We adjust the loudspeaker volume to achieve noise levels around 55 dB, 65 dB, and 72 dB at the participants' location, thereby simulating these different real-world noise conditions. As shown in Figure 19, SwanTrumpet achieves an accuracy of over 88%. These results demonstrate its robustness across environments with varying ambient noise levels. They also provide evidence that earable IMUs are resilient to both airborne and air-muscle-IMU propagated noise.

8.5.2 Impact of Earbud Wearing Position. People have unique habits in how they wear earbuds. These variations in wearing styles are likely to lead to differences in inertial readings related to silent articulatory gestures, thereby affecting the performance of SwanTrumpet.

We first consider the angle at which earbuds are worn. In our experiments, the speech-impaired participants typically wear earbuds whose tails point downwards at angles ranging from 15° ~65° towards the earth's normal direction. We generally divide the wearing styles into three types: low, middle, and high angles. The thresholds are empirically set at 35° and 50°, with user distribution depicted in Figure 20(a). Figure 20(b) demonstrates the robustness of SwanTrumpet.

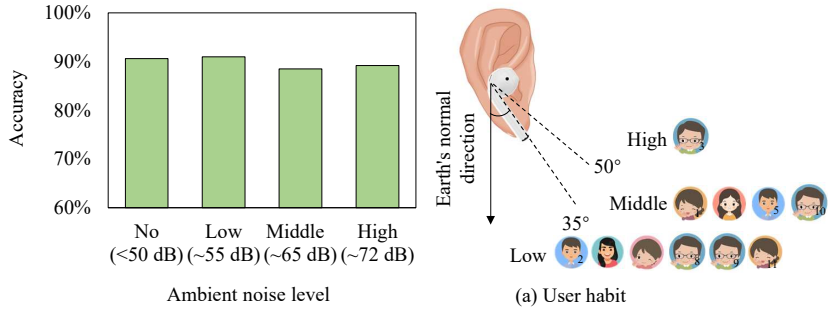


Fig. 19. Impact of environmental noise.

Fig. 20. Robustness to earbuds position (angle).

The tightness with which earbuds are worn also affects the performance. Intuitively, SwanTrumpet might perform poorly if the user wears earbuds loosely, as this position prevents the movement from articulations from being effectively transferred to the earable IMUs. We will quantitatively analyze the impact of tightness in future work. In practice, it is recommended to wear earbuds securely—not too loosely—to avoid accidental detachment as well as performance degradation of SwanTrumpet.

8.5.3 Impact of Speech Prosody Features. Temporal variations in articulation influence the characteristics of the inertial signal. Here, we consider the effects of speaking speed and intonation patterns.

Speaking speed affects the accuracy of segmentation and recognition. We observe that speech-impaired participants tend to 'speak' slowly. Their average speed ranges from 50 to 75 characters per minute (CPM). This is considerably slower than the typical daily speaking speed of voice-able users, approximately 100 CPM. To quantitatively assess the impact of speech speed, we categorize it into three groups: slow (<50 CPM), middle (50~75 CPM), and fast (>75 CPM). As a baseline, the noun-recognition accuracy reaches 90.3% at the middle speed. When two speech-impaired participants intentionally slow down their speech, the accuracy drops slightly by 2.7%. This decrease is considered within the range of random error. However, speech-impaired participants are generally unable or unwilling to articulate too fast. Therefore, we recruit three additional voice-able participants who speak silently at around 100 CPM. Under this setting, the segmentation success rate drops sharply to 63.4%. This occurs because syllables contain fewer samples, and liaisons introduce distortions to the inertial signals. Both factors also contribute to the degradation of recognition accuracy. Even with manual syllable segmentation, recognition accuracy still decreases to 78.0%. One possible solution is to build a library of prerecorded or generated templates tailored to liaison-distorted, fast-speed speech patterns.

Intonation refers to the way our voice pitch changes at the sentence or discourse level [49]. It involves variations in both temporal and frequency characteristics of speech. Generally speaking, intonation patterns can be divided into three categories: flat, falling, and rising. Flat intonation maintains a relatively steady pitch throughout. Falling intonation is most commonly used in declarative sentences to signal certainty or completion, while rising intonation is often found in interrogative sentences, signaling a question or uncertainty. Beyond these basic patterns, intonation can become more complex through combinations. For example, the fall-rise pattern conveys hesitation, mild negativity, or a rebuttal tone, whereas the rise-fall pattern is used to emphasize a point or express strong emotions. In Chinese, individual words or characters typically have a single intonation: flat, rising, or falling [36]. Complex intonation patterns arise from the combination of multiple words in a sentence rather than from changes within a single character. For the same word, falling intonation features a gradual decrease in fundamental frequency and a slight lengthening of duration, while rising intonation shows a gradual increase in frequency and a slight shortening of duration. We ask a participant to repeat a phrase "Shen Me" (meaning "what") five times each in three different intonation patterns: ① declarative (flat), ② interrogative (rising),

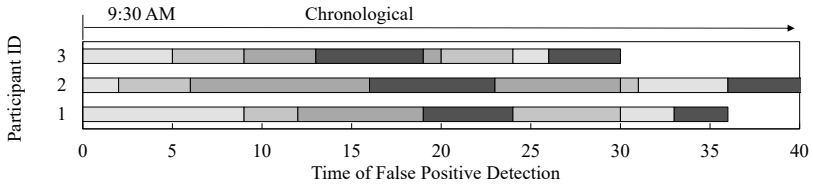


Fig. 21. The number of false positive detections in an eight-hour trace. Each rectangle represents instances of erroneous detections within a one-hour period, and different colors are used to distinguish between the hours.

and ③ exclamatory (rise-fall). As a baseline, the within-pattern DTW distances are 0.672±0.105, 0.678±0.069, and 0.912±0.208, respectively. The between-pattern DTW distances measure 0.720±0.129 (①-②), 0.795±0.222 (①-③), and 0.846±0.220 (②-③). Although the between-pattern DTW distances are generally higher than the within-pattern distances, the differences are not always substantial. This indicates that the intonation patterns in earable inertial readings exhibit considerable similarity, making them difficult to distinguish using DTW-based measures alone. One potential solution is to leverage context-driven sentiment analysis [57] to enhance intonation classification.

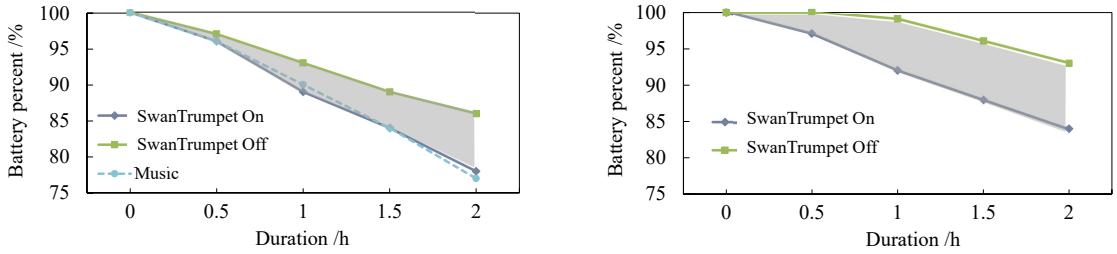
8.6 Real-world False Positive Detection

In real-world applications, there is a possibility of mistakenly detecting an articulatory gesture when a user wears earbuds running SwanTrumpet. The fewer false positive cases, the better the user experience. To evaluate this, three participants[2] wear SwanTrumpet-enabled earbuds (Apple AirPods 3) for 8 hours as they go about their daily lives. In the experiment, we gather the inertial data from earbuds, along with the audio recorded by the on-board microphones as a reference. All data are merely used to estimate the times of false positive detections, without any threat to speech eavesdropping or other privacy leakage. We count the times of false positive detection within each hour over the 8 hours starting from 9:30 AM, with the results in Figure 21. SwanTrumpet is wrongly triggered 4.45 times per hour on average. Users' body and head movements are to blame for these false activations, which have not been included within the training data for motion artifact removal in Section 5.3. A straightforward solution is to involve more movement data for training. Besides, an activation mechanism can be exploited to mitigate erroneous detections, which is a common feature in earable voice assistant services [6, 33]. For example, users often need to press and hold the talk/answer button on their earphones for a few seconds to wake up Siri [14]. Inspired by this, we design a gesture-based activation approach: SwanTrumpet detects a predefined knocking pattern on the earbuds or recognizes a wake-up word through lip-reading detection. In an eight-hour pilot experiment where the activation gesture is set as a double-knock on the earbuds, this approach reduces the false positive rate to zero.

8.7 Power Consumption

Power consumption is an essential issue, particularly when it comes to COTS devices. SwanTrumpet consumes energy on both earable and mobile devices. On the earbuds, the power consumption accounts for the polling and streaming of inertial readings. On mobile devices, the consumption lies in translating lip motions into speech. We conduct two-hour experiments, during which a participant[2] continuously talks while wearing AirPods 3 paired with an iPad 9. We use the 'Battery Usage' function in the iOS/iPadOS to measure energy consumption on both the mobiles and earbuds. The results are shown in Figure 22. Additionally, we measured the power consumed by the miscellany (including the system, screen, Bluetooth, and so on), as the green line in Figure 22.

[2]Since we are unable to conduct long-term experiments on the speech-impaired, the three participants in the false positive detection and the one in the power consumption are without disabilities.



(a) Earbuds (AirPods 3) (b) Mobile device (iPad 9)

Fig. 22. Power Consumption of SwanTrumpet operating on COTS devices.

After the two-hour operation, the earbuds' battery (with a capacity of 345 mAh in AirPods 3) decreases to 78%. Compared to the idle state, SwanTrumpet consumes 8% of the earbuds' battery, as indicated by the grey area in Figure 22(a). SwanTrumpet runs at a low power consumption, approximately 0.053 W (equivalent to 13.8 mAh per hour). As a comparison, we measure the energy consumption of audio played, as the blue dashed line in Figure 22(a). The earbuds play random music at the 50% volume setting for two hours. SwanTrumpet presents the same level of energy consumption as music playback. Considering the energy consumed by the earbud system of approximately 0.1 W, the earbuds can support a long-term operation of SwanTrumpet for over 8 hours. Such a long endurance could meet the daily usage requirements. SwanTrumpet consumes 4.5% of the battery in the iPad 9, which has a capacity of 8557 mAh. As demonstrated by the grey area in Figure 22(b), this corresponds to a power consumption of 1.46 W (or 385.1 mAh per hour). With the aid of the high-capacity batteries and the availability of power banks, the power consumption of SwanTrumpet would not be a big issue in mobile devices.

9 Discussion

In this section, we compare SwanTrumpet with SOTA work on lip language translators, review the datasets on lip language, and discuss potential avenues for future improvements.

9.1 Comparison with Previous Systems

Lip language translation has raised a concern in the literature. However, existing works are primarily oriented to English lip language. Recently, a SOTA research [48] has focused on Chinese lip language. It possesses a substantial potential user base but has received relatively little attention. Table **??** compares SwanTrumpet with SOTA methods.

In the same task of Chinese phrase recognition, a SOTA work [48] reaches an average accuracy of 94.5% across 20 classes. In comparison, SwanTrumpet attains an accuracy of 90.3% in distinguishing the same phrase group among 11 speech-impaired users. Different from the SOTA work [48] that designs a specialized mask integrated with triboelectric sensors, we utilize IMUs with low sampling rates on COTS earbuds. Furthermore, we implement MuteIt [64] on our Chinese dataset, while it is originally designed for English lip language. MuteIt [64] leverages a homemade specialized prototype equipped with two pairs of IMUs, as illustrated in Figure 7. In our experiment, we follow the same setup. In detail, one BMI160 IMU chip is attached to the temporomandibular joint of a speech-impaired participant, while another BMI160 IMU is placed on the user's temporal bone as a reference. The sampling rates of the two IMU chips are configured to 25 Hz. At this sampling rate, the prototype achieves a recognition accuracy of 82.4% for 20 Chinese phrases. In comparison, SwanTrumpet outperforms the baseline by approximately 8%. These results demonstrate that SwanTrumpet performs well in Chinese lip language decoding.

Table 3. Comparison of SwanTrumpet with SOTA works

	Platform			Participant		Lip Reading		ID	New-user Scalability
	Access	Sensor	Sampling Rate	Speech-impaired	Voice-able	Accuracy	Vocabulary		
SwanTrumpet	COTS	IMU	25 Hz	11	26	90.3%	20 phrases	✓	✓
[48]	Homemade	Triboelectric	/	4	/	94.5%	20 phrases	✓	×
MuteIt [64]	Homemade	Two IMUs	250 Hz	0	20	94.8%	100 words	×	×
Unvoiced [65]	Homemade	Two IMUs	250 Hz	0	19	>90%	70 phrases	×	×
LIPCMD [83]	Smartphone	Ultrasound	48 kHz	0	20	91.58%	50 words	×	×
SottoVoce [43]	Peripherals	Ultrasound	3.5 MHz	0	/	90%	4 commands	×	×
HearMe [86]	Peripherals	RFID	/	0	7	>88%	20 words	×	×

/: not mentioned in the original paper.

In general, SwanTrumpet achieves a similar level of accuracy to other approaches that depend on peripherals or hardware modifications. It just requires COTS earbuds (e.g., Apple AirPods) with IMUs sampled at an extremely low rate, i.e., 25 Hz. It enables various functions of lip language decoding, including translation, interaction, and identification. Moreover, it releases the requirement of long-term user registration.

9.2 Datasets on Lip Reading

To advance research on lip reading, many lip reading datasets have been released. We review the publicly available datasets in the lip decoding community and make a comparison with our released datasets [70].

Vision-based datasets. Most publicly available datasets in the lip decoding community are predominantly vision-based. These datasets cover various levels of linguistic granularity, including word-level, phrase-level, and sentence-level collections. A representative large-scale word-level dataset is LRW [17]. It covers 500 words from over 1,000 speakers by clipping videos from BBC television. MODALITY dataset [19] includes 163 commonly used interaction commands, composed of both words and short phrases. GRID [18] provides fixed-length sentences with a single syntactic structure. Each sentence contains six words arranged in a specific order. The words represent verbs, colors, prepositions, letters, numbers, and adverbs. The dataset's total vocabulary includes 51 unique words. In addition, there exist several other audio-visual corpora for lip reading, such as TCD-TIMIT [31], LRS2 [1], and LRS3-TED [2]. In particular, LRW-1000 [81] is a large-scale vision-based dataset specifically designed for Chinese lip reading.

Multimodal datasets. Non-invasive wireless sensing techniques for lip language decoding overcome some limitations of vision-based methods, such as privacy concerns and sensitivity to poor lighting conditions. Existing works have released multimodal datasets. For example, besides visual and audio information, RVTALL dataset [26] includes ultra-wideband (UWB), millimeter wave (mmWave), and laser data. Besides, LIPCMD [83] and SoundLip [85] provide acoustic-based lip reading dataset. However, these techniques rely on specialized peripherals for multimodal data collection.

Earable inertia-based datasets. Before our work, no open-source dataset had leveraged earable inertial sensors for lip language decoding. Although the references [64, 65] design homemade prototypes using earable IMUs to read lip language from voice-able participants who speak silently, these datasets have not yet been released. Our proposed dataset consists of 11 speech-impaired participants, all certified with Level 1 or Level 2 speech disabilities by the China Disabled Persons Federation, along with 26 voice-able participants who speak silently. It represents the first publicly available inertial sensor dataset in this domain and fills an important gap in research resources. While our dataset is relatively small compared to large-scale vision datasets, its unique modality

offers several key advantages. These include improved privacy, greater portability, and increased robustness in environments where visual data is hard to obtain or unreliable. These features make it especially suitable for developing disability-friendly applications, which is a central motivation for our work.

We believe that our dataset will stimulate research and development in this emerging area. In the future, we plan to include more languages into our datasets and evaluate the multilingual capability. We hope our efforts will encourage the release of more inertia-based datasets to broaden the scope and diversity of lip decoding research.

9.3 Limitations and Future Work

Despite the promising results, we acknowledge several limitations of SwanTrumpet and identify areas for future improvement. We will focus our future efforts on the following areas.

9.3.1 Performance Limitations in Diverse Situations. While SwanTrumpet performs well in various real-world scenarios, its performance is subject to certain constraints, along with the possible solutions discussed as follows.

Speaking Speed: The performance of SwanTrumpet varies as the 'speaking' speed. It peaks at a moderate speaking speed (50-75 CPM), which is typical for speech-impaired participants in our experiments. However, performance degrades at faster speeds (>100 CPM), as the segmentation accuracy drops significantly due to distorted inertial signals from liaisons. Future work could involve building a dedicated template library for fast speech patterns to improve the application scope of SwanTrumpet.

Earbud Fit: The tightness of the earbud fit impacts performance. A loose fit can weaken the transfer of articulatory motion to the IMU sensors, potentially degrading accuracy. In particular, users should be advised to wear earbuds securely for optimal performance.

Intonation and Prosody: Our current DTW-based approach shows limitations in distinguishing different intonation patterns (e.g., declarative vs. interrogative), as the inertial data exhibit considerable similarity. While this does not prevent word recognition, it limits the system's ability to capture the full nuance of speech. Future iterations could integrate context-driven sentiment analysis to better classify intonation.

False Positives: In real-world, day-long tests, SwanTrumpet averages 4.45 false positive activations per hour, mainly triggered by unseen user movements. Besides a gesture-based activation mechanism (e.g., double-knock) that effectively eliminates these errors, refining the motion artifact removal model with more diverse movement data could also provide a more seamless user experience.

9.3.2 Cross-Linguistic Applicability. The current implementation of SwanTrumpet is specifically adapted for Mandarin Chinese, leveraging its syllable-timed, single-syllable phonetic structure. Expanding its usage to other languages presents both opportunities and challenges.

Syllable-Timed Languages: Our approach can be readily applied to other syllable-timed languages [27] like Spanish, Thai, or Vietnamese, where syllables serve as isochronous units. The fundamental framework for segmentation and DTW-based recognition would remain effective.

Stress-Timed Languages: For stress-timed languages such as English [27], where rhythm is determined by stressed syllables, the segmentation and recognition logic would need adaptation. Nevertheless, our proposed methods of DTW-based recognition and unvoiced identification remain effective. A preliminary test on two fluent English speakers shows an acceptable accuracy of 91.3% for 20 English words that are mutely repeated 50 times. This indicates the scalability of SwanTrumpet across different languages. A linguistic study on English would benefit the phonetic-level recognition towards English lip language users. We will invite more speech-impaired participants from different languages in the future.

Dialects: Within Chinese, numerous dialects exist. Although participants in our study primarily use Mandarin, the proposed framework can be extended to dialects like Cantonese or Wu [36, 78], which share similar monosyllabic structures as Mandarin. Therefore, our approach can be extended to various dialect groups by leveraging

the corresponding phoneme alphabet. In future work, we plan to recruit speech-impaired volunteers proficient in different dialects to conduct dialect performance analyses.

9.3.3 Scalability improvement. We will enhance scalable applications further. Regarding user diversity, we aim to improve the spectrogram inversion approach. After the encoder-decoder model, we will adopt advanced deep learning techniques, such as WaveNet [74], in place of the Griffin-Lim algorithm [28]. They could better compensate for phase information in generated signals. More speech-impaired participants will be invited to expand our dataset. In terms of device diversity, we will cover various platforms and COTS devices, running different mobile operating systems, e.g., Android OS and HarmonyOS. These systems generally have a measurement architecture similar to iOS/iPadOS for collecting earable inertial data, making it feasible to extend SwanTrumpet to other platforms.

9.3.4 System-level integration. We will conduct system-level integration into mobile devices. By embedding SwanTrumpet into the Framework Layer, lip language can directly interact with applications on mobile devices, such as phone calls, voice assistance services, and voice authentication. This integration holds significant value across various applications.

10 Conclusion

We realize SwanTrumpet, a real-time lip language translator on earbuds. It enhances accessibility by facilitating voice-off communication and interaction for individuals with speech disorders. We have successfully implemented SwanTrumpet on two COTS earbuds, and built a large-scale dataset of lip language from 37 participants. A few-shot encoder-decoder model is exploited to improve the generalization among various users. SwanTrumpet provides those without a voice with a promising way for barrier-free communication. It promotes disability-friendly researches and applications.

Acknowledgments

This paper is partially supported by the National Science Fund for Distinguished Young Scholars of China (Grant No. 62125203), National Natural Science Foundation of China (Grant No. U21A20462, No. 62372400, No. 62502235), Natural Science Foundation of Jiangsu Province of China (Grant No. BK20240615), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 24KJB520028), and Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No. NY224030).

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. LRS2-BBC: A Dataset for Visual Speech Recognition. In *Interspeech*.
[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. LRS3-TED: A Large-Scale Dataset for Visual Speech Recognition. In *arXiv preprint arXiv:1809.00496*.
[3] Analog Devices, Inc. 1999. Anticipating and Managing Critical Noise Sources In MEMS Gyroscopes. https://www.analog.com/en/technical-articles/critical-noise-sources-mems-gyroscopes.html.
[4] S. Abhishek Anand and Nitesh Saxena. 2018. Speechless: Analyzing the Threat to Speech Privacy from Smartphone Motion Sensors. In *Proceedings of IEEE S&P*.
[5] S Abhishek Anand, Chen Wang, Jian Liu, Nitesh Saxena, and Yingying Chen. 2021. Spearphone: A Lightweight Speech Privacy Exploit via Accelerometer-Sensed Reverberations from Smartphone Loudspeakers. In *Proceedings of ACM WiSec*.
[6] Apple Inc. 2022. Use Siri with AirPods. https://support.apple.com/zh-cn/guide/airpods/dev8779e5576/web.
[7] Apple Inc. 2024. Date - Creates and returns a new date object set to the current date and time. https://developer.apple.com/documentation/foundation/nsdate/1591574-date/.
[8] Apple Inc. 2024. Introducing Apple Intelligence, the personal intelligence system that puts powerful generative models at the core of iPhone, iPad, and Mac. https://www.apple.com/newsroom/2024/06/introducing-apple-intelligence-for-iphone-ipad-and-mac/.

[9] Zhongjie Ba, Tianhang Zheng, Xinyu Zhang, Zhan Qin, Baochun Li, Xue Liu, and Kaili Ren. 2020. Learning-based Practical Smartphone Eavesdropping with Built-in Accelerometer. In *Proceedings of NDSS*.

[10] Bot Penguin. 2025. The 8 Most Useful Siri Commands in 2025. https://botpenguin.com/blogs/the-8-most-useful-siri-commands-in-2023.

[11] Yetong Cao, Fan Li, Huijie Chen, Xiaochen Liu, Chunhui Duan, and Yu Wang. 2023. I Can Hear You Without a Microphone: Live Speech Eavesdropping From Earphone Motion Sensors. In *Proceedings of IEEE INFOCOM*.

[12] Stefano Cardarelli, Alessandro Mengarelli, Andrea Tigrini, Annachiara Strazza, Francesco Di Nardo, Federica Verdini, and Sandro Fioretti. 2019. UKF Magnetometer-Free Sensor Fusion for Pelvis Pose Estimation During Treadmill Walking. In *Proceedings of IEEE EMBC*. 1213–1216.

[13] Tao Chen, Yongjie Yang, Xiaoran Fan, Xiuzhen Guo, Jie Xiong, and Longfei Shangguan. 2024. Exploring the Feasibility of Remote Cardiac Auscultation Using Earphones. In *Proceedings of ACM MobiCom*. 357–372.

[14] Tao Chen, Yongjie Yang, Chonghao Qiu, Xiaoran Fan, Xiuzhen Guo, and Longfei Shangguan. 2024. Enabling Hands-Free Voice Assistant Activation on Earphones. In *Proceedings of ACM MobiSys*.

[15] China Disabled Persons' Federation. 2023. National basic database of disabled population. https://www.cdpf.org.cn/zwgk/zccx/cjrgk/15e9ac67d7124f3fb4a23b7e2ac739aa.htm.

[16] Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip Reading Sentences in the Wild. In *Proceedings of IEEE CVPR*.

[17] Joon Son Chung and Andrew Zisserman. 2016. Lip Reading in the Wild. In *Proceedings of ACCV*. 87–103.

[18] M. Cooke, J. Barker, S. Cunningham, and X. 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.

[19] A. Czyzewski, B. Kostek, P. Bratoszewski, J. Kotus, and M. Szykulski. 2017. An audio-visual corpus for multimodal automatic speech recognition. *Journal of Intelligent Information Systems* 49, 2 (2017), 167–192.

[20] Rong Ding, Haiming Jin, Jianrong Ding, Xiaocheng Wang, Guiyun Fan, Fengyuan Zhu, Xiaohua Tian, and Linghe Kong. 2024. Push the Limit of Single-Chip mmWave Radar-Based Egomotion Estimation with Moving Objects in FoV. In *Proceedings of ACM SenSys*. 417–430.

[21] Paulo Sergio Diniz. 2002. *Adaptive Filtering: Algorithms and Practical Implementation*. Kluwer Academic Publishers.

[22] Paulo A. A. Esquef, Luiz W. P. Biscainho, Paulo S. R. Diniz, and Fabio P. Freelanci. 2000. A double-threshold-based approach to impulsive noise detection in audio signals. In *Proceedings of EUSIPCO*.

[23] Ming Gao, Feng Lin, Weiye Xu, Muertikepu Nuermaimaiti, Jinsong Han, Wenyao Xu, and Kui Ren. 2020. Deaf-aid: mobile IoT communication exploiting stealthy speaker-to-gyroscope channel. In *Proceedings of ACM MobiCom*.

[24] Ming Gao, Yajie Liu, Yike Chen, Yimin Li, Zhongjie Ba, Xian Xu, and Jinsong Han. 2022. InertiEAR: Automatic and Device-independent IMU-based Eavesdropping on Smartphones. In *Proceedings of IEEE INFOCOM*.

[25] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-based Silent Speech Interface for Smartphone Users. *Proceedings of IMWUT/Ubicomp* 4, 3 (2020), 80:1–80:27.

[26] Yao Ge, Chong Tang, Haobo Li, Zikang Chen, Jingyan Wang, Wenda Li, Jonathan Cooper, Kevin Chetty, Daniele Faccio, Muhammad Imran, and Qammer H. Abbasi. 2023. A comprehensive multimodal dataset for contactless lip reading and acoustic analysis. *Scientific Data* 10 (2023), 895.

[27] Esther Grabe and Ee Ling Low. 2002. Durational variability in speech and the rhythm class hypothesis. Mouton de Gruyter.

[28] D. Griffin and Jae Lim. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32, 2 (1984), 236–243.

[29] Xudong Guan, Chong Huang, Gaohuan Liu, Xuelian Meng, and Qingsheng Liu. 2016. Mapping Rice Cropping Systems in Vietnam Using an NDVI-Based Time-Series Similarity Measurement Based on DTW Distance. *Remote. Sens.* 8, 1 (2016), 19.

[30] Yunqi Guo, Jinghao Zhao, Boyan Ding, Congkai Tan, Weichong Ling, Zhaowei Tan, Jennifer Miyaki, Hongzhe Du, and Songwu Lu. 2023. Sign-to-911: Emergency Call Service for Sign Language Users with Assistive AR Glasses. In *Proceedings of ACM MobiCom*.

[31] N. Harte and E. Gillen. 2015. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia* 17, 5 (May 2015), 603–615.

[32] Monson H. Hayes. 1996. *Statistical Digital Signal Processing and Modeling*. Wiley press.

[33] Ruiwen He, Xiaoyu Ji, Xinfeng Li, Yushi Cheng, and Wenyuan Xu. 2022. "OK, Siri" or "Hey, Google": Evaluating Voiceprint Distinctiveness via Content-based PROLE Score. In *Proceedings of USENIX Security*.

[34] Jiahui Hou, Xiang-Yang Li, Peide Zhu, Zefan Wang, Yu Wang, Jianwei Qian, and Panlong Yang. 2019. SignSpeaker: A Real-time, High-Precision SmartWatch-based Sign Language Translator. In *Proceedings of ACM MobiCom*.

[35] Pengfei Hu, Hui Zhuang, Panneer Selvam Santhalingam, Riccardo Spolaor, Parth H. Pathak, Guoming Zhang, and Xiuzhen Cheng. 2022. AccEar: Accelerometer Acoustic Eavesdropping with Unconstrained Vocabulary. In *Proceedings of IEEE SP*.

[36] Chu-Ren Huang, Yen-Hwei Lin, I-Hsuan Chen, and Yu-Yin Hsu. 2022. *The Cambridge Handbook of Chinese Linguistics*. Cambridge University Press.

[37] iFLYTEK Co., Ltd. 2021. iFLYTEK Open Platform-An Artificial Intelligence Platform focusing on intelligent speech interaction wihch provides solutions for global developers. https://global.xfyun.cn/.

[38] Chinese Academy of Social Sciences Institute of Linguistics (Ed.). 2020. *Xinhua Dictionary*. The Commercial Press.
[39] R. E. Kalman. 1960. A New Approach To Linear Filtering and Prediction Problems. *Journal of Basic Engineering* 82D (1960), 35–45.
[40] Rachel Kaye, Christopher G Tang, and Catherine F Sinclair. 2017. The electrolarynx: voice restoration after total laryngectomy. *Medical Devices (Auckl)* 10 (2017), 133–140.
[41] Prerna Khanna, Shirin Feiz, Jian Xu, I. V. Ramakrishnan, Shubham Jain, Xiaojun Bi, and Aruna Balasubramanian. 2023. AccessWear: Making Smartphone Applications Accessible to Blind Users. In *Proceedings of ACM MobiCom*.
[42] Prerna Khanna, Tanmay Srivastava, Shijia Pan, Shubham Jain, and Phuc Nguyen. 2021. JawSense: Recognizing Unvoiced Sound using a Low-cost Ear-worn System. In *Proceedings of ACM HotMobile*.
[43] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. In *Proceedings of ACM CHI*.
[44] Gaku Kotani, Daisuke Saito, and Nobuaki Minematsu. 2017. Voice conversion based on deep neural networks for time-variant linear transformations. In *Proceedings of IEEE APSIPA*.
[45] Jianwei Liu, Wenfan Song, Leming Shen, Jinsong Han, and Kui Ren. 2023. Secure User Verification and Continuous Authentication via Earphone IMU. *IEEE Transactions on Mobile Computing* 22, 11 (2023), 6755–6769.
[46] Jianwei Liu, Wenfan Song, Leming Shen, Jinsong Han, Xian Xu, and Kui Ren. 2021. MandiPass: Secure and Usable User Authentication via Earphone IMU. In *Proceedings of IEEE ICDCS*.
[47] Jianwei Liu, Chaowei Xiao, Kaiyan Cui, Jinsong Han, Xian Xu, Kui Ren, and XuFei Mao. 2021. A Behavior Privacy Preserving Method towards RF Sensing. In *Proceedings of IEEE/ACM IWQOS*.
[48] Yijia Lu, Han Tian, Jia Cheng, Fei Zhu, Bin Liu, Shanshan Wei, Linhong Ji, and Zhong Lin Wang. 2022. Decoding lip language using triboelectric sensors with deep learning. *Nature Communications* 13, 1401 (2022), 1–12.
[49] Barney Meekin. 2025. Mastering Intonation in English. https://www.busuu.com/en/english/english-intonation.
[50] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: Recognizing Speech from Gyroscope Signals. In *Proceedings of USENIX Security*.
[51] Ministry of Civil Affairs of the People's Republic of China. 2010. *GB/T 26341-2010: Classification and Grading Criteria of Disability*. Standardization Administration of the People's Republic of China.
[52] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. IMUPoser: Full-Body Pose Estimation using IMUs in Phones, Watches, and Earbuds. In *Proceedings of ACM CHI*.
[53] Nishiki Motokawa, Ami Jinno, Yushi Takayama, Shun Ishii, Anna Yokokubo, and Guillaume Lopez. 2021. Coremoni-WE: Individual Core Training Monitoring and Support System Using an IMU at the Waist and the Ear. In *Proceedings of UbiComp/ISWC Adjunct*.
[54] Gattigorla Nagendar and C. V. Jawahar. 2015. Efficient word image retrieval using fast DTW distance. In *Proceedings of IEEE ICDAR*.
[55] Nobuyuki Otsu. 1979. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics* 9, 1 (1979), 62–66.
[56] Laxmi Pandey and Ahmed Sabbir Arif. 2021. LipType: A Silent Speech Recognizer Augmented with an Independent Repair Model. In *Proceedings of ACM CHI*. 19 pages.
[57] Minlong Peng, Qi Zhang, Yu-Gang Jiang, and Xuanjing Huang. 2018. Cross-Domain Sentiment Classification with Target Domain Specific Information. In *Proceedings of ACL*. 2505–2513.
[58] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. In *Proceedings of ICML*.
[59] Raul Quinonez, Jairo Giraldo, Luis E. Salazar, Erick Bauman, Alvaro A. Cárdenas, and Zhiqiang Lin. 2020. SAVIOR: Securing Autonomous Vehicles with Robust Physical Invariants. In *Proceedings of USENIX Security*.
[60] Julia Romero, Andrea Ferlini, Dimitris Spathis, Ting Dang, Katayoun Farrahi, Fahim Kawsar, and Alessandro Montanari. 2024. OptiBreathe: An Earable-based PPG System for Continuous Respiration Rate, Breathing Phase, and Tidal Volume Monitoring. In *Proceedings of ACM HOTMOBILE*.
[61] 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 ACM ISWC*.
[62] Cong Shi, Xiangyu Xu, Tianfang Zhang, Payton Walker, Yi Wu, Jian Liu, Nitesh Saxena, Yingying Chen, and Jiadi Yu. 2021. Face-Mic: inferring live speech and speaker identity via subtle facial dynamics captured by AR/VR motion sensors. In *Proceedings of ACM MobiCom*.
[63] Sketch Engine. 2024. Wordlist-Chinese web 2017. https://app.sketchengine.eu/#wordlist?corpname=preloaded/zhtenten17_simplified_stf2.
[64] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. MuteIt: Jaw Motion Based Unvoiced Command Recognition Using Earable. *Proceedings of IMWUT/Ubicomp* 6, 3 (2022), 26 pages.
[65] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2024. Unvoiced: Designing an LLM-assisted Unvoiced User Interface using Earables. In *Proceedings of ACM SenSys*. 784–798.
[66] State Language Commission. 1988. *Basic Vocabulary Table of Modern Chinese Characters*. Ministry of Education of the People's Republic of China.

[67] Weigao Su, Daibo Liu, Taiyuan Zhang, and Hongbo Jiang. 2021. Towards Device Independent Eavesdropping on Telephone Conversations with Built-in Accelerometer. *Proceedings of IMWUT/Ubicomp* 5, 4 (2021), 177:1–177:29.
[68] Ke Sun, Chunyu Xia, Songlin Xu, and Xinyu Zhang. 2023. StealthyIMU: Stealing Permission-protected Private Information From Smartphone Voice Assistant Using Zero-Permission Sensors. In *Proceedings of NDSS*.
[69] 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 ACM UIST*. 581–593.
[70] SwanTrumpet. 2025. SwanTrumpet. https://doi.org/10.5281/zenodo.14723896.
[71] Jiayao Tan, Cam-Tu Nguyen, and Xiaoliang Wang. 2017. SilentTalk: Lip reading through ultrasonic sensing on mobile phones. In *Proceedings of IEEE INFOCOM*.
[72] Yun Tang, Anna Y. Sun, Hirofumi Inaguma, Xinyue Chen, Ning Dong, Xutai Ma, Paden Tomasello, and Juan Pino. 2023. Hybrid Transducer and Attention based Encoder-Decoder Modeling for Speech-to-Text Tasks. In *Proceedings of ACL*.
[73] Timothy Trippel, Ofir Weisse, Wenyuan Xu, Peter Honeyman, and Kevin Fu. 2017. WALNUT: Waging Doubt on the Integrity of MEMS Accelerometers with Acoustic Injection Attacks. In *Proceedings of IEEE EuroS&P*.
[74] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. In *Proceedings of ISCA SSW*.
[75] Li Wan, Quan Wang, Alan Papir, and Ignacio López-Moreno. 2018. Generalized End-to-End Loss for Speaker Verification. In *Proceedings of IEEE ICASSP*.
[76] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I. Hong, Carmel Majidi, and Swarun Kumar. 2020. RFID Tattoo: A Wireless Platform for Speech Recognition. *Proceedings of IMWUT/Ubicomp* 3, 4 (2020), 24 pages.
[77] Lei Wang, Meng Chen, Li Lu, Zhongjie Ba, Feng Lin, and Kui Ren. 2023. VoiceListener: A Training-free and Universal Eavesdropping Attack on Built-in Speakers of Mobile Devices. *Proceedings of IMWUT/Ubicomp* 7, 1 (2023), 32:1–32:22.
[78] William S-Y. Wang and Chaofen Sun. 2015. *The Oxford Handbook of Chinese Linguistics*. Oxford University Press.
[79] Norbert Wiener. 1949. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley.
[80] Cong Wu, Kun He, Jing Chen, Ziming Zhao, and Ruiying Du. 2020. Liveness is Not Enough: Enhancing Fingerprint Authentication with Behavioral Biometrics to Defeat Puppet Attacks. In *Proceedings of USENIX Security*.
[81] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. 2019. LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild. In *Proceedings of IEEE FG*. 1–8.
[82] Qingsong Yao, Yuming Liu, Xiongjia Sun, Xuewen Dong, Xiaoyu Ji, and Jianfeng Ma. 2024. Watch the Rhythm: Breaking Privacy with Accelerometer at the Extremely-Low Sampling Rate of 5Hz. In *Proceedings of ACM CCS*.
[83] Yafeng Yin, Zheng Wang, Kang Xia, Lei Xie, and Sanglu Lu. 2024. Acoustic-Based Lip Reading for Mobile Devices: Dataset, Benchmark and a Self Distillation-Based Approach. *IEEE Transactions on Mobile Computing* 23, 5 (2024), 4548–4565.
[84] Li Zhang, Parth H. Pathak, Muchen Wu, Yixin Zhao, and Prasant Mohapatra. 2015. AccelWord: Energy Efficient Hotword Detection through Accelerometer. In *Proceedings of ACM MobiSys*.
[85] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2021. SoundLip: Enabling Word and Sentence-level Lip Interaction for Smart Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1 (2021), 43:1–43:28.
[86] Shigeng Zhang, Zijing Ma, Kaixuan Lu, Xuan Liu, Jia Liu, Song Guo, Albert Y. Zomaya, Jian Zhang, and Jianxin Wang. 2023. HearMe: Accurate and Real-Time Lip Reading Based on Commercial RFID Devices. *IEEE Transactions on Mobile Computing* 22, 12 (2023), 7266–7278.
[87] Shibo Zhang, Ebrahim Nemati, Minh Dinh, Nathan Folkman, Tousif Ahmed, Mahbubur Rahman, Jilong Kuang, Nabil Alshurafa, and Alex Gao. 2022. Coughtrigger: Earbuds IMU Based Cough Detection Activator Using An Energy-Efficient Sensitivity-Prioritized Time Series Classifier. In *Proceedings of IEEE ICASSP*.
[88] Tianfang Zhang, Zhengkun Ye, Ahmed Tanvir Mahdad, Md Mojibur Rahman Redoy Akanda, Cong Shi, Yan Wang, Nitesh Saxena, and Yingying Chen. 2023. FaceReader: Unobtrusively Mining Vital Signs and Vital Sign Embedded Sensitive Info via AR/VR Motion Sensors. In *Proceedings of ACM CCS*.
[89] Yongzhao Zhang, Wei-Hsiang Huang, Chih-Yun Yang, Wen-Ping Wang, Yi-Chao Chen, Chuang-Wen You, Da-Yuan Huang, Guangtao Xue, and Jiadi Yu. 2020. Endophasia: Utilizing Acoustic-Based Imaging for Issuing Contact-Free Silent Speech Commands. *Proceedings of IMWUT/Ubicomp* 4, 1 (2020), 37:1–37:26.
[90] Hao Zhou, Taiting Lu, Kristina Mckinnie, Joseph Palagano, Kenneth DeHaan, and Mahanth Gowda. 2023. SignQuery: A Natural User Interface and Search Engine for Sign Languages with Wearable Sensors. In *Proceedings of ACM MobiCom*.
[91] Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. 2015. Supervised Representation Learning: Transfer Learning with Deep Autoencoders. In *Proceedings of IJCAI*.

A Participant Demographics

The semi-structured interview involves 32 participants with speech disability. Among them, 11 participants pitch in our datasets of lip language and participate in the evaluation. Table 4 shows the demographics of all speech-impaired participants. For more detailed information about our interview and full dataset access, please send the research proposal with the IRB approval to the authors.

Table 4. Participant Information

ID	Gender	Age	Disability	Participation	ID	Gender	Age	Disability	Participation
1	M	20~30	Lvl. 1	◐	17	M	20~30	Lvl. 1	◐
2	F	20~30	Lvl. 1	◐	18	F	20~30	Lvl. 1	◐
3	M	20~30	Lvl. 1	◐	19	M	20~30	Lvl. 1	◐
4	F	20~30	Lvl. 1	◐	20	F	20~30	Lvl. 1	◐
5	F	30~40	Lvl. 1	◐	21	F	20~30	Lvl. 2	●
6	M	20~30	Lvl. 1	◐	22	F	10~20	Lvl. 2	◐
7	F	40~60	Lvl. 1	●	23	M	20~30	Lvl. 1	●
8	M	20~30	Lvl. 1	◐	24	M	40~60	Lvl. 1	◐
9	M	30~40	Lvl. 1	◐	25	F	30~40	Lvl. 2	●
10	M	20~30	Lvl. 2	●	26	M	30~40	Lvl. 2	◐
11	F	20~30	Lvl. 2	◐	27	F	30~40	Lvl. 1	◐
12	M	40~60	Lvl. 1	●	28	F	10~20	Lvl. 2	●
13	F	20~30	Lvl. 1	◐	29	M	40~60	Lvl. 1	●
14	F	20~30	Lvl. 2	◐	30	M	40~60	Lvl. 2	●
15	F	20~30	Lvl. 1	◐	31	M	40~60	Lvl. 2	●
16	M	20~30	Lvl. 1	◐	32	F	40~60	Lvl. 2	●

◐: Participation in the interview only.

●: Participation in the interview and data collection (evaluation).

B Syllable in Chinese Phonetics (Pinyin): Initials and Finals

The initials comprise 21 consonants and a zero initial. There are three types of finals, totaling 39. These include 10 monophthongs (composed of a single vowel), 13 diphthongs (consisting of multiple vowels), and 16 nasal vowels, as listed in Table 5. Here, we merge /n/ and /ŋ as the same classifications (e.g., /en(g)/ən(ŋ)/). This approach is supported by evidence that the voice-able volunteers and Chinese automatic speech recognition (ASR) systems (e.g., [37] can efficiently handle the confusion between /n/ and /ŋ/. The initials are divided into 7 viseme groups, as listed in Table 6. The above approaches simplify the issue of lip motion recognition.

Table 5. Finals in Mandarin Chinese Pinyi and their phonetic symbols

Monophthong	a/A/, o/o/, e/ɤ/, ê/ɛ/ i/i/, u/u/, ü/y/ -i/ɻ/, -i/ʅ/, er/ʅ/
Diphthong	ai/ai/, ei/ei/, ao/ɑu/, ou/ou/, ia/iA/, ie/iɛ/, ua/uA/, uo/uo/, üe/yɛ/,iao/iɑu/, iu/iou/, uai/uɑi/, ui/uei/
Nasal vowel	an/an/, ang/ɑŋ/, en(g)/ən(ŋ)/, ian/ian/, iang/iɑŋ/, in(g))/in(ŋ)/ uan/uan/, uang/uɑŋ/, un(g)/uən(ŋ)/, üan/yan/, ün/yn/, ong/uŋ/, iong/yŋ/

Table 6. Initials in Mandarin Chinese Pinyin and the viseme group with their phonetic symbols

No.	Viseme	Initials	No.	Viseme	Initials
1	Dbilabial	b/p/, p/p^{h}/, m/m/	5	Dental	z/ts/, c/ts^{h}/, s/s/
2	Labiodental	f/f/	6	Blade-palatal	zh/tʂ/, ch/tʂ^{h}/, sh/ʂ/, r/ʒ/
3	Dorsal	j/tɕ/, q/tɕ^{h}/, x//ɕ//	7	Blade-alveolar	d/d/, t/d^{h}/, n/n/, l/l/
4	Velar	g/k/, k/k^{h}/, h/x/			

C Character/Phrase/Sentence List

The character list in the registration phase includes: Ba, De, Shi, You, He, Ren, Wo, Le, Zai, Nü, Liao, Yu, Wei, Jia, Pian, Xiong, Er, Mei, Cheng, Huan, Yun, Qiang, Gao, Xue, Fou, Tian, Kai, Tie, Qiu, Ying.

The phrase list to be recognized in Section 8.2.2 includes: Ping Guo (Apple), Xiang Jiao (Banana), Yang Mei (Bennet), Lan Mei (Blueberry), Ying Tao (Cherry), Ye Zi(Coconut), Hong Zao(Chinese Date), Liu Lian (Durian), Pu Tao (Grape), Ning Meng (Lemon), Mang Guo (Mango), Tian Gua (Melon), Gan Lan (Olive), Ju Zi (Orange), Tao Zi (Peach), Hua Sheng (Peanut), Xue Li (Pear), Bo Luo (Pineapple), You Zi (Pomelo), He Tao (Walnut).

Table 7 lists the 10 commonly used interaction commands for the complete sentence recognition in Section 8.2.2, along with their Chinese pronunciation.

Table 7. Interaction command sentences

No.	English meaning	Chinese pronunciation
1	Call Mom.	Gei Ma Ma Da Dian Hua.
2	Send a text to Mom saying I'll be late.	Fa Duan Xin Gei Ma Ma Shuo Wo Wan Xie Dao.
3	Set an alarm for 7 AM.	She Zhi Yi Ge Zao Shang Qi Dian De Nao Ling.
4	Open the Maps app.	Da Kai Di Tu.
5	Navigate home.	Dao Hang Hui Jia.
6	Turn on Do Not Disturb.	Da Kai Wu Rao Mo Shi.
7	Turn off the lights.	Guan Deng.
8	Play some relaxing music.	Bo Fang Yi Xie Qing Song De Yin Yue.
9	Pause the music.	Zan Ting Yin Yue.
10	What's the weather today?	Jin Tian De Tian Qi Zen Me Yang?