PPGSpeech: A Wearable Silent Speech Interface Leveraging Neck-worn Photoplethysmography

Lingde Hu * *Student Member, IEEE*; Wenbo Zhang * *Student Member, IEEE*; Wenkang Zhang *Student Member, IEEE*; Yu He *Student Member, IEEE*; Seokmin Choi; Yang Gao *Member, IEEE*; Jagmohan Chauhan *Member, IEEE*; Zhanpeng Jin † *Senior Member, IEEE*;

***Abstract*—Silent speech interfaces (SSIs) promise private and noise-immune communication, but current solutions often sacrifice user comfort, mobility, or privacy. This paper introduces PPGSpeech, a novel SSI that overcomes these limitations by pioneering the use of photoplethysmography (PPG) acquired from a comfortable, necklace-style wearable device. Our core discovery is that subtle neck muscle movements during silent articulation induce distinct, measurable modulations in the underlying PPG signal. To harness this phenomenon, we developed a complete end-to-end system featuring (1) a custom neck-worn sensor for multi-wavelength PPG acquisition, (2) a deep learning pipeline that converts 1D PPG signals into 2D time-frequency images via Continuous Wavelet Transform (CWT) and classifies them using a lightweight CNN, and (3) a Pix2Pix GAN model to reconstruct audible speech from the captured signals. In a 16-participant study covering a vocabulary of 15 commands and four confounding actions, our user-dependent model achieved a recognition accuracy of 81.41% ± 9.74. Furthermore, our speech reconstruction achieved a Mean Opinion Score (MOS) of 3.48 and a Word Correct Rate (WCR) of 60.67%, demonstrating that the PPG signal is sufficiently rich to recover intelligible speech. By establishing the viability of neck-based PPG for silent speech, PPGSpeech offers a discreet, privacy-preserving, and continuously wearable paradigm for next-generation human-computer interaction.**

***Index Terms*—PPG, Wearable, Neck-worn Sensor, Silent Speech Recognition.**

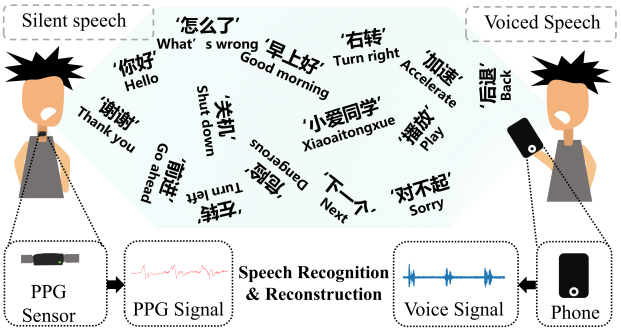


Fig. 1. The system framework of PPGSpeech.

I. Introduction

SILENT Speech Interfaces enable users to communicate without vocalizing, offering significant advantages in privacy, noise immunity, and accessibility. These benefits are particularly valuable in public environments, assistive technologies, and scenarios demanding discreet interaction. Conventional SSIs typically rely on either contactless or contact-based sensing modalities.

Contactless approaches, including acoustic sensing [1], millimeter-wave radar [2], depth cameras [3], and RGB cameras [4], infer speech from visual or physical movements. However, these systems often require bulky external devices and are susceptible to privacy concerns and environmental disturbances such as lighting or background noise. For example, Wang et al. [3] achieved high-precision recognition using depth-based point clouds, but users must face the sensor directly, which limits mobility and comfort.

Contact-based methods, such as those using electromyography (EMG) [5], inertial sensors (IMU) [6], or EEG [7], capture muscle or neural activity during silent articulation. While accurate, these solutions often require expensive equipment, high skin contact, or motion sensitivity, limiting their long-term usability. For instance, Chen et al. [8] used a dense EMG array covering the face and neck, compromising comfort and wearability.

Inspired by the common practice of wearing accessories such as necklaces or scarves, we propose a lightweight, neck-worn solution that leverages PPG signals. PPG is an optical sensing technique that detects changes in blood volume and has been widely used for monitoring heart rate

*Lingde Hu and Wenbo Zhang contributed equally to this work.
†Corresponding author: Zhanpeng Jin.

Received 24 June 2025; revised 22 October 2025; accepted 19 November 2025. This work was supported in part by the Guangdong Provincial Department of Science and Technology under Grant 2023CX10X070; the National Natural Science Foundation of China (NSFC) under Grant 62302168; the Guangdong Provincial Key Laboratory of Human Digital Twin under Grant 2022B1212010004; the Guangzhou Basic Research Program under Grant SL2023A04J00930; and the Shenzhen Holdfound Foundation Endowed Professorship.

Lingde Hu, Wenbo Zhang, Wenkang Zhang, Yu He, Yang Gao, Zhanpeng Jin are with the School of Future Technology, South China University of Technology, Guangzhou, 511442, China (Emails: 202320163218@mail.scut.edu.cn, ftzhangwenbo@mail.scut.edu.cn, ftclound@mail.scut.edu.cn, 202320163202@mail.scut.edu.cn, gaoyang2025@scut.edu.cn, zjin@scut.edu.cn).

Seokmin Choi is with the Samsung Research America, Mountain View, CA, 94043, USA. (Email: smctopp@gmail.com) (This work was done while Seokmin Choi was a graduate student at the University at Buffalo, Buffalo, NY, 14260, US.)

Jagmohan Chauhan is with the Department of Computer Science, University College London, London, WC1E 6BT, UK. (Email: jagmohan.chauhan@ucl.ac.uk)

and respiration. Compared to EMG and IMU, PPG is a low-cost, more comfortable option for long-term wear, and is less affected by transient body motion. Recent studies demonstrate that PPG can also reflect muscle activity, enabling facial expression recognition [9], gesture detection [10], and strength estimation [11]. Building on these insights, we propose **PPGSpeech**, the first system that leverages neck PPG for silent speech interaction. As illustrated in Fig. 1, PPGSpeech captures multi-wavelength PPG signals from the suprasternal notch to sense subtle hemodynamic changes induced by articulation. The system performs two tasks: (i) recognizing silent speech commands, and (ii) reconstructing audible speech from the PPG signal.

Our contributions are summarized as follows:

1) We demonstrate that neck-worn PPG alone enables both silent-speech command recognition and audible speech reconstruction; it reveals phrase-specific hemodynamic modulation at the suprasternal notch.
2) End-to-end PPG-to-speech pipeline: CWT converts 1-D PPG to 2-D time–frequency images; lightweight CNN encoder-decoder achieves 81.41 % user-dependent accuracy; Pix2Pix GAN first recovers high-frequency acoustic features from low-frequency physiological signals.
3) Collar form-factor eliminates facial privacy risks, works under mask occlusion, and shifts the paradigm from facial EMG, cameras, or IMUs to a privacy-preserving, mask-compatible collar.

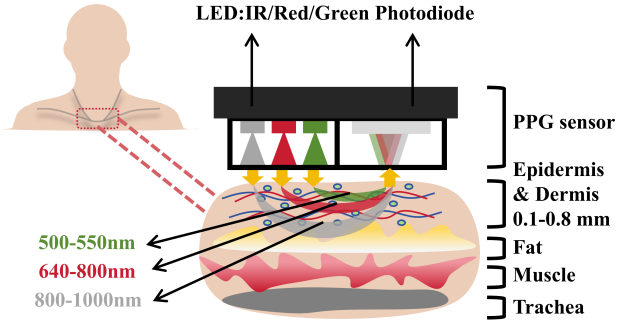


Fig. 2. Working principle of PPG sensing.

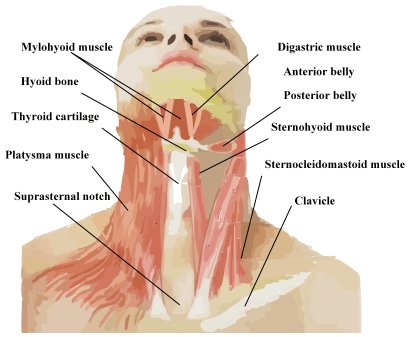


Fig. 3. Major muscles of the neck.

II. Related Work

A. PPG Sensing in HCI

PPG is a non-invasive optical sensing technology that detects changes in blood volume in human tissue, primarily used in health monitoring and biomedical applications. Traditional studies have focused on extracting cardiovascular metrics. For instance, Xiao et al. [12] proposed a multi-task learning model (MDLG-MTLNet) for cuffless blood pressure estimation, and Alessio et al. [13] achieved energy-efficient heart rate monitoring through neural architecture search and model quantization.

While motion artifacts were traditionally considered noise, recent studies have begun to explore their informative potential. Li et al. [10] used a wristband with tri-wavelength PPG sensors to classify four pinch gestures and three force levels. Choi et al. [14] introduced PPGface, an ear-worn device for facial expression recognition, achieving 93.5% accuracy across seven expressions.

B. Neck-worn Device Interaction Application

Despite the popularity of wristbands and earbuds, neck-worn devices remain underutilized. Yet, the neck provides rich physiological signals, such as the carotid pulse and temperature. Huang et al. [15] used a skin-mounted accelerometer at the suprasternal notch to classify cervical spine movements. Zhang et al. [16] identified swallowing activities via inertial and PPG sensors. Grasso et al. [17] reduced motion artifacts in neck-based heart monitoring. Zhang et al. [18] enhanced eating detection via multimodal sensing integrated into a neckband.

C. Silent Speech Interfaces

SSIs capture inaudible articulatory activity to reconstruct or recognize speech. SSIs can be categorized into contactless and contact-based approaches.

Contactless methods rely on visual or wireless sensing. Jin et al. [19] used ear-canal ultrasound to recognize 32 commands with 89.98% accuracy. Gao et al. [1] captured micro-Doppler shifts via ultrasound for 45-word recognition. Wang et al. [3] used depth point clouds with PointVSR for high-accuracy recognition. Zeng et al. [2] applied mmWave radar for SSI. In contrast, Kimura et al. [20] and Zhang et al. [21] used wearable cameras and infrared (IR) imaging, respectively, for silent speech decoding. EarSSR [22] repurposes off-the-shelf headphones to emit 16–22 kHz ultrasound and decode sub-millimeter ear-canal deformations for silent letter/word recognition. EchoSpeech [23] embeds a miniature ultrasonic array on glasses, capturing lip/facial skin motion via active acoustics for continuous silent ASR. Garashi et al. [24] mount IR distance sensors on glasses and ear hooks for silent interaction during wear. SilentMask [25] affixes dual 6-axis IMUs inside a disposable mask to record 12-D mouth-region motion for 21 Alexa commands and mute detection.

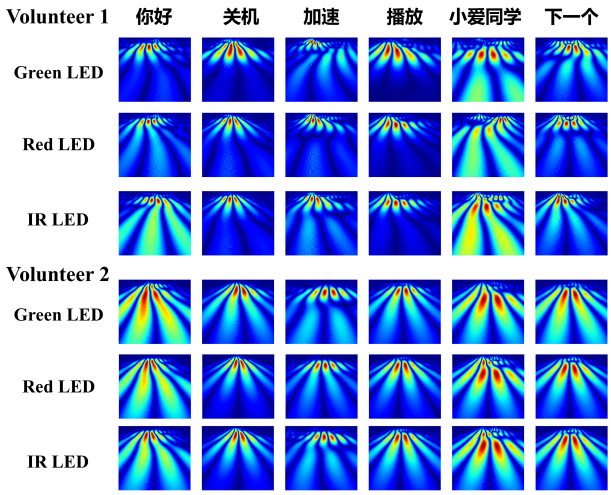


Fig. 4. The continuous wavelet transform of PPG samples collected from different volunteers, LEDs, and Chinese phrases.

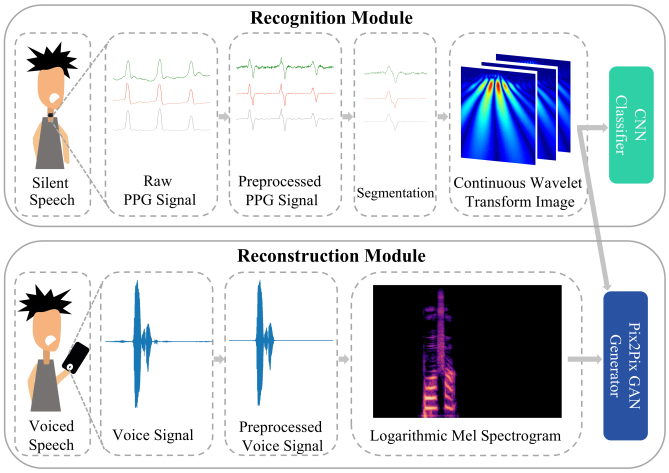


Fig. 5. The overview of PPGSpeech.

Contact-based SSIs rely on body-mounted sensors: Chen et al. [8] used a high-density EMG array on the face/neck; Rekimoto et al. [6] placed 6-axis IMUs under the chin and on the neck to achieve 98.28 % accuracy on 35 commands; Brumberg et al. [26] investigated BCI for silent speech; Kimura et al. [27] developed an EPG-based text-entry system. JawSense [28] employs a low-cost three-axis accelerometer placed on the temporomandibular joint to capture jaw micro-vibrations during silent articulation while MuteIt [29] mounts dual IMUs on the ear to track mandibular micro-movements relative to the temporal bone and reconstruct the signals into complete words via linguistic rules and QuietSync [30] integrates an IMU with novel dry ExG electrodes into off-the-shelf headsets such as headphones glasses and VR headsets to attain 94 % accuracy across 12 commands and Tang [31] presents a necklace-type silent-speech system based on ordered cracked-graphene textile strain sensors that achieves 95.25 % recognition on 20 high-frequency words. Tang et al. [32] further embed four textile-based dry EMG electrodes into standard headphone earmuffs and achieve 96% accuracy on ten control words.

III. Background and Feasibility Study

A. Speech Production Overview

Human speech production involves the coordinated action of the respiratory system, vocal cords, and articulatory organs. Airflow from the lungs causes the vocal cords to vibrate, generating a sound that resonates through the oral and nasal cavities. While this vocal cord vibration defines voiced speech, silent speech depends solely on the movements of articulators such as the lips and tongue.

B. Neck Muscle and Vascular Dynamics

As illustrated in Fig. 3, speech production requires the coordinated activation of intrinsic (e.g., cricothyroid) and extrinsic (e.g., sternohyoid, digastric) laryngeal muscles. These movements alter the shape and tension of the neck, which, in turn, affect the local vasculature, including the carotid arteries and jugular veins. The resulting muscle contractions exert mechanical pressure on these blood vessels, modulating local blood flow patterns in a manner unique to each speech act.

C. PPG Sensing on Neck

PPG sensors emit light absorbed by hemoglobin and measure the reflected light to derive cardiovascular metrics. Light penetration depth varies by wavelength—green light reaches superficial layers, while red and IR light penetrate deeper. Fig. 2 illustrates the working principle.

Compared to finger and wrist sites, the neck provides stable placement with fewer motion artifacts and stronger respiratory components [33]. Neck PPG supports SpO_2 estimation [34] and jugular pulse extraction [35].

During silent speech, articulatory muscle movements deform surrounding tissues and blood vessels, altering optical paths and PPG signals. We hypothesize that such physiological changes produce phrase-specific signatures in neck PPG, enabling silent speech recognition.

IV. System Design

A. Feasibility Analysis

To assess the feasibility of neck-worn PPG for silent speech recognition, we conducted a pilot study with two volunteers. Participants wore a PPG sensor with green, red, and IR LEDs placed at the suprasternal notch and silently articulated six Chinese phrases. The PPG signals were converted into 2D time–frequency representations using CWT, as shown in Fig. 4.

Our analysis of the CWT patterns revealed three key findings. First, different phrases produced distinct patterns, indicating that phrase-specific muscle activation modulate the PPG signal. Second, signals captured at different LED wavelengths exhibited unique characteristics, reflecting their varying tissue penetration depths

and sensitivities to different vascular responses. Third, we observed inter-user variability, where patterns for the exact phrase differed between individuals, likely due to anatomical and pronunciation differences.

Collectively, these findings confirm that neck PPG signals contain rich, speech-specific patterns, thereby validating the core hypothesis of our work. They also emphasize the need for a user-dependent modeling approach to accommodate individual variations.

B. System Overview

As shown in Fig. 5, PPGSpeech consists of two modules: silent speech recognition and speech reconstruction.

For recognition, raw PPG signals are processed through a multi-stage filtering pipeline and segmented into phrase-level samples. Each segment is converted into a two-dimensional CWT image for the extraction of spatial-temporal features. These images are fed into a custom CNN-based encoder-decoder architecture, which outputs the predicted phrase label.

The reconstruction module synthesizes audible speech from silent input. Ground-truth voiced audio is denoised and converted to a mel-spectrogram. The system utilized a U-Net-based image-to-image model (Pix2Pix GAN) to generate a mel-spectrogram directly from the PPG-derived CWT image. Finally, the Griffin-Lim algorithm reconstructs the waveform from the estimated mel-spectrogram.

C. Signal Processing

We implement a multi-stage pipeline to extract informative features from raw PPG and audio signals.

1) Differential Filtering: To eliminate low-frequency baseline drift induced by respiration, temperature, or EM interference, we apply differential filtering:

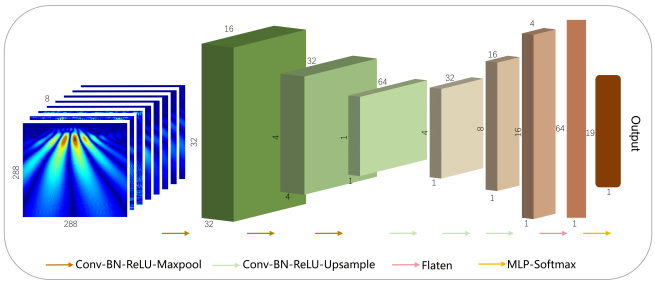
$$S'_j(t) = S_j(t) - S_j(t - \Delta t) \tag{1}$$

where $S_j(t)$ represents the raw PPG measurements at time t, Δt is a short time interval representing the time difference of the filter, and S'_j is the filtered signal of the j-th channel. The Δt in this experiment is 10 ms.

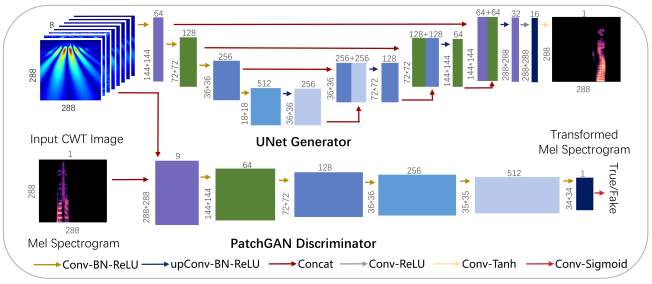
2) Gaussian Filtering: To further remove signal noise, we use a Gaussian filter to smooth the PPG signal. Gaussian filtering suppresses high-frequency noise while preserving meaningful low-frequency signal components. The filtering operation is defined as follows:

$$S''_j(t) = G(S'_j(t), \sigma) \tag{2}$$

where $S'_j(t)$ is the differentially filtered PPG signal at time t, $G(\cdot, \sigma)$ is a Gaussian smoothing function with standard deviation σ, and $S''_j(t)$ is the smoothed signal of the j-th channel. The standard deviation is set to $\sigma = 1$ to balance noise suppression and signal preservation.



(a) Recognition model.



(b) Reconstruction model.

Fig. 6. Recognition model and reconstruction model architecture.

3) Segmentation: To isolate the relevant signal portions, an energy-based thresholding method is applied to the filtered signal $S''_j(t)$ to detect periods of activity corresponding to silent speech. The detected activity is then segmented into fixed-length windows of 288 samples (2.88 seconds), ensuring consistent input dimensions for the subsequent feature extraction and accommodating the duration of all phrases in our command set.

4) Continuous Wavelet Transform: In silent-speech recognition, PPG waveforms exhibit subtle haemodynamic fluctuations—from ultra-low cardiovascular rhythms (<0.5 Hz) to brief articulatory bursts (>2 Hz)—that demand simultaneous resolution of slow trends and fast transients. The CWT meets this need through scale-dependent windows: wide at low frequencies for precise spectral estimation and narrow at high frequencies for sharp temporal localisation. Hence, CWT is expected to surpass STFT in capturing the nuanced PPG signatures of silent speech [36].

To capture the rich time-frequency characteristics of the PPG signal for image-based classification, we transform each 1D signal segment $S''_j(t)$ into a 2D representation using CWT. We employ a Morlet wavelet and compute the transform over 288 scales, yielding a (288, 288) time-frequency image. The transformation is defined as:

$$W_s(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} S''_j(t) \psi^* \left(\frac{t - b}{a} \right) dt \tag{3}$$

where $W_s(a, b)$ denotes the wavelet coefficients at scale a and translation b, and $\psi(t)$ is the Morlet wavelet.

5) Spectral Subtraction: To prepare the ground-truth audio for the reconstruction task, we use spectral subtraction to remove background noise from the original recordings. The process involves transforming the signal

(a) PPG device (b) Sports straps (c) PPGSpeech prototype

Fig. 7. PPG device, sports straps, and PPGSpeech prototype.

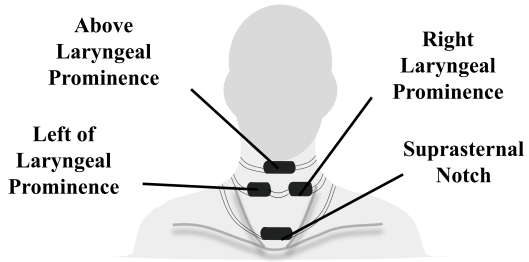


Fig. 8. Four potential neck PPG collection locations

and estimated noise to the frequency domain, subtracting the noise spectrum, and transforming the result back to the time domain:

$$Y'(t) = \mathcal{F}^{-1}\left(\mathcal{F}\{Y(t)\} - \mathcal{F}\{n(t)\}\right) \tag{4}$$

where $Y(t)$ is the original audio waveform, $n(t)$ is the estimated noise, $\mathcal{F}$ is the Fourier Transform, and $\mathcal{F}^{-1}$ is its inverse.

6) Mel spectrum transform: We convert the clean audio signal $Y'(t)$ into a log-Mel spectrogram representation through the following four steps:

(1) Short-Time Fourier Transform (STFT): The signal is windowed and transformed to the frequency domain. We use a Hanning window with an 'nfft' of 2048 and a hop length of 512.

$$X(t,f) = \mathrm{STFT}(Y'(t)) = \int_{-\infty}^{\infty} Y'(t)w(t-\tau)e^{-j2\pi ft}dt \tag{5}$$

(2) Compute Magnitude: The power of the signal at each frequency bin is calculated by taking the magnitude of the complex STFT result.

$$M(t,f) = |X(t,f)| \tag{6}$$

(3) Apply Mel-Scale Mapping: The frequency axis is converted to the Mel scale, which better reflects human auditory perception, using a Mel filter bank.

$$f_{\mathrm{mel}} = 2595\log_{10}\left(1 + \frac{f}{700}\right) \tag{7}$$

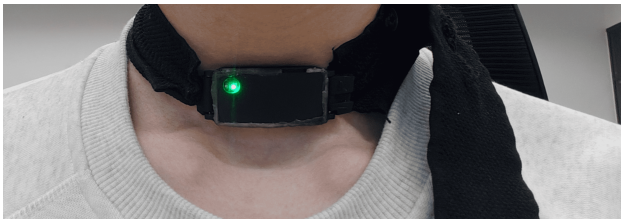


Fig. 9. Experimental setup of the PPGSpeech prototype worn by the user.

(4) Logarithmic Compression: Finally, the amplitude is log-compressed to balance the dynamic range and create the final feature representation.

$$S_{\mathrm{mel}}(t,m) = \log(M(t,f) + \epsilon) \tag{8}$$

where ϵ is a small constant to avoid numerical instability.

D. Silent Speech Recognition Module

To classify silent speech from multi-channel PPG signals, we design a CNN-based encoder-decoder network, illustrated in Fig. 6a. The input consists of eight CWT images with shape (8, 288, 288), representing multi-wavelength and multi-channel features.

The encoder comprises three convolutional blocks, each containing a convolutional layer, batch normalization, ReLU activation, and max pooling. These layers compress spatial features while expanding channel capacity, reducing the feature map to (64, 1, 1). The decoder mirrors this process with upsampling layers to recover temporal resolution, resulting in an output shape of (4, 16, 1). The final fully connected layer flattens the output and maps it to softmax probabilities over silent speech classes.

This architecture captures both spatial and spectral patterns embedded in the CWT image, enabling robust classification across users and phrases.

E. Silent Speech Reconstruction Module

To reconstruct audible speech from PPG, we employ the Pix2Pix framework [37], a conditional Generative Adversarial Network (cGAN) designed for image-to-image translation tasks (see Fig. 6b).

Pix2Pix GAN: The Pix2Pix GAN of PPGSpeech consists of a discriminator and a generator. The CWT image of PPG is used as the input of the generator, and the generated mel spectrogram is output. At the same time, the real mel spectrogram is used as the input of the discriminator to distinguish between true and false images. During the training process, the generator and the discriminator constantly compete with each other, and the network's ability is continuously improved, allowing the generator to establish a mapping between the PPG CWT image and the audio mel spectrogram.

Generator: We employ a U-Net with a symmetric encoder–decoder and skip connections. The encoder extracts deep features from the input CWT image, while the decoder progressively upsamples and reconstructs the

TABLE I
THE COMMANDS SELECTION.

Number	Chinese	Pronunciation in Chinese Pinyin	English
1	你好	[ni'hao]	Hello
2	对不起	[dui'bu'qi]	Sorry
3	谢谢	[xie'xie]	Thank you
4	早上好	[zao'shang'hao]	Good morning
5	危险	[wei'xian]	Dangerous
6	怎么了	[zen'me'le]	What's wrong
7	前进	[qian'jin]	Go ahead
8	后退	[hou'tui]	Back
9	左转	[zuo'zhuan]	Turn left
10	右转	[you'zhuan]	Turn right
11	加速	[jia'su]	Accelerate
12	播放	[bo'fang]	Play
13	关机	[guan'ji]	Shut down
14	小爱同学	[xiao'ai'tong'xue]	Xiaoaitongxue
15	下一个	[xia'yi'ge]	Next
16	（咀嚼）	-	Chew
17	（吸鼻子）	-	Sniffing
18	（咳嗽）	-	Cough
19	（静止）	-	Keep still

corresponding mel-spectrogram. Skip connections preserve local detail, enhancing spectrogram resolution.

Discriminator: A PatchGAN discriminator assesses local regions within the generated spectrogram. It divides the spectrogram into patches and performs binary classification on each patch to encourage local realism.

Loss Functions: The Pix2Pix GAN objective combines three terms: the reconstruction loss L_1, the perceptual loss L_{PAPS}, and the adversarial loss L_{GAN}.

To ensure pixel-wise similarity between the generated spectrogram and the original spectrogram, we employ the L_1 loss, defined as:

$$L_1 = \|G(x) - P(x)\| \tag{9}$$

where $G(x)$ is the generated image and $P(x)$ the target.

To reconstruct more realistic audio features, we use the pre-trained VGG to extract feature representations and define the perceptual loss as:

$$L_{\text{PAPS}} = \mathbb{E}\left[\sum_i \|\phi_i(G(x)) - \phi_i(y)\|_1\right] \tag{10}$$

where $\phi_i(\cdot)$ represents the feature map extracted from the i-th layer of the network.

The adversarial loss follows the standard GAN loss to ensure the consistency of the generated mel-spectrogram with the real mel-spectrogram:

$$L_{\text{GAN}} = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_x[\log(1 - D(x, G(x)))] \tag{11}$$

where $D(x,y)$ is the discriminator output for the real image pair, and $D(x, G(x))$ is the discriminator output for the generated image.

The final objective function of Pix2Pix GAN is a weighted combination of the above losses:

$$L^* = \arg\min_G \max_D L_{\text{GAN}}(G,D) + \lambda_1 L_1 + \lambda_2 L_{\text{PAPS}} \tag{12}$$

where λ_1 and λ_2 are hyperparameters controlling the trade-off between different loss components.

Audio Generation. To recover the audio signal from the reconstructed mel spectrogram, we use the Griffin-Lim (GL) algorithm. The Griffin-Lim algorithm estimates the phase and synthesizes the audio through an iterative optimization method. The iterative process is as follows:

$$X_{n+1}(f,t) = M(f,t)e^{j\angle \text{STFT}(\text{ISTFT}(X_n(f,t)))} \tag{13}$$

where $X_{n+1}(f,t)$ is the frequency domain signal updated after the $n+1$ th iteration, which contains amplitude and phase information.$M(f,t)$ is the target amplitude spectrum, which usually comes from the amplitude spectrum of the input audio signal and represents the intensity information of the audio signal. $e^{j\angle \text{STFT}(\text{ISTFT}(X_n(f,t)))}$ indicates that the phase information is obtained by performing an inverse short-time Fourier transform (ISTFT) on the current frequency-domain signal $X_n(f,t)$ and then applying a short-time Fourier transform (STFT) to the result. This phase information is extracted by and is used to construct the updated frequency-domain signal.

V. EXPERIMENTAL SETUP

A. PPG Sensor

As shown in Fig. 7a, we employ the ADI MAXM86146 [38] PPG sensor module, featuring two photodiodes and four LEDs (two green, one red, and one IR). The sensor is integrated into a 3D-printed enclosure from the official evaluation kit and connects wirelessly via Bluetooth. We set the sampling rate to 100 Hz, exposure integration time to 117.3 µs, and LED wavelengths to 536 nm (green), 655 nm (red), and 940 nm (IR). The drive currents for green, red, and IR LEDs are 1.95 mA, 10.21 mA, and 10.21 mA, respectively.

B. Wearable Form Design

The neck is a sensitive and highly mobile region, making the device form factor critical for stable, comfortable PPG acquisition. We evaluated three neck-wear designs:

(1) Fitting type, commonly used in ECG sensors, offers tight skin contact but suffers from poor durability and potential detachment; *(2) Surrounding type*, used in bone-conduction headphones, offers less skin contact and suffers from signal instability during motion; *(3) Collar type*, which tightly fits the neck and provides consistent skin contact and high signal fidelity.

Based on this comparison, we adopt a collar-style form factor for PPGSpeech to ensure robust PPG capture.

For materials, we selected a commercial antiperspirant neckband [39] composed of 92.2% nylon and 7.8% spandex. It is breathable, skin-friendly, sweat-resistant, and elastic. To ensure user comfort and adaptability, we integrated an adjustable buckle and a knitted non-slip design (Fig. 7b), allowing a stable and customizable fit for users with different neck sizes.

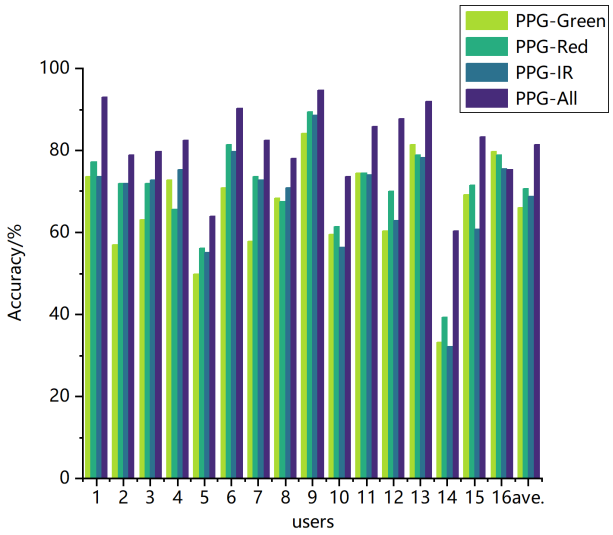


Fig. 10. Comparison of classification results across users and LEDs.

C. Collection Position Selection

As shown in Fig. 8, we evaluated several neck-worn sensor positions. The area above the Adam's apple is susceptible to motion artifacts from swallowing and neck flexion—especially in males due to its prominence—which compromises sensor stability. The left and right sides of the Adam's apple are also susceptible to lateral head rotation and contain thicker subcutaneous tissue, resulting in reduced PPG signal quality and increased noise.

In contrast, the suprasternal notch provides a stable site with thinner skin and less motion interference. It also aligns with common wearable locations such as necklaces, enhancing user comfort and social acceptability. Therefore, PPGSpeech selects the suprasternal notch as the optimal location for PPG signal acquisition.

D. Command Set Selection

We constructed a command set comprising 15 Chinese phrases, as listed in Table I, designed to cover three categories of real-world interaction:

- Phrases 1–6: for daily interpersonal communication;
- Phrases 7–11: for control of smart devices (e.g., drones, robots);
- Phrases 12–15: for interaction with consumer electronics (e.g., phones, speakers).

To evaluate robustness against non-speech activities, we additionally included three everyday actions—chewing, sniffing, and coughing—which involve neck muscle movement and may introduce interference to silent speech recognition. We also recorded baseline signals during rest to test false trigger behavior and system energy efficiency during user inactivity.

E. Data Collection

We recruited 16 healthy volunteers (eight males, eight females; age 18–30 years) in a quiet room. Each participant sat upright with feet flat on the floor and wore the PPGSpeech system at the suprasternal notch via a 3-D-printed jig, as shown in Fig. 9. A MAXM86146 PPG sensor (100 Hz, green/red/IR LEDs) streamed data to a PC over Bluetooth, while an OnePlus Ace3 phone recorded 48 kHz/16-bit audio.

For the 15 Chinese phrases and four confounding actions (chew, sniff, cough, rest), each participant performed 30 silent repetitions, following the on-screen sequence: text with pinyin (1 s), blank (0.5 s), 1 kHz beep (0.2 s), 2.8 s acquisition window. A 3-second inter-trial rest and 30-second breaks every 10 trials prevented fatigue. Immediately after the silent sessions, participants produced 30 voiced repetitions of each phrase in the same posture, speaking clearly at their natural loudness and pace. These recordings were captured with the same OnePlus Ace3 phone (48 kHz/16-bit) and served as the ground-truth audio for PPG-to-speech reconstruction.

All procedures were approved by the Institutional Review Board (IRB) of the host university.

VI. System Evaluation

A. Recognition Performance

1) Evaluation Metrics.: We evaluate silent speech recognition performance using classification accuracy. For each subject, 80% of the samples were randomly selected for training, and 20% for testing. The model was optimized using the AdamW optimizer with a learning rate of 0.001 and weight decay of 0.0001.

2) Within-user Performance.: To assess personalized performance, we conducted within-user classification on 16 participants. Each participant had a user-specific model, and accuracy was computed separately. As shown in Fig. 10, the average within-user accuracy reached 81.41% ± 9.74. Notably, 13 out of 16 users achieved an accuracy rate of 75% or higher, highlighting the system's effectiveness in capturing individual PPG patterns.

This user-dependent strategy is essential in real-world wearable HCI systems, where inter-user variability can challenge generalization. Fig. 11 presents the confusion matrix for User 9, further illustrating the system's high recognition fidelity.

3) Comparison Across Genders: Owing to pronounced anatomical differences in the neck—skin thickness, muscle composition, and the presence or absence of a prominent larynx—we conducted a focused analysis of gender-specific effects on PPG-based silent-speech classification. Female participants attained a mean accuracy of 83.32 %, outperforming their male counterparts who averaged 79.49 %. This modest yet consistent gap suggests that females benefit from smoother, less perturbed PPG signals, likely because the absence of a pronounced Adam' s apple minimizes sensor displacement and tissue-vibration artifacts. These observations highlight the need to integrate physiological diversity into wearable HCI design; future iterations could incorporate gender-aware calibration or adaptive models to maintain high recognition fidelity across all user demographics.

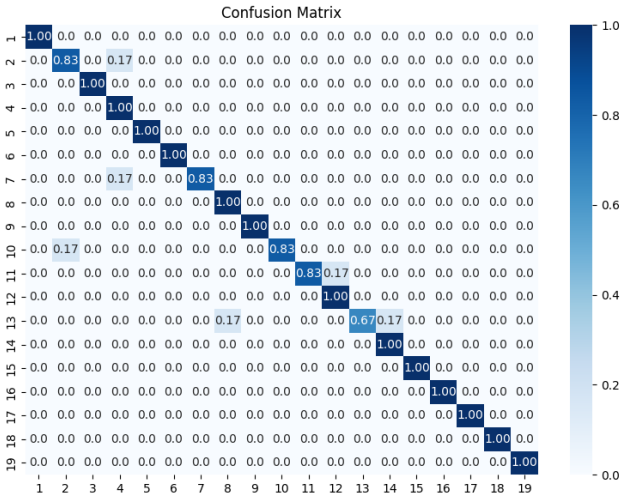


Fig. 11. Confusion Matrix of user 9.

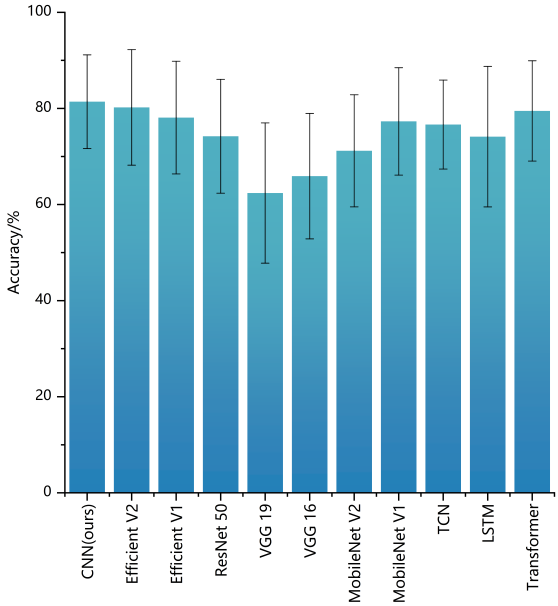


Fig. 12. Comparison of classification results of different models.

4) Comparison Across LED Wavelengths: We compared PPG signals captured from three wavelengths—Green (PPG-Green), Red (PPG-Red), and IR (PPG-IR)—as well as their fusion (PPG-All). As shown in Fig. 10, the fusion model (PPG-All) consistently outperformed all single-wavelength settings. Among individual channels, PPG-Red delivered the best performance, likely due to its deeper tissue penetration. In contrast, PPG-Green, which performs well in wrist-based applications, showed reduced accuracy in this neck-based setup, possibly due to its limited penetration depth. These findings confirm that multi-wavelength fusion provides complementary depth information and enhances robustness.

5) Comparison Across Classification Models: We compared our encoder-decoder CNN with popular deep learning models, including EfficientNet (V1/V2), ResNet-50, VGG (16/19), and MobileNet (V1/V2). As shown in Fig. 12, our tailored CNN achieved the highest accuracy (81.41%), outperforming deeper architectures. To assess the benefit of the CWT-based two-dimensional representation, we re-implemented TCN, Bi-LSTM, and a compact Transformer directly on the filtered one-dimensional PPG streams (288 time-steps, 8 channels); these achieve 76.6 %, 74.1 %, and 79.5 %, respectively, all trailing the 81.4 % of our CWT-CNN pipeline. While MobileNet variants demonstrated competitive performance and efficiency, they fell short of our design. These results demonstrate that lightweight architectures tailored to spatial–temporal features in CWT images are better suited for neck-worn PPG-based SSI tasks than generic image classifiers.

6) Comparison of 1D-to-2D Signal Transformation Methods: To validate the effectiveness of CWT-based feature extraction, we compared it with several alternatives for converting 1D signals into 2D representations: STFT, Recurrence Plot (RP), Markov Transition Field (MTF), and Gramian Angular Field (GAF).

As shown in Table II, the CWT-based approach achieved the highest recognition accuracy and lowest variance, confirming its superior ability to capture localized frequency dynamics and transient patterns. In contrast, GAF exhibited high subject-wise variance (±15.14), whereas STFT struggled with the trade-offs of fixed resolution. RP and MTF, although helpful in visualizing temporal recurrence and transitions, yielded lower accuracy and consistency. These results underscore the advantage of CWT in preserving the rich time–frequency structure of neck PPG signals.

B. Reconstruction Performance

1) Evaluation Metrics.: To evaluate the performance of audio reconstruction, we adopted five metrics: MOS (Table III), Speaker Mean Opinion Score (SMOS) (Table IV), WCR, Short-Time Objective Intelligibility (STOI), and Extended Short-Time Objective Intelligibility (ESTOI). MOS reflects the perceived audio quality, based on a 1–5 subjective rating scale, assessing the degree of restoration of the original speech. SMOS, similar to MOS, focuses on speaker similarity—that is, whether the reconstructed speech retains the original speaker's vocal characteristics. WCR quantifies linguistic intelligibility by measuring the percentage of correctly recognized words in the reconstructed speech. STOI and ESTOI provide objective intelligibility scores ranging from 0 (unintelligible) to 1 (perfectly intelligible), with STOI evaluating short segments and ESTOI extending this to longer durations. To ensure an objective and comprehensive evaluation, we recruited 10 independent raters, each of whom assessed reconstructed samples from all 10 original speakers, thereby ensuring balanced exposure to speaker variability.

2) Within-user Performance.: As shown in Table V. The reconstructed speech achieved a MOS score of 3.48, indicating that the overall audio quality was perceived as intelligible and moderately natural by human raters. The

TABLE II
COMPARISON OF 1D-TO-2D CONVERSION METHODS

Method	CWT (ours)	STFT	GAF	RP	MTF
Accuracy (%)	81.41 ± 9.74	71.18 ± 11.49	69.06 ± 15.14	62.49 ± 17.96	66.27 ± 16.93

TABLE III
MOS

Score	Level
1	Recovered none of the original speech
2	Recovered little of the original speech
3	Recovered half of the original speech
4	Recovered most of the original speech
5	Recovered all of the original speech

SMOS score of 3.21 further demonstrates that the reconstructed audio preserved specific speaker-specific characteristics, although some degradation in speaker identity was noted, likely due to the physiological limitations of PPG signals in conveying fine-grained vocal features. Objective metrics corroborate this impression: the system obtained an STOI of 0.7172 and an ESTOI of 0.6141, both well within the range typically considered acceptable for intelligible speech. In addition, the system achieved a WCR of 60.67%, reflecting a reasonable level of intelligibility from an automatic recognition perspective. This level of performance suggests that the reconstructed speech carries sufficient linguistic information to be useful for downstream tasks. Fig. 13 illustrates the spectrogram comparisons between the original and reconstructed speech for two representative phrases, "你好" (hello) and "对不起" (sorry). The reconstructed spectrograms closely resemble their original counterparts in terms of temporal structure and spectral distribution, indicating that the proposed system can effectively preserve both phonetic content and speaker-specific acoustic patterns during silent speech reconstruction.

3) Comparison Across Genders: In the speech reconstruction task, female participants outperformed male participants across all speech reconstruction metrics: MOS (3.60 vs. 3.36), SMOS (3.40 vs. 3.02), and WCR (63.4% vs. 57.91%). Objective intelligibility also showed this trend, with STOI improving from 0.687 to 0.747 and ESTOI from 0.5908 to 0.6374. This difference is attributed to the thinner neck anatomy and less pronounced Adam's apple in women, which provides cleaner input for Pix2Pix, generating mel-spectrograms that preserve finer harmonic structure and speaker characteristics.

TABLE IV
SMOS

Score	Level
1	Recovered none of the speaker characteristics
2	Recovered little of the speaker characteristics
3	Recovered half of the speaker characteristics
4	Recovered most of the speaker characteristics
5	Recovered all of the speaker characteristics

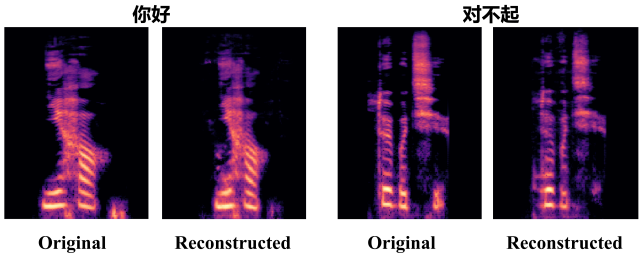


Fig. 13. Comparison of original and reconstructed Mel spectrograms.

TABLE V
PERFORMANCE OF SILENT SPEECH RECONSTRUCTION

	STOI	ESTOI	MOS	SMOS	WCR
Perfomance	0.7172	0.6141	3.48	3.21	60.67

C. User Experience Survey

To assess user acceptance of PPGSpeech, we conducted a user questionnaire focusing on participants' experiences during prototype usage. The results indicate that 93% of participants found the prototype to be lightweight and easy to wear. Furthermore, 75% expressed willingness to use the device in daily life for silent speech interaction. Several participants noted that improvements in the aesthetic design could enhance its suitability for everyday use. Regarding wearability, 80% of users found the device comfortable to wear, and 69% expressed a willingness to wear the prototype for extended periods. However, some participants reported minor discomfort associated with prolonged use. Additionally, 87% of users believed that silent speech interaction via the PPGSpeech device could bring practical benefits to their daily lives.

VII. DISCUSSIONS AND LIMITIONS

A. Discussion

1) Robust Triggering of Silent Speech.: A critical challenge for real-world silent speech interfaces is distinguishing intentional commands from other natural neck muscle activities. Our evaluation demonstrates that PPGSpeech can accurately classify confounding activities such as coughing, chewing, and sniffing, thereby preventing them from being misinterpreted as speech commands. Furthermore, the system robustly differentiates between silent speech and a resting state, which is crucial for avoiding false positives and unintended activations. This dual capability enhances both interaction reliability and user experience. It also improves power efficiency by minimizing energy consumption during user inactivity.

2) Privacy Concerns of PPG Sensing.: A significant advantage of PPGSpeech over visual or wireless sensing-based SSIs is its inherent preservation of privacy regarding facial information. Conventional methods that rely on cameras or sensors aimed at the user's mouth run the risk of capturing facial expressions, identity, or other sensitive visual data. PPGSpeech obviates the need for any facial tracking by capturing signals exclusively from the neck.

TABLE VI
COMPARISON WITH EXISTING SOLUTIONS

Attribute	Ultrasensitive textile strain sensors [31]	Whispering Wearables [40]	Chen et al. [8]	MuteIt [29]	TieLent [20]	**PPGSpeech (ours)**
Sensing Modality	Crack-textile strain	IMU+EXG	HD facial-neck EMG	Dual IMU	Single camera	**PPG**
Placement	Necklace	Ear-/head-worn	Face + neck	Behind-/in-ear	Necklace pendant	**Neck front**
Language	English	English	Mandarin	English	English	**Mandarin**
Vocabulary Size	20 high-freq. words	12 commands	33 words	100 words	15 words	**15 commands**
Accuracy	95.25%	94.2%	82.3%	94.8%	94%	**81.4%**
Speech Reconstruction	✗	✗	✗	✗	✗	**✓ (MOS 3.48)**
Motion Rejection	Only speech	Only speech	9 (e.g., swallow, cough)	Only speech	Only speech	**chew/sniff/cough**
Privacy Level	High	Low	Low	Medium	Low	**High (no visual leak)**
User Friendliness	Good (soft textile)	Fair (cabling)	Poor (64 electrodes)	Fair (ear clips)	Fair (pendant swing)	**Good (collar)**
Subjects	6	9	11	20	3	16

However, the PPG signal itself is a rich source of personal health information, containing sensitive physiological data such as heart rate and potentially blood oxygen saturation. Therefore, while our approach mitigates the risks associated with facial privacy leakage, it underscores the need for robust data governance and security measures to protect users' physiological data during storage, transmission, and processing.

3) Application.: PPGSpeech offers significant practical advantages over conventional silent speech interfaces that rely on mouth-based visual features. Its reliance on neck-based signals, rather than lip movements, ensures robust operation even when the user's mouth is occluded—such as when wearing a face mask or covering the mouth —scenarios where visual-based SSIs would fail. Furthermore, as a self-contained, neck-worn wearable, PPGSpeech enables hands-free interaction, independent of holding a terminal device, such as a smartphone. This design is particularly advantageous in situations where manual operation is complex or socially inappropriate, such as on a crowded subway or in a formal meeting, allowing for discreet interaction in complex environments. In addition to its implications for HCI, PPGSpeech contributes to the broader field of AI-enabled medicine. Because PPG is widely used for clinical monitoring, our findings suggest that articulatory-induced vascular modulation may support future applications in speech rehabilitation, voice disorder assessment, or assistive communication for patients with impaired phonation. Looking forward, the technology can be further miniaturized and integrated into everyday decorative accessories, such as necklaces. Such an implementation would maximize its social acceptability, discretion, and long-term comfort, realizing the vision of a truly seamless and unobtrusive silent speech interface.

4) Comparison with existing solutions.: While PPGSpeech is still in its early stages, it distinguishes itself from the prior art summarized in Table VI through a balanced set of practical merits. By relying on an unobtrusive photoplethysmography collar placed at the neck front, the system avoids the visual privacy risks of camera-based approaches such as TieLent [20] and the cabling or electrode burden associated with Whispering Wearables [40], Chen et al. [8], and MuteIt [29]. PPGSpeech introduces a capability: direct reconstruction of intelligible speech, validated by an MOS of 3.48. Furthermore, the modality naturally tolerates common non-speech artifacts—such as chewing, sniffing, and coughing—without requiring additional hardware. Combined with a comfortable collar form factor evaluated on sixteen Mandarin speakers, these characteristics make PPGSpeech a discreet, user-friendly, and privacy-preserving silent-speech interface.

B. Limitations and Future Work

1) Device Prototype and Form Factor.: While the current prototype, which utilizes a hand-cut sports strap with preset buckles, successfully validates our approach, its form factor presents opportunities for improvement. The current design offers limited adjustability and could be improved in terms of ergonomics and long-term comfort. Future iterations should explore advanced, flexible, or elastic fabrics to improve wearability. Furthermore, incorporating adjustable structures, such as magnetic closures or sliding mechanisms, alongside the potential use of flexible electronics, could provide a more precise, stable, and comfortable fit for a broader range of users.

2) Vocabulary and Language Expansion.: The current study utilized a command set of 15 Chinese phrases and four actions to validate the feasibility of PPGSpeech. To enhance its practical utility, future work should focus on three key areas of expansion. First, the vocabulary size should be increased to include a broader range of phrases and complete sentences. Second, the system's generalizability should be tested by collecting multilingual data, including English and other languages. Finally, to improve robustness in real-world environments, the set of confounding movements should be expanded to include more non-speech neck gestures (e.g., nodding, shaking the head) and environmental interference factors (e.g., adjusting a collar, wearing headphones).

3) Evaluation with Diverse Populations and Scenarios.: Our 16-participant study provided strong proof of concept, but a broader evaluation is needed for generalizability. Future work should recruit a larger and more diverse

cohort to account for physiological differences such as skin tone, neck anatomy, and muscle dynamics. The system should also be tested in more dynamic, ecologically valid settings (e.g., walking or exercise) to assess robustness to motion artifacts. Finally, exploring the system's adaptability across different demographic groups, including children and older adults, will be crucial for understanding its potential as a universal communication aid.

VIII. Conclusion

In this paper, we introduce PPGSpeech, pioneering a method for silent speech recognition using PPG signals from a custom neck-worn device. Our work establishes a new modality for human-computer interaction that prioritizes user comfort, privacy, and discretion. Through our carefully designed wearable system and a lightweight deep learning pipeline, we achieved a user-dependent accuracy of 81.41% in classifying 15 Chinese phrases and four confounding actions. Furthermore, we demonstrated the richness of the captured signal by successfully reconstructing intelligible voiced speech from silent articulations, yielding a MOS of 3.48 and a WCR of 60.67%. These results reveal the significant, previously untapped potential of recovering high-frequency acoustic features from low-frequency physiological signals. PPGSpeech opens new avenues for developing unobtrusive and continuously available interfaces, expanding the possibilities for wearable computing and accessible communication.

References

[1] Y. Gao, Y. Jin, J. Li, S. Choi, and Z. Jin, "EchoWhisper: Exploring an Acoustic-based Silent Speech Interface for Smartphone Users," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 3, pp. 1–27, Sep. 2020.

[2] S. Zeng, H. Wan, S. Shi, and W. Wang, "mSilent: Towards general corpus silent speech recognition using COTS mmWave radar," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 1, pp. 1–28, 2023.

[3] X. Wang, Z. Su, J. Rekimoto, and Y. Zhang, "Watch your mouth: Silent speech recognition with depth sensing," in *Proc. CHI Conf. Hum. Factors Comput. Syst. (CHI)*, 2024, pp. 1–15.

[4] L. Pandey and A. S. Arif, "Liptype: A silent speech recognizer augmented with an independent repair model," in *Proc. CHI Conf. Hum. Factors Comput. Syst. (CHI)*, 2021, pp. 1–19.

[5] H. Ikeda, T. Ohhira, and H. Hashimoto, "Classification of silent speech words considering walking using VMD-applied facial EMG," *Int. Symp. Affective Sci. Eng.*, pp. 1–4, 2023.

[6] J. Rekimoto and Y. Nishimura, "Derma: Silent speech interaction using transcutaneous motion sensing," in *Proc. Augmented Humans Int'l Conf.*, 2021, p. 91–100.

[7] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, Apr. 2019.

[8] X. Chen and X. Chen, "Silent Speech Recognition Based on High-Density Surface Electromyogram Using Hybrid Neural Networks," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 2, pp. 335–345, 2023.

[9] S. Choi, Y. Gao, Y. Jin, S. J. Kim, J. Li, W. Xu, and Z. Jin, "PPGface: Like What You Are Watching? Earphones Can "Feel" Your Facial Expressions," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 2, pp. 1–32, Jul. 2022.

[10] D. Li, P. Kang, K. Zhu, J. Li, and P. B. Shull, "Feasibility of Wearable PPG for Simultaneous Hand Gesture and Force Level Classification," *IEEE Sensors Journal*, vol. 23, no. 6, pp. 6008–6017, Mar. 2023.

[11] X. Liu, F. Li, Y. Cao, S. Zhai, S. Yang, and Y. Wang, "PPGSpotter: Personalized Free Weight Training Monitoring Using Wearable PPG Sensor," in *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications*. Vancouver, BC, Canada: IEEE, May 2024, pp. 2468–2477.

[12] H. Xiao, A. Zhao, W. Song, T. Liu, L. Long, Y. Li, and H. Li, "Advancing cuffless blood pressure estimation: A PPG-based multi-task learning model for enhanced feature extraction and fusion," *Biomed. Signal Process. Control*, vol. 95, p. 106378, 2024.

[13] A. Burrello, D. J. Pagliari, M. Risso, S. Benatti, E. Macii, L. Benini, and M. Poncino, "Q-PPG: Energy-Efficient PPG-based Heart Rate Monitoring on Wearable Devices," Mar. 2022.

[14] S. Choi, Y. Gao, Y. Jin, S. J. Kim, J. Li, W. Xu, and Z. Jin, "Excerpt of PPGface: "Like what you are watching? Earphones can feel your facial expressions"," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. (UbiComp)*, 2022, pp. 233–235.

[15] L. Huang, K. S. Chun, L. Yu, J. Y. Lee, A. Soetikno, H. Chen, H. Jeong, J. Barrett, K. Martell, Y. Kang, A. A. Patel, and S. Xu, "A Novel Method for Tracking Neck Motions Using a Skin-Conformable Wireless Accelerometer: A Pilot Study," *Digital Biomarkers*, vol. 8, no. 1, pp. 40–51, Apr. 2024.

[16] Y. Zhang, H. Zhu, H. Liu, D. Zheng, S. Zhang, and Y. Pan, "A wearable swallowing recognition system based on motion and dual photoplethysmography sensing of laryngeal movements," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2022, pp. 13–16.

[17] A. Lo Grasso, P. Zontone, R. Rinaldo, and A. Affanni, "Advanced Necklace for Real-Time PPG Monitoring in Drivers," *Sensors*, vol. 24, no. 18, p. 5908, Sep. 2024.

[18] S. Zhang, Y. Zhao, D. T. Nguyen, R. Xu, S. Sen, J. Hester, and N. Alshurafa, "NeckSense: A Multi-Sensor Necklace for Detecting Eating Activities in Free-Living Conditions," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 2, pp. 1–26, Jun. 2020.

[19] Y. Jin, Y. Gao, X. Xu, S. Choi, J. Li, F. Liu, Z. Li, and Z. Jin, "EarCommand: "Hearing" Your Silent Speech Commands In Ear," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 2, pp. 1–28, Jul. 2022.

[20] N. Kimura, K. Hayashi, and J. Rekimoto, "TieLent: A Casual Neck-Mounted Mouth Capturing Device for Silent Speech Interaction," 2020.

[21] R. Zhang, M. Chen, B. Steeper, Y. Li, Z. Yan, Y. Chen, S. Tao, T. Chen, H. Lim, and C. Zhang, "Speechin: A smart necklace for silent speech recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 4, pp. 1–23, 2021.

[22] X. Sun, J. Xiong, C. Feng, H. Li, Y. Wu, D. Fang, and X. Chen, "EarSSR: Silent speech recognition via earphones," *IEEE Trans. Mobile Computing*, vol. 23, no. 8, pp. 8493–8507, 2024.

[23] R. Zhang, K. Li, Y. Hao, Y. Wang, Z. Lai, F. Guimbretière, and C. Zhang, "Echospeech: Continuous silent speech recognition on minimally obtrusive eyewear powered by acoustic sensing," in *Proc. CHI Conf. Hum. Factors Comput. Syst. (CHI)*, 2023, pp. 1–18.

[24] Y. Igarashi, K. Futami, and K. Murao, "Silent speech eyewear interface: Silent speech recognition method using eyewear with infrared distance sensors," in *Proc. ACM Int'l Symp. Wearable Computers*, 2022, pp. 33–38.

[25] H. Hiraki and J. Rekimoto, "Silentmask: Mask-type silent speech interface with measurement of mouth movement," in *Proc. Augmented Humans Conf. (AH)*, 2021, pp. 86–90.

[26] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, and F. H. Guenther, "Brain–computer interfaces for speech communication," *Speech Communication*, vol. 52, no. 4, pp. 367–379, Apr. 2010.

[27] N. Kimura, T. Gemicioglu, J. Womack, R. Li, Y. Zhao, A. Bedri, Z. Su, A. Olwal, J. Rekimoto, and T. Starner, "Silentspeller: Towards mobile, hands-free, silent speech text entry using electropalatography," in *Proc. CHI Conf. Hum. Factors Comput. Syst. (CHI)*, 2022, pp. 1–19.

[28] P. Khanna, T. Srivastava, S. Pan, S. Jain, and P. Nguyen, "JawSense: recognizing unvoiced sound using a low-cost ear-worn system," in *Proc. 22nd Int'l Workshop on Mobile Computing Systems and Applications*, 2021, pp. 44–49.

[29] T. Srivastava, P. Khanna, S. Pan, P. Nguyen, and S. Jain, "MuteIt: Jaw Motion Based Unvoiced Command Recognition

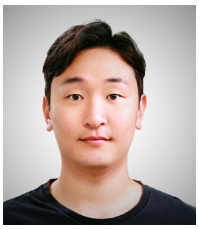
Using Earable," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 3, pp. 1–26, Sep. 2022.
[30] T. Srivastava, R. M. Winters, T. Gable, Y. T. Wang, T. LaScala, and I. J. Tashev, "Whispering wearables: Multimodal approach to silent speech recognition with head-worn devices," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, 2024, pp. 214–223.
[31] C. Tang, M. Xu, W. Yi, Z. Zhang, E. Occhipinti, C. Dong, D. Ravenscroft, S.-M. Jung, S. Lee, S. Gao *et al.*, "Ultrasensitive textile strain sensors redefine wearable silent speech interfaces with high machine learning efficiency," *npj Flex. Electron.*, vol. 8, p. 27, 2024.
[32] C. Tang, J. Mallah, D. Kazieczko, W. Yi, T. R. Kandukuri, E. Occhipinti, B. Mishra, S. Mehta, and L. G. Occhipinti, "Wireless silent speech interface using multichannel textile emg sensors integrated into headphones," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–10, 2025.
[33] I. García-López, R. X. A. Pramono, and E. Rodriguez-Villegas, "Artifacts classification and apnea events detection in neck photoplethysmography signals," *Med. Biol. Eng. Comput.*, vol. 60, no. 12, pp. 3539–3554, 2022.
[34] Y. Zhong, A. Jatav, K. Afrin, T. Shivaram, and S. T. Bukkapatnam, "Enhanced SpO2 estimation using explainable machine learning and neck photoplethysmography," *Artificial Intelligence in Medicine*, vol. 145, p. 102685, Nov. 2023.
[35] I. García-López and E. Rodriguez-Villegas, "Extracting the jugular venous pulse from anterior neck contact photoplethysmography," *Scientific reports*, vol. 10, no. 1, p. 3466, 2020.
[36] E. Mejía-Mejía, J. Allen, K. Budidha, C. El-Hajj, P. A. Kyriacou, and P. H. Charlton, "Photoplethysmography signal processing and synthesis," in *Photoplethysmography*, J. Allen and P. Kyriacou, Eds. Academic Press, 2022, pp. 69–146.
[37] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
[38] A. Devices, "MAXM86146 Datasheet and Product Info | Analog Devices," 2024. [Online]. Available: https://www.analog.com/en/products/maxm86146.html#part-details
[39] ANTA, "Sport headband," 2024. [Online]. Available: https://anta.com/goods-265790.html
[40] T. Srivastava, R. M. Winters, T. Gable, Y. T. Wang, T. LaScala, and I. J. Tashev, "Whispering wearables: Multimodal approach to silent speech recognition with head-worn devices," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, 2024, pp. 214–223.

Lingde Hu received the B.S. degree in Communication Engineering from Changsha University of Science & Technology, Changsha, China, in 2023. He is currently pursuing an M.S. degree in Information and Communication Engineering at the School of Future Technology, South China University of Technology, Guangzhou, China. His primary research interests include human–computer interaction and multimodal sensing.

Wenbo Zhang is currently pursuing the Ph.D. degree in Information and Communication Engineering at the School of Future Technology, South China University of Technology, China. His research focuses on multimodal wearable sensing and human motion representation in wearable and mobile systems. His work explores how inertial, pressure, and physiological signals can be leveraged in wearable and mobile systems to support digital health and interactive intelligent applications.

Wenkang Zhang received the B.S. degree in Electrical Engineering from Hubei Normal University in 2017 and the M.S. degree in Operations Research and Cybernetics from Hubei Normal University in 2020. He was a Technical Engineer with the Shenzhen Metro Operation Group Co., Ltd., China, from 2020 to 2023. He is currently pursuing a Ph.D. degree with the School of Future Technology, South China University of Technology, Guangzhou, China.

Yu He received the B.S. degree in Telecommunications Engineering from Sichuan University, Chengdu, China, in 2023. She is currently pursuing an M.S. degree in Information and Communication Engineering at the School of Future Technology, South China University of Technology, Guangzhou, China. Her primary research interests include human–computer interaction and signal processing.



Seokmin Choi received the Ph.D. degree in Computer Science and Engineering from the University at Buffalo, Buffalo, NY, USA, in 2024. He is now a Senior Machine Learning Research Engineer at Samsung Research America in Mountain View, CA, USA. His research interests include human–computer interaction, ubiquitous sensing, and multimodal systems.

Yang Gao received the PhD degree in Computer Science and Engineering from the University at Buffalo, United States, in 2021. He is currently an Associate Professor with the School of Future Technology, South China University of Technology, Guangzhou, China. He has authored or coauthored more than 40 technical papers in prestigious international journals and conferences, including CHI, UIST, UbiComp, *IEEE TIFS*, *IEEE JBHI*. His research interests include pervasive and mobile computing, AIoT, and human-computer-interaction. He is currently an Associate Editor of *ACM IMWUT*.

Jagmohan Chauhan is an Assistant Professor in the Department of Computer Science and leads the IntellEcT Systems group at University College London. Previously, he was a lecturer at ECS (University of Southampton) from 2021 to 2025. He received his PhD from the School of Electrical Engineering and Telecommunications, UNSW, Australia. He has co-authored more than 50 research papers and has won three best paper awards. His research interests include developing efficient and trustworthy machine learning systems, adaptive and efficient robotics, machine learning for health, and designing and evaluating novel mobile systems and applications.

Zhanpeng Jin is the Xinshi Endowed Professor and Associate Dean of the School of Future Technology at South China University of Technology, China. He also serves as the Deputy Director of both the MOE Engineering Research Center for Human Body Data Sensing and the Guangdong Key Laboratory of Digital Twin Humans. Previously, he was a tenured Associate Professor in the Department of Computer Science and Engineering at the University at Buffalo, SUNY, where he directed the Cyber-Med Lab and served as Director of Graduate Studies. He is a Senior Member of ACM and IEEE. His research focuses on ubiquitous computing, human–computer interaction, intelligent sensing, proactive health, and AI applications in healthcare, biometrics, and IoTs. He has served as a reviewer for agencies including NSF, NIST, NWO, and NSERC. He is currently an Associate Editor for *ACM Computing Surveys*, *ACM IMWUT*, *Computers in Biology and Medicine*, and *CCF Transactions on Pervasive Computing and Interaction*, and a reviewer for more than 30 international venues.