InvisiSpeech: Making Silent Speech Invisible through a Miniature Intraoral Sensor

Yijing Jiang*
Aarhus University
Aarhus, Denmark
Department of Computer Science
ETH Zürich
Switzerland, Zurich
y.jiang@cs.au.dk

Taiting Lu*
Pennsylvania State University
University Park, Pennsylvania, USA
txl5518@psu.edu

Berken Utku Demirel
Department of Computer Science
ETH Zürich
Switzerland, Zurich
berken.demirel@inf.ethz.ch

Zhipeng Li
Department of Computer Science
ETH Zürich
Switzerland, Zurich
zhipeng.li@inf.ethz.ch

Mahanth Gowda
Pennsylvania State University
University Park, Pennsylvania, USA
mahanth.gowda@psu.edu

Michael Wessely
Aarhus University
Aarhus, Denmark
michael.wessely@cs.au.dk

Christian Holz
Department of Computer Science
ETH Zürich
Switzerland, Zurich
christian.holz@inf.ethz.ch

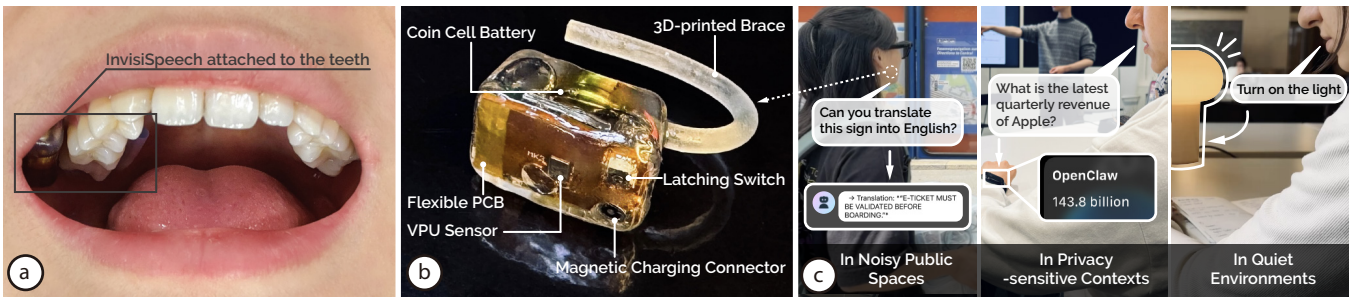


Figure 1: *InvisiSpeech* is the first intraoral interface that enables private, invisible silent speech. (a) A user is wearing *InvisiSpeech* clipped to their upper teeth. (b) *InvisiSpeech* consists of two components: (1) a 3D-printed biocompatible brace and (2) an integrated flexible PCB with a Voice Pickup Unit (VPU, our key input sensing modality), Bluetooth microcontroller, latching switch, magnetic charging connector, and battery. (c) *InvisiSpeech* can power Human–AI interaction, such as silent speech input to interact with agents via voice control in noisy public spaces, privacy-sensitive contexts, and quiet environments.

Abstract

Silent speech promises private interaction, but existing implementations often rely on visible sensors or conspicuous articulation, which limits their discretion in everyday use. We present *InvisiSpeech*, an intraoral interface that remains hidden from outside view and supports silent speech input with minimal lip movement and without drawing visual attention. Our custom flex-PCB device is housed in a biocompatible SLA-printed dental brace and attaches around the back molar. This mount couples a voice-pickup sensor to the tooth for input sensing, which our embedded microcontroller transfers via Bluetooth. Our evaluation with 15 participants shows that *InvisiSpeech* achieves up to 91.5% recognition accuracy on 50 sentence-level commands and maintains robustness in noisy environments. Participants also reported high wearability and perceived social acceptance, highlighting the practicality of our approach for daily use. We demonstrate three applications of our system: hands-free human–AI interaction in public, discreet interaction in a quiet public place, and subtle communication in a privacy-sensitive context. Our results establish invisible silent speech as a practical interaction paradigm for real-world use outside controlled settings.

*Contributed equally.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

UIST '26, Detroit, MI, USA

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2856-3/2026/11
https://doi.org/10.1145/3830398.3830550

CCS Concepts

• **Human-centered computing → Interaction devices**; **Ubiquitous and mobile computing**.

Keywords

Oral Interface, Silent Speech, Wearables, Human–AI Interaction

ACM Reference Format:

Yijing Jiang, Taiting Lu, Berken Utku Demirel, Zhipeng Li, Mahanth Gowda, Michael Wessely, and Christian Holz. 2026. InvisiSpeech: Making Silent Speech Invisible through a Miniature Intraoral Sensor. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26), November 02–05, 2026, Detroit, MI, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3830398.3830550

1 Introduction

Silent speech interfaces (SSIs) enable users to communicate or issue commands without producing audible sound. SSIs instead sense articulatory movements from the tongue or jaw, or detect subtle airflow. SSIs ideally complement situations where silence, privacy, or social discretion is required, such as quiet offices, public transport, or assistive technologies for speech-impaired users. SSIs could thus offer an ideal communication channel for human–AI interaction by supporting private, discreet, and unobtrusive input alongside context-aware interaction.

Recent work has explored a variety of SSIs, such as camera-based lip reading [31, 41, 42, 49] or face-worn EMG sensors and microphones [17, 24, 37, 38]. They reduce the need for audible speech, but the devices and articulatory movements remain *visible*. Bystanders can thus observe interaction, which limits privacy and social acceptability. This highlights the need for SSIs that are invisible at both the device and articulation levels, enabling private, discreet, and socially acceptable human–AI interaction.

We introduce *InvisiSpeech*, an intraoral silent speech interface for truly *invisible* interaction. *InvisiSpeech* is a wireless, waterproof, and portable interface that detects silent speech through minute intraoral activity and the resulting mechanically coupled acoustic vibrations. Our prototype integrates all electronics within a compact, ergonomic form factor that conforms to a person's dental arch for comfort during daily use. Combining a tooth-coupled vibration sensor (Voice Pickup Unit, VPU), a custom flexible PCB, a microprocessor, and Bluetooth Low Energy, *InvisiSpeech* supports hands-free interaction while preserving privacy and minimizing sound leakage. This approach suits both personal and professional settings, positioning *InvisiSpeech* as a private, discreet, and unobtrusive interface for emerging human–AI interactions. More broadly, our design opens opportunities for private human–computer interaction across assistive and general-purpose applications.

Based on the results of a preliminary study that guided sensor selection and placement, we chose a VPU over a conventional microphone and positioned it inside the oral cavity to improve mechanical coupling and to capture the subtle articulatory signals that afford invisible speech detection. Our 15-participant validation study evaluated sensing performance under various conditions and found that participants could comfortably wear *InvisiSpeech* and rated its social acceptability highly, as the device was largely invisible to bystanders. Our system achieved up to 91.5% command recognition accuracy across 50 sentence-level commands and remained robust in noisy environments, achieving 86.3% accuracy. We demonstrate the utility of *InvisiSpeech* in three application scenarios: hands-free interaction in public, discreet interaction in quiet environments, and subtle communication in privacy-sensitive contexts.

In summary, we contribute in this paper:

- the first silent speech interface that is designed for *truly invisible* interaction through an intraoral device;
- a fully wireless, lightweight, and open-source hardware design that is compact, mobile, and discreet;
- a decoding pipeline tailored to silent and invisible speech recognition of sentences across diverse environments;
- an evaluation of *InvisiSpeech*'s sensing performance and user experience in stationary, noisy, and mobile settings.

2 Related Work

Our work is related to silent speech interfaces, interactive oral technologies, and speech recognition.

2.1 Silent Speech Interfaces

HCI research has explored silent speech interaction using camera-based approaches, which infer speech or commands from visible lip and facial movements. These systems typically rely on RGB or depth cameras. For example, Watch Your Mouth leverages depth sensing to recognize 30 silent commands [49]. Lip-Interact employs a smartphone's front camera to recognize 44 commands [42]. MELDER demonstrates real-time silent speech recognition using a mobile phone camera [31], while LipLearner leverages contrastive learning for a customizable vocabulary [41]. Despite their effectiveness, these camera-based systems remain visually observable and depend on special environmental setups, such as stable lighting, camera positioning, and unobstructed views, which constrain their application scenarios and social acceptance.

Beyond camera-based approaches, recent work has also explored silent speech interaction using on-face and in-ear wearable sensing. On-face systems capture articulatory activity via EMG, inertial, acoustic, or RF-based sensing. Examples include AlterEgo, which leverages facial EMG with bone-conduction output [24]. Whispering Wearables combines IMU and ExG on head-worn devices for real-time command recognition [38]. Other approaches include IMU-based jaw motion tracking [37], face-worn RFID tags [48], or an eyewear device with infrared sensors [17]. Reducing device visibility, in-ear interfaces have leveraged ultrasonic or acoustic reflections within the ear canal to recognize silent speech (e.g., EarCommand [23], ReHEarSSE [10]). These approaches reduce environmental constraints, but hardware and users' articulations often remain visually observable and thus limit social discretion.

Intraoral approaches have been less explored. SilentSpeller uses capacitive touch sensing in the mouth to enable large-vocabulary silent spelling [26], but requires hardware extensions outside the mouth for power and data transfer, resulting in a visually conspicuous form factor. Table 1 compares our work with prior silent speech interfaces in HCI. Our work prioritizes the visual invisibility of both the device and the interaction, while supporting hands-free use through a fully wireless, waterproof, and open-source design.

Table 1: Comparison between InvisiSpeech and prior silent speech interfaces.

System	Year	Sensor	Interaction		Device				
			Invisible	Hands-free	Wireless	Invisible	Open Source	Waterproof	Form Factor
LipInteraction [42]	2018	RGB Camera	×	×	✓	×	×	×	Smartphone
SilentVoice [13]	2018	Mic	×	×	×	×	×	×	Hand-held
AlterEgo [24]	2018	sEMG	✓	✓	×	×	×	×	Face-worn
SoundLip [51]	2021	Mic	×	×	✓	×	×	×	Smartphone
Derma [34]	2021	IMU	×	✓	×	×	×	×	Neck-worn
EarCommand [23]	2022	Mic & IMU	×	✓	×	×	×	×	EarBud & Smartphone
SilentSpeller [26]	2022	Capacitive	×	✓	✓	×	×	×	Intraoral
MuteIt [36]	2022	Mic & IMU	×	✓	✓	×	×	×	EarBud
EchoNose [43]	2023	Mic	×	✓	✓	×	×	×	Eyeglasses
LipLearner [41]	2023	RGB Camera	×	×	✓	×	×	×	Smartphone
EchoSpeech [52]	2023	Mic	×	✓	✓	×	×	×	Eyeglasses
ReHEarSSE [10]	2024	Mic	×	✓	×	×	×	×	Headphone & Smartphone
Watch Your Mouth [49]	2024	Depth Camera	×	×	✓	×	×	×	Smartphone
MELDER [31]	2024	RGB Camera	×	×	✓	×	×	×	Smartphone
Whispering Wearables [38]	2024	sEMG & IMU	×	✓	✓	×	×	×	Headphone & Eyeglasses
Unvoiced [37]	2024	Mic & IMU	×	✓	✓	×	×	×	EarBUD
InvisiSpeech (ours)	2026	VPU	✓	✓	✓	✓	✓	✓	**Intraoral**

2.2 Intraoral Interface

Recent research has demonstrated that intraoral interfaces can enable a variety of discreet, hands-free interactions. Both TactTongue [30] and TeethMark [21] introduce intraoral interfaces for tongue input and electrotactile output. Gaze+Mouth [29] combines eye gaze for pointing with intraoral tongue gestures for selection in extended reality. Such interfaces can also greatly enhance accessibility, serving as assistive tools to help people with physical disabilities control wheelchairs, computers, or robotic arms using oral interaction [1, 2, 40]. In addition to interaction, intraoral user interfaces contribute to healthcare by monitoring oral signals. They can track head impacts in sports by mouthguard-embedded accelerometers [6], detect teeth grinding with piezoresistive sensors [9], monitor oral PH level [19] and drug release [22], and support gamified myofunctional therapy [12]. These studies and products highlight the diversity of oral interface technologies.

Fabricating intraoral devices remains challenging due to the humid environment and the oral cavity's complex and compact geometry. These constraints make it difficult to integrate circuits and electronic components, as well as to support reliable wireless data transmission and power delivery, which is a challenge shared with other body-integrated technologies [16]. Recent innovations in HCI have advanced the fabrication of intraoral devices. ChewIt [14] integrates an IMU sensor into a 3D-printed housing that users can place in their mouth for hands-free interaction. Densor [11] presents a battery-free sensing system powered by Near Field Communication (NFC). Meanwhile, MouthIO [20] introduces a prototyping platform that supports various sensors and output components inside the mouth. Building on this prior work, we aim to develop a fully wireless, waterproof, and portable device that users can wear comfortably to support silent speech.

2.3 Speech Recognition

Automatic speech recognition (ASR) has undergone a major shift from traditional pipelines [32] to end-to-end neural architectures [33]. Early deep learning approaches such as encoder-only methods (e.g., wav2vec [5]) demonstrated strong performance. More recently, large-scale sequence-to-sequence models have emerged as a dominant paradigm, leveraging transformer-based [47] encoder–decoder models trained on diverse speech corpora.

A notable milestone in this direction is the Whisper model [33], which demonstrates that scaling weakly supervised speech training to internet-scale data can yield generalizable systems. Whisper is trained on approximately 680,000 hours of audio-text pairs and is capable of performing translation and language identification within a unified framework. Unlike traditional systems that require dataset-specific fine-tuning, Whisper exhibits strong performance, approaching human-level robustness on several benchmarks.

Despite these advances, limitations remain. Whisper and similar models can exhibit sensitivity to some acoustic conditions, under domain shift [53]. Furthermore, their reliance on large-scale paired audio-text data makes adaptation to specialized modalities—such as silent speech interfaces—non-trivial.

In the context of silent speech, these models motivate the exploration of whether Whisper-style representations can be extended beyond acoustic signals. While traditional ASR assumes audible speech, silent speech systems must operate on alternative modalities, raising new challenges for alignment, representation learning, and generalization. As such, adapting modern speech-to-text models to silent speech remains an open research direction.

3 Preliminary Study on Sensor Selection and Placement to Enable *InvisiSpeech*

Invisible Silent Speech Interaction. The goal of our design was to support invisible silent speech—speech that produces minimal visible movement and no output that is audible to bystanders. Unlike prior work that relied on lip reading [31, 41, 42, 49] or on-face IMU and ExG sensors [37, 38], which primarily capture externally visible lip or jaw movements, our approach shifts attention to articulatory behaviors that are less perceptible to bystanders.

Specifically, our interaction technique relies on subtle tongue movements and intraoral airflow rather than audible vocalization.

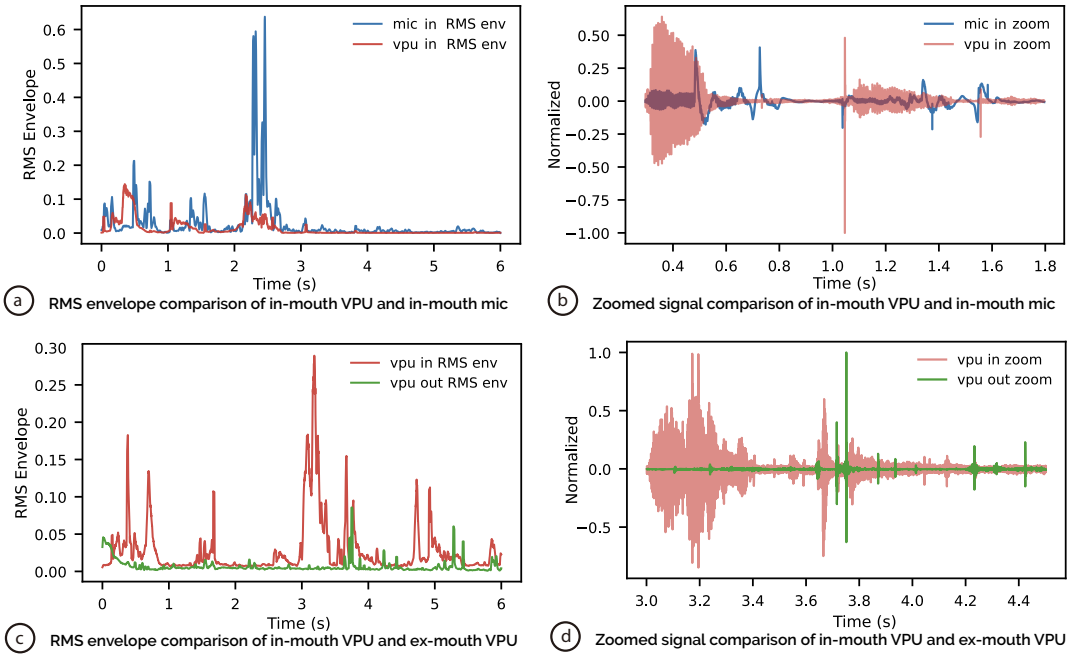


Figure 2: Comparative performance between VPU and microphone sensing inside and outside the mouth environment.

During use, participants keep their mouth in a relaxed, natural state, leaving a slight gap between the lips instead of fully sealing them. This posture allows gentle airflow to pass through the oral cavity while the tongue articulates internally, producing articulatory patterns without generating noticeable sound.

Most phonemes produce little to no visible lip movement when using our interaction technique, relying primarily on tongue positioning and intraoral airflow. A few phonemes, such as bilabial sounds (/m/, /b/, /p/, /w/) and labiodental sounds (/f/, /v/), naturally generate small lip movements [8]. However, even for these sounds, the movements are subtle and unlikely to be noticed by bystanders.

Sensing Approach and Intraoral Sensor Selection. When selecting the sensor for *InvisiSpeech*, we considered several factors to balance sensing performance and user experience: (i) high sensing resolution to capture fine-grained articulatory movements, (ii) compact size to minimize the overall device form factor, (iii) low power consumption without frequent recharging, and (iv) compatibility with a fully sealed enclosure to ensure waterproofing and reliable operation in the humid intraoral environment.

Given the subtle and largely invisible articulatory patterns produced by our interaction technique, acoustic sensing is more suitable than alternatives such as on-face IMUs or ExG sensors. In addition to conventional air-conduction microphones, we evaluated a Voice Pickup Unit (VPU)—a contact-based transducer that captures low-amplitude vibrations transmitted through tissue rather than airborne sound. By sensing mechanical vibrations directly from the body, the VPU is less susceptible to environmental noise and more sensitive to subtle intraoral activity. Few HCI studies have explored the use of VPUs, such as a VPU-embedded ring for recognizing object surface textures [15], wrist-worn VPUs for gesture [25] and event recognition [27], or surface-mounted VPUs for event detection [18]. However, to our knowledge, no prior work has investigated using VPUs for intraoral sensing, such as capturing subtle articulatory movements inside the mouth.

3.1 Comparative Study

We conducted a comparative study of microphones and VPU sensors, evaluating their sensitivity to low-amplitude signals under intraoral and near-mouth placements.

Apparatus. The experimental device consists of a custom mainboard integrating a Teensy 4.0 microcontroller[1] and a PCMD3810 audio codec. Two daughter boards are connected to the mainboard with wires, each equipped with a Sonion Voice Pick-Up (VPU) bone sensor[2] and an SPK0641HT4H MEMS microphone[3]. One daughter board is sealed within an SLA 3D-printed mouthpiece fabricated from biocompatible resin and positioned intraorally between the molars and the inner cheek. The second daughter board is placed externally, approximately 3–5 cm from the mouth for direct comparison between intraoral and extraoral acoustic sensing. This configuration allows us to simultaneously compare three conditions: in-mouth VPU, in-mouth microphone, and ex-mouth VPU.

Speech Input Vocabulary. Our silent speech command set contains 50 commands from Apple's official Siri guide [3]. The set is designed to simulate realistic interactions with a conversational assistant, illustrating how users use short voice commands to control devices or find information. Each command consists of two to seven words, reflecting the concise phrasing typical of natural speech requests. The full 50 commands details are in Appendix A.1.

[1]Teensy 4.0 microcontroller, https://www.pjrc.com/store/teensy40.html
[2]Sonion Voice Pick-Up bone sensor, https://www.sonion.com/products/sensors/
[3]SPK0641HT4H MEMS microphone, https://www.syntiant.com/mems

Participants and Procedure. We recruited five participants (two female, three male). We first introduced invisible silent speech interaction to them and gave them three minutes to practice the invisible articulation of speech. After putting the device on, participants produced each command twice: once with normal speech and once using invisible silent speech. All data were recorded simultaneously from both intraoral and extraoral sensors. Recording all channels in parallel ensured identical articulation conditions, enabling a controlled comparison of sensor modality and placement.

3.2 Results

We analyzed the variance and amplitude envelope of the signals recorded to compare sensing characteristics. In this preliminary study, the in-mouth VPU appeared to capture intraoral vocal vibrations more reliably and to be less affected by noise sources such as tongue movement and saliva-related artifacts. Figure 2 presents an example of the signals and the root-mean-square (RMS) envelope computed for both the in-mouth microphone and in-mouth VPU. The VPU signal shows clearer speech-related temporal patterns and fewer transient spikes than the microphone signal.

To further quantify this difference, we compare the RMS amplitude of signals recorded inside and outside the mouth. As shown in Figure 2, the intraoral VPU exhibits higher RMS values and a more stable amplitude envelope compared to the external VPU. This indicates that the VPU inside the mouth captures stronger and more consistent speech-related vibrations, while the signal outside of the mouth is more susceptible to attenuation. These findings support the use of intraoral VPU sensing, as the higher signal energy provides a more reliable basis for downstream feature extraction.

4 InvisiSpeech Device

We now present the design principles, implementation, and speech recognition method of our *InvisiSpeech* prototype.

4.1 Design Principles

InvisiSpeech enables silent speech interaction through imperceptible oral motions in an invisible and comfortable form factor suitable for daily use without disrupting regular activities. We identified five key design goals that guided the design and development process.

Comfortable Wear. To enable seamless silent speech interaction, the intraoral device must be lightweight, invisible, and minimally invasive. Our system integrates multiple components within a limited volume, balancing size, functionality, and reliability to ensure a comfortable and long-lasting user experience. ❶ **Structural Design.** Our device conforms to the oral cavity, ensuring a secure fit without affecting speech, breathing, or swallowing. The main body follows the dental arch with rounded edges, positioned between the cheeks and molars. A curved hook-like structure wraps around the back molar, allowing easy placement and removal while keeping the device firmly in place during use. This compact, stable design maintains continuous contact with oral tissues for sensing while minimizing bulk and user awareness. ❷ **Miniaturized Circuit Integration.** Our device design achieves compact and miniature integration of sensing, computation, power management, and communication components within a tightly constrained footprint, while ensuring reliable signal acquisition and efficient operation.

Safety. *InvisiSpeech* is a biocompatible intraoral device engineered to ensure safe, comfortable long-term wear. ❶ **Biocompatible Materials.** The use of biocompatible materials is essential to ensure safety and prevent irritation to oral tissues, enabling comfortable long-term wear. We use biocompatible resins as the primary encapsulation materials, which are FDA-certified for contact with bone, tissue, and dentin. ❷ **Water Resistance.** To operate safely in the moist oral environment, the system is fully sealed to prevent moisture ingress, while still supporting essential functions such as power supply and control without compromising the enclosure's integrity. ❸ **Bite-proof.** Before final sealing, the device is fully filled with dental resin to enhance mechanical strength, enabling it to withstand biting forces [20].

Mobility. *InvisiSpeech* enables silent speech interaction via wireless communication, avoiding the usability and mobility limitations of wired connections. The system is optimized for low power consumption to support continuous operation on a small battery while ensuring reliable real-time transmission. ❶ **Wireless Communication.** The system supports seamless Bluetooth connectivity, allowing unrestricted user mobility while maintaining a form factor suitable for intraoral use. ❷ **Power Efficiency.** Acoustic sensing at sufficient fidelity (e.g., 8 kHz) increases energy demands. The system minimizes power consumption to enable continuous operation on a small battery, ensuring reliable real-time data transmission to external devices.

Privacy is a key design goal of *InvisiSpeech*, combining a discreet form factor with silent speech sensing to protect both the user and surrounding individuals. ❶ **Invisible Device.** The device is visually unobtrusive and minimally intrusive, allowing it to be worn discreetly without interfering with natural speech or daily activities, protecting user privacy in public settings. ❷ **Silent Speech Protection.** By capturing speech via internal sensors rather than external microphones, the system prevents eavesdropping, ensuring that speech signals are only sensed within the oral cavity. The embedded VPU does not record ambient sounds, further safeguarding the privacy of surrounding individuals.

Manufacturability. To improve accessibility and affordability for the research community, *InvisiSpeech* is designed with consideration of both cost-efficiency and openness. ❶ **Cost-Effectiveness.** The system leverages commercial off-the-shelf electronic components that are widely available and economically viable. These components are integrated into a custom-designed four-layer PCB, optimized to balance performance while minimizing fabrication expenses. As a result, the overall cost of *InvisiSpeech* remains low, making it practical for broader adoption and research use. ❷ **Openness.** Beyond hardware accessibility, we open-source the software stack for *InvisiSpeech*, including modules for wireless data transmission, signal acquisition, and preprocessing. This allows researchers to access and stream intraoral sensing data in real time and supports the development of new intraoral interaction applications. We hope the community will use this platform to explore more

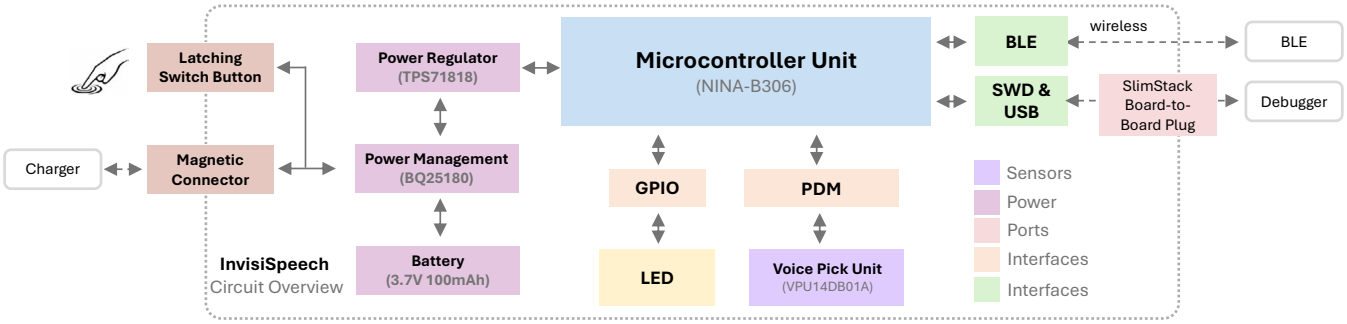


Figure 3: *InvisiSpeech* system architecture. The microcontroller unit is the central hub that communicates with the sensor.

specific use cases of *InvisiSpeech* and further advance research in intraoral interaction.

4.2 Implementation

Our goal is to design a lightweight, invisible, and minimally intrusive intraoral device that integrates input sensing and inter-device communication within a highly constrained form factor, while achieving water resistance, charging, and power control in a sealed housing through appropriate material selection and hardware manufacturing. Figure 4 illustrates the implementation of *InvisiSpeech*'s device, where we optimized circuit design, component layout, and encapsulation to minimize its size and to ensure comfort, stability, and reliable operation during wear in the oral environment.

Circuit Design. As shown in Figure 3 and Figure 4c, we design a compact four-layer PCB that integrates the core electronic components of *InvisiSpeech*, including the VPU sensor, microcontroller unit (MCU), and power management circuitry. The MCU is built around the NINA-B306 [46] module by u-blox for quick prototyping, which provides Bluetooth Low Energy (BLE) communication via a 2.4 GHz radio. The MCU interfaces with the VPU sensor VPU14DB01A [35] by Sonion through a Pulse Density Modulation (PDM) interface. To support components with different voltage requirements, we design a dedicated power management system. A Texas Instruments TPS71818-33 [45] low-dropout (LDO) regulator provides a stable 3.3 V supply for the MCU and VPU sensor. In addition, we incorporate a battery management unit based on the Texas Instruments BQ25180 [44] power management IC for charging control and protection against overcharging and overdischarging. *InvisiSpeech* is powered by a 35 mAh LIR1040 coin cell battery to seamlessly integrate within its compact form factor. Power consumption was approximately 12 mA in standby (BLE disconnected) and 17.17 mA during active sensing and Bluetooth streaming at 3.7 V, supporting up to 2 hours of continuous use.

Magnetic Charging and Latching Switch. To enable charging without compromising the sealed design, we incorporate a magnetic charging interface that allows power transfer without exposed electrical contacts. For power control, we integrate a compact latching mechanism enabled by the BQ25180 [44] power management IC in conjunction with a miniature button. The BQ25180 supports a push-button control interface that allows the system to toggle between power states using momentary input, eliminating the need for bulky mechanical latching switches. This capability enables reliable on/off operation while maintaining a small form factor and preserving the structural integrity of the sealed enclosure.

Wireless Acoustic Sensing. We chose BLE for wireless communication to balance power consumption, communication performance, device form factor, and sensing capability. BLE provides an efficient trade-off for compact, battery-constrained intraoral devices. However, directly streaming raw acoustic data at the required sampling rate (e.g., ~8 kHz) is challenging due to the limited packet transmission frequency of BLE. Through preliminary evaluation, we observe that the maximum effective packet transmission rate is constrained to approximately 100 Hz, making byte-wise or sample-wise transmission infeasible. To address this limitation, we design a packet-based wireless streaming pipeline and employ an Adaptive Differential Pulse Code Modulation (ADPCM)-based [50] encoding scheme to compress the acoustic data. Specifically, the system operates in three stages: (i) collecting vibration-based acoustic data from the VPU sensor at the microcontroller; (ii) encoding the collected samples using ADPCM and aggregating the compressed data into packets; and (iii) transmitting the packets over a BLE connection to a smartphone. The system is implemented in C++ with Arduino and BLE libraries [4, 39]. This reduces bandwidth while preserving signal quality and achieves an effective sampling rate of ~8 kHz, sufficient to capture the characteristics of speech.

Form Factor Design. All hardware, including the flexible PCB and the battery, is enclosed in a smooth, biocompatible, waterproof brace that conforms to the oral cavity, reducing irritation, ensuring comfort during prolonged use, and protecting components from saliva and humidity (Figure 4a).

The brace consists of two parts: a base case and a case lid. We design the case with a curved form to match the dental arch, and incorporate rounded edges to ensure wearing comfort. It is positioned between the face and the back teeth, featuring a curved hook-like structure that securely attaches around the back molar, making it easy to wear and remove (Figure 4b). The lid includes two openings for the magnetic charging connector and the latching switch. The whole device of *InvisiSpeech* measures 22mm × 14mm × 11mm, with a wall thickness of 0.5mm.

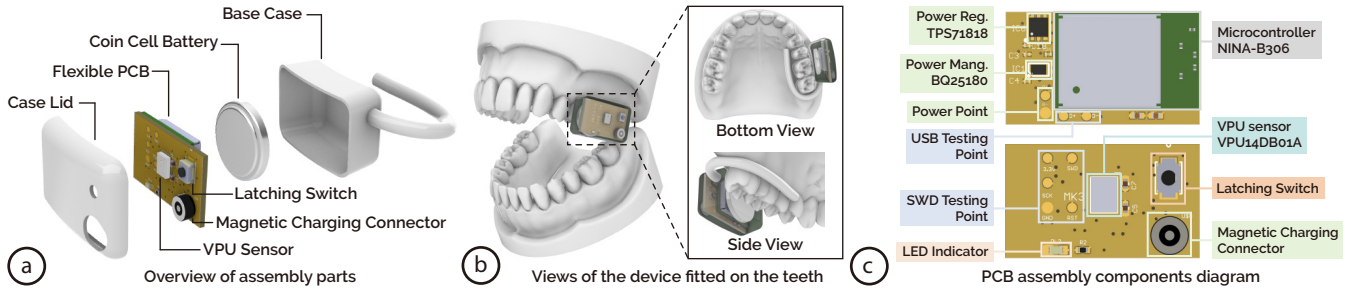


Figure 4: (a) Rendering of *InvisiSpeech* and its embedded components, including the brace, battery, and flexible PCB. (b) Mechanical structure of *InvisiSpeech*, illustrating how the device conforms to the dental arch and the placement of the FPCB within the intraoral form factor. (c) PCB layout of the flexible board, where the front side (top) integrates computing, sensing, and interface components, and the back side (bottom) accommodates the remaining elements.

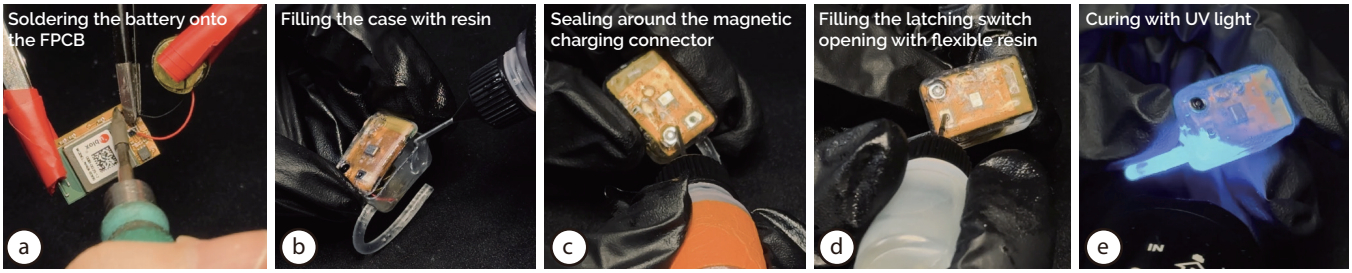


Figure 5: Assembling and fabrication process of our *InvisiSpeech* prototype. (a) Attaching the battery to the flexible circuit board through soldering. (b) Encapsulating the device components by filling the housing with resin. (c) Sealing the charging interface to ensure waterproofing. (d) Applying flexible material into the switch area to preserve functionality. (e) Finalizing the assembly by curing the device under UV light.

Assembly. We first 3D-printed the brace using a Formlabs Form 3b printer[4] and Formlabs BioMed Clear Resin[5], a biocompatible material with FDA registration suitable for tissue and dentin contact. After printing, it is cleaned with isopropyl alcohol and cured under UV light to enhance mechanical stability. Then we programmed the PCB with a bootloader via test points or USB. A coin cell battery is then soldered onto the FPCB (Figure 5a). After the braces and the flexible PCB are ready, we assembled the flexible PCB onto the case lid, and placed it together with the battery into the base case. The base case is then filled with resin (Figure 5b). After closing the lid, resin is applied around the magnetic charging connector to ensure complete waterproofing (Figure 5c). Flexible resin[6] is injected through the opening above the latching switch, allowing it to remain pressable (Figure 5d). We cured the brace with a UV light touch (Figure 5e), then fully cured it in a UV chamber. The final prototype weighs ~4.5 g.

Cost. All firmware packages and programs we used are free. The total cost to produce a single *InvisiSpeech* at wholesale scale is as low as $13.4. A summary of each hardware component's retail and wholesale unit price is shown in Appendix A.2.

[4]Formlabs Form 3b printer, https://formlabs.com/eu/3d-printers/form-3b/
[5]Formlabs BioMed Clear, https://formlabs.com/store/materials/biomed-clear-resin/
[6]FormLabs BioMed Flex 80A Resin, https://formlabs.com/fr/boutique/materials/biomed-flex-80a-resin/

Cleaning and Disinfection. *InvisiSpeech* can be cleaned by standard retainer-cleaning methods, including commercially available cleaning tablets or isopropyl alcohol. For practical use, we recommend removing the brace while eating, similar to other commercial braces. When drinking beverages other than plain water, using a straw is recommended.

Silent Speech Recognition. We adapt a pretrained Whisper [33] to silent speech by learning a residual mapper in the feature space, while keeping the original model frozen. Given paired silent and vocalized recordings of the same command, our aim is to transform silent speech representations so that they align with the corresponding acoustic speech representations expected by Whisper.

We pass each input through the frozen Whisper encoder to obtain latent audio features. For vocalized speech, these encoder outputs are precomputed and cached. For silent speech, the encoder produces a base representation that serves as input to the mapper.

Residual mapper. We learn a small network that operates on the encoder features. The mapper is implemented as a residual module with layer normalization and a two-layer MLP, $\tilde{\mathbf{h}} = \mathbf{h} + \alpha \cdot f(\mathbf{h})$, where $\mathbf{h}$ denotes the silent encoder features, $f(\cdot)$ is a nonlinear transformation, and α is a learned scaling parameter. This residual formulation encourages the model to preserve useful structure from the original representation while applying minimal corrections.

Training objective. We train the mapper on paired silent-vocalized samples. For each pair, the frozen Whisper encoder produces a silent feature sequence $\mathbf{H}^{(s)} \in \mathbb{R}^{T\times d}$ and a vocalized feature sequence $\mathbf{H}^{(v)} \in \mathbb{R}^{T'\times d}$. The mapper $g(\cdot)$ then operates on the silent features to produce a corrected representation, $\widehat{\mathbf{H}}^{(s)} = g\left(\mathbf{H}^{(s)}\right)$.

Since silent and vocalized recordings may not be frame-aligned, we require no frame-wise matching and match pooled feature statistics instead. We define the mean and variance of features as

$$\mu(\mathbf{H}) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{H}_t, \qquad \sigma(\mathbf{H}) = \mathrm{Var}\left(\mathbf{H}_1, \ldots, \mathbf{H}_T\right),$$

where the statistical features are computed along the temporal dimension. Then, we define the feature alignment loss as

$$\mathcal{L}_{\text{feat}} = \left\|\mu\left(\widehat{\mathbf{H}}^{(s)}\right) - \mu\left(\mathbf{H}^{(v)}\right)\right\|_2^2 + \frac{1}{2}\left\|\sigma\left(\widehat{\mathbf{H}}^{(s)}\right) - \sigma\left(\mathbf{H}^{(v)}\right)\right\|_2^2.$$

To further encourage directional agreement in the latent space, we use a cosine loss on the pooled means as in Equation 1.

$$\mathcal{L}_{\cos} = 1 - \frac{\mu\left(\widehat{\mathbf{H}}^{(s)}\right)^{\top}\mu\left(\mathbf{H}^{(v)}\right)}{\left\|\mu\left(\widehat{\mathbf{H}}^{(s)}\right)\right\|_2\left\|\mu\left(\mathbf{H}^{(v)}\right)\right\|_2}. \tag{1}$$

Finally, to preserve compatibility with Whisper decoding, we feed the mapped features into the frozen Whisper decoder and apply teacher-forced token prediction loss as in Equation 2.

$$\mathcal{L}_{\text{CE}} = -\sum_{n=1}^{N}\log p\left(y_n \mid y_{<n}, \widehat{\mathbf{H}}^{(s)}\right), \tag{2}$$

where $y_{1:N}$ is the ground-truth transcription token sequence. The overall training objective is given in Equation 3.

$$\mathcal{L} = \mathcal{L}_{\text{feat}} + \mathcal{L}_{\cos} + \mathcal{L}_{\text{CE}} \tag{3}$$

In our implementation, these terms are combined in this weighted form during training. We train the mapper using AdamW [28] ($1e-3$) while freezing all Whisper parameters. Gradient clipping is applied to stabilize training. Model selection is performed based on validation performance.

At test time, silent speech is passed through the Whisper encoder, transformed by the learned mapper, and decoded using the standard Whisper decoder. Then, we map each decoded transcript to the nearest command in the predefined command set using normalized string similarity. This converts open-ended output into a closed-set prediction. This design leverages the robustness of large-scale pretrained ASR models while requiring only a small number of trainable parameters. By operating in the encoder feature space, the approach avoids retraining the full model and provides a simple mechanism for adapting Whisper to silent speech.

Recognition is performed on a paired phone or laptop. Sensor data are streamed via BLE at 8 kHz using a pipelined acquisition and transmission scheme, making communication latency negligible. Once an input segment is received, the recognition pipeline(Whisper, residual adapter, decoding, and command matching) takes approximately 200 ms per command.

5 Evaluation

To validate the usability and user experience of *InvisiSpeech*, we conducted a study focusing on the performance of speech recognition and participants' perception of wearability and social acceptance.

5.1 Methods

Participants. We recruited 15 participants (7 female, 8 male) from our workplace and wider network, aged 21–30 (M=26.13). We excluded participants with ongoing dental treatment and current dental diseases. Four participants reported having fixed wire retainers. There is no prior experience with intraoral input devices. One participant is a native English speaker, and all others self-reported their English proficiency as fluent to highly fluent. Our study conforms to the ethics regulations at our institution. Participation was voluntary and anonymized, and informed consent was obtained through a signed form.

Apparatus. For the study, we implemented an *InvisiSpeech* prototype with the VPU embedded in a custom 3D-printed dental brace, while the microcontroller was outside the mouth powered by a power bank (Figure 6). The prototype was otherwise identical to the wireless version, sharing the same resin-encapsulated brace design, microcontroller, VPU sensor, Bluetooth stack, ADPCM compression, and signal processing pipeline, resulting in comparable size, weight, and sensing performance. The hardware difference was in the four thin wires that connected our study prototype to the external electronics, whereas the wireless version uses an onboard TPS71818-33 voltage regulator to stabilize battery power.

To maximize participant safety, all electronics embedded in the brace were fully encapsulated in biocompatible resin, while the power source remained external. MouthIO's evaluation demonstrated that the resin-encapsulated device withstands adult bite forces without damaging the internal electronics [20]. The brace was disinfected between participants using isopropyl alcohol and dishwashing detergent. The connecting wires also acted as a physical tether, reducing the already low risk of accidental swallowing.

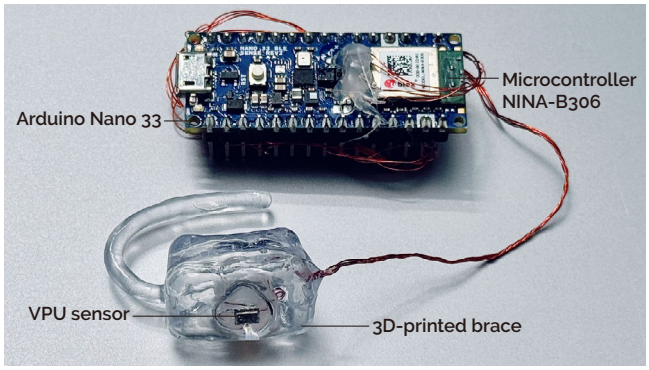


Figure 6: The wired study prototype shared the same hardware and processing pipeline as the wireless version, except that it was powered by an external power bank.

Table 2: ASR performance across three conditions. Lower is better for WER; higher is better for accuracy. Values are reported as mean ± standard deviation across participants.

	Sitting		Noisy Environment		Walking	
Method	**WER ↓**	**Acc. ↑**	**WER ↓**	**Acc. ↑**	**WER ↓**	**Acc. ↑**
Frozen Whisper	0.46 ± 0.05	52.31 ± 4.12	0.49 ± 0.06	48.76 ± 4.85	0.73 ± 0.08	21.44 ± 3.97
LOSO Whisper	0.22 ± 0.04	78.92 ± 3.28	0.27 ± 0.05	74.35 ± 3.91	0.55 ± 0.07	39.18 ± 4.26
Subject-calibrated Whisper	0.09 ± 0.02	91.54 ± 2.11	0.15 ± 0.03	86.27 ± 2.74	0.47 ± 0.06	47.63 ± 3.88

Table 3: Ablation study of the proposed residual-mapper training objectives under evaluation.

	Sitting		Noisy Environment		Walking	
Loss configuration	**WER ↓**	**Acc. ↑**	**WER ↓**	**Acc. ↑**	**WER ↓**	**Acc. ↑**
Full objective	0.09 ± 0.02	91.54 ± 2.11	0.15 ± 0.03	86.27 ± 2.74	0.47 ± 0.06	47.63 ± 3.88
w/o $\mathcal{L}_{\text{CE}}$	0.21 ± 0.04	78.36 ± 3.45	0.29 ± 0.04	71.82 ± 3.92	0.58 ± 0.07	35.91 ± 4.17
w/o $\mathcal{L}_{\cos}$	0.13 ± 0.04	86.74 ± 2.86	0.20 ± 0.04	81.43 ± 3.11	0.51 ± 0.06	42.58 ± 3.95
w/o $\mathcal{L}_{\text{feat}}$	0.11 ± 0.04	88.91 ± 2.48	0.18 ± 0.03	83.67 ± 2.96	0.49 ± 0.08	44.21 ± 3.77

Procedure. Participants were first introduced to invisible silent speech and practiced using the *InvisiSpeech* device. A mirror was provided to allow them to observe their appearance and ensure their articulation remained visually unobtrusive.

The study consisted of three conditions. In Condition A, a quiet seated setting, participants produced 50 full sentences from the comment set, once using normal voice and once using invisible silent speech. In Condition B, a noisy public setting with background noise at approximately 70 dB(A), simulated using recordings of crowd chatter and public environments, participants produced 25 sentences using invisible silent speech. In Condition C, a walking condition, participants walked along a 10-meter path in an office environment at a normal pace while producing 25 sentences using invisible silent speech. During recording, each command was produced individually. After each command, a 6-second segment was recorded to capture the signal.

Finally, participants completed an open-ended survey about their experience. The survey covered three main topics: wearability and social acceptance of the device, ease of learning and intuitiveness of the interaction, and potential applications. The entire study lasted approximately 45 minutes per participant.

5.2 Results

Speech Recognition Performance. We evaluate two training protocols. (i) Leave-one-subject-out (LOSO) cross-validation measures cross-user generalization by holding out one subject at a time. (ii) Personal calibration adapts the mapper using a small subject-specific set of five predefined commands. We find that five commands give a practical balance between calibration effort and performance, where including more commands do not increase the performance significantly.

Table 2 summarizes performance across the three conditions. Adapting Whisper with the proposed residual mapper leads to substantial gains over the frozen baseline. Under LOSO training, the model already improves recognition accuracy. Subject-specific calibration further boosts performance while reaching the lowest WER in the sitting condition. A Friedman test followed by corrected post-hoc pairwise comparisons confirmed that these improvements are statistically significant ($p < 0.05$).

Table 3 presents the ablation study of our method. Removing the decoder loss $\mathcal{L}_{\text{CE}}$ results in the largest performance drop across all conditions, confirming that token-level supervision is critical for aligning mapped features with Whisper's decoding space.

Eliminating the cosine similarity term $\mathcal{L}_{\cos}$ also degrades performance. This indicates that directional alignment in the latent space contributes to stable adaptation. In contrast, removing the feature-matching loss $\mathcal{L}_{\text{feat}}$ leads to a smaller but consistent drop. Overall, these results highlight that while all components contribute, the decoder-level objective plays the dominant role, with auxiliary feature alignment terms providing complementary improvements.

Impact of Background Noises. The performance of our method in the noise remains stable. This is expected given that sensing is performed intra-orally using a VPU, which is largely isolated from ambient interference. As a result, background noise has limited influence, and the learned mapping generalizes well across the conditions. This suggests that the approach is suitable for deployment in public environments where external noise is present.

Impact of Body Motions. Performance degrades significantly in the walking condition. We attribute this to two factors. First, the model is trained only on paired data collected in a sitting position, making walking data out-of-distribution naturally. Second, body motion can introduce mechanical perturbations to the VPU signal, such as footsteps and body vibrations, which are not observable during training. These effects lead to reduced robustness under motion. Despite this, our results demonstrate the potential of *InvisiSpeech* to support interaction while walking. Incorporating motion-diverse training data or explicitly modeling motion-induced variability is a promising direction for improving performance in dynamic, real-world settings.



Figure 7: Applications of *InvisiSpeech*. (a) Subtle communication during a meeting, enabling users to query information without speaking aloud. (b) Hands-free interaction in public, allowing users to interact with AI glasses discreetly. (c) Discreet interaction in quiet environments, enabling silent control of devices without disturbing others.

Sound Leakage. After participants practiced invisible silent speech, they were asked to read five commands before data collection. A sound level meter was positioned 30 cm from the mouth to measure the peak level of leaked sound during articulation.

The sound leakage peak was 36.59 dB(A) on average (min: 34.2 dB(A), max: 39.6 dB(A)), which is below typical ambient noise levels (e.g., 40–60 dB(A) in offices or public spaces). It is thus not perceptible to bystanders. It is also well below the levels of soft normal speech (52.5 dB(A)) and soft whisper (47.4 dB(A)) [13].

User Experience and Wearability. No discomfort was reported from wearing the brace during the study. Seven Participants (P4, P6, P7, P8, P9, P12, P13) believe they can wear *InvisiSpeech* for 2–4 hours without a break, four (P3, P10, P11, and P15) for 1–2 hours, and the remaining four for 15–30 minutes.

We found that the one-size-fits-all design fits some participants well. For example, P15 stated *"I really like how easy it can be put on, it's very ergonomic for my teeth. I found it comfortable."*. Four participants (P3, P7, P9, and P13) found the device slightly loose, while one (P8) reported it as tight, suggesting that a one-size-fits-all design may be insufficient and that multiple sizes (e.g., small, medium, large) could improve fit.

All participants found invisible silent speech easy to learn, and the majority agreed it is natural to use, only P12 noted that *"it's a bit challenging to not move [the] mouth."*. All participants agreed that using the device does not interfere with walking. For normal speaking, some participants reported slight interference, with five participants (P2, P4, P5, P6, and P12) noting they could "feel something in the mouth." However, others (P9 and P15) suggested that *"people can easily get used to it after a while."*

For device improvement, P4 and P6 suggested adding feedback to indicate when they are speaking too loudly or too quietly. Several participants also expressed a preference for a device that is *"smaller"* (P2, P13), *"flatter"* (P4, P6, and P8), or *"more flexible"* (P5, P6).

Social acceptance and privacy. Self-reported perceived social acceptance was high across participants; 8 of them (P2, P4, P6, P7, P9-P12) believed that bystanders would not notice the device at all.

Participants emphasized its strong privacy benefits. For instance, P8 noted *"In public, I can still say something without people understanding"*, and P9 further emphasized its usefulness for private or sensitive communication. P15 highlighted that it captures only the user's input, preventing concerns about recording others.

Potential Use Cases. Participants suggested a wide range of potential applications for the *InvisiSpeech* brace. First, it can be used in noisy environments such as public transportation (P1, P3–P6, P10, P12, P14) or social venues like bars, clubs, and festivals (P4, P10–P12). Second, it supports silent or private communication in settings such as meetings (P2, P3, P11), classes (P2, P11, P13), libraries and museums (P2, P3), shared offices (P4, P7, P11), or when others are sleeping (P7, P15). Third, it enables hands-free interaction (P1, P13, P15), for example, P15 mentioned wanting to use this for recording a podcast while biking. Finally, participants highlighted accessibility benefits for users who have lost their voice due to illness (P4, P11), or whose hands or fingers are unavailable due to motor impairments (P8, P9, P14).

Participants also suggested a variety of daily uses, including interacting with AI assistants (P3, P10), messaging or calling with other people (P6, P7, P10), smart home control (P5, P13, P15), and phone control (P1, P14).

6 Application Examples

Here, we present three application scenarios that demonstrate the potential of *InvisiSpeech*. It enables subtle, hands-free, and noise-robust speech interaction in privacy-sensitive, noisy public, and noise-sensitive environments.

6.1 Discreet in Privacy-sensitive Contexts

InvisiSpeech supports subtle, hands-free communication in professional settings, maintaining both privacy and social discretion. For example, when a user is participating in a team meeting. Instead of interrupting the discussion or speaking aloud, the user can silently issue a query to their AI assistant via *InvisiSpeech*. For instance, while a colleague presents a complex report, the user quietly "asks" the assistant for clarification on a data point (Figure 7a). The device captures invisible speech, transmits it to the AI assistant, and provides a subtle visual response through a paired smartwatch or laptop. This allows users to gather information or take notes in real time without disrupting meetings or drawing any attention.

6.2 Robust in Noisy Public Spaces

InvisiSpeech enables users to interact discreetly yet robustly in noisy public spaces. For example, while standing at a bus station, a user can silently ask AI glasses to translate a sign into English without disturbing nearby people (Figure 7b). The device captures the

user's silent speech, transmits it wirelessly, and delivers the message without requiring audible speech. Because the system relies on contact-based sensing rather than airborne sound, its performance is largely unaffected by environmental noise, enabling reliable interaction even in crowded and noisy settings. This scenario demonstrates how *InvisiSpeech* extends AI glasses with hands-free, private, and noise-robust speech interaction for daily activities such as commuting, traveling, and navigating public spaces.

6.3 Unobtrusive in Quiet Environments

InvisiSpeech allows users to interact in noise-sensitive environments without producing audible speech. For instance, a user can silently control the lighting in a library (Figure 7c). *InvisiSpeech* captures the user's intraoral silent speech and executes the command without disturbing others, eliminating the need to stand up or interrupt other people. This capability demonstrates how *InvisiSpeech* supports subtle, hands-free interaction in quiet public spaces for private information access or environmental control while maintaining social discretion and minimizing auditory disruption.

7 Discussion and Future Work

Although *InvisiSpeech* demonstrates promising performance, several directions remain for future research, including private feedback, open-vocabulary interaction, and broader sensing capabilities.

7.1 Private Feedback

In this paper, we primarily focus on invisible input, leveraging readily available commercial solutions such as earphones for audio feedback and phone or PC screens for visual feedback. However, such feedback mechanisms may not fully align with the goal of subtle and private interaction, as audio can be overheard and visual feedback may require overt attention shifts. As a next step, we plan to explore more discreet and embodied feedback modalities that better match the invisible nature of the input. In particular, we are interested in bone-conduction feedback delivered within the mouth, which could enable highly private audio transmission, as well as on-body electrotactile feedback that can convey information through localized tactile sensations.

7.2 Open Vocabulary Exploration

To explore the potential of our approach beyond the fixed command set, we conducted a preliminary open-vocabulary evaluation using a separate set of ten longer, free-form commands recorded from a single participant. We evaluated the model with the residual mapper on this set without any subject calibration and obtained an accuracy of 0.62. While this setting is more challenging than the command based vocabulary due to increased linguistic variability and the absence of a predefined set, the model is still able to recover meaningful portions of the intended commands.

We emphasize that this evaluation is limited to a single participant and a small number of samples, and therefore should be interpreted as an exploratory result rather than a definitive benchmark. Nonetheless, these findings suggest that the learned adaptation is not restricted to a fixed vocabulary and may generalize to more flexible, open-ended silent speech scenarios. This is an encouraging indication that extending silent speech interfaces beyond command-based interaction is feasible. We leave a more comprehensive evaluation of open-vocabulary settings to future work.

7.3 Beyond Silent and Invisible Speech

The current design of *InvisiSpeech* leverages a VPU sensor, which not only enables invisible silent speech interaction but also holds potential for a broader range of applications. Our VPU sensing design could support continuous respiration monitoring or subtle facial or oral gesture detection. Its sensitivity to intraoral vibrations opens opportunities for exploring other physiological or behavioral signals, suggesting that the device could serve as a versatile interface for both communication and health-related monitoring.

In addition, the fully wireless, waterproof, and compact design of *InvisiSpeech*, along with our fabrication approach for embedding flexible PCBs and batteries within a biocompatible brace, provides a generalizable platform for integrating additional sensors. For example, IMUs could enable detailed head-motion and activity tracking, while PPG sensors could allow continuous monitoring of heart rate or blood oxygen. This combination of hardware, fabrication strategy, and intraoral form factor makes *InvisiSpeech* a flexible foundation for future sensor-enabled wearable applications.

8 Conclusion

We presented *InvisiSpeech*, the first intraoral silent speech interface designed to enable private and socially acceptable human–AI interaction without audible speech or externally visible hardware. *InvisiSpeech* enables users to interact with AI systems through *invisible* silent speech by sensing minute intraoral articulations that remain largely imperceptible to bystanders. Our lightweight (≈4.5 g), low-cost (≈$13.4) prototype demonstrates the feasibility of compact and accessible intraoral interfaces. We further developed a custom signal-processing pipeline that adapts pretrained speech representations for silent speech recognition. In an evaluation with 15 participants, *InvisiSpeech* achieved up to 91.5% sentence-level command recognition accuracy and maintained 86.3% accuracy in noisy environments. Participants also reported high comfort and social acceptability.

Taken together, our results position invisible silent speech as a promising next step for human–AI interaction. Audible voice input can disrupt shared or quiet environments and expose sensitive information to bystanders. By providing a discreet, unobtrusive, and socially acceptable communication channel, *InvisiSpeech* extends interaction with AI systems to settings where speaking aloud would compromise privacy or social discretion. Invisible silent speech may therefore make voice-like human–AI interaction practical across a much broader range of everyday contexts.

Acknowledgments

Yijing Jiang was supported by the Novo Nordisk Foundation grant NNF22OC0081352. Taiting Lu was supported in part by the United States National Science Foundation (NSF) grants CAREER-2046972. Zhipeng Li was in part supported by the Swiss National Science Foundation (Grant No. 10004941).

References

[1] 2023. Augmental — augmental.tech. https://www.augmental.tech/. [Accessed 20-06-2024].

[2] Lotte NS Andreasen Struijk, Line Lindhardt Egsgaard, Romulus Lontis, Michael Gaihede, and Bo Bentsen. 2017. Wireless intraoral tongue control of an assistive robotic arm for individuals with tetraplegia. *Journal of neuroengineering and rehabilitation* 14 (2017), 1–8.

[3] Apple Inc. 2026. What can I ask Siri? - Official Apple Support. https://support.apple.com/siri.

[4] Arduino. 2026. Arduino Nano 33 BLE Sense. https://docs.arduino.cc/hardware/nano-33-ble-sense/. Includes onboard PDM microphone; accessed 2026-04-01.

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 12449–12460. https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[6] David B Camarillo, Pete B Shull, James Mattson, Rebecca Shultz, and Daniel Garza. 2013. An instrumented mouthguard for measuring linear and angular head impact kinematics in American football. *Annals of biomedical engineering* 41 (2013), 1939–1949.

[7] Carnegie Mellon University. 2011. The CMU Pronouncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[8] John Cunnison Catford. 2001. *A practical introduction to phonetics*. Oxford University Press.

[9] Arthur Claude, Olivier Robin, Claudine Gehin, and Bertrand Massot. 2019. Design and evaluation of a novel technology for ambulatory monitoring of bruxism events. *Sensors and Actuators A: Physical* 295 (2019), 532–540.

[10] Xuefu Dong, Yifei Chen, Yuuki Nishiyama, Kaoru Sezaki, Yuntao Wang, Ken Christofferson, and Alex Mariakakis. 2024. ReHEarSSE: Recognizing Hidden-in-the-Ear Silently Spelled Expressions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 321, 16 pages. https://doi.org/10.1145/3613904.3642095

[11] Vivian Dsouza, Jeffrey Pronk, Christian Peppelman, Víctor Ignacio Madariaga, Tatiana Pereira-Cenci, Bas Loomans, and Przemysław Pawełczak. 2024. Densor: An Intraoral Battery-Free Sensing Platform. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 191 (Nov. 2024), 30 pages. https://doi.org/10.1145/3699746

[12] Luke Franzke, Jonas Gartmann, Dominik Bachmann, Tino Töpper, and Karmen Franinović. 2023. TOFI: Designing Intraoral Computer Interfaces for Gamified Myofunctional Therapy. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 401, 8 pages. https://doi.org/10.1145/3544549.3573848

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

[14] Pablo Gallego Cascón, Denys J.C. Matthies, Sachith Muthukumarana, and Suranga Nanayakkara. 2019. ChewIt. An Intraoral Interface for Discreet Interactions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300556

[15] Koushi Hiraoka, Yugo Nakamura, and Yutaka Arakawa. 2026. Exploring Fingertip Interaction Using Surface Acoustic Wave Sensing with a VPU Ring. In *Companion Proceedings of the 27th International Conference on Distributed Computing and Networking (ICDCN Companion '26)*. Association for Computing Machinery, New York, NY, USA, 12–13. https://doi.org/10.1145/3737611.3776956

[16] Christian Holz, Tovi Grossman, George Fitzmaurice, and Anne Agur. 2012. Implanted user interfaces. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 503–512.

[17] Yuya Igarashi, Kyosuke Futami, and Kazuya Murao. 2022. Silent Speech Eyewear Interface: Silent Speech Recognition Method using Eyewear with Infrared Distance Sensors. In *Proceedings of the 2022 ACM International Symposium on Wearable Computers* (Cambridge, United Kingdom) *(ISWC '22)*. Association for Computing Machinery, New York, NY, USA, 33–38. https://doi.org/10.1145/3544794.3558458

[18] Yasha Iravantchi, Yi Zhao, Kenrick Kin, and Alanson P. Sample. 2023. SAWSense: Using Surface Acoustic Waves for Surface-bound Event Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 422, 18 pages. https://doi.org/10.1145/3544548.3580991

[19] Yijing Jiang, Junzhe Jin, Yunhui Song, Haiyang Xu, and Michael Wessely. 2025. Demonstration of BIOral: Fabricating Intraoral pH Sensor for Continuous Health Monitoring. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 721, 5 pages. https://doi.org/10.1145/3706599.3721178

[20] Yijing Jiang, Julia Kleinau, Till Max Eckroth, Eve Hoggan, Stefanie Mueller, and Michael Wessely. 2024. MouthIO: Fabricating Customizable Oral User Interfaces with Integrated Sensing and Actuation. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 1, 16 pages. https://doi.org/10.1145/3654777.3676443

[21] Yijing Jiang, Yunhui Song, Haiyang Xu, Maria Segui, Himani Deshpande, and Michael Wessely. 2026. TeethMark: Enabling Tongue–Tooth Interaction with Capacitive Sensing and Electrotactile Feedback using an Intraoral Wearable. In *Proceedings of the 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26)* (Detroit, MI, USA). Association for Computing Machinery, New York, NY, USA, 14 pages. https://doi.org/10.1145/3830398.3830646

[22] Yijing Jiang, Haiyang Xu, Yunhui Song, Black Sun, Junzhe Jin, and Michael Wessely. 2026. BIOral CapCell: An Intraoral Solution for Drug Release and Monitoring via Colorimetric Sensing. In *Proceedings of the 39th Annual ACM Symposium on User Interface Software and Technology Adjunct (UIST Adjunct '26)*. ACM, New York, NY, USA. https://doi.org/10.1145/3830397.3841851

[23] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: "Hearing" Your Silent Speech Commands In Ear. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 57 (July 2022), 28 pages. https://doi.org/10.1145/3534613

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

[25] Daehwa Kim, Eric Whitmire, Roger Boldu, Wolf Kienzle, and Hrvoje Benko. 2024. SoundScroll: Robust Finger Slide Detection Using Friction Sound and Wrist-Worn Microphones. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers* (Melbourne VIC, Australia) *(ISWC '24)*. Association for Computing Machinery, New York, NY, USA, 63–70. https://doi.org/10.1145/3675095.3676614

[26] Naoki Kimura, Tan Gemicioglu, Jonathan Womack, Richard Li, Yuhui Zhao, Abdelkareem Bedri, Zixiong Su, Alex Olwal, Jun Rekimoto, and Thad Starner. 2022. SilentSpeller: Towards mobile, hands-free, silent speech text entry using electropalatography. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 288, 19 pages. https://doi.org/10.1145/3491102.3502015

[27] Kaylee Yaxuan Li, Yasha Iravantchi, Yichen Zhu, Hyunmin Park, and Alanson P. Sample. 2025. HandSAW: Wearable Hand-based Event Recognition via On-Body Surface Acoustic Waves. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 1, Article 10 (March 2025), 29 pages. https://doi.org/10.1145/3712276

[28] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7

[29] Thorbjørn Mikkelsen, Yijing Jiang, Baris F. Coskun, Michael Wessely, and Ken Pfeuffer. 2025. Gaze+Mouth: Gaze-Directed Tongue Input with Intraoral Device for Selection in XR. In *2025 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 54–58. https://doi.org/10.1109/ISMAR-Adjunct68609.2025.00017

[30] Dinmukhammed Mukashev, Nimesha Ranasinghe, and Aditya Shekhar Nittala. 2023. TactTongue: Prototyping ElectroTactile Stimulations on the Tongue. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 105, 14 pages. https://doi.org/10.1145/3586183.3606829

[31] Laxmi Pandey and Ahmed Sabbir Arif. 2024. MELDER: The Design and Evaluation of a Real-time Silent Speech Recognizer for Mobile Devices. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 320, 23 pages. https://doi.org/10.1145/3613904.3642348

[32] J. R. Pierce. 1969. Whither Speech Recognition? *The Journal of the Acoustical Society of America* 46, 4B (10 1969), 1049–1051. https://doi.org/10.1121/1.1911801

[33] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 1182, 27 pages.

[34] Jun Rekimoto and Yu Nishimura. 2021. Derma: silent speech interaction using transcutaneous motion sensing. In *Proceedings of the Augmented Humans International Conference 2021*. 91–100.

[35] Sonion A/S. 2026. Sensors Product Portfolio. https://www.sonion.com/products/sensors/.

[36] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. Muteit: Jaw motion based unvoiced command recognition using earable. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous*

Technologies 6, 3 (2022), 1–26.
[37] 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 the 22nd ACM Conference on Embedded Networked Sensor Systems* (Hangzhou, China) *(SenSys '24)*. Association for Computing Machinery, New York, NY, USA, 784–798. https://doi.org/10.1145/3666025.3699374
[38] Tanmay Srivastava, R. Michael Winters, Thomas Gable, Yu Te Wang, Teresa LaScala, and Ivan J. Tashev. 2024. Whispering Wearables: Multimodal Approach to Silent Speech Recognition with Head-Worn Devices. In *Proceedings of the 26th International Conference on Multimodal Interaction* (San Jose, Costa Rica) *(ICMI '24)*. Association for Computing Machinery, New York, NY, USA, 214–223. https://doi.org/10.1145/3678957.3685720
[39] STMicroelectronics. 2019. *MP34DT05 Digital MEMS Microphone Datasheet.* STMicroelectronics. https://www.st.com/resource/en/datasheet/mp34dt05.pdf PDM output MEMS microphone.
[40] Lotte NS Andreasen Struijk. 2006. An inductive tongue computer interface for control of computers and assistive devices. *IEEE Transactions on biomedical Engineering* 53, 12 (2006), 2594–2597.
[41] Zixiong Su, Shitao Fang, and Jun Rekimoto. 2023. LipLearner: Customizable Silent Speech Interactions on Mobile Devices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 696, 21 pages. https://doi.org/10.1145/3544548.3581465
[42] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. 2018. Lip-Interact: Improving Mobile Device Interaction with Silent Speech Commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 581–593. https://doi.org/10.1145/3242587.3242599
[43] Rujia Sun, Xiaohe Zhou, Benjamin Steeper, Ruidong Zhang, Sicheng Yin, Ke Li, Shengzhang Wu, Sam Tilsen, Francois Guimbretiere, and Cheng Zhang. 2023. EchoNose: Sensing Mouth, Breathing and Tongue Gestures inside Oral Cavity using a Non-contact Nose Interface. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers* (Cancun, Quintana Roo, Mexico) *(ISWC '23)*. Association for Computing Machinery, New York, NY, USA, 22–26. https://doi.org/10.1145/3594738.3611358
[44] Texas Instruments. 2023. *BQ25180 Ultra-Low IQ, Highly Integrated Battery Charge Management IC Datasheet.* https://www.ti.com/lit/ds/symlink/bq25180.pdf Single-cell Li-Ion/Li-Poly battery charger and power path management IC.
[45] Texas Instruments. 2023. *TPS718xx 150-mA, Low-IQ, Low-Dropout Regulator Datasheet.* Texas Instruments. https://www.ti.com/lit/ds/symlink/tps718.pdf Includes TPS71818-33 fixed 3.3V output variant.
[46] u-blox AG. 2021. *NINA-B3 Series Stand-Alone Bluetooth 5 Low Energy Modules Data Sheet.* u-blox. https://content.u-blox.com/sites/default/files/NINA-B3_DataSheet_UBX-17052099.pdf Includes NINA-B306 module.
[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.
[48] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I. Hong, Carmel Majidi, and Swarun Kumar. 2020. RFID Tattoo: A Wireless Platform for Speech Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 155 (Sept. 2020), 24 pages. https://doi.org/10.1145/3369812
[49] Xue Wang, Zixiong Su, Jun Rekimoto, and Yang Zhang. 2024. Watch Your Mouth: Silent Speech Recognition with Depth Sensing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 323, 15 pages. https://doi.org/10.1145/3613904.3642092
[50] Wikipedia contributors. 2026. Adaptive Differential Pulse-Code Modulation. https://en.wikipedia.org/wiki/Adaptive_differential_pulse-code_modulation.
[51] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2021. SoundLip: Enabling word and sentence-level lip interaction for smart devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–28.
[52] Ruidong Zhang, Ke Li, Yihong Hao, Yufan Wang, Zhengnan Lai, François Guimbretière, and Cheng Zhang. 2023. EchoSpeech: continuous silent speech recognition on minimally-obtrusive eyewear powered by acoustic sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.
[53] Hua Zhong, Fang Qin, and Chunzhi Wu. 2025. Research Status on Performance Degradation of Automatic Speech Recognition Models. In *Proceedings of the 2025 8th International Conference on Computer Information Science and Artificial Intelligence (CISAI '25)*. Association for Computing Machinery, New York, NY, USA, 1209–1214. https://doi.org/10.1145/3773365.3773556

A Appendix

A.1 Command Library

We selected a 50-sentence corpus for silent speech interaction as shown in Figure 8.



Figure 8: Command set used in *InvisiSpeech*. Pink color highlights the five commands used for calibration.

This command set covers seven topics, including Calls and Texts, Everyday tasks, Navigation and Maps, Music and Podcasts, Knowledge and Answers, and Smart Home. This command set is intended to contextualize a scenario where people interact with *InvisiSpeech* to control smart home devices or access information, navigation, and entertainment services. These phrases are partially selected from Apple's official guide to Siri [3]. We include both concise commands and casual expressions, covering diverse phoneme distributions and varying utterance lengths (ranging from 5 to 23 phonemes, with an average length of 12.74 ± 4.14 phonemes). We convert the text into phoneme sequences using the CMU Pronouncing Dictionary [7], which are then used for audio-based analysis and modeling with our intraoral VPU sensing system. Therefore, this corpus is also phonetically well-balanced and suitable for testing the model's capability.

Since recognition is performed over a predefined command set, sequence-alignment methods such as CTC-based alignment or Dynamic Time Warping (DTW) could also be applied. However, DTW scales linearly with the size of the command library and requires pairwise sequence matching during inference, resulting in higher computational cost. Instead, we formulate recognition as constrained decoding with Whisper followed by command matching, which is more computationally efficient while retaining the flexibility of a pretrained ASR model.

A.2 Price Breakdown Table

This table provides a detailed breakdown of the per-unit manufacturing cost of the *InvisiSpeech* device under both retail and wholesale pricing. It highlights the primary cost contributors and shows that the total cost can be reduced to approximately $13.4 at scale.

Table 4: Manufacturing cost breakdown of our *InvisiSpeech* prototype. All prices are per unit and in USD.

Component	Retail	Wholesale
MCU	7.30	5.96
VPU Sensor	3.45	2.29
Power Regulator ICs	1.98	0.97
Power Management ICs	2.69	1.37
Magnetic Connector	0.17	0.17
Latching Switch Button	0.88	0.53
3D-printed Brace	1.31	1.31
Coin Cell Battery	1.20	0.80
Total	**18.98**	**≈13.40**